

UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CAMPUS DE JABOTICABAL

**FUSÃO DE DADOS E APRENDIZAGEM DE MÁQUINA APLICADOS A
PEDOMETRIA**

Gabriela Mourão de Almeida
Mestre em Agronomia

2023

**UNIVERSIDADE ESTADUAL PAULISTA - UNESP
CAMPUS DE JABOTICABAL**

**FUSÃO DE DADOS E APRENDIZAGEM DE MÁQUINA APLICADOS A
PEDOMETRIA**

MSc. Gabriela Mourão de Almeida

Orientador: Prof. Dr. José Marques Júnior

Coorientador: Dra. Kathleen Fernandes Braz

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – FCAV/Unesp, Campus de Jaboticabal, como parte das exigências para a obtenção do título de Doutora em Agronomia (Produção Vegetal).

2023

A447f

Almeida, Gabriela Mourão de

Fusão de dados e aprendizagem de máquina aplicados a pedometria
/ Gabriela Mourão de Almeida. -- Jaboticabal, 2023

139 p. : il., tabs., fotos, mapas

Tese (doutorado) - Universidade Estadual Paulista (Unesp),
Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal

Orientador: José Marques Júnior

Coorientadora: Kathleen Fernandes Braz

1. Ciência de Dados. 2. Mineralogia do solo. 3. Banco de dados
amplos. 4. Inteligência artificial. 5. Agricultura digital. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de
Ciências Agrárias e Veterinárias, Jaboticabal. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: FUSÃO DE DADOS E APRENDIZAGEM DE MÁQUINA APLICADOS A PEDOMETRIA

AUTORA: GABRIELA MOURÃO DE ALMEIDA
ORIENTADOR: JOSÉ MARQUES JUNIOR
COORIENTADORA: KATHLEEN FERNANDES BRAZ

Aprovada como parte das exigências para obtenção do Título de Doutora em Agronomia (Produção Vegetal), pela Comissão Examinadora:

Prof. Dr. JOSÉ MARQUES JUNIOR (Participação Virtual)
Departamento de Ciências da Produção Agrícola / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **JOSÉ MARQUES JUNIOR**
Data: 17/05/2023 18:02:21-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. ZIGOMAR MENEZES DE SOUZA (Participação Virtual)
Faculdade de Engenharia Agrícola / Universidade Estadual de Campinas - UNICAMP

Documento assinado digitalmente
 **ZIGOMAR MENEZES DE SOUZA**
Data: 18/05/2023 10:26:11-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. GLAUCO DE SOUZA ROLIM (Participação Virtual)
Departamento de Engenharia e Ciências Exatas DECEX / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **GLAUCO DE SOUZA ROLIM**
Data: 22/05/2023 15:42:48-0300
Verifique em <https://validar.iti.gov.br>

Prof. Dr. RONNY SOBREIRA BARBOSA (Participação Virtual)
Universidade Federal do Piauí / UFPI - Bom Jesus/PI

Documento assinado digitalmente
 **RONNY SOBREIRA BARBOSA**
Data: 18/05/2023 07:47:51-0300
Verifique em <https://validar.iti.gov.br>

Pós-doutoranda ANGÉLICA SANTOS RABELO DE SOUZA BAHIA (Participação Virtual)
Departamento de Solos e Adubos / FCAV UNESP Jaboticabal

Documento assinado digitalmente
 **ANGELICA SANTOS RABELO DE SOUZA BAHIA**
Data: 25/05/2023 09:54:22-0300
Verifique em <https://validar.iti.gov.br>

Jaboticabal, 28 de abril de 2023

DADOS CURRICULARES DA AUTORA

Gabriela Mourão de Almeida – Filha de Luiz Carlos Carneiro de Almeida e Katia do Socorro da Cunha Mourão de Almeida, nascida na cidade de Belém no estado do Pará, no dia 15 de novembro de 1994. Em abril de 2013 iniciou a graduação em Engenharia Agrônoma na Universidade Federal Rural da Amazônia - campus de Capitão Poço concluiu o curso em fevereiro de 2018, ao longo de sua graduação desenvolveu o projeto *“Avaliação da omissão de nutrientes no desenvolvimento e composição mineral de plantas jovens de cupuaçuzeiro (theobroma grandiflorum) progênie 61”* em parceria com a Embrapa Amazônia Oriental e supervisão do professor Dr. Ismael de Jesus Matos Viégas. Em março de 2018 iniciou o mestrado em Agronomia – Programa de Ciência do Solo na Universidade Estadual Paulista - Faculdade de Ciências Agrárias e Veterinárias (UNESP-FCAV), com bolsa de estudos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes, concluiu o curso em julho de 2019. Durante seu mestrado desenvolveu a pesquisa *“Aprendizagem de Máquina na Determinação de Ambientes de Produção de Cana-de-Açúcar”* sob orientação do professor Dr. Gener Tadeu Pereira e coorientação da professora Dra. Angélica Santos Rabelo de Souza Bahía. Em agosto de 2019 iniciou o doutorado Agronomia – Programa de Produção Vegetal na Universidade Estadual Paulista - Faculdade de Ciências Agrárias e Veterinárias (UNESP-FCAV), com bolsa de estudos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Capes e desenvolveu a pesquisa *“Fusão de dados e aprendizagem de máquina aplicados a pedometria”* sob orientação do professor Dr. José Marques Júnior e coorientação da professora Dra. Kathleen Fernandes Braz. Desde março de 2018 faz parte do grupo de pesquisa Caracterização do solo para fins de manejo específico – CSME localizado na UNESP- campus de Jaboticabal.

“Os de ideias curtas acham que os inconformistas são pragas da sociedade.”

Clarissa Pinkola Estés

“A ciência humana de maneira nenhuma nega a existência de Deus. Quando considero quantas e quão maravilhosas coisas o homem compreende, pesquisa e consegue realizar, então reconheço claramente que o espírito humano é obra de Deus, e a mais notável.”

Galileu Galilei

DEDICATÓRIA

Dedico esta tese Deus, o maior orientador da minha vida. Ele nunca me abandonou nos momentos de necessidade. Aos meus pais Luiz e Kátia, a meu marido Vinicius e à minha irmã Sabrina por todo o apoio e paciência ao longo destes dez anos de carreira acadêmica. Vocês são os maiores incentivadores dos meus sonhos e meu combustível nos momentos difíceis.

AGRADECIMENTOS

Primeiramente agradeço a Deus e Nossa Senhora de Nazaré, por ter me agraciado com o dom da vida e por ao longo desses dez anos de carreira acadêmica ter me dado saúde, persistência, sabedoria e resiliência para perseguir os meus sonhos e por terem me guiado por caminhos que jamais imaginei alcançar.

Aos meus pais, Katia do Socorro da Cunha Mourão de Almeida e Luiz Carlos Carneiro de Almeida, que se sacrificaram, se dedicaram, abdicaram de tempo e de muitos projetos pessoais para que eu tivesse a oportunidade de estudar e de ter uma boa formação profissional, mas também pessoal. Eu devo tudo que sou a vocês, e se sinto orgulho de mim e do lugar onde cheguei, é porque sei que vocês vieram segurando a minha mão.

Ao meu marido, Dr. Vinicius Marchioro por todas as noites em claro que passamos juntos desenvolvendo a nossa pesquisa, por todos aqueles finais de semana que tivemos que abdicar em função das nossas teses, por cada conversa e palavra de incentivo, obrigada por segurar a minha mão a cada nova etapa e suportar as crises de estresse e minha ausência em diversos momentos. Sem você e sua presença este trabalho não seria possível.

Agradeço a minha irmã, Sabrina, por cada palavra de incentivo, por todo o apoio incondicional neste e em todos os anos intensos dentro na universidade, por fim agradeço por sempre ser presente com os nosso pais enquanto eu estava ausente.

Ao orientador professor Dr. José Marques Júnior, agradeço por todos os ensinamentos, conselhos, amizade, pela disponibilidade que sempre manifestou e a empatia com que recebeu as minhas ideias, foram o estímulo que me permitiu vencer as inseguranças deste processo. Obrigada por tudo.

Agradeço à minha coorientadora, professora Dra. Kathleen Fernandes Braz, pelo apoio sempre que precisei, por todos os ensinamentos dados ao longo deste processo e pelos bons momentos que compartilhamos. Obrigada por tudo.

In memoriam ao meu avô Nelson de Azevedo Mourão, a minha avó e madrinha Maria da Cunha Mourão e minha tia Sandra Cristina Mourão que acompanharam o início da minha vida acadêmica em matéria e está acompanhando a conclusão em espírito, pois, a existência material é apenas uma etapa de testes da vida eterna e quando esta se concluiu não quer dizer o fim de tudo e sim o início da verdadeira vida, vocês me fizeram muito feliz na infância. As minhas tias Sandra Maria Mourão e Ana Madalena Mourão por todo o apoio e auxílio dados ao longo da vida desde a ajuda

nas compras do material escolar da educação infantil a companhia para a viagem rumo ao mestrado.

Aos meus avós paternos Maria José Carneiro e Moacyr Almeida por terem dado a vida ao meu pai, Luiz Carlos que é uma pessoa maravilhosa e sou o que sou hoje graças a ele, bem como a todos da família Almeida.

Aos companheiros do grupo de pesquisa CSME por todo o carinho e companheirismo ao longo da caminhada.

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

Ao grupo de pesquisa CSME pela disponibilidade do banco de dados e as agências de pesquisa CAPES (Proc. nº149940); PROPE / UNESP (Pró-reitora de Pesquisa da Universidade Estadual Paulista, Edital Nº 15/2014); CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Edital Universal 01/2016 – Faixa C, Proc.nº 402796/2016-0) e Processo FAPESP nº 2013/17552-6.

Por fim, agradeço a todos que de alguma forma contribuíram para a realização desse sonho.

Sumário

Capítulo 1 – Considerações gerais.....	1
1.1 Introdução	1
1.2 Referencial Teórico	3
1.2.1 Pedometria	3
1.2.2 Uso de sensores na ciência do solo	4
1.2.3 Fusão de dados	6
1.2.4 Aprendizagem de Máquina	8
1.2.5 Extrapolação de dados	11
1.2.6 Análises de mineralogia do solo.....	12
1.3 Referências	13
Capítulo 2 - Fusão de dados multissensor e aprendizagem de máquina na estimativa de atributos mineralógicos em solos do Planalto Ocidental Paulista.	26
Resumo.....	26
Palavras-chave	26
Abstract.....	27
Keywords	27
2.1 Introdução	28
2.2 Material e métodos.....	30
2.2.1 Caracterização da área de estudo	30
2.2.2 Plano amostral de coleta de pontos	31
2.2.3 Análises de laboratório.....	32
2.2.4 Sensores	32
2.2.5 Fusão de Dados.....	33
2.2.6 Análises Estatísticas	34
2.3 Resultados e discussão	37
2.3.1 Estatística exploratória	37
2.3.2 Fusão de dados	40
2.3.3 Análise de predição.....	44
2.3.4 Análise Geoestatística	51
2.4 Conclusões	57
2.5 Referências	58
Capítulo 3 – Uso de aprendizagem de máquina para extrapolação de atributos mineralógicos do solo	75

Resumo.....	75
Palavras-chave	75
Abstract.....	76
Keywords	76
3.1 Introdução	77
3.2 Material e métodos.....	79
3.2.1 Caracterização da área de estudo	79
3.2.2 Plano amostral de coleta de pontos	81
3.2.3 Análises de laboratório – Atributos mineralógicos.....	82
3.2.4 Espectroscopia de Reflectância Difusa (ERD)	82
3.2.5 Aprendizagem de máquina	82
3.2.6 Geoestatística	84
3.3 Resultados e discussão	86
3.4 Conclusões	112
3.5 Referências.....	112

Resumo

Como nos mais diversos setores, a agricultura vem passando por uma revolução nos últimos anos e para os países que são grandes produtores de alimento como o Brasil a adoção de tecnologias inovadoras vem se tornando cada dia mais essencial para alcançar os resultados de produção esperados. Como resultado desta revolução ocorreu um grande aumento na disponibilidade de dados, ampliação da capacidade de processamento computacional e assim os antigos modelos de regressão simples passaram a dar espaço a técnicas mais robustas de análise de dados e assim a aprendizagem de máquina (AM) começou a sair da área de sistema da informação e adentrar nos sistemas agrícolas. A AM é atualmente uma das linhas de pesquisa mais vigoras na agricultura, pois diferente dos modelos clássicos esta tem a capacidade de trabalhar com toda a complexidade que envolve o sistema agrícola, Dentre as variáveis ambientais, o solo é a que mais implica em complexidade, diante disso foi criado uma linha de pesquisa dedicada ao estudo da sua variabilidade que é a pedometria que consiste na aplicação de métodos matemáticos e estatísticos no estudo da distribuição, caracterização e gênese dos solos. Pesquisadores atrelados a essa linha de pesquisa vem afirmando a nova linha de estudo crescente é a utilização de sensores e a aplicação de técnicas de AM, principalmente em países de clima tropical e grande extensão territorial como o Brasil, que tem como característica solos altamente intemperizados e devido a isso demandam uma maior quantidade de amostras para caracterização. Este tipo de solo frequentemente apresenta o ferro na composição de seus minerais e este é um importante indicador analítico do solo, pois ele nos permite a compreender a gênese, classificação e variabilidade do solo.

Palavras-chave: Ciência de Dados; Mineralogia do solo; Banco de dados amplos; Inteligência artificial; Agricultura digital.

Abstract

As in the most diverse sectors, agriculture has been undergoing a revolution in recent years and for countries that are large food producers like Brazil, the adoption of innovative technologies is becoming more and more essential to achieve the expected production results. As a result of this revolution, there was a large increase in data availability, expansion of computational processing capacity and so the old simple regression models began to give way to more robust data analysis techniques and thus machine learning (ML) began to leaving the information system area and entering agricultural systems. AM is currently one of the most vigorous lines of research in agriculture, as unlike classic models it has the ability to work with all the complexity that involves the agricultural system. Among the environmental variables, the soil is the one that most implies complexity, in view of this, a line of research was created dedicated to the study of its variability, which is pedometrics, which consists of the application of mathematical and statistical methods in the study of the distribution, characterization and genesis of soils. Researchers linked to this line of research have been affirming that the new growing line of study is the use of sensors and the application of AM techniques, mainly in countries with a tropical climate and large territorial extension such as Brazil, which is characterized by highly weathered soils and due to this, they demand a larger number of samples for characterization. This type of soil often presents iron in the composition of its minerals and this is an important analytical indicator of the soil, as it allows us to understand the genesis, classification and variability of the soil.

Keywords: Data Science; Soil mineralogy; Large database; Artificial intelligence; Digital agriculture.

Capítulo 1 – Considerações gerais

1.1 Introdução

O Brasil é o maior país de clima tropical e um dos maiores produtores de alimento do mundo, com área agrícola de aproximadamente 60 milhões de hectares e com potencial de expansão (Demattê et al., 2019). Porém, a cada dia este setor vem enfrentando desafios qualitativos e quantitativos, gerados devido a crescimento da população global e a intensificação das mudanças climáticas (Liu et al., 2023)

O desenvolvimento e a adaptação de tecnologias inovadoras podem ser vistos como uma abordagem dinâmica que pode ajudar a aumentar os resultados agrícolas e mitigar a ameaça de escassez de alimentos em todo o mundo (Sarfraz et al., 2023). Nos últimos anos, com o rápido aumento na disponibilidade de dados abertos, com a expansão do poder computacional e com o aperfeiçoamento de algoritmos de aprendizagem de máquina (AM), a inteligência artificial (IA) vem ganhando espaço de destaque nas mais diversas áreas (Jararweh et al., 2023)

A aplicação IA envolve o processo de aprendizagem das máquinas, que tem como objetivo fornecer à máquina dados de experiências passadas e dados estatísticos para que ela possa executar a atividade que lhe foi atribuída para resolver a uma demanda específica. Ou seja, é abordagem matemática para construir máquinas inteligentes e é devido a AM que o domínio de *big data* e da ciência de dados evoluiu tanto nos últimos anos (Jha et al., 2019).

Segundo (Bowden et al., 2023) diferente dos modelos de regressão paramétricos que assumem relacionamentos pré-especificados como lineares ou quadráticos simples, os métodos de aprendizado de máquina (AM), como modelos XGBoost e Florestas Aleatórias vem ganhando bastante espaço na agricultura devido à capacidade de explorar relações sem a necessidade de pré-definir a forma funcional dessas relações, desembaraçando assim a complexidade das interações não lineares

de sistemas agrícolas (Shahhosseini et al., 2021)

Dentre as variações ambientais dos sistemas agrícolas, o solo é um dos que implica em uma maior complexidade, pois é influenciado por atributos físicos, químicos e biológicos, cuja interação condiciona sua variabilidade espacial (Moral et al., 2019a). Devido toda esta complexidade foi desenvolvida uma linha de pesquisa dentro da ciência do solo que utiliza a matemática para explicar a variabilidade do solo, que é a pedometria.

A pedometria reúne um conjunto de técnicas que auxilia a tomada de decisão nos planos de manejo do solo, e é comumente definida como “a aplicação de métodos matemáticos e estatísticos no estudo da distribuição, caracterização e gênese dos solos” (Thompson et al., 2012). Segundo (Rossiter, 2018) a aplicação de técnicas pedométricas demanda avanços tecnológicos de coleta e processamento de grandes conjuntos de dados. Dentro desta evolução, segundo a linha do tempo desenvolvida por McBratney et al. (2019) os sensores e a aprendizagem de máquina (AM) são agora as áreas vigorosas de pesquisa na pedometria e na ciência do solo, devido à enorme demanda de análise de dados, e tendem a ganhar ainda mais importância, principalmente em países de clima tropical como o Brasil (Minasny and McBratney, 2016)

Países de clima tropical tem como característica predominante a presença de Latossolos altamente intemperizados, pois em regiões tropicais há a ocorrência de variações de alta magnitude dos fatores e processos de formação dos solos (Minasny e Hartemink, 2011a). Segundo (Camargo et al., 2009) quanto maior a diversidade destes fatores numa área, principalmente do material de origem e relevo, mais variável são os atributos destes solos.

Devido a este elevado grau de intemperismo, os minerais residuais predominantes na fração argila destes solos são a caulinita e oxi-hidróxidos de ferro e alumínio (goethita, hematita, gibbsita) (Eberhardt et al., 2008a). Estes minerais são

considerados pedoindicadores de processos do solo, e suas características controlam as reações de superfície básicas. Assim, torna-se imprescindível identifica-los para entender as relações de causa e efeito entre os atributos do solo (Bahia et al., 2017a), para entender a variabilidade dos atributos e realizar planejamentos táticos e operacionais de manejo desses solos (Pedron et al., 2018).

1.2 Referencial Teórico

1.2.1 Pedometria

O Brasil é um dos maiores produtores de alimento do mundo, com área agrícola de aproximadamente 60 milhões de hectares e com potencial de expansão. Porém, em apenas 10% destas áreas são utilizadas técnicas de agricultura de precisão (AP), o que, ainda assim, compreende cerca de 200.000 análises de fertilidade do solo por ano (Demattê et al., 2019)

A baixa adoção de AP é associada, dentre outros fatores, ao longo tempo de execução e alto custo das análises convencionais na avaliação da mineralogia e fertilidade do solo (Ramaroson et al., 2018), o que torna a análise pouco atraente para o grande número de amostras necessárias para caracterizar com precisão grandes áreas. Outro fator relevante, quando se considera a avaliação convencional da mineralogia e fertilidade do solo, é a grande quantidade de produtos químicos utilizados durante os procedimentos laboratoriais, que podem vir a ser nocivos ao meio ambiente, além do alto nível de periculosidade para quem realiza a análise (Ng et al., 2020a)

Em consequência às limitações das técnicas convencionais de mapeamento do solo, o nível de detalhamento espacial e a precisão dos mapas convencionais do solo não são adequados para aplicação de manejo que requer o conhecimento da variabilidade dos atributos (Liu et al., 2020). Isto, afeta principalmente os países que

possuem grandes extensões territoriais ocupadas com agricultura (Demattê et al., 2019), como o Brasil. Sendo assim, há necessidade de técnicas alternativas para caracterizar os atributos do solo e suas variações espaciais, por meio de um monitoramento regular, possibilitando tomadas de decisões em tempo real no campo (Ng et al., 2019).

Dentre estas técnicas, pode-se citar a pedometria, que é o ramo da ciência do solo que busca aplicar métodos matemáticos e estatísticos ao estudo de atributos do solo (McBratney et al., 2019). Inicialmente, os estudos da área dedicaram-se à modelagem geoestatística da variação espacial do solo (Rossiter, 2018). Com o passar dos anos a pedometria foi se moldando às novas necessidades do estudo de ciência do solo. Um exemplo é trabalho de (McBratney et al., 2003) Estes autores traçam os caminhos do mapeamento digital do solo atual, passando a ser um espaço grande marco na pedometria. Neste contexto a pedometria passa a ser uma abordagem convencional e praticada por todas as principais instituições mundiais que trabalham com solos (Heuvelink et al., 2019).

Atualmente a pedometria está evoluindo na direção da tecnologia da informação com base no desenvolvimento de computadores mais potentes, associados à grande quantidade de dados (Big Data), justificando a necessidade de métodos mais sofisticados (McBratney et al., 2003). Dentro desta evolução, segundo a linha do tempo desenvolvida por (McBratney et al., 2019), os sensores e a aprendizagem de máquina são agora as áreas vigorosas de pesquisa na ciência do solo e, devido à enorme demanda de análise de dados, tendem a ganhar ainda mais importância (Minasny e McBratney, 2016)

1.2.2 Uso de sensores na ciência do solo

Como consequência das limitações das técnicas convencionais de mapeamento do solo, têm-se o nível de detalhamento espacial e a precisão dos

mapas convencionais do solo não adequados para aplicação de manejo que requer o conhecimento da variabilidade dos atributos (Liu et al., 2020). Isto afeta principalmente os países que possuem grandes extensões territoriais ocupadas com agricultura (Demattê et al., 2019), como o Brasil. Sendo assim, há necessidade de técnicas alternativas para caracterizar os atributos do solo e suas variações espaciais, por meio de um monitoramento regular, possibilitando tomadas de decisões em tempo real no campo (Ng et al., 2020a).

Como a mineralogia do solo é extremamente importante, mas demorada para ser avaliada por análises tradicionais, novas possibilidades surgiram, como métodos baseados em sensores. As medições via sensores proximais não utilizam reagentes químicos e são não destrutivas, proporcionando resultados rápidos e confiáveis (Silva et al., 2022a; Teixeira et al., 2020)

Diversos trabalhos tem sido desenvolvidos utilizando sensores, proximais ou satelitais, individualmente ou em associações, para prever diversos atributos de solo, como grau de intemperismo do solo (Kragt et al., 2012; Teixeira et al., 2022; L. Zhao et al., 2022a), pH do solo (Merl et al., 2022; Silva et al., 2021), Material de origem e mineralogia do solo (Mancini et al., 2022; Mendes et al., 2021; Silva et al., 2022b, 2021; L. Zhao et al., 2022a), atributos e fertilidade do solo (Andrade et al., 2022; Mendes et al., 2021; Silva et al., 2021; Xu et al., 2019; L. Zhao et al., 2022a), carbono orgânico do solo (Faria et al., 2022; L. Zhao et al., 2022a) e metais pesados (Silva et al., 2022; Xu et al., 2020).

Sensores e técnicas utilizando os princípios da espectroscopia de reflectância difusa (ERD) permitem observar as estruturas de óxidos de ferro (hematita e goethita) detectadas no visível (VIS) (400-700 nm), a matéria orgânica, os carbonatos, a água e os minerais argilosos (caulinita, esmectita e illita), pelo infravermelho próximo (NIR) (700–2.500 nm), as assinaturas de minerais de argila, quartzo, óxidos de ferro e

matéria orgânica pelo infravermelho médio (MIR) (2.500–25.000 nm), o que demonstram a potencialidade do uso desta tecnologia na agricultura (Coblinski et al., 2020; Goldshleger et al., 2004; Nocita et al., 2015; Reda et al., 2019). A análise destes dados, de forma conjunta (VIS-NIR-MIR), tem-se mostrado superior àquela quando feita individualmente (Soriano-Disla et al., 2014).

A espectrometria de fluorescência de raios-X (EXRF) funciona por meio da emissão de raios X de alta energia que atingem os átomos do material contido no solo (Weindorf et al., 2013), tornando possível identificar a química elementar dos solos. Como cada elemento possui intensidade de fluorescência característica, a EXRF identifica e quantifica os elementos presentes na amostra (Weindorf et al., 2014a). Segundo (Silva et al., 2020a), esta técnica, assim como a ERD, apresenta baixo custo de execução, fornece informações rápidas e gera um banco de dados robusto, além de ser um método não destrutivo, e necessita de preparação mínima da amostra. As três faixas do ERD e EXRF apresentam pontos fortes e fracos, e as informações contidas em uma faixa espectral específica ou nas intensidades emitidas pela EXRF não contêm informações suficientes para prever todas as propriedades do solo (Ng et al., 2020b).

Entretanto, uma problemática encontrada na utilização de sensores é a inespecificidade e o grande volume de dados gerados, fazendo com que a interpretação e associação dos resultados com atributos desejados não seja feita de forma simples e intuitiva (Almeida et al., 2020).

1.2.3 Fusão de dados

Atualmente, a pedometria está evoluindo na direção da tecnologia da informação, com base no desenvolvimento de computadores mais potentes, associados à grande quantidade de dados (Big Data), justificando a necessidade de

métodos mais sofisticados (McBratney et al., 2003). Dentro desta evolução, segundo a linha do tempo, desenvolvida por (McBratney et al., 2019), a ERD e a aprendizagem de máquina são agora as áreas vigorosas de pesquisa na ciência do solo e, devido à enorme demanda de análise de dados, tendem a ganhar ainda mais importância (Minasny and McBratney, 2016).

Uma das maneiras plausíveis de melhorar a previsão de várias propriedades do solo é fundir os dados das diferentes faixas de ERD com as intensidades da EXRF. A “fusão de dados” consiste na união de dados de diversas fontes e pode ser dividida em três níveis: baixo, médio e alto nível (Hall e Llinas, 2016).

Devido à complexidade e heterogeneidade dos solos, os modelos baseados em sensores únicos são menos precisos do que os modelos que utilizam multissensores (Grunwald et al., 2015). A integração de dados de múltiplos sensores, conhecida como “fusão de sensores” ou “fusão de dados”, está se tornando cada vez mais popular para fins de estimativa de propriedades do solo (Tabatabai et al., 2019; Vasques et al., 2020). A fusão de sensores supera as limitações impostas pelas incertezas contidas em um único conjunto de dados, fornecendo um método mais abrangente, resultando em modelos de previsão mais robustos e precisos de caracterização de solos (Andrade et al., 2022; Grunwald et al., 2015; Mancini et al., 2022; Teixeira et al., 2022; L. Zhao et al., 2022a).

De acordo com (Borràs et al., 2015), o nível mais simples de fusão é o baixo, que consiste simplesmente na concatenação dos dados originários dos diferentes instrumentos. O nível intermediário consiste na extração dos dados de maior relevância, gerados em cada fonte e, em seguida, são convertidos em uma única escala, utilizando análises como a de componentes principais e, por fim, estes dados são concatenados em uma única tabela. O terceiro e último nível é o alto, no qual os dados gerados em cada fonte são submetidos às análises de classificação e de

regressão para o desenvolvimento de um modelo para os dados de cada fonte e, por fim, os modelos são combinados para, assim, obter-se o melhor modelo final.

Como já dito anteriormente, tanto o ERD como o EXRF geram um banco de dados bastante robusto, e a fusão destes dados só aumenta este volume, gerando assim uma “big data” (BD). O BD é um conceito advindo do ramo da tecnologia da informação, que vem sendo amplamente aplicado na agricultura e consiste no processamento de grandes volumes de dados, provenientes de diversas variáveis de entrada, que são processados em alta velocidade (Baseca et al., 2019).

1.2.4 Aprendizagem de Máquina

Diante da grande quantidade de dados e problemas cada vez mais robustos para serem interpretados, há necessidade da utilização de análises de dados mais robustas do que aquelas elaboradas pelos métodos estatísticos tradicionais lineares, que possuem limitações para entender as interações do sistema trifásico do solo (Nawar et al., 2019). Em contraste, as técnicas relacionadas ao aprendizado de máquina (AM) vêm ganhando destaque devido à possibilidade de modelar relações não lineares em dados, prescindindo das premissas restritivas associadas às abordagens tradicionais (Bekker et al., 2019a).

A técnica de AM, originada da interface entre a ciência da computação e a estatística, consiste, basicamente, em utilizar algoritmos matemáticos capazes de processar (aprender) em sistemas, através de análises de regressão e de classificação, à medida que são submetidos a treinamentos (Canizo et al., 2019).

Algoritmos de AM são mais precisos na estimativa de atributos de interesse do que técnicas paramétricas tradicionais, como regressões, por exemplo (Sothe et al., 2019), especialmente quando o objeto de estudo é complexo, e os dados a serem trabalhados são não lineares, com diferentes escalas e de diferentes tipos de fonte. A seleção do algoritmo apropriado para a construção do modelo é fundamental para

otimizar os resultados no contexto dos dados disponíveis (Bui et al., 2020).

Um dos algoritmos de AM mais utilizados atualmente é do Florestas Aleatórias (Random Forest). O modelo de Regressão Florestas Aleatórias (FA) (L. Breiman, 2001) é uma técnica não paramétrica que foi desenvolvida com o intuito de melhorar a previsão dos modelos de Árvores de Classificação e Regressão, pois utiliza árvores de decisão totalmente crescidas e reduz o erro e a variação (Yan et al., 2020a).

Tal algoritmo consiste na combinação de diversas árvores, que são geradas a partir de uma variável de entrada amostrada aleatoriamente, todas as árvores apresentam o mesmo tamanho. Já as subdivisões das árvores são realizadas com base em um subconjunto amostrado aleatoriamente do banco de dados total, o resultado do modelo de FA é composto pela média dos resultados de todas as árvores (Chagas et al., 2016a).

Algoritmos de Random Forest usam estratégias bagging e subespaço aleatório para induzirem diversidade entre os aprendizes base para melhor generalização (L. Breiman, 2001). Bagging, também conhecido como agregação bootstrap, gera vários pacotes de um determinado conjunto de treinamento, de modo que cada árvore seja treinada em um determinado pacote de dados. Cada árvore usa um pacote de dados de treinamento cuja distribuição é semelhante a toda a população e, portanto, cada classificador apresenta bom desempenho de generalização. Dentro de cada árvore de decisão, a abordagem de subespaço aleatório é usada em cada nó não terminal para aumentar ainda mais a diversidade entre os modelos de base (Ganaie et al., 2022).

Para obter o melhor desempenho de generalização, vários hiperparâmetros da floresta aleatória precisam ser escolhidos de maneira ideal. Esses hiperparâmetros incluem número de aprendizes básicos como números de árvores em uma floresta (ntree), número de recursos candidatos para avaliação em um

determinado nó não-folha (mtry) e número de amostras em um nó impuro (nodesize ou minleaf) (L. Breiman, 2001).

Dentre os algoritmos de AM, o XGBoost (Extreme Gradient Boosting), em específico, vem sendo bastante empregado e se destacando pela maior robustez e precisão dos resultados. O XGBoost combina técnicas de otimização de software e hardware para produzir resultados superiores, usa menos recursos de computação em menor período de tempo, tem a vantagem de reduzir superestimação e subestimação de valores, tem a capacidade de minimizar overfitting, trabalha bem com problemas de normalização de dados, é menos influenciado por mínimos locais e permite maior possibilidade de ajuste de hiperparâmetros (Maurya et al., 2022a; D. Zhao et al., 2022a).

Este algoritmo é um modelo de otimização com características tanto de um modelo linear quanto de um modelo de árvore e pode concluir simultaneamente tarefas de regressão e classificação ao mesmo tempo. O algoritmo é composto por várias árvores de decisão e o aprendizado dele é realizado por meio da integração de árvores de decisão, onde os valores previstos de todas as árvores de decisão somam-se aos valores previstos do modelo e o modelo é treinado pelo algoritmo de árvore de decisão de levantamento de gradiente (Wang et al., 2022).

Este algoritmo foi apresentado pela primeira vez por (Chen and Guestrin, 2016b) e tem sido amplamente utilizado devido à sua eficiência robusta. O Xgboost usa um algoritmo de gradiente descendente para atualizar os parâmetros do modelo e tem a capacidade de reconhecer os recursos mais importantes. Devido à natureza estocástica do algoritmo, a sequência de prioridade do recurso pode variar, portanto, o algoritmo é executado várias vezes e a probabilidade de cada recurso para cada posição é calculada e, com base na probabilidade mais alta, as sequências de importância do recurso são decididas (Chen and Guestrin, 2016b), sendo

especialmente usado para lidar com dados espaciais (Xu et al., 2022).

O algoritmo de XGBoost tem sido empregado com sucesso para estimativa de diversos parâmetros ambientais como, cobertura vegetal (Maurya et al., 2022a), qualidade da água (Xu et al., 2022), modelos hidrológicos de bacias (Wang et al., 2022), meteorologia florestal (Ghafarian et al., 2022), água no solo (Karthikeyan and Mishra, 2021), modelos de física do solo (Zhang et al., 2021), salinização do solo (Ma et al., 2021), erosão do solo e da água (Golkarian et al., 2023), carbono do solo (He et al., 2022; Nguyen et al., 2022b; Zhang et al., 2022), parâmetros produtivos de culturas agrícolas (Han et al., 2022), metais pesados no solo (Ye et al., 2023; D. Zhao et al., 2022a), entre outras aplicações, o que tem feito esse algoritmo despontar nos últimos anos.

1.2.5 Extrapolação de dados

Em áreas com bases limitadas de dados de solo, métodos de interpolação espacial podem ser usados para prever solos em locais não visitados, porém, surgem problemas de acurácia quando se pretende prever atributos de solo em áreas com poucas observações de solo (Ruhollah et al., 2022a). Em tais situações, uma abordagem de extrapolação pode ser aplicada para prever os solos na área alvo (locais sem observações) usando os solos da área de referência (área com observações adensadas) se os solos de duas áreas forem controlados principalmente por fatores de formação de solo semelhantes, como bem como processos pedogênicos semelhantes (ten Caten et al., 2011a). Portanto, um modelo AM desenvolvido para uma área de referência, usando abordagens de mapeamento digital do solo, pode ser extrapolado para prever as características do solo da área alvo (Afshar et al., 2018a; Cambule et al., 2013a).

Vários estudos usaram modelos de extrapolação para prever espacialmente

classes de solo ou propriedades em áreas alvo. Em tais estudos, os modelos AM foram treinados usando um conjunto de dados incluindo covariáveis e dados de solo observados de uma área de referência para prever solos em uma área alvo (Angelini et al., 2020a; Silva et al., 2016).

Deve-se notar que existem vários desafios que podem afetar as estimativas para as áreas-alvo. A dissimilaridade nos fatores formadores do solo e nos fatores ambientais (Malone et al., 2016), diferentes usos da terra e histórico de cobertura do solo (Temme and Veldkamp, 2009), entre outros fatores, estão entre os problemas mais importantes que devem ser considerados na modelagem.

1.2.6 Análises de mineralogia do solo

Os fatores e processos de formação do solo influenciam diretamente a variabilidade espacial dos atributos do solo, devendo ser considerado no seu manejo (Gray et al., 2016). Apesar das análises mineralógicas serem essenciais para o entendimento da dinâmica dos solos tropicais (Costa et al., 2020) estas não são realizadas de forma rotineira como as de fertilidade, o que pode ser associado à alta complexidade das análises mineralógicas do solo, o que acaba exigindo maior tempo para análises, profissionais especificamente qualificados e muitos reagentes químicos (Rosin et al., 2022), fazendo com que a tomada de decisão em campo ainda seja complexa, baseada na expertise dos profissionais e pouco assertiva (Moral et al., 2019b; Sumathi, 2018).

Diante desta dificuldade Rennert (2019) tentou emplacar uma abordagem alternativa que apresenta baixo custo e não requer instalações laboratoriais sofisticadas. Este método consiste na separação da substância alvo das amostras de solo e em seguida na execução da análise química do sobrenadante, diminuindo o tempo de análises, a quantidade de reagentes, equipamentos e o número de

profissionais envolvidos.

Para extração de óxidos Fe foram desenvolvidos dois métodos que quantificam Fe totais ('livres') no solo. O primeiro consiste na extração por ditionito–citrato-bicarbonato (FED). Para determinação é realizada a dissolução redutiva através da ditionita, neste processo os íons de Fe^{2+} e Al^{3+} são reduzidos e liberados concomitantemente e permanecem em solução na presença de um agente complexante como o citrato (Holmgren, 1967a). Enquanto a solução de bicarbonato mantém o pH neutro.

O segundo método consiste na extração com oxalato de amônio-ácido oxálico (FEO) que induz a formação complexos solúveis entre o oxalato e os íons Al e Fe. Ao atingir o $\text{pH} < 3.5$, as superfícies de óxidos são protonadas, sendo este o primeiro estágio de dissolução em seguida ocorre a adsorção de oxalato e liberação de íons Al^{3+} e Fe^{3+} (McKeague e Day, 1966). Enquanto o FED mostra os conteúdos de ferros cristalinos que constituem os óxidos, o FEO mostra aqueles de menor cristalinidade. Ambos permitem uma indicação aos conteúdos totais de óxidos de ferro por métodos eficientes e menos morosos (Ker et al., 2012).

Porém, ao longo dos anos mesmo essa abordagem alternativa a quantificação de Fe não ganhou espaço como análise de rotina em laboratórios. O que pode ser atribuído ao fato de as análises laboratoriais serem de alta periculosidade pois utilizam ácidos, sais e soluções de alta periculosidade. Além disso, ainda são necessários técnicos que dominem os métodos para realização das análises, longo tempo para execução e por fim, apresentando alto custo (Mendes et al., 2022a). Fatos que também tornam estas análises pouco escaláveis.

1.3 Referências

Abdikan, S., Sekertekin, A., Narin, O.G., Delen, A., Balik Sanli, F., 2022. A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using Sentinel-1 and Sentinel-2. *Advances in Space Research*. <https://doi.org/10.1016/J.ASR.2022.11.046>

- Afshar, F.A., Ayoubi, S., Jafari, A., 2018a. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Afshar, F.A., Ayoubi, S., Jafari, A., 2018b. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R. de S., Fernandes, K., Marques Júnior, J., 2021. Machine learning in the prediction of sugarcane production environments. *Comput Electron Agric* 190. <https://doi.org/10.1016/j.compag.2021.106452>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R.D.S., Júnior, J.M., 2020. Aprendizagem de máquina na predição de ambientes de produção de cana-de-açúcar. *Comput Electron Agric*.
- Andrade, R., Mancini, M., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/J.GEODERMA.2022.115936>
- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020a. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>
- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020b. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>
- Anzanello, M.J., Ortiz, R.S., Limberger, R.P., Mayorga, P., 2013. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. *J Pharm Biomed Anal* 83, 209–214. <https://doi.org/10.1016/J.JPBA.2013.05.004>
- Bachmann, C.M., Philpot, W., Abelev, A., Korwan, D., 2014. Phase angle dependence of sand density observable in hyperspectral reflectance. *Remote Sens Environ* 150, 53–65. <https://doi.org/10.1016/J.RSE.2014.03.024>
- Bahia, A.S.R. de S., Marques, J., La Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017a. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R. de S., Marques, J., la Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017b. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R.D.S., Marques, J., Siqueira, D.S., 2015. Procedures using diffuse reflectance spectroscopy for estimating hematite and goethite in Oxisols of São Paulo, Brazil. *Geoderma Regional* 5, 150–156. <https://doi.org/10.1016/j.geodrs.2015.04.006>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021a. Soil mineralogical attributes estimated by color as accessed by proximal sensors and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021b. Soil mineralogical attributes estimated by color as accessed by proximal sensors and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Balsam, W.L., Otto-Bliesner, B.L., Deaton, B.C., 1995. Modern and Last Glacial Maximum eolian sedimentation patterns in the Atlantic Ocean interpreted from sediment iron oxide content. *Paleoceanography* 10, 493–507. <https://doi.org/10.1029/95PA00421>
- Baseca, C.C., Sendra, S., Lloret, J., Tomas, J., 2019. A smart decision system for digital farming. *Agronomy* 9. <https://doi.org/10.3390/agronomy9050216>
- Basu, D., Sinha, R., Sahu, S., Malla, J., Chakravorty, N., Ghosal, P.S., 2022. Identification of severity and passive measurement of oxidative stress biomarkers for β -thalassemia patients: K-means, random forest, XGBoost, decision tree, neural network based novel framework. *Advances in Redox Research* 5, 100034. <https://doi.org/10.1016/J.ARRES.2022.100034>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019a. Using machine

- learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582-592.
<https://doi.org/10.1016/j.compag.2019.05.008>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019b. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582-592.
<https://doi.org/10.1016/j.compag.2019.05.008>
- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015a. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1-14. <https://doi.org/10.1016/j.aca.2015.04.042>
- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015b. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1-14. <https://doi.org/10.1016/j.aca.2015.04.042>
- Bowden, C., Foster, T., Parkes, B., 2023. Identifying links between monsoon variability and rice production in India through machine learning. *Sci Rep* 13, 1-12.
<https://doi.org/10.1038/s41598-023-27752-8>
- Breiman, L., 2001. Random Forest. *Mach Learn* 45, 5-32.
<https://doi.org/10.1017/CBO9781107415324.004>
- Breiman, Leo, 2001. Random forests. *Mach Learn* 45, 5-32.
<https://doi.org/10.1023/A:1010933404324>
- Bui, D.T., Tsangaratos, P., Nguyen, V.T., Liem, N. Van, Trinh, P.T., 2020. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena (Amst)* 188.
<https://doi.org/10.1016/j.catena.2019.104426>
- CAMARGO, O.A. de, MONIZ, A.C., JORGE, J.A.J., VALADARES, J.M.A.S., 1986. Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agrônomo de Campinas, Boletim Técnico, 106. ed.
- Camargo, L.A., Marques Júnior, J., Pereira, G.T., Horvat, R.A., 2009. Variabilidade espacial de atributos mineralógicos de um Latossolo sob diferentes formas do relevo. II - Correlação espacial entre mineralogia e agregados. *Rev Bras Cienc Solo* 32, 2279-2288. <https://doi.org/10.1590/s0100-06832008000600007>
- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013a. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341-353.
<https://doi.org/10.1016/J.GEODERMA.2012.08.020>
- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013b. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341-353.
<https://doi.org/10.1016/J.GEODERMA.2012.08.020>
- Camêlo, D. de L., Ker, J.C., Fontes, M.P.F., Corrêa, M.M., da Costa, A.C.S., Melo, V.F., 2017. Pedogenic Iron Oxides in Iron-Rich Oxisols Developed from Mafic Rocks. *Rev Bras Cienc Solo* 41. <https://doi.org/10.1590/18069657RBCS20160379>
- Canizo, B. V., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514-522.
<https://doi.org/10.1016/j.compag.2019.04.043>
- Canizo, B. v., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514-522.
<https://doi.org/10.1016/j.compag.2019.04.043>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016a. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232-240.
<https://doi.org/10.1016/j.catena.2016.01.001>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016b. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232-240.
<https://doi.org/10.1016/J.CATENA.2016.01.001>
- Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7, 1247-

1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, T., Guestrin, C., 2016a. XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T., Guestrin, C., 2016b. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August-2016, 785-794. <https://doi.org/10.1145/2939672.2939785>
- Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., Vašát, R., 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. *Catena (Amst)* 189, 104485. <https://doi.org/10.1016/j.catena.2020.104485>
- Cornell, R.M., Schwertmann, U., 2006. The Iron Oxides: Structure, Properties, Reactions, Occurrences and Uses 707.
- da Costa, A.C.S., Junior, I.G. de S., Canton, L.C., Gil, L.G., Figueiredo, R., 2020. Contribution of the chemical and mineralogical properties of sandy-loam tropical soils to the cation exchange capacity. *Rev Bras Cienc Solo* 44, 1-18. <https://doi.org/10.36783/18069657rbcs20200019>
- de Castro, C.L., Braga, A.P., 2011. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica* 22, 441-466. <https://doi.org/10.1590/S0103-17592011000500002>
- de Motomiya, A.V.A., Corá, J.E., Pereira, G.T., 2006. Uso da krigagem indicatriz na avaliação de indicadores de fertilidade do solo. *Rev Bras Cienc Solo* 30, 485-496. <https://doi.org/10.1590/S0100-06832006000300010>
- Demattê, José Alexandre M, Dotto, A.C., Bedin, L.G., Sayão, V.M., Barros, A., 2019. Geoderma Soil analytical quality control by traditional and spectroscopy techniques : Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma* 337, 111-121. <https://doi.org/10.1016/j.geoderma.2018.09.010>
- Demattê, José A.M., Dotto, A.C., Paiva, A.F.S., Sato, M. V., Dalmolin, R.S.D., de Araújo, M. do S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E. V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C. de V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* 354, 113793. <https://doi.org/10.1016/j.geoderma.2019.05.043>
- Dietterich, T., 1995. Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)* 27, 326-327. <https://doi.org/10.1145/212094.212114>
- dos Santos, F.R., de Oliveira, J.F., Barbosa, G.M.C., Melquiades, F.L., 2021. Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling. *Spectrochim Acta Part B At Spectrosc* 185, 106303. <https://doi.org/10.1016/J.SAB.2021.106303>
- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008a. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009-1016. <https://doi.org/10.1590/s0100-06832008000300010>
- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008b. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009-1016. <https://doi.org/10.1590/s0100-06832008000300010>
- Ebtehaj, I., Bonakdari, H., 2022. A reliable hybrid outlier robust non-tuned rapid machine

- learning model for multi-step ahead flood forecasting in Quebec, Canada. *J Hydrol (Amst)* 614, 128592. <https://doi.org/10.1016/J.JHYDROL.2022.128592>
- EMBRAPA, 2017. Manual de metodos de análises, in: Manual de Métodos de Análise de Solo. Brasília, DF, p. 574.
- Esfandiarpour-Boroujeni, I., Shamsabadi, M.S., Shirani, H., Mosleh, Z., Bagheri Bodaghabadi, M., Salehi, M.H., 2020. Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations. *Catena (Amst)* 191, 104581. <https://doi.org/10.1016/J.CATENA.2020.104581>
- Fan, S.S., Chang, F.H., Hsueh, H.T., Ko, T.H., 2016. Measurement of Total Free Iron in Soils by H₂S Chemisorption and Comparison with the Citrate Bicarbonate Dithionite Method. *J Anal Methods Chem* 2016. <https://doi.org/10.1155/2016/7213542>
- Faria, A.J.G. de, Silva, S.H.G., Andrade, R., Mancini, M., Melo, L.C.A., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Prediction of soil organic matter content by combining data from Nix Pro™ color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional* 28, e00461. <https://doi.org/10.1016/J.GEODRS.2021.E00461>
- Fernandes, L.A., Castro, A.B. de, Basilici, G., 2007. Seismites in continental sand sea deposits of the Late Cretaceous Caiuá Desert, Bauru Basin, Brasil. *Sediment Geol* 199, 61-64. <https://doi.org/10.1016/j.sedgeo.2005.12.030>
- Fernandes, R.B.A., Barrón, V., Torrent, J., Fontes, M.P.F., 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de refletância difusa. *Rev Bras Cienc Solo* 28, 245-257. <https://doi.org/10.1590/S0100-06832004000200003>
- Fink, J.R., Inda, A. v., Bavaresco, J., Barrón, V., Torrent, J., Bayer, C., 2016. Adsorption and desorption of phosphorus in subtropical soils as affected by management system and mineralogy. *Soil Tillage Res* 155, 62-68. <https://doi.org/10.1016/J.STILL.2015.07.017>
- Fonseca, J. da S., Campos, M.C.C., Brito Filho, E.G. de, Mantovanelli, B.C., Silva, L.S., de Lima, A.F.L., Da Cunha, J.M., Simões, E.L., dos Santos, L.A.C., 2021. Soil-landscape relationship in a sandstone-gneiss topolithosequence in the State of Amazonas, Brazil. *Environ Earth Sci* 80, 1-15. <https://doi.org/10.1007/s12665-021-10026-9>
- Ganaie, M.A., Tanveer, M., Suganthan, P.N., Snasel, V., 2022. Oblique and rotation double random forest. *Neural Networks* 153, 496-517. <https://doi.org/10.1016/J.NEUNET.2022.06.012>
- Ghafari, F., Wieland, R., Lüttschwager, D., Nendel, C., 2022. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environmental Modelling & Software* 156, 105466. <https://doi.org/10.1016/J.ENVSOFT.2022.105466>
- Gholizadeh, A., Luboš, B., Saberioon, M., Vašát, R., 2013. Visible, Near-Infrared, and Mid-Infrared Spectroscopy Applications for Soil Assessment with Emphasis on Soil Organic Matter Content and Quality: State-of-the-Art and Key Issues. <http://dx.doi.org/10.1366/13-07288> 67, 1349-1362. <https://doi.org/10.1366/13-07288>
- Gholizadeh, A., Saberioon, M., Pouladi, N., Ben-Dor, E., 2023. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. *International Soil and Water Conservation Research* 11, 112-124. <https://doi.org/10.1016/J.ISWCR.2022.06.004>
- Goldshleger, N., Ben-Dor, E., Benyamini, Y., Agassi, M., 2004. SOIL REFLECTANCE AS A TOOL FOR ASSESSING PHYSICAL CRUST ARRANGEMENT OF FOUR TYPICAL SOILS IN ISRAEL. *Soil Sci* 169, 677-687. <https://doi.org/10.1097/01.ss.0000146024.61559.e2>
- Golkarian, A., Khosravi, K., Panahi, M., Clague, J.J., 2023. Spatial variability of soil water erosion: Comparing empirical and intelligent techniques. *Geoscience Frontiers* 14, 101456. <https://doi.org/10.1016/J.GSF.2022.101456>
- Gómez-Escalonilla, Diancoumba, O., Traoré, D.Y., Montero, E., Martín-Loeches, M., Martínez-Santos, P., 2022. Multiclass spatial predictions of borehole yield in southern Mali by means of machine learning classifiers. *J Hydrol Reg Stud* 44, 101245. <https://doi.org/10.1016/J.EJRH.2022.101245>
- Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. *Catena (Amst)* 147, 429-440.

- <https://doi.org/10.1016/j.catena.2016.07.045>
- Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy* 131, 1-109.
<https://doi.org/10.1016/BS.AGRON.2014.12.004>
- Hall, D.L., Llinas, J., 2016. An introduction to multi-sensor data fusion. *Sensors, Nanoscience, Biomedical Engineering, and Instruments* 85.
- Han, L., Yang, G., Yang, X., Song, X., Xu, B., Li, Z., Wu, Jintao, Yang, H., Wu, Jianwei, 2022. An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images. *Comput Electron Agric* 194, 106804. <https://doi.org/10.1016/J.COMPAG.2022.106804>
- He, M., Tang, L., Li, C., Ren, J., Zhang, L., Li, X., 2022. Dynamics of soil organic carbon and nitrogen and their relations to hydrothermal variability in dryland. *J Environ Manage* 319, 115751. <https://doi.org/10.1016/J.JENVMAN.2022.115751>
- Heuvelink, G.B.M., Brus, D.J., Rossiter, D.G., Shi, Z., 2019. Editorial for pedometrics 2017 special issue. *Eur J Soil Sci* 70, 25-26. <https://doi.org/10.1111/ejss.12772>
- Hikouei, I.S., Kim, S.S., Mishra, D.R., 2021. Machine-learning classification of soil bulk density in salt marsh environments. *Sensors* 21. <https://doi.org/10.3390/s21134408>
- Holmgren, G.G.S., 1967a. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210-211.
<https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Holmgren, G.G.S., 1967b. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210-211.
<https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Hong, H., Ji, K., Hei, H., Wang, C., Liu, C., Zhao, L., Lanson, B., Zhao, C., Fang, Q., Algeo, T.J., 2023. Clay mineral evolution and formation of intermediate phases during pedogenesis on picrite basalt bedrock under temperate conditions (Yunnan, southwestern China). *Catena (Amst)* 220, 106677.
<https://doi.org/10.1016/J.CATENA.2022.106677>
- Isaaks, E.H., Srivastava, R.M., 1989. *An Introduction to Applied Geostatistics, Computers & Geosciences*. Pergamon. [https://doi.org/10.1016/0098-3004\(91\)90055-I](https://doi.org/10.1016/0098-3004(91)90055-I)
- Jararweh, Y., Fatima, S., Jarrah, M., AlZu'bi, S., 2023. Smart and sustainable agriculture: Fundamentals, enabling technologies, and future directions. *Computers and Electrical Engineering* 110, 108799. <https://doi.org/10.1016/j.compeleceng.2023.108799>
- Jha, K., Doshi, A., Patel, P., Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture* 2, 1-12.
<https://doi.org/10.1016/J.AIIA.2019.05.004>
- Kampf, N., Schwertmann, U., 1982. The 5-M-NaOH concentration treatment for iron oxides in soils. *Clays Clay Miner* 30, 401-408.
<https://doi.org/10.1346/CCMN.1982.0300601/METRICS>
- Karthikeyan, L., Mishra, A.K., 2021. Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sens Environ* 266, 112706.
<https://doi.org/10.1016/J.RSE.2021.112706>
- Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. *Inf Sci (N Y)* 585, 609-629. <https://doi.org/10.1016/J.INS.2021.11.036>
- Kaya, F., Başıyigit, L., Keshavarzi, A., Francaviglia, R., 2022. Digital mapping for soil texture class prediction in northwestern Türkiye by different machine learning algorithms. *Geoderma Regional* 31, e00584. <https://doi.org/10.1016/J.GEODRS.2022.E00584>
- Ker, J.C., 1997. LATOSSOLOS DO BRASIL: UMA REVISÃO. *Geonomos* 5, 17-40.
<https://doi.org/10.18285/GEONOMOS.V5I1.187>
- Ker, J.C., Curi, N., Schaefer, C.E., Torrado, P.V., 2012. *Pedologia: fundamentos*. Viçosa.
- Kiangala, S.K., Wang, Z., 2021. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications* 4, 100024. <https://doi.org/10.1016/J.MLWA.2021.100024>
- Kragt, M.E., Pannell, D.J., Robertson, M.J., Thamo, T., 2012. Assessing costs of soil carbon sequestration by crop-livestock farmers in Western Australia. *Agric Syst* 112, 27-37.
<https://doi.org/10.1016/j.agsy.2012.06.005>
- Kuncheva, L.I., Matthews, C.E., Arnaiz-González, Á., Rodríguez, J.J., 2020. Feature

- Selection from High-Dimensional Data with Very Low Sample Size: A Cautionary Tale. Lark, R.M., 2000. Designing sampling grids from imprecise information on soil variability, an approach based on the fuzzy kriging variance. *Geoderma* 98, 35-59. [https://doi.org/10.1016/S0016-7061\(00\)00051-3](https://doi.org/10.1016/S0016-7061(00)00051-3)
- Lee, M.E., Jeon, E.K., Tsang, D.C.W., Baek, K., 2018. Simultaneous application of oxalic acid and dithionite for enhanced extraction of arsenic bound to amorphous and crystalline iron oxides. *J Hazard Mater* 354, 91-98. <https://doi.org/10.1016/J.JHAZMAT.2018.04.083>
- Liu, J. cheng, He, H. ping, Michalski, J., Cuadros, J., Yao, Y. zeng, Tan, W., Qin, X. rong, Li, S. ying, Wei, G. jian, 2021. Reflectance spectroscopy applied to clay mineralogy and alteration intensity of a thick basaltic weathering sequence in Hainan Island, South China. *Appl Clay Sci* 201, 105923. <https://doi.org/10.1016/J.CLAY.2020.105923>
- Liu, J., Yang, K., Tariq, A., Lu, L., Soufan, W., El Sabagh, A., 2023. Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. *Egyptian Journal of Remote Sensing and Space Science* 26, 415-426. <https://doi.org/10.1016/j.ejrs.2023.05.005>
- Liu, Q., He, L., Guo, L., Wang, M., Deng, D., Lv, P., Wang, R., Jia, Z., Hu, Z., Wu, G., Shi, T., 2022. Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. *Catena (Amst)* 219, 106603. <https://doi.org/10.1016/J.CATENA.2022.106603>
- Liu, X., Zhu, A.X., Yang, L., Pei, T., Liu, J., Zeng, C., Wang, D., 2020. A graded proportion method of training sample selection for updating conventional soil maps. *Geoderma* 357, 113939. <https://doi.org/10.1016/j.geoderma.2019.113939>
- Lorz, C., Fruhauf, M., Mailander, R., Phillips, J.D., Kleber, A., 2013. Influence of Cover Beds on Soils, Mid-Latitude Slope Deposits (Cover Beds).
- Lu, W.N., He, Y., Wang (王阳), Y., Ke, S., 2020. Behavior of calcium isotopes during continental subduction recorded in meta-basaltic rocks. *Geochim Cosmochim Acta* 278, 392-404. <https://doi.org/10.1016/J.GCA.2019.09.027>
- Ma, G., Ding, J., Han, L., Zhang, Z., Ran, S., 2021. Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. *Regional Sustainability* 2, 177-188. <https://doi.org/10.1016/J.REGSUS.2021.06.001>
- Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma* 262, 243-253. <https://doi.org/10.1016/J.GEODERMA.2015.08.037>
- Mancini, M., Andrade, R., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for Brazilian soil properties prediction: Exchangeable/available macronutrients, aluminum, and potential acidity. *Geoderma Regional* 30, e00573. <https://doi.org/10.1016/J.GEODRS.2022.E00573>
- Marques, J., Siqueira, D.S., Camargo, L.A., Teixeira, D.D.B., Barrón, V., Torrent, J., 2014. Magnetic susceptibility and diffuse reflectance spectroscopy to characterize the spatial variability of soil properties in a brazilian haplustalf. *Geoderma* 219-220, 63-71. <https://doi.org/10.1016/j.geoderma.2013.12.007>
- Maurya, A.K., Bhargava, N., Singh, D., 2022a. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795-1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- Maurya, A.K., Bhargava, N., Singh, D., 2022b. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795-1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- McBratney, A., Gruijter, J., Bryce, A., 2019. Pedometrics timeline. *Geoderma* 338, 568-575. <https://doi.org/10.1016/j.geoderma.2018.11.048>
- McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, *Geoderma*. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- McKeague, J.A., Day, J.H., 1966. DITHIONITE- AND OXALATE-EXTRACTABLE Fe AND Al AS AIDS IN DIFFERENTIATING VARIOUS CLASSES OF SOILS. *Can J Soil Sci* 46, 13-22. <https://doi.org/10.4141/cjss66-003>
- Mehra, O.P., Jackson, M.L., 1958. Iron Oxide Removal from Soils and Clays by a Dithionite-Citrate System Buffered with Sodium Bicarbonate. *Clays and Clay Minerals* 1958 7:1 7,

- 317-327. <https://doi.org/10.1346/CCMN.1958.0070122>
- Mello, F.A.O., Bellinaso, H., Mello, D.C., Safanelli, J.L., Mendes, W.D.S., Amorim, M.T.A., Gomez, A.M.R., Poppiel, R.R., Silvero, N.E.Q., Gholizadeh, A., Silva, S.H.G., Curi, N., Demattê, J.A.M., 2021. Soil parent material prediction through satellite multispectral analysis on a regional scale at the Western Paulista Plateau, Brazil. *Geoderma Regional* 26, e00412. <https://doi.org/10.1016/J.GEODRS.2021.E00412>
- Mendes, W. de S., Demattê, J.A.M., Bonfatti, B.R., Resende, M.E.B., Campos, L.R., Costa, A.C.S. da, 2021. A novel framework to estimate soil mineralogy using soil spectroscopy. *Applied Geochemistry* 127, 104909. <https://doi.org/10.1016/J.APGEOCHEM.2021.104909>
- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022a. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219. <https://doi.org/10.1016/j.still.2022.105346>
- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022b. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219, 105346. <https://doi.org/10.1016/J.STILL.2022.105346>
- Merl, T., Rasmussen, M.R., Koch, L.R., Søndergaard, J.V., Bust, F.F., Koren, K., 2022. Measuring soil pH at in situ like conditions using optical pH sensors (pH-optodes). *Soil Biol Biochem* 175, 108862. <https://doi.org/10.1016/J.SOILBIO.2022.108862>
- Minasny, B., Hartemink, A.E., 2011a. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52-62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., Hartemink, A.E., 2011b. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52-62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. *Geoderma* 264, 301-311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019a. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194, 104308. <https://doi.org/10.1016/j.still.2019.104308>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019b. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194. <https://doi.org/10.1016/j.still.2019.104308>
- Naimi, S., Ayoubi, S., di Raimo, L.A.D.L., Dematte, J.A.M., 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. *Geoderma Regional* 28, e00484. <https://doi.org/10.1016/J.GEODRS.2022.E00484>
- Nawar, S., Delbecq, N., Declercq, Y., Smedt, P. De, Finke, P., Verdoodt, A., Meirvenne, M. Van, Mouazen, A.M., 2019. Geoderma Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry? *Geoderma* 350, 29-39. <https://doi.org/10.1016/j.geoderma.2019.05.002>
- Nawar, S., Richard, F., Kassim, A.M., Tekin, Y., Mouazen, A.M., 2022. Fusion of Gamma-rays and portable X-ray fluorescence spectral data to measure extractable potassium in soils. *Soil Tillage Res* 223, 105472. <https://doi.org/10.1016/J.STILL.2022.105472>
- Ng, W., Minasny, B., McBratney, A., 2020a. Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Mcbratney, A., 2020b. Science of the Total Environment Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352, 251-267. <https://doi.org/10.1016/j.geoderma.2019.06.016>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022a. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data

- fusion. *Science of the Total Environment* 804. <https://doi.org/10.1016/j.scitotenv.2021.150187>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022b. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of The Total Environment* 804, 150187. <https://doi.org/10.1016/J.SCITOTENV.2021.150187>
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy* 132, 139-159. <https://doi.org/10.1016/bs.agron.2015.02.002>
- Norrish, K., Taylor, R.M., 1961. The isomorphous replacement of iron by aluminium in soil goethites. *Journal of Soil Science* 12, 294-306. <https://doi.org/10.1111/J.1365-2389.1961.TB00919.X>
- Novais, R.F., Alvarez V., V.H., Barros, N.F. de, Fontes, R.L.F., Cantarutti, R.B., Neves, J.C.L., 2007. Fertilidade do solo. Viçosa.
- Nowak-Brzezinska, A., Horyn, C., 2020. Outliers in rules - the comparison of LOF, COF and KMEANS algorithms. *Procedia Comput Sci* 176, 1420-1429. <https://doi.org/10.1016/J.PROCS.2020.09.152>
- Oliveira, C. v., Ker, J.C., Fontes, L.E.F., Curi, N., Pinheiro, J.C., 1998. Química e mineralogia de solos derivados de rochas do Grupo Bambuí no norte de Minas Gerais. *Rev Bras Cienc Solo* 22, 583-593. <https://doi.org/10.1590/S0100-06831998000400003>
- Oliver, M.A., Webster, R., 2014a. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56-69. <https://doi.org/10.1016/J.CATENA.2013.09.006>
- Oliver, M.A., Webster, R., 2014b. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56-69. <https://doi.org/10.1016/j.catena.2013.09.006>
- Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16. <https://doi.org/10.1016/j.geodrs.2018.e00198>
- Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* 12, 2825-2830. <https://doi.org/10.5555/1953048.2078195>
- Pedron, F. de A., Lourenzi, C.R., Ceretta, C.A., Lorensi, J., Cancian, A., 2018. Clay mineralogy of subtropical soils under long-term organic fertilization in no-tillage systems. *Rev Bras Cienc Solo* 42. <https://doi.org/10.1590/18069657rbcs20170092>
- Ramaroson, V.H., Becquer, T., Sá, S.O., Razafimahatratra, H., Delarivière, J.L., Blavet, D., Vendrame, P.R.S., Rabeharisoa, L., Rakotondrazafy, A.F.M., 2018. Mineralogical analysis of ferralitic soils in Madagascar using NIR spectroscopy. *Catena (Amst)* 168, 102-109. <https://doi.org/10.1016/j.catena.2017.07.016>
- Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L., El Hadrami, E.M., 2019. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems* 195. <https://doi.org/10.1016/j.chemolab.2019.103873>
- Rennert, T., 2019. Wet-chemical extractions to characterise pedogenic Al and Fe species-a critical review. *Soil Research* 57, 1-16. <https://doi.org/10.1071/SR18299>
- Rosin, N.A., Demattê, J.A.M., Leite, M.C.A., de Carvalho, H.W.P., Costa, A.C., Greschuk, L.T., Curi, N., Silva, S.H.G., 2022. The fundamental of the effects of water, organic matter, and iron forms on the pXRF information in soil analyses. *Catena (Amst)* 210. <https://doi.org/10.1016/j.catena.2021.105868>
- Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics.

- Geoderma 324, 131-137. <https://doi.org/10.1016/j.geoderma.2018.03.009>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022a. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022b. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruiz, H.A., Ferreira, G.B., Pereira, J.B.M., Resumo,), 2003. Estimativa da capacidade de campo de Latossolos e Neossolos Quartzarênicos pela determinação do equivalente de umidade. *Rev Bras Cienc Solo* 27, 389-393. <https://doi.org/10.1590/S0100-06832003000200019>
- Santos, H.G. dos, Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.Á. de, Lumbrreras, J.F., Coelho, M.R., Almeida, J.A. de, Filho, J.C. de A., Oliveira, J.B. de, Cunha, T.J.F., 2018. Sistema brasileiro de classificação de solos, 5th ed, Embrapa Solos. Embrapa, Brasília.
- Sarfraz, S., Ali, F., Hameed, A., Ahmad, Z., Riaz, K., 2023. Sustainable Agriculture Through Technological Innovations. *Sustainable Agriculture in the Era of the OMICs Revolution* 223-239. https://doi.org/10.1007/978-3-031-15568-0_10
- Schwertmann, U., Taylor, R.M., 2018. Iron oxides. *Minerals in Soil Environments* 379-438. <https://doi.org/10.2136/sssabookser1.2ed.c8>
- Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S. V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep* 11, 1-15. <https://doi.org/10.1038/s41598-020-80820-1>
- Shrestha, G., Calvelo-Pereira, R., Roudier, P., Martin, A.P., Turnbull, R.E., Kereszturi, G., Jeyakumar, P., Anderson, C.W.N., 2022. Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy. *Geoderma* 409, 115649. <https://doi.org/10.1016/J.GEODERMA.2021.115649>
- Silva, F.M., Silva, S.H.G., Acuña-Guzman, S.F., Silva, E.A., Ribeiro, B.T., Fruett, T., Inda, A.V., Teixeira, A.F. dos S., Mancini, M., Guilherme, L.R.G., Curi, N., 2021. Chemical and mineralogical changes in the textural fractions of quartzite-derived tropical soils, along weathering, assessed by portable X-ray fluorescence spectrometry and X-ray diffraction. *J South Am Earth Sci* 112, 103634. <https://doi.org/10.1016/J.JSAMES.2021.103634>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022a. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, 1-10. <https://doi.org/10.1016/j.geodrs.2021.e00465>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022b. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, e00465. <https://doi.org/10.1016/J.GEODRS.2021.E00465>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020a. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/j.catena.2019.104258>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020b. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/J.CATENA.2019.104258>
- Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. *Geoderma* 267, 65-77. <https://doi.org/10.1016/J.GEODERMA.2015.12.025>
- Silvero, N.E.Q., Siqueira, D.S., Coelho, R.M., da Costa Ferreira, D., Marques, J., 2019. Protocol for the use of legacy data and magnetic signature on soil mapping of São Paulo Central West, Brazil. *Science of The Total Environment* 693, 133463. <https://doi.org/10.1016/J.SCITOTENV.2019.07.269>

- Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl Spectrosc Rev* 49, 139-186. <https://doi.org/10.1080/05704928.2013.811081>
- Sothe, C., Almeida, C.M. De, Schimalski, M.B., Liesenberg, V., Rosa, L.E.C. La, Castro, J.D.B., Feitosa, R.Q., 2019. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. <https://doi-org.ez87.periodicos.capes.gov.br/10.1080/01431161.2019.1681600>. <https://doi.org/10.1080/01431161.2019.1681600>
- Sothe, C., de Almeida, C.M., Schimalski, M.B., Liesenberg, V., la Rosa, L.E.C., Castro, J.D.B., Feitosa, R.Q., 2020. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. *Int J Remote Sens* 41, 1943-1969. <https://doi.org/10.1080/01431161.2019.1681600>
- Souza, E.D., Carneiro, M.A.C., Paulino, H.B., 2005. Atributos físicos de um Neossolo Quartzarênico e um Latossolo Vermelho sob diferentes sistemas de manejo. *Pesqui Agropecu Bras* 40, 1135-1139. <https://doi.org/10.1590/S0100-204X2005001100012>
- Stranghoener, M., Dultz, S., Behrens, H., Schippers, A., 2020. Potential mobilizable Fe from secondary phases of differentially altered subsurface basaltic rock- a sequential extraction study on ICDP site Hawaii. *Applied Geochemistry* 121, 104705. <https://doi.org/10.1016/J.APGEOCHEM.2020.104705>
- Suits, D.B., 1957. Use of Dummy Variables in Regression Equations. *J Am Stat Assoc* 52, 548. <https://doi.org/10.2307/2281705>
- Sumathi, K., 2018. Data Analytics platform for intelligent agriculture. 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on 647-650. <https://doi.org/10.1109/I-SMAC.2018.8653740>
- Sun, Z., Wang, J., Wang, Y., Long, L., Luo, Z., Deng, X., Hu, Q., Wang, M., 2020. Sodium-rich volcanic rocks and their relationships with iron deposits in the Aqishan-Yamansu belt of Eastern Tianshan, NW China. *Geoscience Frontiers* 11, 697-713. <https://doi.org/10.1016/J.GSF.2019.06.011>
- Swetha, R.K., Dasgupta, S., Chakraborty, S., Li, B., Weindorf, D.C., Mancini, M., Silva, S.H.G., Ribeiro, B.T., Curi, N., Ray, D.P., 2022. Using Nix color sensor and Munsell soil color variables to classify contrasting soil types and predict soil organic carbon in Eastern India. *Comput Electron Agric* 199, 107192. <https://doi.org/10.1016/J.COMPAG.2022.107192>
- Tabatabai, S., Knadel, M., Thomsen, A., Greve, M.H., 2019. On-the-Go Sensor Fusion for Prediction of Clay and Organic Carbon Using Pre-processing Survey, Different Validation Methods, and Variable Selection. *Soil Science Society of America Journal* 83, 300-310. <https://doi.org/10.2136/SSSAJ2018.10.0377>
- Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma* 428, 116192. <https://doi.org/10.1016/J.GEODERMA.2022.116192>
- Teixeira, A.F. dos S., Andrade, R., Mancini, M., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *J South Am Earth Sci* 116, 103873. <https://doi.org/10.1016/J.JSAMES.2022.103873>
- Teixeira, A.F. dos S., Henrique Procópio Pelegrino, M., Missina Faria, W., Henrique Godinho Silva, S., Gabriela Marcolino Gonçalves, M., Weimar Acerbi Júnior, F., Rezende Gomide, L., Linares Pádua Júnior, A., de Souza, I.A., Chakraborty, S., Weindorf, D.C., Roberto Guimarães Guilherme, L., Curi, N., 2020. Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. *Geoderma* 361, 114132. <https://doi.org/10.1016/j.geoderma.2019.114132>
- Teixeira, R.S., Vendrame, P.R.S., Christoni, A.R.F., SILVA, P.R.C., 2016. QUÍMICA E MINERALOGIA DE UM SOLO DESENVOLVIDO DE BASALTO, COLETADO ATRAVÉS DE SONDAGEM SPT. *Boletim de Geografia* 34, 116-126. <https://doi.org/10.4025/bolgeogr.v34i2.28948>
- Temme, A.J.A.M., Veldkamp, A., 2009. Multi-process Late Quaternary landscape evolution

- modelling reveals lags in climate response over small spatial scales. *Earth Surf Process Landf* 34, 573-589. <https://doi.org/10.1002/ESP.1758>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011a. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812-816. <https://doi.org/10.1590/S0103-84782011000500012>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011b. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812-816. <https://doi.org/10.1590/S0103-84782011000500012>
- Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital Soil Mapping: Interactions with and Applications for Hydropedology, in: *Hydropedology*. Elsevier, pp. 665-709. <https://doi.org/10.1016/B978-0-12-386941-8.00021-6>
- Thorntwaite, C.W., 1948. An Approach toward a Rational Classification of Climate Author (s): C . W . Thorntwaite Published by: American Geographical Society Stable URL : <http://www.jstor.org/stable/210739> Accessed: 11-08-2016 18: 02 UTC. *Geogr Rev* 38, 55-94. <https://doi.org/10.2307/210739>
- Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* 367. <https://doi.org/10.1016/j.geoderma.2020.114208>
- Vasques, G.M., Rodrigues, H.M., Coelho, M.R., Baca, J.F.M., Dart, R.O., Oliveira, R.P., Teixeira, W.G., Ceddia, M.B., 2020. Field Proximal Soil Sensor Fusion for Improving High-Resolution Soil Property Maps. *Soil Systems* 2020, Vol. 4, Page 52 4, 52. <https://doi.org/10.3390/SOILSYSTEMS4030052>
- Vieira, S.R., 2000. Geoestatística em estudos de variabilidade espacial do solo. , in: R.F. Novais, V.H.A.G.R.S. (Ed.), *Tópicos Em Ciência Do Solo*. Viçosa, pp. 1-54.
- Wang, J.H., Jiang, J.H., Yu, R.Q., 1996. Robust back propagation algorithm as a chemometric tool to prevent the overfitting to outliers. *Chemometrics and Intelligent Laboratory Systems* 34, 109-115. [https://doi.org/10.1016/0169-7439\(96\)00005-6](https://doi.org/10.1016/0169-7439(96)00005-6)
- Wang, S., Peng, H., Hu, Q., Jiang, M., 2022. Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *J Hydrol Reg Stud* 42, 101139. <https://doi.org/10.1016/J.EJRH.2022.101139>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014a. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, *Advances in Agronomy*. Elsevier. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014b. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, in: *Advances in Agronomy*. Academic Press Inc., pp. 1-45. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Paulette, L., Man, T., 2013. In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. *Environmental Pollution* 182, 92-100. <https://doi.org/10.1016/j.envpol.2013.07.008>
- Whittig, L.D., 1965. X-Ray Diffraction Techniques for Mineral Identification and Mineralogical Composition. *Methods of Soil Analysis, Part 1: Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling* 671-698. <https://doi.org/10.2134/AGRONMONOGR9.1.C49>
- Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. *Environmental Pollution* 263, 114649. <https://doi.org/10.1016/J.ENVPOL.2020.114649>
- Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. *Eur J Soil Sci* 70, 162-173. <https://doi.org/10.1111/EJSS.12729>
- Xu, G., Fan, H., Oliver, D.M., Dai, Y., Li, H., Shi, Y., Long, H., Xiong, K., Zhao, Z., 2022. Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model. *Environ Res* 214, 113843. <https://doi.org/10.1016/J.ENVRES.2022.113843>
- Xu, S., Zhao, Y., Wang, M., Shi, X., 2018. Quantification of Different Forms of Iron from Intact Soil Cores of Paddy Fields with Vis-NIR Spectroscopy. *Soil Science Society of America Journal* 82, 1497-1511. <https://doi.org/10.2136/SSSAJ2018.01.0014>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020a. Depth-to-bedrock map of China at a

- spatial resolution of 100 meters. *Sci Data* 7, 1-13. <https://doi.org/10.1038/s41597-019-0345-6>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020b. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Scientific Data* 2020 7:1 7, 1-13. <https://doi.org/10.1038/s41597-019-0345-6>
- Ye, M., Zhu, L., Li, X., Ke, Y., Huang, Y., Chen, B., Yu, H., Li, H., Feng, H., 2023. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. *Science of The Total Environment* 858, 159798. <https://doi.org/10.1016/J.SCITOTENV.2022.159798>
- Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G.A., Nannucci, M.S., Leoni, B., Bonomi, T., 2019. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. *J Hydrol (Amst)* 578, 124015. <https://doi.org/10.1016/J.JHYDROL.2019.124015>
- Zeraatpisheh, M., Ayoubi, S., Jafari, A., Finke, P., 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology* 285, 186-204. <https://doi.org/10.1016/J.GEOMORPH.2017.02.015>
- Zevenbergen, C., Honders, A., Orbons, A.J., Viaene, W., Swennen, R., Comans, R.N.J., van Hasselt, H.J., 1997. Immobilisation of heavy metals in contaminated soils by thermal treatment at intermediate temperatures. *Studies in Environmental Science* 71, 661-672. [https://doi.org/10.1016/S0166-1116\(97\)80249-X](https://doi.org/10.1016/S0166-1116(97)80249-X)
- Zhan, Y., Liu, C., Deng, Q., Feng, Q., Qiu, Y., Zhang, A., He, X., 2022. Integrated FFT and XGBoost framework to predict pavement skid resistance using automatic 3D texture measurement. *Measurement* 188, 110638. <https://doi.org/10.1016/J.MEASUREMENT.2021.110638>
- Zhang, R., Li, Y., Goh, A.T.C., Zhang, W., Chen, Z., 2021. Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models. *Journal of Rock Mechanics and Geotechnical Engineering* 13, 1478-1484. <https://doi.org/10.1016/J.JRMGE.2021.08.001>
- Zhang, W. chun, Wan, H. shuang, Zhou, M. hou, Wu, W., Liu, H. bin, 2022. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecol Indic* 143, 109420. <https://doi.org/10.1016/J.ECOLIND.2022.109420>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022a. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022b. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022a. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022b. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>

Capítulo 2 - Fusão de dados multissensor e aprendizagem de máquina na estimativa de atributos mineralógicos em solos do Planalto Ocidental Paulista.

Resumo: A análise dos atributos do solo é de extrema importância para a agricultura. Porém, os métodos de quantificação tradicionais são de altos custos, morosos e geram resíduos, muitas vezes poluentes. A espectroscopia de refletância difusa (ERD), a espectrometria de fluorescência de raios-x (EFRX) e análise de susceptibilidade magnética do solo (SMBx) associados a técnicas de aprendizagem de máquina têm sido utilizadas como alternativas viáveis na determinação de atributos do solo. Neste sentido, o objetivo deste trabalho foi estimar atributos mineralógicos do solo, utilizando-se de técnicas de fusão dos dados obtidos pelos sensores ERD, EFRX e SMBx e algoritmos de aprendizagem de máquina, em solos de diferentes geologias no Planalto Ocidental Paulista. Esta área ocupa cerca de 13 milhões de hectares, onde foram coletadas amostras de solo em 359 pontos em grid irregular. Foram executadas as análises de ERD, EFRX e SMBx, e determinação dos atributos mineralógicos do solo, por metodologias convencionais. Após, os dados foram submetidos a dois níveis de fusão e, em seguida, submetidos às análises de Stepwise e XGBoost para treino, teste e validação dos modelos. Por fim, com o auxílio da geoestatística foram confeccionados mapas da área com os dados obtidos pelo método convencional e preditos pelos modelos. A técnica de “limpeza” das variáveis de entrada por meio da stepwise, permitiu que o resultado do modelo de aprendizagem fosse aprimorado. Foram apresentados os melhores resultados das métricas de avaliação, utilizando apenas cerca de 2% do volume original de dados. A variabilidade geológica impacta diretamente nos dados obtidos pelos sensores e nas previsões. A fusão de dados foi eficiente na estimativa dos atributos mineralógicos. Individualmente os dados obtidos por EFRX foram mais eficientes nas previsões.

Palavras-chave: Pedometria; Mineralogia do solo; Banco de dados amplos; Inteligência artificial; Agricultura digital.

Abstract: The analysis of soil attributes is extremely important for agriculture. However, traditional quantification methods are costly, time-consuming and generate waste, which is often polluting. Diffuse reflectance spectroscopy (ERD), x-ray fluorescence spectrometry (EFRX) and soil magnetic susceptibility analysis (SMBx) have been used as viable alternatives in the determination of soil attributes. In this sense, the objective of this work was to estimate soil mineralogical attributes, using data fusion techniques obtained by ERD, EFRX and SMBx sensors and machine learning algorithms, in soils of different geologies in the Western Plateau of São Paulo. This area occupies approximately 13 million hectares, where soil samples were collected at 359 points in an irregular grid. ERD, EFRX and SMBx analyzes were carried out, as well as determination of soil mineralogical attributes, by conventional methodologies. Afterwards, the data were submitted to two levels of fusion and then submitted to Stepwise and XGBoost analyzes for training, testing and validation of the models. Finally, with the help of geostatistics, maps of the area were made with data obtained by the conventional method and predicted by the models. The technique of “cleaning” the input variables through stepwise allowed the result of the learning model to be improved. The best results of the evaluation metrics were presented, using only about 2% of the original volume of data. Geological variability directly impacts the data obtained by sensors and predictions. Data fusion was efficient in estimating mineralogical attributes. Individually, the data obtained by EFRX were more efficient in the predictions.

Keywords: Pedometry; Soil mineralogy; Large database; Artificial intelligence; Digital agriculture.

2.1 Introdução

A distinta frequência e a intensidade de fatores e processos de formação do solo em zonas tropicais comumente é expressando nas propriedades morfológicas, químicas, físicas e mineralógicas do solo (Fonseca et al., 2021). Resultando na frequente presença de solos intensamente intemperizados nessas regiões (Minasny e Hartemink, 2011b).

Como resultado, os principais minerais de ocorrência na fração argila desses solos são óxi-hidróxidos de alumínio e ferro, como a gibbsita (Gb), hematita (Hm), goethita (Gt) e o aluminossilicato caulinita (Ct) (Eberhardt et al., 2008b). Segundo (Baldo et al., 2021a) esses óxidos apresentam diversas constituições e formas, além de proporcionar uma ampla variação na cor do solo, agregação das partículas, retenção de cátions e ânions e o buffer de elétrons e prótons. Sendo assim influenciam diretamente na resposta do solo às práticas de manejo, mesmo quando presentes em concentrações muito pequenas (Schwertmann and Taylor, 2018).

Os métodos tradicionais de caracterização da mineralogia como: a difração de raios X (DRX) que avalia a estrutura dos cristais no solo (Whittig, 1965), extração por ditionito–cittrato-bicarbonato (FED) que consiste na dissolução redutiva por ditionita, onde íons de Fe^{2+} e Al^{3+} são reduzidos e liberados em solução (Holmgren, 1967b) e na extração por amônio-ácido oxálico (FEO) que induz a formação complexos solúveis entre o oxalato e os íons Al e Fe e ao atingir o $\text{pH} < 3.5$ ocorre a adsorção de oxalato e liberação de íons Al^{3+} e Fe^{3+} (McKeague e Day, 1966) apresentam limitações em sua aplicação.

Dentre essas limitações, destaca-se: a necessidade de técnicos que dominem os métodos para realização das análises, longo tempo para execução, alta periculosidade e por fim, apresentando alto custo (Mendes et al., 2022a). Por consequências tais análises se limitam ao meio acadêmico

A fim de obter um método efetivo para caracterização da mineralogia, a pedometria vem baseando seus avanços no uso de metodologias práticas, de baixo custo e de fácil acesso, como o uso de sensores. A associação de dados obtidos por sensores de espectroscopia de reflectância difusa (ERD), espectrometria de fluorescência de raios-x (EFRX) e susceptibilidade magnética (SMBx) já vem sendo aplicadas com sucesso em várias subáreas da ciência do solo (Bahia et al., 2017b, 2015; Marques et al., 2014; Silva et al., 2020b; Silvero et al., 2019; Weindorf et al., 2014b). Na literatura, já podem ser observadas predições de atributos mineralógicos e físicos como a textura, e químicos (pH, CTC, entre outros), pela fusão dos sensores (Padarian et al., 2019; Tsakiridis et al., 2020) ou pelo uso destes de forma individual (Almeida et al., 2021; Bahia et al., 2017b, 2015; Baldo et al., 2021a; Silva et al., 2020b).

O que há em comum nos dados obtidos a partir desses três sensores (SMBx, EFRX e ERD) é a especificidade e o grande volume de informações geradas. Portanto, são necessárias técnicas de modelagem matemáticas mais robustas, como as técnicas relacionadas ao aprendizado de máquina (AM), que vem ganhando destaque neste âmbito (Almeida et al., 2021). Estas técnicas permitem modelar as relações não lineares dos dados, algo que os métodos estatísticos tradicionais lineares não permitiam (Bekker et al., 2019b).

A AM consiste em utilizar algoritmos matemáticos capazes de processar (aprender) em sistemas, por meio de análises de regressão e de classificação, à medida que são submetidos a treinamentos (Canizo et al., 2019). Dentre os algoritmos de AM o eXtreme Gradient Boosting (XGBoost), algoritmo de última geração, fornece melhores resultados em uma ampla gama de problemas (Chen e Guestrin, 2016). Seu design reduz com sucesso a complexidade da estrutura de uma rede neural e evita o overfitting (Zhan et al., 2022).

Neste contexto, levanta-se a hipótese de que o uso de dados obtidos por sensores (com baixo custo e rapidez) associados a diferentes níveis de fusão de dados e algoritmos de AM podem aprimorar a estimativa de atributos do solo, permitindo a confecção de mapas de manejo específico de alta precisão. Portanto, o objetivo deste trabalho foi estimar atributos mineralógicos (FED e FEO), utilizando diferentes níveis de fusão de dados e o algoritmo XGBoost, a partir de dados obtidos dos sensores ERD, EFRX e SMBx, em solos de diferentes formações geológicas no Planalto Ocidental Paulista.

2.2 Material e métodos

2.2.1 Caracterização da área de estudo

Neste trabalho utilizou-se um banco de dados legados, cedido pelo grupo de pesquisa Caracterização do Solo para Fins de Manejo Específico (CSME) da Unesp-Jaboticabal. O banco de dados refere-se à uma área do Planalto Ocidental Paulista (POP) que corresponde a aproximadamente 48% de todo o Estado de São Paulo, totalizando 13 milhões de hectares (Figura 1a). A área produz cerca de 80% da produção nacional de etanol, açúcar e citros e já vem sendo estudada por vários autores do grupo CSME (Silvero et al., 2019).

A geologia do POP é constituída por dois tipos de rochas: magmática e sedimentares. A rocha magmática extrusiva corresponde a formação Serra Geral. Enquanto as sedimentares constituem-se de várias formações geológicas: Grupo Caiuá composto pelas Formações Santo Anastácio e Rio Paraná no estado de São Paulo, correspondente a depósitos de lençóis de areia, de clima seco, acumulados em extensas e monótonas planícies desérticas, marginais dos grandes complexos de dunas do sand sea (Deserto Caiuá) que se estende para região norte do estado do Paraná; e o Grupo Bauru é composto pelas formações Uberaba, Vale do Rio do Peixe, Araçatuba, São José do Rio Preto, Presidente Prudente e Marília (Fernandes

et al., 2007, 2004).

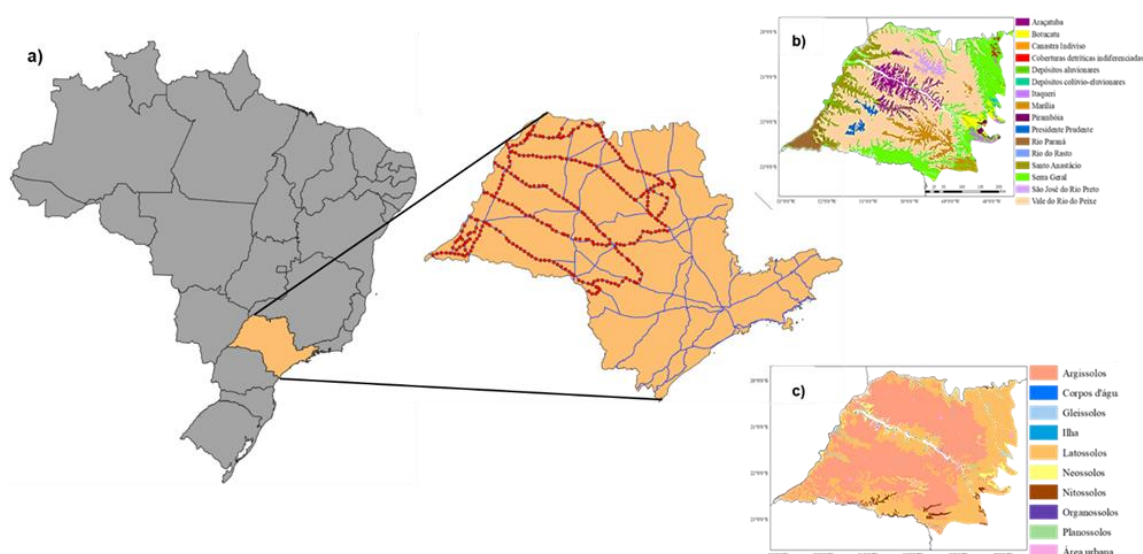


Figura 1. Mapa de localização da área, com destaque para a malha irregular estabelecida no estudo (a), Mapa de distribuição da formação geológica (b) e Mapa da distribuição das classes de solo (c).

Segundo o mapa de solos do Instituto Agrônomo de Campinas, seis classes de solo foram registradas na área de estudo: Argissolo Vermelho-Amarelo, Latossolo Vermelho, Latossolo férrico, Neossolo Litólico, Nitossolo Vermelho e Gleissolo Háplico (Figura 1c). Segundo o padrão geológico os solos de geologia basáltica apresentam maiores conteúdos de Gb e Hm, enquanto os solos originados de arenito apresentam os maiores conteúdos de Ct, conforme pesquisas já publicadas para a área de estudo (Silva et al., 2020b).

Seguindo a classificação de Thornthwaite, o POP apresenta o clima tropical com inverno seco predominante no norte e noroeste, o clima temperado úmido, com verão quente, prevalecendo no sul e o clima temperado úmido com inverno seco e verão quente no leste e sudeste (Thornthwaite, 1948).

2.2.2 Plano amostral de coleta de pontos

Para a amostragem levou-se em consideração a base do arquivo de rodovias do Estado de São Paulo, fornecido pelo Departamento de Estradas de Rodagem (DER). Com isso, criou-se um plano, por meio da ferramenta ET GeoWizards, no

ArcView 9.3. Essa amostragem promoveu representação espacial na área de estudo, especialmente nos compartimentos geológicos e pedológicos do POP (Figura 1a). O total de 359 amostras de solo foram coletadas em áreas preservadas adjacentes às áreas agrícolas com mínima interferência antrópica, tendo como alinhamento as rodovias (Estado SP), em um espaçamento mínimo de 10 quilômetros (km), até o espaçamento máximo de 60 km.

2.2.3 Análises de laboratório

Os atributos mineralógicos, foco deste estudo, são os teores de ferro cristalinos extraídos por ditionito-citrato (FED), que foi determinado seguindo a metodologia de (Mehra and Jackson, 1958). E os teores de ferro extraídos por amônio-ácido oxálico (FEO), relativos aos óxidos de ferro pedogenéticos de baixa cristalinidade, cuja determinação seguiu a metodologia citada por (Camargo et al., 1986).

Para fins de caracterização da área foram citados os teores de Gt e Hm, obtidos utilizando a difratometria de raios-X (DRX), após a separação da argila pelo método de centrifugação (Mehra e Jackson, 1958) e tratamento com NaOH 5 mol L⁻¹ (1g argila/100 ml solução) para a concentração dos mesmos, segundo método de Norrish e Taylor (1961) modificado por Kampf e Schwertmann (1982).

E para determinação da granulometria realizou-se o método da pipeta, com uma solução de NaOH 0,1 mol L⁻¹ como dispersante químico e agitação mecânica em aparato de baixa rotação, por 16 horas, seguindo metodologia preconizada (EMBRAPA, 2017).

2.2.4 Sensores

A análise de ERD foi feita utilizando um sensor Lambda 950 UV/VIS/NIR. As amostras de 0,5 g de terra fina seca ao ar (TFSA) foram acomodadas em porta amostra individual (3,0 × 0,5 cm). Efetuou-se em cada amostra uma varredura de

reflectância na faixa do Visível (VIS), entre 250 e 900 nanômetros (nm) e do infravermelho próximo (NIR) entre 900 e 2500 nanômetros (nm), em um intervalo de 1 nm gerando um banco de dados de 2250 variáveis por amostra de solo.

Para a análise de EFRX foram utilizadas 2 gramas de TFSA, peneirados e homogeneizados. As amostras foram escaneadas no equipamento modelo NexQc e irradiadas na atmosfera do ar, utilizando um tubo de raios X Rh operado na corrente 1 (C1) 15 kV, que efetua a leitura das vibrações que indicam a presença de sódio a escândio (considerando a tabela periódica) e na corrente 2 (C2) 50 kV, que efetua a leitura das vibrações que indicam a presença de titânio a urânio, por 200s e 250s, respectivamente.

A análise de SMBx foi determinada utilizando-se de 10 g de TFSA, utilizando o equipamento Bartington MS2, acoplado ao sensor Bartington MS2B em baixa frequência (0,47 kHz) e foi gerado o total de uma variável por amostra de solo.

2.2.5 Fusão de Dados

Os dados gerados pelos três sensores: ERD (VIS e NIR), EXRF (C1 e C2) e SMBx, foram combinados de diversas formas, com o objetivo de aprimorar as predições e encontrar, assim, a combinação que gera resultados mais precisos nas estimativas de FED e FEO, obtendo o total de 17 combinações (Tabela 1). No presente trabalho foram testados dois níveis de fusão: a de nível baixo (LLF), que consiste na concatenação dos dados gerados pelas três fontes em um único banco de dados, apresentando na linha as amostras de solo analisadas e nas colunas as diferentes variáveis geradas por cada um dos três sensores. Neste nível foi gerado um único banco de dados de 6348 variáveis passando a ser um banco de dados amplo que se caracteriza por um alto número de variáveis e um baixo número de amostras (Borràs et al., 2015b).

Tabela 1. Níveis de fusão de dados e as distintas combinações de sensores testados na pesquisa.

Nível de Fusão	Banco de Dados Gerado
Nível Médio	VIS
Nível Médio	NIR
Nível Médio	C1
Nível Médio	C2
Nível Médio	SBMx
Nível Médio	VIS+NIR
Nível Médio	VIS + C1
Nível Médio	VIS + C2
Nível Médio	VIS + SBMx
Nível Médio	NIR + C1
Nível Médio	NIR + C2
Nível Médio	NIR + SBMx
Nível Médio	C1 + C2
Nível Médio	C1 + SBMx
Nível Médio	C2 + SBMx
Nível Médio	Variáveis relevantes selecionadas pela Stepwise
Nível Baixo	Todas as Variáveis

C1- Corrente 1 de 15 kV; C2 - Corrente 2 de 50 Kv; VIS – reflectância na faixa do visível; NIR - reflectância infravermelho próximo; SBMx –suscetibilidade magnética.

O segundo nível testado foi o nível médio (MLF) que consiste inicialmente na extração de algumas variáveis relevantes do banco de dados de cada sensor separadamente. Em seguida as variáveis selecionadas são concatenadas em uma única matriz, apresentando na linha as amostras de solo analisadas e nas colunas as variáveis selecionadas de cada sensor.

Para este nível foram gerados 17 bancos de dados, sendo 5 com os sensores individuais para avaliação de forma isolada, 10 contendo distintas combinações de sensores e 2 contendo todos os sensores em um banco de dados.

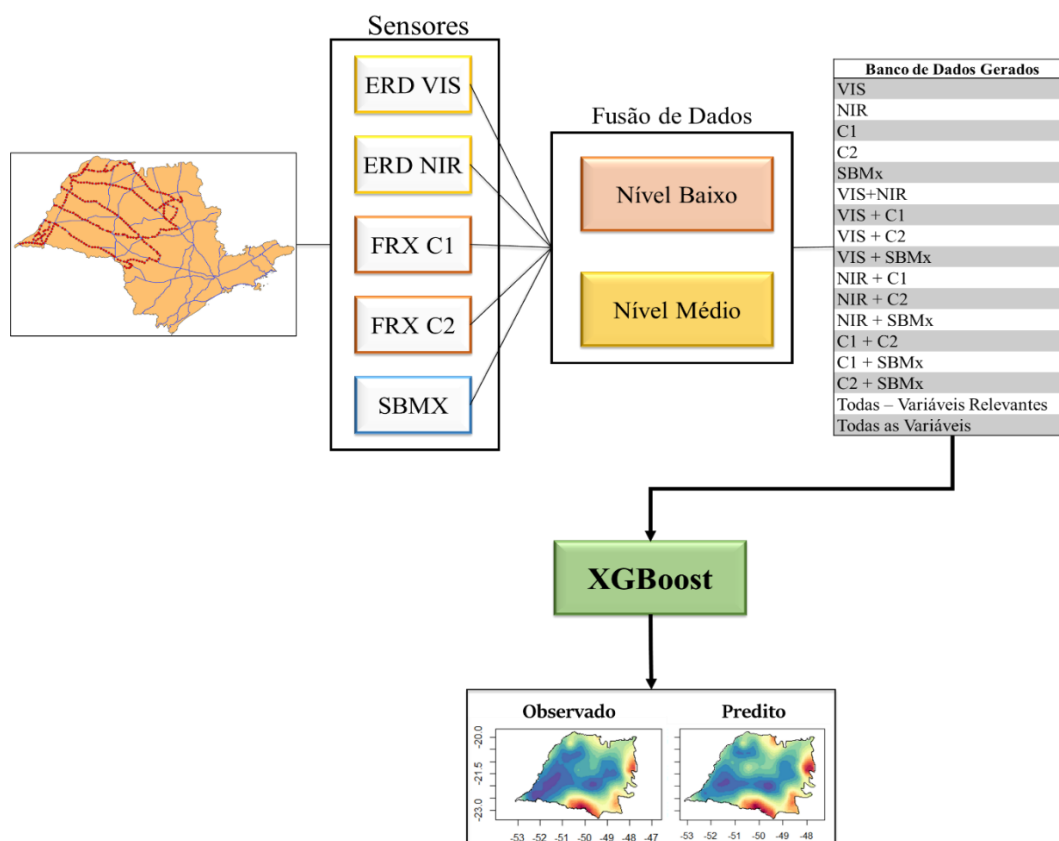
2.2.6 Análises Estatísticas

Em posse do banco de dados, a primeira análise realizada foi a estatística descritiva dos atributos do solo (textura, FED e FEO) e para viabilizar o entendimento dos dados foram desenvolvidos gráficos de violinos, agrupados em função dos compartimentos geológicos existentes na área de estudo. Em seguida os

compartimentos geológicos foram convertidos em variável *dummy* (Suits, 1957), onde solos de compartimento arenítico foram classificados como 0 e solos de compartimento basáltico como 1. Na sequência foi desenvolvida uma matriz de correlação de *spearman* com as variáveis *dummy* e os atributos de solo.

Na etapa seguinte iniciou-se o processo de fusão de dados. A primeira realizada foi a de nível baixo onde as variáveis foram concatenadas em um único banco de dados sem nenhuma etapa de pré-processamento. Já a fusão de nível médio exige uma pré-seleção das variáveis mais relevantes para os atributos de forma individual. E por fim a concatenação das variáveis mais relevantes de todos os sensores.

Para a seleção dos atributos mais relevantes de cada sensor, os dados foram submetidos à análise de regressão linear múltipla com seleção de variáveis pelo método *Stepwise forward - backward*. As variáveis target foram os valores de FED e FEO. Para facilitar o entendimento do processamento dos dados foi confeccionada Figura 2, com um fluxograma de todas as etapas.



C1- Corrente 1 de 15 kV; C2 - Corrente 2 de 50 Kv; VIS – reflectância na faixa do visível; NIR - reflectância infravermelho próximo; SBMx –suceptibilidade magnética, FRX – fluorescência de raio x.

Figura 2. Fluxograma do passo a passo do processamento de dados até a obtenção dos mapas de FED e FEO para o Planalto Ocidental Paulista.

Ao fim da etapa de fusão foram obtidos o total de 17 bancos de dados para cada atributo mineralógico. Para entender como essas variáveis geradas pelos sensores se relacionam com FED e FEO, foram desenvolvidos 4 *heatmaps* para cada atributo, contendo os comprimentos de onda selecionados na fusão de dados nível médio de cada sensor.

Na etapa seguinte deu-se início a predição dos atributos FED e FEO utilizando os 17 banco de dados desenvolvidos nas etapas anteriores. O algoritmo utilizado foi o XGBoost. O XGBoost é uma técnica de aprendizagem de máquina que tem como principal característica a escalabilidade, alta precisão e ampla aplicação (Nguyen et al., 2022). Ela consiste em uma variação dos algoritmos de árvore de decisão, com a técnica de aumento de gradiente que atua ajustando as funções de perda que são

minimizadas pela estrutura de interações sequenciais do algoritmo (Hikouei et al., 2021). Ou seja, o XGBoost usa árvores sequencialmente otimizadas.

Este algoritmo tem capacidade de trabalhar com problemas de classificação e regressão. Neste trabalho, devido à natureza dos dados optamos pelo de regressão. O algoritmo foi aplicado nos dezessete bancos de dados gerados a partir da fusão. Todos os bancos de dados foram divididos em 60% para treino e 40% para teste, pensando que o algoritmo é impulsionado pelo adequado ajuste de parâmetros. Juntamente, foi utilizado um algoritmo de otimização, chamado *RandomizedSearchCV*, no qual foi estabelecido o total de 6 parâmetros (*n_estimators*, *max_depth*, *reg_lambda*, *subsample*, *colsample_bytree*, *booster*) com 10 possibilidades cada, para todos os modelos e foi selecionado o total de 50 árvores com 0,5 de taxa de aprendizagem.

Para avaliar o desempenho do modelo, foram estabelecidas 3 métricas: o coeficiente de determinação (R^2), erro médio absoluto (MAE) e a raiz quadrada do erro médio (RMSE). Por fim para avaliarmos o desempenho do modelo no espaço, os dados observados em campo e o resultado do modelo de melhor desempenho selecionado com base nas métricas a citadas acima, foram interpolados através do método geoestatístico de krigagem (Oliver e Webster, 2014a). Para todas as análises, foram utilizadas as linguagens de programação R e Python, de forma concomitante, no ambiente Jupyter notebook.

2.3 Resultados e discussão

2.3.1 Estatística exploratória

Com o objetivo de caracterização da área, o primeiro atributo avaliado foi a argila. Nota-se que esta apresentou maiores valores no compartimento basáltico, diferente do atributo areia, que se mostrou predominante no compartimento arenítico, associações minerais comuns em solos derivados de basalto e arenito

(Hong et al., 2023). Em ambos os atributos se nota uma grande amplitude de variabilidade dos dados entre os pontos máximos e mínimos e tal fato pode ser atribuído as zonas de transição geológica existentes devido ao tamanho da área de estudo.

Tabela 2. Análise exploratória da granulometria e conteúdo de ferro ditionito e oxalato, para solos de diferente geologia do Planalto Ocidental Paulista

Compartimento Arenítico						
	Argila	Areia	FED	FEO	Teor de Hm	Teor de Gt
Máximo	589,50	926,98	76,38	5,59	105,00	116,00
Q3	176,13	824,25	25,51	1,10	39,00	32,00
Média	156,92	724,33	21,35	1,03	28,34	25,96
Q1	87,37	700,14	12,40	0,45	9,00	14,00
Mínimo	24,00	86,47	4,93	0,16	2,00	3,88
Compartimento Basáltico						
Máximo	644,00	843,25	51,77	5,89	43,92	81,00
Q3	535,50	515,25	49,80	3,54	31,00	18,00
Média	431,38	379,74	43,14	2,53	22,00	18,84
Q1	351,00	184,50	44,72	1,24	14,00	9,00
Mínimo	67,00	102,36	15,60	0,21	3,00	3,00

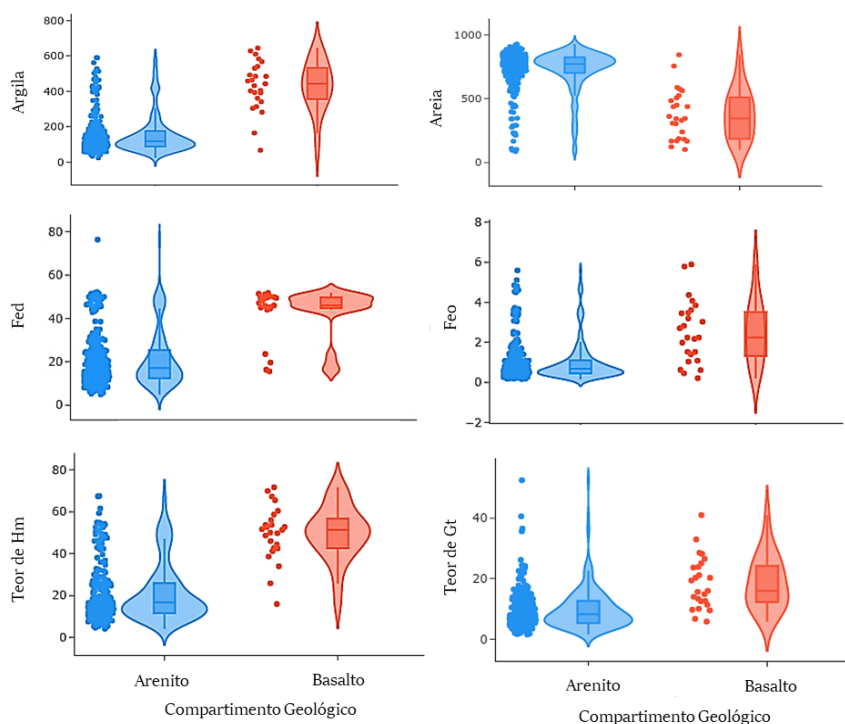


Figura 3. Gráficos do tipo violino para caracterização dos teores dos atributos dos solos do Planalto Ocidental Paulista.

O FED foi a terceira variável analisada, esta apresentou um comportamento semelhante ao da argila, apresentando maiores valores nos compartimentos de

basalto, devido principalmente à maior presença de minerais de Fe mais bem cristalizados nesses ambientes (Teixeira et al., 2016).

Já quando se observa o comportamento de FEO nota-se uma grande variabilidade na distribuição dos valores presentes no compartimento basáltico, o que pode ter sido influenciado devido serem rochas enriquecidas em Fe (até 15% em peso de FEO), porém podem conter grandes quantidades de minerais amorfos ricos em Fe (Lee et al., 2018; Stranghoener et al., 2020).

Após o entendimento da distribuição dos dados, foi realizada a análise de correlação de *spearman* que permitiu constatar que, de fato, a distribuição semelhante representa um alto coeficiente de correlação de FED com o teor de Hm e com o teor de Gt (Teixeira et al., 2016). O atributo FEO também apresentou correlação com os mesmos minerais, porém, inferior à citada anteriormente. Os Minerais Hm e Gt apresentaram correlação direta, porém moderada, tal fato pode ser atribuído as distintas características de formação destes minerais.



Figura 4. Matriz de correlação de spearman entre os compartimentos geológicos do solo, conteúdos de Hematita (Hm) Goethita (Gt) e de ferro cristalino (FED) e ferro não cristalino (FEO), para solos do Planalto Ocidental Paulista.

Geralmente temperaturas médias, menor evapotranspiração, condições úmidas, altas precipitações e solos com baixo teor de ferro, como os arenitos,

favorecem a formação da goethita (Balsam et al., 1995; Cornell e Schwertmann, 2006), enquanto a formação de hematita é favorecida em condições quentes e secas, baixa precipitação, solos ricos em ferro, como o basalto, além de pH neutro e elevado teor de Al (Cornell e Schwertmann, 2006; Fink et al., 2016; Kampf e Schwertmann, 1982).

2.3.2 Fusão de dados

Tendo entendimento de como FED e FEO se relacionam com os compartimentos geológicos e com os teores de Hm e Gt, iniciou-se o processo de fusão dos dados obtidos por meio dos sensores ERD, EFRX e SMBx.

No presente trabalho foram testados dois níveis de fusão, o nível médio exige uma pré-seleção das variáveis mais relevantes, utilizando o algoritmo *Stepwise* para os sensores de forma individual e, por fim, a união de todas as variáveis mais relevantes de cada um. Obteve-se para FED a seleção de 10 comprimentos de onda do espectro VIS, 18 do NIR, 19 pontos de energia para C1 da EFRX e 15 na C2, aproximadamente 1,35% do total do banco de dados inicial. Para a predição do FEO o algoritmo selecionou 25 faixas do VIS, 24 do NIR, 13 pontos de energia de C1 e 27 em C2 da EFRX, aproximadamente 2,15% do total do banco de dados inicial.

Nota-se que para FEO o algoritmo indicou a necessidade de um maior número de variáveis para a sua estimativa, tal fato pode ser atribuído às características de sua formação que são muito específicas. Logo, são necessárias mais variáveis dependentes para capturar as nuances do atributo, uma vez que pode haver grande variabilidade na quantidade desses tipos de Fe na rocha de origem, podendo chegar a 15% (Stranghoener et al., 2020).

Nota-se que o cluster hierárquico foi capaz de realizar o agrupamento com base nos compartimentos geológicos existentes. Ao avaliar os mapas de calor são nítidas as relações de FED com as refletâncias e energias selecionadas pela

Stepwise. Na Figura 5a (FED-VIS) e 5b (FED-NIR) é possível notar que onde existe a ocorrência de maiores teores de FED a refletância dos comprimentos de onda é menor, já no compartimento arenítico, encontram-se os menores teores de FED, que ocasiona em maiores refletâncias.

Liu et al. (2021) observaram menor reflectância em solos de basalto na faixa do NIR, em comparação a solos com maiores quantidades de areia quartzosa. Bachmann et al. (2014) relatam evidências entre o tamanho da partícula e a refletância, à medida que o tamanho da partícula diminui a refletância reduz junto, especialmente para partículas menores que 200 μm . A diminuição é mínima no Visível, mas torna-se mais forte em comprimentos de onda mais longos no infravermelho.

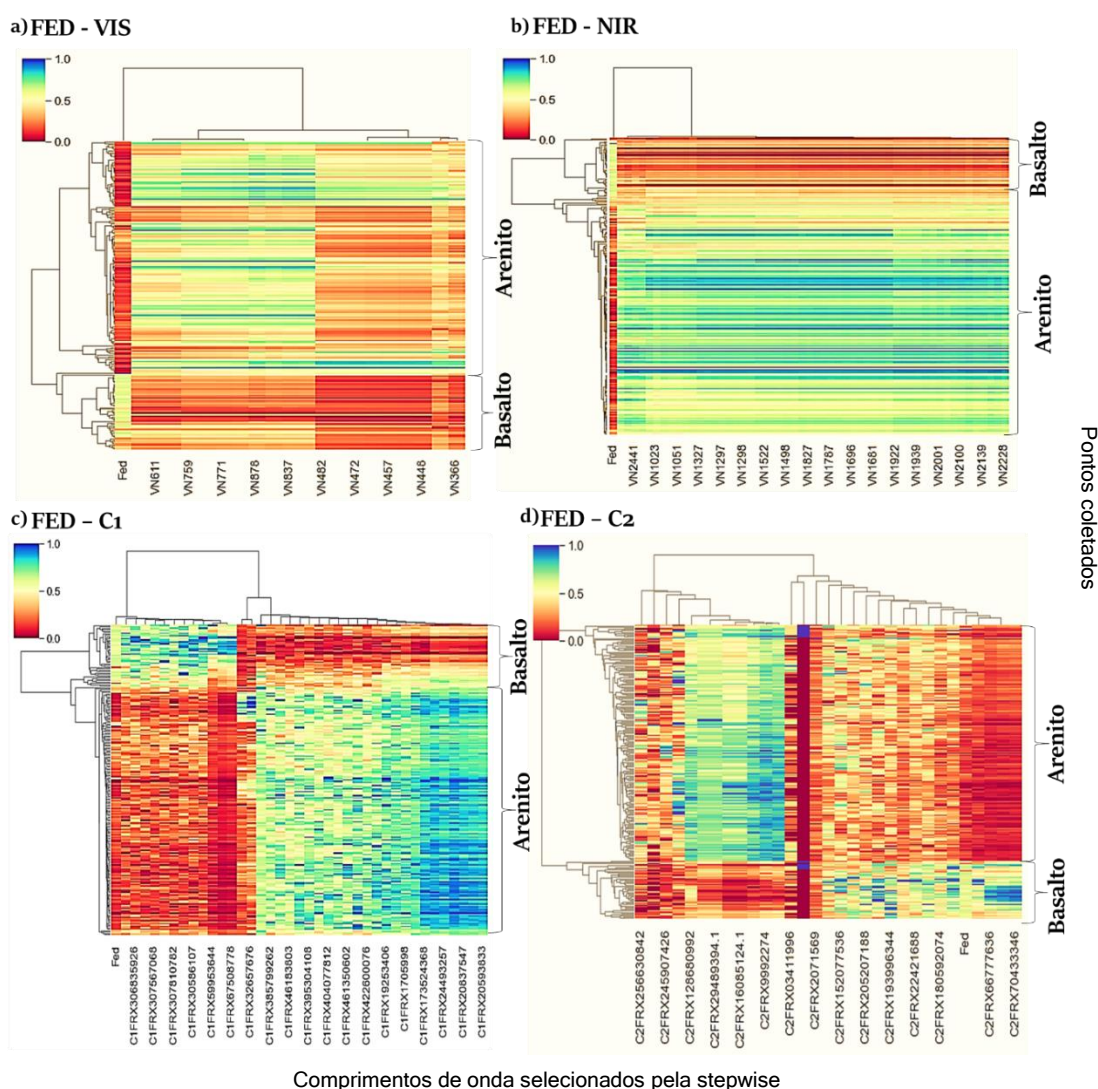


Figura 5. Análise de Cluster com mapa de calor para as faixas de comprimento de onda da ERD (VIS e NIR) e EFRX (correntes 1 e 2), respectivamente, selecionadas pela Stepwise como significativas para a predição de FED, para solos do Planalto Ocidental Paulista

Tal comportamento foi observado tanto no mapa de calor desenvolvido para VIS como para NIR. Porém, na figura 5a (FED-VIS) podemos destacar a ocorrência de dois agrupamentos de comprimentos de onda, realizados pelo cluster hierárquico, onde ocorreu a separação dos comprimentos de onda nos intervalos 366 – 462 nm, que são mais associados às cores roxo e azul, e no intervalo 611 – 837 nm, que compete mais às cores laranja e vermelho.

Pode-se inferir que os solos do agrupamento basalto refletem mais nos comprimentos de onda do laranja e vermelho e refletem menos nos comprimentos de onda do azul e roxo, indicando, assim, solos que variam do vermelho ao alaranjado, cores tipicamente encontradas em solos de origem basáltica (Swetha et al., 2022).

Diante de tudo que foi observado nos mapas de calor de VIS e NIR, nota-se um comportamento semelhante nos picos de energia gerados pela FRX. Na imagem “c” (FED-C1) e “d” (FED-C2) é possível observar que o cluster também dividiu os picos de energia em dois grandes grupos, onde um deles apresenta correlação direta com os teores de FED e o outro correlação inversa, ou seja, à medida que o teor de FED aumenta a energia desses pontos diminuem. Os picos de energia que estão localizados entre 6 e 7 foram os que apresentaram correlação mais alta com os teores de FED. Segundo a literatura, dois picos de ferro são detectados pela fluorescência, este são localizados entre os pontos 6 e 7 do espectrograma da FRX (Santos et al., 2021; Nawar et al., 2022).

Na figura 5c (FED-C1) foi observado que os picos de 3.0 a 3.5 Kev também manifestaram relação direta com o FED, mesmo que menos intensa que a citada

anteriormente. Nesses pontos, segundo a literatura, é comum detectar o elemento cálcio, possivelmente relacionado ao fato de que em solos oriundos de basalto os plagioclásios com alto teor de cálcio serem minerais muito comuns (Lu et al., 2020; Sun et al., 2020). Os pontos que apresentaram relação mais inversa com FED estão localizados entre 1.7 e 2.0 Kev, nesses pontos é comumente detectado silício (dos Santos et al., 2021; Nawar et al., 2022).

Assim como FED, observando a figura 6a (FEO-VIS) e figura 6b (FEO-NIR) nota-se que os solos de compartimento basáltico apresentam a maior ocorrência de FEO e menores valores de refletância. Nota-se que o cluster hierárquico separou as refletâncias em dois grandes grupos. No grupo que compete de 338 a 494 nm observamos relação inversa entre a refletância e o teor de FEO.

Assim como observado em FED, a refletância dos comprimentos de onda associados ao roxo e o azul reflete menos e nos solos com maior teor de FEO, quando comparados com os comprimentos de onda ligados aos espectros vermelho e laranja, devido à maior quantidade de óxidos de ferro que conferem essa cor aos solos oriundos de basalto, como ferridrita, hematita e goethita (Santos et al., 2021; Liu et al., 2021; Swetha et al., 2022).

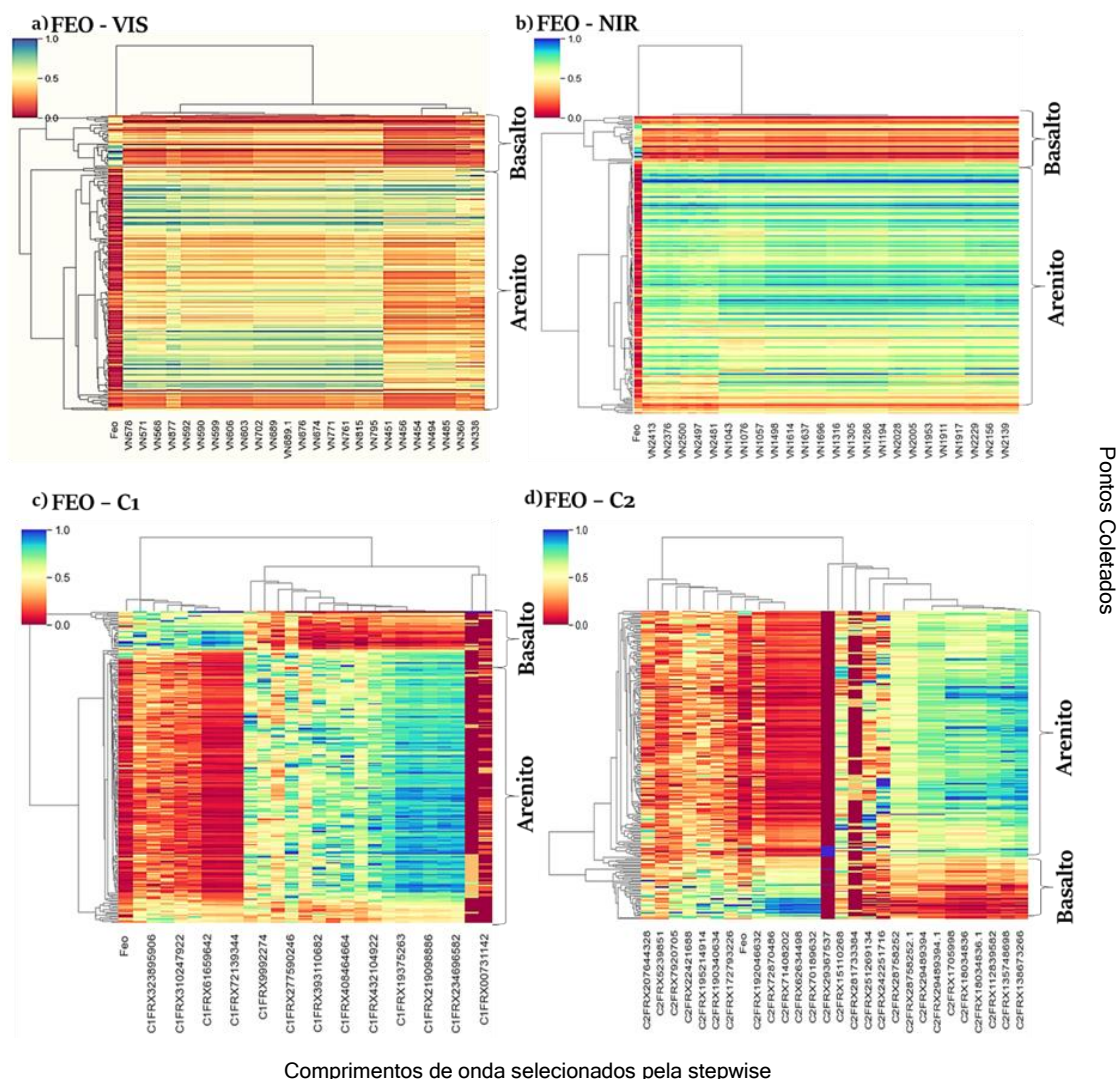


Figura 6. Figuras a, b, c e d representam as faixas de comprimento de onda da ERD (VIS e NIR) e EFRX (correntes 1 e 2), respectivamente, selecionadas pela Stepwise como significativas para a predição de FEO.

Nas figuras 6c e 6d (FEO-C1 e C2), ocorreu a divisão em dois grandes clusters, um deles apresenta correlação direta com os teores de FEO, destes podemos destacar os picos de 6 a 7 que foram os que apresentaram correlação inversa mais intensa o que pode ser associado a conhecida ocorrência de dois picos de ferro neste ponto da curva (Santos et al., 2021; Nawar et al., 2022).

2.3.3 Análise de predição

Na Tabela 3 pode-se observar as métricas utilizadas para avaliar a

performance dos modelos gerados utilizando o algoritmo XGBoost. De modo geral podemos dizer que os modelos desenvolvidos para FED e FEO apresentaram um desempenho semelhante. Os modelos que apresentaram os piores valores das métricas foram os que utilizaram somente a SMBx como variável preditora, tanto para FED quanto para FEO.

Os melhores modelos preditivos, quando se analisa as métricas de treino e teste, foram com a utilização da combinação dos dados dos 3 sensores, selecionados pela Stepwise, ficando com R^2 0,01 e 0,14 pontos acima dos modelos que utilizaram todos os dados, sem nenhuma limpeza.

Vale ressaltar que mesmo apresentando R^2 com valores próximos, a amplitude de MAE e RMSE dos modelos que utilizaram os dados sem nenhuma limpeza prévia, foram sempre superiores aos modelos que utilizaram os dados filtrados. A avaliação das métricas MAE e RMSE se faz necessária na seleção do melhor modelo de predição, uma vez que são indicadores complementares e não excludentes, quando se trata do entendimento dos resultados obtidos (Karunasingha, 2022).

1 Tabela 3. Métricas de avaliação dos modelos preditivos de FED e FEO em função da fonte de dados de Espectroscopia de Reflectância
 2 Difusa (ERD), Espectroscopia de Fluorescência de raios X (EFRX) e Suscetibilidade magnética em baixa frequência (SMBx)
 3

Sensores	Treino			Teste			Treino			Teste		
	R²	MAE	RMSE	R²	MAE	RMSE	R²	MAE	RMSE	R²	MAE	RMSE
	FED						FEO					
VIS	0,86	3,21	4,76	0,69	5,8	8,32	0,87	0,12	0,35	0,7	0,49	0,74
NIR	0,76	4,54	6,28	0,7	8,19	5,57	0,76	0,32	0,47	0,68	0,47	0,77
C1	0,91	2,73	3,85	0,8	3,96	6,74	0,87	0,24	0,35	0,8	0,36	0,8
C2	0,88	2,88	3,68	0,78	3,98	6,71	0,86	0,25	0,36	0,76	0,29	0,46
SBMx	0,28	8,71	11,94	0,16	8,71	12,37	0,25	0,58	0,84	0,2	0,79	1,15
VIS+NIR	0,80	4,11	5,78	0,73	5,35	7,74	0,79	0,29	0,45	0,71	0,45	0,72
VIS + C1	0,90	2,89	3,91	0,82	3,89	6,33	0,81	0,27	0,43	0,75	0,35	0,81
VIS + C2	0,90	2,98	4,1	0,78	4,1	6,95	0,91	0,16	0,28	0,84	0,2	0,38
VIS + SBMx	0,78	4,28	6,03	0,66	6,35	8,47	0,76	0,31	0,48	0,68	0,47	0,77
NIR + C1	0,91	2,79	3,73	0,86	3,4	5,52	0,84	0,26	0,39	0,8	0,33	0,53
NIR + C2	0,90	3,01	3,95	0,86	3,69	5,64	0,81	0,27	0,42	0,78	0,37	0,62
NIR + SBMx	0,91	2,66	3,88	0,82	4,69	6,39	0,8	0,3	0,44	0,75	0,43	0,68
C1 + C2	0,94	2,21	3,09	0,79	4,02	4,89	0,89	0,18	0,31	0,82	0,31	0,57
C1 + SBMx	0,91	2,64	3,42	0,83	3,84	6,16	0,87	0,25	0,35	0,81	0,37	0,59
C2 + SBMx	0,89	2,72	3,75	0,8	3,57	5,82	0,85	0,27	0,37	0,79	0,31	0,52
Variáveis relevantes selecionadas pela Stepwise	0,95	2,18	3,05	0,92	2,55	3,49	0,92	0,14	0,26	0,89	0,19	0,38
Todas as Variáveis	0,94	2,35	3,10	0,89	2,8	4,52	0,83	0,22	0,39	0,75	0,38	0,67

4 *VIS – Faixa do Visível; NIR – faixa do infravermelho próximo; C1 – corrente 1 da EFRX; C2 – corrente 2 da EFRX; SBMx – Susceptibilidade
 5 magnética

Segundo Kuncheva et al. (2020) o desempenho superior dos modelos com dados filtrados pode ser associado alto número de recursos e um baixo número de repetições quando os dados não são filtrados, o que pode prejudicar gravemente o bom desempenho dos algoritmos de aprendizagem, podendo ocasionar problemas de overfitting dos dados, além da baixa representatividade das amostras, o que pode retornar um conjunto de recursos de baixa confiabilidade.

Nesse sentido, o ideal é possibilitar a seleção das variáveis da forma mais simples possível. Essa redução do banco de dados ressalta a importância do uso de técnicas que permitam remover as variáveis não ou pouco relevantes para a predição do atributo de interesse e garantir a construção de modelos consistentes de predição, diminuindo o risco de inferências não confiáveis e reduzindo o custo computacional (Anzanello et al., 2013).

O uso da faixa VIS, mesmo não apresentando as melhores métricas, pode ser uma alternativa interessante pensando em praticidade. A faixa do VIS pode ser facilmente obtida por satélites, que garantem a obtenção dos dados de forma ainda mais fácil e rápida. Como apontam as métricas, os resultados não teriam muita exatidão, mas poderiam facilmente auxiliar em tomadas rápidas de decisões.

Outro ponto que podemos destacar na Tabela 3 é que a nível de sensores, os dados gerados pela EFRX se mostraram variáveis preditoras mais eficientes na predição de FED e FEO, quando comparado com os dados gerados por ERD. Nota-se ainda que a fusão dos sensores ERD e EFRX, independentemente da corrente e faixa, geraram modelos de alta performance, se tornando apenas um pouco menos assertivo que o melhor modelo observado.

Resultados obtidos por Shrestha et al. (2022) e Zhao et al. (2022) corroboram com o que foi demonstrado neste trabalho, no qual a fusão e limpeza dos bancos de dados se mostraram eficientes na melhora da qualidade dos modelos preditivos.

De todos os modelos desenvolvidos destacamos três para FED e três para FEO o primeiro foi o de melhor desempenho (Todas as variáveis selecionadas), o melhor modelo gerado com duas fontes para FED (C1 + C2) e para FEO (VIS + C2). Por fim, selecionamos o modelo gerado pelo VIS por ser um dado de fácil obtenção e baixo custo, podendo ser aplicado até via satélite.

Para agregar no entendimento do desempenho dos modelos foram desenvolvidos modelos de regressão, onde no eixo X são apresentados os dados observados e no Y os dados preditos. Na figura 7 fica nítida a distribuição do banco de dados, principalmente para FED. Nota-se que tem uma grande concentração de dados na parte inferior do eixo, representando os menores valores de FED e FEO, podendo ser associado ao compartimento geológico arenítico e uma concentração na parte superior da reta, representando os maiores valores de FED e FEO, que estão associados ao compartimento basáltico.

É possível observar que o modelo se mostra mais errático quando estão na zona de transição entre os compartimentos geológicos, ou seja, no meio da reta. Tal fato pode ser associado à representatividade da amostra, por se tratar de uma zona de transição, o que gera mudanças abruptas nos valores, além da quantidade de repetições no local não ter sido suficiente para o entendimento adequado do padrão.

Além disso, espera-se que a classe com maior ocorrência seja predita com mais precisão, em comparação com aquela com menor ocorrência, logo, um

tamanho de amostra menor contribui para a incerteza do modelo (Zeraatpisheh et al., 2017).

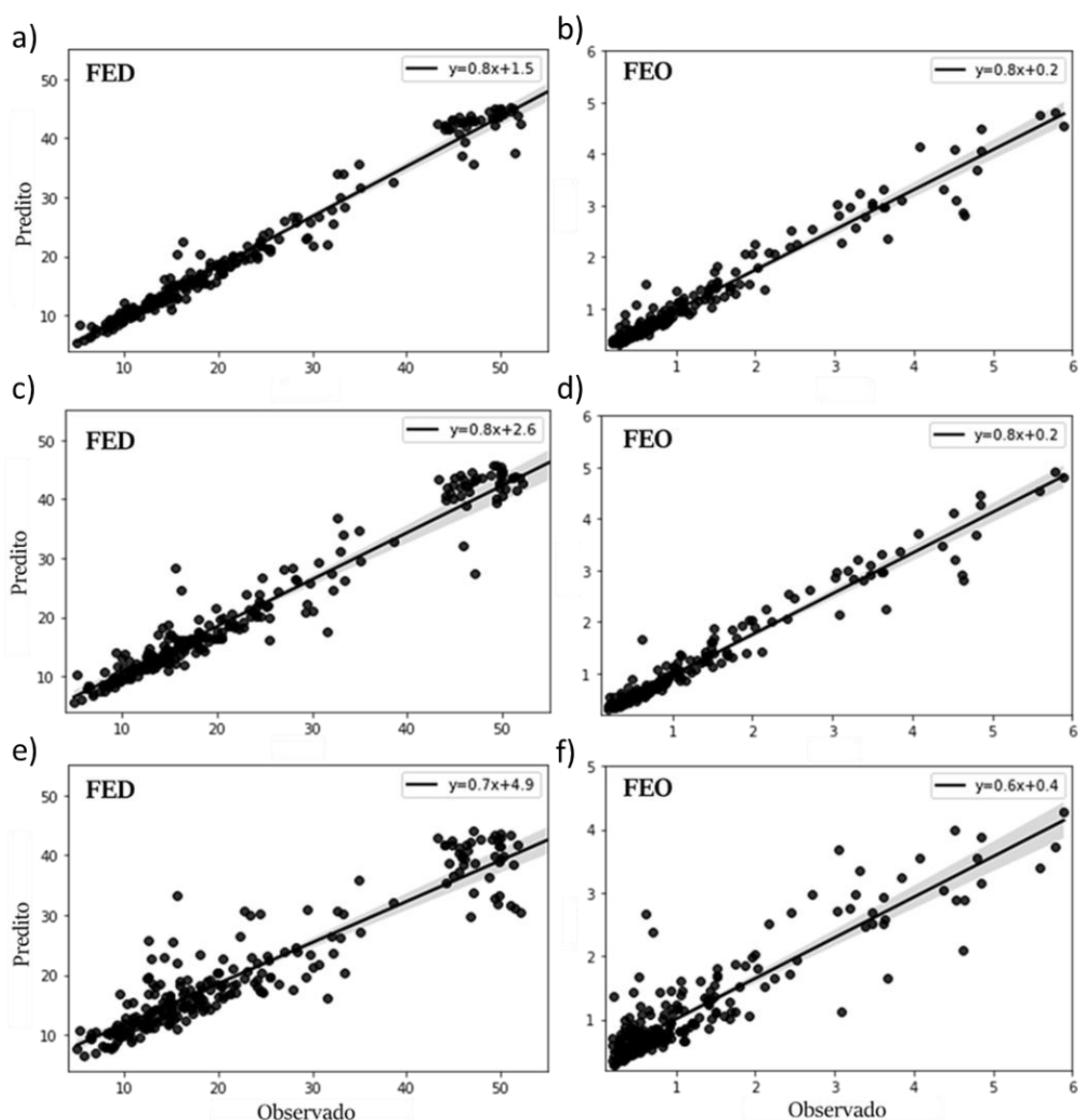


Figura 7. Distribuição de dados preditos e observados para solos do Planalto Ocidental Paulista. a): Variáveis relevantes selecionadas pela Stepwise; b) C1 + C2; c) VIS + C2.

Os modelos desenvolvidos para FEO se mostram mais erráticos em relação aos de FED, tais fatos podem ser associados às características únicas da formação de FEO o que não gera essa clara definição dos compartimentos geológicos, devido principalmente à grande variabilidade na presença de minerais amorfos ricos em Fe (Lee et al., 2018; Stranghoener et al., 2020).

Além disso as regressões de FEO se mostraram mais erráticas nos valores extremos, o que pode ser associado a uma baixa densidade amostral desses valores extremos de FEO no banco de dados. Fato semelhante foi observado por (Baldo et al., 2021b). Os autores observaram que quanto mais extremos os valores de FEO, mais errático o modelo se torna. Ressaltasse que no trabalho dos autores acima citados os dados foram mais erráticos, para FED o modelo utilizando ERD e Random Forest obteve o R^2 de 0,62 e FEO de 0,56 mostrando a diferença e maior eficiência do algoritmo utilizado no presente estudo.

Dentre os modelos, de fato o gerado com base no VIS apresenta maior dispersão nos dados em relação à reta, quando comparado com os demais modelos, justificando o desempenho inferior de suas métricas. Como resultado do trabalho, observa-se que o pré-processamento adequado e a junção de sensores, agregado das técnicas de aprendizagem de máquina, potencializam o desempenho dos modelos.

Porém, a decisão do uso de vários sensores ou da robustez do processamento irá depender do objetivo do estudo, da necessidade de eficiência, da área estudada, entre outros inúmeros fatores. Os resultados mostram qual foi o melhor processamento de dados a ser adotado e quais sensores foram mais eficientes, bem como quais fusões melhoram a performance do modelo em relação ao uso de uma única faixa, com métricas razoáveis.

O objetivo é entender se de fato as fusões e técnicas processamento robusta afetam o desempenho do modelo e como cada sensor agrega esse desempenho e de fato garantir que o gestor tome a decisão de forma mais assertiva possível com base no banco de dados disponíveis, no tamanho da área, na necessidade de trabalho, entre outros fatores.

2.3.4 Análise Geoestatística

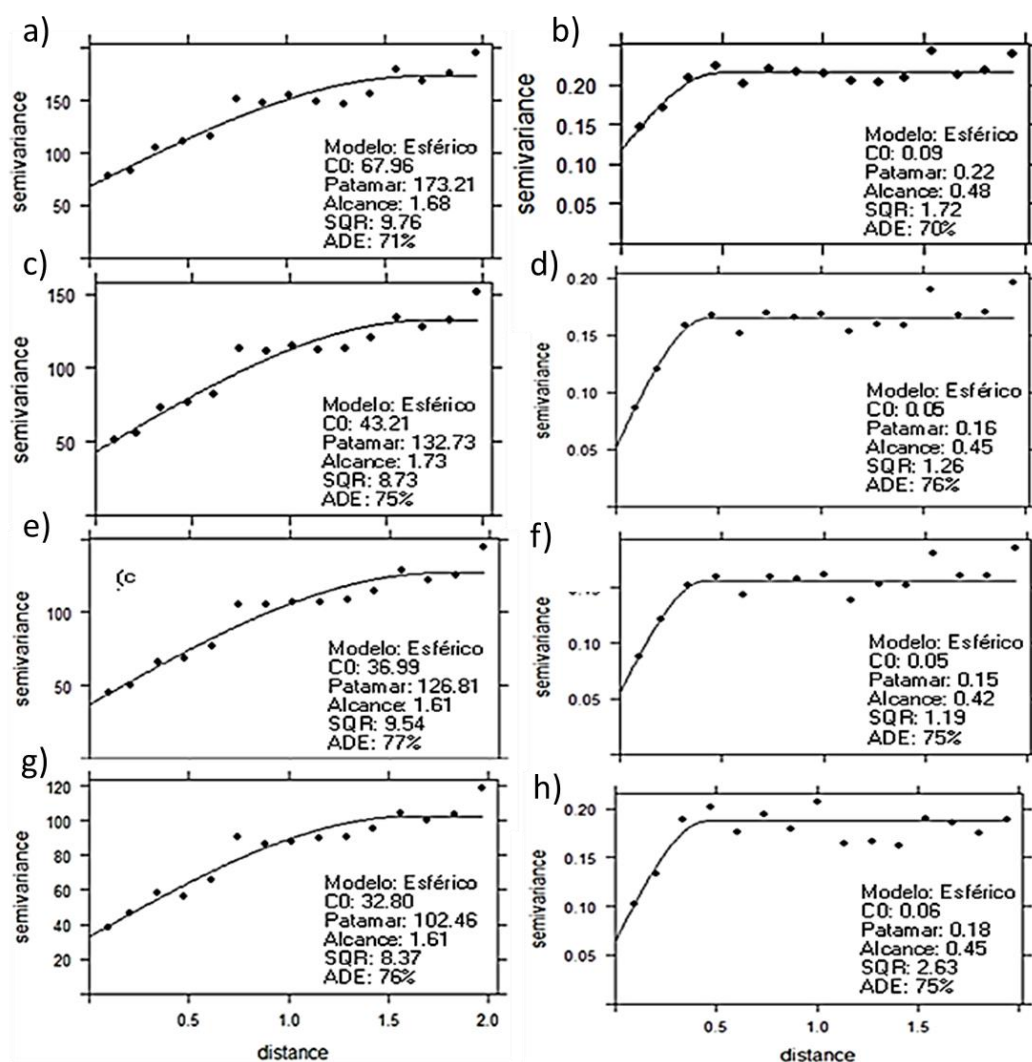
Por fim, foram realizadas as krigagens dos dados preditos e observados de FED e FEO, para os modelos destacados anteriormente na Figura 7. Nos semivariogramas (Figura 8), nota-se comportamento muito semelhante entre os dados preditos e observados, o que é associado à elevada acurácia dos modelos preditivos desenvolvidos. Os dados se ajustaram ao modelo esférico, o que se justifica pela característica pedoindicadora dos óxidos de ferro, como já apontado em outras literaturas (Silva et al., 2020, Bahia et al. 2017) e pela sua transição ocorrer de forma mais lenta, associada à transição geológica.

O semivariograma desenvolvido para o FEO obteve um valor de feito pepita menor quando comparado a FED indicando que devido ao tamanho da área de estudo que é mais difícil captar a variabilidade em pequena escala de FED. Para os semivariogramas desenvolvidos, nota-se que para o que foi desenvolvido utilizando o modelo de variáveis relevantes selecionadas pela stepwise foi o que apresentou patamar mais próximo ao dos dados observados, outro ponto que destacar além disso obteve patamar mais bem estabelecido, quando comparado aos semivariogramas de FED, o que pode ser atribuído aos compartimentos geológicos que conforme observado na figura 7, afetam mais a distribuição de FED, gerando concentração de dados nos extremos.

Outra medida avaliada foi o alcance se mostrou semelhante em todos os semivariogramas, tanto para FED como para FEO e por fim foi calculado ainda o avaliador de dependência espacial e em todos os semivariogramas detectamos elevada dependência espacial entre os dados.

Foram desenvolvidos o total de quatorze mapas, sendo sete para FED e sete para FEO, além de gráficos violino comparando a distribuição dos dados observados e preditos pelo modelo (Figura 9). Na Figura 9, observamos o mapa intitulado como FED-observado. Nota-se a predominância dos maiores teores de FED nos extremos do mapa, o que se relaciona diretamente com o mapa de formação do solo (Figura 1b). Nestes locais o material de origem para a formação do solo é a rocha basáltica, rica em minerais férricos.

No mapa é possível observar uma mudança gradual dos valores de FED até a região central e oeste do POP, onde há predomínio do tom azulado, representando menores valores de FED. É nesta região que temos como rocha de origem o arenito (Fernandes et al., 2007; Mello et al., 2021).



*C0: Efeito pepita. SQR: Soma do quadrado dos resíduos. ADE: Avaliador da dependência espacial

Figura 8. Semivariogramas da krigagem ordinária a) dados observados, c) variáveis relevantes selecionadas pela Stepwise, modelo de C1+C2 (c), modelos VIS (d), para FED e FEO respectivamente.

Analisando o mapa da Figura 9a é possível notar a semelhança com o mapa observado, o que se justifica pela elevada acurácia do modelo desenvolvido, devido à semelhança entre os mapas é necessário recorrer ao mapa de erro (Figura 9d) para detectar as diferenças entre os mapas. No mapa de erro (Figura 9d) vê-se que o modelo subestima os dados no ambiente de arenito, porém, o erro se aproxima a zero.

Nas regiões de maior concentração de basalto nota-se os erros mais

localizados e de maior amplitude, o que pode ser associado à elevada variabilidade dos valores de FED nesses locais, o que pode ter influenciado negativamente na predição.

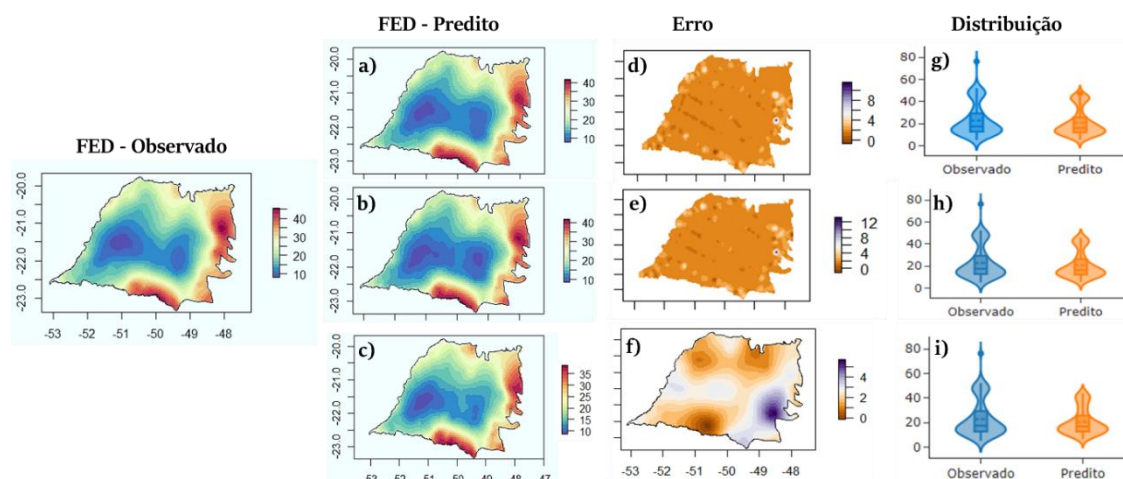


Figura 9. Mapas de FED-Observado, a) modelo de variáveis relevantes selecionadas pela Stepwise, b) modelo de C1+C2, c) modelos VIS, d) mapa erro-modelo de variáveis relevantes selecionadas pela Stepwise, e) mapa erro-modelo de C1+C2, f) mapa erro-modelo de VIS g) gráficos de distribuição-variáveis relevantes selecionadas pela Stepwise, h) gráficos de distribuição-modelo de C1+C2, i) gráficos de distribuição-modelo VIS

para os solos do Planalto Ocidental Paulista.

O mapa obtido com base no modelo, C1+C2, (Figura 9b) também se mostrou muito semelhante ao observado. Vale ressaltar que este modelo utiliza apenas o sensor FRX o que viabiliza a aplicação mais rápida deste modelo em pesquisas. Este modelo assim como o anterior também se mostrou mais errático nos valores extremos, tanto em pontos mínimos como em máximos, o que é natural devida à alta variabilidade local como citado anteriormente.

Por fim, o mapa confeccionado utilizando o modelo obtido com VIS (Figura 9c) foi o que apresentou erros mais evidentes (Figura 9f), superestimando em três as informações na maior parte do mapa. Vale ressaltar que mesmo se mostrando mais errático é possível utilizar este sensor via satélite tal característica viabiliza

a sua adoção em grandes áreas com baixo custo (Mello et al., 2021).

Nos gráficos violinos, é nítida a semelhança na distribuição dos dados observados e preditos, em todos é possível notar a maior concentração dos dados na extremidade inferior e que ocorre uma redução da quantidade de amostras no meio por isso ocorre um afinamento na sobra de distribuição e no extremo superior ela volta a ter um alargamento, e o mesmo comportamento foi observado nos três bancos de dados estimados. Nota-se que nas figuras h e g as linhas das medianas ficaram muito próximas a observada, nota-se que apenas na 9i a linha ficou abaixo do que na observada indicando que o modelo subestimou os dados usando apenas o VIS.

Na Figura 10 é possível observar os mapas referentes ao atributo FEO. Podemos observar os maiores valores de FEO nas extremidades do mapa. Porém, não tão concentrado como os mapas de FED. Nota-se que mais ao centro ocorre uma drástica redução nos valores de FEO, justificado pela predominância de solos proveniente de arenito nesses locais (Fernandes et al., 2007).

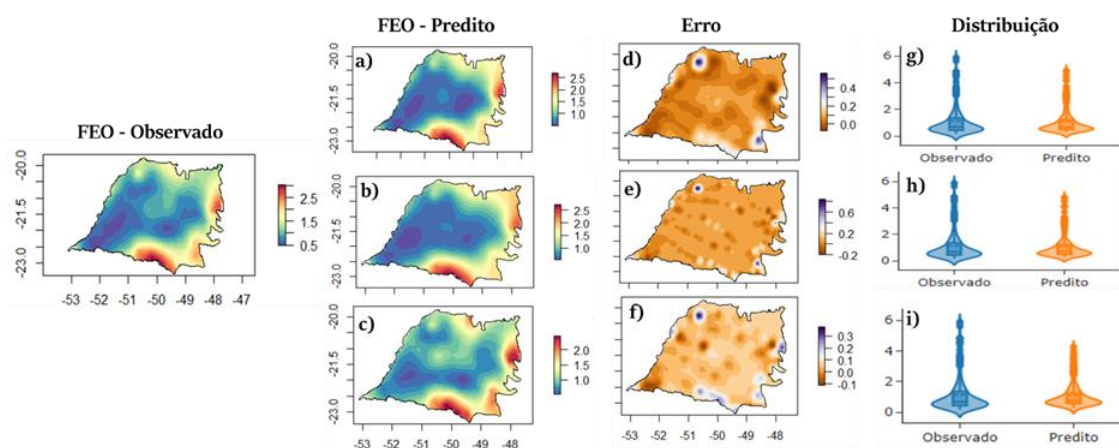


Figura 10. Mapas de FEO-Observado, a) modelo de variáveis relevantes selecionadas pela Stepwise, b) modelo de VIS+C2, c) modelos VIS, d) mapa erro-modelo de variáveis relevantes selecionadas pela Stepwise, e) mapa erro-modelo de VIS+C2, f) mapa erro-modelo de VIS g) gráficos de distribuição-variáveis relevantes selecionadas pela Stepwise, h) gráficos de distribuição-modelo de VIS+C2, i) gráficos de distribuição-modelo VIS

Assim como nos mapas da Figura 9, observamos grande semelhança entre os mapas de FEO predito por todos os sensores com limpeza (Figura 10a) em relação ao observado. Com base no mapa de erro (Figura 10d) é possível notar que o modelo superestimou FEO para a maior parte do mapa, porém, com valores em torno de 0,1 de teor de FEO. Nas regiões de maior ocorrência de basalto foram onde ocorreram os maiores erros, corroborando com o que foi observado nas regressões anteriormente.

Os desempenhos dos modelos, validados pelos valores das métricas apontados na Tabela 2 se mostram semelhantes aos observados nos mapas. O algoritmo utilizado tem eficiência para predição dos teores de FED e FEO, utilizando a fusão dos dados de sensores. Utilizando apenas a faixa do VIS e VIS+NIR, (Bahia et al., 2017b) em uma modelagem matemática diferente, observaram que os valores de FED e FEO são subestimados quando nas amostras de solo há conteúdos maiores que 100 g kg^{-1} e 4 g kg^{-1} destes elementos, respectivamente.

Já Bahia et al. (2017) utilizando apenas SBMx observaram que os mapas preditos são mais próximos aqueles observados para FED e FEO. Como discutido anteriormente, os modelos utilizando SMBx neste estudo podem ter apresentados resultados menos satisfatórios por se tratar de apenas uma leitura, enquanto as curvas de ERD e FRX apresentam vários pontos com influências Fe e outros atributos ligados a presença de óxidos de ferro.

Observando os gráficos violinos da figura 10 é possível notar a semelhança na distribuição dos dados observados e preditos. Nota-se que na figura 10i ocorreu um achatamento do violino dos dados preditos, em relação aos demais, o que corrobora com o mapa subestimado com o modelo gerado usando apenas

o VIS. Nota-se ainda no violino da figura 10g que o modelo com todos os sensores estimou com muita proximidade os *outliers* observados no banco de dados.

A partir de dados de cor obtidos por ERD e utilizando árvore de regressão, Baldo et al. (2021) obtiveram mapas de FED e FEO próximos aos observados pela metodologia convencional, para a área do POP. Todavia, os autores observaram que na região de transição entre arenito e basalto e nos extremos no basalto os conteúdos tendem a ser subestimados.

Neste ponto, o design diferente do algoritmo XGBoost, estruturado para evitar o overfitting (Zhan et al., 2022), pode ter contribuído para os melhores resultados dos mapas. Estes resultados apontam que a evolução dos modelos matemáticos, associados aos novos sensores, vem contribuindo para o avanço da predição de atributos do solo, com maior precisão e facilidade.

A escolha do modelo matemático, do nível de fusão de dados ou do uso dos sensores isoladamente irá depender do objetivo do resultado. Em estudos mais rigorosos para definição de áreas mínimas de manejo e mapeamentos detalhados para compreensão da variabilidade espacial dos atributos do solo, é preferível modelos mais arrojados e com maior rigor matemático, como o modelo utilizando a fusão de todos os dados e aplicando o algoritmo XGBoost.

Em estudos mais simplificados e generalistas, ou com poucos recursos financeiros para investimentos, podem ser utilizados modelos mais flexíveis, como o uso de apenas um sensor. Como o modelo utilizando apenas a faixa do VIS, e o algoritmo XGBoost, ou até algoritmos mais simples, porém que garantam uma predição eficiente de atributos.

2.4 Conclusões

Os atributos FED e FEO podem ser utilizados como pedoindicadores de Hm e Gt.

A fusão de dados nível intermediário, associada ao algoritmo de predição XGBoost se mostrou eficiente na predição de teores de ferro cristalino e não cristalino do solo, obtendo valores de R^2 de 0,95 e 0,92 para dados de treino e 0,92 e 0,89 para dados de teste, respectivamente.

A fusão nível baixo também se mostrou eficiente na estimativa dos teores de ferro cristalino e não cristalino do solo, porém, apresentou menores valores de R^2 e aumento nos valores de MAE e RMSE.

Dados de sensores utilizados de forma isolada na predição apresentam desempenho satisfatório nas predições, porém, os modelos são significativamente melhorados quando há a fusão dos bancos de dados de diversos sensores.

De modo geral a aplicação exclusiva do VIS tende a subestimar os conteúdos de ferro cristalino e não cristalino do solo.

A variação da geologia influencia diretamente nas características do solo, o que impacta nos dados obtidos pelos sensores e na predição deles.

Os modelos de ferro cristalino são mais acurados quando comparados com os de ferro não cristalino.

Destaca-se a maior eficiência da espectroscopia de fluorescência de raios X em relação à espectroscopia de reflectância difusa e suscetibilidade magnética para a predição dos atributos individualmente.

2.5 Referências

Abdikan, S., Sekertekin, A., Narin, O.G., Delen, A., Balik Sanli, F., 2022. A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height

- estimation of sunflower using Sentinel-1 and Sentinel-2. *Advances in Space Research*. <https://doi.org/10.1016/J.ASR.2022.11.046>
- Afshar, F.A., Ayoubi, S., Jafari, A., 2018a. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Afshar, F.A., Ayoubi, S., Jafari, A., 2018b. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R. de S., Fernandes, K., Marques Júnior, J., 2021. Machine learning in the prediction of sugarcane production environments. *Comput Electron Agric* 190. <https://doi.org/10.1016/j.compag.2021.106452>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R.D.S., Júnior, J.M., 2020. Aprendizagem de máquina na previsão de ambientes de produção de cana-de-açúcar. *Comput Electron Agric*.
- Andrade, R., Mancini, M., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/J.GEODERMA.2022.115936>
- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020a. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>
- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020b. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>
- Anzanello, M.J., Ortiz, R.S., Limbergerb, R.P., Mayorga, P., 2013. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. *J Pharm Biomed Anal* 83, 209–214. <https://doi.org/10.1016/J.JPBA.2013.05.004>
- Bachmann, C.M., Philpot, W., Abelev, A., Korwan, D., 2014. Phase angle dependence of sand density observable in hyperspectral reflectance. *Remote Sens Environ* 150, 53–65. <https://doi.org/10.1016/J.RSE.2014.03.024>
- Bahia, A.S.R. de S., Marques, J., La Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017a. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R. de S., Marques, J., la Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017b. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R.D.S., Marques, J., Siqueira, D.S., 2015. Procedures using diffuse reflectance spectroscopy for estimating hematite and goethite in Oxisols of São Paulo, Brazil. *Geoderma Regional* 5, 150–156. <https://doi.org/10.1016/j.geodrs.2015.04.006>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021a. Soil mineralogical attributes estimated by color as accessed by proximal sensors and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021b. Soil mineralogical attributes estimated by color as accessed by proximal sensors

- and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Balsam, W.L., Otto-Bliesner, B.L., Deaton, B.C., 1995. Modern and Last Glacial Maximum eolian sedimentation patterns in the Atlantic Ocean interpreted from sediment iron oxide content. *Paleoceanography* 10, 493–507. <https://doi.org/10.1029/95PA00421>
- Baseca, C.C., Sendra, S., Lloret, J., Tomas, J., 2019. A smart decision system for digital farming. *Agronomy* 9. <https://doi.org/10.3390/agronomy9050216>
- Basu, D., Sinha, R., Sahu, S., Malla, J., Chakravorty, N., Ghosal, P.S., 2022. Identification of severity and passive measurement of oxidative stress biomarkers for β -thalassemia patients: K-means, random forest, XGBoost, decision tree, neural network based novel framework. *Advances in Redox Research* 5, 100034. <https://doi.org/10.1016/J.ARRES.2022.100034>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019a. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582–592. <https://doi.org/10.1016/j.compag.2019.05.008>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019b. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582–592. <https://doi.org/10.1016/j.compag.2019.05.008>
- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015a. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1–14. <https://doi.org/10.1016/j.aca.2015.04.042>
- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015b. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1–14. <https://doi.org/10.1016/j.aca.2015.04.042>
- Bowden, C., Foster, T., Parkes, B., 2023. Identifying links between monsoon variability and rice production in India through machine learning. *Sci Rep* 13, 1–12. <https://doi.org/10.1038/s41598-023-27752-8>
- Breiman, L., 2001. Random Forest. *Mach Learn* 45, 5–32. <https://doi.org/10.1017/CBO9781107415324.004>
- Breiman, Leo, 2001. Random forests. *Mach Learn* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bui, D.T., Tsangaratos, P., Nguyen, V.T., Liem, N. Van, Trinh, P.T., 2020. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena (Amst)* 188. <https://doi.org/10.1016/j.catena.2019.104426>
- CAMARGO, O.A. de, MONIZ, A.C., JORGE, J.A.J., VALADARES, J.M.A.S., 1986. Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agrônomo de Campinas, Boletim Técnico, 106. ed.
- Camargo, L.A., Marques Júnior, J., Pereira, G.T., Horvat, R.A., 2009. Variabilidade espacial de atributos mineralógicos de um Latossolo sob diferentes formas do relevo. II - Correlação espacial entre mineralogia e agregados. *Rev Bras Cienc Solo* 32, 2279–2288. <https://doi.org/10.1590/s0100-06832008000600007>
- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013a. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341–353. <https://doi.org/10.1016/J.GEODERMA.2012.08.020>

- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013b. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341–353. <https://doi.org/10.1016/J.GEODERMA.2012.08.020>
- Camêlo, D. de L., Ker, J.C., Fontes, M.P.F., Corrêa, M.M., da Costa, A.C.S., Melo, V.F., 2017. Pedogenic Iron Oxides in Iron-Rich Oxisols Developed from Mafic Rocks. *Rev Bras Cienc Solo* 41. <https://doi.org/10.1590/18069657RBCS20160379>
- Canizo, B. V., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514–522. <https://doi.org/10.1016/j.compag.2019.04.043>
- Canizo, B. v., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514–522. <https://doi.org/10.1016/j.compag.2019.04.043>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016a. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232–240. <https://doi.org/10.1016/j.catena.2016.01.001>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016b. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232–240. <https://doi.org/10.1016/J.CATENA.2016.01.001>
- Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7, 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, T., Guestrin, C., 2016a. XGBoost: A scalable tree boosting system, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T., Guestrin, C., 2016b. XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 13-17-August-2016, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., Vašát, R., 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. *Catena (Amst)* 189, 104485. <https://doi.org/10.1016/j.catena.2020.104485>
- Cornell, R.M., Schwertmann, U., 2006. The Iron Oxides: Structure, Properties, Reactions, Occurrences and Uses 707.
- da Costa, A.C.S., Junior, I.G. de S., Canton, L.C., Gil, L.G., Figueiredo, R., 2020. Contribution of the chemical and mineralogical properties of sandy-loam tropical soils to the cation exchange capacity. *Rev Bras Cienc Solo* 44, 1–18. <https://doi.org/10.36783/18069657rbcs20200019>
- de Castro, C.L., Braga, A.P., 2011. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica* 22, 441–466. <https://doi.org/10.1590/S0103-17592011000500002>
- de Motomiya, A.V.A., Corá, J.E., Pereira, G.T., 2006. Uso da krigagem indicatriz na avaliação de indicadores de fertilidade do solo. *Rev Bras Cienc Solo* 30, 485–496. <https://doi.org/10.1590/S0100-06832006000300010>

- Demattê, José Alexandre M, Dotto, A.C., Bedin, L.G., Sayão, V.M., Barros, A., 2019. Geoderma Soil analytical quality control by traditional and spectroscopy techniques : Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma* 337, 111–121. <https://doi.org/10.1016/j.geoderma.2018.09.010>
- Demattê, José A.M., Dotto, A.C., Paiva, A.F.S., Sato, M. V., Dalmolin, R.S.D., de Araújo, M. do S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E. V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C. de V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Brighenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* 354, 113793. <https://doi.org/10.1016/j.geoderma.2019.05.043>
- Dietterich, T., 1995. Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)* 27, 326–327. <https://doi.org/10.1145/212094.212114>
- dos Santos, F.R., de Oliveira, J.F., Barbosa, G.M.C., Melquiades, F.L., 2021. Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling. *Spectrochim Acta Part B At Spectrosc* 185, 106303. <https://doi.org/10.1016/J.SAB.2021.106303>
- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008a. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009–1016. <https://doi.org/10.1590/s0100-06832008000300010>
- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008b. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009–1016. <https://doi.org/10.1590/s0100-06832008000300010>
- Ebtehaj, I., Bonakdari, H., 2022. A reliable hybrid outlier robust non-tuned rapid machine learning model for multi-step ahead flood forecasting in Quebec, Canada. *J Hydrol (Amst)* 614, 128592. <https://doi.org/10.1016/J.JHYDROL.2022.128592>
- EMBRAPA, 2017. Manual de metodos de análises, in: Manual de Métodos de Análise de Solo. Brasília, DF, p. 574.
- Esfandiarpour-Boroujeni, I., Shamsabadi, M.S., Shirani, H., Mosleh, Z., Bagheri Bodaghabadi, M., Salehi, M.H., 2020. Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations. *Catena (Amst)* 191, 104581. <https://doi.org/10.1016/J.CATENA.2020.104581>
- Fan, S.S., Chang, F.H., Hsueh, H.T., Ko, T.H., 2016. Measurement of Total Free Iron in Soils by H₂S Chemisorption and Comparison with the Citrate

- Bicarbonate Dithionite Method. *J Anal Methods Chem* 2016. <https://doi.org/10.1155/2016/7213542>
- Faria, A.J.G. de, Silva, S.H.G., Andrade, R., Mancini, M., Melo, L.C.A., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Prediction of soil organic matter content by combining data from Nix ProTM color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional* 28, e00461. <https://doi.org/10.1016/J.GEODRS.2021.E00461>
- Fernandes, L.A., Castro, A.B. de, Basilici, G., 2007. Seismites in continental sand sea deposits of the Late Cretaceous Caiuá Desert , Bauru Basin, Brasil. *Sediment Geol* 199, 61–64. <https://doi.org/10.1016/j.sedgeo.2005.12.030>
- Fernandes, R.B.A., Barrón, V., Torrent, J., Fontes, M.P.F., 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de refletância difusa. *Rev Bras Cienc Solo* 28, 245–257. <https://doi.org/10.1590/S0100-06832004000200003>
- Fink, J.R., Inda, A. v., Bavaresco, J., Barrón, V., Torrent, J., Bayer, C., 2016. Adsorption and desorption of phosphorus in subtropical soils as affected by management system and mineralogy. *Soil Tillage Res* 155, 62–68. <https://doi.org/10.1016/J.STILL.2015.07.017>
- Fonseca, J. da S., Campos, M.C.C., Brito Filho, E.G. de, Mantovanelli, B.C., Silva, L.S., de Lima, A.F.L., Da Cunha, J.M., Simões, E.L., dos Santos, L.A.C., 2021. Soil–landscape relationship in a sandstone-gneiss topolithosequence in the State of Amazonas, Brazil. *Environ Earth Sci* 80, 1–15. <https://doi.org/10.1007/s12665-021-10026-9>
- Ganaie, M.A., Tanveer, M., Suganthan, P.N., Snasel, V., 2022. Oblique and rotation double random forest. *Neural Networks* 153, 496–517. <https://doi.org/10.1016/J.NEUNET.2022.06.012>
- Ghafarian, F., Wieland, R., Lüttschwager, D., Nendel, C., 2022. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environmental Modelling & Software* 156, 105466. <https://doi.org/10.1016/J.ENVSOFT.2022.105466>
- Gholizadeh, A., Luboš, B., Saberioon, M., Vašát, R., 2013. Visible, Near-Infrared, and Mid-Infrared Spectroscopy Applications for Soil Assessment with Emphasis on Soil Organic Matter Content and Quality: State-of-the-Art and Key Issues. <http://dx.doi.org/10.1366/13-07288> 67, 1349–1362. <https://doi.org/10.1366/13-07288>
- Gholizadeh, A., Saberioon, M., Pouladi, N., Ben-Dor, E., 2023. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. *International Soil and Water Conservation Research* 11, 112–124. <https://doi.org/10.1016/J.ISWCR.2022.06.004>
- Goldshleger, N., Ben-Dor, E., Benyamini, Y., Agassi, M., 2004. SOIL REFLECTANCE AS A TOOL FOR ASSESSING PHYSICAL CRUST ARRANGEMENT OF FOUR TYPICAL SOILS IN ISRAEL. *Soil Sci* 169, 677–687. <https://doi.org/10.1097/01.ss.0000146024.61559.e2>
- Golkarian, A., Khosravi, K., Panahi, M., Clague, J.J., 2023. Spatial variability of soil water erosion: Comparing empirical and intelligent techniques. *Geoscience Frontiers* 14, 101456. <https://doi.org/10.1016/J.GSF.2022.101456>
- Gómez-Escalonilla, Diancoumba, O., Traoré, D.Y., Montero, E., Martín-Loeches, M., Martínez-Santos, P., 2022. Multiclass spatial predictions of borehole yield in southern Mali by means of machine learning classifiers. *J Hydrol Reg Stud*

- 44, 101245. <https://doi.org/10.1016/J.EJRH.2022.101245>
- Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. *Catena (Amst)* 147, 429–440. <https://doi.org/10.1016/j.catena.2016.07.045>
- Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy* 131, 1–109. <https://doi.org/10.1016/BS.AGRON.2014.12.004>
- Hall, D.L., Llinas, J., 2016. An introduction to multi-sensor data fusion. *Sensors, Nanoscience, Biomedical Engineering, and Instruments* 85.
- Han, L., Yang, G., Yang, X., Song, X., Xu, B., Li, Z., Wu, Jintao, Yang, H., Wu, Jianwei, 2022. An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images. *Comput Electron Agric* 194, 106804. <https://doi.org/10.1016/J.COMPAG.2022.106804>
- He, M., Tang, L., Li, C., Ren, J., Zhang, L., Li, X., 2022. Dynamics of soil organic carbon and nitrogen and their relations to hydrothermal variability in dryland. *J Environ Manage* 319, 115751. <https://doi.org/10.1016/J.JENVMAN.2022.115751>
- Heuvelink, G.B.M., Brus, D.J., Rossiter, D.G., Shi, Z., 2019. Editorial for pedometrics 2017 special issue. *Eur J Soil Sci* 70, 25–26. <https://doi.org/10.1111/ejss.12772>
- Hikouei, I.S., Kim, S.S., Mishra, D.R., 2021. Machine-learning classification of soil bulk density in salt marsh environments. *Sensors* 21. <https://doi.org/10.3390/s21134408>
- Holmgren, G.G.S., 1967a. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210–211. <https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Holmgren, G.G.S., 1967b. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210–211. <https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Hong, H., Ji, K., Hei, H., Wang, C., Liu, C., Zhao, L., Lanson, B., Zhao, C., Fang, Q., Algeo, T.J., 2023. Clay mineral evolution and formation of intermediate phases during pedogenesis on picrite basalt bedrock under temperate conditions (Yunnan, southwestern China). *Catena (Amst)* 220, 106677. <https://doi.org/10.1016/J.CATENA.2022.106677>
- Isaaks, E.H., Srivastava, R.M., 1989. *An Introduction to Applied Geostatistics, Computers & Geosciences*. Pergamon. [https://doi.org/10.1016/0098-3004\(91\)90055-I](https://doi.org/10.1016/0098-3004(91)90055-I)
- Jararweh, Y., Fatima, S., Jarrah, M., AlZu'bi, S., 2023. Smart and sustainable agriculture: Fundamentals, enabling technologies, and future directions. *Computers and Electrical Engineering* 110, 108799. <https://doi.org/10.1016/j.compeleceng.2023.108799>
- Jha, K., Doshi, A., Patel, P., Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture* 2, 1–12. <https://doi.org/10.1016/J.AIIA.2019.05.004>
- Kampf, N., Schwertmann, U., 1982. The 5-M-NaOH concentration treatment for iron oxides in soils. *Clays Clay Miner* 30, 401–408. <https://doi.org/10.1346/CCMN.1982.0300601/METRICS>
- Karthikeyan, L., Mishra, A.K., 2021. Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sens*

- Environ 266, 112706. <https://doi.org/10.1016/J.RSE.2021.112706>
- Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. *Inf Sci (N Y)* 585, 609–629. <https://doi.org/10.1016/J.INS.2021.11.036>
- Kaya, F., Başayığit, L., Keshavarzi, A., Francaviglia, R., 2022. Digital mapping for soil texture class prediction in northwestern Türkiye by different machine learning algorithms. *Geoderma Regional* 31, e00584. <https://doi.org/10.1016/J.GEODRS.2022.E00584>
- Ker, J.C., 1997. LATOSSOLOS DO BRASIL: UMA REVISÃO. *Geonomos* 5, 17–40. <https://doi.org/10.18285/GEONOMOS.V5I1.187>
- Ker, J.C., Curi, N., Schaefer, C.E., Torrado, P.V., 2012. *Pedologia: fundamentos. Viçosa*.
- Kiangala, S.K., Wang, Z., 2021. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications* 4, 100024. <https://doi.org/10.1016/J.MLWA.2021.100024>
- Kragt, M.E., Pannell, D.J., Robertson, M.J., Thamo, T., 2012. Assessing costs of soil carbon sequestration by crop-livestock farmers in Western Australia. *Agric Syst* 112, 27–37. <https://doi.org/10.1016/j.agsy.2012.06.005>
- Kuncheva, L.I., Matthews, C.E., Arnaiz-González, Á., Rodríguez, J.J., 2020. Feature Selection from High-Dimensional Data with Very Low Sample Size: A Cautionary Tale.
- Lark, R.M., 2000. Designing sampling grids from imprecise information on soil variability, an approach based on the fuzzy kriging variance. *Geoderma* 98, 35–59. [https://doi.org/10.1016/S0016-7061\(00\)00051-3](https://doi.org/10.1016/S0016-7061(00)00051-3)
- Lee, M.E., Jeon, E.K., Tsang, D.C.W., Baek, K., 2018. Simultaneous application of oxalic acid and dithionite for enhanced extraction of arsenic bound to amorphous and crystalline iron oxides. *J Hazard Mater* 354, 91–98. <https://doi.org/10.1016/J.JHAZMAT.2018.04.083>
- Liu, J. cheng, He, H. ping, Michalski, J., Cuadros, J., Yao, Y. zeng, Tan, W., Qin, X. rong, Li, S. ying, Wei, G. jian, 2021. Reflectance spectroscopy applied to clay mineralogy and alteration intensity of a thick basaltic weathering sequence in Hainan Island, South China. *Appl Clay Sci* 201, 105923. <https://doi.org/10.1016/J.CLAY.2020.105923>
- Liu, J., Yang, K., Tariq, A., Lu, L., Soufan, W., El Sabagh, A., 2023. Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. *Egyptian Journal of Remote Sensing and Space Science* 26, 415–426. <https://doi.org/10.1016/j.ejrs.2023.05.005>
- Liu, Q., He, L., Guo, L., Wang, M., Deng, D., Lv, P., Wang, R., Jia, Z., Hu, Z., Wu, G., Shi, T., 2022. Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. *Catena (Amst)* 219, 106603. <https://doi.org/10.1016/J.CATENA.2022.106603>
- Liu, X., Zhu, A.X., Yang, L., Pei, T., Liu, J., Zeng, C., Wang, D., 2020. A graded proportion method of training sample selection for updating conventional soil maps. *Geoderma* 357, 113939. <https://doi.org/10.1016/j.geoderma.2019.113939>
- Lorz, C., Fruhauf, M., Mailander, R., Phillips, J.D., Kleber, A., 2013. Influence of Cover Beds on Soils, Mid-Latitude Slope Deposits (Cover Beds).

- Lu, W.N., He, Y., Wang (王阳), Y., Ke, S., 2020. Behavior of calcium isotopes during continental subduction recorded in meta-basaltic rocks. *Geochim Cosmochim Acta* 278, 392–404. <https://doi.org/10.1016/J.GCA.2019.09.027>
- Ma, G., Ding, J., Han, L., Zhang, Z., Ran, S., 2021. Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. *Regional Sustainability* 2, 177–188. <https://doi.org/10.1016/J.REGSUS.2021.06.001>
- Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma* 262, 243–253. <https://doi.org/10.1016/J.GEODERMA.2015.08.037>
- Mancini, M., Andrade, R., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for Brazilian soil properties prediction: Exchangeable/available macronutrients, aluminum, and potential acidity. *Geoderma Regional* 30, e00573. <https://doi.org/10.1016/J.GEODRS.2022.E00573>
- Marques, J., Siqueira, D.S., Camargo, L.A., Teixeira, D.D.B., Barrón, V., Torrent, J., 2014. Magnetic susceptibility and diffuse reflectance spectroscopy to characterize the spatial variability of soil properties in a brazilian haplustalf. *Geoderma* 219–220, 63–71. <https://doi.org/10.1016/j.geoderma.2013.12.007>
- Maurya, A.K., Bhargava, N., Singh, D., 2022a. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795–1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- Maurya, A.K., Bhargava, N., Singh, D., 2022b. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795–1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- McBratney, A., Gruijter, J., Bryce, A., 2019. Pedometrics timeline. *Geoderma* 338, 568–575. <https://doi.org/10.1016/j.geoderma.2018.11.048>
- McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, *Geoderma*. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- McKeague, J.A., Day, J.H., 1966. DITHIONITE- AND OXALATE-EXTRACTABLE Fe AND Al AS AIDS IN DIFFERENTIATING VARIOUS CLASSES OF SOILS. *Can J Soil Sci* 46, 13–22. <https://doi.org/10.4141/cjss66-003>
- Mehra, O.P., Jackson, M.L., 1958. Iron Oxide Removal from Soils and Clays by a Dithionite-Citrate System Buffered with Sodium Bicarbonate. *Clays and Clay Minerals* 1958 7:1 7, 317–327. <https://doi.org/10.1346/CCMN.1958.0070122>
- Mello, F.A.O., Bellinaso, H., Mello, D.C., Safanelli, J.L., Mendes, W.D.S., Amorim, M.T.A., Gomez, A.M.R., Poppiel, R.R., Silvero, N.E.Q., Gholizadeh, A., Silva, S.H.G., Curi, N., Demattê, J.A.M., 2021. Soil parent material prediction through satellite multispectral analysis on a regional scale at the Western Paulista Plateau, Brazil. *Geoderma Regional* 26, e00412. <https://doi.org/10.1016/J.GEODRS.2021.E00412>
- Mendes, W. de S., Demattê, J.A.M., Bonfatti, B.R., Resende, M.E.B., Campos, L.R., Costa, A.C.S. da, 2021. A novel framework to estimate soil mineralogy using soil spectroscopy. *Applied Geochemistry* 127, 104909. <https://doi.org/10.1016/J.APGEOCHEM.2021.104909>
- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022a. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219. <https://doi.org/10.1016/j.still.2022.105346>

- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022b. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219, 105346. <https://doi.org/10.1016/J.STILL.2022.105346>
- Merl, T., Rasmussen, M.R., Koch, L.R., Søndergaard, J.V., Bust, F.F., Koren, K., 2022. Measuring soil pH at in situ like conditions using optical pH sensors (pH-optodes). *Soil Biol Biochem* 175, 108862. <https://doi.org/10.1016/J.SOILBIO.2022.108862>
- Minasny, B., Hartemink, A.E., 2011a. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52–62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., Hartemink, A.E., 2011b. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52–62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. *Geoderma* 264, 301–311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019a. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194, 104308. <https://doi.org/10.1016/j.still.2019.104308>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019b. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194. <https://doi.org/10.1016/j.still.2019.104308>
- Naimi, S., Ayoubi, S., di Raimo, L.A.D.L., Dematte, J.A.M., 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. *Geoderma Regional* 28, e00484. <https://doi.org/10.1016/J.GEODRS.2022.E00484>
- Nawar, S., Delbecque, N., Declercq, Y., Smedt, P. De, Finke, P., Verdoodt, A., Meirvenne, M. Van, Mouazen, A.M., 2019. Geoderma Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry? *Geoderma* 350, 29–39. <https://doi.org/10.1016/j.geoderma.2019.05.002>
- Nawar, S., Richard, F., Kassim, A.M., Tekin, Y., Mouazen, A.M., 2022. Fusion of Gamma-rays and portable X-ray fluorescence spectral data to measure extractable potassium in soils. *Soil Tillage Res* 223, 105472. <https://doi.org/10.1016/J.STILL.2022.105472>
- Ng, W., Minasny, B., McBratney, A., 2020a. Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Mcbratney, A., 2020b. Science of the Total Environment Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352, 251–267. <https://doi.org/10.1016/j.geoderma.2019.06.016>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022a. A novel intelligence approach based

- active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of the Total Environment* 804. <https://doi.org/10.1016/j.scitotenv.2021.150187>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022b. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of The Total Environment* 804, 150187. <https://doi.org/10.1016/J.SCITOTENV.2021.150187>
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy* 132, 139–159. <https://doi.org/10.1016/bs.agron.2015.02.002>
- Norrish, K., Taylor, R.M., 1961. The isomorphous replacement of iron by aluminium in soil goethites. *Journal of Soil Science* 12, 294–306. <https://doi.org/10.1111/J.1365-2389.1961.TB00919.X>
- Novais, R.F., Alvarez V., V.H., Barros, N.F. de, Fontes, R.L.F., Cantarutti, R.B., Neves, J.C.L., 2007. Fertilidade do solo. Viçosa.
- Nowak-Brzezinska, A., Horyn, C., 2020. Outliers in rules - the comparision of LOF, COF and KMEANS algorithms. *Procedia Comput Sci* 176, 1420–1429. <https://doi.org/10.1016/J.PROCS.2020.09.152>
- Oliveira, C. v., Ker, J.C., Fontes, L.E.F., Curi, N., Pinheiro, J.C., 1998. Química e mineralogia de solos derivados de rochas do Grupo Bambuí no norte de Minas Gerais. *Rev Bras Cienc Solo* 22, 583–593. <https://doi.org/10.1590/S0100-06831998000400003>
- Oliver, M.A., Webster, R., 2014a. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56–69. <https://doi.org/10.1016/J.CATENA.2013.09.006>
- Oliver, M.A., Webster, R., 2014b. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56–69. <https://doi.org/10.1016/j.catena.2013.09.006>
- Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16. <https://doi.org/10.1016/j.geodrs.2018.e00198>
- Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* 12, 2825–2830. <https://doi.org/10.5555/1953048.2078195>
- Pedron, F. de A., Lourenzi, C.R., Ceretta, C.A., Lorensi, J., Cancian, A., 2018. Clay mineralogy of subtropical soils under long-term organic fertilization in no-tillage systems. *Rev Bras Cienc Solo* 42. <https://doi.org/10.1590/18069657rbcs20170092>
- Ramaroson, V.H., Becquer, T., Sá, S.O., Razafimahatratra, H., Delarivière, J.L., Blavet, D., Vendrame, P.R.S., Rabeharisoa, L., Rakotondrazafy, A.F.M., 2018. Mineralogical analysis of ferrallitic soils in Madagascar using NIR spectroscopy. *Catena (Amst)* 168, 102–109. <https://doi.org/10.1016/j.catena.2017.07.016>
- Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L., El Hadrami, E.M.,

2019. A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems* 195. <https://doi.org/10.1016/j.chemolab.2019.103873>
- Rennert, T., 2019. Wet-chemical extractions to characterise pedogenic Al and Fe species-a critical review. *Soil Research* 57, 1–16. <https://doi.org/10.1071/SR18299>
- Rosin, N.A., Demattê, J.A.M., Leite, M.C.A., de Carvalho, H.W.P., Costa, A.C., Greschuk, L.T., Curi, N., Silva, S.H.G., 2022. The fundamental of the effects of water, organic matter, and iron forms on the pXRF information in soil analyses. *Catena (Amst)* 210. <https://doi.org/10.1016/j.catena.2021.105868>
- Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. *Geoderma* 324, 131–137. <https://doi.org/10.1016/j.geoderma.2018.03.009>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022a. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022b. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruiz, H.A., Ferreira, G.B., Pereira, J.B.M., Resumo,), 2003. Estimativa da capacidade de campo de Latossolos e Neossolos Quartzarênicos pela determinação do equivalente de umidade. *Rev Bras Cienc Solo* 27, 389–393. <https://doi.org/10.1590/S0100-06832003000200019>
- Santos, H.G. dos, Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.Á. de, Lumberras, J.F., Coelho, M.R., Almeida, J.A. de, Filho, J.C. de A., Oliveira, J.B. de, Cunha, T.J.F., 2018. Sistema brasileiro de classificação de solos, 5th ed, Embrapa Solos. Embrapa, Brasília.
- Sarfraz, S., Ali, F., Hameed, A., Ahmad, Z., Riaz, K., 2023. Sustainable Agriculture Through Technological Innovations. *Sustainable Agriculture in the Era of the OMICs Revolution* 223–239. https://doi.org/10.1007/978-3-031-15568-0_10
- Schwertmann, U., Taylor, R.M., 2018. Iron oxides. *Minerals in Soil Environments* 379–438. <https://doi.org/10.2136/sssabookser1.2ed.c8>
- Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S. V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep* 11, 1–15. <https://doi.org/10.1038/s41598-020-80820-1>
- Shrestha, G., Calvelo-Pereira, R., Roudier, P., Martin, A.P., Turnbull, R.E., Kereszturi, G., Jeyakumar, P., Anderson, C.W.N., 2022. Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy. *Geoderma* 409, 115649. <https://doi.org/10.1016/J.GEODERMA.2021.115649>
- Silva, F.M., Silva, S.H.G., Acuña-Guzman, S.F., Silva, E.A., Ribeiro, B.T., Fruett, T., Inda, A.V., Teixeira, A.F. dos S., Mancini, M., Guilherme, L.R.G., Curi, N., 2021. Chemical and mineralogical changes in the textural fractions of quartzite-derived tropical soils, along weathering, assessed by portable X-ray fluorescence spectrometry and X-ray diffraction. *J South Am Earth Sci* 112, 103634. <https://doi.org/10.1016/J.JSAMES.2021.103634>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf,

- D.C., Guilherme, L.R.G., Curi, N., 2022a. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, 1–10. <https://doi.org/10.1016/j.geodrs.2021.e00465>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022b. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, e00465. <https://doi.org/10.1016/J.GEODRS.2021.E00465>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020a. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/j.catena.2019.104258>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020b. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/J.CATENA.2019.104258>
- Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. *Geoderma* 267, 65–77. <https://doi.org/10.1016/J.GEODERMA.2015.12.025>
- Silvero, N.E.Q., Siqueira, D.S., Coelho, R.M., da Costa Ferreira, D., Marques, J., 2019. Protocol for the use of legacy data and magnetic signature on soil mapping of São Paulo Central West, Brazil. *Science of The Total Environment* 693, 133463. <https://doi.org/10.1016/J.SCITOTENV.2019.07.269>
- Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl Spectrosc Rev* 49, 139–186. <https://doi.org/10.1080/05704928.2013.811081>
- Sothe, C., Almeida, C.M. De, Schimalski, M.B., Liesenberg, V., Rosa, L.E.C. La, Castro, J.D.B., Feitosa, R.Q., 2019. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. <https://doi.org/ez87.periodicos.capes.gov.br/10.1080/01431161.2019.1681600>. <https://doi.org/10.1080/01431161.2019.1681600>
- Sothe, C., de Almeida, C.M., Schimalski, M.B., Liesenberg, V., la Rosa, L.E.C., Castro, J.D.B., Feitosa, R.Q., 2020. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. *Int J Remote Sens* 41, 1943–1969. <https://doi.org/10.1080/01431161.2019.1681600>
- Souza, E.D., Carneiro, M.A.C., Paulino, H.B., 2005. Atributos físicos de um Neossolo Quartzarênico e um Latossolo Vermelho sob diferentes sistemas de manejo. *Pesqui Agropecu Bras* 40, 1135–1139. <https://doi.org/10.1590/S0100-204X2005001100012>
- Stranghoener, M., Dultz, S., Behrens, H., Schippers, A., 2020. Potential mobilizable Fe from secondary phases of differentially altered subsurface basaltic rock– a sequential extraction study on ICDP site Hawaii. *Applied Geochemistry* 121, 104705. <https://doi.org/10.1016/J.APGEOCHEM.2020.104705>
- Suits, D.B., 1957. Use of Dummy Variables in Regression Equations. *J Am Stat Assoc* 52, 548. <https://doi.org/10.2307/2281705>

- Sumathi, K., 2018. Data Analytics platform for intelligent agriculture. 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on 647–650. <https://doi.org/10.1109/I-SMAC.2018.8653740>
- Sun, Z., Wang, J., Wang, Y., Long, L., Luo, Z., Deng, X., Hu, Q., Wang, M., 2020. Sodium-rich volcanic rocks and their relationships with iron deposits in the Aqishan–Yamansu belt of Eastern Tianshan, NW China. *Geoscience Frontiers* 11, 697–713. <https://doi.org/10.1016/J.GSF.2019.06.011>
- Swetha, R.K., Dasgupta, S., Chakraborty, S., Li, B., Weindorf, D.C., Mancini, M., Silva, S.H.G., Ribeiro, B.T., Curi, N., Ray, D.P., 2022. Using Nix color sensor and Munsell soil color variables to classify contrasting soil types and predict soil organic carbon in Eastern India. *Comput Electron Agric* 199, 107192. <https://doi.org/10.1016/J.COMPAG.2022.107192>
- Tabatabai, S., Knadel, M., Thomsen, A., Greve, M.H., 2019. On-the-Go Sensor Fusion for Prediction of Clay and Organic Carbon Using Pre-processing Survey, Different Validation Methods, and Variable Selection. *Soil Science Society of America Journal* 83, 300–310. <https://doi.org/10.2136/SSSAJ2018.10.0377>
- Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma* 428, 116192. <https://doi.org/10.1016/J.GEODERMA.2022.116192>
- Teixeira, A.F. dos S., Andrade, R., Mancini, M., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *J South Am Earth Sci* 116, 103873. <https://doi.org/10.1016/J.JSAMES.2022.103873>
- Teixeira, A.F. dos S., Henrique Procópio Pelegrino, M., Missina Faria, W., Henrique Godinho Silva, S., Gabriela Marcolino Gonçalves, M., Weimar Acerbi Júnior, F., Rezende Gomide, L., Linares Pádua Júnior, A., de Souza, I.A., Chakraborty, S., Weindorf, D.C., Roberto Guimarães Guilherme, L., Curi, N., 2020. Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. *Geoderma* 361, 114132. <https://doi.org/10.1016/j.geoderma.2019.114132>
- Teixeira, R.S., Vendrame, P.R.S., Christoni, A.R.F., SILVA, P.R.C., 2016. QUÍMICA E MINERALOGIA DE UM SOLO DESENVOLVIDO DE BASALTO, COLETADO ATRAVÉS DE SONDAGEM SPT. *Boletim de Geografia* 34, 116–126. <https://doi.org/10.4025/bolgeogr.v34i2.28948>
- Temme, A.J.A.M., Veldkamp, A., 2009. Multi-process Late Quaternary landscape evolution modelling reveals lags in climate response over small spatial scales. *Earth Surf Process Landf* 34, 573–589. <https://doi.org/10.1002/ESP.1758>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011a. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812–816. <https://doi.org/10.1590/S0103-84782011000500012>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011b. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812–816. <https://doi.org/10.1590/S0103-84782011000500012>
- Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital Soil

- Mapping: Interactions with and Applications for Hydropedology, in: *Hydropedology*. Elsevier, pp. 665–709. <https://doi.org/10.1016/B978-0-12-386941-8.00021-6>
- Thornthwaite, C.W., 1948. An Approach toward a Rational Classification of Climate Author (s): C. W. Thornthwaite Published by: American Geographical Society Stable URL : <http://www.jstor.org/stable/210739> Accessed : 11-08-2016 18 : 02 UTC. *Geogr Rev* 38, 55–94. <https://doi.org/10.2307/210739>
- Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* 367. <https://doi.org/10.1016/j.geoderma.2020.114208>
- Vasques, G.M., Rodrigues, H.M., Coelho, M.R., Baca, J.F.M., Dart, R.O., Oliveira, R.P., Teixeira, W.G., Ceddia, M.B., 2020. Field Proximal Soil Sensor Fusion for Improving High-Resolution Soil Property Maps. *Soil Systems* 2020, Vol. 4, Page 52 4, 52. <https://doi.org/10.3390/SOILSYSTEMS4030052>
- Vieira, S.R., 2000. Geoestatística em estudos de variabilidade espacial do solo. , in: R.F. Novais, V.H.A.G.R.S. (Ed.), *Tópicos Em Ciência Do Solo*. Viçosa, pp. 1–54.
- Wang, J.H., Jiang, J.H., Yu, R.Q., 1996. Robust back propagation algorithm as a chemometric tool to prevent the overfitting to outliers. *Chemometrics and Intelligent Laboratory Systems* 34, 109–115. [https://doi.org/10.1016/0169-7439\(96\)00005-6](https://doi.org/10.1016/0169-7439(96)00005-6)
- Wang, S., Peng, H., Hu, Q., Jiang, M., 2022. Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *J Hydrol Reg Stud* 42, 101139. <https://doi.org/10.1016/J.EJRH.2022.101139>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014a. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, *Advances in Agronomy*. Elsevier. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014b. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, in: *Advances in Agronomy*. Academic Press Inc., pp. 1–45. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Paulette, L., Man, T., 2013. In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. *Environmental Pollution* 182, 92–100. <https://doi.org/10.1016/j.envpol.2013.07.008>
- Whittig, L.D., 1965. X-Ray Diffraction Techniques for Mineral Identification and Mineralogical Composition. *Methods of Soil Analysis, Part 1: Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling* 671–698. <https://doi.org/10.2134/AGRONMONOGR9.1.C49>
- Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. *Environmental Pollution* 263, 114649. <https://doi.org/10.1016/J.ENVPOL.2020.114649>
- Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. *Eur J Soil Sci* 70, 162–173. <https://doi.org/10.1111/EJSS.12729>
- Xu, G., Fan, H., Oliver, D.M., Dai, Y., Li, H., Shi, Y., Long, H., Xiong, K., Zhao, Z., 2022. Decoding river pollution trends and their landscape determinants in an

- ecologically fragile karst basin using a machine learning model. *Environ Res* 214, 113843. <https://doi.org/10.1016/J.ENVRES.2022.113843>
- Xu, S., Zhao, Y., Wang, M., Shi, X., 2018. Quantification of Different Forms of Iron from Intact Soil Cores of Paddy Fields with Vis-NIR Spectroscopy. *Soil Science Society of America Journal* 82, 1497–1511. <https://doi.org/10.2136/SSSAJ2018.01.0014>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020a. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Sci Data* 7, 1–13. <https://doi.org/10.1038/s41597-019-0345-6>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020b. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Scientific Data* 2020 7:1 7, 1–13. <https://doi.org/10.1038/s41597-019-0345-6>
- Ye, M., Zhu, L., Li, X., Ke, Y., Huang, Y., Chen, B., Yu, H., Li, H., Feng, H., 2023. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. *Science of The Total Environment* 858, 159798. <https://doi.org/10.1016/J.SCITOTENV.2022.159798>
- Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G.A., Nannucci, M.S., Leoni, B., Bonomi, T., 2019. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. *J Hydrol (Amst)* 578, 124015. <https://doi.org/10.1016/J.JHYDROL.2019.124015>
- Zeraatpisheh, M., Ayoubi, S., Jafari, A., Finke, P., 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in Iran. *Geomorphology* 285, 186–204. <https://doi.org/10.1016/J.GEOMORPH.2017.02.015>
- Zevenbergen, C., Honders, A., Orbons, A.J., Viaene, W., Swennen, R., Comans, R.N.J., van Hasselt, H.J., 1997. Immobilisation of heavy metals in contaminated soils by thermal treatment at intermediate temperatures. *Studies in Environmental Science* 71, 661–672. [https://doi.org/10.1016/S0166-1116\(97\)80249-X](https://doi.org/10.1016/S0166-1116(97)80249-X)
- Zhan, Y., Liu, C., Deng, Q., Feng, Q., Qiu, Y., Zhang, A., He, X., 2022. Integrated FFT and XGBoost framework to predict pavement skid resistance using automatic 3D texture measurement. *Measurement* 188, 110638. <https://doi.org/10.1016/J.MEASUREMENT.2021.110638>
- Zhang, R., Li, Y., Goh, A.T.C., Zhang, W., Chen, Z., 2021. Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models. *Journal of Rock Mechanics and Geotechnical Engineering* 13, 1478–1484. <https://doi.org/10.1016/J.JRMGE.2021.08.001>
- Zhang, W. chun, Wan, H. shuang, Zhou, M. hou, Wu, W., Liu, H. bin, 2022. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecol Indic* 143, 109420. <https://doi.org/10.1016/J.ECOLIND.2022.109420>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022a. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022b. Reflectance spectroscopy for assessing heavy metal pollution indices in

- mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022a. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022b. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>

Capítulo 3 – Uso de aprendizagem de máquina para extrapolação de atributos mineralógicos do solo

Resumo: Os principais solos agricultáveis do Brasil são caracterizados como altamente intemperizados, tendo os minerais de ferro desempenhando papel fundamental na dinâmica dos solos, porém suas quantificações tradicionais implicam em altos custos e maior tempo para análise. Diante disso sensores como Espectroscopia de Reflectância Difusa (ERD) vem ganhando espaço no mercado porém em países de grandes extensões territoriais como o Brasil tamanho da área ainda é uma variável limitante, devido a isso técnicas de extrapolação de dados é uma promissora área de estudo pois ela nos permite estimar dados de áreas não visitadas. Diante desse contexto, objetivou-se (i) estimar atributos mineralógicos do solo utilizando algoritmos de aprendizagem de máquina (AM), com base nas curvas espectrais geradas pela Espectroscopia de Reflectância Difusa (ERD) em uma Área de Referência (AR) menor e com pontos de coleta mais adensados e, (ii) extrapolar os parâmetros da predição para uma Área Alvo (AA) maior, e com menos pontos de coleta. A AR está localizada no município de Guataporã, no nordeste do estado de São Paulo, Brasil, com aproximadamente 870 hectares e 360 pontos de coleta. A AA é o Planalto Ocidental Paulista (POP), que corresponde a aproximadamente 48% de todo o Estado de São Paulo, totalizando 13 milhões de hectares e com 231 pontos de coleta. Utilizamos os algoritmos de AM de Florestas Aleatórias (FA) e de XGBoost (XGB) para prever os atributos Ferro Ditionito (FED) e Ferro Oxalato (FEO) no solo e treinar os algoritmos com os parâmetros da AR e extrapolar para a AA. O algoritmo XGB foi mais eficiente na extrapolação dos parâmetros da AR para a predição dos valores de FED na AA com R^2 de 85,30% contra 76,33% do FA. Para FEO, o algoritmo de FA apresentou maior valor de $R^2=57,88\%$ contra 54,67% do algoritmo de XGB, entretanto, os valores de erro foram menores para o algoritmo de XGB. Algoritmos de AM são eficientes na predição e extrapolação de valores de atributos mineralógicos do solo, desde que sejam utilizados algoritmos robustos e que a base de dados tenha observações suficientes para treinamento adequado dos modelos. O uso do sensor ERD, associado à modelos de aprendizagem de máquina é eficiente na predição de atributos mineralógicos do solo, sendo essa associação de técnicas de grande potencial para a ciência do solo.

Palavras-chave: Mapeamento digital do solo, Florestas Aleatórias, XGBoost, Ciência de dados, Ferro Ditionito, Ferro Oxalato.

Abstract: The main arable soils in Brazil are characterized as highly weathered, with iron minerals playing a fundamental role in soil dynamics, acting as a pedoindicator for various purposes. The indirect estimation of these minerals is extremely important since traditional quantification methods imply high analysis costs and longer time to obtain data. Given this context, the objective was (i) to estimate soil mineralogical attributes using machine learning (ML) algorithms, based on the spectral curves generated by Diffuse Reflectance Spectroscopy (DRS) in a smaller Reference Area (RA) and with denser collection points, and (ii) extrapolate the prediction parameters to a larger Target Area (TA), with fewer collection points. RA is located in the municipality of Guatapar, in the northeast of the state of So Paulo, Brazil, with approximately 870 hectares and 360 collection points. The TA is the Western Paulista Plateau (WPP), which corresponds to approximately 48% of the entire State of So Paulo, totaling 13 million hectares and with 231 collection points. We used the Random Forests (RF) and XGBoost (XGB) ML algorithms to predict the Iron Dithionite (FED) and Iron Oxalate (FEO) attributes in the soil and train the algorithms with the RA parameters and extrapolate to the TA. The XGB algorithm was more efficient in extrapolating the RA parameters to predict the FED values in the TA with an $R^2 = 85.30\%$ against 76.33% for the RF. For FEO, the RF algorithm presented a higher value of $R^2 = 57.88\%$ against 54.67% for the XGB algorithm, however, the error values were lower for the XGB algorithm. ML algorithms are efficient in predicting and extrapolating soil mineralogical attribute values, provided that robust algorithms are used and that the database has enough observations for adequate training of the models. The use of the DRS sensor, associated with machine learning models, is efficient in predicting soil mineralogical attributes, and this combination of techniques has great potential for soil science.

Keywords: Digital Soil Mapping, Random Forests, XGBoost, Data Science, Dithionite iron, Oxalate iron.

3.1 Introdução

O ferro é o elemento mais frequentemente encontrado nos minerais de solos tropicais e sua ocorrência e concentração são influenciados diretamente pelos fatores e processos de formação do solo (Mendes et al., 2022b). O ferro livre é um importante indicador analítico do solo, pois ele nos permite compreender a gênese, classificação e variabilidade do solo (Fan et al., 2016).

A principal forma de obtenção dos óxidos de ferro consiste na ligação do ácido oxálico aos óxidos de ferro amorfos pela complexação entre oxalato e ferro férrico, extraíndo o atributo denominado como “Ferro Oxalato”. Para aumentar a extração óxidos de ferro cristalinos, foi proposta a aplicação simultânea de ácido oxálico e da Ditionita que é forte agente redutor, extraíndo então o atributo denominado como “Ferro Ditionito” (Zevenbergen et al., 1997; Lee et al., 2018). Ambos os procedimentos de dissolução de Oxalato e Ditionito para a extração de Fe têm ampla aplicação na classificação do solo e no estudo de sua gênese (Zevenbergen et al., 1997; Lorz et al., 2013).

Logo compreender as inter-relações e dinâmicas espaciais dos óxidos de ferro com o ambiente é necessário quando se pensa em correlacionar esses indicadores com o manejo do solo. A forma convencional de determinação dos teores de óxidos de ferro envolve extensas observações no campo associadas à complexas análises laboratoriais, onde são realizadas múltiplas extrações individuais utilizando ácidos fortes. Tais análises são destrutivas, além de implicarem em um alto investimento, morosidade na confecção dos resultados e necessidade de uma equipe altamente especializada para execução (Xu et al., 2018).

Visando tornar a obtenção dessas informações mais eficiente, o sensor de

espectroscopia de reflectância difusa (ERD) vem se mostrando uma excelente alternativa para prever com precisão e agilidade diversos atributos do solo, reduzindo assim o custo com amostragem e análise de solo (Gholizadeh et al., 2023, 2013; Naimi et al., 2022). Segundo Silva et al. (2020), a ERD, apresenta baixo custo de execução, fornece informações rápidas e gera um banco de dados robusto, além de ser um método não destrutivo e de necessitar de preparação mínima da amostra.

Entretanto, análises utilizando sensores geram grande volume de informações com dados de natureza específica de cada sensor (Almeida et al., 2021). Devido a isso e a complexidade do sistema agrícola, algoritmos de aprendizagem de máquina (AM) vem a cada dia sendo mais aplicado a ciência do solo, pois são mais precisos na estimativa de atributos de interesse do que técnicas paramétricas tradicionais, como regressões simples (Sothe et al., 2020).

Dentre os algoritmos de AM, destacam-se : a Regressão Florestas Aleatórias (FA) que é uma técnica não paramétrica que foi desenvolvida com o intuito de melhorar a previsão dos modelos de árvores de regressão, pois utiliza árvores totalmente crescidas e reduz o erro e a variação (Yan et al., 2020a) e o Extreme Gradient Boosting (XGBoost) que é algoritmo otimizado sua aprendizagem é realizada por meio da integração de árvores de decisão e do otimizado gradiente descendente (Shahhosseini et al., 2021), estes algoritmos estão entre os mais utilizados da atualidade devido a sua capacidade de trabalhar com dados complexos.

Porém em países de grande extensão territorial como o Brasil, ocorre uma limitação nas bases de dados de solo. Sendo assim métodos de interpolação espacial podem ser usados para prever solos em locais não visitados, porém,

surgem problemas de acurácia quando se pretende prever atributos de solo em áreas com poucas observações amostrais (Ruhollah et al., 2022b).

Em tais situações, quando uma coleta maior e mais robusta de amostras de solo se torna inviável financeira e operacionalmente, uma abordagem de extrapolação pode ser aplicada para prever os solos na área alvo (loais sem observações) usando os solos da área de referência (área com observações adensadas) se os solos de duas áreas forem controlados principalmente por fatores semelhantes de formação de solo, bem como processos pedogênicos semelhantes (ten Caten et al., 2011b). Portanto, um modelo AM desenvolvido para uma área de referência, usando abordagens de mapeamento digital do solo (MDS), pode ser extrapolado para prever as características do solo da área alvo (Cambule et al., 2013; Afshar et al., 2018).

Nesse contexto, partindo da hipótese de que os algoritmos RF e XGBoost são capazes de estimar e extrapolar atributos de solo, a partir de dados provenientes do sensor ERD. Objetivou-se (i) estimar atributos mineralógicos do solo utilizando algoritmos de aprendizagem de máquina (AM), com base nas curvas espectrais geradas pela Espectroscopia de Reflectância Difusa (ERD) em uma Área de Referência (AR) menor e com pontos de coleta mais adensados, (ii) extrapolar os parâmetros da predição para uma Área Alvo (AA) maior, e com menos pontos de coleta no Planalto Ocidental Paulista e (iii) Entender qual dentre os algoritmos utilizados (RF e XGBoost) é mais eficiente na predição dos atributos de solo.

3.2 Material e métodos

3.2.1 Caracterização da área de estudo

Neste trabalho utilizamos dois bancos de dados, cedidos pelo grupo de

pesquisa em Caracterização do Solo para Fins de Manejo Específico (CSME) da FCAV-UNESP, Campus de Jaboticabal, São Paulo. Este estudo envolveu duas áreas (Área de Referência e Área Alvo) (Figura 1). A Área de Referência está localizada no município de Guataporá, no nordeste do estado de São Paulo, Brasil (-21.475394, -48.017398), com aproximadamente 870 hectares, na área é cultivada com cana-de-açúcar utilizando o sistema de colheita mecanizada há mais de 10 anos.

Na Área de Referência foram registrados a ocorrência de duas classes de solos e cinco sub-classes: Latossolo Vermelho eutroférico textura argilosa, Latossolo Vermelho distrófico textura média, Latossolo Vermelho distroférico textura argilosa, Latossolo Vermelho-Amarelo distrófico textura média, Neossolo Quartzarênico órtico distrófica textura arenosa. Para esta área foi definida uma grade amostral com aproximadamente 1 ponto a cada 2,5 hectares, totalizando 360 pontos, nos quais foram coletadas amostras de solo.

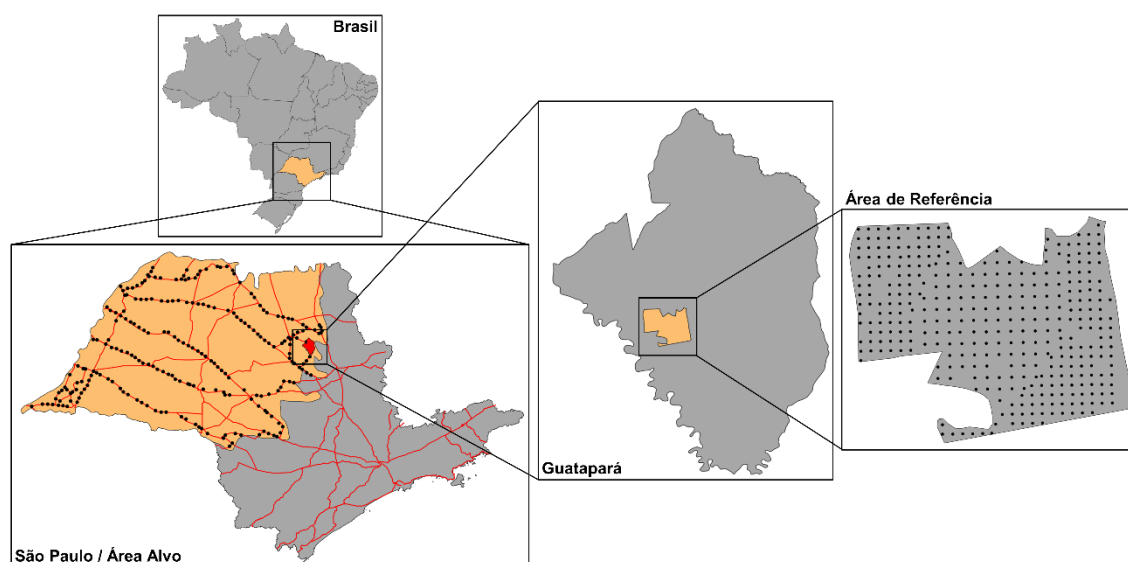


Figura 1. Ilustração da localização das áreas experimentais e dos pontos de coleta. No quadro “São Paulo/Área Alvo” as linhas vermelhas indicam a malha rodoviária utilizada como base para alocação dos pontos amostrais.

A Área Alvo está inserida no Planalto Ocidental Paulista (POP), que

corresponde a aproximadamente 48% de todo o Estado de São Paulo, totalizando 13 milhões de hectares. Esta área é responsável por cerca de 80% da produção nacional de etanol, açúcar e citros e vem sendo estudada por vários outros pesquisadores do grupo CSME (Bahia et al., 2015; Silvero et al., 2019; Silva et al., 2020; Baldo et al., 2021).

Na Área Alvo foram classificadas duas classes de solos, sendo elas Latossolo e Neossolo. Na área do POP, segundo o mapa de solos da Embrapa, seis classes de solo foram registradas, sendo elas: Argissolo Vermelho-Amarelo, Latossolo Vermelho, Latossolo férrico, Neossolo Litólico, Nitossolo Vermelho e Gleissolo Háptico (Figura 2) (Santos et al., 2018).

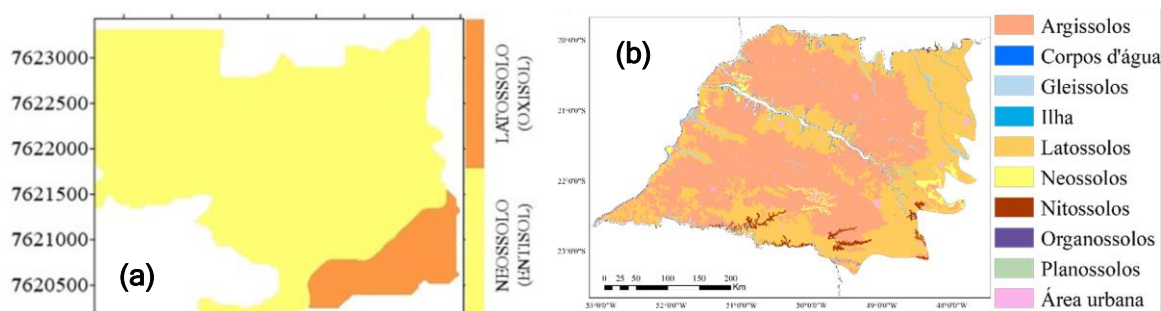


Figura 2. Mapa pedológico da Área de Referência (a), adaptado de Almeida et al. (2021). Mapa pedológico do Planalto Ocidental Paulista (b), Instituto Agronômico de Campinas (2017).

3.2.2 Plano amostral de coleta de pontos

Para a amostragem de solo da Área Alvo, levamos em consideração a base do arquivo de rodovias do Estado de São Paulo, fornecido pelo Departamento de Estradas de Rodagem (DER). A partir desse arquivo criamos um mapa por meio da ferramenta ET GeoWizards, no ArcView 9.3. Nessa área foram coletadas no total 231 amostras de solo na profundidade de 0-20 cm, realizando as coletas preferencialmente, em áreas com mínima interferência antrópica, em áreas preservadas adjacentes às áreas agrícolas, em espaçamento mínimo de 10

quilômetros (km), até o espaçamento máximo de 60 km.

3.2.3 Análises de laboratório – Atributos mineralógicos

Os atributos mineralógicos avaliados foram os teores de ferro cristalinos extraídos por ditionito-citrato (FED) (Mehra and Jackson, 1958) e os teores de ferro extraídos por amônio-ácido oxálico (FEO), relativos aos óxidos de ferro pedogenéticos de baixa cristalinidade, determinados conforme metodologia proposta por Camargo et al. (1986).

3.2.4 Espectroscopia de Reflectância Difusa (ERD)

A análise de Espectroscopia de Reflectância Difusa (ERD) foi realizada utilizando-se de um sensor Lambda 950 UV/VIS/NIR. As amostras de 0,5 g de terra fina seca ao ar (TFSA) foram acomodadas em porta amostra individual (3,0 × 0,5 cm). Efetuou-se em cada amostra uma varredura de reflectância na faixa do VISível (VIS), entre 250 e 900 nanômetros (nm) e do infravermelho próximo (NIR) entre 900 e 2500 nanômetros (nm), em um intervalo de 1 nm, gerando um banco de dados de 2.250 variáveis por amostra de solo, que corresponde a reflectância em cada comprimento de onda do espectro.

3.2.5 Aprendizagem de máquina

Em posse do banco de dados, a primeira análise realizada foi a estatística descritiva dos atributos do solo FED e FEO. Na sequência, para viabilizar o entendimento do comportamento dos dados foi desenvolvido um gráfico de distribuição. Na etapa seguinte iniciou-se o processo de predição dos atributos FED e FEO utilizando os algoritmos XGBoost e Florestas Aleatórias

O XGBoost é um algoritmo de aprendizagem de máquina que tem como principal característica a escalabilidade, alta precisão e ampla aplicação (Nguyen et al., 2022a). Ele consiste em uma variação dos algoritmos de árvore de decisão

com a técnica de aumento de gradiente, que atua ajustando as funções de perda, que são minimizadas pela estrutura de interações sequenciais do algoritmo (Hikouei et al., 2021), ou seja, o XGBoost usa *ensembles* de árvores sequencialmente otimizadas.

Este algoritmo tem capacidade de trabalhar com problemas de classificação e regressão, neste trabalho devido à natureza dos dados optamos pela regressão. O algoritmo, aplicado nos bancos de dados, foram divididos em 70% para treino e 30% para teste. Pensando que o algoritmo é impulsionado pelo adequado ajuste de hiperparâmetros, juntamente na análise foi utilizado um algoritmo de otimização, chamado *RandomizedSearchCV*. Foi estabelecido o total de 6 hiperparâmetros (*n_estimators*, *max_depth*, *reg_lambda*, *subsample*, *colsample_bytree*, *booster*) com 10 possibilidades cada, para todos os modelos foi selecionado o total de 50 árvores com 0.5 de taxa de aprendizagem.

O modelo de Regressão Florestas Aleatórias (FA) (Breiman, 2001) é uma técnica não paramétrica que foi desenvolvida com o intuito de melhorar a previsão dos modelos de Árvores de Classificação e Regressão, pois utiliza árvores de decisão totalmente crescidas e reduz o erro e a variação (Yan et al., 2020b). Tal algoritmo consiste na combinação de diversas árvores, que são geradas a partir de uma variável de entrada amostrada aleatoriamente, todas as árvores apresentam o mesmo tamanho. Já as subdivisões das árvores são realizadas com base em um subconjunto amostrado aleatoriamente do banco de dados total, o resultado do modelo de FA é composto pela média dos resultados de todas as árvores (Chagas et al., 2016b).

Após a otimização dos modelos de FA foram definidos os seguintes parâmetros: número de árvores (*n-estimados*) = 100; aleatoriedade *bootstrapping*

das amostras (*random-state*) = 50; profundidade máxima da árvore (*max_depth*) = 20; número mínimo de amostras necessárias para nó por folha (*min_samples_leaf*) = 5. Todas as análises de FA foram implementadas utilizando o algoritmo *RandomForestRegressor* do pacote Scikit-Learn (Pedregosa et al., 2011).

Para avaliar o desempenho dos modelos, foram estabelecidas 3 métricas: o coeficiente de determinação (R^2), erro médio absoluto (MAE) e a raiz quadrada do erro médio (RMSE).

3.2.6 Geoestatística

Para a determinação da existência da dependência espacial foram modelados semivariogramas simples, com base na teoria das variáveis regionalizadas e nos princípios da hipótese intrínseca (Isaaks; Srivastava, 1989). O variograma descreve a continuidade espacial das variáveis como função das distâncias entre duas localizações, sendo estimado por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \quad (1)$$

em que, $\hat{\gamma}(h)$ é a semivariância experimental para uma distância de separação h , $z(x_i)$ é o valor da propriedade no ponto i , e $N(h)$ é o número de pares de pontos separados pela distância h . Observada a existência da dependência espacial entre as amostras definida pelo semivariograma, foram estimados valores onde a variável não foi medida, fazendo uso da krigagem (Oliver; Webster, 2014). Para todas as análises foram utilizadas as linguagens de programação R e Python, de forma concomitante, no ambiente Jupyter notebook, seguindo o resumo gráfico apresentado na Figura 3.

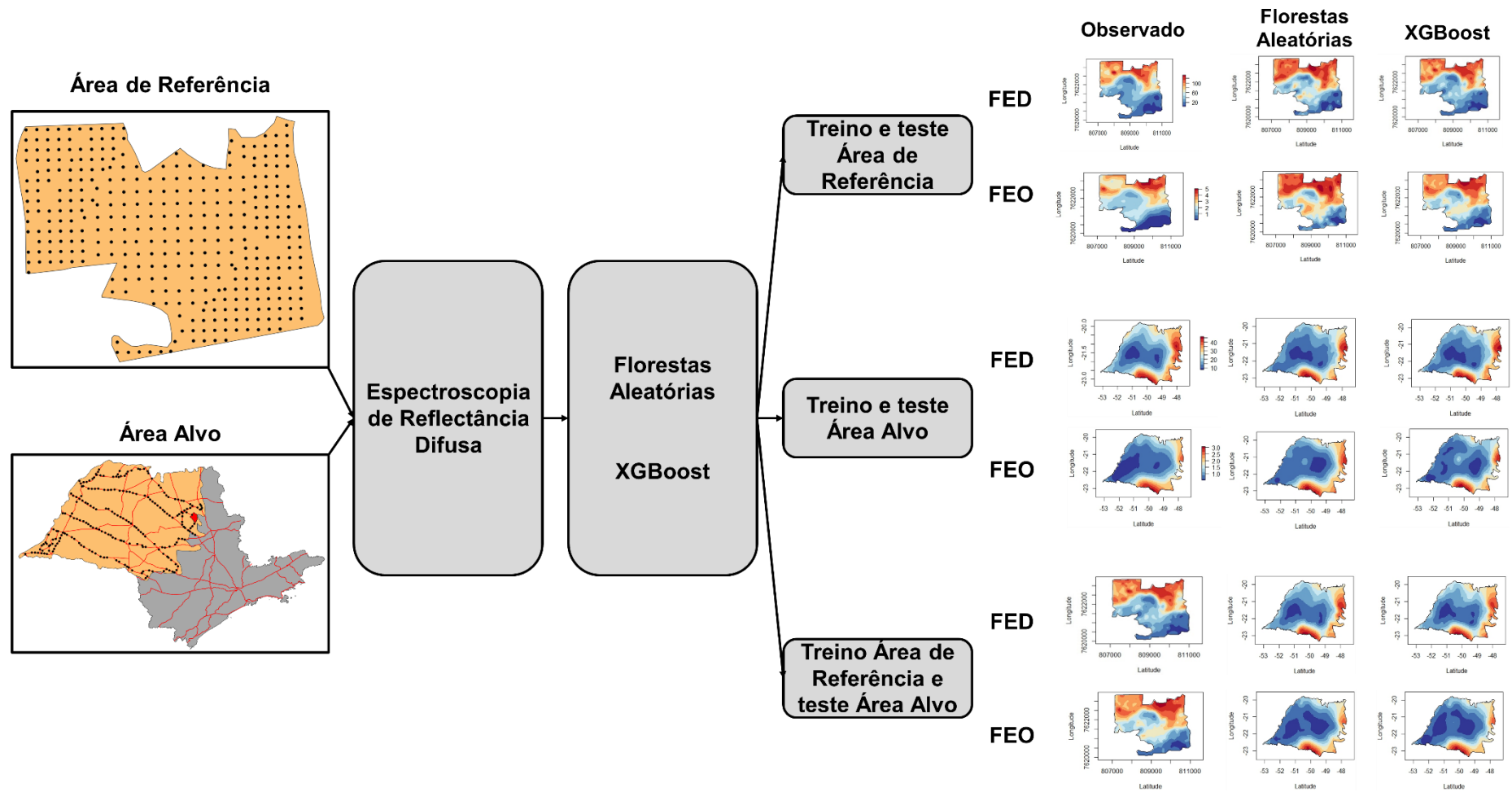


Figura 3. Fluxograma do passo a passo do processamento de dados até a obtenção dos mapas dos atributos analisados para a área de referência e área alvo.

3.3 Resultados e discussão

Para o melhor entendimento do comportamento da amplitude de variação dos dados foi realizada a estatística descritiva para o FED e FEO (Tabela 1). Podemos observar que de modo geral os dados de ambas as áreas apresentaram comportamento muito heterogêneo. Esta variabilidade pode ser associada às transições geológicas existentes em ambas as áreas, que apresentam classes de solos com características totalmente distintas como os Latossolos e Neossolos.

Tabela 1. Análise descritiva dos dados de FED e FEO das áreas de referência e alvo.

Área de Referência	FED	FEO
CV	71,24%	68,42%
Desvio Padrão	37,75	1,40
Máximo	177,71	7,44
Mediana	40,64	1,93
Média	53,00	2,05
Mínimo	1,99	0,04
Área Alvo	FED	FEO
CV	62,29%	100,25%
Desvio Padrão	14,32	1,17
Máximo	76,38	5,89
Mediana	17,69	0,70
Média	22,99	1,17
Mínimo	4,93	0,19

FED: Ferro Ditionito; FEO: Ferro Oxalato; CV: coeficiente de variação. N (número de observações) na área de referência: 360; na área alvo: 231.

Latossolos são caracterizados pela intemperização e lixiviação intensas e pela baixa atividade das argilas e capacidade de troca de cátions. Além disso, são caracterizados como solos profundos, de coloração relativamente homogênea e com distribuição uniforme de argila ao longo do perfil, elevada estabilidade de agregados e baixo conteúdo de silte em relação à argila (Ker, 1997), sendo os principais solos utilizados para produção agrícola no Brasil (Novais et al., 2007).

Neossolos quartzarênicos são solos originados principalmente a partir de rochas de arenitos ou sedimentos arenosos não consolidados. São caracterizados

pela baixa fertilidade e capacidade de retenção de água. São muito permeáveis, mal estruturados, com limitações ao uso agrícola e alta suscetibilidade à erosão (Ruiz et al., 2003; Souza et al., 2005).

Oliveira et al. (1998) relatam que a baixa relação de FEO/FED (ferro oxalato/ferro ditionito) no solo revelou predomínio de óxidos mais cristalinos como goethita e hematita, formados principalmente em solos situados em posições mais elevadas na paisagem, formando solos mais avermelhados e amarelo-avermelhados, indicando solos com melhor drenagem.

A primeira métrica apresentada na Tabela 1 é o coeficiente de variação (CV), observa-se que em ambas as áreas os dados apresentam comportamento heterogêneo. Para FED, a área de referência apresentou maior CV e desvio padrão indicando que os dados dessa área são mais dispersos do que o da área alvo. Corroborando com tal afirmação, ao observar os valores mínimos e máximos nota-se que na área de referência a amplitude do atributo é maior quando comparado com os valores da área alvo. Estes resultados podem ser associados à transição dos tipos de solo e geologias de origem desses solos, uma vez que pode haver grande variabilidade na quantidade de Fe na rocha de origem (Stranghoener et al., 2020).

Diferente do observado para FED, para FEO observa-se que na área alvo o valor de CV é consideravelmente maior do que o observado na área de referência, embora a área de referência apresente maior valor de desvio padrão e amplitude de variação entre os valores máximo e mínimo. De modo geral, a área de referência apresenta maior variabilidade dos atributos quando comparada a área alvo, tal fato pode ser atribuído ao grid de coleta de solo mais denso estabelecido para a área de referência, fazendo com que maior variabilidade espacial dos atributos seja

mapeada (Lark, 2000).

Para a primeira parte da análise de dados foram apresentadas as medidas de centro e de dispersão e na sequência para o entendimento mais aprofundado foram observadas as medidas de forma (curtose e assimetria) (Tabela 2). Para FED a curtose se mostrou negativa na área alvo, indicando um alto nível de achatamento da curva, ou seja, distribuição platicúrtica. Na área de referência nota que a curtose de FED ficou muito próxima de zero, indicando assim uma curva de padrão normal, ou seja, mesocúrtica.

Tabela 2. Valores de curtose a assimetria dos dados de FED e FEO das áreas de referência e alvo.

Área de Referência	FED	FEO
Curtose	-0,84	0,03
Assimetria	0,49	0,45
Área Alvo	FED	FEO
Curtose	0,03	3,76
Assimetria	1,04	2,05

FED: Ferro Ditionito; FEO: Ferro Oxalato.

Uma questão metodológica importante é a distribuição não normal dos dados, pois problemas com dados não normais acrescentam dificuldades adicionais quando comparados à predição de resultados com base em conjuntos de dados com distribuições normais (Esfandiarpour-Boroujeni et al., 2020; Gómez-Escalonilla et al., 2022).

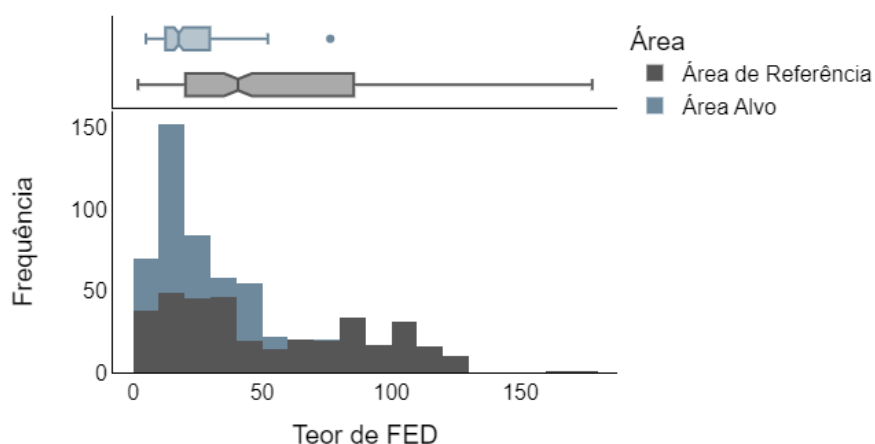
Já para FEO observa-se o padrão contrário ao observado para FED. Na área alvo a distribuição apresentou o padrão mesocúrtico e na área alvo apresentou achatamento baixo, ou seja, curva leptocúrtica, o que pode estar associado à distribuição bimodal dos dados de FED da área de referência, o que não ocorre para os dados de FED da área alvo nem para os dados de FEO de nenhuma das áreas. Em relação a assimetria, de modo geral em todos os atributos avaliados foi

observada presença de simetria positiva, sendo essa mais intensa na área alvo quando comparada a área de referência, o que pode ser associado a densidade amostral de ambas as áreas, como na área alvo foram coletadas menos amostras, o que ocasionou na formação de caldas devido a baixa representatividade.

Para apoiar o entendimento das medidas de centro, dispersão e forma, foi desenvolvido um gráfico de distribuição para cada atributo comparando os dados da área de referência e área alvo (Figura 4). Ambos os atributos apresentam na área de referência maior amplitude de dados do que na área alvo, tal diferença fica muito nítida nos box-plot apresentados.

Nota-se quem em ambos o gráfico a área de referência apresenta distribuição mais homogênea quando comparada a alvo, o que pode ser atribuído a maior densidade amostral desta área quando comparada com a área alvo, logo o menor espaçamento entre pontos da área de referência capta muito melhor a dependência espacial entre os atributos gerando uma distribuição mais homogênea.

(a)



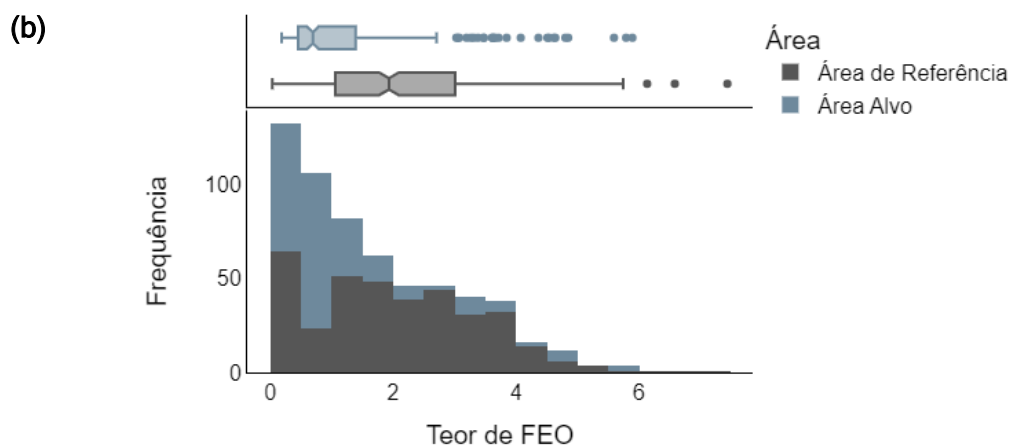


Figura 4. Distribuição dos dados de FED (a) e FEO (b) das áreas de referência e alvo. FED: Ferro Ditionito; FEO: Ferro Oxalato.

Tal fato fica ainda mais nítido no gráfico referente a FED (Figura 4a), no qual notam-se picos de frequência dos menores valores de FED na área alvo e na sequência reduções drásticas. Similar do padrão observado para FED, no gráfico de distribuição para FEO também se observa maior concentração do elemento em valores baixos na área alvo, porém com distribuição mais homogênea e parelha, nota-se que ela não reduz de forma abrupta como para FED. Além disso ambas as áreas apresentam um padrão semelhante de comportamento, sendo que a área de referência apresenta alguns valores máximos mais extremos do que a área alvo.

As áreas com maiores teores de FED são as argilosas, quem ficam mais nas bordas dos mapas da área alvo, com menor representatividade, já na área de referência, áreas com solos mais argilosos estão presentes em maior parte da área, o que pode ocasionar a distribuição mais bem distribuída desses elementos, uma vez que já se espera maior teor desses óxidos quanto maior o teor de argila dos solos (Camêlo et al., 2017; Silva et al., 2020b). A presença mais homogênea de FEO nas áreas pode ser associada as características únicas da formação de FEO, que não apresentando tão grande diferenciação em sua formação em função dos compartimentos geológicos, devido principalmente à grande variabilidade na presença de minerais amorfos ricos em Fe nesses solos (Lee et al., 2018;

Stranghoener et al., 2020).

De modo geral o comportamento de FEO apresentou maior semelhança de distribuição entre as áreas quando comparado com FED, tal afirmação corrobora diretamente com os box-plot apresentados nas Figuras 4a e 4b. Entendendo o comportamento dos atributos em ambas as áreas, foi possível observar que o intervalo de variação da área alvo está contido dentro do intervalo da área de referência, logo entendem-se que o padrão da área alvo existe dentro da base de dados da área de referência. Entendendo que as áreas apresentam padrões em comum, partiu-se para o desenvolvimento dos modelos de aprendizagem de máquina.

Na Tabela 3 estão as métricas utilizadas para avaliar a performance dos algoritmos. O modelo desenvolvido utilizando o algoritmo XGBoost apresentou desempenho superior ao algoritmo de Florestas Aleatórias, o que pode ser atribuído a robustez do algoritmo, pois usa em sua arquitetura sistema de gradiente descendente para atualizar os parâmetros do modelo ter a capacidade de reconhecer os recursos mais importantes (Chen; Guestrin, 2016).

A primeira métrica desenvolvida foi o coeficiente de determinação (R^2) que representa o percentual de variância dos dados que é explicado pelo modelo. Observando os resultados de FEO nota-se que de modo geral este apresentou R^2 inferiores aos dos modelos desenvolvidos para FED.

Para FED e FEO os valores de R^2 do algoritmo XGBoost mostrou-se mais eficiente em estimar os valores observados. O modelo que melhor performou nessa métrica para ambos os atributos foi o da área de referência, tanto na etapa de treino como na de teste, além de ter apresentado os maiores valores de R^2 . Para FED, podemos observar menor amplitude de variação entre as etapas de treino e teste,

indicando menor probabilidade de ocorrência de *overfitting*.

Tabela 3. Métricas de avaliação dos modelos de aprendizagem de máquina.

Modelo		Métrica	Florestas Aleatórias		XGBoost	
			FED	FEO	FED	FEO
Área de Referência	Treino	R ²	82,53%	80,84%	94,56%	81,41%
		MAE	10,39	0,41	5,8	0,44
		RMSE	15,64	0,61	8,83	0,62
	Teste	R ²	76,33%	57,88%	85,30%	54,67%
		MAE	12,98	0,63	9,82	0,42
		RMSE	18,47	0,89	14,24	0,83
Área Alvo	Treino	R ²	71,44%	61,16%	90,01%	72,68%
		MAE	4,71	0,4	2,85	0,34
		RMSE	7,61	0,68	4,62	0,59
	Teste	R ²	56,71%	44,05%	79,44%	67,63%
		MAE	6,49	0,53	4,62	0,43
		RMSE	9,42	0,92	6,16	0,68
Modelo de Extrapolação	Treino	R ²	86,18%	77,72%	93,50%	94,79%
		MAE	8,82	0,44	6,21	0,47
		RMSE	14,01	0,65	9,6	0,62
	Teste	R ²	72,05%	62,14%	88,36%	81,18%
		MAE	4,47	0,4	3,29	0,55
		RMSE	7,55	0,71	4,87	0,63

FED: Ferro Ditionito; FEO: Ferro Oxalato; R²: coeficiente de determinação; MAE: erro absoluto médio; RMSE: raiz quadrática média.

O *overfitting* é definido como o problema que ocorre quando um modelo foi treinado com muita dificuldade para ajustar os dados de treinamento, resultando em um modelo que aprendeu a reproduzir ruídos e peculiaridades nos dados de treinamento, em vez de encontrar uma regra preditiva geral (Dietterich, 1995). Caso ocorra o *overfitting* a precisão dos modelos desenvolvidos pode decair significativamente em um banco de dados de teste (Zanotti et al., 2019).

O modelo de pior performance tanto para FED como para FEO foi o de Florestas Aleatórias da área alvo. Além de apresentar elevada variação de valores entre as etapas de treino e teste, os resultados foram significativamente inferiores quando comparados com os demais modelos. Este resultado pode ser atribuído à

baixa densidade amostral existente na base de dados da área alvo. Além disso, o algoritmo de Florestas Aleatórias desenvolve diversas árvores de regressão e por fim realiza uma média dos resultados de todas.

Como a média é muito afetada pela presença de *outliers* na base de dados, a performance do algoritmo pode ter sido afetada pela presença dos *outliers*, uma vez que o desempenho de algoritmos preditivos pode ser severamente afetado pela presença de *outliers* nos dados experimentais, o que torna o algoritmo não robusto (Wang et al., 1996).

Para FED os menores valores de MAE foram observados no modelo desenvolvido usando o algoritmo XGBoost para a área de referência. Este resultado pode ser atribuído à robustez da base de dados da área de referência, que por consequência submete o algoritmo a maior quantidade de observações dos padrões da área e aperfeiçoamento da aprendizagem do modelo. Isso ocorre porque por diminuir a densidade amostral dos dados a serem previstos reduz naturalmente o número de exemplos em cada ambiente para os algoritmos aprenderem e, quanto mais limitado for o número de pontos em cada ambiente, mais difícil será ajustar a capacidade preditiva dos algoritmos (Esfandiarpour-Boroujeni et al., 2020; Gómez-Escalonilla et al., 2022).

Já para a métrica RMSE, o modelo que apresentou os menores valores foi o da área alvo. Este resultado indica que esse modelo foi o menos impactado pela presença de *outliers* nas bases de dados, uma vez que por haver maior densidade amostral e maior quantidade de observações dentro de cada ambiente os algoritmos foram melhor treinados, o que faz com que a menor ocorrência de *outliers* seja menos prejudicial para os valores de RMSE (Chai; Draxler, 2014).

Seguindo com a análise de FED, o modelo que apresentou menor

performance nas métricas de MAE e RMSE foi o modelo de extrapolação utilizando o algoritmo de Florestas Aleatórias. Este resultado era esperado, partindo da premissa que o modelo foi treinado e testado em áreas diferentes, logo a margem de erro da predição seria maior, similar aos resultados apresentados por Angelini et al. (2020) e Ruhollah et al. (2022) que, trabalhando com modelos de predição e extrapolação de dados de solo obtiveram métricas inferiores para os modelos de extrapolação, em comparação com os modelos das áreas de referência, independentemente do algoritmo utilizado.

Para FEO, nota-se que as métricas MAE e RMSE apresentaram resultados melhores do que para FED, o que pode ser associado à distribuição mais homogênea dos dados como observado nas Figuras 4a e 4b. Os modelos desenvolvidos apresentaram métricas de erro bem similares entre si. O modelo de Florestas Aleatórias da área de referência apresentou uma pequena diferença em relação aos demais na métrica MAE.

Porém, nota-se que este modelo foi muito penalizado pelos *outliers*, apresentando um RMSE ligeiramente mais alto do que nos modelos das áreas de referência e alvo, desenvolvidos com o algoritmo XGBoost. Entretanto, os modelos que tiveram RMSE mais penalizados pelos *outliers* foram os de extrapolação, independentemente do algoritmo utilizado.

De modo geral, a métrica RMSE apresentou resultados superiores ao MAE nos modelos. Porém, nota-se maiores diferenças nos modelos de FED, o que pode ser atribuído à distribuição dos dados desse atributo, que é mais heterogênea do que a de FEO, logo, sendo os modelos desenvolvidos para FEO mais penalizados pela presença de *outliers* (Chai; Draxler, 2014).

Observando especificamente os modelos de extrapolação quando

comparados aos demais desenvolvidos, nota-se maiores valores das métricas de erro, mesmo apresentando valores satisfatórios de R^2 . Estes maiores valores de erro podem ser atribuídos às diferentes bases apresentadas nas etapas de treino e teste. Por mais que com base na distribuição dos dados das bases utilizadas os padrões existentes na área alvo também estão contidos na área de referência, estas são áreas distintas. Entretanto, mesmo com valores de erro mais elevados, é possível inferir que os resultados foram satisfatórios, principalmente no modelo que foi utilizado o algoritmo XGBoost.

O algoritmo de XGBoost tem sido empregado com sucesso para estimativa de diversos parâmetros ambientais como, cobertura vegetal (Maurya et al., 2022a), qualidade da água (Xu et al., 2022), modelos hidrológicos de bacias (Wang et al., 2022), meteorologia florestal (Ghafarian et al., 2022), água no solo (Karthikeyan; and Mishra, 2021), modelos de física do solo (Zhang et al., 2021), salinização do solo (Ma et al., 2021), erosão do solo e da água (Golkarian et al., 2023), carbono do solo (He et al., 2022; Nguyen et al., 2022b; Zhang et al., 2022), parâmetros produtivos de culturas agrícolas (Han et al., 2022), metais pesados no solo (Ye et al., 2023; D. Zhao et al., 2022a), entre outras aplicações, o que tem feito esse algoritmo despontar nos últimos anos, sendo na grande maioria das vezes superior quando comparado com quaisquer que sejam os algoritmos.

Para viabilizar o entendimento do comportamento dos dados para cada modelo preditivo foram desenvolvidas regressões comparando os dois algoritmos (Figura 5). Observando inicialmente os modelos desenvolvidos para FED (Figuras 5a, 5b e 5c) é possível notar maior espalhamento dos dados preditos pelo modelo de Florestas Aleatórias ao redor da reta. Estes resultados corroboram com o que foi citado anteriormente, que os modelos desenvolvidos utilizando o XGBoost

performaram melhor, quando comparados com modelos desenvolvidos com o algoritmo Florestas Aleatórias, independentemente do campo de trabalho, seja na medicina (Basu et al., 2022), na indústria (Kiangala; Wang, 2021) ou em atributos ligados à mineralogia do solo, a exemplo do presente estudo.

Outro ponto que é possível observar é a continuidade da distribuição dos dados de FED da área de referência. Na área alvo nota-se uma concentração de pontos na parte inferior da reta e algumas poucas amostras na parte superior da reta, ficando um *gap* entre as extremidades das retas.

Esta condição pode ser atribuída à baixa densidade amostral nas áreas de transição e é possível notar ainda que nesta região ocorre maior concentração de pontos dispersos, tanto na Figura 5b bem como na Figura 5c, indicando que o algoritmo não foi capaz de estimar corretamente estes pontos, logo, tornando-se mais errático nessas regiões.

Geralmente, a tendência é que os modelos de aprendizagem favoreçam as áreas com mais observações, pois há maior quantidade de exemplos de ocorrência, resultando em baixas taxas de assertividade para locais pouco amostrados, uma vez que o algoritmo supõe que as distribuições são relativamente equilibradas (Castro; Braga, 2011).

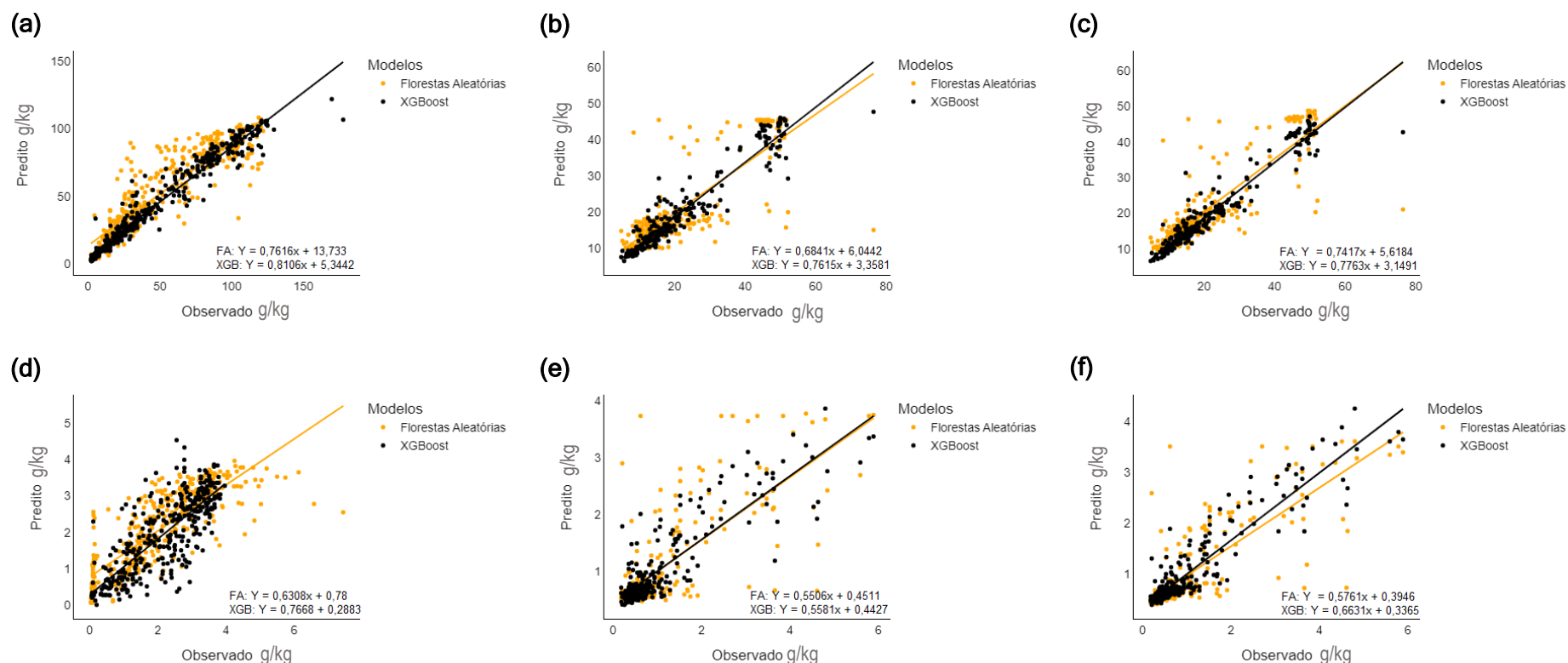


Figura 5. Análise de regressão dos dados preditos e observados de FED e FEO. (a) FED Área de referência; (b) FED Área Alvo; (c) FED extrapolado da Área de Referência para a Área Alvo; (d) FEO Área de referência; (e) FEO Área Alvo; (f) FEO extrapolado da Área de Referência para a Área Alvo; FED: Ferro Ditionito; FEO: Ferro Oxalato; FA: Florestas Aleatórias; XGB: XGBoost.

Observando as regressões desenvolvidas para FEO é possível observar que os dados estimados ficaram ainda mais espalhados, principalmente nas Figuras 5e e 5f, sendo, os modelos desenvolvidos com o algoritmo XGBoost foram mais eficientes nas estimativas.

Na regressão desenvolvida para área de referência é possível notar que os dados preditos pelo modelo de Florestas Aleatórias se distanciaram mais da reta, além de terem subestimado os dados, ocasionando o empilhamento e sobreposição de dados no ponto zero do eixo Y. Dentre todas as regressões é possível notar que a referente à extrapolação de FEO foi a que houve maior descolamento das duas retas, indicando a maior divergência entre as predições de FA e XGBoost para o FEO, possivelmente devido à menor amplitude de valores desse atributo.

Para avaliarmos espacialmente a performance dos modelos foram realizadas as krigagem dos dados observados e dos dados preditos por cada modelo, sendo as Figuras 6a, 6b e 6c – FED área de referência; Figuras 7a, 7b e 7c – FED área alvo; Figuras 8a, 8b e 8c – FED extrapolado; Figuras 9a, 9b e 9c – FEO área de referência; Figuras 10a, 10b e 10c – FEO área alvo; Figuras 11a, 11b e 11c – FEO extrapolado.

De modo geral tanto para FED como para FEO os semivariogramas dos dados observados e preditos apresentaram o mesmo comportamento esférico. Este modelo de ajuste é comumente associado a dados ligados à atributos do solo, pois a sua transição ocorre de forma mais rápida (Motomiya et al., 2006). O modelo esférico é comumente obtido quando se modela atributos do solo com variações mais abruptas (Isaaks and Srivastava, 1989; Vieira, 2000).

Dentre os semivariogramas desenvolvidos, os referentes à área alvo

apresentaram patamares e alcance mais bem definidos, quando comparados aos desenvolvidos para a área de referência. Isso pode ser associado ao tamanho da base de dados e à maior densidade amostral, pois a área de referência apresenta uma base de dados mais robusta o que ocasiona no aumento da amplitude de máximos e mínimos e por consequência no aumento da variabilidade. Oliver e Webster (2014) relatam que existem vários fatores que prejudicam a confiabilidade dos ajustes dos semivariogramas experimentais, sendo eles o tamanho da amostra, pois quanto mais dados você tiver, maior será a precisão, porém, mais difícil tende a ser o ajuste, uma vez que a amostragem capta melhor a variabilidade a curtas distâncias.

Nota-se que tanto os semivariogramas de FED como de FEO da área alvo e da extrapolação apresentaram parâmetros como efeito pepita (C_0), patamar e alcance muito semelhantes entre si, podendo inferir que os modelos apresentaram comportamento semelhantes entre si. Por fim, foram confeccionados o total de dezoito mapas e seis gráficos violinos comparando a distribuição dos dados observados e preditos por ambos os modelos.

O primeiro mapa refere-se ao atributo FED na área de referência (Figura 6). Assim como constatado nos semivariogramas é possível observar a ocorrência de mudanças abruptas nos valores de FED ao longo da área, o que pode ser atribuído à elevada densidade amostral da base de dados, o que traz mais detalhes às mudanças nos valores de FED a curtas distâncias.

De modo geral os mapas apresentaram semelhança entre si. Porém, isso fica ainda mais claro quando comparamos o mapa desenvolvido utilizando o algoritmo XGBoost (Figura 6c). Neste mapa é possível visualizar com precisão o padrão das transições geológicas, diferente do mapa confeccionado utilizando o

algoritmo de Florestas Aleatórias (Figura 6b).

Ao centro do mapa da Figura 6b é possível notar que o algoritmo de Floresta Aleatória superestima os valores de FED, assim como nas bordas da parte superior do mapa. Este resultado não é observado no mapa gerado com base no modelo do algoritmo XGBoost. O XGBoost, por ser um algoritmo que combina técnicas de otimização tem a vantagem de reduzir superestimação e subestimação de valores e a capacidade de minimizar *overfitting*, trabalhando bem com problemas de normalização de dados e sendo menos influenciado pela presença de conjuntos de dados desbalanceados (Maurya et al., 2022b; D. Zhao et al., 2022b).

No gráfico violino (Figura 6d) é possível notar no box-plot que as três bases de dados apresentaram comportamentos muito semelhantes. Porém, a que mais se distinguiu foi a desenvolvida com o modelo de Florestas Aleatórias. Nota-se um ligeiro achatamento nos extremos do contorno, quando comparado com a base de dados observada.

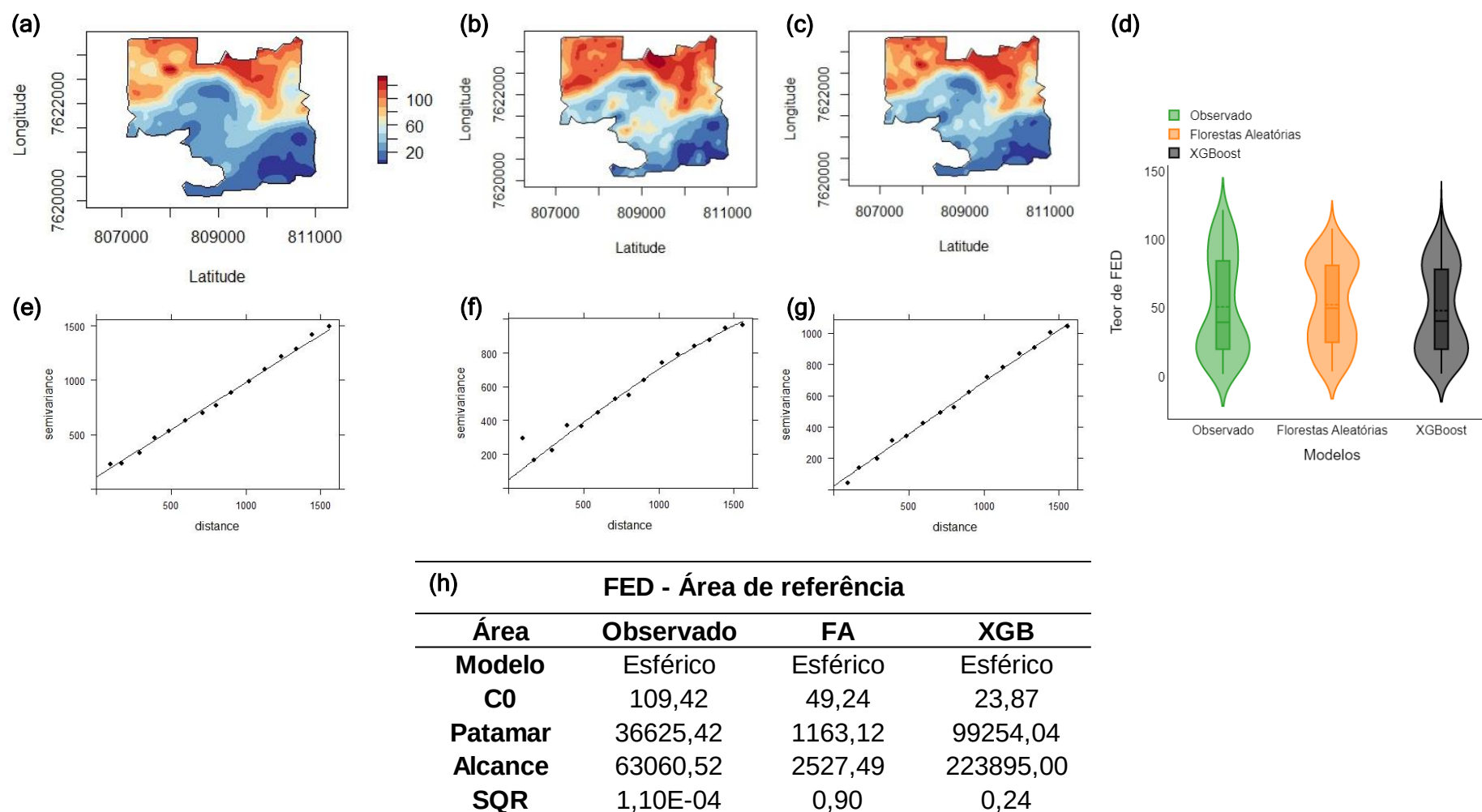


Figura 6. Mapas, semivariogramas e gráficos violinos referentes ao atributo FED no solo da Área de Referência: (a) dados observados; (b) dados previstos usando Florestas Aleatórias; (c) dados previstos usando XGBoost; (d) gráficos de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas de ajuste dos semivariogramas. FED: Ferro Ditionito.

Além disso, o conjunto de dados preditos pelo algoritmo de Florestas Aleatórias apresentou a média mais alta em relação aos conjuntos de dados observado e estimados pelo algoritmo XGBoost, nos outros dois gráficos violinos. O violino que representa o modelo gerado pelo algoritmo XGBoost apresentou distribuição muito semelhante aos constatado no observado.

Na sequência foram desenvolvidos os mapas dos teores de FED para a área alvo (Figura 7). Nota-se uma grande semelhança entre eles, porém, assim como nos mapas anteriores, a semelhança maior entre os mapas dos dados observados e o mapa dos dados gerados utilizando o algoritmo XGBoost, corroborando com o que foi observado na tabela de métricas (Tabela 3) e nas regressões (Figura 5).

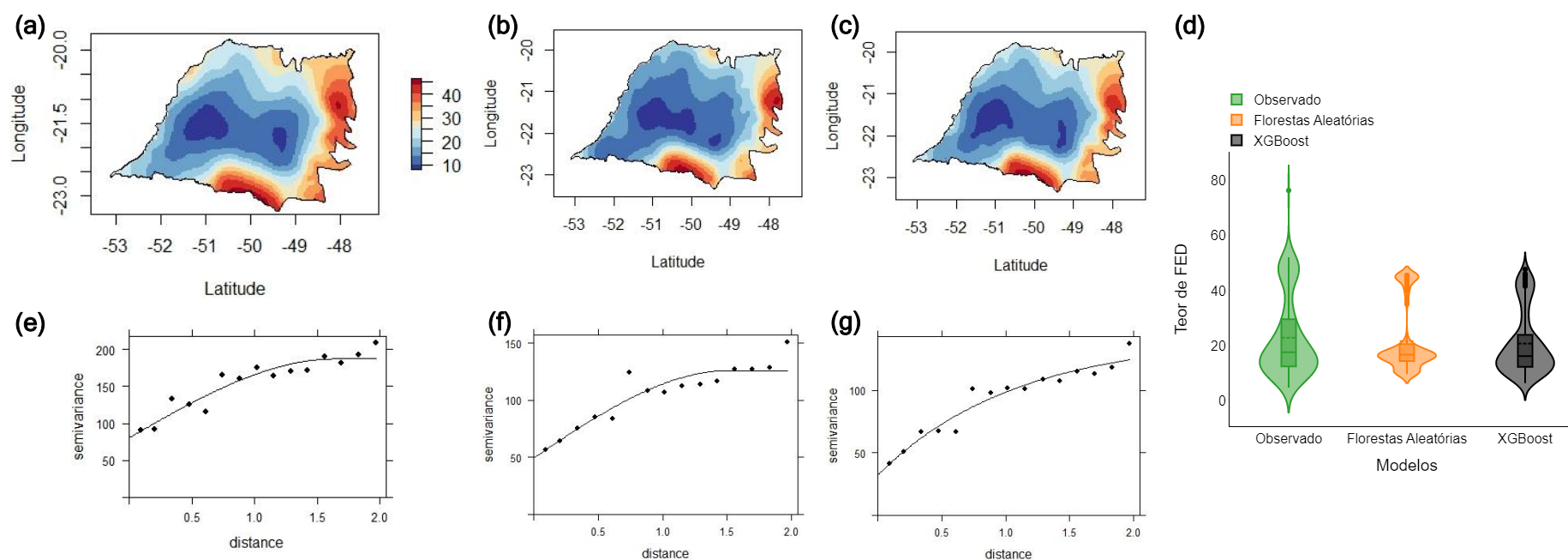
É possível observar que na Figura 7b desenvolvida utilizando Florestas Aleatórias o mapa perdeu uma de suas áreas de transição, o que ocasionou a perda de qualidade da informação em uma parte do mapa. Este efeito fica ainda mais claro ao analisar o gráfico violino, no qual pode-se observar um afunilamento dos dados. Este resultado pode ser explicado pelas regressões da Figura 5, onde é possível observar o espalhamento dos dados na zona de transição da base de dados. Logo, o modelo não foi capaz de estimar com precisão os dados dessas áreas de transição, ocasionando perda de informação da camada e o afunilamento do violino.

O oposto foi observado no mapa dos dados preditos pelo algoritmo XGBoost, que não foi tão assertivo em valores extremos como podemos notar no violino, mas que conseguiu lidar com a zona de transição de forma mais apropriada. Abdikan et al. (2022) comparando diversos algoritmos de aprendizagem de máquina (Regressão Linear Simples, Regressão Linear Múltipla, Rede Neural Artificial, XGBoost e Rede Neural Convolucional) obteve o resultado que o XGBoost foi mais

eficiente para a estimativa da altura da cultura do girassol em diferentes cenários.

Nos mapas criados com os modelos extrapolados (Figuras 8b e 8c) é possível observar que o algoritmo de Florestas Aleatórias lidou melhor com a transição dos dados, o que pode ser atribuído ao incremento da base de dados e por consequência ao aumento de padrões observados. Entretanto, mesmo sendo mais assertivo o mapa segue sendo errático em algumas regiões de transição.

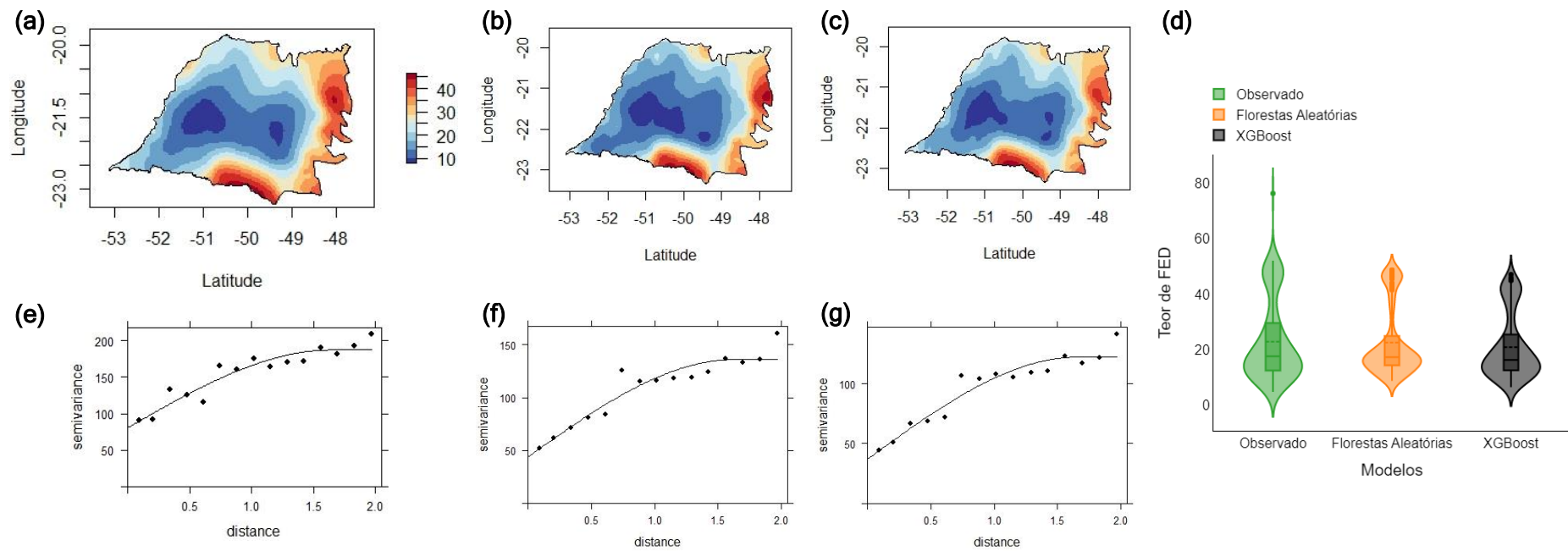
O mapa dos dados preditos pelo XGBoost (Figura 8c) mostrou-se muito semelhante ao mapa dos dados observados, errando apenas em alguns valores extremos, considerados como *outliers*. Nowak-Brzezinska e Horyn (2020) relatam a importância da detecção de *outliers*, para que os padrões nos dados possam ser melhor estudados e melhores modelos sejam desenvolvidos.



(h) FED - Área Alvo

Área	Observado	FA	XGB
Modelo	Esférico	Esférico	Esférico
C0	80,66	49,53	34,72
Patamar	106,66	76,54	82,24
Alcance	1,64	1,51	1,61
SQR	1,84	2,44	1,06

Figura 7. Mapas, semivariogramas e gráficos violinos referentes ao atributo FED no solo da Área Alvo: (a) dados observados; (b) dados preVISTos usando Florestas Aleatórias; (c) dados preVISTos usando XGBoost; (d) gráficos de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas de ajuste dos semivariogramas. FED: Ferro Ditionito.



(h) FED - Extrapolado

Área	Observado	FA	XGB
Modelo	Esférico	Esférico	Esférico
C0	80,66	43,88	37,01
Patamar	106,66	92,07	85,2
Alcance	1,64	1,62	1,66
SQR	1,84	1,27	1,27

3 Figura 8. Mapas, semivariogramas e gráficos violinos referentes ao atributo FED no solo extrapolado da Área de Referência para a
 4 Área Alvo: (a) dados observados; (b) dados preVISTos usando Florestas Aleatórias; (c) dados preVISTos usando XGBoost; (d) gráficos
 5 de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas
 6 de ajuste dos semivariogramas. FED: Ferro Ditionito.

Os mapas de FEO da área de referência, desenvolvidos a partir dos modelos de Florestas Aleatórias (Figura 9b) e XGBoost (Figuras 9c), apresentaram algumas nítidas diferenças em relação ao mapa elaborado com os dados observados (Figura 9a). Em ambos os mapas desenvolvidos com base nos modelos os valores foram superestimados em algumas regiões. Nota-se que isso foi mais evidente nos dados gerados utilizando o algoritmo de Florestas Aleatórias.

Corroborando com estes resultados é possível observar os gráficos violinos (Figura 9d), onde nota-se que o correspondente ao modelo de Floresta Aleatória apresentou maior média, quando comparado com a distribuição dos dados observados. Resultados similares foram apresentados por Takoutsing e Heuvelink (2022) que demonstraram a limitação do algoritmo de Florestas Aleatórias em prever dados extremos em bancos de dados desbalanceados.

A diferença do comportamento das bases de dados fica muito nítida ao observar o gráfico violino (Figura 9d), o contorno que corresponde a frequência dos três apresenta uma considerável diferença, além disso os modelos não conseguiram lidar com os *outliers* existentes nas bases de dados observados.

No mapa de FEO da área de referência (Figura 9a) é possível notar a presença de transições abruptas, o que justifica a indefinição do patamar e do alcance. Essas áreas de transição abruptas não foram satisfatoriamente preditas pelos modelos, resultando em mapas mais erráticos nas áreas de transição. Resultados similares foram observados por Liu et al. (2022) que utilizando de diferentes algoritmos para mapeamento digital de carbono do solo obtiveram modelos mais erráticos sempre nos valores mais extremos do banco de dados.

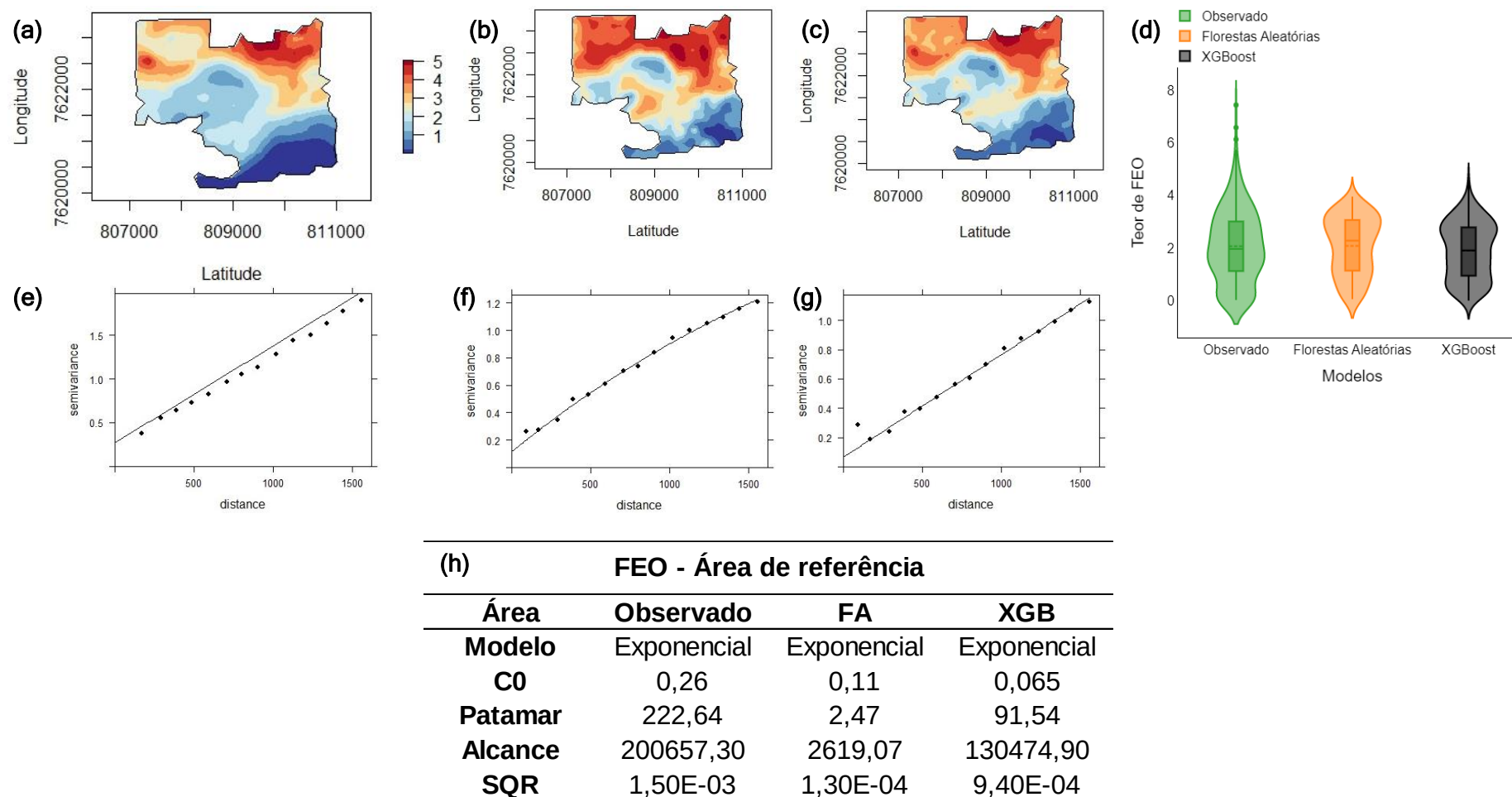


Figura 9. Mapas, semivariogramas e gráficos violinos referentes ao atributo FEO no solo da Área de Referência: (a) dados observados; (b) dados preVISTos usando Florestas Aleatórias; (c) dados preVISTos usando XGBoost; (d) gráficos de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas de ajuste dos semivariogramas. FEO: Ferro Oxalato.

Na sequência foram desenvolvidos os mapas referentes ao FEO da área alvo (Figura 10). Nota-se que no mapa desenvolvido com os dados preditos pelo modelo de Florestas Aleatórias (Figura 10b) o algoritmo superestimou os valores em determinados pontos, como é possível observar na região mais central e nas bordas à direita do mapa.

O mapa gerado utilizando os dados preditos pelo XGBoost (Figura 10c) também superestimou os valores de FEO na região mais central do mapa, porém, nas demais partes apresentou valores muito similares com os observados no mapa dos dados reais (Figura 10a).

Ao observar os gráficos violino do FEO da área alvo (Figura 10d) nota-se distribuição semelhante dos dados, sendo que apenas o violino referente aos dados gerados pelo modelo de Florestas Aleatórias foi penalizado em ambos os extremos e os dados gerados pelo modelo de XGBoost foi penalizado apenas no extremo superior.

Kaya et al. (2022) utilizando diferentes algoritmos para performar técnicas de mapeamento digital do solo observou que o algoritmo de Florestas Aleatórias foi ineficiente para prever atributos de solo com bases de dados desbalanceados, não sendo eficiente na estimativa de valores extremos.

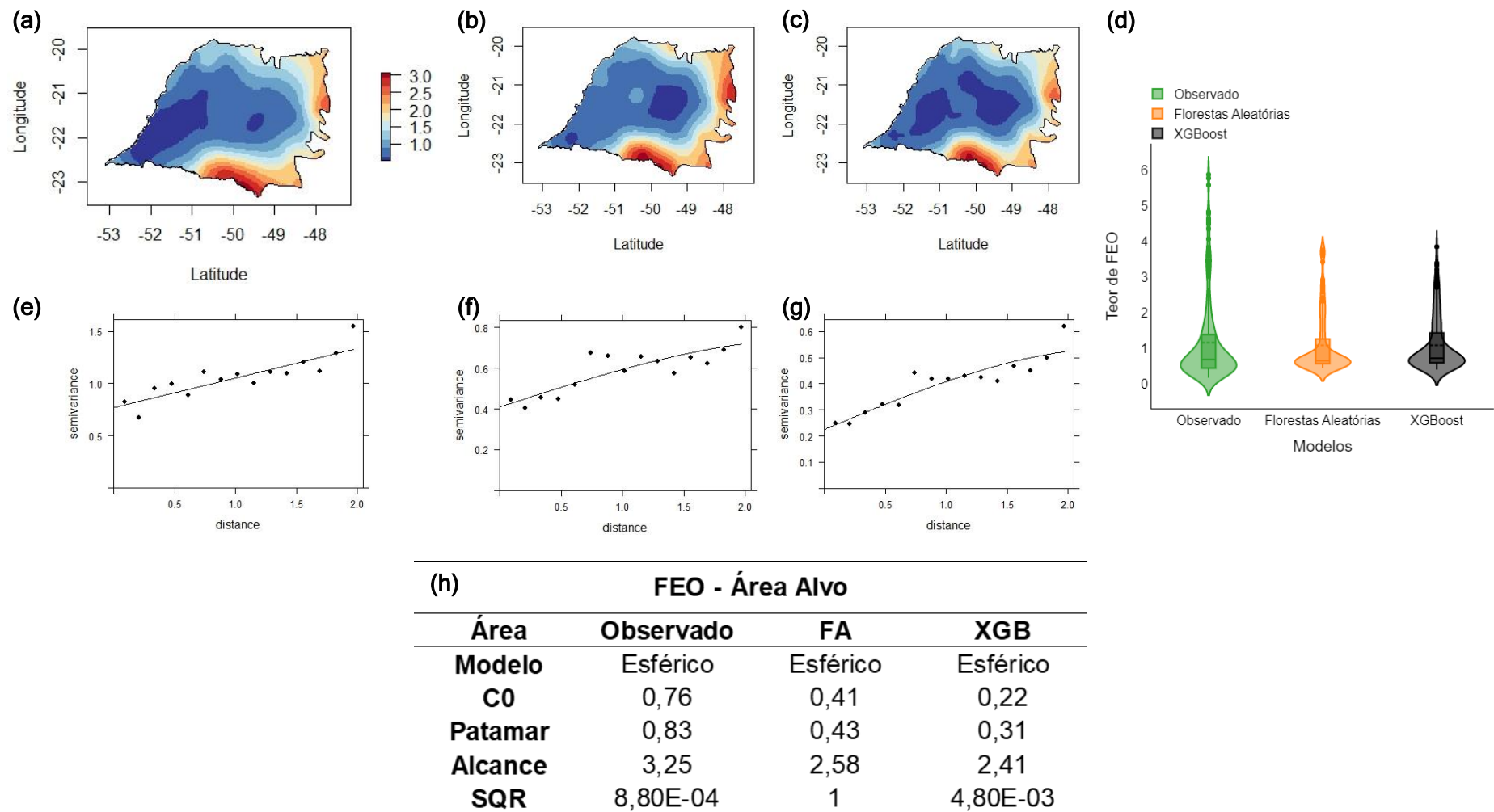


Figura 10. Mapas, semivariogramas e gráficos violinos referentes ao atributo FEO no solo da Área Alvo: (a) dados observados; (b) dados preVISTos usando Florestas Aleatórias; (c) dados preVISTos usando XGBoost; (d) gráficos de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas de ajuste dos semivariogramas. FEO: Ferro Oxalato.

Por fim, foram confeccionados os mapas referentes ao modelo de extrapolação de FEO (Figura 11). No mapa referente aos dados estimados pelo modelo de Florestas Aleatórias (Figura 11b), observa-se a diferença ao centro e nas bordas do mapa, locais onde os dados foram superestimados.

O mapa referente aos dados estimados pelo modelo XGBoost (Figura 11c) apresentou maior semelhança com o mapa dos dados observados (Figura 11a), apenas subestimando os valores de FEO na região central do mapa. Observa-se também maior assertividade do modelo de XGBoost para prever corretamente todas as transições no mapa.

Semelhante ao que foi citado anteriormente, os gráficos violino (Figura 11d) apresentaram a semelhança entre as bases de dados observada e preditas, sendo os modelos de predição penalizados pelos valores extremos (máximos e mínimos), uma vez que modelos de aprendizagem de máquina são negativamente influenciados por bancos de dados desbalanceados e pela presença de *outliers* no conjunto de dados (Wang et al., 1996; Nowak-Brzezinska; Horyn, 2020; Ebtehaj; Bonakdari, 2022).

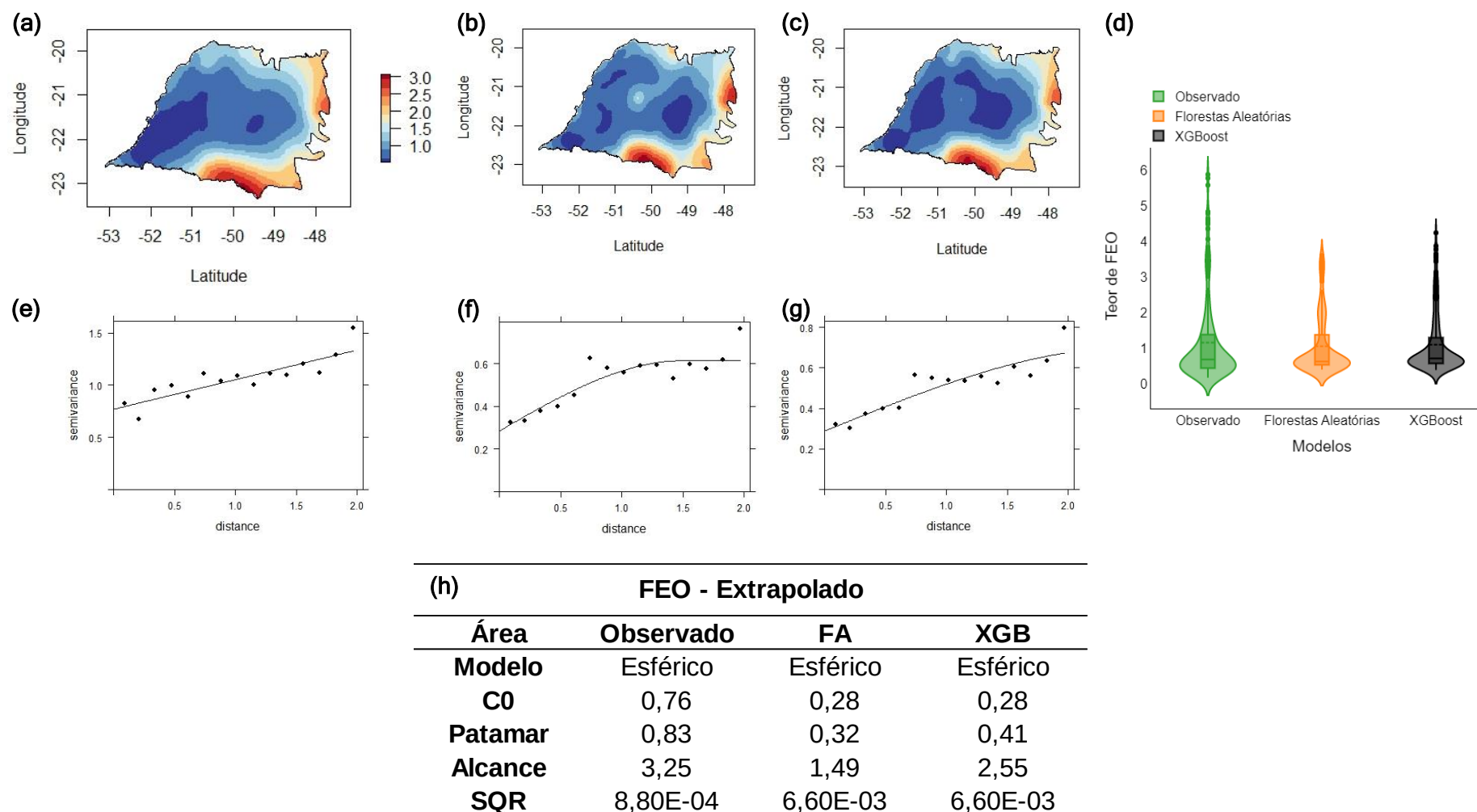


Figura 11. Mapas, semivariogramas e gráficos violinos referentes ao atributo FEO no solo extrapolado da Área de Referência para a Área Alvo: (a) dados observados; (b) dados preVISTos usando Florestas Aleatórias; (c) dados preVISTos usando XGBoost; (d) gráficos de violino; (e) semivariograma dados observados; (f) semivariograma Florestas Aleatórias; (g) semivariograma XGBoost; (h) métricas de ajuste dos semivariogramas. FEO: Ferro Oxalato.

3.4 Conclusões

O algoritmo de XGBoost mostrou-se mais eficiente que o algoritmo de Florestas Aleatórias para predizer e extrapolar os parâmetros de predição de FED e FEO do solo da área de referência para a área alvo.

O processo de extrapolação de parâmetros para predição de valores de FED e FEO é possível e assertivo, desde que haja uma base de dados robusta e que as áreas de referência e alvo tenham amplitudes de valores semelhantes.

O uso do sensor ERD, associados à modelos de aprendizagem de máquina é eficiente na predição de atributos mineralógicos do solo, sendo essa associação de técnicas de grande potencial para a ciência do solo.

A presença de *outliers* e banco de dados desbalanceados comprometem a qualidade dos modelos de aprendizagem de máquina.

3.5 Referências

- Abdikan, S., Sekertekin, A., Narin, O.G., Delen, A., Balik Sanli, F., 2022. A comparative analysis of SLR, MLR, ANN, XGBoost and CNN for crop height estimation of sunflower using Sentinel-1 and Sentinel-2. *Advances in Space Research*. <https://doi.org/10.1016/J.ASR.2022.11.046>
- Afshar, F.A., Ayoubi, S., Jafari, A., 2018a. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Afshar, F.A., Ayoubi, S., Jafari, A., 2018b. The extrapolation of soil great groups using multinomial logistic regression at regional scale in arid regions of Iran. *Geoderma* 315, 36–48. <https://doi.org/10.1016/J.GEODERMA.2017.11.030>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R. de S., Fernandes, K., Marques Júnior, J., 2021. Machine learning in the prediction of sugarcane production environments. *Comput Electron Agric* 190. <https://doi.org/10.1016/j.compag.2021.106452>
- Almeida, G.M. de, Pereira, G.T., Bahia, A.S.R.D.S., Júnior, J.M., 2020. Aprendizagem de máquina na predição de ambientes de produção de cana-de-açúcar. *Comput Electron Agric*.
- Andrade, R., Mancini, M., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion and auxiliary information for tropical soil property prediction: Soil texture. *Geoderma* 422, 115936. <https://doi.org/10.1016/J.GEODERMA.2022.115936>
- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020a. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>

- Angelini, M.E., Kempen, B., Heuvelink, G.B.M., Temme, A.J.A.M., Ransom, M.D., 2020b. Extrapolation of a structural equation model for digital soil mapping. *Geoderma* 367, 114226. <https://doi.org/10.1016/J.GEODERMA.2020.114226>
- Anzanello, M.J., Ortiz, R.S., Limbergerb, R.P., Mayorga, P., 2013. A multivariate-based wavenumber selection method for classifying medicines into authentic or counterfeit classes. *J Pharm Biomed Anal* 83, 209–214. <https://doi.org/10.1016/J.JPBA.2013.05.004>
- Bachmann, C.M., Philpot, W., Abelev, A., Korwan, D., 2014. Phase angle dependence of sand density observable in hyperspectral reflectance. *Remote Sens Environ* 150, 53–65. <https://doi.org/10.1016/J.RSE.2014.03.024>
- Bahia, A.S.R. de S., Marques, J., La Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017a. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R. de S., Marques, J., la Scala, N., Pellegrino Cerri, C.E., Camargo, L.A., 2017b. Prediction and Mapping of Soil Attributes using Diffuse Reflectance Spectroscopy and Magnetic Susceptibility. *Soil Science Society of America Journal* 81, 1450–1462. <https://doi.org/10.2136/sssaj2017.06.0206>
- Bahia, A.S.R.D.S., Marques, J., Siqueira, D.S., 2015. Procedures using diffuse reflectance spectroscopy for estimating hematite and goethite in Oxisols of São Paulo, Brazil. *Geoderma Regional* 5, 150–156. <https://doi.org/10.1016/j.geodrs.2015.04.006>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021a. Soil mineralogical attributes estimated by color as accessed by proximal sensors and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Baldo, D., Marques, J., Fernandes, K., de Almeida, G.M., Siqueira, D.S., 2021b. Soil mineralogical attributes estimated by color as accessed by proximal sensors and machine learning. *Soil Science Society of America Journal* 85, 2185–2200. <https://doi.org/10.1002/saj2.20309>
- Balsam, W.L., Otto-Bliesner, B.L., Deaton, B.C., 1995. Modern and Last Glacial Maximum eolian sedimentation patterns in the Atlantic Ocean interpreted from sediment iron oxide content. *Paleoceanography* 10, 493–507. <https://doi.org/10.1029/95PA00421>
- Baseca, C.C., Sendra, S., Lloret, J., Tomas, J., 2019. A smart decision system for digital farming. *Agronomy* 9. <https://doi.org/10.3390/agronomy9050216>
- Basu, D., Sinha, R., Sahu, S., Malla, J., Chakravorty, N., Ghosal, P.S., 2022. Identification of severity and passive measurement of oxidative stress biomarkers for β -thalassemia patients: K-means, random forest, XGBoost, decision tree, neural network based novel framework. *Advances in Redox Research* 5, 100034. <https://doi.org/10.1016/J.ARRES.2022.100034>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019a. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582–592. <https://doi.org/10.1016/j.compag.2019.05.008>
- Bekker, G.F.H. van G., Addison, M., Addison, P., van Niekerk, A., 2019b. Using machine learning to identify the geographical drivers of *Ceratitis capitata* trap catch in an agricultural landscape. *Comput Electron Agric* 162, 582–592. <https://doi.org/10.1016/j.compag.2019.05.008>
- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015a. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1–14. <https://doi.org/10.1016/j.aca.2015.04.042>

- Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Busto, O., 2015b. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Anal Chim Acta* 891, 1–14. <https://doi.org/10.1016/j.aca.2015.04.042>
- Bowden, C., Foster, T., Parkes, B., 2023. Identifying links between monsoon variability and rice production in India through machine learning. *Sci Rep* 13, 1–12. <https://doi.org/10.1038/s41598-023-27752-8>
- Breiman, L., 2001. Random Forest. *Mach Learn* 45, 5–32. <https://doi.org/10.1017/CBO9781107415324.004>
- Breiman, Leo, 2001. Random forests. *Mach Learn* 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bui, D.T., Tsangaratos, P., Nguyen, V.T., Liem, N. Van, Trinh, P.T., 2020. Comparing the prediction performance of a Deep Learning Neural Network model with conventional machine learning models in landslide susceptibility assessment. *Catena (Amst)* 188. <https://doi.org/10.1016/j.catena.2019.104426>
- CAMARGO, O.A. de, MONIZ, A.C., JORGE, J.A.J., VALADARES, J.M.A.S., 1986. Métodos de Análise Química, Mineralógica e Física de Solos do Instituto Agrônomo de Campinas, Boletim Técnico, 106. ed.
- Camargo, L.A., Marques Júnior, J., Pereira, G.T., Horvat, R.A., 2009. Variabilidade espacial de atributos mineralógicos de um Latossolo sob diferentes formas do relevo. II - Correlação espacial entre mineralogia e agregados. *Rev Bras Cienc Solo* 32, 2279–2288. <https://doi.org/10.1590/s0100-06832008000600007>
- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013a. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341–353. <https://doi.org/10.1016/J.GEODERMA.2012.08.020>
- Cambule, A.H., Rossiter, D.G., Stoorvogel, J.J., 2013b. A methodology for digital soil mapping in poorly-accessible areas. *Geoderma* 192, 341–353. <https://doi.org/10.1016/J.GEODERMA.2012.08.020>
- Camêlo, D. de L., Ker, J.C., Fontes, M.P.F., Corrêa, M.M., da Costa, A.C.S., Melo, V.F., 2017. Pedogenic Iron Oxides in Iron-Rich Oxisols Developed from Mafic Rocks. *Rev Bras Cienc Solo* 41. <https://doi.org/10.1590/18069657RBCS20160379>
- Canizo, B. V., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514–522. <https://doi.org/10.1016/j.compag.2019.04.043>
- Canizo, B. v., Escudero, L.B., Pellerano, R.G., Wuilloud, R.G., 2019. Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes. *Comput Electron Agric* 162, 514–522. <https://doi.org/10.1016/j.compag.2019.04.043>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016a. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232–240. <https://doi.org/10.1016/j.catena.2016.01.001>
- Chagas, C. da S., de Carvalho Junior, W., Bhering, S.B., Calderano Filho, B., 2016b. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena (Amst)* 139, 232–240. <https://doi.org/10.1016/J.CATENA.2016.01.001>
- Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 7, 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>
- Chen, T., Guestrin, C., 2016a. XGBoost: A scalable tree boosting system, in: *Proceedings of the ACM SIGKDD International Conference on Knowledge*

- Discovery and Data Mining. Association for Computing Machinery, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, T., Guestrin, C., 2016b. XGBoost: A scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August-2016, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Coblinski, J.A., Giasson, É., Demattê, J.A.M., Dotto, A.C., Costa, J.J.F., Vašát, R., 2020. Prediction of soil texture classes through different wavelength regions of reflectance spectroscopy at various soil depths. *Catena (Amst)* 189, 104485. <https://doi.org/10.1016/j.catena.2020.104485>
- Cornell, R.M., Schwertmann, U., 2006. The Iron Oxides: Structure, Properties, Reactions, Occurrences and Uses 707.
- da Costa, A.C.S., Junior, I.G. de S., Canton, L.C., Gil, L.G., Figueiredo, R., 2020. Contribution of the chemical and mineralogical properties of sandy-loam tropical soils to the cation exchange capacity. *Rev Bras Cienc Solo* 44, 1–18. <https://doi.org/10.36783/18069657rbcs20200019>
- de Castro, C.L., Braga, A.P., 2011. Aprendizado supervisionado com conjuntos de dados desbalanceados. *Sba: Controle & Automação Sociedade Brasileira de Automatica* 22, 441–466. <https://doi.org/10.1590/S0103-17592011000500002>
- de Motomiya, A.V.A., Corá, J.E., Pereira, G.T., 2006. Uso da krigagem indicatriz na avaliação de indicadores de fertilidade do solo. *Rev Bras Cienc Solo* 30, 485–496. <https://doi.org/10.1590/S0100-06832006000300010>
- Demattê, José Alexandre M, Dotto, A.C., Bedin, L.G., Sayão, V.M., Barros, A., 2019. Geoderma Soil analytical quality control by traditional and spectroscopy techniques : Constructing the future of a hybrid laboratory for low environmental impact. *Geoderma* 337, 111–121. <https://doi.org/10.1016/j.geoderma.2018.09.010>
- Demattê, José A.M., Dotto, A.C., Paiva, A.F.S., Sato, M. V., Dalmolin, R.S.D., de Araújo, M. do S.B., da Silva, E.B., Nanni, M.R., ten Caten, A., Noronha, N.C., Lacerda, M.P.C., de Araújo Filho, J.C., Rizzo, R., Bellinaso, H., Francelino, M.R., Schaefer, C.E.G.R., Vicente, L.E., dos Santos, U.J., de Sá Barretto Sampaio, E. V., Menezes, R.S.C., de Souza, J.J.L.L., Abrahão, W.A.P., Coelho, R.M., Grego, C.R., Lani, J.L., Fernandes, A.R., Gonçalves, D.A.M., Silva, S.H.G., de Menezes, M.D., Curi, N., Couto, E.G., dos Anjos, L.H.C., Ceddia, M.B., Pinheiro, É.F.M., Grunwald, S., Vasques, G.M., Marques Júnior, J., da Silva, A.J., Barreto, M.C. de V., Nóbrega, G.N., da Silva, M.Z., de Souza, S.F., Valladares, G.S., Viana, J.H.M., da Silva Terra, F., Horák-Terra, I., Fiorio, P.R., da Silva, R.C., Frade Júnior, E.F., Lima, R.H.C., Alba, J.M.F., de Souza Junior, V.S., Brefin, M.D.L.M.S., Ruivo, M.D.L.P., Ferreira, T.O., Brait, M.A., Caetano, N.R., Bringhenti, I., de Sousa Mendes, W., Safanelli, J.L., Guimarães, C.C.B., Poppiel, R.R., e Souza, A.B., Quesada, C.A., do Couto, H.T.Z., 2019. The Brazilian Soil Spectral Library (BSSL): A general view, application and challenges. *Geoderma* 354, 113793. <https://doi.org/10.1016/j.geoderma.2019.05.043>
- Dietterich, T., 1995. Overfitting and undercomputing in machine learning. *ACM Computing Surveys (CSUR)* 27, 326–327. <https://doi.org/10.1145/212094.212114>
- dos Santos, F.R., de Oliveira, J.F., Barbosa, G.M.C., Melquiades, F.L., 2021. Comparison between energy dispersive X-ray fluorescence spectral data and elemental data for soil attributes modelling. *Spectrochim Acta Part B At Spectrosc* 185, 106303. <https://doi.org/10.1016/J.SAB.2021.106303>
- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008a. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009–1016. <https://doi.org/10.1590/s0100-06832008000300010>

- Eberhardt, D.N., Vendrame, P.R.S., Becquer, T., De Fátima Guimarães, M., 2008b. Influência da granulometria e da mineralogia sobre a retenção do fósforo em latossolos sob pastagens no cerrado. *Rev Bras Cienc Solo* 32, 1009–1016. <https://doi.org/10.1590/s0100-06832008000300010>
- Ebtehaj, I., Bonakdari, H., 2022. A reliable hybrid outlier robust non-tuned rapid machine learning model for multi-step ahead flood forecasting in Quebec, Canada. *J Hydrol (Amst)* 614, 128592. <https://doi.org/10.1016/J.JHYDROL.2022.128592>
- EMBRAPA, 2017. Manual de metodos de análises, in: Manual de Métodos de Análise de Solo. Brasília, DF, p. 574.
- Esfandiarpour-Boroujeni, I., Shamsabadi, M.S., Shirani, H., Mosleh, Z., Bagheri Bodaghabadi, M., Salehi, M.H., 2020. Comparison of error and uncertainty of decision tree and learning vector quantization models for predicting soil classes in areas with low altitude variations. *Catena (Amst)* 191, 104581. <https://doi.org/10.1016/J.CATENA.2020.104581>
- Fan, S.S., Chang, F.H., Hsueh, H.T., Ko, T.H., 2016. Measurement of Total Free Iron in Soils by H₂S Chemisorption and Comparison with the Citrate Bicarbonate Dithionite Method. *J Anal Methods Chem* 2016. <https://doi.org/10.1155/2016/7213542>
- Faria, A.J.G. de, Silva, S.H.G., Andrade, R., Mancini, M., Melo, L.C.A., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022. Prediction of soil organic matter content by combining data from Nix Pro™ color sensor and portable X-ray fluorescence spectrometry in tropical soils. *Geoderma Regional* 28, e00461. <https://doi.org/10.1016/J.GEODRS.2021.E00461>
- Fernandes, L.A., Castro, A.B. de, Basilici, G., 2007. Seismites in continental sand sea deposits of the Late Cretaceous Caiuá Desert, Bauru Basin, Brasil. *Sediment Geol* 199, 61–64. <https://doi.org/10.1016/j.sedgeo.2005.12.030>
- Fernandes, R.B.A., Barrón, V., Torrent, J., Fontes, M.P.F., 2004. Quantificação de óxidos de ferro de Latossolos brasileiros por espectroscopia de refletância difusa. *Rev Bras Cienc Solo* 28, 245–257. <https://doi.org/10.1590/S0100-06832004000200003>
- Fink, J.R., Inda, A. v., Bavaresco, J., Barrón, V., Torrent, J., Bayer, C., 2016. Adsorption and desorption of phosphorus in subtropical soils as affected by management system and mineralogy. *Soil Tillage Res* 155, 62–68. <https://doi.org/10.1016/J.STILL.2015.07.017>
- Fonseca, J. da S., Campos, M.C.C., Brito Filho, E.G. de, Mantovanelli, B.C., Silva, L.S., de Lima, A.F.L., Da Cunha, J.M., Simões, E.L., dos Santos, L.A.C., 2021. Soil–landscape relationship in a sandstone-gneiss topolithosequence in the State of Amazonas, Brazil. *Environ Earth Sci* 80, 1–15. <https://doi.org/10.1007/s12665-021-10026-9>
- Ganaie, M.A., Tanveer, M., Suganthan, P.N., Snasel, V., 2022. Oblique and rotation double random forest. *Neural Networks* 153, 496–517. <https://doi.org/10.1016/J.NEUNET.2022.06.012>
- Ghafarian, F., Wieland, R., Lüttschwager, D., Nendel, C., 2022. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environmental Modelling & Software* 156, 105466. <https://doi.org/10.1016/J.ENVSOFT.2022.105466>
- Gholizadeh, A., Luboš, B., Saberioon, M., Vašát, R., 2013. Visible, Near-Infrared, and Mid-Infrared Spectroscopy Applications for Soil Assessment with Emphasis on Soil Organic Matter Content and Quality: State-of-the-Art and Key Issues. <http://dx.doi.org/10.1366/13-07288> 67, 1349–1362. <https://doi.org/10.1366/13-07288>

07288

- Gholizadeh, A., Saberioon, M., Pouladi, N., Ben-Dor, E., 2023. Quantification and depth distribution analysis of carbon to nitrogen ratio in forest soils using reflectance spectroscopy. *International Soil and Water Conservation Research* 11, 112–124. <https://doi.org/10.1016/J.ISWCR.2022.06.004>
- Goldshleger, N., Ben-Dor, E., Benyamini, Y., Agassi, M., 2004. SOIL REFLECTANCE AS A TOOL FOR ASSESSING PHYSICAL CRUST ARRANGEMENT OF FOUR TYPICAL SOILS IN ISRAEL. *Soil Sci* 169, 677–687. <https://doi.org/10.1097/01.ss.0000146024.61559.e2>
- Golkarian, A., Khosravi, K., Panahi, M., Clague, J.J., 2023. Spatial variability of soil water erosion: Comparing empirical and intelligent techniques. *Geoscience Frontiers* 14, 101456. <https://doi.org/10.1016/J.GSF.2022.101456>
- Gómez-Escalonilla, Diancoumba, O., Traoré, D.Y., Montero, E., Martín-Loeches, M., Martínez-Santos, P., 2022. Multiclass spatial predictions of borehole yield in southern Mali by means of machine learning classifiers. *J Hydrol Reg Stud* 44, 101245. <https://doi.org/10.1016/J.EJRH.2022.101245>
- Gray, J.M., Bishop, T.F.A., Wilford, J.R., 2016. Lithology and soil relationships for soil modelling and mapping. *Catena (Amst)* 147, 429–440. <https://doi.org/10.1016/j.catena.2016.07.045>
- Grunwald, S., Vasques, G.M., Rivero, R.G., 2015. Fusion of Soil and Remote Sensing Data to Model Soil Properties. *Advances in Agronomy* 131, 1–109. <https://doi.org/10.1016/BS.AGRON.2014.12.004>
- Hall, D.L., Llinas, J., 2016. An introduction to multi-sensor data fusion. *Sensors, Nanoscience, Biomedical Engineering, and Instruments* 85.
- Han, L., Yang, G., Yang, X., Song, X., Xu, B., Li, Z., Wu, Jintao, Yang, H., Wu, Jianwei, 2022. An explainable XGBoost model improved by SMOTE-ENN technique for maize lodging detection based on multi-source unmanned aerial vehicle images. *Comput Electron Agric* 194, 106804. <https://doi.org/10.1016/J.COMPAG.2022.106804>
- He, M., Tang, L., Li, C., Ren, J., Zhang, L., Li, X., 2022. Dynamics of soil organic carbon and nitrogen and their relations to hydrothermal variability in dryland. *J Environ Manage* 319, 115751. <https://doi.org/10.1016/J.JENVMAN.2022.115751>
- Heuvelink, G.B.M., Brus, D.J., Rossiter, D.G., Shi, Z., 2019. Editorial for pedometrics 2017 special issue. *Eur J Soil Sci* 70, 25–26. <https://doi.org/10.1111/ejss.12772>
- Hikouei, I.S., Kim, S.S., Mishra, D.R., 2021. Machine-learning classification of soil bulk density in salt marsh environments. *Sensors* 21. <https://doi.org/10.3390/s21134408>
- Holmgren, G.G.S., 1967a. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210–211. <https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Holmgren, G.G.S., 1967b. A Rapid Citrate-Dithionite Extractable Iron Procedure. *Soil Science Society of America Journal* 31, 210–211. <https://doi.org/10.2136/sssaj1967.03615995003100020020x>
- Hong, H., Ji, K., Hei, H., Wang, C., Liu, C., Zhao, L., Lanson, B., Zhao, C., Fang, Q., Algeo, T.J., 2023. Clay mineral evolution and formation of intermediate phases during pedogenesis on picrite basalt bedrock under temperate conditions (Yunnan, southwestern China). *Catena (Amst)* 220, 106677. <https://doi.org/10.1016/J.CATENA.2022.106677>
- Isaaks, E.H., Srivastava, R.M., 1989. *An Introduction to Applied Geostatistics, Computers & Geosciences*. Pergamon. [https://doi.org/10.1016/0098-3004\(91\)90055-I](https://doi.org/10.1016/0098-3004(91)90055-I)
- Jararweh, Y., Fatima, S., Jarrah, M., AlZu'bi, S., 2023. Smart and sustainable

- agriculture: Fundamentals, enabling technologies, and future directions. *Computers and Electrical Engineering* 110, 108799. <https://doi.org/10.1016/j.compeleceng.2023.108799>
- Jha, K., Doshi, A., Patel, P., Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. *Artificial Intelligence in Agriculture* 2, 1–12. <https://doi.org/10.1016/J.AIIA.2019.05.004>
- Kampf, N., Schwertmann, U., 1982. The 5-M-NaOH concentration treatment for iron oxides in soils. *Clays Clay Miner* 30, 401–408. <https://doi.org/10.1346/CCMN.1982.0300601/METRICS>
- Karthikeyan, L., Mishra, A.K., 2021. Multi-layer high-resolution soil moisture estimation using machine learning over the United States. *Remote Sens Environ* 266, 112706. <https://doi.org/10.1016/J.RSE.2021.112706>
- Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. *Inf Sci (N Y)* 585, 609–629. <https://doi.org/10.1016/J.INS.2021.11.036>
- Kaya, F., Başayığit, L., Keshavarzi, A., Francaviglia, R., 2022. Digital mapping for soil texture class prediction in northwestern Türkiye by different machine learning algorithms. *Geoderma Regional* 31, e00584. <https://doi.org/10.1016/J.GEODRS.2022.E00584>
- Ker, J.C., 1997. LATOSSOLOS DO BRASIL: UMA REVISÃO. *Geonomos* 5, 17–40. <https://doi.org/10.18285/GEONOMOS.V5I1.187>
- Ker, J.C., Curi, N., Schaefer, C.E., Torrado, P.V., 2012. *Pedologia: fundamentos. Viçosa.*
- Kiangala, S.K., Wang, Z., 2021. An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment. *Machine Learning with Applications* 4, 100024. <https://doi.org/10.1016/J.MLWA.2021.100024>
- Kragt, M.E., Pannell, D.J., Robertson, M.J., Thamo, T., 2012. Assessing costs of soil carbon sequestration by crop-livestock farmers in Western Australia. *Agric Syst* 112, 27–37. <https://doi.org/10.1016/j.agry.2012.06.005>
- Kuncheva, L.I., Matthews, C.E., Arnaiz-González, Á., Rodríguez, J.J., 2020. Feature Selection from High-Dimensional Data with Very Low Sample Size: A Cautionary Tale.
- Lark, R.M., 2000. Designing sampling grids from imprecise information on soil variability, an approach based on the fuzzy kriging variance. *Geoderma* 98, 35–59. [https://doi.org/10.1016/S0016-7061\(00\)00051-3](https://doi.org/10.1016/S0016-7061(00)00051-3)
- Lee, M.E., Jeon, E.K., Tsang, D.C.W., Baek, K., 2018. Simultaneous application of oxalic acid and dithionite for enhanced extraction of arsenic bound to amorphous and crystalline iron oxides. *J Hazard Mater* 354, 91–98. <https://doi.org/10.1016/J.JHAZMAT.2018.04.083>
- Liu, J. cheng, He, H. ping, Michalski, J., Cuadros, J., Yao, Y. zeng, Tan, W., Qin, X. rong, Li, S. ying, Wei, G. jian, 2021. Reflectance spectroscopy applied to clay mineralogy and alteration intensity of a thick basaltic weathering sequence in Hainan Island, South China. *Appl Clay Sci* 201, 105923. <https://doi.org/10.1016/J.CLAY.2020.105923>
- Liu, J., Yang, K., Tariq, A., Lu, L., Soufan, W., El Sabagh, A., 2023. Interaction of climate, topography and soil properties with cropland and cropping pattern using remote sensing data and machine learning methods. *Egyptian Journal of Remote Sensing and Space Science* 26, 415–426. <https://doi.org/10.1016/j.ejrs.2023.05.005>
- Liu, Q., He, L., Guo, L., Wang, M., Deng, D., Lv, P., Wang, R., Jia, Z., Hu, Z., Wu, G.,

- Shi, T., 2022. Digital mapping of soil organic carbon density using newly developed bare soil spectral indices and deep neural network. *Catena (Amst)* 219, 106603. <https://doi.org/10.1016/J.CATENA.2022.106603>
- Liu, X., Zhu, A.X., Yang, L., Pei, T., Liu, J., Zeng, C., Wang, D., 2020. A graded proportion method of training sample selection for updating conventional soil maps. *Geoderma* 357, 113939. <https://doi.org/10.1016/j.geoderma.2019.113939>
- Lorz, C., Fruhauf, M., Mailander, R., Phillips, J.D., Kleber, A., 2013. Influence of Cover Beds on Soils, Mid-Latitude Slope Deposits (Cover Beds).
- Lu, W.N., He, Y., Wang (王阳), Y., Ke, S., 2020. Behavior of calcium isotopes during continental subduction recorded in meta-basaltic rocks. *Geochim Cosmochim Acta* 278, 392–404. <https://doi.org/10.1016/J.GCA.2019.09.027>
- Ma, G., Ding, J., Han, L., Zhang, Z., Ran, S., 2021. Digital mapping of soil salinization based on Sentinel-1 and Sentinel-2 data combined with machine learning algorithms. *Regional Sustainability* 2, 177–188. <https://doi.org/10.1016/J.REGSUS.2021.06.001>
- Malone, B.P., Jha, S.K., Minasny, B., McBratney, A.B., 2016. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. *Geoderma* 262, 243–253. <https://doi.org/10.1016/J.GEODERMA.2015.08.037>
- Mancini, M., Andrade, R., Teixeira, A.F. dos S., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for Brazilian soil properties prediction: Exchangeable/available macronutrients, aluminum, and potential acidity. *Geoderma Regional* 30, e00573. <https://doi.org/10.1016/J.GEODRS.2022.E00573>
- Marques, J., Siqueira, D.S., Camargo, L.A., Teixeira, D.D.B., Barrón, V., Torrent, J., 2014. Magnetic susceptibility and diffuse reflectance spectroscopy to characterize the spatial variability of soil properties in a brazilian haplustalf. *Geoderma* 219–220, 63–71. <https://doi.org/10.1016/j.geoderma.2013.12.007>
- Maurya, A.K., Bhargava, N., Singh, D., 2022a. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795–1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- Maurya, A.K., Bhargava, N., Singh, D., 2022b. Efficient selection of SAR features using ML based algorithms for accurate FVC estimation. *Advances in Space Research* 70, 1795–1809. <https://doi.org/10.1016/J.ASR.2022.06.039>
- McBratney, A., Gruijter, J., Bryce, A., 2019. Pedometrics timeline. *Geoderma* 338, 568–575. <https://doi.org/10.1016/j.geoderma.2018.11.048>
- McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping, *Geoderma*. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4)
- McKeague, J.A., Day, J.H., 1966. DITHIONITE- AND OXALATE-EXTRACTABLE Fe AND Al AS AIDS IN DIFFERENTIATING VARIOUS CLASSES OF SOILS. *Can J Soil Sci* 46, 13–22. <https://doi.org/10.4141/cjss66-003>
- Mehra, O.P., Jackson, M.L., 1958. Iron Oxide Removal from Soils and Clays by a Dithionite-Citrate System Buffered with Sodium Bicarbonate. *Clays and Clay Minerals* 1958 7:1 7, 317–327. <https://doi.org/10.1346/CCMN.1958.0070122>
- Mello, F.A.O., Bellinaso, H., Mello, D.C., Safanelli, J.L., Mendes, W.D.S., Amorim, M.T.A., Gomez, A.M.R., Poppiel, R.R., Silvero, N.E.Q., Gholizadeh, A., Silva, S.H.G., Curi, N., Demattê, J.A.M., 2021. Soil parent material prediction through satellite multispectral analysis on a regional scale at the Western Paulista Plateau, Brazil. *Geoderma Regional* 26, e00412. <https://doi.org/10.1016/J.GEODRS.2021.E00412>
- Mendes, W. de S., Demattê, J.A.M., Bonfatti, B.R., Resende, M.E.B., Campos, L.R., Costa, A.C.S. da, 2021. A novel framework to estimate soil mineralogy using soil

- spectroscopy. *Applied Geochemistry* 127, 104909. <https://doi.org/10.1016/J.APGEOCHEM.2021.104909>
- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022a. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219. <https://doi.org/10.1016/j.still.2022.105346>
- Mendes, W. de S., Demattê, J.A.M., Minasny, B., Silvero, N.E.Q., Bonfatti, B.R., Safanelli, J.L., Rizzo, R., Costa, A.C.S. da, 2022b. Free iron oxide content in tropical soils predicted by integrative digital mapping. *Soil Tillage Res* 219, 105346. <https://doi.org/10.1016/J.STILL.2022.105346>
- Merl, T., Rasmussen, M.R., Koch, L.R., Søndergaard, J.V., Bust, F.F., Koren, K., 2022. Measuring soil pH at in situ like conditions using optical pH sensors (pH-optodes). *Soil Biol Biochem* 175, 108862. <https://doi.org/10.1016/J.SOILBIO.2022.108862>
- Minasny, B., Hartemink, A.E., 2011a. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52–62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., Hartemink, A.E., 2011b. Predicting soil properties in the tropics. *Earth Sci Rev* 106, 52–62. <https://doi.org/10.1016/j.earscirev.2011.01.005>
- Minasny, B., McBratney, A.B., 2016. Digital soil mapping: A brief history and some lessons. *Geoderma* 264, 301–311. <https://doi.org/10.1016/j.geoderma.2015.07.017>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019a. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194, 104308. <https://doi.org/10.1016/j.still.2019.104308>
- Moral, F.J., Rebollo, F.J., Campillo, C., Serrano, J.M., 2019b. Using an objective and probabilistic model to delineate homogeneous zones in hedgerow olive orchards. *Soil Tillage Res* 194. <https://doi.org/10.1016/j.still.2019.104308>
- Naimi, S., Ayoubi, S., di Raimo, L.A.D.L., Dematte, J.A.M., 2022. Quantification of some intrinsic soil properties using proximal sensing in arid lands: Application of Vis-NIR, MIR, and pXRF spectroscopy. *Geoderma Regional* 28, e00484. <https://doi.org/10.1016/J.GEODRS.2022.E00484>
- Nawar, S., Delbecque, N., Declercq, Y., Smedt, P. De, Finke, P., Verdoodt, A., Meirvenne, M. Van, Mouazen, A.M., 2019. Geoderma Can spectral analyses improve measurement of key soil fertility parameters with X-ray fluorescence spectrometry? *Geoderma* 350, 29–39. <https://doi.org/10.1016/j.geoderma.2019.05.002>
- Nawar, S., Richard, F., Kassim, A.M., Tekin, Y., Mouazen, A.M., 2022. Fusion of Gamma-rays and portable X-ray fluorescence spectral data to measure extractable potassium in soils. *Soil Tillage Res* 223, 105472. <https://doi.org/10.1016/J.STILL.2022.105472>
- Ng, W., Minasny, B., McBratney, A., 2020a. Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Mcbratney, A., 2020b. Science of the Total Environment Convolutional neural network for soil microplastic contamination screening using infrared spectroscopy. *Science of the Total Environment* 702, 134723. <https://doi.org/10.1016/j.scitotenv.2019.134723>
- Ng, W., Minasny, B., Montazerolghaem, M., Padarian, J., Ferguson, R., Bailey, S., McBratney, A.B., 2019. Convolutional neural network for simultaneous prediction of several soil properties using visible/near-infrared, mid-infrared, and their combined spectra. *Geoderma* 352, 251–267. <https://doi.org/10.1016/j.geoderma.2019.06.016>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang,

- N.B., Guo, W., Ngo, H.H., 2022a. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of the Total Environment* 804. <https://doi.org/10.1016/j.scitotenv.2021.150187>
- Nguyen, T.T., Pham, T.D., Nguyen, C.T., Delfos, J., Archibald, R., Dang, K.B., Hoang, N.B., Guo, W., Ngo, H.H., 2022b. A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. *Science of The Total Environment* 804, 150187. <https://doi.org/10.1016/J.SCITOTENV.2021.150187>
- Nocita, M., Stevens, A., van Wesemael, B., Aitkenhead, M., Bachmann, M., Barthès, B., Dor, E. Ben, Brown, D.J., Clairotte, M., Csorba, A., Dardenne, P., Demattê, J.A.M., Genot, V., Guerrero, C., Knadel, M., Montanarella, L., Noon, C., Ramirez-Lopez, L., Robertson, J., Sakai, H., Soriano-Disla, J.M., Shepherd, K.D., Stenberg, B., Towett, E.K., Vargas, R., Wetterlind, J., 2015. Soil Spectroscopy: An Alternative to Wet Chemistry for Soil Monitoring. *Advances in Agronomy* 132, 139–159. <https://doi.org/10.1016/bs.agron.2015.02.002>
- Norrish, K., Taylor, R.M., 1961. The isomorphous replacement of iron by aluminium in soil goethites. *Journal of Soil Science* 12, 294–306. <https://doi.org/10.1111/J.1365-2389.1961.TB00919.X>
- Novais, R.F., Alvarez V., V.H.;, Barros, N.F. de;, Fontes, R.L.F.;, Cantarutti, R.B.;, Neves, J.C.L., 2007. Fertilidade do solo. Viçosa.
- Nowak-Brzezinska, A., Horyn, C., 2020. Outliers in rules - the comparison of LOF, COF and KMEANS algorithms. *Procedia Comput Sci* 176, 1420–1429. <https://doi.org/10.1016/J.PROCS.2020.09.152>
- Oliveira, C. v., Ker, J.C., Fontes, L.E.F., Curi, N., Pinheiro, J.C., 1998. Química e mineralogia de solos derivados de rochas do Grupo Bambuí no norte de Minas Gerais. *Rev Bras Cienc Solo* 22, 583–593. <https://doi.org/10.1590/S0100-06831998000400003>
- Oliver, M.A., Webster, R., 2014a. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56–69. <https://doi.org/10.1016/J.CATENA.2013.09.006>
- Oliver, M.A., Webster, R., 2014b. A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena (Amst)* 113, 56–69. <https://doi.org/10.1016/j.catena.2013.09.006>
- Padarian, J., Minasny, B., McBratney, A.B., 2019. Using deep learning to predict soil properties from regional spectral data. *Geoderma Regional* 16. <https://doi.org/10.1016/j.geodrs.2018.e00198>
- Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., 2011. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research* 12, 2825–2830. <https://doi.org/10.5555/1953048.2078195>
- Pedron, F. de A., Lourenzi, C.R., Ceretta, C.A., Lorensi, J., Cancian, A., 2018. Clay mineralogy of subtropical soils under long-term organic fertilization in no-tillage systems. *Rev Bras Cienc Solo* 42. <https://doi.org/10.1590/18069657rbcs20170092>
- Ramaroson, V.H., Becquer, T., Sá, S.O., Razafimahatratra, H., Delarivière, J.L., Blavet, D., Vendrame, P.R.S., Rabeharisoa, L., Rakotondrazafy, A.F.M., 2018. Mineralogical analysis of ferrallitic soils in Madagascar using NIR spectroscopy. *Catena (Amst)* 168, 102–109. <https://doi.org/10.1016/j.catena.2017.07.016>
- Reda, R., Saffaj, T., Ilham, B., Saidi, O., Issam, K., Brahim, L., El Hadrami, E.M., 2019. A comparative study between a new method and other machine learning

- algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. *Chemometrics and Intelligent Laboratory Systems* 195. <https://doi.org/10.1016/j.chemolab.2019.103873>
- Rennert, T., 2019. Wet-chemical extractions to characterise pedogenic Al and Fe species-a critical review. *Soil Research* 57, 1–16. <https://doi.org/10.1071/SR18299>
- Rosin, N.A., Demattê, J.A.M., Leite, M.C.A., de Carvalho, H.W.P., Costa, A.C., Greschuk, L.T., Curi, N., Silva, S.H.G., 2022. The fundamental of the effects of water, organic matter, and iron forms on the pXRF information in soil analyses. *Catena (Amst)* 210. <https://doi.org/10.1016/j.catena.2021.105868>
- Rossiter, D.G., 2018. Past, present & future of information technology in pedometrics. *Geoderma* 324, 131–137. <https://doi.org/10.1016/j.geoderma.2018.03.009>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022a. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruhollah, T.-M., Sheikhpour, R., Zeraatpisheh, M., Amirian-Chakan, A., Toomanian, N., Kerry, R., Scholten, T., 2022b. Semi-supervised learning for the spatial extrapolation of soil information. *Geoderma* 426, 116094. <https://doi.org/10.1016/J.GEODERMA.2022.116094>
- Ruiz, H.A., Ferreira, G.B., Pereira, J.B.M., Resumo,), 2003. Estimativa da capacidade de campo de Latossolos e Neossolos Quartzarênicos pela determinação do equivalente de umidade. *Rev Bras Cienc Solo* 27, 389–393. <https://doi.org/10.1590/S0100-06832003000200019>
- Santos, H.G. dos, Jacomine, P.K.T., Anjos, L.H.C. dos, Oliveira, V.Á. de, Lumbrreras, J.F., Coelho, M.R., Almeida, J.A. de, Filho, J.C. de A., Oliveira, J.B. de, Cunha, T.J.F., 2018. Sistema brasileiro de classificação de solos, 5th ed, Embrapa Solos. Embrapa, Brasília.
- Sarfraz, S., Ali, F., Hameed, A., Ahmad, Z., Riaz, K., 2023. Sustainable Agriculture Through Technological Innovations. *Sustainable Agriculture in the Era of the OMICs Revolution* 223–239. https://doi.org/10.1007/978-3-031-15568-0_10
- Schwertmann, U., Taylor, R.M., 2018. Iron oxides. *Minerals in Soil Environments* 379–438. <https://doi.org/10.2136/sssabookser1.2ed.c8>
- Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S. V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. *Sci Rep* 11, 1–15. <https://doi.org/10.1038/s41598-020-80820-1>
- Shrestha, G., Calvelo-Pereira, R., Roudier, P., Martin, A.P., Turnbull, R.E., Kereszturi, G., Jeyakumar, P., Anderson, C.W.N., 2022. Quantification of multiple soil trace elements by combining portable X-ray fluorescence and reflectance spectroscopy. *Geoderma* 409, 115649. <https://doi.org/10.1016/J.GEODERMA.2021.115649>
- Silva, F.M., Silva, S.H.G., Acuña-Guzman, S.F., Silva, E.A., Ribeiro, B.T., Fruett, T., Inda, A.V., Teixeira, A.F. dos S., Mancini, M., Guilherme, L.R.G., Curi, N., 2021. Chemical and mineralogical changes in the textural fractions of quartzite-derived tropical soils, along weathering, assessed by portable X-ray fluorescence spectrometry and X-ray diffraction. *J South Am Earth Sci* 112, 103634. <https://doi.org/10.1016/J.JSAMES.2021.103634>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022a. Using proximal sensors to assess pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, 1–10. <https://doi.org/10.1016/j.geodrs.2021.e00465>
- Silva, F.M., Silva, S.H.G., Teixeira, A.F. dos S., Inda, A.V., Fruett, T., Weindorf, D.C., Guilherme, L.R.G., Curi, N., 2022b. Using proximal sensors to assess

- pedogenetic development of Inceptisols and Oxisols in Brazil. *Geoderma Regional* 28, e00465. <https://doi.org/10.1016/J.GEODRS.2021.E00465>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020a. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/j.catena.2019.104258>
- Silva, L.S., Marques Júnior, J., Barrón, V., Gomes, R.P., Teixeira, D.D.B., Siqueira, D.S., Vasconcelos, V., 2020b. Spatial variability of iron oxides in soils from Brazilian sandstone and basalt. *Catena (Amst)* 185, 104258. <https://doi.org/10.1016/J.CATENA.2019.104258>
- Silva, S.H.G., Menezes, M.D. de, Owens, P.R., Curi, N., 2016. Retrieving pedologist's mental model from existing soil map and comparing data mining tools for refining a larger area map under similar environmental conditions in Southeastern Brazil. *Geoderma* 267, 65–77. <https://doi.org/10.1016/J.GEODERMA.2015.12.025>
- Silvero, N.E.Q., Siqueira, D.S., Coelho, R.M., da Costa Ferreira, D., Marques, J., 2019. Protocol for the use of legacy data and magnetic signature on soil mapping of São Paulo Central West, Brazil. *Science of The Total Environment* 693, 133463. <https://doi.org/10.1016/J.SCITOTENV.2019.07.269>
- Soriano-Disla, J.M., Janik, L.J., Viscarra Rossel, R.A., MacDonald, L.M., McLaughlin, M.J., 2014. The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties. *Appl Spectrosc Rev* 49, 139–186. <https://doi.org/10.1080/05704928.2013.811081>
- Sothe, C., Almeida, C.M. De, Schimalski, M.B., Liesenberg, V., Rosa, L.E.C. La, Castro, J.D.B., Feitosa, R.Q., 2019. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. <https://doi-org.ez87.periodicos.capes.gov.br/10.1080/01431161.2019.1681600>. <https://doi.org/10.1080/01431161.2019.1681600>
- Sothe, C., de Almeida, C.M., Schimalski, M.B., Liesenberg, V., la Rosa, L.E.C., Castro, J.D.B., Feitosa, R.Q., 2020. A comparison of machine and deep-learning algorithms applied to multisource data for a subtropical forest area classification. *Int J Remote Sens* 41, 1943–1969. <https://doi.org/10.1080/01431161.2019.1681600>
- Souza, E.D., Carneiro, M.A.C., Paulino, H.B., 2005. Atributos físicos de um Neossolo Quartzarênico e um Latossolo Vermelho sob diferentes sistemas de manejo. *Pesqui Agropecu Bras* 40, 1135–1139. <https://doi.org/10.1590/S0100-204X2005001100012>
- Stranghoener, M., Dultz, S., Behrens, H., Schippers, A., 2020. Potential mobilizable Fe from secondary phases of differentially altered subsurface basaltic rock— a sequential extraction study on ICDP site Hawaii. *Applied Geochemistry* 121, 104705. <https://doi.org/10.1016/J.APGEOCHEM.2020.104705>
- Suits, D.B., 1957. Use of Dummy Variables in Regression Equations. *J Am Stat Assoc* 52, 548. <https://doi.org/10.2307/2281705>
- Sumathi, K., 2018. Data Analytics platform for intelligent agriculture. 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on 647–650. <https://doi.org/10.1109/I-SMAC.2018.8653740>
- Sun, Z., Wang, J., Wang, Y., Long, L., Luo, Z., Deng, X., Hu, Q., Wang, M., 2020. Sodium-rich volcanic rocks and their relationships with iron deposits in the Aqishan–Yamansu belt of Eastern Tianshan, NW China. *Geoscience Frontiers* 11, 697–713. <https://doi.org/10.1016/J.GSF.2019.06.011>
- Swetha, R.K., Dasgupta, S., Chakraborty, S., Li, B., Weindorf, D.C., Mancini, M., Silva,

- S.H.G., Ribeiro, B.T., Curi, N., Ray, D.P., 2022. Using Nix color sensor and Munsell soil color variables to classify contrasting soil types and predict soil organic carbon in Eastern India. *Comput Electron Agric* 199, 107192. <https://doi.org/10.1016/J.COMPAG.2022.107192>
- Tabatabai, S., Knadel, M., Thomsen, A., Greve, M.H., 2019. On-the-Go Sensor Fusion for Prediction of Clay and Organic Carbon Using Pre-processing Survey, Different Validation Methods, and Variable Selection. *Soil Science Society of America Journal* 83, 300–310. <https://doi.org/10.2136/SSSAJ2018.10.0377>
- Takoutsing, B., Heuvelink, G.B.M., 2022. Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. *Geoderma* 428, 116192. <https://doi.org/10.1016/J.GEODERMA.2022.116192>
- Teixeira, A.F. dos S., Andrade, R., Mancini, M., Silva, S.H.G., Weindorf, D.C., Chakraborty, S., Guilherme, L.R.G., Curi, N., 2022. Proximal sensor data fusion for tropical soil property prediction: Soil fertility properties. *J South Am Earth Sci* 116, 103873. <https://doi.org/10.1016/J.JSAMES.2022.103873>
- Teixeira, A.F. dos S., Henrique Procópio Pelegrino, M., Missina Faria, W., Henrique Godinho Silva, S., Gabriela Marcolino Gonçalves, M., Weimar Acerbi Júnior, F., Rezende Gomide, L., Linares Pádua Júnior, A., de Souza, I.A., Chakraborty, S., Weindorf, D.C., Roberto Guimarães Guilherme, L., Curi, N., 2020. Tropical soil pH and sorption complex prediction via portable X-ray fluorescence spectrometry. *Geoderma* 361, 114132. <https://doi.org/10.1016/j.geoderma.2019.114132>
- Teixeira, R.S., Vendrame, P.R.S., Christoni, A.R.F., SILVA, P.R.C., 2016. QUÍMICA E MINERALOGIA DE UM SOLO DESENVOLVIDO DE BASALTO, COLETADO ATRAVÉS DE SONDAGEM SPT. *Boletim de Geografia* 34, 116–126. <https://doi.org/10.4025/bolgeogr.v34i2.28948>
- Temme, A.J.A.M., Veldkamp, A., 2009. Multi-process Late Quaternary landscape evolution modelling reveals lags in climate response over small spatial scales. *Earth Surf Process Landf* 34, 573–589. <https://doi.org/10.1002/ESP.1758>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011a. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812–816. <https://doi.org/10.1590/S0103-84782011000500012>
- ten Caten, A., Dalmolin, R.S.D., Pedron, F. de A., Santos, M. de L.M., 2011b. Extrapolação das relações solo-paisagem a partir de uma área de referência. *Ciência Rural* 41, 812–816. <https://doi.org/10.1590/S0103-84782011000500012>
- Thompson, J.A., Roecker, S., Grunwald, S., Owens, P.R., 2012. Digital Soil Mapping: Interactions with and Applications for Hydropedology, in: *Hydropedology*. Elsevier, pp. 665–709. <https://doi.org/10.1016/B978-0-12-386941-8.00021-6>
- Thornthwaite, C.W., 1948. An Approach toward a Rational Classification of Climate Author (s): C . W . Thornthwaite Published by : American Geographical Society Stable URL : <http://www.jstor.org/stable/210739> Accessed : 11-08-2016 18 : 02 UTC. *Geogr Rev* 38, 55–94. <https://doi.org/10.2307/210739>
- Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B., Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. *Geoderma* 367. <https://doi.org/10.1016/j.geoderma.2020.114208>
- Vasques, G.M., Rodrigues, H.M., Coelho, M.R., Baca, J.F.M., Dart, R.O., Oliveira, R.P., Teixeira, W.G., Ceddia, M.B., 2020. Field Proximal Soil Sensor Fusion for Improving High-Resolution Soil Property Maps. *Soil Systems* 2020, Vol. 4, Page 52 4, 52. <https://doi.org/10.3390/SOILSYSTEMS4030052>
- Vieira, S.R., 2000. Geoestatística em estudos de variabilidade espacial do solo. , in: R.F. Novais, V.H.A.G.R.S. (Ed.), *Tópicos Em Ciência Do Solo*. Viçosa, pp. 1–54.

- Wang, J.H., Jiang, J.H., Yu, R.Q., 1996. Robust back propagation algorithm as a chemometric tool to prevent the overfitting to outliers. *Chemometrics and Intelligent Laboratory Systems* 34, 109–115. [https://doi.org/10.1016/0169-7439\(96\)00005-6](https://doi.org/10.1016/0169-7439(96)00005-6)
- Wang, S., Peng, H., Hu, Q., Jiang, M., 2022. Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *J Hydrol Reg Stud* 42, 101139. <https://doi.org/10.1016/J.EJRH.2022.101139>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014a. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, *Advances in Agronomy*. Elsevier. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Bakr, N., Zhu, Y., 2014b. Advances in portable X-ray fluorescence (PXRF) for environmental, pedological, and agronomic applications, in: *Advances in Agronomy*. Academic Press Inc., pp. 1–45. <https://doi.org/10.1016/B978-0-12-802139-2.00001-9>
- Weindorf, D.C., Paulette, L., Man, T., 2013. In-situ assessment of metal contamination via portable X-ray fluorescence spectroscopy: Zlatna, Romania. *Environmental Pollution* 182, 92–100. <https://doi.org/10.1016/j.envpol.2013.07.008>
- Whittig, L.D., 1965. X-Ray Diffraction Techniques for Mineral Identification and Mineralogical Composition. *Methods of Soil Analysis, Part 1: Physical and Mineralogical Properties, Including Statistics of Measurement and Sampling* 671–698. <https://doi.org/10.2134/AGRONMONOGR9.1.C49>
- Xu, D., Chen, S., Xu, H., Wang, N., Zhou, Y., Shi, Z., 2020. Data fusion for the measurement of potentially toxic elements in soil using portable spectrometers. *Environmental Pollution* 263, 114649. <https://doi.org/10.1016/J.ENVPOL.2020.114649>
- Xu, D., Zhao, R., Li, S., Chen, S., Jiang, Q., Zhou, L., Shi, Z., 2019. Multi-sensor fusion for the determination of several soil properties in the Yangtze River Delta, China. *Eur J Soil Sci* 70, 162–173. <https://doi.org/10.1111/EJSS.12729>
- Xu, G., Fan, H., Oliver, D.M., Dai, Y., Li, H., Shi, Y., Long, H., Xiong, K., Zhao, Z., 2022. Decoding river pollution trends and their landscape determinants in an ecologically fragile karst basin using a machine learning model. *Environ Res* 214, 113843. <https://doi.org/10.1016/J.ENVRES.2022.113843>
- Xu, S., Zhao, Y., Wang, M., Shi, X., 2018. Quantification of Different Forms of Iron from Intact Soil Cores of Paddy Fields with Vis-NIR Spectroscopy. *Soil Science Society of America Journal* 82, 1497–1511. <https://doi.org/10.2136/SSSAJ2018.01.0014>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020a. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Sci Data* 7, 1–13. <https://doi.org/10.1038/s41597-019-0345-6>
- Yan, F., Shangguan, W., Zhang, J., Hu, B., 2020b. Depth-to-bedrock map of China at a spatial resolution of 100 meters. *Scientific Data* 2020 7:1 7, 1–13. <https://doi.org/10.1038/s41597-019-0345-6>
- Ye, M., Zhu, L., Li, X., Ke, Y., Huang, Y., Chen, B., Yu, H., Li, H., Feng, H., 2023. Estimation of the soil arsenic concentration using a geographically weighted XGBoost model based on hyperspectral data. *Science of The Total Environment* 858, 159798. <https://doi.org/10.1016/J.SCITOTENV.2022.159798>
- Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G.A., Nannucci, M.S., Leoni, B., Bonomi, T., 2019. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. *J Hydrol (Amst)* 578, 124015. <https://doi.org/10.1016/J.JHYDROL.2019.124015>
- Zeraatpisheh, M., Ayoubi, S., Jafari, A., Finke, P., 2017. Comparing the efficiency of digital and conventional soil mapping to predict soil types in a semi-arid region in

- Iran. *Geomorphology* 285, 186–204.
<https://doi.org/10.1016/J.GEOMORPH.2017.02.015>
- Zevenbergen, C., Honders, A., Orbons, A.J., Viaene, W., Swennen, R., Comans, R.N.J., van Hasselt, H.J., 1997. Immobilisation of heavy metals in contaminated soils by thermal treatment at intermediate temperatures. *Studies in Environmental Science* 71, 661–672. [https://doi.org/10.1016/S0166-1116\(97\)80249-X](https://doi.org/10.1016/S0166-1116(97)80249-X)
- Zhan, Y., Liu, C., Deng, Q., Feng, Q., Qiu, Y., Zhang, A., He, X., 2022. Integrated FFT and XGBoost framework to predict pavement skid resistance using automatic 3D texture measurement. *Measurement* 188, 110638.
<https://doi.org/10.1016/J.MEASUREMENT.2021.110638>
- Zhang, R., Li, Y., Goh, A.T.C., Zhang, W., Chen, Z., 2021. Analysis of ground surface settlement in anisotropic clays using extreme gradient boosting and random forest regression models. *Journal of Rock Mechanics and Geotechnical Engineering* 13, 1478–1484. <https://doi.org/10.1016/J.JRMGE.2021.08.001>
- Zhang, W. chun, Wan, H. shuang, Zhou, M. hou, Wu, W., Liu, H. bin, 2022. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecol Indic* 143, 109420.
<https://doi.org/10.1016/J.ECOLIND.2022.109420>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022a. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, D., Wang, Junjie, Jiang, X., Zhen, J., Miao, J., Wang, Jingzhe, Wu, G., 2022b. Reflectance spectroscopy for assessing heavy metal pollution indices in mangrove sediments using XGBoost method and physicochemical properties. *Catena (Amst)* 211, 105967. <https://doi.org/10.1016/J.CATENA.2021.105967>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022a. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>
- Zhao, L., Fang, Q., Hong, H., Algeo, T.J., Lu, A., Yin, K., Wang, C., Liu, C., Chen, L., Xie, S., 2022b. Pedogenic-weathering evolution and soil discrimination by sensor fusion combined with machine-learning-based spectral modeling. *Geoderma* 409, 115648. <https://doi.org/10.1016/J.GEODERMA.2021.115648>