CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

ESTIMATIVA DO CARBONO ORGÂNICO DO SOLO POR ABORDAGENS QUANTÍLICAS E INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL

Edson Landim de Almeida

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ, como parte dos requisitos necessários à obtenção do grau de mestre.

Orientadores:
Diego Nunes Brandão
Jorge de Abreu Soares

Rio de Janeiro,

6 de agosto de 2026

Estimativa do Carbono Orgânico do Solo por Abordagens Quantílicas e Inteligência Artificial Explicável

Dissertação de Mestrado em Ciência da Computação, Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ.

Edson Landim de Almeida

Aprovada por:

Presidente, Prof. Diego Nunes Brandão, D.Sc. (orientador)

Jorge de Abreu Soares, D.Sc. (coorientador)

Kele Teixeira Belloze, D.Sc. - CEFET/RJ

Marcos Bacis Ceddia, D.Sc. - UFRRJ

Rio de Janeiro,

6 de agosto de 2026

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

A447 Almeida, Edson Landim de
Estimativa do carbono orgânico do solo por abordagens quantílicas e inteligência artificial explicável / Edson Landim de Almeida – 2026.
61f : il. (algumas color.) , enc.

Dissertação (Mestrado). Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2026.
Bibliografia : f. 55-61.
Orientador: Diego Nunes Brandão.
Orientador: Jorge de Abreu Soares.

1. Aprendizado de máquina. 2. Carbono. 3. Solos. 4. Bioquímica ambiental. 5. Inteligência artificial. I. Brandão, Diego Nunes (Orient.). II. Soares, Jorge de Abreu (Orient.). III. Título.

CDD 006.31

Elaborada pelo bibliotecário Leandro Mota de Menezes – CRB/7 nº 5281

RESUMO

Estimativa do Carbono Orgânico do Solo por Abordagens Quantílicas e Inteligência Artificial Explicável

Edson Landim de Almeida

Orientadores:
Diego Nunes Brandão
Jorge de Abreu Soares

Resumo da Dissertação submetida ao Programa de Pós-graduação em Ciência da Computação do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ como parte dos requisitos necessários à obtenção do grau de mestre.

O Carbono Orgânico do Solo (COS) desempenha papel central nos ciclos biogeoquímicos, na fertilidade dos solos e na mitigação das mudanças climáticas, constituindo um dos maiores reservatórios de carbono do sistema terrestre. A quantificação direta do COS por métodos laboratoriais, embora precisa, é onerosa, demorada e limita a densidade amostral em levantamentos de grande escala. Nesse contexto, técnicas estatísticas e de aprendizado de máquina têm sido utilizadas para estimar o teor de carbono a partir de atributos edáficos e ambientais mais facilmente obtidos. Este trabalho investiga a estimativa do carbono orgânico do solo a partir de dados pedológicos disponibilizados pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa), integrando regressão linear múltipla, *Random Forest*, Florestas Aleatórias Quantílicas (*Quantile Regression Forests* – QRF) e técnicas de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) baseadas em *SHapley Additive exPlanations* (*SHAP*). A regressão linear múltipla foi utilizada como linha de base interpretável, enquanto o *Random Forest* foi empregado para capturar relações não lineares e interações entre atributos do solo. A abordagem quantílica baseada em florestas aleatórias permitiu analisar diferentes regiões da distribuição condicional do COS, considerando os quantis 0,1, 0,5 e 0,9. Os resultados evidenciaram que a estimativa do COS é marcada por elevada variabilidade, assimetria e heterocedasticidade, com diferenças relevantes entre os usos da terra avaliados. O modelo *Random Forest* convencional foi capaz de representar tendências gerais dos dados, mas apresentou maior dispersão dos erros em faixas superiores de carbono. O modelo QRF ampliou a análise ao permitir a representação de limites inferiores, medianas e limites superiores da resposta, oferecendo uma leitura mais informativa da incerteza associada às predições. A análise com SHAP indicou a relevância de atributos relacionados à cor do solo, profundidade e textura, em coerência com o conhecimento pedológico. Em conjunto, os resultados indicam que a integração entre modelos baseados em árvores, abordagem quantílica e técnicas de explicabilidade constitui uma estratégia promissora para a estimativa e interpretação do carbono orgânico do solo, contribuindo para o avanço de metodologias aplicadas ao mapeamento digital do solo e ao monitoramento ambiental.

Palavras-chave: carbono no solo; métodos quantílicos; método explicável; pedologia

Rio de Janeiro,

6 de agosto de 2026

ABSTRACT

Estimativa do Carbono Orgânico do Solo por Abordagens Quantílicas e Inteligência Artificial Explicável

Edson Landim de Almeida

Advisors:
Diego Nunes Brandão
Jorge de Abreu Soares

Abstract of dissertation submitted to Programa de Pós-graduação em Ciência da Computação - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca CEFET/RJ as partial fulfillment of the requirements for the degree of master.

Soil organic carbon (SOC) plays a central role in biogeochemical cycles, soil fertility, and climate change mitigation, representing one of the largest carbon reservoirs in the terrestrial system. Direct SOC quantification via laboratory methods, although accurate, is costly and time-consuming and limits sampling density in large-scale surveys. In this context, statistical and machine learning techniques have been used to estimate carbon content from more easily obtainable edaphic and environmental attributes. This work investigates the estimation of soil organic carbon using pedological data provided by the Brazilian Agricultural Research Corporation (Embrapa), integrating multiple linear regression, *Random Forest*, Quantile Regression Forests (QRF), and Explainable Artificial Intelligence (XAI) techniques based on SHAP. Multiple linear regression was used as an interpretable baseline, while *Random Forest* was employed to capture nonlinear relationships and interactions among soil attributes. The quantile-based forest approach enabled the analysis of different regions of the conditional distribution of SOC, considering the 0.1, 0.5, and 0.9 quantiles. The results showed that SOC estimation is characterized by high variability, asymmetry, and heteroscedasticity, with relevant differences among the evaluated land uses. The conventional *Random Forest* model captured general trends in the data but exhibited greater error dispersion in higher-carbon ranges. The QRF model extended the analysis by representing lower, median, and upper bounds on the response, providing a more informative view of the uncertainty associated with the predictions. The SHAP analysis indicated the relevance of attributes related to soil color, depth, and texture, in agreement with pedological knowledge. Overall, the results suggest that integrating tree-based models, quantile-based modeling, and explainability techniques is a promising strategy for estimating and interpreting soil organic carbon, thereby advancing methodologies for digital soil mapping and environmental monitoring.

Key-words: soil organic carbon; quantile approach; explainable artificial intelligence; soil science

Rio de Janeiro,

6 de agosto de 2026

Lista de Figuras

II.1 Ciclo da dinâmica do carbono no solo. Fonte: [EMBRAPA HORTALIÇAS, 2018] . . 6

IV.1 Fluxo geral da metodologia adotada no estudo. Fonte: O próprio autor. 28
IV.2 Imputação e fusão de dados. Fonte: O próprio autor 29
IV.3 Fluxo detalhado da metodologia. Fonte: O próprio autor 37

V.1 Distribuição do teor de carbono orgânico por uso da terra. Fonte: O próprio autor . 40
V.2 Matriz de correlação e significância estatística calculadas com o método de Spearman e aproximação assintótica por distribuição t. Fonte: O próprio autor 42
V.3 Regressão linear para o COS usando os atributos selecionados para os vários usos da terra. Fonte: O próprio autor 42
V.4 Importância relativa dos atributos no modelo *Random Forest*. Fonte: O próprio autor 43
V.5 Otimização do número de atributos por meio de RFE. Fonte: O próprio autor 44
V.6 Comparação entre valores observados e preditos pelo modelo *Random Forest* convencional. Fonte: O próprio autor 45
V.7 Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso Agrícola. Fonte: O próprio autor. 46
V.8 Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso Floresta. Fonte: O próprio autor. 47
V.9 Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso em Pastagem. Fonte: O próprio autor. 47
V.10 Importância global dos atributos segundo valores SHAP para o método *Random Forest* convencional. Fonte: O próprio autor. 48
V.11 Importância global dos atributos segundo os valores de SHAP para o método *Random Forest* quantílico. Fonte: O próprio autor. 49
V.12 Importância dos atributos em cada quantil. Fonte: O próprio autor 50

Lista de Tabelas

III.1 Síntese dos estudos incluídos na revisão da literatura 18
III.2 Resultados de acurácia reportados na literatura 21
III.3 Abordagens de quantificação de incerteza 21
III.4 Técnicas de interpretabilidade e principais achados 22

IV.1 Quantidades de registros no conjunto de dados antes e depois da limpeza . Fonte: O próprio autor 30
IV.2 Atributos selecionados. Fonte: O próprio autor 30
IV.3 Comparação entre os modelos utilizados na estimativa do carbono orgânico do solo. . 35

V.1 Principais bibliotecas usadas na implementação do código em Python. Fonte : O próprio autor 38
V.2 Número de registros por bioma e uso da terra. 41

Lista de Acrônimos

COS Carbono Orgânico no Solo.

GLM Modelos Lineares Generalizados.

MAE Erro Absoluto Médio.

MPIW *Mean Prediction Interval Width*.

ODS Objetivos de Desenvolvimento Sustentável.

PICP *Prediction Interval Coverage Probability*.

QRF *Quantile Regression Forests*.

RFE *Recursive Feature Elimination*.

RLM Regressão Linear Múltipla.

RMSE Erro Quadrático Médio.

SHAP *SHapley Additive exPlanations*.

XAI *Explainable Artificial Intelligence*.

Sumário

I Introdução **1**

II Fundamentação Teórica **5**

II.1 Carbono Orgânico do Solo 5

II.2 Técnicas Tradicionais de Regressão para Predição do Carbono do Solo 7

II.2.1 Regressão Linear Múltipla 7

II.2.2 Modelos Lineares Generalizados 8

II.2.3 Geoestatística e Regressão Espacial 8

II.3 Aprendizado de Máquina na Ciência do Solo 9

II.4 Regressão Quantílica 9

II.5 Florestas Aleatórias Quantílicas 10

II.6 Inteligência Artificial Explicável 11

II.7 Métricas de Avaliação de Modelos 12

II.7.1 Erro Quadrático Médio (RMSE) 12

II.7.2 Erro Absoluto Médio (MAE) 12

II.7.3 Coeficiente de Determinação (R^2) 13

II.7.4 Pinball Loss 13

II.7.5 D^2 Pinball Loss 13

II.8 Incerteza em Modelos Preditivos 14

II.9 Integração entre Pedologia, Aprendizado de Máquina, Incerteza e Explicabilidade 15

III Trabalhos Relacionados **16**

III.1 Metodologia da Revisão 17

III.2 Caracterização dos Estudos 18

III.3 Acurácia dos Modelos 20

III.4 Quantificação de Incerteza 21

III.5 Interpretabilidade e Inteligência Artificial Explicável 22

III.6 Comparação entre QRF e Regressão Quantílica Tradicional 23

III.7 Análise Crítica dos Estudos Selecionados 23

III.8 Resumo da Revisão 27

IV Metodologia 28

IV.1 Aquisição e preparação de dados 28

IV.2 Pré-processamento de dados 29

IV.2.1 Codificação de Variáveis 30

IV.2.2 Estratificação por Uso da Terra 31

IV.2.3 Normalização das Variáveis 31

IV.2.4 Detecção de Outliers 31

IV.2.5 Análise de Correlação 32

IV.3 Seleção de Atributos 32

IV.4 Treinamento do modelo de regressão 32

IV.4.1 Modelagem com Regressão Linear Múltipla 32

IV.4.2 Modelagem com Random Forest 33

IV.4.3 Florestas Aleatórias Quantílicas 33

IV.4.4 Configuração Experimental 34

IV.5 Avaliação dos Modelos de Regressão e Análise Explicável 34

IV.5.1 Comparação entre os modelos utilizados 35

IV.5.2 Análise de Explicabilidade 36

V Resultados e Discussão 38

V.1 Pré-processamento e caracterização dos dados 39

V.2 Regressão linear múltipla 41

V.3 Modelagem do carbono orgânico com Random Forest 41

V.4 Validação do modelo Random Forest 44

V.5 Validação do modelo Random Forest quantílico 46

V.6 Análise explicável dos modelos 48

V.7 Síntese dos resultados 50

VI Considerações Finais 52

Referências 55

Capítulo I Introdução

O carbono orgânico do solo (COS) desempenha papel central nos ciclos biogeoquímicos, na manutenção da fertilidade, na estruturação física do solo e na regulação do clima, constituindo um dos maiores reservatórios de carbono do sistema terrestre [Lal, 2018]. Estima-se que os solos armazenem mais carbono do que a atmosfera e a vegetação combinadas, o que torna o seu manejo adequado um componente estratégico nas políticas de mitigação das mudanças climáticas [Lal, 2013]. Além disso, o teor de carbono orgânico está diretamente associado à capacidade de retenção de água, à disponibilidade de nutrientes, à atividade biológica e à estabilidade dos agregados, influenciando de forma decisiva a produtividade agrícola e a resiliência dos sistemas de uso da terra.

Nesse contexto, a estimativa precisa e confiável do carbono orgânico do solo é fundamental tanto para a agricultura sustentável quanto para a conservação ambiental, em alinhamento direto com os Objetivos de Desenvolvimento Sustentável (ODS) da Organização das Nações Unidas [Carlsen and Bruggemann, 2022], em especial os ODS 2 (Fome Zero e Agricultura Sustentável), 13 (Ação Contra a Mudança Global do Clima) e 15 (Vida Terrestre).

Tradicionalmente, a quantificação do carbono orgânico do solo é realizada por meio de análises laboratoriais que envolvem coleta de amostras em campo, preparo físico-químico do material e aplicação de métodos analíticos específicos, como combustão seca ou oxidação úmida [Teixeira et al., 2017]. Embora esses procedimentos sejam considerados referência em termos de precisão, apresentam custos elevados, demandam tempo, infraestrutura especializada e mão de obra qualificada, além de limitarem a densidade amostral em levantamentos de maior escala. Mais especificamente , o forno utilizado na medição do teor de carbono no solo através de queima da amostra tem um custo elevado de aquisição e de manutenção , e portanto, não está amplamente disponível. Outra limitação é a coleta de amostras deformadas e indeformadas do solo , no caso de se medir o estoque de carbono. Essas amostras são trabalhosas de se obter no campo [Teixeira et al., 2017]. Como consequência, estudos regionais, nacionais ou globais frequentemente enfrentam restrições logísticas e financeiras para obter dados suficientemente densos e representativos.

Esse cenário tem impulsionado o desenvolvimento de abordagens baseadas em modelagem estatística, sensoriamento remoto e aprendizado de máquina, capazes de estimar o teor de carbono orgânico do solo a partir de atributos mais facilmente obtidos, como cor, textura, profundidade, relevo, clima, uso e manejo do solo. Essas abordagens ampliam a viabilidade de aplicações em larga

escala e em contextos com recursos limitados [Pardelinha et al., 2024; Ziviani et al., 2024].

Na interface entre a pedologia e a ciência da computação, modelos preditivos têm sido empregados para capturar relações entre atributos edáficos (variáveis relacionadas às características do solo, como textura, pH, matéria orgânica e densidade), frequentemente utilizadas como preditoras em modelos pedométricos, e o teor de carbono orgânico [Pardelinha et al., 2024]. Modelos estatísticos clássicos, como a regressão linear, oferecem uma primeira aproximação para esse problema, pois permitem avaliar relações globais entre variáveis explicativas e resposta. Entretanto, esses modelos frequentemente apresentam limitações para representar a complexidade dos processos pedogenéticos e a elevada variabilidade espacial e vertical dos solos [Somarathna et al., 2017].

Dados pedológicos caracterizam-se por não linearidades, interações entre variáveis, assimetrias e heterocedasticidade, resultantes da atuação conjunta de fatores de formação do solo, como material de origem, clima, relevo, organismos, tempo e uso antrópico. Assim, modelos baseados apenas em relações lineares globais podem capturar tendências gerais, mas tendem a apresentar dificuldades para representar a variabilidade do carbono orgânico em diferentes usos da terra e em faixas distintas da distribuição da variável resposta.

Nesse contexto, técnicas de aprendizado de máquina, como florestas aleatórias e outros métodos baseados em árvores de decisão, têm sido amplamente utilizadas para modelar relações não lineares entre atributos ambientais e o teor de carbono no solo [McNicol et al., 2019; Emadi et al., 2020; Szatmári et al., 2021]. Esses métodos apresentam flexibilidade para lidar com dados complexos, capturar interações entre variáveis e acomodar relações não lineares, sendo amplamente empregados em estudos de mapeamento digital do solo.

Apesar de seu desempenho preditivo, muitos desses modelos são frequentemente considerados modelos de "caixa-preta", o que dificulta a compreensão dos fatores associados às predições. Na ciência do solo, essa limitação é particularmente relevante, uma vez que a interpretação dos resultados deve dialogar com o conhecimento pedológico e com processos relacionados à formação, à dinâmica e ao manejo dos solos. Assim, embora o aprendizado de máquina permita capturar padrões complexos nos dados, a falta de interpretabilidade pode limitar a confiança e a adoção desses modelos em aplicações científicas e ambientais.

Além da não linearidade, outro desafio importante na estimativa do carbono orgânico do solo é a heterocedasticidade. Em muitos conjuntos de dados pedológicos, a variabilidade do teor de carbono não permanece constante ao longo do domínio dos atributos explicativos. Isso significa que os erros de predição podem ser menores em determinadas faixas de carbono e maiores em outras, especialmente em valores mais elevados ou em contextos de maior heterogeneidade ambiental. Modelos que produzem apenas estimativas pontuais podem não representar adequadamente essa variação da resposta.

Para lidar com esse desafio, extensões quantílicas baseadas em aprendizado de máquina, como as Florestas Aleatórias Quantílicas (*Quantile Regression Forests* – QRF), têm sido propostas [Meinshausen and Ridgeway, 2006]. A QRF estende o modelo de florestas aleatórias ao permitir a estimação de quantis da distribuição condicional da variável resposta, possibilitando analisar diferentes regiões da distribuição do carbono orgânico do solo. Dessa forma, em vez de considerar apenas uma predição média ou pontual, a abordagem quantílica baseada em florestas permite investigar a variabilidade associada às estimativas e compreender como a importância dos atributos pode mudar entre diferentes faixas da resposta [Meinshausen and Ridgeway, 2006; Theres and Sakthivel, 2022a,b].

Neste trabalho, a perspectiva quantílica é incorporada por meio do modelo de *Random Forest* Quantílico, e não pela aplicação de regressão quantílica tradicional paramétrica. Essa escolha metodológica está associada à necessidade de combinar a flexibilidade dos modelos baseados em árvores com a capacidade de analisar a variabilidade condicional do teor de carbono orgânico, especialmente diante da heterocedasticidade observada nos dados pedológicos.

Para aumentar a interpretabilidade dos modelos, técnicas de Inteligência Artificial Explicável (*Explainable Artificial Intelligence* – XAI) podem ser utilizadas em conjunto com modelos de aprendizado de máquina. Métodos baseados em valores de SHAP (*SHapley Additive exPlanations*), por exemplo, permitem decompor as predições do modelo em contribuições individuais das variáveis explicativas, possibilitando analisar tanto explicações locais quanto globais e avaliar se os padrões aprendidos são consistentes com o conhecimento pedológico estabelecido.

Diante desse cenário, emerge a seguinte pergunta de pesquisa: *Como integrar regressão linear, modelos baseados em florestas aleatórias, modelagem quantílica baseada em Random Forest e técnicas de inteligência artificial explicável para estimar e interpretar o carbono orgânico do solo em dados pedológicos heterogêneos?*

No contexto brasileiro, essa discussão é particularmente relevante. O Brasil apresenta grande diversidade de solos, climas, formas de relevo e sistemas de uso da terra, o que resulta em elevada variabilidade espacial do carbono orgânico do solo. Ao mesmo tempo, o país ocupa posição de destaque na produção agrícola mundial e enfrenta desafios relacionados à degradação do solo, às mudanças no uso da terra e às emissões de gases de efeito estufa. A capacidade de estimar e interpretar o carbono orgânico do solo é, portanto, estratégica para a agricultura sustentável, para a gestão ambiental e para o cumprimento de compromissos internacionais relacionados ao clima.

Diante desse cenário, esta dissertação tem como objetivo geral investigar a relação entre atributos diagnósticos do solo e o teor de carbono orgânico, integrando regressão linear, aprendizado de máquina baseado em florestas aleatórias, modelagem quantílica baseada em *Random Forest* e técnicas de interpretabilidade com SHAP.

Como objetivos específicos, busca-se: (i) analisar a distribuição e a estrutura dos dados pedológicos utilizados na estimativa do carbono orgânico do solo; (ii) empregar a regressão linear múltipla como linha de base para avaliar relações iniciais entre os atributos diagnósticos e o teor de carbono; (iii) modelar o teor de carbono orgânico utilizando o algoritmo *Random Forest*; (iv) aplicar o modelo de *Random Forest* Quantílico para investigar diferentes regiões da distribuição condicional do carbono orgânico; e (v) interpretar a contribuição dos atributos nas predições por meio de técnicas de explicabilidade baseadas em SHAP, incluindo a análise da importância dos atributos em diferentes quantis.

A principal contribuição deste trabalho consiste em integrar modelagem preditiva, análise quantílica baseada em florestas aleatórias e interpretação pedológica na investigação do carbono orgânico do solo. Ao combinar regressão linear, *Random Forest*, *Random Forest* Quantílico e SHAP, o estudo busca oferecer uma estrutura analítica capaz de avaliar tendências gerais, capturar relações não lineares, analisar a heterocedasticidade dos dados e interpretar a contribuição dos atributos diagnósticos na predição do carbono orgânico.

Por fim, esta dissertação está estruturada da seguinte forma. O Capítulo II apresenta os fundamentos teóricos relacionados ao carbono orgânico do solo, às técnicas de regressão, ao aprendizado de máquina, à modelagem quantílica e aos métodos de explicabilidade. O Capítulo III apresenta uma breve revisão da literatura sobre estimativa de carbono orgânico em solos. Em seguida, o Capítulo IV descreve os materiais e métodos utilizados, incluindo a base de dados, o pré-processamento e os modelos adotados. O Capítulo V apresenta e discute os resultados obtidos. Por fim, o Capítulo VI apresenta as considerações finais, limitações do estudo e perspectivas para trabalhos futuros.

Capítulo II Fundamentação Teórica

Este capítulo apresenta os fundamentos teóricos que sustentam o desenvolvimento desta dissertação. A organização do capítulo acompanha a estrutura metodológica adotada no trabalho, partindo dos conceitos relacionados ao carbono orgânico do solo e aos atributos pedológicos utilizados na modelagem, avançando para os modelos estatísticos de regressão, os métodos de aprendizado de máquina, as Florestas Aleatórias Quantílicas, as métricas de avaliação, a quantificação de incerteza e as técnicas de Inteligência Artificial Explicável.

A proposta é articular os fundamentos da ciência do solo com os métodos computacionais empregados na estimativa do carbono orgânico. Do ponto de vista pedológico, busca-se compreender como atributos físicos, morfológicos e ambientais se relacionam com a dinâmica do carbono no solo. Do ponto de vista computacional, são apresentados os fundamentos dos modelos utilizados para estimar, avaliar e interpretar essa relação em dados heterogêneos.

II.1 Carbono Orgânico do Solo

O carbono orgânico do solo (COS) constitui a fração orgânica da matéria orgânica do solo e desempenha papel central na estruturação física, na fertilidade química e na atividade biológica dos solos. Do ponto de vista do ciclo global do carbono, os solos representam um dos maiores reservatórios de carbono do sistema terrestre, superando os estoques combinados da atmosfera e da biomassa vegetal [Liang et al., 2019]. Essa característica confere ao solo um papel estratégico na mitigação das mudanças climáticas e na regulação do equilíbrio global de gases de efeito estufa [Rodrigues et al., 2023].

A dinâmica do carbono no solo é controlada por processos de entrada, transformação, estabilização e perda, conforme apresentado na Figura II.1. As entradas ocorrem principalmente por meio de resíduos vegetais, compostos orgânicos liberados pelas raízes e aportes orgânicos externos. A transformação é mediada por organismos do solo, especialmente microrganismos, que decompõem a matéria orgânica e promovem sua humificação [Cotrufo et al., 2013]. A estabilização do carbono ocorre por mecanismos físicos, como a proteção em agregados, e por mecanismos químicos, como a associação com minerais de argila, óxidos de ferro e alumínio. As perdas ocorrem devido à mineralização, à erosão e à lixiviação [Six et al., 2002].

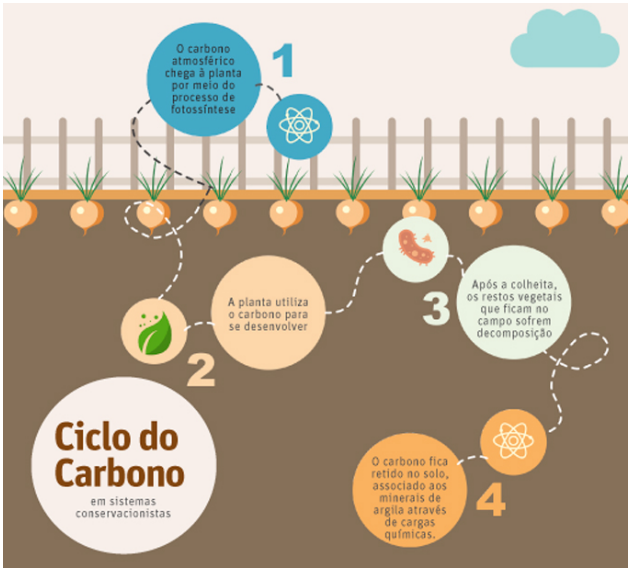


Figura II.1: Ciclo da dinâmica do carbono no solo. Fonte: [EMBRAPA HORTALIÇAS, 2018]

A formação do solo é influenciada por fatores como clima, material de origem, relevo, organismos, tempo e uso antrópico [McBratney et al., 2003]. Dados representativos desses fatores constituem dados pedológicos e podem ser utilizados como variáveis explicativas em modelos de estimativa do carbono orgânico do solo.

A interação entre esses fatores determina a distribuição e os estoques de carbono. A textura do solo influencia diretamente a capacidade de estabilização do carbono, pois partículas finas apresentam maior área superficial, o que favorece a formação de complexos organominerais [Six et al., 2002]. A profundidade está associada a gradientes verticais de matéria orgânica, com maiores teores nas camadas superficiais e redução à medida que se avança no perfil do solo [Jobbágy and Jackson, 2000]. A drenagem influencia as condições de oxirredução e a atividade microbiana, afetando a decomposição da matéria orgânica [Davidson and Janssens, 2006]. A cor do solo, frequentemente expressa em sistemas como o Munsell[1], está relacionada à presença de matéria orgânica e de óxidos de ferro, sendo utilizada como indicador indireto do teor de carbono [Schulze et al., 1993].

A quantificação do COS é tradicionalmente realizada por meio de métodos laboratoriais, como a combustão seca e a oxidação úmida. Embora precisos, esses métodos são custosos e limitam a densidade amostral, o que motiva o uso de abordagens computacionais para estimar o carbono indiretamente com base em atributos edáficos e ambientais.

[1]O sistema de cores Munsell é uma escala padronizada utilizada na descrição morfológica de solos, baseada em três componentes: matiz, valor e croma. O matiz representa a tonalidade dominante, o valor indica a claridade ou escuridão, e o croma expressa a intensidade da cor. Essas informações auxiliam na interpretação de atributos pedológicos, como matéria orgânica, drenagem e presença de óxidos de ferro.

II.2 Técnicas Tradicionais de Regressão para Predição do Carbono do Solo

Antes da consolidação de métodos baseados em aprendizado de máquina, a estimativa do carbono orgânico do solo era amplamente realizada por meio de modelos estatísticos clássicos de regressão [Tomasella and Hodnett, 1998; Rawls, 1983; Viscarra Rossel et al., 2006]. Esses modelos estabelecem relações paramétricas explícitas entre a variável resposta, neste caso o teor de carbono, e um conjunto de variáveis explicativas.

No contexto desta dissertação, os modelos de regressão são importantes porque fornecem uma linha de base interpretável para avaliar a relação entre atributos diagnósticos do solo e o teor de carbono orgânico. Entretanto, como discutido na Seção II.1, os dados pedológicos resultam da interação entre diversos fatores de formação do solo, o que frequentemente produz relações não lineares, assimetrias e heterocedasticidade [McBratney et al., 2003].

A homocedasticidade ocorre quando a variância dos erros permanece aproximadamente constante ao longo do domínio das variáveis explicativas. Em muitos dados de solo, essa condição não é satisfeita, pois a variabilidade do carbono orgânico pode aumentar em determinadas faixas da resposta ou em contextos ambientais específicos. Essa característica limita o desempenho de modelos lineares globais, embora eles continuem úteis como referência inicial [Pearson, 1905].

II.2.1 Regressão Linear Múltipla

A Regressão Linear Múltipla (RLM) assume que a variável resposta Y pode ser descrita como uma combinação linear das variáveis explicativas [Tomasella and Hodnett, 1998]. Sua formulação é apresentada na Equação II.1, em que β_0 representa o intercepto, β_j são os coeficientes da regressão, X_j são as variáveis explicativas e ε é o erro associado ao modelo:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \tag{II.1}$$

A principal vantagem da regressão linear múltipla é sua interpretabilidade. Os coeficientes estimados permitem avaliar a direção e a magnitude da associação entre cada variável explicativa e a resposta, considerando as demais variáveis no modelo. Por essa razão, a RLM é frequentemente utilizada como ponto de partida em problemas de predição.

No entanto, essa abordagem pressupõe linearidade entre as variáveis explicativas e a resposta, além de homocedasticidade dos resíduos. Essas condições podem ser violadas em dados pedológicos, nos quais os atributos do solo interagem de forma complexa e a variabilidade do carbono orgânico pode mudar ao longo da distribuição. Assim, no presente trabalho, a regressão linear múltipla é utilizada como linha de base, permitindo comparar seus resultados com modelos mais flexíveis.

II.2.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (GLM) estendem a regressão linear ao permitir diferentes distribuições para a variável resposta e diferentes funções de ligação [Nelder and Wedderburn, 1972]. Sua formulação geral é dada na equação II.2:

$$g(\mathbb{E}[Y|X]) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{II.2}$$

Nessa formulação, $g(\cdot)$ representa a função de ligação entre a esperança condicional da variável resposta e a combinação linear das variáveis explicativas. Essa estrutura torna os GLM mais flexíveis que a regressão linear clássica, pois permite adaptar o modelo a diferentes tipos de resposta [Nelder and Wedderburn, 1972] .

Apesar dessa maior flexibilidade, os GLM ainda dependem da especificação prévia da estrutura funcional das relações entre variáveis. Em sistemas pedológicos complexos, essa limitação pode dificultar a representação de interações não lineares, efeitos de limiar e padrões heterogêneos de variabilidade.

II.2.3 Geoestatística e Regressão Espacial

Em estudos de mapeamento de solos, técnicas geoestatísticas, como a krigagem e a regressão-krigagem, são frequentemente empregadas para incorporar a dependência espacial das observações. A regressão-krigagem combina um modelo determinístico, responsável por explicar parte da variação da resposta, com a interpolação espacial dos resíduos [McBratney et al., 2003]. Sua expressão é dada pela equação II.3:

$$\hat{Y}(s) = \hat{m}(s) + \hat{\varepsilon}(s) \tag{II.3}$$

Nessa expressão, $\hat{m}(s)$ representa a tendência estimada no ponto espacial s, enquanto $\hat{\varepsilon}(s)$ representa o componente residual interpolado espacialmente. Essas abordagens são úteis quando a localização espacial das amostras contém informação relevante para explicar a distribuição do atributo de interesse.

Apesar de sua importância no mapeamento de solos, métodos geoestatísticos podem apresentar dificuldades para representar interações complexas entre múltiplos atributos ambientais quando usados isoladamente [McBratney et al., 2003]. Essa limitação motiva o uso de métodos de aprendizado de máquina, especialmente em bases de dados heterogêneas e com relações não lineares entre atributos e resposta.

II.3 Aprendizado de Máquina na Ciência do Solo

O aprendizado de máquina consiste em um conjunto de técnicas computacionais capazes de aprender padrões a partir dos dados e realizar predições sem a necessidade de especificar explicitamente uma forma funcional fixa para a relação entre variáveis . Formalmente, dado um conjunto de dados $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, busca-se encontrar uma função $f : \mathbb{R}^p \to \mathbb{R}$ que minimize uma função de perda $L(y, f(x))$ [James et al., 2021].

Na ciência do solo, técnicas como árvores de decisão, florestas aleatórias, máquinas de vetores de suporte, redes neurais artificiais e métodos de *boosting* têm sido amplamente utilizadas para a predição de atributos relacionados aos solos. Esses métodos são capazes de capturar relações não lineares e interações complexas entre variáveis, características comuns em dados pedológicos [McBratney et al., 2003], [Grimm et al., 2008], [Padarian et al., 2020].

As florestas aleatórias, em particular, consistem em um conjunto de árvores de decisão $\{T_b\}_{b=1}^{B}$, em que cada árvore é treinada sobre uma amostra *bootstrap* dos dados e, em cada divisão, considera-se apenas uma subamostra aleatória dos atributos. A amostragem *bootstrap* é construída por meio de sorteios com reposição, gerando diferentes conjuntos de treinamento a partir do conjunto original. Essa estratégia reduz a variância do modelo e contribui para evitar sobreajuste (*overfitting*) [Breiman, 2001].

A predição final do modelo é dada pela média das predições individuais das árvores, conforme a Equação II.4:

$$\hat{f}_{RF}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x). \tag{II.4}$$

Essa estratégia torna o *Random Forest* adequado para dados ruidosos, correlacionados e com interações complexas, como frequentemente ocorre em bases de dados de solo. No contexto desta dissertação, o modelo é utilizado para estimar o teor de carbono orgânico a partir de atributos diagnósticos, funcionando como uma alternativa mais flexível em relação à regressão linear múltipla.

Entretanto, embora o *Random Forest* possa apresentar bom desempenho preditivo, sua predição convencional é pontual, isto é, baseada em uma estimativa central da resposta. Em problemas marcados por heterocedasticidade, como a estimativa do carbono orgânico do solo, essa predição pontual pode ser insuficiente para representar a variabilidade da resposta ao longo da distribuição.

II.4 Regressão Quantílica

Embora métodos de aprendizado de máquina permitam modelar relações complexas entre variáveis, muitos deles concentram-se na estimativa de valores médios ou centrais. Em problemas ambientais e pedológicos, entretanto, é frequentemente necessário modelar não apenas a tendência

central, mas também diferentes regiões da distribuição da variável resposta.

A regressão quantílica é uma generalização da regressão clássica que permite modelar diferentes quantis da distribuição condicional da variável resposta [Koenker, 2005]. Enquanto a regressão linear ordinária estima a esperança condicional $E(Y|X)$, a regressão quantílica estima o quantil τ-ésimo da distribuição condicional $Q_Y(\tau|X)$, para $\tau \in (0,1)$.

Formalmente, seja (x_i, y_i), $i = 1, \ldots, n$, um conjunto de observações, em que $x_i \in \mathbb{R}^p$ representa o vetor de variáveis explicativas e $y_i \in \mathbb{R}$ representa a variável de resposta. A regressão quantílica resolve o seguinte problema de otimização [Koenker, 2005]:

$$\hat{\beta}_\tau = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^\top \beta), \tag{II.5}$$

onde $\rho_\tau(u)$ é a função de perda assimétrica, conhecida como *pinball loss*, definida como:

$$\rho_\tau(u) = \begin{cases} \tau u, & \text{se } u \geq 0, \\ (\tau - 1)u, & \text{se } u < 0. \end{cases} \tag{II.6}$$

Essa função penaliza de forma distinta resíduos positivos e negativos, permitindo controlar o quantil estimado [Koenker, 2005]. Para $\tau = 0{,}5$, obtém-se a regressão pela mediana, que é mais robusta à influência de *outliers* do que a regressão baseada na média.

No contexto do carbono orgânico do solo, a perspectiva quantílica é relevante porque os dados frequentemente apresentam heterocedasticidade, assimetria e caudas longas. A possibilidade de modelar diferentes quantis permite analisar como os atributos edáficos e ambientais se relacionam não apenas com valores típicos de carbono, mas também com regiões inferiores e superiores da distribuição.

No entanto, a regressão quantílica linear assume relações lineares entre as variáveis explicativas e a resposta. Essa suposição pode ser inadequada em sistemas pedológicos complexos, nos quais predominam interações não lineares, efeitos de limiar e relações dependentes do uso da terra. Por essa razão, neste trabalho a perspectiva quantílica é incorporada por meio de Florestas Aleatórias Quantílicas, que combinam a flexibilidade dos modelos baseados em árvores com a estimativa de quantis condicionais.

II.5 Florestas Aleatórias Quantílicas

As Florestas Aleatórias Quantílicas (QRF) estendem as florestas aleatórias tradicionais para estimar a distribuição condicional completa de $Y|X = x$ [Meinshausen and Ridgeway, 2006]. Em vez de armazenar apenas a média da variável resposta em cada nó terminal, a QRF armazena o conjunto de observações associado aos nós terminais das árvores. Dessa forma, torna-se possível

estimar não apenas um valor pontual, mas diferentes quantis da distribuição condicional da resposta.

Seja $w_i(x)$ o peso da observação i na predição para x, definido como a proporção de árvores em que x_i e x caem no mesmo nó terminal. A distribuição condicional é estimada pela equação II.7:

$$\hat{F}(y|x) = \sum_{i=1}^{n} w_i(x)\mathbb{I}(y_i \leq y), \tag{II.7}$$

em que $\mathbb{I}(\cdot)$ é a função indicadora. O quantil τ-ésimo é então obtido pela equação II.8 :

$$\hat{Q}_\tau(x) = \inf\{y : \hat{F}(y|x) \geq \tau\}. \tag{II.8}$$

Essa formulação permite estimar quantis específicos, como $\hat{Q}_{0.1}(x)$, $\hat{Q}_{0.5}(x)$ e $\hat{Q}_{0.9}(x)$, além de intervalos preditivos, como $[\hat{Q}_{0.05}(x), \hat{Q}_{0.95}(x)]$. Com isso, o modelo fornece uma medida direta da variabilidade associada à predição.

No contexto pedológico, a QRF é particularmente adequada porque permite capturar interações complexas entre textura, profundidade, relevo, clima, uso da terra e outros atributos ambientais. Além disso, a abordagem quantílica é útil para representar a heterocedasticidade observada nos dados de carbono orgânico do solo, permitindo analisar separadamente as regiões inferior, mediana e superior da distribuição [Meinshausen and Ridgeway, 2006].

No presente trabalho, a QRF é empregada para complementar o modelo *Random Forest* convencional. Enquanto o *Random Forest* fornece uma predição pontual, a QRF permite investigar como a resposta varia em diferentes quantis e como a importância dos atributos pode mudar ao longo da distribuição do carbono orgânico.

II.6 Inteligência Artificial Explicável

A Inteligência Artificial Explicável (XAI) busca tornar os modelos de aprendizado de máquina mais transparentes e interpretáveis [Arrieta et al., 2020]. Em modelos complexos, como florestas aleatórias, a relação entre atributos explicativos e predições nem sempre é diretamente observável. Por isso, técnicas de explicabilidade são importantes para avaliar quais variáveis contribuíram mais para o comportamento do modelo.

Formalmente, dado um modelo f e uma predição $f(x)$, uma explicação pode ser entendida como uma função g que fornece uma representação compreensível da contribuição de cada atributo x_j para a predição. Técnicas como SHAP [Lundberg and Lee, 2017] baseiam-se na teoria dos valores de Shapley [Shapley, 1953], oriunda da teoria dos jogos, e decompõem a predição de acordo com a equação II.9 :

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j, \tag{II.9}$$

onde ϕ_j representa a contribuição do atributo j para a predição. Essa decomposição permite analisar explicações locais, voltadas a uma observação específica, e explicações globais, obtidas a partir da agregação das contribuições ao longo do conjunto de dados.

Na ciência do solo, a XAI permite avaliar se o modelo atribui maior importância a atributos pedológicos relevantes, como cor, textura, profundidade e drenagem. Isso contribui para verificar a coerência dos padrões aprendidos pelo modelo com o conhecimento pedológico. Entretanto, é importante destacar que os valores SHAP explicam o comportamento do modelo ajustado e não devem ser interpretados, isoladamente, como evidência causal direta dos processos que controlam o carbono orgânico do solo.

II.7 Métricas de Avaliação de Modelos

A avaliação de modelos preditivos envolve a quantificação do erro entre os valores observados y_i e os valores preditos $\hat{y}_i$. No contexto desta dissertação, as métricas são utilizadas para avaliar tanto modelos de predição pontual, como a regressão linear múltipla e o *Random Forest* convencional, quanto modelos quantílicos, como a QRF.

II.7.1 Erro Quadrático Médio (RMSE)

O erro quadrático médio, ou RMSE, é definido pela equação II.10 :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{II.10}$$

O RMSE penaliza erros maiores de forma mais intensa, sendo sensível à presença de *outliers*. Em problemas de estimativa do COS, essa métrica é útil para identificar situações em que o modelo apresenta erros elevados, especialmente em faixas superiores da variável resposta.

II.7.2 Erro Absoluto Médio (MAE)

O erro absoluto médio, ou MAE, é definido em II.11:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{II.11}$$

O MAE fornece uma medida direta do erro médio absoluto e é menos sensível a valores extremos do que o RMSE. Por isso, pode complementar a avaliação do desempenho preditivo em bases com distribuições assimétricas.

II.7.3 Coeficiente de Determinação (R^2)

O coeficiente de determinação [Wooldridge, 2020] é definido em II.12:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{II.12}$$

O R^2 indica a proporção da variância da variável resposta explicada pelo modelo. Em modelos de predição pontual, valores mais altos indicam maior capacidade de representar a variação observada nos dados.

II.7.4 Pinball Loss

Para modelos quantílicos, a função de perda *pinball* é dada em II.13 :

$$L_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}), & \text{se } y \geq \hat{y}, \\ (1 - \tau)(\hat{y} - y), & \text{se } y < \hat{y}. \end{cases} \tag{II.13}$$

Essa métrica avalia diretamente a qualidade das estimativas de quantis, penalizando de forma assimétrica erros acima ou abaixo do quantil estimado. Dessa forma, a *pinball loss* é adequada para avaliar modelos como a QRF, nos quais o objetivo não é apenas predizer uma média, mas estimar diferentes regiões da distribuição condicional da resposta.

II.7.5 D^2 Pinball Loss

Além da função de perda *pinball*, modelos quantílicos também podem ser avaliados por meio do D^2 *pinball loss* [McFadden, 1974]. Essa métrica funciona como uma medida relativa de desempenho para modelos quantílicos, de modo análogo ao papel desempenhado pelo coeficiente de determinação R^2 em modelos de predição pontual.

O D^2 *pinball loss* compara a perda *pinball* obtida pelo modelo ajustado com a perda obtida por um modelo de referência. De forma geral, pode ser expresso na equação II.14:

$$D^2_{\text{pinball}} = 1 - \frac{\sum_{i=1}^{n} L_\tau(y_i, \hat{q}_\tau(x_i))}{\sum_{i=1}^{n} L_\tau(y_i, \tilde{q}_\tau)} \tag{II.14}$$

onde L_τ representa a função de perda *pinball*, $\hat{q}_\tau(x_i)$ é o quantil estimado pelo modelo para a observação i, e $\tilde{q}_\tau$ representa o quantil empírico utilizado como modelo de referência.

Valores de D^2_{pinball} mais próximos de 1 indicam melhor desempenho relativo do modelo quantílico. Valores próximos de 0 indicam que o modelo apresenta desempenho semelhante ao modelo de referência. Valores negativos indicam que o modelo ajustado apresenta desempenho inferior ao modelo de referência.

No contexto desta dissertação, o D^2 *pinball loss* é utilizado para avaliar o desempenho do *Random Forest* quantílico nos quantis 0,1, 0,5 e 0,9. Essa métrica permite verificar o ganho preditivo relativo do modelo em diferentes regiões da distribuição condicional do carbono orgânico do solo [Gneiting, 2011].

II.8 Incerteza em Modelos Preditivos

A incerteza em modelos preditivos pode ser decomposta em duas componentes principais: incerteza aleatória, também chamada de irredutível, e incerteza epistêmica, também chamada de redutível [Hüllermeier and Waegeman, 2021]. A incerteza aleatória está associada à variabilidade intrínseca do fenômeno observado, enquanto a incerteza epistêmica está relacionada a limitações dos dados disponíveis, da estrutura do modelo ou do processo de aprendizagem.

Em modelos quantílicos, a incerteza pode ser representada diretamente por meio de intervalos preditivos. Um intervalo de predição para um nível de confiança $1 - \alpha$ é dado por II.15 :

$$[\hat{Q}_{\alpha/2}(x), \hat{Q}_{1-\alpha/2}(x)] \tag{II.15}$$

A qualidade desses intervalos pode ser avaliada por métricas de calibração, como a taxa de cobertura, também conhecida como PICP (*Prediction Interval Coverage Probability*), e a largura média do intervalo, conhecida como MPIW(*Mean Prediction Interval Width*) [Khosravi et al., 2011]. A taxa de cobertura é definida na equação II.16:

$$PICP = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_i \in [\hat{Q}_{\alpha/2}(x_i), \hat{Q}_{1-\alpha/2}(x_i)]) \tag{II.16}$$

A PICP mede a proporção de valores observados que ficam dentro do intervalo preditivo estimado pelo modelo. Em um intervalo nominal de 90%, por exemplo, espera-se que aproximadamente 90% das observações estejam contidas entre os quantis inferior e superior estimados. Valores muito baixos de PICP indicam que os intervalos são estreitos demais ou mal calibrados, enquanto valores muito altos podem indicar intervalos excessivamente amplos.

A largura média do intervalo de predição, ou MPIW, mede a amplitude média dos intervalos gerados pelo modelo. Em modelos quantílicos, essa métrica pode ser calculada pela diferença entre os quantis superior e inferior estimados para cada observação, como dado na equação II.17 :

$$MPIW = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{Q}_{1-\alpha/2}(x_i) - \hat{Q}_{\alpha/2}(x_i) \right) \tag{II.17}$$

Valores menores de MPIW indicam intervalos mais estreitos e, portanto, menor incerteza associada às predições. No entanto, essa métrica deve ser analisada em conjunto com a PICP, pois

intervalos muito estreitos podem apresentar baixa cobertura dos valores observados. Assim, um bom modelo quantílico deve buscar equilíbrio entre alta cobertura e intervalos não excessivamente largos.

No contexto do carbono orgânico do solo, a quantificação da incerteza é fundamental, pois decisões relacionadas ao manejo, às políticas públicas e à contabilização de carbono devem considerar não apenas valores pontuais, mas também a variabilidade e o risco associados às estimativas.

II.9 Integração entre Pedologia, Aprendizado de Máquina, Incerteza e Explicabilidade

A integração entre pedologia, aprendizado de máquina, quantificação de incerteza e explicabilidade constitui o eixo metodológico desta dissertação. Do ponto de vista computacional, essa integração permite desenvolver modelos capazes de lidar com dados complexos, heterogêneos e não lineares. Do ponto de vista pedológico, permite interpretar os resultados à luz dos processos de formação do solo, da dinâmica da matéria orgânica e das características ambientais associadas ao carbono orgânico.

A regressão linear múltipla fornece uma linha de base interpretável para avaliar relações globais entre atributos diagnósticos e COS. O *Random Forest* convencional permite capturar relações não lineares e interações entre variáveis. A QRF amplia essa análise ao estimar diferentes quantis da distribuição condicional, permitindo representar a heterocedasticidade e a variabilidade associada às predições. Por fim, as técnicas de XAI, especialmente SHAP, contribuem para interpretar o comportamento dos modelos e verificar se os padrões aprendidos são coerentes com o conhecimento pedológico.

Essa abordagem integrada é particularmente relevante em contextos de elevada diversidade edafoclimática, como o Brasil, nos quais a distribuição do carbono orgânico do solo varia em função do bioma, do uso da terra, da profundidade, da textura, da drenagem e de outros fatores ambientais. Assim, a fundamentação teórica apresentada neste capítulo sustenta a proposta desta dissertação de combinar regressão linear, *Random Forest*, Florestas Aleatórias Quantílicas e Inteligência Artificial Explicável na estimativa e interpretação do carbono orgânico do solo.

Capítulo III Trabalhos Relacionados

A estimativa do carbono orgânico do solo (COS) constitui um tema central na ciência do solo, em razão de sua relação direta com a mitigação das mudanças climáticas, a segurança alimentar, a qualidade ambiental e o manejo sustentável dos ecossistemas. O COS atua como um importante reservatório terrestre de carbono, e sua dinâmica é influenciada por processos biogeoquímicos complexos associados ao clima, ao relevo, ao material de origem, aos organismos, ao uso e à cobertura da terra e às práticas de manejo. Apesar de sua relevância, a quantificação direta do COS ainda depende, em grande parte, de análises laboratoriais onerosas, demoradas e espacialmente limitadas. Esse contexto tem motivado o desenvolvimento de modelos preditivos baseados em dados ambientais, produtos de sensoriamento remoto, técnicas geoestatísticas e métodos de aprendizado de máquina.

Nos últimos anos, abordagens baseadas em *Quantile Regression Forests* (QRF) têm recebido atenção crescente no mapeamento digital de solos, especialmente por permitirem estimar não apenas valores pontuais, mas também quantis da distribuição condicional da variável resposta. Essa característica torna a QRF particularmente relevante para problemas nos quais a quantificação da incerteza é tão importante quanto a acurácia preditiva. Paralelamente, técnicas de Inteligência Artificial Explicável, do inglês *Explainable Artificial Intelligence* (XAI), têm sido empregadas para aumentar a transparência de modelos complexos, permitindo identificar a contribuição relativa das covariáveis ambientais e interpretar padrões espaciais e pedológicos associados às estimativas de COS. Em contraste, a regressão quantílica tradicional permanece relevante por sua capacidade de modelar diferentes pontos da distribuição condicional e por oferecer maior interpretabilidade paramétrica e suporte à inferência estatística formal.

Diante desse cenário, este capítulo apresenta trabalhos relacionados à estimativa do carbono orgânico do solo com foco na comparação entre regressão quantílica tradicional, Florestas de Regressão Quantílica e abordagens de Inteligência Artificial Explicável. A análise busca discutir como essas abordagens se diferenciam em termos de acurácia, quantificação da incerteza e interpretabilidade, bem como identificar lacunas metodológicas que justifiquem a investigação desenvolvida nesta dissertação.

III.1 Metodologia da Revisão

A revisão da literatura foi orientada por uma busca semântica realizada na plataforma Elicit[1], a partir da seguinte questão de pesquisa: "Como as Florestas de Regressão Quantílica e as abordagens de inteligência artificial explicável se comparam à regressão quantílica tradicional em termos de acurácia, quantificação da incerteza e interpretabilidade na estimativa do carbono orgânico do solo?". A busca foi conduzida em uma base com mais de 138 milhões de artigos acadêmicos indexados pela plataforma, incluindo registros provenientes do *Semantic Scholar* e do OpenAlex. Essa estratégia permitiu identificar estudos relacionados à modelagem do carbono orgânico do solo, à regressão quantílica, aos métodos de aprendizado de máquina e às técnicas de inteligência artificial explicável.

A busca semântica retornou os 495 artigos considerados mais relevantes para a questão de pesquisa. Esses trabalhos foram submetidos a uma etapa de triagem com base nos títulos e resumos, considerando conjuntamente critérios metodológicos e temáticos. Foram priorizados estudos que: aplicassem QRF, regressão quantílica tradicional ou técnicas de XAI no contexto da estimativa de COS; utilizassem dados reais de solo ou dados sintéticos realistas; apresentassem métricas de acurácia, procedimentos de quantificação da incerteza ou análises de interpretabilidade; fossem publicados como artigos revisados por pares, anais de conferência ou revisões sistemáticas; e fornecessem detalhamento metodológico suficiente para avaliar a implementação dos modelos utilizados.

Também foram considerados critérios de escopo, de modo a excluir estudos que tratassem exclusivamente de outros atributos do solo, como pH, nutrientes ou textura, sem incluir estimativas de carbono orgânico do solo. Do mesmo modo, estudos que apenas mencionavam regressão quantílica, QRF ou XAI, sem aplicação direta à modelagem de COS, não foram priorizados. A decisão de inclusão foi realizada de forma holística, considerando o alinhamento geral de cada estudo com a questão de pesquisa, e não apenas o atendimento isolado a um critério específico.

Após a triagem, foi realizada uma extração estruturada das informações dos estudos selecionados. Essa extração contemplou os métodos de modelagem empregados, as estratégias de validação, as métricas de desempenho, os procedimentos de quantificação da incerteza, as técnicas de interpretabilidade utilizadas, o contexto geográfico e ambiental, as características dos conjuntos de dados, os resultados específicos relacionados à estimativa de COS e as limitações metodológicas reportadas pelos autores. Ao final desse processo de busca, triagem e extração, 36 fontes foram selecionadas para compor a análise detalhada desta revisão da literatura.

[1] https://elicit.com/

III.2 Caracterização dos Estudos

As fontes analisadas abrangem diferentes contextos geográficos, incluindo aplicações na Europa, Ásia, África, América do Sul, América do Norte e Oceania. A maior parte dos trabalhos concentra-se na estimativa do COS em camadas superficiais do solo, especialmente entre 0 e 30 cm, embora alguns estudos abordem perfis mais profundos, estoques totais ou frações específicas do carbono, como o carbono orgânico associado a minerais.

A Tabela III.1 apresenta uma síntese dos estudos incluídos na revisão. A seleção desses trabalhos ilustra a diversidade de métodos, regiões e métricas de carbono orgânico do solo considerados.

Tabela III.1: Síntese dos estudos incluídos na revisão da literatura

Estudo	**Método Principal**	**Região**	**Métrica de COS**
Helfenstein et al. [2021]	QRF	Países Baixos	Carbono orgânico do solo; métrica específica não informada no material
Lee and Lee [2024]	DT, RF, GBoost e SHAP	Coreia do Sul	Matéria orgânica do solo
Hasanawi and Purwanto [2024]	QRF	Ilhas de Sumatra e Java, Indonésia	Concentração de COS
Ugbaje et al. [2024]	QRF e XAI	Austrália	Estoque de COS e mudança temporal
Salgado et al. [2025]	QRF, PLSR e Extra-Trees	Allande, Astúrias, noroeste da Espanha	Frações de carbono orgânico do solo
van der Westhuizen et al. [2023]	RF, QRF e krigagem	Não especificada no material	COS e nitrogênio total
Wadoux et al. [2023]	RF e SHAP	França continental	Estoque de COS
Haithem et al. [2022]	QRF	Tunísia	Estoque de COS no topsoil
Nikou and Tziachris [2022]	QRF, RF e krigagem	Norte da Grécia, unidade regional de Kastoria	Matéria orgânica do solo
Bryan et al. [2024]	QRF, RNAs, SVM e AD	Sub-bacia do rio Chambo, província de Chimborazo	Conteúdo de COS
Dharumarajan et al. [2020]	QRF	Platô Norte de Karnataka, Índia	Carbono orgânico do solo
Gavilán-Acuña et al. [2024]	QRF	Região do Maule, centro-sul do Chile	Matéria orgânica do solo

Estudo	Método Principal	Região	Métrica de COS
Zhang et al. [2023]	QRF com Boruta	Planície do Nordeste da China	Concentração de COS
Durante et al. [2024]	QRF, SVM e RF	Região de Murcia, sul da Península Ibérica	Concentração e estoque de COS
Baveye [2021]	RQ	Global	COS; métrica específica não informada no material
Santos [nd]	QRF e SHAP	Escala global	COS, quantis e intervalos de incerteza
Heuvelink et al. [2020]	QRF e IPCC	Argentina	Estoque de COS e mudança temporal
Baltensweiler et al. [2021]	QRF, krigagem, T-cubist e SVM	Áreas florestais da Suíça	Conteúdo de COS
Lamichhane et al. [2021]	RF e QRF	Bacia de Palung, região montanhosa central do Nepal	Conteúdo de COS
Mendes et al. [2025]	QRF, R-Bayes e GLMs	Não especificada no material	Estoque de COS em perfil
Heuvelink et al. [2020]	QRF	Argentina, com foco na região dos Pampas	Estoque de COS e mudança temporal
Szatmári et al. [2019]	QRF	Hungria	Estoque de COS no topsoil
Cheng et al. [2024]	X-SSHM, RF e GSHAP	Não especificada no material	COS; métrica específica não informada no material
Urbina-Salazar et al. [2023]	QRF	Região de Beauce, França	Conteúdo de COS
Ding et al. [2025]	XGBoost e SHAP	Xinjiang, China	Concentração e densidade de COS
Dong et al. [2025]	RF e SHAP	Nordeste da China	Concentração de COS
Wadoux [2023]	RF e SHAP	34 países europeus	Carbono orgânico total
Szatmári and Pásztor [2020]	QRF, krigagem, Gauss e RF-krigagem	Hungria	Estoque de COS
Castaldi et al. [2023]	QRF, QRF e LM	Hemisfério Norte; áreas agrícolas na Türkiye, Itália, Lituânia e EUA	Concentração de COS
Baake et al. [2018]	QRF, krigagem e RF	Edgeroi, leste da Austrália	Conteúdo de COS

Estudo	Método Principal	Região	Métrica de COS
Uduwerella et al. [2023]	QRF	Sri Lanka	COS
Masobeng et al. [2025]	QRF	Sub-bacias selecionadas do Lesoto: Maletsunyane e Makhalaneng	Estoque de COS e mudança temporal
Teleki et al. [2025]	QRF	Seis sub-bacias do Lesoto	Concentração de COS em 0–20 cm e 20–40 cm
Shi et al. [2025]	RQ e RF	Metade norte da Irlanda	MAOC e potencial de sequestro
Theres and Sakthivel [2022b]	Não especificado no material	Blocos de Salem e Rasipuram, Tamil Nadu, Índia	Carbono orgânico

De modo geral, os estudos diferem substancialmente quanto ao tamanho amostral, à resolução espacial das covariáveis, à profundidade do solo considerada, ao tipo de validação adotado e à natureza da variável de resposta. Essa heterogeneidade limita comparações diretas entre resultados absolutos de desempenho, mas permite identificar tendências metodológicas relevantes, especialmente quanto ao uso crescente de QRF para a quantificação de incerteza e de técnicas de XAI para a interpretação de modelos preditivos.

III.3 Acurácia dos Modelos

Os resultados de acurácia reportados para modelos baseados em QRF apresentam variação expressiva, com valores de R^2 entre 0,19 e 0,86. Essa amplitude reflete diferenças entre as regiões de estudo, tipos de solo, profundidades avaliadas, densidade amostral, qualidade dos dados e conjuntos de covariáveis ambientais utilizados. Em geral, os modelos tendem a apresentar melhor desempenho em camadas superficiais e em contextos com maior densidade de amostras e maior disponibilidade de covariáveis ambientais relevantes.

A Tabela III.2 resume alguns dos principais resultados de acurácia reportados na literatura, destacando estudos representativos entre as fontes analisadas.

Tabela III.2: Resultados de acurácia reportados na literatura

Estudo	R^2	**RMSE**	**Validação**
Nikou and Tziachris [2022]	0,53	0,63	*Hold-out*
Teleki et al. [2025]	0,72	1,45	*10-fold Cross-validation*
Zhang et al. [2023]	0,68	5,10	*Cross-validation + Out-of-Bag*
Haithem et al. [2022]	0,44	1,94	Externa
Hasanawi and Purwanto [2024]	0,71	0,68	*Cross-validation*
Shi et al. [2025]	0,86	4,41	Reamostragem

Os estudos revisados indicam que a QRF apresenta desempenho competitivo em relação a métodos como krigagem, PLSR, SVM, redes neurais artificiais e outras abordagens de aprendizado de máquina. No entanto, a literatura não permite concluir, de forma definitiva, que a QRF seja superior em todos os contextos. O desempenho dos modelos depende fortemente da estrutura dos dados, da escala espacial, das covariáveis utilizadas e do procedimento de validação. Além disso, a regressão quantílica tradicional aparece em um número reduzido de estudos, geralmente associada à investigação de limites de saturação de carbono ou à modelagem de quantis específicos, o que dificulta comparações sistemáticas com QRF.

III.4 Quantificação de Incerteza

Uma das principais vantagens da QRF é sua capacidade de estimar quantis da distribuição condicional da variável resposta, permitindo construir intervalos de predição sem pressupor uma forma paramétrica específica para os resíduos. Essa característica é particularmente relevante na estimativa de COS, uma vez que os erros de predição podem variar espacialmente e ser influenciados por fatores como heterogeneidade ambiental, densidade amostral e resolução das covariáveis.

Diversos estudos relatam valores de *Prediction Interval Coverage Probability* (PICP) entre 83% e 92%, sugerindo boa calibração dos intervalos de predição em determinados contextos. A Tabela III.3 apresenta uma síntese das principais abordagens de quantificação de incerteza identificadas.

Tabela III.3: Abordagens de quantificação de incerteza

Estudo	**Método de Incerteza**	**Resultado**
Dharumarajan et al. [2020]	PICP	83–92%
Zhang et al. [2023]	Intervalos 90%	PICP = 0,88
Baake et al. [2018]	Intervalos QRF	Boa calibração
Szatmári and Pásztor [2020]	QRF vs SGS	QRF e SGS superiores
Shi et al. [2025]	Regressão Quantílica Local	Limites realistas de MAOC

Os resultados indicam que QRF e Simulação Gaussiana Sequencial (SGS) podem apresentar estimativas de incerteza mais bem calibradas do que abordagens baseadas exclusivamente em krigagem universal ou combinações de floresta aleatória com krigagem. Ainda assim, a comparação entre métodos de incerteza deve ser interpretada com cautela, pois os estudos variam quanto ao desenho amostral, às métricas utilizadas, à escala espacial e à forma de validação dos intervalos de predição.

III.5 Interpretabilidade e Inteligência Artificial Explicável

A incorporação de técnicas de XAI tem se tornado uma estratégia recorrente para lidar com uma limitação importante dos modelos de aprendizado de máquina: a dificuldade de interpretar os mecanismos subjacentes às predições. Entre as técnicas identificadas, destacam-se os valores de Shapley, frequentemente implementados por meio de SHAP, além de medidas de importância de variáveis, gráficos de dependência parcial e abordagens voltadas à análise da heterogeneidade espacial das contribuições das covariáveis.

A Tabela III.4 resume algumas das principais técnicas de interpretabilidade utilizadas nos estudos revisados.

Tabela III.4: Técnicas de interpretabilidade e principais achados

Estudo	**Técnica XAI**	**Principais Variáveis**
Wadoux et al. [2023]	SHAP	Vegetação, relevo
Ding et al. [2025]	SHAP + XGBoost	Condutividade, salinidade
Dong et al. [2025]	SHAP + RF	Cobertura de resíduos
Lee and Lee [2024]	SHAP, PDP	Uso da terra, NDVI
Cheng et al. [2024]	GSHAP	Heterogeneidade espacial

Os estudos indicam que o clima, a vegetação, o relevo, o uso e a cobertura da terra, a textura, a mineralogia e as propriedades químicas do solo, como a CEC, são frequentemente entre os principais preditores do COS. Entretanto, a importância relativa dessas variáveis varia espacialmente e depende do contexto ambiental analisado. Assim, os resultados produzidos por técnicas de XAI devem ser interpretados como evidências associativas e explicativas do comportamento do modelo, e não como demonstrações causais dos processos pedológicos subjacentes.

Essa distinção é importante porque métodos como SHAP podem revelar padrões coerentes com o conhecimento do domínio, mas não substituem desenhos experimentais ou análises causais. Ainda assim, sua utilização contribui para aumentar a transparência dos modelos e pode auxiliar na formulação de hipóteses sobre os fatores ambientais que condicionam a distribuição espacial do COS.

III.6 Comparação entre QRF e Regressão Quantílica Tradicional

A comparação direta entre QRF e a regressão quantílica tradicional ainda é pouco explorada na literatura sobre a estimativa do carbono orgânico do solo. Entre os estudos analisados, apenas dois empregam métodos tradicionais de regressão quantílica de forma mais direta. O estudo de Baveye [2021] utiliza regressão quantílica global para discutir limites associados à iniciativa "4p1000", enquanto Shi et al. [2025] propõe uma regressão quantílica local baseada em espectroscopia MIR para estimar limites de saturação de MAOC.

Os resultados disponíveis sugerem que a regressão quantílica tradicional pode apresentar vantagens em contextos específicos, especialmente quando o objetivo é interpretar coeficientes, investigar limites condicionais ou realizar inferência estatística formal. Por exemplo, a regressão quantílica local permitiu estimar limites de saturação de carbono mais realistas, entre 44 e 82 g C kg^{-1}, em comparação com uma estimativa teórica global constante. Esse resultado indica que abordagens quantílicas tradicionais ainda podem oferecer contribuições relevantes para problemas específicos de ciência do solo.

Por outro lado, a QRF apresenta maior flexibilidade para lidar com relações não lineares, interações complexas e conjuntos de covariáveis de alta dimensionalidade. Além disso, sua capacidade de gerar distribuições condicionais empíricas torna o método atrativo para a quantificação de incerteza em aplicações de mapeamento digital de solos. Contudo, essa flexibilidade vem acompanhada de menor interpretabilidade paramétrica e de limitações na inferência estatística formal.

Dessa forma, as evidências atualmente disponíveis não sustentam uma conclusão definitiva quanto à superioridade da QRF em relação à regressão quantílica tradicional. A ausência de estudos comparativos diretos, conduzidos sobre os mesmos conjuntos de dados e avaliados pelas mesmas métricas, constitui uma lacuna importante na literatura. Essa lacuna justifica a necessidade de investigações que comparem sistematicamente essas abordagens em termos de acurácia, incerteza e interpretabilidade.

Embora os estudos selecionados forneçam evidências relevantes sobre o uso de QRF e XAI na estimativa do COS, a análise crítica da literatura revela limitações metodológicas importantes, especialmente quanto à comparabilidade dos resultados, à validação da incerteza e à ausência de confrontos diretos com a regressão quantílica tradicional.

III.7 Análise Crítica dos Estudos Selecionados

A análise dos estudos revisados indica que a literatura sobre a estimativa do carbono orgânico do solo tem avançado significativamente no uso de métodos baseados em aprendizado de máquina, especialmente as Florestas de Regressão Quantílica. No entanto, esse avanço não ocorre de ma-

neira homogênea. A maior parte dos trabalhos utiliza QRF como ferramenta preditiva ou como mecanismo de quantificação de incerteza, mas poucos estudos discutem de forma aprofundada as implicações estatísticas, interpretativas e pedológicas das escolhas metodológicas adotadas. Além disso, a comparação direta entre QRF e regressão quantílica tradicional permanece praticamente ausente, o que limita a obtenção de conclusões mais robustas sobre a superioridade de uma abordagem em relação à outra.

Um primeiro aspecto crítico refere-se à heterogeneidade dos contextos de aplicação. Estudos como os de Durante et al. [2024], Castaldi et al. [2023], Heuvelink et al. [2020], Zhang et al. [2023] e Teleki et al. [2025] mostram que o desempenho dos modelos depende fortemente da escala espacial, da profundidade do solo, da densidade amostral e da disponibilidade de covariáveis ambientais. Essa variação dificulta comparações diretas entre os valores de acurácia reportados. Por exemplo, resultados obtidos em áreas agrícolas relativamente homogêneas ou em camadas superficiais do solo tendem a não ser diretamente comparáveis a estudos realizados em áreas extensas, com maior heterogeneidade ambiental ou com presença de solos orgânicos e turfeiras. Assim, embora a QRF apresente desempenho competitivo em diferentes aplicações, os resultados devem ser interpretados como dependentes do contexto, e não como evidência de superioridade universal do método.

O estudo de Hasanawi and Purwanto [2024] é particularmente ilustrativo dessa limitação. Ao aplicar QRF para mapear COS nas ilhas de Sumatra e Java, os autores reportam desempenho substancialmente distinto entre as duas regiões, com melhor ajuste em Java, dominada por solos minerais, e desempenho inferior em Sumatra, onde há maior presença de solos turfosos. Esse resultado sugere que a QRF pode apresentar dificuldades em contextos caracterizados por distribuições altamente assimétricas, valores extremos ou regimes pedológicos muito distintos. Portanto, a adoção de QRF em áreas com forte heterogeneidade do COS exige cuidado adicional, seja por meio de estratificação espacial, aumento da densidade amostral ou uso de modelos específicos para diferentes classes de solo.

Outro ponto relevante diz respeito aos estudos que comparam o QRF a métodos geoestatísticos ou a outros algoritmos de aprendizado de máquina. Trabalhos como Nikou and Tziachris [2022], Salgado et al. [2025], Baltensweiler et al. [2021] e Haithem et al. [2022] indicam que a QRF frequentemente apresenta desempenho igual ou superior a métodos como krigagem ordinária, krigagem com deriva externa, PLSR, SVM e produtos globais como o SoilGrids. No entanto, em alguns casos, a superioridade da QRF é apenas marginal. Em Nikou and Tziachris [2022], por exemplo, a floresta aleatória apresentou um valor de R^2 ligeiramente superior ao da QRF, ainda que ambos tenham desempenho semelhante. Esse tipo de resultado reforça que a principal contribuição da QRF pode não estar apenas na melhoria da acurácia pontual, mas, sobretudo, na possibilidade de estimar distribuições condicionais e intervalos de predição.

A quantificação da incerteza é, de fato, um dos aspectos mais fortes da QRF na literatura analisada. Estudos como Dharumarajan et al. [2020], Zhang et al. [2023], Baake et al. [2018] e Szatmári and Pásztor [2020] destacam o uso de quantis, intervalos de predição e métricas, como a PICP, para avaliar a confiabilidade das estimativas. Essa característica é especialmente importante no mapeamento digital de solos, pois mapas de COS sem incerteza associada podem induzir interpretações excessivamente determinísticas. Ainda assim, a literatura mostra que a incerteza nem sempre é avaliada com o mesmo rigor. Em alguns estudos, a incerteza é apenas mencionada ou apresentada visualmente, sem avaliação formal de calibração. Em outros, a validação dos intervalos é realizada por validação cruzada, mas sem um conjunto externo independente. Isso limita a capacidade de avaliar a generalização espacial das estimativas.

A comparação entre QRF e métodos geoestatísticos de incerteza também merece atenção. Szatmári and Pásztor [2020] sugere que QRF e Simulação Gaussiana Sequencial apresentam melhor desempenho na quantificação da incerteza do que krigagem universal e combinações de floresta aleatória com krigagem. Entretanto, Baake et al. [2018] aponta que a QRF pode gerar intervalos excessivamente otimistas em determinadas situações, enquanto a regressão de krigagem pode ser mais conservadora. Essa divergência indica que a qualidade da incerteza estimada não deve ser assumida automaticamente pelo método utilizado. Pelo contrário, deve ser avaliada empiricamente em cada aplicação, preferencialmente com dados independentes e métricas específicas de calibração.

No campo da interpretabilidade, os estudos baseados em Inteligência Artificial Explicável representam um avanço importante, mas também introduzem novas limitações. Trabalhos como Wadoux [2023], Lee and Lee [2024], Dong et al. [2025] e Ding et al. [2025] utilizam valores de Shapley, gráficos de dependência parcial ou estratégias similares para identificar os principais fatores associados à variação do COS. Esses estudos contribuem para aproximar modelos de aprendizado de máquina do conhecimento pedológico, ao indicar a relevância de variáveis como vegetação, clima, relevo, textura, condutividade elétrica, salinidade e cobertura de resíduos culturais. No entanto, os resultados de XAI devem ser interpretados como explicações do comportamento do modelo, e não como evidência causal dos processos que controlam o COS.

Essa distinção é fundamental. Valores de Shapley indicam como determinadas covariáveis contribuíram para as predições de um modelo treinado em um conjunto de dados específico. Eles não demonstram, por si só, que uma variável seja causalmente responsável pelo aumento ou pela redução do COS. Além disso, técnicas como SHAP podem ser sensíveis à correlação entre preditores, à escolha do algoritmo, à escala espacial dos dados e à distribuição das amostras. Portanto, embora XAI aumente a transparência dos modelos, ela não elimina a necessidade de interpretação pedológica cuidadosa nem substitui desenhos experimentais ou análises causais.

O trabalho de Cheng et al. [2024], ao propor uma abordagem espacialmente explícita baseada

em GSHAP, aponta uma direção promissora para lidar com a heterogeneidade espacial. A principal contribuição desse tipo de abordagem está em reconhecer que a importância das covariáveis não é necessariamente estacionária no espaço. Em outras palavras, uma variável pode contribuir positivamente para a predição do COS em uma região e negativamente em outra, dependendo das condições climáticas, do uso do solo, do relevo e das propriedades edáficas locais. Essa perspectiva é especialmente relevante para o mapeamento de COS em grandes extensões territoriais, mas também aumenta a complexidade computacional e interpretativa dos modelos.

Em contraste com a expansão dos métodos de QRF e XAI, a regressão quantílica tradicional aparece de forma limitada nos estudos analisados. Baveye [2021] utiliza regressão quantílica em escala global para discutir limites associados ao sequestro de carbono, enquanto Shi et al. [2025] propõe uma regressão quantílica local baseada em espectroscopia MIR para estimar limites de saturação do carbono orgânico associado a minerais. O estudo de Shi et al. [2025] é especialmente relevante porque sugere que abordagens quantílicas locais podem capturar melhor a variabilidade associada à mineralogia e ao manejo do solo do que modelos globais baseados em limites fixos. Esse resultado indica que a regressão quantílica tradicional, embora menos explorada, ainda possui valor metodológico em problemas nos quais a interpretação dos quantis, dos limites condicionais e dos coeficientes é central.

A ausência de comparações diretas entre QRF e regressão quantílica tradicional constitui, portanto, uma lacuna importante. A QRF apresenta vantagens evidentes em cenários com relações não lineares, interações complexas e grande número de covariáveis ambientais. Por outro lado, a regressão quantílica tradicional oferece maior transparência paramétrica, permite interpretação direta dos efeitos das covariáveis em diferentes pontos da distribuição condicional e possibilita inferência estatística formal. A literatura revisada ainda não esclarece em quais situações a flexibilidade preditiva da QRF compensa a perda de interpretabilidade estatística da regressão quantílica tradicional.

Também é necessário destacar limitações metodológicas da própria literatura revisada. Muitos estudos utilizam validação cruzada, mas poucos recorrem à validação externa independente. Além disso, há variação substancial nas métricas reportadas, nos procedimentos de amostragem, nas profundidades consideradas e nas covariáveis utilizadas. Essa heterogeneidade dificulta a realização de uma síntese quantitativa mais rigorosa. Outro aspecto importante é que parte das informações disponíveis na revisão foi extraída de resumos ou de descrições parciais dos trabalhos, o que pode limitar a avaliação detalhada de hiperparâmetros, procedimentos de pré-processamento, critérios de validação e limitações reconhecidas pelos próprios autores.

Dessa forma, a literatura atual permite afirmar que a QRF é uma abordagem robusta e amplamente utilizada para a estimativa e o mapeamento do COS, especialmente quando a quantificação da incerteza é relevante. No entanto, as evidências ainda não permitem concluir que ela seja supe-

rior à regressão quantílica tradicional em todos os aspectos. A principal lacuna está na ausência de estudos que comparem, sobre os mesmos dados e sob o mesmo protocolo de validação, a acurácia, a calibração da incerteza e a interpretabilidade de QRF, regressão quantílica tradicional e modelos combinados com XAI. Essa lacuna justifica a necessidade de estudos comparativos mais controlados, capazes de avaliar não apenas o desempenho preditivo, mas também a utilidade científica e interpretativa de cada abordagem.

III.8 Resumo da Revisão

A revisão realizada evidencia que as Florestas de Regressão Quantílica têm se consolidado como uma abordagem relevante para a estimativa do carbono orgânico do solo, especialmente em aplicações que demandam quantificação da incerteza. Os estudos analisados indicam um desempenho preditivo geralmente competitivo e a capacidade de produzir intervalos de predição bem calibrados em diferentes contextos. No entanto, os resultados também demonstram que a acurácia dos modelos é sensível à densidade amostral, à profundidade do solo, ao tipo de solo, à escala espacial e à qualidade das covariáveis ambientais disponíveis.

As técnicas de Inteligência Artificial Explicável, em especial SHAP, agregam valor interpretativo aos modelos de aprendizado de máquina ao permitir a identificação de variáveis influentes e a análise de suas contribuições locais e globais. Ainda assim, tais técnicas devem ser interpretadas com cautela, pois explicam o comportamento do modelo e não necessariamente os mecanismos causais que regem a dinâmica do carbono orgânico do solo.

A regressão quantílica tradicional, embora menos frequente nos estudos revisados, permanece metodologicamente relevante por sua interpretabilidade, por sua capacidade de modelar diferentes pontos da distribuição condicional e por seu potencial para análises inferenciais. A literatura, contudo, carece de comparações diretas entre a regressão quantílica tradicional, o QRF e abordagens integradas com XAI.

Assim, a principal lacuna identificada nesta revisão é a ausência de estudos que comparem, de forma sistemática e com os mesmos dados, essas abordagens quanto à acurácia preditiva, à qualidade da quantificação da incerteza e ao nível de interpretabilidade oferecido. A presente dissertação insere-se precisamente nesse contexto, contribuindo para o avanço metodológico da estimativa do carbono orgânico do solo ao avaliar comparativamente modelos quantílicos tradicionais, modelos baseados em florestas e estratégias de explicabilidade aplicadas a bases de dados nacionais e internacionais.

Capítulo IV Metodologia

Este capítulo apresenta os procedimentos metodológicos adotados para a construção do conjunto de dados, o pré-processamento das variáveis, o treinamento dos modelos de regressão e a análise interpretável dos resultados. A estratégia metodológica foi concebida para integrar princípios da pedologia e da ciência de dados, permitindo estimar o teor de carbono orgânico do solo a partir de atributos edáficos e ambientais, ao mesmo tempo em que se busca garantir rigor científico, robustez estatística e reprodutibilidade dos experimentos.

De forma geral, o fluxo metodológico foi organizado em quatro etapas principais: (i) obtenção, preparação e análise exploratória dos dados pedológicos; (ii) construção do conjunto de dados para modelagem; (iii) treinamento e avaliação dos modelos de regressão; e (iv) interpretação dos resultados por meio de técnicas de Inteligência Artificial Explicável. A Figura IV.1 apresenta uma visão geral desse processo.

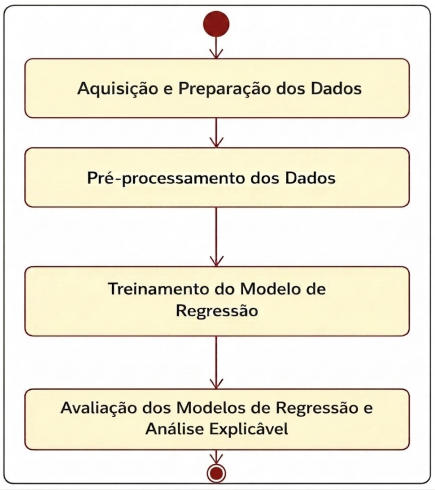


Figura IV.1: Fluxo geral da metodologia adotada no estudo. Fonte: O próprio autor.

IV.1 Aquisição e preparação de dados

Os dados utilizados neste estudo foram obtidos a partir do banco de dados de solos da Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), denominado BDSolos [EMBRAPA SOLOS, 2014]. Esse repositório reúne informações detalhadas sobre perfis de solo coletados em diferentes regiões do território brasileiro, incluindo atributos morfológicos, físicos e químicos relevantes

para a caracterização pedológica.

No BDSolos, um ponto de amostragem pode representar tanto um perfil completo quanto uma tradagem. Para os objetivos desta pesquisa, foram considerados exclusivamente os perfis completos, uma vez que a análise do teor de carbono orgânico exige a preservação da estrutura vertical do perfil do solo. Essa escolha permite uma interpretação pedológica mais consistente das relações entre profundidade, propriedades físicas, atributos morfológicos e concentração de carbono.

Os dados do BDSolos foram disponibilizados em arquivos no formato CSV (*Comma Separated Values*), os quais foram consolidados em uma base única para análise. Essa consolidação foi feita usando o nome do município em cada registro como referência e a posterior consulta a uma planilha do MapBiomas [Souza et al., 2020] para a obtenção do bioma brasileiro correspondente. A figura IV.2 representa essa sequência de atividades.

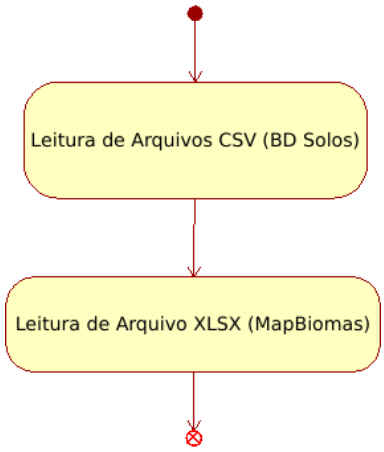


Figura IV.2: Imputação e fusão de dados. Fonte: O próprio autor

IV.2 Pré-processamento de dados

Após a consolidação dos dados, realizou-se um processo de validação e limpeza dos dados, incluindo a identificação de registros duplicados, a verificação de inconsistências e o tratamento de valores ausentes. Registros com informações críticas incompletas foram removidos, de modo a assegurar a consistência do conjunto de dados utilizado nas etapas posteriores. Por exemplo, registros do BDSolos com ausência do teor de carbono ou sem alguma componente da cor foram removidos. Assim como registros com medidas inconsistentes , provavelmente resultantes de erros de medida. A Tabela IV.1 mostra as quantidades de registros pré- e pós-limpeza.

Uma vez obtido um conjunto de dados com registros íntegros , conduziu-se uma Análise Exploratória de Dados (EDA) com o objetivo de compreender as características fundamentais do conjunto de dados. Essa etapa envolveu a análise das distribuições das variáveis, a identificação de possíveis *outliers*, a avaliação das correlações entre atributos e a inspeção de padrões de heterogeneidade presentes nos dados pedológicos. Os resultados dessa análise subsidiaram as decisões relativas à

Tabela IV.1: Quantidades de registros no conjunto de dados antes e depois da limpeza . Fonte: O próprio autor

Antes da limpeza		**Depois da limpeza**
Quantidade de registros com profundidade inconsistente	19	0
Quantidade de registros com algum valor nulo	10770	0
Total de registros	32147	21358

Tabela IV.2: Atributos selecionados. Fonte: O próprio autor

Nome	**Relação**	**Tipo/Valores**	**Unidade**
Valor amostra úmida	Cor	Numérico (entre 2 e 8)	Escala Munsell
Croma amostra úmida	Cor	Numérico (entre 1 e 8)	Escala Munsell
Ponto médio espessura	Profundidade	Numérico	centímetros
Fração Total de Argila	Textura	Numérico	grama/kilograma
pH	Química	Numérico	sem dimensão
Drenagem	Física	Categórico (0 - Boa, 1 - Média, 2 - Ruim)	sem dimensão
Altitude	Ambiental	Número	metros
bioma	Ambiental	Categórico (1 -Amazônia ,2-Cerrado, 3-Caatinga, 4 -Mata Atlântica, 5 - Pantanal, 6 -Pampa)	sem dimensão

seleção de atributos, à transformação da variável alvo, à estratificação por uso da terra e à escolha dos modelos empregados.

A escolha das variáveis explicativas foi orientada pelos atributos diagnósticos descritos no Sistema Brasileiro de Classificação de Solos (SiBCS) [Santos et al., 2018]. Dessa forma, buscou-se garantir que o conjunto de atributos utilizados na modelagem apresentasse coerência pedológica e refletisse propriedades reconhecidas como relevantes para a dinâmica da matéria orgânica do solo.

A Tabela IV.2 apresenta os atributos utilizados na modelagem.

IV.2.1 Codificação de Variáveis

A classe de drenagem foi codificada numericamente por meio de *label encoding*, preservando a ordenação natural das categorias. Essa técnica realiza o mapeamento de atributos categóricos para valores numéricos, sendo adequada quando existe uma relação ordinal entre as categorias [Müller and Guido, 2016]. Posteriormente, essas classes foram agrupadas em três categorias principais: boa, média e ruim. Essa agregação reduziu a granularidade da variável e favoreceu a representação de padrões gerais associados às condições de drenagem.

O atributo bioma foi transformado por meio de *one-hot encoding*, uma vez que suas categorias não possuem relação ordinal. Essa técnica cria variáveis indicadoras binárias, atribuindo valor 1 para a presença de uma categoria e 0 para sua ausência [Zheng and Casari, 2018]. Dessa forma, evita-se impor uma ordem artificial entre biomas distintos.

IV.2.2 Estratificação por Uso da Terra

Considerando a influência do uso da terra sobre a dinâmica do carbono orgânico, os dados foram segmentados em três subconjuntos principais: (i) Floresta; (ii) Pastagem; e (iii) Agricultura.

Essa segmentação foi adotada para reduzir a heterogeneidade do conjunto de dados e permitir avaliar o comportamento dos modelos em diferentes contextos de uso do solo. A separação por uso da terra também possibilita interpretar os resultados de forma mais coerente com os processos pedológicos e antrópicos associados ao acúmulo, à perda e à estabilização do carbono orgânico.

Além disso, foram consideradas apenas amostras com profundidade inferior ou igual a 30 cm, uma vez que a maior parte da matéria orgânica do solo tende a se concentrar nas camadas superficiais. Essa decisão também contribui para reduzir a variabilidade decorrente de diferenças muito acentuadas de profundidade entre os registros.

IV.2.3 Normalização das Variáveis

Antes do treinamento dos modelos, as variáveis numéricas foram submetidas a procedimentos de normalização e padronização. Para variáveis com distribuição aproximadamente simétrica, foi utilizado o *StandardScaler* [Pedregosa et al., 2011], que transforma os dados para média zero e desvio padrão unitário. Para variáveis com presença de *outliers* ou distribuições assimétricas, foi utilizado o *RobustScaler* [Pedregosa et al., 2011], que realiza a transformação com base em estatísticas robustas, como a mediana e o intervalo interquartil.

A variável alvo, correspondente ao carbono orgânico do solo, apresentou distribuição assimétrica. Para reduzir esse efeito, foi aplicada a transformação logarítmica dada pela Equação IV.1

$$y' = \log(y + 1) \tag{IV.1}$$

Essa transformação reduz a influência de valores extremos e torna a distribuição da variável resposta mais adequada ao treinamento dos modelos. A adição de 1 evita problemas matemáticos em casos de valores próximos de zero.

IV.2.4 Detecção de Outliers

A identificação de observações discrepantes foi realizada utilizando o algoritmo *Local Outlier Factor* (LOF). Esse método é baseado em densidade local e identifica como possíveis *outliers* os registros cuja densidade local é substancialmente menor do que a densidade observada em sua vizinhança [Breunig et al., 2000].

A adoção do LOF é adequada para dados pedológicos, pois permite identificar observações anômalas sem eliminar automaticamente registros raros, mas potencialmente válidos. Assim, a remoção

de observações foi realizada de forma criteriosa. Registros identificados como discrepantes foram removidos apenas quando sua exclusão resultou em melhoria superior a 5% nas métricas de desempenho do modelo. Esse critério buscou evitar a exclusão indevida de observações pedologicamente relevantes.

IV.2.5 Análise de Correlação

A relação entre as variáveis explicativas foi analisada utilizando o coeficiente de correlação de Spearman [Zar, 2010], apropriado para dados não gaussianos, relações monotônicas e variáveis ordinais. Apenas correlações com significância estatística ($p < 0.05$) foram consideradas relevantes.

Essa etapa teve como objetivo avaliar possíveis associações entre os atributos, identificar redundâncias e orientar a interpretação posterior dos modelos. Embora métodos baseados em árvores, como o *Random Forest*, sejam menos sensíveis à multicolinearidade do que modelos lineares, a presença de correlação entre atributos pode influenciar medidas de importância e explicações baseadas em SHAP.

IV.3 Seleção de Atributos

Mesmo após a análise de correlação, a presença de múltiplas variáveis pode aumentar a complexidade do modelo e introduzir redundâncias. Para reduzir esse efeito, foi utilizada a técnica de Eliminação Recursiva de Atributos (*Recursive Feature Elimination* – RFE) [Guyon et al., 2002]. Esse método remove iterativamente os atributos menos relevantes com base em sua importância no modelo.

A aplicação do RFE teve como objetivo identificar um subconjunto de atributos capaz de preservar a capacidade preditiva dos modelos, reduzindo a dimensionalidade e favorecendo a interpretabilidade. A seleção de atributos também contribui para evitar a inclusão de variáveis pouco informativas ou redundantes, o que pode comprometer a generalização dos modelos.

IV.4 Treinamento do modelo de regressão

O conjunto de dados final obtido após a etapa de pré-processamento foi utilizado para treinar os modelos de Regressão Linear Múltipla, *Random Forest* Padrão e o *Random Forest* Quantílico. Esses modelos servirão de base para a posterior análise explicativa.

IV.4.1 Modelagem com Regressão Linear Múltipla

A regressão linear múltipla foi utilizada como modelo de referência para a estimativa do carbono orgânico do solo. Essa escolha se justifica pela interpretabilidade do método e por sua utilidade

como linha de base para comparação com modelos mais flexíveis.

O modelo considera que a variável resposta pode ser representada como uma combinação linear dos atributos explicativos selecionados. A formulação geral adotada é a mesma apresentada na Equação II.1, descrita no Capítulo II. No contexto desta dissertação, a regressão linear múltipla foi empregada para avaliar a capacidade inicial dos atributos diagnósticos de explicar a variação do carbono orgânico do solo.

Apesar de sua simplicidade, esse modelo é importante porque permite verificar se relações globais entre atributos pedológicos e carbono orgânico podem ser capturadas por uma estrutura linear. Os resultados obtidos com essa abordagem serviram como referência para a avaliação dos modelos baseados em florestas aleatórias.

IV.4.2 Modelagem com Random Forest

O algoritmo *Random Forest* foi utilizado como modelo não paramétrico de regressão. Esse método combina múltiplas árvores de decisão treinadas a partir de amostras *bootstrap* do conjunto de dados, reduzindo a variância do modelo e aumentando sua capacidade de generalização. A Equação IV.2 apresenta o modelo preditivo deste algoritmo.

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{IV.2}$$

onde $T_b(x)$ representa a predição da árvore b e B representa o número total de árvores da floresta.

A utilização do *Random Forest* se justifica pela sua capacidade de capturar relações não lineares e interações entre atributos pedológicos. Essa característica é particularmente relevante para a estimativa do carbono orgânico do solo, uma vez que sua dinâmica depende da interação entre textura, profundidade, drenagem, cor do solo, bioma e outros fatores ambientais.

Neste estudo, o *Random Forest* foi utilizado para gerar predições pontuais do teor de carbono orgânico. Seus resultados foram avaliados por métricas de desempenho e pela análise da relação entre valores observados e preditos.

IV.4.3 Florestas Aleatórias Quantílicas

Para representar explicitamente a variabilidade associada às predições, foi utilizado o método de Florestas Aleatórias Quantílicas (QRF). Esse método estende o *Random Forest* convencional ao permitir a estimação da distribuição condicional da variável resposta. Sua expressão é dada pela Equação II.7, e ela não fornece apenas uma estimativa pontual.

Neste estudo, foram avaliados os quantis 0,1, 0,5 e 0,9. O quantil 0,1 representa uma estimativa associada à região inferior da distribuição; o quantil 0,5 corresponde à mediana; e o quantil 0,9 re-

presenta a região superior da distribuição. Essa abordagem permite analisar não apenas a tendência central do carbono orgânico, mas também a variabilidade e a incerteza associadas às predições em diferentes usos da terra.

A adoção do QRF está diretamente relacionada à presença de heterocedasticidade nos dados. Como a variabilidade do carbono orgânico pode mudar ao longo da distribuição, modelos que fornecem apenas uma predição pontual podem não representar adequadamente os limites inferiores e superiores da resposta.

IV.4.4 Configuração Experimental

Os experimentos foram realizados utilizando a divisão dos dados em conjuntos de treinamento e de teste na proporção de 80/20. Essa divisão permitiu treinar os modelos em uma parcela dos dados e avaliar sua capacidade de generalização em dados não utilizados durante o ajuste.

Os hiperparâmetros dos modelos baseados em florestas foram ajustados por meio de busca em grade (*grid search*), considerando diferentes valores para o número de árvores, a profundidade máxima e o número mínimo de amostras por nó. Esse procedimento teve como objetivo selecionar configurações adequadas para melhorar o desempenho preditivo e reduzir o risco de sobreajuste.

Todos os experimentos foram conduzidos utilizando a linguagem Python e bibliotecas comumente utilizadas em ciência de dados, incluindo Scikit-learn, Pandas, NumPy e SHAP.

IV.5 Avaliação dos Modelos de Regressão e Análise Explicável

O desempenho dos modelos de regressão foi avaliado utilizando métricas amplamente empregadas em problemas de predição contínua. Essas métricas permitem quantificar a diferença entre os valores observados de carbono orgânico do solo e os valores estimados pelos modelos, fornecendo uma medida objetiva da qualidade das predições.

Para os modelos de predição pontual, foram utilizadas três métricas principais: Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE) e coeficiente de determinação (R^2). A utilização conjunta dessas métricas permite obter uma avaliação mais completa do desempenho dos modelos. Enquanto o RMSE enfatiza erros de grande magnitude, o MAE fornece uma medida robusta do erro médio, e o coeficiente R^2 permite avaliar a capacidade global do modelo de representar a variabilidade dos dados.

Além da avaliação quantitativa por meio dessas métricas, foi realizada análise visual da relação entre valores observados e preditos. Essa análise permite identificar possíveis padrões sistemáticos de erro, como tendência de superestimação ou subestimação em determinadas faixas de carbono orgânico.

Para os modelos quantílicos, a avaliação foi realizada por meio da função de perda *pinball* e

do D^2 *pinball loss*. A função de perda *pinball* avalia diretamente a qualidade das estimativas de quantis, penalizando de forma assimétrica erros acima ou abaixo do quantil estimado. Já o D^2 *pinball loss* mede o ganho preditivo relativo do modelo quantílico em comparação a um modelo de referência baseado no quantil empírico.

Valores mais próximos de 1 para o D^2 *pinball loss* indicam melhor desempenho relativo do modelo quantílico. Valores próximos de 0 indicam desempenho semelhante ao modelo de referência, enquanto valores negativos indicam desempenho inferior. Neste estudo, essa métrica foi utilizada para avaliar o desempenho do QRF nos quantis 0,1, 0,5 e 0,9.

No contexto deste estudo, a avaliação dos modelos também considerou a calibração visual dos quantis. Para um quantil τ, espera-se que aproximadamente uma proporção τ das observações esteja abaixo das predições correspondentes. Dessa forma, a análise conjunta do *pinball loss*, do D^2 *pinball loss* e da posição dos valores observados em relação aos quantis preditos permitiu avaliar a utilidade dos modelos quantílicos para representar regiões inferiores, medianas e superiores da distribuição do carbono orgânico do solo.

IV.5.1 Comparação entre os modelos utilizados

Com o objetivo de analisar diferentes estratégias de modelagem para a estimativa do carbono orgânico do solo, foram empregados três tipos principais de modelos: regressão linear múltipla, *Random Forest* convencional e Florestas Aleatórias Quantílicas. Cada abordagem possui características distintas em termos de capacidade de modelagem, interpretabilidade e representação da incerteza.

A regressão linear múltipla representa uma abordagem paramétrica e interpretável, útil como linha de base para avaliar relações globais entre os atributos diagnósticos e o carbono orgânico. O *Random Forest* convencional constitui um método de aprendizado de máquina não paramétrico, capaz de capturar relações não lineares e interações complexas entre os atributos. Por fim, as Florestas Aleatórias Quantílicas estendem o *Random Forest* ao permitir a estimativa de diferentes quantis da distribuição condicional da variável resposta, possibilitando uma representação mais direta da variabilidade e da incerteza. A Tabela IV.3 apresenta um resumo das principais características dessas abordagens.

Tabela IV.3: Comparação entre os modelos utilizados na estimativa do carbono orgânico do solo.

Característica	**Regressão Linear**	**Random Forest**	**QRF**
Tipo de modelo	Paramétrico	Não paramétrico	Não paramétrico
Relações não lineares	Limitadas	Sim	Sim
Captura de interações	Limitada	Sim	Sim
Interpretabilidade	Alta	Média	Média
Estimativa de incerteza	Limitada	Limitada	Direta (quantis)
Complexidade computacional	Baixa	Moderada	Moderada
Adequação para dados pedológicos	Moderada	Alta	Alta

IV.5.2 Análise de Explicabilidade

Para interpretar os resultados dos modelos, foi utilizado o método SHAP (*SHapley Additive exPlanations*). Esse método permite decompor a predição do modelo em contribuições individuais de cada atributo, conforme apresentado na Equação IV.3.

$$f(x) = \phi_0 + \sum_{j=1}^{p} \phi_j \tag{IV.3}$$

em que ϕ_0 representa o valor base da predição e ϕ_j representa a contribuição do atributo j para a predição do modelo.

Essa abordagem permite identificar os atributos mais relevantes para a estimativa do carbono orgânico do solo e verificar se os padrões aprendidos pelo modelo são coerentes com o conhecimento pedológico. No presente estudo, a análise SHAP foi aplicada tanto ao modelo *Random Forest* convencional quanto ao modelo *Random Forest* quantílico.

Além da importância global dos atributos, também foi analisada a contribuição dos atributos em diferentes quantis da distribuição condicional do carbono orgânico. Essa análise permitiu verificar se a importância das variáveis se mantém constante ou varia entre regiões inferiores, medianas e superiores da resposta.

Cabe destacar que os valores SHAP foram utilizados como ferramenta de interpretação do comportamento dos modelos ajustados. Portanto, seus resultados devem ser compreendidos como explicações das predições realizadas pelos modelos, e não como evidência causal direta dos processos pedológicos.

Por fim, a Figura IV.3 apresenta o fluxo completo da metodologia adotada neste estudo.

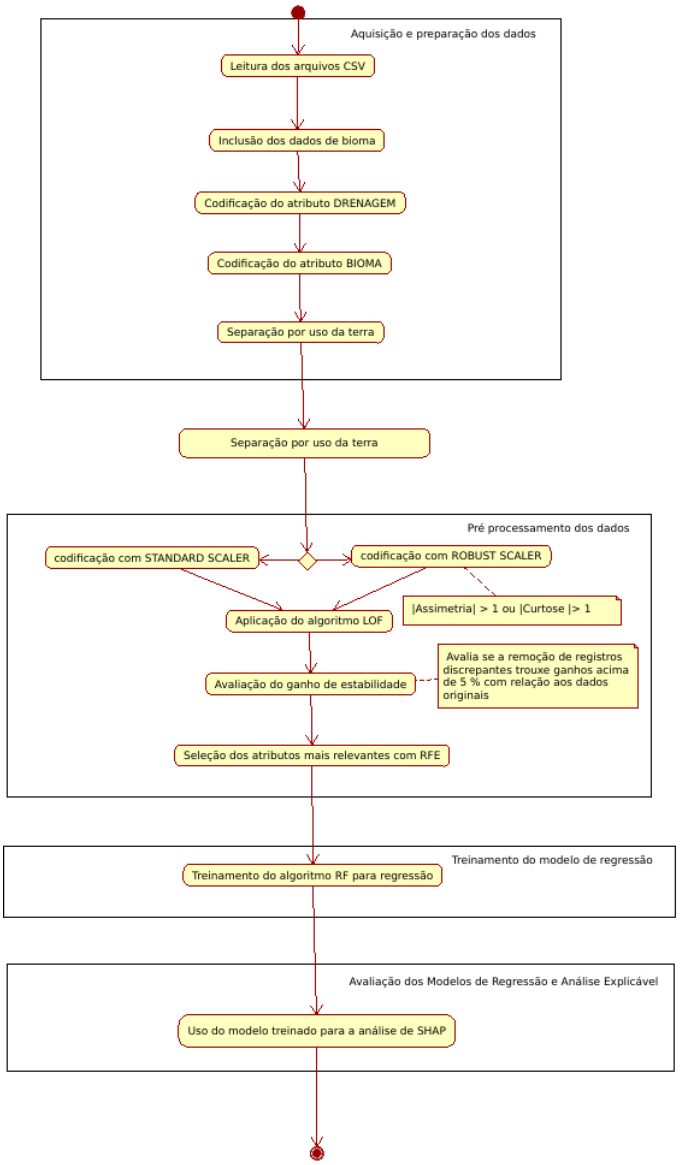


Figura IV.3: Fluxo detalhado da metodologia. Fonte: O próprio autor

Capítulo V Resultados e Discussão

Este capítulo apresenta e discute os resultados obtidos a partir das etapas metodológicas descritas no Capítulo IV. A análise foi organizada de forma progressiva, partindo da caracterização dos dados e avançando para modelos de complexidade crescente. Inicialmente, são apresentados os resultados do pré-processamento, da distribuição do carbono orgânico do solo e da estrutura dos dados por bioma e uso da terra. Em seguida, a regressão linear múltipla é utilizada como modelo de referência, permitindo avaliar relações globais entre os atributos diagnósticos e o teor de carbono orgânico do solo.

Na sequência, são discutidos os resultados obtidos com o modelo *Random Forest*, incluindo a importância relativa dos atributos, a seleção de variáveis e a comparação entre os valores observados e os preditos. Posteriormente, apresenta-se a análise com *Random Forest* quantílico, cuja finalidade é investigar diferentes regiões da distribuição condicional do carbono orgânico, considerando os quantis 0,1, 0,5 e 0,9. Por fim, são analisados os resultados de explicabilidade com SHAP, tanto para o modelo *Random Forest* convencional quanto para a abordagem quantílica. O código[1] foi implementado em Python usando as bibliotecas listadas na Tabela V.1 em uma máquina AMD Ryzen 7 5800X com 8 *Cores*, 32 Gb de RAM e sistema operacional Linux Ubuntu 24.04.4 LTS .

Tabela V.1: Principais bibliotecas usadas na implementação do código em Python. Fonte : O próprio autor

Biblioteca	**Versão**
matplotlib	3.11.0
numpy	2.5.0
pandas	3.0.3
piecewise_regression	1.5.0
quantile_forest	1.4.1
scikit_learn	1.9.0
scikit_optimize	0.10.2
scipy	1.18.0
seaborn	0.13.2
shap	0.49.1

[1] https://github.com/SCICOM-CEFET-RJ/Quantil-Carbon

V.1 Pré-processamento e caracterização dos dados

A Figura V.1 apresenta a distribuição das variáveis explicativas, e do teor de carbono orgânico do solo para cada uso da terra após a aplicação da transformação logarítmica na variável resposta. Essa transformação foi empregada com o objetivo de reduzir a assimetria da variável resposta e tornar a distribuição dos dados mais adequada às etapas de modelagem.

Observa-se que o teor de carbono orgânico apresenta padrões distintos entre os usos da terra. Tradicionalmente , as áreas de floresta tendem a concentrar valores mais elevados de carbono, o que é coerente com a maior presença de matéria orgânica, maior cobertura vegetal e menor intensidade de revolvimento do solo. No entanto, esse fato não foi observado no uso dos dados empregados nesta dissertação.

Os dados com o uso de floresta apresentam curtose e assimetria que mostram uma distribuição de valores mais equilibrada. Isso também é de se esperar em um fluxo de carbono mais homogêneo e estável em uma condição de floresta. Por outro lado, os usos agrícolas e pastagem apresentam distribuições mais dispersas, o que é apontado pela curtose e assimetria dessas distribuições . A heterogeneidade das condições de manejo, das propriedades físicas do solo e dos diferentes contextos ambientais presentes nos registros da base de dados pertencentes a esses usos introduz valores elevados de teor de carbono, o que faz com que os usos agrícola e pastagem tenham o valor central da distribuição superior ao uso floresta. Áreas de lavoura muito produtivas ou áreas de pastagens bem manejadas provocam esse efeito.

Essa variabilidade inicial indica que a estimativa do carbono orgânico do solo não deve ser tratada como um problema homogêneo. As diferenças observadas entre os usos da terra justificam a análise estratificada adotada no estudo e reforçam a necessidade de modelos capazes de capturar relações não lineares, interações entre atributos e variações na dispersão dos dados.

A Tabela V.2 apresenta a quantidade de registros disponíveis por bioma e uso da terra. Verifica-se uma distribuição desigual entre os biomas, com maior concentração de registros na Mata Atlântica, no Cerrado e na Caatinga. Essa desigualdade deve ser considerada na interpretação dos resultados, uma vez que modelos treinados com maior número de observações tendem a representar melhor os padrões presentes nesses grupos.

Também se observa que alguns biomas apresentam baixa representatividade em determinados usos da terra, como o Pantanal, que possui poucos registros e ausência de observações para os usos agrícolas e pastagem. Essa limitação pode afetar a capacidade de generalização dos modelos para contextos pouco representados na base.

A Figura V.2 apresenta as matrizes de correlação e significância estatística entre os atributos utilizados na modelagem. De modo geral, as correlações observadas não indicam multicolinearidade

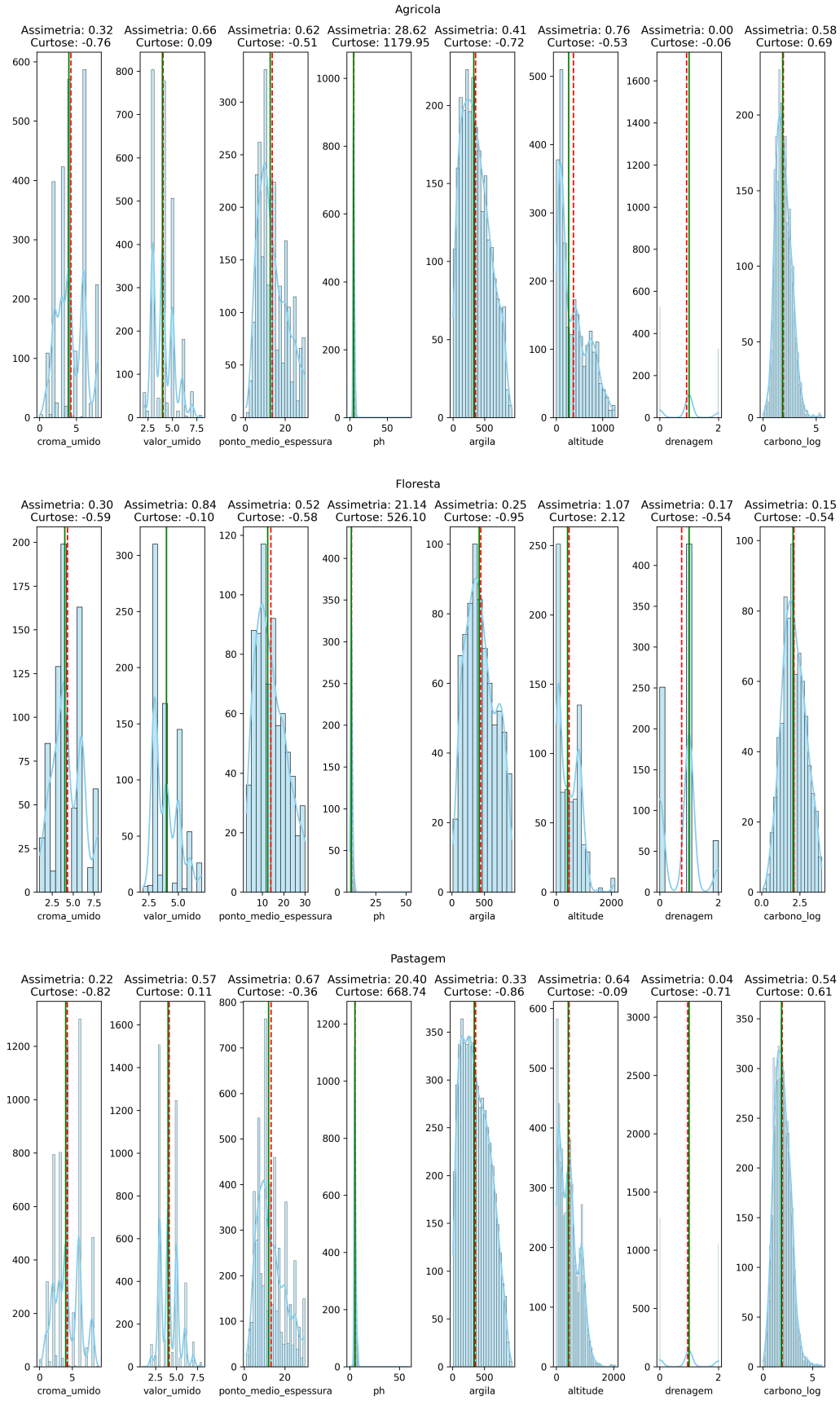


Figura V.1: Distribuição do teor de carbono orgânico por uso da terra. Fonte: O próprio autor

Tabela V.2: Número de registros por bioma e uso da terra.

Bioma	Agrícola	Floresta	Pastagem
Amazônia	63	153	452
Caatinga	1050	38	1203
Cerrado	218	127	1681
Mata Atlântica	1136	405	1886
Pampa	39	12	96
Pantanal	0	5	0

severa entre os atributos selecionados. Esse resultado sugere que as variáveis mantidas representam dimensões distintas das propriedades do solo e podem contribuir de maneira complementar para a predição do carbono orgânico.

A análise de correlação também é importante para orientar a interpretação posterior dos modelos. Embora modelos baseados em árvores sejam menos sensíveis à multicolinearidade do que modelos lineares, a presença de correlações entre atributos pode influenciar medidas de importância e explicações baseadas em SHAP. Dessa forma, a leitura dos resultados deve considerar que os atributos pedológicos frequentemente expressam processos interdependentes.

V.2 Regressão linear múltipla

A regressão linear múltipla foi empregada como modelo de referência para avaliar a relação global entre os atributos selecionados e o teor de carbono orgânico do solo. Essa etapa fornece uma linha de base interpretável, permitindo observar até que ponto uma combinação linear dos atributos diagnósticos é capaz de explicar a variação do carbono.

O modelo foi ajustado com os atributos selecionados na Seção IV.3, conforme a formulação apresentada na Equação II.1. A Figura V.3 apresenta os resultados obtidos para os diferentes usos da terra.

Observa-se que a regressão linear é capaz de capturar parte da tendência geral dos dados, mas apresenta limitações na representação da variabilidade do carbono orgânico, especialmente em faixas superiores da variável resposta. Esse comportamento é coerente com a natureza dos dados pedológicos, nos quais são comuns relações não lineares, interações entre atributos e heterocedasticidade. Assim, a regressão linear cumpre o papel de modelo inicial de comparação, mas não é suficiente para representar toda a complexidade do problema.

V.3 Modelagem do carbono orgânico com Random Forest

Após a análise inicial com regressão linear, foi ajustado o modelo *Random Forest*. A adoção desse método justifica-se por sua capacidade de modelar relações não lineares e interações entre variáveis, características comuns em dados pedológicos. Diferentemente da regressão linear, o *Random Forest*

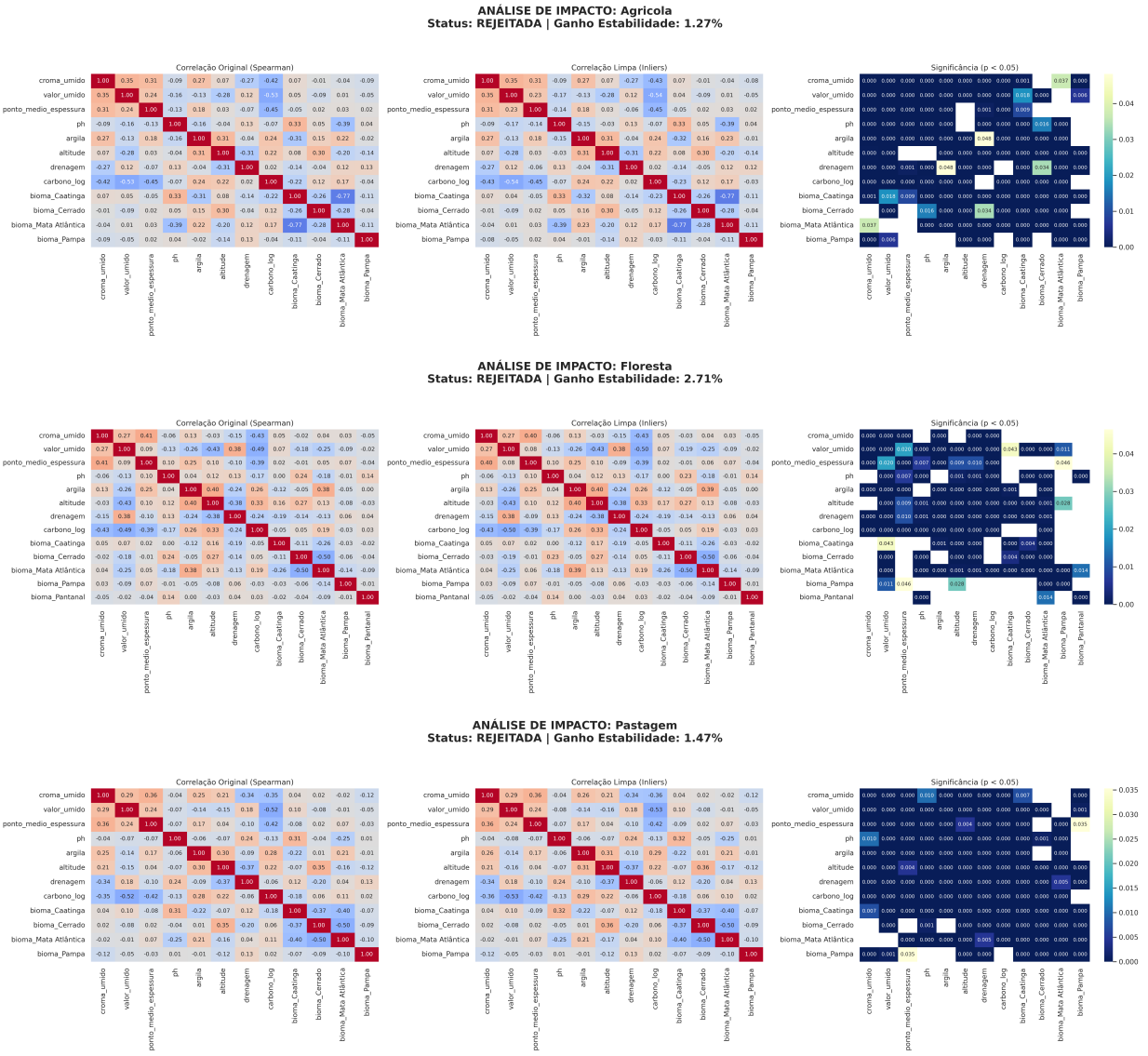


Figura V.2: Matriz de correlação e significância estatística calculadas com o método de Spearman e aproximação assintótica por distribuição t. Fonte: O próprio autor

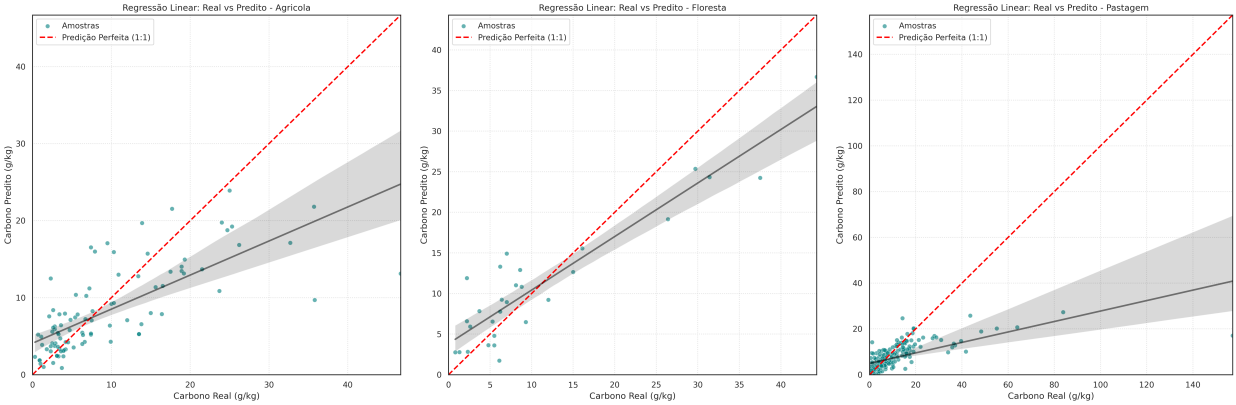


Figura V.3: Regressão linear para o COS usando os atributos selecionados para os vários usos da terra. Fonte: O próprio autor

não impõe uma forma funcional fixa para a relação entre os atributos explicativos e o carbono orgânico do solo.

A Figura V.4 apresenta a importância relativa dos atributos no modelo. Os resultados indicam que variáveis associadas à cor do solo, à profundidade e à textura desempenham papel relevante na predição do carbono orgânico. Esse resultado é coerente com o conhecimento pedológico, uma vez que solos mais escuros tendem a apresentar maior conteúdo de matéria orgânica, enquanto a profundidade está relacionada à distribuição vertical do carbono e a textura influencia a estabilização da matéria orgânica no solo.

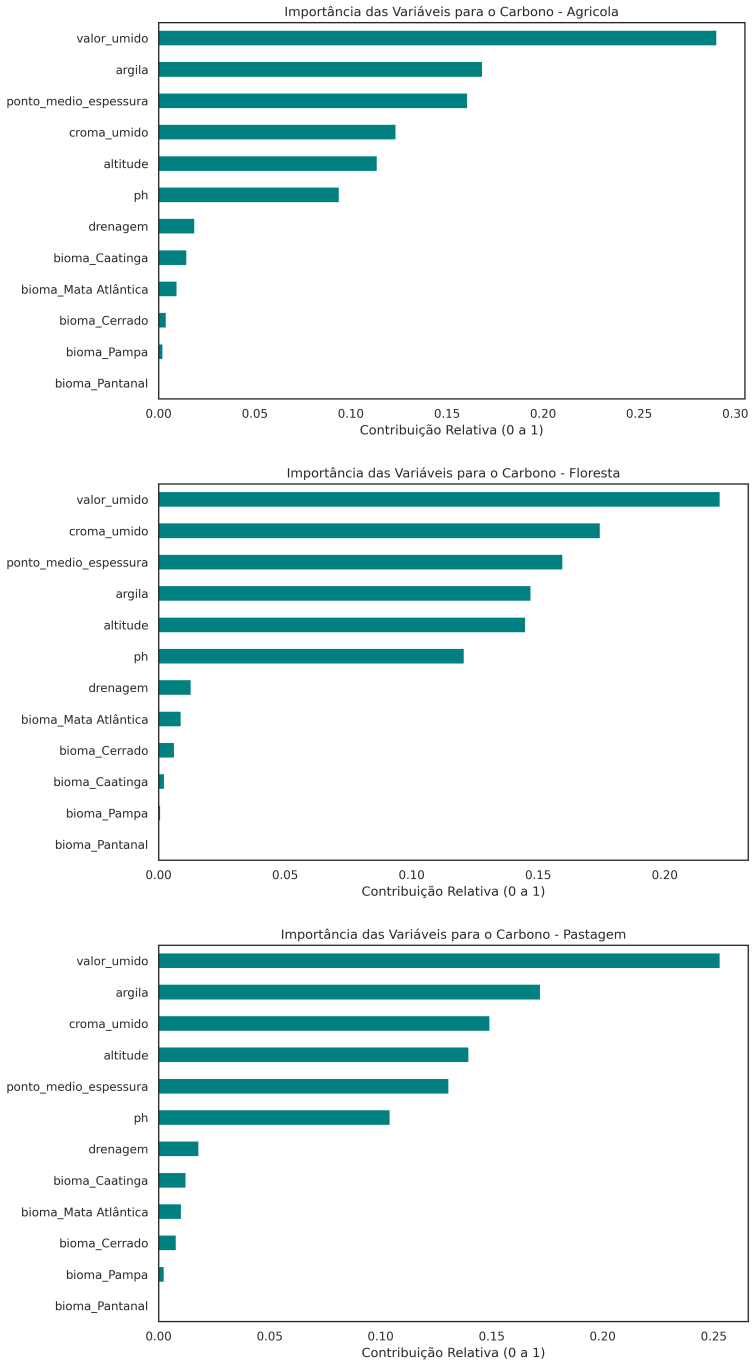


Figura V.4: Importância relativa dos atributos no modelo *Random Forest*. Fonte: O próprio autor

Para refinar o conjunto de atributos utilizado na modelagem, aplicou-se o método de Eliminação Recursiva de Variáveis (RFE) com validação cruzada. A Figura V.5 apresenta o comportamento do erro médio quadrático em função do número de atributos considerados. Essa análise permite identificar um conjunto de variáveis que preserva a capacidade preditiva do modelo, evitando a inclusão de atributos pouco informativos ou redundantes. A seleção de atributos também contribui para melhorar a interpretabilidade dos resultados, pois reduz a complexidade do modelo e favorece uma leitura mais direta das variáveis mais relevantes para a predição do carbono orgânico. A figura V.5 mostra a redução do número de atributos pelo método RFE no conjunto de treinamento do algoritmo de *Random Forest*, e a figura V.4 mostra como as importâncias dos atributos restantes são ordenadas pelo próprio método. Pode-se observar que o atributo categórico bioma, na sua forma codificada, sempre se mantém como a menor importância. Isso revela uma colinearidade nos dados, ou seja, a informação trazida pelo fato de um determinado registro pertencer a um bioma específico, já está expressa nos outros atributos de solo como cor, textura , etc. Além disso, o algortimo de *Random Forest*, devido a sua natureza, prioriza atributos numéricos contínuos em detrimento de atributos categóricos, pois, ele é mais eficiente na divisão dos nós das árvores de classificação criando regras do tipo maior ou menor que , dentro de cada nó, do que regras do tipo pertence ou não pertence.

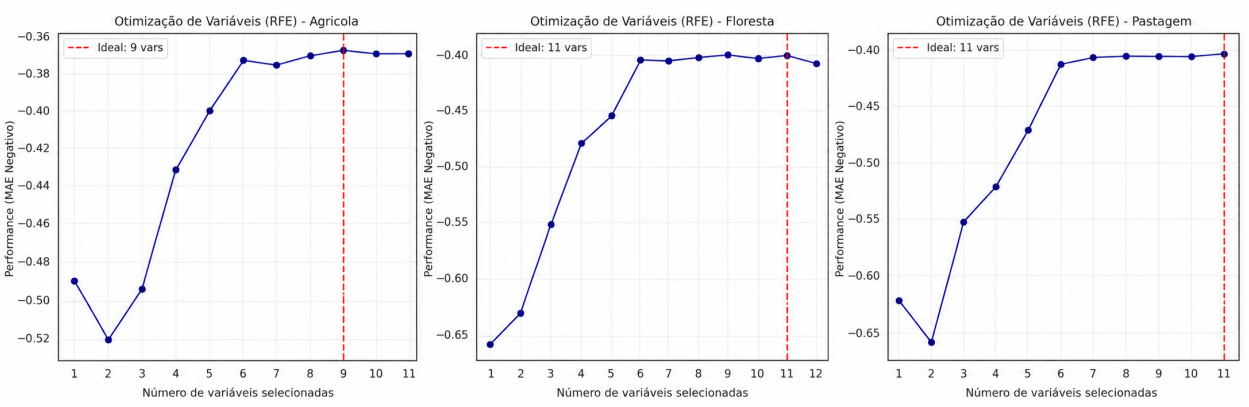


Figura V.5: Otimização do número de atributos por meio de RFE. Fonte: O próprio autor

V.4 Validação do modelo Random Forest

A Figura V.6 apresenta a comparação entre os valores observados e preditos pelo modelo *Random Forest* convencional para os diferentes usos da terra. A linha de referência indica a situação ideal, na qual os valores preditos coincidem com os valores observados.

A avaliação do modelo combina a inspeção visual da relação entre valores observados e preditos com métricas quantitativas de desempenho. A proximidade dos pontos em relação à linha de referência indica maior aderência entre observação e predição, enquanto a dispersão em torno dessa

linha evidencia os erros do modelo.

De modo geral, observa-se que o modelo captura as tendências principais dos dados, especialmente nas faixas intermediárias de carbono orgânico. No entanto, há maior dispersão dos pontos em faixas superiores de carbono, indicando aumento do erro de predição nessas regiões. Esse comportamento sugere a presença de heterocedasticidade, isto é, a variabilidade dos erros não permanece constante ao longo de toda a distribuição do carbono orgânico.

Esse resultado é particularmente importante para a sequência da análise, pois evidencia uma limitação dos modelos que produzem apenas estimativas pontuais. Mesmo quando o modelo apresenta desempenho adequado para valores centrais, ele pode ter maior dificuldade para representar regiões extremas da distribuição. Essa constatação motiva o uso do *Random Forest* quantílico, que permite avaliar não apenas uma predição central, mas diferentes quantis da distribuição condicional.

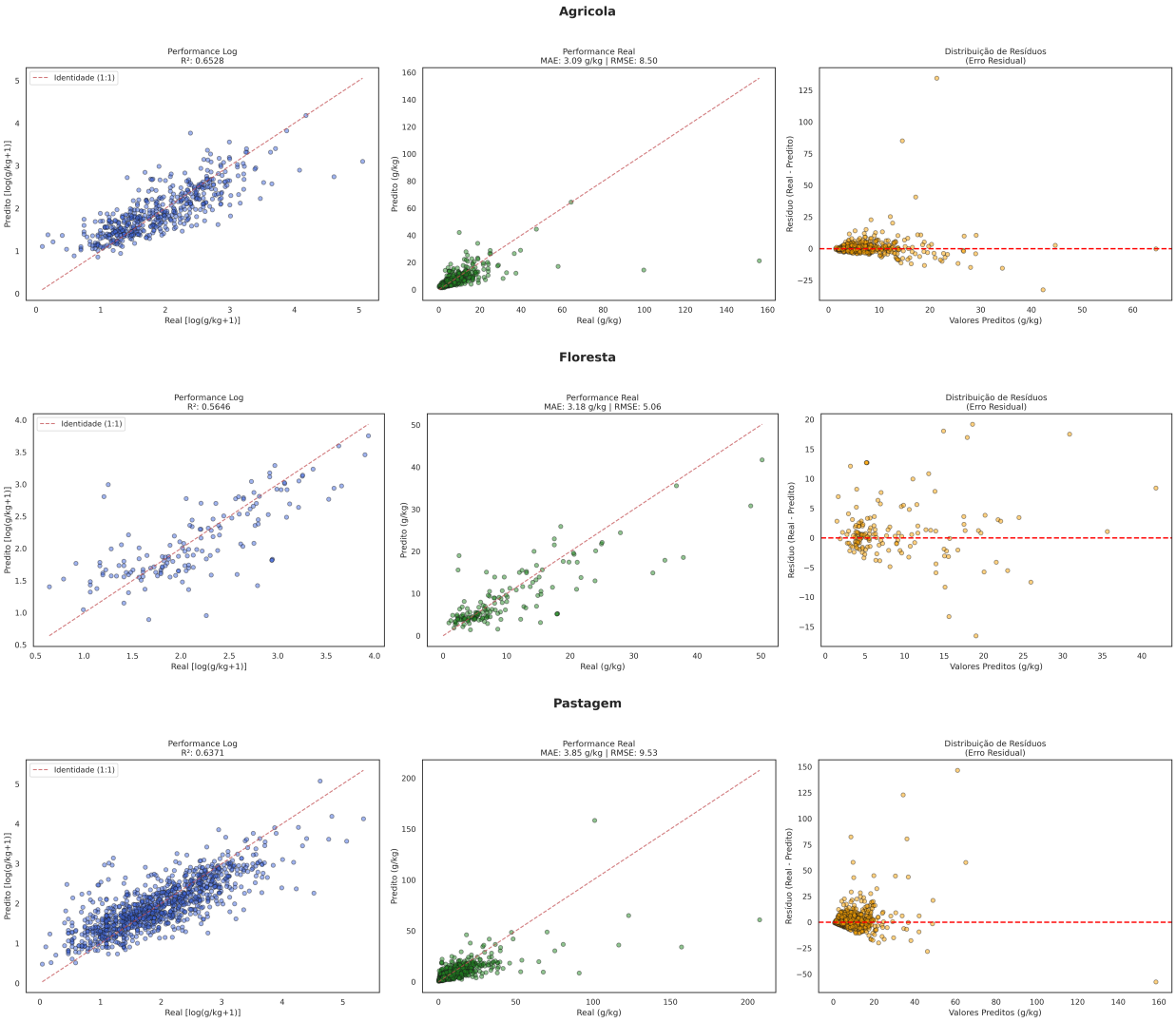


Figura V.6: Comparação entre valores observados e preditos pelo modelo *Random Forest* convencional. Fonte: O próprio autor

V.5 Validação do modelo Random Forest quantílico

Com o objetivo de investigar a variabilidade das predições em diferentes regiões da distribuição do carbono orgânico, foi aplicado o modelo *Random Forest* quantílico. Diferentemente do *Random Forest* convencional, que retorna uma estimativa pontual, a abordagem quantílica permite estimar diferentes quantis da distribuição condicional da variável resposta.

Neste estudo, foram avaliados os quantis 0,1, 0,5 e 0,9. O quantil 0,1 representa uma estimativa associada à região inferior da distribuição; o quantil 0,5 corresponde à mediana; e o quantil 0,9 representa a região superior da distribuição. Dessa forma, a análise permite avaliar não apenas a tendência central, mas também os limites inferiores e superiores esperados para o carbono orgânico do solo em cada uso da terra.

A Figura V.7 apresenta os resultados para o uso agrícola. Observa-se que o quantil 0,5 apresenta melhor alinhamento com a linha de referência, indicando maior capacidade do modelo em representar a tendência central dos dados. Nos quantis 0,1 e 0,9, a dispersão é mais evidente, o que indica maior dificuldade em representar os extremos da distribuição. Ainda assim, a separação entre os quantis fornece uma leitura mais informativa da variabilidade associada às predições.

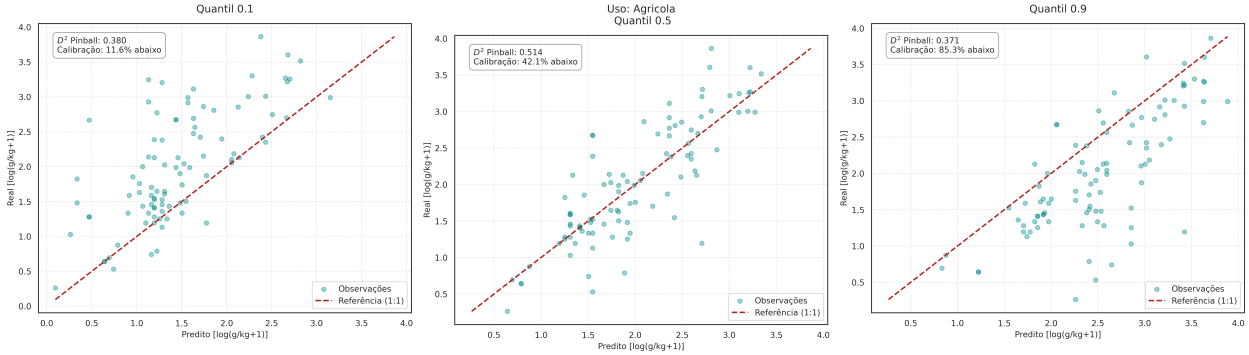


Figura V.7: Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso Agrícola. Fonte: O próprio autor.

A Figura V.8 apresenta os resultados para o uso floresta. Nesse caso, observa-se maior dispersão dos pontos em relação à linha de referência, especialmente nos quantis extremos. O desempenho mais limitado no quantil 0,1, indicado pelo menor valor de D^2 *pinball*, sugere que o modelo teve maior dificuldade em representar a região inferior da distribuição do carbono orgânico em áreas florestais.

Esse comportamento pode estar associado à maior heterogeneidade dos ambientes florestais, ao menor número de observações disponíveis ou à presença de padrões não completamente capturados pelos atributos utilizados. Assim, embora o modelo quantílico forneça uma representação mais rica da incerteza, os resultados para floresta indicam maior instabilidade em comparação aos demais usos da terra.

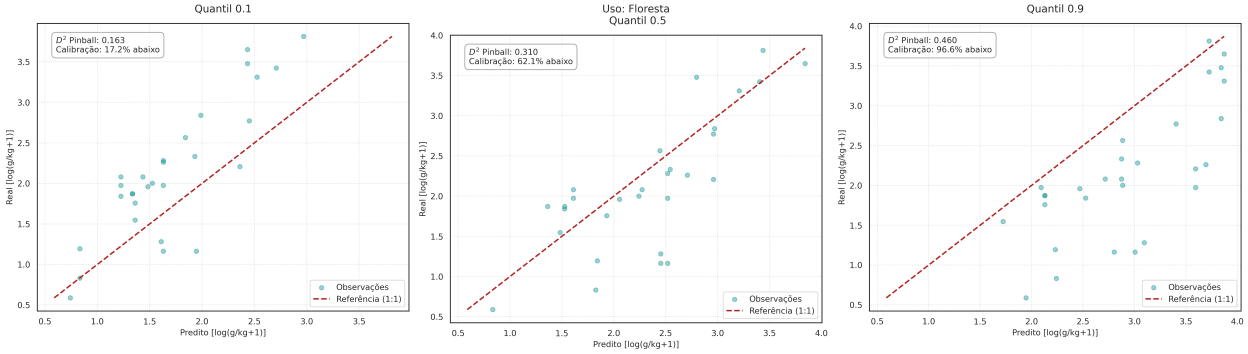


Figura V.8: Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso Floresta. Fonte: O próprio autor.

A Figura V.9 apresenta os resultados para o uso de pastagem. Observa-se um comportamento mais estável, com maior concentração dos pontos próximos à linha de referência, principalmente no quantil 0,5. Os quantis 0,1 e 0,9 também apresentam comportamento coerente, delimitando regiões inferiores e superiores da distribuição prevista.

Entre os usos avaliados, a pastagem apresenta uma resposta visualmente mais consistente do que o modelo quantílico. Esse resultado sugere que, para esse grupo, os atributos disponíveis foram capazes de representar de forma mais adequada a variação do carbono orgânico, tanto na tendência central quanto nas regiões inferiores e superiores da distribuição.

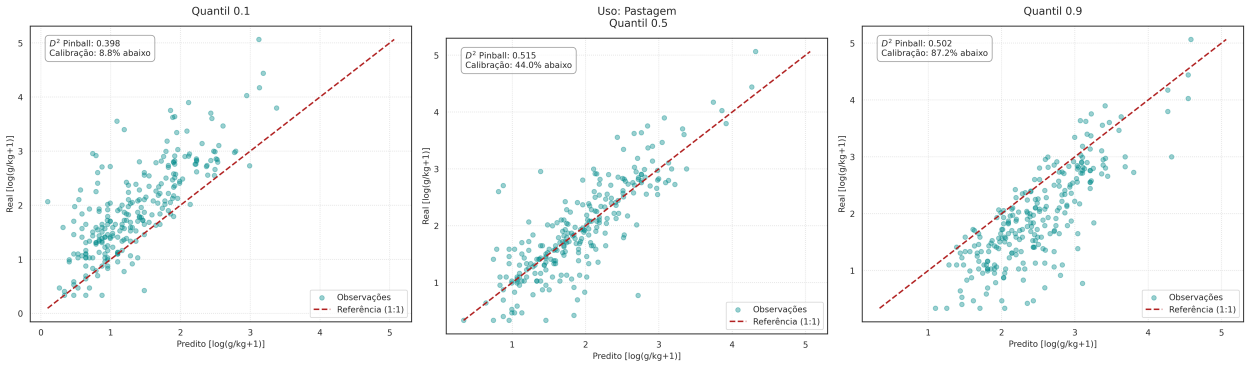


Figura V.9: Comparação entre valores observados e preditos pelo modelo *Random Forest* quantílico para uso em Pastagem. Fonte: O próprio autor.

De forma geral, as Figuras V.7, V.8 e V.9 mostram que a abordagem quantílica acrescenta uma dimensão importante à análise. Enquanto o modelo convencional avalia apenas a proximidade entre valores observados e uma predição pontual, o modelo quantílico permite observar como a incerteza se comporta ao longo da distribuição do carbono orgânico.

Os valores de D^2 *pinball* indicam o ganho preditivo do modelo quantílico em relação a uma previsão quantílica de referência. Valores mais próximos de 1 indicam melhor desempenho relativo, enquanto valores próximos de 0 sugerem menor ganho em relação ao modelo de referência. No entanto, a avaliação do modelo quantílico não deve ser feita apenas por essa métrica. A calibração

dos quantis também é relevante, pois indica se os quantis estimados estão coerentes com a proporção esperada de observações abaixo de cada limite. Assim, um modelo pode apresentar D^2 *pinball* modesto, mas ainda fornecer quantis úteis para representar limites inferiores, medianas e limites superiores da distribuição do carbono orgânico.

V.6 Análise explicável dos modelos

Para interpretar o comportamento dos modelos, foi realizada uma análise explicável com base em valores SHAP. Essa etapa permite avaliar a contribuição relativa dos atributos nas predições e verificar se os padrões aprendidos pelos modelos são coerentes com o conhecimento pedológico.

A Figura V.10 apresenta a importância global dos atributos no modelo *Random Forest* convencional. Observa-se que atributos relacionados à cor do solo, à profundidade e à textura aparecem entre os mais relevantes. Esse resultado reforça a interpretação pedológica dos modelos, pois tais atributos estão diretamente relacionados à presença, à distribuição e à estabilização do carbono orgânico no solo.

Cabe destacar que os valores SHAP explicam o comportamento do modelo ajustado e não devem ser interpretados como evidência causal direta dos processos pedológicos. Assim, a análise deve ser entendida como uma forma de identificar quais atributos contribuíram mais para as predições realizadas pelo modelo.

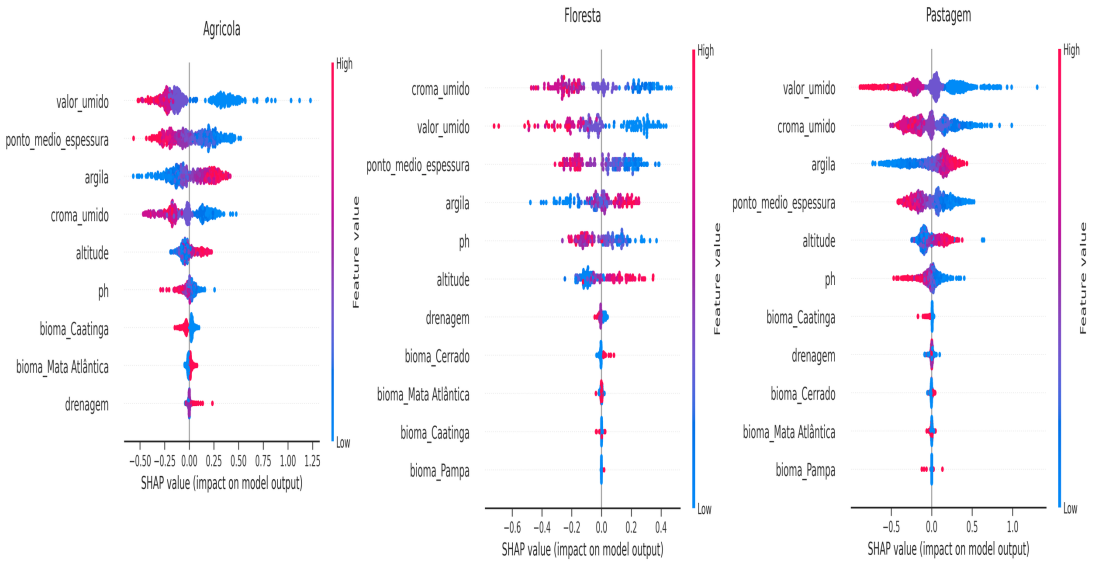


Figura V.10: Importância global dos atributos segundo valores SHAP para o método *Random Forest* convencional. Fonte: O próprio autor.

Além da importância global, os valores SHAP permitem analisar a direção do efeito dos atributos sobre as predições. Em geral, menores valores de croma e de valor na escala Munsell estão associados a maiores estimativas de carbono orgânico. Esse comportamento é coerente com a interpretação

morfológica dos solos, segundo a qual solos mais escuros tendem a apresentar maior teor de matéria orgânica.

A Figura V.11 apresenta a importância global dos atributos para o modelo *Random Forest* quantílico. A comparação com o modelo convencional indica que a mudança para a abordagem quantílica não altera substancialmente o conjunto de atributos mais relevantes. Esse resultado sugere consistência entre os modelos e reforça que as variáveis associadas à cor (valor_umido e croma_umido) , profundidade (ponto_medio_espessura) e textura (argila) continuam sendo importantes mesmo quando a análise passa a considerar diferentes regiões da distribuição do carbono orgânico.

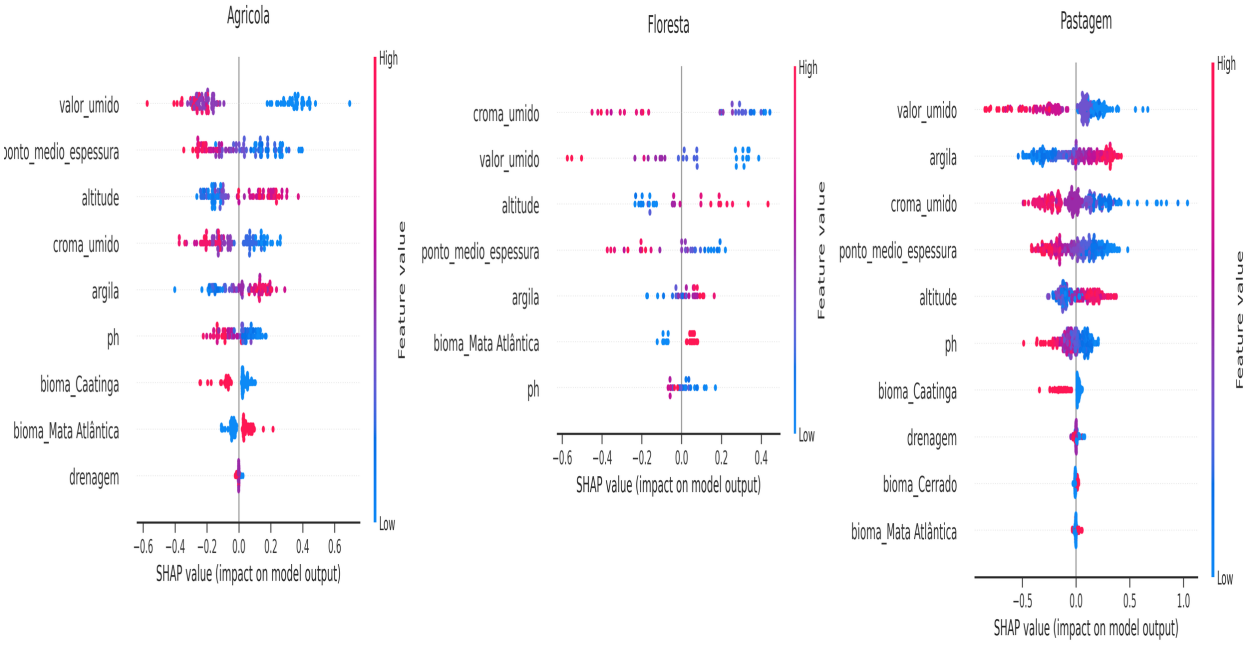


Figura V.11: Importância global dos atributos segundo os valores de SHAP para o método *Random Forest* quantílico. Fonte: O próprio autor.

A Figura V.12 apresenta a importância dos atributos em cada quantil. Essa análise complementa a importância global, pois permite observar se a contribuição dos atributos varia entre a região inferior, a mediana e a região superior da distribuição do carbono orgânico.

Os resultados indicam que a importância dos atributos não é necessariamente constante entre os quantis. Além disso, observa-se que o padrão de importância também difere entre os usos da terra. Isso indica que os fatores associados à predição do carbono orgânico não são uniformes entre agricultura, floresta e pastagem, reforçando a pertinência da análise estratificada.

Essa variação é coerente com a presença de heterocedasticidade nos dados, pois diferentes faixas de carbono orgânico podem ser influenciadas com intensidades distintas pelos mesmos atributos. Assim, a análise por quantis contribui para uma interpretação mais detalhada do modelo, mostrando que os fatores associados a valores baixos, medianos e altos de carbono podem não atuar da mesma forma.

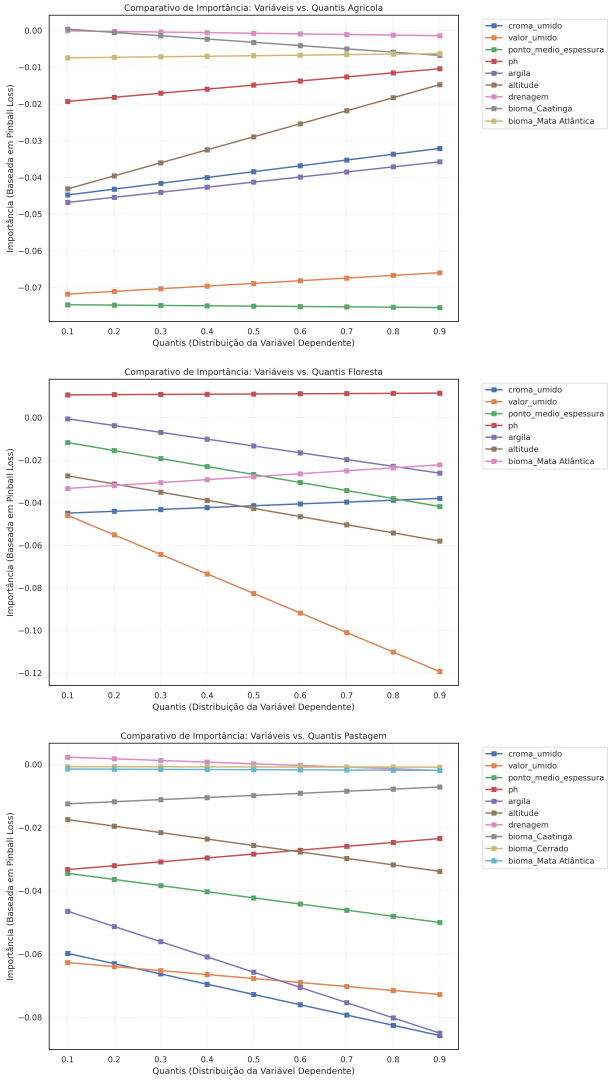


Figura V.12: Importância dos atributos em cada quantil. Fonte: O próprio autor

V.7 Síntese dos resultados

Os resultados apresentados neste capítulo mostram que a estimativa do carbono orgânico do solo é marcada por elevada variabilidade e por diferenças relevantes entre os usos da terra. A análise exploratória indicou distribuições assimétricas, heterogeneidade entre biomas e padrões distintos para agricultura, floresta e pastagem. Esses aspectos justificam o uso de modelos capazes de lidar com relações não lineares e com variação não constante dos erros.

A regressão linear múltipla permitiu estabelecer uma linha de base interpretável, mas apresentou limitações para representar a complexidade dos dados. O modelo *Random Forest* convencional avançou em relação a essa abordagem ao capturar interações e relações não lineares entre os atributos pedológicos. Ainda assim, a validação mostrou maior dispersão dos erros em faixas superiores de carbono, indicando que a predição pontual não é suficiente para representar toda a variabilidade do fenômeno.

O modelo *Random Forest* quantílico ampliou a análise ao estimar diferentes regiões da distribuição condicional do carbono orgânico. As Figuras V.7, V.8 e V.9 mostram que os quantis 0,1, 0,5 e 0,9 permitem representar, respectivamente, faixas inferiores, medianas e superiores da resposta. A pastagem apresentou comportamento mais estável, enquanto a floresta apresentou maior dispersão e menor desempenho no quantil inferior. Esses resultados indicam que a qualidade da predição quantílica depende do uso da terra e da estrutura dos dados disponíveis para cada grupo.

A análise explicável com SHAP confirmou a relevância de atributos relacionados à cor, profundidade e textura do solo. Esses resultados são coerentes com o conhecimento pedológico, uma vez que a cor do solo expressa indiretamente a presença de matéria orgânica, a profundidade representa a variação vertical do carbono e a textura influencia a capacidade de estabilização da matéria orgânica. Ao mesmo tempo, os resultados devem ser interpretados como explicações do comportamento dos modelos ajustados, e não como evidência causal direta.

A análise por quantis mostrou ainda que a importância dos atributos pode variar ao longo da distribuição e entre os usos da terra. Esse resultado reforça a utilidade da abordagem quantílica para compreender a heterogeneidade dos dados e para avaliar de forma mais detalhada a contribuição dos atributos na predição de diferentes faixas de carbono orgânico.

Em conjunto, os resultados demonstram que a integração entre regressão linear, *Random Forest*, *Random Forest* quantílico e SHAP constitui uma estratégia adequada para estimar e interpretar o carbono orgânico do solo. Essa combinação permite avaliar tendências gerais, capturar relações não lineares, representar a incerteza associada às predições e interpretar os fatores pedológicos mais relevantes para o comportamento do modelo.

O *Random Forest* quantílico permite uma estimativa mais útil, em termos de piso e teto, do teor de carbono no solo. Se, por exemplo, for adotado um piso de 10 % de carbono no solo, o que equivale à previsão do quantil 0,1, tem-se uma estimativa de 88,4 % de confiança de que 10 % é o piso da concentração de carbono para um uso agrícola. Para os usos de floresta e de pastagem, essa confiança é de 82,8 % e 91,2 % respectivamente . Um raciocínio análogo pode ser empregado para o teto da estimativa usando o quantil 0,9. Nesse caso, as confianças são de 85,3 % para o uso agrícola , 96,6 % para o uso de floresta e 87,2 % para o uso de pastagem . Esses valores derivam da calibração do *Random Forest* quantílico mostrada nas Figuras V.7, V.8 e V.9.

Capítulo VI Considerações Finais

Esta dissertação teve como objetivo investigar a relação entre atributos diagnósticos do solo e o teor de carbono orgânico do solo (COS), integrando regressão linear múltipla, *Random Forest* convencional, *Random Forest* quantílico e técnicas de Inteligência Artificial Explicável baseadas em SHAP. A proposta central consistiu em avaliar a estimativa do COS a partir de atributos pedológicos e em interpretar a contribuição desses atributos nos modelos ajustados, considerando diferentes usos da terra.

Os resultados obtidos evidenciaram que a estimativa do COS é um problema marcado por elevada variabilidade, assimetria na distribuição dos dados e presença de heterocedasticidade. A etapa de caracterização dos dados mostrou diferenças relevantes entre agricultura, floresta e pastagem, justificando a análise estratificada adotada ao longo do estudo. Essa segmentação permitiu avaliar o comportamento dos modelos em contextos distintos de uso da terra, reduzindo parte da heterogeneidade observada no conjunto de dados.

A regressão linear múltipla foi utilizada como modelo de referência para avaliar a relação global entre os atributos selecionados e o teor de carbono orgânico. Embora tenha permitido capturar parte da tendência geral dos dados, esse modelo apresentou limitações para representar a complexidade do problema, especialmente em faixas superiores de carbono. Esse resultado reforça que relações lineares globais são úteis como linha de base, mas não são suficientes para descrever integralmente a variabilidade do COS em dados pedológicos heterogêneos.

O modelo *Random Forest* convencional representou um avanço em relação à regressão linear múltipla, pois permitiu capturar relações não lineares e interações entre os atributos do solo. Os resultados indicaram que o modelo foi capaz de representar tendências gerais presentes nos dados, sobretudo nas faixas intermediárias de carbono orgânico. No entanto, a maior dispersão dos erros em faixas superiores da variável resposta indicou que a predição pontual apresenta limitações para representar toda a variabilidade do COS.

A aplicação do *Random Forest* quantílico ampliou a análise ao considerar diferentes regiões da distribuição condicional do carbono orgânico. Os quantis 0,1, 0,5 e 0,9 permitiram representar, respectivamente, faixas inferiores, medianas e superiores da resposta. Essa abordagem mostrou-se relevante porque a variabilidade do COS não se distribui de maneira uniforme entre os usos da terra nem ao longo de toda a faixa de valores observados.

Entre os usos avaliados, a pastagem apresentou comportamento mais estável no modelo quantílico, com maior proximidade visual entre valores observados e preditos, especialmente no quantil mediano. O uso agrícola apresentou desempenho intermediário, com representação razoável da tendência central, mas maior dispersão nos quantis extremos. Já o uso floresta apresentou maior instabilidade, principalmente no quantil inferior, indicando maior dificuldade do modelo em representar essa região da distribuição do carbono orgânico.

Os valores de D^2 *pinball* indicaram ganhos preditivos distintos entre usos da terra e quantis. Assim, os resultados sugerem que o desempenho do modelo quantílico depende tanto da região da distribuição analisada quanto da estrutura dos dados em cada uso da terra. Essa constatação reforça a importância de avaliar modelos quantílicos de forma estratificada, evitando interpretações agregadas que possam ocultar diferenças relevantes entre grupos.

A análise de importância dos atributos e os resultados obtidos com SHAP indicaram a relevância de variáveis relacionadas à cor do solo, à profundidade e à textura na predição do COS. Esse resultado é coerente com o conhecimento pedológico, uma vez que a cor do solo expressa indiretamente a presença de matéria orgânica, a profundidade representa a variação vertical do carbono e a textura influencia a estabilização da matéria orgânica. Ao mesmo tempo, os resultados mostraram que a importância dos atributos pode variar entre usos da terra e entre diferentes quantis da distribuição.

Cabe destacar que os valores SHAP foram utilizados como ferramenta de interpretação do comportamento dos modelos ajustados. Portanto, seus resultados não devem ser interpretados como evidência causal direta dos processos pedológicos que controlam o carbono orgânico do solo. Ainda assim, a análise contribuiu para verificar se os padrões aprendidos pelos modelos são coerentes com interpretações pedológicas esperadas.

Uma contribuição central desta dissertação foi propor uma estrutura analítica integrada para estimar e interpretar o teor de carbono orgânico do solo. Nessa estrutura, a regressão linear múltipla atuou como linha de base interpretável; o *Random Forest* convencional permitiu modelar relações não lineares; o *Random Forest* quantílico possibilitou analisar diferentes regiões da distribuição condicional da resposta; e o SHAP contribuiu para interpretar a influência dos atributos nos modelos ajustados.

Além da contribuição metodológica, esta pesquisa também gerou produção científica associada ao tema da dissertação. Como resultado parcial do trabalho desenvolvido, foi publicado o artigo intitulado *"Quantile Regression for Soil Organic Carbon: A Distributional Analysis of Soil Variability"* no evento *2026 International Conference on Computational Science and Its Applications*. Essa publicação explicita a relevância da perspectiva distribucional adotada na pesquisa, especialmente para a análise da variabilidade do carbono orgânico do solo e da heterocedasticidade presente nos dados pedológicos.

Entre as limitações deste estudo, destaca-se a dependência da qualidade, da completude e da representatividade dos dados disponíveis. A distribuição desigual de registros entre biomas e usos da terra pode influenciar o desempenho dos modelos, especialmente em grupos com menor número de observações. Além disso, a ausência de algumas variáveis potencialmente relevantes, como atributos climáticos, informações detalhadas de manejo, dados temporais e produtos derivados de sensoriamento remoto, pode ter limitado a capacidade dos modelos de representar parte da variabilidade observada.

Também se reconhece que a análise foi conduzida a partir de associações estatísticas e padrões aprendidos pelos modelos. Embora os resultados sejam coerentes com o conhecimento pedológico, eles não demonstram causalidade entre os atributos analisados e o teor de carbono orgânico. Essa limitação é própria de estudos baseados em dados observacionais e deve ser considerada na interpretação dos resultados.

Apesar dessas limitações, os resultados indicam que a combinação entre modelos baseados em florestas aleatórias, abordagem quantílica e técnicas de explicabilidade constitui uma estratégia adequada para a estimativa e interpretação do teor de carbono orgânico do solo. Essa integração permitiu avaliar tendências gerais, representar diferentes faixas da distribuição da resposta e interpretar a contribuição dos atributos em diferentes contextos de uso da terra.

Como trabalhos futuros, recomenda-se incorporar novas covariáveis ambientais, especialmente dados climáticos, atributos de relevo em maior resolução, informações provenientes de sensoriamento remoto e variáveis associadas ao manejo do solo. Também se recomenda ampliar a validação dos modelos em bases independentes, de modo a avaliar sua capacidade de generalização em diferentes regiões, biomas e condições pedológicas.

Além disso, futuras pesquisas podem aprofundar a comparação entre *Random Forest* convencional, *Random Forest* quantílico e outras abordagens capazes de lidar com heterocedasticidade. Também seria relevante explorar estratégias de explicabilidade específicas por uso da terra e por faixa de carbono orgânico, permitindo compreender com maior detalhe como os atributos pedológicos contribuem para a predição em diferentes contextos.

Em síntese, esta dissertação demonstra que a integração entre aprendizado de máquina, modelagem quantílica baseada em florestas aleatórias e Inteligência Artificial Explicável oferece uma abordagem consistente para estimar e interpretar o teor de carbono orgânico do solo. Ao combinar desempenho preditivo, análise de diferentes quantis e interpretação dos atributos, o trabalho contribui para o avanço de metodologias computacionais aplicadas à análise de dados pedológicos heterogêneos.

Referências

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges towards responsible AI. *Information Fusion*, 58:82–115, 2020.

Baake, K., Heuvelink, G., and de Bruin, S. Quantifying uncertainty of random forest predictions. *Computers & Geosciences*, 2018.

Baltensweiler, A., Walthert, L., Hanewinkel, M., Zimmermann, S., and Nussbaum, M. Machine learning based soil maps for a wide range of soil properties for the forested area of switzerland. *Geoderma Regional*, 2021.

Baveye, P. The 4p1000 initiative: Opportunities, limitations and challenges for implementing soil organic carbon sequestration as a sustainable development strategy. *Soil Security*, 2021.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. LOF: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 93–104, New York, NY, USA. ACM, 2000.

Bryan, B., Ordóñez Echeverría, A., Echeverría Guadalupe, M., and Toulkeridis, T. Integrating machine learning models for enhanced soil organic carbon estimation: A multi-model fusion approach. *Fusion: Practice and Applications*, 2024.

Carlsen, L. and Bruggemann, R. The 17 united nations' sustainable development goals: A status by 2020. *International Journal of Sustainable Development & World Ecology*, 29(3):219–229, 2022.

Castaldi, F., Koparan, M. H., Wetterlind, J., Žydelis, R., Vinci, I., Savaş, A. , and Kıvrak, C. e. a. Assessing the capability of sentinel-2 time-series to estimate soil organic carbon and clay content at local scale in croplands. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2023.

Cheng, L. et al. Interpretable digital soil organic matter mapping based on ggp-gam. *Agriculture*, 2024.

Cotrufo, M. F., Wallenstein, M. D., Boot, C. M., Denef, K., and Paul, E. A. The microbial efficiency-matrix stabilization (mems) framework integrates microbial metabolic efficiency and physiological traits to predict soil organic matter stabilization. *Global Change Biology*, 19(4):988–995, 2013.

Davidson, E. A. and Janssens, I. A. Temperature sensitivity of soil carbon decomposition and feedbacks to climate change. *Nature*, 440(7081):165–173, 2006.

Dharumarajan, S., Kalaiselvi, B., Suputhra, A., Lalitha, M., Hegde, R., Singh, S. K., and Lagacherie, P. Digital soil mapping of key globalsoilmap properties in northern karnataka plateau. *Geoderma*, 2020.

Ding, L., Zhang, X., Wu, B., Chen, H., Luo, X., Lu, L., and Zhang, J. Moss powder amendment modifies organic and inorganic carbon of saline-alkaline soil: Insights from conventional statistics and artificial intelligence. *Polish Journal of Environmental Studies*, 2025.

Dong, Y. et al. Enhancing soil organic carbon prediction using interpretable machine learning. *Geoderma*, 2025.

Durante, P., Guevara, M., Vargas, R., and Oyonarte, C. Predicting soil organic carbon with different approaches and spatial resolutions for the southern iberian peninsula, spain. *Geoderma Regional*, 2024.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., and Scholten, T. Predicting and mapping of soil organic carbon using machine learning algorithms in northern iran. *Remote Sensing*, 12(14), 2020.

EMBRAPA HORTALIÇAS. Ciclo do Carbono em sistemas conservacionistas. 1 ilustração, 2018.

EMBRAPA SOLOS. Banco de dados de solos - BD Solos, 2014.

Gavilán-Acuña, G., Coops, N., Olmedo, G. F., Tompalski, P., Roeser, D., and Varhola, A. Assessing soil prediction distributions for forest management using digital soil mapping. *Soil Systems*, 2024.

Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(495):746–762, 2011.

Grimm, R., Behrens, T., Märker, M., and Elshereif, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forest analysis. *Geoderma*, 146(1–2):102–113, 2008.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1):389–422, 2002.

Haithem, B., Raclot, D., Barbouchi, M., Lagacherie, P., and Mohamed, A. Mapping soil organic carbon stocks in tunisian topsoils. *Geoderma Regional*, 2022.

Hasanawi, A. and Purwanto, S. Analyze qrf model for soil organic carbon map building with digital soil mapping (case study: Sumatra island and java island). *IOP Conference Series: Earth and Environmental Science*, 2024.

Helfenstein, A., Mulder, V. L., Heuvelink, G., and Okx, J. Bis-3d: High resolution 3d soil maps for the netherlands using accuracy thresholds. Technical report, Wageningen University, 2021.

Heuvelink, G. B. M., Angelini, M., Poggio, L., Bai, Z., Batjes, N., van den Bosch, R., and Bossio, D. e. a. Machine learning in space and time for modelling soil organic carbon change. *European Journal of Soil Science*, 2020.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R.* Springer, New York, 2 edition, 2021.

Jobbágy, E. G. and Jackson, R. B. The vertical distribution of soil organic carbon and its relation to vegetation and climate. *Ecological Applications*, 10(2):423–436, 2000.

Khosravi, A., Nahavandi, S., Creighton, D., and Atariya, S. F. Lower Upper Bound Estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2011.

Koenker, R. *Quantile Regression.* Cambridge University Press, New York, 2005.

Lal, R. Soil carbon management and climate change. *Carbon Management*, 4(4):439–462, 2013.

Lal, R. Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. *Global change biology*, 24(8):3285–3301, 2018.

Lamichhane, S., Adhikari, K., and Kumar, L. Use of multi-seasonal satellite images to predict soil organic carbon from cultivated lands in a montane ecosystem. *Remote Sensing*, 2021.

Lee, W. and Lee, J. Tree-based modeling for large-scale management in agriculture: Explaining organic matter content in soil. *Applied Sciences*, 2024.

Liang, Z., Chen, S., Yang, Y., Ruiying, Z., Shi, Z., and Viscarra Rossel, R. Baseline map of soil organic matter in china and its associated uncertainty. *Geoderma*, 335:47–56, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 4765–4774. Curran Associates, Inc., 2017.

Masobeng, T., Mapeshoane, B. E., Marake, M., Ramakhanna, S., Nkunyane, K., Motšoane, M., and Moleleki, M. e. a. Spatio-temporal modelling of soil organic carbon stock change in relation to land use changes in selected sub-catchments of lesotho. *Soil Advances*, 2025.

McBratney, A. et al. On digital soil mapping. *Geoderma*, 117:3–52, 2003.

McFadden, D. Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York, 1974.

McNicol, G., Bulmer, C., D'Amore, D. V., Sanborn, P., Saunders, S. C., Giesbrecht, I., Arriola, S. G., Bidlack, A., Butman, D., and Buma, B. Large, climate-sensitive soil carbon stocks mapped with pedology-informed machine learning in the north pacific coastal temperate rainforest. *Environmental Research Letters*, 2019.

Meinshausen, N. and Ridgeway, G. Quantile regression forests. *Journal of machine learning research*, 7(6), 2006.

Mendes, W. d. S., Roudier, P., Mudge, P., McNeill, S., and Goodrich, J. P. Comparison of different approaches for estimating profile soil organic carbon stocks from topsoil samples. *Geoderma*, 2025.

Müller, A. C. and Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, Inc., Sebastopol, CA, 2016.

Nelder, J. A. and Wedderburn, R. W. M. Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Nikou, M. and Tziachris, P. Prediction and uncertainty capabilities of quantile regression forests in estimating spatial distribution of soil organic matter. *ISPRS International Journal of Geo-Information*, 2022.

Padarian, J., McBratney, A., and Minasny, B. Game theory interpretation of digital soil mapping convolutional neural networks. *The Soil*, 6:389–400, 2020.

Pardelinha, A., Ceddia, M. B., Gervasio, R., Belloze, K., Aguilar, C. d. L., Assis, L., and Brandao, D. Aplicaçao de técnicas de aprendizado de máquina na determinaçao de estoque de carbono no solo. In *Simpósio Brasileiro de Banco de Dados (SBBD)*, pages 673–679. SBC, 2024.

Pearson, K. *Mathematical contributions to the theory of evolution. XIV. On the general theory of skew correlation and non-linear regression*, volume 2 of *Studies in National Deterioration*. Dulau and Co., London, 1905.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rawls, W. J. Estimating soil bulk density from particle size analysis and organic matter content. *Soil Science*, 135(2):123–125, 1983.

Rodrigues, C. I. D., Brito, L. M., and Nunes, L. J. Soil carbon sequestration in the context of climate change mitigation: A review. *Soil Systems*, 7(3):64, 2023.

Salgado, L., Forján, R., Rodríguez-Pérez, J. R., Colina, A., Mejía-Correal, K. B., López-Sánchez, C., and Gallego, J. R. Soil organic carbon fractionation assessment in areas with high fire activity using diffuse spectroscopy and tree-based machine learning algorithms. *Earth Systems and Environment*, 2025.

Santos, C. M. M. d. *Inteligência Artificial Aplicada à Dimensão Pedológica Global Para Modelar, Mapear e Compreender Interações Biota-Solo-Clima e Dinâmicas do Carbono Orgânico*. PhD thesis, Universidade, n.d.

Santos, H. G. d., Jacomine, P. K. T., Anjos, L. H. C. d., Oliveira, V. d., Lumbreras, J. F., Coelho, M. R., Almeida, J. A. d., Araújo Filho, J. C. d., Oliveira, J. B. d., and Cunha, T. J. F. *Sistema Brasileiro de Classificação de Solos*. Embrapa, Brasília, DF, 5 edition, 2018.

Schulze, D. G., Nagel, J. L., Van Scoyoc, G. E., Henderson, T. L., Baumgardner, M. F., and Stott, D. E. Significance of soil organic matter in determining soil color. In *Soil Color*, pages 71–90. Soil Science Society of America, Madison, WI, 1993.

Shapley, L. S. A value for n-person games. In Kuhn, H. W. and Tucker, A. W., editors, *Contributions to the Theory of Games (AM-28), Volume II*, pages 307–317. Princeton University Press, 1953.

Shi, L., Daly, K., and O'Rourke, S. Estimating mineral-associated organic carbon saturation and sequestration potential using mir spectral based local quantile regression. *Geoderma*, 2025.

Six, J., Conant, R. T., Paul, E. A., and Paustian, K. Stabilization mechanisms of soil organic matter: Implications for c-saturation of soils. *Plant and Soil*, 241(2):155–176, 2002.

Somarathna, P., Minasny, B., and Malone, B. P. More data or a better model? figuring out what matters most for the spatial prediction of soil carbon. *Soil Science Society of America Journal*, 81(6):1413–1426, 2017.

Souza, C. M., Z. Shimbo, J., Rosa, M. R., Parente, L. L., A. Alencar, A., Rudorff, B. F. T., Hasenack, H., Matsumoto, M., G. Ferreira, L., Souza-Filho, P. W. M., de Oliveira, S. W., Rocha, W. F., Fonseca, A. V., Marques, C. B., Diniz, C. G., Costa, D., Monteiro, D., Rosa, E. R., Vélez-Martin, E., Weber, E. J., Lenti, F. E. B., Paternost, F. F., Pareyn, F. G. C., Siqueira, J. V., Viera, J. L., Neto, L. C. F., Saraiva, M. M., Sales, M. H., Salgado, M. P. G., Vasconcelos, R., Galano, S., Mesquita, V. V., and Azevedo, T. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17), 2020.

Szatmári, G., Pirkó, B., Koós, S., Laborczi, A., Bakacsi, Z., Szabó, J., and Pásztor, L. Spatio-temporal assessment of topsoil organic carbon stock change in hungary. *Geoderma*, 2019.

Szatmári, G. and Pásztor, L. Comparison of uncertainty quantification methods on the example of soil organic carbon stock mapping in hungary. *Geoderma*, 2020.

Szatmári, G., Pásztor, L., and Heuvelink, G. B. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. *Geoderma*, 403:115356, 2021.

Teixeira, P. C., Donagemma, G. K., Fontana, A., and Teixeira, W. G. *Manual de métodos de análise de solo*. Embrapa Solos, Brasília, DF, 3. rev. e ampl. edition, 2017.

Teleki, M., Marake, M., Mapeshoane, B. E., Motšoane, M., Moleleki, M., Nkunyane, K., and Lekokotoana, M. e. a. Spatial distribution of soil organic carbon along a toposequence using quantile regression forest in lesotho. *Applied and Environmental Soil Science*, 2025.

Theres, B. L. and Sakthivel, R. Prediction of soil properties using quantile regression forest machine learning algorithm – a case study of salem and rasipuram block, tamil nadu, india. *International Journal of Environment and Climate Change*, 2022a.

Theres, B. L. and Sakthivel, R. Prediction of soil properties using quantile regression forest machine learning algorithm – a case study of salem and rasipuram block, tamil nadu, india. *International Journal of Environment and Climate Change*, 2022b.

Tomasella, J. and Hodnett, M. G. Estimating soil water retention characteristics from limited data in the Brazilian Amazon. *Soil Science*, 163(3):190–202, 1998.

Uduwerella, H., Iddawela, A. U., and Kadupitiya, H. K. R-modelling-based quantile regression approach for spatial prediction and mapping soil properties in sri lanka. *Tropical Agriculture*, 2023.

Ugbaje, S. U., Karunaratne, S., Bishop, T. F. A., Gregory, L., Searle, R., Coelli, K., and Farrell, M. Space-time mapping of soil organic carbon stock and its local drivers: Potential for use in carbon accounting. *Geoderma*, 2024.

Urbina-Salazar, D., Vaudour, E., Richer-de Forges, A., Chen, S., Martelet, G., Baghdadi, N., and Arrouays, D. Sentinel-2 and sentinel-1 bare soil temporal mosaics of 6-year periods for soil organic carbon content mapping in central france. *Remote Sensing*, 2023.

van der Westhuizen, S., Heuvelink, G., and Hofmeyr, D. Multivariate random forest for digital soil mapping. *SSRN Electronic Journal*, 2023.

Viscarra Rossel, R. A., Walvoort, D. J. J., McBratney, A. B., Janik, L. J., and Skjemstad, J. O. Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131(1-2):59–75, 2006.

Wadoux, A. Interpretable spectroscopic modelling of soil with machine learning. *European Journal of Soil Science*, 2023.

Wadoux, A., Saby, N., and Martin, M. Shapley values reveal the drivers of soil organic carbon stock prediction. *The Soil*, 2023.

Wooldridge, J. M. *Introductory Econometrics: A Modern Approach*. Cengage Learning, Boston, 7 edition, 2020.

Zar, J. H. *Biostatistical Analysis*. Prentice Hall, Upper Saddle River, NJ, 5th edition, 2010.

Zhang, X., Xue, J., Chen, S., Wang, N., Xie, T., Xiao, Y., and Chen, X. e. a. Fine resolution mapping of soil organic carbon in croplands with feature selection and machine learning in northeast plain china. *Remote Sensing*, 2023.

Zheng, A. and Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, Inc., Sebastopol, CA, 2018.

Ziviani, M. M., Pinheiro, É. F. M., Ceddia, M. B., Ferreira, A. C. S., and Machado, F. S. Carbon and nitrogen stocks and soil organic matter persistence under native vegetation along a topographic and vegetation gradient in the central amazon region. *Soil Systems*, 8(2):65, 2024.