

UNIVERSIDADE FEDERAL DE PELOTAS
Faculdade de Agronomia Eliseu Maciel
Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água



Tese

**Mapeamento digital de atributos do solo em planície costeira arenosa:
seleção de covariáveis e predição espacial com *Random Forest* no Litoral
Sul do Rio Grande do Sul**

Ana Paula Knapp

Pelotas, 2026

Ana Paula Knapp

**Mapeamento digital de atributos do solo em planície costeira arenosa:
seleção de covariáveis e predição espacial com *Random Forest* no Litoral
Sul do Rio Grande do Sul**

Tese apresentada ao Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água da Faculdade de Agronomia Eliseu Maciel da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutora em Ciências.

Orientador: Pablo Miguel

Coorientadora: Juliana Brito da Silva

Pelotas, 2026

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

K67m Knapp, Ana Paula

Mapeamento digital de atributos do solo em planície costeira arenosa: [recurso eletrônico] : seleção de covariáveis e predição espacial com *Random Forest* no litoral sul do Rio Grande do Sul / Ana Paula Knapp ; Pablo Miguel, orientador ; Juliana Brito da Silva, coorientadora. — Pelotas, 2026.
217 f.

Tese (Doutorado) — Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, 2026.

1. Atributos do solo. 2. Aprendizado de máquina. 3. Validação cruzada espacial. 4. Planície costeira. 5. Incerteza preditiva. I. Miguel, Pablo, orient. II. Silva, Juliana Brito da, coorient. III. Título.

CDD 631.478165

Elaborada por Maria Beatriz Vaghetti Vieira CRB: 10/1032

Ana Paula Knapp

Mapeamento digital de atributos do solo em planície costeira arenosa: seleção de covariáveis e predição espacial com *Random Forest* no Litoral Sul do Rio Grande do Sul

Tese aprovada, como requisito parcial para obtenção do grau de Doutora em Ciências, Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas.

Data da defesa: 28/08/2026

Banca examinadora:

Prof. Dr. Pablo Miguel (Orientador)

Doutor em Ciência do Solo pela Universidade Federal de Santa Maria

Profa. Dra. Lizete Stumpf

Doutora em Agronomia - Solos pela Universidade Federal de Pelotas

Profa. Dra. Mélory Maria Fernandes de Araujo

Doutora em Ciências pela Universidade Federal de Pelotas

Dr. Stefan Domingues Nachtigall

Doutor em Ciências pela Universidade Federal de Pelotas

Agradecimentos

Agradeço, primeiramente, à minha família, em especial meus pais Gilberto e Lonise e ao meu irmão, João Vitor, por todo apoio, compreensão, incentivo e carinho dedicados a mim, pois sem o apoio incondicional de vocês, nada disso seria possível. E ao meu companheiro, Darlei, por estar ao meu lado, me apoiando e incentivando, nos momentos felizes e também nos dias mais difíceis! O carinho e encorajamento de vocês foram essenciais para que eu chegasse até aqui!

Ao meu orientador, Prof. Pablo, muito obrigada por me guiar nessa jornada, por todo o incentivo, otimismo, comprometimento e atenção dedicados durante esse processo.

A minha coorientadora, Profa. Juliana, agradeço pela disponibilidade, pelo apoio e pela atenção dedicados.

Gostaria de expressar um agradecimento especial aos colegas, estagiários e amigos do grupo de pesquisa Pedologia/UFPel, pela convivência diária, pela troca constante de conhecimentos e pela colaboração ao longo desta jornada. A parceria, o companheirismo e o ambiente de apoio mútuo tornaram os desafios da pesquisa mais leves e contribuíram significativamente para meu crescimento acadêmico, profissional e pessoal. Sou grata pelos momentos de aprendizado, pelas conversas, pelo incentivo e pela amizade construída durante esse período.

Aos demais professores e colegas do PPG MACSA, pelo conhecimento compartilhado e pela convivência diária.

Estendo meus agradecimentos aos membros da banca examinadora, pelas sugestões, críticas construtivas e pelo tempo dedicado à avaliação deste trabalho, contribuindo para seu aprimoramento.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) pelo suporte através da concessão de bolsas de pesquisa.

Agradeço ao Departamento de Solos da Universidade Federal de Santa Maria, na pessoa do Prof. Fabrício Pedron, pela parceria e valiosa colaboração no desenvolvimento desta pesquisa.

À Universidade Federal de Pelotas, à Faculdade de Agronomia Eliseu Maciel e ao Programa de Pós-Graduação em Manejo e Conservação do Solo e

da Água, pela infraestrutura, recursos oferecidos e a oportunidade de realização deste Curso.

E a todos que me incentivaram e torceram por mim.

Muito obrigada.

Resumo

KNAPP, Ana Paula. **Mapeamento digital de atributos do solo em planície costeira arenosa: seleção de covariáveis e predição espacial com *Random Forest* no Litoral Sul do Rio Grande do Sul**. Orientador: Pablo Miguel. 2026. 217 p. Tese (Doutorado em Manejo e Conservação do Solo e da Água) – Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, Pelotas, 2026.

O conhecimento da variabilidade espacial dos atributos do solo é essencial para o manejo sustentável, a conservação ambiental e a mitigação dos impactos das mudanças climáticas, particularmente em regiões de elevada fragilidade ambiental como as planícies costeiras arenosas. O mapeamento digital de solos, fundamentado no uso de covariáveis ambientais e algoritmos de aprendizado de máquina, configura-se como uma abordagem eficiente para a predição espacial desses atributos. Contudo, sua aplicação em ambientes costeiros de baixa variação topográfica representa um desafio devido à reduzida amplitude dos gradientes ambientais disponíveis para a modelagem preditiva, especialmente em áreas de solos arenosos com elevada suscetibilidade à degradação. A compreensão da distribuição espacial de atributos como o estoque de carbono orgânico do solo, a densidade do solo, o teor de areia e o pH é fundamental para subsidiar o manejo sustentável e a conservação desses solos. Este estudo tem como objetivo modelar e mapear espacialmente o estoque de carbono orgânico do solo (ECOS), a densidade do solo, o teor de areia e o pH em uma planície costeira arenosa no Litoral Sul do Rio Grande do Sul, utilizando o algoritmo *Random Forest* e covariáveis ambientais topográficas e espectrais. O trabalho utiliza 64 amostras de solo distribuídas em cinco áreas definidas pelo método do Hipercubo Latino Condicionado (cLHS), nas profundidades padrão do *GlobalSoilMap* (0–5, 5–15, 15–30, 30–60 e 60–100 cm). O conjunto de preditores compreende 48 covariáveis ambientais, sendo 33 topográficas/morfométricas derivadas de modelo digital de elevação SRTM e 15 espectrais derivadas de imagens Landsat. A modelagem emprega validação cruzada espacial do tipo Leave-One-Cluster-Out, com otimização do hiperparâmetro *mtry* por busca em grade. O estudo avalia três estratégias de seleção de covariáveis, sendo elas: conjunto completo, seleção por correlação e seleção por correlação e importância preditiva, seguidas de redução dimensional para as 15 covariáveis mais importantes. Os modelos apresentam capacidade preditiva limitada, com R^2 máximo de 0,103 para ECOS, 0,225 para densidade do solo, 0,175 para teor de areia e 0,229 para pH. A abordagem sem remoção prévia de covariáveis predomina para ECOS, densidade do solo e areia, enquanto a seleção combinada por correlação e importância destaca-se para pH nas camadas superficiais. A redução dimensional produz ganhos preditivos na maioria das combinações avaliadas, mesmo que de pequena magnitude. As covariáveis topográficas apresentam maior contribuição preditiva, contrariando a expectativa de predominância de preditores espectrais em ambientes de baixo relevo. Os mapas preditivos reproduzem a tendência geral dos dados observados, com suavização espacial característica de modelos de

aprendizado de máquina e a incerteza aumenta progressivamente com a profundidade. Os resultados demonstram que o desempenho dos modelos depende mais da capacidade das covariáveis em representar os processos pedogenéticos do que da quantidade de preditores utilizados, e que mesmo em paisagens aparentemente homogêneas existem gradientes ambientais capazes de influenciar a distribuição espacial dos atributos do solo.

Palavras-chave: atributos do solo; aprendizado de máquina; validação cruzada espacial; planície costeira; incerteza preditiva.

Abstract

KNAPP, Ana Paula. **Digital mapping of soil attributes in a sandy coastal plain: covariate selection and spatial prediction with *Random Forest* on the Southern Coast of Rio Grande do Sul**. Advisor: Pablo Miguel. 2026. 217 p. Thesis (Doctorate in Soil and Water Management and Conservation) – Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, Pelotas, 2026.

Knowledge of the spatial variability of soil attributes is essential for sustainable management, environmental conservation, and mitigation of climate change impacts, particularly in high environmental fragility regions such as sandy coastal plains. Digital soil mapping, based on environmental covariates and machine learning algorithms, is an efficient approach for the spatial prediction of these attributes. However, its application in coastal environments of low topographic variation represents a challenge due to the reduced amplitude of environmental gradients available for predictive modeling, especially in areas of sandy soils with high susceptibility to degradation. Understanding the spatial distribution of attributes such as soil organic carbon stock, bulk density, sand content, and pH is essential to support sustainable management and conservation of these soils. This study aims to model and spatially map soil organic carbon stock (SOCS), soil bulk density, sand content, and soil pH in a sandy coastal plain on the Southern Coast of Rio Grande do Sul, Brazil, using the Random Forest algorithm and topographic and spectral environmental covariates. The work uses 64 soil samples distributed across five areas defined by the conditioned Latin Hypercube Sampling (cLHS) method, at GlobalSoilMap standard depths (0–5, 5–15, 15–30, 30–60, and 60–100 cm). The predictor set comprises 48 environmental covariates, consisting of 33 topographic/morphometric variables derived from SRTM digital elevation model and 15 spectral variables derived from Landsat imagery. Modeling employs spatial cross-validation with Leave-One-Cluster-Out strategy, with mtry hyperparameter optimization via grid search. The study evaluates three covariate selection strategies, namely: full set, correlation-based selection, and correlation and importance-based selection, followed by dimensional reduction to the 15 most important covariates. The models show limited predictive performance, with maximum R^2 of 0.103 for SOCS, 0.225 for bulk density, 0.175 for sand content, and 0.229 for pH. The approach without prior covariate removal predominates for SOCS, bulk density, and sand, while the combined correlation and importance selection stands out for pH in the surface layers. Dimensional reduction yields predictive gains in most evaluated combinations, even if of small magnitude. Topographic covariates show the highest predictive contribution, contrary to the expectation of spectral predictor predominance in low-relief environments. The predictive maps reproduce the general trend of observed data, with spatial smoothing characteristic of machine learning models and uncertainty increases progressively with depth. The results demonstrate that model performance depends more on the ability of covariates to represent pedogenetic processes than on the number of predictors used, and that even in

apparently homogeneous landscapes there are environmental gradients capable of influencing the spatial distribution of soil attributes.

Keywords: soil attributes; machine learning; spatial cross-validation; coastal plain; predictive uncertainty.

Lista de Figuras

Figura 1	Delimitação e localização geográfica da área de estudo no litoral sul do estado do Rio Grande do Sul, Brasil	46
Figura 2	Distribuição das cinco áreas de amostragem delimitadas pelo algoritmo cLHS na área de estudo, cada uma com raio de 15 km	49
Figura 3	Distribuição espacial dos 64 pontos de coleta de solo nas cinco áreas de amostragem da área de estudo	50
Figura 4	Representação esquemática do processo de validação cruzada espacial leave-one-cluster-out	58
Figura 5	Representação esquemática das estratégias de seleção de covariáveis utilizadas na modelagem Random Forest	60
Figura 6	Esquema metodológico para geração dos mapas de predição e incerteza de ECOS	64
Figura 7	Distribuição do estoque de carbono orgânico do solo (ECOS) em diferentes profundidades de amostragem	68
Figura 8	Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0–5, 5-15, 15-30, 30-60 e 60-100 cm do solo pela abordagem RF_all	87
Figura 9	Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0–5, 5-15, 15-30, 30-60 e 60-100 cm do solo pela abordagem RF_Corr	88
Figura 10	Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0–5, 5-15, 15-30, 30-60 e 60-100 cm do solo pela abordagem RF_CorrImp	90
Figura 11	Distribuição espacial dos estoques de carbono orgânico do solo (ECOS) estimados pelo modelo Random Forest nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm	102
Figura 12	Distribuição espacial da incerteza associada às estimativas de ECOS, expressa pelo coeficiente de variação (CV), nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm	104
Figura 13	Localização da área de estudo no litoral sul do estado do Rio Grande do Sul, Brasil	113

Figura 14	Delimitação dos locais de amostragem da área de estudo pelo método cLHS	115
Figura 15	Distribuição espacial dos pontos de coleta de solo na área de estudo	117
Figura 16	Fluxograma da validação cruzada espacial com otimização do hiperparâmetro mtry no modelo Random Forest	123
Figura 17	Fluxograma da metodologia de seleção de covariáveis e comparação das abordagens de Random Forest	125
Figura 18	Fluxo metodológico para geração dos mapas de predição espacial e incerteza	127
Figura 19	Distribuição da densidade do solo (g cm^{-3}), teor de areia (%) e pH por profundidade (0–5, 5–15, 15–30, 30–60 e 60–100 cm), para os 64 pontos amostrais	130
Figura 20	Distribuição vertical da densidade do solo ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos	132
Figura 21	Distribuição vertical do teor de areia ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos	134
Figura 22	Distribuição vertical do pH ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos	136
Figura 23	Diferença entre os modelos ajustados com as 15 covariáveis mais importantes (Top15) e aqueles ajustados com todas as covariáveis para os atributos areia, densidade do solo (BD) e pH, em diferentes profundidades	144
Figura 24	Heatmap da frequência de inclusão de cada covariável no conjunto Top15, para todas as combinações de abordagem x profundidade	150
Figura 25	Índice de Jaccard entre pares de abordagens (RF_all x RF_Corr, RF_all x RF_CorrImp, RF_Corr x RF_CorrImp; Top15) para BD, areia e pH por profundidade	154
Figura 26	Índice de Jaccard entre profundidades adjacentes (RF_all	

	Top15) para densidade do solo, areia e pH	155
Figura 27	Índice de Jaccard entre atributos (RF_all Top15) por profundidade	157
Figura 28	Distribuição espacial dos valores preditos de densidade do solo (BD) nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade	161
Figura 29	Distribuição espacial dos valores preditos de teor de areia nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade	162
Figura 30	Distribuição espacial dos valores preditos de pH nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade	164
Figura 31	Distribuição espacial da incerteza das predições do teor de areia nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)	167
Figura 32	Distribuição espacial da incerteza das predições da densidade do solo nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)	169
Figura 33	Distribuição espacial da incerteza das predições do pH do solo nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)	171

Lista de Tabelas

Tabela 1	Caracterização morfológica dos 64 pontos de coleta de solo distribuídos na área de estudo, conforme classes de solo do SiBCS	51
Tabela 2	Densidade do solo média (g cm^{-3}) por classe de solo e profundidade na área de estudo	52
Tabela 3	Distribuição dos 64 pontos de coleta de solo conforme uso da terra na área de estudo	52
Tabela 4	Covariáveis ambientais utilizadas na modelagem do estoque de carbono orgânico do solo (ECOS)	54
Tabela 5	Caracterização estatística do estoque de carbono orgânico do solo (ECOS) por camada de profundidade na área de estudo	65
Tabela 6	Covariáveis ambientais removidas durante o processo de seleção de preditores com base no critério de correlação simples (RF_Corr)	71
Tabela 7	Covariáveis ambientais removidas durante o processo de seleção de preditores com base no critério de correlação e importância (RF_CorrImp)	73
Tabela 8	Valores otimizados do hiperparâmetro mtry e número de covariáveis utilizadas nas diferentes abordagens de modelagem Random Forest para predição do estoque de carbono orgânico do solo	73
Tabela 9	Métricas de desempenho dos modelos Random Forest obtidas por validação cruzada espacial para diferentes abordagens de seleção de covariáveis e profundidades do solo	77
Tabela 10	Métricas de desempenho dos modelos Random Forest com redução dimensional do conjunto de covariáveis (Top15) para diferentes profundidades do solo	82
Tabela 11	Efeito da seleção das 15 covariáveis mais importantes sobre o desempenho dos modelos Random Forest em diferentes profundidades do solo	85

Tabela 12	Melhor abordagem de modelagem Random Forest para estimativa do ECOS em diferentes profundidades	97
Tabela 13	Estatística descritiva dos mapas preditivos de estoques de carbono orgânico do solo (ECOS) espacializados pelo modelo Random Forest em diferentes profundidades do solo	99
Tabela 14	Estatísticas descritivas da incerteza dos mapas preditivos de ECOS, expressa pelo coeficiente de variação (CV), em diferentes profundidades do solo	100
Tabela 15	Covariáveis ambientais utilizadas na modelagem preditiva dos atributos do solo	119
Tabela 16	Distribuição das classes de solo e parâmetros dos horizontes pedogenéticos (n = 64 perfis)	131
Tabela 17	Abordagem selecionada e métricas de desempenho dos modelos de Random Forest para cada atributo e profundidade	138
Tabela 18	Valores de mtry ótimo dos modelos vencedores por atributo e profundidade	147
Tabela 19	Frequência de inclusão no Top15 das quinze covariáveis mais recorrentes	149
Tabela 20	Estatísticas descritivas dos mapas preditivos e comparação da variabilidade relativa com os dados observados para cada atributo e profundidade	159
Tabela 21	Estatísticas descritivas dos mapas de incerteza obtidos a partir do coeficiente de variação (CV%) entre as predições individuais das árvores do modelo Random Forest para cada atributo e profundidade	165

Sumário

1 Introdução	17
1.1 Objetivos	19
1.1.1 Geral	19
1.1.2 Específicos	19
1.2 Hipóteses	20
2 Revisão de literatura	21
2.1 Mapeamento digital de solos (MDS)	21
2.1.1 Origem e evolução do MDS	21
2.1.2 Covariáveis ambientais	23
2.1.3 <i>Random Forest</i> no MDS	25
2.2 Validação cruzada e seleção de covariáveis ambientais em MDS	27
2.3 Mudanças climáticas e o estoque de carbono orgânico do solo	30
2.4 Mapeamento digital de solos para predição de estoque de carbono orgânico do solo	31
2.5 Mapeamento digital de solos para predição de pH do solo	34
2.6 Mapeamento digital de solos para predição da densidade do solo	35
2.7 Mapeamento digital de solos para predição da granulometria do solo	37
2.8 Contexto geomorfológico e dinâmica costeira do litoral sul do Rio Grande do Sul	39
3 Capítulo I – Desempenho de estratégias de seleção de covariáveis no mapeamento digital do estoque de carbono orgânico do solo no Litoral Sul do Rio Grande do Sul	41
3.1 Introdução	41
3.2 Metodologia	45
3.2.1 Área de estudo	45
3.2.2 Amostragem de solo	48
3.2.3 Análises laboratoriais	53
3.2.4 Covariáveis ambientais	53
3.2.5 Calibração e validação	56
3.2.6 Abordagens de seleção de covariáveis e redução dimensional do número de covariáveis	59
3.2.6.1 Abordagens de seleção de covariáveis	60
3.2.6.2 Redução dimensional do número de covariáveis	62

3.2.7 Predição espacial e incerteza.....	63
3.2.8 Software e ferramentas utilizadas	65
3.3 Resultados e discussão.....	65
3.3.1 Distribuição vertical do ECOS observado.....	65
3.3.2 Análise exploratória das covariáveis ambientais	69
3.3.3 Desempenho preditivo dos modelos <i>Random Forest</i>	73
3.3.3.1 Ajuste dos hiperparâmetros.....	73
3.3.3.2 Comparação entre estratégias de seleção de covariáveis	76
3.3.3.3 Efeito da redução dimensional de covariáveis	81
3.3.4 Importância das covariáveis ambientais.....	86
3.3.5 Melhor abordagem	96
3.3.6 Espacialização do ECOS	98
3.4 Conclusões.....	106
4 Capítulo II – Seleção de covariáveis e predição espacial de densidade do solo, teor de areia e pH no Litoral Sul do Rio Grande do Sul.....	108
4.1 Introdução	108
4.2 Metodologia.....	112
4.2.1 Área de estudo	112
4.2.2 Amostragem de solo.....	115
4.2.3 Análises laboratoriais	118
4.2.4 Modelagem preditiva	119
4.2.4.1 Modelagem com <i>Machine Learning</i>	119
4.2.4.2 Covariáveis ambientais	119
4.2.4.3 Calibração e validação do modelo	122
4.2.5 Estratégias de seleção de covariáveis	124
4.2.6 Redução dimensional do conjunto de covariáveis.....	125
4.2.7 Mapas preditivos e de incerteza.....	126
4.2.8 Análise de importância e consistência das covariáveis	128
4.2.9 Software e ferramentas utilizadas	129
4.3 Resultados e discussão.....	129
4.3.1 Caracterização descritiva dos dados de solo	129
4.3.2 Desempenho preditivo e efeito da seleção de covariáveis.....	137
4.3.2.1 Desempenho geral	137
4.3.2.2 Efeito da redução dimensional (Top15).....	143

4.3.2.3 Ajuste do hiperparâmetro mtry	146
4.3.2.4 Importância das covariáveis ambientais.....	148
4.3.3 Análise de sobreposição de covariáveis.....	153
4.3.3.1 Entre estratégias de seleção	153
4.3.3.2 Entre profundidades	155
4.3.3.3 Entre atributos	157
4.3.4 Espacialização dos atributos de solo	159
4.4 Conclusões.....	173
5 Considerações finais	175
Referências	176
Apêndices	194

1 Introdução

A compreensão aprofundada da variabilidade espacial das propriedades do solo é um pilar fundamental para o avanço da sustentabilidade ambiental e da segurança alimentar em escala global. Solos são sistemas complexos e dinâmicos, e suas características físicas e químicas influenciam diretamente a produtividade agrícola, a qualidade dos ecossistemas e a capacidade de resposta a desafios como as mudanças climáticas. Em regiões com características pedológicas específicas, como os solos arenosos do Litoral Sul do Brasil, essa compreensão torna-se ainda mais crítica devido à sua elevada fragilidade e suscetibilidade à degradação.

O carbono orgânico do solo (COS) e o estoque de carbono orgânico do solo (ECOS) são reconhecidos como elementos cruciais para a saúde do solo e para o monitoramento do impacto das atividades humanas no clima. Solos com maiores teores de matéria orgânica apresentam melhor fertilidade, estrutura aprimorada, ciclagem de nutrientes eficiente e maior atividade biológica, promovendo o desenvolvimento vegetal. Em solos arenosos, a matéria orgânica atua como um agente agregador, aumentando a resistência à erosão hídrica e eólica e melhorando a capacidade de armazenamento de água. Contudo, a dinâmica do COS é sensível a mudanças no uso e manejo do solo, especialmente em biomas como o Pampa, no sul do Brasil, que passaram por intensas transformações de campos naturais para áreas de cultivo.

Além do carbono, outras propriedades do solo, como o pH, a composição granulométrica e a densidade, desempenham papéis igualmente vitais. A heterogeneidade espacial dessas propriedades afeta a disponibilidade de nutrientes, a atividade microbiana, a retenção de água e a aeração, influenciando diretamente a produtividade e a qualidade do ecossistema. O pH regula a disponibilidade de nutrientes e a atividade microbiana, enquanto a granulometria determina a estrutura física do solo e a capacidade de retenção de água, impactando a estabilização do carbono orgânico. A densidade do solo, por sua vez, está diretamente ligada à compactação e porosidade, com solos compactados apresentando restrições ao crescimento radicular e à infiltração de água.

A região litorânea do sul do Rio Grande do Sul, caracterizada por solos predominantemente de textura arenosa, é particularmente vulnerável a

processos erosivos e apresenta baixa resiliência, reforçando a necessidade de um entendimento aprofundado de suas características para o uso e manejo adequados. Diante da complexidade e da importância dessas propriedades, o Mapeamento Digital de Solos (MDS) é uma abordagem promissora para aprimorar a compreensão da variabilidade espacial do solo, permitindo um planejamento mais eficiente do uso e manejo sustentável dos recursos naturais.

As técnicas de aprendizado de máquina, como o *Random Forest* (RF), integradas a covariáveis ambientais derivadas de sensoriamento remoto e Modelo Digital de Elevação (MDE), oferecem ferramentas robustas para prever a distribuição espacial desses atributos com alta precisão. Essa abordagem permite não apenas a criação de mapas preditivos detalhados, mas também a identificação das covariáveis-chave que mais influenciam a variabilidade de cada propriedade do solo, otimizando a coleta de dados e a eficiência dos modelos. Além disso, a capacidade de quantificar e avaliar a incerteza associada às previsões é fundamental para a confiabilidade das informações geradas, permitindo um planejamento mais seguro e assertivo.

Com base nesse contexto, este trabalho tem como objetivo principal prever propriedades químicas (estoque de carbono orgânico e pH do solo) e físicas (densidade e teor de areia) do solo em solos arenosos da região litorânea do sul do Rio Grande do Sul, utilizando técnicas de aprendizado de máquina baseado em *Random Forest* (RF) e covariáveis ambientais. Tais atributos foram selecionados para serem mapeados por representarem propriedades relacionadas à qualidade química e física do solo, apresentarem variabilidade vertical e horizontal e possuírem reconhecida importância para o uso e manejo de solos arenosos.

Buscou-se, com este estudo, contribuir para o avanço do conhecimento sobre a distribuição espacial dessas propriedades, avaliando o desempenho e a incerteza das previsões, identificando as covariáveis mais importantes e analisando o impacto da simplificação dos modelos. Os resultados gerados visam fornecer subsídios para práticas agrícolas mais sustentáveis, o manejo adequado do solo e a formulação de políticas públicas voltadas à conservação ambiental e à mitigação dos impactos das mudanças climáticas em uma região ecologicamente sensível. Ademais, para além das aplicações para manejo

sustentável, o estudo contribui para o avanço metodológico do mapeamento digital de solos em ambientes costeiros de baixa variação topográfica.

1.1 Objetivos

1.1.1 Geral

Predizer e mapear propriedades químicas (estoque de carbono orgânico do solo e pH) e físicas (densidade do solo e teor de areia) do solo em cinco profundidades padrão do *GlobalSoilMap*, utilizando o algoritmo *Random Forest* e covariáveis ambientais derivadas de sensoriamento remoto (topográficas e espectrais), avaliando o efeito de diferentes estratégias de seleção e redução de covariáveis sobre o desempenho preditivo da modelagem e a consistência dessas estratégias entre diferentes atributos do solo em ambiente costeiro de baixa variação topográfica.

1.1.2 Específicos

- Construir modelos *Random Forest* para prever o estoque de carbono orgânico do solo (ECOS), a densidade do solo, o teor de areia e o pH em cinco profundidades padrão do *GlobalSoilMap* (0-5, 5-15, 15-30, 30-60 e 60-100 cm).
- Comparar o desempenho preditivo de diferentes estratégias de seleção de covariáveis ambientais para cada atributo do solo.
- Avaliar o impacto da redução dimensional por meio da construção de modelos utilizando apenas as 15 covariáveis mais importantes para cada atributo do solo.
- Identificar as covariáveis ambientais mais relevantes para a predição dos atributos do solo nas diferentes profundidades.
- Verificar a consistência das estratégias de seleção de covariáveis que apresentam melhor desempenho entre os diferentes atributos do solo e entre profundidades.
- Gerar mapas digitais dos atributos do solo para as cinco profundidades padrão do *GlobalSoilMap*, utilizando a estratégia de seleção de covariáveis que apresentar o melhor desempenho preditivo para cada variável resposta.

- Quantificar e mapear a incerteza das predições por meio do coeficiente de variação associado às estimativas do modelo para cada atributo e profundidade.

1.2 Hipóteses

- As estratégias de seleção de covariáveis ambientais reduzem a redundância entre preditores sem comprometer significativamente o desempenho dos modelos *Random Forest*.
- Modelos construídos com conjuntos reduzidos e mais informativos de covariáveis ambientais apresentam desempenho igual ou superior aos modelos que utilizam todas as covariáveis.
- As estratégias de seleção de covariáveis ambientais que apresentam melhor desempenho variam entre diferentes atributos do solo e são mais consistentes entre profundidades do que entre atributos.
- A importância relativa das covariáveis ambientais varia entre as profundidades e entre os atributos do solo avaliados, refletindo diferentes processos pedogenéticos e fatores de controle ao longo do perfil.
- Em ambientes costeiros de baixa variabilidade topográfica, covariáveis espectrais derivadas de sensoriamento remoto apresentam maior contribuição para a predição dos atributos do solo do que covariáveis derivadas do relevo, independentemente do atributo analisado.

2 Revisão de literatura

2.1 Mapeamento digital de solos (MDS)

2.1.1 Origem e evolução do MDS

O mapeamento de solos é uma ferramenta fundamental para uso e manejo adequados do recurso solo, por isso, a integração entre o uso de geotecnologias e coletas de campo é essencial para suprir a carência de informações mais detalhadas acerca do solo (Dalmolin; Ten Caten, 2015). O avanço da tecnologia propiciou a geração de dados em grande escala e novas ferramentas como o surgimento de bancos de dados de informações de solos e, com isso, o desafio de desenvolver ferramentas e métodos estatísticos que culminaram na mineração de dados e aprendizado de máquina. O advento do sensoriamento remoto e das ferramentas de sistemas de informações geográficas (SIG) também se somaram a esses avanços, atendendo à demanda crescente de informações relacionadas ao solo para fins de monitoramento ambiental (McBratney *et al.*, 2003).

O mapeamento digital de solos (MDS), que é parte da pedologia, que é o ramo da ciência do solo que se ocupa do estudo da gênese, morfologia, classificação e levantamento de solos, representa uma evolução ao método convencional de mapeamento de solos, que tornou possível a integração entre dados de campo, laboratório e covariáveis ambientais. Entretanto, assim como os conhecimentos da pedologia e da pedogênese são importantes para o MDS, as informações geradas pelo MDS podem trazer avanços para os estudos de pedogênese através do entendimento da distribuição vertical e lateral de propriedades e atributos do solo (Ma *et al.*, 2019).

O mapeamento de dados relacionados aos solos, na sua forma convencional, exige elevados custos, necessidade de mão-de-obra qualificada em campo para coleta manual de dados e amostras, bem como para proceder às análises laboratoriais necessárias, tornando-se assim uma atividade demorada. Nesse contexto, surge a pedometria, reunindo a pedologia e a geoestatística, tornando o levantamento de informações acerca dos solos mais rápido e barato, resultando em amostragens mais eficientes e mapas e previsões mais precisas, permitindo assim prever e modelar propriedades do solo. Portanto, a pedometria emerge como uma ferramenta de gestão ambiental (McBratney *et al.*, 2019).

O MDS se baseia em sete fatores preditivos, sendo eles: o solo, o clima, os organismos, a topografia, o material de origem, a idade ou fator tempo e a posição geográfica. A proposta de uma estrutura geral para o MDS foi construída por McBratney *et al.* (2003), o chamado método SCORPAN, que propõe que o próprio solo e sua posição geográfica sejam considerados para prever o comportamento do solo. Esse método tem sua origem nos cinco fatores de formação do solo propostos por Jenny (1941) no qual o solo seria função das condições de clima (cl), de relevo e topografia (r), dos organismos (o), do material de origem (p) e do tempo (t) ou abordagem *clorpt*. A equação de Jenny é um modelo mecanicista que era utilizado para entender a formação do solo e não necessariamente para prever o solo a partir de outros fatores (como é o pressuposto do MDS) enquanto com o modelo SCORPAN, o solo passa a ser considerado um fator de predição porque as propriedades do solo podem ser utilizadas para prever outros atributos dele mesmo (McBratney *et al.*, 2003).

Os avanços na área do MDS se deram, principalmente, pelo desenvolvimento de ferramentas de SIG, GPS, sensores remotos e informações de modelos digitais de elevação (MDE), atendendo assim às demandas da ciência pelo monitoramento e modelagem ambiental. Contudo, ainda se percebe, em nível mundial, que há uma limitação importante na qual a evolução do MDS esbarra, que é a falta de infraestrutura de dados, isso ocorre porque os levantamentos de solos convencionais demandam elevados recursos financeiros e são demorados. Entretanto, prever o solo a partir de outros atributos do próprio solo é possível a partir da utilização de variáveis secundárias que estejam disponíveis com densidade adequada para este fim (McBratney *et al.*, 2003).

Um avanço promovido pela pedometria e pelo MDS foi a melhoria das estratégias de amostragem de solo, que antes eram definidas a campo selecionando locais que eram representativos da paisagem, exigindo conhecimento e prática dos agentes de campo. O advento da geoestatística permitiu que se avançasse no entendimento de como os locais de amostragem devem ser escolhidos, por exemplo, com o advento do método cLHS (*Conditioned Latin Hypercube Sampling*) que permite que se use um conjunto

de variáveis para definição dos pontos de amostragem mais adequados dentro da área de interesse (McBratney *et al.*, 2019).

O principal avanço que a aplicação da geoestatística traz para o mapeamento de solos é a possibilidade de, a partir de dados limitados, estimar valores em locais onde não foi feita amostragem de campo e também de quantificar as incertezas dessas estimativas. A pedometria, portanto, trouxe à luz da ciência do solo o uso de diversas ferramentas para tratar da natureza contínua do solo. Com isso, a pedometria deixa como legado para a ciência do solo a popularização do uso de modelos digitais de elevação, o uso de sensoriamento remoto e proximal do solo, a espectroscopia, as funções de pedotransferência e avanços na amostragem do solo (McBratney *et al.*, 2019).

2.1.2 Covariáveis ambientais

A utilização de covariáveis ambientais no mapeamento digital de solos tem se destacado como uma abordagem eficaz para compreender a variabilidade espacial das propriedades do solo. A integração entre dados de campo e de laboratório com covariáveis ambientais é um dos avanços do mapeamento digital de solos (Ma *et al.*, 2019). As covariáveis ambientais são informações usadas no MDS por estarem associadas aos processos de formação dos solos. Por exemplo, dados geológicos que remetem ao material de origem que por sua vez afetam a mineralogia e a textura do solo; informações topográficas extraídas de modelos digitais de elevação; informações de sensoriamento remoto que refletem o uso e cobertura do solo e as condições da vegetação que estão associadas à atividade humana e dos organismos e ainda dados climáticos (Lamichhane *et al.*, 2019). Contudo, covariáveis extraídas de MDE e de imagens de sensoriamento remoto têm sido as mais populares para predição de propriedades do solo (Pouladi *et al.*, 2023).

Inúmeros estudos de mapeamento digital de solos apontam a importância das diferentes covariáveis ambientais na predição de atributos dos solos. A precipitação e a reflectância superficial demonstraram maior relevância na predição de propriedades químicas do solo, ao passo que o relevo, o material de origem e a vegetação apresentaram maior peso para os atributos físicos (Hengl *et al.*, 2017). No que diz respeito às frações granulométricas, as covariáveis de sensoriamento remoto se destacaram entre as mais

importantes, com a temperatura exercendo influência específica sobre a predição dos teores de argila. Enquanto o mapeamento do carbono orgânico do solo foi predominantemente influenciado por covariáveis de relevo e de sensoriamento remoto (Forkuor *et al.*, 2017).

Achados de Parvizi e Fatehi (2025) evidenciaram que a partir de um conjunto de covariáveis preditoras constituído por 15 covariáveis topográficas, 2 de vegetação, 6 de material de origem e 4 climáticas, foram as covariáveis relacionadas ao material de origem e à topografia que tiveram o maior efeito na predição do carbono orgânico do solo. Já John *et al.* (2020), em área de relevo plano, observaram fraca correlação entre as covariáveis de sensoriamento remoto e o carbono orgânico do solo. Ainda em áreas de relevo plano, Li *et al.* (2020) encontraram uma boa relação entre o carbono orgânico e índice de umidade topográfica (covariável derivada de modelo digital de elevação).

Em duas áreas sob diferentes condições de inclinação em uma planície aluvial datada do início do Pleistoceno, na Hungria, Hateffard *et al.* (2023) observaram diferentes covariáveis ambientais influenciando com maior intensidade o COS, o ECOS e a densidade do solo. Em uma das áreas (com inclinação média de 0,5-1%), foi observada maior influência de covariáveis de terreno na predição de COS e ECOS e a interação entre terreno e vegetação para a densidade do solo; na outra área (com inclinação média de 1,5-5%), o comportamento inverso foi observado para o COS e ECOS, sendo mais influenciado por índice de vegetação e a densidade do solo sendo mais influenciada por covariáveis de terreno. Na América Latina, Guevara *et al.* (2018) identificaram como principais preditores do carbono orgânico do solo fatores como temperatura, tipo de solo, índices de vegetação e topografia.

Com relação à quantidade de covariáveis ambientais utilizadas para modelagens preditivas de solo, Poggio *et al.* (2021) concluíram que o desempenho dos modelos preditivos melhora ao utilizar-se conjuntos que contenham de 15 a 20 covariáveis. Nussbaum *et al.* (2018) obtiveram melhores desempenhos para seus modelos preditivos com conjuntos reduzidos de covariáveis ambientais, representando de 3 a 6% de todas as covariáveis utilizadas. Com base no exposto, vale ressaltar que McBratney *et al.* (2003) já chamavam atenção para a importância da seleção de dados de entrada nas modelagens preditivas em qualidade e quantidade adequadas, ideia essa

reforçada por Batjes *et al.* (2020) ao chamarem atenção para o uso de informações padronizadas para as modelagens.

A escala espacial também influencia a relevância das covariáveis na predição de atributos do solo. Em áreas menores, índices de vegetação e uso da terra mostraram-se mais representativos, enquanto em escalas regionais, fatores como clima, material de origem, topografia e uso da terra tiveram maior impacto na predição do estoque de carbono orgânico. A geologia, por sua vez, tem papel essencial na predição do carbono orgânico do solo, sendo ainda mais relevante em profundidades maiores, enquanto o uso da terra afeta principalmente camadas superficiais. Dessa forma, os dados geológicos mostram grande potencial para modelagem em áreas extensas e em profundidades (Lamichhane *et al.*, 2019).

Além disso, aprimorar a resolução espacial dos dados pode impactar significativamente o desempenho preditivo dos modelos. Hengl *et al.* (2017) verificaram que a alteração da resolução espacial de 1 km para 250 m resultou em um aumento expressivo na precisão das estimativas, melhorando em 200% a predição do carbono orgânico do solo, 65% para o pH do solo, 234% para a fração areia, 127% para a fração silte, 198% para a fração argila e 138% para a densidade do solo.

2.1.3 Random Forest no MDS

O *Random Forest* (RF) é um modelo de aprendizado de máquina baseado em múltiplas árvores de decisão. Cada árvore dentro da floresta é criada a partir de um vetor aleatório, amostrado independentemente e seguindo a mesma distribuição. Essas árvores são treinadas com diferentes subconjuntos dos dados, permitindo que cada uma faça previsões individuais. Após a construção de um número significativo de árvores, a predição final se baseia na média das previsões individuais das árvores. Esse mecanismo torna o RF um modelo robusto e eficiente para diversas aplicações, reduzindo o risco de *overfitting* ao combinar múltiplas previsões (Breiman, 2001).

Apesar de sua eficiência, é fundamental selecionar cuidadosamente as covariáveis utilizadas na modelagem. A inclusão excessiva de covariáveis pode aumentar as incertezas, reduzindo a influência de preditores críticos e comprometendo a capacidade de generalização dos modelos (Huang *et al.*,

2025). No entanto, diversos estudos demonstram a eficácia do RF em diferentes cenários. Lamichhane *et al.* (2019) concluíram que RF apresentou bom desempenho em diversas condições de paisagem, sendo especialmente útil em áreas extensas e de difícil acesso, onde o número de amostras é reduzido. Hateffard *et al.* (2023) também verificaram que o RF teve resultados satisfatórios mesmo com dados de campo limitados.

Na modelagem do pH do solo, a parametrização adequada do RF influencia diretamente os resultados. Huang *et al.* (2025) determinaram que os valores ideais de *ntree* e *mtry* para a predição do pH são 200 e 15, respectivamente, resultando em um R^2 de 0,41. Dharumarajan *et al.* (2017), utilizando *ntree* = 200, obtiveram R^2 de 0,30, enquanto Mousavi *et al.* (2023), com *ntree* = 1000 e *mtry* = 23, alcançaram R^2 de 0,82, evidenciando que a escolha de parâmetros impacta significativamente a precisão dos modelos.

Além da predição do pH, o RF tem sido amplamente aplicado na estimativa do ECOS. Em uma revisão que avaliou 120 estudos ao redor do mundo, Lamichhane *et al.* (2019) verificaram que o RF superou a regressão linear múltipla e outras técnicas de aprendizado de máquina. Na América Latina, Guevara *et al.* (2018) identificaram melhores resultados para RF no Brasil, onde o método apresentou maior R^2 e menor RMSE na predição do ECOS.

A validação dos modelos preditivos também influencia os resultados. Hengl *et al.* (2017) observaram que o RF obteve R^2 de 0,85 ao utilizar amostras *out-of-bag* (OOB) e R^2 de 0,60 com validação cruzada repetida, destacando a importância da estratégia de validação escolhida. Forkuor *et al.* (2017), ao mapear propriedades do solo em uma bacia agrícola de Burkina Faso (África Ocidental), concluíram que RF foi o método de aprendizado de máquina mais preciso, mas ressaltaram que a divisão da área de estudo em regiões homogêneas pode melhorar ainda mais os resultados preditivos.

A escolha da abordagem de amostragem também impacta o desempenho do RF. Castro-Franco *et al.* (2015) demonstraram que RF, combinado com amostragem baseada em cLHS, explicou 71% da variabilidade da argila e 65% da variabilidade da matéria orgânica do solo. Já a abordagem Fuzzy C-Means, combinada ao RF, resultou em 68% e 57% de explicação para argila e matéria orgânica, respectivamente, evidenciando que a estratégia de

amostragem pode influenciar a qualidade da predição. Além disso, a avaliação da incerteza representa um aspecto essencial em qualquer modelo preditivo, especialmente quando os resultados são utilizados para orientar processos decisórios (Padarian *et al.*, 2020). Garantir que as previsões sejam acompanhadas por métricas de incerteza ajuda a aumentar a confiabilidade das análises e aprimorar a aplicabilidade dos modelos em contextos científicos e práticos.

2.2 Validação cruzada e seleção de covariáveis ambientais em MDS

Em mapeamento digital de solos (MDS), os modelos preditivos são calibrados a partir de dados pontuais de solo coletados em campo e analisados em laboratório, combinados com informações ambientais derivadas de covariáveis disponíveis para toda a área de estudo. A partir dessas relações, são gerados mapas contínuos de atributos ou classes de solo (Padarian *et al.*, 2019; Piikki *et al.*, 2021).

Como esses produtos representam uma generalização da variabilidade espacial dos solos, é fundamental que sejam submetidos a processos de validação que permitam avaliar sua qualidade e confiabilidade. De forma geral, a validação consiste na comparação entre valores observados em campo e valores preditos pelo modelo em pontos não utilizados durante a calibração. Entre os métodos de validação utilizados em MDS, a validação cruzada é amplamente empregada, com destaque para as abordagens *leave-one-out* e *k-fold*. Essa popularidade está relacionada à sua eficiência no aproveitamento de conjuntos de dados reduzidos, comuns em estudos de solos, além de fornecer estimativas estáveis de desempenho (Piikki *et al.*, 2021). Especificamente no mapeamento digital de carbono orgânico do solo, Lamichhane *et al.* (2019) observaram que a divisão simples de dados e a validação cruzada são as estratégias mais frequentemente utilizadas na literatura.

Entretanto, a validação cruzada tradicional assume independência entre observações, o que nem sempre é válido em dados espaciais. Em MDS, a proximidade geográfica entre amostras pode gerar autocorrelação espacial, de modo que pontos próximos tendem a apresentar características semelhantes. Quando essa estrutura não é considerada, a divisão aleatória dos dados entre calibração e validação pode comprometer a independência entre os conjuntos,

resultando em estimativas de desempenho excessivamente otimistas (Roberts *et al.*, 2017). Por isso, em estudos de mapeamento digital de solos, em algumas situações, pode ser necessário manter dados geograficamente próximos juntos no mesmo conjunto no momento da validação, por exemplo, deixando um cluster fora por vez (*leave-one-cluster-out*).

Nesse contexto, diferentes estudos têm demonstrado o impacto da autocorrelação espacial na avaliação de modelos preditivos. Schratz *et al.* (2019) verificaram que modelos avaliados sem considerar a estrutura espacial dos dados tendem a apresentar métricas de desempenho mais elevadas do que aquelas obtidas com validação cruzada espacial. De forma semelhante, Ploton *et al.* (2020) observaram reduções expressivas no desempenho preditivo ao comparar validação aleatória com validação baseada em agrupamentos espaciais, com o coeficiente de determinação (R^2) diminuindo de 0,53 para 0,14. Diante disso, a validação cruzada espacial tem sido proposta como uma alternativa mais robusta, baseada na separação de blocos ou agrupamentos espaciais inteiros entre calibração e validação. Essa abordagem busca reduzir a dependência espacial entre os conjuntos de dados, minimizando os efeitos da autocorrelação espacial e proporcionando estimativas mais realistas da capacidade de generalização dos modelos. Com isso, sua adoção atende à recomendação de Hengl *et al.* (2018), segundo a qual a obtenção de estimativas mais realistas da acurácia de mapeamentos pode exigir o emprego de estratégias de validação cruzada mais rigorosas.

Por outro lado, a validação cruzada espacial também apresenta limitações. Wadoux *et al.* (2021) destacam que a exclusão de regiões inteiras do espaço geográfico pode levar à sub-representação de condições ambientais no conjunto de treinamento. Como consequência, o modelo pode ser avaliado em condições ambientais não observadas previamente, sendo forçado a extrapolar no espaço de covariáveis. Nesses casos, as métricas de desempenho tendem a ser excessivamente pessimistas, subestimando a real capacidade preditiva do modelo.

Além da validação adequada, o desempenho dos modelos em MDS depende fortemente da qualidade e representatividade dos dados utilizados na calibração. McBratney *et al.* (2003) ressaltam a importância de uma amostragem adequada em campo e da seleção criteriosa de covariáveis

ambientais, uma vez que esses elementos influenciam diretamente a capacidade de generalização dos modelos e ajudam a reduzir problemas de sobreajuste.

A seleção de covariáveis ambientais constitui uma etapa fundamental no MDS, pois busca identificar os preditores mais relevantes para explicar a variabilidade dos atributos do solo, reduzindo redundâncias e potenciais fontes de erro. Segundo Bouslihim *et al.* (2024), esse processo contribui para a diminuição da dimensionalidade dos dados e para o aumento da robustez dos modelos. Em seu estudo, a aplicação do algoritmo Boruta reduziu significativamente o conjunto inicial de covariáveis, selecionando apenas aquelas com maior relevância para a predição da matéria orgânica e do pH do solo. Diversos métodos têm sido utilizados para seleção de covariáveis em MDS. Manteghi *et al.* (2024) aplicaram análise de componentes principais (PCA), reduzindo um conjunto de 46 covariáveis para nove componentes que explicaram cerca de 85% da variância total. Zhang *et al.* (2023) também demonstraram reduções substanciais na dimensionalidade, variando de 392 covariáveis iniciais para subconjuntos entre 9 e 76 covariáveis, dependendo da abordagem adotada.

No contexto brasileiro, Meier *et al.* (2018) utilizaram eliminação recursiva de características para reduzir um conjunto de 73 covariáveis para apenas 10 covariáveis mais relevantes para o mapeamento de solos. De forma semelhante, Campos *et al.* (2019) avaliaram diferentes estratégias de seleção em uma bacia hidrográfica no Rio Grande do Sul e verificaram que apenas cerca de metade das covariáveis derivadas de modelo digital de elevação contribuíam efetivamente para a predição dos atributos do solo.

Além disso, a combinação de técnicas de aprendizado de máquina com seleção de covariáveis tem demonstrado bons resultados em MDS. Brungard *et al.* (2015) observaram que o uso do algoritmo *Random Forest* associado à eliminação recursiva de características resultou em alta precisão na predição de classes taxonômicas de solo.

Por fim, é importante reconhecer que algumas propriedades do solo ainda apresentam limitações para estimativa direta a partir de dados remotos. Dewitte *et al.* (2012) destacam que atributos como pH, capacidade de troca catiônica, saturação por bases e porosidade são difíceis de serem inferidos por

sensoriamento remoto, devido à natureza tridimensional do solo e à ausência de assinaturas espectrais suficientemente específicas para sua caracterização direta.

2.3 Mudanças climáticas e o estoque de carbono orgânico do solo

Frente ao cenário atual de preocupação com as mudanças climáticas, a avaliação do ECOS e os fatores que o afetam têm tido importância crucial no entendimento de como o solo pode atuar na mitigação dos impactos negativos da emissão de gases de efeito estufa (Xia *et al.*, 2010). O ciclo do carbono no sistema solo-planta-atmosfera é composto pelo carbono adicionado ao solo por meio dos resíduos vegetais (assimilado durante o processo de fotossíntese), o carbono emitido para a atmosfera na forma de CO₂ e o estoque de carbono orgânico que permanece no solo (COS). Com isso, as práticas de manejo têm papel fundamental na dinâmica deste elemento, uma vez que é possível reter carbono atmosférico no solo através dos sistemas de manejo com o intuito de auxiliar na mitigação do aquecimento global.

O plantio direto acoplado ao uso de culturas com elevada adição de massa verde ao solo contribui para o acréscimo de carbono no solo e o sistema de preparo convencional do solo e o uso de culturas de baixo aporte de massa verde resultam em balanço negativo de carbono no solo. Além disso, as emissões do elemento do solo para a atmosfera apresentam variação sazonal e relacionada à temperatura do solo (Costa *et al.*, 2008). Em locais onde os teores de carbono na camada mais superficial do solo (10 cm) são inicialmente baixos, não há influência significativa do aquecimento no estoque de carbono; entretanto, em áreas onde o estoque de carbono no solo é inicialmente maior, há perdas significativas desse elemento quando submetido ao aquecimento. Isso ocorre porque esses locais, quando submetidos a temperaturas mais elevadas, apresentam decomposição acelerada dos resíduos orgânicos, de forma que essa decomposição acelerada resulta em perdas de carbono mais expressivas do que a capacidade potencial de aporte de carbono gerado pelo aumento do crescimento vegetal. Dessa forma, regiões de alta latitude, que apresentam estoques iniciais de carbono maiores, tornam-se mais vulneráveis às mudanças climáticas (Crowther *et al.*, 2016).

A temperatura é apontada por Xia *et al.* (2010) como o fator climatológico que controla os estoques de carbono nos solos, de forma que a perda potencial de carbono do solo para a atmosfera é maior quando submetido a mudanças na temperatura do que quando submetido às mudanças no uso da terra; entretanto, o efeito aditivo de aquecimento e mudança no uso do solo torna as perdas de carbono no solo comparáveis às perdas por combustão de combustíveis fósseis.

Contudo, a avaliação do sequestro de carbono nos solos e o potencial de mitigação dos efeitos das mudanças climáticas devem levar em conta um aspecto importante, sendo ele o fato de que adotar, em escala global, medidas efetivas de sequestro de carbono no solo demanda tempo. Além disso, o potencial dos solos em armazenar carbono é limitado, pois o solo atingiria um novo ponto de equilíbrio, e a partir daí não seria mais tão efetivo em contribuir para a mitigação das mudanças climáticas (Sommer; Bossio, 2014). Por isso, os estudos que visam mapear os ECOS são de extrema importância para apoiar ações de conservação do solo, bem como promover o uso sustentável do recurso solo visando reduzir e/ou evitar as perdas do carbono que se acumulou no perfil de solo ao longo do tempo (Vasques *et al.*, 2010).

2.4 Mapeamento digital de solos para predição de estoque de carbono orgânico do solo

No Brasil, foi identificado maior acúmulo de COS por área nas regiões altas do sul e do sudeste do país, indicando que em regiões de clima mais ameno, há estabilização do COS a longo prazo (Gomes *et al.*, 2019). Guevara *et al.* (2018) estimaram um estoque de carbono médio no Brasil de aproximadamente 27 Pg. Na América Latina, zonas costeiras, úmidas, pastagens subtropicais, pastagens temperadas e áreas de cultivo apresentam estoques de carbono médios de 1,78 Pg, 0,35 Pg, 1,12 Pg, 3,04 Pg e 12,42 Pg, respectivamente. Já os estoques de carbono na região do Cerrado brasileiro encontram-se em torno de 37,5 Mg ha⁻¹ na camada de até 10 cm e de 129 Mg ha⁻¹ na camada até 100 cm em áreas de vegetação nativa e a conversão dessas áreas para outros usos tem ligação com a perda de carbono, principalmente na camada de 0 a 10 cm do solo (Oliveira *et al.*, 2023).

Ao analisar a distribuição de carbono em profundidade em perfis de solos em uma bacia hidrográfica na Flórida (EUA), Vasques *et al.* (2010) identificaram, além da tendência de redução nos teores de carbono total com o aumento da profundidade, que esse também foi influenciado pelo uso e tipo de solo, condições de drenagem e formação geológica. Em condições de drenagem deficiente houve maior acúmulo de carbono no solo quando comparado a áreas melhor drenadas, pois nas áreas onde a drenagem se dá de forma mais rápida, a falta de água pode prejudicar o crescimento vegetal, bem como promover a lixiviação de nutrientes e de material orgânico. Assim, os solos que se encontravam em áreas mais úmidas foram os que mais acumularam carbono no perfil, enquanto os Neossolos foram os que menos tiveram acúmulo do elemento no perfil, principalmente por serem solos arenosos e com baixa concentração de minerais reativos, o que dificulta a retenção de material orgânico.

Oliveira *et al.* (2023) argumentam que, em estudos que avaliem as variações do estoque de carbono no solo em função do uso e cobertura da terra, os estoques de carbono sob boas práticas de manejo sejam comparados com o uso anterior do solo, diferentemente do que tem sido adotado em muitas pesquisas, em que se comparam as diferentes práticas de uso e manejo do solo com áreas de vegetação nativa, além de que a maioria dos estudos de modelagem de estoque de carbono no solo apresentam carência no que diz respeito a dados de teor de carbono e densidade nas diferentes camadas amostradas. Outra dificuldade relatada por D'Ávila *et al.* (2019) na predição dos estoques de carbono na região dos Campos de Cima da Serra no RS foi a incerteza associada a áreas de transição entre diferentes classes de solos, de forma que as variações entre classes que ocorrem em uma curta distância só seriam captadas em levantamentos de solos mais detalhados.

Buscando prever os estoques de carbono dos solos do Brasil, Gomes *et al.* (2019) utilizaram dados legados de mais de 8000 perfis de solo e banco de dados com 74 covariáveis (que incluíam dados climáticos, mapa de solos, índices de vegetação e informações do relevo derivadas de MDE). O algoritmo *Random Forest* apresentou o melhor desempenho para prever os estoques de carbono no solo, sendo que as covariáveis mais importantes foram a classe de solo, precipitação, temperatura média mensal, declividade e índices de

vegetação. Cerca de 31% do carbono armazenado em solos brasileiros se encontra em áreas de preservação (onde não há pressão de uso exercida pelo ser humano), enquanto os maiores teores do elemento encontram-se em Organossolos e cerca de 57% do estoque de carbono encontra-se em solos intemperizados e pobres quimicamente, como é o caso dos Latossolos e Argissolos. Com esses achados, os autores concluíram que a metodologia utilizada para a predição de ECOS pode ser utilizada para realizar o inventário do mesmo em larga escala e ainda, destacam a importância das áreas de preservação ambiental na proteção do carbono do solo.

A análise da variação temporal dos estoques de carbono orgânico do solo (ECOS) é fundamental para o planejamento de estratégias de manejo sustentável, especialmente em áreas submetidas a intensas pressões antrópicas. No sudoeste de Camarões, Nguemezi *et al.* (2021) constataram uma redução de 34% nos ECOS ao longo de 32 anos, atribuída principalmente à urbanização acelerada e à expansão de áreas agrícolas. Os autores ressaltam que a compreensão dessas variações é crucial, uma vez que tanto os atributos climáticos quanto os relacionados ao relevo influenciam significativamente a predição dos estoques de carbono no solo. Oliveira *et al.* (2023) pontuam que o plantio direto auxilia na preservação do carbono presente no solo em profundidade, de forma a exercer um papel importante evitando a decomposição do elemento, ao invés de efetivamente aportar maior quantidade dele no solo. Desse modo, ao associar o uso de plantas de cobertura com o plantio direto pode-se ter efeito positivo no acréscimo de carbono no solo. Da mesma forma, a adoção de sistemas de produção integrados, como a integração lavoura-pecuária e lavoura-pecuária-floresta, se mostra eficiente no sequestro de carbono e na mitigação das mudanças climáticas. Ademais, áreas de pastagens naturais apresentaram maiores estoques de carbono até 30 cm do que áreas com pastagens intensivas e manejadas; além disso, áreas de pastagem com menor lotação animal podem ser pontos-chave para a preservação do carbono no solo (D'Ávila *et al.*, 2019).

Contudo, o desenvolvimento de sistemas de manejos voltados para a conservação dos solos esbarra na falta de difusão de informações e de mapeamentos detalhados de solos nos diferentes biomas do país. Sendo o Brasil um país de alta variabilidade de solos, que resulta em respostas variadas

dos solos frente às questões de manejo e de influência na dinâmica do carbono no solo e na atividade microbiológica, faz-se essencial a obtenção de informações com maior nível de detalhamento acerca dos solos para que se possam ter as melhores condições para apoiar as tomadas de decisões sobre planejamento agrícola, manejo e conservação do solo (Demattê *et al.*, 2022).

O *Random Forest* foi o algoritmo de mapeamento digital de solos mais utilizado no período compreendido entre 2010 e 2023 (Pouladi *et al.*, 2023). Esse algoritmo é conhecido por reduzir o ruído experimental e ter melhor precisão nas predições, isso ocorre porque, no treinamento do modelo, é gerado um número significativo de árvores aleatórias que contribuem para a previsão final (Gomes *et al.*, 2019). Enquanto covariáveis extraídas de Modelos Digitais de Elevação (MDE) e de imagens de sensoriamento remoto foram mais utilizadas como dados de entrada dos modelos de predição de carbono orgânico do solo, ao passo que as métricas mais populares para avaliar os modelos de predição foram o coeficiente de determinação (R^2) e a raiz do erro quadrático médio (RMSE) (Pouladi *et al.*, 2023).

2.5 Mapeamento digital de solos para predição de pH do solo

O pH do solo é um dos fatores fundamentais para o desenvolvimento das plantas e a sustentabilidade dos ecossistemas agrícolas. Ele influencia a disponibilidade de nutrientes, a atividade microbiológica e a estrutura do solo, afetando diretamente a produtividade das culturas. Portanto, compreender e monitorar essa variável é essencial para práticas agrícolas mais eficientes e sustentáveis.

É comum encontrar em áreas agrícolas, camadas superficiais e subsuperficiais do solo com pH ácido, resultando na redução da produtividade. Entretanto, a depender do tipo de solo e do manejo empregado, pode ser um desafio corrigir a acidez (Butterly *et al.*, 2022). Todavia, vale ressaltar que as causas da acidificação dos solos podem ser de ordem natural ou antropogênica, sendo que os maiores problemas da acidificação dos solos é a toxicidade por alumínio, que afeta o sistema radicular das plantas comprometendo o desenvolvimento, e a deficiência de cálcio, que afeta de maneira generalizada o crescimento vegetal (Hue, 2022).

Tradicionalmente, a análise do pH do solo envolve métodos laboratoriais convencionais, que, embora precisos, podem ser demorados. O avanço das tecnologias digitais têm permitido a automação e otimização desse processo por meio do mapeamento digital do pH do solo, que utiliza sensores, aprendizado de máquina e técnicas geoespaciais para gerar mapas detalhados da variabilidade química do solo.

Buscando otimização do uso de covariáveis ambientais para modelagem do pH do solo, Huang *et al.* (2025) verificaram que fatores topográficos, climáticos e biológicos são bons preditores da distribuição espacial do pH. Enquanto as temperaturas médias, evaporação, umidade média, elevação, e rede de drenagem tiveram grande influência nos valores de pH, o uso e tipo de solo não demonstraram efeito sobre o pH. Sendo possível concluir que em locais de maior variabilidade de terreno e melhores condições de drenagem, há menor lixiviação de cátions básicos, mantendo o pH mais elevado, enquanto em áreas mais influenciadas pela presença de água, o pH é menor devido ao excesso hídrico e acúmulo de substâncias ácidas.

De forma similar, Mousavi *et al.* (2023) também não encontraram diferença significativa de pH em áreas sob diferentes usos do solo, indicando que mudanças no uso do solo não levam a mudanças de pH, contudo seus achados demonstraram que a utilização do *Random Forest* combinado com krigagem ordinária apresenta bom desempenho preditivo para o pH do solo ($R^2 = 0,82$). Nesse estudo a predição do pH também foi altamente impactada por condições topográficas, principalmente elevação e índice de umidade topográfica, evidenciando mais uma vez a importância da redistribuição da água e da lixiviação na predição do atributo. Em estudo realizado no sul da Índia, Dharumarajan *et al.* (2017) identificaram o NDVI (*Normalized Difference Vegetation Index*), índice relacionado à cobertura vegetal, como covariável de maior importância para predição do pH, entretanto, o desempenho do modelo foi abaixo do esperado, com R^2 de 0,30.

2.6 Mapeamento digital de solos para predição da densidade do solo

A densidade do solo influencia processos essenciais, como retenção de água, crescimento das plantas e compactação do solo. Sua modelagem preditiva permite estimá-la sem análises laboratoriais demoradas, utilizando

dados ambientais e algoritmos avançados. Isso otimiza decisões em agricultura, conservação ambiental e geotecnia, promovendo a gestão sustentável dos solos. Contudo, a densidade do solo é uma propriedade menos visada para fins de predição no mapeamento digital de solos, todavia sabe-se que sua predição é altamente influenciada por condições climáticas (temperatura e precipitação), características de solo (classe), uso da terra e material parental (Chen *et al.*, 2022).

Em um estudo realizado na Irlanda, Bondi *et al.* (2018) propuseram dois modelos para estimar essa propriedade com base em descritores visuais, obtidos a partir de observações de campo sobre os perfis de solo. A estimativa da densidade por meio desses modelos poderia substituir análises laboratoriais tradicionais, considerando características húmicas, porosidade e pedogênese como principais variáveis preditoras. O primeiro modelo, baseado em árvore de decisão, apresentou uma precisão de aproximadamente 60%, enquanto o segundo, fundamentado em uma equação linear, obteve um coeficiente de correlação de cerca de 0,65. Em escala global, Hengl *et al.* (2017) alcançaram um desempenho preditivo de R^2 próximo a 75% para a densidade do solo.

Já na Hungria, Hateffard *et al.* (2023) investigaram a densidade do solo em duas áreas com características distintas. Na primeira área (com solo mais desenvolvido, bem agregado e com acúmulo de carbonatos), na etapa de calibração dos modelos preditivos, a metodologia *Random Forest* (RF) apresentou um desempenho inferior à regressão linear múltipla e às redes neurais, com R^2 de 80%, enquanto a regressão linear múltipla atingiu 99%. No entanto, na validação, o RF demonstrou melhor desempenho, com R^2 de 88%, além de registrar o menor erro médio quadrático (RMSE). Na segunda área (com solo arenoso, pouco agregado e pobre em carbono), durante a etapa de calibração, a regressão linear múltipla foi o modelo mais eficaz, alcançando R^2 acima de 90%, seguido pelo RF, que obteve R^2 em torno de 80% para a densidade do solo. Por fim, na validação dessa área, o RF se destacou, atingindo R^2 superior a 80% e registrando o menor RMSE, evidenciando sua eficiência na modelagem da densidade do solo. Zizala *et al.* (2022), por outro lado, constataram maior ocorrência de erros na predição em situações onde há baixa cobertura de dados, sendo a predição da densidade do solo particularmente afetada por isso. Isso resultou em menor previsibilidade do

atributo, culminando em $R^2=0,45$ para camadas superficiais do solo e $R^2=0,17$ para camadas mais profundas.

Em escalas mais detalhadas, como em um vinhedo no Chile, Mashalaba *et al.* (2020) alcançaram desempenho do modelo preditivo com R^2 de 0,26 a 0,51 em diferentes profundidades e mapas preditivos com baixa variação para o atributo. Com a utilização de redes neurais para realizar a predição da densidade do solo em nível nacional, na Escócia, Aitkenhead e Coull (2020) obtiveram elevada precisão, com R^2 de 0,79 e RMSE de 0,17. O modelo de *Random Forest* apresentou o melhor desempenho na predição da densidade do solo em zonas ripárias no Irã, superando outros algoritmos de aprendizado de máquina, incluindo redes neurais, com um coeficiente de determinação (R^2) de 0,62 (Zolfaghari Nia *et al.*, 2022).

2.7 Mapeamento digital de solos para predição da granulometria do solo

Diversos estudos têm explorado a predição das frações granulométricas do solo em diferentes regiões e profundidades, utilizando abordagens estatísticas e de aprendizado de máquina. Mousavi *et al.* (2023) identificaram que a predição apresentou melhor desempenho para o silte ($R^2 = 0,89$), seguido pela argila ($R^2 = 0,87$) e pela areia ($R^2 = 0,84$). Em relação à incerteza preditiva, os mesmos autores observaram que as frações areia e silte apresentaram os maiores níveis de incerteza, atingindo 35% e 46%, respectivamente, enquanto a predição da argila mostrou menor variabilidade, com incerteza máxima em torno de 18%.

Estudos anteriores corroboram esse comportamento. Forkuor *et al.* (2017), ao realizarem validação interna, identificaram que o silte teve o melhor desempenho preditivo ($R^2 = 53,8\%$), seguido pela areia ($R^2 = 34,2\%$) e pela argila ($R^2 = 18,5\%$), apontando uma menor precisão na estimativa das frações de argila. Esse padrão de desempenho foi semelhante ao obtido por Hengl *et al.* (2017), que alcançaram coeficientes de determinação superiores para a fração silte ($R^2 = 79,4\%$), seguida pela areia ($R^2 = 78,6\%$) e pela argila ($R^2 = 72,6\%$). Contudo, nesse caso, os valores de R^2 foram mais elevados.

Para predizer as frações granulométricas do solo, no Planalto de Loess, na China, He *et al.* (2024) utilizaram um modelo de floresta aleatória multivariada para que fosse possível captar as interrelações entre areia, silte e

argila. Resultados satisfatórios foram obtidos com essa abordagem, alcançando coeficientes de determinação (R^2) de 0,62, 0,53 e 0,73, respectivamente, para as frações. Nesse caso, as covariáveis climáticas foram as melhores preditoras da granulometria. Esses achados podem dar suporte efetivo para a gestão da erosão do solo em uma região crítica para a perda de solo no país.

A qualidade das predições das frações granulométricas do solo varia conforme a profundidade e o tipo de fração analisada. Jena *et al.* (2023) se propuseram a mapear as frações granulométricas do solo na Índia nas profundidades padronizadas do *GlobalSoilMap*, segundo Arrouays *et al.* (2014). Para a predição da areia, o desempenho do modelo piorava com o aumento da profundidade, sendo que na camada mais superficial de 0 a 5 cm, o R^2 foi de 0,30, passando para 0,28 na camada de 5 a 15 cm, 0,21 na camada de 15 a 30 cm, depois disso o modelo *Random Forest* teve dificuldade para fazer boas predições. Comportamento semelhante foi observado para o silte e para a argila. Contudo, a modelagem da fração silte mostrou melhores desempenhos preditivos e consistência até 60 cm de profundidade, diferente das demais frações onde os resultados são consistentes apenas até a camada de 15 a 30 cm. Com relação aos fatores mais importantes para as predições, houve o predomínio de variáveis climáticas nas predições da areia e da argila, enquanto para o silte, covariáveis derivadas de sensoriamento remoto e topográficas foram mais relevantes.

A análise comparativa de metodologias de modelagem preditiva das propriedades do solo em escala nacional, realizada por Patriche *et al.* (2023) para a Romênia, demonstrou que o pH e o carbono orgânico apresentam os mais altos níveis de previsibilidade. As frações granulométricas areia e silte exibiram desempenho preditivo intermediário, enquanto a argila apresentou baixa acurácia na modelagem. Contudo, na predição de atributos do solo, valores de R^2 na faixa de 30% a 50% são frequentemente reportados na literatura. O modelo *Random Forest* foi utilizado para predizer as frações granulométricas do solo nas seis camadas padrão do *GlobalSoilMap*, em escala nacional, para a Nigéria (Akpa *et al.*, 2014). Os melhores desempenhos preditivos foram observados nas camadas mais superficiais, em comparação com as de maior profundidade. Os resultados indicam que a integração das

diferentes variáveis preditoras pelo modelo foi capaz de explicar 16% a 53% da variação nos teores de argila, 21% a 48% da variação na areia e 21% a 26% na fração de silte.

2.8 Contexto geomorfológico e dinâmica costeira do litoral sul do Rio Grande do Sul

O litoral do estado do Rio Grande do Sul está localizado na província geomorfológica da Planície Costeira e é um ecossistema sedimentar extenso e dinâmico. A província costeira do estado é composta por dois grandes domínios geológicos: o embasamento e a Bacia de Pelotas; sendo esta última responsável pela configuração do litoral sul, cuja formação teve início no período Cretáceo, com a deposição contínua de sedimentos provenientes de áreas adjacentes. Ao longo do período Terciário, variações relativas do nível do mar intensificaram esse processo, resultando na formação de uma espessa sequência sedimentar, superior a 5 km, composta por materiais de diferentes granulometrias. Esses processos deram origem a uma diversidade de ambientes deposicionais, incluindo depósitos marinhos e eólicos, feições praias, sistemas lagunares e leques aluviais, caracterizando uma extensa planície de baixa altitude, geralmente inferior a 200 metros e datada da Era Cenozoica, com predomínio de sedimentos arenosos e forte atuação de ventos e ondas. Nesse contexto, destaca-se a presença do sistema lagunar Patos-Mirim, que ocupa uma área aproximada de 11.000 km² e constitui um dos principais elementos fisiográficos da região (Villwock, 1984).

A configuração atual da planície costeira está associada a quatro ciclos transgressivos e regressivos do nível do mar ocorridos durante o Quaternário, os quais foram responsáveis pela formação de sistemas deposicionais do tipo laguna-barreira, desenvolvidos durante os máximos transgressivos e preservados nas fases regressivas subsequentes (Barboza *et al.*, 2009). Essa dinâmica evidencia que os processos erosivos observados ao longo da costa são, em grande parte, naturais e resultam da própria interação entre oscilações do nível do mar e a dinâmica marinha. A costa é composta majoritariamente por depósitos quaternários pouco consolidados, o que contribui para sua suscetibilidade a mudanças, embora em alguns trechos, como na porção sul, haja relativo equilíbrio entre o aporte e a remoção de sedimentos, com

variações inferiores a 20 metros em determinadas décadas (Baitelli *et al.*, 1999). Ainda assim, extensas áreas do litoral apresentam erosão ativa, evidenciando a complexidade dos processos costeiros, influenciados pela ação de ondas, tempestades e variações do nível do mar.

Do ponto de vista geomorfológico, a linha de costa apresenta uma morfologia suavemente ondulada, sendo os campos de dunas as principais formas de relevo mais elevadas (Barboza; Rosa, 2014). Em alguns setores, há registros expressivos de recuo da linha de costa, como nas proximidades do farol da Conceição, onde se observou retração de cerca de 50 metros entre 1975 e 1995, correspondendo a uma taxa média de aproximadamente 2,5 metros por ano (Tomazelli *et al.*, 1998). Além das características físicas, o litoral sul do estado abriga ecossistemas de grande relevância ambiental, com destaque para banhados e áreas úmidas associadas às lagoas e cursos d'água, que dominam a paisagem regional. No entanto, essa área apresenta elevada fragilidade ambiental devido à sua formação geológica recente e à presença de solos arenosos pouco espessos, tornando-se suscetível a impactos decorrentes do uso intensivo da terra, especialmente por atividades como agricultura, pecuária e silvicultura com espécies exóticas (Portz *et al.*, 2011).

3 Capítulo I – Desempenho de estratégias de seleção de covariáveis no mapeamento digital do estoque de carbono orgânico do solo no Litoral Sul do Rio Grande do Sul

3.1 Introdução

A compreensão da dinâmica do carbono orgânico do solo (COS) é fundamental sob o ponto de vista da sustentabilidade ambiental de forma geral; entretanto, há grande destaque no seu estudo no ramo da agropecuária, já que os estoques de carbono orgânico do solo (ECOS) são suscetíveis às mudanças no uso e ocupação dos solos e aos manejos empregados visando à produção agropecuária. Além disso, do ponto de vista de mudanças climáticas, o COS e seu estoque são cruciais para o monitoramento do impacto das atividades humanas no clima em escalas local e global (Smith *et al.*, 2008; Stockmann *et al.*, 2013).

O COS também tem função vital na qualidade dos solos, uma vez que auxilia na melhoria da fertilidade e estrutura do solo, ciclagem de nutrientes e atividade biológica, promovendo ambientes mais propícios ao desenvolvimento de plantas, podendo ser um fator que contribui de forma expressiva para a garantia da segurança alimentar. Em áreas com solos arenosos, o carbono orgânico pode contribuir para fins de conservação do solo, uma vez que a matéria orgânica age como agente de agregação das partículas de solo, promovendo assim, melhoria da resistência do solo frente à erosão hídrica e eólica e da capacidade de armazenamento de água (Reichert *et al.*, 2016). Frente a isso, a promoção do aumento do ECOS nos solos é importante para reduzir os problemas relacionados à erosão e perda de solo, além de auxiliar na promoção da melhoria no armazenamento de água nesses solos.

Simulações a nível global sustentam a hipótese de que o aumento da temperatura no planeta favorecerá a perda líquida de carbono do solo para a atmosfera, acelerando ainda mais o aquecimento global (Crowther *et al.*, 2016). A região do bioma Pampa, no sul do Brasil, sofreu intensas modificações no seu uso, passando de áreas com predomínio de campos naturais e herbáceos, para áreas de cultivos anuais, com destaque para a soja (Overbeck *et al.*, 2007; Kuplich *et al.*, 2018; Lemos; Rizzi, 2020). Mesmo que o sul do Brasil tenha solos que apresentam maiores teores de carbono do que em outras

regiões do país, que têm clima mais quente, essa modificação no uso do solo, impulsionada pelas *commodities* agrícolas, resulta em impactos ambientais que ainda necessitam de estudos para serem melhor entendidos (Demattê *et al.*, 2022). Contudo, práticas como o plantio direto, uso de cultura de cobertura, adubação com corretivos orgânicos e sistemas de integração (lavoura-pecuária e lavoura-pecuária-floresta) podem compensar de forma parcial as perdas de carbono sofridas anteriormente pelos solos e com isso, auxiliar na mitigação dos efeitos do aquecimento global. Já que, de forma geral, essas práticas de agricultura climaticamente inteligentes tendem a resultar em acréscimo de carbono no solo (Oliveira *et al.*, 2023). Entretanto, o potencial dos solos em armazenar carbono é limitado, pois atingiria um novo ponto de equilíbrio a partir do qual não seria mais tão efetivo em contribuir para a mitigação das mudanças climáticas (Sommer; Bossio, 2014). Por isso, estudos voltados ao mapeamento do ECOS são fundamentais para subsidiar estratégias de conservação e promover o uso sustentável do solo, pois a disponibilidade de informações espaciais sobre os atributos do solo é essencial para embasar tomadas de decisão mais assertivas (Hateffard *et al.*, 2024).

Neste contexto, o mapeamento digital de solo (MDS) se apresenta como uma técnica de aplicação viável para estudos de análise e monitoramento espaciais e temporais de atributos e propriedades do solo, permitindo que se façam previsões com base em covariáveis ambientais que estejam relacionadas com o fenômeno que se deseja estudar. O MDS fundamenta-se no paradigma SCORPAN, proposto por McBratney *et al.* (2003), que estabelece o solo como função de outros atributos do solo (s), clima (c), organismos (o), relevo (r), material de origem (p), idade (a) e posição geográfica (n). Essa estrutura, derivada dos fatores de formação do solo propostos por Jenny (1941), tornou possível a integração entre dados de campo, laboratório e covariáveis ambientais derivadas de sensoriamento remoto e modelos digitais de elevação (MDE), que têm sido as fontes de preditores mais populares para a modelagem de propriedades do solo, enquanto o modelo de aprendizado de máquina *Random Forest* tem sido o mais popularmente empregado para fins de previsão de atributos do solo no contexto do MDS (Pouladi *et al.*, 2023; Lamichhane *et al.*, 2019), destacando-se por sua robustez, capacidade de lidar com relações não-lineares entre

covariáveis e bom desempenho mesmo com conjuntos de dados reduzidos (Breiman, 2001).

A região do Litoral Sul do Rio Grande do Sul, inserida na província geomorfológica da Planície Costeira, é caracterizada por uma extensa área de baixa altitude formada pela deposição de sedimentos marinhos e eólicos ao longo do Quaternário, com predomínio de solos arenosos (Villwock, 1984). Essa origem geológica recente confere à região elevada fragilidade ambiental, pois os solos apresentam baixa capacidade de retenção de água e nutrientes, elevada suscetibilidade à erosão e reduzida resiliência frente a distúrbios antrópicos (Portz *et al.*, 2011). Apesar dessa vulnerabilidade e da importância ecológica da região, que abriga o sistema lagunar Patos-Mirim (um dos maiores sistemas lagunares do mundo), não existem mapeamentos pedológicos detalhados que deem suporte ao planejamento e ao manejo adequado dos solos, lacuna que se insere em um contexto mais amplo de escassez de informações detalhadas sobre os solos brasileiros (Demattê *et al.*, 2022). No entanto, a aplicação do MDS nessa região esbarra em um desafio específico, pois em ambientes de baixa variação topográfica, a reduzida variabilidade das covariáveis de terreno pode limitar o poder preditivo dos modelos. Estudos conduzidos em áreas de relevo plano têm reportado resultados divergentes quanto à capacidade preditiva das covariáveis ambientais; enquanto alguns indicam fraca correlação entre covariáveis de sensoriamento remoto e o COS (John *et al.*, 2020), outros evidenciam boa relação entre o carbono orgânico e o índice de umidade topográfica derivado de MDE (Li *et al.*, 2020). Hateffard *et al.* (2023), em áreas de planície aluvial com diferentes gradientes de inclinação, observaram que as covariáveis mais influentes na predição do COS e ECOS variaram conforme as características do terreno, evidenciando que a importância dos preditores é dependente do contexto geomorfológico.

A incerteza quanto ao desempenho das covariáveis em áreas planas reforça a necessidade de investigar quais preditores são mais adequados para a modelagem do ECOS nesse contexto específico. Por isso, a seleção de covariáveis ambientais constitui uma etapa fundamental do MDS (McBratney *et al.*, 2003), pois permite identificar os preditores mais relevantes, reduzir a redundância entre covariáveis e obter modelos mais parcimoniosos e

interpretáveis. Estudos demonstram que conjuntos reduzidos de covariáveis podem apresentar desempenho igual ou superior aos modelos com conjuntos completos, sendo que conjuntos contendo de 15 a 20 covariáveis tendem a apresentar bons resultados (Poggio *et al.*, 2021). Contudo, a eficácia dessas estratégias de seleção ainda é pouco explorada em ambientes costeiros de baixa variação topográfica, onde a contribuição relativa das covariáveis topográficas e espectrais pode diferir substancialmente do observado em áreas de relevo mais movimentado.

Portanto, entender a dinâmica e distribuição do ECOS nos solos é crucial para subsidiar o planejamento estratégico nos diferentes níveis organizacionais visando à promoção do sequestro de carbono no solo. Dessa forma, será promovido o armazenamento de carbono de forma mais estável, auxiliando na mitigação dos efeitos do aquecimento global. Além disso, a elaboração de metodologias para predição desses estoques e o desenvolvimento de mapas preditivos são fundamentais para dar suporte a uma agricultura mais sustentável, já que o solo é o maior reservatório terrestre de carbono orgânico, superior ao carbono acumulado na atmosfera e na vegetação (Lal, 2004; Scharlemann *et al.*, 2014). Nesse sentido, o estudo adota as profundidades padrão estabelecidas pela iniciativa *GlobalSoilMap*, que fornece um referencial metodológico internacional para a padronização do mapeamento de solos em profundidade, atendendo à recomendação de que o uso de métodos padronizados de amostragem e modelagem é essencial para a comparabilidade entre estudos em diferentes regiões do mundo (Arrouays *et al.*, 2014; Scharlemann *et al.*, 2014).

Apesar da importância do ECOS e dos avanços do MDS, estudos que abordem conjuntamente a predição do ECOS e a avaliação de estratégias de seleção de covariáveis em solos arenosos costeiros de baixa variação topográfica são escassos. Diante disso, este estudo propõe-se a modelar e mapear o estoque de carbono orgânico do solo (ECOS) na região do Litoral Sul do Rio Grande do Sul. Hipotetiza-se que seja possível predizer, com precisão aceitável, o ECOS para áreas costeiras de relevo plano e solos predominantemente arenosos, a partir de algoritmo *Random Forest* (RF) e com o uso de covariáveis ambientais topográficas e espectrais derivadas de sensoriamento remoto. Além disso, este capítulo avalia diferentes estratégias

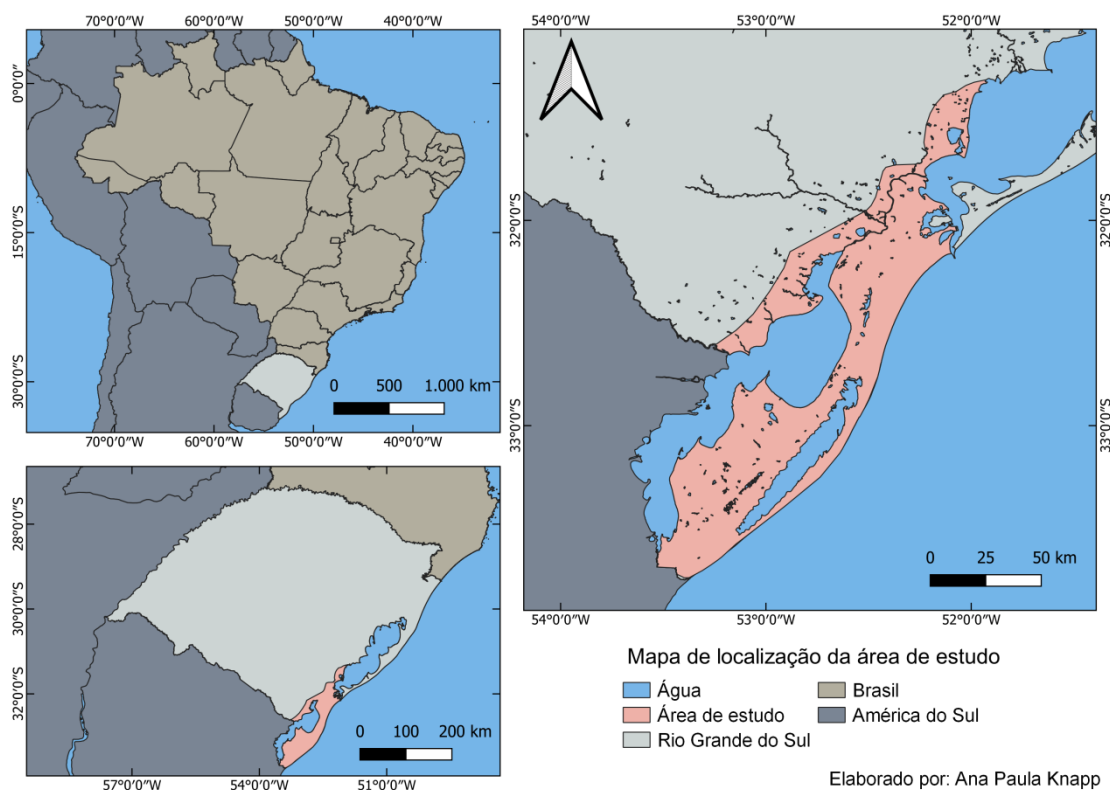
de seleção de covariáveis ambientais e o efeito da redução do número de preditores sobre o desempenho dos modelos preditivos. Com isso, busca-se ampliar o conhecimento metodológico relacionado ao mapeamento digital do ECOS e à compreensão de sua distribuição espacial na região estudada, proporcionando subsídios para manejo sustentável do solo e a formulação de políticas voltadas à conservação do solo e à mitigação das mudanças climáticas.

3.2 Metodologia

3.2.1 Área de estudo

Localizada no Litoral Sul do estado do Rio Grande do Sul, no Brasil, a área de estudo (Figura 1) possui uma extensão de 10.393,9 km² e um perímetro de 1.190,5 km, estende-se do município de Pelotas ao município do Chuí, abrangendo as regiões da Lagoa Mirim, Lagoa Mangueira e parte da Laguna dos Patos. Segundo a classificação climática de Köppen, o clima da região é temperado quente (Cfb) e subtropical (Cfa), com temperaturas médias anuais entre 17,1 e 19 °C e precipitação anual variando de 1.200 a 1.400 mm (Wrege *et al.*, 2012). Mais informações sobre a área de estudo podem ser visualizadas na seção de apêndices (Apêndice A, Apêndice B e Apêndice C).

Figura 1 - Delimitação e localização geográfica da área de estudo no litoral sul do estado do Rio Grande do Sul, Brasil



Fonte: Elaborado pela autora (2025).

A região apresenta relevância ecológica, abrigando importantes ecossistemas costeiros como as áreas de banhados, além de apresentar diversidade no uso da terra, incluindo atividades agrícolas e áreas de conservação, que podem influenciar significativamente no ciclo do carbono e, consequentemente, no ECOS. Contudo, apesar da importância ambiental e agrícola dessa região, ainda existem poucas informações especializadas sobre a distribuição do ECOS em seus solos predominantemente arenosos.

A área pertence à província geomorfológica da Planície Costeira, caracterizada por altitudes inferiores a 40 metros. Sua formação resulta de sucessivos ciclos de avanço e recuo do nível do mar, moldando uma paisagem dominada por feições praias, lagunas e cordões arenosos. Esse contexto geoclimático deu origem a depósitos de sedimentos inconsolidados, em áreas consideradas as mais recentes formações do estado. A região inclui a Planície Costeira Interna, situada entre o Escudo Sul-Rio-Grandense e a Laguna dos Patos, e a Planície Costeira Externa, localizada junto ao Oceano Atlântico. Essa divisão influencia a variabilidade dos solos de modo que em altitudes mais elevadas e terrenos suavemente ondulados, predominam Argissolos e

Chernossolos, enquanto em áreas de relevo plano predominam Planossolos, Gleissolos e Neossolos Quartzarênicos (Streck *et al.*, 2018).

Os solos de terras baixas, como aqueles encontrados na Planície Costeira do Rio Grande do Sul, ocupam áreas de relevo plano a suavemente ondulado e apresentam características que variam em função do material de origem e da posição na paisagem. Esses solos são marcados pela presença de hidromorfismo, condicionado principalmente pelo relevo e pela proximidade do lençol freático, o que resulta em diferentes graus de redução e oxidação, refletidos em atributos morfológicos como colorações acinzentadas, mosqueados e tonalidades brunadas. Os solos de terras baixas geralmente apresentam características comuns, como horizonte A pouco espesso, elevada densidade natural, presença de camadas adensadas, drenagem deficiente e baixa fertilidade natural. No Rio Grande do Sul, as principais classes de solos de terras baixas são os Planossolos e os Gleissolos, que frequentemente ocorrem associados e representam mais de 60% das ocorrências nesses ambientes (Sousa *et al.*, 2023).

Na área de estudo, predominam as Unidades de Mapeamento (UM) Tuia, Pelotas e Lagoa, que correspondem, respectivamente, ao Argissolo Vermelho-Amarelo, Planossolo Háplico e Neossolo Quartzarênico. Os solos da UM Tuia desenvolvem-se, em geral, nas posições mais elevadas da paisagem, em cotas superiores às de ocorrência da UM Pelotas. Originados a partir de sedimentos arenosos, apresentam textura arenosa nos horizontes superficiais e textura variando de franco-arenosa a franco-argiloarenosa no horizonte B. A sequência típica de horizontes é A/E/Bt/C, sendo comum horizonte A espesso e horizonte E bem desenvolvido. Devido ao material de origem, esses solos apresentam baixos teores de matéria orgânica, reduzida capacidade de troca de cátions (CTC) e baixa disponibilidade natural de nutrientes. Os solos da UM Pelotas são derivados de sedimentos lagunares e lacustres e ocorrem em relevo plano a suavemente ondulado. Sua principal característica é a presença de horizonte B plânico, um tipo especial de horizonte B textural de baixa permeabilidade, responsável pela drenagem deficiente característica desses solos. As sequências de horizontes mais frequentes são A/Btg/Cg e A/E/Btg/Cg, evidenciando a influência do hidromorfismo em sua gênese. Por sua vez, os solos da UM Lagoa são profundos, arenosos e apresentam

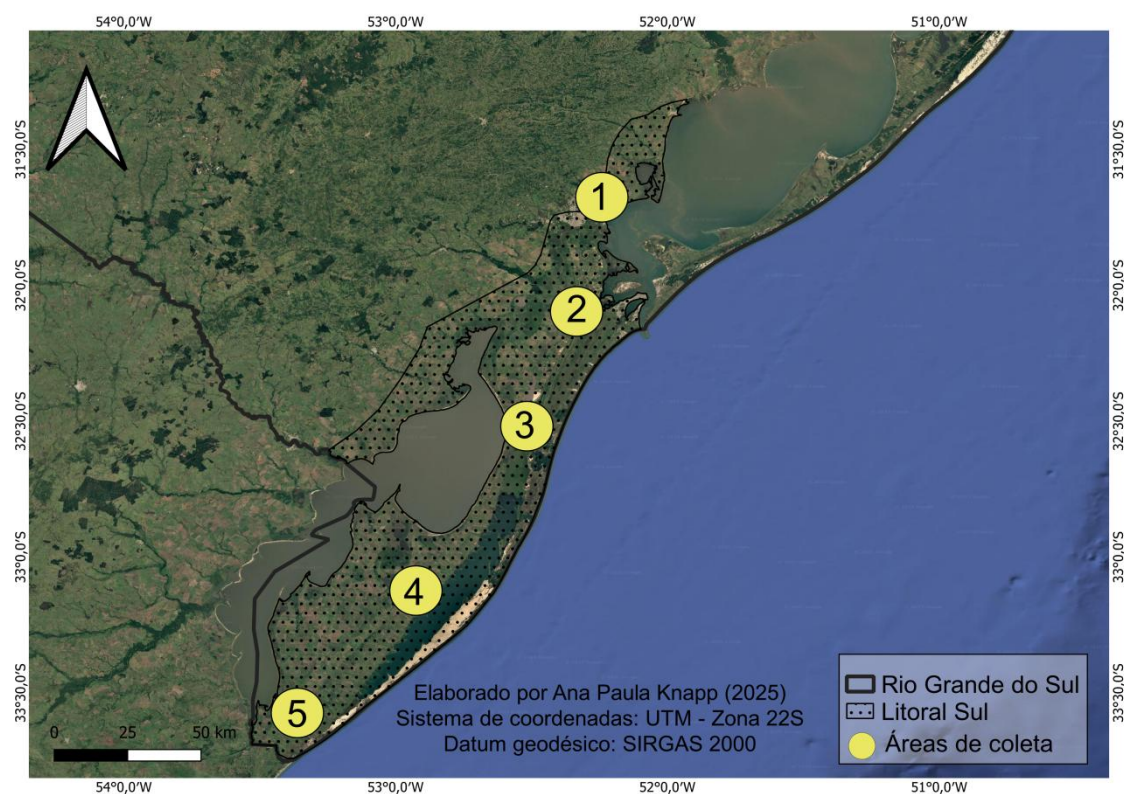
drenagem deficiente. Sua sequência de horizontes é A/C, sendo o horizonte superficial mais escuro em razão do maior acúmulo de matéria orgânica. Embora sejam solos com teores maiores de matéria orgânica, elevando a CTC, a saturação por bases é baixa, refletindo a baixa fertilidade natural (Miguel *et al.*, 2025).

3.2.2 Amostragem de solo

Os pontos de amostragem foram definidos utilizando o método do Hipercubo Latino Condicionado (cLHS) (Minasny; McBratney, 2006), implementado no software R (R Core Team, 2018) por meio do pacote cLHS (Roudier, 2014). Essa metodologia considerou o Modelo Digital de Elevação (MDE) *Shuttle Radar Topography Mission* (SRTM) (Farr *et al.*, 2007), com 30 metros de resolução espacial, o mapa geológico do estado do Rio Grande do Sul, em escala 1:750.000 (CPRM, 2008) e imagens interpretadas de uso e cobertura da terra, em escala 1:250.000, obtidas através da plataforma MapBiomas (2023), como referência para a seleção dos pontos de amostragem. O método cLHS foi adotado neste trabalho por sua capacidade de selecionar um conjunto reduzido de amostras que, ainda assim, representa adequadamente a variabilidade ambiental e a distribuição espacial da área de estudo (Gelsleichter *et al.*, 2025).

Devido à grande extensão da área de estudo, delimitaram-se 5 áreas de amostragem com raio de 15 km por meio do cLHS (Figura 2). Dentro de cada uma dessas áreas, determinaram-se os pontos de coleta também por meio do método cLHS. A definição dos pontos de coleta dentro de cada área baseou-se em um arquivo contendo estradas primárias, secundárias e terciárias, de modo que os pontos de coleta estivessem a até 10 metros do eixo das estradas.

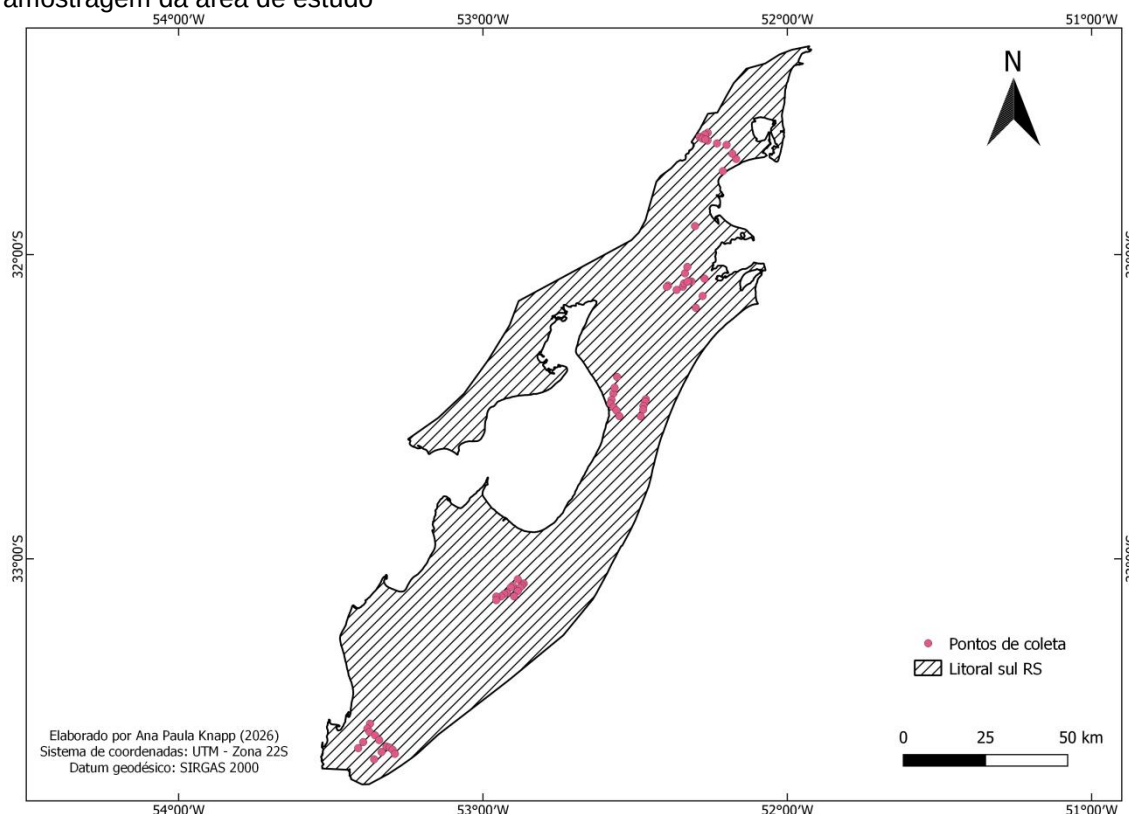
Figura 2 - Distribuição das cinco áreas de amostragem delimitadas pelo algoritmo cLHS na área de estudo, cada uma com raio de 15 km.



Fonte: Elaborado pela autora (2025).

Foram coletadas 64 amostras de solo na área de estudo (Figura 3), distribuídas entre as cinco áreas de amostragem, com 11 pontos na área 1, 14 pontos na área 2 e 13 pontos em cada uma das áreas 3, 4 e 5. A densidade amostral nas áreas de coleta variou de 0,0156 a 0,0198 pontos por km^2 (um ponto para cada 64,3 a 50,5 km^2), resultando em densidade amostral média de 0,0062 pontos por km^2 para a área de estudo, equivalente a um ponto para cada 162,4 km^2 . Nos casos em que o acesso aos pontos previamente selecionados pelo cLHS não foi possível, a localização foi ajustada em campo com auxílio de GPS, mantendo-se a maior proximidade possível em relação ao ponto original.

Figura 3 - Distribuição espacial dos 64 pontos de coleta de solo nas cinco áreas de amostragem da área de estudo



Fonte: Elaborado pela autora (2026).

Em cada ponto, foram coletadas amostras de solo em cinco profundidades: 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm e 60-100 cm, conforme especificado por Arrouays *et al.* (2014), resultando em um total de 320 amostras. Em cada ponto, a cada profundidade, foram coletadas amostras deformadas para as análises de determinação de carbono orgânico e amostras indeformadas para determinação da densidade do solo, visando o posterior cálculo do ECOS. As amostras indeformadas foram coletadas com o auxílio de anéis volumétricos com 2 cm de altura e 6 cm de diâmetro.

Para cada ponto de amostragem, foi realizada uma descrição detalhada do local, contemplando informações como a classe de solo até o segundo nível categórico do SiBCS (Santos *et al.*, 2025), uso atual da terra, o relevo e a declividade, a espessura do horizonte superficial, a profundidade do horizonte B e a espessura do solum, com o objetivo de caracterizar de forma abrangente as condições específicas de cada local de coleta.

A caracterização pedológica dos pontos de coleta revelou predomínio de Planossolos (n=32), seguidos por Argissolos (n=22) e Neossolos (n=7), com

ocorrência pontual de Espodossolo, Chernossolo e Gleissolo. Essa distribuição reflete a diversidade pedológica da Planície Costeira, onde a variação da relação solo-paisagem, mesmo em condições de relevo predominantemente plano, condiciona a formação de diferentes classes de solo. A espessura média do horizonte A variou expressivamente entre as classes, de 15 cm (Chernossolo) a 90 cm (Espodossolo), enquanto o limite superior do horizonte B variou de 20 a 155 cm (Tabela 1). Os Planossolos, classe mais representada pelos pontos de coleta, apresentaram horizonte A com espessura média de 39,3 cm e limite superior do horizonte B em torno de 48,8 cm, característica típica da sequência A/E/Bt desses solos em áreas de relevo plano. Os Argissolos, por sua vez, exibiram horizonte A mais espesso (média de 53,6 cm) e limite superior do B mais profundo (média de 67,0 cm), coerente com a maior diferenciação textural entre os horizontes superficiais e subsuperficiais dessa classe.

Tabela 1 - Caracterização morfológica dos 64 pontos de coleta de solo distribuídos na área de estudo, conforme classes de solo do SiBCS

Classe de solo (SiBCS)	n	Relevo predominante	Espessura do horizonte A (cm)	Limite superior do horizonte B (cm)
Planossolo	32	Plano	39,3 (20–80)*	48,8 (20–130)
Argissolo	22	Plano a suave ondulado	53,6 (25–120)*	67,0 (25–155)
Neossolo	7	Plano	54,1 (30–90)*	NA *
Espodossolo	1	Plano	90	90
Chernossolo	1	Plano	15	35
Gleissolo	1	Plano	40	40

* Valores apresentados como média (amplitude). *Neossolos não apresentam horizonte B diagnóstico.

Fonte: Elaborado pela autora (2026).

A densidade do solo, determinada pelo método do anel volumétrico em cada uma das cinco profundidades avaliadas, apresentou comportamento distinto entre as classes pedológicas (Tabela 2). De modo geral, observou-se aumento da densidade do solo com a profundidade, refletindo o adensamento natural dos horizontes subsuperficiais. Os Argissolos atingiram valores médios

de $1,66 \text{ g cm}^{-3}$ na camada de 30–60 cm, associados à presença de horizontes B texturais mais argilosos. Os Planossolos, por sua vez, apresentaram densidade média de $1,65 \text{ g cm}^{-3}$ nessa mesma camada, coerente com a mudança textural que caracteriza essa classe. Os Neossolos exibiram valores menores de densidade (média de $1,31$ a $1,47 \text{ g cm}^{-3}$), atribuído à textura arenosa e à menor diferenciação entre horizontes.

Tabela 2 - Densidade do solo média (g cm^{-3}) por classe de solo e profundidade na área de estudo

Classe de solo	0–5 cm	5–15 cm	15–30 cm	30–60 cm	60–100 cm
Planossolo (n=32)	1,43	1,59	1,63	1,65	1,62
Argissolo (n=22)	1,39	1,57	1,63	1,66	1,60
Neossolo (n=7)	1,31	1,40	1,39	1,47	1,47
Espodossolo (n=1)*	1,39	1,60	1,52	1,53	1,55
Chernossolo (n=1)*	1,37	1,56	1,39	1,51	1,64
Gleissolo (n=1)*	1,25	1,47	1,39	1,68	1,62

Fonte: Elaborado pela autora (2026). * Classes com n=1 representam observações individuais.

O uso da terra nos pontos de coleta reflete o padrão característico da região da Planície Costeira Sul (Tabela 3). Áreas de campo predominaram, abrangendo 60,9% dos pontos, seguido pela agricultura (31,3%), representada principalmente pelo cultivo de soja e arroz. Os demais usos (silvicultura, integração lavoura-pecuária (ILP) e pastagem) somaram 7,8% das amostras. Essa diversidade de usos é relevante para a modelagem do ECOS, uma vez que o manejo e a cobertura do solo influenciam diretamente os aportes de matéria orgânica e a dinâmica do carbono nas diferentes camadas avaliadas.

Tabela 3 - Distribuição dos 64 pontos de coleta de solo conforme uso da terra na área de estudo

Uso da terra	n	%
Campo	39	60,9
Agricultura	20	31,3
Outros	5	7,8

Fonte: Elaborado pela autora (2026).

Informações complementares sobre os pontos de coleta estão disponíveis na seção de apêndices (Apêndice D, Apêndice E, Apêndice F, Apêndice G e Apêndice H).

3.2.3 Análises laboratoriais

O conteúdo de carbono orgânico do solo (COS) foi determinado pelo método de oxidação úmida de Walkley-Black (Tedesco *et al.*, 1995). O ECOS para cada uma das profundidades analisadas foi calculado pela Equação 1, conforme metodologia descrita por Veldkamp (1994), Ellert e Bettany (1995), Sisti *et al.* (2004), Fernandes e Fernandes (2008) e Hateffard *et al.* (2023).

$$ECOS = (C * BD * SD)/10 \quad (1)$$

Em que:

ECOS: Estoque de carbono orgânico na profundidade considerada ($Mg\ ha^{-1}$);

C: Teor de carbono orgânico na profundidade analisada ($g\ kg^{-1}$);

BD: Densidade do solo na profundidade analisada ($g\ cm^{-3}$); e

SD: Espessura da camada amostrada (cm).

A densidade do solo foi determinada pelo método do anel volumétrico (Teixeira *et al.*, 2017), de acordo com a equação 2.

$$BD = \frac{m}{V} \quad (2)$$

Em que:

BD: Densidade do solo ($g\ cm^{-3}$);

m: massa seca da amostra (g); e

V: Volume do anel (cm^3).

3.2.4 Covariáveis ambientais

Foram geradas 48 covariáveis ambientais para a predição do estoque de carbono orgânico do solo (ECOS). Destas, 33 correspondem a covariáveis de terreno derivadas do MDE *Shuttle Radar Topography Mission* (SRTM), com resolução espacial de 30 metros (Farr *et al.*, 2007). Adicionalmente, foram

obtidas 15 covariáveis espectrais derivadas de imagens de sensoriamento remoto dos satélites Landsat 8-9/OLI, sendo nove índices e seis bandas espectrais. As imagens Landsat utilizadas para a extração das covariáveis ambientais datam de janeiro de 2025. As covariáveis de sensoriamento remoto foram obtidas a partir de imagens multiespectrais do satélite Landsat, compreendendo seis bandas espectrais: azul (B2), verde (B3), vermelho (B4), infravermelho próximo - NIR (B5) e infravermelho de ondas curtas - SWIR1 (B6) e SWIR2 (B7).

O processamento e a padronização das covariáveis (Tabela 4) foram conduzidos em ambiente R (R Core Team, 2024), utilizando os pacotes *raster* (Hijmans, 2022), *rgdal* (Bivand et al., 2021) e *RSAGA* (Brenning et al., 2025).

Tabela 4 - Covariáveis ambientais utilizadas na modelagem do estoque de carbono orgânico do solo (ECOS)

Covariável	Sigla	Categoria	Descrição
<i>Aspect</i>	asp	Morfométrica	Orientação do declive
<i>Bare Soil Index</i>	BSI	Índice espectral	Índice de solo exposto
<i>Blue</i>	b2blue	Banda espectral	Banda azul do sensor do satélite
<i>Clay Minerals Ratio</i>	CMR	Índice espectral	Razão de minerais de argila
<i>Convergence Index</i>	ci	Morfométrica	Índice de convergência/divergência do escoamento
<i>Cross-Sectional Curvature</i>	csc	Morfométrica	Curvatura perpendicular ao declive
<i>Elevation</i>	mde	Morfométrica	Elevação
<i>Diurnal Anisotropic Heating</i>	dah	Morfométrica	Medição contínua da energia dependente da exposição
<i>Easternness</i>	e	Morfométrica	Seno de aspecto
<i>Flow Line Curvature</i>	flc	Morfométrica	Curvatura ao longo da linha de fluxo
<i>General Curvature</i>	gc	Morfométrica	Curvatura combinada do plano e perfil
<i>Green</i>	b3green	Banda espectral	Banda verde do sensor do satélite
<i>Longitudinal Curvature</i>	lc	Morfométrica	Curvatura na direção do declive
<i>LS Factor</i>	ls	Morfométrica	Fator topográfico

<i>Mass Balance Index</i>	mbi	Morfométrica	Índice de equilíbrio entre erosão e deposição
<i>Maximal Curvature</i>	maxc	Morfométrica	Curvatura máxima na seção normal local
<i>Mid-Slope Position</i>	msh	Morfométrica	Distância relativa entre topo e vale
<i>Minimal Curvature</i>	minc	Morfométrica	Curvatura mínima na seção normal local
<i>Multiresolution Index of Ridge Top Flatness</i>	mrrtf	Morfométrica	Índice de planicidade no topo de cristas
<i>Multiresolution Index of Valley Bottom Flatness</i>	mrvmf	Morfométrica	Índice de planicidade do fundo do vale
<i>Near Infrared</i>	b5nir	Banda espectral	Banda infravermelho próximo
<i>Northernness</i>	n	Morfométrica	Cosseno de aspecto
<i>Normalized Difference Moisture Index</i>	NDMI	Índice espectral	Índice de umidade por diferença normalizada
<i>Normalized Difference Tillage Index</i>	NDTI	Índice espectral	Índice de preparo do solo
<i>Normalized Difference Vegetation Index</i>	NDVI	Índice espectral	Índice de vegetação
<i>Normalized Difference Water Index</i>	NDWI	Índice espectral	Índice de água
<i>Normalized Height</i>	nh	Morfométrica	Altura normalizada entre base e cume
<i>Normalized Pigment Chlorophyll Ratio Index</i>	NPCRI	Índice espectral	Índice de clorofila
<i>Plan Curvature</i>	planc	Morfométrica	Curvatura da linha de contorno hipotética
<i>Profile Curvature</i>	proc	Morfométrica	Curvatura ao longo da inclinação mais íngreme
<i>Real Surface Area</i>	rsa	Morfométrica	Área real da superfície
<i>Red</i>	b4red	Banda espectral	Banda vermelha do sensor do satélite
<i>Short-Wave Infrared</i>	b6swir1	Banda espectral	Banda infravermelho de ondas curtas
<i>Short-Wave Infrared</i>	b7swir2	Banda espectral	Banda infravermelho de ondas curtas
<i>Slope</i>	s	Morfométrica	Declividade angular local
<i>Slope Height</i>	sh	Morfométrica	Distância vertical entre base e cume da encosta
<i>Soil Adjusted Vegetation Index</i>	SAVI	Índice espectral	Índice de vegetação ajustado para solo
<i>Soil Brightness Index</i>	SBI	Índice espectral	Índice de brilho do solo

<i>Standardized Height</i>	stanh	Morfométrica	Altura padronizada
<i>Surface Specific Points</i>	ssp	Morfométrica	Diferenças entre pontos de deslocamento
<i>Tangential Curvature</i>	tanc	Morfométrica	Curvatura perpendicular ao gradiente
<i>Terrain Ruggedness Index</i>	tri	Morfométrica	Índice de rugosidade do terreno
<i>Terrain Surface Convexity</i>	tsc	Morfométrica	Índice de convexidade da superfície
<i>Topographic Position Index</i>	tpi	Morfométrica	Diferença entre elevação de um ponto e vizinhança
<i>Topographic Wetness Index</i>	twi	Morfométrica	Índice de umidade baseado no relevo
<i>Total Curvature</i>	tc	Morfométrica	Medida geral da curvatura
<i>Valley Depth</i>	vd	Morfométrica	Profundidade relativa do vale
<i>Vector Ruggedness Measure</i>	vrn	Morfométrica	Mede a variação da rugosidade do terreno

Fonte: Elaborado pela autora (2026), adaptado de Nachtigall (2022).

Todas as 48 covariáveis do tipo raster foram compiladas abrangendo toda a extensão da área de estudo. Todas as camadas foram padronizadas previamente às análises, sendo reamostradas para uma mesma resolução espacial e ajustadas a uma grade comum. Além disso, as covariáveis foram reprojetadas para um sistema de coordenadas único, garantindo consistência espacial entre as variáveis. Por fim, foi verificada a cobertura completa da área de estudo por todas as camadas, assegurando a ausência de lacunas espaciais nos dados utilizados na modelagem.

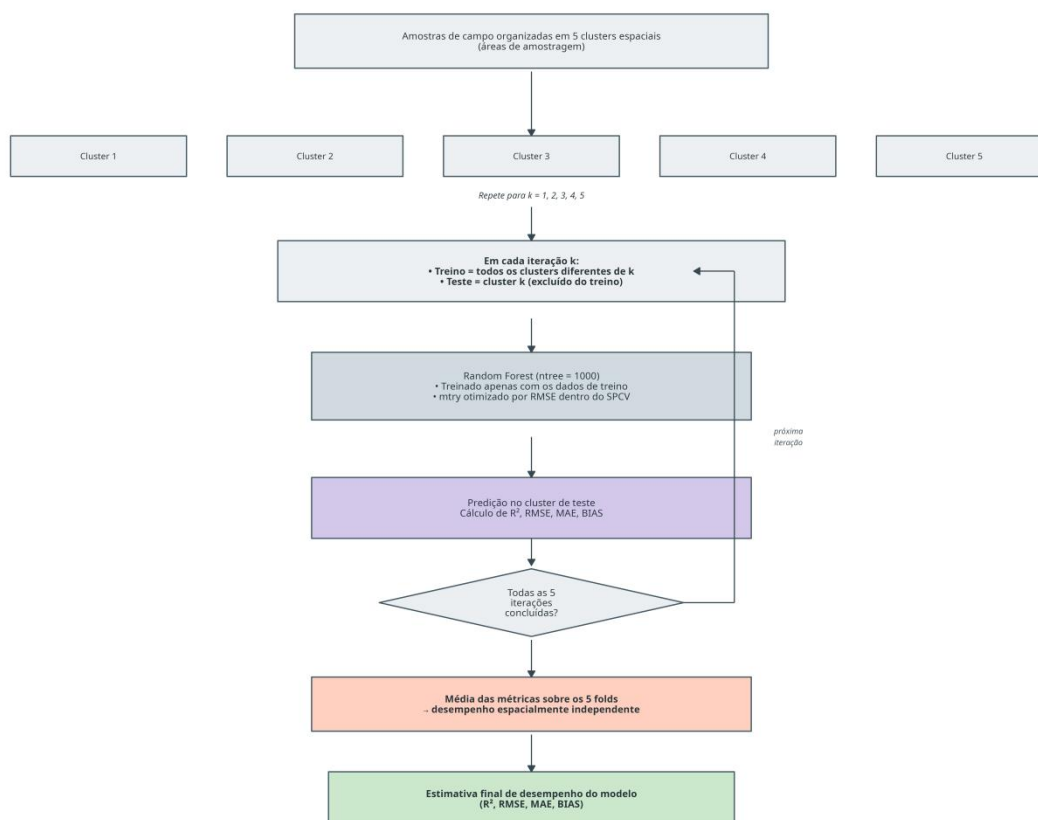
3.2.5 Calibração e validação

A modelagem do ECOS foi realizada por meio do método *Random Forest* (RF) (Breiman, 2001; Liaw; Wiener, 2002), um método de aprendizado de máquina supervisionado do tipo *ensemble* que constrói um conjunto de árvores de decisão independentes. Cada árvore é treinada a partir de uma subamostra aleatória dos dados originais (*bootstrap*), e, em cada nó de divisão, apenas um subconjunto aleatório de covariáveis (*mtry*) é considerado como candidato, o que reduz a correlação entre as árvores e aumenta a robustez das predições. A predição final é obtida pela agregação das predições individuais (média).

Os modelos foram ajustados com o pacote *randomForest* em ambiente R, fixando $n_{tree} = 1.000$. O hiperparâmetro $mtry$ foi otimizado por *grid search* sob validação cruzada espacial, minimizando o RMSE médio entre os folds.

A avaliação da capacidade preditiva dos modelos foi realizada por meio de validação cruzada espacial (*Spatial Cross-Validation* – SPCV), empregando a estratégia *Leave-One-Cluster-Out* (LOCO). Inicialmente, as amostras foram agrupadas em clusters espaciais, de forma que cada cluster representasse uma das cinco áreas de coleta para fins de validação. Em cada iteração, um cluster (área de coleta) foi completamente removido do conjunto de treinamento e utilizado exclusivamente como conjunto de teste, enquanto todos os demais clusters (áreas de coleta) foram empregados para ajustar o modelo RF. Esse procedimento foi repetido sucessivamente até que todos os clusters tivessem sido utilizados uma vez como conjunto de validação (Figura 4). A adoção da validação cruzada espacial visa reduzir os efeitos da autocorrelação espacial entre amostras de treinamento e teste, fornecendo uma estimativa mais realista da capacidade de generalização do modelo em áreas não amostradas (Roberts *et al.*, 2017).

Figura 4 - Representação esquemática do processo de validação cruzada espacial *leave-one-cluster-out*



Fonte: Elaborado pela autora (2026).

Para cada iteração foram obtidas previsões para as observações do cluster de validação, sendo posteriormente calculadas métricas de desempenho, incluindo o coeficiente de determinação (R^2) (Equação 3), a raiz do erro quadrático médio (RMSE) (Equação 4), o erro absoluto médio (MAE) (Equação 5) e o viés (BIAS) (Equação 6). Os valores finais de desempenho foram obtidos a partir da média das métricas calculadas em todos os ciclos da validação cruzada espacial. Sendo assim, todas as métricas de desempenho reportadas neste estudo correspondem à capacidade preditiva dos modelos na validação (SPCV).

$$R^2 = \frac{\sum_{i=1}^n (\text{pred} - \text{mobs})^2}{\sum_{i=1}^n (\text{obs} - \text{mobs})^2} \quad (3)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{pred} - \text{obs})^2} \quad (4)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |\text{obs} - \text{pred}| \quad (5)$$

$$\text{BIAS} = \frac{1}{n} \sum_{i=1}^n (\text{pred} - \text{obs}) \quad (6)$$

Em que:

Pred: é o valor predito;

Mobs: é o valor médio observado;

Obs: é o valor observado;

n: é o número de amostras.

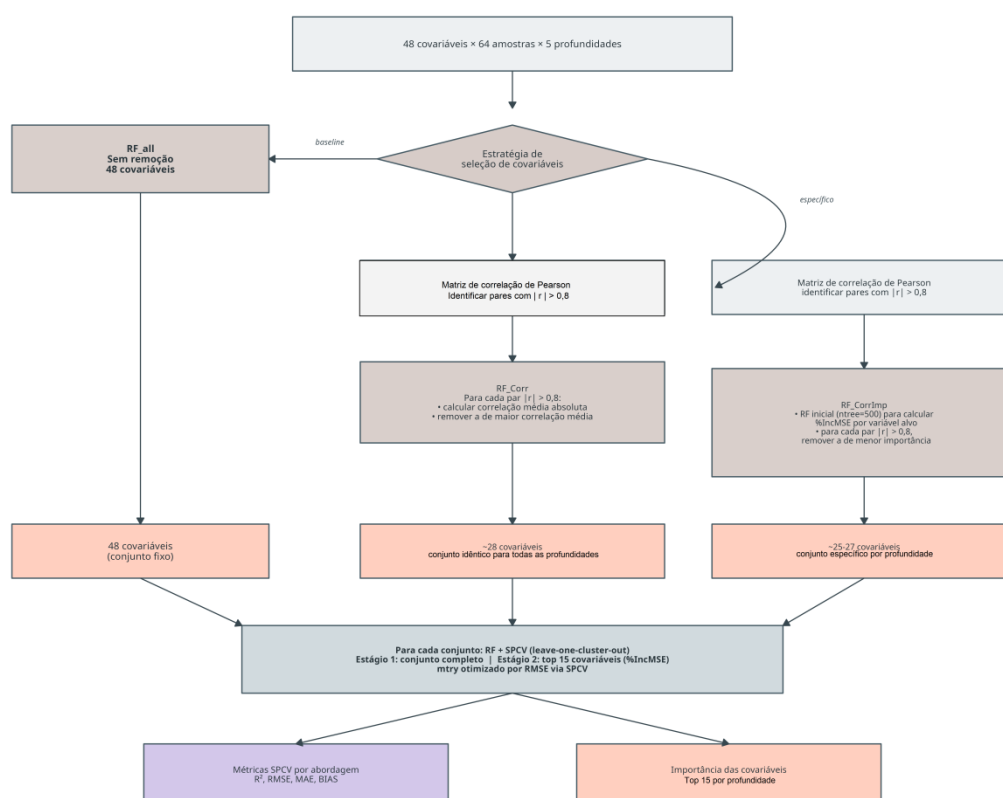
A otimização do hiperparâmetro *mtry* do algoritmo *Random Forest* foi realizada utilizando o mesmo esquema de validação cruzada espacial. O parâmetro *mtry* corresponde ao número de variáveis preditoras selecionadas aleatoriamente e avaliadas em cada divisão dos nós das árvores de decisão. Foram testados diferentes valores de *mtry*, limitados ao número total de covariáveis disponíveis no conjunto de dados (os valores candidatos testados foram 1, 2, 3, 5, 8, 12, 16, 20, 25 e 30). Para cada valor avaliado, o modelo foi submetido à validação cruzada espacial completa, sendo calculadas as métricas de desempenho correspondentes. A seleção do valor ótimo de *mtry* foi baseada na minimização do RMSE médio obtido durante a SPCV, uma vez que essa métrica expressa diretamente a magnitude dos erros de predição na mesma unidade da variável resposta.

3.2.6 Abordagens de seleção de covariáveis e redução dimensional do número de covariáveis

Foram testadas três abordagens de seleção de covariáveis ambientais para cada profundidade de solo estudada, sendo elas: completa (sem remoção

de covariáveis), baseada em correlação simples entre covariáveis e correlação de importância específica. Para cada uma das abordagens de seleção de covariáveis foi realizada, em seguida, a redução dimensional do conjunto de covariáveis, mantendo apenas as 15 covariáveis mais importantes para a predição. Portanto, a modelagem se deu em dois estágios, nos quais, para cada abordagem em cada profundidade, foi produzido um modelo com todas as covariáveis disponíveis de acordo com o mecanismo de seleção de covariáveis e um modelo reduzido com as 15 covariáveis mais importantes, conforme ilustrado pela Figura 5.

Figura 5 - Representação esquemática das estratégias de seleção de covariáveis utilizadas na modelagem *Random Forest*



Fonte: Elaborado pela autora (2026).

3.2.6.1 Abordagens de seleção de covariáveis

- **Abordagem 1 – Sem remoção (referência – RF_all)**

Neste caso, optou-se por utilizar o conjunto completo de covariáveis disponíveis para a predição da variável resposta. Essa abordagem teve como

objetivo maximizar a capacidade do modelo em capturar relações complexas entre a variável de interesse e os diferentes fatores ambientais representados pelas covariáveis. A utilização de todas as covariáveis é adequada para modelos baseados em *Random Forest*, uma vez que esse tipo de algoritmo apresenta elevada robustez à multicolinearidade e à presença de variáveis pouco informativas (Breiman, 2001). Além disso, a inclusão de todas as covariáveis permite avaliar o potencial preditivo máximo do conjunto de dados disponível, sem impor restrições prévias quanto à relevância individual das covariáveis. Dessa forma, o modelo é capaz de explorar possíveis relações não lineares e interações entre os preditores, que poderiam ser perdidas em procedimentos de seleção prévia de covariáveis.

- **Abordagem 2** – Seleção de covariáveis baseada em análise de correlação (RF_Corr)

Neste caso, com o objetivo de reduzir a redundância das informações e minimizar a presença de multicolinearidade entre os preditores, foi realizada uma análise de correlação entre as covariáveis utilizando o coeficiente de correlação de Pearson. A matriz de correlação foi calculada a partir das observações completas do conjunto de dados, e os pares de covariáveis que apresentaram coeficiente de correlação absoluto superior a 0,80 ($|r| > 0,80$) foram identificados como altamente correlacionados (Guyon; Elisseeff, 2003; Dormann *et al.*, 2013). Para cada par identificado, foi removida a covariável que apresentou maior correlação média absoluta com as demais covariáveis do conjunto, assumindo-se que essa covariável continha maior redundância de informação. As covariáveis remanescentes foram utilizadas nas etapas subsequentes de calibração, validação cruzada espacial e predição por meio do algoritmo *Random Forest*.

- **Abordagem 3** – Seleção de covariáveis baseada em correlação e importância preditiva (RF_CorrImp)

Neste caso, com o objetivo de reduzir a redundância de informação e minimizar a multicolinearidade entre os preditores, foi realizada uma análise de correlação utilizando o coeficiente de correlação de Pearson e, paralelamente, foi ajustado um modelo RF preliminar (ntree=500) utilizando todas as

covariáveis disponíveis, a partir do qual foi estimada a importância relativa de cada covariável por meio da métrica de incremento percentual do erro quadrático médio (%IncMSE). Essa métrica quantifica a contribuição de cada covariável para a capacidade preditiva do modelo em relação à variável resposta.

Foram identificados os pares de covariáveis com correlação absoluta superior a 0,80 ($|r| > 0,80$) (Guyon; Elisseeff, 2003; Dormann *et al.*, 2013) e, para cada par altamente correlacionado, foi mantida a covariável que apresentou maior importância preditiva no modelo RF preliminar, enquanto a covariável menos importante foi removida do conjunto de preditores.

Essa abordagem permitiu reduzir a redundância entre covariáveis preservando aquelas com maior relevância para a predição da variável resposta, resultando em um conjunto de preditores mais parcimonioso e potencialmente mais informativo para a modelagem final.

3.2.6.2 Redução dimensional do número de covariáveis

Para as três abordagens de seleção de covariáveis, após o ajuste do modelo RF utilizando o conjunto completo de covariáveis disponíveis e previamente selecionadas, foi realizada uma análise da importância dos preditores com base na métrica de incremento percentual do erro quadrático médio (%IncMSE). As covariáveis foram ordenadas de acordo com seus valores de importância, e as quinze covariáveis com maior contribuição preditiva foram selecionadas para compor um conjunto reduzido de preditores (RF_Top15). Embora em MDS seja comum utilizar conjuntos grandes de covariáveis ambientais, isso pode gerar maior custo computacional e deixar os modelos mais complexos levando a dificuldade de interpretação (Kasraei *et al.*, 2024).

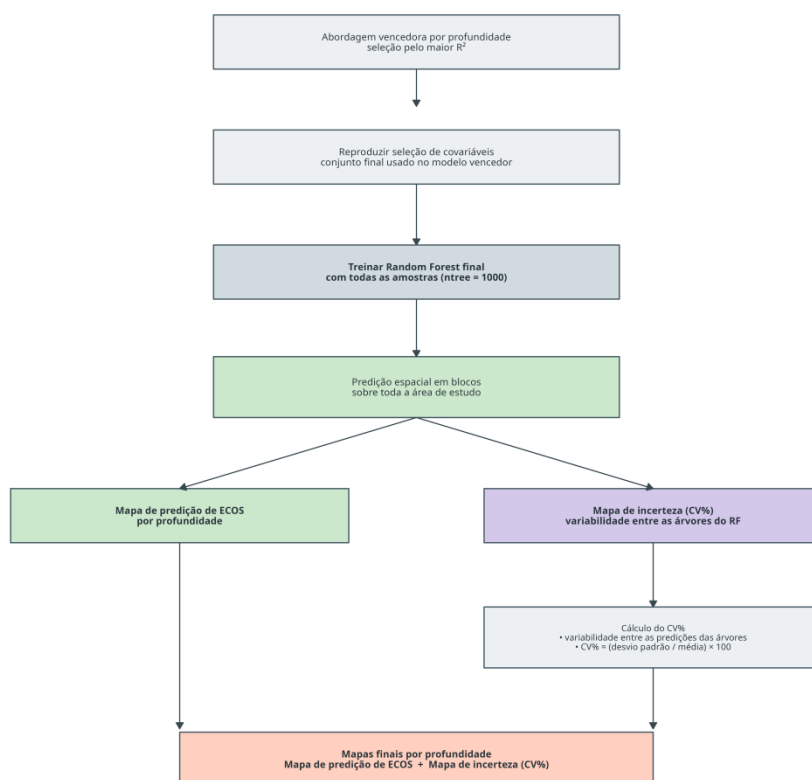
Uma vez estruturado o subconjunto RF_Top15, realizou-se uma nova rodada independente de otimização do hiperparâmetro *mtry* via busca em grade (*grid search*), restringindo-se os valores candidatos aos limites do novo modelo reduzido (*mtry*= 1, 2, 3, 5, 8 e 12). O valor final de *mtry* para o modelo simplificado foi igualmente definido pelo menor RMSE médio obtido na validação cruzada espacial. Essa abordagem teve como objetivo avaliar se um modelo mais simplificado, contendo apenas as covariáveis mais relevantes,

poderia manter ou melhorar o desempenho preditivo em relação ao modelo completo.

3.2.7 Predição espacial e incerteza

Após a etapa de calibração e validação dos modelos, e da comparação entre as diferentes estratégias de seleção de covariáveis ambientais, foi selecionado, para cada profundidade do solo, o modelo que apresentou o melhor desempenho preditivo durante a validação cruzada espacial, considerando principalmente o coeficiente de determinação (R^2) e, complementarmente, os valores de RMSE e MAE. O modelo selecionado foi então recalibrado utilizando a totalidade das observações disponíveis e aplicado às camadas *raster* das covariáveis ambientais para a geração dos mapas preditivos do estoque de carbono orgânico do solo (ECOS) (Figura 6). Esse procedimento foi realizado individualmente para cada profundidade analisada, resultando na obtenção dos mapas finais de distribuição espacial do ECOS na área de estudo.

Figura 6 - Esquema metodológico para geração dos mapas de predição e incerteza de ECOS



Fonte: Elaborado pela autora (2026).

Para cada profundidade, foram gerados ainda mapas de incerteza expressos por meio do coeficiente de variação (CV) entre as predições individuais das árvores do RF (Equação 7).

$$CV = \frac{\sigma}{\mu} \quad (7)$$

Em que:

σ : é o desvio padrão calculado para cada pixel; e

μ : é a média calculada para cada pixel.

Para o cálculo do CV, foram utilizadas 200 árvores (ntree=200) em vez das 1.000 empregadas na predição. Essa redução no número de árvores se justifica, pois as métricas de desempenho convergem para um valor limite à medida que o número de árvores aumenta. Sendo assim, é possível utilizar valores de ntree viáveis computacionalmente quando a convergência é

alcançada, já que o maior ganho de desempenho seria alcançado ao treinar as primeiras 100 árvores (Probst; Boulesteix, 2018). Valores elevados de CV indicam maior discordância entre as árvores, refletindo regiões com maior incerteza preditiva.

Um fluxograma detalhado da metodologia empregada para a modelagem preditiva pode ser consultado no Apêndice I.

3.2.8 Software e ferramentas utilizadas

As etapas de processamento de dados, modelagem e validação foram realizadas em ambiente R, versão 4.3.3 (R Core Team, 2024). Adicionalmente, foi utilizado o software QGIS em sua versão 3.22 (QGIS Development Team, 2022), para o processamento e manipulação de dados geoespaciais, incluindo visualização, organização de camadas raster e vetoriais, bem como suporte às etapas de preparação das variáveis ambientais.

3.3 Resultados e discussão

3.3.1 Distribuição vertical do ECOS observado

Com base nos dados de estoque de carbono orgânico do solo (ECOS) das 64 amostras coletadas a campo, observou-se uma distribuição vertical distinta para o atributo ao longo do perfil do solo (Tabela 5).

Tabela 5 - Caracterização estatística do estoque de carbono orgânico do solo (ECOS) por camada de profundidade na área de estudo

Profundidade	N	Média (Mg ha ⁻¹)	Mediana (Mg ha ⁻¹)	Mínimo	Máximo	DP	CV (%)
0-5 cm	64	8,48	8,04	3,4	22,65	3,4	40,1
5-15 cm	64	11,17	10,63	4,19	27,33	4,35	38,9
15-30 cm	64	12,54	10,94	4,67	27,56	5,48	43,7
30-60 cm	64	20,78	20,8	2,67	41,43	8,81	42,4
60-100 cm	64	17,71	16,59	1,38	50,08	9,23	52,1

Fonte: Elaborado pela autora (2026).

O ECOS apresentou aumento progressivo entre as camadas superficiais e subsuperficiais, passando de 8,48 Mg ha⁻¹ na camada de 0-5 cm para o valor máximo de 20,78 Mg ha⁻¹ na camada de 30-60 cm. Posteriormente, observou-se redução para 17,71 Mg ha⁻¹ na camada de 60-100 cm. Os estoques de carbono observados nas camadas superficiais estão de acordo com os valores

relatados por Reis *et al.* (2026) em sistemas produtivos baseados em integração lavoura-pecuária na região.

O maior ECOS observado na camada de 30–60 cm representa um comportamento que merece atenção, uma vez que a tendência geral descrita para a maioria dos solos é de redução do conteúdo de carbono orgânico com o aumento da profundidade (Akpa *et al.*, 2016). Entretanto, esse comportamento não é uniforme entre diferentes classes de solos e ambientes, especialmente em áreas sujeitas à saturação hídrica. Nesse mesmo estudo, os autores demonstraram que os Gleysols, conforme a classificação da WRB/FAO, apresentaram os maiores estoques de carbono em profundidades superiores a 30 cm, evidenciando que condições hidromórficas podem favorecer o acúmulo e preservação do carbono em camadas subsuperficiais. O mesmo padrão foi observado por Vasques *et al.* (2010), em que o conteúdo de carbono teve tendência a diminuir com a profundidade e os maiores valores de COS foram detectados em solos de áreas úmidas.

Além disso, em solos hidromórficos, a matéria orgânica pode estar predominantemente associada à fração pesada, considerada mais estável e menos suscetível à decomposição. Em estudo realizado em Planossolo de várzea, a maior parte do carbono orgânico presente ao longo do perfil foi atribuída a essa fração, enquanto a proteção física por oclusão em agregados mostrou-se pouco expressiva em razão das condições de má drenagem (Rosa *et al.*, 2011). O predomínio de Planossolos na área (50% dos pontos) sugere que o horizonte B mais adensado desses solos, com limite superior médio em 48,8 cm, atua como barreira física que favorece o acúmulo de carbono orgânico nessa camada. Adicionalmente, a presença de Argissolos com horizonte B textural (limite superior médio de 67,0 cm) e o Gleissolo com horizonte Btg contribuem para a retenção de matéria orgânica em subsuperfície, seja por proteção física em horizontes mais argilosos ou por condições de anaerobiose associadas à saturação hídrica.

Observa-se um coeficiente de variação elevado, que varia de 38,9% na camada de 5-15 cm a 52,1% na camada de 60-100 cm. Isso indica alta variabilidade espacial do ECOS em todas as profundidades, ou seja, há alta heterogeneidade espacial entre os pontos amostrados, mesmo dentro de uma mesma profundidade. Esse comportamento é esperado em ambientes onde

fatores como relevo, regime hídrico, cobertura vegetal e dinâmica sedimentar variam em curtas distâncias, pois esses fatores atuam de forma direta nos processos de produção, deposição, decomposição e estabilização da matéria orgânica, corroborando os achados de John *et al.* (2020) para uma planície aluvial, em que o COS até 30 cm de profundidade apresentou CV de 47%. O aumento no CV em camadas mais profundas pode refletir a influência de processos pedogenéticos e deposicionais no estoque de carbono, evidenciando diferenças históricas de sedimentação, variação no nível do lençol freático e padrões de acúmulo de matéria orgânica, resultando na maior dispersão dos valores observados.

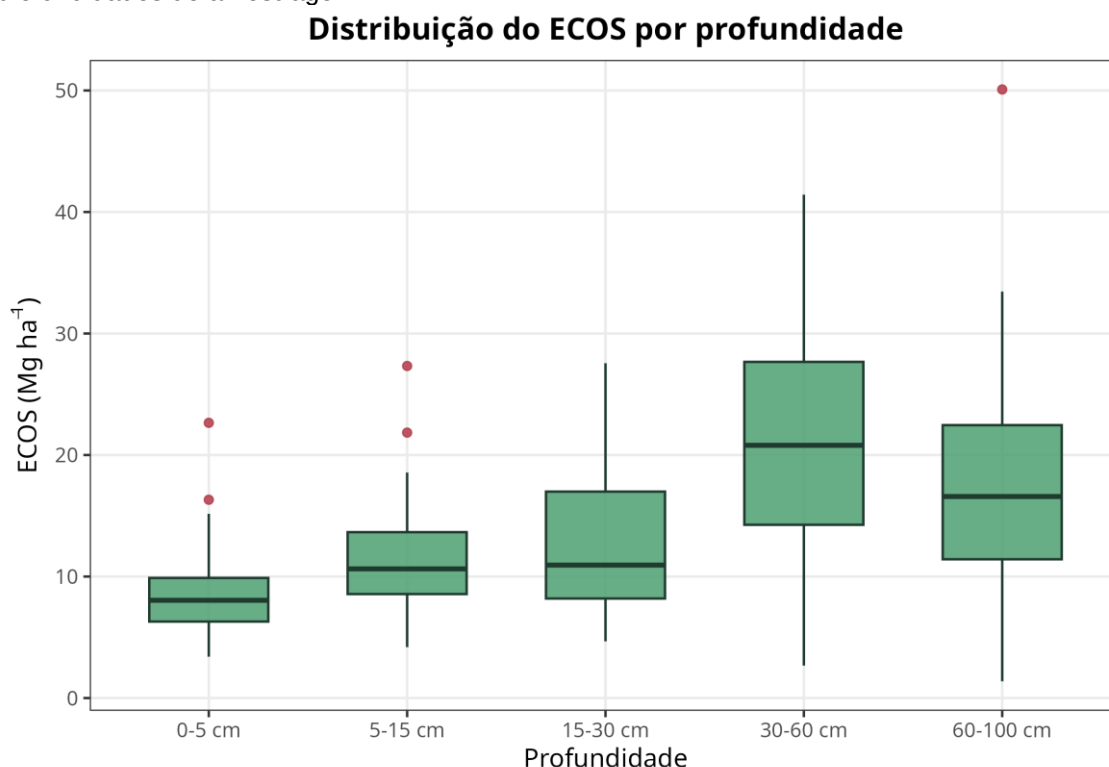
A diversidade pedológica da área de estudo, com predomínio de Planossolos (n=32) e Argissolos (n=22), além de Neossolos (n=7), Espodossolo, Chernossolo e Gleissolo, fornece base para interpretar esse aumento da variabilidade, pois nas camadas superficiais (0–15 cm), todos os pontos encontram-se no horizonte A, relativamente homogêneo. A partir de 15–30 cm, inicia-se a transição para horizontes E ou B em Planossolos (limite superior médio de 48,8 cm) e Argissolos (67,0 cm), enquanto Neossolos mantêm a sequência A/C sem horizonte B diagnóstico. Na camada de 30–60 cm, a diversidade entre classes atinge seu ponto máximo, pois Planossolos e Argissolos já estão no horizonte B (densidade do solo de 1,65 e 1,66 g cm⁻³, respectivamente), Neossolos permanecem com densidade inferior (1,47 g cm⁻³), e o Gleissolo apresenta horizonte Btg com densidade de 1,68 g cm⁻³ (Tabela 1 e Tabela 2).

A amplitude entre os valores mínimos e máximos é pronunciada em todas as profundidades, especialmente na camada de 60-100 cm, sugerindo alta variabilidade espacial no estoque de carbono. Além disso, a camada de 30-60 cm apresenta média 2,45 vezes superior à camada 0-5 cm, indicando acúmulo de carbono em profundidade. A alta variabilidade observada (CV > 30% em todas as profundidades) representa um desafio para a modelagem preditiva, exigindo covariáveis ambientais mais robustas para capturar a heterogeneidade espacial do estoque de carbono (Lamichhane *et al.*, 2019).

O boxplot permite a visualização da distribuição do ECOS observado para as cinco profundidades analisadas (Figura 7). A camada de 0-5 cm apresenta a menor mediana e amplitude interquartil reduzida, indicando

concentração de valores em uma faixa estreita. A presença de outliers superiores sugere alguns locais com ECOS excepcionalmente alto. A camada de 5-15 cm apresenta mediana intermediária, com aumento da amplitude interquartil em relação à camada de 0-5 cm. Já na camada de 15-30 cm, observa-se mediana mais elevada, com dispersão similar à camada anterior. O boxplot confirma a tendência de aumento do ECOS até esta profundidade. Na camada de 30-60 cm, há a maior mediana e a maior amplitude interquartil, confirmando o pico de ECOS observado nas estatísticas descritivas. Por fim, na camada de 60-100 cm, a mediana reduz em relação à camada anterior, mas a amplitude interquartil continua ampla, confirmando a alta variabilidade indicada pelo CV de 52,1%.

Figura 7 - Distribuição do estoque de carbono orgânico do solo (ECOS) em diferentes profundidades de amostragem



Fonte: Elaborado pela autora (2026).

A camada de 0-5 cm apresenta distribuição distinta das demais, com menor mediana e menor dispersão sugerindo processos de controle diferenciados (maior influência de fatores de superfície). Visualmente são identificáveis outliers em múltiplas profundidades, corroborando a presença de valores extremos identificados nas estatísticas descritivas. Diante da presença

de valores extremos em todas as profundidades, é importante ressaltar que valores muito altos ou muito baixos podem afetar o treinamento do *Random Forest*, embora o algoritmo seja relativamente robusto a outliers.

A análise exploratória dos dados revelou que as áreas 1 e 5 concentraram a maior frequência de valores elevados de ECOS, especialmente nas camadas mais profundas do perfil. Na área 1, a maioria dos pontos amostrais está localizada em relevo suave ondulado e sob cobertura de campo nativo, em ambientes com menor intensidade de uso agrícola. De forma semelhante, os pontos da área 5 encontram-se predominantemente associados a áreas de campo nativo (Apêndice D e Apêndice H). Esse padrão sugere que a maior acumulação de carbono observada nessas áreas pode estar relacionada às condições de uso e cobertura da terra, uma vez que ecossistemas menos perturbados tendem a favorecer o aporte contínuo de resíduos vegetais e a conservação da matéria orgânica no solo. Essa interpretação é corroborada por Parvizi e Fatehi (2025), que observaram maiores teores de carbono orgânico em áreas de pastagem e os menores valores em áreas agrícolas, evidenciando a influência do manejo e das mudanças no uso da terra sobre a dinâmica do carbono.

3.3.2 Análise exploratória das covariáveis ambientais

A análise da matriz de correlação das covariáveis ambientais utilizadas para a modelagem evidenciou padrões de redundância entre as 48 covariáveis consideradas (Apêndice J). Dentro do conjunto completo de covariáveis, foram identificados pares de covariáveis altamente correlacionados nos três blocos de covariáveis disponíveis: bandas espectrais Landsat, índices espectrais e covariáveis topográficas.

Entre as bandas espectrais das imagens Landsat, foram identificadas correlações extremamente altas entre bandas do visível e infravermelho, ilustrados por pares de covariáveis como b2blue-b3green ($r=0,98$), b3green-b4red ($r=0,97$), b2blue-b4red ($r=0,95$) e b6swir1-b7swir2 ($r=0,97$). Nos índices espectrais também foi observado pares de covariáveis altamente correlacionados como BSI-NDVI ($r=-0,87$), NDVI-SAVI ($0,99$), NDVI-NDWI ($r=-0,96$), BSI-NDMI ($r=-0,99$) e CMR-NDTI ($r=0,99$). Enquanto entre as covariáveis topográficas/morfométricas, foram identificados grupos de covariáveis com

elevada correlação como, por exemplo, entre covariáveis relacionadas a curvaturas (maxc, proc, planc, tanc, csc, gc) com correlações entre 0,60 e 0,99; bem como para covariáveis relacionadas à posição no relevo (rsa, sh, s) com correlações entre 0,59 e 0,98 e também entre índices de rugosidade (vrn e tri) com correlação de 0,76.

A elevada correlação e, conseqüentemente, redundância entre as covariáveis tem origem, principalmente, da derivação de múltiplos índices espectrais a partir das mesmas bandas Landsat e da dependência matemática entre os atributos topográficos derivados do MDE. Embora a forte correlação entre atributos morfométricos reflita sua dependência matemática, a definição das covariáveis utilizadas na modelagem preditiva deve priorizar sua relevância física e evitar a inclusão simultânea de informações redundantes (Florinsky, 2017). Heung *et al.* (2016), por exemplo, na predição de classes de solos no Canadá, de um conjunto inicial de 26 covariáveis topográficas, após utilizar o recurso da análise de componentes principais (PCA) para reduzir o número de covariáveis e eliminar multicolinearidade, concluíram que 13 covariáveis do grupo inicial eram responsáveis por explicar 95% da variância acumulada. Frente à complexa interação entre as covariáveis disponíveis para a modelagem, justifica-se a necessidade de seleção de covariáveis buscando simplificar os modelos, torná-los mais explicativos e computacionalmente amigáveis, bem como reduzir o estabelecimento de relações redundantes entre as covariáveis durante o processo de treinamento do modelo (Guyon; Elisseeff, 2003).

As covariáveis removidas nas estratégias de seleção aplicadas no estudo evidenciam que a abordagem RF_Corr (correlação simples) removeu consistentemente 20 covariáveis em todas as profundidades (Tabela 6), incluindo bandas espectrais (b3green, b4red, b6swir1, b7swir2), índices espectrais (ndvi, savi, bsi), além de covariáveis derivadas de mde (nh, ssp, gc, csc).

Tabela 6 - Covariáveis ambientais removidas durante o processo de seleção de preditores com base no critério de correlação simples (RF_Corr)

Covariável	Grupo	Profundidade de remoção	Critério de remoção
ndvi	Espectral	Todas	Correlação simples
cmr	Espectral	Todas	Correlação simples
bsi	Espectral	Todas	Correlação simples
gc	Morfométrica	Todas	Correlação simples
lc	Morfométrica	Todas	Correlação simples
csc	Morfométrica	Todas	Correlação simples
b3green	Espectral	Todas	Correlação simples
tri	Morfométrica	Todas	Correlação simples
b7swir2	Espectral	Todas	Correlação simples
savi	Espectral	Todas	Correlação simples
sbi	Espectral	Todas	Correlação simples
b4red	Espectral	Todas	Correlação simples
s	Morfométrica	Todas	Correlação simples
mbi	Morfométrica	Todas	Correlação simples
b6swir1	Espectral	Todas	Correlação simples
tanc	Morfométrica	Todas	Correlação simples
minc	Morfométrica	Todas	Correlação simples
npcri	Espectral	Todas	Correlação simples
nh	Morfométrica	Todas	Correlação simples
ssp	Morfométrica	Todas	Correlação simples

Fonte: Elaborado pela autora (2026).

Na estratégia de seleção da abordagem RF_CorrImp (correlação-importância), houve a remoção de 20 a 23 covariáveis, dependendo da profundidade (Tabela 7). Por exemplo, na profundidade de 5-15 cm, foi removida a maior quantidade de covariáveis desta abordagem, sendo 23 covariáveis removidas, incluindo maxc, proc e ci, que são covariáveis que aparecem como importantes em outras profundidades. A estratégia RF_CorrImp, ao considerar a importância preditiva, preservou covariáveis de maior relevância para o ECOS, embora ainda tenha removido bandas espectrais importantes para as camadas superficiais.

Tabela 7 - Covariáveis ambientais removidas durante o processo de seleção de preditores com base no critério de correlação e importância (RF_CorrImp)

Covariável	Frequência de remoção	Profundidades
ndvi	100%	Todas
ndmi	100%	Todas
ndwi	100%	Todas
mbi	100%	Todas
bsi	100%	Todas
gc	80%	exceto 5-15 cm
lc	80%	exceto 60-100 cm
csc	80%	exceto 60-100 cm
s	80%	exceto 30-60 cm
b6swir1	80%	exceto 30-60 cm
b4red	80%	exceto 15-30 cm
planc	80%	exceto 30-60 cm
tanc	80%	exceto 15-30 cm
ssp	80%	exceto 5-15 cm
tri	80%	exceto 0-5 cm
b7swir2	80%	exceto 0-5 cm
proc	80%	exceto 0-5 cm
cmr	60%	0-5 cm; 15-30 cm; 60-100 cm
b3green	60%	0-5 cm; 15-30 cm; 30-60 cm
npcri	60%	0-5 cm; 5-15 cm; 30-60 cm
minc	60%	0-5 cm; 5-15 cm; 60-100 cm
stanh	60%	0-5 cm; 5-15 cm; 30-60 cm
b2blue	60%	5-15 cm; 15-30 cm; 60-100 cm
maxc	60%	5-15 cm; 15-30 cm; 60-100 cm
sbi	60%	15-30 cm; 30-60 cm; 60-100 cm
b5nir	40%	0-5 cm; 5-15 cm
rsa	40%	0-5 cm; 30-60 cm
ndti	40%	5-15 cm; 30-60 cm
nh	40%	15-30 cm; 60-100 cm
ci	20%	5-15 cm
savi	20%	15-30 cm

Fonte: Elaborado pela autora (2026).

A alta correlação entre as covariáveis espectrais justifica a remoção de, em média, 10 covariáveis deste grupo pelas estratégias RF_Corr e RF_CorrImp. Contudo, a seleção baseada apenas em correlação (sem considerar a importância preditiva) pode eliminar covariáveis que, embora redundantes, carregam informações complementares sobre o ECOS.

Dentre as três abordagens utilizadas para seleção de covariáveis, a abordagem RF_all mantém equilíbrio mais representativo, enquanto as abordagens de seleção favorecem fortemente covariáveis topográficas; o que

corroborar a conclusão de Mosleh *et al.* (2016), de que mesmo em áreas de baixo relevo, os atributos de terreno podem ser de grande importância para prever os atributos do solo.

3.3.3 Desempenho preditivo dos modelos *Random Forest*

3.3.3.1 Ajuste dos hiperparâmetros

Neste estudo, a otimização do hiperparâmetro *mtry* foi conduzida por busca em grade (*grid search*) sobre o conjunto de valores candidatos (1, 2, 3, 5, 8, 12, 16, 20, 25, 30), restrito aos valores menores ou iguais ao número de covariáveis disponíveis em cada abordagem. Para cada valor candidato, o desempenho foi avaliado pelo RMSE médio obtido na validação cruzada espacial (SPCV), sendo selecionado o valor de *mtry* que minimizou esse critério. O hiperparâmetro *ntree* foi fixado em todas as abordagens e profundidades em 1.000. Na Tabela 8 podem ser visualizados os valores de *mtry* otimizados para cada abordagem de seleção de covariáveis e profundidade avaliada, incluindo o número de covariáveis disponíveis em cada configuração.

Tabela 8 - Valores otimizados do hiperparâmetro *mtry* e número de covariáveis utilizadas nas diferentes abordagens de modelagem *Random Forest* para predição do estoque de carbono orgânico do solo

Abordagem	Covariáveis disponíveis	Profundidade (cm)				
		0-5	5-15	15-30	30-60	60-100
RF_all	48	1	1	1	1	3
RF_all Top15	15	1	1	1	1	2
RF_Corr	28	1	1	1	1	3
RF_Corr Top15	15	1	1	1	1	1
RF_CorrImp	25-28*	1	1	1	1	1
RF_CorrImp Top15	15	1	12	1	1	1

* Número de covariáveis restantes após seleção por correlação e importância, variável por profundidade (27 para 0–5 cm; 25 para 5–15 cm; 26 para 15–30 cm; 28 para 30–60 cm; 27 para 60–100 cm).

Fonte: Elaborado pela autora (2026).

A otimização do hiperparâmetro *mtry* mostra convergência para *mtry* = 1 em 26 das 30 combinações avaliadas nas camadas de 0–5, 5–15, 15–30, 30–60 e 60–100 cm. Isso indica que, para a maioria dos casos, nenhum valor maior da grade testada (2, 3, 5, 8, 12, 16, 20, 25, 30) reduziu o RMSE-SPCV em

relação à $mtry = 1$, sugerindo que o sinal preditivo para o ECOS encontra-se distribuído entre múltiplas covariáveis e que o aumento do número de candidatos por nó não se traduz em ganho de desempenho preditivo. Esse padrão é consistente com a elevada multicolinearidade interna do conjunto de covariáveis utilizado, composto predominantemente por índices espectrais derivados de imagens de satélite Landsat e atributos topográficos derivados de modelo digital de elevação (declividade, curvatura de plano, curvatura mínima e máxima, índice de posição topográfica, entre outros). Nesse sentido, utilizar valores baixos de $mtry$ pode levar à melhor utilização das informações disponíveis, dando a chance de preditores com efeitos moderados serem selecionados em vez de serem mascarados por preditores mais fortes (Boulesteix *et al.*, 2012).

Covariáveis altamente correlacionadas entre si podem carregar informação redundante (Dormann *et al.*, 2013), embora a correlação entre preditores geralmente não exerça grande influência no desempenho preditivo (Boulesteix *et al.*, 2012), de modo que mesmo avaliando apenas uma covariável candidata por nó, como ocorre com $mtry = 1$, o algoritmo encontra sinal preditivo suficiente para realizar divisões informativas. Sendo assim, elevar $mtry$ não amplia substancialmente a informação disponível por nó, mas aumenta a correlação entre as árvores individuais (Probst *et al.*, 2019). Frente à variabilidade observada do ECOS nas camadas de 0–5 a 30–60 cm (CV entre 38,9% e 43,7%), o comportamento dos valores de $mtry$ otimizados corrobora com a ausência de um preditor dominante isolado, sustentando a otimização para $mtry = 1$.

A camada de 60–100 cm apresenta comportamento distinto nas abordagens RF_all e RF_Corr, com $mtry$ otimizado em 3 para o conjunto completo de covariáveis e em 2 para o subconjunto Top15 da RF_all. Esse padrão não se repete nas abordagens com seleção combinada de correlação e importância (RF_CorrImp e RF_CorrImp Top15), que mantiveram $mtry = 1$.

O aumento do $mtry$ nas abordagens com maior número de covariáveis disponíveis (48 para RF_all e 28 para RF_Corr) na camada de 60–100 cm pode ser interpretado como resposta do algoritmo à redução da relação sinal-ruído nessa profundidade. Com CV de 52,1%, o maior entre as cinco camadas avaliadas, e amplitude máxima de 1,38 a 50,08 Mg ha⁻¹, o ECOS na camada

mais profunda apresenta alta variabilidade e, ao mesmo tempo, menor relação com as covariáveis de superfície. Nesse caso, para que o sorteio aleatório por nó capture regularmente as poucas covariáveis informativas, o algoritmo eleva *mtry*, aumentando a probabilidade de inclusão dessas covariáveis em cada divisão (Genuer *et al.*, 2010). Contudo, quando a seleção por correlação e importância (RF_CorrImp) remove as covariáveis de menor relevância para essa profundidade (21 covariáveis removidas para 60–100 cm), o conjunto residual é suficientemente informativo para que *mtry* = 1 seja mantido. Esse resultado sugere que a filtragem prévia de covariáveis por importância atuou de forma a eliminar o ruído que forçava a elevação do *mtry* nas abordagens menos seletivas, corroborando a literatura que sugere que quando se têm muitos preditores informativos, deve-se optar por um valor de *mtry* baixo (Boulesteix *et al.*, 2012; Probst *et al.*, 2019).

Valores ótimos para *mtry* mais baixos em mapeamento preditivo de atributos do solo são regularmente relatados como em Bouslihim *et al.* (2024) em que o *mtry* ótimo para predição de MOS foi 3, Yang *et al.* (2020) com *mtry*=2 para mapeamento de classes de solo e Wang *et al.* (2018) com *mtry*=3 para mapeamento de ECOS. Embora valores baixos de *mtry* possam resultar em árvores individuais menos precisas, eles podem melhorar o desempenho global do *Random Forest* ao aumentar a diversidade entre as árvores e reduzir sua correlação (Probst *et al.*, 2019).

O valor *mtry* = 12 obtido para a abordagem RF_CorrImp Top15 na camada de 5–15 cm foi um achado que destoou dos demais resultados. Porém, nesse caso, com 15 covariáveis disponíveis após remoção de 23 preditores pela filtragem combinada de correlação e importância, a grade de candidatos efetivamente avaliada foi restrita a 1, 2, 3, 5, 8, 12, sendo *mtry* = 12 o maior valor testado dentro dessa grade. O valor foi selecionado porque produziu o menor RMSE-SPCV entre os seis candidatos avaliados, configurando um resultado legítimo dentro dos critérios da metodologia adotada. Um valor de *mtry* = 12 é equivalente a avaliar 80% das covariáveis disponíveis em cada nó de divisão, podendo ser entendido de forma que após dupla filtragem (correlação-importância e Top15), as 15 covariáveis restantes apresentam alta relevância individual e baixa redundância entre si, sendo assim, com preditores pouco correlacionados, qualquer subconjunto sorteado

em um nó tende a conter variáveis informativas. Nesse caso, a vantagem de $mtry = 12$ sobre valores menores pode ser modesta, embora suficiente para ser detectada pelo critério de minimização do RMSE com 64 observações. Contudo, esses resultados devem ser interpretados com cautela, uma vez que a redução do conjunto de preditores a um grupo estritamente independente, ou seja, com baixa correlação, nem sempre é desejável. Além disso, os resultados dependem do tamanho amostral, de modo que, em determinadas situações, pode ser mais adequado selecionar poucos preditores representativos de grupos altamente correlacionados (Boulesteix *et al.*, 2012).

Embora atípico no contexto deste estudo, onde $mtry = 1$ predominou em 26 das 30 combinações avaliadas, valores elevados de $mtry$ não são incomuns na literatura. Gelsleichter *et al.* (2025), utilizando busca em grade com conjunto de dados de COS e número de amostras semelhante ($n = 70$), também obtiveram $mtry = 12$, assim como Poggio *et al.* (2021) para o mapeamento de COS em escala global. Heung *et al.* (2014), com 27 covariáveis e $ntree$ fixado em 750, obtiveram $mtry = 11$, enquanto Heung *et al.* (2016), com 20 covariáveis ambientais para predição de classes de solo, reportaram $mtry$ variando entre 9 e 12. Taghizadeh-Mehrjardi *et al.* (2015) otimizaram $mtry$ em 9 com $ntree = 800$, e Hateffard *et al.* (2023), com $ntree$ fixado em 50, obtiveram $mtry$ entre 6 e 8 para modelagem de ECOS e outros atributos do solo.

3.3.3.2 Comparação entre estratégias de seleção de covariáveis

Foram avaliadas três estratégias de seleção de covariáveis: RF_all que consiste na utilização de todas as 48 covariáveis disponíveis sem remoção prévia de covariáveis; RF_Corr que consiste na remoção de covariáveis por correlação de Pearson $|r| > 0,80$, neste caso, foram removidas 20 covariáveis, restando 28 covariáveis para modelagem; e, RF_CorrImp que consiste na remoção de covariáveis combinadas por correlação e por importância preditiva, nesse caso, restaram de 25 a 28 covariáveis para modelagem, variando conforme a profundidade avaliada. A magnitude da quantidade de covariáveis removidas está de acordo com a literatura, em que é relatado que a seleção de covariáveis baseada em correlação, normalmente, elimina em torno de metade das covariáveis (Taghizadeh-Mehrjardi *et al.*, 2015).

As métricas de desempenho são médias das métricas calculadas por fold da SPCV e estão representadas na Tabela 9.

Tabela 9 - Métricas de desempenho dos modelos *Random Forest* obtidas por validação cruzada espacial para diferentes abordagens de seleção de covariáveis e profundidades do solo

Profundidade	Abordagem	R ²	RMSE (Mg ha ⁻¹)	MAE (Mg ha ⁻¹)	BIAS (Mg ha ⁻¹)
0–5 cm	RF_all	0,043	3,43	2,68	0,044
0–5 cm	RF_Corr	0,030	3,45	2,69	0,000
0–5 cm	RF_CorrImp	0,035	3,46	2,70	0,070
5–15 cm	RF_all	0,074	4,49	3,56	0,089
5–15 cm	RF_Corr	0,089	4,52	3,58	-0,002
5–15 cm	RF_CorrImp	0,103	4,50	3,59	0,061
15–30 cm	RF_all	0,103	5,85	5,05	-0,155
15–30 cm	RF_Corr	0,066	5,87	5,06	-0,163
15–30 cm	RF_CorrImp	0,074	5,82	4,99	-0,303
30–60 cm	RF_all	0,064	9,35	7,98	-0,197
30–60 cm	RF_Corr	0,064	9,45	8,01	-0,414
30–60 cm	RF_CorrImp	0,044	9,40	7,93	-0,363
60–100 cm	RF_all	0,047	9,48	7,93	-0,355
60–100 cm	RF_Corr	0,028	9,63	7,99	-0,621
60–100 cm	RF_CorrImp	0,069	9,41	7,83	-0,565

Fonte: Elaborado pela autora (2026).

Na camada de 0-5 cm, as três abordagens apresentaram desempenho preditivo muito baixo e praticamente equivalente com R² variando de 0,030 (RF_Corr) a 0,043 (RF_all) e RMSE entre 3,43 e 3,46 Mg ha⁻¹. A remoção de covariáveis por correlação simples (RF_Corr) ou por correlação e importância (RF_CorrImp) não resultou em ganho de desempenho nessa camada. As diferenças absolutas entre abordagens são pequenas (Δ RMSE < 0,03 Mg ha⁻¹) e não são suficientes para sustentar superioridade de qualquer estratégia.

Na camada de 5-15 cm, a abordagem RF_CorrImp apresentou o maior R² (0,103) entre as três estratégias, embora as diferenças sejam de pequena magnitude (Δ R² máximo de 0,029 em relação à RF_all). A remoção combinada de covariáveis redundantes e de baixa importância (RF_CorrImp) foi, entre as estratégias testadas, a que melhor performou para essa camada.

Na camada de 15-30 cm, a abordagem RF_all obteve o maior R² (0,103) e o RF_CorrImp o menor RMSE (5,82 Mg ha⁻¹) e MAE (4,99 Mg ha⁻¹). As diferenças entre abordagens são pequenas com Δ R² de 0,037 entre RF_all e

RF_Corr; Δ RMSE de 0,05 Mg ha⁻¹ entre RF_Corr e RF_CorrImp. Um comportamento que chama atenção é que RF_Corr, que remove apenas covariáveis altamente correlacionadas, apresentou o pior desempenho nessa camada, sugerindo que algumas das covariáveis eliminadas por correlação simples poderiam ser preditivamente relevantes para o ECOS entre 15 e 30 cm de profundidade.

Na camada de 30-60 cm, RF_all e RF_Corr apresentaram os mesmos valores de R², enquanto RF_CorrImp registrou R² inferior (0,044). O menor RMSE foi obtido por RF_all (9,35 Mg ha⁻¹). A remoção mais agressiva de covariáveis (RF_CorrImp) resultou em perda relativa de desempenho nessa camada, indicando que algumas covariáveis eliminadas pelo critério combinado poderiam contribuir para a predição do ECOS entre 30 e 60 cm. O BIAS foi negativo nas três abordagens, com maior magnitude em RF_Corr (-0,414 Mg ha⁻¹), indicando tendência de subestimação sistemática.

Na camada de 60-100 cm, RF_CorrImp apresentou o maior R² (0,069) e o menor RMSE (9,41 Mg ha⁻¹). RF_Corr apresentou o pior desempenho com R² = 0,028 e RMSE = 9,63 Mg ha⁻¹. A remoção de 20 covariáveis por correlação simples sem critério de importância (RF_Corr) mostrou-se particularmente inadequada para a camada mais profunda, possivelmente eliminando covariáveis relevantes para o ECOS nessa profundidade. A filtragem adicional por importância (RF_CorrImp) recuperou parte do desempenho ao reter preditores com maior relevância específica para essa camada.

Em síntese, os modelos com conjunto completo de covariáveis apresentaram desempenho preditivo consistentemente baixo em todas as profundidades e abordagens, com R² máximo de 0,103 (RF_all, 15–30 cm; RF_CorrImp, 5–15 cm). Esse padrão indica que nenhuma das três estratégias de seleção produziu modelos capazes de prever de forma satisfatória o ECOS para a área de estudo, o que pode ser atribuído tanto à elevada variabilidade espacial do ECOS (CV entre 38,9 e 52,1%) quanto ao tamanho amostral limitado (n = 64). Vale ressaltar aqui que existem recomendações de que a densidade amostral adequada para aplicação da técnica do MDS gira em torno de 10 pontos por km² (Somarathna *et al.*, 2017; Yang *et al.*, 2020), o que, de certa forma, poderia dificultar estudos em grandes escalas. Adicionalmente, a diversidade pedológica da área, com seis classes de solo distintas quanto à

espessura do horizonte A (15 a 120 cm), profundidade do limite superior do B (20 a 155 cm) e densidade do solo (1,25 a 1,68 g cm⁻³), representa um fator de controle não capturado pelas 48 covariáveis ambientais utilizadas, todas derivadas de atributos da superfície da paisagem. As covariáveis topográficas e espectrais representam indiretamente parte dessa variabilidade, mas não discriminam, por exemplo, a presença de um horizonte E em Planossolos ou a profundidade do horizonte B textural em Argissolos, que controlam diretamente o acúmulo de carbono em subsuperfície.

O desempenho dos modelos RF apresentado na Tabela 9, é comparável com os resultados obtidos por Mosleh *et al.* (2016) para a predição de COS, em uma planície Iraniana, utilizando redes neurais artificiais ($R^2=0,03$) e de Zeraatpisheh *et al.* (2022), utilizando o método da Regressão Parcial dos Mínimos Quadrados (PLSR), que atingiram valores de R^2 que variam de 0,05 a 0,11 na camada de 0-20 cm para diferentes conjuntos de covariáveis ambientais. Adhikari *et al.* (2014) relataram, ao mapear COS e ECOS nas mesmas profundidades padrões do *GlobalSoilMap*, melhor desempenho preditivo na camada de 5-15 cm (R^2 de teste=0,63 e R^2 de validação=0,42) e o pior desempenho preditivo na camada de 60-100 cm (R^2 de teste=0,28 e R^2 de validação=0,23), com o modelo *Cubist*. Akpa *et al.* (2016), utilizando um número semelhante de covariáveis ambientais (23 variáveis) e avaliando as mesmas cinco camadas de profundidade, observaram redução da capacidade preditiva dos modelos com o aumento da profundidade. Embora os maiores valores de R^2 tenham sido obtidos pelo *Random Forest*, variando de 0,34 na camada superficial a 0,18 em subsuperfície (60–100 cm), os demais algoritmos testados apresentaram desempenhos bastante restritos nas camadas mais profundas, com valores de R^2 de 0,13 a 0,07 para o método *Cubist* e de 0,11 a 0,07 para *Boosted Regression Trees*. Esses resultados demonstram que, mesmo com uma quantidade semelhante de covariáveis e diferentes algoritmos de aprendizado de máquina, a explicação da variabilidade espacial do ECOS permanece desafiadora, especialmente em subsuperfície.

De maneira semelhante, Bócoli *et al.* (2025), em condições brasileiras, trabalhando com elevada densidade amostral (uma amostra a cada 3,5 ha) e resolução espacial de 12,5 m, também reportaram desempenho limitado dos modelos *Random Forest* para a predição do carbono do solo. Embora o melhor

resultado tenha alcançado R^2 de 0,47 para o COS na camada de 0–20 cm, esse desempenho representou uma condição específica entre os diversos conjuntos de covariáveis avaliadas. De forma geral, os valores médios de R^2 foram de apenas 0,18 para o COS superficial e de 0,01 para a camada de 80–100 cm. Para o estoque de carbono do solo (ECOS), o desempenho foi ainda mais restrito, com valores médios de R^2 de aproximadamente 0,03 na camada de 0–20 cm e 0,01 na camada de 80–100 cm, além da ocorrência frequente de coeficientes de determinação negativos. Esses resultados evidenciam a baixa capacidade dos modelos em explicar a variabilidade espacial do carbono, especialmente em profundidade, mesmo sob condições de alta densidade amostral e elevada resolução espacial. Nesse sentido, os autores destacam que uma elevada densidade amostral não garante, necessariamente, alto desempenho preditivo.

Métodos de redução do número de covariáveis são comumente utilizados em estudos de MDS com o objetivo de selecionar conjuntos mais informativos de preditores e reduzir redundâncias entre variáveis. Contudo, nem sempre o emprego de técnicas de filtragem e remoção resulta em melhorias na capacidade preditiva dos modelos. Song *et al.* (2017), por exemplo, relataram que a redução do conjunto inicial de covariáveis por meio de técnicas baseadas em regressão linear e análise de componentes principais não produziu efeitos significativos sobre os valores de R^2 . De forma semelhante, Xiong *et al.* (2014) utilizaram o algoritmo Boruta para identificar as covariáveis mais relevantes para explicar a variabilidade do COS na Flórida, Estados Unidos, reduzindo o conjunto inicial de 210 para 43 covariáveis sem alterações expressivas no desempenho final dos modelos. Por outro lado, em estudos conduzidos em escalas menores, a seleção de covariáveis pode representar ganhos substanciais em eficiência computacional. Ao mapear propriedades do solo em escala global, Poggio *et al.* (2021) reduziram significativamente a complexidade dos modelos preditivos para COS, diminuindo o conjunto inicial de aproximadamente 400 covariáveis para apenas 40. Para isso, os autores aplicaram inicialmente uma análise de desconexão, que reduziu o número de covariáveis para 150, seguida da técnica de eliminação recursiva de características, resultando no conjunto final de preditores.

Em contrapartida, a simples adição de novas fontes de informação também não garante melhorias substanciais na predição. Dai *et al.* (2026) avaliaram o efeito da incorporação de covariáveis ambientais a modelos de matéria orgânica do solo baseados em espectroscopia VIS-NIR e verificaram que a inclusão dessas covariáveis produziu ganhos limitados para o algoritmo *Random Forest*. Embora tenha sido observada melhora na camada de 20–40 cm, com aumento do R^2 de 0,44 para 0,57, as demais profundidades apresentaram incrementos modestos ou até mesmo redução do desempenho preditivo. Em conjunto, esses estudos indicam que tanto a remoção quanto a inclusão de covariáveis podem ter impacto limitado sobre a qualidade das predições quando as covariáveis disponíveis não acrescentam informações efetivamente relacionadas aos processos que controlam a variabilidade espacial do atributo de interesse. Nesses casos, a capacidade preditiva dos modelos parece depender mais da relevância e complementaridade das informações utilizadas do que da quantidade de covariáveis empregadas.

Dessa forma, os baixos valores de R^2 encontrados neste estudo podem refletir não apenas limitações do modelo empregado, mas também a complexidade dos fatores que controlam a distribuição do carbono no solo e a dificuldade das covariáveis ambientais em capturar esses processos em diferentes profundidades. De acordo com os dados analisados, não foi possível selecionar uma estratégia unânime para a área de estudo, já que não houve evidência de superioridade consistente de uma única abordagem em todas as profundidades, pois RF_all foi mais eficiente nas camadas 0–5, 15–30 e 30–60 cm; RF_CorrImp foi superior nas camadas 5–15 e 60–100 cm; RF_Corr não apresentou vantagem em nenhuma profundidade.

3.3.3.3 Efeito da redução dimensional de covariáveis

Após avaliadas as três estratégias de seleção de covariáveis (RF_all, RF_Corr e RF_CorrImp), foi realizada uma redução dimensional no conjunto de covariáveis disponíveis, de forma que apenas as 15 covariáveis mais importantes para cada abordagem e cada profundidade fossem utilizadas para a nova modelagem (Top15). O critério de seleção foi o %IncMSE (*Percent Increase in Mean Squared Error*), métrica calculada internamente pelo *Random Forest* a partir das amostras *out-of-bag* (OOB), onde para cada covariável,

seus valores são permutados aleatoriamente nas amostras OOB de cada árvore e mede-se o percentual de aumento do MSE em relação ao MSE OOB original, sendo o valor final a média desse aumento sobre todas as árvores. Sendo assim, quanto maior o aumento percentual do MSE, maior é a importância da covariável para a predição.

Após a definição das 15 covariáveis mais importantes para cada abordagem e profundidade, novos modelos *Random Forest* foram ajustados utilizando apenas esse conjunto reduzido de preditores (Top15). Os resultados obtidos para as diferentes estratégias de seleção de covariáveis após a redução dimensional são apresentados na Tabela 10.

Tabela 10 - Métricas de desempenho dos modelos *Random Forest* com redução dimensional do conjunto de covariáveis (Top15) para diferentes profundidades do solo

Profundidade	Abordagem	R ²	RMSE (Mg ha ⁻¹)	MAE (Mg ha ⁻¹)	BIAS (Mg ha ⁻¹)
0–5 cm	RF_all Top15	0,066	3,45	2,71	0,116
0–5 cm	RF_Corr Top15	0,010	3,33	2,60	0,024
0–5 cm	RF_CorrImp Top15	0,022	3,38	2,62	0,033
5–15 cm	RF_all Top15	0,102	4,22	3,24	0,062
5–15 cm	RF_Corr Top15	0,085	4,21	3,25	-0,042
5–15 cm	RF_CorrImp Top15	0,073	4,28	3,27	0,048
15–30 cm	RF_all Top15	0,041	5,57	4,75	-0,106
15–30 cm	RF_Corr Top15	0,012	5,67	4,91	-0,237
15–30 cm	RF_CorrImp Top15	0,008	5,73	4,85	-0,277
30–60 cm	RF_all Top15	0,028	9,02	7,69	0,056
30–60 cm	RF_Corr Top15	0,027	9,24	7,81	-0,235
30–60 cm	RF_CorrImp Top15	0,019	9,22	7,79	-0,512
60–100 cm	RF_all Top15	0,093	8,73	7,27	-0,522
60–100 cm	RF_Corr Top15	0,062	8,96	7,52	-0,493
60–100 cm	RF_CorrImp Top15	0,062	8,84	7,38	-0,550

Fonte: Elaborado pela autora (2026).

Na camada de 0-5 cm, a abordagem RF_all Top15 apresentou maior R² (0,066), mas RF_Corr Top15 obteve menor RMSE (3,33 Mg ha⁻¹) e menor MAE (2,60 Mg ha⁻¹). Esse resultado indica que as 15 covariáveis selecionadas pelo modelo treinado sem remoção prévia (RF_all) têm maior correlação com o ECOS, porém as 15 selecionadas pela RF_Corr (após remoção da redundância) produzem predições com menor erro absoluto.

Na camada 5-15 cm, RF_all Top15 e RF_Corr Top15 apresentaram desempenho muito próximo com R² de 0,102 e 0,085; RMSE de 4,22 e 4,21 Mg

ha⁻¹, respectivamente. RF_CorrImp Top15 registrou menor R² (0,073) e maior RMSE (4,28 Mg ha⁻¹), apesar de operar com mtry = 12 (o maior valor da grade testada para essa configuração) o que não se traduziu em vantagem preditiva mensurável.

Na camada 15-30 cm, RF_all Top15 apresentou desempenho consistentemente superior nas três métricas principais: maior R² (0,041), menor RMSE (5,57 Mg ha⁻¹) e menor MAE (4,75 Mg ha⁻¹). RF_Corr Top15 e RF_CorrImp Top15 tiveram desempenho inferior, com R² abaixo de 0,02. A redução de covariáveis por correlação e importância antes da seleção do Top15 mostrou-se inadequada para essa camada, pois ao filtrar preditores previamente, a RF_CorrImp restringiu a quantidade de covariáveis candidatas ao Top15, podendo ter excluído covariáveis que seriam relevantes nessa profundidade específica.

Na camada 30-60 cm, RF_all Top15 foi a melhor abordagem com R² = 0,028, RMSE = 9,02 Mg ha⁻¹ e BIAS positivo mínimo (+0,056 Mg ha⁻¹). As abordagens RF_Corr Top15 e RF_CorrImp Top15 apresentaram RMSE entre 9,22 e 9,24 Mg ha⁻¹ e BIAS negativo de maior magnitude (-0,235 e -0,512 Mg ha⁻¹, respectivamente), indicando subestimação sistemática do ECOS nessa camada quando a seleção prévia de covariáveis é aplicada.

Na camada de 60-100 cm, RF_all Top15 apresentou o maior R² (0,093), sendo o segundo maior entre todos os modelos Top15 em todas as profundidades, e o menor RMSE (8,73 Mg ha⁻¹) para esta profundidade. Esse resultado revela que ao selecionar as 15 covariáveis mais importantes diretamente do conjunto completo (48 covariáveis), o modelo captura de forma mais eficiente os preditores para o ECOS em subsuperfície sem o viés introduzido pela filtragem prévia. Enquanto RF_Corr Top15 e RF_CorrImp Top15 apresentaram o mesmo desempenho para o R² (0,062).

De modo geral, a redução do conjunto total de covariáveis para o subconjunto Top15 resultou em melhorias nos valores de RMSE em comparação ao conjunto completo de preditores, sugerindo que a remoção de covariáveis de baixa relevância contribuiu para reduzir o ruído preditivo dos modelos. Entre as estratégias testadas, a abordagem RF_all Top15 apresentou o comportamento mais consistente ao longo das profundidades, indicando que a seleção direta das 15 covariáveis mais importantes com base no %IncMSE,

sem filtrações prévias por correlação ou importância, foi eficiente para concentrar o poder preditivo do modelo e minimizar potenciais perdas de informação decorrentes da exclusão antecipada de preditores.

Entretanto, embora a redução dimensional tenha contribuído para pequenas melhorias no RMSE, os ganhos observados em termos de coeficiente de determinação (R^2) foram limitados. Esse resultado está de acordo com os achados de Xiong *et al.* (2014), que verificaram que a simplificação do conjunto preditor reduziu a complexidade computacional do modelo, mas teve pouco impacto sobre seu desempenho preditivo. Cabe destacar que o estudo foi conduzido em escala regional, com densidade amostral de aproximadamente 0,0072 pontos por km² (um ponto para cada 138,9 km²), valor próximo ao observado no presente trabalho, além de ter sido desenvolvido em ambiente caracterizado por solos predominantemente arenosos e relevo suave. Resultado semelhante foi reportado por Heung *et al.* (2014), que também não observaram melhorias expressivas no desempenho do *Random Forest* após a redução do número de covariáveis.

Por outro lado, Parvizi e Fatehi (2025) obtiveram resultados positivos ao selecionar covariáveis ambientais para modelar o COS nas cinco profundidades padrão do *GlobalSoilMap*. Os autores reduziram um conjunto inicial de 27 covariáveis para conjuntos com 10, 12, 9, 4 e 4 covariáveis nas camadas de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, respectivamente, alcançando R^2 de validação de 0,64 com o algoritmo *Random Forest*. Esses resultados indicam que os benefícios da redução dimensional dependem não apenas da quantidade de covariáveis removidas, mas principalmente da capacidade das covariáveis remanescentes em representar os fatores que controlam a variabilidade espacial do carbono no solo.

Assim, a eficácia da seleção de covariáveis tende a ser dependente do contexto ambiental e da qualidade da relação entre os preditores disponíveis e o atributo a ser modelado. Quando as covariáveis removidas contêm informações redundantes, a simplificação do conjunto preditor pode melhorar a eficiência do modelo sem comprometer sua capacidade preditiva; entretanto, em ambientes onde as covariáveis disponíveis já apresentam baixa capacidade explicativa, a redução do número de preditores tende a produzir apenas ganhos marginais.

A comparação entre "Todas as covariáveis" e "Top 15 covariáveis" (Tabela 11) mostra que em alguns casos, o Top15 apresenta desempenho ligeiramente superior (ex: 0-5 cm RF_all e 60-100 cm RF_all) e em outros casos, o conjunto completo apresenta desempenho superior (ex: 5-15 cm RF_CorrImp, 15-30 cm RF_all, 30-60 cm RF_all), sendo as diferenças pequenas e inconsistentes entre profundidades e abordagens, sugerindo que a redução dimensional para 15 covariáveis não degrada significativamente o desempenho, mas também não oferece ganhos consistentes.

Tabela 11 - Efeito da seleção das 15 covariáveis mais importantes sobre o desempenho dos modelos *Random Forest* em diferentes profundidades do solo

Profundidade	Abordagem	R ² (todas)	R ² (Top15)	Δ R ²
0-5 cm	RF_all	0,043	0,066	0,023
	RF_Corr	0,03	0,01	-0,02
	RF_CorrImp	0,035	0,022	-0,013
5-15 cm	RF_all	0,074	0,102	0,028
	RF_Corr	0,089	0,085	-0,004
	RF_CorrImp	0,103	0,073	-0,03
15-30 cm	RF_all	0,103	0,041	-0,062
	RF_Corr	0,066	0,012	-0,054
	RF_CorrImp	0,074	0,008	-0,066
30-60 cm	RF_all	0,064	0,028	-0,036
	RF_Corr	0,064	0,027	-0,037
	RF_CorrImp	0,044	0,019	-0,025
60-100 cm	RF_all	0,047	0,093	0,046
	RF_Corr	0,028	0,062	0,034
	RF_CorrImp	0,069	0,062	-0,007

Fonte: Elaborado pela autora (2026).

Esse comportamento difere de achados da literatura como, por exemplo, Brungard *et al.* (2015) que concluíram que *Random Forest* com covariáveis selecionadas por eliminação recursiva de características (culminando em conjuntos com 6 a 10 covariáveis ambientais) foram consistentemente mais precisas para predição de classes de solo. Por outro lado, Zeraatpisheh *et al.*

(2022) afirmam que nem sempre adicionar mais covariáveis à modelagem resulta, de fato, em melhoria na precisão dos modelos, mas podem auxiliar a produzir mapas com menor incerteza.

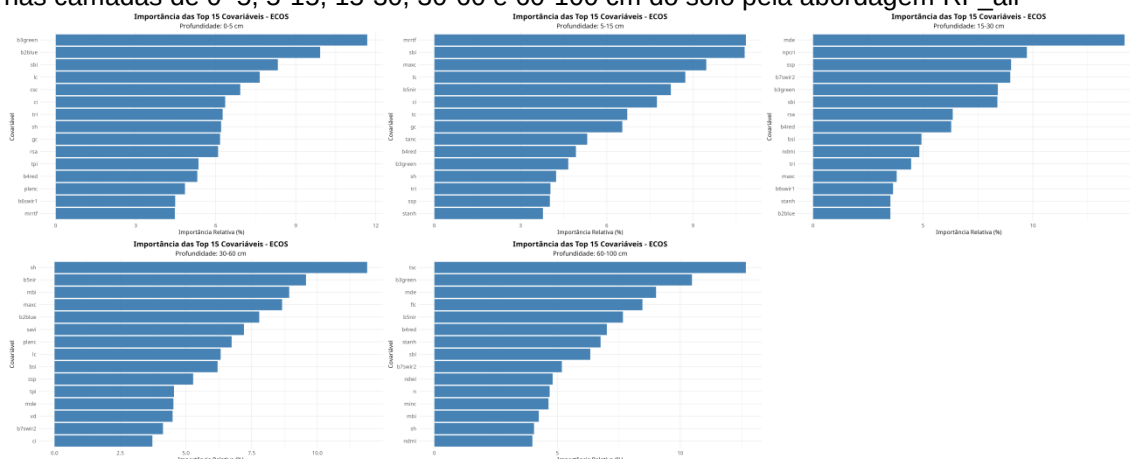
3.3.4 Importância das covariáveis ambientais

A abordagem RF_Corr removeu 20 covariáveis em todas as profundidades, resultando em um conjunto final de 28 covariáveis para a modelagem preditiva. Em contrapartida, a abordagem RF_CorrImp promoveu uma redução variável entre as profundidades analisadas. Nas camadas de 0-5 cm e 60-100 cm foram removidas 21 covariáveis, restando 27 para modelagem. Na camada de 5-15 cm, foram removidas 23 covariáveis, resultando em 25 covariáveis remanescentes. Já na camada de 15-30 cm foram excluídas 22 covariáveis, permanecendo 26, enquanto na camada de 30-60 cm foram removidas 20 covariáveis, restando 28.

Quanto ao tipo de covariável removida, a abordagem RF_Corr eliminou igualmente 10 covariáveis espectrais e 10 covariáveis topográficas/morfométricas. Esse resultado indica que a estratégia baseada exclusivamente na correlação promoveu uma remoção equilibrada entre os grupos de covariáveis correlacionadas. Entretanto, considerando o conjunto inicial composto por 48 covariáveis (33 topográficas/morfométricas e 15 espectrais), observa-se um desequilíbrio na composição do subconjunto final disponibilizado para a modelagem. Comportamento semelhante foi observado para a abordagem RF_CorrImp, na qual foram removidas entre 20 e 23 covariáveis e, destas, em média, 10 eram covariáveis espectrais.

A seleção das 15 covariáveis mais importantes para a predição do ECOS sob a abordagem RF_all é apresentada na Figura 8.

Figura 8 - Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0-5, 5-15, 15-30, 30-60 e 60-100 cm do solo pela abordagem RF all



Fonte: Elaborado pela autora (2026).

Nas camadas superficiais (0–5 e 5–15 cm), observa-se transição na composição do subconjunto Top15. Na camada de 0–5 cm, cinco covariáveis espectrais (b3green, b2blue, sbi, b4red e b6swir1) figuram entre as 15 mais importantes, representando um terço do subconjunto. As três covariáveis de maior importância são derivadas de imagens Landsat, evidenciando a relevância das informações espectrais para a predição do ECOS em superfície. Após essas três primeiras posições, observa-se maior frequência de covariáveis topográficas relacionadas à curvatura, declividade e escoamento superficial. Já na camada de 5–15 cm, há predominância de covariáveis topográficas/morfométricas, com apenas quatro covariáveis derivadas de Landsat, destacando-se o índice sbi, associado à reflectância da superfície, que ocupa a segunda posição em importância relativa.

Na camada de 15–30 cm, destaca-se a elevada relevância da elevação (mde), que representa a covariável mais importante do subconjunto. Além disso, essa camada apresenta nove covariáveis derivadas de imagens Landsat entre as 15 mais importantes (npcri, b7swir2, b3green, sbi, b4red, bsi, ndmi, b6swir1 e b2blue), indicando forte contribuição das informações espectrais para a predição do ECOS nessa profundidade.

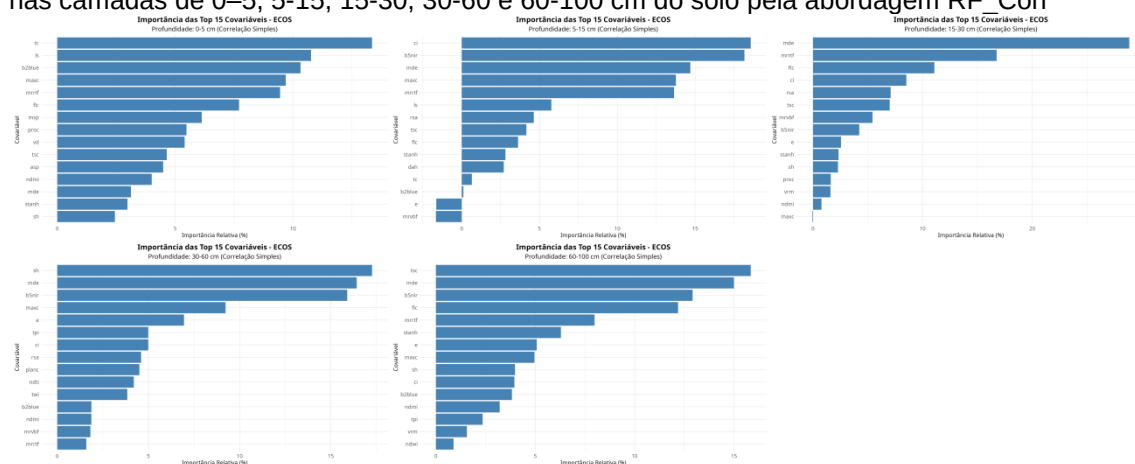
Nas camadas mais profundas (30–60 e 60–100 cm), o equilíbrio entre covariáveis topográficas e espectrais varia. Na camada de 30–60 cm, apenas cinco covariáveis espectrais (b5nir, b2blue, savi, bsi e b7swir2) aparecem entre as 15 mais importantes, observando-se predominância de covariáveis topográficas, com destaque para a altura do declive (sh), que apresenta a

maior importância relativa do subconjunto. Na camada de 60–100 cm, verifica-se maior equilíbrio entre a origem das covariáveis selecionadas, sendo sete derivadas de imagens Landsat e oito derivadas de MDE. Entre as covariáveis topográficas, destacam-se a convexidade do terreno (tsc), a elevação (mde) e o comprimento de fluxo (flc), todas posicionadas entre as cinco mais importantes.

De modo geral, os resultados evidenciam que a contribuição relativa das covariáveis espectrais e topográficas varia ao longo do perfil do solo. As covariáveis espectrais têm maior proeminência na camada superficial (0–5 cm) e na intermediária (15–30 cm), enquanto as topográficas predominam em subsuperfície (5–15, 30–60 e 60–100 cm), com a elevação (mde) ganhando importância progressivamente à medida que a profundidade aumenta.

A seleção das 15 covariáveis mais importantes para a predição do ECOS sob a abordagem RF_Corr é apresentada na Figura 9.

Figura 9 - Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0–5, 5–15, 15–30, 30–60 e 60–100 cm do solo pela abordagem RF_Corr



Fonte: Elaborado pela autora (2026).

Nas camadas superficiais (0–5 e 5–15 cm), observa-se forte predomínio de covariáveis topográficas/morfométricas, com apenas duas covariáveis derivadas de Landsat em cada subconjunto Top15. Na camada de 0–5 cm, as covariáveis espectrais b2blue e ndmi figuram entre as 15 mais importantes, enquanto entre as topográficas destacam-se as relacionadas à curvatura da superfície (tc e maxc) e o fator topográfico (ls), associado aos processos erosivos. Na camada de 5–15 cm, apenas duas covariáveis derivadas de Landsat permanecem, sendo todas as demais derivadas de MDE. Destaca-se

o fato de que as duas últimas covariáveis do ranking apresentam importância relativa negativa, indicando que sua inclusão tende a introduzir ruído ao modelo. Conforme discutido por Boulesteix *et al.* (2012), as medidas de importância de variáveis do *Random Forest* fornecem apenas um ranqueamento relativo dos preditores, de modo que sempre produzem uma classificação, mesmo quando os preditores não contêm informação útil para a predição. Consequentemente, a seleção das 15 covariáveis mais importantes não garante que todas elas possuam sinal preditivo real, sendo esperado que variáveis ruidosas também possam aparecer entre as mais bem classificadas (comportamento observado também na Figura 10). Nessa profundidade, sobressaem as covariáveis relacionadas ao escoamento superficial (ci), à banda de infravermelho próximo (b5nir) e à elevação (mde).

Na camada de 15–30 cm, novamente apenas duas covariáveis derivadas de Landsat aparecem entre as 15 mais importantes, ambas com baixa importância relativa (b5nir e ndmi). O subconjunto é dominado pelas covariáveis topográficas/morfométricas, tendo a elevação (mde) como covariável mais relevante, seguida por outras covariáveis relacionadas à forma da superfície, curvatura e escoamento superficial.

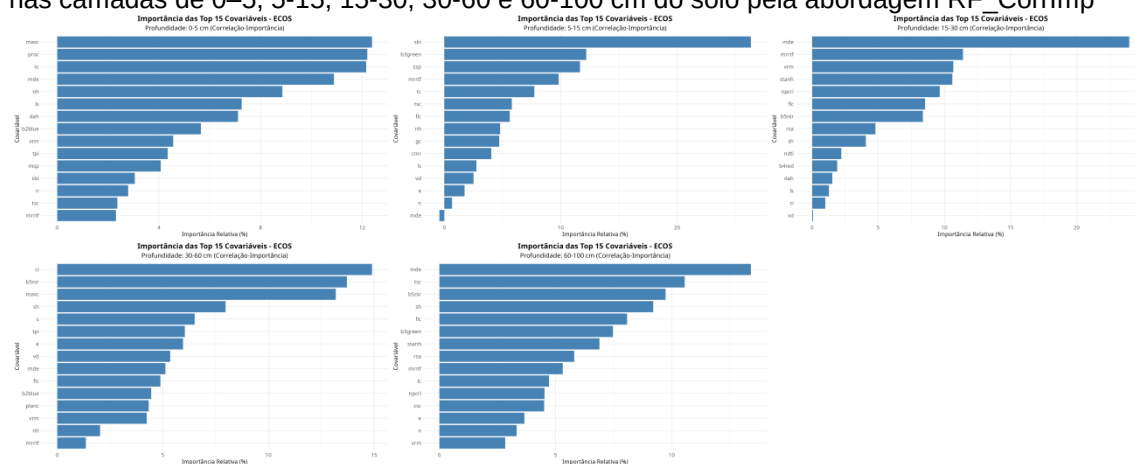
Nas camadas mais profundas (30–60 e 60–100 cm), a participação de covariáveis espectrais aumenta em relação às camadas superficiais, mas permanece limitada. Na camada de 30–60 cm, quatro covariáveis derivadas de Landsat figuraram entre as 15 mais importantes, embora apenas a b5nir tenha apresentado elevada importância relativa. Entre as demais covariáveis mais relevantes destacaram-se a altura do declive (sh) e a elevação (mde). Nessa camada, o TWI (índice de umidade topográfica) figurou entre as 15 covariáveis mais importantes pela única vez em todos os casos avaliados. Ainda assim, a covariável não foi removida por nenhuma das abordagens de seleção empregadas, contrastando com resultados obtidos em áreas de relevo plano e ambientes úmidos nos Estados Unidos, em que o TWI foi identificado como um importante preditor do COS (Li *et al.*, 2020). Esse contraste sugere que os preditores e os processos que controlam a distribuição do COS e do ECOS na área de estudo diferem substancialmente daqueles observados em outras regiões, mesmo sob condições ambientais semelhantes. Na camada de 60–100 cm, quatro covariáveis derivadas de Landsat compõem o subconjunto

Top15, sendo novamente a banda b5nir a única com elevada importância relativa. Nessa profundidade, atributos relacionados à conformação do terreno e à elevação apresentam os maiores valores de importância.

De modo geral, a abordagem RF_Corr favorece fortemente as covariáveis topográficas/morfométricas em detrimento das espectrais, refletindo a remoção prévia de covariáveis correlacionadas, que impactou mais intensamente o grupo espectral. Ainda assim, a banda b5nir mantém-se como a covariável espectral mais recorrente nas camadas subsuperficiais, e a elevação (mde) consolida-se como preditor de destaque a partir de 15 cm de profundidade.

A seleção das 15 covariáveis mais importantes para predição do ECOS sob a abordagem RF_CorrImp é apresentada na Figura 10.

Figura 10 - Importância relativa das 15 principais covariáveis utilizadas na predição do ECOS nas camadas de 0–5, 5–15, 15–30, 30–60 e 60–100 cm do solo pela abordagem RF_CorrImp



Fonte: Elaborado pela autora (2026).

Nas camadas superficiais (0–5 e 5–15 cm), observa-se predomínio de covariáveis topográficas/morfométricas, com participação limitada de covariáveis espectrais. Na camada de 0–5 cm, apenas duas covariáveis derivadas de Landsat figuram entre as 15 mais importantes, evidenciando o predomínio de covariáveis topográficas mesmo em superfície. Entre estas, destacam-se aquelas relacionadas à curvatura da superfície (maxc, proc, tc) e à elevação (mde). Na camada de 5–15 cm, a elevação (mde) apresenta importância relativa negativa, indicando que sua presença no subconjunto selecionado contribui para o aumento do ruído na predição e, consequentemente, sua remoção poderia beneficiar o desempenho do modelo.

Nessa camada, três covariáveis derivadas de Landsat permanecem entre as mais importantes, sendo sbi e b3green as de maior relevância.

Na camada de 15–30 cm, a elevação (mde) apresenta a maior importância relativa entre todas as covariáveis selecionadas. Quatro covariáveis derivadas de Landsat aparecem no subconjunto Top15, destacando-se npcri (índice de clorofila) como a covariável espectral mais relevante.

Nas camadas mais profundas (30–60 e 60–100 cm), a banda do infravermelho próximo (b5nir) emerge como a covariável espectral de maior destaque. Na camada de 30–60 cm, apenas duas covariáveis derivadas de Landsat figuram entre as 15 mais importantes, com b5nir apresentando a segunda maior importância relativa, ficando atrás apenas do índice de convergência de escoamento (ci). Na camada de 60–100 cm, a mais profunda, a elevação (mde) apresenta a maior importância relativa. O subconjunto Top15 inclui ainda três covariáveis derivadas de Landsat (b5nir, b3green e npcri).

De modo geral, a abordagem RF_CorrImp apresenta padrão similar ao da RF_Corr, com predomínio de covariáveis topográficas em todas as profundidades. A elevação (mde) consolida-se como o preditor mais relevante a partir de 15 cm, enquanto a banda b5nir e o índice npcri destacam-se como as covariáveis espectrais mais consistentes em subsuperfície. O número de covariáveis espectrais no subconjunto Top15 varia de duas a quatro conforme a profundidade, sendo mais expressivo a partir de 15–30 cm.

Destaca-se a semelhança entre os resultados obtidos neste estudo e aqueles reportados por Guo *et al.* (2019) na predição do COS. Os autores avaliaram a importância de covariáveis derivadas do MDE em diferentes resoluções espaciais e verificaram que o índice de convergência (ci), a elevação (mde), a profundidade do vale (vd), a altura normalizada (nh) e a altura padronizada (stanh) estiveram entre os preditores mais consistentes e importantes. De forma semelhante, no presente estudo, essas covariáveis também apresentaram elevada frequência de seleção entre os preditores mais importantes, representando aproximadamente 17% das ocorrências no conjunto Top15. Entre elas, a elevação (mde) apresentou a maior recorrência, com 13 aparições, seguida pelo índice de convergência (ci) e pela altura padronizada (stanh), ambos com 9 aparições. A recorrência dessas covariáveis

em diferentes estudos evidencia o potencial delas em representar gradientes ambientais associados à variabilidade espacial do carbono orgânico no solo. Em uma ampla revisão de literatura sobre mapeamento digital de solos, Chen *et al.* (2022) constataram que os estoques de carbono geralmente se relacionam melhor com covariáveis ambientais climáticas (ligadas à temperatura e precipitação), à elevação e ao material de origem. Embora o presente estudo não tenha se utilizado de preditores climáticos e de material de origem, a covariável topográfica de elevação teve sua elevada importância confirmada mostrando conformidade com os relatos da literatura.

Após a redução dimensional para as 15 covariáveis mais importantes em cada abordagem (RF_all, RF_Corr e RF_CorrImp), foram identificados padrões relacionados à contribuição da elevação, das bandas espectrais e dos atributos de curvatura. A elevação (mde) destacou-se como uma das covariáveis mais relevantes em todas as abordagens. No RF_all, foi a covariável mais importante nas camadas de 15-30 e 60-100 cm. Na abordagem RF_Corr, figurou entre as três covariáveis mais importantes nas camadas abaixo de 5 cm. Já em RF_CorrImp, sua relevância tornou-se mais evidente a partir da camada de 15-30 cm. Esses resultados sugerem que a influência da elevação aumenta progressivamente com a profundidade do solo.

As covariáveis espectrais derivadas de bandas Landsat também apresentaram presença consistente nos subconjuntos Top15. Na abordagem RF_all, elas estiveram presentes em todas as profundidades analisadas. Na abordagem RF_Corr, mantiveram-se presentes em todas as camadas, embora em menor quantidade. Por sua vez, na RF_CorrImp, a participação das covariáveis espectrais torna-se mais expressiva a partir da camada de 15-30 cm. Mesmo após a remoção média de 10 covariáveis espectrais durante as etapas de seleção, as bandas Landsat permaneceram presentes em todos os subconjuntos Top15, independentemente da abordagem empregada. Esse resultado reforça a relevância das informações desse grupo de covariáveis na predição do ECOS ao longo de todo o perfil do solo.

Hateffard *et al.* (2023), em uma planície aluvial no leste da Hungria, concluíram que o NDVI (Índice de Vegetação por Diferença Normalizada), relacionado à densidade e vigor da vegetação, foi uma das covariáveis ambientais de maior importância para a predição do carbono orgânico do solo e

seu estoque em superfície (0-10 cm). Neste estudo, o NDVI foi removido em todas as profundidades tanto na abordagem RF_Corr quanto em RF_CorrImp (Tabela 6 e Tabela 7) e não apareceu no subconjunto de Top15 covariáveis na abordagem RF_all. Em contrapartida, as bandas espectrais do vermelho (b4red) (inferência sobre vigor vegetal) e do infravermelho próximo (b5nir) (inferência sobre vigor vegetal, densidade vegetal e delimitação de áreas úmidas e corpos hídricos), que são utilizadas para o cálculo do NDVI, figuram nos subconjuntos Top15 das três abordagens testadas.

A relevância dessas duas covariáveis está de acordo com os achados de Mahdavi *et al.* (2018), que relatam as bandas do vermelho e do infravermelho próximo com as mais relevantes para identificação de áreas úmidas. Como essas bandas respondem ao vigor e à densidade da vegetação, bem como à presença de umidade superficial, sua importância sugere que características associadas à vegetação e ao regime hidrológico exerceram influência significativa sobre a variabilidade espacial do ECOS na área de estudo. Essa interpretação é consistente com a literatura que demonstra que áreas úmidas e planícies de inundação tendem a apresentar maiores estoques de carbono orgânico (Mitsch *et al.*, 2013; Wang *et al.*, 2019; Ouyang; Lee, 2020).

Resultados semelhantes foram reportados por Forkuor *et al.* (2017), que, ao modelarem o COS em escala de bacia hidrográfica utilizando o algoritmo *Random Forest*, identificaram as bandas do vermelho visível (b4red) e do infravermelho próximo (b5nir), juntamente com a elevação (mde) e o Índice de Brilho do Solo (SBI), como as covariáveis de maior importância para a predição do COS na camada de 0–30 cm. Da mesma forma, Song *et al.* (2017) verificaram que a elevação e a banda do vermelho visível foram os principais fatores associados à variabilidade espacial da matéria orgânica do solo.

As covariáveis topográficas relacionadas à curvatura (maxc, proc, tc, flc, lc, csc, tsc), também demonstraram elevada consistência, estando presentes em todas as abordagens e profundidades avaliadas. Apesar de ser esperado que covariáveis ligadas à curvatura tenham alta correlação entre si devido às suas dependências matemáticas (Florinsky, 2017), a consistência delas nos subconjuntos Top15 indica que os atributos de curvatura exercem papel importante tanto nas camadas superficiais quanto nas mais profundas,

refletindo a influência da microtopografia e dos processos hidrológicos na distribuição do ECOS, corroborando os achados de Hateffard *et al.* (2023), em que a curvatura do plano foi um dos fatores dominantes para explicar a variabilidade espacial do ECOS.

Com relação ao balanço entre covariáveis topográficas e espectrais nos subconjuntos Top15, observa-se que a abordagem RF_all preserva melhor a contribuição de covariáveis espectrais, enquanto as abordagens de seleção favorecem fortemente as covariáveis topográficas. Contudo, esse resultado é coerente com a remoção sistemática de covariáveis espectrais por correlação. Além disso, a maior relevância das covariáveis topográficas observada neste estudo está de acordo com os resultados de Piccini *et al.* (2014), que, ao utilizarem a análise de componentes principais (PCA) acoplada ao processo de predição de carbono no solo, verificaram que o primeiro componente principal (responsável pela maior parcela da variabilidade dos dados) estava predominantemente associado a atributos derivados do MDE, enquanto os índices espectrais derivados de bandas Landsat apresentaram maior correlação com componentes secundários, de menor importância para a predição. Em conjunto, esses resultados sugerem que as covariáveis topográficas capturam uma parcela mais expressiva da variabilidade relacionada à distribuição espacial do carbono do que as covariáveis espectrais na área de estudo, uma vez que as covariáveis topográficas atuam como indicadores da dinâmica da paisagem, refletindo processos de redistribuição de água, erosão, deposição de sedimentos, drenagem e estabilidade das superfícies, os quais condicionam o acúmulo e a preservação da matéria orgânica.

Conforme destacado anteriormente, a persistência de covariáveis espectrais (bandas e índices) em todas as profundidades, mesmo nas abordagens de seleção, sugerindo a influência da vegetação sobre o ECOS em todo o perfil do solo é um achado importante do presente estudo, pois as covariáveis espectrais representam condições da cobertura vegetal e do uso da terra, que afetam o aporte de resíduos orgânicos, a proteção da superfície do solo, a ciclagem de nutrientes e o microclima, fatores que controlam indiretamente o acúmulo e a estabilização do carbono orgânico. Essa relação já foi evidenciada na literatura por Bhunia *et al.* (2019), que, por meio de uma

abordagem multivariada, verificaram associação significativa entre o carbono orgânico do solo e os índices espectrais NDVI e BSI. Os autores obtiveram um R^2 ajustado de 0,71, indicando que aproximadamente 71% da variabilidade do COS pode ser explicada por essas duas variáveis. Embora o estudo tenha sido conduzido na camada superficial do solo (0–20 cm), com 50 amostras em uma região da Índia, seus resultados reforçam o potencial dos atributos espectrais da vegetação como preditores do carbono do solo, corroborando a relevância dessas covariáveis observada no presente estudo. Resultados semelhantes foram relatados por Bouslihim *et al.* (2024) no mapeamento da matéria orgânica do solo (MOS) em Marrocos, onde após a aplicação do algoritmo Boruta, os autores reduziram o conjunto de covariáveis ambientais de 29 para 20 covariáveis, sendo que apenas uma delas não possuía origem espectral. Mesmo com essa redução, o modelo *Random Forest* apresentou elevado desempenho preditivo ($R^2 = 0,90$), evidenciando a forte contribuição das informações espectrais para a modelagem dos atributos do solo.

Em ambientes de topografia plana, pode ser desafiador identificar relações significativas entre covariáveis ambientais derivadas de MDE e de sensoriamento remoto. Por essa razão, alguns autores sugerem ampliar o conjunto de preditores e incluir dados litológicos, propriedades químicas e físicas, tipo de solo e de paisagem para melhorar a predição espacial (Bou Kheir *et al.*, 2010; John *et al.*, 2020). Na disponibilidade de covariáveis climáticas, uso do solo e tipo de solo, os parâmetros de terreno têm importância menor na predição do carbono no solo, com a precipitação sendo a covariável mais dominante, a geologia tornando-se mais importante com o aumento da profundidade e o uso do solo sendo mais importante para predições superficiais (Adhikari *et al.*, 2014). Ao passo que para Xiong *et al.* (2014), o uso e cobertura da terra e propriedades ligadas à dinâmica da água no solo foram os principais fatores que explicaram a variação espacial do COS, seguidos pela classificação taxonômica dos solos e índices de vegetação enquanto topografia e clima mostraram efeitos moderados sobre o COS. De forma semelhante, Adeniyi *et al.* (2024), ao mapear o COS em uma planície aluvial com solos de textura arenosa na Itália, destacaram que covariáveis derivadas de MDE relacionadas aos processos hidrológicos apresentam maior

potencial para representar a variabilidade espacial do carbono orgânico do solo em áreas de planície.

3.3.5 Melhor abordagem

Os valores de R^2 máximos observados (0,103 em 15-30 cm para RF_all e RF_CorrImp, Tabela 12) indicam que os modelos explicam apenas cerca de 10% da variabilidade do ECOS (Tabela 9), o que reforça a ideia de que o melhor desempenho preditivo não implica, necessariamente, estimativas acuradas (Mosleh *et al.*, 2016). Este baixo desempenho é consistente com a alta variabilidade observada nos dados (CV 40-52%) (Tabela 5) e sugere que as covariáveis ambientais utilizadas não capturam completamente os fatores que controlam o ECOS no ambiente costeiro estudado. Ademais, Hateffard *et al.* (2024) chamam atenção para o fato de que mapear áreas não contempladas nos dados de treinamento (extrapolação geográfica) pode comprometer a precisão da predição, aspecto particularmente relevante considerando que a validação cruzada espacial *Leave-One-Cluster-Out* empregada neste estudo simula condições de extrapolação entre as cinco áreas de coleta. O BIAS é geralmente pequeno em magnitude ($< 0,5$ para a maioria dos casos), indicando que os modelos não apresentam tendência sistemática de superestimação ou subestimação. Valores negativos predominam, sugerindo leve tendência de subestimação em alguns casos (Tabela 9 e Tabela 10). Não havendo um padrão claro de superioridade de uma abordagem sobre as outras.

As diferenças entre estratégias são pequenas em termos absolutos (diferenças de $R^2 < 0,05$ na maioria dos casos), sugerindo que a seleção de covariáveis não teve impacto substancial no desempenho preditivo. Os valores atingidos pelo coeficiente de determinação (R^2) nas melhores abordagens selecionadas para cada profundidade evidenciam o desafio que é modelar o ECOS em grandes áreas, com número limitado de amostras e em ambiente de baixa variação topográfica (Tabela 12). Nesse contexto, Carbajal *et al.* (2025) também observaram desempenho limitado do *Random Forest* em bacias hidrográficas dos Andes peruanos, com valores de R^2 de 0,11 e até mesmo negativos, indicando baixa capacidade preditiva do modelo. Os autores atribuíram esses resultados a fatores como seleção inadequada de covariáveis,

ajuste insuficiente dos hiperparâmetros, reduzida disponibilidade de dados de treinamento e ausência de covariáveis espaciais relevantes.

Tabela 12 - Melhor abordagem de modelagem *Random Forest* para estimativa do ECOS em diferentes profundidades

Profundidade	Melhor abordagem	R ²	RMSE	mtry
0-5 cm	RF_all (Top15)	0,066	3,453	1
5-15 cm	RF_CorrImp (todas)	0,103	4,496	1
15-30 cm	RF_all (todas)	0,103	5,846	1
30-60 cm	RF_all (todas)	0,064	9,346	1
60-100 cm	RF_all (Top15)	0,093	8,735	2

Fonte: Elaborado pela autora (2026).

Os resultados obtidos neste estudo atingiram valores semelhantes aos encontrados por John *et al.* (2020) na predição do conteúdo de COS até a profundidade de 30 cm, onde o R² de validação com o método de Regressão Linear Múltipla (MLR) foi de 0,17, ao passo que o R² obtido com a aplicação de *Random Forest*, foi de 0,68. Contudo, nesse caso, que foi modelado para uma planície aluvial na Nigéria, a densidade amostral era de uma observação por km². Gelsleichter *et al.* (2025) com um número semelhante de amostras (n=70), porém em uma área de 3500 km², conseguiram, através da utilização da validação cruzada, métricas estáveis tanto para validação quanto para calibração do modelo RF para COS, atingindo R²=0,43 e 0,41, respectivamente. Enquanto Vaysse *et al.* (2017), em uma área de estudo de, aproximadamente, 27000 km², obtiveram resultados de R² de validação de 0,34 para a predição de COS na França.

Há preferência pela abordagem RF_all em 4 das 5 profundidades (Tabela 12), com apenas uma profundidade (5-15 cm) usando a abordagem RF_CorrImp. Isso sugere que a abordagem com todas as covariáveis originais (RF_all) foi considerada geralmente adequada, possivelmente por manter o equilíbrio entre covariáveis espectrais e topográficas. O hiperparâmetro mtry = 1 em quatro das cinco profundidades indica baixa complexidade dos modelos, consistente com o baixo número de amostras (n=64).

O baixo desempenho preditivo observado neste estudo pode estar relacionado às características geomorfológicas da área avaliada, marcada por relevo predominantemente plano e baixa variabilidade dos atributos derivados do terreno. Nessas condições, a reduzida amplitude dos gradientes

topográficos pode limitar a capacidade das covariáveis de relevo em representar a variabilidade espacial do ECOS, restringindo a capacidade dos modelos em estabelecer relações preditivas robustas. Contudo, é importante destacar que, embora a variabilidade topográfica seja reduzida, a área não se apresenta homogênea do ponto de vista pedológico. A caracterização dos 64 pontos de coleta revelou seis classes de solo distintas, com espessura média do horizonte A variando de 15 a 90 cm e limite superior do horizonte B entre 20 e 155 cm (Tabela 1). Essa heterogeneidade pedológica indica que processos de formação do solo (sedimentação, oscilação do lençol freático e iluviação) representam uma variabilidade que as covariáveis de terreno não conseguem capturar.

Entretanto, limitações associadas à complexidade do relevo também são observadas em ambientes com características opostas. Konra *et al.* (2026), ao avaliar a predição espacial do ECOS em uma região altamente montanhosa, atribuíram as limitações dos modelos à elevada heterogeneidade do terreno e à complexidade dos processos ambientais associados ao relevo. Os autores obtiveram valores de R^2 de 0,12 e 0,08 para redes neurais artificiais nas camadas de 0–15 e 15–30 cm, respectivamente, enquanto o algoritmo *Random Forest* apresentou desempenho superior, com R^2 de 0,41 e 0,36 para as mesmas profundidades. Embora os valores obtidos por *Random Forest* tenham sido superiores aos encontrados neste estudo, os baixos valores observados por alguns modelos reforçam que a predição do ECOS permanece desafiadora mesmo em ambientes com maior disponibilidade de gradientes ambientais. Dessa forma, tanto ambientes extremamente homogêneos, como áreas planas com baixa variação topográfica, quanto ambientes altamente heterogêneos, como regiões montanhosas, podem impor limitações ao MDS, embora por mecanismos distintos.

3.3.6 Espacialização do ECOS

A média do ECOS dos mapas preditivos (Tabela 13) é muito próxima da média de ECOS dos pontos de coleta (Tabela 5), indicando que os modelos não apresentam viés sistemático de superestimação ou subestimação em termos de média. Os mapas apresentam amplitude menor que os dados observados (aproximadamente 50-60% da amplitude observada). Este é um

padrão típico de modelos de aprendizado de máquina baseados em árvores, que tendem a suavizar previsões e não capturar valores extremos (Konra *et al.*, 2026), uma vez que não são capazes de prever valores fora da faixa dos mínimos e máximos observados nos dados de treinamento (Hateffard *et al.*, 2024). O CV dos mapas é menor que o CV dos dados observados (aproximadamente 20-25% do CV observado). Isso indica que os mapas são muito mais homogêneos espacialmente do que os dados observados, refletindo a suavização característica de modelos preditivos. Além disso, a média dos mapas reproduz a tendência de aumento do ECOS com a profundidade até 30-60 cm, com redução em 60-100 cm.

Tabela 13 - Estatística descritiva dos mapas preditivos de estoques de carbono orgânico do solo (ECOS) espacializados pelo modelo *Random Forest* em diferentes profundidades do solo

Profundidade	Média (Mg ha ⁻¹)	Mínimo (Mg ha ⁻¹)	Máximo (Mg ha ⁻¹)	DP (Mg ha ⁻¹)	CV%
0-5 cm	8,85	5,80	14,55	0,88	9,96
5-15 cm	11,51	7,19	18,48	0,74	6,39
15-30 cm	12,97	7,94	20,96	1,16	8,91
30-60 cm	20,65	12,18	30,61	1,54	7,45
60-100 cm	16,70	7,69	34,27	2,34	13,99

Fonte: Elaborado pela autora (2026).

Apesar das limitações associadas à predição espacial do ECOS, as estimativas de ECOS obtidas neste estudo apresentaram elevada concordância com os valores reportados por Akpa *et al.* (2016). O estoque médio de carbono até 1 m de profundidade foi semelhante entre os estudos, com 70,68 Mg ha⁻¹ estimados neste trabalho e 71,6 Mg ha⁻¹ encontrados pelos autores. Além disso, o padrão de distribuição vertical do ECOS foi consistente entre as áreas avaliadas. Na camada de 0–30 cm, este estudo estimou uma contribuição relativa de 47,1% do estoque total, valor próximo aos 48% reportados por Akpa *et al.* (2016). Para as camadas de 30–60 cm e 60–100 cm, as proporções estimadas foram de 29,2% e 23,6%, respectivamente, enquanto os autores encontraram valores de 27% e 25%. Essa proximidade nos valores absolutos e na distribuição ao longo do perfil indica que, apesar das incertezas inerentes ao processo de modelagem, as estimativas geradas foram capazes de representar adequadamente a magnitude e o padrão vertical de armazenamento de carbono no solo descritos na literatura.

A incerteza (CV médio) aumenta progressivamente com a profundidade (Tabela 14), sendo coerente com aumento da variabilidade dos dados observados (CV de 40,1% em 0-5 cm para 52,1% em 60-100 cm) (Tabela 5), aumento do RMSE dos modelos (3,4 em 0-5 cm para 9,3 em 30-60 cm) (Tabela 12) e a maior dificuldade de predição em camadas profundas, onde os fatores de controle são mais complexos e menos relacionados às covariáveis superficiais.

Tabela 14 - Estatísticas descritivas da incerteza dos mapas preditivos de ECOS, expressa pelo coeficiente de variação (CV), em diferentes profundidades do solo

Profundidade	CV médio (%)	DP (%)	CV do CV (%)	Mínimo (%)	Máximo (%)
0-5 cm	32,6	5,7	17,5	15,7	66,3
5-15 cm	33,4	4	12,1	18,6	57,3
15-30 cm	38,3	4,8	12,5	22,2	63,3
30-60 cm	38,7	4,5	11,5	21	75,1
60-100 cm	46,8	6,5	14	19,9	102,7

Fonte: Elaborado pela autora (2026).

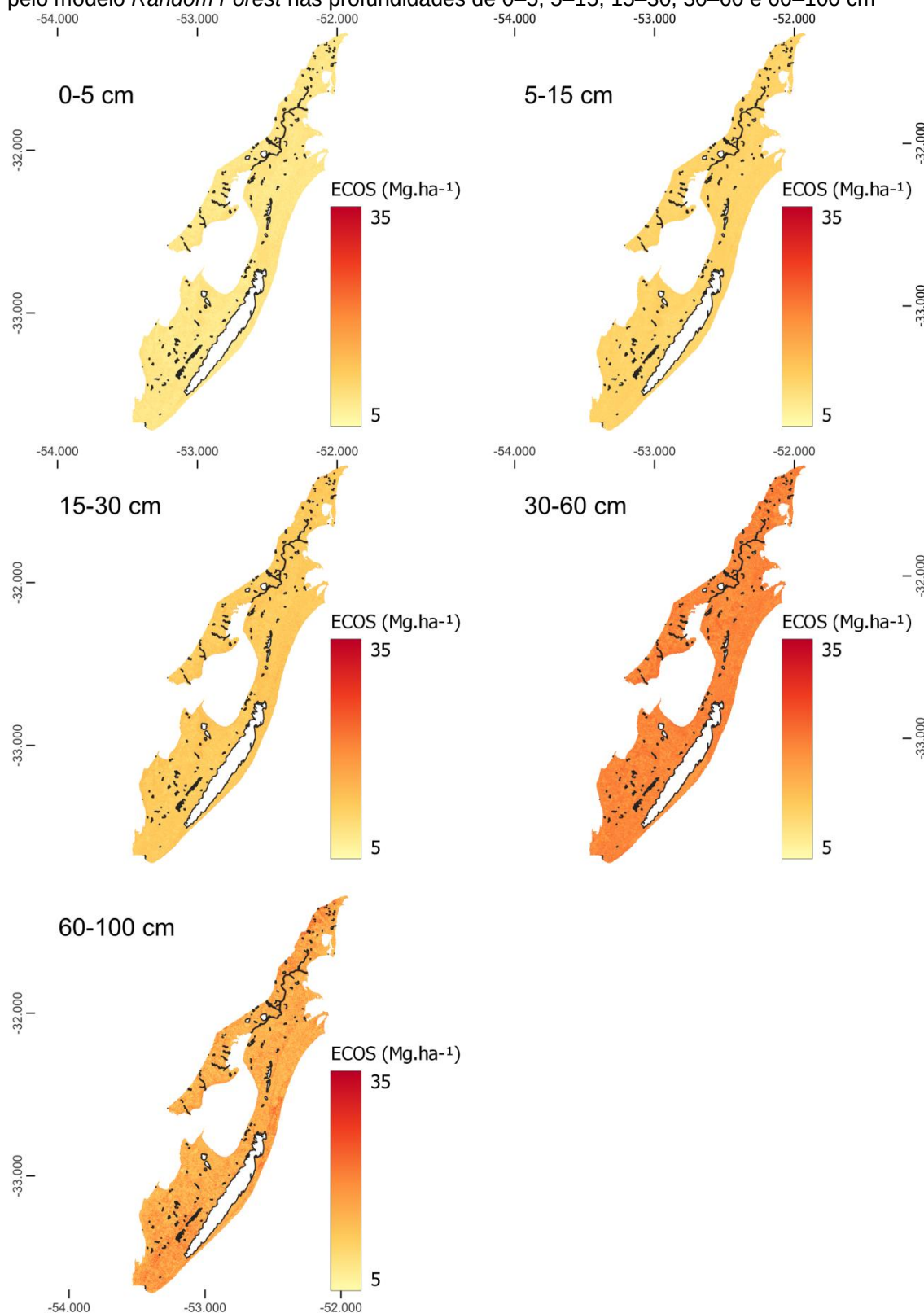
A redução da precisão das predições de atributos do solo em profundidade é um fenômeno frequentemente relatado na literatura (Adhikari *et al.*, 2014; Zhang *et al.*, 2022). Entretanto, resultados apresentados por Somarathna *et al.* (2017) indicam que uma densidade amostral adequada e representativa pode contribuir para a melhoria das predições, mesmo em camadas mais profundas. Os autores observaram que o aumento do tamanho da amostra de treinamento resultou em um incremento de 14% na precisão das predições na camada superficial e de 8% na camada subsuperficial, utilizando regressão linear múltipla. Esses resultados sugerem que, embora a adição de amostras promova melhorias na acurácia dos modelos, os ganhos tendem a ser menores em profundidade, possivelmente devido à maior complexidade dos padrões de distribuição do carbono e à menor relação entre os preditores ambientais e os teores de COS e ECOS nas camadas subsuperficiais.

Os valores de CV médio > 30% em todas as profundidades indicam predições com baixa precisão relativa, comportamento compatível com o baixo R^2 dos modelos (Tabela 12). Valores de CV do coeficiente de variação dos mapas (CV do CV) entre 11 e 18% indicam que a incerteza é moderadamente variável espacialmente, mas não extremamente heterogênea. A camada superficial (0-5 cm) apresenta maior variabilidade espacial da incerteza

(17,5%), possivelmente refletindo maior heterogeneidade das condições superficiais. O valor máximo do CV de 102,7% em 60-100 cm indica que há áreas onde a incerteza é maior que a própria predição média, sugerindo predições pouco confiáveis em algumas localidades da camada mais profunda. A incerteza (CV médio) dos mapas é menor que a variabilidade dos dados observados, sendo este um resultado esperado, pois a incerteza mede a variabilidade das predições entre árvores do *Random Forest*, não a variabilidade total dos dados. No entanto, a incerteza aumenta com a profundidade de forma coerente com o aumento da variabilidade observada.

A espacialização do ECOS pode ser observada na Figura 11. A partir dos mapas preditivos de ECOS, pode-se observar que na faixa de 0-30 cm existe um padrão de valores mais elevados de ECOS em áreas de baixa elevação com solos hidromórficos e uso do solo associados à água (cultivo irrigado) e à vegetação campestre/alagada (Apêndice K). Esse padrão está associado à maior permanência de umidade na superfície do solo e menor taxa de decomposição da matéria orgânica. A partir da profundidade de 30 cm, manchas com maiores valores de ECOS se tornam mais dispersas, exigindo maior cautela na interpretação dos dados (Apêndice L). Além disso, é possível observar que áreas de transição entre classes de ECOS ocorrem preferencialmente em zonas que correspondem a feições de transição geomorfológica (sistema laguna-barreira).

Figura 11 - Distribuição espacial dos estoques de carbono orgânico do solo (ECOS) estimados pelo modelo *Random Forest* nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm

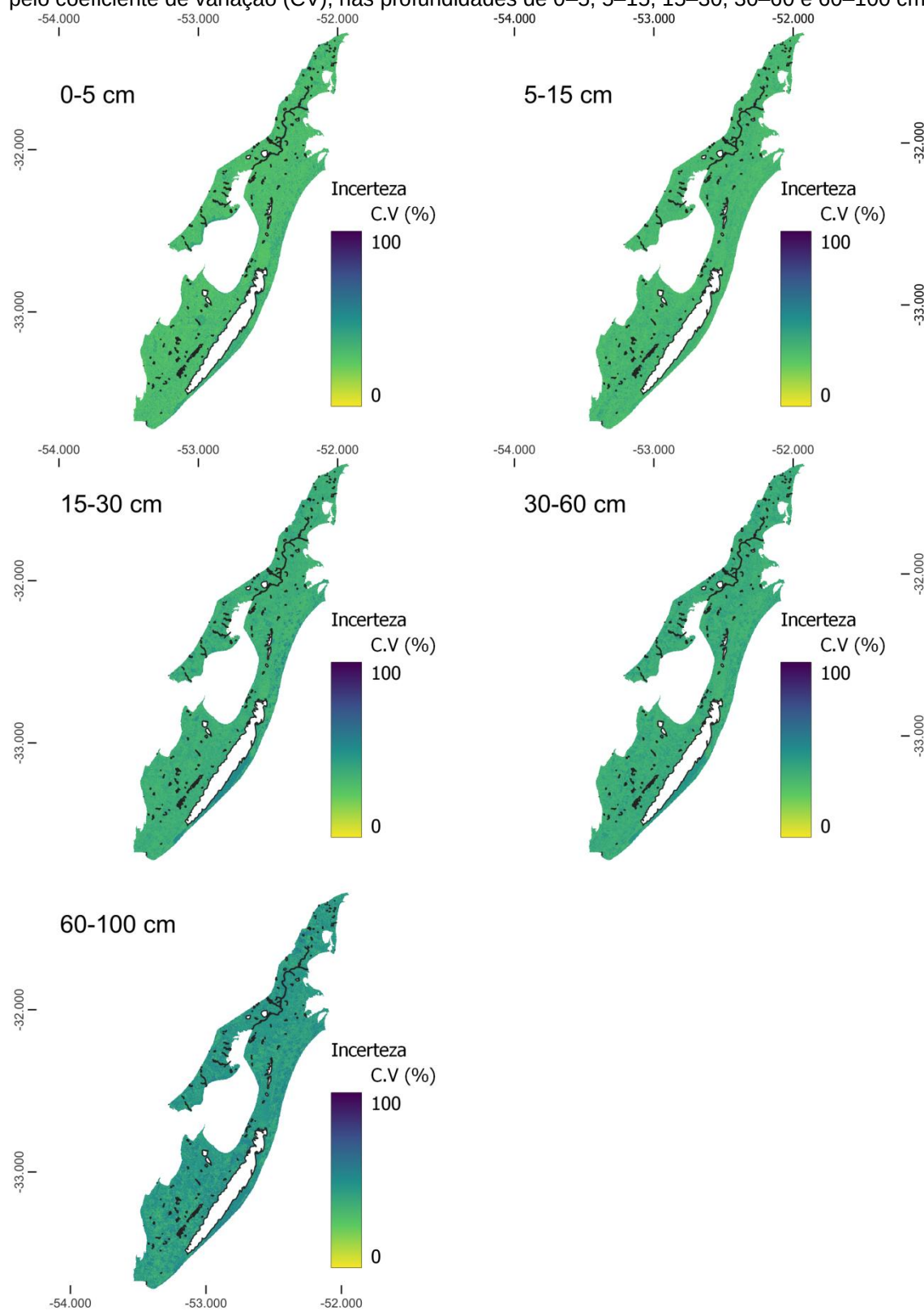


Fonte: Elaborado pela autora (2026).

Os mapas de incerteza preditiva de ECOS mostram que a confiabilidade das predições de ECOS varia espacialmente e em profundidade (Figura 12).

Considerando que a quantificação da incerteza constitui uma etapa essencial na avaliação de modelos, sobretudo quando seus resultados são empregados como suporte à tomada de decisões (Padarian *et al.*, 2020), a ocorrência de zonas de maior CV indica áreas onde a interpretação dos mapas preditivos deve ser mais cautelosa, sendo que a camada 60-100 cm exige atenção adicional pela maior amplitude de CV (Apêndice M e Apêndice N). Além da seleção de covariáveis, outros fatores podem contribuir para a redução das incertezas associadas ao mapeamento do ECOS. Szatmári *et al.* (2019) destacaram três estratégias principais para aumentar a confiabilidade das predições: a adoção de técnicas de modelagem espacial que produzam menores níveis de incerteza, o emprego de covariáveis ambientais com maior resolução espacial e temática e o aprimoramento do desenho amostral por meio da otimização da densidade e da distribuição espacial dos pontos de amostragem.

Figura 12 - Distribuição espacial da incerteza associada às estimativas de ECOS, expressa pelo coeficiente de variação (CV), nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm



Fonte: Elaborado pela autora (2026).

Como interpretação técnica, os mapas indicam que a confiabilidade espacial das predições de ECOS não é homogênea. As áreas com CV mais alto devem ser tratadas com maior cautela em análises quantitativas, comparação entre profundidades, extrapolação espacial e tomada de decisão. No contexto do Litoral Sul do Rio Grande do Sul, a ocorrência de manchas de incerteza pode ser compatível com transições ambientais associadas a variações, mesmo que sutis, no relevo, variações de drenagem, umidade, solos hidromórficos, feições lagunares ou gradientes costeiros.

Em suma, em áreas de baixa variabilidade ambiental, o desempenho do MDS pode ser limitado pela reduzida capacidade das covariáveis em explicar a variabilidade dos atributos do solo, reforçando a necessidade de novas fontes de informação e métodos capazes de capturar relações mais complexas entre covariáveis ambientais e propriedades do solo (Mosleh *et al.*, 2016). A baixa precisão dos modelos deve ser interpretada à luz da escala e da complexidade do estudo, que envolveu o mapeamento do ECOS em cinco camadas de profundidade em uma área extensa e com número limitado de amostras de referência. Conforme discutido por Piikki *et al.* (2021), o mapeamento de propriedades do solo em múltiplas profundidades requer uma base amostral e de covariáveis compatível com a complexidade do problema, de modo que limitações nesses recursos podem comprometer a qualidade das predições.

3.4 Conclusões

O mapeamento digital do estoque de carbono orgânico do solo (ECOS) no ambiente costeiro de baixa variação topográfica do Litoral Sul do Rio Grande do Sul mostrou-se desafiador e complexo, com os modelos *Random Forest* apresentando capacidade preditiva limitada ($R^2 \leq 0,103$) em todas as profundidades avaliadas. A hipótese de que seria possível prever o ECOS com precisão aceitável não foi confirmada, uma vez que as covariáveis ambientais utilizadas, embora descrevam adequadamente características superficiais da paisagem, representam apenas parcialmente os processos pedogenéticos, geológicos e hidrológicos responsáveis pelo acúmulo de carbono em subsuperfície. Ainda assim, o estoque médio estimado até 1 m de profundidade ($70,68 \text{ Mg ha}^{-1}$) e o padrão de distribuição vertical ao longo do perfil mostraram-se consistentes com valores reportados na literatura.

A análise das estratégias de seleção de covariáveis evidenciou que aproximadamente metade dos preditores originais pôde ser removida sem comprometer significativamente o desempenho dos modelos, confirmando a forte multicolinearidade do conjunto de dados. Contudo, nenhuma estratégia de seleção resultou em ganhos preditivos consistentes em comparação ao uso de todas as covariáveis. A seleção baseada apenas em correlação apresentou os piores resultados, indicando que a eliminação de covariáveis sem considerar sua relevância preditiva pode reduzir a capacidade de modelagem. A abordagem RF_all (sem filtragens prévias) foi a melhor estratégia em quatro das cinco profundidades, sugerindo que a remoção antecipada de covariáveis pode eliminar preditores relevantes. Esses resultados indicam que, em ambientes costeiros de baixa variação topográfica, o desempenho dos modelos depende menos da quantidade de covariáveis e mais da capacidade dessas covariáveis de representar os processos que controlam a distribuição do carbono no solo.

A importância relativa das covariáveis variou ao longo do perfil, refletindo mudanças nos mecanismos de controle do ECOS em diferentes profundidades. Nas camadas superficiais, o ECOS esteve mais associado a fatores dinâmicos como microtopografia e cobertura vegetal, enquanto nas camadas mais profundas destacaram-se atributos relacionados à posição na paisagem e à elevação, comportamento associado à relação solo-paisagem e aos processos

de formação da planície costeira. Mesmo em relevo de baixa amplitude, as covariáveis topográficas apresentaram contribuição preditiva semelhante ou superior à das covariáveis espectrais. Os mapas produzidos reproduziram a tendência geral dos dados observados, mas apresentaram suavização espacial típica de modelos de aprendizado de máquina, com incerteza aumentando progressivamente com a profundidade e atingindo os maiores valores na camada de 60–100 cm.

Apesar das limitações observadas, este trabalho contribui para o entendimento dos processos que controlam o ECOS em ambientes costeiros e demonstra que, mesmo em paisagens aparentemente homogêneas, existem gradientes ambientais capazes de influenciar a distribuição espacial do carbono no solo. O aprimoramento da capacidade preditiva depende da incorporação de informações complementares (dados geológicos e litológicos, dinâmica do lençol freático, histórico de uso da terra e atributos obtidos diretamente em campo), do aumento da densidade amostral e do uso de séries temporais de imagens. A variação dos fatores controladores ao longo do perfil reforça ainda a importância de adotar abordagens específicas para cada profundidade, incorporando a dimensão vertical nos estudos de mapeamento digital do solo.

4 Capítulo II – Seleção de covariáveis e predição espacial de densidade do solo, teor de areia e pH no Litoral Sul do Rio Grande do Sul

4.1 Introdução

A heterogeneidade dos solos em escala espacial influencia significativamente sua função ecológica e produtiva, justificando assim a necessidade crescente de informações detalhadas sobre a variabilidade dos solos para enfrentar desafios globais como as mudanças climáticas, segurança alimentar e gestão ambiental (Minasny; McBratney, 2016). A compreensão dessa variabilidade espacial é essencial para o manejo sustentável dos solos, pois permite intervenções estratégicas que minimizem impactos negativos, otimizem a produtividade agrícola e que contribuam com a conservação ambiental.

O mapeamento digital de solos (MDS) permite predições mais precisas e acessíveis sobre a variabilidade do solo (Hengl *et al.*, 2017; Poggio *et al.*, 2021). Com o avanço das técnicas preditivas, tornou-se possível preencher lacunas, fornecendo informações valiosas para a gestão do recurso solo. Modelos baseados em aprendizado de máquina e estatística espacial permitiram que fatores ambientais fossem utilizados para explicar a variabilidade espacial dos solos, revelando padrões até então difíceis de detectar (McBratney *et al.*, 2019). Portanto, estimar com maior precisão a distribuição de atributos do solo como pH, composição granulométrica e densidade do solo não apenas melhora o planejamento agrícola, mas também contribui para o uso mais sustentável dos recursos naturais.

O pH dos solos é uma variável fundamental na caracterização de suas propriedades (Thomas, 1996). A acidez do solo é determinada pela composição mineral do mesmo e por reações de troca iônica e hidrólise de seus componentes (Thomas; Hargrove, 1984). Por isso, o pH do solo é um dos principais reguladores da disponibilidade de nutrientes e da atividade microbiana, influenciando diretamente o metabolismo das plantas e a interação dos organismos no solo. Em MDS, o pH está entre os atributos químicos mais frequentemente modelados, com desempenhos que variam amplamente conforme a escala, a densidade amostral e o contraste ambiental da área de estudo. Enquanto alguns trabalhos reportam elevada capacidade preditiva,

como por exemplo, R^2 de 0,76 na camada de 5–15 cm (Vaysse *et al.*, 2017) e R^2 entre 0,45 e 0,54 com diferentes algoritmos de aprendizado de máquina (Ma *et al.*, 2021); outros obtêm resultados modestos, como R^2 de 0,26 associado à baixa variabilidade do atributo (Bouslih *et al.*, 2024) ou R^2 entre 0,03 e 0,11 em condições subtropicais brasileiras (Carvalho Junior *et al.*, 2014). Assim, embora a modelagem do pH seja consolidada na literatura, o bom desempenho preditivo depende fortemente do contexto pedológico e da representatividade das covariáveis ambientais.

A granulometria do solo, que é caracterizada pela proporção de areia, silte e argila, afeta a estrutura física e capacidade de retenção de água dos solos, bem como outros processos físicos que ocorrem no sistema solo-planta-atmosfera, além de contribuir com aspectos químicos e nutricionais, como por exemplo, a proteção dos estoques de carbono. Plante *et al.* (2006) evidenciam que solos mais argilosos tendem a estabilizar melhor o carbono orgânico devido à maior superfície específica e à formação de microagregados, enquanto solos arenosos possuem menor capacidade de retenção de matéria orgânica. Além disso, a granulometria afeta a interação dos microrganismos com o solo e sua capacidade de promover processos bioquímicos essenciais. Solos de textura fina, devido à sua maior retenção de água, apresentam produtividade superior em comparação aos de textura grossa. Além disso, a eficiência no uso da água tende a ser mais elevada nesses solos, reforçando sua importância na manutenção da produtividade agrícola (Jalota *et al.*, 2010). Essas características influenciam diretamente a composição vegetal, favorecendo maior quantidade de vegetação arbustiva em solos arenosos, enquanto solos argilosos beneficiam gramíneas e herbáceas (Sanaei *et al.*, 2019).

A relação entre textura do solo, disponibilidade hídrica e distribuição das plantas destaca o papel da granulometria na regulação dos ecossistemas agrícolas e naturais. Além disso, a textura do solo influencia diretamente o crescimento das raízes e a interação das plantas com microrganismos benéficos. Solos argilosos, devido à sua estrutura compacta e menor aeração, podem restringir o desenvolvimento radicular e induzir uma maior liberação de exsudatos, favorecendo a absorção de nutrientes e atraindo microrganismos benéficos. Por outro lado, solos arenosos, mais soltos, permitem uma

penetração radicular facilitada, resultando em menor produção de exsudatos (Adeniji *et al.*, 2024). No que se refere à modelagem, as frações granulométricas também são amplamente previstas no MDS, embora areia, silte e argila não apresentem desempenho equivalente. Em escala global, produtos como o SoilGrids indicam boa capacidade de predição textural quando há gradientes climáticos, litológicos e topográficos amplos e grande volume de dados de treinamento (Hengl *et al.*, 2014, 2017). Em escalas regionais, o desempenho é mais heterogêneo: Forkuor *et al.* (2017) obtiveram R^2 de 0,34 para o teor de areia em bacia hidrográfica; Adhikari *et al.* (2013) e Akpa *et al.* (2014) observaram declínio da capacidade preditiva da areia com a profundidade; e, em solos com predomínio de areia, Fathizad *et al.* (2025) reportaram R^2 de 0,76 para essa fração, enquanto em alguns contextos sul-americanos os valores foram mais baixos, na ordem de 0,13 a 0,27 (Ließ *et al.*, 2012; Carvalho Junior *et al.*, 2014).

Já a densidade do solo é um atributo físico diretamente relacionado à compactação e à porosidade. Solos compactados tendem a apresentar menor taxa de infiltração e disponibilidade de oxigênio, impactando negativamente o crescimento radicular e o metabolismo dos microorganismos do solo. Estudos indicam que práticas agrícolas intensivas, sejam elas pelo preparo do solo ou pelo tráfego de máquinas e animais, podem aumentar a densidade do solo, reduzindo sua capacidade de suporte para culturas agrícolas e elevando os riscos de degradação estrutural (Pulido *et al.*, 2018; Gao *et al.*, 2019; Bondi *et al.*, 2021; Topa *et al.*, 2021). A densidade do solo figura entre os atributos físicos frequentemente modelados no MDS, porém costuma ser considerada mais desafiadora do que o pH ou a textura quando se utilizam apenas covariáveis ambientais, pois é um atributo do solo que responde a processos relacionados ao uso e manejo do solo (tráfego, preparo, pastejo) nem sempre capturados pelos preditores mais comumente utilizados em MDS. Resultados na literatura variam de R^2 elevados em áreas pequenas e densamente amostradas (Hateffard *et al.*, 2023) a valores moderados em escalas nacionais (Akpa *et al.*, 2016), sendo o desempenho geralmente superior quando se incluem preditores intrínsecos ao solo, como carbono orgânico e CTC (Carvalho Junior *et al.*, 2016).

De modo geral, a literatura demonstra que pH, densidade do solo e teor de areia já foram amplamente modelados no MDS, sobretudo com algoritmos de aprendizado de máquina como o *Random Forest*. Contudo, o desempenho preditivo é atributo-específico e fortemente dependente do contexto ambiental. Boa parte dos estudos com resultados mais favoráveis concentra-se em áreas com gradientes ambientais contrastantes como relevo mais movimentado, diversidade climática e litológica, ou escalas nacionais e globais com grande volume amostral (Hengl *et al.*, 2017; Akpa *et al.*, 2014, 2016; Hateffard *et al.*, 2023). Ambientes com solos de textura média a argilosa e variação topográfica expressiva tendem a oferecer maior contraste nas covariáveis de relevo e nas assinaturas espectrais. Em contrapartida, planícies costeiras arenosas de baixa variação topográfica permanecem menos representadas, e carecem de um melhor entendimento no processo de modelagem, pois nesses ambientes, a homogeneidade do material de origem e a menor amplitude dos preditores morfométricos podem reduzir o sinal ambiental disponível para o modelo.

Os solos do Litoral Sul do Rio Grande do Sul apresentam características únicas devido à sua gênese sedimentar e à influência de processos fluviais e marinhos. Por apresentar solos predominantemente arenosos, é uma área de elevada fragilidade ambiental, principalmente pela suscetibilidade à erosão e baixa resiliência. Portanto, o conhecimento de suas características tanto físicas quanto químicas é essencial para o correto uso e manejo desses solos (Streck *et al.*, 2018; Pedron; Dalmolin, 2019). Diante dessa necessidade de integrar as informações acerca dos solos da região, se justifica a utilização de técnicas avançadas de mapeamento e modelagem das propriedades do solo.

O MDS tem se consolidado como uma abordagem eficiente na modelagem de atributos do solo, permitindo previsões precisas sobre a distribuição espacial do pH, granulometria e densidade do solo (Poggio *et al.*, 2021). A integração de dados ambientais em algoritmos de aprendizado de máquina tem potencial para otimizar decisões sobre manejo sustentável e conservação dos solos. Ainda assim, permanecem abertas questões sobre o desempenho dessas previsões em ambientes costeiros de baixa variação topográfica e sobre a consistência das estratégias de seleção de covariáveis ambientais entre diferentes atributos do solo.

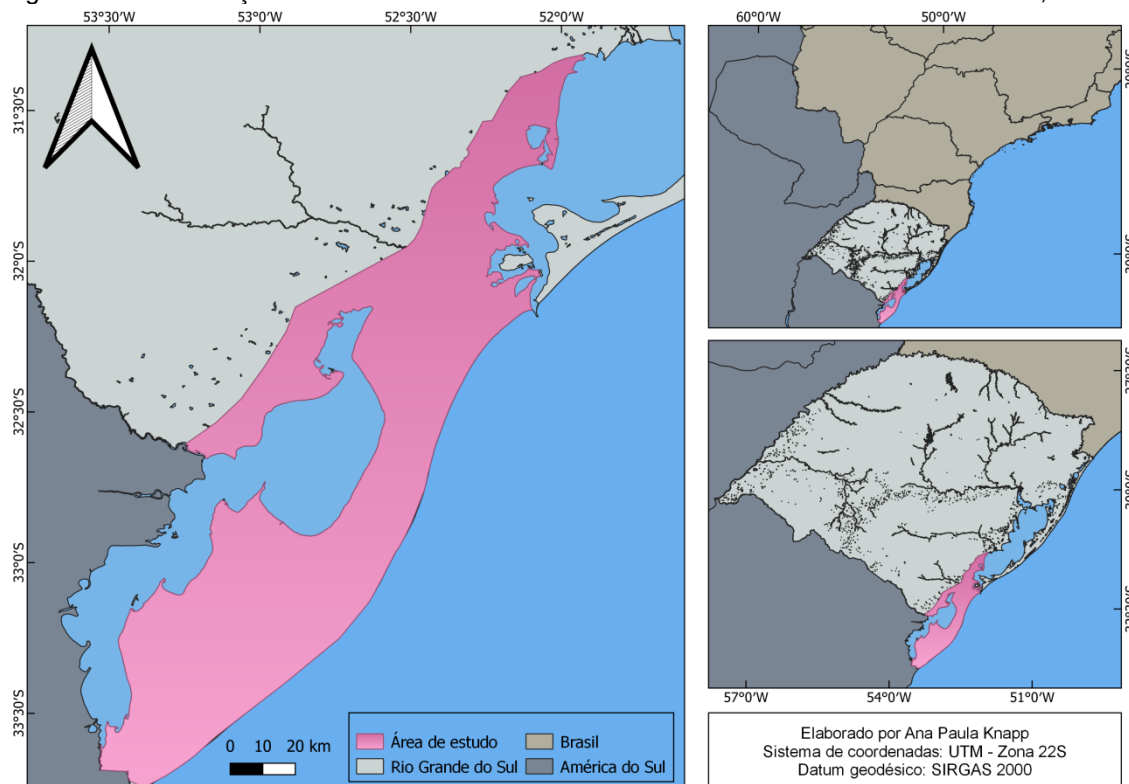
Diante do exposto, esse estudo propõe-se a modelar e mapear o pH, a densidade do solo e o teor de areia dos solos da região do Litoral Sul do Rio Grande do Sul. Acredita-se que seja possível predizer esses atributos do solo com eficiência a partir do algoritmo *Random Forest* (RF), utilizando covariáveis ambientais topográficas e espectrais derivadas de sensoriamento remoto. Com isso, busca-se contribuir para o avanço no conhecimento da distribuição espacial desses atributos, utilizando métodos avançados de predição para fornecer subsídios a práticas agrícolas mais sustentáveis e políticas de conservação ambiental.

4.2 Metodologia

4.2.1 Área de estudo

A área de estudo localiza-se na porção sul do litoral do estado do Rio Grande do Sul (Figura 13), região caracterizada por extensas planícies arenosas. De acordo com Tomazelli *et al.* (2000), o litoral do Rio Grande do Sul possui uma extensão total de 33 mil km², sendo formado por sedimentos arenosos inconsolidados associados a depósitos eólicos, lagunares, praias, paludiais e fluviolagunares (Pedron; Dalmolin, 2019). A área de estudo abrange uma área correspondente a 10.393,9 km², com um perímetro de 1.190,5 km desse ambiente litorâneo.

Figura 13 - Localização da área de estudo no litoral sul do estado do Rio Grande do Sul, Brasil



Fonte: Elaborado pela autora (2026)

A região apresenta solos predominantemente arenosos, incluindo Neossolos Quartzarênicos, Planossolos, Gleissolos e Argissolos. Parte considerável desses solos apresenta a condição de hidromorfismo, ou seja, permanecem saturados com água por longos períodos durante o ano. Esse fenômeno ocorre devido à baixa altitude e ao relevo plano, somados à presença de lagunas distribuídas pela paisagem (Streck *et al.*, 2018).

Os solos da área de estudo são representados pelas Unidades de Mapeamento (UM) Lagoa, Tuia e Pelotas, correspondentes, respectivamente, ao Neossolo Quartzarênico Hidromórfico típico, ao Argissolo Vermelho-Amarelo Distrófico arênico abrupático e ao Planossolo Háptico Eutrófico solódico, conforme Brasil (1973). O Neossolo Quartzarênico Hidromórfico apresenta sequência de horizontes superficiais e subsuperficiais do tipo A e C, com predomínio da fração areia ao longo do perfil, baixa participação de argila e elevada saturação por bases nos horizontes subsuperficiais. O Argissolo Vermelho-Amarelo Distrófico caracteriza-se pela sucessão de horizontes A/E/Bt/C, evidenciando contraste textural decorrente do aumento do teor de argila no horizonte Bt, além de baixa fertilidade natural, expressa pela reduzida

saturação por bases. O Planossolo Háptico Eutrófico apresenta sequência de horizontes A/Bt/C, com incremento do teor de argila no horizonte Bt, refletindo diferenciação textural no perfil, e fertilidade natural relativamente mais elevada em comparação ao Argissolo, evidenciada pelos maiores teores de bases trocáveis e maior saturação por bases. Em termos de distribuição na paisagem, os Neossolos Quartzarênicos Hidromórficos e os Planossolos ocupam, predominantemente, áreas de terras baixas sujeitas à influência do lençol freático e à ocorrência de hidromorfismo, enquanto os Argissolos ocorrem em posições mais elevadas, apresentando melhor drenagem natural. Em função dessas características, os Neossolos e Planossolos apresentam restrições relacionadas ao excesso de água, ao passo que os Argissolos são mais suscetíveis a processos erosivos devido ao contraste textural e ao predomínio de frações arenosas nos horizontes superficiais (Miguel *et al.*, 2025).

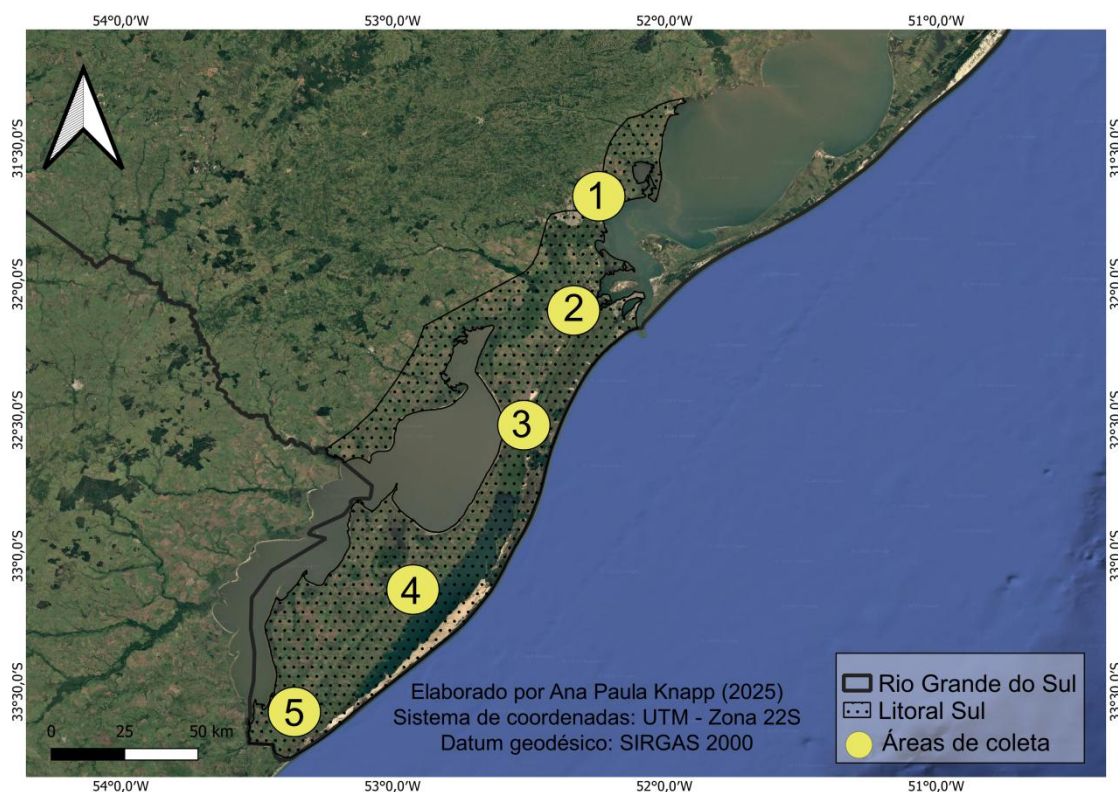
Os depósitos arenosos da área são originários do período Quaternário, caracterizando-se como formações geológicas recentes. Conforme Pedron e Dalmolin (2019), os solos arenosos do bioma Pampa são naturalmente frágeis, apresentando baixa fertilidade e suscetibilidade a processos erosivos. Devido ao elevado teor de areia, esses solos apresentam baixa capacidade de retenção de água, baixa capacidade de troca catiônica (CTC) e baixos teores de matéria orgânica (M.O). Além disso, a reduzida reatividade química da fração arenosa e a baixa estabilidade estrutural conferem a esses solos uma baixa resistência à erosão hídrica e eólica, exigindo práticas específicas de manejo sustentável.

O clima da área de estudo é classificado como subtropical (Cfa) e temperado quente (Cfb) segundo a classificação climática de Köppen, caracterizado por temperatura média anual de 17,1 a 19°C e precipitação de 1200 a 1400 mm/ano (Wrege *et al.*, 2012). A paisagem apresenta relevo predominantemente plano, situando-se a uma altitude inferior a 40 metros acima do nível do mar (Streck *et al.*, 2018). Mais informações sobre a área de estudo podem ser visualizadas na seção de apêndices (Apêndice A, Apêndice B e Apêndice C).

4.2.2 Amostragem de solo

A definição dos pontos de coleta foi apoiada pela aplicação do método de amostragem hipercubo latino condicionado (cLHS) (Minasny; McBratney, 2006), implementado via software R (R Core Team, 2018) por meio do pacote cLHS (Roudier, 2014). Foram utilizados como dados de entrada para a seleção dos pontos de amostragem o Modelo Digital de Elevação (MDE) *Shuttle Radar Topography Mission* (SRTM) (Farr *et al.*, 2007), com 30 metros de resolução espacial, o mapa geológico do estado do Rio Grande do Sul, em escala 1:750.000 (CPRM, 2008), e imagens interpretadas de uso e cobertura da terra, em escala 1:250.000, obtidas através da plataforma MapBiomas (2023). O cLHS é um método de amostragem capaz de detectar de forma eficiente a variabilidade das condições de campo com base em covariáveis ambientais (Minasny; McBratney, 2006). Dessa forma, foram definidas 5 áreas de amostragem que representam de maneira abrangente as condições da região de estudo (Figura 14).

Figura 14 - Delimitação dos locais de amostragem da área de estudo pelo método cLHS

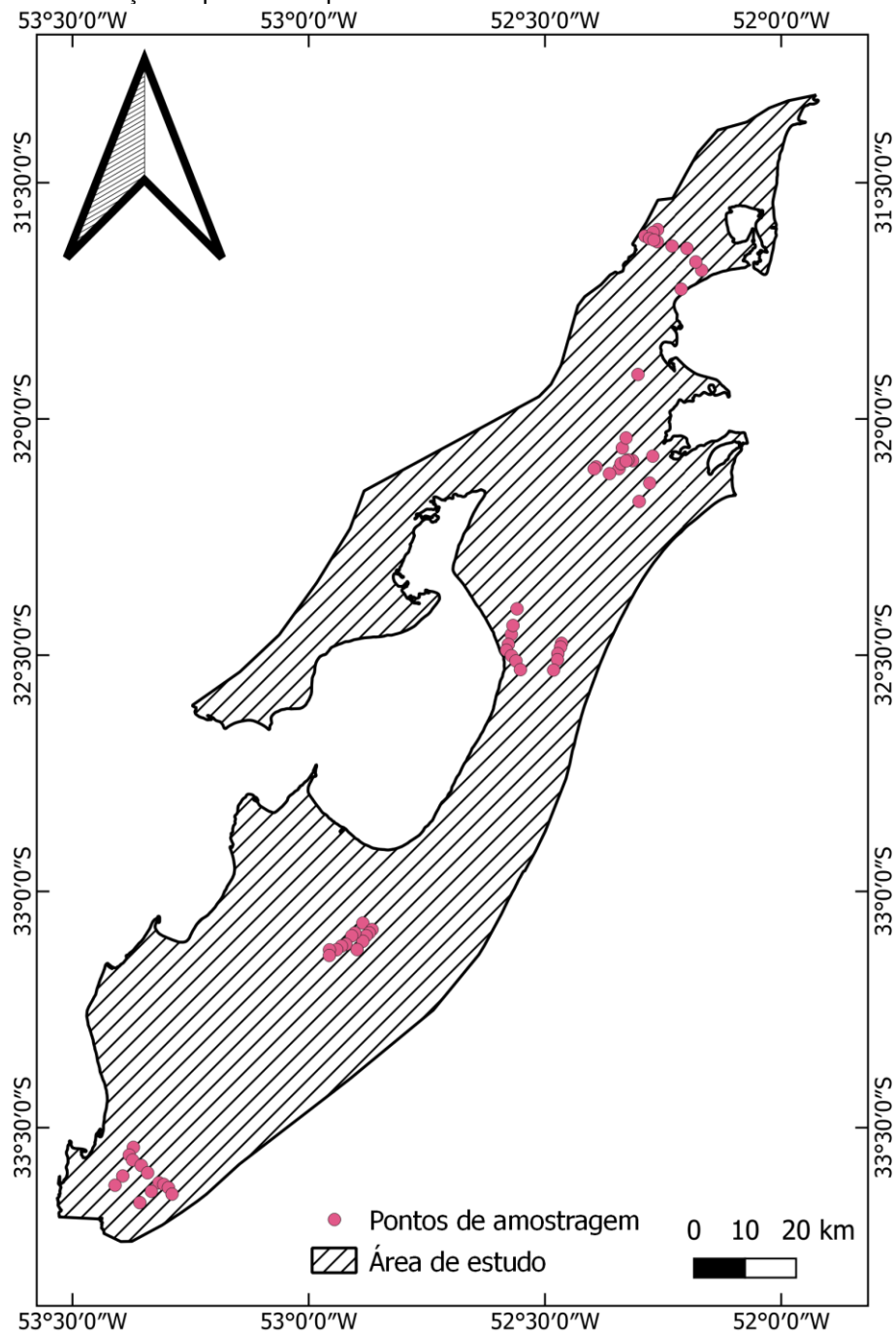


Fonte: Elaborado pela autora (2025).

Em cada área selecionada pelo método cLHS foi delimitado um raio de 15 km, dentro do qual foram definidos os pontos de coleta também por meio do método cLHS. No total foram coletados 64 pontos (Figura 15): 11 pontos na área 1, 13 pontos nas áreas 3, 4 e 5 e 14 pontos na área 2.

A escolha dos locais de coleta levou em consideração a proximidade com estradas, garantindo acesso adequado para a realização da amostragem. Para isso, foi utilizado um arquivo com as estradas primárias, secundárias e terciárias, que foi usado como balizador para a distribuição dos pontos, considerando uma distância de 10 metros a partir do centro das vias. Com isso foi possível otimizar a distribuição das amostras, assegurando uma cobertura eficiente das condições ambientais.

Figura 15 - Distribuição espacial dos pontos de coleta de solo na área de estudo



Fonte: Elaborado pela autora (2026).

Em cada ponto de amostragem, as amostras foram coletadas em duplicata, utilizando anéis metálicos para determinação da densidade do solo. As amostras foram coletadas nas profundidades de 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm e 60-100 cm, conforme definido por Arrouays *et al.* (2014), sendo armazenadas em latas metálicas, devidamente identificadas e transportadas ao laboratório para análise. Também foram coletadas amostras deformadas para a

determinação de pH em água e análise granulométrica em cada uma das profundidades.

Além da coleta das amostras, cada ponto foi descrito para fins de caracterização do local de coleta, incluindo informações como a classificação do solo até o segundo nível categórico do SiBCS (Santos *et al.*, 2025), o uso atual da terra, características de relevo e declividade, a espessura do horizonte A, a profundidade do horizonte B e a espessura do solum. Um resumo destas informações pode ser encontrado na seção de apêndices (Apêndice D, Apêndice E, Apêndice F, Apêndice G e Apêndice H).

4.2.3 Análises laboratoriais

A densidade do solo (BD) foi determinada pelo método do cilindro volumétrico, conforme descrito por Teixeira *et al.* (2017), sendo as amostras secas em estufa a 105°C até atingirem peso constante, permitindo o cálculo da densidade aparente (Equação 8).

$$BD = \frac{m}{V} \quad (8)$$

Em que:

BD: Densidade do solo (g cm⁻³);

m: Massa seca da amostra (g); e

V: Volume do anel (cm³).

A granulometria foi determinada utilizando o método da pipeta (Teixeira *et al.*, 2017) adaptado, que envolve dispersão química e agitação mecânica para separar as frações de argila, silte e areia. Além disso, o pH do solo foi determinado pelo método da acidez ativa (pH em água), seguindo os procedimentos descritos em Tedesco *et al.* (1995), com razão solo/água de 1:1 e medição realizada com potenciômetro calibrado. Todas as análises citadas foram conduzidas em duplicata.

4.2.4 Modelagem preditiva

4.2.4.1 Modelagem com *Machine Learning*

A relação entre os atributos do solo e as covariáveis ambientais foi analisada por meio do algoritmo *Random Forest* (RF), uma técnica de aprendizado de máquina supervisionado. O *Random Forest* consiste em um conjunto de árvores de decisão, em que cada árvore é treinada de forma independente com uma subamostra aleatória do conjunto de dados obtida por *bootstrap* (reamostragem com reposição). A predição final é obtida pela média dos resultados individuais, minimizando a influência de *outliers* e proporcionando maior robustez ao modelo (Wadoux *et al.*, 2019). O algoritmo foi implementado no software R (R Core Team, 2024) por meio do pacote "*randomForest*" (Liaw; Wiener, 2002), baseado no algoritmo original proposto por Breiman (2001).

4.2.4.2 Covariáveis ambientais

Para calibração dos modelos RF para predição das propriedades químicas e físicas do solo, foram geradas covariáveis ambientais que auxiliaram na predição da variabilidade espacial do pH, teor de areia e densidade do solo. No total foram geradas 33 covariáveis topográficas/morfométricas derivadas de modelo digital de elevação (MDE) e 15 covariáveis espectrais derivadas de imagens de sensoriamento remoto, totalizando 48 covariáveis de entrada (Tabela 15). O MDE utilizado foi derivado da missão *Shuttle Radar Topography* (SRTM), com resolução espacial de 30 metros (Farr *et al.*, 2007), sendo utilizado para caracterizar a topografia e as condições de relevo. Os atributos de terreno foram extraídos em ambiente R (R Core Team, 2024) acoplados à interface SAGA GIS (Olaya; Conrad, 2009). As covariáveis espectrais foram obtidas a partir de imagens de satélite Landsat 8-9/OLI, para o mês de janeiro de 2025.

Tabela 15 - Covariáveis ambientais utilizadas na modelagem preditiva dos atributos do solo

Fonte da Informação	Variável	Sigla	Descrição
Modelo digital de elevação	Aspect	asp	Orientação do declive
	Convergence Index	ci	Índice de convergência/divergência do escoamento

Modelo digital de elevação	Cross-Sectional Curvature	csc	Curvatura perpendicular ao declive
	Diurnal Anisotropic Heating	dah	Medição contínua da energia dependente da exposição
	Elevation	mde	Elevação
	Easternness	e	Seno de aspecto
	Flow Line Curvature	flc	Curvatura ao longo da linha de fluxo
	General Curvature	gc	Curvatura combinada do plano e perfil
	Longitudinal Curvature	lc	Curvatura na direção do declive
	LS Factor	ls	Fator topográfico
	Mass Balance Index	mbi	Índice de equilíbrio entre erosão e deposição
	Maximal Curvature	maxc	Curvatura máxima na seção normal local
	Mid-Slope Position	msh	Distância relativa entre topo e vale
	Minimal Curvature	minc	Curvatura mínima na seção normal local
	Multiresolution Index of Ridge Top Flatness	mrrtf	Índice de planicidade no topo de cristas
	Multiresolution Index of Valley Bottom Flatness	mrvbf	Índice de planicidade do fundo do vale
	Northernness	n	Cosseno de aspecto
	Normalized Height	nh	Altura normalizada entre base e cume
	Plan Curvature	planc	Curvatura da linha de contorno hipotética
	Profile Curvature	proc	Curvatura ao longo da inclinação mais íngreme
	Real Surface Area	rsa	Área real da superfície
	Slope	s	Declividade angular local
	Slope Height	sh	Distância vertical entre base e cume da encosta
	Standardized Height	stanh	Altura padronizada
	Surface Specific Points	ssp	Diferenças entre pontos de deslocamento
	Tangential Curvature	tanc	Curvatura perpendicular ao gradiente

Imagens Landsat	Terrain Ruggedness Index	tri	Índice de rugosidade do terreno
	Terrain Surface Convexity	tsc	Índice de convexidade da superfície
	Topographic Position Index	tpi	Diferença entre elevação de um ponto e vizinhança
	Topographic Wetness Index	twi	Índice de umidade baseado no relevo
	Total Curvature	tc	Medida geral da curvatura
	Valley Depth	vd	Profundidade relativa do vale
	Vector Ruggedness Measure	vrn	Mede a variação da rugosidade do terreno
	Bare Soil Index	BSI	Índice de solo exposto
	Blue	b2blue	Banda azul do sensor do satélite
	Clay Minerals Ratio	CMR	Razão de minerais de argila
	Green	b3green	Banda verde do sensor do satélite
	Red	b4red	Banda vermelha do sensor do satélite
	Near Infrared	b5nir	Banda infravermelho próximo
	Normalized Difference Moisture Index	NDMI	Índice de umidade por diferença normalizada
	Normalized Difference Tillage Index	NDTI	Índice de preparo do solo
	Normalized Difference Vegetation Index	NDVI	Índice de vegetação
	Normalized Difference Water Index	NDWI	Índice de água
	Normalized Pigment Chlorophyll Ratio Index	NPCRI	Índice de clorofila
	Soil Adjusted Vegetation Index	SAVI	Índice de vegetação ajustado para solo
	Short-Wave Infrared	b6swir1	Banda infravermelho de ondas curtas
	Short-Wave Infrared	b7swir2	Banda infravermelho de ondas curtas
	Soil Brightness Index	SBI	Índice de brilho do solo

Fonte: Elaborado pela autora (2026), adaptado de Nachtigall (2022).

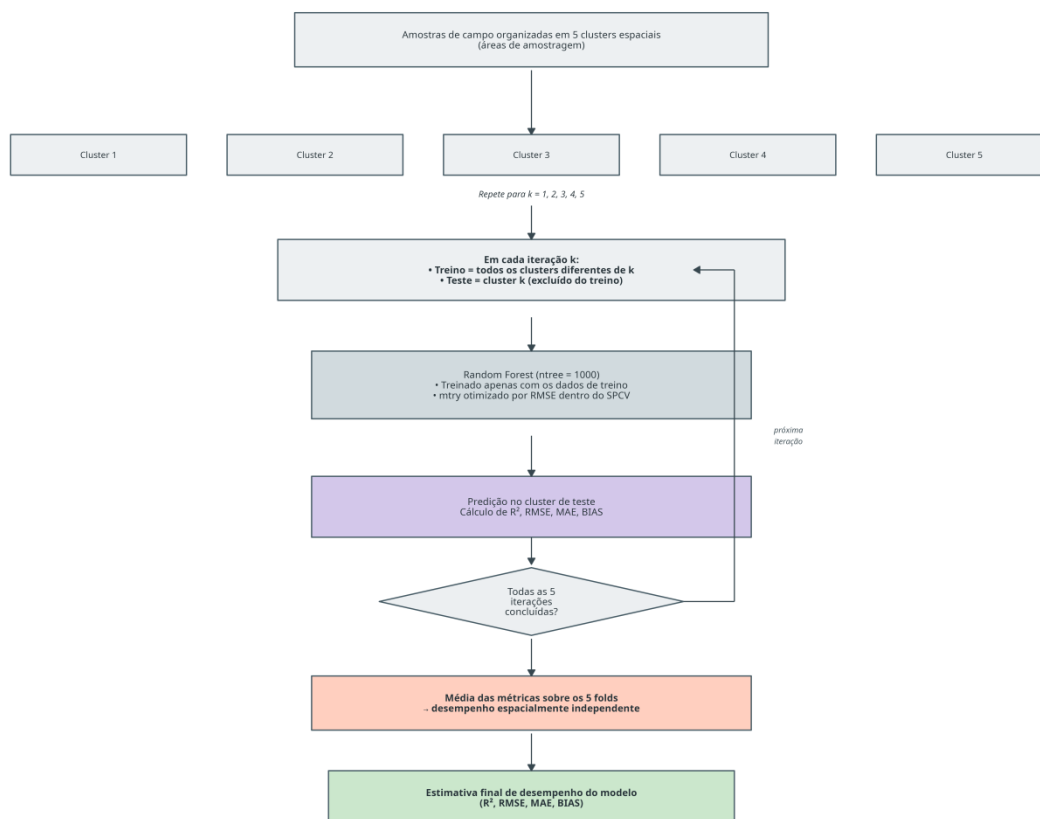
Todas as covariáveis em formato *raster* foram reprojetaadas para um sistema de coordenadas comum e reamostradas para uma resolução espacial uniforme, garantindo consistência espacial entre as camadas. Em seguida, foram extraídos os valores das covariáveis nos pontos amostrais. O processamento dessas informações foi conduzido em ambiente R (R Core Team, 2024), utilizando os pacotes *raster* (Hijmans, 2022), *rgdal* (Bivand *et al.*, 2021) e RSAGA (Brenning *et al.*, 2025).

4.2.4.3 Calibração e validação do modelo

A avaliação dos modelos foi realizada por validação cruzada espacial do tipo *Leave-One-Cluster-Out* (SPCV-LOCO) (Figura 16). Os dados foram previamente agrupados em clusters espaciais correspondentes às áreas de coleta. Em cada iteração da validação cruzada, um cluster foi utilizado como conjunto de teste, enquanto os demais foram utilizados para treinamento do modelo. Esse processo foi repetido até que todos os clusters tivessem sido utilizados como teste. Essa abordagem reduz o efeito da autocorrelação espacial entre amostras de treinamento e teste, fornecendo uma estimativa mais realista da capacidade de generalização do modelo em áreas não amostradas (Roberts *et al.*, 2017).

Para a construção dos modelos RF, o número de árvores (*ntree*) foi fixado em 1.000. O hiperparâmetro *mtry*, que define o número de covariáveis testadas em cada divisão dos nós das árvores, foi otimizado por meio de busca em grade (*grid search*). No modelo inicial (abordagens de seleção de covariáveis), avaliou-se uma grade de valores candidatos para *mtry* compreendendo os valores de 1, 2, 3, 5, 8, 12, 16, 20, 25 e 30 (limitados ao número total de preditores disponíveis). Para o modelo reduzido, contendo apenas as 15 covariáveis mais importantes, a grade de busca foi restrita aos valores menores ou iguais a 15 (1, 2, 3, 5, 8 e 12). Diferentes valores de *mtry* foram avaliados dentro do esquema de validação cruzada espacial, sendo selecionado o valor que minimizou o RMSE médio em cada uma das respectivas etapas.

Figura 16 - Fluxograma da validação cruzada espacial com otimização do hiperparâmetro *mtry* no modelo *Random Forest*



Fonte: Elaborado pela autora (2026).

O desempenho dos modelos foi avaliado por meio das seguintes métricas: coeficiente de determinação (R^2) (Equação 9), raiz do erro quadrático médio (RMSE) (Equação 10), erro absoluto médio (MAE) (Equação 11) e viés (BIAS) (Equação 12). As métricas foram calculadas em cada iteração da validação cruzada espacial e posteriormente agregadas por média.

$$R^2 = \frac{\sum_{i=1}^n (\text{pred} - \text{mobs})^2}{\sum_{i=1}^n (\text{obs} - \text{mobs})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\text{pred} - \text{obs})^2} \quad (10)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\text{obs} - \text{pred}| \quad (11)$$

$$BIAS = \frac{1}{n} \sum_{i=1}^n (\text{pred} - \text{obs}) \quad (12)$$

Em que:

Pred: é o valor predito;

Mobs: é o valor médio observado;

Obs: é o valor observado;

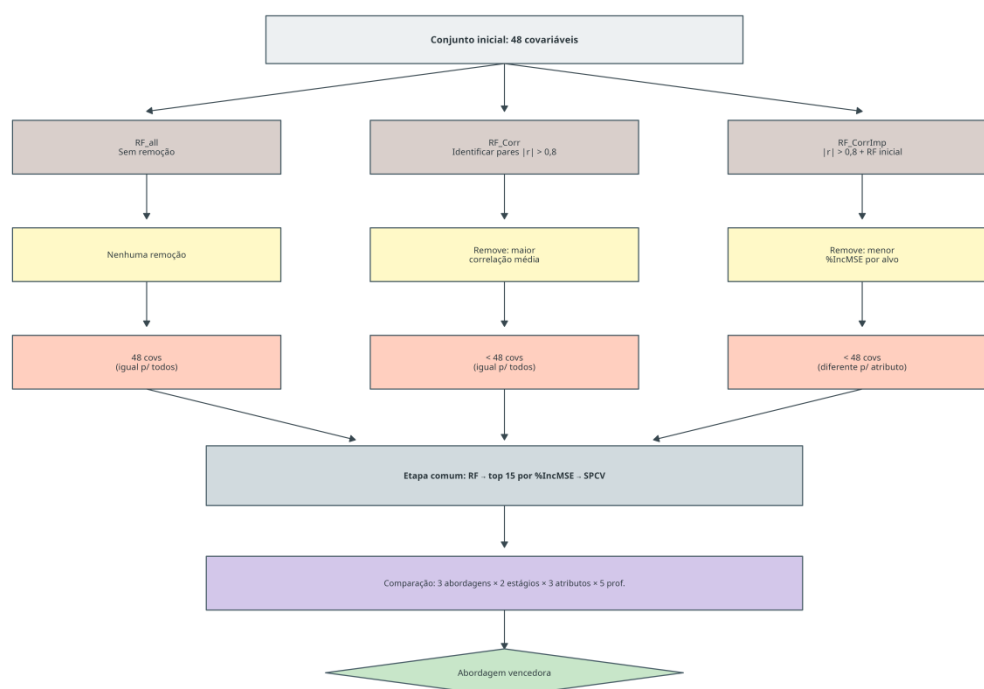
n: é o número de amostras.

4.2.5 Estratégias de seleção de covariáveis

Com o objetivo de avaliar o efeito de diferentes estratégias de seleção de covariáveis sobre o desempenho dos modelos de predição dos atributos do solo, foram testadas três abordagens distintas: utilização do conjunto completo de covariáveis disponíveis (RF_all), seleção baseada na correlação entre covariáveis (RF_Corr) e seleção baseada na combinação entre correlação e importância preditiva estimada pelo *Random Forest* (RF_CorrImp) (Breiman, 2001; Guyon; Elisseeff, 2003; Dormann *et al.*, 2013) (Figura 17). Na abordagem RF_all, todas as covariáveis ambientais foram mantidas no processo de modelagem, permitindo explorar o potencial preditivo máximo do conjunto de dados. Na abordagem RF_Corr, foi calculada a matriz de correlação de Pearson entre as covariáveis, sendo identificados os pares com correlação absoluta superior a 0,80 ($|r| > 0,80$). Para cada par altamente correlacionado, foi removida a covariável que apresentou maior correlação média absoluta com as demais covariáveis, visando reduzir a redundância de informação e a dependência linear entre os preditores. Na abordagem RF_CorrImp, os pares altamente correlacionados foram igualmente identificados, porém a remoção das covariáveis foi baseada em sua importância preditiva, estimada por meio

da métrica de incremento percentual do erro quadrático médio (%IncMSE) obtida em um modelo RF preliminar (ntree=500). Nesse caso, para cada par correlacionado, foi mantida a covariável de maior importância e removida aquela com menor contribuição preditiva.

Figura 17 - Fluxograma da metodologia de seleção de covariáveis e comparação das abordagens de *Random Forest*



Fonte: Elaborado pela autora (2026).

4.2.6 Redução dimensional do conjunto de covariáveis

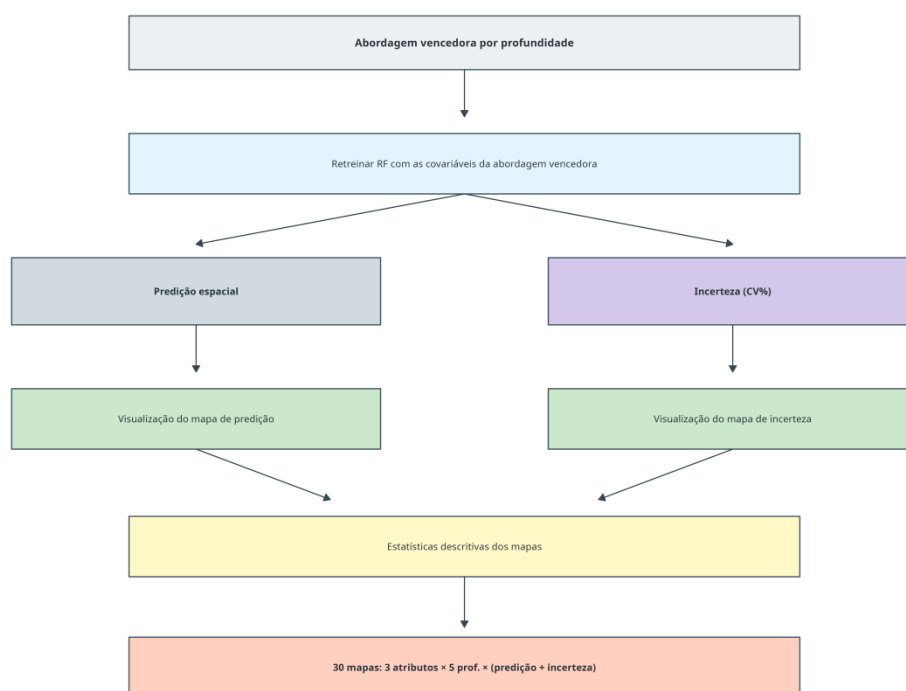
Após a aplicação de cada estratégia de seleção, foi realizada uma etapa adicional de redução dimensional das covariáveis. Para isso, a importância das covariáveis foi quantificada a partir dos modelos ajustados com os conjuntos de preditores resultantes do pré-processamento de cada abordagem, por meio da métrica %IncMSE (Aumento Percentual do Erro Quadrático Médio), calculada pela permutação aleatória dos valores de cada preditor nos dados *out-of-bag*. Em seguida, as covariáveis foram ordenadas de forma decrescente de acordo com sua importância, sendo selecionadas as quinze covariáveis mais relevantes para compor um conjunto reduzido de preditores (RF_Top15). Essa etapa teve como finalidade avaliar se modelos mais simples e com menor número de covariáveis seriam capazes de manter desempenho preditivo semelhante aos modelos completos (Kasraei *et al.*, 2024), além de favorecer a

interpretação dos fatores ambientais associados à variabilidade dos atributos do solo e reduzir a complexidade computacional da modelagem.

4.2.7 Mapas preditivos e de incerteza

Após a etapa de calibração, validação e comparação das diferentes estratégias de seleção e redução dimensional de covariáveis, foi selecionado, para cada atributo do solo e profundidade avaliada, o modelo que apresentou o melhor desempenho preditivo durante a validação cruzada espacial. A seleção dos modelos foi baseada no maior valor de coeficiente de determinação (R^2), sendo as métricas RMSE e MAE utilizadas como informação complementar. Uma vez identificado o modelo de melhor desempenho para cada combinação atributo–profundidade, esse modelo foi recalibrado utilizando a totalidade das observações disponíveis e aplicado às camadas *raster* das covariáveis do conjunto selecionado (Figura 18). As predições foram realizadas pixel a pixel em toda a extensão da área de estudo, resultando em mapas contínuos de densidade do solo, teor de areia e pH para as cinco profundidades padrão do *GlobalSoilMap* (0–5, 5–15, 15–30, 30–60 e 60–100 cm). Esse procedimento permitiu caracterizar a distribuição espacial dos atributos do solo, bem como comparar a variabilidade espacial entre atributos e profundidades na área de estudo.

Figura 18 - Fluxo metodológico para geração dos mapas de predição espacial e incerteza



Fonte: Elaborado pela autora (2026).

Para cada profundidade, foram gerados mapas de incerteza expressos por meio do coeficiente de variação (CV%) entre as predições individuais das árvores do RF (Equação 13).

$$CV = \frac{\sigma}{\mu} \quad (13)$$

Em que:

σ : é o desvio padrão calculado para cada pixel; e

μ : é a média calculada para cada pixel.

Para o cálculo do CV, foram utilizadas 200 árvores (ntree=200) em vez das 1.000 empregadas na predição. Essa redução no número de árvores se justifica, pois as métricas de desempenho convergem para um valor limite à medida que o número de árvores aumenta. Valores elevados de CV indicam maior discordância entre as árvores, refletindo regiões com maior incerteza preditiva.

Um fluxograma detalhado da metodologia empregada para a modelagem preditiva pode ser consultado no Apêndice O.

4.2.8 Análise de importância e consistência das covariáveis

A importância das covariáveis foi quantificada pela métrica %IncMSE, que mede o aumento percentual no erro quadrático médio *out-of-bag* quando os valores de uma covariável são permutados aleatoriamente. Quanto maior o %IncMSE, maior a contribuição da covariável para a capacidade preditiva do modelo. Nos gráficos de importância, os valores de %IncMSE das 15 covariáveis selecionadas foram normalizados de forma relativa, de modo que cada barra representa a proporção da contribuição da covariável dentro do conjunto Top15, expressa em porcentagem (Equação 14).

$$Importância\ Relativa\ (\%)_i = \frac{\%IncMSE_i}{\sum_{j=1}^{15} \%IncMSE_j} * 100 \quad (14)$$

Em que:

%IncMSE_i é o valor da métrica para a covariável *i*, e o denominador representa a soma dos %IncMSE de todas as 15 covariáveis selecionadas.

Vale ressaltar que, por se tratar de uma normalização interna ao conjunto Top15, os valores percentuais não são diretamente comparáveis entre modelos de atributos, profundidades ou abordagens distintas.

Para facilitar a interpretação das covariáveis selecionadas, cada covariável foi classificada em um dos dois grupos funcionais: topográfico (33 covariáveis) ou espectral (15 covariáveis). Essa classificação permitiu avaliar a contribuição relativa de cada grupo para a predição de cada atributo do solo e identificar se diferentes estratégias de seleção de covariáveis alteram o equilíbrio entre os dois grupos.

A análise de frequência contabilizou, para cada covariável, em quantas combinações de profundidade ela integrou o conjunto Top15, gerando matrizes de frequência por atributo e por abordagem. Essa análise permite identificar covariáveis persistentemente importantes e covariáveis que apareceram pontualmente em profundidades específicas.

A análise de sobreposição foi realizada por meio do índice de Jaccard (Equação 15), uma métrica utilizada para quantificar a similaridade entre conjuntos (Chust *et al.*, 2006; Pueyo *et al.*, 2008; Hwang *et al.*, 2018). O índice foi empregado para avaliar a estabilidade das escolhas de covariáveis: 1) entre as três abordagens de seleção para o mesmo atributo e profundidade, verificando se RF_Corr e RF_CorrImp produzem conjuntos semelhantes ao RF_all; 2) entre os atributos do solo para uma mesma profundidade, investigando se pH, areia e densidade do solo compartilham covariáveis preditoras; e 3) entre profundidades adjacentes, avaliando a continuidade vertical dos preditores selecionados.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (15)$$

Em que:

A e B: conjuntos de covariáveis.

4.2.9 Software e ferramentas utilizadas

As etapas de processamento de dados e modelagem foram realizadas em ambiente R, versão 4.3.3 (R Core Team, 2024). Adicionalmente, foi utilizado o software QGIS em sua versão 3.22 (QGIS Development Team, 2022), para o processamento e manipulação de dados geoespaciais, incluindo visualização, organização de camadas raster e vetoriais, bem como suporte às etapas de preparação das covariáveis ambientais.

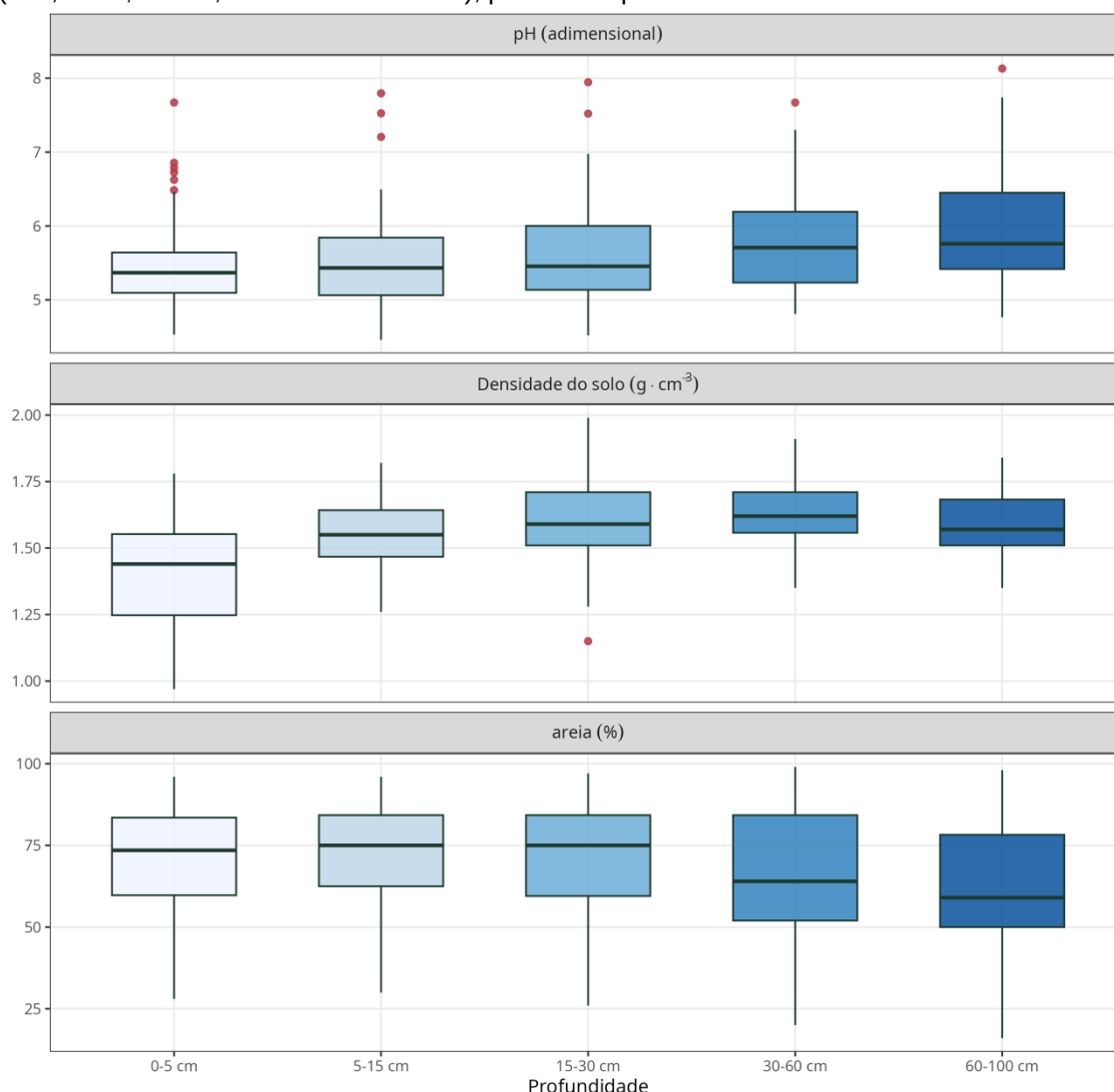
4.3 Resultados e discussão

4.3.1 Caracterização descritiva dos dados de solo

Os atributos analisados (densidade do solo, teor de areia e pH em água) variaram de forma distinta entre si e ao longo do perfil (Figura 19). Ao considerar todas as profundidades (n=320), o teor de areia apresentou o maior coeficiente de variação (CV=26,6%), classificado como moderado segundo a escala proposta por Warrick e Nielsen (1980), enquanto pH (CV=11,8%) e densidade do solo (CV=9,4%) se enquadraram na categoria de baixa variabilidade. O comportamento heterogêneo entre os atributos do solo reflete

os diferentes mecanismos de controle pedogenético de cada um no ambiente estudado, podendo estar associado a processos deposicionais, material de origem, processos biogeoquímicos, compactação do solo, estrutura e teor de matéria orgânica (Castilhos *et al.*, 2002; Lima *et al.*, 2008; Rosa *et al.*, 2011; Ribeiro *et al.*, 2016; Silva; Rehbein, 2018). Os dados detalhados da caracterização descritiva dos atributos nos pontos de coleta encontram-se no Apêndice P.

Figura 19 - Distribuição da densidade do solo (g cm^{-3}), teor de areia (%) e pH por profundidade (0–5, 5–15, 15–30, 30–60 e 60–100 cm), para os 64 pontos amostrais



Fonte: Elaborado pela autora (2026).

A caracterização pedológica dos 64 perfis amostrais (Tabela 16) evidencia o predomínio de Planossolos (50%), seguidos por Argissolos (34,4%), Neossolos (10,9%) e, em menor proporção, Espodossolos, Chernossolos e Gleissolos (1,6% cada). A espessura média do horizonte A foi

de 46 cm com amplitude de 15 a 120 cm; enquanto o limite superior do horizonte B ocorreu em média a 56 cm, sendo ausente nos 7 perfis classificados como Neossolos que, por definição, não possuem horizonte B diagnóstico (Santos *et al.*, 2025). Dentre as classes de solos mais representativas da área de estudo, os Argissolos apresentaram horizonte A mais espesso (média = 53,6 cm) e limite superior do B mais profundo (média = 67,0 cm), refletindo o desenvolvimento pedogenético mais avançado dessa ordem. Os Planossolos, por sua vez, exibiram horizonte A menos espesso (média = 39,3 cm) e transição A/B mais rasa (média = 48,8 cm), coerente com a mudança textural abrupta que caracteriza a classe (Santos *et al.*, 2025). Os Neossolos apresentaram espessura média do A de 54,1 cm.

Tabela 16 - Distribuição das classes de solo e parâmetros dos horizontes pedogenéticos (n = 64 perfis)

Classe de solo	n	%	Espessura A — média (cm)	Limite superior B — média (cm)
Planossolo	32	50,0	39,3	48,8
Argissolo	22	34,4	53,6	67
Neossolo	7	10,9	54,1	Não possui
Espodossolo	1	1,6	90	90
Chernossolo	1	1,6	15	35
Gleissolo	1	1,6	40	40
Geral	64	100	46,3	56,2

Fonte: Elaborado pela autora (2026).

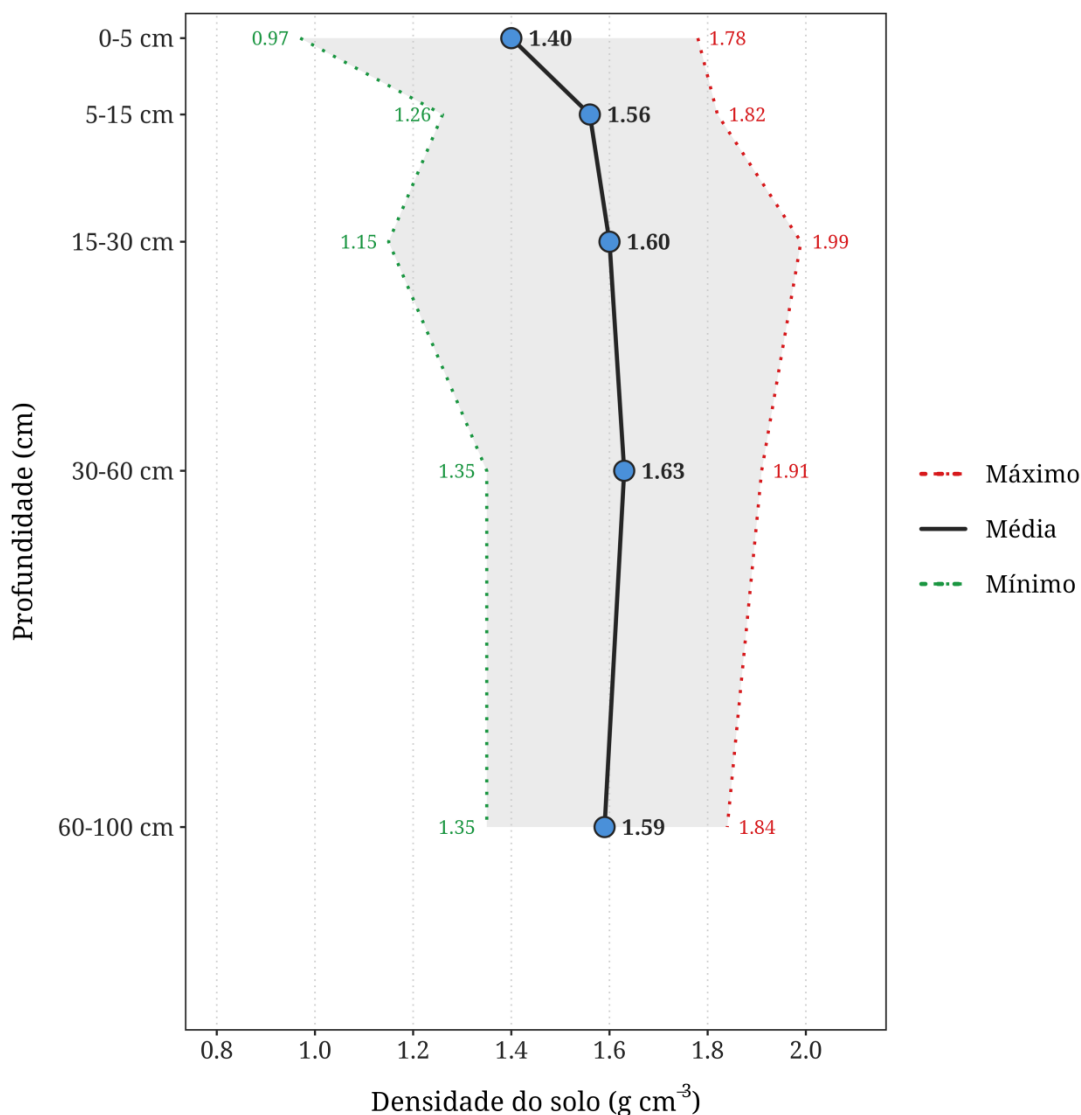
A diferença entre a espessura do horizonte A e o limite superior do horizonte B indica a presença de um horizonte E (horizonte de eluviação) entre eles, observável principalmente em Planossolos e Argissolos. Nos Planossolos, a diferença média entre o limite do horizonte A (39,3 cm) e o horizonte do B (48,8 cm) corresponde a um horizonte E com espessura em torno de 9,5 cm. O horizonte E é caracteristicamente empobrecido em argila, com predomínio de areia, resultante da migração de partículas finas para o horizonte B. O horizonte B que recebe a argila eluviada é denominado B textural (Bt), característica diagnóstica de ambas as classes. Essa sequência pedogenética A/E/Bt tem implicações diretas na interpretação dos padrões verticais dos atributos do solo (Santos *et al.*, 2025).

Esses parâmetros indicam que as profundidades padronizadas de 0–5, 5–15 e 15–30 cm situam-se majoritariamente dentro do horizonte A, a camada

de 30–60 cm intercepta a transição dos horizontes E/Bt (particularmente nos Planossolos, cujo limite superior médio do B está nessa faixa) e a camada de 60–100 cm encontra-se predominantemente no horizonte Bt, exceto nos Neossolos (ausência de hz B diagnóstico) e em parte dos Planossolos com transição mais rasa.

A densidade do solo (BD) aumentou de $1,40 \text{ g cm}^{-3}$ em 0–5 cm até $1,63 \text{ g cm}^{-3}$ em 30–60 cm, com leve redução em 60–100 cm ($1,59 \text{ g cm}^{-3}$) (Figura 20). Esse gradiente vertical pode ser explicado pelo menor adensamento da camada superficial, decorrente do maior acúmulo de matéria orgânica e da atividade biológica.

Figura 20 - Distribuição vertical da densidade do solo ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos



Fonte: Elaborado pela autora (2026).

Em contraste, as camadas subsuperficiais apresentam maior densidade em função do adensamento natural do solo, da compactação gerada pelo uso do solo e dos processos pedogenéticos de eluviação-iluviação, sendo compatível com a sequência pedogenética A/E/Bt. O horizonte A, menos adensado e mais rico em matéria orgânica, em alguns casos, é sucedido pelo horizonte E, pobre em argila em decorrência da eluviação, e posteriormente pelo horizonte Bt, onde a argila eluviada acumula e promove adensamento, favorecendo assim o aumento da densidade do solo. O fato de o maior valor de BD ocorrer em 30–60 cm, faixa que coincide com a zona de transição E/Bt (limite superior médio do B = 56,2 cm), reforça essa interpretação. Os valores médios observados para BD estão de acordo com aqueles relatados por Lima *et al.* (2008) para um Planossolo Háplico sob diferentes condições de uso em terras baixas no Rio Grande do Sul.

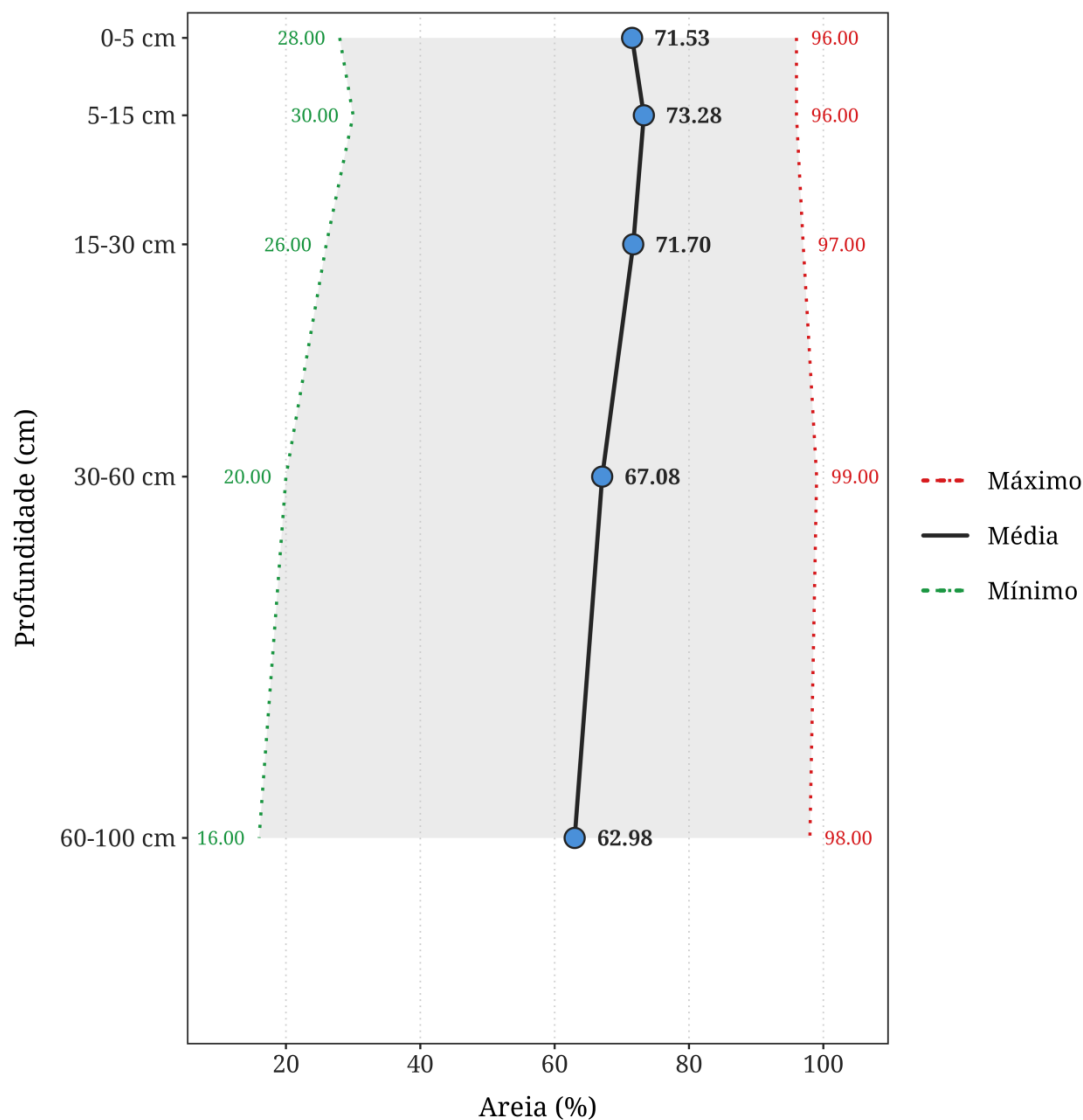
A redução da BD em 60–100 cm, embora pequena, pode estar relacionada à ocorrência de materiais de origem mais arenosos em profundidade, característica comum dos depósitos sedimentares da planície costeira do sul do Rio Grande do Sul (Tomazelli *et al.*, 2000). Além disso, essa redução pode refletir a heterogeneidade dos horizontes subsuperficiais entre as classes de solos presentes na área de estudo; uma vez que Planossolos e Argissolos apresentam horizonte B mais adensado, enquanto os Neossolos (10,9% da amostra), por não possuírem horizonte B diagnóstico, tendem a apresentar menores valores de densidade nas camadas mais profundas.

O coeficiente de variação (CV) mais elevado em 0–5 cm (13,9%) reflete maior heterogeneidade da camada superficial em resposta à cobertura vegetal, ao uso e manejo e às condições hidrológicas locais. Em profundidade, observa-se redução da variabilidade (CV = 7,0–10,1%), sugerindo maior homogeneidade em decorrência da menor influência dos fatores superficiais e da predominância dos processos pedogenéticos (Apêndice P).

O teor de areia apresentou tendência de redução com a profundidade, passando de 71,5% na camada de 0–5 cm para 63,0% em 60–100 cm, enquanto o coeficiente de variação (CV) aumentou progressivamente de 23,8% para 33,3% (Figura 21, Apêndice P). Esse comportamento é compatível com a sequência pedogenética A/E/Bt. O horizonte A, predominantemente arenoso, em alguns casos, é sucedido pelo horizonte E, pobre em argila em decorrência

da eluviação, e posteriormente pelo horizonte Bt, onde ocorre o acúmulo de argila iluviada, reduzindo o teor de areia. Essa transição é particularmente abrupta nos Planossolos e Argissolos, que representam cerca de 85% dos perfis e apresentam mudança textural marcante entre os horizontes E e Bt.

Figura 21 - Distribuição vertical do teor de areia ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos



Fonte: Elaborado pela autora (2026).

O aumento da variabilidade nas camadas de 30–60 cm e 60–100 cm decorre principalmente da heterogeneidade da profundidade em que ocorre a transição entre os horizontes E e B, cujo limite superior do horizonte B varia de aproximadamente 20 a 155 cm entre os perfis. Além disso, a presença de Neossolos (10,9% da amostra), que não apresentam horizonte B diagnóstico e mantêm textura arenosa ao longo de todo o perfil, contribui para ampliar a

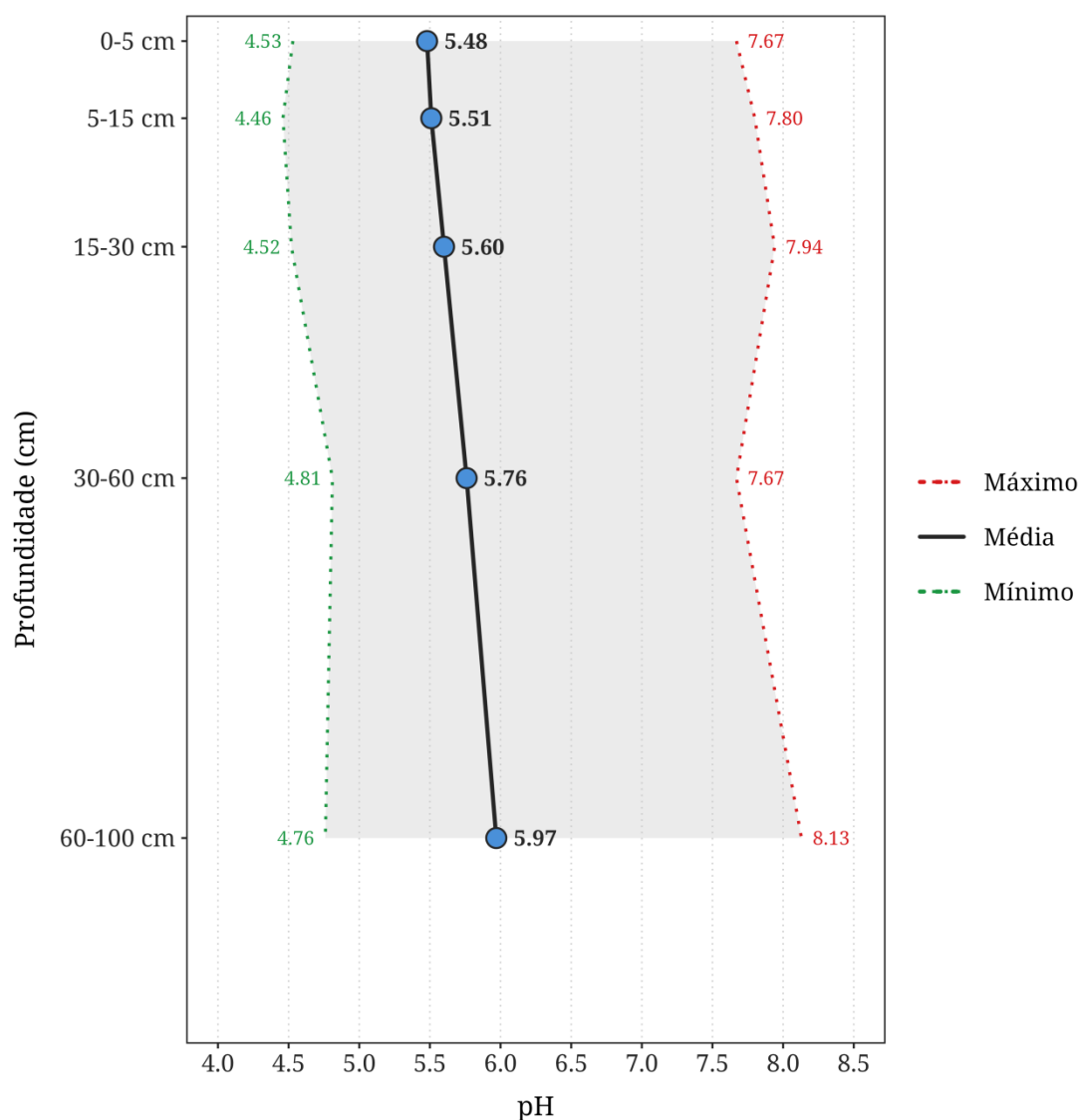
dispersão dos dados. Dessa forma, uma mesma camada de amostragem pode incluir perfis ainda posicionados no horizonte E, de textura arenosa, e outros já inseridos no horizonte Bt, mais argiloso e, ainda, alguns no horizonte A, resultando na elevada variabilidade observada.

O padrão de redução do teor de areia e aumento da dispersão relativa dos dados em profundidade também evidencia a heterogeneidade estratigráfica das camadas subsuperficiais, onde a influência dos processos superficiais é menor e a composição do solo reflete mais diretamente a variabilidade dos materiais de origem já que, no litoral do Rio Grande do Sul, as oscilações do nível relativo do mar durante o período Quaternário originaram unidades sedimentares marinhas, lagunares e fluviomarinhas com diferentes composições granulométricas, produzindo perfis com estratificação complexa (Tomazelli *et al.*, 2000; Castilhos *et al.*, 2002).

Além disso, a ocorrência simultânea de valores extremamente elevados de areia (até 99%) e de teores reduzidos (mínimos entre 16 e 20%) indica que a área de estudo engloba desde solos muito arenosos, como Neossolos Quartzarênicos, até solos desenvolvidos sobre planícies de inundação, impondo desafios à modelagem digital do solo, uma vez que um mesmo conjunto de covariáveis ambientais deve representar ambientes pedológicos e sedimentares bastante distintos.

O pH apresentou tendência de aumento com a profundidade, passando de 5,48 na camada de 0–5 cm para 5,97 em 60–100 cm (Figura 22). Esse padrão é consistente com a maior acidez das camadas superficiais, decorrente da decomposição da matéria orgânica, da liberação de íons H^+ e da lixiviação de bases; processos particularmente intensos em solos arenosos devido à sua elevada condutividade hidráulica e baixa capacidade de retenção de cátions (Kämpf; Curi, 2015). O horizonte E, caracterizado pela eluviação de argila e bases, tende a manter essa condição mais ácida, enquanto os valores mais elevados de pH em profundidade são compatíveis com o horizonte Bt de Planossolos e Argissolos, onde o acúmulo de argila favorece maior retenção de cátions básicos. Em posições mais baixas da paisagem, a influência da saturação hídrica/encharcamento pode contribuir para a manutenção de valores de pH relativamente mais elevados (Camargo *et al.*, 1999).

Figura 22 - Distribuição vertical do pH ao longo do perfil nas cinco profundidades avaliadas, com valores médios, mínimos e máximos



Fonte: Elaborado pela autora (2026).

O coeficiente de variação permaneceu relativamente estável ao longo do perfil (10,9–13,1%) (Apêndice P), indicando que a variabilidade espacial do pH é semelhante entre as diferentes profundidades. Esse comportamento sugere que, embora os processos pedogenéticos promovam um gradiente vertical de acidez, a coexistência de diferentes condições de drenagem e classes de solo mantém uma elevada heterogeneidade geoquímica na área de estudo, o que é evidenciado pela ampla variação entre os valores máximos e mínimos de pH.

Portanto, a variabilidade moderada da densidade do solo e do pH e a maior dispersão da areia nos dados observados para a área de estudo são reflexo da coexistência de classes de solo com diferentes graus de

desenvolvimento pedogenético, desde Neossolos pouco evoluídos, sem horizonte B diagnóstico, até Planossolos e Argissolos com sequência A/E/Bt bem definida, o que contribui para a variabilidade observada em subsuperfície, particularmente nas camadas que interceptam a zona de transição entre horizontes distintos.

4.3.2 Desempenho preditivo e efeito da seleção de covariáveis

4.3.2.1 Desempenho geral

Os valores de R^2 foram baixos em todas as combinações atributo x profundidade x abordagem avaliadas, com máximo de 0,225 para densidade do solo (15–30 cm, RF_CorrImp Top15), 0,175 para o teor de areia (0–5 cm, RF_Corr Top15) e 0,229 para pH (5–15 cm, RF_CorrImp Top15) (Tabela 17). O erro médio absoluto (MAE) dos modelos vencedores variou de 0,086 a 0,152 g cm^{-3} para a densidade, de 14,0 a 19,0% para areia e de 0,423 a 0,707 para pH. Esse resultado, embora indique capacidade preditiva limitada, é esperado para a validação cruzada espacial empregada (SPCV-LOCO), na qual os modelos são avaliados em clusters espacialmente independentes. Essa estratégia tende a produzir estimativas de desempenho mais conservadoras do que a validação cruzada aleatória, por impor um cenário mais rigoroso de extrapolação espacial (Roberts *et al.*, 2017; Meyer *et al.*, 2018; Wadoux *et al.*, 2021).

Contudo, embora o *Random Forest* seja amplamente reconhecido por sua robustez em problemas de regressão, seu desempenho foi inferior ao esperado neste estudo. Uma das possíveis explicações é o reduzido tamanho amostral, uma vez que Liu *et al.* (2026) argumentam que a natureza das predições por partição independentes pode limitar o desempenho do algoritmo em cenários com poucas observações. Além disso, conforme destacado por Mosleh *et al.* (2016), o desempenho de modelos preditivos em áreas de baixo relevo não deve ser generalizado, pois sua eficácia depende de fatores como as características dos solos, o delineamento amostral e a qualidade e representatividade das covariáveis ambientais utilizadas.

Tabela 17 - Abordagem selecionada e métricas de desempenho dos modelos de *Random Forest* para cada atributo e profundidade

Atributo	Profundidade	Abordagem	Estágio	R ²	RMSE*	MAE*	BIAS*
BD	0–5 cm	RF_all	Top15	0,149	0,186	0,152	-0,011
BD	5–15 cm	RF_all	Top15	0,135	0,119	0,097	-0,003
BD	15–30 cm	RF_CorrImp	Top15	0,225	0,146	0,117	-0,001
BD	30–60 cm	RF_all	Top15	0,139	0,107	0,086	0,000
BD	60–100 cm	RF_Corr	Todas	0,200	0,132	0,110	0,000
Areia	0–5 cm	RF_Corr	Top15	0,175	16,40	14,00	0,62
Areia	5–15 cm	RF_all	Todas	0,159	17,90	15,20	0,41
Areia	15–30 cm	RF_all	Top15	0,173	17,60	14,50	0,19
Areia	30–60 cm	RF_all	Top15	0,136	19,90	17,40	0,01
Areia	60–100 cm	RF_CorrImp	Todas	0,090	21,30	19,00	1,75
pH	0–5 cm	RF_CorrImp	Top15	0,216	0,550	0,423	0,014
pH	5–15 cm	RF_CorrImp	Top15	0,229	0,592	0,478	0,036
pH	15–30 cm	RF_CorrImp	Top15	0,133	0,688	0,563	0,006
pH	30–60 cm	RF_Corr	Top15	0,096	0,688	0,576	0,007
pH	60–100 cm	RF_all	Todas	0,117	0,851	0,707	0,041

* unidades de medida das métricas: BD (g cm⁻³), areia (%) e pH (adimensional). Fonte: Elaborado pela autora (2026).

Os valores de R² para a densidade do solo (BD) variaram de 0,135 (5–15 cm) a 0,225 (15–30 cm), com o melhor desempenho observado na camada de 15–30 cm (RF_CorrImp Top15; R² = 0,225). Em geral, espera-se maior capacidade preditiva nas camadas superficiais, onde a influência da cobertura vegetal, do uso da terra e dos processos superficiais é mais intensa. Entretanto, o melhor desempenho observado na camada de 15–30 cm pode indicar que essa profundidade corresponde a uma zona de transição pedogenética na qual as covariáveis morfométricas apresentam maior poder de discriminação das variações da densidade do solo.

Resultados superiores foram reportados por Hateffard *et al.* (2023), que obtiveram valores de R² entre 0,79 e 0,88 para treinamento e entre 0,82 e 0,88 para validação utilizando *Random Forest* em planícies aluviais na Hungria, em áreas de 47 e 160 ha. Em escala nacional, Akpa *et al.* (2016) encontraram R² variando de 0,25 na camada de 0–5 cm a 0,48 na camada de 60–100 cm, evidenciando aumento da capacidade preditiva com a profundidade do solo. Esse padrão é comparável com os resultados obtidos neste estudo, onde a camada de 60–100 cm apresentou maior R² em comparação às camadas de 0–

5 cm e 5–15 cm (Tabela 17), embora com magnitudes inferiores às relatadas na literatura.

Por outro lado, o desempenho das predições pode ter sido fortemente influenciado pelo tipo de covariável utilizado na predição. Carvalho Junior *et al.* (2016), por exemplo, alcançaram desempenho preditivo superior (R^2 entre 0,42 e 0,48) para a predição da densidade do solo utilizando informações intrínsecas ao solo como preditores, com a justificativa de que esse tipo de informação é de fácil acesso em estudos de levantamento de solos. Nesse contexto, os autores observaram elevada importância do carbono orgânico, saturação de bases, CTC e hidrogênio como preditores da densidade do solo.

O teor de areia apresentou valores de R^2 variando de 0,090 (60–100 cm) a 0,175 (0–5 cm), evidenciando um declínio progressivo da capacidade preditiva com o aumento da profundidade. Esse comportamento também foi observado por Adhikari *et al.* (2013) para as frações de areia fina e areia grossa e pode ser explicado pela redução da influência das covariáveis relacionadas aos processos superficiais nas camadas mais profundas.

Em geral, os valores de R^2 obtidos neste estudo foram inferiores aos reportados na literatura. Em escala de bacia hidrográfica, Forkuor *et al.* (2017) obtiveram R^2 de 0,34 para o teor de areia (RMSE = 7,6), desempenho intermediário em relação às demais frações granulométricas, uma vez que a argila apresentou R^2 de 0,18 e o silte de 0,54. De forma semelhante, Mosleh *et al.* (2016) alcançaram R^2 de 0,54 utilizando um modelo linear generalizado, mesmo em um conjunto de dados com elevada variabilidade (CV = 77%).

Na predição dos teores de areia em escala nacional na Nigéria, nas profundidades padronizadas do *GlobalSoilMap*, Akpa *et al.* (2014) observaram melhor desempenho preditivo em superfície devido a maior relação dos preditores com os processos de superfície. Valores de R^2 variaram de 0,49 nas camadas superiores até 0,33 na camada de 60-100 cm, porém ao utilizar a profundidade como preditor no *Random Forest*, os valores de R^2 aumentaram para 0,92 em superfície e 0,76 em profundidades maiores. Os autores atribuíram o bom desempenho preditivo nesse caso ao efeito da utilização da transformação aditiva de razão logarítmica (ALR), apropriada para o tratamento de dados composicionais. Resultados igualmente superiores foram obtidos por Fathizad *et al.* (2025), que apesar do conjunto de dados apresentar

características semelhantes às observadas neste estudo, com teor médio de areia de 73% e coeficiente de variação de aproximadamente 20% (Apêndice P), os autores obtiveram R^2 de 0,76 na predição do teor de areia utilizando *Random Forest*, sendo a areia a fração granulométrica mais bem predita, superando a argila ($R^2 = 0,28$) e o silte ($R^2 = 0,56$).

Por outro lado, alguns estudos reportaram desempenhos mais próximos aos observados neste trabalho. Em condições andinas no Equador, Ließ *et al.* (2012) obtiveram R^2 de 0,23 para a areia na camada superficial e de 0,13 em subsuperfície, valores semelhantes aos encontrados neste estudo. Em condições brasileiras, Carvalho Junior *et al.* (2014) relataram R^2 variando de 0,15 a 0,27 para a predição do teor de areia, enquanto na Índia, Jena *et al.* (2023) obtiveram valores de R^2 variando de 0,30 a 0,21 até 30 cm de profundidade, depois disso houve queda pronunciada de precisão e valores de $R^2=0,02$ foram reportados nas camadas entre 30 e 100 cm. Em contraste, Dharumarajan *et al.* (2020), utilizando *Quantile Regression Forest* (QRF), explicaram entre 42% e 49% da variabilidade do teor de areia, evidenciando desempenho superior.

O pH apresentou um padrão marcante, com valores de R^2 variando de 0,096 (30–60 cm) a 0,229 (5–15 cm). Os maiores valores foram observados nas camadas superficiais (0–5 cm: 0,216 e 5–15 cm: 0,229), seguidos por uma redução acentuada em profundidade (15–30 cm: 0,133; 30–60 cm: 0,096; 60–100 cm: 0,117). O melhor desempenho nas camadas superficiais é coerente com o fato de que o pH é fortemente influenciado pela cobertura vegetal, deposição de serrapilheira, teor de matéria orgânica e condições de umidade, processos que apresentam assinaturas espectrais detectáveis através das imagens Landsat (Mulder *et al.*, 2011; Dewitte *et al.*, 2012). Em profundidade, o pH passa a ser controlado predominantemente por processos geoquímicos relacionados ao material de origem e às condições de drenagem, fatores que não são plenamente representados pelas covariáveis utilizadas, justificando a redução da capacidade preditiva. Embora o R^2 na camada de 60–100 cm (0,117) tenha sido ligeiramente superior ao observado em 30–60 cm (0,096), esse resultado foi obtido utilizando todas as covariáveis, sem filtragem prévia, sugerindo que, nessa profundidade, a etapa de seleção não proporcionou ganhos de desempenho.

De modo geral, assim como para a densidade do solo e para o teor de areia, os valores de R^2 obtidos para a predição do pH foram inferiores aos reportados na literatura. Vaysse *et al.* (2017) obtiveram R^2 de 0,76 para a predição do pH na camada de 5–15 cm. Ao passo que, Vasundhara *et al.* (2025) alcançaram R^2 de 0,61 na camada de 0–15 cm utilizando *Random Forest*, que teve desempenho superior aos observados para *Support Vector Machine* (SVM) e *Cubist*. Cabe destacar que os dados empregados por esses autores apresentavam variabilidade espacial (CV=14%) semelhante à observada neste estudo (Apêndice P). Ma *et al.* (2021) reportaram R^2 entre 0,45 e 0,54 utilizando *Cubist*, *Quantile Regression Forest* (QRF) e redes neurais, enquanto Bouslihim *et al.* (2024), obtiveram R^2 de 0,26 para predição do pH e atribuíram o desempenho relativamente baixo ao fato de que dados de pH não apresentavam variação suficiente para que o modelo fosse capaz de aprender e fazer as predições com maior precisão.

Por outro lado, alguns estudos reportaram desempenhos semelhantes aos observados neste trabalho, como Carvalho Junior *et al.* (2014) que, em condições subtropicais brasileiras, obtiveram valores de R^2 entre 0,03 e 0,11 para a predição do pH e Dharumarajan *et al.* (2020) que alcançaram desempenho preditivo capaz de explicar entre 8% e 23% da variabilidade do pH.

Os valores de R^2 obtidos neste estudo (máximo de 0,229) são inferiores aos reportados em estudos de MDS conduzidos em escalas mais amplas e em ambientes com maior heterogeneidade ambiental. Por exemplo, Hengl *et al.* (2017), no SoilGrids250m, relataram valores de R^2 entre 0,70 e 0,80 para atributos texturais em escala global e Hengl *et al.* (2014), no SoilGrids1km, obtiveram explicação entre 20 e 35% da variabilidade das frações granulométricas por meio de modelos de regressão. Em escala nacional no México, Abbruzzini *et al.* (2026) obtiveram valores de R^2 de 0,59 para o teor de areia, 0,40 para a densidade do solo e 0,73 para o pH. No entanto, estudos como estes se beneficiam de gradientes climáticos, litológicos e topográficos mais amplos do que os de uma planície costeira homogênea.

De forma geral, a redução do desempenho preditivo com o aumento da profundidade é condizente com o fato de que as covariáveis ambientais utilizadas (derivadas de imagens orbitais e do modelo digital de elevação)

representam principalmente processos atuantes na superfície do terreno. Consequentemente, seu poder explicativo diminui nas camadas mais profundas, cujas propriedades passam a ser controladas em maior intensidade pelo material de origem e por processos pedogenéticos menos relacionados às condições superficiais. Esse comportamento está de acordo com o que foi apresentado por Chen *et al.* (2022) em seu artigo de revisão, segundo o qual as propriedades do solo mapeadas por meio de MDS tendem a apresentar redução da capacidade preditiva com o aumento da profundidade. Os autores, entretanto, destacam o pH como uma exceção a essa tendência.

O BIAS foi próximo de zero em todas as profundidades e abordagens para os modelos vencedores (Tabela 17), indicando ausência de viés sistemático nas predições. Esse resultado é importante, pois demonstra que, mesmo diante de valores relativamente baixos de R^2 , os modelos não tendem a superestimar ou subestimar sistematicamente os atributos, preservando a representatividade estatística das médias dos mapas produzidos. Para o teor de areia, entretanto, observou-se aumento do BIAS com a profundidade, passando de aproximadamente 0,6% na camada de 0–5 cm para cerca de 1,8% em 60–100 cm. Embora esse viés positivo seja de pequena magnitude, ele indica uma tendência crescente de superestimação do teor de areia nas camadas subsuperficiais.

A abordagem RF_all foi dominante para BD (3/5 profundidades: 0–5, 5–15 e 30–60 cm) e teor de areia (3/5 profundidades: 5–15, 15–30 e 30–60 cm), enquanto a abordagem RF_CorrImp foi vencedora para pH em 3/5 profundidades (0–5, 5–15 e 15–30 cm). Esse padrão diferenciado sugere que o pH nas camadas superficiais e intermediárias beneficiou-se da remoção de covariáveis redundantes ponderada pela importância, possivelmente porque o sinal preditivo para esse atributo está concentrado em um subconjunto mais restrito de covariáveis, tornando vantajosa a eliminação de preditores redundantes e potencialmente ruidosos. Em contrapartida, para BD e para a areia, o conjunto completo de covariáveis (RF_all) proporcionou melhor desempenho, sugerindo que a informação preditiva desses atributos está distribuída entre um conjunto mais amplo de preditores. Nesses casos, a remoção prévia de covariáveis correlacionadas pode ter eliminado informações complementares relevantes para a predição, uma vez que preditores altamente

correlacionados nem sempre são redundantes do ponto de vista preditivo (Dormann *et al.*, 2013). Essa interpretação está de acordo com Boulesteix *et al.* (2012), que destacam que, embora a correlação entre preditores geralmente tenha impacto limitado sobre a capacidade preditiva dos modelos de *Random Forest*, a seleção de um subconjunto representativo de covariáveis correlacionadas pode ser útil em algumas situações específicas. Os autores ressaltam, contudo, que os benefícios dessa estratégia dependem de fatores como o tamanho amostral e que a redução do conjunto de dados para um conjunto de covariáveis estritamente independentes não é, em geral, desejável. Assim, a remoção prévia de covariáveis correlacionadas pode ter eliminado informações complementares relevantes para a predição, o que pode explicar o melhor desempenho da RF_all para BD e teor de areia.

A lista completa das covariáveis excluídas e mantidas em cada abordagem, por atributo e profundidade, é apresentada no Apêndice Q.

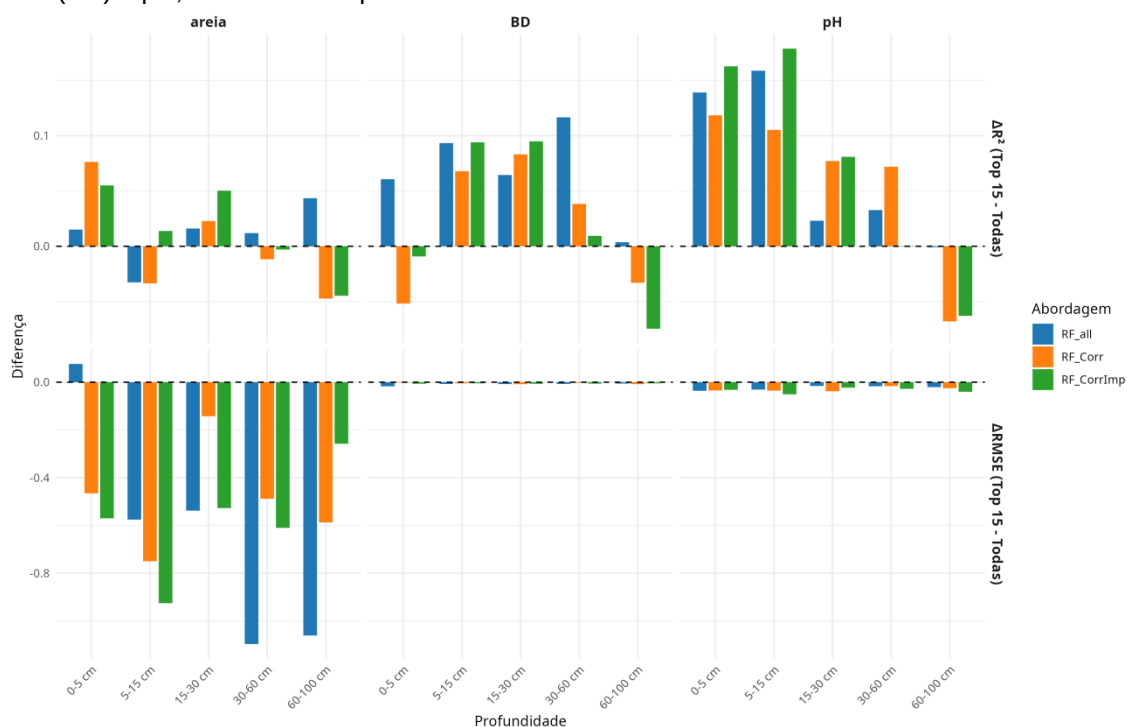
4.3.2.2 Efeito da redução dimensional (Top15)

O estágio de seleção das 15 covariáveis mais importantes (%IncMSE) foi o vencedor em 11 das 15 combinações atributo x profundidade, superando o uso do conjunto completo de covariáveis (Tabela 17). Os maiores ganhos foram observados para o pH com RF_CorrImp nas camadas superficiais, em que o R^2 aumentou de 0,053 para 0,216 em 0–5 cm (+0,163) e de 0,050 para 0,229 em 5–15 cm (+0,179). Para a densidade do solo (BD), o maior incremento ocorreu na camada de 30–60 cm com RF_all, em que o R^2 passou de 0,022 para 0,139 (+0,117). Em contraste, os casos em que a utilização de todas as covariáveis superou a abordagem Top15 foram associados a diferenças reduzidas de desempenho (Apêndice R). Além disso, a redução dimensional foi acompanhada, na maior parte das condições, por diminuição do RMSE, comportamento também relatado por Heinze *et al.* (2018), indicando que os ganhos observados não se restringiram apenas ao aumento do coeficiente de determinação, mas também resultaram em menor erro de predição (Apêndice S).

A Figura 23 evidencia que a redução dimensional para as 15 covariáveis mais importantes promoveu aumento do desempenho preditivo na maioria das

combinações atributo-profundidade, embora com magnitude distinta entre atributos e profundidades. Os maiores ganhos foram observados para o pH nas camadas superficiais, sugerindo que seu sinal preditivo está concentrado em um conjunto reduzido de covariáveis altamente informativas.

Figura 23 - Diferença entre os modelos ajustados com as 15 covariáveis mais importantes (Top15) e aqueles ajustados com todas as covariáveis para os atributos areia, densidade do solo (BD) e pH, em diferentes profundidades



* Valores positivos de ΔR^2 indicam aumento do coeficiente de determinação após a redução dimensional, enquanto valores negativos de $\Delta RMSE$ indicam redução do erro de predição, representando melhora no desempenho do modelo. Fonte: Elaborado pela autora (2026).

Bouslihim *et al.* (2024) aplicaram o algoritmo Boruta para seleção de covariáveis, com o objetivo de reduzir a redundância entre preditores, onde a partir de um conjunto inicial de 29 covariáveis ambientais, 11 foram selecionadas como relevantes para a predição do pH, incluindo tanto preditores morfométricos quanto espectrais, o que evidencia a contribuição conjunta desses diferentes tipos de preditores para a explicação do atributo. Já Zhao *et al.* (2023), ao aplicarem diferentes técnicas de seleção de covariáveis ambientais para a predição do pH na China, observaram melhorias no desempenho dos modelos em relação ao uso de todas as covariáveis disponíveis. No conjunto de treino, os valores de R^2 variaram entre 0,57 e 0,61 com seleção de preditores, enquanto o modelo sem seleção apresentou R^2 de 0,56. Já no conjunto de teste, o desempenho aumentou de 0,45 (sem seleção)

para 0,65, 0,71 e 0,69 nas diferentes abordagens de seleção, evidenciando maior impacto positivo da redução dimensional na capacidade preditiva dos modelos.

Para a densidade do solo (BD), observaram-se ganhos em praticamente todas as profundidades, com destaque para a camada de 30–60 cm, correspondente à transição E/Bt. Para o teor de areia, as diferenças entre os modelos com Top15 e todas as covariáveis foram, em geral, de pequena magnitude, variando entre profundidades e abordagens, sem um padrão consistente de ganho associado à redução dimensional. Nas combinações em que a seleção das Top15 apresentou desempenho inferior ao conjunto completo de covariáveis (ΔR^2 negativo), as reduções de R^2 variaram entre atributos e profundidades, concentrando-se principalmente nos modelos de areia.

Esse padrão de melhoria do desempenho após a redução dimensional sugere que a utilização de um conjunto mais restrito de covariáveis pode ter reduzido a influência de preditores pouco informativos, favorecendo a capacidade preditiva do modelo através da construção de árvores mais informativas e podendo favorecer a interpretabilidade dos modelos (Svetnik *et al.*, 2004). No contexto do mapeamento de índices de qualidade do solo (IQS), que são compostos por atributos químicos e físicos do solo, Baghche-Maryam *et al.* (2025) avaliaram quatro métodos de seleção de preditores (eliminação recursiva de características, Boruta, PCA e importância de variáveis acoplada ao *Random Forest*). A redução dos conjuntos iniciais de covariáveis resultou em desempenhos preditivos para o modelo *Random Forest* variando de 0,35 a 0,77, com os maiores valores de R^2 associados aos conjuntos selecionados por importância de variáveis e PCA. Por outro lado, Carvalho Junior *et al.* (2016), na predição de BD usando como covariáveis dados de propriedades do solo comumente obtidas em levantamentos de solos, não obtiveram melhoria no desempenho preditivo ao reduzir o número de preditores usados. Os valores de R^2 obtidos pelos autores variou entre 0,42 e 0,47 utilizando grupos contendo entre 6 e 32 preditores.

O efeito positivo da redução dimensional do conjunto de preditores foi particularmente evidente para o pH, no qual o conjunto completo de covariáveis, mesmo após a pré-filtragem realizada pelo RF_CorrImp, ainda

continha preditores de baixa relevância que competiam com as covariáveis realmente informativas durante a construção das árvores. Esse efeito tende a ser intensificado quando o valor ótimo de *mtry* é baixo (predominantemente entre 1 e 3 neste estudo) pois, com poucos candidatos avaliados em cada divisão, a chance de que preditores fracos sejam selecionados aumenta, reduzindo a consistência do ensemble (Breiman, 2001). A remoção prévia dessas covariáveis concentra a seleção nos preditores de maior importância, produzindo árvores mais consistentes e previsões mais estáveis. Esse padrão está de acordo com estudos que demonstram que a eliminação de preditores pouco informativos pode reduzir a variância do modelo sem aumento expressivo do viés, especialmente em situações em que a seleção de preditores é crítica para o desempenho do *Random Forest* (Genuer *et al.*, 2010; Probst *et al.*, 2019).

4.3.2.3 Ajuste do hiperparâmetro *mtry*

O hiperparâmetro *mtry* define o número de covariáveis selecionadas aleatoriamente como candidatas em cada divisão das árvores do *Random Forest*, controlando o equilíbrio entre a diversidade das árvores do modelo e a quantidade de informação utilizada por cada árvore. Valores baixos de *mtry* aumentam a diversidade do *ensemble*, reduzindo a correlação entre as árvores e, conseqüentemente, a variância das previsões. Por outro lado, valores elevados aproximam o modelo do *bagging* tradicional ($mtry = p$), aumentando a força individual das árvores, mas também elevando a correlação entre elas, o que pode aumentar a variância do *ensemble* (Breiman, 2001).

Para problemas de regressão, o valor padrão é $mtry = p/3$ (Liaw; Wiener, 2002; Probst *et al.*, 2019), em que p representa o número de covariáveis disponíveis. Assim, seriam esperados valores de $mtry = 5$ para o estágio Top15 ($p = 15$) e $mtry = 16$ para o conjunto completo de covariáveis ($p = 48$). O número de árvores (*ntree*) foi fixado em 1.000 para todos os modelos, valor suficiente para garantir a estabilização do erro *out-of-bag* (Breiman, 2001; Probst; Boulesteix, 2018).

Neste estudo, o *mtry* foi otimizado por validação cruzada SPCV-LOCO para cada combinação atributo x profundidade x abordagem x estágio. Os

modelos vencedores apresentaram valores ótimos baixos, variando entre 1 e 3, inferiores aos valores recomendados pela regra prática (Tabela 18).

Tabela 18 - Valores de mtry ótimo dos modelos vencedores por atributo e profundidade

Atributo	0–5 cm	5–15 cm	15–30 cm	30–60 cm	60–100 cm
Densidade do solo	3	2	1	3	1
Teor de areia	2	2	1	1	1
pH	1	3	1	1	1

Fonte: Elaborado pela autora (2026).

Esse resultado é coerente com as características do conjunto de dados. Primeiro, o sinal preditivo encontra-se distribuído entre diversas covariáveis de importância moderada, sem predominância de um único preditor. Nessa condição, valores reduzidos de mtry forçam cada árvore a explorar diferentes subconjuntos de covariáveis, aumentando a diversidade do *ensemble* (Breiman, 2001; Probst *et al.*, 2019). Em segundo lugar, o conjunto de dados apresenta 64 observações para 15 e 48 covariáveis, de modo que o número de covariáveis é comparável ao tamanho amostral (mas não caracteriza um cenário de alta dimensionalidade, onde $p \gg n$). Nessa situação, árvores individuais tornam-se mais instáveis e a utilização de mtry baixo aumenta sua independência, permitindo que a agregação das árvores reduza a variância global do modelo de forma mais eficiente. Esse comportamento está alinhado à literatura que destaca a necessidade de ajustar mtry aos dados e à relação entre preditores e observações, com valores abaixo do padrão frequentemente sendo ótimos em conjuntos de pequeno tamanho amostral (Genuer *et al.*, 2010; Probst *et al.*, 2019).

Em contraste, Capitaine *et al.* (2021) destacam que valores mais elevados de mtry tendem a ser vantajosos em cenários de alta dimensionalidade, nos quais o número de preditores é muito superior ao número de observações, aumentando a probabilidade de que covariáveis informativas sejam consideradas em cada divisão da árvore. Como essa condição não se aplica ao presente estudo, os valores ótimos reduzidos observados para mtry são consistentes com a estrutura dos dados e com a necessidade de aumentar a diversidade das árvores.

A otimização do hiperparâmetro mtry em valores entre 1 e 3, está de acordo com relatos de Bouslihim *et al.* (2024) que reportaram valor ótimo de

mtry igual a 2 para predição do pH e de Zhao *et al.* (2023) que reportaram valores entre 1 e 5 em diferentes configurações de modelagem de pH. De forma semelhante, Baghche-Maryam *et al.* (2025), ao modelarem índices de qualidade do solo (IQS) compostos por propriedades químicas, físicas e biológicas, obtiveram valor otimizado de mtry igual a 2.

Os resultados também evidenciam uma interação favorável entre a otimização de mtry e a seleção das 15 covariáveis mais importantes. A combinação da redução do conjunto de covariáveis com os baixos valores de mtry permite aumentar a frequência com que as covariáveis mais informativas foram avaliadas nas divisões das árvores. Por exemplo, quando $mtry = 1$, essa probabilidade aumenta de $1/48$ ($\approx 2\%$) para $1/15$ ($\approx 7\%$), intensificando a utilização das covariáveis mais informativas durante a construção do modelo. Esse mecanismo ajuda a explicar por que a seleção das Top15 produziu os maiores ganhos de desempenho, especialmente para o pH nas camadas superficiais, onde o sinal preditivo mostrou-se concentrado em um conjunto reduzido de covariáveis relevantes. Em conjunto, esses resultados indicam que a combinação entre a seleção de covariáveis e valores reduzidos de mtry concentrou a construção das árvores nas covariáveis mais informativas, reduzindo a influência de preditores redundantes ou pouco relevantes e produzindo árvores mais diversas e predições mais estáveis.

4.3.2.4 Importância das covariáveis ambientais

A análise da frequência de inclusão das covariáveis no conjunto Top15, considerando todas as combinações de atributo, profundidade e abordagem de modelagem (15 modelos por atributo), revelou predominância de covariáveis derivadas do modelo digital de elevação entre os preditores mais recorrentes (Tabela 19). Das quinze covariáveis com maior frequência de seleção, apenas três pertencem ao grupo espectral (b2blue, b5nir e ndmi), enquanto as demais doze são derivadas do relevo, o que contrasta com a expectativa de maior contribuição de covariáveis espectrais em ambientes de baixa variação topográfica. A elevada frequência de inclusão desse tipo de covariável corrobora sua relevância na modelagem das propriedades do solo, uma vez que as covariáveis morfométricas representam quantitativamente a

configuração do relevo e permitem inferir a influência da topografia sobre a distribuição espacial e variabilidade de propriedades do solo (Florinsky, 2017).

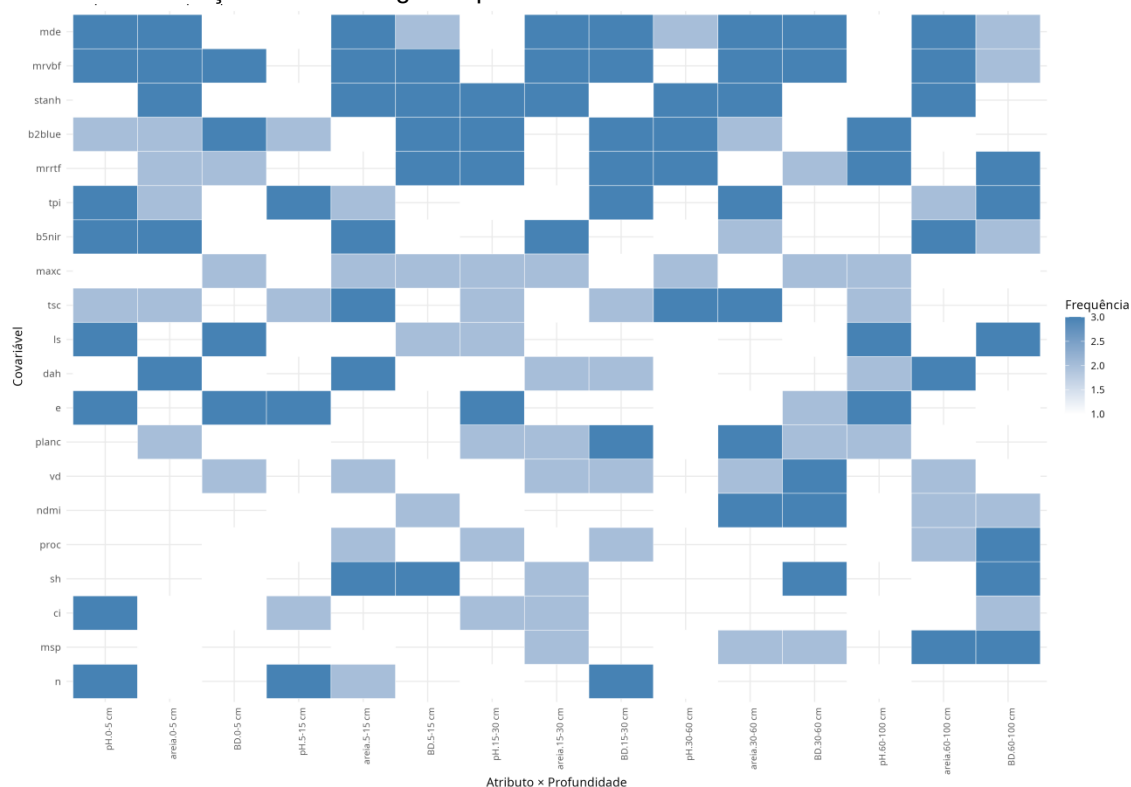
Tabela 19 - Frequência de inclusão no Top15 das quinze covariáveis mais recorrentes

Covariável	Grupo	Densidade	Areia	pH	Total
mde	Topográfica	11	15	8	34
mrvmf	Topográfica	14	15	4	33
stanh	Topográfica	6	15	9	30
b2blue	Espectral	10	6	13	29
mrrtf	Topográfica	13	3	9	25
tpi	Topográfica	7	10	7	24
b5nir	Espectral	4	14	5	23
maxc	Topográfica	8	7	8	23
tsc	Topográfica	2	9	11	22
ls	Topográfica	9	2	10	21
dah	Topográfica	4	11	5	20
e	Topográfica	6	1	13	20
planc	Topográfica	6	8	6	20
vd	Topográfica	9	8	1	18
ndmi	Espectral	7	6	4	17

Fonte: Elaborado pela autora (2026).

A Figura 24 complementa a Tabela 19 ao evidenciar diferenças na importância relativa das covariáveis entre os atributos avaliados. O *heatmap* mostra que covariáveis topográfica/morfométricas, especialmente mde, mrvmf e stanh, apresentam elevada frequência de seleção para o teor de areia, indicando maior contribuição na explicação da variabilidade granulométrica, corroborando Adhikari *et al.* (2013). Em contraste, essas mesmas covariáveis apresentam menor frequência de seleção para o pH, sugerindo que os controles espaciais associados ao relevo possuem menor capacidade explicativa para a acidez ativa do solo.

Figura 24 - *Heatmap* da frequência de inclusão de cada covariável no conjunto Top15, para todas as combinações de abordagem x profundidade



Fonte: Elaborado pela autora (2026).

O comportamento inverso é observado para a covariável espectral *b2blue*, que apresenta maior frequência de seleção nos modelos de pH em comparação aos modelos de teor de areia. Isso sugere que covariáveis espectrais podem capturar indiretamente condições superficiais associadas à vegetação e à matéria orgânica, processos que influenciam o pH superficial, mas que apresentam menor relação com a distribuição granulométrica. Contudo, resultados apresentados por Akpa *et al.* (2014) evidenciam contribuição relevante de bandas espectrais (bandas 1, 2, 3, 4 e 7) na predição de frações granulométricas, em conjunto com covariáveis de relevo como elevação, declividade e curvatura, com destaque para as bandas do azul (2), verde (3) e o NDVI entre os preditores espectrais. Já Vasundhara *et al.* (2025) identificaram maior importância de preditores bioclimáticos, especialmente relacionados à temperatura; para a predição do pH; seguidos por bandas no infravermelho de ondas curtas e covariáveis de terreno como *vd*, *mrvmf* e *mde*, com menor contribuição relativa.

Para a densidade do solo, o padrão de seleção de covariáveis apresentou distribuição mais homogênea entre diferentes grupos de preditores,

sem predominância evidente de um único fator explicativo. Covariáveis como *mrrtf* e *mrvmf* apresentaram participação relevante, indicando a influência de características morfométricas; entretanto, a dispersão das frequências sugere que a densidade do solo resulta da interação entre múltiplos fatores ambientais, corroborando os achados de Akpa *et al.* (2016), nos quais a densidade do solo foi explicada por preditores de diferentes naturezas, incluindo variáveis climáticas (temperatura e precipitação), tipo de solo, bandas espectrais (principalmente as bandas 2 e 3) e elevação.

O padrão observado para as covariáveis topográficas mais importantes na predição de BD neste estudo também converge com os resultados de Hateffard *et al.* (2023). Esses autores, ao compararem diferentes algoritmos de aprendizado de máquina para a predição da densidade do solo superficial (0–10 cm), identificaram como principais covariáveis *planc*, *vd*, *mrvmf* e *mde*, os quais também foram destacados neste estudo (Tabela 19), reforçando a importância de covariáveis derivadas da morfometria do terreno e da topografia na explicação da variabilidade espacial da densidade do solo. De forma complementar, Ließ *et al.* (2012) também observaram elevada importância da curvatura do plano (*planc*) para predição da areia.

As covariáveis relacionadas à posição na paisagem apresentaram as maiores frequências de seleção. Os índices *mrvmf* (*Multi-Resolution Valley Bottom Flatness*), *mrrtf* (*Multi-Resolution Ridge Top Flatness*) e *stanh* (*Standardized Terrain Height*) descrevem diferentes aspectos da posição topográfica relativa. O *mrvmf* identifica superfícies planas associadas a fundos de vale, enquanto o *mrrtf* caracteriza superfícies planas localizadas em posições mais elevadas. A elevada frequência de seleção do *mrvmf* pode estar relacionada ao fato de esse índice ter sido desenvolvido para identificar diferenças sutis de posição na paisagem em áreas de baixo relevo, nas quais a elevação absoluta e os divisores de drenagem apresentam baixa capacidade discriminatória (Gallant; Dowling, 2003). Já o *stanh* representa a posição relativa de um ponto dentro do gradiente altimétrico local, sendo particularmente útil em áreas de relevo suavizado, onde pequenas diferenças altimétricas podem refletir mudanças importantes na dinâmica da paisagem (Guo *et al.*, 2019). A elevada frequência de seleção dessas covariáveis, especialmente para areia e BD, indica que a posição topográfica exerce forte

influência sobre processos deposicionais, drenagem, acúmulo de umidade e matéria orgânica, fatores que condicionam tanto a distribuição granulométrica quanto a organização estrutural do solo.

Entre as covariáveis espectrais, destacaram-se b2blue e b5nir. A banda azul apresentou elevada frequência de seleção para o pH e participação relevante para BD, possivelmente por refletir variações associadas à matéria orgânica superficial, à cobertura vegetal e à umidade do solo, características que influenciam indiretamente processos pedogenéticos e a dinâmica da acidez. Por sua vez, a banda do infravermelho próximo (b5nir) foi uma das covariáveis mais importantes para a areia, indicando que diferenças granulométricas influenciam a resposta espectral da superfície, resultado compatível com estudos que demonstram o potencial de covariáveis espectrais para a predição de atributos granulométricos no mapeamento digital de solos (Rossel; Webster, 2012; Forkuor *et al.*, 2017; Demattê *et al.*, 2020; Safanelli *et al.*, 2021; Pelegrino *et al.*, 2025).

Outra variável de destaque foi tsc (*Terrain Surface Convexity*), cuja maior frequência de seleção ocorreu para o pH. Como esse índice está relacionado aos padrões de convergência e dispersão do fluxo superficial, sua importância sugere que processos hidrológicos influenciam a distribuição espacial da acidez do solo, possivelmente por meio de diferenças na lixiviação, na redistribuição de bases e no acúmulo de matéria orgânica ao longo da paisagem.

De forma geral, os padrões observados reforçam que os três atributos apresentam controles espaciais distintos: a areia apresenta maior associação com covariáveis relacionadas à geomorfologia e à posição na paisagem; o pH responde a uma combinação de informações espectrais e topográficas, especialmente relacionadas às condições superficiais; e a densidade do solo apresenta comportamento mais complexo, com contribuição distribuída entre diferentes grupos de covariáveis. Contudo, vale ressaltar que as medidas de importância de covariáveis em modelos de aprendizado de máquina devem ser interpretadas como indicadores da contribuição relativa dos preditores para o processo preditivo do modelo, e não como evidências diretas de relações causais entre os preditores e a variável resposta (Grömping, 2015). Além disso, essas medidas fornecem um ranqueamento relativo entre as covariáveis, sem

garantir que todas as mais bem classificadas apresentem contribuição preditiva efetiva, uma vez que o algoritmo sempre estabelece uma ordenação dos preditores (Boulesteix *et al.*, 2012). Esse aspecto é observado na seção de apêndices (Apêndice T, Apêndice U e Apêndice V), nos quais algumas covariáveis incluídas entre as 15 de maior importância apresentam valores negativos de importância, indicando ausência de contribuição preditiva consistente ou mesmo um efeito potencialmente prejudicial ao desempenho do modelo.

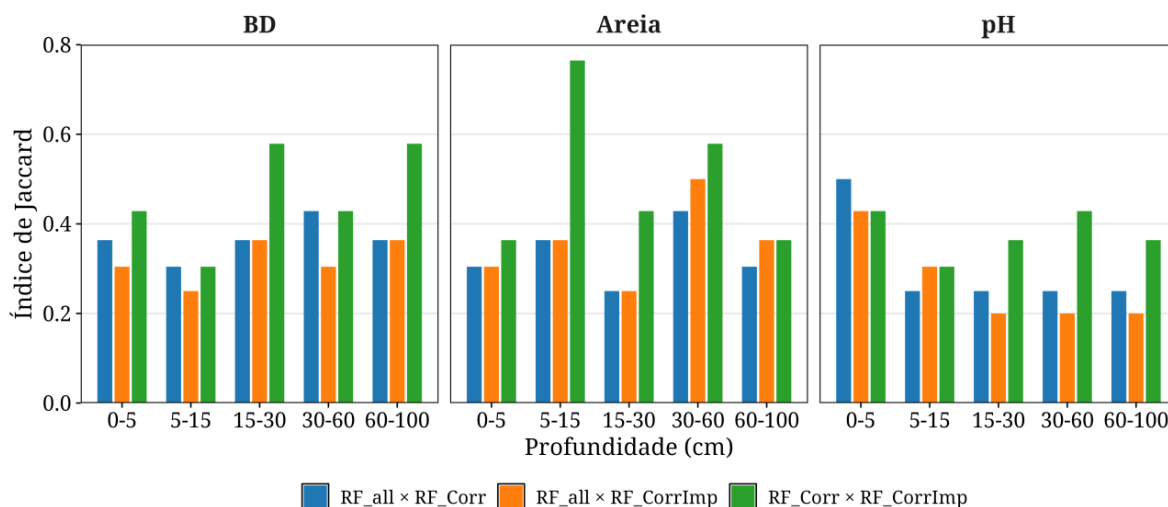
4.3.3 Análise de sobreposição de covariáveis

4.3.3.1 Entre estratégias de seleção

A sobreposição entre os conjuntos das 15 covariáveis mais importantes, medida pelo índice de Jaccard, variou entre atributos e profundidades (Figura 25). O par RF_Corr x RF_CorrImp apresentou, de forma consistente, os maiores valores de sobreposição para os três atributos, enquanto os pares envolvendo RF_all exibiram valores menores. Para areia em 5–15 cm, a sobreposição entre RF_Corr e RF_CorrImp atingiu 0,76, o maior valor observado, ao passo que, nas demais profundidades, variou entre 0,36 e 0,58. Para BD e pH, os valores de Jaccard entre RF_Corr e RF_CorrImp situaram-se predominantemente entre 0,30 e 0,58, enquanto os pares RF_all x RF_Corr e RF_all x RF_CorrImp apresentaram valores entre 0,20 e 0,50.

Esse padrão é consistente com a forma como as abordagens são construídas, pois tanto RF_Corr quanto RF_CorrImp partem de conjuntos de covariáveis previamente reduzidos pela remoção de covariáveis altamente correlacionadas, resultando em um *pool* parcialmente comum de preditores disponíveis para seleção. Como as covariáveis topográficas tendem a apresentar menor redundância entre si do que as espectrais, ambas as abordagens convergem para um conjunto semelhante de preditores topográficos de elevada importância, aumentando a sobreposição entre seus conjuntos Top15.

Figura 25 - Índice de Jaccard entre pares de abordagens (RF_all x RF_Corr, RF_all x RF_CorrImp, RF_Corr x RF_CorrImp; Top15) para BD, areia e pH por profundidade



Fonte: Elaborado pela autora (2026).

Em contraste, RF_all mantém todas as covariáveis candidatas, incluindo grupos de preditores espectrais altamente correlacionados que são removidos nas abordagens de seleção. Quando essas covariáveis preservam informação preditiva relevante, podem ser selecionadas entre as mais importantes apenas em RF_all, reduzindo a interseção com os conjuntos obtidos por RF_Corr e RF_CorrImp. Esse comportamento também é coerente com o melhor desempenho observado da RF_all para densidade do solo e areia (Tabela 17), indicando que parte das covariáveis espectrais eliminadas durante a filtragem ainda continha informação preditiva complementar.

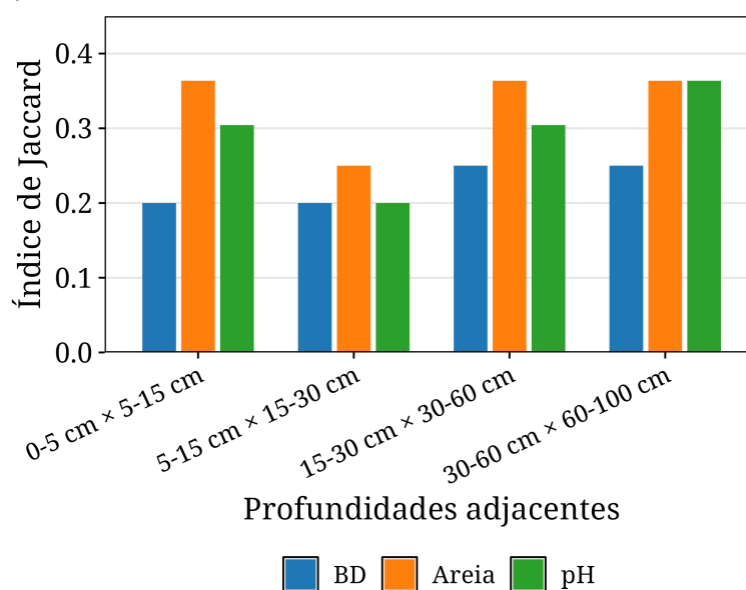
De encontro a isso, Akpa *et al.* (2014) salientam que é importante evitar a eliminação de covariáveis mesmo que elas sejam altamente correlacionadas entre si, pois elas podem carregar informações pedologicamente relevantes e úteis para as predições. Essa perspectiva indica que a redundância estatística nem sempre implica redundância informacional. Por outro lado, Zhao *et al.* (2023), que avaliaram diferentes métodos de seleção de covariáveis para predição de pH, não obtiveram ganhos pronunciados de desempenho com a seleção de preditores, pois no conjunto de treino, o R^2 variou de 0,56 (sem seleção) para até 0,61 com seleção de covariáveis, enquanto no conjunto de teste o desempenho aumentou de 0,45 (sem seleção) para até 0,71 dependendo do método de seleção de preditores.

O caso da areia na camada de 5–15 cm ilustra esse comportamento de forma clara. O índice de Jaccard de 0,76 entre RF_Corr e RF_CorrImp indica convergência quase completa entre as duas abordagens filtradas (13 de 15 covariáveis em comum), sugerindo que o poder preditivo estava concentrado em um conjunto reduzido de covariáveis, predominantemente topográficas. Esse resultado é compatível com a seleção recorrente de mde, mrvbf e stanh em todas as 15 execuções para esse atributo, sendo a elevada sobreposição explicada pela coincidência desses e de outros preditores que compõem o núcleo de covariáveis mais relevantes.

4.3.3.2 Entre profundidades

Os valores do índice de Jaccard entre profundidades adjacentes foram baixos para os três atributos, variando de 0,20 a 0,36 (Figura 26) sob abordagem RF_all. Considerando conjuntos de 15 covariáveis, esses valores correspondem aproximadamente ao compartilhamento de apenas três a seis preditores entre profundidades consecutivas, indicando baixa estabilidade do conjunto de covariáveis mais importantes ao longo do perfil do solo. A areia apresentou a maior consistência média entre profundidades ($J = 0,333$), seguida pelo pH ($J = 0,290$) e pela BD ($J = 0,225$).

Figura 26 - Índice de Jaccard entre profundidades adjacentes (RF_all Top15) para densidade do solo, areia e pH



Fonte: Elaborado pela autora (2026).

Para o teor de areia, os índices permaneceram entre 0,25 e 0,36, sugerindo maior persistência dos principais controles ambientais ao longo do perfil. Em contraste, a densidade do solo (BD) apresentou os menores valores de consistência (0,20–0,25), indicando maior variação dos preditores relevantes entre as camadas amostradas. O pH exibiu comportamento intermediário, com maior sobreposição entre 30–60 e 60–100 cm ($J = 0,36$), sugerindo que os fatores que controlam sua variabilidade tendem a ser mais semelhantes nas camadas mais profundas do que nas superficiais.

No mapeamento da textura do solo na Índia, Chattaraj *et al.* (2025), avaliaram a importância das covariáveis ambientais e para a fração areia foi possível visualizar um certo padrão de sobreposição entre as covariáveis mais importantes nas diferentes profundidades. Nas camadas superficiais (0–15 cm), houve predominância de covariáveis espectrais, responsáveis por cerca de 30% dos preditores mais relevantes. Na camada de 15–30 cm, observou-se uma transição, com aumento da importância das covariáveis de terreno, que passaram a ter contribuição equivalente às espectrais. Já nas camadas mais profundas (30–100 cm), as covariáveis climáticas tornaram-se predominantes, seguidas pelas covariáveis de terreno, evidenciando uma mudança gradual dos controles ambientais com a profundidade.

Os padrões observados na Figura 26 são consistentes com a natureza dos atributos do solo analisados e com os processos pedogenéticos que os controlam ao longo do perfil. A distribuição da fração areia é fortemente condicionada pela posição na paisagem e por processos geomorfológicos relativamente persistentes, favorecendo a recorrência dos mesmos preditores entre profundidades. Em contrapartida, a densidade do solo é influenciada por fatores que podem variar ao longo do perfil, incluindo estrutura, matéria orgânica, compactação e condições de drenagem, resultando em maior substituição das covariáveis mais importantes. Para o pH, a maior consistência nas camadas profundas pode refletir a predominância de controles geoquímicos mais estáveis, enquanto nas camadas superficiais a influência de processos biológicos e do manejo tende a aumentar a variabilidade dos preditores selecionados.

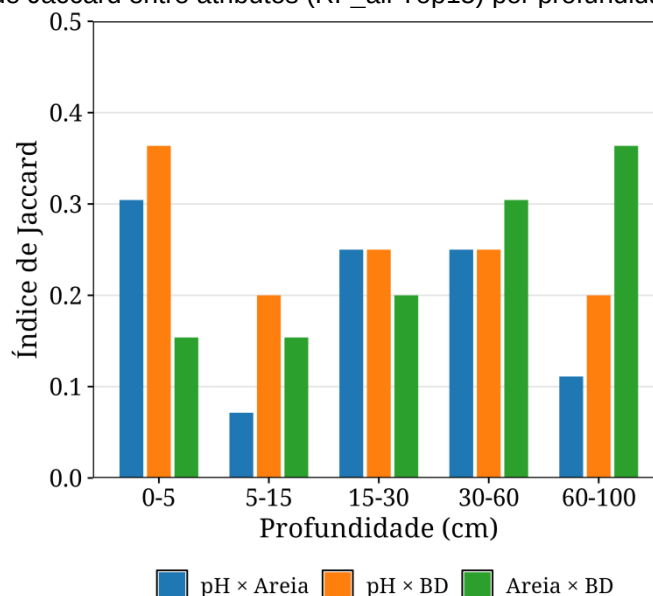
Entretanto, essa interpretação deve ser feita com cautela, pois parte da baixa sobreposição observada pode refletir instabilidade inerente ao

ranqueamento de importância de covariáveis no *Random Forest* (Nicodemus; Malley, 2009), podendo ser mais pronunciada quando aplicado a conjuntos de dados relativamente pequenos ($n = 64$). Nesse contexto, pequenas variações nas amostras utilizadas para treinamento podem modificar a posição relativa das covariáveis sem necessariamente indicar mudanças reais nos processos que controlam os atributos do solo. Assim, a baixa consistência entre profundidades provavelmente resulta da combinação entre mudanças efetivas nos controles ambientais ao longo do perfil e da variabilidade estatística associada às estimativas de importância das covariáveis.

4.3.3.3 Entre atributos

A sobreposição entre os conjuntos das 15 covariáveis mais importantes (RF_all) foi, em geral, baixa entre os diferentes atributos do solo (Figura 27). Os índices de Jaccard variaram de 0,07 a 0,36, com média geral de 0,23. Entre os pares de atributos, pH x BD apresentou a maior sobreposição média ($J = 0,25$), seguido por areia x BD ($J = 0,23$) e pH x areia ($J = 0,20$).

Figura 27 - Índice de Jaccard entre atributos (RF_all Top15) por profundidade



Fonte: Elaborado pela autora (2026).

O menor valor observado ocorreu para pH x areia na profundidade de 5–15 cm ($J = 0,07$), indicando compartilhamento mínimo entre os conjuntos de covariáveis selecionadas. Nessa camada, o conjunto de preditores do pH incluiu covariáveis espectrais e índices relacionados à curvatura e posição na

paisagem, enquanto o conjunto da areia foi composto principalmente por índices topográficos associados à posição na paisagem e à morfologia do terreno. Apenas duas covariáveis (b2blue e sbi) foram comuns aos dois conjuntos, sugerindo que os fatores ambientais mais relevantes para explicar a variabilidade desses atributos diferem substancialmente nessa profundidade.

Em contraste, a sobreposição entre areia e BD aumentou progressivamente com a profundidade, passando de $J = 0,15$ na camada de 0–5 cm para $J = 0,36$ em 60–100 cm. Esse comportamento é consistente com uma influência relativamente maior do material de origem nas camadas mais profundas, onde tanto a textura quanto a densidade do solo tendem a refletir mais diretamente características sedimentares e geomorfológicas da paisagem.

A implicação prática dos baixos índices de sobreposição entre atributos é que estratégias de seleção de covariáveis desenvolvidas para um atributo não são diretamente transferíveis para outro, ainda que sob as mesmas condições ambientais. Nesse sentido, Akpa *et al.* (2014) demonstram que diferentes grupos de covariáveis ambientais contribuem para explicar a variabilidade espacial das frações granulométricas do solo. Entretanto, a relevância de cada covariável não foi constante, variando conforme a fração granulométrica e a profundidade analisada. Esse comportamento evidencia que, mesmo entre atributos estritamente relacionados, a importância das covariáveis pode diferir, indicando que a transferência dos preditores mais relevantes entre diferentes atributos do solo deve ser realizada de forma criteriosa e fundamentada nos processos que controlam cada propriedade.

Isso reforça a necessidade de adoção de procedimentos de seleção específicos para cada atributo em estudos de mapeamento digital do solo, especialmente quando diferentes propriedades do solo apresentam mecanismos ambientais de controle distintos, pois aplicações que utilizam um único conjunto de covariáveis para prever múltiplos atributos podem não explorar adequadamente as relações específicas entre os atributos do solo e seus principais controles ambientais. Sendo assim, a especificidade dos conjuntos de preditores por atributo justifica abordagens de seleção independente para cada variável-alvo, auxiliando assim, inclusive, na melhoria da interpretabilidade dos resultados (Padarian *et al.*, 2020).

4.3.4 Espacialização dos atributos de solo

Os mapas de predição foram gerados utilizando a abordagem de melhor desempenho para cada atributo e profundidade (Tabela 17). As estatísticas descritivas dos mapas evidenciaram menor variabilidade em relação aos dados amostrais em todas as combinações de atributo e profundidade (Tabela 20).

Tabela 20 - Estatísticas descritivas dos mapas preditivos e comparação da variabilidade relativa com os dados observados para cada atributo e profundidade

Atributo	Profundidade	Média	Mín	Máx	DP mapa	CV mapa (%)	CV obs. (%)	Razão CV
BD	0–5 cm	1,396	1,131	1,618	0,054	3,83	13,9	0,28
BD	5–15 cm	1,559	1,361	1,696	0,044	2,80	8,4	0,33
BD	15–30 cm	1,577	1,374	1,812	0,046	2,93	10,1	0,29
BD	30–60 cm	1,610	1,482	1,800	0,031	1,94	7,0	0,28
BD	60–100 cm	1,600	1,474	1,732	0,021	1,29	7,7	0,17
Areia	0–5 cm	74,00	45,3	89	4,31	5,84	23,8	0,25
Areia	5–15 cm	72,50	47,7	87	3,57	4,92	22,6	0,22
Areia	15–30 cm	71,80	46,4	86,7	3,63	5,06	23,8	0,21
Areia	30–60 cm	69,00	40,4	87,9	4,63	6,71	29,5	0,23
Areia	60–100 cm	65,2	36,1	83,9	3,72	5,71	33,3	0,17
pH	0–5 cm	5,51	4,90	6,58	0,144	2,62	10,9	0,24
pH	5–15 cm	5,54	4,88	6,91	0,221	3,98	11,7	0,34
pH	15–30 cm	5,64	4,94	6,99	0,166	2,95	11,9	0,25
pH	30–60 cm	5,78	5,17	6,76	0,131	2,27	11,5	0,2
pH	60–100 cm	5,98	5,31	6,93	0,130	2,18	13,1	0,17

Fonte: Elaborado pela autora (2026).

O coeficiente de variação (CV) dos mapas variou de 1,29 a 3,83% para densidade do solo (BD), de 4,92 a 6,71% para o teor de areia e de 2,18 a 3,98% para pH. Em contraste, os CVs observados nos dados amostrais variaram de 7,0 a 13,9%, de 22,6 a 33,3% e de 10,9 a 13,1%, respectivamente. Consequentemente, a razão entre o CV dos mapas e o CV dos dados observados variou de 0,17 a 0,34, indicando que a dispersão relativa das predições correspondeu a aproximadamente 17–34% da dispersão relativa observada nas amostras. Esse resultado evidencia um efeito de suavização das predições, caracterizado pela redução da amplitude da variabilidade representada nos mapas em relação à variabilidade observada nos dados amostrais.

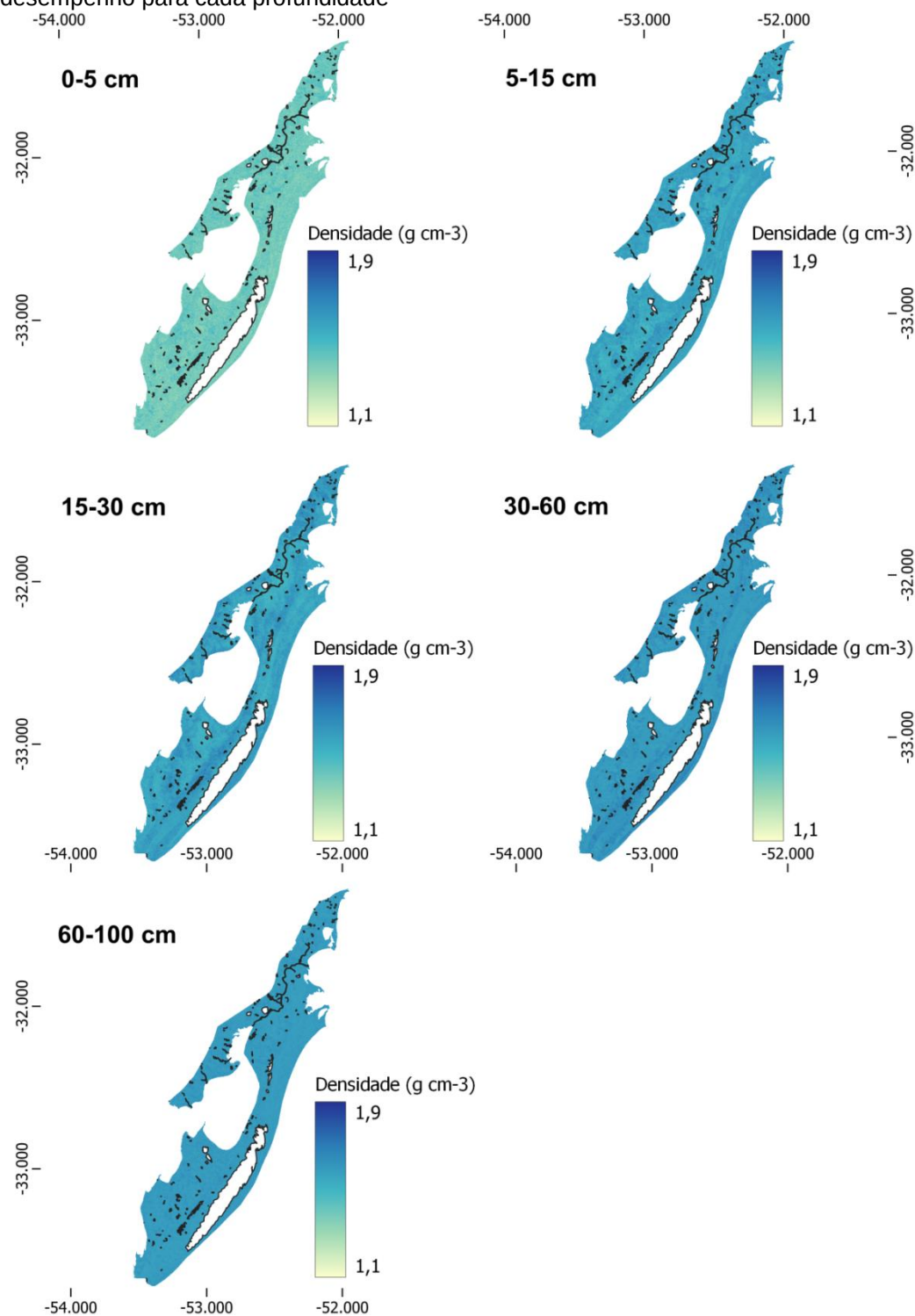
A redução da variabilidade observada nos mapas preditivos, frequentemente descrita na literatura como *smoothing*, é um comportamento característico de modelos baseados em *Random Forest* (Wyner et al., 2017).

Como as predições são obtidas pela agregação das estimativas de um grande número de árvores de decisão, o modelo tende a gerar superfícies espacialmente mais suaves, com menor amplitude de variação e, conseqüentemente, a subestimar valores extremos, tanto altos quanto baixos (Breiman, 2001). Esse comportamento é compatível com os baixos coeficientes de determinação obtidos neste estudo ($R^2 < 0,23$), sugerindo uma capacidade limitada dos modelos em reproduzir a variabilidade observada.

Os mapas de predição (Figura 28, Figura 29 e Figura 30) evidenciam padrões espaciais relativamente homogêneos para os três atributos, em consonância com a baixa variabilidade observada na Tabela 20.

Para BD, os mapas apresentam valores predominantemente entre 1,4 e 1,6 g cm⁻³ na maior parte da área de estudo. Com o aumento da profundidade, os mapas tornam-se progressivamente mais homogêneos, refletindo a redução da variabilidade espacial capturada pelo modelo. Esse comportamento é corroborado pela diminuição do coeficiente de variação (CV) dos mapas, de 3,83% em 0–5 cm para 1,29% em 60–100 cm (Figura 28).

Figura 28 - Distribuição espacial dos valores preditos de densidade do solo (BD) nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade

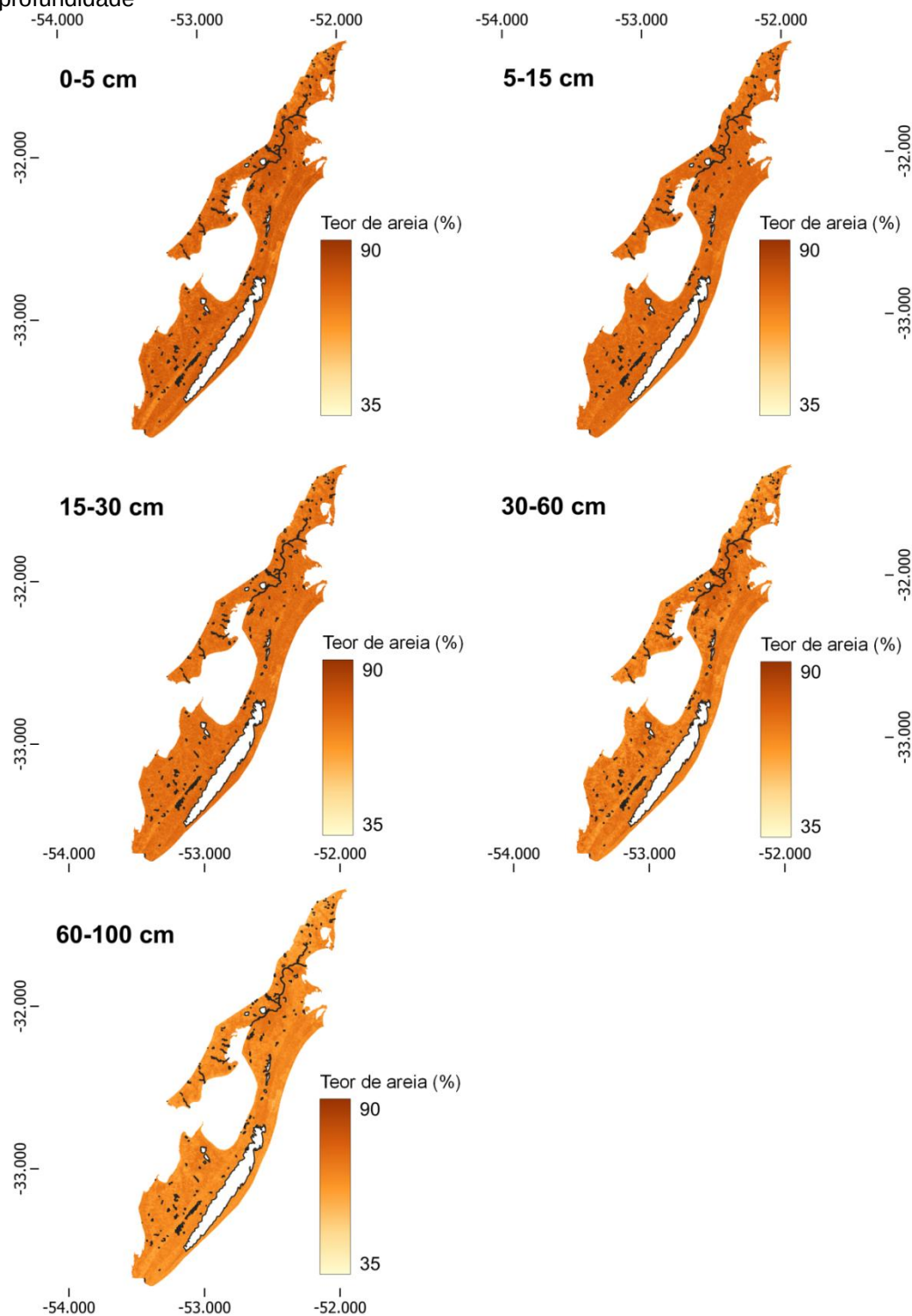


Fonte: Elaborado pela autora (2026).

Os mapas de areia (Figura 29) apresentam contraste espacial ligeiramente superior ao observado para densidade do solo. A tendência de

redução do teor médio de areia com a profundidade, observada nos dados amostrais, também foi reproduzida pelos mapas. O maior contraste espacial ocorreu na camada de 30–60 cm (CV = 6,71%).

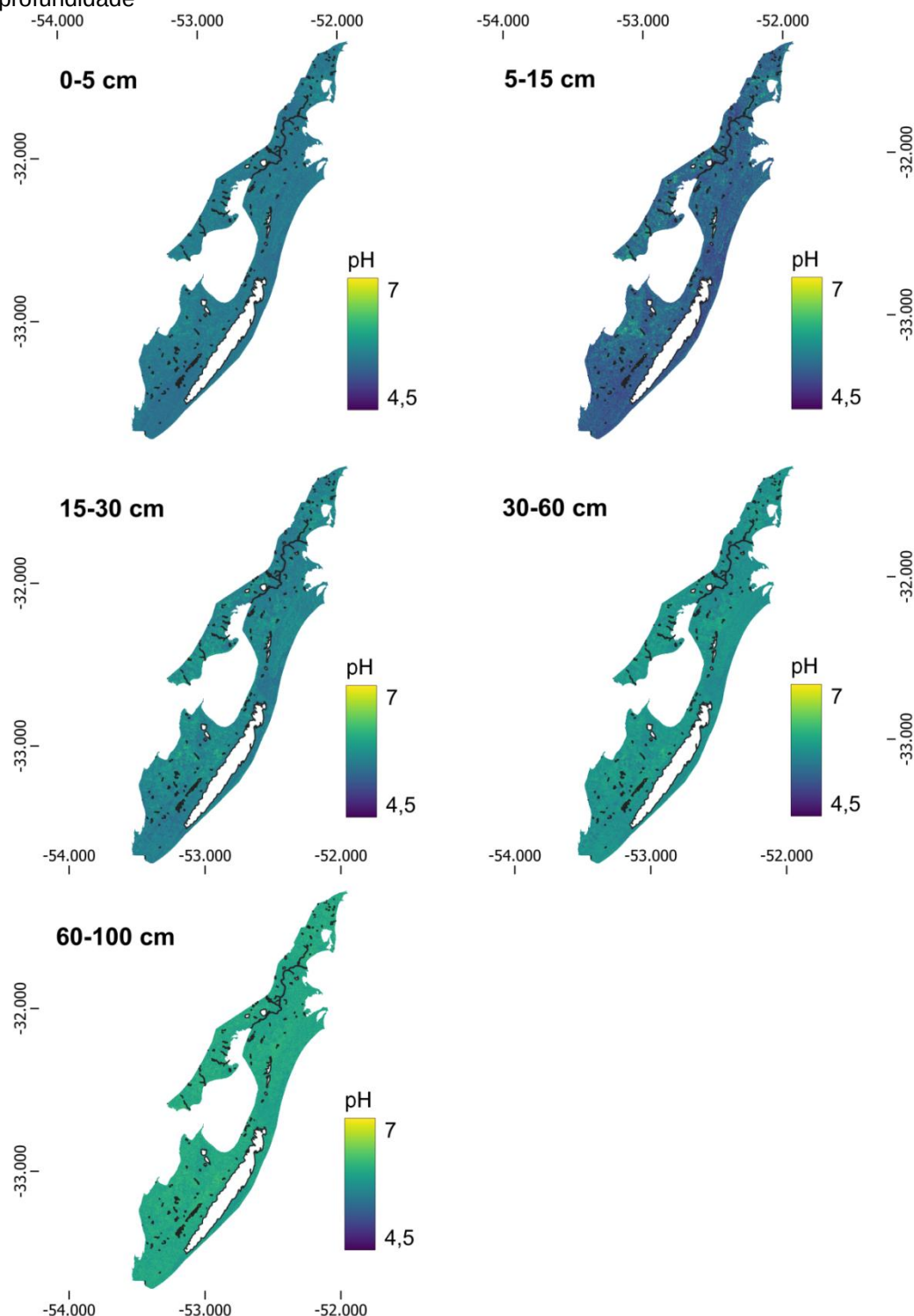
Figura 29 - Distribuição espacial dos valores preditos de teor de areia nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade



Fonte: Elaborado pela autora (2026).

Os mapas de pH (Figura 30) apresentaram valores predominantemente entre 5,4 e 5,7 e coeficientes de variação entre 2,18% e 3,98%. Assim como observado nos dados de campo, os valores médios de pH aumentam com a profundidade, passando de 5,51 na camada de 0–5 cm para 5,98 em 60–100 cm. A camada de 5–15 cm apresentou a maior variabilidade entre os mapas de pH ($CV = 3,98\%$), em concordância com o melhor desempenho obtido para essa profundidade ($R^2 = 0,229$), indicando maior capacidade do modelo em representar a variabilidade espacial relativa nessa camada.

Figura 30 - Distribuição espacial dos valores preditos de pH nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, obtida com o modelo de melhor desempenho para cada profundidade



Fonte: Elaborado pela autora (2026).

Do ponto de vista prático, esses resultados indicam que os mapas produzidos são mais adequados para representar tendências espaciais médias e gradientes regionais do que para identificar áreas de valores extremos de

densidade do solo, teor de areia ou pH, uma vez que esses extremos são significativamente atenuados pelo processo de predição.

Os mapas de incerteza, representados pelo coeficiente de variação (CV%), calculado a partir das predições individuais das árvores do *Random Forest*, evidenciaram diferenças marcantes entre os atributos (Tabela 21). Em média, a incerteza foi mais elevada para o teor de areia (22,3–29,7%), intermediária para o pH (9,9–11,9%) e menor para a densidade do solo (BD) (6,5–13,6%). Esses resultados indicam menor concordância entre as árvores da floresta para as predições do teor de areia, sugerindo maior dificuldade do modelo em representar sua distribuição espacial.

Tabela 21 - Estatísticas descritivas dos mapas de incerteza obtidos a partir do coeficiente de variação (CV%) entre as predições individuais das árvores do modelo *Random Forest* para cada atributo e profundidade

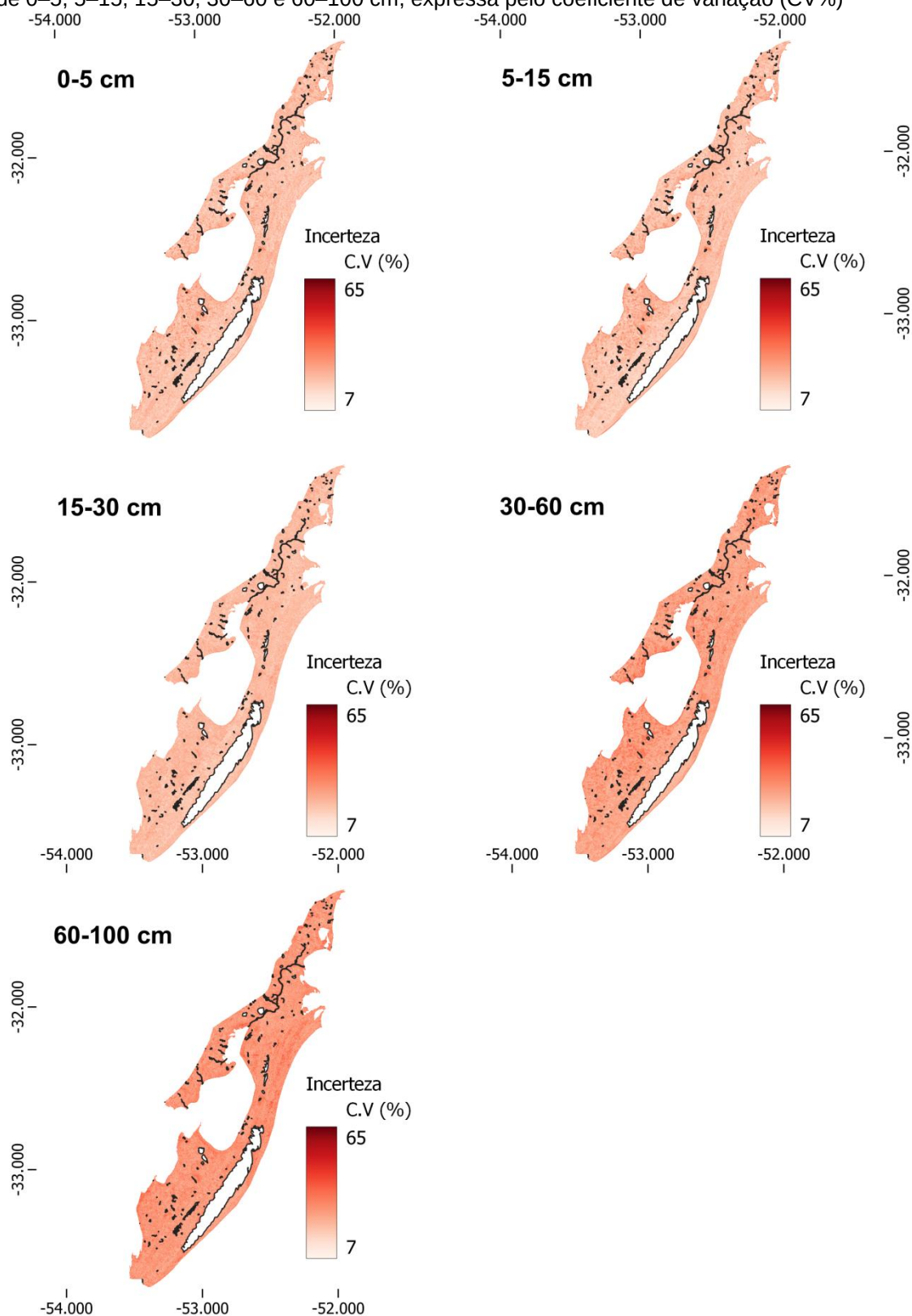
Atributo	Profundidade	CV% médio	Mín	Máx	DP	CV do CV
BD	0–5 cm	13,6	4,1	21,3	2,16	15,9
BD	5–15 cm	8	3,7	13,8	1,24	15,5
BD	15–30 cm	10,7	3,9	17,8	1,86	17,3
BD	30–60 cm	6,5	2,8	11,6	1,17	18,1
BD	60–100 cm	7,1	2,5	11,1	0,97	13,7
Areia	0–5 cm	22,3	7,5	40,6	4,21	18,9
Areia	5–15 cm	22,5	9,3	37,9	4,73	21
Areia	15–30 cm	22,5	10,4	41,8	3,2	14,2
Areia	30–60 cm	28	15,1	48	4,12	14,7
Areia	60–100 cm	29,7	14	63,2	3,9	13,2
pH	0–5 cm	9,9	3,6	16,7	1,34	13,6
pH	5–15 cm	11,3	4,7	21,6	2,39	21,1
pH	15–30 cm	11,9	4,8	20,2	2,37	19,8
pH	30–60 cm	11,5	5,5	15,6	1,39	12
pH	60–100 cm	11,6	5,4	17,1	1,39	12

Fonte: Elaborado pela autora (2026).

Para o teor de areia (Figura 31), o CV% médio aumentou progressivamente com a profundidade, passando de 22,3% em 0–5 cm para 29,7% em 60–100 cm, com valores máximos de até 63,2%. Esse comportamento indica redução da concordância entre as árvores do *Random Forest* nas camadas mais profundas, sugerindo maior incerteza nas predições em subsuperfície. A tendência de aumento da incerteza foi acompanhada pela redução do desempenho preditivo do modelo, com diminuição do R^2 de 0,175 na camada superficial para 0,090 em 60–100 cm e aumento do RMSE de 16,4

para 21,3% (Tabela 17). Esses resultados indicam que as covariáveis ambientais utilizadas apresentaram menor capacidade de explicar a distribuição espacial da areia nas camadas profundas, resultando em maior divergência entre as árvores e, conseqüentemente, maior incerteza preditiva. Além disso, a elevada incerteza observada para esse atributo é consistente com a maior compressão da variabilidade nos mapas preditivos, evidenciada pelas menores razões entre o CV dos mapas e o CV observado. Esse comportamento é coerente com os resultados de Greve *et al.* (2012), que relataram maior complexidade estrutural (maior quantidade e complexidade de nós) e maiores erros de predição em modelos desenvolvidos para areia grossa e areia fina em comparação com aqueles ajustados para argila e silte. Segundo os autores, esse resultado reflete a menor capacidade das covariáveis ambientais em representar a variabilidade espacial das frações de areia, o que pode contribuir para maior incerteza nas predições desse atributo.

Figura 31 - Distribuição espacial da incerteza das predições do teor de areia nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)



Fonte: Elaborado pela autora (2026).

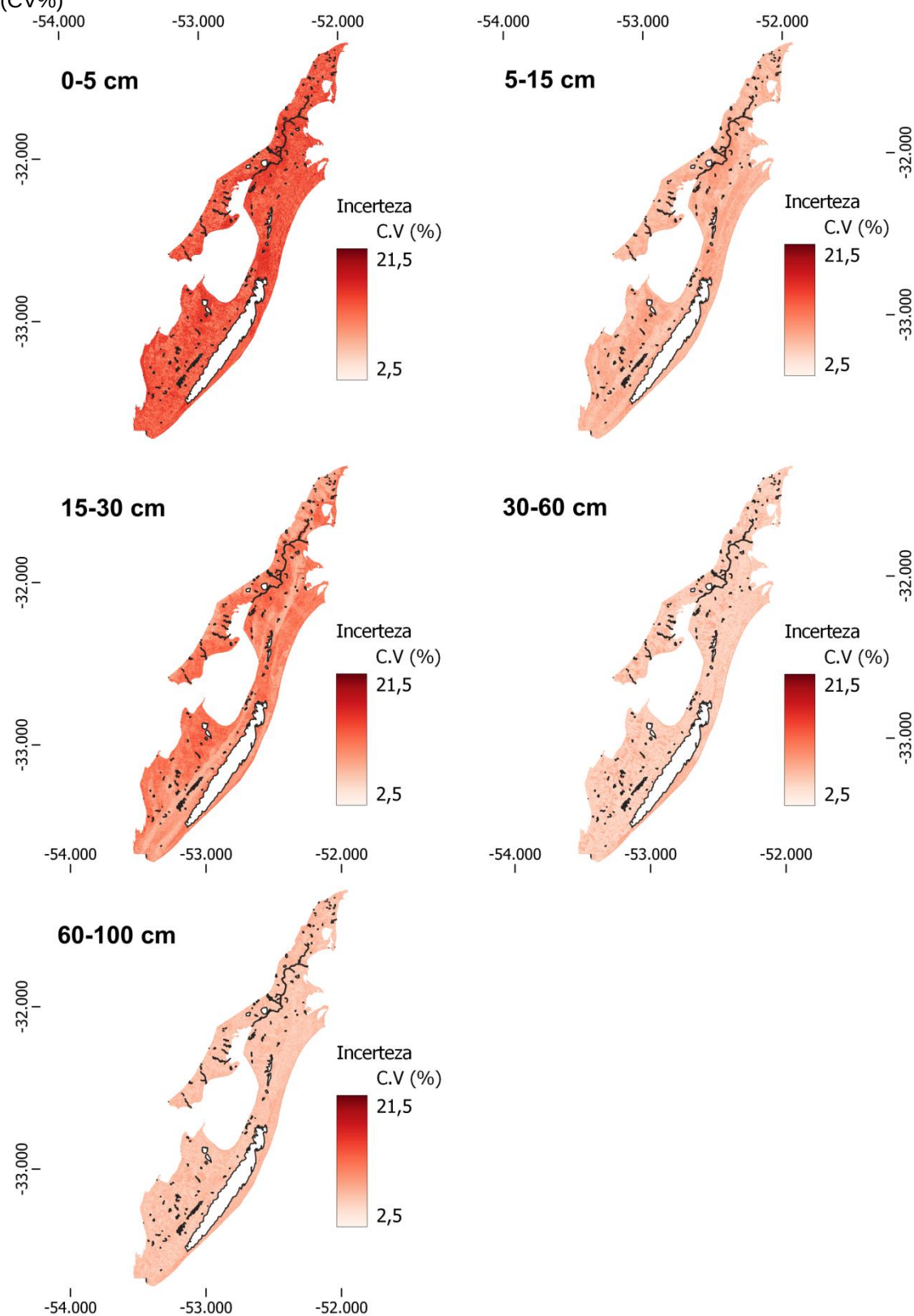
Para BD (Figura 32), a maior incerteza foi observada na camada superficial (CV médio = 13,6%), com redução até 30–60 cm (6,5%) e leve

aumento na camada de 60–100 cm (7,1%). Esse padrão indica maior divergência entre as árvores do *Random Forest* nas camadas superficiais, possivelmente associada à maior heterogeneidade espacial da densidade do solo nessa profundidade, onde o atributo é mais influenciado por fatores como teor de matéria orgânica, estrutura do solo e efeitos do uso e manejo.

A menor incerteza relativa da densidade do solo, quando comparada ao teor de areia, é consistente com o desempenho preditivo dos modelos, que apresentaram valores de R^2 entre 0,135 e 0,225 (Tabela 17). O melhor desempenho ocorreu em 15–30 cm ($R^2 = 0,225$; RMSE = 0,146; MAE = 0,117), justamente na profundidade em que foi observado aumento da incerteza, indicando que a maior variabilidade local dessa camada não necessariamente comprometeu a capacidade média do modelo em reproduzir a distribuição espacial do atributo.

Os valores de CV do mapa de incerteza (13,7–18,1%) indicam variabilidade espacial moderada da incerteza, evidenciando que a confiabilidade das estimativas não é homogênea em toda a área mapeada. Essa distribuição espacial da incerteza deve ser considerada na interpretação dos mapas preditivos, especialmente em regiões onde a divergência entre as árvores do modelo foi mais elevada.

Figura 32 - Distribuição espacial da incerteza das predições da densidade do solo nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)

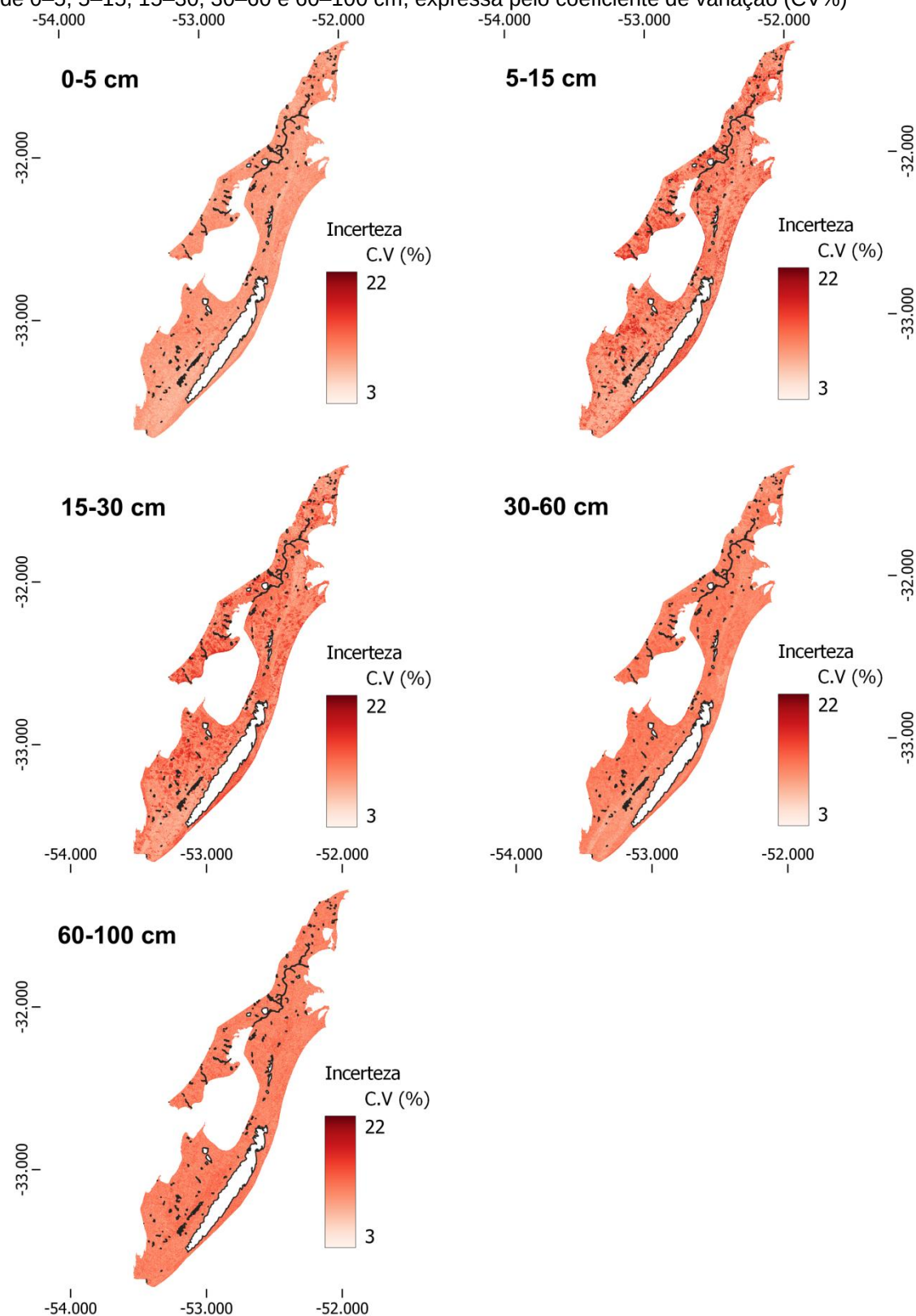


Fonte: Elaborado pela autora (2026).

Os mapas de pH (Figura 33) apresentaram os menores níveis de incerteza entre os atributos avaliados, com valores de CV médio relativamente estáveis entre 9,9% e 11,9%. A camada de 15–30 cm apresentou o maior valor médio de incerteza (11,9%), coincidindo com a redução do desempenho preditivo do modelo nessa profundidade ($R^2 = 0,133$; RMSE = 0,688; MAE = 0,563) (Tabela 17). Esse comportamento indica maior divergência entre as árvores do *Random Forest* nessa camada e sugere que a menor capacidade das covariáveis ambientais em explicar a distribuição espacial do pH resultou em maior dispersão das predições.

Apesar disso, a incerteza permaneceu relativamente baixa em todas as profundidades, indicando maior estabilidade das estimativas de pH em comparação aos demais atributos avaliados. Esse comportamento é consistente com o desempenho preditivo relativamente superior do pH nas camadas superficiais, especialmente em 5–15 cm, onde foi obtido o maior R^2 (0,229), associado ao menor grau de incerteza entre as profundidades avaliadas. Esse comportamento é coerente com Hateffard *et al.* (2024), que reportaram melhor desempenho do pH na extrapolação espacial com *Random Forest* em comparação a outros atributos do solo.

Figura 33 - Distribuição espacial da incerteza das predições do pH do solo nas profundidades de 0–5, 5–15, 15–30, 30–60 e 60–100 cm, expressa pelo coeficiente de variação (CV%)



Fonte: Elaborado pela autora (2026).

A distribuição espacial da incerteza não foi uniforme entre a área de estudo. Visualmente, os maiores valores de CV% concentraram-se na porção

sul da área (região adjacente à Lagoa Mirim) para os três atributos avaliados. Esse padrão indica menor concordância entre as árvores do *Random Forest* nessas regiões, resultando em maior dispersão entre as estimativas do modelo e, conseqüentemente, maior incerteza associada às predições. A concentração espacial de maiores níveis de incerteza sugere que essas áreas apresentam maior dificuldade de representação pelas covariáveis utilizadas, devendo ser consideradas com maior cautela na interpretação dos mapas preditivos. Além disso, esses resultados podem auxiliar no direcionamento de futuras campanhas de amostragem, priorizando regiões com maior incerteza para melhorar a caracterização espacial dos atributos e reduzir a incerteza dos modelos.

4.4 Conclusões

O desempenho dos modelos *Random Forest* para a predição da densidade do solo, teor de areia e pH evidenciou a complexidade das relações entre atributos do solo e covariáveis ambientais em uma planície costeira de baixa variabilidade topográfica. Embora os modelos tenham sido capazes de reconhecer padrões espaciais dos atributos avaliados, os resultados indicaram que a variabilidade pedológica existente no ambiente estudado é apenas parcialmente explicada pelas covariáveis utilizadas.

A comparação entre as estratégias de seleção de covariáveis mostrou que a escolha dos preditores exerce influência distinta conforme o atributo analisado. A redução do conjunto de preditores para as covariáveis mais importantes apresentou ganhos em parte significativa das combinações avaliadas, principalmente para o pH, demonstrando que a eliminação de informações redundantes pode favorecer a modelagem quando o número de observações é limitado. Entretanto, a resposta diferenciada entre os atributos evidencia que não há uma estratégia de seleção universalmente superior. Enquanto a densidade do solo e o teor de areia mantiveram relações mais distribuídas entre diferentes grupos de covariáveis, o pH apresentou maior dependência de um conjunto reduzido de covariáveis ambientais.

A elevada contribuição das covariáveis topográficas para os modelos revelou um comportamento não esperado para um ambiente de baixa variação do relevo. Embora áreas planas sejam frequentemente associadas a uma menor influência da topografia sobre a distribuição dos solos, os índices relacionados à posição na paisagem e à morfometria estiveram entre os preditores mais importantes, especialmente para o teor de areia. Esse resultado indica que pequenas variações na superfície, mesmo quando pouco perceptíveis em escala regional, carregam informações importantes.

A variação das covariáveis selecionadas entre atributos e profundidades evidencia que os controles ambientais responsáveis pela distribuição dos solos não são constantes ao longo do perfil. A densidade do solo, o teor de areia e o pH responderam de maneira distinta às informações ambientais disponíveis, refletindo diferenças nos processos que controlam sua formação e distribuição espacial. Esse resultado reforça a necessidade de abordagens específicas para cada atributo e profundidade, especialmente em ambientes costeiros

planos, onde pequenas mudanças nas condições deposicionais e hidrológicas podem produzir contrastes significativos nas propriedades do solo.

Os mapas preditivos produzidos permitiram representar a distribuição espacial geral da densidade do solo, teor de areia e pH, mas apresentaram a tendência esperada de redução da amplitude dos valores em relação aos dados observados. Essa suavização espacial evidencia uma característica comum dos modelos de aprendizado de máquina aplicados ao mapeamento digital de solos, especialmente quando a disponibilidade de amostras é limitada. A inclusão dos mapas de incerteza permitiu complementar a interpretação dos resultados, indicando áreas onde as estimativas apresentam maior variabilidade e auxiliando na identificação de regiões que podem ser priorizadas em futuras campanhas de amostragem.

5 Considerações finais

Os resultados obtidos ao longo desta pesquisa mostram que o mapeamento digital de solos em ambientes costeiros de baixa variação topográfica ainda enfrenta desafios relacionados à representação dos processos que controlam a distribuição espacial dos atributos do solo. Embora os modelos desenvolvidos tenham permitido identificar tendências espaciais para o estoque de carbono orgânico do solo, densidade do solo, teor de areia e pH, a capacidade preditiva variou entre os atributos e profundidades, evidenciando a complexidade das relações entre solo, relevo, vegetação e demais fatores ambientais nesse tipo de ambiente.

Em conjunto, os resultados demonstram que o desempenho dos modelos RF depende menos da quantidade de covariáveis utilizadas e mais da capacidade dessas covariáveis em representar os processos pedogenéticos, hidrológicos e deposicionais que condicionam a variabilidade dos solos. A redução dimensional e a filtragem de preditores redundantes mostraram-se viáveis, mas seus benefícios não foram uniformes. Além disso, a importância da microtopografia, mesmo em uma paisagem de baixa amplitude altimétrica e a variação dos fatores controladores entre atributos e profundidades reforçam que abordagens específicas para cada atributo tendem a ser mais adequadas do que estratégias generalistas.

Os mapas preditivos e de incerteza produzidos constituem uma base espacial explícita para a região, ainda que devam ser interpretados com cautela diante do desempenho preditivo limitado e da suavização típica dos modelos de aprendizado de máquina. A quantificação da incerteza, em particular, aponta áreas prioritárias para futuras amostragens e para o uso criterioso das estimativas em decisões de manejo e conservação.

De forma geral, este trabalho contribui para ampliar o conhecimento sobre as potencialidades e limitações do mapeamento digital de solos em planícies costeiras arenosas, oferecendo subsídios para o aprimoramento de futuras aplicações do mapeamento digital de solos nesses ambientes e reforçando a importância da integração entre conhecimento pedológico, seleção criteriosa de covariáveis e avaliação da incerteza na produção de mapas digitais mais confiáveis.

Referências

ABBRUZZINI, T. F.; FIGUEROA, D.; CRUZ, C. E. A.; VARÓN-RAMÍREZ, V. M.; SANTAMARIA, M. A. G.; PRADO, B. F. *et al.* Assessing soil nutrient spatial variability in Mexico through digital soil mapping: Implications for sustainable agriculture. **Geoderma Regional**, p. e01081, 2026.

ADENIJI, A.; HUANG, J.; LI, S.; LU, X.; GUO, R. Hot viewpoint on how soil texture, soil nutrient availability, and root exudates interact to shape microbial dynamics and plant health. **Plant and Soil**, p. 1-22, 2024.

ADENIYI, O. D.; BRENNING, A.; MAERKER, M. Spatial prediction of soil organic carbon: Combining machine learning with residual kriging in an agricultural lowland area (Lombardy region, Italy). **Geoderma**, v. 448, p. 116953, 2024.

ADHIKARI, K.; KHEIR, R. B.; GREVE, M. B.; BØCHER, P. K.; MALONE, B. P.; MINASNY, B.; MCBRATNEY, A. B.; GREVE, M. H. High-resolution 3-D mapping of soil texture in Denmark. **Soil Science Society of America Journal**, v. 77, n. 3, p. 860-876, 2013.

ADHIKARI, K.; HARTEMINK, A. E.; MINASNY, B.; BOU KHEIR, R.; GREVE, M. B.; GREVE, M. H. Digital mapping of soil organic carbon contents and stocks in Denmark. **PloS one**, v. 9, n. 8, p. e105519, 2014.

AITKENHEAD, M.; COULL, M. Mapping soil profile depth, bulk density and carbon stock in Scotland using remote sensing and spatial covariates. **European Journal of Soil Science**, v. 71, n. 4, p. 553-567, 2020.

AKPA, S. I.; ODEH, I. O.; BISHOP, T. F.; HARTEMINK, A. E. Digital mapping of soil particle-size fractions for Nigeria. **Soil Science Society of America Journal**, v. 78, n. 6, p. 1953-1966, 2014.

AKPA, S. I.; ODEH, I. O.; BISHOP, T. F.; HARTEMINK, A. E.; AMAPU, I. Y. Total soil organic carbon and carbon sequestration potential in Nigeria. **Geoderma**, v. 271, p. 202-215, 2016.

ARROUAYS, D.; GRUNDY, M.G.; HARTEMINK, A.E.; HEMPEL, J.W.; HEUVELINK, G.B.M.; HONG, S.Y.; LAGACHERIE, P.; LELYK, G.; MCBRATNEY, A.B.; MCKENZIE, N.J.; MENDONÇA-SANTOS, M.L.; MINASNY, B.; MONTANARELLA, L.; ODEH, I.O.A.; SANCHEZ, P.A.; THOMPSON, J.A.; ZHANG, G.L. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. **Advances in Agronomy**, v. 125, p. 93–134, 2014.

BAGHCHE-MARYAM, M. Z.; SHEKLABADI, M.; AYOUBI, S. Covariate selection approaches in spatial prediction of soil quality indices using machine learning models at the watershed scale, west of Iran. **Soil and Tillage Research**, v. 252, p. 106571, 2025.

BAITELLI, R.; ALMEIDA, L. E. S. B.; TOLDO Jr, E. E.; BARROS, C. E.; MARTINS, L. R.; NICOLODI, J. L. Retreat of the Rio Grande do Sul coastal zone, Brazil. Non Living Resources of the Southern Brazilian Coastal Zone and Continental Margin. Porto Alegre, Brazil: Editora da UFRGS, p. 62-68, 1999.

BARBOZA, E. G.; ROSA, M. L. C. C. Indicadores geológicos e geomorfológicos de setores em erosão na costa do Rio Grande do Sul. Nuevas miradas a la Problemática de los ambientes costeros Sur de Brasil, Uruguay y Argentina. 1º ed. Montevideo: DIRAC Facultad de Ciencias, v. 1, p. 83-98, 2014.

BARBOZA, E. G.; TOMAZELLI, L. J.; DILLENBURG, S. R.; ROSA, M. L. C. C. Planície costeira do Rio Grande Do Sul: erosão em longo período. **Revista Sociedad Uruguaya de Geología-SUG**, n. 15, 2009.

BATJES, N. H.; RIBEIRO, E.; VAN OOSTRUM, A. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). **Earth System Science Data**, v. 12, n. 1, p. 299-320, 2020.

BHUNIA, G. S.; KUMAR SHIT, P.; POURGHASEMI, H. R. Soil organic carbon mapping using remote sensing techniques and multivariate regression model. **Geocarto international**, v. 34, n. 2, p. 215-226, 2019.

BIVAND, R.; KEITT, T.; ROWLINGSON, B. rgdal: Bindings the “Geospatial” Data Abstraction Library. **R package version 1.5-23.**, [s. l.], 2021.

BOCOLI, F. A. *et al.* Can environmental variables, high sampling density and machine learning deliver detailed maps of soil organic carbon and carbon stock in tropical regions?. **Catena**, v. 249, p. 108718, 2025.

BONDI, G.; CREAMER, R.; FERRARI, A.; FENTON, O.; WALL, D. Using machine learning to predict soil bulk density on the basis of visual parameters: Tools for in-field and post-field evaluation. **Geoderma**, v. 318, p. 137-147, 2018.

BONDI, G.; FENTON, O.; CREAMER, R.; MARONGIU, I.; WALL, D. P. Trafficking intensity index for soil compaction management in grasslands. **Soil Use and Management**, v. 37, n. 3, p. 504-518, 2021.

BOU KHEIR, R.; GREVE, M. H.; BØCHER, P. K.; GREVE, M. B.; LARSEN, R.; MCCLOY, K. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: The case study of Denmark. **Journal of Environmental Management**, v. 91, n. 5, p. 1150-1160, 2010.

BOULESTEIX, A. L.; JANITZA, S.; KRUPPA, J.; KÖNIG, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 2, n. 6, p. 493-507, 2012.

BOUSLIHIM, Y.; JOHN, K.; MIFTAH, A.; AZMI, R.; ABOUTAYEB, R.; BOUASRIA, A.; RAZOUK, R.; HSSAINI, L. The effect of covariates on Soil Organic Matter and pH variability: A digital soil mapping approach using random forest model. **Annals of GIS**, v. 30, n. 2, p. 215-232, 2024.

BRASIL. **Levantamento de reconhecimento dos solos do Rio Grande do Sul**. Recife: Ministério da Agricultura, 1973. 431 p. (Boletim Técnico, 30)

BRASIL. **Projeto RADAMBRASIL: Folha SH.22 Porto Alegre e parte das Folhas SH.21 Uruguaiana e SI.22 Lagoa Mirim**. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 1986. v. 33. Disponível em: <https://biblioteca.ibge.gov.br/visualizacao/livros/liv19048.pdf>.

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

BRENNING, A.; BANGS, D.; BECKER, M. RSAGA: SAGA Geoprocessing and Terrain Analysis. R package version 1.4.2. 2025.

BRUNGARD, C. W.; BOETTINGER, J. L.; DUNIWAY, M. C.; WILLS, S. A.; EDWARDS Jr, T. C. Machine learning for predicting soil classes in three semi-arid landscapes. **Geoderma**, v. 239, p. 68-83, 2015.

BUTTERLY, C. R.; AMADO, T. J. C.; TANG, C. Soil acidity and acidification. In: **Subsoil constraints for crop production**. Cham: Springer International Publishing, 2022. p. 53-81.

CAMARGO, F. A. O.; SANTOS, G. A.; ZONTA, E. Alterações eletroquímicas em solos inundados. **Ciência Rural**, v. 29, n. 1, p. 171-180, 1999.

CAMPOS, A. R.; GIASSON, E.; COSTA, J. J. F.; MACHADO, I. R.; SILVA, E. B. D.; BONFATTI, B. R. Selection of environmental covariates for classifier training applied in digital soil mapping. **Revista Brasileira de Ciência do Solo**, v. 42, p. e0170414, 2019.

CAPITAINE, L.; GENUER, R.; THIÉBAUT, R. Random forests for high-dimensional longitudinal data. **Statistical methods in medical research**, v. 30, n. 1, p. 166-184, 2021.

CARBAJAL, C.; TUMBALOBOS-DEXTRE, M.; CONDORI-ATAUPILLCO, T.; CUELLAR-CONDORI, N.; GAVILAN, C. Spatial prediction of soil organic carbon

stocks across contrasting Andean basins, Peru. **Geoderma Regional**, p. e01026, 2025.

CARVALHO JUNIOR, W.; LAGACHERIE, P.; SILVA CHAGAS, C.; CALDERANO FILHO, B.; BHERING, S. B. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. **Geoderma**, v. 232, p. 479-486, 2014.

CARVALHO JUNIOR, W.; CALDERANO FILHO, B.; CHAGAS, C. S.; BHERING, S. B.; PEREIRA, N. R.; PINHEIRO, H. S. K. Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas. **Pesquisa Agropecuária Brasileira**, v. 51, n. 9, p. 1428-1437, 2016.

CASTILHOS, R. M.; MEURER, E. J.; KÄMPF, N.; PINTO, L. F. D. S. Mineralogia e fontes de potássio em solos no Rio Grande do Sul cultivados com arroz irrigado. **Revista brasileira de ciência do solo**, v. 26, n. 3, p. 579-587, 2002.

CASTRO-FRANCO, M.; COSTA, J. L.; PERALTA, N.; APARICIO, V. Prediction of soil properties at farm scale using a model-based soil sampling scheme and random forest. **Soil science**, v. 180, n. 2, p. 74-85, 2015.

CHATTARAJ, S.; DARIPA, A.; REZA, S. K.; GHOSH, B. N.; MAJHI, S.; MALLICK, D.; PAUL, S.; DEY, S.; RAHMAN, F. H. Soil texture and lithological discontinuity mapping for sustainable land use planning: An application of digital soil mapping in a part of Eastern Himalayan Foothills, India. **Catena**, v. 258, p. 109253, 2025.

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

CHUST, G.; PÉREZ-HAASE, A.; CHAVE, J.; PRETUS, J. L. Floristic patterns and plant traits of Mediterranean communities in fragmented habitats. **Journal of Biogeography**, v. 33, n. 7, p. 1235-1245, 2006.

COSTA, F. D. S.; BAYER, C.; ZANATTA, J. A.; MIELNICZUK, J. Estoque de carbono orgânico no solo e emissões de dióxido de carbono influenciadas por sistemas de manejo no sul do Brasil. **Revista Brasileira de Ciência do Solo**, v. 32, p. 323-332, 2008.

CPRM – COMPANHIA DE PESQUISA DE RECURSOS MINERAIS. **Projeto: Mapa Geológico e de Recursos Minerais do Estado do Rio Grande do Sul**. Escala: 1:750.000, 2008.

CROWTHER, T. W. *et al.* Quantifying global soil carbon losses in response to warming. **Nature**, v. 540, n. 7631, p. 104-108, 2016.

D'ÁVILA, G. A. J.; TORNQUIST, C. G.; HERMANN, J. M.; OVERBECK, G. E.; INDA, A. V. Refining regional soil C stocks estimates in temperate highlands of Southern Brazil. **Geoderma Regional**, v. 17, p. e00224, 2019.

DAI, J.; YANG, H.; MAIMAITITUERSUN, A.; MAIMAITIALI, K. Development and contribution analysis of a multi-depth soil organic matter prediction model integrating VIS–NIR spectroscopy and environmental covariates. **Plant and Soil**, v. 520, n. 2, p. 1249-1272, 2026.

DALMOLIN, R. S. D.; TEN CATEN, A. Mapeamento digital: nova abordagem em levantamento de solos. **Investigación Agraria**, v. 17, n. 2, p. 77-86, 2015.

DEMATTE, J. A. M. *et al.* Bare earth's surface spectra as a proxy for soil resource monitoring. **Scientific reports**, v. 10, n. 1, p. 4461, 2020.

DEMATTE, J. A. M. *et al.* The Brazilian soil priorities. **Geoderma Regional**, v. 29, p. e00503, 2022.

DEWITTE, O.; JONES, A.; ELBELRHITI, H.; HORION, S.; MONTANARELLA, L. Satellite remote sensing for soil mapping in Africa: An overview. *Progress in physical geography*, v. 36, n. 4, p. 514-538, 2012.

DHARUMARAJAN, S.; HEGDE, R.; SINGH, S. K. Spatial prediction of major soil properties using Random Forest techniques-A case study in semi-arid tropics of South India. **Geoderma Regional**, v. 10, p. 154-162, 2017.

DHARUMARAJAN, S.; KALAISELVI, B.; SUPUTHRA, A.; LALITHA, M.; HEGDE, R.; SINGH, S. K.; LAGACHERIE, P. Digital soil mapping of key GlobalSoilMap properties in Northern Karnataka Plateau. **Geoderma Regional**, v. 20, p. e00250, 2020.

DORMANN, C. F. *et al.* Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. **Ecography**, v. 36, n. 1, p. 27-46, 2013.

ELLERT, B. H.; BETTANY, J. R. Calculation of organic matter and nutrients stored in soils under contrasting management regimes. **Canadian Journal of Soil Science**, v. 75, n. 4, p. 529-538, 1995.

FARR, T. G.; ROSEN, P. A.; CARO, E.; CRIPPEN, R. *et al.* The shuttle radar topography mission. **Reviews of geophysics**, v. 45, n. 2, 2007.

FATHIZAD, H.; HAKIMZADEH ARDAKANI, M. A.; ESHGHI, B. Digital mapping of soil texture with the use of MLP, Decision Tree and RF models in Iran's desert region. **Modeling Earth Systems and Environment**, v. 11, n. 6, p. 414, 2025.

FERNANDES, F. A.; FERNANDES, A. H. B. M. Cálculo dos estoques de carbono do solo sob diferentes condições de manejo. Corumbá: Embrapa Pantanal. 4p. **Comunicado Técnico**, 69. 2008.

FLORINSKY, I. V. An illustrated introduction to general geomorphometry. **Progress in Physical Geography**, v. 41, n. 6, p. 723-752, 2017.

FORKUOR, G.; HOUNKPATIN, O. K.; WELP, G.; THIEL, M. High resolution mapping of soil properties using remote sensing variables in south-western Burkina Faso: a comparison of machine learning and multiple linear regression models. **PloS one**, v. 12, n. 1, p. e0170478, 2017.

GALLANT, J. C.; DOWLING, T. I. A multiresolution index of valley bottom flatness for mapping depositional areas. **Water resources research**, v. 39, n. 12, 2003.

GAO, L. *et al.* Soil wet aggregate distribution and pore size distribution under different tillage systems after 16 years in the Loess Plateau of China. **Catena**, v. 173, p. 38-47, 2019.

GELSLEICHTER, Y. A.; TOTH, A. J.; MICHELI, E.; MUTUMA, E.; WAWIRE, A.; CSORBA, A. Digital Soil Mapping of Soil Organic Carbon at Eastern Slopes of Mount Kenya. **Eurasian Soil Science**, v. 58, n. 11, p. 140, 2025.

GENUER, R.; POGGI, J. M.; TULEAU-MALOT, C. Variable selection using random forests. **Pattern recognition letters**, v. 31, n. 14, p. 2225-2236, 2010.

GOMES, L. C.; FARIA, R. M.; SOUZA, E.; VELOSO, G. V.; SCHAEFER, C. E. G.; FERNANDES FILHO, E. I. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337-350, 2019.

GREVE, M. H.; KHEIR, R. B.; GREVE, M. B.; BØCHER, P. K. Quantifying the ability of environmental parameters to predict soil texture fractions using regression-tree model with GIS and LIDAR data: The case study of Denmark. *Ecological Indicators*, v. 18, p. 1-10, 2012.

GRÖMPING, U. Variable importance in regression models. **Wiley interdisciplinary reviews: Computational statistics**, v. 7, n. 2, p. 137-152, 2015.

GUEVARA, M. *et al.* No silver bullet for digital soil mapping: country-specific soil organic carbon estimates across Latin America. **SOIL Discussions**, v. 2018, p. 1-20, 2018.

GUO, Z.; ADHIKARI, K.; CHELLASAMY, M.; GREVE, M. B.; OWENS, P. R.; GREVE, M. H. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. **Geoderma**, v. 340, p. 303-312, 2019.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of machine learning research**, v. 3, n. Mar, p. 1157-1182, 2003.

HATEFFARD, F.; SZATMÁRI, G.; NOVÁK, T. Applicability of machine learning models for predicting soil organic carbon content and bulk density under different soil conditions. **SOIL SCIENCE ANNUAL**, v. 74, n. 1, p. 165879, 2023.

HATEFFARD, F.; STEINBUCH, L.; HEUVELINK, G. B. M. Evaluating the extrapolation potential of random forest digital soil mapping. **Geoderma**, v. 441, p. 116740, 2024.

HE, W.; XIAO, Z.; LU, Q.; WEI, L.; LIU, X. Digital Mapping of Soil Particle Size Fractions in the Loess Plateau, China, Using Environmental Variables and Multivariate Random Forest. **Remote Sensing**, v. 16, n. 5, p. 785, 2024.

HEINZE, G.; WALLISCH, C.; DUNKLER, D. Variable selection—a review and recommendations for the practicing statistician. **Biometrical journal**, v. 60, n. 3, p. 431-449, 2018.

HENGL, T.; DE JESUS, J. M.; MACMILLAN, R. A.; BATJES, N. H.; HEUVELINK, G. B. M.; RIBEIRO, E.; SAMUEL-ROSA, A.; KEMPEN, B.; LEENAARS, J. G. B.; WALSH, M. G.; RUIPEREZ GONZALEZ, M. SoilGrids1km—global soil information based on automated mapping. **PloS one**, v. 9, n. 8, p. e105992, 2014.

HENGL, T. *et al.* SoilGrids250m: Global gridded soil information based on machine learning. **PLoS one**, v. 12, n. 2, p. e0169748, 2017.

HENGL, T.; NUSSBAUM, M.; WRIGHT, M. N.; HEUVELINK, G. B.; GRÄLER, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. **PeerJ**, v. 6, p. e5518, 2018.

HEUNG, B.; BULMER, C. E.; SCHMIDT, M. G. Predictive soil parent material mapping at a regional-scale: A Random Forest approach. **Geoderma**, v. 214, p. 141-154, 2014.

HEUNG, B.; HO, H. C.; ZHANG, J.; KNUDBY, A.; BULMER, C. E.; SCHMIDT, M. G. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma**, v. 265, p. 62-77, 2016.

HIJMANS, R. J. raster: Geographic Data Analysis and Modeling. R package version 3.5-15. 2022.

HUANG, H.; LIU, Y.; TONG, Z.; REN, Z.; XIE, Y. Digital Mapping of Soil pH and Driving Factor Analysis Based on Environmental Variable Screening. **Sustainability**, v. 17, n. 7, p. 3173, 2025.

HUE, N. Soil acidity: Development, impacts, and management. In: **Structure and functions of pedosphere**. Singapore: Springer Nature Singapore, 2022. p. 103-131.

HWANG, C.M.; YANG, M. S.; HUNG, W. L. New similarity measures of intuitionistic fuzzy sets based on the Jaccard index with its application to clustering. **International Journal of Intelligent Systems**, v. 33, n. 8, p. 1672-1688, 2018.

JALOTA, S. K.; SINGH, S.; CHAHAL, G. B. S.; RAY, S. S.; PANIGRAGHY, S.; SINGH, K. B. Soil texture, climate and management effects on plant growth, grain yield and water use by rainfed maize–wheat cropping system: Field and simulation study. **Agricultural water management**, v. 97, n. 1, p. 83-90, 2010.

JENA, R. K. *et al.* Spatial prediction of soil particle-size fractions using digital soil mapping in the North Eastern Region of India. **Land**, v. 12, n. 7, p. 1295, 2023.

JENNY, H. **Factors of Soil Formation: A System of Quantitative Pedology**. New York: McGraw-Hill Book Company, 1941. 281 p.

JOHN, K.; ABRAHAM ISONG, I.; MICHAEL KEBONYE, N.; OKON AYITO, E.; CHAPMAN AGYEMAN, P.; MARCUS AFU, S. Using machine learning algorithms to estimate soil organic carbon variability with environmental variables and soil nutrient indicators in an alluvial soil. **Land**, v. 9, n. 12, p. 487, 2020.

KÄMPF, N.; CURI, N. Formação e evolução do solo (pedogênese). In: KER, J. C.; CURI, N.; SCHAEFER, C. E. G. R.; VIDAL-TORRADO, P. (ed.). *Pedologia*:

fundamentos. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2015. p. 207–302.

KASRAEI, B.; SCHMIDT, M. G.; ZHANG, J.; BULMER, C. E.; FILATOW, D. S.; PENNELL, T.; HEUNG, B. A framework for optimizing environmental covariates to support model interpretability in digital soil mapping. **Geoderma**, v. 445, p. 116873, 2024.

KONRA, N.; KALAMBUKATTU, J. G.; KUMAR, S. Spatial Prediction of Soil Organic Carbon Stock Using Machine Learning in a Himalayan Watershed. **Journal of the Indian Society of Remote Sensing**, p. 1-11, 2026.

KUPLICH, T. M.; CAPOANE, V.; COSTA, L. F. F. O avanço da soja no bioma Pampa. **Boletim Geográfico do Rio Grande do Sul**, v. 31, n. 89, p. 53, 2018.

LAL, R. Soil carbon sequestration to mitigate climate change. **Geoderma**, v. 123, n. 1-2, p. 1-22, 2004.

LAMICHHANE, S.; KUMAR, L.; WILSON, B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. **Geoderma**, v. 352, p. 395-413, 2019.

LE MOS, G. S.; RIZZI, R. A expansão da soja no bioma pampa brasileiro e sua interação espaço-temporal com arroz e campo. **Boletim Geográfico do Rio Grande do Sul**, n. 35, p. 9-26, 2020.

LI, X.; MCCARTY, G. W.; DU, L.; LEE, S. Use of topographic models for mapping soil properties and processes. **Soil Systems**, v. 4, n. 2, p. 32, 2020.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R news**, v. 2, n. 3, p. 18-22, 2002.

LIEß, M.; GLASER, B.; HUWE, B. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. **Geoderma**, v. 170, p. 70-79, 2012.

LIMA, C. L. R. D.; PILLON, C. N.; SUZUKI, L. E. A. S.; CRUZ, L. E. C. D. Atributos físicos de um Planossolo Háplico sob sistemas de manejo comparados aos do campo nativo. **Revista Brasileira de Ciência do Solo**, v. 32, p. 1849-1855, 2008.

LIU, Z.; LUONG, P.; BOLEY, M.; SCHMIDT, D. F. Improving random forests by smoothing. **Machine Learning**, v. 115, n. 7, p. 152, 2026.

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

- MA, Y.; MINASNY, B.; MCBRATNEY, A.; POGGIO, L.; FAJARDO, M. Predicting soil properties in 3D: Should depth be a covariate?. **Geoderma**, v. 383, p. 114794, 2021.
- MAHDAVI, S.; SALEHI, B.; GRANGER, J.; AMANI, M.; BRISCO, B.; HUANG, W. Remote sensing for wetland classification: A comprehensive review. **GIScience & remote sensing**, v. 55, n. 5, p. 623-658, 2018.
- MANTEGHI, S.; MORAVEJ, K.; MOUSAVI, S. R.; DELAVAR, M. A.; MASTINU, A. Digital soil mapping for soil types using machine learning approaches at the landscape scale in the arid regions of Iran. **Advances in Space Research**, v. 74, n. 1, p. 1-16, 2024.
- MAPBIOMAS. Coleção 9 (2023) de Mapas de Cobertura e Uso de Solo do Brasil. Disponível em: <https://brasil.mapbiomas.org>
- MASHALABA, L.; GALLEGUILLOS, M.; SEGUEL, O.; POBLETE-OLIVARES, J. Predicting spatial variability of selected soil properties using digital soil mapping in a rainfed vineyard of central Chile. **Geoderma Regional**, v. 22, p. e00289, 2020.
- MCBRATNEY, A. B.; SANTOS, M. L. M.; MINASNY, B. On digital soil mapping. **Geoderma**, v. 117, n. 1-2, p. 3-52, 2003.
- MCBRATNEY, A.; DE GRUIJTER, J.; BRYCE, A. Pedometrics timeline. **Geoderma**, v. 338, p. 568-575, 2019.
- MEIER, M.; SOUZA, E. D.; FRANCELINO, M. R.; FERNANDES FILHO, E. I.; SCHAEFER, C. E. G. R. Digital soil mapping using machine learning algorithms in a tropical mountainous area. **Revista Brasileira de Ciência do Solo**, v. 42, p. e0170421, 2018.
- MEYER, H.; REUDENBACH, C.; HENGL, T.; KATURJI, M.; NAUSS, T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. **Environmental Modelling & Software**, v. 101, p. 1-9, 2018.
- MIGUEL, P.; PEDRON, F. A.; MOURA-BUENO, J. M.; DALMOLIN, R. S. D. Solos da região de Pelotas no Litoral Sul do Rio Grande do Sul. In: FONTANA, A.; LOSS, A.; PEDRON, F. A. (Org.). **Compêndio de Solos do Brasil: volume 2**. 2. ed. Santa Maria, RS: Sociedade Brasileira de Ciência do Solo - Núcleo Regional Sul, 2025. p. 27-45.
- MINASNY, B.; MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. **Computers & geosciences**, v. 32, n. 9, p. 1378-1388, 2006.
- MINASNY, B.; MCBRATNEY, A. B. Digital soil mapping: A brief history and some lessons. **Geoderma**, v. 264, p. 301-311, 2016.

- MITSCH, W. J.; BERNAL, B.; NAHLIK, A. M.; MANDER, Ü.; ZHANG, L.; ANDERSON, C. J.; JORGENSEN, S. E.; BRIX, H. Wetlands, carbon, and climate change. **Landscape ecology**, v. 28, n. 4, p. 583-597, 2013.
- MOSLEH, Z.; SALEHI, M. H.; JAFARI, A.; BORUJENI, I. E.; MEHNATKESH, A. The effectiveness of digital soil mapping to predict soil properties over low-relief areas. **Environmental monitoring and assessment**, v. 188, n. 3, p. 195, 2016.
- MOUSAVI, A.; KARIMI, A.; MALEKI, S.; SAFARI, T.; TAGHIZADEH-MEHRJARDI, R. Digital mapping of selected soil properties using machine learning and geostatistical techniques in Mashhad plain, northeastern Iran. **Environmental Earth Sciences**, v. 82, n. 9, p. 234, 2023.
- MULDER, V. L.; DE BRUIN, S.; SCHAEPMAN, M. E.; MAYR, T. R. The use of remote sensing in soil and terrain mapping—A review. **Geoderma**, v. 162, n. 1-2, p. 1-19, 2011.
- NACHTIGALL, S. D. **Modelagem para predição de classes de solo em escala de município**. Tese (Doutorado) – Manejo e Conservação do Solo e da Água, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, 2022.
- NGUEMEZI, C.; TEMATIO, P.; SILATSA, F. B.; YEMEFACK, M. Spatial variation and temporal decline (1985–2017) of soil organic carbon stocks (SOCS) in relation to land use types in Tombel area, South-West Cameroon. **Soil and Tillage Research**, v. 213, p. 105114, 2021.
- NICODEMUS, K. K.; MALLEY, J. D. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, v. 25, n. 15, p. 1884-1890, 2009.
- NUSSBAUM, M.; SPIESS, K.; BALTENSWEILER, A.; GROB, U.; KELLER, A.; GREINER, L.; SCHAEPMAN, M. E.; PAPRITZ, A. Evaluation of digital soil mapping approaches with large sets of environmental covariates. *Soil*, v. 4, n. 1, p. 1-22, 2018.
- OLAYA, V.; CONRAD, O. Geomorphometry in SAGA. **Developments in soil science**, v. 33, p. 293-308, 2009.
- OLIVEIRA, D. M. S. *et al.* Climate-smart agriculture and soil C sequestration in Brazilian Cerrado: a systematic review. **Revista Brasileira de Ciência do Solo**, v. 47, p. e0220055, 2023.
- OUYANG, X.; LEE, S. Y. Improved estimates on global carbon stock and carbon pools in tidal wetlands. **Nature communications**, v. 11, n. 1, p. 317, 2020.
- OVERBECK, G. E.; MÜLLER, S. C.; FIDELIS, A.; PFADENHAUER, J.; PILLAR, V. D.; BLANCO, C. C.; BOLDRINI, I. I.; BOTH, R.; FORNECK, E. D. Brazil's

neglected biome: the South Brazilian Campos. **Perspectives in Plant Ecology, Evolution and Systematics**, v. 9, n. 2, p. 101-116, 2007.

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

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

PARVIZI, Y.; FATEHI, S. Geospatial digital mapping of soil organic carbon using machine learning and geostatistical methods in different land uses. **Scientific Reports**, v. 15, n. 1, p. 4449, 2025.

PATRICHE, C. V.; ROȘCA, B.; PÎRNĂU, R. G.; VASILINIUC, I. Spatial modelling of topsoil properties in Romania using geostatistical methods and machine learning. **Plos one**, v. 18, n. 8, p. e0289286, 2023.

PEDRON, F. A.; DALMOLIN, R. S. D. Solos arenosos do bioma Pampa brasileiro. Santa Maria: Editora UFSM, 2019. 280p.

PELEGIRINO, M. H. P.; GUILHERME, L. R. G.; LIMA, G. O.; POPPIEL, R.; ADHIKARI, K.; DEMATTE, J. M.; CURTI, N.; MENEZES, M. D. Optimizing soil texture spatial prediction in the Brazilian Cerrado: Insights from random forest and spectral data. **Geoderma Regional**, v. 40, p. e00922, 2025.

PICCINI, C.; MARCHETTI, A.; FRANCAVIGLIA, R. Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. **Ecological indicators**, v. 36, p. 301-314, 2014.

PIIKKI, K.; WETTERLIND, J.; SÖDERSTRÖM, M.; STENBERG, B. Perspectives on validation in digital soil mapping of continuous attributes—A review. **Soil Use and Management**, v. 37, n. 1, p. 7-21, 2021.

PLANTE, A. F.; CONANT, R. T.; STEWART, C. E.; PAUSTIAN, K.; SIX, J. *et al.* Impact of soil texture on the distribution of soil organic matter in physical and chemical fractions. **Soil Science Society of America Journal**, v. 70, n. 1, p. 287-296, 2006.

PLOTON, P. *et al.* Spatial validation reveals poor predictive performance of large-scale ecological mapping models. **Nature communications**, v. 11, n. 1, p. 4540, 2020.

POGGIO, L. *et al.* SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. **Soil**, v. 7, n. 1, p. 217-240, 2021.

PORTZ, L.; MANZOLLI, R. P.; CORRÊA, I. C. S. Ferramentas de gestão ambiental aplicadas na zona costeira do Rio Grande do Sul, Brasil. **Revista de Gestão Costeira Integrada-Journal of Integrated Coastal Zone Management**, v. 11, n. 4, p. 459-470, 2011.

POULADI, N.; GHOLIZADEH, A.; KHOSRAVI, V.; BORUVKA, L. Digital mapping of soil organic carbon using remote sensing data: A systematic review. **Catena**, v. 232, p. 107409, 2023.

PROBST, P.; BOULESTEIX, A.L. To tune or not to tune the number of trees in random forest. **Journal of Machine Learning Research**, v. 18, n. 181, p. 1-18, 2018.

PROBST, P.; WRIGHT, M. N.; BOULESTEIX, A.L. Hyperparameters and tuning strategies for random forest. **Wiley Interdisciplinary Reviews: data mining and knowledge discovery**, v. 9, n. 3, p. e1301, 2019.

PUEYO, Y.; ALADOS, C. L.; BARRANTES, O.; KOMAC, B.; RIETKERK, M. Differences in gypsum plant communities associated with habitat fragmentation and livestock grazing. **Ecological Applications**, v. 18, n. 4, p. 954-964, 2008.

PULIDO, M.; SCHNABEL, S.; LAVADO CONTADOR, J. F.; LOZANO-PARRA, J.; GONZÁLEZ, F. The impact of heavy grazing on soil quality and pasture production in rangelands of SW Spain. **Land Degradation & Development**, v. 29, n. 2, p. 219-230, 2018.

QGIS DEVELOPMENT TEAM. QGIS Geographic Information System. Version 3.22. Open Source Geospatial Foundation Project, 2022.

R CORE TEAM. A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**. 2018. Vienna, Austria: R Foundation for Statistical Computing.

R CORE TEAM. A Language and Environment for Statistical Computing. **R Foundation for Statistical Computing**. 2024. Vienna, Austria: R Foundation for Statistical Computing.

REICHERT, J. M.; AMADO, T. J. C.; REINERT, D. J.; RODRIGUES, M. F.; SUZUKI, L. E. A. S. Land use effects on subtropical, sandy soil under sandization/desertification processes. **Agriculture, Ecosystems & Environment**, v. 233, p. 370-380, 2016.

REIS, R. B.; OLIVEIRA, J. R.; LIMA, C. L. R.; BADINELLI, P. G.; TRINDADE, J. K.; CARMONA, F. C.; SOUSA, R. O.; ANGHINONI, I.; CARVALHO, P. C. F.; MARTINS, A. P.; BAYER, C.; CARLOS, F. S. Integrated crop-livestock systems with crop diversification increase soil C and N levels and stocks in subtropical paddy fields in Brazil. **Agriculture, Ecosystems & Environment**, v. 407, p. 110436, 2026.

RIBEIRO, P. L.; BAMBERG, A. L.; REIS, D. A.; OLIVEIRA, A. C. B. D. Condições físico-hídricas de Planossolo cultivado com soja em plantio direto e preparo convencional. **Pesquisa Agropecuária Brasileira**, v. 51, p. 1484-1491, 2016.

ROBERTS, D. R. *et al.* Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. **Ecography**, v. 40, n. 8, p. 913-929, 2017.

ROSA, C. M. D.; CASTILHOS, R. M. V.; PAULETTO, E. A.; PILLON, C. N.; LEAL, O. D. A. Conteúdo de carbono orgânico em Planossolo Háplico sob sistemas de manejo do arroz irrigado. **Revista Brasileira de Ciência do Solo**, v. 35, n. 5, p. 1769-1776, 2011.

ROSSEL, R. A. V.; WEBSTER, R. Predicting soil properties from the Australian soil visible–near infrared spectroscopic database. **European journal of soil science**, v. 63, n. 6, p. 848-860, 2012.

ROUDIER, P. cLHS - A R Package for Conditioned Latin Hypercube Sampling. 2014.

SAFANELLI, J. L.; DEMATTÊ, J. A. M.; CHABRILLAT, S.; POPPIEL, R. R.; RIZZO, R.; DOTTO, A. C.; SILVERO, N. E. Q.; MENDES, W. S.; BONFATTI, B. R.; RUIZ, L. F. C.; TEN CATEN, A.; DALMOLIN, R. S. Leveraging the application of Earth observation data for mapping cropland soils in Brazil. **Geoderma**, v. 396, p. 115042, 2021.

SANAEI, A.; LI, M.; ALI, A. Topography, grazing, and soil textures control over rangelands' vegetation quantity and quality. **Science of the Total Environment**, v. 697, p. 134153, 2019.

SANTOS, H. G. dos; JACOMINE, P. K. T.; ANJOS, L. H. C. dos; OLIVEIRA, V. A. de; LUMBRERAS, J. F.; COELHO, M. R.; ALMEIDA, J. A. de; ARAUJO FILHO, J. C. de; LIMA, H. N.; MARQUES, F. A.; OLIVEIRA, J. B. de; CUNHA, T. J. F. **Sistema Brasileiro de Classificação de Solos**. 6ª ed. Brasília, DF, Embrapa, 2025, 393p.

SCHARLEMANN, J. P.; TANNER, E. V.; HIEDERER, R.; KAPOV, V. Global soil carbon: understanding and managing the largest terrestrial carbon pool. **Carbon management**, v. 5, n. 1, p. 81-91, 2014.

SCHRATZ, P.; MUENCHOW, J.; ITURRITXA, E.; RICHTER, J.; BRENNING, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. **Ecological Modelling**, v. 406, p. 109-120, 2019.

SILVA, A. R. E.; REHBEIN, M. O. Análise e mapeamento geomorfológico da área de influência da Planície Costeira de Pelotas (Rio Grande do Sul, Brasil). **Revista Brasileira de Geomorfologia**, v. 19, n. 3, 2018.

SISTI, C. P.; SANTOS, H. P.; KOHHANN, R.; ALVES, B. J.; URQUIAGA, S.; BODDEY, R. M. Change in carbon and nitrogen stocks in soil under 13 years of conventional or zero tillage in southern Brazil. **Soil and tillage research**, v. 76, n. 1, p. 39-58, 2004.

SMITH, P.; FANG, C.; DAWSON, J. J.; MONCRIEFF, J. B. Impact of global warming on soil organic carbon. **Advances in agronomy**, v. 97, p. 1-43, 2008.

SOMARATHNA, P. D. S. N.; MINASNY, B.; 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**, v. 81, n. 6, p. 1413-1426, 2017.

SOMMER, R.; BOSSIO, D. Dynamics and climate change mitigation potential of soil organic carbon sequestration. **Journal of environmental management**, v. 144, p. 83-87, 2014.

SONG, Y. Q.; YANG, L. A.; LI, B.; HU, Y. M.; WANG, A. L.; ZHOU, W.; CUI, X. S.; LIU, Y. L. Spatial prediction of soil organic matter using a hybrid geostatistical model of an extreme learning machine and ordinary kriging. **Sustainability**, v. 9, n. 5, p. 754, 2017.

SOUSA, R. O.; SILVA, L. S.; CARLOS, F. S.; GONÇALVES, G. K. Manejo de solo em terras baixas. In: **Manejo e Conservação do Solo e da Água**. 1. ed. Santa Maria, RS: Núcleo Regional Sul da Sociedade Brasileira de Ciência do Solo, 2023. p. 457-483.

STOCKMANN, U.; ADAMS, M.A.; CRAWFORD, J.W.; FIELD, D.J.; HENAKAARCHCHI, N.; JENKINS, M.; MINASNY, B.; MCBRATNEY, A.B.; DE COURCELLES, V.R.; SINGH, K.; WHEELER, I.; ABBOTT, L.; ANGERS, D.A.; BALDOCK, J.; BIRD, M.; BROOKES, P.C.; CHENU, C.; JASTROW, J.D.; LAL, R.; LEHMANN, J.; O'DONNELL, A.G.; PARTON, W.J.; TORN, D.; TORN, M.S. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. **Agriculture, Ecosystems & Environment**, v. 164, p. 80–99, 2013.

STRECK, E. V. *et al.*; FLORES, C. A.; SCHNEIDER, P. (clbs.) **Solos do Rio Grande do Sul**. 3. Ed., ver. e ampl. Porto Alegre: Emater/RS-Ascar, 2018. 252p.

SVETNIK, V.; LIAW, A.; TONG, C. Variable selection in random forest with application to quantitative structure-activity relationship. **Proceedings of the 7th Course on Ensemble Methods for Learning Machines**, 2004.

SZATMÁRI, G.; PIRKÓ, B.; KOÓS, S.; LABORCZI, A.; BAKACSI, Z.; SZABÓ, J.; PÁSZTOR, L. Spatio-temporal assessment of topsoil organic carbon stock change in Hungary. **Soil and Tillage Research**, v. 195, p. 104410, 2019.

TAGHIZADEH-MEHRJARDI, R.; NABIOLLAHI, K.; MINASNY, B.; TRIANTAFILIS, J. Comparing data mining classifiers to predict spatial

distribution of USDA-family soil groups in Baneh region, Iran. **Geoderma**, v. 253, p. 67-77, 2015.

TEDESCO, M.J.; GIANELLO, C.; BISSANI, C.A.; BOHNEN, H.; VOLKWEISS, S.J. *Análise de solo, plantas e outros materiais*. 2.ed. Porto Alegre, Universidade Federal do Rio Grande do Sul, 1995. 174p.

TEIXEIRA, P. C. *et al.* **Manual de métodos de análise de solo**. Brasília: Embrapa, p. 573, 2017.

THOMAS, G. W. Soil pH and soil acidity. **Methods of soil analysis: part 3 chemical methods**, v. 5, p. 475-490, 1996.

THOMAS, G. W.; HARGROVE, W. L. The chemistry of soil acidity. **Soil acidity and liming**, v. 12, p. 3-56, 1984.

TOMAZELLI, L. J.; DILLENBURG, S. R.; VILLWOCK, J. A. Late Quaternary geological history of Rio Grande do Sul coastal plain, southern Brazil. **Brazilian Journal of Geology**, v. 30, n. 3, p. 474-476, 2000.

TOMAZELLI, L. J.; VILLWOCK, J. A.; DILLENBURG, S. R.; BACHI, F. A.; DEHNHARDT, B. A. Significance of present-day coastal erosion and marine transgression, Rio Grande do Sul, southern Brazil. **Anais-Academia Brasileira de Ciências**, v. 70, p. 221-230, 1998.

TOPA, D.; CARA, I. G.; JIȚĂREANU, G. Long term impact of different tillage systems on carbon pools and stocks, soil bulk density, aggregation and nutrients: A field meta-analysis. **Catena**, v. 199, p. 105102, 2021.

VASQUES, G. M.; GRUNWALD, S.; COMERFORD, N. B.; SICKMAN, J. O. Regional modelling of soil carbon at multiple depths within a subtropical watershed. **Geoderma**, v. 156, n. 3-4, p. 326-336, 2010.

VASUNDHARA, R. *et al.* Modeling Soil pH at regional scale using environmental covariates and machine learning algorithm. **Environmental Monitoring and Assessment**, v. 197, n. 7, p. 799, 2025.

VAYSSE, K.; HEUVELINK, G. B. M.; LAGACHERIE, P. Spatial aggregation of soil property predictions in support of local land management. **Soil Use and Management**, v. 33, n. 2, p. 299-310, 2017.

VELDKAMP, E. Organic carbon turnover in three tropical soils under pasture after deforestation. **Soil Science Society of America Journal**, v. 58, n. 1, p. 175-180, 1994.

VILLWOCK, J. A. Geology of the Coastal Province of Rio Grande do Sul, Southern Brazil. A Synthesis. **Pesquisas em Geociências**, v. 16, n. 16, p. 5-49, 1984.

WADOUX, A. M. C.; HEUVELINK, G. B.; DE BRUIN, S.; BRUS, D. J. Spatial cross-validation is not the right way to evaluate map accuracy. **Ecological Modelling**, v. 457, p. 109692, 2021.

WADOUX, A. M. J. C.; BRUS, D. J.; HEUVELINK, G.B.M. Sampling design optimization for soil mapping with random forest. **Geoderma**, v. 355, p. 113913, 2019.

WANG, B.; WATERS, C.; ORGILL, S.; COWIE, A.; CLARK, A.; LI LIU, D.; SIMPSON, M.; MCGOWEN, I.; SIDES, T. Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia. **Ecological indicators**, v. 88, p. 425-438, 2018.

WANG, F.; LU, X.; SANDERS, C. J.; TANG, J. Tidal wetland resilience to sea level rise increases their carbon sequestration capacity in United States. **Nature Communications**, v. 10, n. 1, p. 5434, 2019.

WARRICK, A.W.; NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D. (Ed.). Applications of soil physics. New York: Academic Press, 1980.

WREGE, M. S.; STEINMETZ, S.; REISSER JUNIOR, C.; ALMEIDA, I. R. **Atlas climático da região sul do Brasil: estados do Paraná, Santa Catarina e Rio Grande do Sul**. 2ª Ed. Pelotas: Embrapa Clima Temperado; Colombo: Embrapa Florestas, 2012.

WYNER, A. J.; OLSON, M.; BLEICH, J.; MEASE, D. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, v. 18, n. 48, p. 1-33, 2017.

XIA, X.; YANG, Z.; LIAO, Y.; CUI, Y.; LI, Y. Temporal variation of soil carbon stock and its controlling factors over the last two decades on the southern Song-nen Plain, Heilongjiang Province. **Geoscience Frontiers**, v. 1, n. 1, p. 125-132, 2010.

XIONG, X.; GRUNWALD, S.; MYERS, D. B.; KIM, J.; HARRIS, W. G.; COMERFORD, N. B. Holistic environmental soil-landscape modeling of soil organic carbon. **Environmental Modelling & Software**, v. 57, p. 202-215, 2014.

YANG, L. *et al.* Evaluation of conditioned Latin hypercube sampling for soil mapping based on a machine learning method. **Geoderma**, v. 369, p. 114337, 2020.

ZERAATPISHEH, M.; GAROSI, Y.; OWLIAIE, H. R.; AYOUBI, S.; TAGHIZADEH-MEHRJARDI, R.; SCHOLTEN, T.; XU, M. Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. **Catena**, v. 208, p. 105723, 2022.

ZHANG, X. *et al.* Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. **Geoderma**, v. 432, p. 116383, 2023.

ZHANG, Y.; SAURETTE, D. D.; EASHER, T. H.; JI, W.; ADAMCHUK, V. I.; BISWAS, A. Comparison of sampling designs for calibrating digital soil maps at multiple depths. **Pedosphere**, v. 32, n. 4, p. 588-601, 2022.

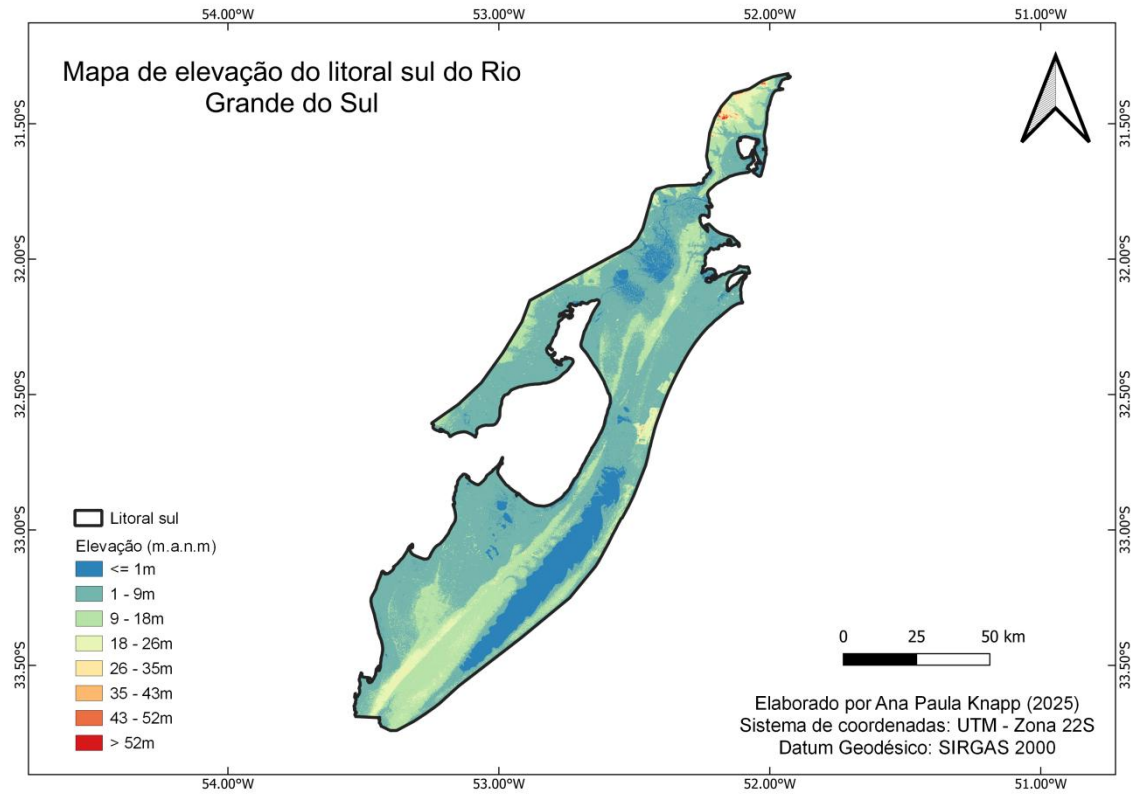
ZHAO, Z. D.; ZHAO, M. S.; LU, H. L.; WANG, S. H.; LU, Y. Y. Digital mapping of soil pH based on machine learning combined with feature selection methods in east China. **Sustainability**, v. 15, n. 17, p. 12874, 2023.

ŽÍŽALA, D. *et al.* High-resolution agriculture soil property maps from digital soil mapping methods, Czech Republic. **Catena**, v. 212, p. 106024, 2022.

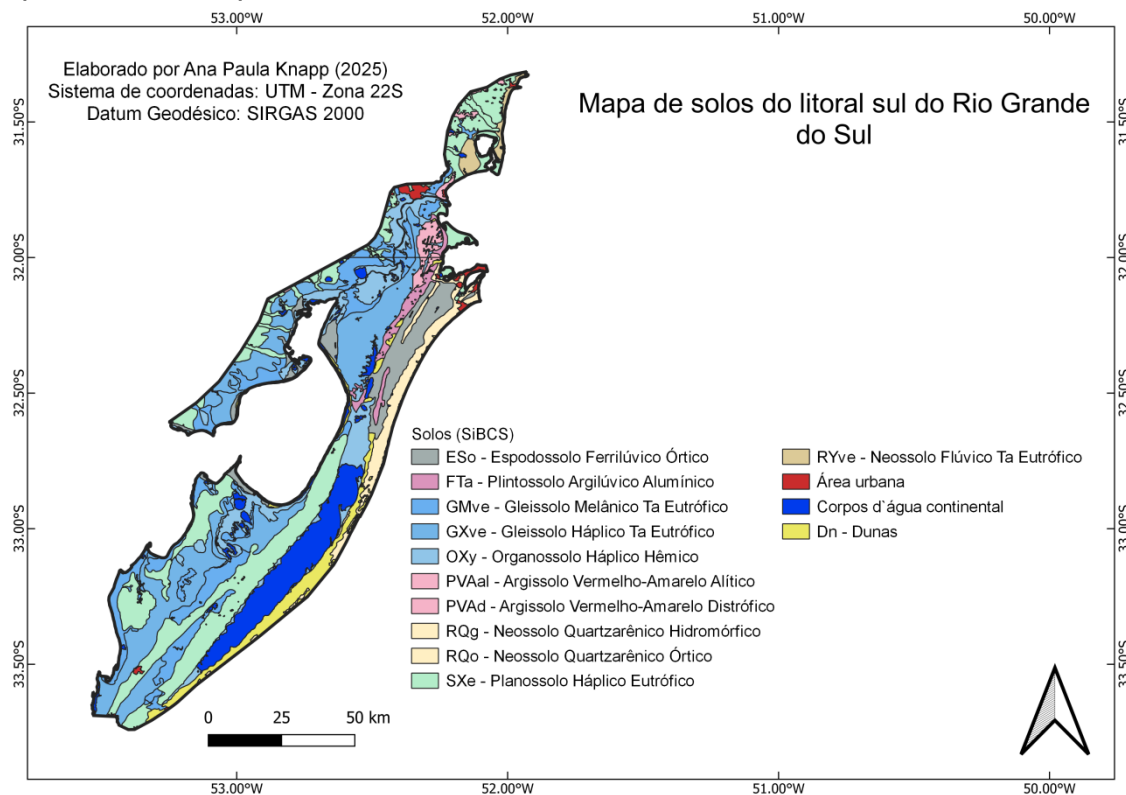
ZOLFAGHARI NIA, M.; MORADI, M.; MORADI, G.; TAGHIZADEH-MEHRJARDI, R. Machine learning models for prediction of soil properties in the Riparian forests. **Land**, v. 12, n. 1, p. 32, 2022.

Apêndices

Apêndice A - Mapa de elevação da área de estudo

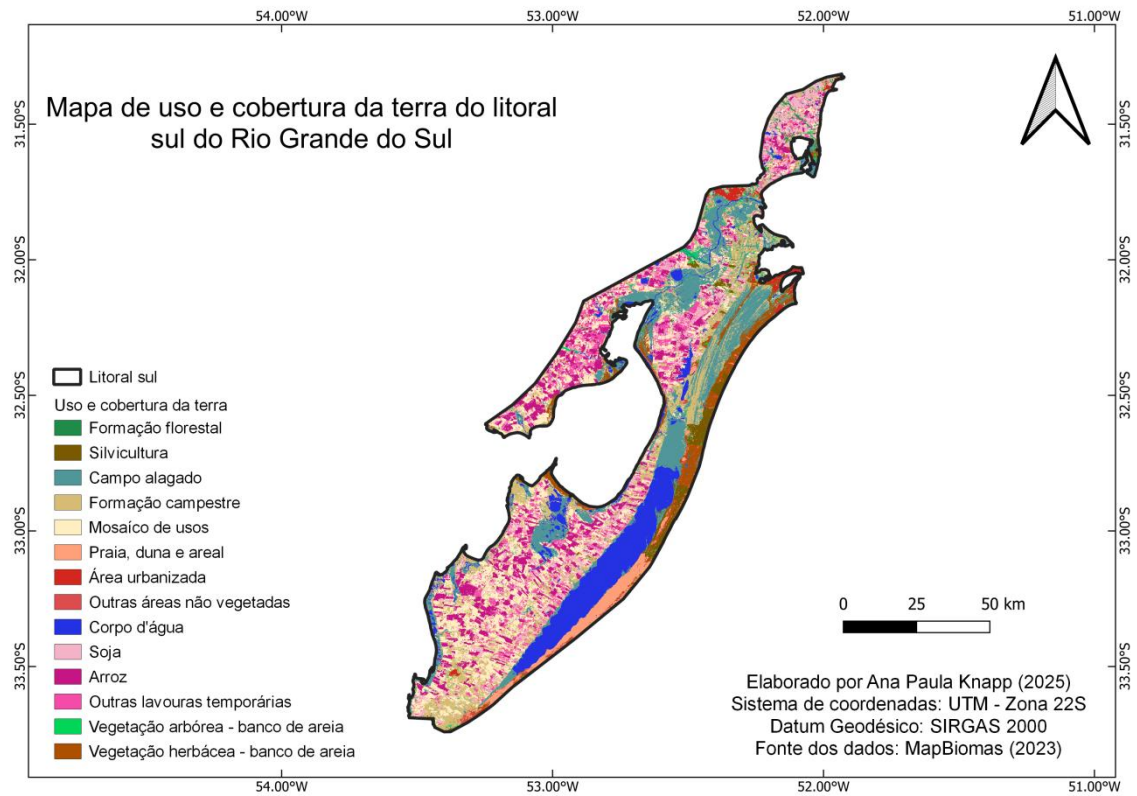


Apêndice B - Mapa de solos da área de estudo



Fonte primária dos dados: BRASIL (1986)

Apêndice C - Mapa de uso e cobertura da terra da área de estudo



Apêndice D - Resumo das informações de campo dos pontos coletados na área 1

Ponto	Relevo	Uso atual	Classificação SiBCS (Santos <i>et al.</i>, 2025)
P3	Plano	Campo	Planossolo Háplico
P4	Suave ondulado	Campo	Planossolo Háplico
P5	Suave ondulado	Agricultura – pousio	Planossolo Háplico
P6	Suave ondulado	Agricultura – pousio	Planossolo Háplico
P7	Suave ondulado	Silvicultura	Planossolo Háplico
P8	Suave ondulado	Campo	Planossolo Háplico
P9	Suave ondulado	Campo – melhorado	Argissolo Vermelho- Amarelo
P10	Suave ondulado	ILP	Planossolo Háplico
P11	Plano	Campo	Planossolo Háplico
P12	Suave ondulado	Campo	Argissolo Vermelho- Amarelo
P13	Suave ondulado	Campo	Planossolo Háplico

Apêndice E - Resumo das informações de campo dos pontos coletados na área
2

Ponto	Relevo	Uso atual	Classificação SiBCS (Santos <i>et al.</i>, 2025)
P1	Plano	Campo	Planossolo Háplico
P14	Plano	Campo	Espodossolo
P15	Plano	Campo	Planossolo Háplico
P16	Plano	Campo	Planossolo Háplico
P17	Plano	Agricultura – pousio	Argissolo Vermelho- Amarelo
P18	Plano	Agricultura – arroz	Argissolo Vermelho- Amarelo
P19	Plano	Agricultura – arroz	Argissolo Vermelho- Amarelo
P20	Plano	Agricultura	Argissolo Vermelho- Amarelo
P21	Plano	Campo	Argissolo Vermelho- Amarelo
P22	Plano	Campo	Neossolo Quartzarênico
P23	Plano	Campo	Neossolo Quartzarênico
P24	Plano	Campo	Planossolo Háplico
P25	Plano	Agricultura – soja	Planossolo Háplico
P26	Plano	Campo	Planossolo Háplico

Apêndice F - Resumo das informações de campo dos pontos coletados na área
3

Ponto	Relevo	Uso atual	Classificação SiBCS (Santos <i>et al.</i>, 2025)
P27	Plano	Campo	Planossolo Háplico
P28	Plano	Campo	Planossolo Háplico
P29	Plano	Agricultura – pousio	Planossolo Háplico
P30	Plano	Agricultura – soja	Planossolo Háplico
P31	Plano	Agricultura – soja	Planossolo Háplico
P32	Suave ondulado	Campo	Planossolo Háplico
P33	Suave ondulado	Agricultura	Planossolo Háplico
P34	Ondulado	Campo	Argissolo Vermelho- Amarelo
P35	Plano	Campo	Neossolo Quartzarênico
P36	Plano	Campo	Neossolo Quartzarênico
P37	Plano	Campo	Neossolo Quartzarênico
P38	Plano	Campo	Neossolo Quartzarênico
P39	Plano	Campo	Neossolo Quartzarênico

Apêndice G - Resumo das informações de campo dos pontos coletados na área 4

Ponto	Relevo	Uso atual	Classificação SiBCS (Santos <i>et al.</i>, 2025)
P40	Plano	Agricultura – soja	Argissolo Amarelo
P41	Plano	Agricultura – pousio	Planossolo Háplico
P42	Plano	Agricultura	Planossolo Háplico
P43	Plano	Silvicultura	Planossolo Háplico
P44	Plano	Agricultura – soja	Planossolo Háplico
P45	Plano	Campo	Argissolo Vermelho- Amarelo
P46	Plano	Campo	Planossolo Háplico
P47	Plano	Campo	Planossolo Háplico
P48	Plano	Agricultura	Chernossolo Argilúvico
P49	Plano	Agricultura	Argissolo Vermelho- Amarelo
P56	Plano	Campo	Planossolo Háplico
P57	Plano	Agricultura – soja	Argissolo Vermelho- Amarelo
P71	Plano	Campo	Argissolo Vermelho- Amarelo

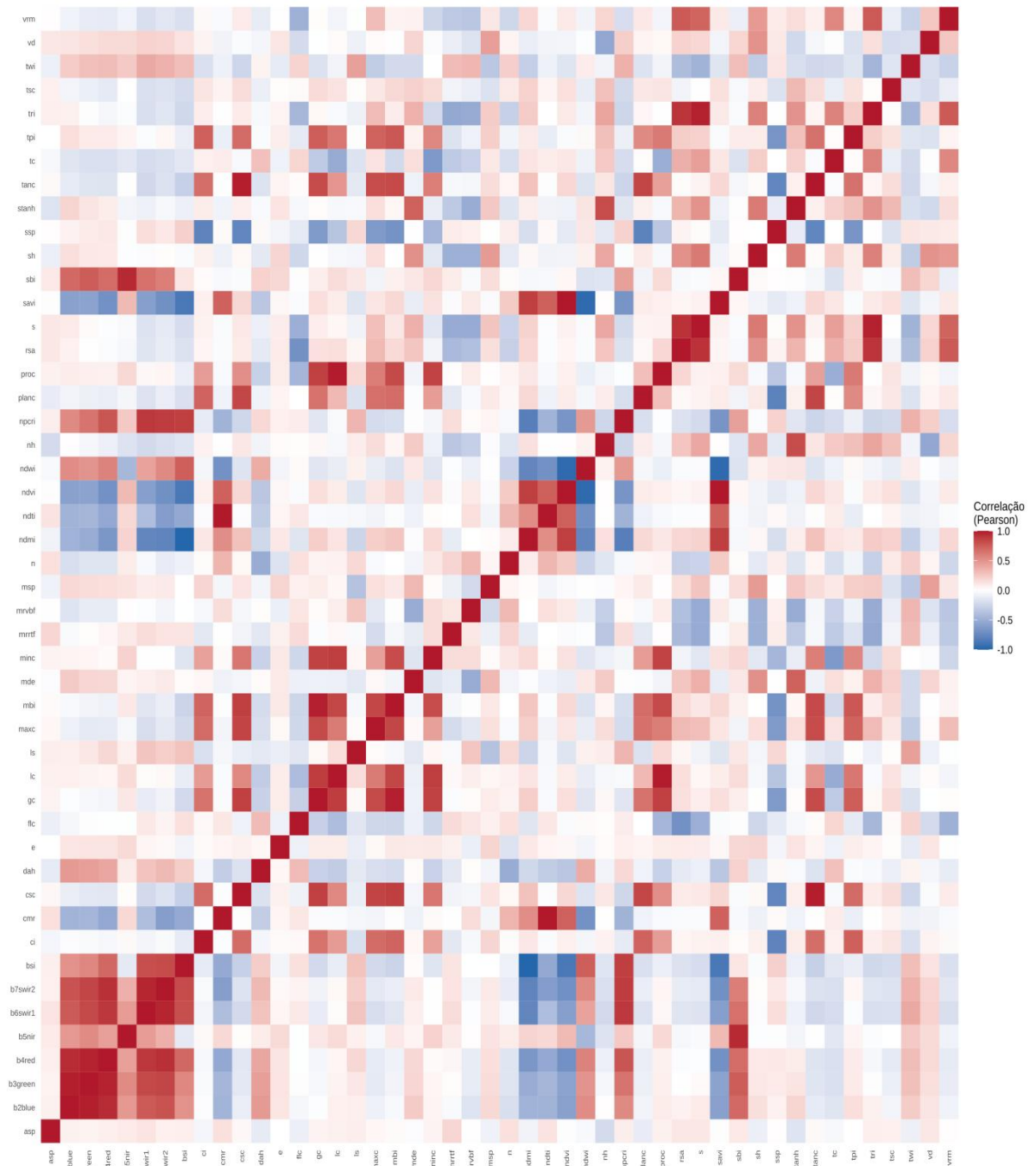
Apêndice H – Resumo das informações de campo dos pontos coletados na área 5

Ponto	Relevo	Uso atual	Classificação SiBCS (Santos <i>et al.</i>, 2025)
P58	Plano	Agricultura – soja	Argissolo Vermelho-Amarelo
P59	Plano	Agricultura – arroz	Argissolo Vermelho-Amarelo
P60	Plano	Campo	Planossolo Háplico
P61	Plano	Campo	Planossolo Háplico
P62	Plano	Campo	Gleissolo Háplico
P63	Plano	Campo	Argissolo Vermelho-Amarelo
P64	Plano	Campo	Argissolo Vermelho-Amarelo
P65	Plano	Campo	Argissolo Vermelho-Amarelo
P66	Plano	Pastagem	Argissolo Vermelho-Amarelo
P67	Plano	Pastagem	Argissolo Vermelho-Amarelo
P68	Plano	Campo	Argissolo Vermelho-Amarelo
P69	Plano	Campo	Planossolo Háplico
P70	Suave ondulado	Campo	Argissolo Vermelho-Amarelo

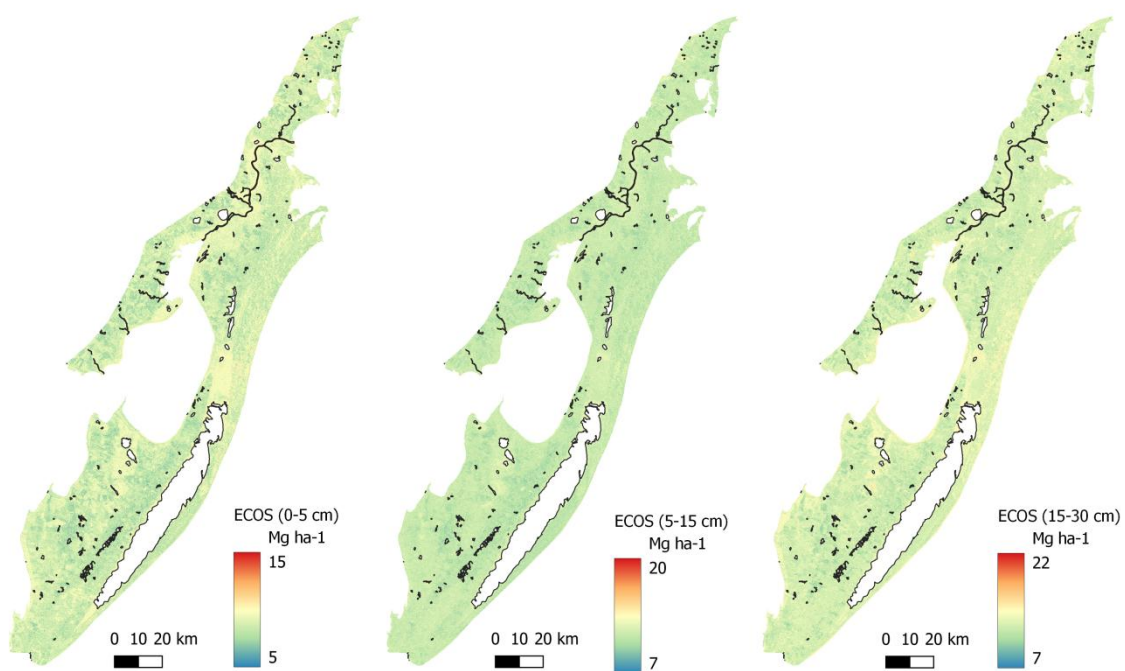
Apêndice I - Fluxograma metodológico detalhado do capítulo 1



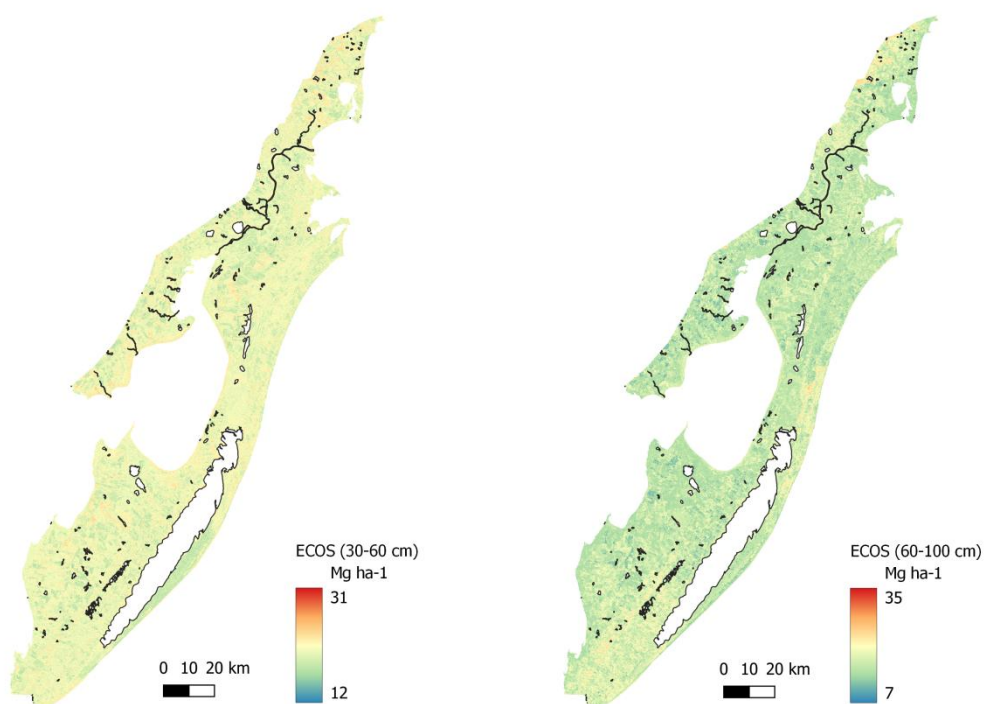
Apêndice J – Heatmap da correlação entre as covariáveis ambientais



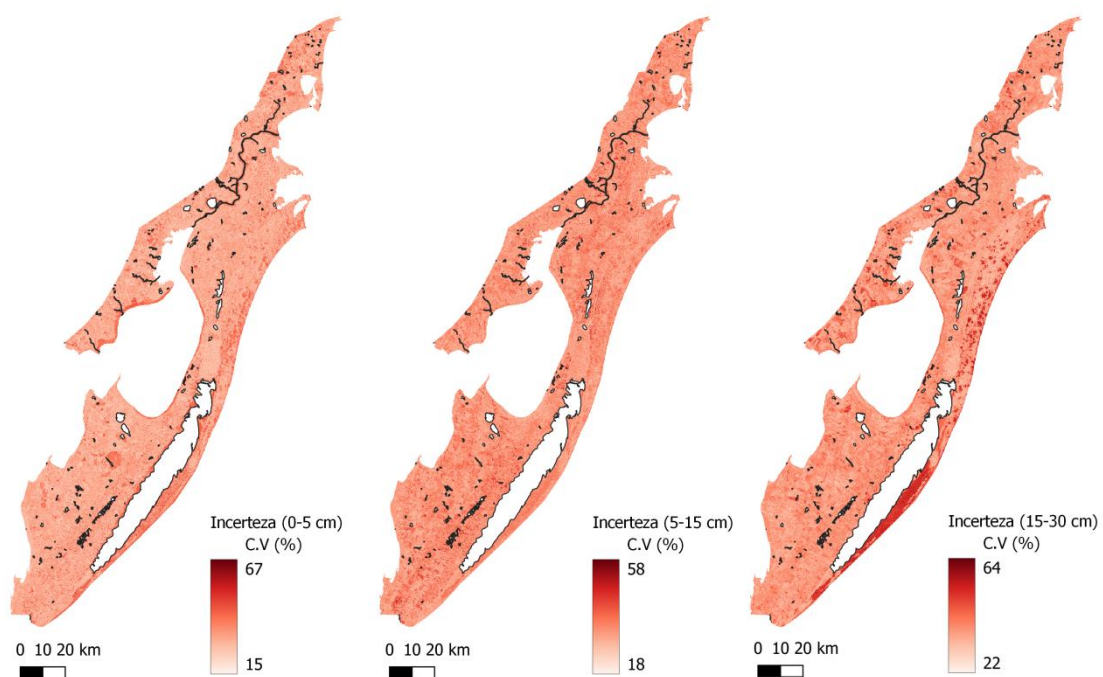
Apêndice K - Espacialização do ECOS nas camadas de 0-5, 5-15 e 15-30 cm



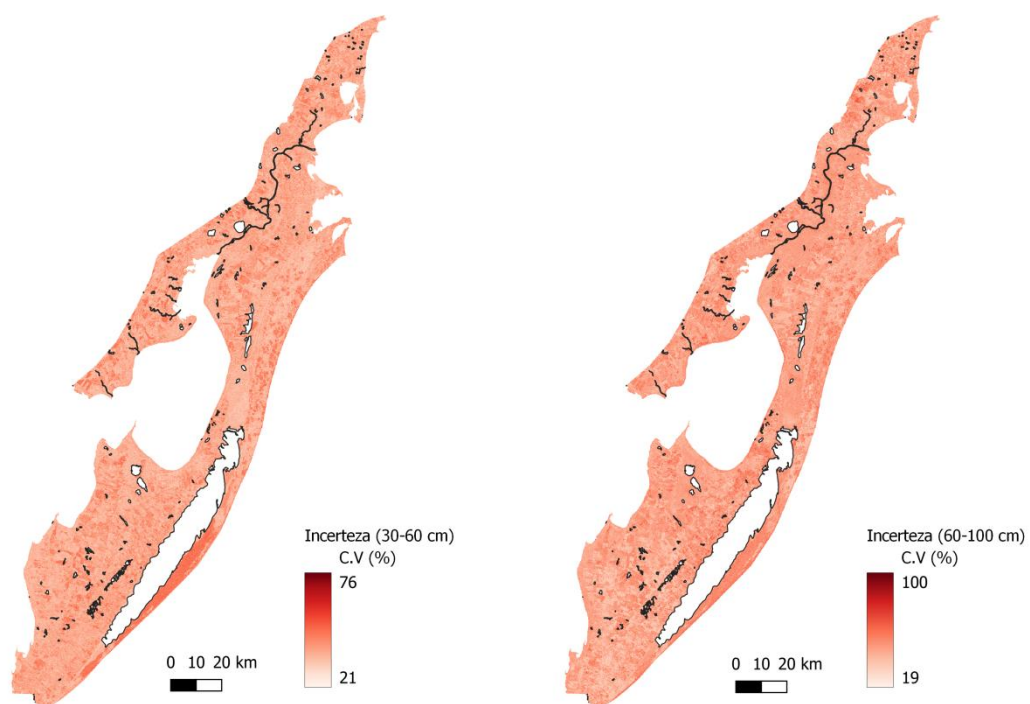
Apêndice L - Espacialização do ECOS nas camadas de 30-60 e 60-100 cm



Apêndice M - Espacialização da incerteza nas camadas de 0-5, 5-15 e 15-30 cm



Apêndice N - Espacialização da incerteza nas camadas de 30-60 e 60-100 cm



Apêndice O - Fluxograma geral do mapeamento digital de pH, areia e densidade do solo (BD) em cinco profundidades, com três abordagens de seleção de covariáveis e validação cruzada espacial



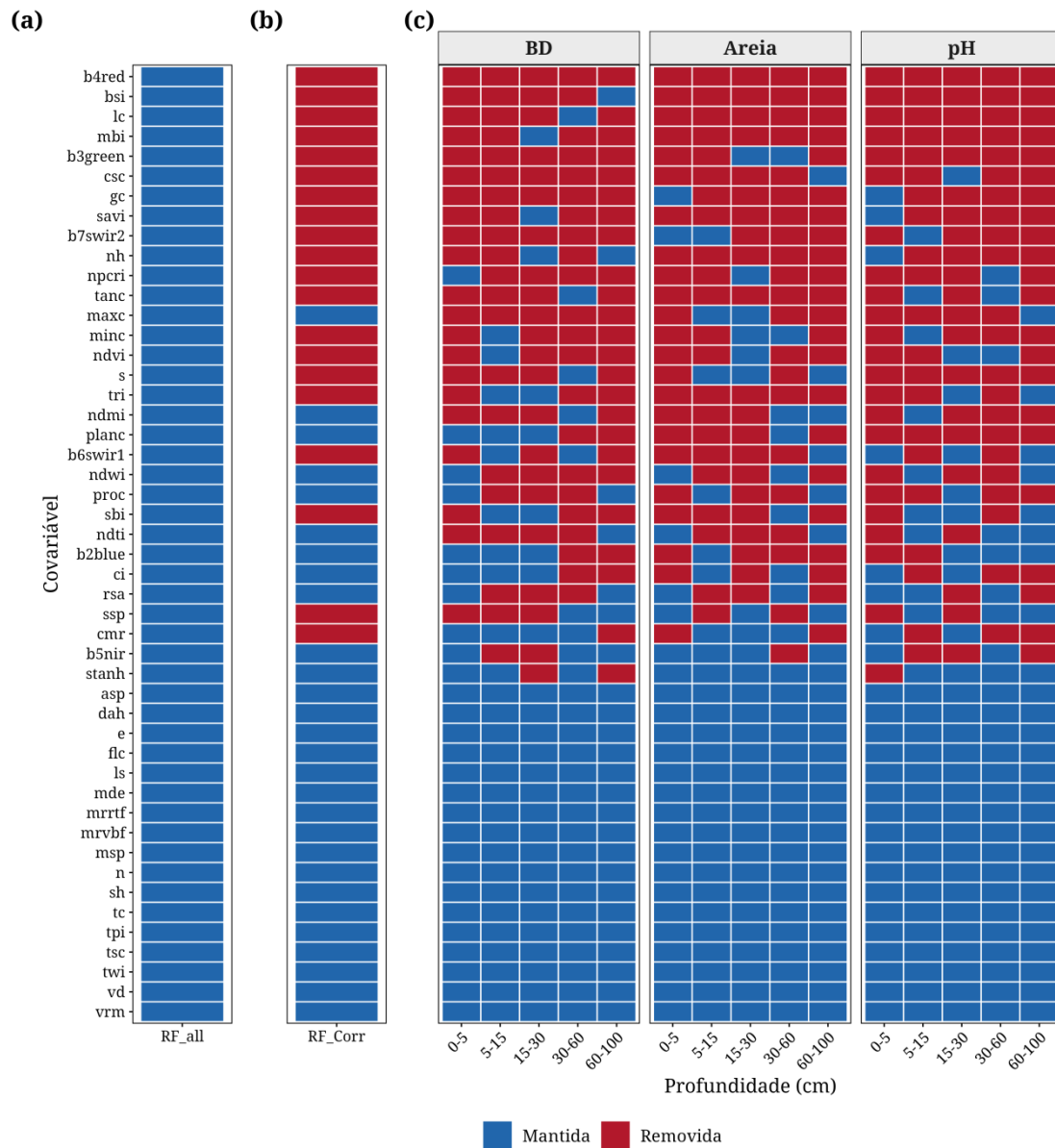
Apêndice P - Estatísticas descritivas de densidade do solo, teor de areia e pH em água por profundidade (n = 64 por profundidade)

Densidade do solo — BD (g cm⁻³)					
Estatística	0–5 cm	5–15 cm	15–30 cm	30–60 cm	60–100 cm
Média	1,40	1,56	1,60	1,63	1,59
Mediana	1,44	1,55	1,59	1,62	1,57
Mínimo	0,97	1,26	1,15	1,35	1,35
Máximo	1,78	1,82	1,99	1,91	1,84
Desvio padrão	0,20	0,13	0,16	0,11	0,12
CV (%)	13,9	8,4	10,1	7,0	7,7

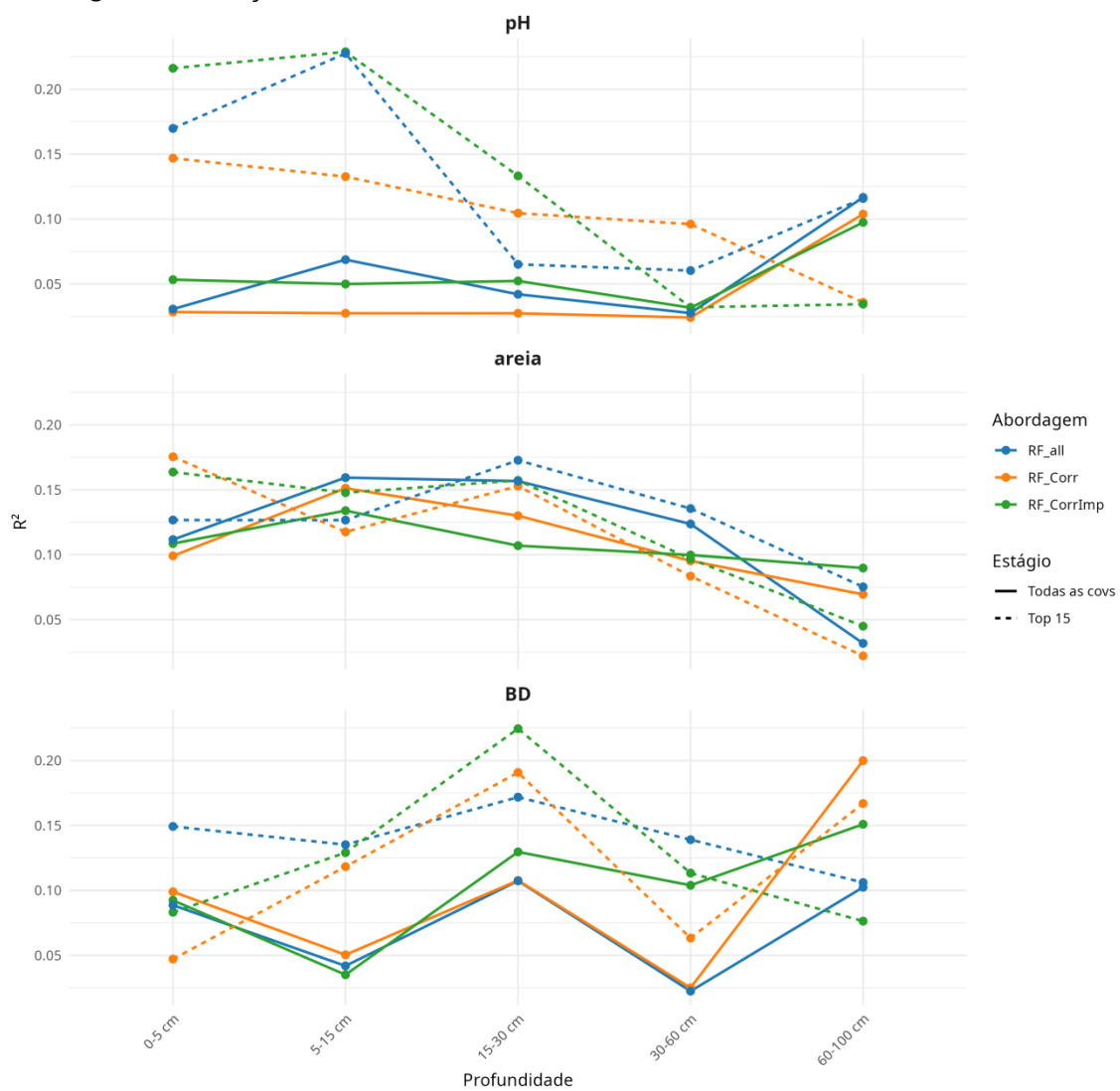
Teor de areia (%)					
Estatística	0–5 cm	5–15 cm	15–30 cm	30–60 cm	60–100 cm
Média	71,5	73,3	71,7	67,1	63,0
Mediana	73,5	75,0	75,0	64,0	59,0
Mínimo	28,0	30,0	26,0	20,0	16,0
Máximo	96,0	96,0	97,0	99,0	98,0
Desvio padrão	17,0	16,6	17,1	19,8	21,0
CV (%)	23,8	22,6	23,8	29,5	33,3

pH em água					
Estatística	0–5 cm	5–15 cm	15–30 cm	30–60 cm	60–100 cm
Média	5,48	5,51	5,6	5,76	5,97
Mediana	5,37	5,43	5,46	5,71	5,76
Mínimo	4,53	4,46	4,52	4,81	4,76
Máximo	7,67	7,8	7,94	7,67	8,13
Desvio padrão	0,60	0,65	0,67	0,66	0,78
CV (%)	10,9	11,7	11,9	11,5	13,1

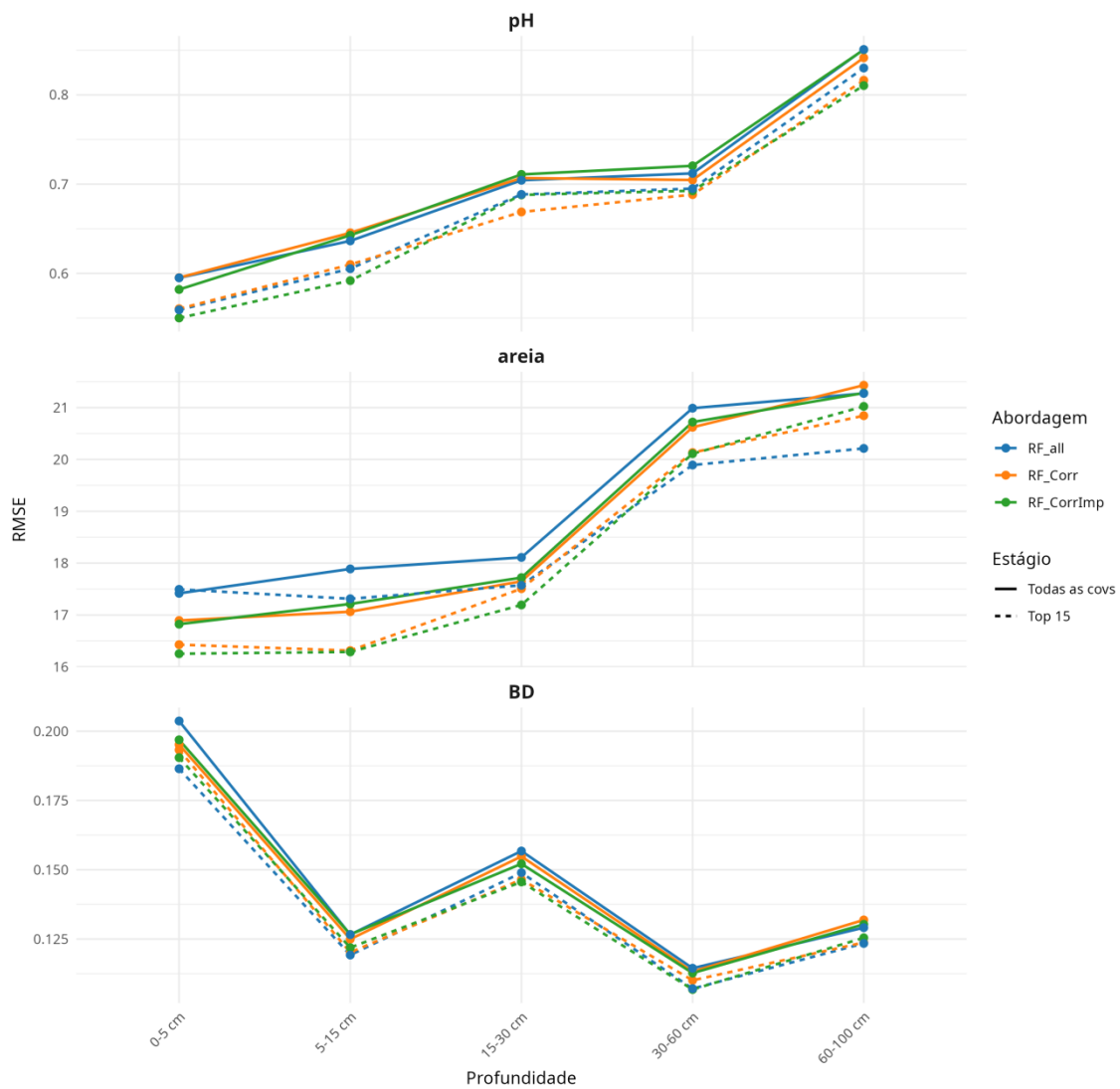
Apêndice Q – Lista de covariáveis mantidas e removidas para as três abordagens de seleção: a) RF_all, b) RF_Corr e c) RF_CorrImp; nas cinco profundidades e para os três atributos avaliados



Apêndice R - Variação do coeficiente de determinação (R^2) dos modelos Random Forest por atributo e profundidade, para cada abordagem de seleção e estágio de redução de covariáveis



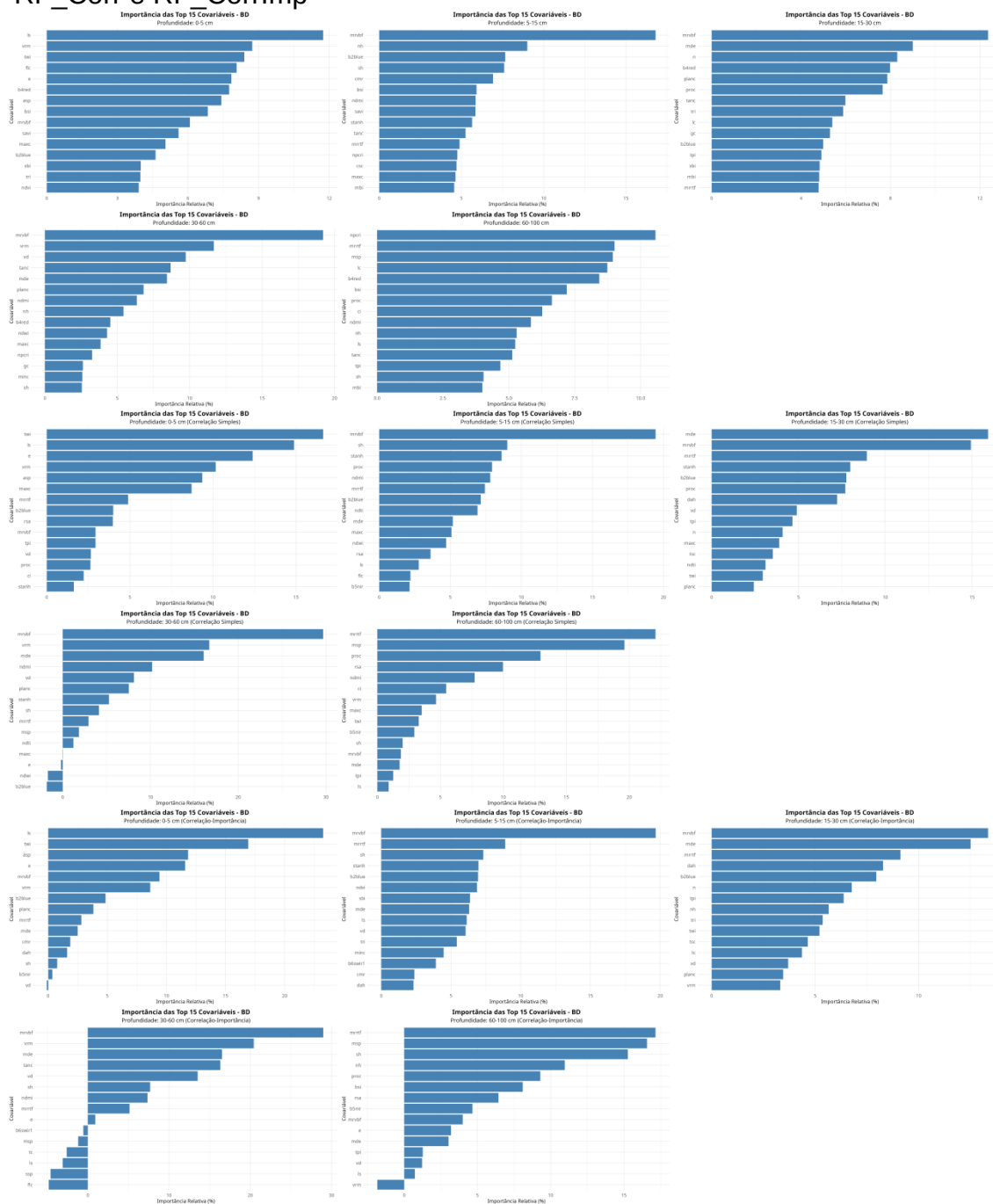
Apêndice S - Variação do RMSE dos modelos Random Forest por atributo e profundidade, para cada abordagem de seleção e estágio de redução de covariáveis



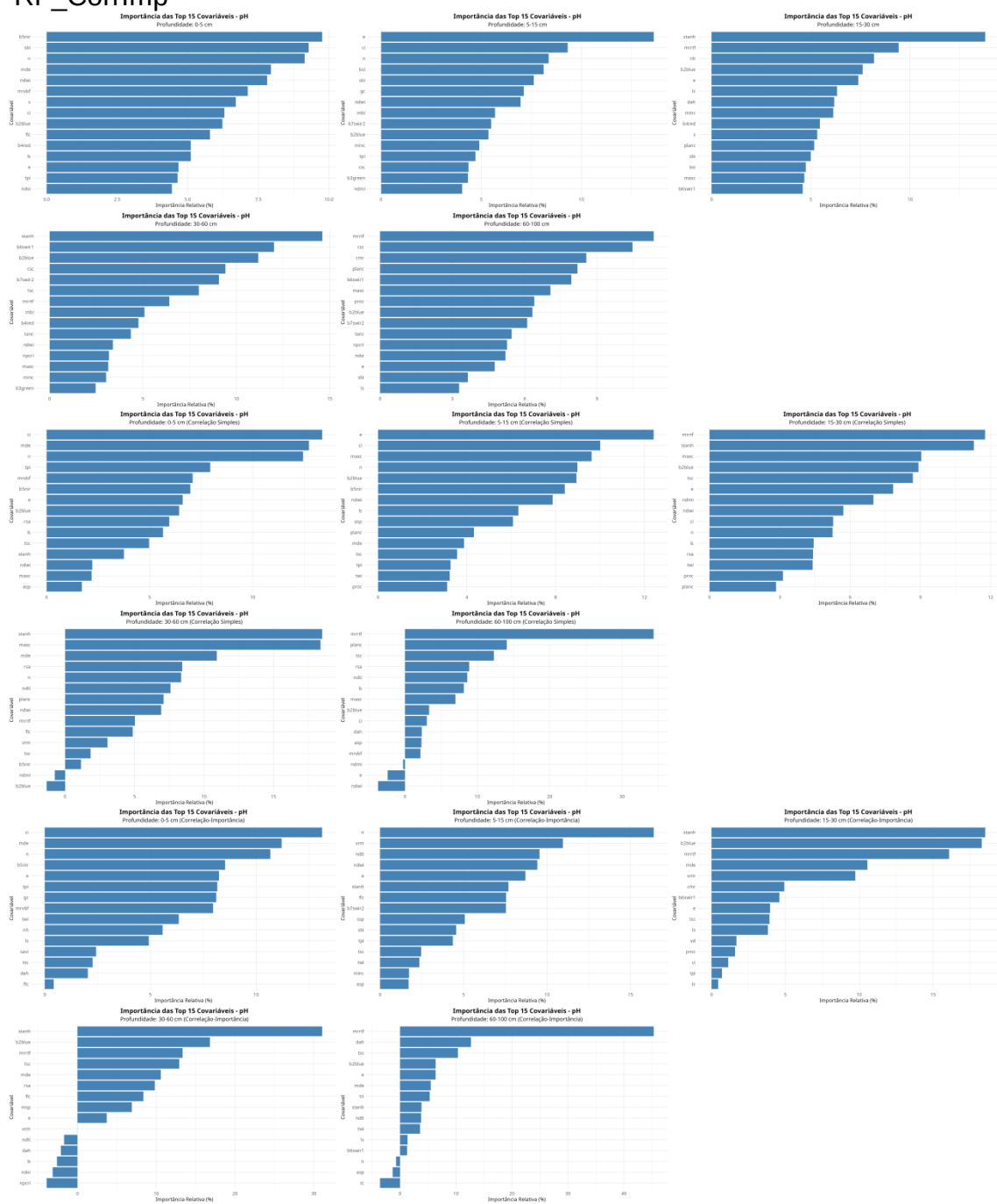
Apêndice T - Importância relativa das 15 covariáveis mais importantes para o teor de areia nas cinco profundidades, segundo as abordagens RF_all, RF_Corr e RF_CorrImp



Apêndice U - Importância relativa das 15 covariáveis mais importantes para a densidade do solo nas cinco profundidades, segundo as abordagens RF_all, RF_Corr e RF_CorrImp



Apêndice V - Importância relativa das 15 covariáveis mais importantes para o pH nas cinco profundidades, segundo as abordagens RF_all, RF_Corr e RF_CorrImp



DECLARAÇÃO DE USO DE FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

Declaro, para os devidos fins, que foram utilizadas ferramentas de Inteligência Artificial Generativa, especificamente os modelos ChatGPT (GPT-5, desenvolvido pela OpenAI), SWE (SWE-1.7, desenvolvido pela Cognition), Claude (Claude Sonnet 5, desenvolvido pela Anthropic) e Gemini (Gemini 3 Flash, desenvolvido pelo Google), de forma exclusivamente auxiliar, com os seguintes propósitos:

- Apoio na revisão linguística e ortográfica do texto;
- Apoio na escrita e refatoração de scripts em R usados nas análises (mediado pela IDE - interface de desenvolvimento integrado - Devin, desenvolvida pela empresa Cognition);
- Tradução assistida do resumo (abstract) para o inglês;
- Busca pontual de referências bibliográficas sobre conceitos específicos, sempre acompanhada de verificação manual da fonte primária.

Ressalto que todas as análises, interpretações, conclusões e formulações científicas contidas neste trabalho são de autoria exclusiva da pesquisadora, e que não houve geração automatizada de conteúdo inédito, de dados empíricos, nem de resultados por meio de IA.

Esta declaração tem por objetivo garantir transparência e integridade acadêmica, em conformidade com as recomendações da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) sobre o uso responsável de tecnologias de Inteligência Artificial em pesquisas científicas.

Pelotas, agosto de 2026.

Ana Paula Knapp
Autora da Tese