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

FABRÍCIO DO CARMO FARIAS

MAPEAMENTO PEDOLÓGICO DIGITAL E USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA EM TRACUATEUA, PARÁ

BELÉM-PA
2023

FABRÍCIO DO CARMO FARIAS

MAPEAMENTO PEDOLÓGICO DIGITAL E USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA EM TRACUATEUA, PARÁ

Dissertação apresentada à Universidade Federal Rural da Amazônia, como parte das exigências do Programa de Pós-graduação em Agronomia, visando a obtenção do título de mestre em agronomia.
Linha de pesquisa: Manejo e Conservação de Recursos Ambientais.
Orientador: Prof. Dr. João Fernandes da Silva Júnior.

BELÉM-PA

2023

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

D631m Farias, Fabrício do Carmo
Mapeamento pedológico digital e uso de algoritmos de aprendizado de máquina em Tracuateua, Pará / Fabrício do Carmo Farias. - 2023.
92 f.: il. color.

Dissertação (Mestrado) - Programa de PÓS-GRADUAÇÃO em Agronomia (PPGA), Campus Universitário de Belém, Universidade Federal Rural da Amazônia, Belém, 2023.
Orientador: Prof. Dr. João Fernandes da Silva Júnior

1. Pedometria. 2. Inteligência artificial. 3. Scorpan. 4. Manejo sustentável do solo. 5. Dados legados. I. Silva Júnior, João Fernandes da, *orient.* II. Título

CDD

FABRÍCIO DO CARMO FARIAS

MAPEAMENTO PEDOLÓGICO DIGITAL E USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA EM TRACUATEUA, PARÁ

Dissertação apresentada à Universidade Federal Rural da Amazônia, como parte das exigências do Programa de Pós-graduação em Agronomia, visando a obtenção do título de mestre em agronomia. Linha de pesquisa: Manejo e Conservação de Recursos Ambientais.

Data de aprovação: 13/06/2023

BANCA EXAMINADORA

Prof. Dr. João Fernandes da Silva Júnior
Universidade Federal Rural da Amazônia – UFRA Campus Capanema
(Orientador)

Prof. Dr. Gener Tadeu Pereira
Universidade Estadual Paulista - UNESP

Profª. Dra. Suzana Romeiro Araújo
Universidade Federal Rural da Amazônia – UFRA

Prof. Dr. Daniel Pereira Pinheiro
Universidade Federal Rural da Amazônia – UFRA Campus Capanema

Aos meus pais, Iraildes Medeiros do Carmo Farias e Ronaldo de Jesus dos Santos Farias, esposa Tainara Lino Ribeiro e irmãos Geane Medeiros, Jeziane Medeiros, Johnny Farias e Taiana Farias.

Fabrício do Carmo Farias

DEDICO.

AGRADECIMENTOS

Ao senhor nosso Deus, pai e criador, pois sem ele nada seria possível.Aos meus pais, Ronaldo Farias e Iraildes Farias que nunca deixaram de acreditar: me incentivaram, lutaram e deram apoio em todos os momentos de minha vida para alcançar meus objetivos.

Ao Prof. Dr. João Fernandes da Silva Júnior, da Universidade Federal Rural da Amazônia (UFRA), *Campus* Capanema, por sua orientação, amizade, carinho, paciência, atenção e todos os ensinamentos e contribuições indispensáveis para a realização deste trabalho.

À minha querida esposa Tainara Lino, pelo incentivo, paciência e companheirismo durante todo período da pós-graduação e por estar presente em minha vida nos momentos bons e difíceis que enfrentamos durante este projeto.

Ao Programa de Pós-Graduação de Agronomia (PGAgro), pela oportunidade de ser um dos primeiros discentes de pós-graduação do *Campus* fora de sede da UFRA Capanema e por toda estrutura disponibilizada para que eu pudesse aprender um pouco mais.

A todos aqueles que contribuíram direta ou indiretamente para a realização deste trabalho.

Muitíssimo obrigado!

“Meu filho, não te esqueças de meu ensinamento e guarda meus preceitos em teu coração porque eles aumentarão os teus dias, e te acrescentarão anos de vida e paz. Não te desamparem a benignidade e a fidelidade; ata-as ao teu pescoço, grava-as em teu coração. E acharás graça e boa compreensão diante de DEUS e dos homens.

Confia no SENHOR de todo o teu coração, e não te estribes no teu próprio entendimento. Reconhece-o em todos os teus caminhos e ele endireitará as tuas veredas. Não sejas sábio aos teus próprios olhos: teme ao SENHOR e aparta-te do mal; será isto saúde para teu corpo, e refrigério para teus ossos. Honra o SENHOR com os teus bens e com as primícias de toda tua renda; e se encherão fartamente os teus celeiros, e transbordarão de vinhos os teus lagares. Filho meu, não rejeites a disciplina do SENHOR, nem te enfades da sua repreensão. Porque o senhor repreende a quem ama, assim como o pai ao filho a quem quer bem. Feliz o homem que acha sabedoria e o homem que adquire conhecimento; porque melhor é o lucro que ela dá do que o da prata, e melhor a sua renda do que o ouro mais fino.”

(Provérbios 3:1-14).

RESUMO

A necessidade de mapeamento de solos é contínua. Porém, o método tradicional não permite uma execução rápida e econômica. Logo, o Mapeamento Digital de Solos (MDS) é capaz de melhorar esse processo ao espacializar o solo com o uso de modelos que quantificam a variabilidade espacial. Nesta perspectiva, o objetivo deste estudo foi realizar o MDS da cidade de Tracuateua, no Norte do Brasil, avaliando o desempenho de algoritmos de aprendizado de máquina, utilizando um conjunto de atributos derivados do Modelo Digital de Elevação (MDE), imagens de satélite e dados legados como parâmetros de entrada. Foram selecionados aleatoriamente 244 pixels, distribuídos em nove Unidades de Mapeamento (UM), onde foram realizadas prospecção em campo, com a finalidade de evitar zonas de transição e confirmar dados de treinamento. Os dados foram organizados e processados em *software* QGIS, que é um Sistema de Informação Geográfica (SIG) e linguagem de programação R. O MDE derivou as covariáveis preditoras e posteriormente selecionadas as significativas pelo *Recursive Feature Elimination* (RFE), função do pacote *caret*. A avaliação do desempenho dos algoritmos foi realizada por meio da matriz de confusão, do índice *Kappa* e acurácia global. Os algoritmos *Random Forest*, *Ranger* e C5.0 foram considerados moderados ao mapear os solos da área de estudo, sendo o *Ranger* com o melhor desempenho no mapeamento pedológico em Tracuateua, no nordeste paraense; tanto na etapa de modelagem com índice *Kappa* de 0,71 e acurácia global de 0,74, quanto na comparação com o mapa convencional (referência), com resultados de índice *Kappa* global 0,49 e acurácia global de 0,56. O algoritmo *Artificial Neural Networks* (ANN) apresentou os menores resultados na modelagem, com índice *Kappa* de 0,14 e acurácia de 0,26 e na comparação com o mapa de referência, tendo o índice *Kappa* global 0,35 e acurácia global de 0,48. Os algoritmos de árvore de decisão (*Ranger,* RF e C5.0) mostraram potencial moderado para o mapeamento digital de solos em escala de 1:100.000 no município de Tracuateua, no nordeste paraense.

Palavras-chave: Pedometria; Inteligência artificial; Scorpan; Manejo sustentável do solo; Dados legados.

ABSTRACT

The need for soil mapping is continuous. However, the traditional method does not allow for a quick and economical execution. Therefore, Digital Soil Mapping (DSM) is able to improve this process by spatializing the soil using models that quantify spatial variability. In this perspective, the objective of this study was to carry out the DSM of the city of Tracuateua, in the North of Brazil, evaluating the performance of machine learning algorithms, using a set of attributes derived from the Digital Elevation Model (DEM), satellite images and legacy data as input parameters. 244 pixels were randomly selected, distributed in nine Mapping Units (MU), where field prospecting was carried out, in order to avoid transition zones and confirm training data. The data were organized and processed in QGIS software, which is a Geographic Information System (SIG) and R programming language. The DEM derived the predictive covariates and subsequently selected the significant ones by the Recursive Feature Elimination (RFE), function of the caret package. The evaluation of the performance of the algorithms was performed using the confusion matrix, the Kappa index and global accuracy. The Random Forest, Ranger and C5.0 algorithms were considered moderate when mapping the soils of the study area, with Ranger having the best performance in pedological mapping in Tracuateua, in the northeast of Pará; both in the modeling stage with a Kappa index of 0.71 and global accuracy of 0.74, and in the comparison with the conventional map (reference), with results of a global Kappa index of 0.49 and global accuracy of 0.56. The Artificial Neural Networks (ANN) algorithm presented the lowest results in the modeling, with a Kappa index of 0.14 and accuracy of 0.26 and in comparison, with the reference map, with a global Kappa index of 0.35 and global accuracy of 0.48. Decision tree algorithms (Ranger, RF and C5.0) showed moderate potential for digital soil mapping at a scale of 1:100,000 in the municipality of Tracuateua, in northeastern Pará.

Keywords: Pedometry; Artificial intelligence; Scorpan; Sustainable soil management; Legacy data.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

AH-Analytical Hillshading
ANN-Artificial Neural Network
AS-Aspect
CD-Closed Depressions
CI-Convergence Index
CNBL-Channel Network Base Level
CSC-Cross Sectional Curvature
DP-Desvio Padrão
ESo-ESPODOSSOLOS FERRILÚVICOS Órticos
FA-Flow Accumulation
GC-General Curvature
GD-Gradient
GXve2-GLEISSOLOS HÁPLICOS Ta Eutróficos
GZn1-GLEISSOLOS SÁLICOS Sódicos.
LAd-LATOSSOLOS AMARELO Distróficos
LCU-Local Curvature
LGC-Longitudinal Curvature
LSF-LS Factor
MBI-Mass Balance Index
MDE-Modelo Digital de Elevação
MDS-Mapeamento digital de solo
MRRTF-Multiresolution Index of Ridge Top Flatness
MRVBF-Multiresolution Index of Valley Bottom Flatness
NB-*Naive Bayes*
PAd-ARGISSOLOS AMARELOS Distróficos
PC-Plan Curvature
PFC-Profile Curvature
RF-*Random Forest*
RFE-Recursive Feature Elimination
RQg1-NEOSSOLOS QUARTZARÊNICOS Hidromórficos
RQg5-NEOSSOLOS QUARTZARÊNICOS Hidromórficos + Neossolos Quartzarênicos Órticos + Gleissolos Háplicos Tb Distróficos
RQo1-NEOSSOLOS QUARTZARÊNICOS Órticos
RQo5-NEOSSOLOS QUARTZARÊNICOS Órticos + Argissolos Vermelho–Amarelos Distróficos
RSP-Relative Slope Position
SIG-Sistema de Informações Geográficas
SP-Slop
svmL-Support Vector Machine Linear Kernel
svmP-Support Vector Machine Polynomial Kernel
SRTM-Shuttle Radar Topographic Mission
TC-Tangential Curvature
TI-Total Insolation
TPI-Topographic Position Index
TRI-Terrain Ruggedness Index
TWI-Topographic Wetness Index
TX-Texture
UM-Unidade de Mapeamento
USC-Upslope Curvature
UTM-Universal transversa de Mercator
VDCN-Vertical Distace to Channel Network
UM-Unidade de Mapeamento
VP-Valley Depth
VRM-Vector Ruggedness Measure

SUMÁRIO

1 CONTEXTUALIZAÇÃO

O município de Tracuateua está localizado na mesorregião nordeste paraense, e compõe a microrregião bragantina com os municípios: Augusto Corrêa, Bonito, Bragança, Capanema, Igarapé-Açu, Nova Timboteua, Peixe-Boi, Primavera, Quatipuru, Santa Maria do Pará, Santarém Novo e São Francisco do Pará (LUZ, 2013; CORDEIRO; ARBAGE, SCHWARTZ, 2017).

Dentre os municípios que compõe a microrregião, é o único a possuir um mapa de classes de solos em escala de 1:100.000, fruto do trabalho desenvolvido pelo Companhia de Pesquisa de Recursos Minerais (CPRM), em parceria com a Empresa Brasileira de Pesquisa Agropecuária (Embrapa) e a Prefeitura do município, sendo uma importante ferramenta de planejamento e desenvolvimento, desde a sua produção (CPRM, 1998).

É importante ressaltar, que poucos são os trabalhos de mapeamento de classes de solos em escalas semelhantes na região norte do país, isso devido ao elevado custo e tempo de execução dos procedimentos de mapeamento, e em alguns casos, escassez de profissionais dedicados a esta área da pedologia.

Com a demanda governamental e atentos aos objetivos do Programa Nacional de Levantamento e Interpretação de Solos do Brasil (PronaSolos), o emprego de técnicas de modelagem, o uso de algoritmos de aprendizado de máquinas e os sensores modernos aplicados ao mapeamento de solos, contribuem para obtenção de mapas digitais de classes de solos, bem como diminuem o tempo e os custos para obtenção dos mesmos (POLIDORO *et al*., 2021).

O município de Tracuateua/PA está situado em uma posição estratégica dentre os demais integrantes da microrregião Bragantina e compartilha características semelhantes com municípios vizinhos, como exemplo: geologia, topografia e vegetação. Os resultados deste trabalho podem, posteriormente, ajudar no mapeamento dessas classes de solos, em municípios próximos, aplicando técnicas de área de referência, por exemplo.

Neste sentido, faz necessário o emprego das técnicas de Mapeamento Digital de Solos (MDS) propostas McBratney; Mendonça-Santos; Minasny (2003), com o intuito de identificar quais os melhores preditores de classes de solos em Tracuateua e, futuramente, possam ser utilizados para a aplicação nos demais municípios da microrregião bragantina.

1.1 Hipóteses da pesquisa

Usando algoritmos de aprendizado de máquina, é possível prever a distribuição espacial das classes de solos em Tracuateua, no nordeste paraense, com rapidez e baixo custo através do uso de técnicas de mapeamento digital.

Por meio do conhecimento tácito do pedólogo, da seleção criteriosa das variáveis preditoras e da avaliação dos métodos quantitativos de predição, é possível melhorar o mapeamento, diminuindo a subjetividade da interpretação do pedólogo e fornecendo um caráter mais quantitativo ao produto final.

1.2 Objetivos

1.2.1 Geral

Avaliar a eficiência de algoritmos de aprendizado de máquina (*Machine Learning*) para produção de mapa pedológico digital do município de Tracuateua, no nordeste paraense, utilizando dados de legados, sistemas de informações geográficas e variáveis geomorfométricas.

1.2.2 Específicos

- Produzir um mapa digital de solos em escala 1:100.000 utilizando o método de algoritmos de aprendizado de máquina, tomando como referência o mapa de solos de Tracuateua, Pará, Brasil;
- Selecionar o conjunto de covariáveis mais importantes para o mapeamento pedológico digital, usando algoritmos de aprendizado de máquina;
- Comparar algoritmos para mapeamento de solos em uma paisagem tropical entre a bacia do rio Tracuateua, rio Caeté e Reserva extrativista de Tracuateua.

2 MAPEAMENTO PEDOLÓGICO DIGITAL E USO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA EM TRACUATEUA, PARÁ

2.1 Introdução

Com a evolução da Inteligência Artificial (IA), o aprendizado de máquinas, também conhecido como *Machine Learning* (ML) e a demanda cada vez maior por dados de solos em escala detalhadas, permite que as técnicas computacionais se apresente como uma ferramenta promissora para o Mapeamento Digital do Solo (MDS), com a obtenção do mapa em menor tempo, baixo custo e erros conhecidos, que são vistos como as maiores vantagens em relação aos métodos convencionais de mapeamento (PINHEIRO *et al*., 2012). Devido suas vantagens e as diversas aplicações, os métodos digitais de mapeamento podem contribuir significativamente no planejamento do uso e ocupação do solo, principalmente na região amazônica.

De modo geral, o MDS é capaz de espacializar a distribuição dos solos por meio da análise das relações entre as características do solo e as variáveis ambientais, com o uso de modelos de aprendizado de máquina (*Machine Learning* - ML), geoestatísticos e uma combinação deles, os chamados *ensemble* (conjunto). As variáveis ambientais desempenham um papel importante na previsão dos solos ou de suas propriedades do solo em diferentes paisagens, especialmente em terrenos complexos (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003).

Assim, alguns estudos de MDS passaram a testar o uso de algoritmos para quantificar a espacialidade e distribuição das classes dos solos (MALONE, *et al*., 2017; BRUNGARD, *et al*., 2021; TAGHIZADEH-MEHRJARDI, R. *et al*., 2021). Isso levou à adoção de modelos híbridos que combinam diferentes métodos, como o uso de algoritmos de classificação, os quais são combinadas com o objetivo de aprimorar ambas as técnicas, incluindo sempre o componente espacial (ARRUDA; DEMATTÊ; CHAGAS, 2013).

E dentre as variáveis utilizadas nas abordagens de MDS, as derivadas do Modelo Digital de Elevação (MDE) têm sido testadas como variáveis preditoras (MENEZES *et al*., 2014; OLIVEIRA *et al*., 2012; PINHEIRO *et al*., 2012; YANG *et al*., 2016) em diferentes métodos de geração de mapas digitais e apresentado bons resultados nos trabalhos de mapeamento.

Diversas técnicas têm sido avaliadas para mapeamento de solos no Brasil e no mundo, dentre elas redes neurais artificiais, árvores de decisão e regressão linear múltipla. Entretanto, são poucos os trabalhos que mapearam solos ou atributos de solos na região norte do país, utilizando, além das covariáveis ambientais, variáveis espaciais como preditoras.

Diante desse contexto, este estudo pretende avaliar o desempenho do ML no mapeamento digital de solos. Serão avaliados algoritmos de ML para a classificação de solos em Tracuateua, no nordeste paraense, utilizando uma variedade de dados geoespaciais e técnicas de processamento de informações em SIG.

2.2 Revisão de literatura

2.2.1 Mapeamento convencional

O mapeamento pedológico convencional é uma ilustração da distribuição geográfica dos solos, estabelecida por um conjunto de relações e atributos ambientais que identifica e separa as unidades de mapeamento, além de prever e delimitar suas áreas na paisagem. Os dados obtidos a partir de um levantamento pedológico desempenham um papel crucial na avaliação das potencialidades e limitações de uma área, estabelecendo-se como um recurso fundamental para a realização de estudos de viabilidade técnica e econômica de projetos, além de auxiliar no planejamento do uso, manejo e conservação do solo (IBGE, 2015).

No Brasil, os primeiros trabalhos de mapeamento de solo datam do início do século XX. Eram realizados de forma manual e envolviam a coleta de amostras de solo, seguida de análises laboratoriais e produção de mapas manuais (CAMARGO *et al*., 2010; CARVALHO *et al*., 2013; LEPSCH, 2013; EMBRAPA, 2016). A partir da década de 1950, as técnicas de mapeamento evoluíram e incluíram o uso de fotografias aéreas e de tecnologias de georreferenciamento para produzir mapas de solos (FLACH, 2017). O Instituto Agronômico de Campinas (IAC) e outras instituições científicas desempenharam um papel importante na evolução do mapeamento de solos no Brasil.

Foi com essa evolução, que em 1970, o governo brasileiro, iniciou o Projeto Radar na Amazônia (RADAM). Este projeto teve como objetivo a coleta de dados sobre recursos minerais, solos, vegetação, uso da terra e a cartografia da Amazônia e que depois se estendeu a áreas adjacentes da região nordeste. E, posteriormente, a todo território nacional, quando passou a ser chamado de RADAMBRASIL, tornando-se na época, o maior projeto de mapeamento efetuado com radar aerotransportado (BECKER, 1996; PEREIRA; MENEZES, 2007) (Figura 1).

Figura 1 - Aeronave Caravelle - PTDUW utilizada no Projeto RADAM.

Fonte: Azevedo (2009).

Todos os registros obtidos pelos projetos RADAM e RADAMBRASIL foram organizados e disponibilizados em 550 mosaicos de radar na escala 1:250.000, constituindo um importante legado repleto de experiências e sendo a base para tomadas de decisões nas mais diversas áreas do conhecimento (BRASIL, 1973).

Atualmente, o mapeamento de solos pelo método convencional continua sendo a mais popular forma de mapeamento no mundo (MENDONÇA-SANTOS; SANTOS, 2003). Este método de mapeamento tem suas raízes na taxonomia biológica, mapeamentos geológicos e nos fundamentos teóricos básicos do método, destacados nos trabalhos de Dokuchaev (1883) e Jenny (1941), onde o segundo aborda de maneira quantitativa as complexas relações entre os fatores de formação dos solos, como representado na Equação 1.

$$Solo = f(cl, o, r, p, t, \ldots.) \quad (1)$$

Em que:

cl = representa a variável clima;

o = organismos;

r = relevo;

p = material de origem;

t = tempo;

.... = fatores não determinados.

De acordo com Hudson (1992), o mapeamento de solos consiste em uma estratégia científica baseada nos conceitos de fatores de formação dos solos acoplados com relações solo-paisagem. Trata-se, portanto, de um estudo do ambiente baseado em dois paradigmas complementares: o dos fatores de formação, para reconhecer e explicar os solos; e das relações solo-paisagem, para inferir e descrever sua distribuição espacial.

De posse das informações obtidas no levantamento pedológico é possível entender a qualidade do solo e suas características, bem como o potencial agrícola e a aptidão para usos específicos. Estas informações são úteis para a gestão de terras e recursos, a gestão de água, a conservação, a prevenção de erosão e a recuperação de áreas degradadas (IBGE, 2015).

Os resultados dos mapeamentos incluem um mapa de solos e um relatório textual que sintetiza o conhecimento adquirido, contendo a descrição e caracterização das classes de solo identificadas, a composição das unidades de mapeamento, a descrição da ocorrência dos solos na paisagem e dados analíticos e descritivos de perfis representativos (BUI, 2004).

Em geral, o mapeamento de solos começa com a coleta e estudo dos dados disponíveis, em seguida, são definidas as unidades básicas de mapeamento, a partir das quais os locais são escolhidos e visitados em campo para observar e descrever perfis representativos. Então, a partir da interpretação dos dados de campo, um modelo conceitual de solo-paisagem é desenvolvido para deduzir as variações espaciais do solo. Finalmente, o modelo conceitual é aplicado para deduzir a distribuição espacial dos solos no restante da área e delinear as unidades cartográficas definitivas (HUDSON, 1992; BUI, 2004).

A forma de execução praticamente não mudou desde os primeiros mapeamentos sistemáticos de solos (MILLER; SCHAETZL, 2014). A necessidade de recursos financeiros, equipe técnica treinada e tempo limitou o nível de detalhamento e extensão dos mapeamentos, o que provocou um declínio acentuado a partir de meados da década de 1980, quando os mapeamentos de solos praticamente estagnaram (MCBRATNEY; SANTOS; MINASNY, 2003).

Este método de mapeamento tradicional é o mais utilizado no mundo e é utilizado para separar em grupos os diferentes tipos de solos na paisagem. Porém, ele tem sido criticado por alguns autores, por seus aspectos subjetivos, pois, não considera a dependência espacial entre as unidades de mapeamento, a qual pode ser forte, principalmente em se tratando de

levantamentos detalhados ou em áreas onde os limites entre os solos não são óbvios (McBRATNEY; WEBSTER, 1981; BURGESS; WEBSTER, 1984; ODEH *et al*., 1990).

Essas limitações do método tradicional de mapeamento de solo foram parcialmente superadas com o desenvolvimento de novas tecnologias, como sensores remotos e técnicas de processamento de dados, que permitem a produção de mapas de solos com uma escala espacial maior e com uma quantidade mais completa de informações.

Do ponto de vista metodológico, o método convencional de mapeamento de solos consiste de três etapas: 1) Observação de dados auxiliares de campo e de perfis de solo descritos; 2) Incorporação dos atributos dos solos em um modelo conceitual implícito que é usado para inferir a variação do solo; 3) Aplicação do modelo conceitual, que é usado para inferir o solo em locais não observados. Logo, o modelo conceitual de variação do solo é então transformado em modelo cartográfico, delimitando unidades de mapeamento de solos em fotos aéreas e mapas topográficos. Nesses casos, as escalas das fotografias e/ou do mapa topográfico, bem como a densidade de pontos amostrais, definem a escala final do mapa de solos (VILLELA, 2013).

O Brasil é totalmente mapeado em 1:5.000.000 e 1:1.000.000, mas apenas 35% do território está coberto com escalas entre 1:100.000 e 1:600.000, e uma porção muito pequena tem mapas de solos mais detalhados (MENDONÇA-SANTOS; SANTOS, 2007). Neste sentido, a realização de mapeamentos de solos com mais detalhes possibilitará uma gestão mais eficaz dos recursos naturais, como a prevenção de erosão, o monitoramento da qualidade do solo, a identificação e recuperação de áreas degradadas. Por fim, a realização de mapeamentos de solos com mais detalhes contribuirá para o avanço da pesquisa científica sobre a dinâmica dos solos.

2.2.2 Avanços nacionais em mapeamento de solos - PronaSolos

O Programa Nacional de Levantamento e Interpretação de Solos do Brasil (PronaSolos) é uma iniciativa do governo federal brasileiro e tem como objetivo principal gerar informações sobre as características dos solos do país e sua distribuição geográfica. O programa foi criado em 1979 e coordenado pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) Solos, que é responsável por realizar o levantamento, a classificação e a interpretação dos solos brasileiros, além de produzir mapas e relatórios técnicos sobre o assunto (POLIDORO *et al*., 2021).

Atualmente, o PronaSolos está em sua terceira fase, que começou em 2017 e terá vigência de 30 anos, podendo o prazo ser prorrogado, dependendo de fatores incontroláveis.

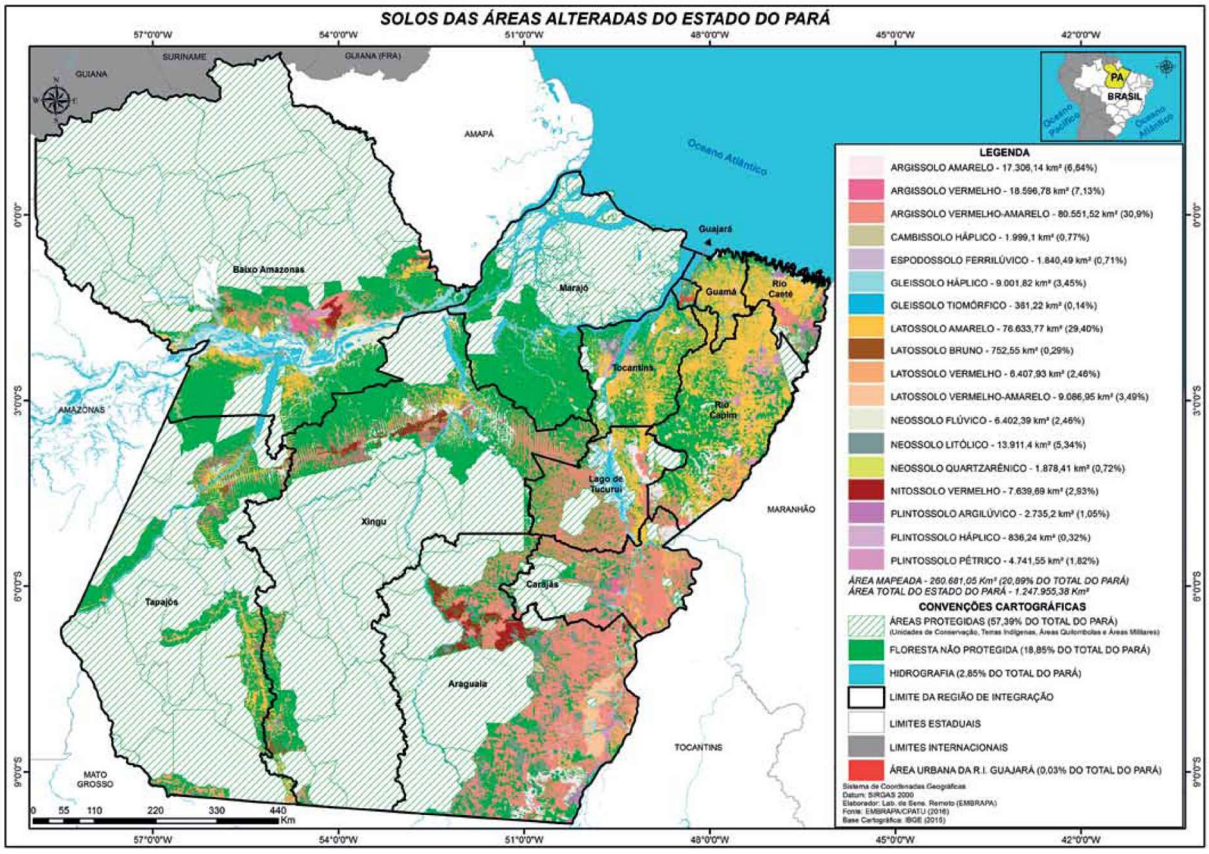
Nesta fase, o programa tem como objetivo principal atualizar as informações sobre os solos brasileiros e produzir novos mapas, além de desenvolver tecnologias e metodologias inovadoras para a análise dos solos (POLIDORO *et al*., 2021).

O programa surgiu com o objetivo de fornecer subsídios técnicos à formulação de políticas públicas de gestão territorial, visando à expansão e produção sustentável a partir de zoneamento ecológico e econômico de determinadas culturas (zoneamentos) em conformidade à legislação ambiental vigente (COELHO *et al*., 2014).

Seguindo exemplos de sucesso, como os Estados Unidos da América, que contam com um programa de levantamento de solos permanente, iniciado em 1896 (DITZLER *et al*., 2017), com recursos humanos e econômicos que possibilitam a sua execução em escalas cartográficas detalhadas (1:35.000 ou maior) em aproximadamente 95% do território (ARNOLD, 2016). Com expectativa de que em breve alcançarão 100% de seu território (DITZLER *et al*., 2017). Esta iniciativa, transformou o país em uma grande potência mundial em tecnologias inovadoras de mapeamento de solos e em produção sustentável agrícola.

No Brasil, um exemplo da aplicação e importância do levantamento de solos é a expansão das fronteiras agrícolas do país, mais especificamente a região reconhecida por MATOPIBA, uma junção das siglas dos estados do Maranhão (MA), Tocantins (TO), Piauí (PI) e Bahia (BA). E também de estados que fazem parte da região norte, numa parte da região amazônica, como o estado do Pará (Figura 2), a qual, tem experimentado diversas transformações socioeconômicas na última década (LUMBRERAS *et al*., 2015).

Figura 2 - Mapas de Solos e de Aptidão Agrícola das Áreas Alteradas do Pará.



Fonte: Embrapa Amazônia Oriental (2016).

Atualmente, o Brasil está mapeado na escala de 1:5.000,000 e 1:1.000.000, onde aproximadamente 8,6% do território brasileiro está coberto por mapas de solos entre as escalas 1:100.000 e 1:250.000; e somente 0,6% de seu território está mapeado em escalas mais detalhadas que 1:100.000 (SANTOS *et al*., 2013) (Figura 3). A demanda atual do país, coincide com o objetivo principal do PronaSolos, que é aumentar e suprir a carência por informações de solos no Brasil por meio de levantamentos de solos e interpretações de uso em escalas iguais ou mais detalhadas que 1:100.000 a fim de prover informações adequadas para as tomadas de decisão em níveis estaduais, municipais e de microbacia hidrográfica (POLIDORO *et al*., 2021).

Figura 3 - Tipos e escalas cartográficas de mapeamentos de solos existentes no Brasil.

Fonte: Santos (2013).

A recente expansão e intensificação da agricultura e pecuária, associada a sérios problemas de degradação de solos (compactação, contaminação, erosão, desertificação, arenização, salinização), bem como a questões ambientais globais, têm aumentado a conscientização e preocupação dos órgãos públicos e da sociedade civil sobre a gestão dos recursos solo e água (MONTANARELLA *et al*., 2016). Diante do exposto, é evidente que os levantamentos de solos se destinam a ajudar todas as pessoas interessadas a entender os solos e a melhor utilizá-los. Sendo justificativa suficiente para o fomento do programa pela iniciativa do poder público em parcerias de outras instituições (POLIDORO *et al*., 2016).

A partir de sua publicação no Diário Oficial da União, sessão 1, edição nº 117 de 20 de junho de 2018 (DOU, 2018), as ações estão programadas para serem executadas num prazo de 30 anos e em etapas de curto (1 a 4 anos), médio (5 a 10 anos) e longo (11 a 30 anos) prazo, sendo que todas devem ser submetidas a revisões periódicas em intervalos de dois anos a fim de adequá-las às demandas da sociedade, às novas tecnologias e às políticas governamentais (POLIDORO *et al*., 2021).

Segundo Polidoro *et al*. (2021), o período de vigência do programa está estimado em 30 anos, mas, existe a possibilidade que a sua duração seja inferior ou superior, devido a fatores difíceis de prever, como a evolução de técnicas de mapeamento de solos, que podem reduzir os custos para execução dos trabalhos de mapeamento, bem como o tempo dos mesmos.

Diante dessa situação, a pedometria constitui uma importante ferramenta, que poderá ajudar no processo de obtenção de informação quantitativa com uma incerteza mensurável e de alta qualidade, por meio da aplicação de novos métodos, procedimentos e ferramentas. Procedimentos de análise de solos usando sensores proximais, agilizarão e diminuirão os custos aplicados aos mapeamentos (CARVALHO JUNIOR; MENDONÇA SANTOS; ANJOS, 2017).

2.2.3 Pedogênese e relação solo-paisagem

A noção de solo como um resultado das interações dos fatores ambientais inspirou diferentes abordagens matemáticas, sendo que Dokuchaev propôs sua própria equação em 1899. Isso levou Chas F. Shaw, em 1930, a formular a primeira equação que relaciona os fatores de formação do solo e apresentou-a durante o Congresso Internacional de Ciência do Solo de 1932, conforme descrito na Equação 2.

$$S = m\,(c + v)t + d \tag{2}$$

Em que S = solo, m = material de origem, c = clima, v = organismos vivos, t = tempo e d = processos de erosão e deposição. Naquele momento, considerava-se os fatores, as causas da formação dos solos e as propriedades, seus efeitos (SARMENTO, 2015).

De acordo com Jenny (1941), em seu livro *Factors of Soil Formation*, o solo é um sistema aberto, resultado dos fatores de formação do solo e estes atuam de maneira independente, reescrevendo a função de Shaw (1932), assim o solo é função do clima, dos organismos, do relevo, do material de origem, do tempo, conforme representado na Equação 1.

Essa teoria implica em que, se a distribuição espacial dos fatores de formação for conhecida, o solo ou suas propriedades podem ser estimados. Isso significa, que as relações entre eles podem ser quantificadas (SCULL *et al*., 2003; BUOL *et al*., 2011).

2.2.4 Material de origem

Na formação do solo, o material de origem influencia diversos atributos e pode ser dividido em dois grupos: rochas e sedimentos. As principais características das rochas que influenciam nos atributos do solo são: composição química, mineralogia, cor e textura (BRADY; WEIL, 2013). Já os sedimentos são formados a partir da intemperização das rochas e atuação de processos erosivos, sendo transportados e depositados ao longo da paisagem. Podem ser classificados em coluviais, ou seja, resultado da intemperização e erosão nos pontos altos da paisagem e depositados ao longo da encosta ou aluviais, depositados em ocasião de transbordamento dos rios (SUGUIO, 2003).

2.2.5 Relevo

O relevo é responsável por promover no solo diferenças facilmente perceptíveis ao observar um perfil de solo (LEPSCH, 2010). É um fator importante na formação do solo, pois é responsável pelo controle de toda dinâmica dos fluxos de água na paisagem, como lixiviação de solutos, atuação de processos erosivos e condições de drenagem (ANJOS *et al*., 1998). Dependendo do tipo de relevo, sendo inclinado ou plano, a água da chuva poderá infiltrar no solo, escoar pela superfície ou se acumular (RODRIGUES, 2018).

O relevo plano facilita a infiltração da água, propiciando condições para a formação de solos profundos, a exemplo, temos os Latossolos. Em relevos inclinados, o escoamento superficial da água é facilitado, favorecendo a erosão e dificultando a pedogênese, nessas áreas geralmente formam-se solos rasos. Em relevo abaciado, tende-se haver maior acúmulo de água de chuva ou provenientes de regiões mais altas, característicos de áreas de várzeas, onde se formam geralmente solos hidromórficos (PEREIRA *et al*., 2019).

2.2.6 Clima

A atuação do clima na pedogênese está associada aos atributos precipitação pluvial, as taxas de evaporação e a temperatura, tendo em vista a influência dos mesmos no intemperismo e evolução dos solos (KÄMPF, CURI, 2012). Climas úmidos e quentes são fatores favoráveis à formação de solos muito intemperizados, bem drenados e profundos, resultando em acidez elevada e baixa fertilidade natural. Essas características devem-se ao fato de geralmente apresentarem maior cobertura vegetal, associada a mais agentes de intemperismo (físico e químico). Este fator pode explicar, por exemplo, que em regiões de climas úmidos, ácidos orgânicos e argilas podem ser transportadas pela água para os horizontes mais profundos do solo, pelo processo denominado eluviação, esse deslocamento pode ser tanto horizontal, quanto vertical e segue o fluxo hídrico (LEPSCH, 2011).

2.2.7 Organismos

Os organismos possuem relação íntima com o fator clima na formação do solo, considerando a adaptabilidade da fauna e da flora às condições de umidade e temperatura de um determinado ambiente. São considerados condicionantes para a pedogênese - a ação dos organismos no substrato representa a diferença entre os processos de pedogênese e intemperismo (LEPSCH, 2011).

A matéria orgânica adicionada ao solo pelos vegetais, seja pelos resíduos de folhas ou de raízes, ou seja pela decomposição, por meio da ação da fauna, como formigas, minhocas e microrganismos, participa de diversos processos no solo e influencia na agregação de partículas, no escurecimento do horizonte superficial, na infiltração da água, minimizando a erosão e, na retenção de nutrientes fundamentais ao desenvolvimento das plantas (PAVINATO; RESOLEM, 2008).

2.2.8 Tempo

O fator tempo apresenta uma relação não apenas de cronologia, mas também de maturidade e evolução (KÄMPF, CURI, 2012). Em ambientes de clima árido e semiárido, com baixa precipitação pluviométrica, mesmo com o material de origem exposto por um longo tempo, a baixa intensidade de intemperização formará solos jovens, pouco evoluídos. Por outro lado, condições de intenso intemperismo e alteração do material de origem, mesmo com exposição recente deste, formará solos maduros e evoluídos do ponto de vista da pedogênese (PEREIRA *et al*., 2019).

2.2.9 Conhecimento tácito e mapeamento digital de solos

Na visão de Bennett (2001), o conhecimento tácito é um tipo de conhecimento que é difícil de ser expresso verbalmente ou formalizado por meio de regras e procedimentos. É o conhecimento implícito nas experiências, crenças, valores e percepções de uma pessoa, que são muitas vezes inconscientes e não articuladas (DAVENPORT; PRUSAK, 2004). É o tipo de conhecimento que as pessoas adquirem ao longo do tempo por meio da prática e da experiência, e que muitas vezes é difícil de ser transferido para outras pessoas (HUDSON, 1992).

Embora o conhecimento tácito seja difícil de ser expresso e formalizado, ele é muitas vezes considerado mais valioso do que o conhecimento explícito ou formalizado, uma vez que é baseado na experiência real e na prática, e pode levar a uma compreensão mais profunda e holística do mundo. No entanto, é importante notar que o conhecimento tácito pode ser complementado e aprimorado pelo conhecimento explícito, como manuais e guias, para melhorar a compreensão e a transferência do conhecimento entre as pessoas (NONAKA *et al*., 2000).

No mapeamento tradicional de solos, ao fazer a fotointerpretação, o pedólogo usa seu conhecimento tácito para separar polígonos homogêneos, estabelecendo assim a relação solo-paisagem, definindo as classes de solos que estão associados àquele polígono, o que depois é validado em campo (FRANCO *et al*., 2015). A posição espacial dos polígonos do solo relaciona-se aos diferentes tipos de solo e as suas condições ambientais subjacentes. A delimitação de polígonos está integrada aos múltiplos conhecimentos do pedólogo (QI; ZHU, 2003).

Os autores também explicam que a ideia básica de obter informações de polígonos exibidos em um mapa de solo, é reverter o processo de mapeamento. As relações entre as características do solo e da paisagem podem ser reveladas através da abordagem de descobrimento de conhecimento (conhecimento tácito), analisando mapas de solos com características da paisagem capturadas por Sistema de Informação Geográfica (FRANCO *et al*., 2015).

A consolidação do conhecimento tácito nas relações solo-paisagem está baseada nas covariáveis ambientais que efetivamente descrevem os fatores de formação do solo. Na maioria dos trabalhos de mapeamento as covariáveis comumente utilizadas na avaliação do modelo relação solo-paisagem são geologia, topografia (MDE) e índices relacionados à vegetação (QI; ZHU, 2003).

O uso combinado do conhecimento tácito especializado e técnicas de MDS, revelou-se capaz de alocar com sucesso solos individuais dentro das unidades do mapa, mesmo quando há pouca informação contextual disponível (SARMENTO *et al.*, 2017).

O trabalho realizado por Silva Júnior *et al*, (2021) demonstrou que a utilização combinada de janela móvel bipartida (SMW) e janelas móveis bipartidas multivariadas (MSMW) juntamente com unidades geomorfológicas e conhecimento tácito se mostraram ferramentas promissoras para a delimitação dos limites do solo. Ainda segundo os autores, a utilização da técnica (MSMW) durante mapeamentos convencionais e digitais de solos é fundamental para os pedólogos validarem seu conhecimento tácito sobre as relações solo-paisagem, além de proporcionar uma ampliação da visão pedológica.

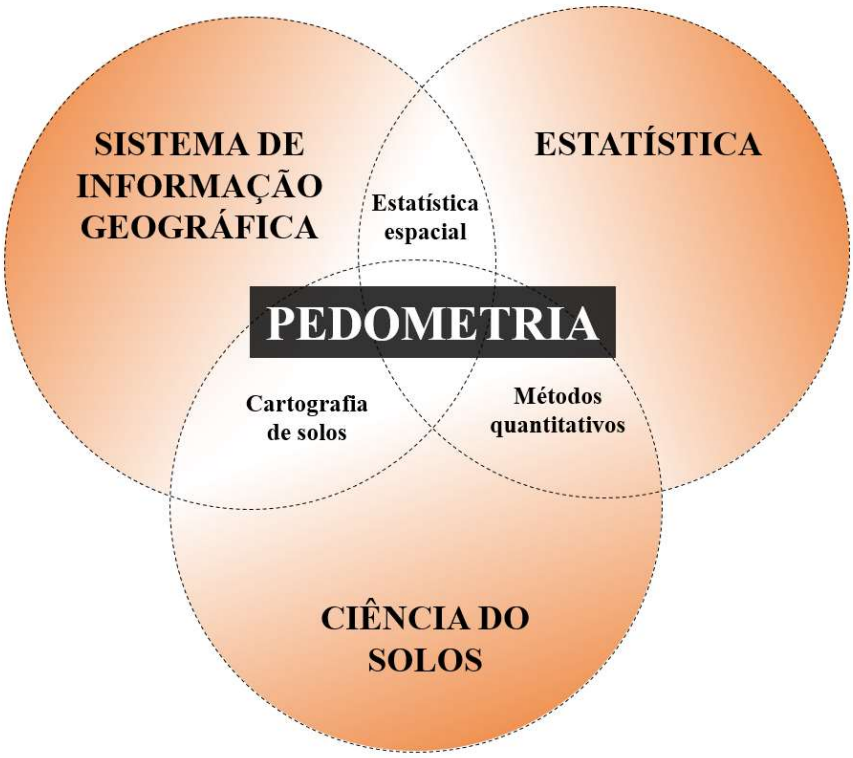
Dessa forma, o conhecimento tácito de um pedólogo é fundamental no mapeamento digital de solos, por diversos motivos, uma vez que a construção desse tipo de mapa requer a integração de diferentes tipos de informações, incluindo dados de sensoriamento remoto, informações geológicas, hidrológicas e topográficas, além de conhecimentos específicos sobre os solos da região em questão (BUI, 2003).

2.2.10 Pedometria

A pedologia é usualmente conceituada como sendo o estudo do solo, comumente subdividida em classificação, fatores, mapeamento, morfologia e processos de formação (BOCKHEIN *et al*., 2005). Contudo, nos últimos anos, os pesquisadores buscando formular e resolver questões de mapeamento de solos e atributos através de técnicas matemáticas e estatística, surgiu a pedometria como alternativa a essa demanda (MINASNY *et al*., 2014).

A pedometria teve origem no trabalho de Richard Webster (1994), e foi oficialmente cunhada nos trabalhos de McBratney *et al*, (2003) e Hengl (2003). Neste sentido, o estudo da distribuição, organização e gênese de solos se dá por meio da aplicação de métodos quantitativos, sendo esse o foco principal da pedometria. Sendo bastante aplicada nos trabalhos de mapeamento digital de solos (classes e atributos), ela continua a se desenvolver consistentemente, tornando-se uma ciência interdisciplinar que integra os campos da ciência do solo, matemática e estatística aplicada, bem como geoinformação (WADOUX *et al*., 2021) (Figura 4).

Figura 4 - A pedometria como uma ciência interdisciplinar.



Fonte: Hengl (2003).

Logo, a representação das feições do terreno de forma digital permitiu classificar as formas da paisagem e definir os limites de classes dos solos de maneira rápida e com menor subjetividade (TEN CATEN *et al*., 2012; CUNHA, 2013).

2.2.11 Mapeamento digital de solos

A técnica do mapeamento digital de solos (MDS) é aplicada para modelar e mapear as classes ou atributos do solo. Esse procedimento envolve a manipulação de dados coletados no campo e laboratório, bem como inferências baseadas nos fatores que influenciam a formação do solo e nas variáveis ambientais, juntamente com o uso de algoritmos de aprendizado de máquina (LAGACHERIE; MCBRATNEY, 2007).

Ao considerar as variações a partir da utilização de dados obtidos do solo, a modelagem aplicada em solos pode ser entendida como o "desenvolvimento de modelos matemáticos que tentam simular, da melhor forma possível, os processos de formação do solo" (VEREECKEN *et al*., 2016).

É importante ressaltar que o procedimento de mapeamento digital de solos e de atributos de solos, por vezes adotem a mesma metodologia, tratam-se de objetos distintos. O mapeamento digital de solos, refere-se às classes de solos (ordens, subordens, grandes grupos,

subgrupos, famílias e séries), conforme os estudos de Sarmento *et al*., 2012; Meier *et al*., 2018 e Coelho *et al*., 2021, enquanto que o mapeamento de atributos de solos, refere-se às propriedades físicas e químicas do solo (carbono orgânico, areia, silte e argila) (CAMPBELL *et al*., 2019; BELLINASO, H. *et al*., 2021; ROSIN *et al*., 2023).

Considerando as limitações (subjetividade, custo de execução, tempo) do método de mapeamento convencional de solos e o desenvolvimento de algoritmos de aprendizado de máquinas e ferramentas computacionais, a partir da década de 70, os avanços tecnológicos possibilitaram o surgimento de uma nova perspectiva metodológica de mapeamento, em que a variabilidade espacial dos solos pudesse ser mapeada e representada por expressões numéricas. A evolução dessas tecnologias possibilitou a execução de diversos trabalhos de mapeamento digital de classes de solos (MINASNY; MCBRATNEY, 2016; CHAGAS *et al*., 2017; ZHANG *et al*., 2022; COELHO *et al*. 2021; NASCIMENTO *et al*., 2022; LEMERCIER *et al*. 2022).

Adaptando a equação proposta por Jenny (1941), McBratney, Santos e Minasny (2003) propuseram então um modelo determinístico, para correlacionar os processos de formação e distribuição dos solos na paisagem com técnicas quantitativas usadas em pedometria, a proposta resultou na Equação 3, que ficou conhecida como modelo SCORPAN.

$$S_{c,a} = f(s, c, o, r, p, a, n) \quad (3)$$

Em que S_c = classes de solo e S_a = atributos de solo;

s = solo, outras propriedades do solo em um ponto;

c = clima, propriedades climáticas do ambiente em um ponto;

o = organismos, vegetação ou fauna ou humanos atividade;

r = topografia, atributos da paisagem;

p = material original, litologia;

a = idade, o fator tempo;

n = espaço, posição espacial.

Basicamente, a diferença entre a equação de Jenny e a apresentada pelos autores, é a introdução dos fatores solo e espaço. A introdução da informação solo tem por objetivo associar ao modelo, o conhecimento existente sobre os solos, na região que se pretende mapear digitalmente. Esse conhecimento pode ser incorporado de diferentes formas, como, por exemplo, a classe de solo do local e/ou atributos do solo medidos que auxiliem no mapeamento. O espaço se refere às coordenadas x, y e outras medidas de distância, da informação de solo associada (VILLELA, 2013).

Contudo, Sarmento (2010) ao analisar a rotina dos trabalhos de mapeamentos digitais, constatou que a maioria dos trabalhos seguem algumas etapas que podem ser consideradas equivalentes àquelas que são usadas nos levantamentos convencionais.

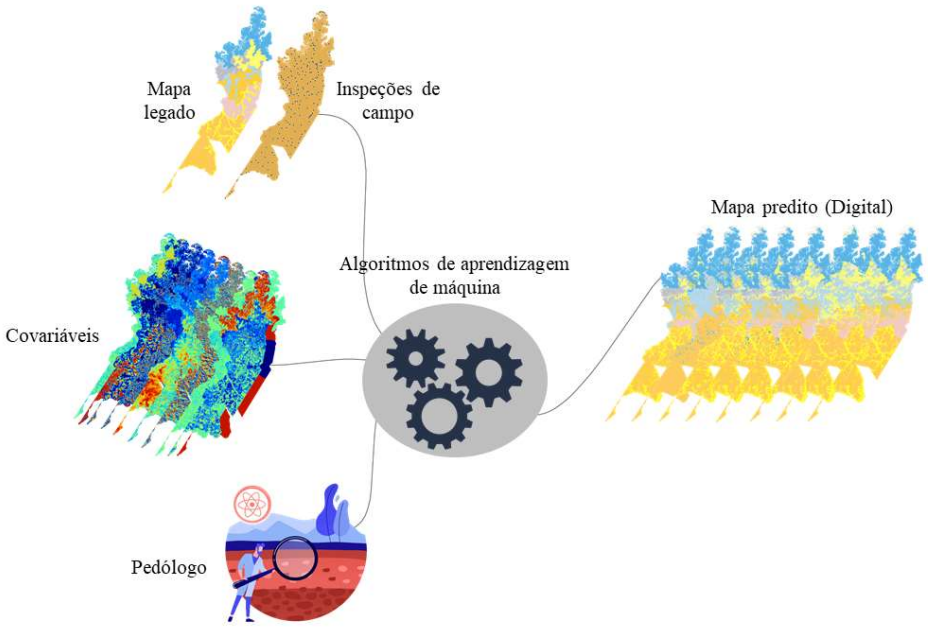
O mapeamento digital, assim como o convencional, utiliza técnicas similares, com informações coletadas ou disponíveis em pontos de observação de solos. Essas informações são utilizadas para ajustar um modelo quantitativo com variáveis relativas às condições do ambiente nos mesmos locais e o modelo ajustado é depois empregado para predizer propriedades do solo ou classes de solos para o restante da área (MCBRATNEY; SANTOS; MINASNY, 2003) e (LAGACHERIE, 2008).

Na prática, outro diferencial entre o método digital e o convencional, é a produção de expressões quantitativas da variabilidade do solo, medidas pelo índice *Kappa* e acurácia. Neste sentido, o índice *Kappa* e a acurácia funcionam como parâmetros de avaliação do grau de incertezas do método e permitem a interpretação do mapa digital, na figura 5, apresenta-se um resumo geral do processo de mapeamento digital de solos.

Outras técnicas também estão sendo empregadas na tentativa de melhorar a exatidão das predições espaciais de classes de solos e/ou atributos, como exemplo temos os métodos híbridos de mapeamento, que combinam dados geoestatísticos e estatístico ou multivariados, por krigagem universal, cokrigagem, krigagem-regressão, krigagem com tendência externa e krigagem fatorial (MCBRATNEY *et al.*, 2000).

Este método tem sido empregado nos trabalhos de mapeamento digital de solos por ser de fácil implementação e demonstrarem bons resultados quando comparado com outros (MINASNY e MCBRATNEY, 2007; AKSOY *et al*., 2012).

Figura 5 - Fluxograma das etapas do método de MDS.



Fonte: O autor (2023).

2.2.12 Covariáveis ambientais utilizadas em mapeamento de solos

É notório que um dos avanços da ciência do solo no Brasil e no mundo, a partir dos anos de 1980, foi a possibilidade de aplicação de tecnologias informação no Mapeamento Digital de Solos (MDS), sendo possível o mapeamento de pequenas e grandes áreas, aplicando as diversas ferramentas disponíveis para análise, modelagem e predição, aos conjunto de dados de campo para a inferência de variações espaciais dos tipos de solos e suas propriedades (TEN CATEN *et al*., 2012; DEMATTÊ *et al*., 2014; DALMOLIN *et al*., 2017).

As covariáveis ambientais são usadas como preditores em algoritmos denominados de *Machine Learning* (WADOUX *et al*., 2020). Eles podem ajudar a explicar parte dos processos que promovem a variação espacial do solo na paisagem, sendo que a maioria dos estudos usaram cerca de 20 covariáveis. Apenas alguns usaram menos de cinco (DAI *et al*., 2014; PADARIAN *et al*., 2019), enquanto outros usaram mais de 100 (HENGL *et al*., 2017; RAMCHARAN *et al*., 2018).

Uma vez que essas tais covariáveis representam os fatores formadores do solo, muitos estudos (ROSSEL; CHEN, 2011; WANG *et al*., 2018; GOMES *et al*., 2019; SZATMÁRI; PÁSZTOR, 2019) utilizaram métodos para selecionar covariáveis representativas dos fatores-chave do modelo SCORPAN de variação espacial do solo.

Os mais comuns são propriedades existentes do solo ou mapas de classe, média anual de precipitação e temperatura, imagens de sensoriamento remoto, imagens de satélite ou índices de vegetação derivados de imagens de satélite, elevação, atributos do terreno (declive, curvatura local, índice de umidade topográfica) e mapas geológicos existentes (WADOUX *et al*., 2020).

Muitas vezes os dados que representam os fatores SCORPAN podem não estar disponíveis ou facilmente obtidos. Vários estudos, calibraram os algoritmos de *Machine Learning* usando conjuntos de variáveis climáticas, imagens de sensoriamento remoto ou atributos de terreno apenas, ou uma combinação deles (WADOUX *et al*., 2020).

No trabalho de Mansuy *et al*. (2014) usaram um conjunto de oito variáveis de atributos climáticos e oito de terreno para mapear carbono, nitrogênio e textura do solo em uma área extensa no Canadá. Já Sharififar *et al*. (2019) usaram seis atributos de terreno como covariáveis. Estas foram escolhidas de um grande conjunto de covariáveis ambientais usando conhecimento sobre a relação esperada entre a covariável e a propriedade do solo a ser mapeada. Outros estudos (HENGL *et al*., 2018; MILLER *et al*., 2015) consideraram um número muito grande (> 100) de covariáveis para calibração dos algoritmos.

A quantidade de covariáveis depende principalmente de imagens de sensoriamento remoto, como produtos terrestres MODIS ou produtos Landsat. Usar covariáveis temporais em um algoritmo de *Machine Learning* também é uma forma de mapear a dinâmica temporal das propriedades do solo. Os trabalhos de Heuvelink *et al*. (2021) usaram séries temporais de produtos MODIS para mapear a dinâmica temporal do carbono orgânico na Argentina entre 1982 e 2007.

Alguns estudos explicam a variação multiescala das covariáveis ambientais. Em outras palavras, os derivados do terreno podem ser agregados para explicar os processos físicos no solo que não são visíveis em uma escala mais fina. Exemplos de estudos usando covariáveis multiescala para mapeamento com algoritmos de *Machine Learning* são Behrens *et al*. (2010), Miller *et al*. (2015) e mais recentemente Behrens *et al*. (2018).

É neste contexto, que as variáveis do terreno, tidas como explicativas para a distribuição de classes de solos ou de seus atributos são obtidas através do Modelo Digital de Elevação (MDE) (COELHO; GIASSON, 2010; TEN CATEN *et al*., 2012; COSTA, 2016).

O MDE é definido como a variação contínua do relevo sobre o espaço, que nas etapas de levantamento de dados no MDS, pode ser utilizado para obter variáveis geomorfométricas ou atributos topográficos (WILSON; GALLANT, 2000). Sendo uma alternativa rápida e econômica para a quantificação e classificação do relevo, permitindo a obtenção de dados precisos e, que possibilitam a definição das unidades morfológicas da paisagem (IPPOLITI *et*

al., 2005). Permitindo assim, uma melhor compreensão do comportamento da superfície do terreno e indicativo de tendências no comportamento dos atributos do solo, e, desta forma, auxilia nos levantamentos de solos. Essas são, algumas das razões pela qual as variáveis geomorfométricas, são bastante utilizadas como preditoras nos trabalhos de MDS (COSTA, 2016).

Essas variáveis, podem ser classificadas de duas formas: atributos topográficos primários e os atributos topográficos secundários. O primeiro, corresponde literalmente ao MDE, e representa as características básicas como a altitude, declividade, orientação e forma da vertente (OLIVEIRA, 2019). Dentre estes, a declividade possui papel importante na formação dos solos, uma vez que, áreas de maior declividade, favorecem a formação de solos menos desenvolvidos pedogeneticamente, pois a ação do fator escoamento superficial é mais intensa e há retirada de material. Ao passo que, áreas de menor declividade, há pouco escoamento superficial e maior percolação de água no solo, o que intensifica os processos de intemperismo, logo ocorrem solos mais desenvolvidos (LEPSCH, 2010; KAMPF & CURI, 2012).

Já os atributos topográficos secundários são obtidos a partir de um ou mais atributos primários e representam características mais complexas, esses dados são importantes no estudo de evolução da vertente, pois podem indicar a formação de bacias de ordem zero (MINELLA; MERTEN, 2012). Logo, as covariáveis topográficas oriundas do MDE, juntamente com dados legados, dados climáticos e outras variáveis preditoras, quando, associadas aos algoritmos de aprendizado de máquina (*machine learning*), principal técnica de MDS, constituem-se em importantes variáveis no conhecimento da espacialização e predição de classes de solos na paisagem. (IPPOLITI *et al*., 2005; OLIVEIRA, 2019).

2.2.13 Técnicas de predição espacial de classes de solos

As diversas técnicas de MDS buscam correlacionar as variáveis geomorfométricas, ditas como preditoras, de forma lógica, identificando as regiões onde há semelhança de parâmetros e características do relevo que permitam cada classe de solo (GONÇALVES, 2019). Há diversos métodos que são empregados, a depender do tipo de dado a ser mapeado, para especializar a distribuição de classes ou atributos de solo que podem ser processadas e manipuladas em ambiente computacional. As técnicas de processamento e análise de dados mais utilizados são: geoestatística (*kriging*), lógica de conjuntos difusos (*lógica Fuzzy*), redes neurais artificiais, regressões múltiplas e árvores de decisão (SARMENTO, 2010).

A geoestatística surgiu na África do Sul, quando Krige (1951), ao analisar as variâncias das concentrações de ouro, precisou levar em consideração as distâncias entre as amostras e, assim, surgiu a teoria das variáveis regionalizadas, que considera a localização geográfica e a dependência espacial (GREGO; OLIVEIRA, 2015), ou seja, quanto mais próximas são as amostras, mais semelhanças entre si elas se apresentam.

O método de krigagem, leva este nome em homenagem ao seu autor, Daniel Krige e consiste em ponderar os vizinhos mais próximos do ponto a ser estimado, obedecendo os critérios não tendenciosidade, que significa que, em média, a diferença entre valores estimados e observados para o mesmo ponto deve ser nula e ter mínima variância (BOISVERT; DEUTSCH, 2011).

As principais limitações do método estão relacionadas: a) à hipótese de estacionariedade que geralmente não é encontrada em campo; b) a grande quantidade de dados requeridos para definir a autocorrelação espacial, e c) às situações de complexidade do terreno e dos processos de formação do solo (MCBRATNEY *et al*., 2000).

No trabalho de Yoshida e Stolf (2016), os autores utilizaram a krigagem para o levantamento das características geomorfológicas e os atributos físicos e químicos do horizonte superficial dos solos. Os modelos permitiram destacar as regiões onde os atributos são fortemente encontrados, e as regiões menos expressivas, com mudança gradual dos padrões, assim como ocorre no ambiente.

A lógica *Fuzzy* por sua vez é considerada como análise algébrica de mapas não cumulativos (NEUMANN, 2012). A aplicação dos métodos *Fuzzy* permite manipulação de informações incertas e imprecisas, simulando o comportamento de raciocínio humano (Figura 13). No MDS o método possibilita a espacialização de classes de solos na paisagem sem o estabelecimento de limites rígidos. Logo, é mais adequada às mudanças graduais que ocorrem na transição de uma classe de solo para outra (GONÇALVES, 2019).

Dessa forma, quando observada alteração do relevo é perceptível a passagem gradual das características do solo, sem mudança brusca. Esta funcionalidade do método tem atraído a atenção de estudiosos devido à habilidade em capturar e representar a natureza contínua da variação espacial do solo na paisagem (NEUMANN, 2012; LIMA, 2013).

Para o MDS, a Lógica Fuzzy possibilita a alocação de indivíduos (*pedons*), de acordo com o grau de pertinência do indivíduo em relação a cada classe de solo mapeada. O fator limitante do método, está na escala de trabalho, assim, se a escala for pequena, é viável trabalhar com lógica booleana. Se o trabalho for em grande escala será indicado o uso da Lógica Fuzzy, atrelado ao bom conhecimento do comportamento da variável ambiental (LIMA, 2013).

Outra técnica muito empregada na obtenção de mapas de solos é o uso das Redes Neurais Artificiais (RNA). Este algoritmo realiza o processamento de um grande conjunto de dados que operam em paralelo, permitindo o estabelecimento de relações matemáticas entre as variáveis ambientais e as classes ou atributos de solos (NEUMANN, 2012; ARRUDA; DEMATTÊ; CHAGAS, 2013).

As vantagens deste método, incluem a possibilidade de manipulação de grande número de dados e a capacidade de generalização. O método tem a habilidade de manipular os dados adquiridos de diferentes fontes e com diferentes níveis de precisão (NEUMANN, 2012). As RNA possuem capacidade de aprender e ser treinada por meio de exemplos, por apresentar bom desempenho em tarefas mal definidas e não requerem conhecimento a respeito dos modelos matemáticos (LIMA, 2013).

As regressões lineares múltiplas, também são empregadas em trabalhos de MDS, pois assumem a existência de uma relação linear entre a variável dependente e duas ou mais variáveis independentes (preditoras). Já as regressões logísticas, são calculadas pelo método da máxima verossimilhança. O método assume que a variável dependente é categórica (NEUMANN, 2012). O trabalho de Coelho e Giasson (2010) afirma que por ser utilizada mais de uma variável independente, o modelo é considerado múltiplo e, pelo fato de predizer mais que duas classes de solo, é considerado multinomial.

Nos algoritmos de árvore de decisão, o modo de operação simula o processo de abstração humana através de uma categorização hierárquica, obtendo regras similares a uma chave de classificação. O método é baseado em algoritmos de aprendizagem de máquina que estabelecem modelos através de exemplos (SARMENTO, 2010).

Uma árvore de decisão geralmente começa com um único nó que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais que se ramificam em outras possibilidades. Um exemplo deste modo de operação é o *Random Forest*. Esta técnica de processamento é abordada por Zhang, Liu e Song (2017) como sendo uma das técnicas mais empregadas no mapeamento de solos, devido a sua robustez e elevada performance nos trabalhos de MDS (WOLSKI *et al*., 2017).

2.2.14 Inteligência Artificial (IA)

A ideia de criar máquinas capazes de pensar como seres humanos e realizar tarefas inteligentes remonta a antigas histórias mitológicas e filosóficas (BARBOSA, 2020). No entanto, o termo "inteligência artificial" foi cunhado somente em 1956, durante a Conferência

de Dartmouth, que reuniu pesquisadores de diferentes áreas para discutir o tema (PEREIRA ; NORVIG, 2009; GANASCIA, 2018).

Neste sentido, a Inteligência Artificial (IA) pode ser caracterizada pela capacidade do sistema de interpretar corretamente dados externos, aprender a partir desses dados e utilizar essa aprendizagem para atingir objetivos e tarefas específicas, por meio de adaptação flexível (KAPLAN; HAENLEIN, 2019). Assim, as técnicas de IA como redes neurais artificiais e clustering são utilizadas para analisar os dados e auxiliar na tomada de decisões inteligentes (BU; WANG, 2019).

A IA origina-se da capacidade de simular o comportamento do cérebro humano para resolver problemas complexos e poder apresentar uma estratégia eficiente de simulação e otimização de processos (CHEN; JAKEMAN; NORTON, 2008).Um dos campos mais explorados da IA é o aprendizado de máquina, que compreende algoritmos capazes de "aprender", ou seja, tomar decisões e/ou fazer previsões a partir de dados fornecidos (KOHAVI; PROVOST, 1998). Dentre as diversas técnicas de aprendizado de máquina existentes atualmente, duas das mais aplicadas na resolução de problemas são as redes neurais artificiais e algoritmos genéticos (FATEHNIA; AMIRINIA, 2018).

2.2.15 Aprendizado de máquina (*Machine Learning*)

O Aprendizado de máquina ou *machine learning* é uma subárea da inteligência artificial (IA) e tem por objetivo desenvolver algoritmos e técnicas computacionais capazes de adquirir conhecimento de maneira automática (GOLDSCHIMIDT, 2010). O conceito de IA, faz referência a ideia de que um sistema inteligente é capaz de reconhecer aspectos específicos e pode ser realizado através de programação específica (GARBADE, 2018). Já na aprendizagem de máquina, há uma função matemática de aprendizagem que é definida de acordo com os dados apresentados ao modelo (*Input data*) (BROWNLEE, 2016; KANIOURA; EITEL-PORTER, 2020) e pode ajudar no processo de mapeamento de solos, por auxiliar o pedólogo nas avaliações de relação solo-paisagem.

Os algoritmos de aprendizado de máquina são construídos de forma a aprender e fazer previsões a partir de dados. Ao contrário dos algoritmos de programação, os algoritmos de aprendizado de máquina realizam uma análise automatizada dos dados e realizam previsões a partir de amostras dos dados (CHOUDHARY; GIANEY, 2017).

O aprendizado de máquina é eficaz para regressão e classificação (supervisionada ou não-supervisionada) de sistemas não lineares, possibilitando utilizar diversas de variáveis (LARY *et al*., 2016). Há quatro diferentes tipos de aprendizagem: a supervisionada, não

supervisionada, semi-supervisionada e o aprendizado por reforço (RUSSELL; NORVIG, 2013).

Nos algoritmos de aprendizado supervisionado, os atributos de saída e os dados de treinamento são conhecidos, com possibilidade de avaliar a capacidade do resultado de predizer os dados. Já no aprendizado não supervisionado, não se conhece os atributos de saída e objetiva-se estabelecer a existência de grupos ou encontrar padrões de associação dos dados (MONTAÑO, 2016).

No aprendizado semi-supervisionado há um conjunto de dados rotulados e um conjunto maior de dados não rotulados, sendo um tipo intermediário entre o aprendizado supervisionado e não supervisionado, e no aprendizado de máquina por reforço, o agente aprende a partir de um conjunto de reforços, após decidir quais ações realizadas foram responsáveis pelo resultado predito (RUSSELL; NORVIG, 2013).

O aprendizado supervisionado divide-se em classificação e regressão. Se os rótulos forem conjuntos discretos de valores (categorias) o problema de aprendizado é chamado de classificação, se os rótulos forem números o problema de aprendizagem é a regressão (RUSSELL; NORVIG, 2013). Já o aprendizado não supervisionado, é dividido em agrupamento, onde os dados são agrupados conforme a similaridade ou associação, que consiste em encontrar padrões nos atributos de um conjunto de dados, e sumarização, que consiste em encontrar uma descrição simples para um conjunto de dados (FACELI *et al*., 2011).

2.2.16 *Support Vector Machine* (SVM)

O *Support Vector Machine* (SVM) foi fundamentado na teoria da aprendizagem estatística e é um algoritmo de aprendizado de máquina, que foi desenvolvido por Vapnik (VAPNIK, 1995), com o intuito de resolver problemas de classificação de padrões. A máquina de vetores suporte é uma outra categoria das redes neurais, cujas saídas dos neurônios de uma camada alimentam os neurônios da camada seguinte, não ocorrendo realimentação (HAYKIN, 2001).

Esta técnica originalmente desenvolvida para classificação binária, consiste na construção de um hiperplano como superfície de decisão, de tal forma que a separação entre um conjunto de dados seja máxima (Figura 16). Isso considerando padrões linearmente separáveis (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O desempenho do classificador é mensurado pelo número de predições incorretas que o mesmo obtém durante o treinamento, sendo o erro o menor possível. Sendo assim definimos

como risco empírico, como sendo a medida de perda entre a resposta desejada e a resposta real (LORENA; CARVALHO, 2007).

Devido a sua eficiência na classificação de dados de alta dimensionalidade, na literatura a técnica é reportada como altamente robusta, muitas vezes equiparada às Redes Neurais (SUNG; MUKKAMALA, 2003, DING; DUBCHAK, 2001 e BRAGA; CARVALHO, 2000). E ultimamente, vem sendo amplamente utilizada em trabalhos de mapeamento de classes e atributos de solos, como os de BRUNGARD *et al*. (2015), ZHU, *et al*. (2021), ZHAI, *et al*. (2021), PEREIRA, *et al*. (2022), SONG, *et al*. (2022).

2.2.17 *Random Forest* (RF)

O algoritmo proposto por Breiman (2001), consiste em uma técnica de agregação de algoritmos do tipo árvore de decisão (*Decision Trees*), construídos de forma que sua estrutura seja composta de maneira aleatória. Para determinar a classe de uma instância, o método combina o resultado de várias árvores de decisão, por meio de um mecanismo de votação.

Melhorias significativas na precisão da classificação resultaram da manutenção de um conjunto de árvores e permitem que elas votem na classe mais popular. Para que aumente o conjunto de dados, são gerados vetores aleatórios que conduzem o crescimento de cada árvore no conjunto. A exemplo temos o conjunto de treinamento (BREIMAN, 1996), onde o crescimento de cada árvore é controlado por uma seleção aleatória.

O algoritmo RF vem apresentando bons desempenhos em estudos de predição e distribuição das classes e atributos de solos Dharumarajan e Hegde (2022), Lagacherie *et al*. (2020) e Machado *et al*. (2019).

2.2.18 *Ranger*

O *Ranger* é uma implementação rápida de florestas aleatórias (BREIMAN, 2001) ou particionamento recursivo, particularmente adequado para dados de alta dimensão. O *Ranger* é um algoritmo de floresta aleatória (*Random forest*), que é um tipo de modelo de aprendizado de máquina que combina vários modelos de árvore de decisão para melhorar a precisão da previsão. A floresta aleatória trabalha criando várias árvores de decisão aleatórias a partir do conjunto de treinamento e, em seguida, combinando as previsões de todas essas árvores para chegar a uma previsão final (WRIGHT; ZIEGLER, 2015; WRIGHT, 2017).

Uma das principais vantagens do *Ranger* é a sua capacidade de trabalhar com grandes conjuntos de dados. Ele é capaz de lidar com conjuntos de dados com milhões de exemplos e centenas de características e é otimizado para trabalhar com conjuntos de dados esparsos. Além

disso, ele é capaz de treinar modelos em vários núcleos de processamento simultaneamente, o que acelera o processo de treinamento em grandes conjuntos de dados.

2.2.19 *Recursive PARTitioning* (Rpart)

Do inglês "*Recursive PARTitioning*" (Rpart), este algoritmo constrói modelos de classificação ou regressão de um conjunto de dados, usando um procedimento em duas etapas: os modelos resultantes podem ser representados como árvores binárias (THERNEAU, 1997). A árvore é construída pelo seguinte processo: primeiro é encontrada a variável única que melhor divide os dados em dois grupos. Os dados são separados e então este processo é aplicado separadamente a cada subgrupo. E assim por diante recursivamente até os subgrupos atingirem um tamanho mínimo ou até que nenhuma melhoria possa ser feita. A segunda etapa do procedimento consiste na validação cruzada para aparar a árvore inteira (THERNEAU, 1997).

A modelagem baseada em árvores é uma técnica exploratória para descobrir estrutura em dados. Especificamente, a técnica é útil para problemas de classificação e regressão onde se tem um conjunto de variáveis de classificação ou preditor e uma variável de resposta única, no caso a classe ou atributo de solo (CLAR; PREGIBON, 2017).

2.2.20 C5.0

Este modelo é uma versão mais avançada do modelo de classificação C4.5 de Quinlan (1992), que possui recursos adicionais, como aumento e custos desiguais para diferentes tipos de erros (KUHN, 2013). Os detalhes das extensões são em grande parte não documentadas. O modelo pode assumir a forma de uma árvore de decisão completa ou uma coleção de regras (versões aprimoradas). Ao usar o método, os fatores e outras classes são preservados (as variáveis fictícias não são criadas automaticamente). Este modelo em particular lida com dados não numéricos de alguns tipos (como caractere, fator e dados ordenados).

2.2.21 Redes Neurais Artificiais

Segundo Silva (2018), as RNA consistem em um modelo matemático que interliga unidades, chamadas de neurônios, e estima as correlações entre as variáveis. Cada uma das entradas, as quais representam as variáveis preditoras, será inicialmente ponderada pelos pesos sinápticos a fim de quantificar sua importância aos objetivos funcionais do neurônio, cujo propósito será então mapear o comportamento (entrada/saída) do processo.

Em seguida, o valor resultante da composição de todas as entradas já devidamente ponderadas pelos seus respectivos pesos, adicionado ainda do limiar de ativação, é repassado

como argumento da função de ativação, cujo resultado de retorno será a saída, os quais representam as variáveis preditas (SILVA, 2016).

O algoritmo RNA é um método de previsão eficaz para lidar com complexas relações não lineares entre a propriedade do solo e as covariáveis preditoras. E têm sido amplamente utilizadas para estimar atributos dos solos (MINASNY *et al*., 2004), predição e distribuição das classes de solos, como nos trabalhos de Calderano Filho *et al*. (2014), Arruda *et al*. (2016), Heung *et al*. (2016) e Chagas *et al*. (2017), Li e Wang (2019) e Shao *et al*. (2022).

2.2.22 *Naive Bayes*

Naive Bayes (NB) é um dos algoritmos de mineração de dados mais conhecidos para classificação (WU *et al*., 2008). Um termo mais descritivo para o modelo de probabilidade subjacente seria "independente modelo de recurso". Em termos simples, um classificador *Naive Bayes* assume que a presença (ou ausência) de uma característica particular de uma classe não está relacionada à presença (ou ausência) de qualquer outra característica (BHARGAVI; JYOTHI, 2009).

Dependendo da natureza precisa do modelo de probabilidade, os algoritmos *Naive Bayes* podem ser treinados de forma muito eficiente em um ambiente de aprendizado supervisionado. Em muitas aplicações práticas, a estimativa de parâmetros para modelos ingênuos de Bayes usa o método de máxima verossimilhança; em outras palavras, pode-se trabalhar com o modelo ingênuo de Bayes sem acreditar na probabilidade bayesiana ou usar qualquer método bayesiano (BHARGAVI; JYOTHI, 2009).

Os algoritmos *Naive Bayes* geralmente funcionam muito em situações complexas. Recentemente, uma análise cuidadosa do problema de classificação bayesiana mostrou que existem algumas razões teóricas para a eficácia aparentemente irracional dos algoritmos *Naive Bayes* (IAN; EIBE; MARK, 2011). Uma vantagem do algoritmo *Naive Bayes* é que ele requer uma pequena quantidade de dados de treinamento para estimar os parâmetros necessários para a classificação. Como as variáveis independentes são assumidas, apenas as variâncias das variáveis para cada classe precisam ser determinadas e não toda a matriz de covariância (BHARGAVI; JYOTHI, 2009).

2.2.23 *Overfitting*

O conceito de *overfitting* na modelagem pode ser entendido quando o algoritmo de aprendizado de máquina modela dados de treinamento muito bem, mas não é capaz de repetir a mesma precisão no conjunto de dados de teste (SINGH *et al*., 2021). Durante a etapa de

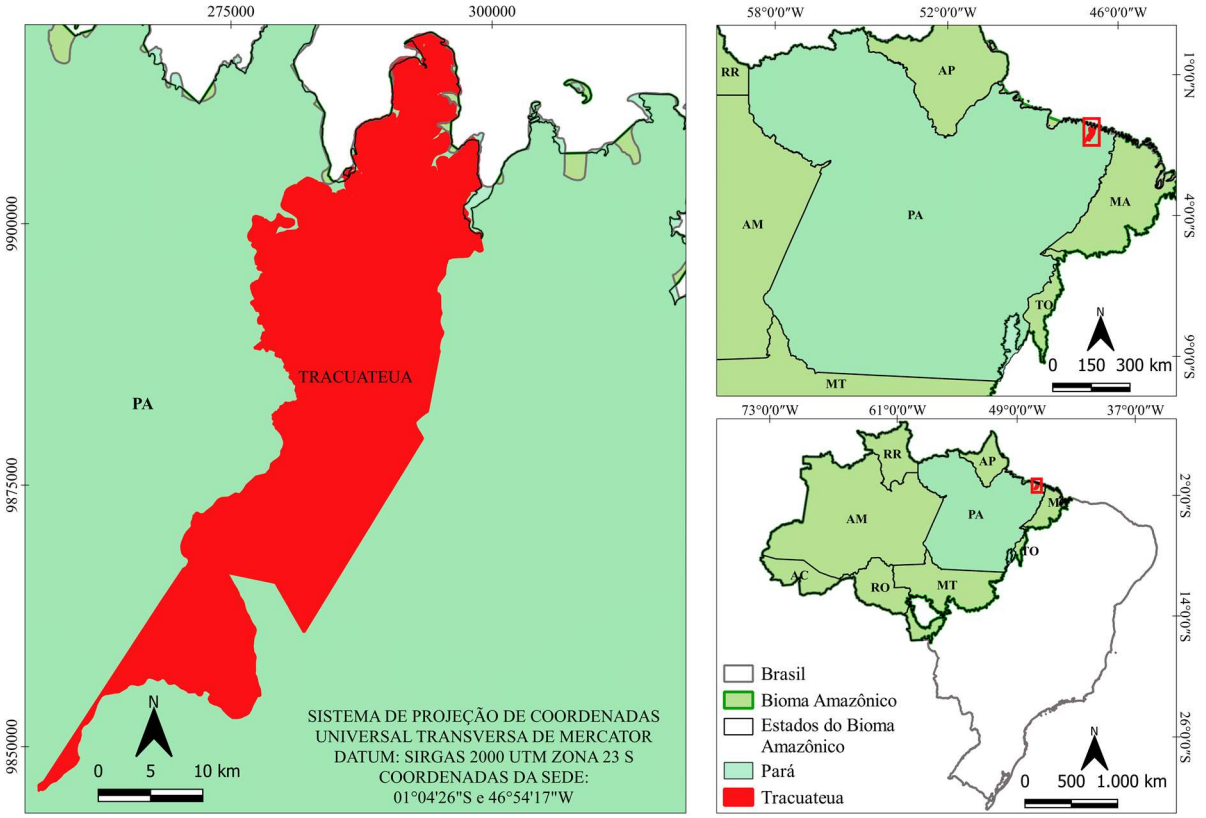
treinamento do conjunto de dados, algumas vezes, o modelo aprende o ruído e a flutuação presentes nos dados de treinamento e tenta aplicá-los em dados de teste não vistos, o que afeta negativamente a performance do modelo. Logo, o *overfitting* permite que o ruído e a flutuação aleatória nos dados de treinamento sejam apreendidos como um conceito e adotados no estágio de modelagem (NASR; SHOKRI, 2018).

3 MATERIAL E MÉTODOS

3.1 Caracterização da área de estudo

O município de Tracuateua está localizado na região nordeste do Estado do Pará, na mesorregião do nordeste paraense, microrregião do Salgado, ocupando uma área de 900,76 km^2 e coordenadas geográficas de 00º46’18” de latitude sul e 47º10’35” de longitude oeste de Greenwich (Figura 6). Possui limites ao norte com o Oceano Atlântico, a leste com o município de Bragança, a oeste com o município de Quatipuru e Capanema e ao sul com os municípios de Capanema e Primavera (OLIVEIRA JÚNIOR, *et al*., 1999).

Figura 6 - Mapa de localização de Tracuateua, Pará, Brasil.



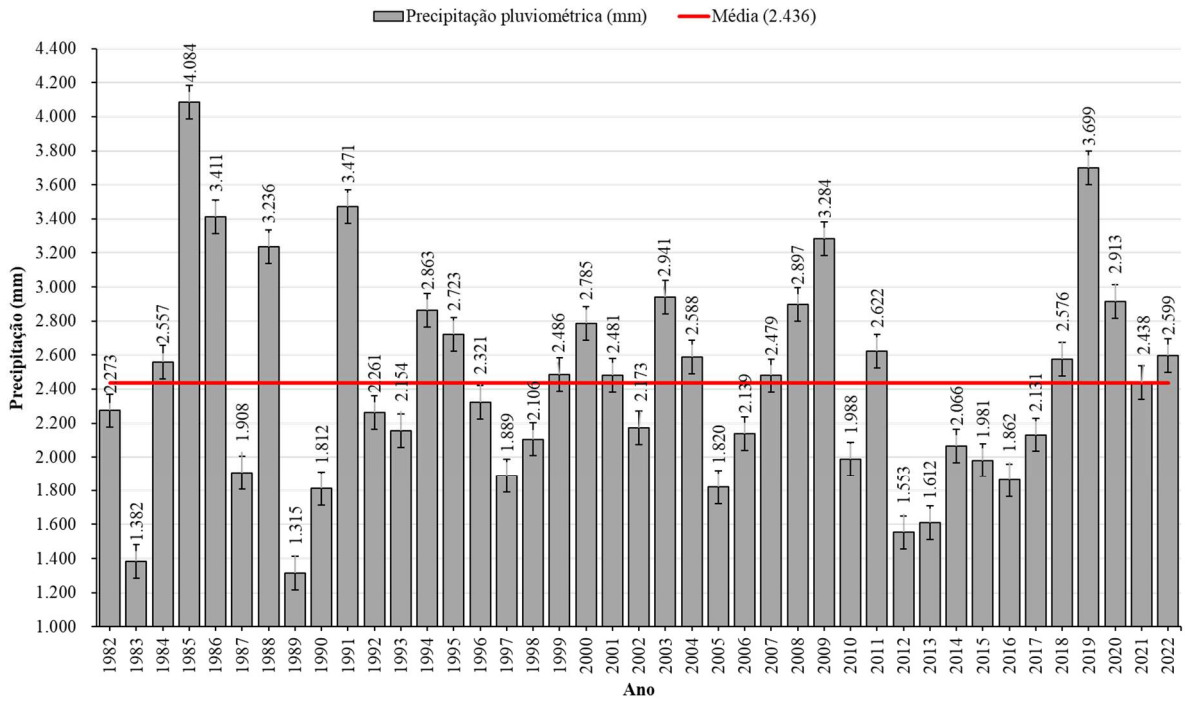
Fonte: O autor (2023).

A cobertura vegetal do município de Tracuateua está composta por seis formações bem definidas: Floresta Equatorial Subperenifólia, Floresta Equatorial Hidrófila e Higrófila de Várzea, Campos Equatoriais Higrófilos de Várzea, Formações de Praias e Dunas e Manguezal (OLIVEIRA JÚNIOR, *et al*., 1999).

O clima da região é classificado como Ami (sistema de classificação de Köppen), com precipitação média anual superior a 2.000 mm e 85% de umidade relativa do ar (ALVARES *et al*., 2013). A temperatura média durante o ano está em torno de 26 º C e a precipitação média

anual é de 2.436 mm. Segundo a série de precipitação de 1982 a 2022 (estação do Instituto Nacional de Meteorologia em Tracuateua-INMET), o regime de precipitação é caracterizado por um período chuvoso, o qual inicia em janeiro e termina em julho, com a média da precipitação máxima mensal de 520,72 mm (Figura 7).

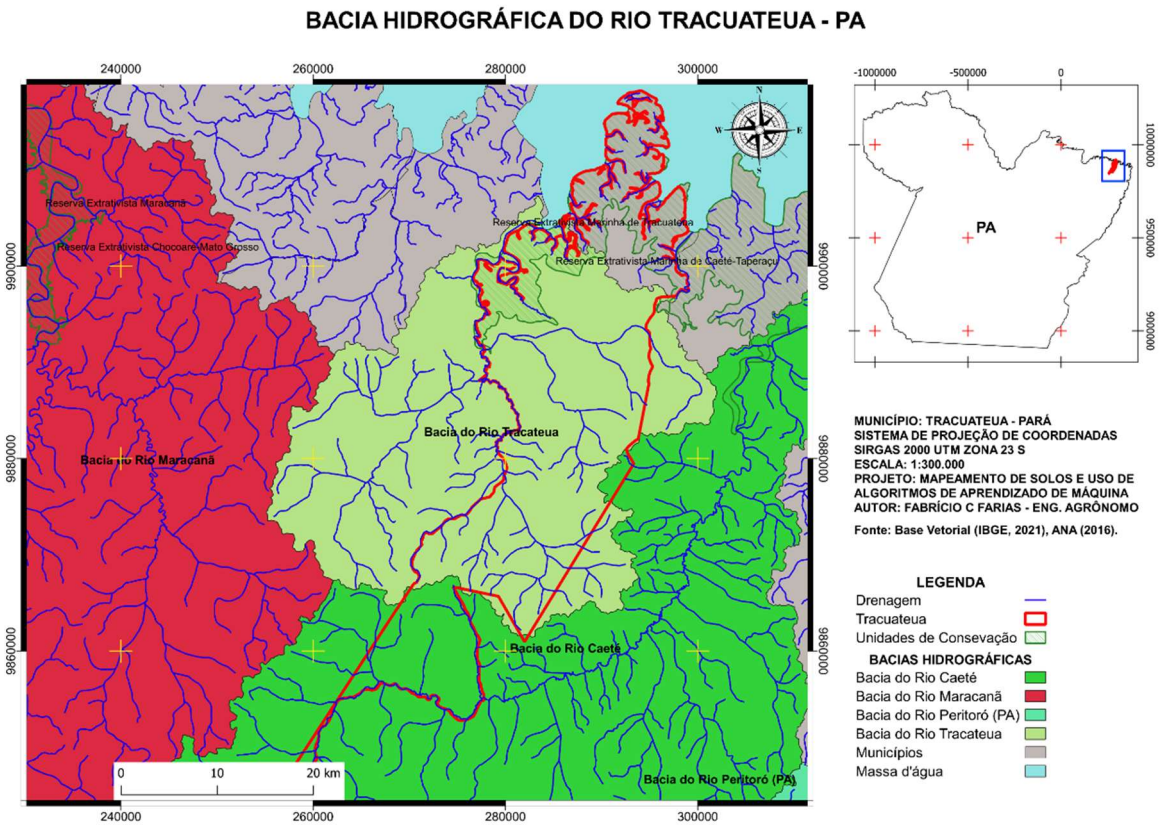
Figura 7 - Distribuição da precipitação ao longo de 40 anos em Tracuateua, Pará, Brasil.



Fonte: CPTEC/INPE (2023).

No município existem duas bacias hidrográficas, sendo a do Rio Tracuateua de maior predominância (Figura 8), compreendendo parte da Região Hidrográfica Atlântico Nordeste Ocidental, sub-região Costa Atlântica. É uma bacia de 4ª ordem, com altitudes superiores a 50 m. Possui uma área aproximada de 300 km², sendo considerada uma sub-bacia do rio Quatipuru (MARTINS *et al*., 2005).

Figura 8 - Hidrografia do município de Tracuateua, Pará, Brasil.

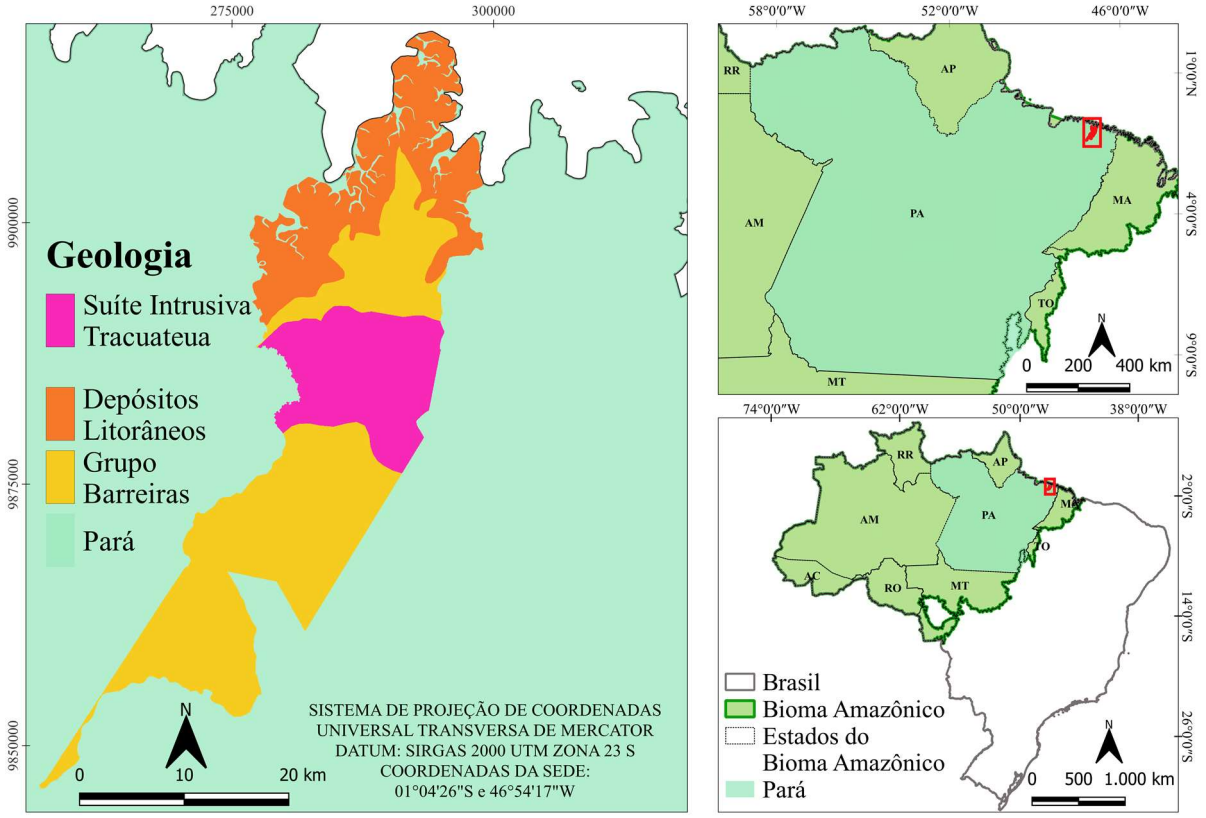


Fonte: O autor (2023).

Para caracterização da geologia local (Figura 9), tomou-se como base trabalhos existentes sobre a região de estudo (BRASIL, 1973). Logo, a geologia está representada por depósitos aluviais recentes, constituídos por cascalhos, areias e argilas inconsolidadas. Aparecem como faixa estreita e, às vezes, descontínuas, ao longo dos rios mais importantes, como o rio Japerica. Ocorre também em todo o litoral da área estudada, constituindo as praias e mangues. Nessa unidade são encontrados solos desenvolvidos desse material geológico, quais sejam: Gleissolo Háplico, Neossolo Quartzarênico Hidromórfico e Gleissolos Háplico Salino (OLIVEIRA JÚNIOR, *et al*., 1999).

A Formação Barreiras, é constituída por sedimentos clásticos, mal selecionados, variando de siltitos a conglomerados. As cores predominantes são o amarelo e o vermelho, porém variam muito de local. Os arenitos, em geral, são cauliníticos, com lentes de folhelhos. A sua sedimentação inicia-se com um calcário fossilífero, o qual em alguns locais pode não existir. Já a Formação Pirabas está bem representada em afloramentos do litoral paraense, onde são encontrados os Latossolos e os Argissolos (OLIVEIRA JÚNIOR, *et al*., 1999).

Figura 9 - Caracterização geológica de Tracuateua, Pará, Brasil.

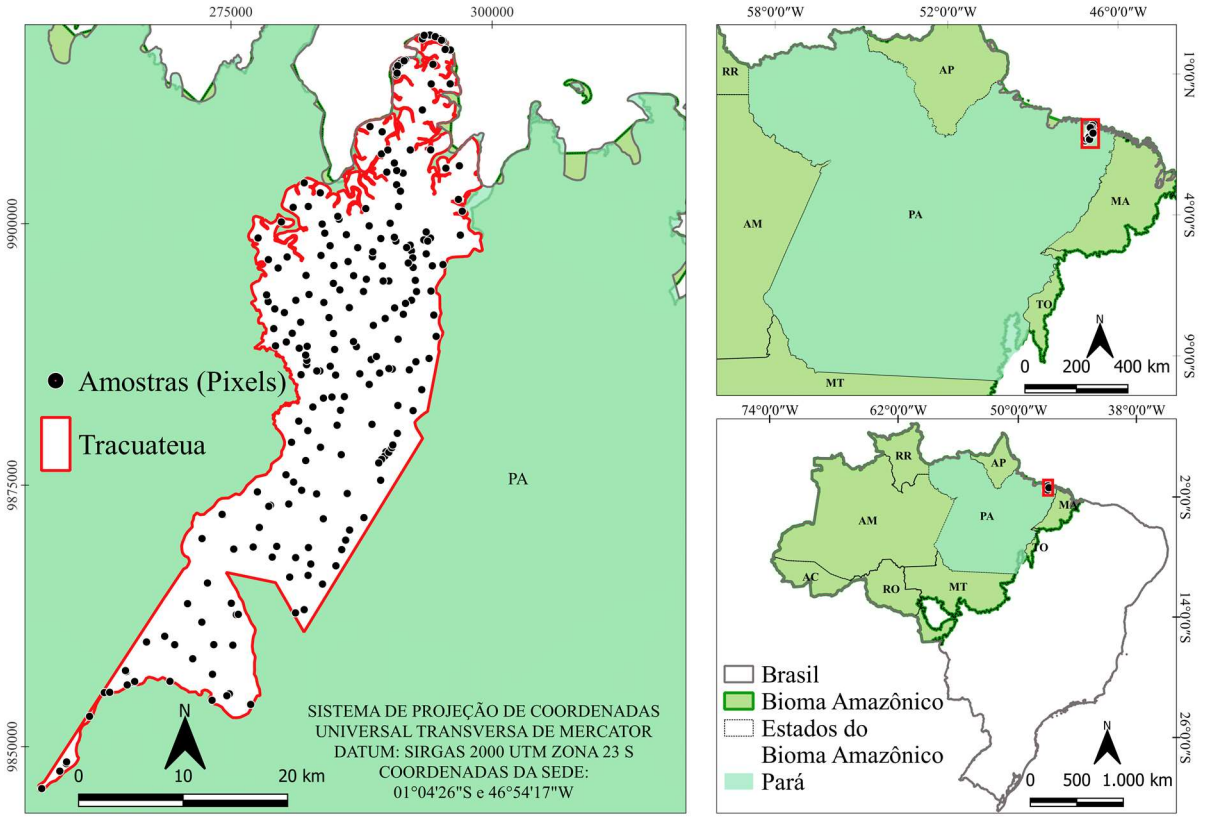


Fonte: Adaptado do IBGE (2021).

3.2 Composição da unidade de mapeamento

Foram selecionados 244 pixels, representativos das unidades de mapeamento do mapa de solos do município, para alimentação do banco de dados e para o treinamento da máquina (Figura 10). Foram realizadas tradagens dentro dos geoambientes, estratificados de acordo com possibilidade de acesso e sempre evitando zonas de transição entre as unidades de mapeamento.

Figura 10 - Distribuição dos pontos amostrais na área de estudo, Tracuateua, Pará, Brasil.



Fonte: O autor (2023).

Para definição das unidades de mapeamento (UM) foram utilizados os dados legados do levantamento realizado por CPRM (1999), na escala 1:100.000, disponíveis para a área de estudo. Onde foram realizadas prospecção em campo e análise de padrões baseados nos paradigmas solo-paisagem de Tracuateua.

Foram realizadas expedições de campo e tradagens entre as profundidades de 0 até 1,20 m, nos centros de cada classe taxonômica tiveram o objetivo de confirmar os dados legados e evitar a zona de transição entre unidades de mapeamento (Figura 11).

Figura 11 - Expedições de campo para coleta de dados em Tracuateua, Pará, Brasil.

Fonte: O autor (2023).

3.3 Covariáveis usadas para mapeamento pedológico digital

Um conjunto de 43 covariáveis relacionadas aos fatores de formação do solo (SCORPAN) foram utilizadas no mapeamento podológico digital em Tracuateua/PA. Os dados foram obtidos das seguintes bases de dados disponíveis gratuitamente: Imagens orbitais, com resolução espacial de 30 m, obtidas de plataforma USGS (2023); Mapa de solos de Tracuateua/PA na escala 1:100.000 (CPRM, 1998); Mapa geológico na escala 1:2500.000 (IBGE, 2021); Imagens do sensor OLI, do satélite Landsat-9 (7 bandas) com resolução espacial de 30 m; Modelo Digital de Elevação (MDE) SRTM (*Shuttle Radar Topographic Mission*) com resolução espacial original 90 m, reamostrada para 30 m. E a partir do MDE foram obtidas as covariáveis morfométricas usando SAGA GIS v.7.8.2, conforme descrito na Tabela 1.

Tabela 1 - Descrição das covariáveis utilizadas no mapeamento pedológico digital em Tracuateua, Pará, Brasil.

	COVARIÁVEL	
	ORIUNDAS DO MODELO DIGITAL DE ELEVAÇÃO-MDE	
01	*Analytical Hillshading* (AH)	Fornece visualização dos alinhamentos e da densidade textural dos elementos do relevo presentes na área.
02	*Aspect* (AS)	Cria um grid de exposição de vertentes.
03	*Channel Network Base Level* (CNBL)	Distância vertical para o nível de base da rede do canal.
04	*Closed Depressions* (CD)	Trata-se de uma área fechada que não tem saída de drenagem superficial e da qual a água escapa apenas por evaporação ou drenagem subterrânea.
05	*Convergence Index* (CI)	Calcula o índice de convergência/divergência em relação ao escoamento superficial.
06	*Cross Sectional Curvature* (CSC)	Curvatura relativa ao plano vertical das diferenças de nível.
07	*Flow Accumulation* (FA)	Áreas onde o fluxo se concentra e permite identificar os caminhos ou fluxo da água.
08	*General Curvature* (GC)	Curvatura geral.
09	*Gradient* (GD)	Corresponde ao gradiente hidráulico.
10	*Local Curvature* (LCU)	Curvatura local.
11	*Longitudinal Curvature* (LGC)	Taxa de mudança de inclinação, representa o desvio do gradiente ao longo do fluxo (é negativo se o gradiente aumentar).
12	*LS Factor* (LSF)	Representa o cálculo do fator de comprimento de declive.
13	*Mass Balance Index* (MBI)	Índice de balanço de massa.
14	*Multiresolution Index of Ridge Top Flatness* (MRRTF)	Índice de Multiresolução de topo de vale (MRRTF).
15	*Multiresolution Index of Valley Bottom Flatness* (MRVBF)	Índice de Multiresolução do Fundo do Vale (MRVBF).
16	*Plan Curvature* (PC)	Refere-se à curvatura plana do local. Serve para previsão de risco de movimento de massa.
17	*Profile Curvature* (PFC)	Formas das feições das paisagens. Descreve o segundo mecanismo de acumulação.
18	*Relative Slope Position* (RSP)	Representa a posição do declive da célula e sua posição relativa entre o vale e o cume.
19	*Slope* (SP)	Inclinação que a superfície do terreno possui em relação ao plano horizontal.
20	*Tangential Curvature* (TC)	Descreve o primeiro mecanismo de acumulação
21	*Terrain Ruggedness Index* (TRI)	Calcula a diferença dos valores de elevação a partir de uma célula central e as oito células vizinhas.
22	*Texture* (TX)	Textura de superfície de terreno.
23	*Topographic Position Index* (TPI)	Índice topográfico baseado na geomorfologia.

24	*Topographic Wetness Index* (TWI)	Relação entre a declividade local e a área de contribuição específica de montante, é uma medida relativa da disponibilidade em longo prazo de umidade do solo de um determinado local na paisagem.
25	*Total Insolation* (TI)	Cálculo de potencial de entrada de radiação solar (insolação).
26	*Upslope Curvature* (USC)	É a curvatura local média ponderada de distância na área de contribuição de uma célula com base na direção de fluxo múltiplo.
27	*Valley Depth* (VD)	A profundidade do vale refere-se à distância vertical a um nível de base da rede do canal.
28	*Vector Ruggedness Measure* (VRM)	Rugosidade de superfície.
29	*Vertical Distace to Channel Network* (VDCN)	Distância vertical para o nível de base da rede do canal.
DERIVADAS DA IMAGEM LANDSAT-9 (OLI/TIRS)		
01	Banda 2-Blue (B2)	Pode ser usada para detectar corpos d'água, como rios, lagos e oceanos. A água absorve a luz na região do azul, resultando em valores de reflectância mais baixos. Isso pode ajudar na identificação e mapeamento de corpos d'água em imagens de satélite.
02	Banda 3-Green (B3)	É muito útil para a detecção e análise da vegetação. Pode ser usada para avaliar a saúde e o vigor da vegetação, detectar áreas cobertas por vegetação e distinguir diferentes tipos de cobertura vegetal.
03	Banda 4-Red (B3)	A vegetação verde, densa e uniforme, apresenta grande absorção, ficando escura, permitindo bom contraste entre as áreas ocupadas com vegetação (ex.: solo exposto, estradas e áreas urbanas).
04	Banda 5-NIR (B4)	É sensível a diferentes características da superfície terrestre e possui várias aplicações em monitoramento ambiental, mapeamento de vegetação e estudos de uso da terra.
05	Banda 6 - SWIR1 (B6)	A radiação emitida pela superfície terrestre na faixa do infravermelho termal está diretamente relacionada à temperatura do objeto. Assim, é possível estimar a temperatura da superfície, o que é útil em estudos climáticos, hidrológicos, agrícolas e de recursos naturais.
06	Banda 7 - SWIR2 (B7)	Apresenta sensibilidade à morfologia do terreno, permitindo obter informações sobre geomorfologia, solos e geologia.
07	*Clay Minerals Ratio* (CMR)	Esta razão de banda destaca rochas hidrotermicamente alteradas contendo argila e alunita. Este índice atenua as mudanças de iluminação devido ao terreno, uma vez que é uma razão
08	*Enhanced Vegetation Index* (EVI)	É um índice otimizado projetado para aprimorar o sinal de vegetação com sensibilidade aprimorada em regiões de alta biomassa e monitoramento de vegetação.
09	*Ferrous Minerals Ratio* (FMR)	Esta relação de banda destaca minerais de rolamento de ferro. Ele usa a relação entre a banda SWIR e a banda NIR
10	*Iron Oxide Ratio* (IOR)	A razão de óxido de ferro é uma razão dos comprimentos de onda vermelho e azul. Isso faz com que áreas com forte

		alteração de ferro sejam brilhantes. A natureza da razão permite que este índice atenue as diferenças de iluminação causadas pela sombra do terreno.
11	*Normalized Difference Vegetation Index* (NDVI)	Analisa a resposta espectral das plantas nas bandas do vermelho e do infravermelho próximo, com valores possíveis variando de -1 a 1.
12	*Simple Ratio* (SR)	É um índice baseado na razão entre a banda infravermelho próximo (NIR) e a banda do vermelho (RED), produzindo um parâmetro que é altamente sensível à presença de vegetação.
13	*Soil Adjusted Vegetation Index* (SAVI)	É um índice de vegetação que tenta minimizar as influências do brilho do solo usando um fator de correção do brilho do solo.
DERIVADAS DE MAPAS (DADOS LEGADOS)		
01	Geologia (G)	Mapa Geológico do estado do Pará (Escala: 1:250.000).
02	Mapa Convencional (MC)	Mapa de solos de Tracuateua, Pará (Escala: 1:100.000).

Fonte: O autor (2023).

3.4 Seleção de covariáveis preditoras (*data mining*)

Devido à quantidade de dados para uso na modelagem foi necessário o uso de uma técnica de mineração de dados que tem por finalidade, selecionar características importantes em um conjunto de dados, pois, elimina as características menos relevantes uma a uma, permitindo que o modelo se adapte melhor ao conjunto de dados restante, evitar sobre ajustes e consequentemente erro da etapa de execução do modelo.

Para isso, utilizou-se um processo de seleção em que as covariáveis foram classificadas com base em seu grau de importância, selecionando-se as dez mais importantes para o mapeamento digital do solo. Para a seleção, utilizou-se o recurso *Recursive Feature Elimination* (RFE), componente do pacote *caret* (KUHN, 2022), que combinados com análises de correlação removeu as covariáveis com correlação igual ou superior a 95%.

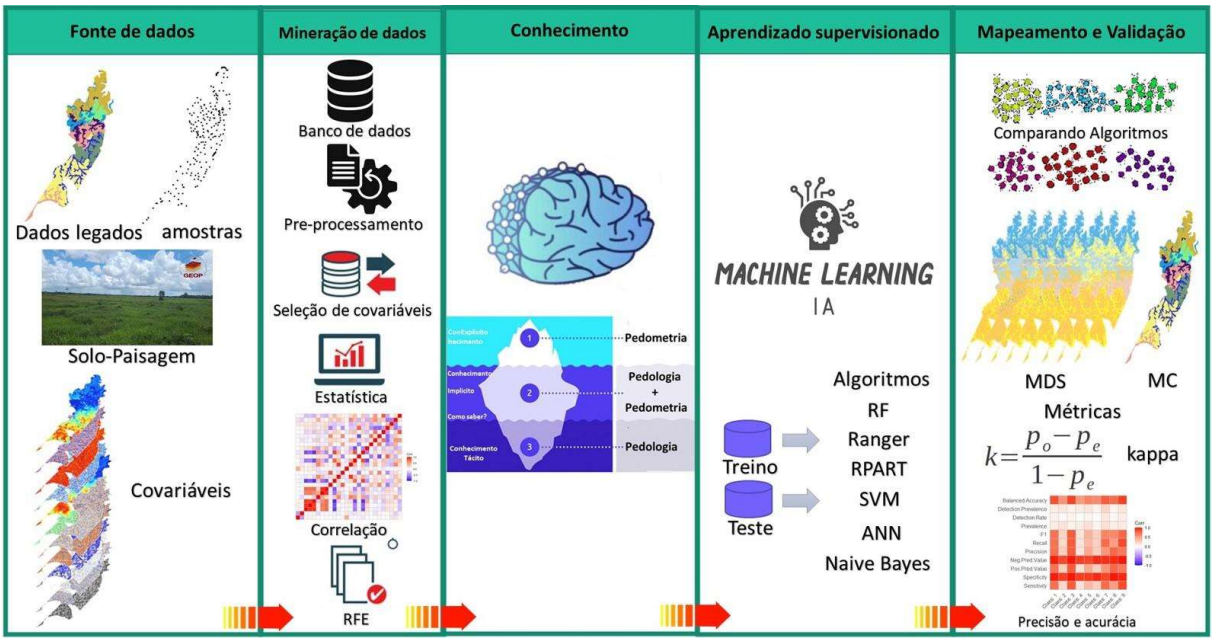
Recursive Feature Elimination é uma função do pacote *caret* (KUHN, 2022) que funciona eliminando recursivamente as características menos importantes do conjunto de dados, até que o conjunto final de características seja alcançado, essa seleção evita o reajuste de muitos modelos em cada etapa da pesquisa (KUHN; JOHNSON, 2013). O processo de eliminação foi repetido até que um número desejado de características seja alcançado ou até que todas as características sejam eliminadas. Quando o modelo completo é criado, os dados são classificados e a importância é calculada do mais para o menos importante (KUHN, 2013). Neste trabalho, utilizou-se o RFE baseado em *Random Forest* (RF), que classificou as covariáveis preditores com base no índice de Gini, conforme implementado no algoritmo *Random Forest* (BREIMAN, 2001).

A seleção foi realizada em duas etapas: a primeira, na qual foram identificadas as covariáveis altamente correlacionadas e removidas (KUHN, 2017), e a segunda, na qual valores de importância determinados pela RFE, permitindo a avaliação e seleção dos mais importantes.

3.5 Treinamento dos algoritmos de classificação

Para o estabelecimento das relações entre as variáveis preditoras e a distribuição espacial dos solos, foram utilizados oito algoritmos aprendizagem de máquina para classificação e predição, sendo um de redes neurais artificiais quatro de árvore de decisão, um de classificação probabilística que se baseia no teorema de *Bayes*, dois de vetores de suporte. Um resumo das etapas que foram adotados no mapeamento pedológico digital em Tracuateua está descrito na figura (12).

Figura 12 - Resumo esquemático do mapeamento pedológico digital em Tracuateua, Pará, Brasil.



Fonte: O autor (2023).

O treinamento dos algoritmos foi realizado implementando os pontos de amostragem (pixels) classificados com o código correspondente a ordem de solo e/ou unidade de mapeamento e os mapas das covariáveis preditoras. Sendo o conjunto de treinamento, inicialmente definido como 70% dos dados de treinamento e 30% validação. Durante a etapa de treinamento, nos pontos amostrados são extraídos os valores de cada covariável preditora utilizada, visando o estabelecimento das relações entre as covariáveis e a distribuição espacial solos da área de estudo.

Todos os algoritmos foram treinados com o mesmo conjunto de pontos amostrais classificados. Depois da etapa de treinamento, cada algoritmo foi então usado para classificar

as ordens de solos no restante da área de estudo, tomando-se como base os valores das covariáveis preditoras.

3.6 Algoritmos de aprendizado de máquina

Oito algoritmos foram avaliados: *Random Forest* (RF), *Ranger* (*Fast Implementation of Random Forests*), *Support Vector Machine polynomial kernel* (*SVMPoly*), *Support Vector Machine linear kernel* (*SVMLinear)*, *Rpart*, *Naive Bayes*, *C5.0* e o *Artificial Neural Network* (*ANN*). Os algoritmos foram implementados no pacote *CARET* (KUHN, 2013), e executados no programa R (R Team, 2022) utilizando os parâmetros disponíveis.

O princípio de funcionamento da operação de RF é um algoritmo de aprendizado de máquina que utiliza a técnica de *ensemble learning* para construir um modelo de classificação ou regressão. Ele combina várias árvores de decisão independentes e aleatórias em uma única previsão, com o objetivo de reduzir a super parametrização (*overfitting*) e melhorar a precisão do modelo.

Já o princípio de funcionamento da operação do *Ranger* é um algoritmo de floresta aleatória utilizado para tarefas de classificação e regressão, é uma técnica de aprendizado de máquina popular devido à sua eficácia em lidar com conjuntos de dados grandes e complexos, e sua capacidade de lidar com dados categóricos e numéricos. Ele também pode lidar com dados faltantes e é menos sensível a dados ruidosos em comparação com outros algoritmos de aprendizado de máquina. No entanto, o *Ranger* pode ser sensível a desequilíbrios de classe e requer ajuste de parâmetros adequados para evitar o *overfitting*.

O SVMPoly é um algoritmo de aprendizado de máquina que utiliza uma técnica de classificação não linear para separar os dados de entrada em diferentes classes. Ele é baseado no conceito de margem máxima e utiliza uma função de kernel polinomial para transformar o espaço de entrada em um espaço de dimensão superior, onde os dados podem ser mais facilmente separados.

O Rpart é um algoritmo de árvore de decisão utilizado para tarefas de classificação e regressão. É uma técnica de aprendizado de máquina popular devido à sua interpretabilidade e facilidade de implementação. Ele também pode lidar com dados categóricos e numéricos, valores faltantes e pode ser usado para tarefas de classificação e regressão.

O *Naive Bayes* é um algoritmo de classificação probabilístico que utiliza o teorema de Bayes para estimar a probabilidade de uma amostra pertencer a uma determinada classe. O *Naive Bayes* é um algoritmo de aprendizado de máquina conhecido devido à sua simplicidade

e eficiência computacional. Ele pode ser utilizado para problemas de classificação binária e multiclasse e pode lidar com dados categóricos e numéricos.

O C5.0 é uma técnica de aprendizado de máquina popular devido à sua alta precisão e interpretabilidade. Ele também pode lidar com dados categóricos e numéricos, valores faltantes e pode ser usado para tarefas de classificação e regressão. No entanto, o C5.0 pode ser sensível a dados ruidosos e requer ajuste de parâmetros adequados para evitar o *overfitting*.

Tabela 2 - Resumo dos hiperparâmetros de cada algoritmo de Machine Learning-ML, usados neste estudo em ambiente de software R.

Algoritmos	Pacote R	Hiperparâmetros
Random Forest	*randomFores*t (LIAW; WIENER, 2002)	*mtry, ntree*
Ranger	*Ranger* (WRIGHT; ZIEGLER, 2017)	*mtry, min.node.size, splitrule*
C5.0	*C50* (KUHN; QUINLAN, 2022)	*trails, rules, control (CF, minCases, earlyStopping)*
Rpart	*rpart* (THERNEAU; ATKINSON; RIPLEY, 2022)	*minspilt; cp*
SVM Polynomial	*e1071* (MEYER *et al.,* 2022)	*Default*
SVM Linear	e1071 (MEYER *et al.,* 2022)	Default
Naive Bayes	naivebayes (MAJKA, 2019)	*Default*
ANN	*neuralnet (*FRITSCH; GUENTHER; GUENTHER, 2019)*, NeuralNetTools* (BECK, 2018)	*Default*

Fonte: O autor (2023).

3.7 Avaliação da performance do treinamento e do modelo

A análise dos resultados foi realizada por meio da comparação dos mapas digitais produzidos com o mapa original (referência), pixel a pixel, utilizando-se da ferramenta *Semi-Automatic Classification Plugin* (SCP) (CONGEDO, 2021) implementada no software QGIS Desktop v.3.28.4 (2023). Após o processamento dos dados, foi determinado o índice *Kappa* (nível de confiança) e acurácia global (exatidão global) da classificação dos algoritmos, através da análise da matriz de confusão (CONGALTON; GREEN, 2019).

O índice *Kappa* (K) de Cohen (COEHN, 1960) mede a concordância real (indicada pelos elementos diagonais da matriz de confusão) menos a concordância potencial (representada pelo produto do total de linhas e colunas, excluindo entradas não reconhecidas). E indica quanto os dados classificados concordam com os dados de referência e pode ser calculado por meio da Equação (4):

$$K = \frac{n\sum_{i=1}^{C} n_{ii} - \sum_{i=1}^{C} n_{i+} + n_{+i}}{n^2 - \sum_{i=1}^{C} n_{i+} + n_{+i}} \quad (4)$$

Em que K é o índice *Kappa*; n_{ii} é o valor na linha i e na coluna i; n_{i+} é a soma da linha *i*, e n_{+i} é a soma da coluna i da matriz de confusão; *n* é o número total de amostras; e *c* é o número total de classes.

Os critérios estabelecidos para o índice *Kappa* se enquadram numa escala compreendida entre zero e um (Tabela 3). Em específico, quando os algoritmos apresentam uma pontuação igual ou superior a 0,75 isso indica um desempenho muito bom ou excelente. No entanto, quando os valores variam entre 0,4 e 0,74 isso mostra um desempenho moderado. Por último, valores inferiores a 0,4 refletem um baixo desempenho na classificação (LANDIS; KOCH, 1977).

Tabela 3 - Valores de referência para comparação do índice Kappa.

Índice *Kappa*	**Desempenho**
0	Péssimo
0-0,20	Ruim
0,21-0,40	Regular
0,41-0,60	Moderado
0,61-0,80	Muito bom
0,81-1,00	Excelente

Fonte: Landis e Koch (1977).

A acurácia global (P_o) foi calculada dividindo-se a soma do número de pixels corretamente classificados na diagonal principal (concordância real) pelo total de pixels, fornecendo a proporção de pixels corretamente classificados quando comparada com os dados de referência da área mapeada, com valores em percentagem (%), variando entre 0 até 100 e pode ser calculada conforme a Equação (5).

$$P_o = \frac{\sum_{i=1}^{m} n_{ii}}{N} \tag{5}$$

Em que, P_o é o valor de Acurácia global, *N* representa o número total de pixels contemplados pela matriz de confusão; n_{ii} representa os elementos (pixels) da diagonal principal e *m* o número de classes presentes na matriz.

Para realizar uma comparação estatística e detectar diferenças entre os índices *Kappa* obtidos pelos algoritmos, ao mapear o solo, utilizou-se o teste T, que compara pares de algoritmos conforme a Equação 6:

$$T = \frac{|K_1 - K_2|}{\sqrt{\frac{sd1^2}{n} + \frac{sd2^2}{n}}} \tag{6}$$

Em que T é o valor que define o teste T, K_1 e K_2 são os valores do índice *Kappa* dos algoritmos 1 e 2 tomados para a comparação; e "*sd*" é o desvio padrão do índice *Kappa* 1 e 2, respectivamente e *n* é o tamanho da amostra. O valor de T permite avaliar a diferença entre os algoritmos na execução do mapeamento com os dados disponíveis.

3.8 Variabilidade de unidades de mapeamento entre modelos de MDS

Utilizando a função condicional IF da calculadora raster do software QGIS, foram analisados os oito mapas gerados pelos algoritmos de *machine learning* para identificar áreas no mapa com maior variabilidade nas unidades de mapeamento (UM) de solo.

A análise de variabilidade executou um cálculo por pixel a pixel entre os mapas de solo para calcular a "Variedade" UM (ou seja, calcular o número de UMs exclusivos em cada pixel). Se os algoritmos concordarem em mapear a UM em um pixel, a função retornará o número um (UM exclusivo); para dois UMs diferentes atribuídos, ele retorna o número dois e assim por diante. Assim, quanto maior o número de discordâncias do classificador, maior a variabilidade do mapeamento.

3.9 Concordância do mapa convencional e mapas digitais de solos

A análise de concordância entre o mapa de solo convencional e os mapas de solo dos algoritmos de aprendizado de máquina foi realizada para fornecer uma medida comparativa em termos de área mapeada para cada UM. O solo no mapa convencional foi convertido em um arquivo raster. Para calcular a concordância pixel a pixel, um pixel concordante retorna o valor 1 e um não concordante retorna o valor zero.

4 RESULTADOS E DISCUSSÃO

4.1 Mapeamento pelo método convencional

O trabalho de mapeamento convencional dos solos de Tracuateua/PA, foi realizado pela Empresa Brasileira de Pesquisa Agropecuária através do Centro de Pesquisa Agropecuária do Trópico Úmido (CPATU) que atualmente é conhecido como Embrapa Amazônia Oriental, em parceria com a Companhia de Pesquisa de Recursos Minerais (CPRM) e a Prefeitura Municipal de Tracuateua/PA, como parte dos trabalhos do Programa de Integração Mineral em Municípios da Amazônia (PRIMAZ), coordenado pela CPRM.

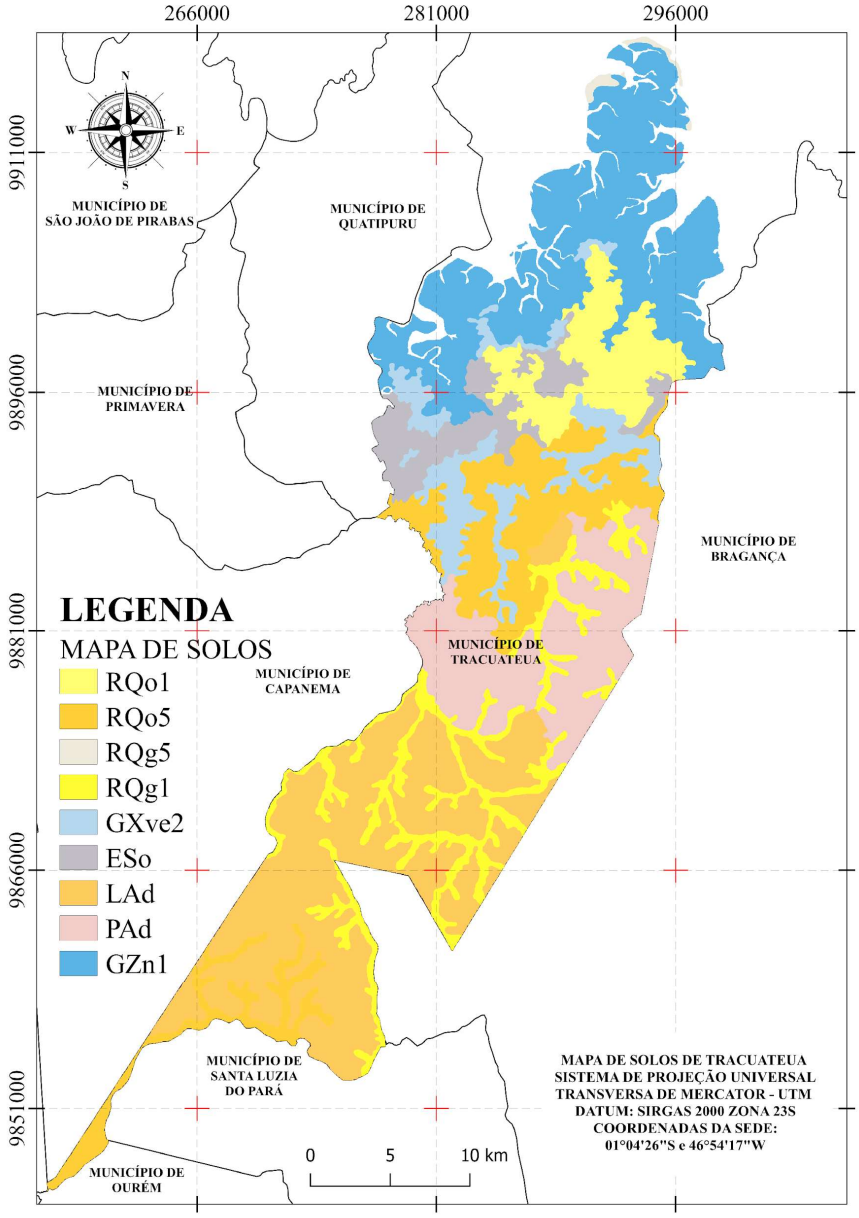
Parte do trabalho se deu por interpretações preliminares de imagens de satélite Landsat TM-5 e mosaicos semicontrolados de radar, todos com escala 1:100.000, delineando-se as unidades fisiográficas, levando-se em consideração a uniformidade de relevo, geologia, vegetação, tipos de drenagem e tonalidade.

A descrição morfológica e coleta de amostras de solos dos perfis obedeceram aos procedimentos adotados pela Embrapa Solos e constantes em Estados Unidos (1951, 1975) e Embrapa (1997, 1988a e 1988b). As cores das amostras de solos dos horizontes dos perfis foram determinadas por meio de comparação com as cores da *Munsell Soil Color Charts* (Munsell, 1954). Os solos foram classificados conforme as normas em uso pela Embrapa Solos (Embrapa, 1995 e 1999).

Os solos descritos no município de Tracuateua, nordeste paraense, foram classificados no terceiro nível categórico pelo Sistema de Classificação de Solos da Embrapa 1998, revelando predomínio das seguintes classes: LATOSSOLOS AMARELOS Distróficos (LAd), GLEISSOLOS SÁLICOS Sódicos (GZn1), NEOSSOLOS QUARTZARÊNICOS Órticos + Argissolos Amarelos (RQo5) e ARGISSOLO AMARELO Distrófico (PAd). Um total de nove UMs foram formados para fins de mapeamento do solo (Figura 13).

Os LATOSSOLOS AMARELOS Distróficos (LAd) correspondem a 25,3 (%) da área de estudo (Figura 13), esses solos são encontrados em relevos que variam de plano a suave ondulado, não tendo sido observada a ocorrência de erosão intensa, principalmente, quando sob proteção da vegetação (capoeira). Independente da textura, são aproveitados agricolamente com pastagens e plantios de dendê, pimenta-do-reino, mamão, maracujá e culturas de subsistência (OLIVEIRA JUNIOR *et al*., 1999).

Figura 13 - Mapa convencional de solos de Tracuateua, Pará, Brasil.



UNIDADES DE MAPEAMENTO - UM	ÁREA ha	(%)
NEOSSOLOS QUARTZARÊNICOS Órticos – RQo1	5.818	(6,9)
NEOSSOLOS QUARTZARÊNICOS Órticos Argissolos Vermelho-Amarelos Distróficos – RQo5	10.779	(12,9)
NEOSSOLOS QUARTZARÊNICOS Hidromórficos Neossolos Quartzarênicos Órticos Gleissolos Háplicos Tb Distróficos – RQg5	327	(0,4)
NEOSSOLOS QUARTZARÊNICOS Hidromórficos – RQg1	7.417	(8,9)
GLEISSOLOS HÁPLICOS Ta Eutróficos – GXve2	5.618	(6,7)
ESPODOSSOLOS FERRILÚVICOS Órticos – ESo	4.559	(5,4)
LATOSSOLOS AMARELO Distróficos – LAd	21.226	(25,3)
ARGISSOLOS AMARELOS Distróficos – PAd	10.294	(12,3)
GLEISSOLOS SÁLICOS Sódicos – GZn1	17.733	(21,3)
Total	83.771	(100)

Fonte: Adaptado de CPRM (1998).

Já os ARGISSOLOS AMARELOS Distróficos (PAd) (12,3%), foram encontrados regionalmente em áreas com relevo plano, suave ondulado e raramente em ondulado, sob vegetação de floresta equatorial subperenifólia primária e secundária (capoeiras). De acordo com Oliveira Júnior *et al*. (1999), os fatores limitantes destes solos quanto ao uso agrícola, são principalmente, a fertilidade natural baixa e a susceptibilidade à erosão. São utilizados com pastagens, culturas de subsistência e plantações de dendê, pimenta- do-reino e fruteiras regionais.

Os GLEISSOLOS SÁLICOS Sódicos (GZn1), normalmente são de origem aluvial, possuem boas características físicas que permitem a sua recuperação com lavagem e manejos apropriados. Estes solos quase sempre se apresentam floculados devido ao excesso de sais e ausência de quantidades significativas de sódio trocável. Em consequência, a permeabilidade é maior ou igual a dos solos normais (OLIVEIRA JUNIOR *et al*., 1999).

Do ponto de vista do uso e manejo, esses solos apresentam limitações tanto no comportamento físico por excesso de água como pelos elevados teores com saturação por sódio e sódio solúveis. O uso desses solos é limitado ao manejo e à preservação do meio ambiente e estão sob vegetação constituída dominantemente de mangue. A natureza é oriunda da deposição de material holocênico com influência marcante das águas do mar subsistência (OLIVEIRA JUNIOR *et al*., 1999).

Os NEOSSOLOS QUARTZARÊNICOS Órticos (RQo5), são solos de textura arenosa (classes texturais areia e areia-franca), essencialmente quartzosos, excessivamente drenados, praticamente sem estrutura, com ausência de materiais primários menos resistentes ao intemperismo. As Areias correspondem a 12,9 (%) ocorrem na área e apresentam semelhança de cor com os Argissolos Amarelos de textura arenosa/média, por isso, foram classificadas como Areias Quartzosas podzólicas. Ocorrem em contato com o Argissolo Amarelo, em área plana sob vegetação de floresta equatorial subperenifólia (OLIVEIRA JUNIOR *et al*., 1999).

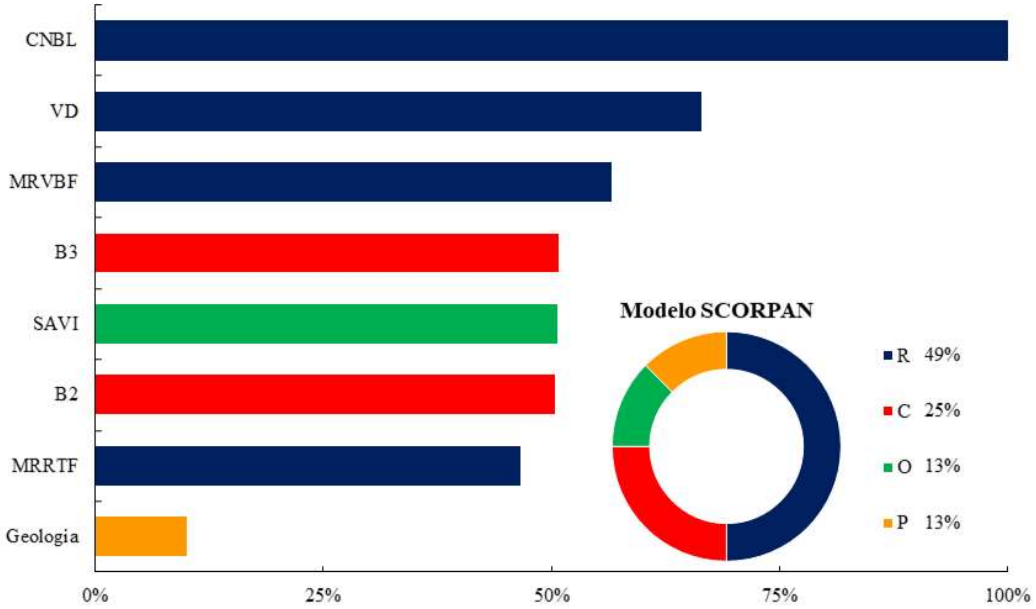
4.2 Covariáveis selecionadas no mapeamento pedológico digital

O mapeamento digital de solo no município de Tracuateua/PA contou com a seleção de 14 covariáveis, classificadas com base em seu grau de importância, entre elas: seis covariáveis morfométricas (*Channel Network Base Level* - CNBL, *Valley Depression* - VD, *Multiresolution Index of Valley Bottom Flatness* - MRVBF, *Multiresolution Ridge Top Flatness* - MRRTF, *Relative Slope Position* - RSP e *Profile Curvature* - PFC); dois índice de vegetação (*Soil Adjusted Vegetation Index* - SAVI e *Iron Oxide Ratio* - IOR); cinco imagens do Landsat-9 (bandas 2, 3 4, 5 e 6) e um mapa (Geologia do local). O maior número de covariáveis

relacionadas à forma de relevo deve-se à influência do relevo no processo de formação do solo na área, onde a topografia e a forma de relevo são fatores-chave para a diversidade do solo (MEIER *et al*., 2018).

A contribuição de cada covariável para o modelo de mapeamento do solo (Figuras 14, 15, 16, 17 e 18), medida como porcentagem de acurácia decrescente, destaca as covariáveis de importância no mapeamento da distribuição espacial do solo de Tracuateua/PA.

Figura 14 - Importância das covariáveis, pelo índice de Gini, implementado no modelo (RF).

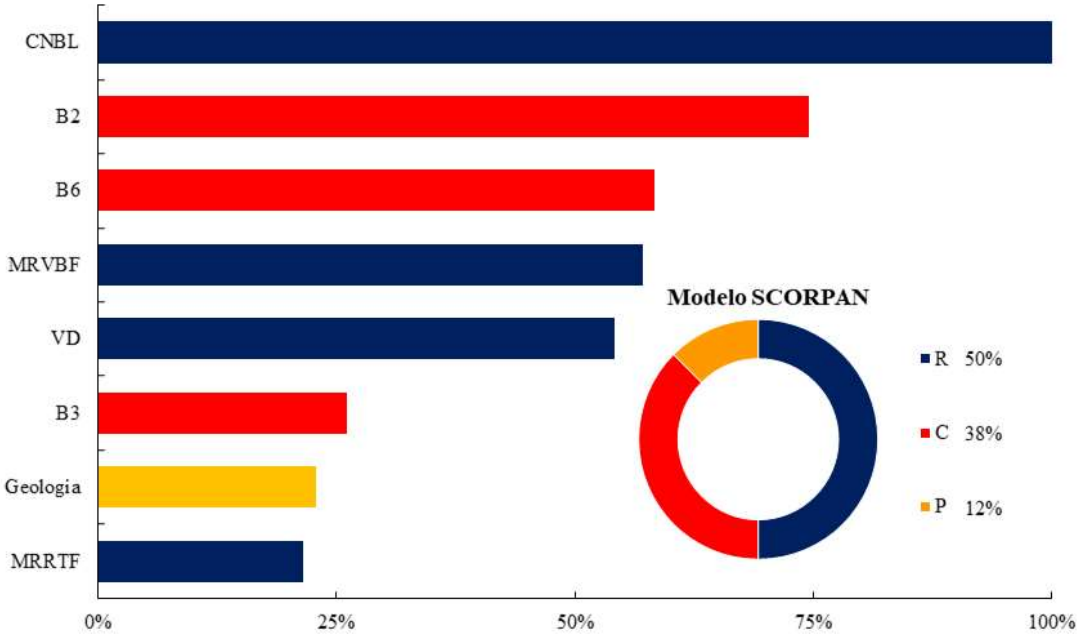


Nota: R -Relevo; C - Clima; O-Organismos; P - Material Parental; CNBL -Channel Network Base Level; VD - *Valley Depth*; MRRTF - *Multiresolution Index of Ridge Top Flatness*; MRVBF - *Multiresolution Index of Valley Bottom Flatness*; B2 - Banda b2 Landsat 9 (*Blue*); B3 - Banda b3 Landsat 9 (*Green*); SAVI - *Soil Adjusted Vegetation Index*.
Fonte: O autor (2023).

O *Channel Network Base Level* (CNBL) foi identificado como mais importante para o algoritmo RF (Figura 14) contribuindo com 100%. A variável VD contribuiu com 66%, MRVBF com 57% e SAVI correspondeu a 51% de importância. As três covariáveis oriundas de imagens do Landsat-9 com resolução espacial de 30 m, tiveram performance entre 51% e 45%. A Geologia apresentou uma importância medida em 10%, evidenciando que a variabilidade geológica da área de estudo é um dado essencial ao processo de mapeamento, seja ele convencional ou digital (Figura 14).

Para o algoritmo Rpart, o CNBL foi identificado como o mais importante contribuindo com 100%. As bandas B2 e B6 oriundas de imagens do Landsat-9, contribuíram com 75% e 58% respectivamente. MRVBF contribuiu com 57% e VD correspondeu a 54% de importância. A geologia apresentou uma importância medida em 23% (Figura 15).

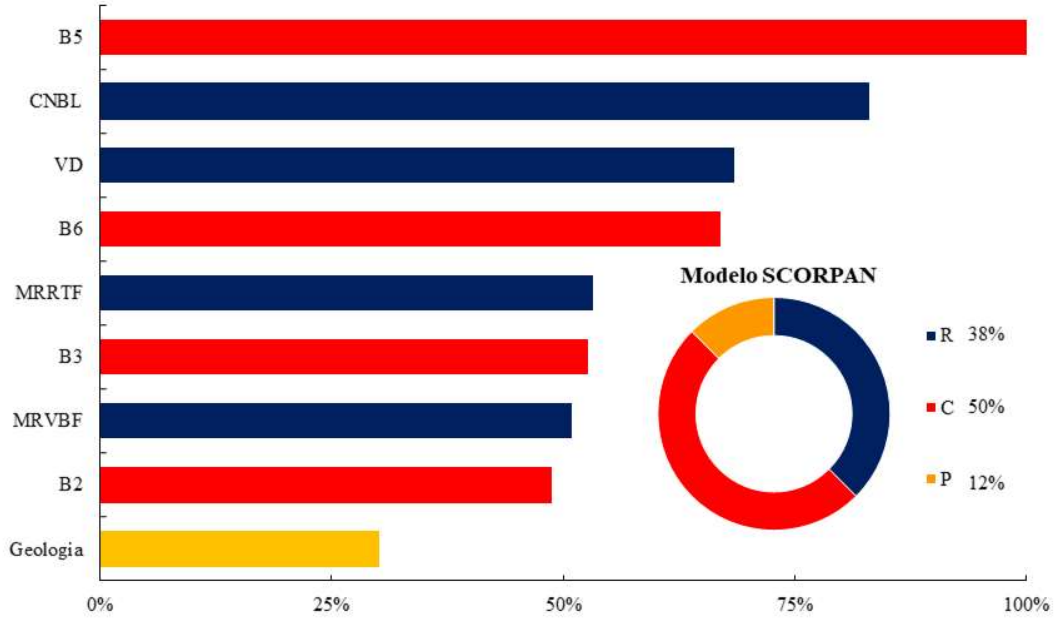
Figura 15 - Importância das covariáveis para o algoritmo Rpart.



Nota: R - Relevo; C - Clima; O-Organismos; P - Material Parental; CNBL - Channel Network Base Level; VD - Valley Depth; MRVBF - Multiresolution Index of Valley Bottom Flatness; B2 - Banda b2 Landsat 9 (Blue); B3 - Banda b3 Landsat 9 (Green); B6 - Banda b6 Landsat 9 (Infravermelho médio - SWIR1).
Fonte: O autor (2023).

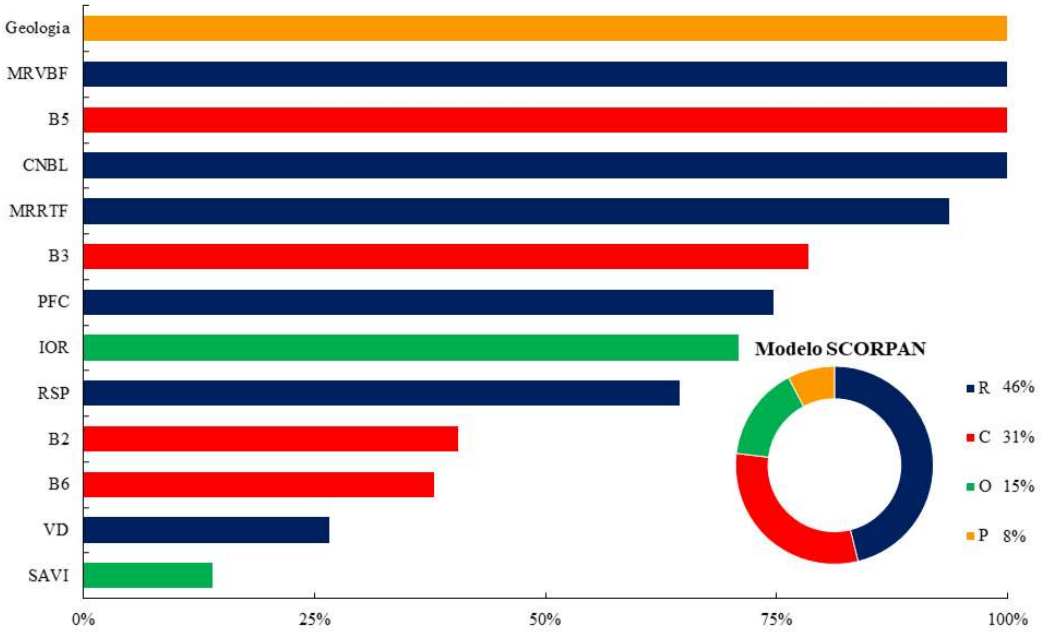
Para o algoritmo ANN, a banda B5 foi identificada como mais importante (Figura 16) contribuindo com 100 %, seguidas pelas variáveis CNBL, VD, B6 que contribuíram com 83%, 69% e 67% respectivamente. Já o algoritmo C5.0 (Figura 17), a geologia, MRVBF E B5 teve mais importância no mapeamento pedológico digital em Tracuateua, com valor de 100%.

Figura 16 - Importância das covariáveis para o algoritmo ANN.



Nota: R - Relevo; C - Clima; P - Material Parental; CNBL - *Channel Network Base Level*; VD - *Valley Depth*; MRRTF - *Multiresolution Index of Ridge Top Flatness*; MRVBF - *Multiresolution Index of Valley Bottom Flatness*; B2 - Banda b2 Landsat 9 (*Blue*); B3 - Banda b3 Landsat 9 (*Green*); B5 - Banda b5 Landsat 9 (Infravermelho próximo - NIR);B6 - Banda b6 Landsat 9 (Infravermelho médio - SWIR1).
Fonte: O autor (2023).

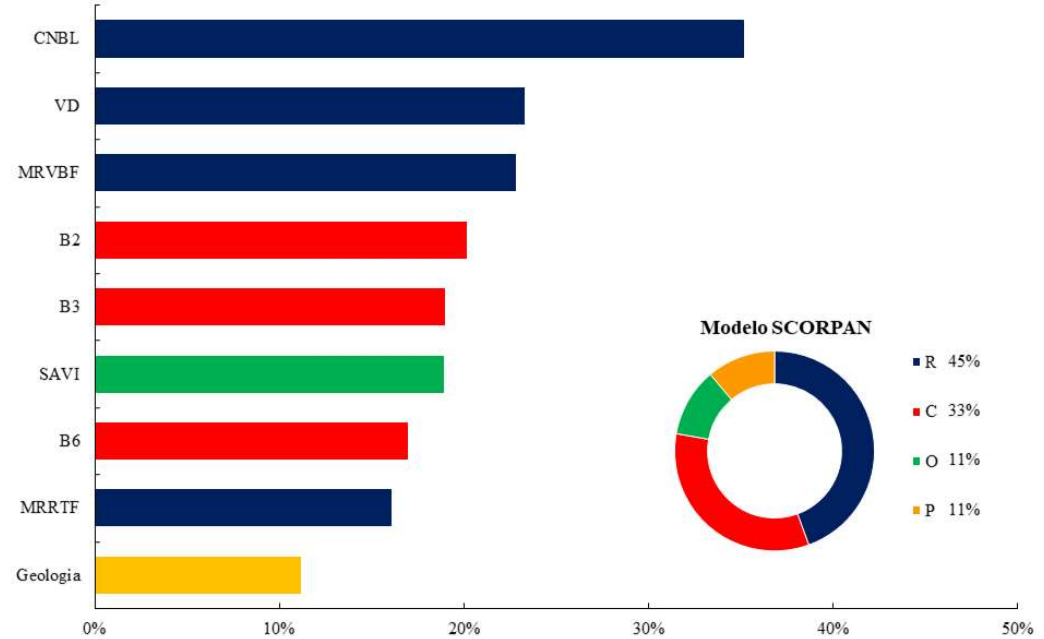
Figura 17 - Importância das covariáveis para o algoritmo C5.0.



Nota: R - Relevo; C - Clima; O-Organismos; P - Material Parental; CNBL - *Channel Network Base Level*; VD - *Valley Depth*; MRRTF - *Multiresolution Index of Ridge Top Flatness*; MRVBF - *Multiresolution Index of Valley Bottom Flatness*; B2 - Banda b2 Landsat 9 (*Blue*); B3 - Banda b3 Landsat 9 (*Green*); B5 - Banda b5 Landsat 9 (Infravermelho próximo - NIR); B6 - Banda b6 Landsat 9 (Infravermelho médio - SWIR1); SAVI - *Soil Adjusted Vegetation Index;* RSP- *Relative Slope Position;* IOR- *Iron Oxide Ratio;* PFC- *Profile Curvature.*
Fonte: O autor (2023).

Para o algoritmo *Ranger*, as variáveis relacionadas ao relevo (R) CNBL, VD, MRVBF apresentaram-se como mais importantes no mapeamento dos solos na área de estudo (Figura 18), com valores de importância de 35%, 23% e 22%, respectivamente.

Figura 18 - Importância das covariáveis para o algoritmo Ranger.



Nota: R - Relevo; C - Clima; O-Organismos; P - Material Parental; CNBL - *Channel Network Base Level*; VD - *Valley Depth*; MRRTF - *Multiresolution Index of Ridge Top Flatness*; MRVBF - *Multiresolution Index of Valley Bottom Flatness*; B2 - Banda b2 Landsat 9 (*Blue*); B3 - Banda b3 Landsat 9 (*Green*); B6 - Banda b6 Landsat 9 (Infravermelho médio - SWIR1); SAVI - *Soil Adjusted Vegetation Index.*
Fonte: O autor (2023).

Vale ressaltar que não foram gerados os gráficos de importância para os algoritmos svmP, svmL e *Naive Bayes*, pois o modelo correspondente aos mesmos, não geram essas informações. Outro ponto a observar é que as covariáveis relacionadas ao relevo (R), clima (C) e organismos (O) do modelo SCORPAN, tiveram uma expressiva importância nas etapas do MDS em Tracuateua, fato este também evidenciado no trabalho de Meier *et al*. (2018).

A geologia local esteve entre as variáveis mais importantes nos algoritmos avaliados. Em estudos, com a utilização de diferentes escalas entre o mapa legado e o mapa geológico, os autores verificaram que a geologia influencia na predição da distribuição das classes de solos (CRIVELENTI *et al*., 2009; LEMERCIER *et al*., 2012).

O uso de dados geológicos como covariável categórica, melhorou o desempenho dos classificadores baseados em pixels e indicam um caminho promissor a ser mais explorado no mapeamento digital de classes de solos (DORNIK *et al*.,2018; CREMON *et al*., 2021).

Cada área de estudo possui uma relação solo-paisagem única. Então, não há covariáveis que sejam mais importantes em todas as ocasiões, já que as classes taxonômicas do solo são influenciadas por múltiplos fatores de formação de solo ao longo de extensos períodos de tempo e em diferentes escalas (VALADARES *et al*., 2019).

O trabalho de Brungard *et al*. (2015), indica que é necessário realizar mais pesquisas sobre as relações entre covariáveis ambientais específicas e processos de formação do solo, além de mostrar que algoritmos de aprendizado de máquina podem ampliar nossa compreensão científica da distribuição e gênese.

Desta forma, o debate acerca dos nossos resultados aborda como as covariáveis mais importantes possam ter afetado a distribuição das classes de solo, tendo como referência um mapa legado de solos existente da área de estudo.

4.3 Avaliação dos algoritmos no mapeamento pedológico

Os ajustes dos modelos de classificação possuem como métrica universal o índice *Kappa* e acurácia. Dos oito algoritmos avaliados, foram selecionados os que resultaram nas melhores performance, ou seja, aqueles cujo o desempenho no procedimento de extração, mineração de dados, seleção de variáveis de treino e validação do classificador, mostraram o mínimo de acurácia (0,45) e índice *Kappa* (0,40) conforme recomendado por (LANDIS; KOCH, 1977) (Tabela 4).

Tabela 4 - Valores do índice Kappa e acurácia da etapa de modelagem (treinamento e validação) e comparação com o mapa convencional.

Métricas	**Algoritmos**							
	RF	*Ranger*	C5.0	Rpart	svmP	svmL	ANN	NB
Kappa[1]	0,60	0,71	0,63	0,46	0,47	0,60	0,14	0,55
Acurácia[1]	0,65	0,74	0,68	0,53	0,53	0,65	0,26	0,55
Comparação com o mapa convencional								
Kappa Global[2]	0,48	0,49	0,48	0,43	0,46	0,45	0,35	0,42
Acurácia Global[2]	0,56	0,56	0,56	0,52	0,54	0,53	0,48	0,50

[1] Conjunto de treino e validação; [2] Comparação com o mapa de referência; RF-*Random Forest*; NB-*Naive Bayes*; ANN-*Artificial Neural Network*; svmP-*Support Vector Machine Polynomial Kernel*; svmL - *Support Vector Machine Linear Kernel*.
Fonte: O autor (2023).

Nesta etapa do mapeamento de solo usando os algoritmos de aprendizado de máquina, exibiram um índice *Kappa* global variando de 0,35 a 0,51 e uma acurácia global variando de 0,48 a 0,56 (Tabela 4), que pode ser considerado moderado (LANDIS; KOCH, 1977).No geral, o algoritmo *Ranger* apresentou melhor desempenho, com índice *Kappa* mais alto (0,49). Este algoritmo apresentou uma baixa precisão no mapeamento de NEOSSOLOS QUARTZARÊNICOS Órticos+Argissolos Amarelos-RQo5 (33%) e alto desempenho no mapeamento de NEOSSOLOS QUARTZARÊNICOS Hidromórficos-RQg5 (99%) e GLEISSOLOS SÁLICOS Sódicos-GZn1 (81%) (Tabela 5). Desempenho semelhante foi obtido pelo algoritmo RF, cujo a baixa performance ocorreu na unidade de mapeamento RQo5 (26%) e melhor desempenho no mapeamento de RQg5 (97%) e GZn1 (82%).

Tabela 5 - Matriz de confusão do mapeamento digital de solos em Tracuateua com cada algoritmos de ML.

Random Forest											
UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**39965**	19488	0	1096	6000	5918	488	1556	6935	81446	7330
RQo5	295	**31088**	0	13392	3624	333	44956	12656	137	106481	9583
RQg5	142	52	**3578**	0	79	1270	0	1	10794	15916	1432
RQg1	113	12237	0	**40600**	473	0	47701	16824	77	118025	10622
GXve2	21353	31027	57	2376	**37417**	16781	514	3939	12911	126375	11374
ESo	1923	2685	27	31	7481	**22477**	4	233	3998	38859	3497
LAd	10	14043	0	18792	300	21	**140390**	40230	171	213957	19256
PAd	0	9310	0	6303	1403	1	1968	**38964**	164	58113	5230
GZn1	844	51	8	22	5646	3924	0	43	**163500**	174038	15663
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,62	0,26	0,97	0,49	0,60	0,44	0,59	0,34	0,82	--	--
PG (%)						0,57					
Kappa						0,48					

Ranger

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**40860**	8197	0	532	5043	5980	44	1036	7307	68999	6210
RQo5	77	**40191**	0	14464	4281	419	48964	12472	4	120872	10878
RQg5	498	20	**3619**	0	653	3032	0	0	13730	21552	1940
RQg1	5	10709	0	**36016**	76	0	46454	12180	0	105440	9490
GXve2	18593	32307	24	2553	**38545**	14601	428	4472	12310	123833	11145
ESo	2861	1288	27	0	6503	**23122**	0	3	5122	38926	3503
LAd	113	13530	0	18224	264	26	**137721**	38179	5	208062	18726
PAd	0	13739	0	10823	1791	11	2410	**46104**	0	74878	6739
GZn1	1638	0	0	0	5267	3534	0	0	**160209**	170648	15358
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,63	0,33	0,99	0,44	0,62	0,46	0,58	0,40	0,81	--	--
PG (%)						0,58					
Kappa						0,51					

SVMPoly

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**40754**	9194	0	1151	6601	7067	107	932	6581	72387	6515
RQo5	25	**41335**	0	20605	3805	367	56715	20548	7	143407	12907
RQg5	529	121	**3486**	5	2217	2494	13	29	15775	24669	2220
RQg1	9	9751	0	**34107**	62	0	46086	11061	0	101076	9097
GXve2	17239	34825	153	4220	**36788**	17813	1434	7913	11498	131883	11869
ESo	3785	1555	31	132	6303	**18426**	168	474	9308	40182	3616
LAd	23	9778	0	14397	53	20	**126924**	29803	5	181003	16290
PAd	0	13403	0	6923	1470	18	2403	**41968**	0	66185	5957
GZn1	2281	19	0	1072	5124	4520	2171	1718	**155513**	172418	15518
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,63	0,34	0,95	0,41	0,59	0,36	0,54	0,37	0,78		--
PG (%)						0,55					
Kappa						0,46					

SVMLinear

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**41348**	9598	0	1341	7298	7301	90	1028	6397	74401	6696
RQo5	10	**41168**	0	25383	3739	344	55454	29924	3	156025	14042
RQg5	567	275	**3498**	4	2479	3008	0	0	16331	26162	2355
RQg1	0	10377	0	**31897**	17	0	45562	7446	0	95299	8577
GXve2	16130	34454	141	3622	**34939**	16046	1171	7050	9736	123289	11096
ESo	4219	1909	31	243	7435	**20335**	465	874	11076	46587	4193
LAd	4	10484	0	14630	93	20	**130127**	29810	0	185168	16665
PAd	0	11703	0	4606	1006	18	1599	**37106**	0	56038	5043
GZn1	2367	13	0	886	5417	3653	1553	1208	**155144**	170241	15322
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,64	0,34	0,95	0,39	0,56	0,40	0,55	0,32	0,78		--
PG (%)						0,55					
Kappa						0,45					

---------- Rpart ----------

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**36400**	38655	27	5288	9411	9374	1895	7831	27420	136301	12267
RQo5	0	**0**	0	0	0	0	0	0	0	0	0
RQg5	398	638	**3154**	0	690	3757	0	4	12216	20857	1877
RQg1	6	9888	0	**29637**	458	0	29068	11755	5	80817	7274
GXve2	26730	36597	297	1649	**42360**	34169	787	6745	12441	161775	14560
ESo	0	0	0	0	0	**0**	0	0	0	0	0
LAd	1	27685	0	40381	415	0	**203484**	58893	867	331726	29855
PAd	0	6509	0	5500	3247	198	740	**29134**	6796	52124	4691
GZn1	1110	9	192	157	5842	3227	47	84	**138942**	149610	13465
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,56	0,00	0,86	0,36	0,68	0,00	0,86	0,25	0,70	--	--
PG (%)						0,48					
Kappa						0,43					

---------- C5.0 ----------

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**39839**	10133	0	584	6610	7795	40	857	7182	73040	6574
RQo5	513	**37443**	0	13693	3439	406	46174	8971	61	110700	9963
RQg5	1160	20	**3610**	0	468	2912	0	0	14058	22228	2001
RQg1	66	12907	0	**38592**	359	0	52171	16452	0	120547	10849
GXve2	19461	31606	23	2408	**33494**	12937	422	5328	10200	115879	10429
ESo	2421	2819	35	17	9980	**24243**	1	272	7903	47691	4292
LAd	2	14103	0	16974	290	0	**135150**	33500	0	200019	18002
PAd	0	10950	0	10324	1619	0	2063	**49065**	26	74047	6664
GZn1	1183	0	2	20	6164	2432	0	1	**159257**	169059	15215
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	
AC (%)	0,62	0,31	0,98	0,47	0,54	0,48	0,57	0,43	0,80	--	--
PG (%)						0,58					
Kappa						0,48					

---------- ANN ----------

UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**0**	0	0	0	0	0	0	0	0	0	0
RQo5	685	**2105**	0	288	208	80	67	504	83	4020	362
RQg5	309	842	**2891**	484	1731	1567	1116	1076	14367	24383	2194
RQg1	0	6	0	**54**	9	0	58	5	3	135	12
GXve2	52044	54437	567	4838	**49913**	41214	2022	6373	16669	228077	20527
ESo	0	0	0	0	0	**0**	0	0	0	0	0
LAd	8386	60772	3	73276	6212	1715	**227509**	101815	3792	483480	43513
PAd	0	0	0	0	0	0	0	**0**	0	0	0
GZn1	3221	1819	209	3672	4350	6149	5249	4673	**163773**	193115	17380
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	
AC (%)	0,00	0,02	0,79	0,00	0,80	0,00	0,96	0,00	0,82	--	--
PG (%)						0,48					
Kappa						0,35					

---------- *Naive Bayes* ----------											
UM	RQo1	RQo5	RQg5	RQg1	GXve2	ESo	LAd	PAd	GZn1	Total	Área (ha)
RQo1	**36766**	22660	0	1081	7383	6157	596	2116	10272	87031	7833
RQo5	4033	**27208**	0	14572	3094	1461	45174	18812	1299	115653	10409
RQg5	569	241	**3481**	216	1067	1546	309	756	18875	27060	2435
RQg1	518	14773	1	**44983**	1036	307	60321	19592	1782	143313	12898
GXve2	19983	31909	135	2514	**37120**	18590	820	3554	15337	129962	11697
ESo	2128	2160	50	33	7534	**20020**	47	172	6914	39058	3515
LAd	148	13072	0	15146	216	40	**120541**	33294	78	182535	16428
PAd	0	7945	0	3988	1915	21	8211	**36128**	687	58895	5301
GZn1	500	13	3	79	3058	2583	2	22	**143443**	149703	13473
Total	64645	119981	3670	82612	62423	50725	236021	114446	198687	**933210**	--
Área (ha)	5818	10798	330	7435	5618	4565	21242	10300	17882	**83989**	--
AC (%)	0,57	0,23	0,95	0,54	0,59	0,39	0,51	0,32	0,72	--	--
PG (%)						0,54					
Kappa						0,42					

Nota: UM = Unidade de mapeamento; AC = Acurácia do produtor; PG = Precisão geral; RQo1-NEOSSOLOS QUARTZARÊNICOS Órticos; RQo5-NEOSSOLOS QUARTZARÊNICOS Órticos + Argissolos Vermelho-Amarelos Distróficos; RQg5-NEOSSOLOS QUARTZARÊNICOS Hidromórficos + Neossolos Quartzarênicos Órticos + Gleissolos Háplicos Tb Distróficos; RQg1-NEOSSOLOS QUARTZARÊNICOS Hidromórficos; GXve2-GLEISSOLOS HÁPLICOS Ta Eutróficos; ESo-ESPODOSSOLOS FERRILÚVICOS Órticos; LAd-LATOSSOLOS AMARELO Distróficos; PAd-ARGISSOLOS AMARELOS Distróficos; GZn1-GLEISSOLOS SÁLICOS Sódicos.
Fonte: O autor (2023).

De acordo com a tabela 5, todos os algoritmos testados não foram precisos ao mapear as UMs: NEOSSOLOS QUARTZARÊNICOS Órticos+Argissolos Amarelos (RQo5), ESPODOSSOLOS FERRILÚVICOS Órticos (ESo) e ARGISSOLO AMARELO Distrófico (PAd), com precisão variando de 0 a 50%. Destes, o algoritmo Rpart e o algoritmo o ANN, não mapearam as UM (ESo e RQo5), sendo o ANN foi menos preciso ainda, considerando os oito algoritmos testados, uma vez que classificou apenas as UM, NEOSSOLOS QUARTZARÊNICOS Hidromórficos-RQg5, GLEISSOLOS HÁPLICOS Ta Eutróficos-GXve2, LATOSSOLO AMARELO distrófico-LAd e GLEISSOLOS SÁLICOS Sódicos-GZn1.

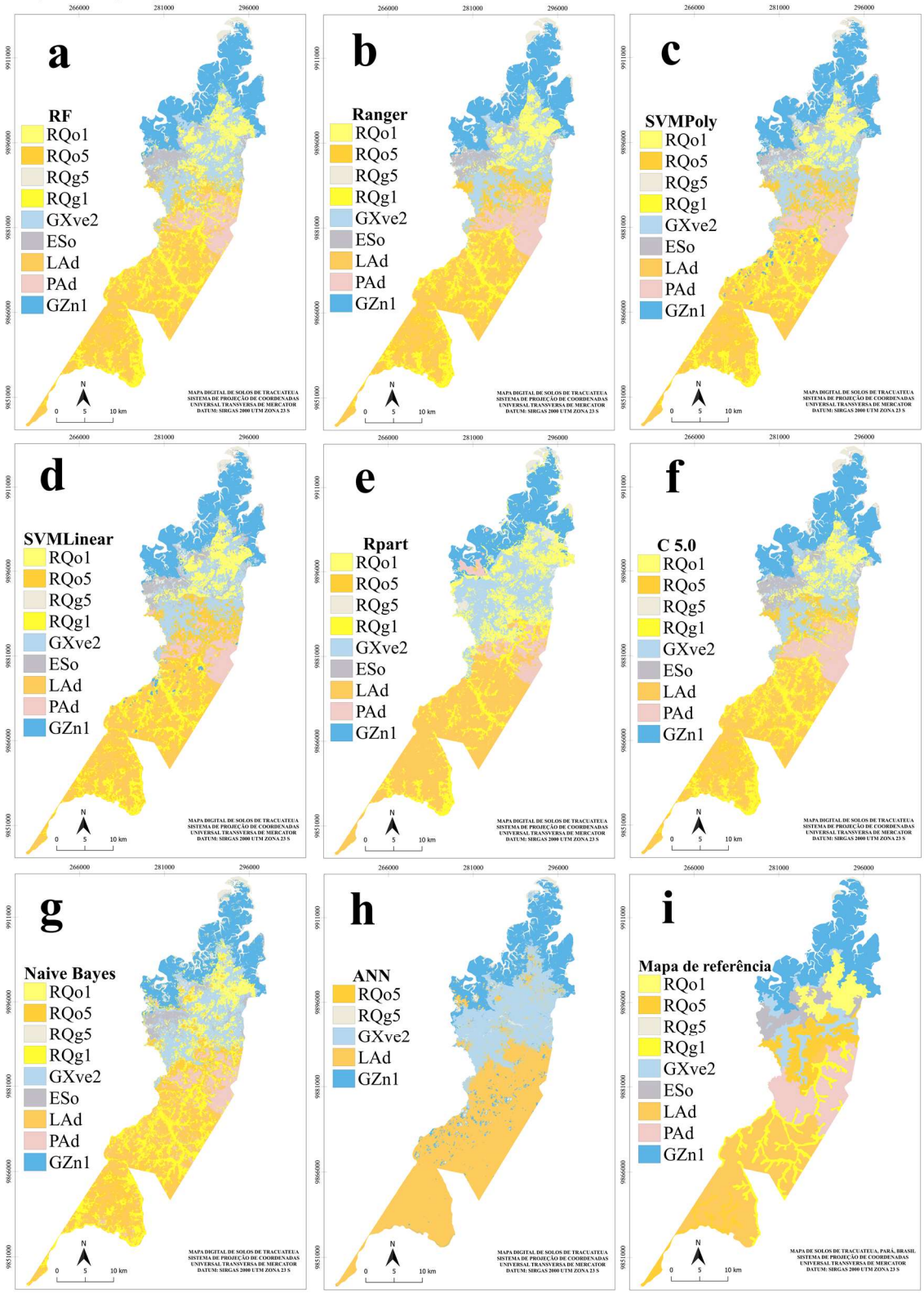
Nos estudos de mapeamento digital de solos de Giasson *et al*,. (2011); Caten *et al*., (2012), onde também foram observadas as limitações na capacidade dos algoritmos em mapear os solos em áreas menos extensas na região estudada. Logo, as unidades de mapeamento com menor representatividade espacial foram as que apresentaram menores valores de acurácia (HÖFIG; GIASSON; VENDRAME, 2014).

É importante observar que os algoritmos (RF, *Ranger*, SVMPoly e SVMLinear) testados mostraram desempenho moderado na distinção das unidades de mapeamento (RQo1, GXve2 e LAd). Por outro lado, todos os algoritmos tiveram bom desempenho no mapeamento

(RQg5 e GZn1), em que a menor acurácia foi de 70% no Rpart, e os maiores percentuais (95-99%) nos algoritmos RF, *Ranger*, SVMPoly, SVMLinear, C5.0 e *Naive Bayes*.

Os mapas gerados pelos algoritmos de aprendizado de máquina, estão representados na Figura 19. Em uma rápida análise visual não é possível observar diferenças entre os mapas digitais, com exceção do classificador (ANN), que aumentou consideravelmente a área mapeada de LAd, GXve2 e RQg5 (Figura 19h).

Figura 19 - Mapas digitais gerados pelos algoritmos: a) RF, b) Ranger, c) SVMPoly, d) SVMLinear, e) Rpart, f) C5.0, g) Naive Bayes, h) ANN e i) Mapa Convencional.



Nota: RQo1-NEOSSOLOS QUARTZARÊNICOS Órticos; RQo5-NEOSSOLOS QUARTZARÊNICOS Órticos + Argissolos Vermelho-Amarelos Distróficos; RQg5-NEOSSOLOS QUARTZARÊNICOS Hidromórficos + Neossolos Quartzarênicos Órticos + Gleissolos Háplicos Tb Distróficos; RQg1-NEOSSOLOS QUARTZARÊNICOS Hidromórficos; GXve2-GLEISSOLOS HÁPLICOS Ta Eutróficos; ESo-ESPODOSSOLOS FERRILÚVICOS Órticos; LAd-LATOSSOLOS AMARELO Distróficos; PAd-ARGISSOLOS AMARELOS Distróficos; GZn1-GLEISSOLOS SÁLICOS Sódicos.
Fonte: O autor (2023).

Tabela 6 - Área das unidades de mapeamento de solos no município de Tracuateua, nordeste paraense, mapeadas por algoritmos de aprendizado de máquina e por mapeamento de solo convencional.

	RF	*Ranger*	C5.0	Rpart	svmP	svmL	ANN	NB	DP	MC	Área
UM	---------- (ha) ----------										
RQo1	7330	6210	6574	12267	6515	6696	0	7833	3335	RQo1	5818
RQo5	9583	10878	9963	0	12907	14042	362	10409	5362	RQo5	10798
RQg5	1432	1940	2001	1877	2220	2355	2194	2435	320	RQg5	330
RQg1	10622	9490	10849	7274	9097	8577	12	12898	3858	RQg1	7435
GXve2	11374	11145	10429	14560	11869	11096	20527	11697	3342	GXve2	5618
ESo	3497	3503	4292	0	3616	4193	0	3515	1773	ESo	4565
LAd	19256	18726	18002	29855	16290	16665	43513	16428	9636	LAd	21241
PAd	5230	6739	6664	4691	5957	5043	0	5301	2136	PAd	10300
GZn1	15663	15358	15215	13465	15518	15322	17380	13473	1259	GZn1	17882
Total	83989	83989	83989	83989	83989	83989	83989	83989		Total’	3989
UM	---------- (%) ----------										
RQo1	8,73	7,39	7,83	14,61	7,76	7,97	0,00	9,33	3,97	RQo1	6,93
RQo5	11,41	12,95	11,86	0,00	15,37	16,72	0,43	12,39	6,38	RQo5	12,86
RQg5	1,71	2,31	2,38	2,23	2,64	2,80	2,61	2,90	0,38	RQg5	0,39
RQg1	12,65	11,30	12,92	8,66	10,83	10,21	0,01	15,36	4,59	RQg1	8,85
GXve2	13,54	13,27	12,42	17,34	14,13	13,21	24,44	13,93	3,98	GXve2	6,69
ESo	4,16	4,17	5,11	0,00	4,31	4,99	0,00	4,19	2,11	ESo	5,44
LAd	22,93	22,30	21,43	35,55	19,40	19,84	51,81	19,56	11,47	LAd	25,29
PAd	6,23	8,02	7,93	5,59	7,09	6,00	0,00	6,31	2,54	PAd	12,26
GZn1	18,65	18,29	18,12	16,03	18,48	18,24	20,69	16,04	1,50	GZn1	21,29
Total	100,00	100,00	100,00	100,00	100,00	100,00	100,00	100,00		Total	100,00

Nota: RF-*Random Forest*; NB-*Naive Bayes*; ANN-*Artificial Neural Network*; svmP-*Support Vector Machine Polynomial Kernel*; svmL - *Support Vector Machine Linear Kernel;* UM-Unidade de mapeamento; DP-Desvio Padrão; MC-Mapa Convencional; RQo1-NEOSSOLOS QUARTZARÊNICOS Órticos; RQo5-NEOSSOLOS QUARTZARÊNICOS Órticos + Argissolos Vermelho-Amarelos Distróficos; RQg5-NEOSSOLOS QUARTZARÊNICOS Hidromórficos + Neossolos Quartzarênicos Órticos + Gleissolos Háplicos Tb Distróficos; RQg1-NEOSSOLOS QUARTZARÊNICOS Hidromórficos; GXve2-GLEISSOLOS HÁPLICOS Ta Eutróficos; ESo-ESPODOSSOLOS FERRILÚVICOS Órticos; LAd-LATOSSOLOS AMARELO Distróficos; PAd-ARGISSOLOS AMARELOS Distróficos; GZn1-GLEISSOLOS SÁLICOS Sódicos.
Fonte: O autor (2023).

Dentre os algoritmos avaliados, o pior desempenho foi observado para o ANN, que mapeou apenas seis unidades de mapeamento, das nove descritas pelo mapa de referência. O trabalho de Silveira *et al*. (2013), demonstrou que o uso de ANN possui potencial na classificação de unidades de mapeamento de solo usando técnicas de geomorfometria, com valores de acurácia global e índice *Kappa* de 0,72 e 0,56, respectivamente. Ressaltando que a classificação ANN permite a diminuição da subjetividade na determinação dos limites entre as unidades de mapeamento.

O trabalho de Clemente; Francelino; Melo (2018) avaliou o desempenho de 21 algoritmos no mapeamento de classes de solos e obtiveram um bom desempenho no MDS no

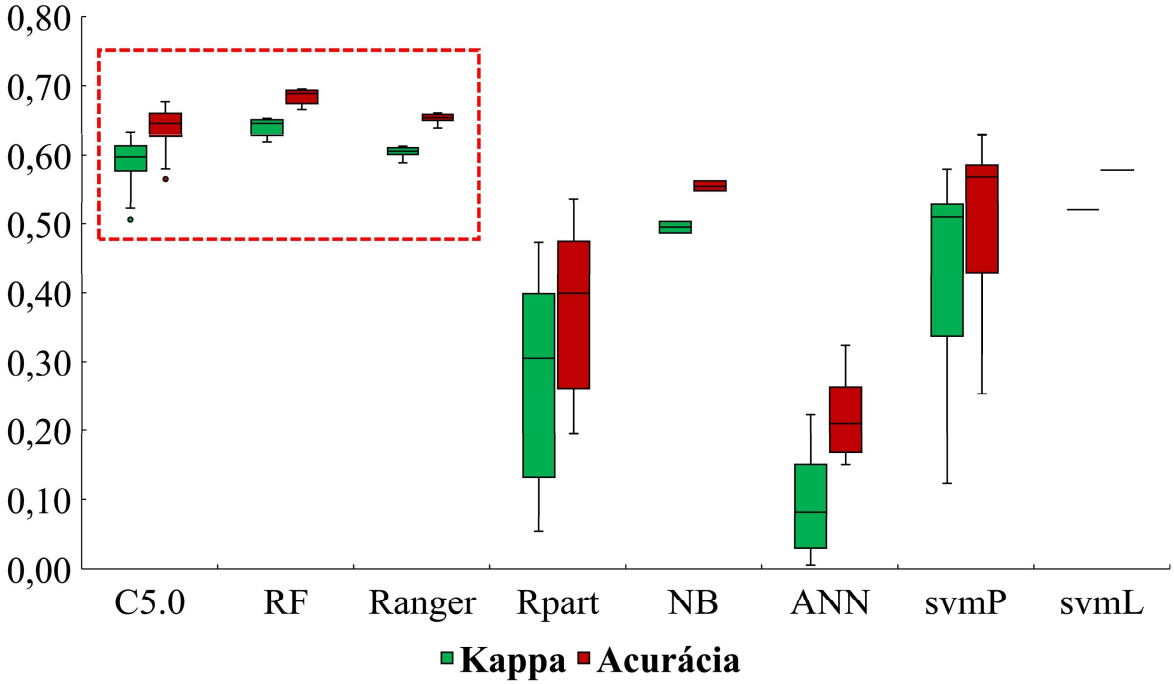
extremo norte da Amazônia com os algoritmos RF, com *Kappa* de 0,58 e acurácia de 0,69 e *Ranger* que obteve índice *Kappa* de 0,49 e acurácia 0,62. Isso evidencia que os resultados alcançados no MDS em Tracuateua, Pará, com uso dos algoritmos de aprendizado de máquina, foram semelhantes aos resultados supracitados, tendo melhores desempenhos com os algoritmos de árvore de decisão.

Outro estudo realizado por Heung *et al*., (2016) também mostrou resultados comparáveis aos obtidos nesta pesquisa ao comparar diferentes técnicas de aprendizado de máquina no mapeamento digital de solos em Vancouver, Canadá, utilizando dez algoritmos. Ao analisarem a performance dos algoritmos RF, Rpart e SVMRadial no mapeamento das classes de solos em nível de ordem, verificou-se que o método RF se destacou com um índice *Kappa* (0,63).

Valadares *et al*. (2019), mapearam solos em diversas regiões do estado de São Paulo (Brasil) e o algoritmo RF teve o melhor resultado na predição de solos do segundo ao quarto nível de classificação do Sistema Brasiliero de Classificação de Solos, com uma acurácia global de 0,78 e índice kappa de 0,67 com dados baseado em pixels; enquanto os algoritmos ANN (Redes Neurais Artificiais) e Rede Bayesiana (Redes Neurais Bayesianas) apresentaram pior desempenho, apesar dos índices Kappa de 0,57 e 0,50, respectivamente.

A Figura 20, apresenta os boxplots dos valores do índice *Kappa* e acurácia dos algoritmos testados neste estudo. É possível notar que os algoritmos de árvore de decisão (C5.0, RF e *Ranger*), apresentaram na etapa de modelagem (treino e validação) menor variabilidade e maiores valores dos índices *Kappa* e acurácia, o quadrado tracejado em vermelho indica essa distribuição. Neste estudo os algoritmos *Rpart*, SVMPoly e ANN apresentaram maior dispersão dos valores do índice *Kappa* e acurácia, sendo o ANN com valores abaixo de 0,30 conforme evidenciado na Figura 20.

Figura 20 - Distribuição dos valores do índice Kappa e acurácia dos algoritmos avaliados



Nota: RF-*Random Forest*; NB-*Naive Bayes*; ANN-*Artificial Neural Network*; svmP-*Support Vector Machine Polynomial Kernel*; svmL - *Support Vector Machine Linear Kernel.*
Fonte: O autor (2023).

No caso em questão, a linha contida dentro da caixa simboliza a mediana dos dados, a altura da caixa denota o grau de variação dos dados; quanto mais extensa for a caixa, maior será a dispersão dos dados. As hastes representam os valores mínimos e máximos dos dados e, por fim, valores atípicos em relação ao restante dos dados são representados por pontos individuais que se encontram fora das hastes, conhecidos como *outliers*.

O desempenho dos algoritmos de árvore de decisão, confirma a robustez desses algoritmos ao lidar com uma grande quantidade de dados e apresentarem resultados bons em trabalhos de mapeamento (HENGL *el at*., 2017).

Para avaliar diferenças estatísticas entre os valores do índice Kappa, dos mapas digitais gerados pelos algoritmos, utilizou-se o teste T, ao nível de significância de 5,00%. De acordo com o teste T e pelos valores apresentados na Tabela 7, houve diferenças significativas entre os algoritmos avaliados neste estudo. Quando avaliados os algoritmos de árvore de decisão (RF, Ranger, C5.0 e Rpart) o RF deferiu de todos os demais, os resultados do Ranger não diferiu do algoritmo NB, para o algoritmo C5.0 não foram identificados diferença significativa quando comparado com o desempenho dos algoritmos SVMLinear e NB, enquanto que o Rpart diferiu apenas do ANN, que dentre os oito avaliados, apresentou os piores resultados no mapeamento.

Sendo evidente, o maior desempenho dos algoritmos baseados em árvores (RF, Ranger) apresentando os melhores resultados no mapeamento dos solos em comparação com os demais modelos.

Tabela 7 - Desempenho de precisão e significância do teste T para os índices Kappa de cada classificador.

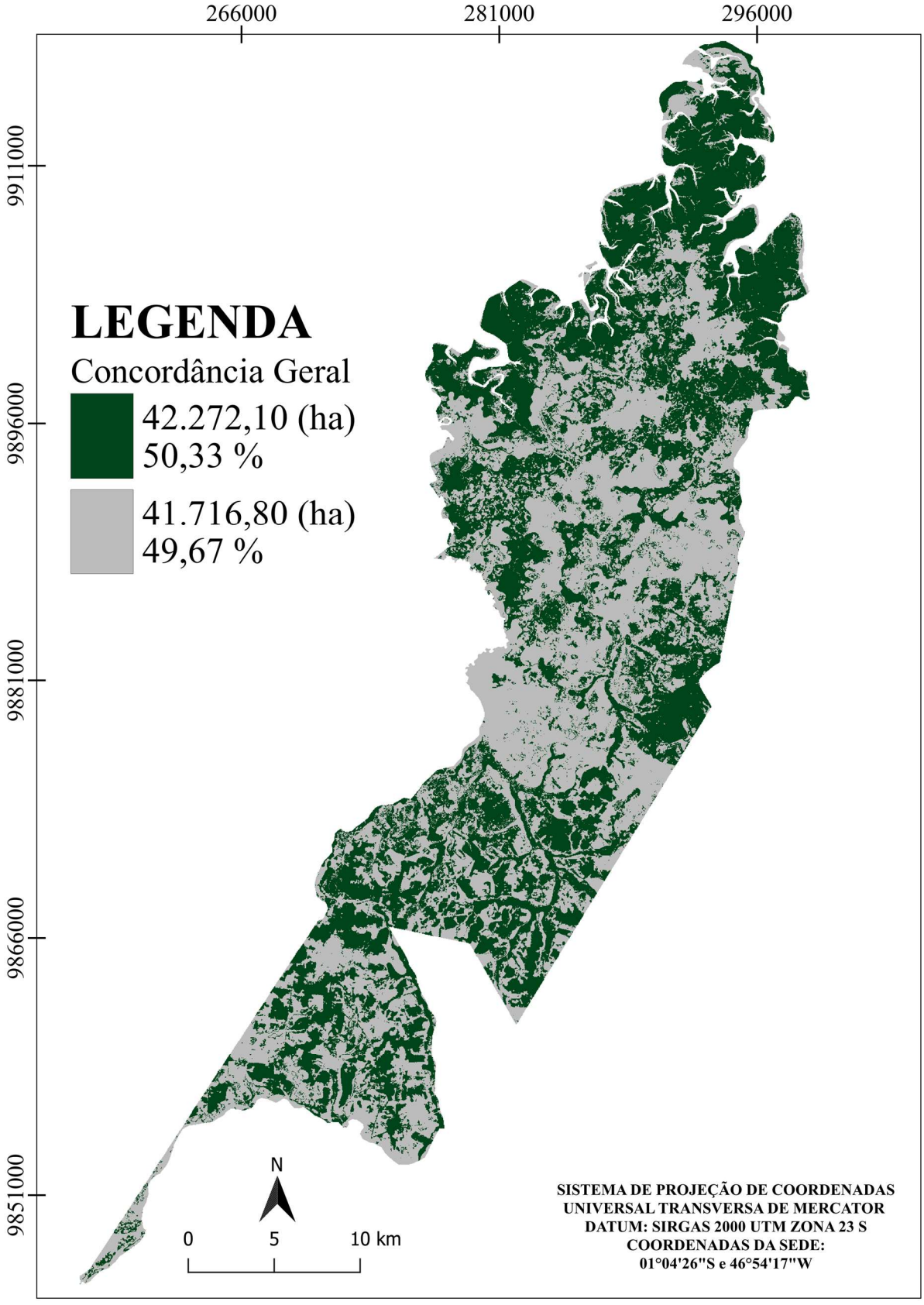
	RF	*Ranger*	C5.0	Rpart	svmP	svmL	ANN	NB
Kappa global	0,48	0,49	0,48	0,43	0,46	0,45	0,35	0,42
Acurácia global	0,56	0,56	0,56	0,52	0,54	0,53	0,48	0,50
				Teste T				
RF	-	**6,97**	**3,18**	**5,30**	**3,14**	**12,26**	**19,93**	**13,48**
Ranger		-	**1,40**	**7,04**	**3,70**	**15,57**	**26,49**	17,77^{ns}
C5.0			-	**8,71**	**4,75**	2,62^{ns}	**32,44**	3,58^{ns}
Rpart				-	-2,05^{ns}	-2,13^{ns}	**4,69**	-1,91^{ns}
svmP					-	-1,26	**10,43**	-0,67
svmL						-	**9,85**	2,97^{ns}
ANN							-	**-9,27**
NB								-

^{ns}: Diferença não significativa; Valores em negrito indicam diferenças significativas entre as médias do índice Kappa pelo teste T ($p \leq 0,05$); RF-*Random Forest*; NB-*Naive Bayes*; ANN-*Artificial Neural Network*; svmP-*Support Vector Machine Polynomial Kernel*; svmL - *Support Vector Machine Linear Kernel.*Fonte: O autor (2023).

4.4 Avaliação da variabilidade das UMs no MDS

A variabilidade das UMs dos oito mapas digitais, foram avaliadas pixel por pixel (Figura 21), e demonstram a concordância ou discordância entre os algoritmos de aprendizado de máquina no mapeamento das classes de solos em Tracuateua, Pará. Em pelo menos 50,33% da área de estudo, equivalente a 42.272,10 ha, ou seja, na maior parte da área, os oito algoritmos apresentaram concordância no mapeamento das classes de solos, indicada pela cor verde no mapa. Já as cores em cinza, indicam áreas onde todos os algoritmos divergiram na classificação no mapeamento das UM. Isso representa um total de 49,67% da área total avaliada.

Figura 21 - Mapa da variabilidade, pixel a pixel, entre os mapas digitais de solo gerados com os algoritmos de ML em Tracuateua, Pará, Brasil.



Fonte: O autor (2023).

Esses resultados indicam locais em que os algoritmos apresentaram baixo desempenho no mapeamento e necessitam de maior atenção em trabalhos futuros. Neste caso, uma das alternativas que podem ser adotadas para melhorar o mapeamento dessas regiões é a combinação de métodos de mapeamento, os chamados métodos híbridos (SILVA JÚNIOR *et al*., 2012; NOVAIS *et al*., 2021), aumento do número de amostras de treinamento e uso de dados de melhor resolução espacial (MDE, Geologia e mapas legados).

E considerando que os maiores índices de concordância foram observados em áreas correspondentes às unidades de mapeamento de maior representatividade no mapa, como os Latossolos, Gleissolos e Neossolos. Neste trabalho, constatou-se que os algoritmos testados no MDS, apresentaram boa concordância no mapeamento, com exceção do algoritmo ANN, que obteve o menor desempenho para mapeamento no município.

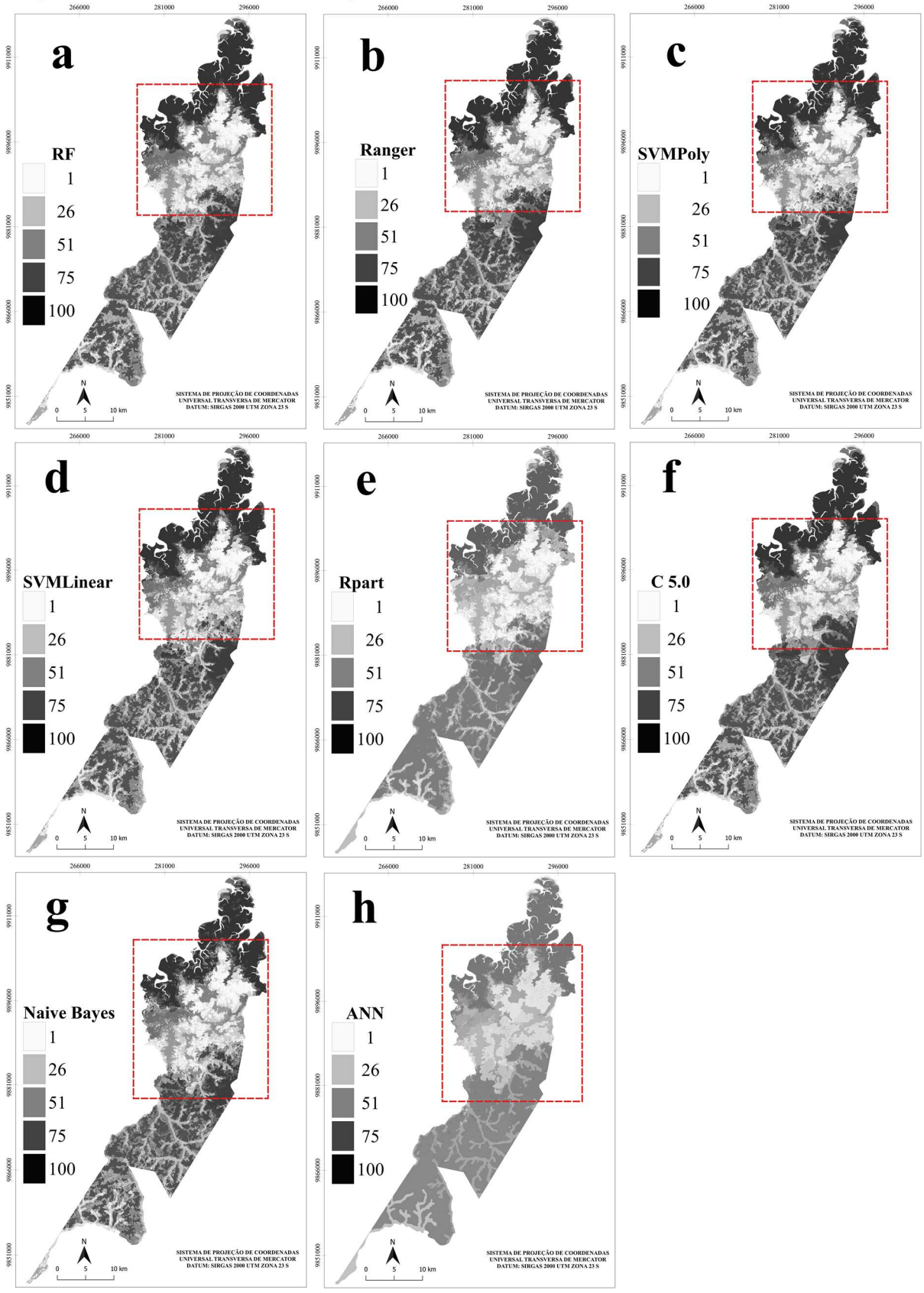
4.5 Análise da concordância do mapa

Os mapas que mostram a concordância entre o mapa de solo convencional (referência) e os mapas gerados pelos algoritmos de aprendizado de máquina estão apresentados na figura 22. A comparação evidencia boa concordância em áreas de predomínio de relevo plano - suave ondulado e áreas de mangues, com relevante influência marinha. Contudo, as maiores divergências ocorreram nas áreas de declives baixos, onde predominam áreas de campos naturais que passam a maior parte do ano inundados, com ocorrência de solos de características hidromórficas. Nessas áreas ocorre uma grande diversidade de tipos de solo, o que dificultou a classificação pelos algoritmos, como destacados nos quadros tracejados em vermelho nos mapas da Figura 22.

As características do solo nesses pedoambientes podem ser influenciadas por diversos fatores, tais como a vegetação e a topografia. Esse comportamento se deve, principalmente, à posição rebaixada e a ocorrência em depressões topográficas que favorecem a deposição de sedimentos mais finos, de origem aluviais. Esses fatores podem ter influenciado diretamente na classificação dos solos pelos algoritmos avaliados neste estudo (ROSOLEN; HERPIN, 2008) (Figura 22).

Na figura 22 é possível observar que onde a coloração é mais escura a proporção de concordância entre o mapeamento digital e o convencional foi maior, alcançando percentuais que variaram de 54% a 81%. Contudo, as cores mais claras nos mapas, apresentados na figura 22, indicam regiões onde ocorreram as maiores divergências na classificação dos algoritmos testados e o mapa de referência, com valores que variaram de 1% a 30%, e ocorreram justamente nas regiões de ocorrência de campos naturais existentes no município.

Figura 22 - Concordância/discordância entre o mapa de solos convencional e os mapas digitais de solos a) RF, b) Ranger, c) SVMPoly, d) SVMLinear, e) Rpart, f) C5.0, g) Naive Bayes e h) ANN.



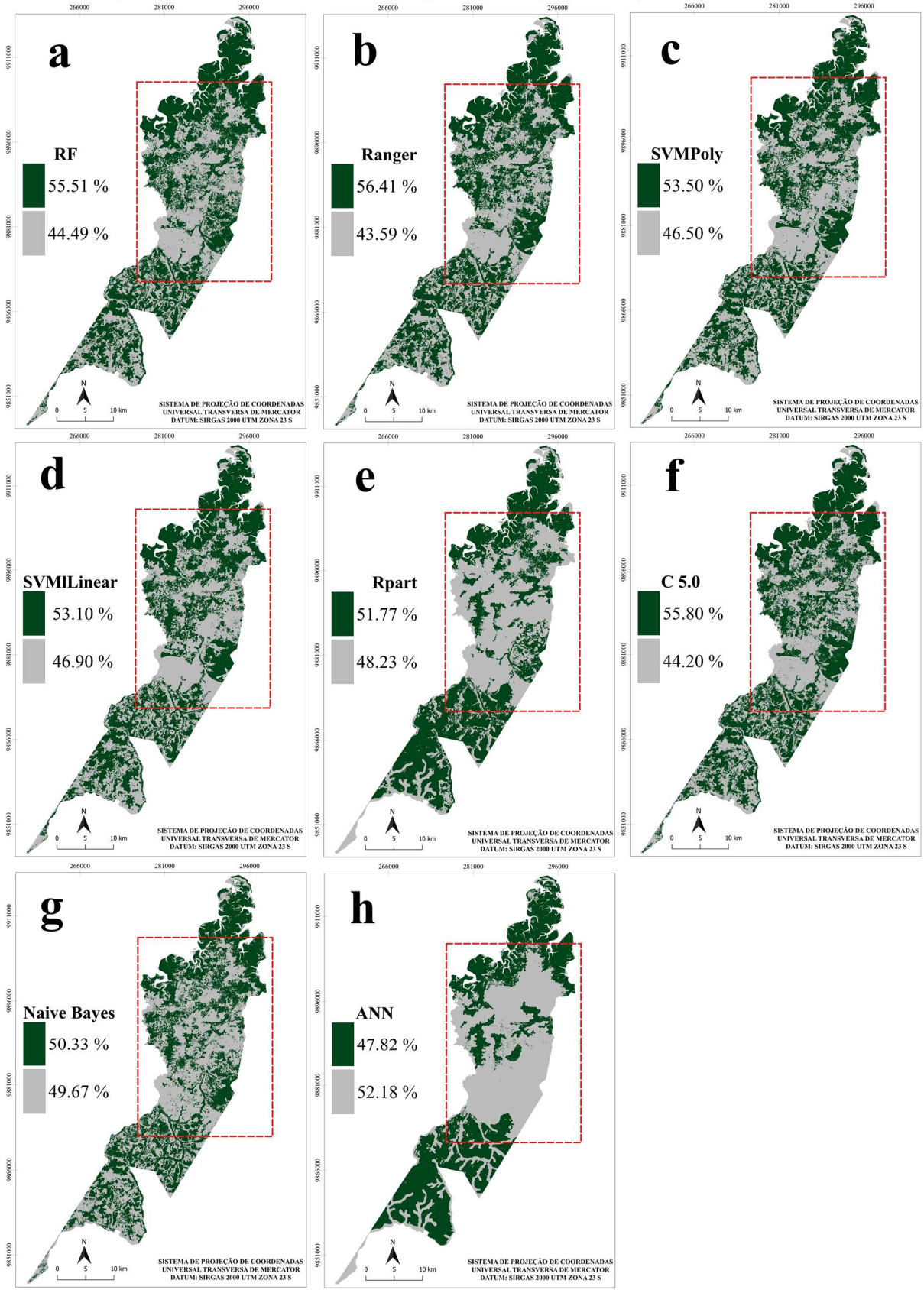
Fonte: O autor (2023).

Neste sentido, é importante observar que todos os algoritmos avaliados neste estudo, o fator geologia, após a aplicação do recurso RFE, classificou como uma variável importante no MDS em Tracuateua/PA. Contudo, a escala do mapa geológico (1:250.000) da área de estudo, usado para gerar os mapas digitais pode não ter sido a mais adequada, o que fez com que as informações geológicas contribuíssem menos do que poderiam para o mapeamento das classes de solos do município.

Ao analisar os mapas digitais gerados pelos oito algoritmos apresentados na figura 23, é possível observar em uma visão geral, que no pixel onde houve a correta classificação (cor verde), independentemente da classe de solo e tomando como base o mapa convencional, os melhores resultados foram observados no *Ranger*, C5.0 e RF, com valores de concordância global de 56,41; 55,80; 55,51, respectivamente.

Já os algoritmos ANN e Rpart, apresentaram valores de concordância global inferiores aos demais, tendo como região de maiores índices de desacordo, onde há maior variabilidade de UM, como já relatado anteriormente (Figura 23). Os quadros tracejados em cor vermelha nos mapas de concordância global destacam as regiões de maior desacordo.

Figura 23 - Erro e acerto entre o mapa de solos convencional e o mapeamento digital de solos com os algoritmos a) RF, b) Ranger, c) SVMPoly, d) SVMLinear, e) Rpart, f) C5.0, g) Naive Bayes e h) ANN.



Fonte: O autor (2023).

Os algoritmos de árvore de decisão (RF, *Ranger*, C5.0) apresentaram resultados que foram considerados moderados, conforme Landis e Koch (1977), na predição das classes de solos com valores do índice *Kappa* superiores a 0,48 e acurácia global maiores que 0,55. Os valores do índice *Kappa* apresentados indicam o nível de confiança ou concordância ao comparar dois ou mais conjuntos de dados, ou seja, comparação entre o mapa predito (digital) e o mapa de referência (mapa de solos convencional) e quanto mais próximos os valores estiverem de um, maior será a concordância entre os conjuntos avaliados.

Este estudo representa um esforço para mapear solos no bioma amazônico com o auxílio de algoritmos de *Machine Learning* e pode contribuir de maneira significativa para o PronaSolos (Programa Nacional de Solos do Brasil). Embora o programa tenha uma abrangência nacional, a Amazônia apresenta uma grande extensão geográfica com características únicas, que pode apresentar alguns desafios para os trabalhos de mapeamento de classes de solos. Estes achados constituem um importante passo no mapeamento digital de solos na Amazônia que, caracteristicamente, é uma região de difícil acesso, com uma área de floresta tropical muito extensa.

Notadamente, a área de estudo apresenta uma grande diversidade de tipos de solo, como evidenciado no mapa de solos do município, resultantes de diferentes processos de interação dos fatores de formação de solos (OLIVEIRA JÚNIOR *et al*., 1999). A diversidade de solos no município dificultou a padronização dos métodos de coleta e análise de amostras de solo, pois cada tipo de solo pode exigir abordagens específicas, principalmente nas unidades de mapeamento que correspondem às classes de difícil acesso, como os solos hidromórficos.

Para contornar esses desafios é fundamental uma abordagem integrada que envolva a cooperação entre instituições governamentais, cientistas, comunidades locais e outras partes interessadas. O investimento em tecnologias de sensoriamento remoto, o uso de novos métodos de coleta de dados em campo e o engajamento da população local podem contribuir para o sucesso do mapeamento dos solos na Amazônia, em escala que possibilitem o desenvolvimento local, com menos impacto neste que é considerado um recurso não renovável a curto prazo.

Dentre as vantagens do método digital de mapeamento pedológico está a possibilidade de se obter mapas de solos em escalas que possibilitem o planejamento dos usos dos recursos em nível local, bem como a diminuição do tempo e custos do trabalho de mapeamento. Por meio do uso de algoritmos aplicados neste estudo é possível melhorar a precisão da carta de solos. Ou ainda aplicar as técnicas para predizer unidade de mapeamento e/ou classes de solos em municípios próximos, que por sua vez possuem características semelhantes, tais como topografia e geologia.

Outro fator limitante também observado neste estudo foi a ocorrência do efeito visual conhecido como sal e pimenta, no raster de saída (mapa digital), que degrada a qualidade visual da imagem para fins de interpretações em toda a área mapeada. Nestes casos é necessário a aplicação de algumas operações de pós-processamento, como filtragens, para a eliminação desses ruídos.

Os algoritmos de árvore de decisão se mostraram eficientes no MDS. Logo, é possível usar esses modelos combinados a outros algoritmos ou métodos de mapeamento, técnica conhecida como *ensemble learning* para melhorar sua performance do mapeamento, ou seja pode ser usada para melhorar a precisão de modelos de ML no mapeamento digital de solos (BRUNGARD, 2021; TAGHIZADEH-MEHRJARDI, 2021).

5 CONCLUSÕES

Os algoritmos RF, *Ranger*, SVMPoly, SVMLinear, *Rpart*, C 5.0, *Naive Bayes* apresentaram desempenho moderado no mapeamento digital de solos até o terceiro nível categórico no município de Tracuateua, Pará. Sendo o algoritmo *Ranger* com maior desempenho tanto na etapa de modelagem quanto na etapa de comparação, mostrando viabilidade no mapeamento de solos na região bragantina, no nordeste paraense.

As covariáveis mais importantes para o MDS em Tracuateua/PA, foram: CNBL, VD, MRVBF, MRRTF, RSP, PFC, SAVI, IOR, as bandas b2, b3, b4, b5 e b6 do Landsat-9 e o mapa geológico. Destas, a maioria se refere ao relevo evidenciando a importância deste fator do modelo SCORPAN no processo de formação dos solos na área estudada.

Os algoritmos *Ranger*, RF, SVMLinear e C5.0 foram os mais eficientes na classificação de Latossolos, Gleissolos e Neossolos. E todos os algoritmos avaliados tiveram baixa performance na classificação dos solos hidromórficos que ocorrem nas extensões dos campos naturais do município.

6 REFERÊNCIAS

AKSOY, E.; PANAGOS, P.; MONTANARELLA, L. Spatial prediction of soil organic carbon of Crete by using geostatistics. *In*: **Digital soil assessments and beyond-proceedings of the Fifth Global Workshop on digital soil mapping**. 2012. p. 149-153

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

ANJOS, L. H. *et al*. Landscape and Pedogenesis of an Oxisol-Inceptisol-Ultisol Sequence in Southeastern Brazil. **Soil Science Society of America Journal**, v. 62, n. 6, p. 1651-1658, 1998.

ARRUDA, G. P. DE *et al*. Digital soil mapping using reference area and artificial neural networks. **Scientia Agricola**, v. 73, n. 3, p. 266–273, 2016.

ARRUDA, G. P. DE; DEMATTÊ, J. A. M.; CHAGAS, C. DA S. Mapeamento digital de solos por redes neurais artificiais com base na relação solo-paisagem. **Revista brasileira de ciencia do solo**, v. 37, n. 2, p. 327–338, 2013.

BARBOSA, V. A. DE F. *et al*. Heg.IA: an intelligent system to support diagnosis of Covid-19 based on blood tests. **Research on Biomedical Engineering**, v. 38, n. 1, p. 99–116, 2020.

BECK, M. W. NeuralNetTools: Visualization and analysis tools for neural networks. **Journal of statistical software**, v. 85, n. 11, p. 1-20, 2018.

BECKER, B. **Redefinindo a Amazônia: o vetor tecno-ecológico. Brasil: questões atuais da reorganização do território**. Rio de Janeiro: Bertrand Brasil, 1996. p. 223-244

BEHRENS, T. *et al*. Multi-scale digital terrain analysis and feature selection for digital soil mapping. **Geoderma**, v. 155, n. 3-4, p. 175-185, 2010.

BEHRENS, T. *et al*. Spatial modelling with Euclidean distance fields and machine learning. **European journal of soil science**, v. 69, n. 5, p. 757-770, 2018.

BENNETT, R. "Ba" as a determinant of salesforce effectiveness: an empirical assessment of the applicability of the Nonaka-Takeuchi model to the management of the selling function. **Marketing Intelligence & Planning**, v. 19, n. 3, p. 188–199, 2001.

BHARGAVI, P.; JYOTHI., S. Applying Naive Bayes Data Mining Technique for Classification of Agricultural Land Soils. **International Journal of Computer Science and Network Security**, v. 9, n. 8, p. 117–122, 2009.

BOISVERT, J. B.; DEUTSCH, C. Programs for kriging and sequential Gaussian simulation with locally varying anisotropy using non-Euclidean distances. **Computers & Geosciences**, v.37, n.4, p.495-510, 2011.

BRADY, N.C; WEIL, R.R. **Elementos da Natureza e Propriedades dos Solos**. 3. ed. Tradução técnica: Igo Fernando Lepsch. Editora Bookman, Porto Alegre, RS, 2013. 685 p.

BRAGA, A. de P.; LUDERMIR, T. B.; CARVALHO, A. C. P. L. F. Redes neurais artificiais: teoria e aplicações. 2000.

BRASIL. Ministério das Minas e Energia. Departamento Nacional da Produção Mineral. Projeto RADAMBRASIL. Folha SA. 23 São Luis e parte da folha SA. 24 Fortaleza; geologia, geomorfologia, solos, vegetação, uso potencial da terra. Rio de Janeiro: O Projeto, 1973. (Projeto RADAMBRASIL. Levantamento de Recursos Naturais, 3).

BREIMAN, L. Bagging predictors. **Machine learning**, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BROWNLEE, J. **Parametric and Nonparametric Machine Learning Algorithms**. 2016. Disponível em: https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms.

BRUNGARD, C. *et al*. Regional ensemble modeling reduces uncertainty for digital soil mapping. **Geoderma**, v. 397, n. 114998, p. 114998, 2021.

BRUNGARD, C. W. *et al*. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239–240, p. 68–83, 2015.

BU, F.; WANG, X. A smart agriculture IoT system based on deep reinforcement learning. **Future generations computer systems: FGCS**, v. 99, p. 500–507, 2019.

BUI, E. N. Soil survey as a knowledge system. **Geoderma**, v. 120, n. 1–2, p. 17–26, 2004.

BUOL, S. W. *et al*. **Soil genesis and classification**. John Wiley & Sons, 2011.

CALDERANO FILHO, B. *et al*. Artificial neural networks applied for soil class prediction in mountainous landscape of the Serra do Mar. **Revista Brasileira de Ciência do Solo**, v. 38, n. 6, p.1681-1693, 2014.CAMARGO, F. A. DE O.; ALVAREZ V., V. H.; BAVEYE, P. C. Brazilian soil science: from its inception to the future, and beyond. **Revista Brasileira de Ciência do Solo**, v. 34, n. 3, p. 589–599, 2010.

CARVALHO JUNIOR, W. de; MENDONÇA SANTOS, M. de L.; ANJOS, L. H. C. Contribuições da pedometria para a governança de solos e o Pronasolos. 2017.

CARVALHO, C. C. N. de.; NUNES, F. C.; ANTUNES, M. A. H. Histórico do levantamento de Solos no Brasil da industrialização brasileira à era da informação. **Revista Brasileira de Cartografia**, v. 65, n. 5, p. 997-1013, 2013.

CHAGAS, C. DA S. *et al*. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. **Pesquisa Agropecuária Brasileira**, v. 45, n. 5, p. 497–507, 2010.

CHAGAS, C. DA S. *et al*. Data mining methods applied to map soil units on tropical hillslopes in Rio de Janeiro, Brazil. **Geoderma Regional**, v. 9, p. 47–55, 2017.

CHEN, S. H.; JAKEMAN, A. J.; NORTON, J. P. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. **Mathematics and computers in simulation**, v. 78, n. 2–3, p. 379–400, 2008.

CHOUDHARY, R.; GIANEY, H. K. Comprehensive review on supervised machine learning algorithms. *In*: **2017 International Conference on Machine Learning and Data Science (MLDS)**. IEEE, 2017. p. 37-43.

CLARK, LINDA A.; PREGIBON, D. Tree-based models. In: **Statistical models in S**. Routledge, 2017. p. 377-419.

CLEMENTE, N.; FRANCELINO, M.; MELO, V. F. Mapeamento digital do solo como ferramenta de análise ambiental: caso de estudo na Amazônia. **Fonte**, n. 20, p. 97–105, 2018.

COELHO, F. F.; GIASSON, E. Métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. **Revista Ciência Rural**, Santa Maria, v. 40, n.10, 2010.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, v. 20, n. 1, p. 37–46, 1960.

COMPANHIA DE PESQUISA DE RECURSOS MINERAIS - CPRM. Mapa de solos de Tracuateua, Pará. Belém, 1998. 1 atlas. Escala 1:100.000.

CONGALTON, R. G.; GREEN, Kass. **Assessing the accuracy of remotely sensed data: principles and practices**. 3. ed. Boca Raton: CRC press, 2019. 346.

CONGEDO, L. 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, 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, 2015.

CORDEIRO, I. M. C. C. *et al*. **Nordeste Paraense: panorama geral e uso sustentável das florestas secundárias**. Belém: EDUFRA, 2017. p. 323

CORDEIRO, I. M. C. C.; ARBAGE, M. J. C.; SCHWARTZ, G. Nordeste do Pará: configuração atual e aspectos identitários. *In:* CORDEIRO, I. M. C. C; VASCONCELOS, L. G. T; SCHWARTZ, G**. Nordeste Paraense: panorama geral e uso sustentável das florestas secundárias**. Belém: EDUFRA. 2017. p. 323

COSTA, J. J. F. **Mapeamento digital de solos com uso de árvores de decisão na microbacia córrego Tarumãzinho, Águas Frias, SC**. Orientador: Elvio Giasson. 2016. 70 f. Dissertação (Mestrado em Ciência do Solo) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2016.

CREMON, É. 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, n. 115315, p. 115315, 2021.

CRIVELENTI, R. C. *et al*. Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo. **Pesquisa Agropecuária Brasileira**, v. 44, n. 12, p. 1707–1715, 2009.

DAI, F. *et al*. Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. **Ecological indicators**, v. 45, p. 184–194, 2014.

DALMOLIN, R. S. D.; TEN CATEN, A.; DOTTO, A. C. Pedometria: uma breve contextualização nacional e mundial. **Boletim Informativo da Sociedade Brasileira de Ciência do Solo**, v. 43, p. 18-21, 2017.

DAVENPORT, T.; PRUSAK, L. **Conhecimento empresarial. Como as organizações gerenciam o seu capital intelectual**. 8. ed. Rio de Janeiro: Campus, 2004, 237 p.

DEMATTÊ, J. A. M. *et al*. Detecção de limites de solos por dados espectrais e de relevo. **Revista Brasileira de Ciência do Solo,** v. 38, n. 3, p. 718–729, 2014.

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, 2022.

DING, C. H.; DUBCHAK, I. Multi-class protein fold recognition using support vector machines and neural networks. **Bioinformatics**, v. 17, n. 4, p. 349–358, 2001.

DING, D. *et al*. Using the double-exponential water retention equation to determine how soil pore-size distribution is linked to soil texture. **Soil and Tillage Research**, v. 156, p. 119–130, 2016.

DOKUCHAEV, V.V. 1883. Russian Chernozem. *In*: V.V. Dokuchaev. Selected Papers, 1: 14-419. (Translated into English by N. Kander – Jerusalem: Israel Program for Scientific Translations).

DORNIK, A.; DRĂGUŢ, 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.

EMBRAPA. Empresa Brasileira de Pesquisa Agropecuária. **Programa Nacional de Solos do Brasil (PronaSolos)**. Rio de Janeiro: Embrapa Solos, 2016.

FACELI, K. *et al*. **Inteligência Artificial: Uma abordagem de Aprendizado de Máquina**. Rio de Janeiro: LTC, 2011.

FATEHNIA, M.; AMIRINIA, G. A review of Genetic Programming and Artificial Neural Network applications in pile foundations. **International journal of geo-engineering**, v. 9, n. 1, 2018.

FLACH, C. W.; CORRÊA, E. A. Levantamento de solos no Brasil: métodos, práticas e dificuldades. **Geographia Meridionalis**, v. 3, n. 3, p. 420-431, 2017.

FRANCO, A. M. P. *et al*. Delineamento das unidades de mapeamento de solos utilizando o Google Earth. **Geociências**, v. 34, n. 4, p. 861-871, 2015.

FRITSCH, S.; GUENTHER, F.; GUENTHER, M. F. Package 'neuralnet'. **Training of Neural Networks**, v. 2, p. 30, 2019.

GALLANT, J.; DOWLING, T.; AUSTIN, J. **Multi-resolution Ridge Top Flatness (MrRTF)**. V. 2CSIRO, 2013.

GANASCIA, J. G. Inteligência artificial: entre o mito e a realidade. **O Correio da UNESCO**, n. 3, p. 7-9, 2018.

GARBADE, M. J. **Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences**. 2018. Disponível em: https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb.

GIASSON, E. *et al*. Decision trees for digital soil mapping on subtropical basaltic steeplands. **Scientia Agricola**, v.68, p.167-174, 2011.

GOLDSCHIMIDT, R. R. **Inteligência Computacional**. 1ª ed. Rio de Janeiro: Instituto Superior de Tecnologia - Rio, 2010. p.143.

GOMES, L. C. *et al*. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337-350, 2019.

GONÇALVES, T. G. **Mapeamento digital de solos: Predição de classes e atributos para o município de Itajubá, Minas Gerais**. Orientadora: Nívea Adriana Dias Pons. 2019. 129 f. Dissertação (Mestrado em Ciências em Meio Ambiente e Recursos Hídricos) – Universidade Federal de Itajubá, Itajubá, 2019.

GREGO, C. R.; OLIVEIRA, R. P. Conceitos Básicos da Geoestatística. In: OLIVEIRA, R. P.; GREGO, C. R.; BRANDÃO, Z. N. (Org). **Geoestatística aplicada na agricultura de precisão utilizando o VESPER**. Brasília, Embrapa, 2015.

HALEVY, A.; NORVIG, P.; PEREIRA, F. The unreasonable effectiveness of data. **IEEE intelligent systems**, v. 24, n. 2, p. 8–12, 2009.

HASTIE, T.R; TIBSHIRANI, R.; FRIEDMAN, J. Overview of supervised learning. In: **The elements of statistical learning**. Springer, New York, NY, 2009. p. 9-41

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

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

HENGL, T. *et al*. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. **Nutrient cycling in agroecosystems**, v. 109, n. 1, p. 77–102, 2017.

HEUNG, B. *et al*. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62–77, 2016.

HEUVELINK, G. B. M. *et al*. Machine learning in space and time for modelling soil organic carbon change. **European Journal of Soil Science**, v. 72, n. 4, p.1607-1623, 2021.

HÖFIG, P.; GIASSON, E.; VENDRAME, P. R. S. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. **Pesquisa Agropecuária Brasileira**, v. 49, p. 958-966, 2014.

HUDSON, B. D. The soil survey as paradigm-based science. Soil Science Society of America journal. **Soil Science Society of America**, v. 56, n. 3, p. 836–841, 1992.

IBGE, Coordenação de Recursos Naturais e Estudos Ambientais. **Manual técnico de pedologia**. 3. ed. Rio de Janeiro: IBGE, 2015. 430 p.

IBGE. **Manual técnico de pedologia: guia prático de campo**. Rio de Janeiro: IBGE, 2015. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv95015.pdf

Instituto Brasileiro de Geografia e Estatística - IBGE. **Base dados espacial, geologia, escala 1:250.000, no recorte ao milionésimo (Vetores)**. Rio de Janeiro: IBGE. 2021. https://geoftp.ibge.gov.br/informacoes_ambientais/geologia/levantamento_geologico/vetores/escala_250_mil/versao_2021/geol_area.zip

IPPOLITI R, G. A. *et al*. Análise digital do terreno: ferramenta na identificação de pedoformas em microbacia na região de" mar de morros. **Revista Brasileira de Ciência do Solo**, v. 29, n. 2, 2005.

JENNY, H. **Factors of soil formation: a system of quantitative pedology**. New York: McGraw Hill, 1941. p. 320

JENNY, H. Factors of soil formation: a system of quantitative pedology. New York: McGraw Hill, 1941. p. 320

KÄMPF, N; CURI, N. **Formação e evolução do solo (Pedogênese)**. Pedologia: fundamentos. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, p. 207-302, 2012.

KANIOURA, A.; EITEL-PORTER, R. **What is AI exacly?** 2020. Disponível em: https://www.accenture.com/hk-en/insights/artificial-intelligence/what-ai-exactly.KAPLAN, A.; HAENLEIN, M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. **Business Horizons**, v. 62, n. 1, p. 15-25, 2019.

KOHAVI, R.; PROVOST, F. Glossary of terms. Machine Learning-Special Issue on Applications of Machine Learning and the Knowledge Discovery Process. **Machine learning**, v. 30, n. 3, p. 271–274, 1998.

KRIGE, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. **Journal of the Southern African Institute of Mining and Metallurgy**, v. 52, n. 6, p. 119–139, 1951.

KUHN, M.; JOHNSON, K. Classification trees and rule-based models. *In*: **Applied Predictive Modeling**. New York, NY: Springer New York, 2013. p. 369–413.

KUHN, Max *et al*. **Applied predictive modeling**. New York: Springer, 2013.

KUHN, Max *et al*. Classification trees and rule-based models. **Applied predictive modeling**, p. 369-413, 2022.

KUHN. M. caret: Classification and Regression Training. R package Version 6.0-92, 2022.

LAGACHERIE, P. Digital soil mapping: A state of the art. *In*: **Digital Soil Mapping with Limited Data**. Dordrecht: Springer Netherlands, 2008. p. 3-14

LAGACHERIE, P. *et al*. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: A numerical experiment on Quantile Random Forest using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. **Geoderma**, v. 375, p. 114503, 2020.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159, 1977.

LARY, D. J. *et al*. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3-10, 2016.

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

LEPSCH, I. F. **Formação e conservação dos solos**. 2.ed. São Paulo: Oficina de textos, 2010. 216 p.

LEPSCH, I. F. Status of soil surveys and demand for soil series descriptions in Brazil. **Soil horizons**, v. 54, n. 2, p. 1-5, 2013.

LI, J.; WANG, J. Optimal sampling design for reclaimed land management in mining area: An improved simulated annealing approach. **Journal of cleaner production**, v. 231, p. 1059-1069, 2019.

LIAW, A; WIENER M. "Classification and Regression by randomForest.". ***R News***, **2**(3), 18-22, 2002.

LIMA, L. A. S. **Aplicação dos métodos semi-automático e lógica Fuzzy para o mapeamento de solos da bacia do Sarandi**. Orientador: Henrique Llacer Roig. 2013. 124 f. Dissertação (Mestrado em Geoprocessamento e Análise Ambiental) – Universidade de Brasília, Brasília, 2013.

LORENA, A. C.; DE CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 43–67, 2007.

LUZ, L. M. *et al*. **Atlas Geográfico Escolar do Estado do Pará**. 1. ed. Belém: GAPTA/UFPA. 2013. p. 64

MACHADO, D. F. T. *et al*. Soil type spatial prediction from Random Forest: different training datasets, transferability, accuracy and uncertainty assessment. **Scientia Agricola**, v. 76, n. 3, p. 243–254, 2019.

MAJKA, M. **naivebayes: High Performance Implementation of the Naive Bayes**. R package version 0.9.7: Algorithm in R, 2019.

MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B. **Using R for digital soil mapping**. Cham: Springer International Publishing, 2017.

MANSUY, N. *et al*. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. **Geoderma**, v. 235, p. 59-73, 2014.

MARTINS, A. A. V.; COSTA, R. A. M. DA; PEREIRA, L. C. C. Distribuição espaço-temporal da comunidade zooplanctônica de uma lagoa costeira artificial na região amazônica, Bragança, Pará, Brasil. **Boletim do Museu Paraense Emilio Goeldi. Ciencias naturais**, v. 1, n. 3, p. 103–111, 2005.

MCBRATNEY, A. B. *et al*. An overview of pedometric techniques for use in soil survey. **Geoderma**, v. 97, n. 3-4, p. 293-327, 2000.

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

MCBRATNEY, A. B.; WEBSTER, R.; BURGESS, T. M. The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I. **Computers & geosciences**, v. 7, n. 4, p. 331–334, 1981.

MEIER, M. *et al*. Digital Soil Mapping Using Machine Learning Algorithms in a Tropical Mountainous Area. **Revista Brasileira de Ciência do Solo**, v. 42, p. e0170421, 2018.

MENDONÇA-SANTOS, M. L.; DOS SANTOS, H. G. Chapter 3 the state of the art of Brazilian soil mapping and prospects for digital soil mapping. In: **Developments in Soil Science**. Amesterdã: Elsevier, 2006. p. 39–601

MENDONÇA-SANTOS, M. L.; DOS SANTOS, H. G. Mapeamento digital de classes e atributos de solos: métodos, paradigmas e novas técnicas. Rio de Janeiro: Embrapa Solos, 2003. p. 19

MENEZES, M. D. DE *et al*. Solum depth spatial prediction comparing conventional with knowledge-based digital soil mapping approaches. **Scientia agricola**, v. 71, n. 4, p. 316–323, 2014.

MEYER, D. *et al.* **e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien**. R package version, v. 1.7-13, n. 2, 2022.

MILLER, B. A. *et al*. Impact of multi-scale predictor selection for modeling soil properties. **Geoderma**, v. 239, p. 97-106, 2015.

MILLER, B. A.; SCHAETZL, R. J. The historical role of base maps in soil geography. **Geoderma**, v. 230–231, p. 329–339, 2014.

MILLER, B. A.; SCHAETZL, R. J. The historical role of base maps in soil geography. **Geoderma**, v. 230, p. 329-339, 2014.

MINASNY, B. *et al*. Neural networks prediction of soil hydraulic functions for alluvial soils using multistep outflow data. **Soil Science Society of America journal**, v. 68, n. 2, p. 417–429, 2004.

MINASNY, B.; MCBRATNEY, A. B. Spatial prediction of soil properties using EBLUP with the Matérn covariance function. **Geoderma**, v. 140, n. 4, p. 324–336, 2007.

MINELLA, J. P. G.; MERTEN, G. H. Índices topográficos aplicados à modelagem agrícola e ambiental. **Ciência Rural**, v. 42, n. 9, 2012.

MONTANARELLA, L. *et al*. World's soils are under threat. **Soil**, v. 2, n. 1, p. 79-82, 2016.

MONTAÑO, R. A. N. R. **Aplicação de técnicas de aprendizado de máquina na mensuração florestal**. Orientador: Eduardo Todt, 2016. 102 f. Tese (Doutorado em Informática) – Universidade Federal do Paraná, Curitiba, 2016.

NASR, M.; SHOKRI, R.; HOUMANSADR, A. Machine learning with membership privacy using adversarial regularization. *In*: **Proceedings of the 2018 ACM SIGSAC conference on computer and communications security**. 2018. p. 634-646.

NEUMANN, M. R. B. **Mapeamento Digital de Solos no Distrito Federal**. Orientador: Henrique Llacer Roig. 2012. 123 f. Tese (Doutorado em Geociências Aplicadas) – Universidade de Brasília, Brasília, 2012.

NONAKA, I.; TOYAMA, R.; KONNO, N. SECI, *Ba* and Leadership: a Unified Model of Dynamic Knowledge Creation. **Long Range Planning**, v.33, p. 5-34, 2000.

NOVAIS, J. J. *et al*. Digital soil mapping using multispectral modeling with Landsat time series cloud computing based. **Remote Sensing**, v. 13, n. 6, p. 1181, 2021.

OLIVEIRA JÚNIOR, R. C. de; SANTOS, P. L. dos; RODRIGUES, T. E.; VALENTE, M. A. Zoneamento agroecológico do município de Tracuateua, Estado do Pará. Belém: Embrapa Amazônia Oriental. p. 45, 1999. (Embrapa Amazônia Oriental. Documentos, 15).

OLIVEIRA, P. A. **Relação solo-relevo assistida por árvore de decisão.** Orientador: Ericson Hideki Hayakawa. 2019. 128 f. Dissertação (Mestrado em Geografia) – Universidade Estadual do Oeste do Paraná, Marechal Cândido Rondon, 2019.

OLIVEIRA, S. *et al*. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. **Forest ecology and management**, v. 275, p. 117–129, 2012.

PADARIAN, J; MINASNY, B.; MCBRATNEY, A. B. Using deep learning for digital soil mapping. **Soil**, v. 5, n. 1, p. 79-89, 2019.

PAVINATO, P. S.; RESOLEM, C. A. Disponibilidade de nutrientes no solo – -decomposição e liberação de compostos orgânicos de resíduos vegetais. **Revista Brasileira de Ciência do Solo**, v. 32, n. 3, p. 911-920. 2008.

PEREIRA, G. W. *et al*. Smart-Map: An open-source QGIS plugin for digital mapping using machine Learning techniques and Ordinary Kriging. **Agronomy**, v. 12, n. 6, p. 1350, 2022b.

PEREIRA, G. W. *et al*. Soil mapping for precision agriculture using support vector machines combined with inverse distance weighting. **Precision Agriculture**, v. 23, n. 4, p. 1189–1204, 2022a.

PEREIRA, I. C. N.; MENEZES, PML de. O radar como instrumento de geração da informação espacial para a gestão do território na Amazônia: uma análise do Projeto Radam. In: **Anais XIII Simpósio Brasileiro de Sensoriamento Remoto**. 2007. p. 6913-6920.

PEREIRA, M. G. *et al*. Formação e Caracterização de Solos. *In*: **Formação, Classificação e Cartografia dos Solos**. Ponta Grossa, PR: Atena Editora, 2019, 1–20 p.

PINHEIRO, H. S. K. *et al*. Modelos de elevação para obtenção de atributos topográficos utilizados em mapeamento digital de solos. **Pesquisa Agropecuária Brasileira**, v. 47, n. 9, p. 1384-1394, 2012.

QI, F.; ZHU, A.-X. Knowledge discovery from soil maps using inductive learning. **Geographical Information Systems**, v. 17, n. 8, p. 771–795, 2003.

QUINLAN, J. R. **C4.5 Programs for Machine Learning**, San Mateo, CA: Morgan Kaufmann, 1992.

RAMCHARAN, A. *et al*. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. **Soil Science Society of America Jornal**. v. 82, n. 1, p. 186–201, 2018.

RODRIGUES, R. A. S. **Ciência do Solo: Morfologia e Gênese**. 1. ed. Londrina: Editora e Distribuidora Educacional S.A., 2018. 264 p.

ROSOLEN, V.; HERPIN, U. Expansão dos solos hidromórficos e mudanças na paisagem: um estudo de caso na região Sudeste da Amazônia Brasileira. **Acta Amazônica**, v. 38, n. 03, p. 483-490, 2008.

ROSSEL, R. A. V.; CHEN, C. Digitally mapping the information content of visible–near infrared spectra of surficial Australian soils. **Remote Sensing of Environment**, v. 115, n. 6, p. 1443-1455, 2011.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. Tradução Regina Célia Simille, 3.ed. Rio de Janeiro: **Elsevier**, 2013.

SARMENTO, E. C. **Comparação entre quatro algoritmos de aprendizagem de máquina no Mapeamento Digital de Solos no Vale dos Vinhedos, RS, Brasil.** Orientador: Elvio Giasson. 2010. 109 f. Dissertação (Mestrado em Ciência do Solo) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.

SARMENTO, E. C. **Predição de classes de solos em diferentes escalas na Serra Gaúcha usando mapeamento digital de solos a partir de dados legados.** Orientador: Elvio Giasson. 2015. 116 f. Tese (Doutorado em Ciência do Solo) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015.

SCHAETZL, R.; ANDERSON, S. **Soils: Genesis and Geomorphology**. England: Cambridge University Press. p. 833, 2005.

SCULL, P. *et al*. Predictive soil mapping: a review. **Progress in physical geography**, v. 27, n. 2, p. 171–197, 2003.

SHAO, S. *et al*. Sample design optimization for soil mapping using improved artificial neural networks and simulated annealing. **Geoderma**, v. 413, n. 115749, p. 115749, 2022.

SHARIFIFAR, A. *et al.* Addressing the issue of digital mapping of soil classes with imbalanced class observations. **Geoderma**, v. 350, p. 84-92, 2019.

SILVA JÚNIOR, J. F. *et al*. Classificação numérica e modelo digital de elevação na caracterização espacial de atributos dos solos. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 16, n. 4, p. 415–424, 2012.

SILVA JÚNIOR, J. F. *et al*. Multivariate split moving windows and magnetic susceptibility for locating soil boundaries of São Paulo, Brazil. **Geoderma Regional**, v. 26, n. e00418, p. e00418, 2021.

SILVA, E. **Mapeamento de solos e uso de algoritmos de aprendizagem em Lavras (MG)**. Orientador: Nilton Curi. 2018. f 194. Tese (Doutorado em Ciência do Solo) – Universidade Federal de Lavras, Lavras, 2018.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas**. ed. 2, São Paulo: Artliber, 2016.

SINGH, C. *et al*. Imodels: A python package for fitting interpretable models. **Journal of open source software**, v. 6, n. 61, p. 3192, 2021.

SONG, J. *et al*. Estimation of Soil Organic Carbon Content in Coastal Wetlands with Measured VIS-NIR Spectroscopy Using Optimized Support Vector Machines and *Random Forest*s. **Remote Sensing**, v. 14, n. 17, p. 4372, 2022.

SUGUIO, K. **Geologia Sedimentar**. Edgard Blücher ltda./EDUSP, São Paulo, SP, 2003, 400 p.SUNG, A. H.; MUKKAMALA, S. Identificação de recursos importantes para detecção de intrusões usando máquinas vetoriais de suporte e redes neurais. In: **2003 Simpósio sobre Aplicações e Internet, 2003. Procedimento.** IEEE, 2003. p. 209-216.

SZATMÁRI, G.; PÁSZTOR, L. Comparison of various uncertainty modelling approaches based on geostatistics and machine learning algorithms. **Geoderma**, v. 337, p. 1329-1340, 2019.

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

TEN CATEN, A. *et al*. Mapeamento digital de classes de solos: características da abordagem brasileira. **Ciência rural**, v. 42, n. 11, p. 1989–1997, 2012.

THERNEAU, T. M. *et al*. **Uma introdução ao particionamento recursivo usando as rotinas RPART**. Fundação Mayo: Relatório técnico, 1997.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B. **Rpart: Recursive Partitioning**. R Package Version 4.119, 2022.

US Geological Survey - USGS. **EarthExplorer**. 2023. https://earthexplorer.usgs.gov

VALADARES, A. P.; COELHO, R. M.; OLIVEIRA, S. R. DE M. Preprocessing procedures and supervised classification applied to a database of systematic soil survey. **Scientia Agricola**, v. 76, n. 5, p. 439–447, 2019.

VAPNIK, V. N. **The nature of statistical learning theory**. New York, NY: Springer New York, 1995.

VILLELA, A. L. O. **Mapeamento Digital de Solos da Formação Solimões Sob Floresta Tropical Amazônica**. Orientador: Marcos Bacis Ceddia. 2013. 112 f. Tese (Doutorado em Agronomia – Ciência do Solo) – Universidade Federal Rural do Rio de Janeiro, Seropédica, 2013.

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.

WANG, B. *et al*. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. **Science of The Total Environment**, v.630, p.367-378, 2018.

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

WEBSTER, R.; BURGESS, T. M. Sampling and bulking strategies for estimating soil properties in small regions. **Journal of Soil Science**, v. 35, n. 1, p. 127-140, 1984.

WILSON J. P.; GALLANT J. C. Digital terrain analysis. In: **Terrain Analysis: Principles and Applications**, Wilson JP, Gallant JC (eds). John Wiley: New York; 1–27, 2000.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data mining: Practical machine learning tools and techniques**. 3. ed. Oxford, England: Morgan Kaufmann, 2011.

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

WRIGHT, M. N.; ZIEGLER, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. **Journal of statistical software**, v. 77, n. 1, 2017.

WU, X. *et al*. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, n. 1, p. 1–37, 2008.

YANG, R. M. *et al*. Comparison of boosted regression tree and *Random Forest* models for mapping topsoil organic carbono concentration in an alpine ecosystem. **Ecological Indicators,** v.60, p.870-878, 2016.

ZHAI, R. *et al*. Wet Aggregate Stability Predicting of Soil in Multiple Land-Uses Based on Support Vector Machine. In: **2021 International Conference on Networking and Network Applications (NaNA)**. IEEE, 2021. p. 527-531.

ZHANG, G.; LIU, F.; SONG, X. Recent progress and future prospect of digital soil mapping: A review. **Journal of Integrative Agriculture**, v. 16, n. 12, p. 2871–2885, 2017.

ZHU, Q.; WANG, Y.; LUO, Y. Improvement of multi-layer soil moisture prediction using support vector machines and ensemble Kalman filter coupled with remote sensing soil moisture datasets over an agriculture dominant basin in China. **Hydrological processes**, v. 35, n. 4, 2021.