UNIVERSIDADE ESTADUAL PAULISTA (UNESP)
"JÚLIO DE MESQUITA FILHO"

Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal
Departamento de Ciências Exatas
(FCAV - DCE)

Relatório Final
Triênio 2021-2023

MODELOS DE APRENDIZADO DE MÁQUINA PARA A EMISSÃO DE CO_2 DO SOLO EM ÁREAS AGRÍCOLAS NO BRASIL

Pesquisador responsável: Prof. Dr. Alan Rodrigo Panosso
Departamento de Ciências Exatas

Jaboticabal – SP – Brasil
Abril de 2024

SUMÁRIO

IDENTIFICAÇÃO DA PROPOSTA

Título

MODELOS DE APRENDIZADO DE MÁQUINA PARA A EMISSÃO DE CO_2 DO SOLO EM ÁREAS AGRÍCOLAS NO BRASIL

Proponente (Coordenador do Projeto)

Prof. Assistente Doutor Alan Rodrigo Panosso; CPF 290.833.258-22.

Professor MS-3.1, contrato desde março de 2013.

E-mail: alan.panosso@unesp.br

Tel: (16) 3209-7210

Nome da Instituição

Universidade Estadual Paulista (UNESP) "Júlio de Mesquita Filho", Faculdade de Ciências Agrárias e Veterinárias (FCAV), Campus de Jaboticabal, SP.

Local de desenvolvimento da proposta

Departamento de Engenharia e Ciências Exatas (antigo Ciências Exatas) da Faculdade de Ciências Agrárias e Veterinárias de Jaboticabal.

Comitê assessor

AG – Agronomia.

Área de conhecimento da chamada CNPq Nº 09/2020

Manejo e Conservação do Solo

Pesquisadores Associados (colaboradores)

Prof. Dr. Newton La Scala Júnior; Professor Titular; Departamento de Engenharia e Ciências Exatas-UNESP/FCAV Jaboticabal - SP; E-mail: la.scala@unesp.br; Especialidade: Aspectos diversos da emissão de CO_2 do solo em áreas agrícolas.

Prof. Dr. Glauco de Souza Rolim; Professor Adjunto; Departamento de Engenharia e Ciências Exatas-UNESP/FCAV Jaboticabal - SP; E-mail: glauco.rolim@unesp.br; Especialidade: Agrometeorologia, Modelagem na Agricultura e Ciência dos Dados.

Prof. Dr. Fábio Roberto Chavarette; Professor Adjunto; Departamento de Engenharia, Física e Matemática-UNESP/IQ Araraquara - SP; E-mail: fabio.chavarette@unesp.br; Especialidade: Simulações numéricas, Suporte computacional, Interfaces e Criação de Bancos de Dados.

Dr. Falberni de Souza Costa; Pesquisador; Centro de Pesquisa Agroflorestal do Acre, Embrapa Acre, Rio Branco - AC; E-mail: falberni.costa@embrapa.br; Especialidade:

Manejo e conservação de água e solo em agroecossistemas tropicais, com ênfase em sistemas conservacionistas de manejo do solo e fluxo de gases de efeito estufa.

Prof. Dr. Rafael Montanari; Professor Adjunto; Departamento de Fitossanidade, Engenharia Rural e Solos-UNESP/FE Ilha Solteira - SP; E-mail: r.montanari@unesp.br; Especialidade: Geoestatística, Amostragem, Atributos Físicos e Químicos do solo e Agricultura Sustentável.

Prof. Dr. Milton César Costa Campos; Professor Associado II; Universidade Federal do Amazonas, UFAM Campus Vale do Rio Madeira, Humaitá - AM; E-mail: mcesarsolos@gmail.com; Especialidade: Gênese e Morfologia do Solo; Mineralogia do Solo e Relação Solo-Paisagem.

Metas atingidas (311981/2020-8)

O presente projeto tem como hipótese básica que a dinâmica espaçotemporal da emissão de CO_2 do solo (FCO_2) pode ser descrita como um fenômeno com estruturas multidimensionais que evoluem no tempo, aptas a serem modeladas por algoritmos de aprendizado de máquina e aprendizado profundo. Como principais resultados, temos que a modelagem de FCO_2 por meio de algoritmos de aprendizado de máquina pode contribuir significativamente para reduzir as incertezas associadas às fontes e sumidouros de carbono no solo. Os modelos de aprendizado de máquina permitiram avaliações explícitas e rigorosas usando vários indicadores de erro e avaliando as incertezas associadas ao ciclo global do carbono. Essas informações podem subsidiar cientificamente a formulação de planos estratégicos que aproveitem a agricultura de baixo carbono para mitigar e adaptar-se às mudanças climáticas globais. Além disso, devemos salientar que modelos de visão computacional, utilizando redes neurais artificiais, não apresentaram melhores desempenhos preditivos quando comparados a algoritmos baseados em árvores de decisão como florestas aleatórias (*Random Forest)* e *XGBoost*, sendo esses mais simples, de fácil implementação e *tuning* de hiperparâmetros. O modelo *Random Forest* foi o mais preciso e robusto para prever FCO_2, incluindo sua distribuição espacial em microescalas.

Em adição, informamos os seguintes quantitativos relativos à proposta anterior:

a) 04 trabalhos de iniciação científica concluídas (acadêmicos Letícia Roberta de Lima, Lucas de Oliveira Gonçalves, Vinicius De Sousa Crispim e Luís Miguel da Costa);

b) 06 dissertações de mestrado concluídas (acadêmicos Angélica da Silva, Laís de Souza Teixeira e Leandro Alves Pinto, como orientador principal e os acadêmicos Arianis Ibeth Santos Nicolella, Marcelo Laranjeira Pimentel e Hugo Guiné Pinto Ferreira como co-orientador);

c) 06 teses de doutorado concluídas (acadêmicas Deise Cristina Santos Nogueira, Maria Elisa Vicentini e Ludhanna Marinho Veras, como orientador principal e os acadêmicos Adriano Maltezo da Rocha, Paulo Alexandre da Silva e Roque Flôr dos Santos Júnior como co-orientador);

d) 01 dissertação de mestrado em andamento (acadêmica Renata Amaral da Silva);

e) 01 tese de doutorado em andamento (acadêmico e Kleve Freddy Ferreira Canteral);

f) 07 artigos científicos publicados em periódicos associados à proposta (até o momento):

i) da Costa, L. M.; de Araújo Santos, G. A.; de Mendonça, G. C.; de Souza Maria, L.; da Silva JR, C. A.; Panosso, A. R.; La Scala JR, N. Exploring CO_2 anomalies in Brazilian biomes combining OCO-

2 & 3 data: Linkages to wildfires patterns. Advances in Space Research, v. 73, n. 8, p. 4158-4174, 2024.

ii) Vicentini, M. E.; da Silva, P. A.; Canteral, K. F. F.; de Lucena, W. B.; de Moraes, M. L. T.; Montanari, R.; Filho, M. C. M. T.; Peruzzi, N. J.; La Scala, N.; de Souza Rolim, G.; Panosso, A. R. Artificial neural networks and adaptive neuro-fuzzy inference systems for prediction of soil respiration in forested areas southern Brazil. Environmental Monitoring and Assessment, v. 195, n. 9, p. 1074, 2023.

iii) Canteral, K. F. F.; Vicentini, M. E.; de Lucena, W. B.; de Moraes, M. L. T.; Montanari, R.; Ferraudo, A. S.; Peruzzi, N. J.; La Scala, N.; Panosso, A. R. Machine learning for prediction of soil CO_2 emission in tropical forests in the Brazilian Cerrado. Environmental Science and Pollution Research, v., n., p., 2023;

iv) de Souza Maria, L.; Rossi, F. S.; Costa, L. M. D.; Campos, M. O.; Blas, J. C. G.; Panosso, A. R.; Silva, J. L. D.; Silva Junior, C. A. D.; LA SCALA JR, N. Spatiotemporal analysis of atmospheric XCH_4 as related to fires in the Amazon biome during 2015–2020. Remote Sensing Applications: Society and Environment, v. 30, n., p. 100967, 2023;

v) de Lucena, W. B.; Vicentini, M. E.; DE Araújo Santos, G. A.; de Oliveira Silva, B.; Mesquita da Costa, D. V.; Ferreira Canteral, K. F.; Neira Román, J. A.; de Souza Rolim, G.; Panosso, A. R.; La Scala, N. Temporal variability of the CO_2 emission and the O2 influx in a tropical soil in contrasting coverage conditions. Journal of South American Earth Sciences, v. 121, n. 1, p. 104120, 2023;

vi) da Costa, L. M.; de Mendonça, G. C.; Araújo Santos, G. A. D.; Moraes, J. R. D. S. C. D.; Colombo, R.; Panosso, A. R.; La Scala JR, N. High spatial resolution solar-induced chlorophyll fluorescence and its relation to rainfall precipitation across Brazilian ecosystems. Environmental Research, v. 218, n., p. 114991, 2023;

vii) da Costa, L. M.; Santos, G. A. D.; Panosso, A. R.; Rolim, G. D.; La Scala, N. An empirical model for estimating daily atmospheric column-averaged CO_2 concentration above Sao Paulo state, Brazil. Carbon Balance and Management, v. 17, n. 1, p., 2022.

g) 02 manuscritos já submetidos para publicação em periódicos internacionais;

Por fim, gostaríamos de destacar que durante o período de vigência de nossa bolsa de Produtividade em Pesquisa-PQ, nossos índices de acadêmicos melhoraram significativamente. O índice-*h* na base de dados *Web of Science* aumentou de *h* = 13, em 2020 para *h* = 18, em 2024 (formato do nome da consulta: panosso ar). Atualmente, temos um total de 1059 citações na base do *Institute for Scientific Information (ISI) Knowledge - Web of Science*, com total de 82 artigos científicos publicados (dados atualizados em 09/04/2024). Além disso, temos 1168 citações na base *Scopus – Elsevier*, com um total de 85 artigos científicos publicados e um índice *h* = 18 (dados atualizados em 09/04/2024).

MODELOS DE APRENDIZADO DE MÁQUINA PARA A EMISSÃO DE CO_2 DO SOLO EM ÁREAS AGRÍCOLAS NO BRASIL

RESUMO – Modelar a dinâmica do carbono em áreas agrícolas é uma ação estratégica para diminuir as incertezas associados aos processos de mitigação de gases do efeito estufa (GEE) e melhorar a capacidade de análise para construção de cenários mais acurados. Nas últimas décadas, técnicas de inteligência artificial e mineração de dados têm sido aplicadas com sucesso na modelagem de inúmeros atributos nas ciências agrárias e ambientais. A hipótese principal do trabalho é que a transferência de carbono do solo via dinâmica espaçotemporal da emissão de CO_2 do solo (FCO_2) pode ser descrita como um fenômeno com estruturas multidimensionais que evoluem no tempo, aptas a serem modeladas por algoritmos de aprendizado de máquina estatístico. Portanto, o objetivo do estudo foi avaliar o desempenho preditivo de algoritmos de aprendizado de máquina estatístico para FCO_2 em áreas agrícolas dos estados de São Paulo e Mato Grosso do Sul. As técnicas utilizadas foram: árvore de decisão (DT), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost). Os dados de FCO_2 e atributos do solo são provenientes 19 experimentos de campo conduzidos nos últimos de 21 anos. A base de dados pode ser acessada no repositório GitHub desenvolvido e disponibilizado em https://github.com/arpanosso/tese-fco2-ml-2023. Para a modelagem temporal foram utilizados dados de múltiplas fontes: Solo, Sensoriamento Remoto e Estação Agrometeorológica. Para a modelagem espacial a dependência espacial de foi avaliada e foram construídos mapas de padrão espacial, por meio da krigagem ordinária (KO), para todos os atributos do solo. Os mapas foram utilizados no processo de aprendizagem para modelagem de FCO_2 e são provenientes de experimento realizado no ano 2017, em áreas de floresta plantada de eucalipto (EU) e sistema de consórcio de aroeira-vermelha e capim braquiária, denominado silvipastoril (SI). De maneira geral 70-80% das observações foram utilizadas para a aprendizagem estatística dos modelos (processo de treinamento) e 30-20% para validação (processo de teste). O processo de treinamento e seleção dos melhores modelos para cada algoritmo envolveu o ajuste de hiperparâmetros e validação cruzada com a utilização de 5 *folds*. Nossos resultados indicam que o algoritmo RF obteve os maiores valores de R^2 e os menores valores de RMSE (R^2 = 0,75 e RMSE = 1,00 μmol m^{-2} s^{-1}) quando considerado o grupo de variáveis Solo e Solo + Sensoriamento Remoto. Quando considerado o grupo de variáveis Solo + Sensoriamento Remoto + Estação Agrometeorológica, o algoritmo XGBoost apresentou melhor performance preditiva (R^2 = 0,83 e RMSE = 0,93 μmol m^{-2} s^{-1}) em comparação com os demais métodos. As variáveis de concentração de CO_2 atmosférico (X_{CO2}) e fluorescência de clorofila induzida pelo sol (SIF), obtidas por sensoriamento remoto, mantiveram-se consistentes em todos os melhores modelos encontrados, contribuindo de maneira significativa para o processo de aprendizagem. Na modelagem espacial, o algoritmo RF teve um desempenho superior na previsão de FCO_2 na área SI, enquanto o algoritmo XGBoost se destacou na previsão de FCO_2 na área EU. A seleção de variáveis indicou que a variabilidade espacial de FCO_2 no sistema silvipastoril foi principalmente controlada por atributos relacionados à atividade microbiana do solo e respiração heterotrófica. Por outro lado, na área de eucalipto, a variabilidade espacial de FCO_2 foi influenciada principalmente pela interação entre a temperatura a umidade do solo, atributos

diretamente associados à respiração autotrófica do solo, incluindo a respiração das raízes. A comparação entre os padrões espaciais de FCO_2 gerados pela KO e aqueles estimados pelos diferentes métodos testados, revelou que todos os algoritmos apresentaram um ótimo desempenho, com valores de R^2 entre 0,92 e 0,99% e medidas de RMSE com pequenas variações entre 0,03 e 0,55 μmol m^{-2} s^{-1}. Apesar de todos os esforços e do notável poder das técnicas de aprendizado estatístico, como o avançado algoritmo XGBoost, que aprimora seu processo de aprendizado por meio de inúmeras iterações, é fundamental destacar que, embora os modelos de Árvores de Decisão (DT) apresentem um desempenho inferior, eles oferecem as vantagens de custos computacionais mais baixos, ajuste simples dos hiperparâmetros e, principalmente, interpretabilidade do modelo. Portanto, eles continuam altamente recomendados para o processo de modelagem espacial da emissão de CO_2 em áreas agrícolas.

Palavras-chave: respiração do solo, modelagem, mudanças climáticas, efeito estufa.

MACHINE LEARNING MODELS FOR SOIL CO2 EMISSION IN AGRICULTURAL AREAS IN BRAZIL

ABSTRACT – Modeling carbon dynamics in agricultural areas is a strategic action to reduce uncertainties associated with greenhouse gas (GHG) mitigation processes and improve the analytical capacity for constructing more accurate scenarios. In recent decades, artificial intelligence and data mining techniques have been successfully applied in modeling numerous attributes in agricultural and environmental sciences. The main hypothesis of this work is that the transfer of carbon from the soil through the spatiotemporal dynamics of soil CO_2 emissions (FCO_2) can be described as a phenomenon with multidimensional structures that evolve over time, suitable for modeling by statistical machine learning algorithms. Therefore, the objective of the was to evaluate the predictive performance of statistical machine learning algorithms for FCO_2 in agricultural areas in the states of São Paulo and Mato Grosso do Sul. The techniques used were decision tree (DT), Random Forest (RF), and Extreme Gradient Boosting (XGBoost). FCO_2 data and soil attributes are derived from 19 field experiments conducted over the past 21 years. The database can be accessed on the GitHub repository developed and made available at https://github.com/arpanosso/tese-fco2-ml-2023. For temporal modeling, data from multiple sources were used: Soil, Remote Sensing, and Agrometeorological Station. For spatial modeling, spatial dependence was assessed, and spatial pattern maps were constructed using ordinary kriging (OK) for all soil attributes. The maps were used in the learning process for FCO_2 modeling and originate from an experiment conducted in 2017 in areas of eucalyptus planted reforest (EU) and a consortium system of red gum and Brachiaria grass, known as silvopastoral (SI). In general, 70-80% of the observations were used for statistical model training (training process), while 30-20% were reserved for validation (testing process). The training and selection process of the best models for each algorithm involved hyperparameter tuning and cross-validation using 5 folds. Our results indicate that the RF algorithm achieved the highest R^2 values and the lowest RMSE values (R^2 = 0.75 and RMSE = 1.00 μmol m^{-2} s^{-1}) when considering the Soil and Soil + Remote Sensing variable groups. When considering the Soil + Remote Sensing + Agrometeorological Station variable group, the XGBoost algorithm exhibited better predictive performance (R^2 = 0.83 and RMSE = 0.93 μmol m^{-2} s^{-1}) compared to the other methods. The atmospheric CO_2 concentration (X_{CO2}) and sun-induced chlorophyll fluorescence (SIF) variables obtained through remote sensing remained consistent in all the best models found, significantly contributing to the learning process. In spatial modeling, the RF algorithm exhibited superior performance in predicting FCO_2 in the SI area, while the XGBoost algorithm excelled in predicting FCO_2 in the EU area. The variable selection indicated that the spatial variability of FCO_2 in the silvopastoral system was primarily controlled by attributes related to soil microbial activity and heterotrophic respiration. Conversely, in the eucalyptus area, FCO_2 spatial variability was predominantly influenced by the interaction between soil temperature and soil moisture, attributes directly associated with autotrophic soil respiration, including root respiration. The comparison between the spatial patterns of FCO_2 generated by OK and those produced by the different tested methods indicated that all algorithms demonstrated excellent performance, with R^2 values ranging from 0.92 to 0.99% and RMSE measurements showing slight variations between 0.03 and 0.55 μmol m^{-2} s^{-1}. Despite all efforts and the

remarkable power of statistical learning techniques, such as the more advanced XGBoost algorithm, which refines its learning process through countless iterations, it is essential to emphasize that, although Decision Tree (DT) models exhibit lower performance, they offer the advantages of lower computational costs, easy fine-tuning of hyperparameters, and, the invaluable attribute of model interpretability. Thus, they remain highly recommended for the FCO_2 modeling process in agricultural areas.

Keywords: soil respiration, modeling, climate change, greenhouse effect.

1 INTRODUÇÃO

1.1 Enunciado do problema

As mudanças climáticas e a ocorrência de eventos climáticos extremos afetam a qualidade de vida das pessoas em todo o planeta. As emissões de gases de efeito estufa (GEE) representam um dos principais desafios ambientais enfrentados em escala global. Essas emissões têm impactos significativos contribuindo diretamente para o aquecimento do planeta.

De acordo com o último boletim da Organização Meteorológica Mundial publicado em outubro de 2022, as concentrações atmosféricas de dióxido de carbono (CO_2), metano (CH_4) e óxido nitroso (N_2O) atingiram os patamares de: 415,7±0,2 ppm, 1908±2 ppb e 334,5±0,1 ppb, respectivamente. Esses valores representam um aumento de 149%, 262% e 124% em relação aos níveis pré-industriais (WMO, 2022). Ainda, de acordo com o mesmo relatório, no período de 1990 a 2019, a forçante radiativa ocasionada pela concentração de GGE de longa duração na atmosfera aumentou em 45%, sendo o CO_2 responsável por cerca de 80% desse aumento.

Em âmbito global, o desmatamento e outras atividades relacionadas às mudanças do uso da terra resultaram em uma média de emissão de 5,5 Gt de CO_2 ano^{-1} no período de 10 anos, em comparação com as emissões estimadas de 36,6 Gt de CO_2 provenientes da queima de combustíveis fósseis e da produção de cimento, no mesmo período. Nesse cenário o Brasil ocupa o 6º lugar entre os países emissores de GEE, representando 3,2% do total mundial (SEEG, 2020). A principal fonte de emissão de GEE no Brasil é a agropecuária e o desmatamento associado às mudanças no uso e cobertura da terra. Em contrapartida, muitas práticas agrícolas têm potencial de mitigar as emissões GEE, das quais as mais proeminentes são: o melhor manejo dos solos agrícolas e pastagens, a restauração de solos degradados e orgânicos, a mudança no uso da terra e agroflorestas (Smith et al., 2008; Ussiri; Lal, 2009; Warner et al., 2019). Nessa conjuntura, aumentar o armazenamento de carbono nos solos é estratégico para compensar o aumento das emissões antropogênicas dos GEE.

O processo de perda de carbono do solo para a atmosfera é denominado respiração do solo, ou emissão de CO_2 do solo (FCO_2), resultante da atividade

microbiana (oxidação química) e respiração das raízes. Tal processo é considerado a segunda maior fonte de CO_2 para atmosfera, ficando atrás apenas dos oceanos.

Os níveis de carbono orgânico nos solos são os resultados das complexas interações entre variáveis relacionadas aos processos de produção e transporte do CO_2 do solo para a atmosfera. Pequenos incrementos nas taxas de FCO_2 podem ser suficientes para um ecossistema mudar de sumidouro para fonte de carbono para a atmosfera (Sartori et al., 2006; Laganiere; Angers; Pare, 2010). Entretanto, os fatores ambientais e atributos do solo que controlam a magnitude de FCO_2 continuam a ser de difícil separação e interpretação. Mesmo após décadas de pesquisa, ainda são poucos modelos matemáticos robustos desenvolvidos para prever o efeito de fatores bióticos e abióticos no equilíbrio do carbono do solo (Sierra et al., 2015; Farhate et al., 2018a; Grunwald, 2022).

As razões para essa dificuldade devem-se às numerosas interações entre tais fatores, em combinação com diferentes escalas de variação no tempo e no espaço (Allaire et al., 2012; Graf et al., 2012; Bicalho et al., 2014). Na escala temporal a emissão de CO_2 do solo é controlada por atributos como a umidade do solo (Leon et al., 2014; Santos et al., 2019a), porosidade livre de água (Linn; Doran, 1984; Warner et al., 2018), oxigenação do solo (Chen; Chen, 2010; Vicentini et al., 2019) e variáveis climáticas como a precipitação, a temperatura do ar e a radiação solar (La Scala Jr.; Panosso; Pereira, 2003; Warner et al., 2019). Na escala espacial, as variações de FCO_2 são controladas principalmente por atributos físicos, químicos e biológicos do solo, tais como densidade do solo (Saiz et al., 2006), textura (Dilustro et al., 2005; Tavares et al., 2018), teor de matéria orgânica (Soe; Buchmann, 2005; Tavares et al., 2018), estoque de carbono (Panosso et al., 2011; da Silva et al., 2020), pH e acidez potencial (Reth; Reichstein; Falge, 2005; Farhate et al., 2018b).

A complexidade dessas relações exige o uso de técnicas matemáticas e estatísticas sólidas para extrair informações cada vez mais precisas e acuradas. Diante dessa demanda, as técnicas de aprendizado de máquina (*Machine Learning*) têm ganhado destaque devido à sua eficiência, robustez e versatilidade. Por exemplo, modelos de ML para o estoque de carbono no solo têm sido construídos usando fatores ambientais que representam domínios como: solo, topografia, ecologia,

litologia, atmosfera, hidrologia, biota e atividades humanas (McBratney; Santos; Minasny, 2003; McBratney; de Gruijter; Bryce, 2019; Grunwald, 2022).

Quanto à FCO_2, essa abordagem torna-se bastante desafiadora, uma vez que vários atributos do solo, apresentam dependências espacial (autocorrelação espacial), e em muitas vezes anisotropia, ou seja, os padrões de variabilidade espacial são diferentes em função das diferentes direções em campo. Nesse contexto a junção das técnicas de geoestatística e aprendizado de máquina tem potencial de aplicação na área, pois, de acordo com Reichstein et al. (2019) o aprendizado de máquina é capaz de extrair padrões de dados espaçotemporais automaticamente. Esta característica possibilita a melhor compreensão de fenômenos diversos, resultando em ganho da capacidade preditiva de processos sazonais e na modelagem das relações espaciais em várias escala de tempo.

Assim, a caracterização de como os diferentes usos e manejos dos solos brasileiros afetam a dinâmica FCO_2 é de grande importância para a determinação de forma quantitativa, do potencial de mitigação desse gás, refletindo no impacto dessas práticas no clima do planeta. O melhor entendimento deste fenômeno e suas interações são essenciais para a construção de mapas de padrões espaciais com menores incertezas associadas, podendo ser utilizados na elaboração de estratégias para apoiar ações de mitigação, especialmente em regiões agrícolas tropicais, onde as emissões são frequentemente mais altas associadas aos baixos valores de estoque de carbono do solo.

1.2 Hipótese e objetivo

A hipótese principal do trabalho é que a dinâmica espaçotemporal da emissão de CO_2 do solo pode ser descrita como um fenômeno com estruturas multidimensionais que evoluem no tempo, aptas a serem modeladas por algoritmos de aprendizado de máquina, haja visto a complexidade de suas interações. Diante do exposto, o objetivo do estudo foi avaliar o desempenho preditivo de algoritmos de aprendizado de máquina para emissão de CO_2 do solo em diferentes usos, manejos e regiões agrícolas do Brasil, mais especificamente, dos estados de São Paulo e Mato Grosso do Sul. As técnicas utilizadas foram: árvore de decisão (DT), *Random Fores*t (RF) e *Extreme Gradient Boosting* (XGBoost). Especificamente, pretende-se usar a

abordagem de aprendizado de estatístico para investigar padrões espaçotemporais de FCO_2, avaliando as fontes de incerteza de previsão do atributo em diferentes regiões e manejos e usos do solo.

2 REVISÃO DE LITERATURA

2.1 Inteligência artificial

Inteligência artificial (IA) é a capacidade de máquinas (computadores e sistemas de software) assimilarem informações por meio de algoritmos, para realizar tarefas características da inteligência humana, como reconhecer objetos e sons, contextualizar a linguagem, aprender com o ambiente e resolver problemas. A etapa de assimilação de informação é denominada aprendizado de máquina (*Machine Learning*). Assim, a IA pode ser entendida como o potencial da máquina em tomar a melhor decisão possível, considerando a quantidade de informação disponível e demonstrando habilidade de adaptação a diferentes situações (Kuhn; Johnson, 2013; Chollet; Allaire, 2017).

Nos últimos anos a IA tem sido utilizada em áreas como educação, saúde, direito, entretenimento, indústria e ambiental (Ge et al., 2017; Chiavegatto Filho et al., 2018; Grunwald, 2022), se tornando cada vez mais corriqueira e presente na vida cotidiana da sociedade contemporânea (Di Minin et al., 2018). Estudos a respeito da IA iniciaram-se na década de 1950 dentro da área da ciência da computação com o intuito de automatizar tarefas intelectuais, normalmente executadas por seres humanos. Até o final dos anos 80 acreditava-se que a IA, em nível humano, poderia ser alcançada mediante a programação de um grande conjunto de regras explícitas para manipular o conhecimento (Chollet; Allaire, 2017). Esse paradigma ficou conhecido como IA simbólica e, embora capaz de solucionar uma série de problemas lógicos bem definidos, não se mostrou eficiente na solução de problemas mais complexos e confusos como os inerentes à classificação de imagens, ao reconhecimento de linguagem natural e à tradução linguística. Isso porque a IA simbólica não consegue derivar as regras explícitas para a solução de um problema, uma vez que as regras devem ser previamente programadas no sistema, e os dados são processados de acordo com essas mesmas regras (Kuhn; Johnson, 2013; Bruce; Bruce, 2017; Chollet; Allaire, 2017).

Em meados dos anos 80 e início dos anos 90 uma nova abordagem surgiu, o aprendizado de máquina estatístico, ou simplesmente, aprendizagem de máquina. Nesse novo paradigma os dados a respeito de um problema são fornecidos à máquina, bem como as respostas esperadas, e o aprendizado consiste na geração

das regras de decisão como o resultado final do sistema computacional, ou seja, a máquina aprende por conta própria as regras para a tomada de decisão (Bruce; Bruce, 2017). Nesse momento da história, a aprendizagem de máquina se concentrou no desenvolvimento de algoritmos eficientes que escalonam grandes quantidades de dados, visando otimizar modelos preditivos para compor as regras de predição a partir da compreensão da estrutura dos dados. Portanto, o aprendizado de máquina estatístico é um subcampo da IA cujo objetivo é o desenvolvimento de algoritmos e modelos que permitam que os computadores aprendam a partir de dados e façam previsões ou tomem decisões com base nessas mesmas observações. É um campo de pesquisa estatística para o treinamento de algoritmos computacionais cuja estratégia é dividir, agrupar e transformar a base de dados para maximizar a capacidade de classificar, prever, ou descobrir padrões em um conjunto de dados de destino (Reichstein et al., 2019).

Pode-se entender o aprendizado de máquina como o campo de estudo que possibilita aos computadores a capacidade de aprender sem serem explicitamente programados. O termo "aprendizagem de máquina" foi cunhado pela primeira vez por Arthur Samuel em 1959, no seu artigo intitulado "*Some Studies in Machine Learning Using the Game of Checkers*", onde o autor descreve como um programa de computador conseguiu aprender a jogar damas por meio da prática e da experiência, ao invés de ser programado de forma explícita com as regras do jogo (Samuel, 1959). Neste estudo, o autor faz a distinção entre dois métodos gerais de abordagem no processo de aprendizado. Na primeira abordagem, conhecida como *abordagem de redes neurais*, explora-se a possibilidade de induzir comportamentos aprendidos em uma rede de comutação conectada aleatoriamente (ou simulada computacionalmente), como resultado de um processo de recompensa e punição. Embora essa abordagem não fosse computacionalmente possível para os padrões da época, ela poderia levar ao desenvolvimento de máquinas de aprendizado de propósito geral. Por outro lado, na segunda abordagem descrita, explorou-se a criação de *redes altamente organizadas* projetadas para aprender apenas tarefas específicas. Além de ser viável para a época, essa abordagem era muito mais eficiente, porém exigia reprogramação para cada nova aplicação (Samuel, 1959). Atualmente, ambas as abordagens são amplamente utilizadas no campo da aprendizado de máquina.

Como exemplo da abordagem de *redes neurais*, tem-se as Redes Neurais Convolucionais (CNNs), Redes Neurais Recorrentes (RNNs) e Redes Neurais de Transformadores (RNT). Já como exemplo da abordagem de *redes altamente organizadas*, temos técnicas como Regressão Logística, *Support Vector Machines* (SVM), Árvores de Decisão (*Decision Tree* - DT), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost).

Um exemplo que merece atenção atualmente é o ChatGPT, pois foi observado um notável aumento em sua popularidade durante o ano de 2023. O chat representa um, entre vários, *chatbots*, ou seja, Modelos de Linguagem Natural de Grande Escala (*Large Language Model - LLM*), que exibe a capacidade de fornecer respostas sofisticadas e aparentemente inteligentes durante interações com seus usuários (Bockting et al., 2023). O cerne do ChatGPT é o modelo GPT (*Generative Pre-trained Transformer*), fundado em RNT que, após um treinamento massivo com extensos conjuntos de dados de texto, foram capazes de adquirir um aprimorado entendimento das estruturas e padrões linguísticos, permitindo-lhe prever com elevada precisão a próxima palavra ou *token* em uma sequência de texto (Bockting et al., 2023; Stokel-Walker, 2023).

2.2 Métodos de aprendizado de máquina

Os métodos de aprendizado de máquina podem ser divididos em quatro categorias: a) *aprendizado não supervisionado*; b) *aprendizado supervisionado*; c) *aprendizado semi-supervisionado* e; d) *aprendizado por reforço* (Bruce; Bruce, 2017; Santos et al., 2019b; Bruce; Bruce; Gedeck, 2020):

No *aprendizado supervisionado* os dados abarcam variáveis de entrada (preditoras ou *features*) em conjunto com suas respectivas variáveis de saída (respostas ou rótulos) (Kroese et al., 2019). As variáveis de saída podem ser categóricas ou contínuas, portanto, existe uma distinção na nomenclatura para cada um desses casos. Por exemplo. ao examinar os padrões de degradação do solo em uma área específica, com o objetivo de classificar os distintos níveis de degradação, tem-se como resposta uma variável categórica como resposta (Das et al., 2023). Nesse exemplo, dizemos que a aplicação é uma tarefa de aprendizado supervisionado de *classificação*. Por outro lado, se a resposta for uma ou mais variáveis contínuas, a

tarefa de aprendizado supervisionado é então denominada de *regressão*. Um exemplo ilustrativo do problema de regressão é a previsão do coeficiente de escoamento e da produção de sedimentos em processo de erosão do solo em áreas de pastagem, em que as entradas englobam variáveis como atributos do solo, chuva e informações específicas como cobertura do solo e rugosidade da superfície (Martinez et al., 2017).

No *aprendizado não supervisionado,* os dados consistem apenas em um conjunto de entradas sem alvos correspondentes, ou seja, não é apresentado uma variável resposta, os dados não possuem rótulos (Kroese et al., 2019). O objetivo principal desses métodos é explorar os dados e encontrar alguma estrutura oculta entre eles, geralmente procuram-se grupos de amostras, ou seja, acessos similares nos dados, chamado de agrupamento (*clustering*). Outro objetivo pode ser determinar a distribuição dos dados no espaço de entrada, conhecida como estimativa de densidade, ou projetar os dados com o propósito de redução de dimensionalidade e visualização deles (Kroese et al., 2019). A redução de dimensionalidade é uma análise que é realizada em ambas as famílias de tipos de *aprendizado supervisionado* e *não supervisionado*, com o objetivo de fornecer uma representação mais compacta e de menor dimensionalidade de um conjunto de dados para preservar o máximo de informações possível dos dados originais (Ge et al., 2017; Liakos et al., 2018). Métodos de *aprendizado não supervisionado* comumente empregados nesses contextos incluem a análise de componentes principais (PCA), regressão de mínimos quadrados parciais (PLSR), análise discriminante linear (LDA), análise de componentes independentes (ICA), técnicas de clusterização hierárquicas e não hierárquicas (*k-means*), estimativa de densidade de kernel e o uso de mapas auto-organizáveis (Liakos et al., 2018; Padarian; Minasny; McBratney, 2020; Grunwald, 2022).

O *aprendizado semi-supervisionado* pode ser considerado o *aprendizado supervisionado* onde os dados de treinamento contêm poucos acessos rotulados, e muitos acessos não rotulados. De acordo com Ge et al. (2017) o método de aprendizado *semi-supervisio*nado é particularmente útil quando o custo de rotular amostras de dados é oneroso ou demorado. Categorização de texto e diagnósticos médicos são alguns exemplos desse tipo de tarefa (Chiavegatto Filho et al., 2018; Santos et al., 2019b). Na agricultura, o aprendizado *semi-supervisionado* geralmente

tem sido aplicado para classificação de culturas em imagens de satélite; monitoramento de doenças e pragas, identificação de plantas daninhas e mesmo em previsão de rendimento de culturas (Amorim et al., 2019; Jiang et al., 2020; Tetila et al., 2020; Jiang et al., 2021; Khan et al., 2021; Shorewala et al., 2021).

Como última categoria de aprendizado, temos o *aprendizado por reforço.* Essa categoria utiliza uma abordagem que permite programar agentes por meio de recompensas e penalizações, sem a necessidade de explicitar como uma tarefa deve ser realizada (Kaelbling; Littman; Moore, 1996). O *aprendizado por reforço* é muito utilizado em áreas como robótica, automação, jogos e navegação (Ge et al., 2017). Uma das principais diferenças entre *o aprendizado supervisionado* e o *aprendizado por reforço* é a ausência de apresentação de pares de entrada/saída. Ao invés disso, após selecionar uma ação, o agente recebe a recompensa imediata, mas não obtém instruções sobre a ação mais benéfica a longo prazo (Kaelbling; Littman; Moore, 1996). Assim, é necessário que o agente ativamente acumule experiência útil sobre os estados potenciais do sistema, ações, transições e recompensas para otimizar suas decisões (Kaelbling; Littman; Moore, 1996). A estrutura desse método é intermediária, onde diversas entradas resultam em várias saídas que são avaliadas para determinar sua eficácia. Essa etapa de avaliação, geralmente, é realizada por humanos, medindo a precisão da previsão e fornecendo ao agente um indicativo de erro. Ademais, em contraste com o *aprendizado supervisionado*, o desempenho em tempo real é essencial, implicando em avaliações simultâneas, durante o processo de aprendizado. Consequentemente, o *aprendizado por reforço* se mostra eficaz mesmo com menos dados, diferenciando-se da abordagem supervisionada que exige um volume substancial de informações. Um exemplo desse paradigma é o, já mencionado, ChatGPT. No caso do ChatGPT, o primeiro passo do processo de aprendizado foi *supervisionado*, onde vários estímulos (*prompts, ou seja,* perguntas realizadas durante a fase de treinamento) foram atribuídos e um humano forneceu as respostas correspondentes. O agente, então, estabeleceu uma conexão entre esses estímulos humanos e buscou gerar respostas coerentes em conformidade. Após aprender a gerar as primeiras respostas, ocorreu o *aprendizado por reforço*. Nesse segundo passo o agente foi treinado com *prompts* repetidos várias vezes (ou com novos prompts), gerando um conjunto de saída de várias respostas. Então um humano

classificou as respostas em termos de qualidade, ou seja, o agente recebeu feedback em forma de recompensas positivas e negativas. Por meio de muitas interações, os modelos internos foram ajustados para aprender a melhorar as respostas com base nas recompensas recebidas.

2.3 Aprendizado estatístico

O processo de aprendizado, conhecido como aprendizado estatístico, requer a compreensão de certos conceitos fundamentais. De forma geral, suponha que em um estudo foi observado uma resposta quantitativa Y e p diferentes variáveis preditoras denotadas por $\boldsymbol{X}_{n\times p}$. É assumido então que existe uma relação entre Y e $\boldsymbol{X}$ = ($\boldsymbol{X}_1$, $\boldsymbol{X}_2$,...,$\boldsymbol{X}_p$), a qual pode ser expressa na forma bastante geral (James et al., 2013; Kroese et al., 2019):

$$Y = f(\boldsymbol{X}) + \epsilon. \quad (1)$$

Em que f é alguma função fixa, mas não conhecida de $\boldsymbol{X}_1$, $\boldsymbol{X}_2$,...,$\boldsymbol{X}_p$, e ϵ é o termo de erro aleatório, que é independente de $\boldsymbol{X}$ e possui média zero. Em geral, x_{ij} representa o valor da j-ésima variável para a i-ésima observação, onde $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$. Portanto $\boldsymbol{X}$ representa uma matriz $n \times p$, cujo elemento (i, j) é representado por x_{ij},

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Nesta formulação f representa a informação sistemática que $\boldsymbol{X}$ fornece sobre Y (Izbicki; Santos, 2020). No entanto, a função f que conecta as variáveis de entrada à variável de saída é não conhecida, sendo necessário estimá-la com base nas observações amostrais. Dados esses elementos iniciais, começamos a vislumbrar o aprendizado de máquina estatístico, pois, de acordo com James et al. (2013), o termo *aprendizado estatístico* se refere a um conjunto de abordagens cujo objetivo principal é encontrar f que expresse a relação existente entre as variáveis de entrada e a variável de saída, ou seja, entre as variáveis preditoras e a resposta.

Na literatura científica existem duas motivações principais para estimar f: uma para *inferência* e outra para *previsão* (Breiman, 2001; James et al., 2013). No objetivo inferencial o interesse é identificar quais são as entradas mais importantes para uma

saída, bem como entender qual a relação entre cada $\boldsymbol{X}_j$ e a variável alvo Y. Já, quando o objetivo é preditivo, deseja-se criar uma função que tenha um bom poder de previsão, ou seja, dadas novas observações das variáveis preditoras será realizada uma boa previsão da variável resposta (Izbicki; Santos, 2020). Nestas situações, uma vez que o termo de erro tem média zero, pode-se prever Y por meio de:

$$\hat{Y} = \hat{f}(\boldsymbol{X}), \tag{2}$$

em que $\hat{f}$ representa a estimativa de f e $\hat{Y}$ o resultado da predição de Y. Nesse contexto, normalmente, não se procura a forma exata de $\hat{f}$, desde que ela forneça previsões precisas para Y (James et al., 2013). Assim, não se assume que o modelo utilizado para os dados é correto; o modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações (Shalev-Shwartz; Ben-David, 2014; Izbicki; Santos, 2020).

Nesse contexto, é oportuno ressaltar que a acurácia de $\hat{Y}$ como uma previsão de Y depende de duas quantidades denominadas como o *erro redutível* e o *erro irredutível*. De maneira geral $\hat{f}$ não será uma estimativa perfeita de f e a sua falta de acurácia adicionará erro nas estimativas. Esse erro é *redutível* pois pode-se potencialmente melhorar a precisão de $\hat{f}$ ao utilizar técnicas de aprendizado estatístico mais apropriados para estimar f. Destaca-se que, mesmo que fosse possível obter uma estimativa perfeita para f ou seja, $\hat{Y} = f(\boldsymbol{X})$, a predição ainda conteria algum erro. Isso ocorre pois Y também é uma função de ϵ que, por definição, não pode ser previsto usando $\boldsymbol{X}$. Portanto, a variabilidade associada ao termo ϵ da equação (1) também afeta a precisão das previsões, isso é conhecido como o *erro irredutível*, pois, não importa o quão bem estima-se f, não é possível reduzir o erro introduzido pelo termo ϵ. Ainda, de acordo com James et al. (2013), a quantidade ϵ pode conter variáveis não medidas que são úteis para prever Y, inviabilizando o uso dessas informações por f para realizar a boas previsões.

Assim, considerando uma dada função estimada $\hat{f}$ e um conjunto de preditores $\boldsymbol{X}$, que resultam na previsão $\hat{Y} = \hat{f}(\boldsymbol{X})$. Ao assumirmos que ambos $\hat{f}$ e X são fixos, pode-se demonstrar que (James et al., 2013; Shalev-Shwartz; Ben-David, 2014; Kroese et al., 2019):

$$E\left(Y - \hat{Y}\right)^2 = E\left[f(\boldsymbol{X}) + \epsilon - \hat{f}(\boldsymbol{X})\right]^2$$

$$E\left(Y - \hat{Y}\right)^2 = \left[f(\boldsymbol{X}) - \hat{f}(X)\right]^2 + \mathrm{Var}(\epsilon), \quad (3)$$

em que $E\left(Y - \hat{Y}\right)^2$ representa o valor esperado do quadrado da diferença entre o valor predito e o valor real, observado de Y, o termo $\left[f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\right]^2$ é o *erro redutível* e $\mathrm{Var}(\epsilon)$ representa a variância associada ao termo de erro ϵ, ou seja, o *erro irredutível*.

Em resumo, o aprendizado estatístico acontece por meio de técnicas que visam estimar f, com o objetivo de minimizar o *erro redutível*, sendo que o *erro irredutível* sempre estabelecerá um limite superior para a acurácia e precisão das mesmas estimativas de Y (Ng, 2018). Nesse contexto são empregadas diferentes abordagens, mas que sempre compartilharão algumas características em comum, que serão apresentadas a seguir.

No início do processo de aprendizado estatístico uma parte dos dados deverá ser separada, tais observações são chamadas de *dados de treinamento* e serão utilizadas para ensinar o método em questão a estimar f não conhecida. Costuma-se deixar de 70 a 80% da base de dados total para essa etapa. De acordo com James et al. (2013), o objetivo será encontrar $\hat{f}$ tal que $Y = \hat{f}(\boldsymbol{X})$, para qualquer observação $(\boldsymbol{X}, Y)$. Nos dados de treinamento x_{ij} representa o valor do j-ésimo preditor, para a observação i, onde $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, p$. De maneira análoga, y_i representa a variável resposta para a i-ésima observação. Então, os dados de treinamento consistem em $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$, em que $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{ip}\right)^T$.

O próximo passo na estimativa de funções de predição envolve a criação de um critério para avaliar o desempenho de uma previsão específica, o que pode ser feito pelo *risco quadrático*, por exemplo, em um contexto de regressão.

$$R_{\mathrm{pred}}(\hat{f}) = E\left[(Y - \hat{f}(\boldsymbol{X}))^2\right], \quad (4)$$

$R_{\mathrm{pred}}(\hat{f})$ é conhecido como o *risco de previsão* (ou erro de generalização), em que $(\boldsymbol{X}, Y)$ representa uma observação que não foi utilizada previamente no processo de estimativa de $\hat{f}$. Assim, define-se uma métrica de qualidade das previsões realizadas por um modelo específico de aprendizado estatístico, aplicado em novos dados que não foram utilizados durante a etapa de treinamento. O *risco de previsão* pode ser entendido como uma medida do quanto o modelo consegue generalizar o conhecimento adquirido no treinamento para fazer previsões precisas em dados ainda

não observados. Assim, quanto menor o *risco de previsão*, melhor será a função de predição $\hat{f}$ (Kroese et al., 2019). Portanto pode-se agora definir:

$$L(\hat{f}(\boldsymbol{X}); Y) = (Y - \hat{f}(\boldsymbol{X}))^2, \tag{5}$$

a equação (5) é denominada de *função de perda quadrática*, embora outras funções possam ser utilizadas como critério de seleção. Geralmente, as funções de perda são específicas para as diferentes técnicas de aprendizado estatístico.

Uma vez definido o *risco de previsão*, outro termo deve ser definido, o *risco de regressão*. Esse termo é uma medida mais ampla que o *risco de previsão*, pois indica a qualidade do ajuste do modelo em relação aos dados de treinamento. Lembrando que, em problemas de regressão, o objetivo é encontrar uma função que relacione as variáveis (entrada e saída) com o menor erro possível. O *risco de regressão* mede o erro entre as previsões do modelo e os valores observados nos dados da base de treinamento. Observe que, o *erro redutível* está associado à falta de acurácia do modelo em relação aos dados de treinamento e, consequentemente, está relacionado ao *risco de regressão*. Podemos medir o risco de um estimador da função de regressão por meio da *perda quadrática* por:

$$R_{\text{reg}}(\hat{f}) = E\left[(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2\right], \tag{6}$$

e é possível demonstrar que:

$$R_{\text{pred}}(\hat{f}) = R_{\text{reg}}(\hat{f}) + E[\text{Var}(Y|\boldsymbol{X})], \tag{7}$$

Portanto, criar uma função de predição ($\hat{f}(\boldsymbol{X})$) é equivalente a encontrar um bom estimador para a função de regressão $f(\boldsymbol{X})$ (Izbicki; Santos, 2020). Assim temos que:

$$E\left[\left(Y - \hat{f}(\boldsymbol{X})\right)^2\right] = E[(f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X}))^2] + E[\text{Var}(Y|\boldsymbol{X})], \tag{8}$$

então, a melhor função de predição para Y é a função de regressão $f(\boldsymbol{X})$, ou seja, $\hat{f}$ é a função que minimiza o *risco de previs*ão que também minimiza o *risco de regressão* quando comparado à função de referência. No processo de aprendizado de máquina estatístico é o computador que procura por uma $\hat{f}(\boldsymbol{X})$ que faz com que tanto o *risco de previsão* quanto o risco de regressão sejam mínimos em relação à $f(\boldsymbol{X})$.

A seguir, se faz necessário discorrer sobre as diferentes formas que $\hat{f}$ pode assumir e as implicações inerentes a essas aproximações. De maneira geral, a

maioria dos métodos de aprendizado estatístico para encontrar a função $\hat{f}$ tal que $Y \approx \hat{f}(\boldsymbol{X})$ para qualquer observação $(\boldsymbol{X}, Y)$, podem ser caracterizados como métodos *paramétricos* ou métodos *não paramétricos* (James et al., 2013).

Métodos paramétricos assumem que a função de regressão pode ser parametrizada com um número finito de parâmetros (James et al., 2013; Izbicki; Santos, 2020), a regressão linear utiliza uma forma linear para a função de predição, escrita como:

$$f(\boldsymbol{X}) = \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \cdots + \beta_p \boldsymbol{X}_p, \quad (9)$$

tal suposição sobre a forma funcional da função é uma estratégia simplista pois, ao invés de estimar-se uma função $f(\boldsymbol{X})$ inteiramente arbitrária de p dimensões, é necessário apenas estimar os $p + 1$ coeficientes $\beta_0, \beta_1, \ldots, \beta_p$. Nesse contexto, deve-salientar que $\boldsymbol{X}_1$, por exemplo, não será necessariamente a j-ésima variável original, pois pode-se criar novas covariáveis que serão funções das originais (James et al., 2013; Izbicki; Santos, 2020). A desvantagem potencial de uma abordagem paramétrica é que o modelo que escolhemos geralmente não corresponderá à verdadeira forma não conhecida de f (James et al., 2013). Se o modelo escolhido estiver muito distante do verdadeiro f, então nossa estimativa será imprecisa. Pode-se contornar essa desvantagem a partir da estimativa de modelos mais flexíveis capazes de se ajustar a muitas formas de f. Porém, maior flexibilidade implica na estimativa de um maior número de parâmetros, aumentando assim o grau de complexidade do modelo. Esses modelos mais complexos podem levar a um fenômeno conhecido como sobreajuste aos dados (*overfitting*), o que significa que os modelos acabam seguindo muito de perto os erros ou ruídos (Dutt; Chandramouli; Das, 2018). Assim, qualquer abordagem paramétrica traz consigo a possibilidade de que a forma funcional usada para estimar f seja muito diferente da verdadeira f, caso em que o modelo resultante não se ajusta bem aos dados (Shalev-Shwartz; Ben-David, 2014).

Em contraste, métodos *não paramétricos* evitam esse problema na medida em que nenhuma suposição é feita sobre a forma funcional f, ao invés disso, tais métodos procuram uma estimativa de f que se aproxime dos pontos de dados (observações) de forma suave e consistente, evitando irregularidades excessivas. Métodos *não paramétricos* têm o potencial de se ajustar com precisão a uma variedade mais ampla

de possíveis formas para f. Contudo, abordagens *não paramétricas* sofrem de uma desvantagem significativa, pois ao não reduzir o problema de estimar f para um pequeno número de parâmetros, é necessário um grande número de observações para obter uma estimativa precisa de f em comparação ao número de observações necessário na abordagem paramétrica (James et al., 2013). Além disso, o fenômeno de *overfitting* ainda pode ocorrer nesses métodos quando é definido um baixo nível de suavização, resultando em ajustes mais irregulares, fazendo com que o modelo, mais uma vez, siga de perto os erros ou ruídos presentes na base de dados (James et al., 2013; Shalev-Shwartz; Ben-David, 2014).

O fenômeno do *overfitting* é uma preocupação recorrente em diversos métodos de modelagem e previsão e a compreensão de suas origens é de extrema relevância para elucidar as medidas necessárias para prevenir e mitigar seus efeitos. Como dito anteriormente, o *risco de previsão* é frequentemente avaliado usando métricas de desempenho, como o erro quadrático médio (MSE), calculado por:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2, \tag{10}$$

em que $\hat{f}(x_i)$ é o valor predito pela função $\hat{f}$ para a i-ésima observação. Observe que MSE apresentará valor pequeno se a previsão apresentar valor próximo do verdadeiro valor da resposta y_i. Durante o processo de aprendizagem, pode-se calcular essa métrica a partir da base de treinamento que foi separada para ajustar o modelo, denominado *MSE de treinamento.* No entanto, o objetivo é determinar se $\hat{f}(x_0)$ é aproximadamente igual a y_0, onde (x_0, y_0) é uma observação do conjunto de teste que não foi utilizada na prévia etapa de treinamento do método de aprendizado estatístico. Então, deve-se selecionar o método que resulta no menor *MSE de teste*, em contraste ao menor *MSE de treinamento* (James et al., 2013). Contudo, não há garantias de que o método com o menor *MSE de treinamento* também terá o menor *MSE de teste*. Geralmente, o problema é que muitos métodos estatísticos estimam os coeficientes de forma a minimizar o *MSE de treinamento*, mas o *MSE de teste* é frequentemente muito maior, ou seja, o MSE é um estimador muito otimista do *risco de previsão*.

Por exemplo, dado um cenário em que vários modelos são ajustados a um conjunto de dados de treinamento, com a complexidade dos modelos aumentando gradualmente, ou seja, os modelos têm diferentes níveis de flexibilidade, como

exemplos: modelos lineares, quadráticos, cúbicos, até polinômios de ordem superior (Figura 1). A análise dos ajustes indica com o aumento da flexibilidade dos modelos (reta e curvas) os valores de $\hat{f}(x_i)$ se aproximem cada vez mais dos valores reais (círculos), ou seja, as previsões ficam próximas das respostas y_i (James et al., 2013; Kroese et al., 2019).

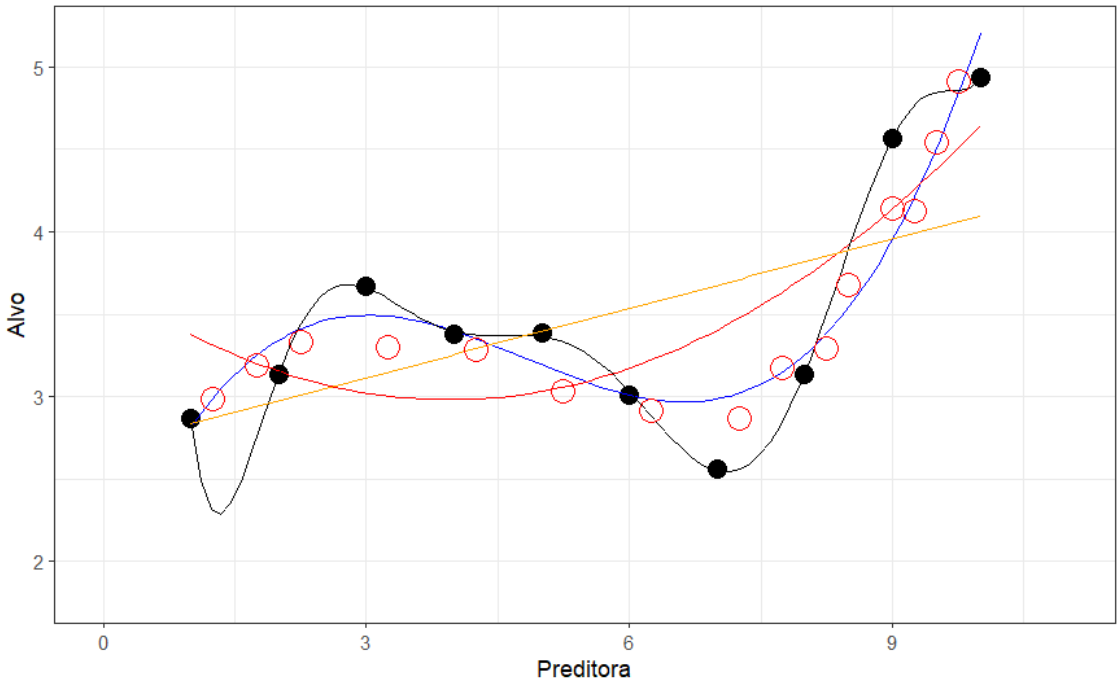

Figura 1. Dados simulados de f e ajuste de modelos polinomiais. Dados do conjunto de treinamento (●); dados do conjunto de teste (○). Regressão linear (linha laranja), ajustes polinomiais quadrático (linha vermelha); cúbico (linha azul); grau 10 (linha preta). Fonte: Adaptado de Shalev-Shwartz; Ben-David (2014).

Consequentemente, o valor do *MSE de treinamento* tende a diminuir com o aumento da flexibilidade dos modelos (Figura 2). No entanto, esse mesmo comportamento não se verifica para *MSE de teste*. Inicialmente, pode-se esperar que o aumento na complexidade dos modelos resulte em previsões mais próximas dos valores observados no conjunto de teste (círculo vermelhos da Figura 1). Contudo, a partir de um certo grau de flexibilidade, as estimativas $\hat{f}(x_i)$ começam a se ajustar em excesso aos dados específicos do conjunto de treinamento (círculos pretos da Figura 1), seguindo os erros aleatórios na tentativa de reduzir ainda mais o *MSE de treinamento.* Como consequência, $\hat{f}(x_i)$ começa a se afastar do verdadeiro formato de f, levando a um aumento no *MSE de teste* (linha azul na Figura 2). Nas situações

em que os modelos começam a perder a capacidade de generalização, tem-se o fenômeno conhecido como sobreajuste (*overfitting*). Isso ocorre porque o procedimento de aprendizado estatístico está se esforçando demasiadamente a encontrar padrões nos dados de treinamento e pode estar identificando alguns padrões resultantes do acaso, ao invés de propriedades verdadeiras da função não conhecida f (Ng, 2018). Então, o *MSE de teste* aumentará, pois os supostos padrões que o método encontrou nos dados de treinamento simplesmente não existem nos dados de teste, como exemplificado na Figura 2.

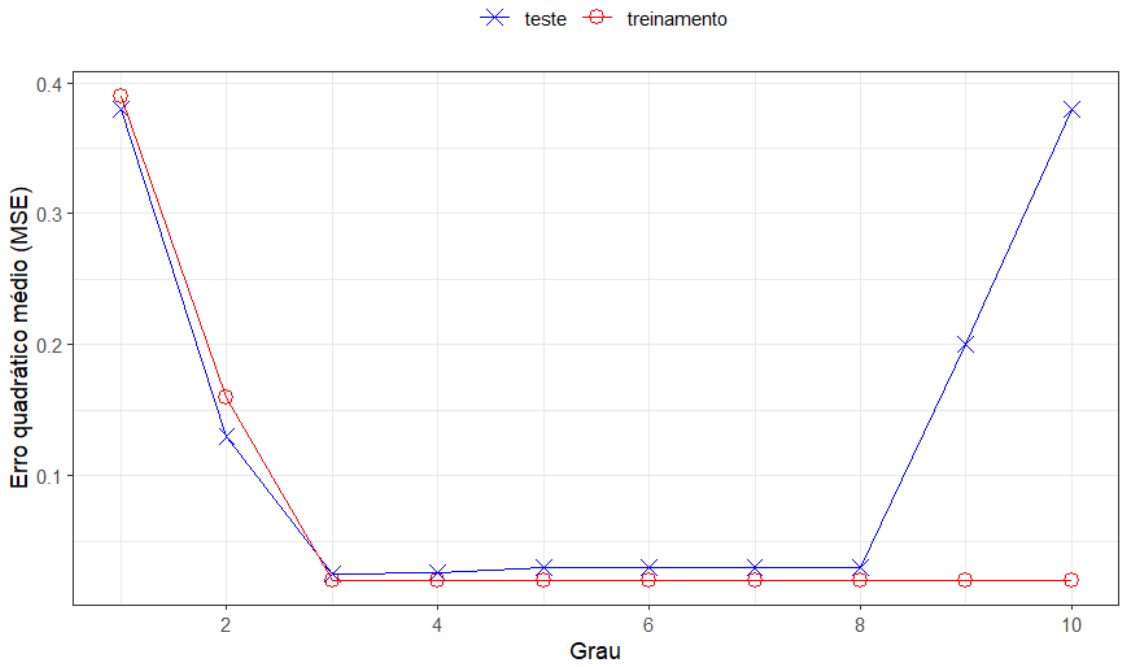

Figura 2. MSE de treino (linha vermelha) e MSE de treinamento. Fonte: Adaptado de Shalev-Shwartz; Ben-David (2014).

Esta é uma propriedade fundamental do aprendizado estatístico, que se mantém independentemente do conjunto de dados utilizados ou dos métodos de aprendizado praticados. Assim, pode-se dizer que o *overfitting* ocorre quando determinado modelo tem um ótimo desempenho no conjunto de treinamento, mas não consegue fazer previsões precisas no conjunto de teste. Portanto, separar o conjunto de dados em duas partes (*data splitting*), *treinamento* e *teste,* é a maneira mais eficiente de contornar o problema acima descrito. Nessa divisão, o conjunto de treinamento é usado para estimar $\hat{f}$ e o conjunto de teste é usado para estimar MSE,

na prática, avalia-se o erro quadrático médio no conjunto de teste (James et al., 2013; Kroese et al., 2019; Izbicki; Santos, 2020).

Ademais, o valor esperado de *MSE de teste* para um dado x_0 pode ser decomposto na soma de três quantidade fundamentais: a variância de $\hat{f}(x_0)$, o quadrado do viés de $\hat{f}(x_0)$ (Bias) e a variância dos termos de erro (ϵ). De acordo com James et al. (2013) e Kroese et al. (2019) temos:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}(\hat{f}(x_0))\right]^2 + \mathrm{Var}(\epsilon), \quad (11)$$

em que $E\left(y_0 - \hat{f}(x_0)\right)^2$ define o valor esperado do *MSE de teste* e se refere ao erro quadrático médio obtido ao realizar estimativas repetidas de f a partir de uma grande quantidade de conjuntos de treinamento, sendo estes testados em cada valor de x_0. O erro quadrático médio esperado geral de teste pode ser calculado pela média de $E\left(y_0 - \hat{f}(x_0)\right)^2$ em todos os possíveis valores de x_0 no conjunto de teste. Isso implica que, para minimizar o erro de teste esperado, é necessário selecionar um método de aprendizado estatístico que alcance simultaneamente baixa variância e baixo viés (Bias). Note que, como a variância e o viés são quantidades não negativas, o valor esperado para *MSE de teste* não poderá ser inferior à $\mathrm{Var}(\epsilon)$, o erro irredutível apresentado na equação (3).

Em geral, métodos estatísticos mais flexíveis tendem a exibir uma variância mais alta (James et al., 2013). A variância se refere à quantidade pela qual $\hat{f}$ mudaria se fosse estimado em diferentes conjuntos de dados de treinamento. O ideal é que essa variação seja mínima entre conjuntos de treinamento, entretanto, quando um método possui alta variância, pequenas alterações nos dados de treinamento podem resultar em variações significativas em $\hat{f}$. Em contraste, o viés se refere ao erro que é introduzido ao aproximar um problema da vida real, que pode ser extremamente complicado, por um modelo muito mais simples (Dutt; Chandramouli; Das, 2018). Nessa situação, não importa a quantidade de observação da base de treino, não será possível produzir uma estimativa acurada a partir de um modelo simples, se a verdadeira forma de f for demasiadamente complexa. De modo geral, modelos mais flexíveis resultam em menores valores de viés e, a partir da relação entre essas duas quantidades (viés e variância), dá-se que à medida que a flexibilidade de uma classe de métodos é aumentada, o viés tende a diminuir mais rapidamente do que a variância

aumentar e, consequentemente, o *MSE de teste* esperado diminui (Shalev-Shwartz; Ben-David, 2014; Dutt; Chandramouli; Das, 2018). No entanto, em algum grau específico de complexidade, aumentar a flexibilidade dos modelos tem pouco impacto no viés, mas, em contrapartida, a variância começa a aumentar significativamente, consequentemente, o *MSE de teste* aumenta (James et al., 2013; Shalev-Shwartz; Ben-David, 2014).

A relação entre essas duas quantidades e o erro quadrático médio no conjunto de teste é conhecida como *trade-off* entre viés e variância (Shalev-Shwartz; Ben-David, 2014). É um equilíbrio desejado durante o processo de treinamento e seleção dos modelos no aprendizado estatístico, ou seja, procura-se um modelo que seja suficientemente complexo para ajustar os dados de treinamento razoavelmente bem, mas não tão complexo a ponto de ser excessivamente sensível a pequenas variações nos dados de treinamento (Ng, 2018).

Apresentada a importância do equilíbrio entre viés e variância, são necessárias técnicas que auxiliem na avaliação da performance de modelos de aprendizado de máquina estatístico que levam em consideração esse equilíbrio (Ng, 2018). Três técnicas se destacam na literatura, sendo amplamente utilizadas em vários trabalhos de modelagem, são elas: a *abordagem de conjunto de validação*, *leave-one-out cross-validation* (*LOOCV*) e k*-fold cross-validation* (k*-fold CV*). A seguir será apresentado uma breve explicação sobre cada uma dessas abordagens.

Na *abordagem de conjunto de validação*, o banco de dados é dividido aleatoriamente em duas partes (geralmente 50%): um conjunto de treinamento e um conjunto de validação (teste), o modelo é ajustado no conjunto de treinamento, e posteriormente utilizado prever as respostas das observações no conjunto de validação. A taxa de erro resultante no conjunto de validação (*MSE de teste* no caso de uma resposta quantitativa) fornece uma estimativa da taxa de erro no teste. Essa abordagem, apesar de simples e de fácil implementação, possui algumas desvantagens. Observe que se o processo de dividir aleatoriamente a base de dados em duas partes for repetido, será obtido uma estimativa diferente de *MSE de teste*. Em adição, somente o conjunto de treino é utilizado para ajustar o modelo, e desde que os métodos estatísticos tendem a performar pior quando treinados em poucas observações, espera-se a taxa de erro do conjunto de validação superestime a taxa

de erro no teste quando comparado ao modelo ajustado no conjunto de dados completo (James et al., 2013).

A técnica *LOOCV* avalia o desempenho de um modelo treinado em um conjunto de dados ligeiramente diferente do original, removendo um ponto de dados por vez. Tal procedimento infere se o modelo em teste é excessivamente sensível aos dados de treinamento, ou seja, se possui alta variância. Então, ao invés de criar-se dois subconjuntos de tamanho semelhantes, como na *abordagem de conjunto de validação*, uma única observação (x_1, y_1) é usada para o conjunto de validação enquanto as demais observações $\{(x_2, y_2), \dots, (x_n, y_n)\}$ compõem o conjunto de treinamento. O método, então, será ajustado às $n - 1$ observações de treinamento e a previsão de $\hat{y}_1$ será realizada para a observação excluída a partir de x_1. Calcula-se então $\mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$, que apesar de baixo viés, apresenta alta variância, uma vez que seu valor é baseado em uma única observação. O procedimento é repetido para cada uma das demais observações do banco de dados, assim, a estimativa do *LOOCV* para o *MSE de teste* é a média dessas n estimativas de erro de teste:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i. \qquad (12)$$

No *LOOCV*, ajusta-se repetidamente o método de aprendizado estatístico usando conjuntos de treinamento com uma quantidade de observações semelhante à do conjunto de dados inteiro (apenas $n - 1$ menor). Consequentemente, a abordagem do *LOOCV* tende a não superestimar a taxa de erro no teste em comparação à *abordagem de conjunto de validação*. Outra diferença, em comparação ao observado na *abordagem de conjunto de validação*, é que realizar *LOOCV* várias vezes sempre produzirá os mesmos resultados pois não há aleatoriedade nas divisões entre conjunto de treinamento e validação. Contudo, a principal desvantagem do *LOOCV* ocorre quando se tem n muito grande ou quando o modelo individual demora para ser ajustado aos dados, nessas situações o custo computacional é alto, e o processo torna-se muito demorado.

Como alternativa ao *LOOCV* tem-se o, *k-fold CV*, que divide os dados em k grupos menores (ou dobras) de tamanho equivalentes, sempre com $k < n$ (Kroese et al., 2019) O primeiro *fold* é tratado como o conjunto de validação e ajusta-se o modelo nos $k - 1$ *folds* restantes e o MSE_1 é calculado nesse primeiro *fold* de validação (Shalev-Shwartz; Ben-David, 2014). Este procedimento é repetido k vezes

e em cada uma das vezes, um grupo diferente de observações é tratado como um conjunto de validação, resultando então em k estimativas do erro de teste, $MSE_1, MSE_2, \dots, MSE_k$. Então, a estimativa de k-*folds CV* é calculada pela média desses valores (Kroese et al., 2019):

$$\mathrm{CV}_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i. \tag{13}$$

Observe que o *LOOCV* é um caso específico do k-*folds CV*, em que k é ajustado para ser igual a n. Na prática, geralmente opta-se por realizar o k-*fold CV*, o que pode ser muito menos oneroso computacionalmente que o *LOOCV* que demanda o ajuste do método de aprendizado estatístico n vezes. De acordo com James et al. (2013), devido à variabilidade na forma como as observações são divididas em k diferentes conjuntos a variabilidade da estimativa do erro de teste de k-*fold CV* é geralmente menor do que a variabilidade nas estimativas de erro de teste resultantes da *abordagem de conjunto de validação*. Apesar dessa subestimação, a propriedade anteriormente observada associada ao aumento da complexidade (flexibilidade) dos modelos ainda é aplicável a essa abordagem. Ou seja, apesar das subestimações nos valores de *MSE de teste*, ainda é possível identificar o ponto mínimo na curva estimada do *MSE de teste*, quando diferentes modelos, ou complexidades, são comparados.

Devido a equilíbrio entre viés e variância, a técnica k-*fold CV* frequentemente fornece estimativas mais precisas da taxa de erro de teste do que o *LOOCV*. Sendo a técnica recomendada, e amplamente utilizada nos trabalhos de aprendizado de máquina observados na literatura científica. Levando em consideração o viés, a *abordagem de conjunto de validação* pode levar a superestimações da taxa de erro de teste, como apresentado anteriormente. Já, a técnica *LOOCV* fornecerá estimativas não viesadas do erro de teste, uma vez que cada conjunto de treinamento contém $n - 1$ observações, o que é quase o mesmo número de observações no conjunto de dados original. Por sua vez, o k-*fold CV* levará a um nível intermediário de viés, uma vez que cada conjunto de treinamento contém menos observações do que na abordagem do *LOOCV*, mas substancialmente mais do que na *abordagem de conjunto de validação*. Portanto, quanto à redução do viés, *LOOCV* é preferível ao k-*fold CV*. Em contrapartida, a variâncias desses procedimentos também devem ser consideradas nesse equilíbrio. De acordo com James et al. (2013), No *LOOCV*

calcula-se a média dos resultados gerados por n modelos ajustados. Cada um desses modelos é treinado em um conjunto de observações praticamente idêntico, resultando em uma alta correlação positiva entre suas saídas. Em contraste, no k-*fold CV* com, calcula-se a média das saídas de k modelos ajustados que exibem uma correlação um pouco menor, quando comparada àquela do *LOOCV*. Essa redução na correlação é devida à menor sobreposição entre os conjuntos de treinamento usados em cada modelo. Uma vez que a média de muitas quantidades altamente correlacionadas possui uma variância maior do que a média de muitas quantidades não tão correlacionadas, a estimativa de erro de teste resultante do *LOOCV* tende a ter uma variância mais elevada do que a estimativa de erro de teste resultante do k-*fold CV* (James et al., 2013). Recomenda-se que a validação cruzada k-*fold* seja realizada usando $k = 5$ ou $k = 10$, pois esses valores foram demonstrados empiricamente como aqueles que fornecem estimativas de taxa de erro de teste que não sofrem de viés variância excessivamente altos (James et al., 2013; Izbicki; Santos, 2020).

2.4 Aprendizado de máquina estatístico no mapeamento do solo

A área da Geociência em muito se beneficiou com esses avanços. Algoritmos de aprendizado de máquina supervisionado, como as máquinas de vetores de suporte (*support vector machines* - SVM), árvores de classificação e regressão (CARTs) e redes neurais artificiais (RNAs) são consideradas ferramentas promissoras e robustas, para interpretar dados de alta variabilidade e indicar relações não-lineares no espaço multivariado (McBratney et al., 2000; Warner et al., 2019). Estas técnicas têm sido satisfatoriamente empregadas no desenvolvimento de modelos preditivos para diversas variáveis nas ciências climatológicas, ambientais e agrárias (Yilmaz; Kaynar, 2011; Besalatpour et al., 2013; Hengl et al., 2017; Hengl et al., 2018; Reichstein et al., 2019; Shadrin et al., 2020). Como exemplos recentes de aplicações dos algoritmos de aprendizado de máquina temos: previsão e mapeamento de vários atributos do solo, como o teor de carbono orgânico e o nitrogênio total (Hengl et al., 2017; Hengl et al., 2018; Heuvelink et al., 2020; Zhou et al., 2020), potencial de sequestro de carbono do solo (Mahmoudzadeh et al., 2020) e estimativas de produtividade de culturas agrícolas (Bocca; Rodrigues, 2016; Ferraciolli; Bocca;

Rodrigues, 2019; Sanches; Graziano Magalhães; Junqueira Franco, 2019; Yang; Zhao; Cai, 2020).

Muitos estudos têm sido conduzidos para analisar e compreender a relação entre a emissão de GEE e o aumento da temperatura global usando abordagens estatísticas convencionais. Entretanto, essas técnicas seguem suposições de modelagem probabilística, em que os resultados podem ser associados a incertezas de várias magnitudes (Khan; Khan, 2019). As técnicas de aprendizado de máquina oferecem grande potencial para refinar as estimativas a respeito ciclo de carbono em escala global (Tramontana et al., 2016; Reichstein et al., 2019). Essas técnicas permitem a superação dos desafios estatísticos impostos por um número cada vez maior de covariáveis espaciais disponíveis, muitos das quais podem estar correlacionados entre si, ou ter relações não-lineares, incluso a dependência espacial, violando a importante suposição de dados distribuídos de forma idêntica e independente impostos pela abordagem probabilística (Reichstein et al., 2019; Warner et al., 2019).

De tal modo, na última década foi observado um aumento da utilização das técnicas de aprendizado de máquina para criação de modelos preditivos da emissão de CO_2 do solo em diferentes localidades do planeta (Freitas et al., 2018; Vargas et al., 2018; Tang et al., 2020). Utilizando RNAs para prever FCO_2 em áreas de cana-de-açúcar, Farhate et al. (2018a), observaram que a arquitetura de rede *Multilayer Perceptron* (MLP) com o método de seleção de atributos *Wrapper* apresentou alta precisão para classificação das taxas de perda de carbono no solo via FCO_2. Similarmente, Freitas et al. (2018), obtiveram resultados satisfatórios ao estimarem a variabilidade espaçotemporal de FCO_2 em áreas de cultivo de cana-de-açúcar por meio de rede MLP e pelo algoritmo de retro propagação do erro, com ótimo desempenho preditivo, indicado pelos valores preditos próximos aos valores medidos experimentalmente, conforme o MAPE (18,29%) e o coeficiente de determinação obtidos ($R^2 = 0,92$).

Apesar de grandes avanços na área, as abordagens de aprendizado de máquina atuais podem não ser ideais quando o comportamento do sistema é dominado pelo contexto espaçotemporal (Song et al., 2017; Ferraciolli; Bocca;

Rodrigues, 2019), como no caso de fluxos de GEE para a atmosfera e da dinâmica de gases no solo (FCO_2).

Na estatística clássica, assume-se que a variabilidade ao redor da média é aleatória e espacialmente independente. Contudo, a variabilidade de vários atributos do solo, frequentemente, apresenta uma componente espacialmente dependente. Em outras palavras, essa variabilidade pode ser descrita como uma função da distância de separação entre as amostras, a independência entre elas não é verificada pois o atributo apresenta autocorrelação espacial (Trangmar; Yost; Uehara, 1985; Isaaks; Srivastava, 1989). A análise geoestatística fornece as bases para uma descrição quantitativa da variação espacial de um atributo no solo e pode ser utilizada para estimativas desse atributo em locais não amostrados, gerando os mapas de padrões espaciais (Webster, 1985; Webster; Oliver, 1990). Nos últimos anos, vários trabalhos utilizando técnicas geoestatísticas têm sido conduzidos no Brasil com o objetivo de entender a variabilidade espacial da FCO_2, gerado informações que auxiliem na escolha de práticas agrícolas mais sustentáveis para a mitigação das emissões de GEE (La Scala et al., 2000; La Scala et al., 2009; Panosso et al., 2009a; Panosso et al., 2012; Teixeira et al., 2013a; Bicalho et al., 2014; Leon et al., 2014; Mantovanelli et al., 2016; da Cunha et al., 2018; Tavanti et al., 2020b).

Nesse contexto a união de técnicas geoestatísticas e aprendizado de máquina é promissora para a melhoria do processo de previsão dos atributos do solo (Hengl et al., 2017; Hengl et al., 2018), pois no processo de aprendizado, a autocorrelação espacial deve ser levada em consideração. Ferraciolli; Bocca; Rodrigues (2019) utilizando diferentes algoritmos de aprendizado de máquina na modelagem da produtividade da cana-de-açúcar no Brasil, concluíram que negligenciar a autocorrelação espacial causa subestimação do erro dos modelos de produtividade, causando inconsistências na etapa de seleção de modelos. Em estudo realizado na região central da China (província de Shaanxi), Song et al. (2017) propuseram um método estatístico híbrido de krigagem ordinária (KO) e aprendizado de máquina para prever a variabilidade espacial do teor de matéria orgânica do solo (MOS) utilizando covariáveis proveniente de sensoriamento remoto e fatores como topografia, clima e atributos do solo. Os autores concluíram que o modelo híbrido apresentou melhor desempenho quando comparado aos modelos de regressão linear múltipla e mesmo

redes neurais artificiais. Os mesmos autores destacam que a aplicação de dados de múltiplas fonte, como imagens de sensoriamento remoto e covariáveis ambientais expressam uma relação de hierarquia não-linear e multidimensional no entendimento da variabilidade espacial do atributo em questão, tornando possível a construção de mapas de alta qualidade, mais precisos e acurados quando comparados aos construídos pela estatística clássica, ou somente a geoestatística.

3 MATERIAL E MÉTODOS

3.1 Construção da base de dados

O presente trabalho compila resultados anteriores e alguns novos, sobre a modelagem espaçotemporal da emissão de CO_2 do solo (FCO_2), e atributos associados, proveniente de 19 ensaios de campo realizados em diferentes regiões do Brasil, especificamente, nos estados de São Paulo e Mato Grosso do Sul (Tabela 1). Os ensaios fizeram parte de trabalhos de iniciação científica, mestrado e doutorado conduzidos nos últimos 21 anos por alunos e professores da Universidade Estadual Paulista "Júlio de Mesquita Filho" – UNESP, dos Câmpus de Jaboticabal e Ilha Solteira (Figura 3). A variabilidade espacial da FCO_2 foi avaliada em alguns ensaios juntamente à variabilidade temporal, avaliada em todos os experimentos realizados. As principais informações a respeito das localidades onde os ensaios foram realizados, como o clima, o tipo de solo o uso do solo, entre outras, encontra-se listadas na Tabela 1. Os experimentos selecionados representam usos da terra típicos da região específica onde os ensaios foram conduzidos, bem como onde alguns processos envolvidos na mudança do uso da terra ocorreram (La Scala Jr; Panosso; Pereira, 2003; La Scala Jr.; Panosso; Pereira, 2003; La Scala; Bolonhezi; Pereira, 2006; Panosso et al., 2008; La Scala et al., 2009; Panosso et al., 2009a; Panosso et al., 2011; Panosso et al., 2012; Moitinho et al., 2013; Teixeira et al., 2013b; Bicalho et al., 2014; Bicalho et al., 2017; de Figueiredo et al., 2017; Almeida et al., 2018; Pinheiro da Silva et al., 2019; Vicentini et al., 2019; Canteral et al., 2023a; Vicentini et al., 2023).

A base de dados, contendo 15.397 observações e 39 variáveis pode ser encontrada no endereço: https://github.com/arpanosso/tese-fco2-ml-2023. Durante a condução do estudo foi desenvolvido um pacote personalizado em R (R Development Core Team, 2023) para facilitar a divulgação e análise dos resultados obtidos. Pacotes, ou bibliotecas, podem ser definidos como coleções de procedimentos, exemplos e documentação de funções implementadas em linguagem R. Esse pacote, nomeado `fco2r`[1], permite a visualização dos dados, a execução de análises estatísticas avançadas e a geração de gráficos interativos para tornar os resultados mais acessíveis e compreensíveis para a comunidade científica, alunos e profissionais interessados no campo da modelagem da emissão de CO_2 do solo (`fco2r::`

[1] https://github.com/arpanosso/fco2r

`data_fco2`). O pacote representa uma contribuição significativa para a disseminação das aplicações deste estudo e está disponível para uso público, promovendo a reprodutibilidade e a colaboração em pesquisa (Hanson; Sugden; Alberts, 2011). É importante ressaltar que o ambiente de desenvolvimento R é uma linguagem livre amplamente reconhecida e utilizada na comunidade científica, especialmente, na área de estatística, ciência de dados e aprendizado de máquina. A escolha do R para criar este pacote não apenas facilita a disseminação dos resultados, mas também promove a transparência e a replicabilidade, alinhadas com os princípios fundamentais científicos atrelados às linguagens de código aberto (Ince; Hatton; Graham-Cumming, 2012).

Tabela 1. Informações geográficas, clima e histórico das áreas onde foram conduzidos os experimentos de variabilidade espaçotemporal da emissão de CO_2 e atributos físicos e químicos do solo ao longo dos últimos 21 anos, juntamente com suas respectivas publicações.

Cidade	Estado	Coordenadas Geográficas	Elevação (m)[1]	Clima[2]	Cultura	Manejo	TC	Ano	Solo	N	Referência Bibliográfica
Jaboticabal	SP	21º15' S; 48º18' O	580	Aw	Milho/Soja	Convencional	31	2001	Latossolo Vermelho eutrófico	65	(La Scala Jr.; Panosso; Pereira, 2003; La Scala et al., 2009)
Jaboticabal	SP	21º14' S; 48º17' O	605	Aw	Milho/Soja	Mínimo	33	2003	Latossolo Vermelho eutrófico	70	(Panosso et al., 2006)
Jaboticabal	SP	21º14' S; 48º17' O	605	Aw	Milho/Soja	Mínimo	34	2004	Latossolo Vermelho eutrófico	70	(La Scala; Bolonhezi; Pereira, 2006)
Jaboticabal	SP	21º14' S; 48º17' O	614	Aw	Feijão	Convencional	34	2004	Latossolo Vermelho eutrófico	64	(La Scala et al., 2005; Panosso et al., 2006)
Guariba	SP	21º19' S; 48º13' O	600	Aw	Cana-de-açúcar	Cana queimada	30	2005	Latossolo Vermelho eutroférrico	60	(Panosso et al., 2008)
Guariba	SP	21º19' S; 48º13' O	600	Aw	Cana-de-açúcar	Cana crua	10	2005	Latossolo Vermelho eutroférrico	60	(Panosso et al., 2008)
Rincão	SP	21º24' S; 48º09' O	550	Aw	Cana-de-açúcar	Cana crua	7	2007	Latossolo Vermelho eutroférrico	60	(Panosso et al., 2009b; Panosso et al., 2011)
Rincão	SP	21º24' S; 48º09' O	550	Aw	Cana-de-açúcar	Cana queimada	30	2007	Latossolo Vermelho eutroférrico	60	(Panosso et al., 2009b; Panosso et al., 2011)
Rincão	SP	21º24' S; 48º09' O	550	Aw	Cana-de-açúcar	Cana crua	7	2008	Latossolo Vermelho eutroférrico	89	(Panosso et al., 2012)
Guariba	SP	21º21' S; 48º11' O	620	Aw	Cana-de-açúcar	Cana crua	8	2010	Latossolo Vermelho eutroférrico	141	(Teixeira et al., 2011a; Teixeira et al., 2011b; Bicalho et al., 2014)
Dourados	MS	21º14' S; 54º49' O	452	Am	Cana-de-açúcar	Cana crua	5	2011	Latossolo Vermelho distroférrico	45	(Moitinho et al., 2013; Moitinho et al., 2015a; Moitinho et al., 2015b)

TC = Tempo de conversão; N = número de pontos amostrais; Ano = ano de condução do experimento. [1] Elevação em metros acima do nível do mar. [2] Classificação climática de acordo com Köppen.

Tabela 1 (Continuação). Informações geográficas, clima e histórico das áreas onde foram conduzidos os experimentos de variabilidade espaçotemporal da emissão de CO_2 e atributos físicos e químicos do solo ao longo dos últimos 21 anos, juntamente com suas respectivas publicações.

Cidade	Estado	Coordenadas Geográficas	Elevação (m)[1]	Clima[2]	Cultura	Manejo	TC	Ano	Solo	N	Referência Bibliográfica
Pradópolis	SP	21º20' S; 48º08' O	515	Aw	Cana-de-açúcar	Cana crua	15	2012	Latossolo Vermelho eutroférrico	133	(Bicalho et al., 2017)
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Milho/Soja/Feijão	Plantio Direto	10	2013	Latossolo Vermelho distroférrico	133	(Terçariol et al., 2016)
Mococa	SP	21º21' S; 47º04' O	673	Aw	Pastagem	Manjada	3	2013	Latossolo Vermelho eutroférrico	102	(de Figueiredo et al., 2016)
Mococa	SP	21º21' S; 47º04' O	673	Aw	Pastagem	Degradada	30	2013	Latossolo Vermelho eutroférrico	102	(de Figueiredo et al., 2016)
Aparecida do Taboado	MS	20º19' S; 51º13' O	380	Aw	Cana-de-açúcar	Cana crua	5	2014	Latossolo Vermelho distroférrico	102	(Terçariol et al., 2016; Almeida et al., 2018)
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Eucalipto	Reflorestamento	29	2015-2017	Latossolo Vermelho distroférrico	102	(Vicentini et al., 2019; Canteral et al., 2023a; Vicentini et al., 2023)
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Pinus	Reflorestamento	29	2015-2017	Latossolo Vermelho distroférrico	15	Vicentini et al., 2019; Canteral et al., 2023; Vicentini et al., 2023)
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Silvipastoril	iLPF	29	2015-2017	Latossolo Vermelho distroférrico	15	---
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Nativas Cerrado	Reflorestamento	29	2015-2016	Latossolo Vermelho distroférrico	15	(Vicentini et al., 2019)
Selvíria	MS	20º20' S; 51º24' O	362	Aw	Cerrado	Mata nativa	60	2017	Latossolo Vermelho distroférrico	15	Vicentini et al., 2019; Canteral et al., 2023; Vicentini et al., 2023)
Selvíria	MS	20º20' S; 51º24' O	335	Aw	Pastagem	Degradada	31	2017-2019	Latossolo Vermelho Amarelo distrófico	86	(Tavanti et al., 2020a)
Selvíria	MS	20º20' S; 51º24' O	335	Aw	Pastagem	Manejada	31	2017-2019	Latossolo Vermelho Amarelo distrófico	86	(Tavanti et al., 2020a)

TC = Tempo de conversão; N = número de pontos amostrais; Ano = ano de condução do experimento. [1] Elevação em metros acima do nível do mar. [2] Classificação climática de acordo com Köppen.

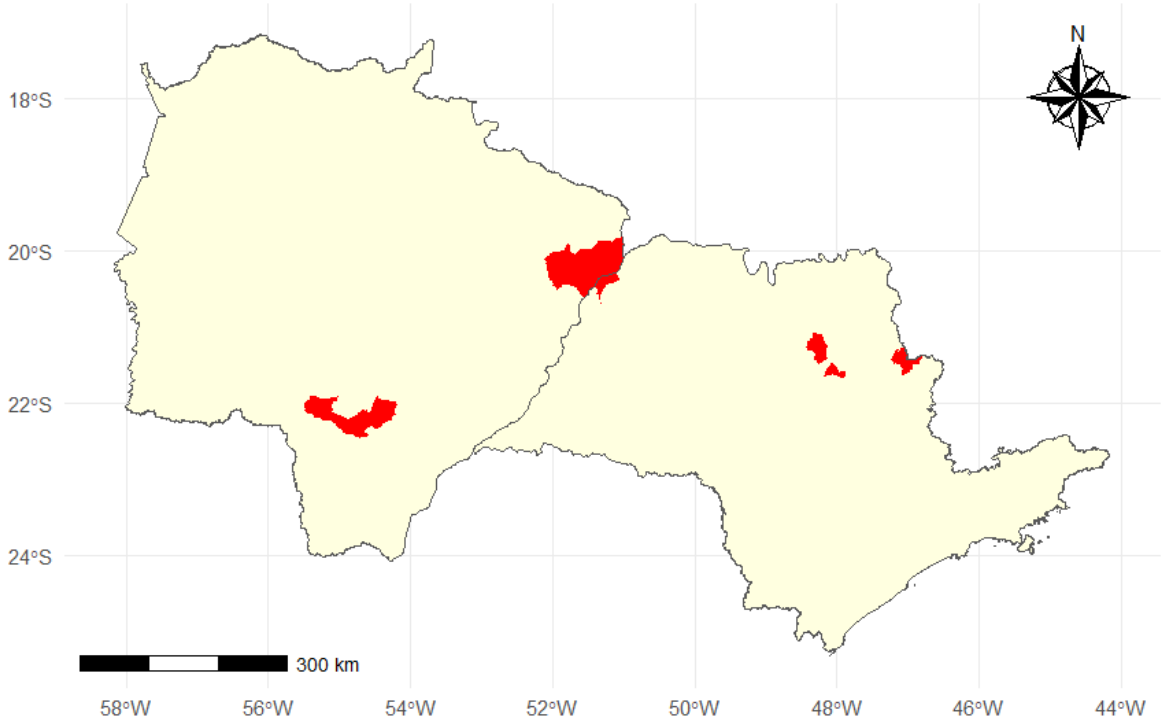

Figura 3. Distribuição espacial das áreas experimentais que compuseram a base de dados de emissão de CO_2 e atributos do solo para aplicação dos modelos de aprendizado de máquina estatísticos.

3.2 Aquisição de dados de sensoriamento remoto X_{CO2} e SIF

Os dados de concentração de CO_2 atmosférico (X_{CO2}) e fluorescência de clorofila induzida pelo sol (SIF) foram obtidos a partir do satélite *Orbiting Carbon Observatory*-2 (OCO-2) para todo o território brasileiro no período de 2014 a 2020 (Figura 4). O sensor orbital mediu indiretamente a concentração de CO_2 atmosférico por meio da intensidade da radiação solar refletida, em função da presença de dióxido de carbono (CO_2) em uma coluna de ar. Essas leituras foram realizadas em três faixas de comprimento de onda: a do oxigênio (O_2), na faixa de 0,757 a 0,775 µm, e as do CO_2, que são subdividas em banda fraca (1,594 – 1,627 µm) e banda forte (2,043 – 2,087 µm). A fluorescência de clorofila induzida pelo sol foi obtida devido à sobreposição que ocorre nas faixas de comprimento de onda do SIF com a faixa de absorção de O_2 (680 - 850 nm) (Crisp et al., 2012; O'Dell et al., 2012; Mohammed et al., 2019). Quanto ao SIF, levou-se em consideração uma combinação dos comprimentos e onda de 757 nm e 775 nm devido a estudo anterior realizado em escala global onde foram utilizados algoritmos de aprendizado de máquina, como redes neurais artificiais (ANN), para treinar as observações nativas de SIF do OCO-2

com a reflectância de superfície de sete bandas corrigidas pelo MODIS BRDF (Yu et al., 2019). A metodologia de aquisição dos dados de X_{CO2} e SIF, bem como o tratamento inicial dos dados, pode ser acessada em repositório GitHub disponível no endereço https://github.com/arpanosso/projetofinal_r4ds2. Em adição, essa base de dados também se encontra disponível no pacote `fco2r`, anteriormente apresentado, e possui 37.387 observações e 18 colunas (`fco2r::oco2_br`).

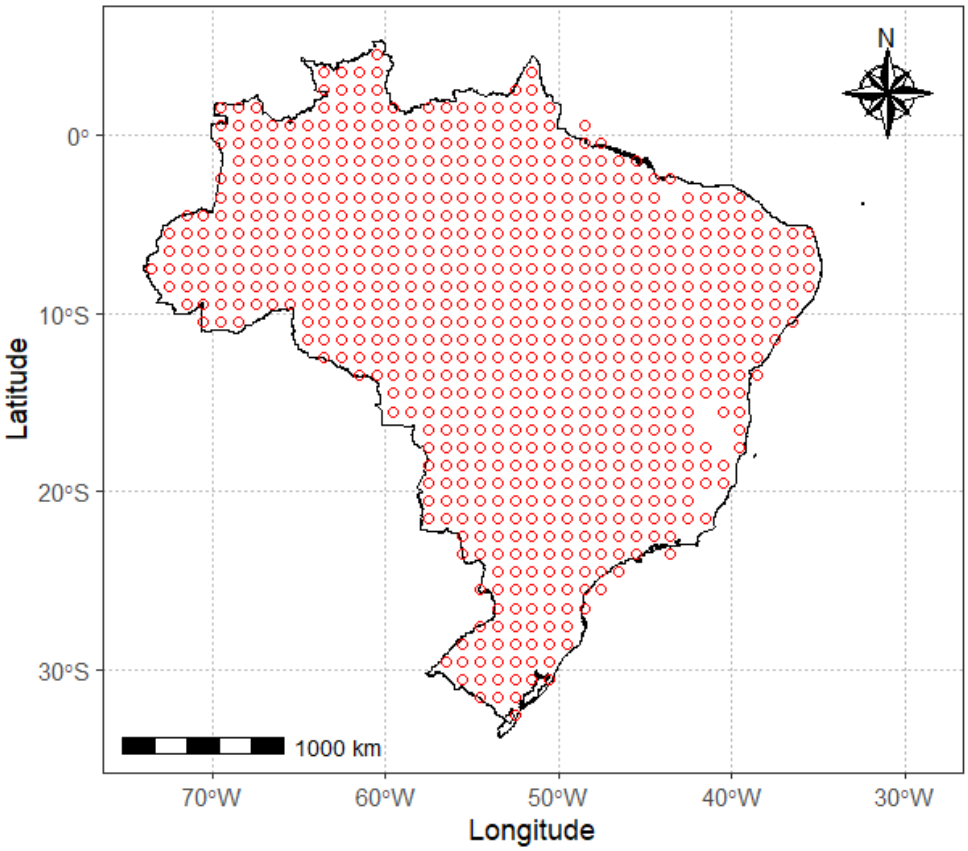

Figura 4. As localizações geográficas representadas pelos círculos em vermelhos no mapa (amostra aleatória de 1000 observações) indicam os locais onde as leituras de X_{CO2} e SIF foram obtidas pelo satélite *Orbiting Carbon Observatory-2* (OCO-2) da NASA, no período de 2014 a 2020.

3.3 Aquisição de dados da estação agrometeorológica

Em adição, as seguintes variáveis ambientais foram selecionadas: velocidades do vento- (Velmax e Velmin em m s^{-1}); evapotranspiração por Penman_Monteith (Eto em mm dia^{-1}); radiação fotossinteticamente ativa (PAR em µmol m^{-2} s^{-1}); radiação global (Rad em MJ m^2 dia $^{-1}$); pressão atmosférica (Pkpa em kpa); temperaturas do ar média, máxima e mínima (Tmed, Tmax e Tmin,

respectivamente, em ºC), precipitação atmosférica (chuva em mm), umidade relativas do ar média, máxima e mínima (Umed, Umax e Umin, respectivamente, em %), insolação (inso em h dia $^{-1}$). A radiação (PAR) é a radiação incidente na faixa de ondas de 400 a 700 nm, que pode ser absorvida pelo sistema fotossintético das plantas. Seu valor é responsável por aproximadamente 50% da radiação solar. As variáveis foram adquiridas a partir de estação agrometeorológica localizada no município de Ilha Solteira-SP.

3.4 Determinação da emissão de CO_2, temperatura e umidade do solo

Em todos os ensaios a emissão de CO_2 do solo (FCO_2) foi registrada por meio do sistema LI-COR (LI-8100 – Figura 5a), que em seu modo de medição monitora as mudanças na concentração de CO_2 dentro da câmara para solos (Figura 5b) por meio de espectroscopia na região do infravermelho. A câmara para solos apresenta volume interno de 854,2 cm^3 com área de contato circular com o solo de 83,7 cm^2. Essa câmara é acoplada sobre colares de PVC (Figura 5c) previamente inseridos no solo em cada ponto amostral na profundidade de 0,03 m. A emissão de CO_2 (ou fluxo de CO_2) foi avaliado em cada ponto por um ajuste da concentração de CO_2 do ar dentro da câmara em função de uma regressão exponencial no tempo após o fechamento da câmara. A temperatura do solo foi monitorada, concomitantemente às avaliações de respiração do solo, utilizando-se um sensor de temperatura que é parte integrante do sistema ao LI-8100. Tal sensor consiste em uma haste de 0,2 m que foi inserida no interior do solo próximos ao local onde foram instalados os colares de PVC para a avaliação. Em alguns ensaios, a temperatura do solo foi determinada com auxílio de um termômetro digital do tipo espeto (Figura 5d). A umidade do solo foi determinada por meio de um equipamento de TDR (*Time Domain Reflectometry - Hydrosense* TM, *Campbell Scientific*, Austrália - Figura 4 e). O aparelho de TDR é constituído por uma sonda, apresentando duas hastes de 0,12 m, que devem ser inseridas no interior do solo próximos aos colares de PVC.

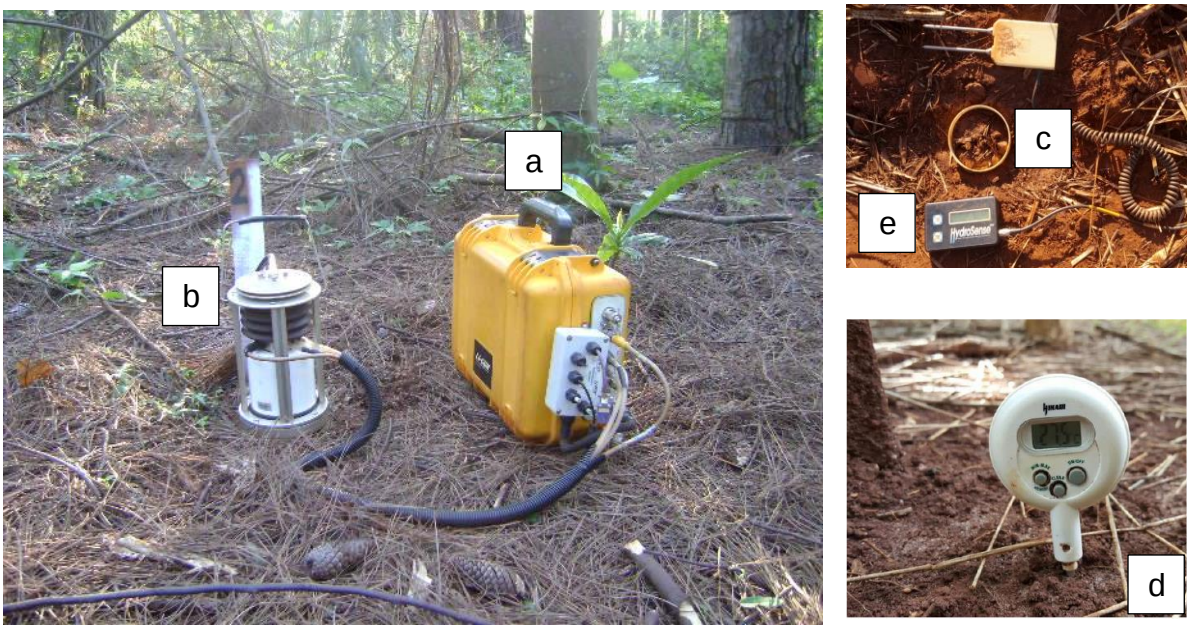

Figura 5. Sistema LI-8100 (a) interligado à câmara para solos (b), câmara para solo é inserida sobre o colar de PVC (c), sensor de temperatura do solo do tipo espeto (d), TDR - *Hydrosense system*, sistema portátil utilizado para avaliação da umidade do solo (e).

3.5 Determinação dos atributos do solo

Os dados referentes aos atributos físico e químicos do solo foram determinados na profundidade de 0 a 0,20 m. Foram realizadas as seguintes análises de rotina: pH, determinação do teor de matéria orgânica do solo (MO), teor de Fósforo disponível (P), Potássio (K), Cálcio (Ca), Magnésio (Mg), soma de bases (SB), acidez potencial (H_Al). Os teores de cálcio, magnésio e potássio trocáveis e fósforo disponível foram extraídos utilizando-se o método da resina trocadora de íons (Raij, 2001); a capacidade de troca de cátions (CTC) e a saturação por base (V) também foram calculados. O carbono orgânico foi determinado pelo método da combustão úmida, via colorimétrica (Raij et al., 1987), a determinação do teor de nitrogênio foi realizada por meio de digestão sulfúrica (Malavolta; Vitti; Oliveira, 1997). As amostras indeformadas foram coletadas com amostrador adaptado a cilindros com dimensões médias de 0,05 m de diâmetro interno e 0,04 m de altura (EMBRAPA, 1997). A macroporosidade (Macro) e microporosidade (Micro) foram determinadas por meio de mesa de tensão com 0,60 m de altura de coluna d'água em amostras previamente saturadas. O volume de água retido na amostra nesta condição corresponde à Macro.

A Micro foi determinada após a retirada dos anéis da estufa à 105 ºC num período de 24 h e posterior pesagem. Já a porosidade total (volume total de poros - VTP) foi calculada pela soma dos macroporos e microporos. A porosidade livre de água (PLA) foi calculada como a diferença entre o volume total de poros e a fração de porosidade preenchida com água, que é equivalente à umidade do solo (Us). Também foi determinado a densidade do solo (Ds) utilizando-se a seguinte equação (EMBRAPA, 1997):

$$Ds = \frac{MS_{se}}{Va}, \quad (14)$$

em que MS_{se} é massa de solo seco em estufa e Va é volume do anel.

Uma vez que as amostras de solo foram coletadas em uma camada fixa, os estoques de carbono do solo (EstC) foram ajustados para as mudanças na densidade do solo (Ds) que ocorrem após as mudanças do uso da terra. Para isso, foi utilizada a metodologia descrita por Ellert; Bettany (1995) e Moraes et al. (1996) para corrigir os estoques de carbono do solo em uma profundidade de massa equivalente, ou seja, a profundidade do solo das áreas manejadas que contém a mesma massa de solo como a camada correspondente (0 – 0,20 m) na área de vegetação nativa (área de referência específica para cada região). Os cálculos da camada de solo equivalente, foram realizados de acordo com o apresentado por Carvalho et al. (2009) e Segnini Segnini et al. (2013).

$$CE = \frac{M_{CE}}{M_{área}} 20 \ , \quad (15)$$

em que CE é a camada de solo equivalente em centímetro, M_{CE} é a média ponderada da densidade do solo (*Ds*) nas respectivas camadas de solo na área de vegetação nativa, específica para cada ensaio. $M_{área}$ é a média ponderada de *Ds* nas respectivas camadas do solo em cada área e, o valor 20 é relacionado com a profundidade do solo de 0 a 20 cm na área de referência (vegetação nativa específica). O estoque de carbono (Mg ha^{-1}) foi calculado multiplicando a concentração de Carbono (%) pela densidade do solo *Ds* (g cm^{-3}) e pela espessura da camada de solo equivalente (cm).

Para a determinação do grau de humificação da matéria orgânica do solo e do teor de carbono das amostras de solos, foram utilizadas as análises de fluorescência induzida por laser (LIFS) e espectroscopia de emissão óptica com plasma induzido por laser (LIBS), respectivamente. Os espectros LIBS foram capturados utilizando

um sistema comercial modelo LIBS2500, da *Ocean Optics* (USA). LIBS é uma técnica analítica avançada para análise elementar semi-quantitativa, baseada na medida da emissão de espécies excitadas em um plasma produzido por um laser (Ferreira et al., 2009). A técnica de fluorescência induzida por laser (LIFS) tem como princípio básico a excitação das amostras do solo com um laser de emissão, na região do ultravioleta/azul, resultando na fluorescência de grupos funcionais da matéria orgânica, relacionados com o processo de humificação (Milori et al., 2006). A fluorescência total (área em baixo da curva) correlaciona-se aos teores de carbono do solo e, quando ponderada a partir dos teores de C orgânico da amostra (obtidos pelo LIBS), traz informações a respeito do grau de humificação da matéria orgânica. Os espectros LIFS foram capturados utilizando um sistema montado pela Embrapa Instrumentação Agropecuária. A área do espectro LIFS de cada amostra de solo foi dividida pelo teor de carbono correspondente, obtida no LIBS, calculando-se assim os sinais de fluorescência normalizada, que então foram definidos como o grau de humificação da matéria orgânica do solo (HLIFS).

3.6 Formas de análises dos resultados.

Todas as análises estatísticas e as etapas de seleção de variáveis, pré-processamento, aprendizado de máquina estatístico, validação e avaliação do desempenho dos modelos, foram realizadas a partir de implementações em linguagem R (R Development Core Team, 2023) utilizando o ambiente de desenvolvimento RStudio (Posit team, 2023). Foram utilizados os meta-pacotes `tidyverse` (Wickham et al., 2019) e `tidymodels` (Kuhn; Wickham, 2020) e os pacotes: `ggstats` (Larmarange, 2023), `sp` (Pebesma; Bivand, 2015), `geobr` (Pereira et al., 2023), `skimr` (Waring et al., 2022), `geoR` (Diggle; Ribeiro, 2010), `patchwork` (Pedersen, 2023), `ggspatial` (Dunnington; Thorne; Hernangómez, 2023), `parsnip` (Kuhn et al., 2023a), `vip` (Greenwell; Boehmke, 2023), `future` (Bengtsson, 2023), `visdat` (Tierney, 2017), `corrplot` (Wei; Simko, 2021), `Metrics` (Hamner; Frasco; LeDell, 2018), e `caret` (Kuhn et al., 2023b). Muitos desses pacotes são conjuntos de funções implementadas para simplificação de todo o processo de criação e avaliação dos modelos preditivos além de muitos serem recomendados para o pré-processamento e visualização dos resultados.

Para realizar todas as etapas das análises foram utilizados três microcomputadores (Figura 6) com as seguintes configurações de hardware: Aspire Nitro 5 Notebook (processador AMD Ryzen 7 4800H com Radeon Graphics 16 CPUs, 2.9GHz, memória RAM de 15742MB, GPU discreta AMD Radeon(TM) Graphics, armazenamento em unidade de estado sólido (SSD) de 931GB e sistema operacional: Windows 11 Pro 64-bit versão 10.0 – Figura 6a); Samsung Notebook (processador Intel(R) Core(TM) i7-5500U 4 CPUs, 2.40GHz, memória RAM de 8086MB, GPU integrada NVIDIA GeForce 920M, armazenamento em unidade de estado sólido (SSD) de 446GB e sistema operacional: Windows 10 Pro 64-bit Versão 10.0 – Figura 6b; LG All in one (processador Intel(R) Core(TM) i5-5200U 4 CPUs, 2.20GHz, memória RAM de 4002MB, GPU integrada Intel(R) HD Graphics 5500, armazenamento em unidade de estado sólido (SSD) de 446GB e sistema operacional Windows 10 Pro 64 bits Versão 10.0 – Figura 6c).

Figura 6. Três microcomputadores utilizados na pesquisa: Aspire Nitro 5 Notebook (a); Samsung Notebook (b) e LG All in one (c), os quais desempenharam um papel fundamental no treinamento de modelos de aprendizado de máquina para análise de dados.

As etapas do processo de aprendizado de máquina estatístico são apresentadas na Figura 7. O protocolo adotado foi adaptado de Chapman et al. (2000), constituído das seguintes etapas:

Compreensão: definição dos objetivos e necessidades iniciais do estudo que se pretende atender com a modelagem e mineração de dados, compreende as etapas iniciais do projeto;

Entendimento: os dados foram de diversas fontes e localidades, alguns provenientes de sensores em diferentes formatos, assim, compreendeu-se as estruturas dos diferentes tipos de dados junto às suas especificidades com o principal objetivo de identificação de problemas e testes da qualidade das bases. Na modelagem espacial, apenas os atributos do solo foram considerados como *features* (covariáveis). Para a modelagem temporal, foram empregadas três estratégias de agrupamento de covariáveis (recortes do banco de dados), denominada configurações de *features*: atributos do solo (Solo) compreendendo o período de 2001 a 2020, atributos do solo e sensoriamento remoto (Solo + Sensoriamento Remoto), compreendendo o período de 2015 a 2019 e atributos do solo e sensoriamento remoto e dados da estação agrometeorológica (Solo + Sensoriamento Remoto + Estação Agrometeorológica) compreendendo o período de 2015 a 2019. Salientamos que esse recorte temporal foi necessário uma vez que o sensor orbital *Orbiting Carbon Observatory*-2 iniciou suas atividades a partir do mês de setembro de 2014.

c) *Pré-processamento*: incorporação dos dados no banco de dados final e separação dos dados dos ensaios espaciais e temporais. Para isso, a estrutura do banco de dados foi modificada, para o ajuste da resolução temporal de cada banco de dados. Algumas covariáveis foram eliminadas, algumas foram incluídas e foram realizadas mudanças de escala, eliminação de inconsistências e dados faltantes. As variáveis categóricas cultura, manejo e cobertura do solo, foram transformadas em variáveis binárias, permitindo que essas variáveis fossem incorporadas pelos algoritmos de aprendizado estatístico utilizados. Cada categoria presente nas variáveis categóricas foi transformada em uma variável binária separada (0 ou 1), onde 0 representou a ausência da categoria e 1 representou a presença da categoria. Isso foi feito para evitar que o algoritmo intérprete erroneamente a ordem ou importância das categorias e para permitir a inclusão dessas variáveis nos modelos de maneira apropriada. Ainda nessa etapa foi realizado a construção dos mapas de padrões espaciais para os ensaios geoestatísticos. Finalmente, o banco de dados foi dividido em duas partes para a modelagem espacial e temporal.

Modelagem: etapa de aprendizado estatístico propriamente dita. Atendimento das exigências específicas de cada técnica de aprendizado e realizada de forma independente para cada conjunto (espacial e temporal) e para cada estratégia de agrupamento de variáveis;

Avaliação: após o desenvolvimento do modelo, foram revisados os processos e verificado se os objetivos foram atingidos para ao final da etapa ser tomada a decisão de utilização ou não dos resultados. Para avaliação dos modelos, foi feita a separação do banco de dados em dois conjuntos, sendo um para o processo de aprendizado (treinamento) e outro para o processo de avaliação (teste) dos modelos gerados a qual feita pelo processo k-fold CV com $k = 5$;

Implantação: proposição de aplicação e distribuição do conhecimento gerado durante o processo. Os resultados publicados de forma aberta para que este possa ser utilizado pela comunidade acadêmica e científica no repositório desenvolvido disponível em: https://github.com/arpanosso/tese-fco2-ml-2023.

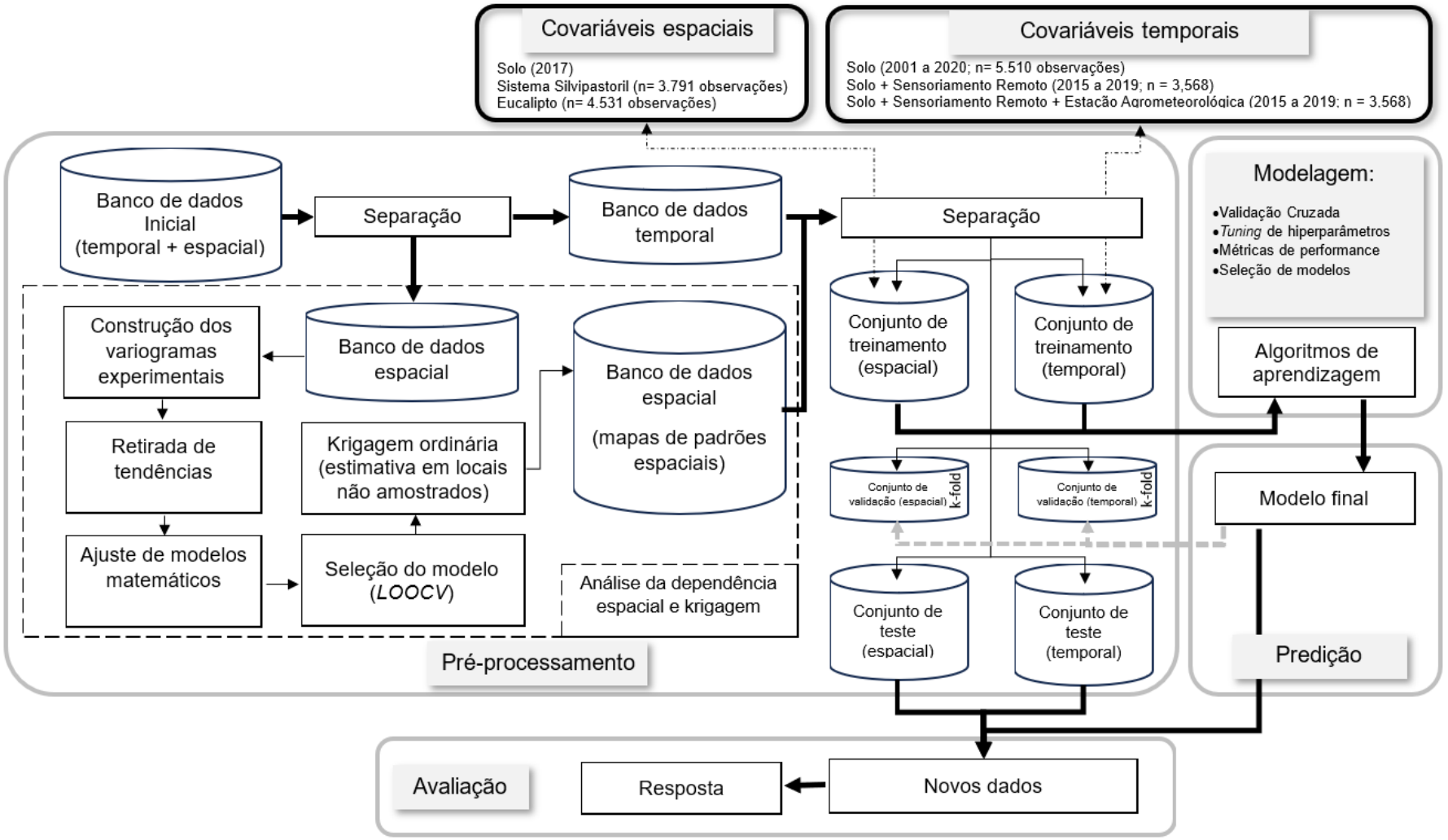

Figura 7. Sequência metodológica para modelagem da emissão de CO_2 do solo em dinâmica temporal e espacial. Destaca-se que o procedimento de aprendizado foi realizado de maneira independente em relação aos dados espaciais e temporais.

Para os dados especializados, os quais foram observados em gradeados regulares e irregulares instalados em áreas específicas, foi realizada a análise de dependência espacial para posterior estimativa dos valores em locais não amostrados por interpolação via krigagem ordinária (KO), etapa de Pré-processamento apresentado na Figura 7. A análise geoestatística é baseada na teoria das variáveis regionalizadas. Uma variável regionalizada é uma variável aleatória que assume diferentes valores, de acordo com a sua posição na área de estudo, e é considerada a realização de um conjunto de variáveis aleatórias.

Na prática, quando retiramos uma amostra de solo em um local com coordenadas definidas, temos apenas uma única realização da função aleatória. Para estimar valores em locais não amostrados, deve-se introduzir as restrições de estacionaridade estatística. Em outras palavras, a existência de estacionaridade permite que o ensaio seja repetido mesmo que se as amostras forem coletadas em pontos diferentes, pois elas pertencem à mesma população, com os mesmos momentos estatísticos (Vieira, 2000). Assim, tem-se a estacionaridade de primeira ordem (estacionaridade da média). Contudo, é também necessária a estacionaridade de segunda ordem, implicando que para cada par de uma variável aleatória, a função de covariância $Cov(h)$ existe e é dependente da distância h (Vauclin et al., 1983), sendo h o vetor de distância entre as amostras.

A estacionaridade de segunda ordem não é uma condição fácil de ser satisfeita na prática pois exige uma variância finita dos valores medidos, suposição essa difícil de ser verificada. Portanto, uma suposição alternativa e mais simples é assumida, a denominada de hipótese intrínseca. A hipótese intrínseca requer que, para todo vetor h, a variância do incremento $Z(x_i) - Z(x_i + h)$ seja finita e independente da posição dentro da área de estudo (Trangmar; Yost; Uehara, 1985), e temos assim a função:

$$Var\ [Z(x_i) - Z(x_i + h)] = E\ [Z(x_i) - Z(x_i + h)]^2 = 2\gamma(h) \quad (16)$$

que é denominada de variograma. Na prática, a forma do variograma não é muito utilizada e sim a forma $\gamma(h)$, denominada de semivariograma, que é estimado como a média do quadrado das diferenças entre todas as observações separadas pela distância h. Portanto, quando a pressuposição da hipótese intrínseca é satisfeita, o semivariograma apresenta a seguinte forma (Burrough; McDonnell, 1998):

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2, \quad (17)$$

em que, $\hat{\gamma}(h)$ é a estimativa da semivariância na distância de separação h; N é o número de pares de pontos separados pela distância h; $Z(x_i)$ é o valor da variável Z no ponto x_i; $Z(x_i + h)$ é o valor da variável Z no ponto $x_i + h$. O gráfico de $\hat{\gamma}(h)$, em função de h, é o chamado semivariograma experimental, o qual exibe um comportamento puramente aleatório ou sistemático que pode ser descrito por modelos teóricos. Os semivariogramas experimentais foram ajustados pelo método dos mínimos quadrados, aos seguintes modelos matemáticos teóricos:

Modelo exponencial:

$$\hat{\gamma}(h) = C_0 + C_1\left[1 - e^{-3\left(\frac{h}{a}\right)}\right] \quad , h > 0 \tag{18}$$

Modelo esférico:

$$\hat{\gamma}(h) = C_0 + C_1[3/2(h/a) - 1/2(h/a)^3],$$
$$0 \leq h \leq a \text{ e } \hat{\gamma}(h) = C_0 + C_1,\ h > a; \tag{19}$$

Modelo Gaussiano:

$$\hat{\gamma}(h) = C_0 + C_1[1 - e^{-3\left(\frac{h}{a}\right)^2}]\ ,\ 0 < h < d, \tag{20}$$

sendo d a máxima distância na qual o semivariogramas experimental foi definido. Caso a hipótese intrínseca não tenha sido satisfeita, foi realizada a retirada de tendência (em função das coordenadas longitude e latitude) e o semivariograma experimental foi novamente construído a partir dos resíduos. O valor de semivariância, na qual ocorre a estabilização do semivariograma, é denominado de patamar, representado pelo símbolo $C_0 + C_1$, sendo aproximadamente igual ao valor da estimativa da variância dos dados analisados. A distância na qual ocorre a estabilização do semivariograma é denominada de alcance, simbolizado por a, e define o limite da dependência espacial (autocorrelação). O valor C_1representa a estrutura de variabilidade espacial dos dados. O efeito pepita, representado pelo símbolo C_0, é o valor de semivariância encontrado no intercepto do modelo ajustado com o eixo Y. A escolha do melhor modelo ajustado aos semivariogramas utilizados no método da krigagem ordinária (KO) foi realizada por meio das validações cruzada (*LOOCV*). O processo *LOOCV* foi utilizado para a verificação da confiabilidade do modelo matemático ajustado (Isaaks; Srivastava, 1989). Os parâmetros dos modelos ajustados aos semivariogramas experimentais foram utilizados na estimativa dos atributos estudados em locais não amostrados por meio da técnica de KO ordinária.

O modelo escolhido foi aquele que melhor estimou os valores observados, ou seja, aquele que produziu uma equação de regressão linear entre os valores observados, em função dos valores estimados o mais próximo da bissetriz (intercepto igual a zero e coeficiente angular = 1).

Em um local não amostrado e para um determinado semivariograma, uma estimativa de KO pode ser considerada a estimativa não viesada e de variância mínima, ou seja, uma média ponderada ótima dos dados vizinhos (Cressie, 1990). A KO estima o valor de um campo aleatório em um local não amostrado x_0 baseado em um valor medido na forma linear:

$$\hat{z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i), \quad (21)$$

sendo: $\hat{z}(x_0)$ a estimativa da KO no ponto x_0, $z(x_i)$ os valores medidos em x_i, $i = 1, 2, \ldots, N$ e λ_i os pesos da KO atribuídos aos $z(x_i)$ para estimar $\hat{z}(x_0)$ (Trangmar; Yost; Uehara, 1985; Webster, 1985; Webster; Oliver, 1990; Vieira et al., 1997).

Para o processo de modelagem dos dados, foram utilizados três algoritmos de aprendizado estatístico:

Árvore de decisão (DT): As árvores de decisão são uma técnica de aprendizado de máquina estatístico amplamente utilizada para problemas de classificação e regressão. A implementação foi realizada a partir do pacote em R `rpart` (Therneau; Atkinson; Ripley, 2022). Em resumo, o algoritmo vai construindo uma estrutura de árvore que divide os dados em subconjuntos com base em suas *features* que, além de capturar não linearidade nos dados permitem a utilização de variáveis categóricas. Isso é feito de forma recursiva, onde o algoritmo escolhe a melhor *feature* para dividir os dados em cada etapa, com base na redução do erro de regressão (James et al., 2013; Kroese et al., 2019). Grande vantagem dessa técnica é a facilidade de Interpretação e robustez a *outliers* uma vez não são tão afetadas por pontos extremos. Os hiperparâmetros ajustados no processo de aprendizado foram: custo de complexidade (`cost_complexity`), profundidade da árvore (`tree_depth`) e o número mínimo de observações por nó (`min_n`). Nesse momento, devemos ressaltar que as árvores de decisão foram a base para os demais algoritmos de ensemble, ou seja, as árvores de decisão foram combinação de vários outros modelos de aprendizado estatístico para melhorar o desempenho geral dos modelos testados nesse estudo.

Random Forest (RF): São modelos estruturados em árvore utilizados tanto para classificação, quanto para regressão, sendo compostos por uma combinação de árvores preditoras (floresta) geradas a partir de um vetor aleatório, utilizando uma técnica chamada *Bagging (Bootstrap Aggregating),* amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta (James et al., 2013; Kroese et al., 2019). De forma genérica, o *Bagging* baseia-se na construção de múltiplas árvores de decisão independentes, onde cada árvore é treinada em um subconjunto aleatório e com substituição do conjunto de dados de treinamento original. A construção de uma árvore de regressão o modelo baseia-se na divisão das *features*, conforme suas semelhanças e dissimilaridades e o seu resultado é formado pela média dos resultados de todas as árvores. Em outras palavras o algoritmo treina múltiplas árvores de decisão em subconjuntos aleatórios dos dados e combina suas previsões. É uma técnica não paramétrica de aprendizagem estatística. A implementação foi realizada a partir do pacote em R `randomForest` (Breiman et al., 2022). No seu processo de aprendizagem, três hiperparâmetros foram otimizados: o número de árvores no *random forest* (`tree`); número mínimo de observações por folha (`min_n`) e o número de *features* consideradas em cada nó de divisão (`mtry`).

Extreme Gradient Boosting (XGBoost)*:* O *Boosting* baseia-se no conceito de melhorar algoritmos de classificação/regressão fracos, transformando-os em algoritmos mais fortes (James et al., 2013; Kroese et al., 2019). Assim como no algoritmo *Random Forest*, ele distribui pesos para as árvores individuais e, em seguida, gera um consenso combinando suas saídas. Nessa abordagem, um classificador recebe um peso maior se seu antecessor apresentar um desempenho pior, e esses pesos são ajustados durante o processo de aprendizado estatístico. O método funciona aplicando-se sequencialmente os classificadores às versões re-ponderadas do conjunto de treinamento, atribuindo maior peso às observações classificadas incorretamente no passo imediatamente anterior e menor peso às classificadas corretamente. Isso resulta em uma sequência de árvores que utilizam informações das árvores anteriores, permitindo ao algoritmo aprender com seus erros. A implementação foi realizada a partir do pacote em R `xgboost` (Chen et al., 2023). No seu processo de aprendizagem, sete hiperparâmetros foram otimizados, divididos em cinco etapas de ajuste: etapa 1: taxa de aprendizado (`learn_rate`) e o número

de árvores a serem treinadas (`trees`); etapa 2: profundidade de árvore (`tree_depth`) e número mínimo de observações por folhas (`min_n`); etapa 3: somente a redução de perda mínima para divisão (`loss_reduction`); etapa 4: o número de *features* consideradas em cada nó de divisão (`mtry`) e o tamanho da amostra aleatória usada em cada iteração (`sample_size`); etapa 5: os hiperparâmetros `trees` e `learn_rate` foram novamente otimizados.

Para cada um dos algoritmos anteriormente descritos nesta aplicação (Árvore de Decisão, *Random Forest* e XGBoost), uma lista de valores candidatos para seus hiperparâmetros foi definida. Em seguida, utilizando a validação cruzada k*-fold* com $k = 5$, realizou-se uma análise de desempenho preditivo para cada modelo, avaliando a raiz do erro quadrático médio (RMSE). Isso permitiu selecionar o modelo com o melhor desempenho em cada iteração, considerando os dados espaciais e temporais em diferentes configurações de *features,* passos referentes ao aprendizado estatístico e predição apresentados na Figura 7. Posteriormente, o modelo selecionado foi aplicado aos dados de teste para avaliar seu desempenho na previsão de observações futuras, utilizando novamente o RMSE como métrica de avaliação (Etapa de Avaliação apresentada na Figura 7). Foram ajustados, em média, 50 modelos para DT, 50 modelos para RF e 1850 modelos para o XGBoost. Isso ocorreu devido ao XGBoost ter um maior número de hiperparâmetros em comparação com os outros algoritmos, totalizando 1950 modelos ajustados. Considerando três aproximações (diferentes combinações de *features*), o número total de modelos ajustados foi de 5.850 na modelagem temporal. Para a modelagem espacial, os 1950 modelos foram ajustados para dois usos do solo em cinco dias de avaliação, totalizando 19.500 modelos. Por fim, temos o total de 25.350 testados e avaliados durante todo o processo de aprendizado de máquina. Os diferentes modelos foram comparados pelas seguintes métricas de performance (indicadores estatísticos): coeficiente de correlação (r), coeficiente de determinação (R^2), erro quadrático médio (MSE), raiz do erro quadrático médio (RMSE), erro médio absoluto (MAE) e erro absoluto médio percentual (MAPE), todos calculados a partir do conjunto de teste para os melhores modelos selecionados.

$$r = \frac{\sum_{i=1}^{n}(y_{i\times}y_i') - \mathrm{n}\bar{y} \times \bar{y}'}{(n-1)\, \mathrm{S}_y\, \mathrm{S}_{y'}}, \tag{22}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i')^2}{\sum_i (y_i - \bar{y})^2}, \quad (23)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2, \quad (24)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2}, \quad (25)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|, \quad (26)$$

$$MAPE = \left[\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_i'}{y_i}\right|\right] \times 100, \quad (27)$$

em que y e y' são, respectivamente, o valor observado e o valor estimado para FCO_2, n é o número de observações no conjunto de teste, $\bar{y}$ e $\bar{y}'$ são os valores médios dos observados e estimados, respectivamente e S_y e $S_{y'}$ são os respectivos desvios padrão dos valores observados e preditos.

4 RESULTADOS E DISCUSSÃO

4.1 Modelagem temporal

As métricas de desempenho dos melhores modelos selecionados para cada método de aprendizado de máquina, observadas no conjunto de teste, estão apresentadas na Tabela 2. Os algoritmos de aprendizado de máquina testados exibiram comportamentos diferenciados para cada conjunto de covariáveis modelado, sem necessariamente ser observado um método que apresentou melhor performance em todas as configurações de *features.* Para o conjunto de variáveis Solo o algoritmo de RF apresentou desempenho superior em todas as métricas avaliadas, quando comparados aos demais algoritmos, com os menores valores de RMSE (1,19 μmol m^{-2} s^{-1}) e MAPE (35,15%). Por outro lado, DT apresentou os maiores valores de RMSE (1,33 μmol m^{-2} s^{-1}) e MAPE (39,01%) enquanto o XGBoost apresentou valores de RMSE (1,21 μmol m^{-2} s^{-1}) e MAPE (37,17%), ligeiramente superiores aos observados em RF. Esses resultados indicam que o algoritmo de *Random Forest* pode ser considerado a escolha mais robusta e eficaz entre os algoritmos estudados para configuração de covariáveis formadas apenas pelos atributos físicos e químicos do solo. Como uma média de FCO_2 observada para o banco de dados foi de 1,33 μmol m^{-2} s^{-1}, observou-se valor de RMSE menor do que a média da variável alvo, indicando boa precisão do modelo RF. Contudo, o valor de MAPE de 35% indica uma discrepância moderada entre as previsões do modelo e os valores reais.

As variáveis selecionadas e seus respectivos valores de importância para DT, RF e XGBoost foram semelhantes quando considerado o conjunto Solo (Figura 8). Os três modelos estudados incluíram as variáveis temperatura do solo (Ts), macroporosidade (Macro), microporosidade (Micro), pH, teor de Fósforo disponível no solo e o manejo com destaque às ações de reflorestamento, embora tenham sido classificadas como níveis de importância diferentes, dependendo do algoritmo.

A Figura 8 apresenta a performance dos modelos, com destaque para o algoritmo de *Random Forest,* o qual apresentou o maior valor de coeficiente de determinação (R^2 = 0,64) quando comparado a DT (0,57) e XGBoost (0,64). Em adição, a análise dos resultados da regressão dos observados versus preditos (estimados), aponta um valor de coeficiente linear α não diferente de zero (teste t; p >

0,05), indicando uma baixa tendência de viés aliado a um coeficiente angular β não diferente de 1 (teste t; p > 0,05).

Tabela 2. Métricas de desempenho dos algoritmos de aprendizado de máquina, Árvore de Decisão (DT), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost) aplicados aos diferentes conjuntos de covariáveis estudadas na abordagem temporal.

Métrica	DT	RF	XGBoost
	Solo (n= 5.510)		
r	0,76	**0,81**	0,80
R^2	0,57	**0,66**	0,64
MSE	1,77	**1,41**	1,47
RMSE	1,33	**1,19**	1,21
MAE	0,85	**0,74**	0,77
MAPE	39,01	**35,15**	37,17
	Solo + Sensoriamento Remoto (n = 3,568)		
r	0,78	**0,87**	0,86
R^2	0,60	**0,75**	0,73
MSE	1,61	**1,00**	1,07
RMSE	1,27	**1,00**	1,03
MAE	0,83	**0,65**	0,67
MAPE	32,69	**25,93**	26,72
	Solo + Sensoriamento Remoto + Estação Agrometeorológica (n = 3,568)		
r	0,83	0,90	**0,91**
R^2	0,69	0,80	**0,83**
MSE	1,59	1,01	**0,87**
RMSE	1,26	1,00	**0,93**
MAE	0,83	0,64	**0,59**
MAPE	27,33	21,34	**19,78**

r= coeficiente de correlação linear; R^2= coeficiente de determinação; MSE = erro quadrático médio ($\mu mol\ m^{-2}\ s^{-2}$)²; RMSE raiz do erro quadrático médio ($\mu mol\ m^{-2}\ s^{-2}$); MAE erro médio absoluto ($\mu mol\ m^{-2}\ s^{-2}$) e MAPE erro médio absoluto percentual (%). Os valores destacados em negrito e sublinhado representam os melhores resultados obtidos para cada métrica de desempenho.

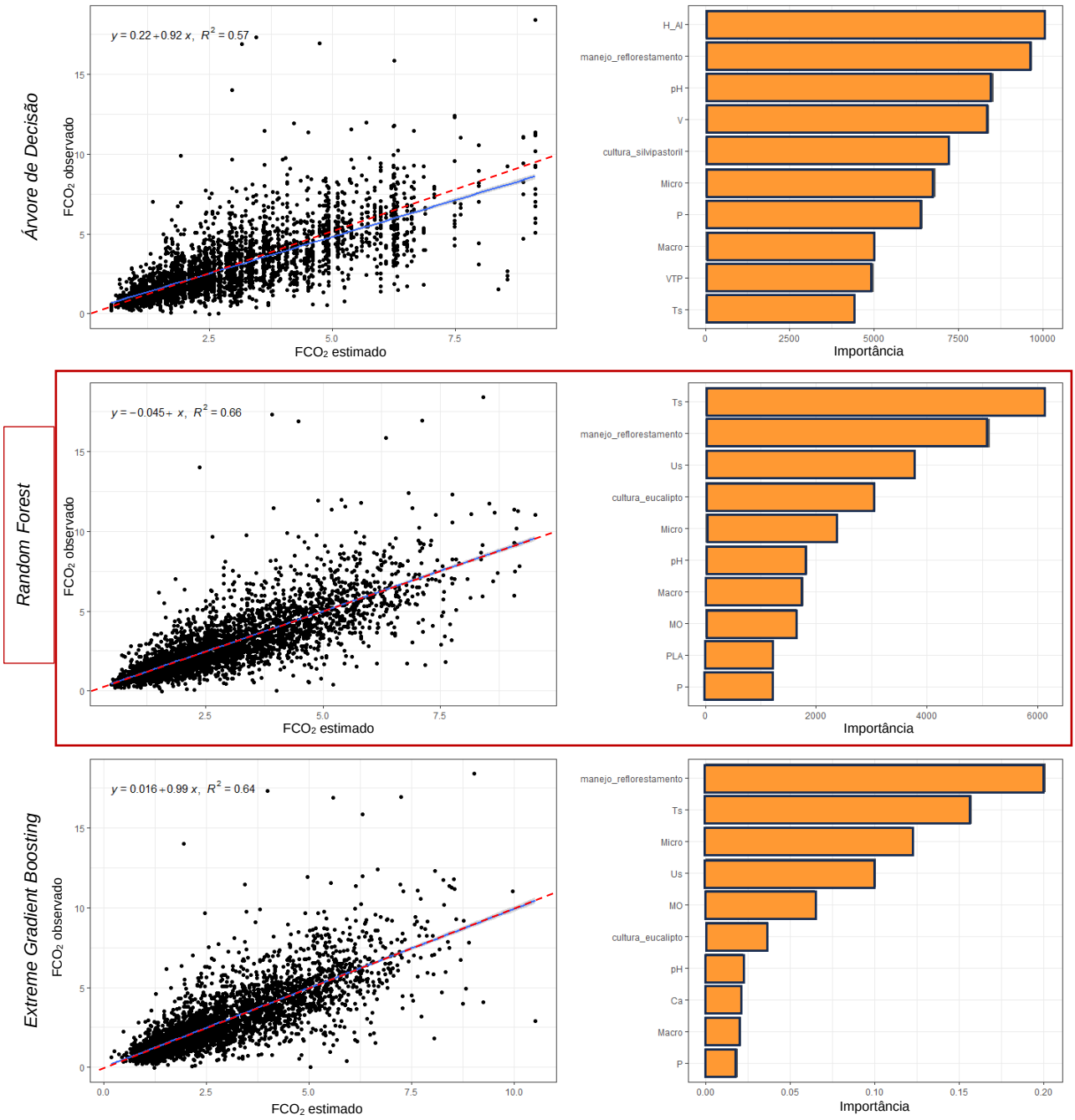

Figura 8. Desempenho e importância de variáveis dos melhores modelos ajustados em cada algoritmo de aprendizado testado, para o conjunto de covariáveis Solo (n= 5.510), abrangendo o período de 2001 a 2020. Atributos do solo representados pela cor laranja. O melhor modelo foi destacado pelo retângulo vermelho.

Semelhante ao observado anteriormente, para o conjunto de variáveis Solo + Sensoriamento Remoto o algoritmo de RF também apresentou desempenho superior em todas as métricas avaliadas, quando comparados aos demais algoritmos, exibindo os menores valores de RMSE (1,00 μmol m^{-2} s^{-1}) e MAPE (25,93%). Em relação aos demais algoritmos, DT apresentou os maiores valores de RMSE (1,26 μmol m^{-2} s^{-1}) e MAPE (27,33%) enquanto o algoritmo XGBoost apresentou valores de RMSE (1,03 μmol m^{-2} s^{-1}) e MAPE (26,72%). A análise da Figura 9 indica que os três algoritmos de aprendizado de máquina testados selecionaram as variáveis X_{CO2} e SIF, com elevado grau de importância. Além dessas variáveis, a temperatura do solo (Ts), a microporosidade (Micro), e porosidade livre de água no solo (PLA) foram consistentemente incluídas em todos os modelos. Esses resultados indicam que independentemente do algoritmo utilizado, é notório uma melhoria nas métricas de performance ao serem adicionadas as variáveis de Sensoriamento Remoto ao conjunto de dados. X_{CO2} e SIF carregam informações significativas para a modelagem da emissão de CO_2 do solo, principalmente em áreas agrícolas, indicando que os modelos estão se aproximando cada vez mais dos valores reais sendo essas, portanto, essenciais para o processo de estimativa de FCO_2. Além disso, deve-se ressaltar que nessa configuração de *features,* RF selecionou quatro atributos químicos do solo: teor de matéria orgânica do solo (MO), capacidade de troca de cátion (CTC), soma de bases do solo (SB) e o teor de Magnésio do solo (Mg). Por sua vez, XGBoost selecionou dois atributos químicos do solo (MO e teor de fósforo disponível no solo - P), enquanto DT selecionou apenas um atributo químico do solo (acidez potencial trocável - H_Al).

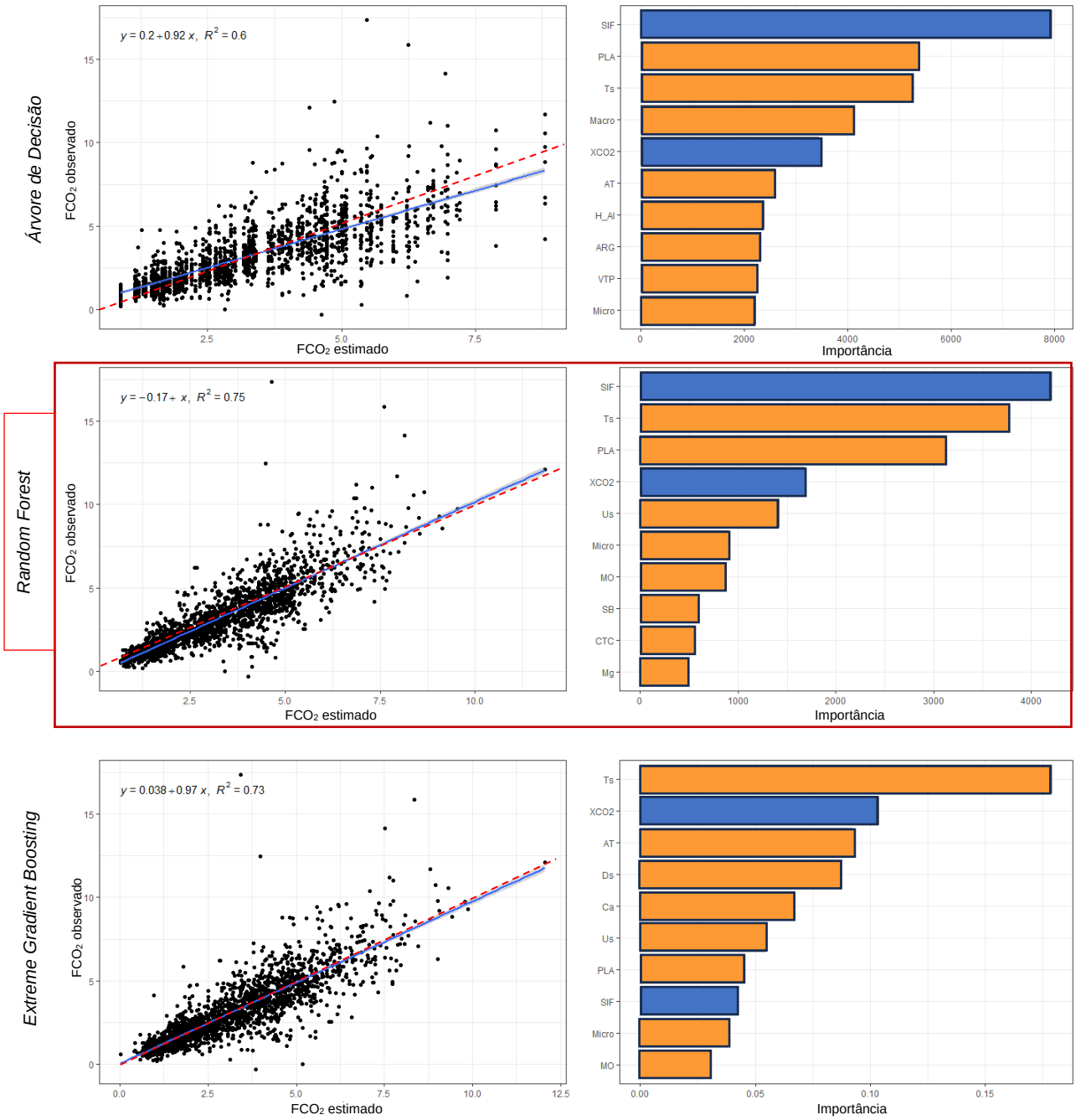

Figura 9. Desempenho e importância de variáveis dos melhores modelos ajustados em cada algoritmo de aprendizado testado, para o conjunto de covariáveis Solo + Sensoriamento Remoto (n= 3.568), abrangendo o período de 2015 a 2019. Atributos do solo representados pela cor laranja e variáveis de sensoriamento são representados pela cor azul. O melhor modelo foi destacado pelo retângulo vermelho.

Para o conjunto de covariáveis Solo + Sensoriamento Remoto + Estação Agrometeorológica, foram observadas diferenças no desempenho entre os algoritmos, agora com inversão de papéis, com o algoritmo XGBoost se destacando em todas as métricas de performance avaliadas. O XGBoost obteve valores de RMSE e MAPE iguais a 0,93 µmol m^{-2} s^{-1} e 19,78%, respectivamente (Tabela 2). Em contrapartida, a árvore de decisão (DT) apresentou os maiores valores de RMSE (1,26 µmol m^{-2} s^{-1}) e MAPE (27,33%), enquanto o Random Forest (RF) não teve mudança significativa no valor de RMSE, permanecendo em 1,00 µmol m^{-2} s^{-2}, mas mostrou uma diminuição no valor de MAPE atingindo o valor de 21,34%. A abordagem de aprendizado estatístico tem se mostrado eficiente para a modelagem de FCO_2. Resultados recentes sugerem que a modelagem mecanicista de FCO_2 apresentou uma maior incerteza associada aos seus resultados quando comparados a modelos de aprendizado de máquina (Lu et al., 2021). Huang et al. (2020) investigaram as variações espaciais e temporais da respiração do solo global, bem como sua relação com fatores climáticos e cobertura do solo, compararam diversos algoritmos de aprendizado de máquina (*Random Forest*, *Support Vector Machine* e Redes Neurais Artificiais), e um método tradicional (regressão não linear multivariada. Os autores observaram que o modelo *Random Forest* teve o melhor desempenho em seis biomas avaliados (com R^2 variando entre 0,47 e 0,68), seguido pelo algoritmo de *Support Vector Machine*, que se destacou em quatro biomas (R^2 variando entre 0,41 e 0,69). O modelo baseado em Redes Neurais Artificiais, por sua vez, apresentou um desempenho moderado, com R^2 variando entre 0,35 e 0,62.

Nossos resultados sugerem que, apesar da boa performance preditiva observada para o RF, o algoritmo XGBoost pode ser o mais robusto em algumas ocasiões para estimativas da emissão de CO_2 do solo em configurações de *features* formadas por sinais integrados a partir de múltiplas fontes de dados. No entanto, ainda é escassa a literatura que compara o desempenho preditivo de modelos baseados em *Gradient Boosted Trees* (XGBoost) em relação aos algoritmos mais utilizados, como o *Random Forest* quanto *à* FCO_2. Em um estudo recente, com objetivo de estimar a respiração edáfica total de um agroecossistema (cultura do trigo), diversas técnicas, incluindo regressão linear múltipla, *Support Vector Regression*, *Random Forest, Extreme Gradient Boosting* e Redes Neurais Artificiais, foram testadas. Os resultados

indicaram que o algoritmo XGBoost superou os outros métodos em termos de desempenho preditivo, observando-se valores de R^2 variando de 0,85 a 0,88 (Lu et al., 2023).

Quanto à importância das variáveis, na Figura 10, DT selecionou, em sua maioria, as variáveis provenientes de estação agrometeorológica, com destaque para as temperaturas do ar (Tmed, Tmax e Tmin), juntamente com a evapotranspiração (Eto). Variáveis essas, reconhecidamente como as controladoras do FCO_2 ao longo do tempo (La Scala Jr.; Panosso; Pereira, 2003; Canteral et al., 2023b). Além disso, X_{CO2} e SIF não foram selecionadas por esse modelo, sendo o teor de Calcio no solo (Ca) o único atributo químico do solo selecionado. Por outro lado, as variáveis de sensoriamento remoto foram selecionadas sempre pelos algoritmos RF e XGBoost, os quais apresentaram maior diversidade quanto à fonte das variáveis. Entre as variáveis de solo, a temperatura do solo (Ts) foi selecionada por ambos os algoritmos. Entre as variáveis agroclimatológicas, a Tmin foi selecionada por ambos os algoritmos, com maior destaque para o RF. Por outro lado, a radiação global (Rad) foi a variável com maior importância para o XGBoost, sendo este último o algoritmo que apresentou o maior ganho percentual para essa métrica, correspondendo a um aumento de 10% no poder de previsão com a adição desse grupo de variáveis.

Além disso, a análise dos resultados da regressão entre os valores de FCO_2 observados e os valores previstos indica melhores desempenhos para esses dois algoritmos quando comparados ao DT. Segundo Huang et al. (2020), utilizando dados de sensoriamento remoto e modelos estatísticos específicos para diferentes biomas, as mudanças no uso da terra desempenharam um papel fundamental na regulação das variações no fluxo de CO_2 do solo em escala anual, especialmente em regiões de clima temperado e boreal. Os autores observaram variações significativas no FCO_2 em áreas onde houve alterações na cobertura vegetal baixa (isto é, vegetação com menos de 5 metros de altura), em contraste com áreas que apresentaram mudanças climáticas significativas. Vale ressaltar que as regiões boreais, temperadas e tropicais contribuíram com 15%, 24% e 61%, respectivamente, para a média anual global total de respiração do solo, com a cobertura do solo emergindo como a variável explicativa primordial para o FCO_2 global (Huang et al., 2020; Grunwald, 2022).

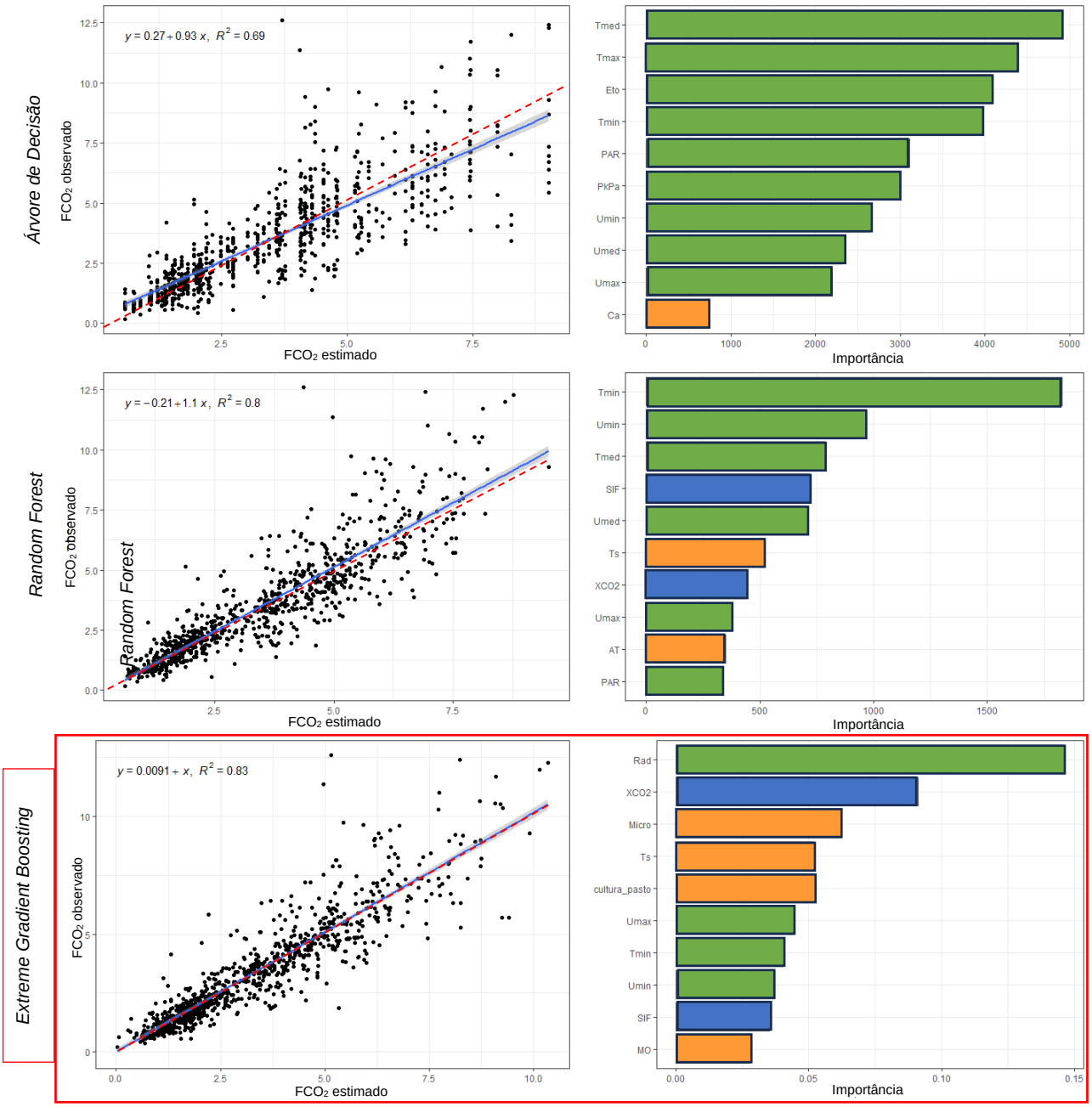

Figura 10. Desempenho e importância de variáveis dos melhores modelos ajustados em cada algoritmo de aprendizado testado, para o conjunto de covariáveis Solo + Sensoriamento Remoto + Estação Agrometeorológica (n= 3.568), abrangendo o período de 2015 a 2019. Atributos do solo representados pela cor laranja, variáveis de sensoriamento são representados pela cor azul e variáveis agrometeorológica são representadas pela cor verde. O melhor modelo foi destacado pelo retângulo vermelho.

4.2 Modelagem espacial

A base de dados utilizada neste estudo consiste em 12 experimentos de campos georreferenciados conduzidos no período de 2001 a 2018, abrangendo diversas áreas, culturas, manejos de solo e configurações de amostragem. Alguns desses resultados foram previamente publicados em revistas científicas (ver Tabela 1). No entanto, para este recorte da base de dados, foi selecionado experimento que ainda não publicados, originado do trabalho de mestrado de Oliveira (2018). Para esse estudo, as áreas selecionadas pertencem à Fazenda de Ensino, Pesquisa e Extensão (FEPE), da Faculdade de Engenharia de Ilha Solteira (FEIS - UNESP), localizada no município de Selvíria, estado do Mato Grosso do Sul (Tabela 1). No seu histórico, as áreas eram originalmente cobertas por vegetação nativa do Cerrado até a década de 1970, quando, em 1978, foram desmatadas e passaram a ser usadas para culturas anuais, como milho, soja, algodão e adubos verdes, até 1986. Durante os anos de 1986-1987, as áreas foram convertidas para os seguintes usos: floresta plantada de eucalipto (EU) e sistema silvipastoril (SI), uma floresta plantada de aroeira-vermelha (*Myracrodruon urundeuva*) em consórcio com capim braquiária (*Brachiaria decumbens*). A área de EU (*Eucalyptus camaldulensis*) foi formada em 26 de abril de 1986 e a área de SI foi formada em dezembro de 1987. No momento da realização das avaliações de FCO_2, as áreas já haviam passado por mais de 30 anos de mudança de conversão. As determinações foram realizadas no período de 03 de fevereiro a 17 de junho de 2017, durante as manhãs, entre 7 e 10 h. Foram selecionadas as seguintes datas: 17/02, 15/03, 03/06, 10/06 e 17/06 para EU e as datas 22/02, 17/03, 03/06, 10/06 e 17/06 para SI. A Figura 11 indica os dias de avaliação e os valores de precipitação (chuva em mm). Ao final do período de determinação da emissão de CO_2 do solo, as amostras de solo foram coletadas e todos os atributos físicos e químicos foram determinados.

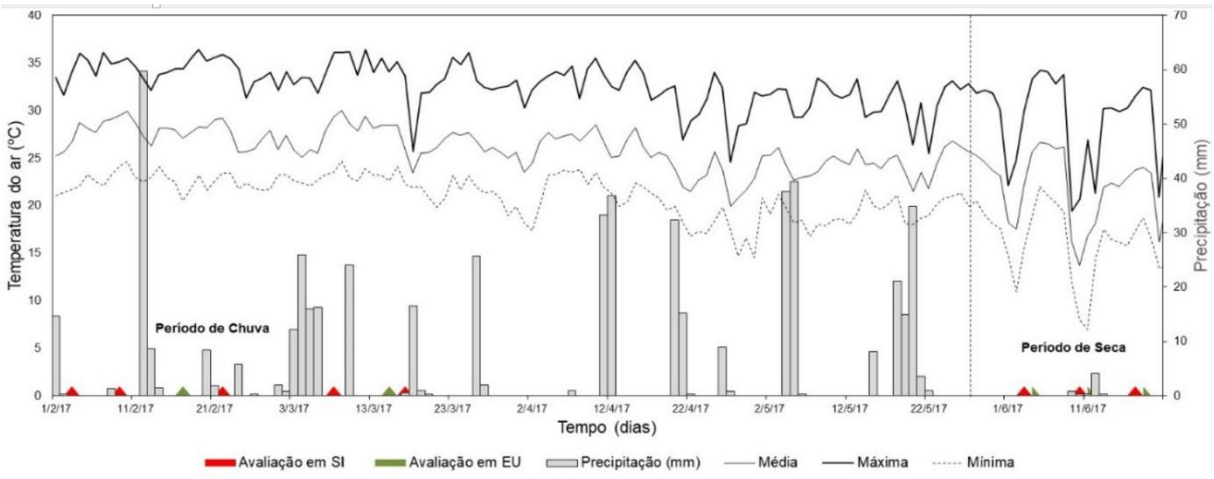

Figura 11. Precipitação pluviométrica durante o período de avaliação do experimento na área de sistema silvipastoril (SI) e floresta plantada de eucalipto (EU), localizada no município de Selvíria-MS, durante o ano de 2017. Fonte: Oliveira (2018).

Como estratégia de aprendizado de máquina, os algoritmos foram testados individualmente para cada dia de avaliação apresentados na Figura 11, e para cada uso da terra. Essa abordagem de aprendizado foi adotada com base no histórico de conversão de longa data observado para as áreas, que sugere que os atributos do solo, especialmente aqueles relacionados à dinâmica espaçotemporal de FCO_2 exibem uma alta especificidade em termos de magnitude, variabilidade e dependência espacial, justificando a análise separada por área. Os resultados detalhados dessas análises estão resumidos na Tabela 3. Em geral, foi observado que todos os algoritmos apresentaram um alto desempenho, conforme evidenciado pelos elevados valores de R^2 (0,92 – 0,99). Além disso, as medidas de RMSE exibiram uma pequena variação, com valores oscilando entre 0,03 e 0,55 μmol m^{-2} s^{-1}, indicando que as previsões dos modelos foram próximas das observações estimadas pela krigagem ordinária (KO). A avaliação do MAPE também corroborou essa conclusão, com porcentagens de erro médias variando de 0,50% a 5,37%. O algoritmo RF demonstrou um desempenho superior na previsão de FCO_2 na área SI. Ele se destacou como o melhor modelo em três dos cinco dias estudados. Por outro lado, o algoritmo XGBoost mostrou-se mais eficaz em prever a emissão de CO_2 do solo em EU, liderando como o melhor modelo em quatro dos cinco dias avaliados. Esta variação no desempenho entre os algoritmos indica que a escolha do modelo pode depender da área de estudo,

destacando a importância da adaptação dos métodos de aprendizado de máquina às características específicas do solo e do ambiente.

As variáveis selecionadas pelos modelos em diferentes dias de avaliação são apresentadas nas Figuras 12 e 13. Na área SI (Figura 12) a variável estoque de carbono no solo (EstC) se destacou, sendo selecionada em todos os modelos para todos os dias estudados, indicando a importância desse atributo para estimativas mais acuradas da emissão de CO_2 do solo no sistema silvipastoril. Os resultados apontam que o EstC não apenas desempenha um papel fundamental na determinação da magnitude das emissões em diferentes ecossistemas, mas também é essencial para compreender os padrões de variabilidade espacial dessas emissões. Essa seleção consistente sugere que o teor de carbono no solo é um fator influente sobre as próprias taxas de perda de carbono do solo via emissões de CO_2 nessa área ao longo do tempo e do espaço, como esperado para um modelo de decaimento de primeira ordem (La Scala et al., 2008):

$$\frac{\mathrm{d}C_{solo}(t)}{\mathrm{d}t} = -kC_{solo}(t), \tag{28}$$

em que C_{solo} é a quantidade de carbono lábil da matéria orgânica prontamente decomponível, k é a constante de decaimento, e t é o tempo.

No entanto, na área de eucalipto (Figura 13), o estoque de carbono do solo foi selecionado por todos os modelos, mas apenas em dois dos cinco dias avaliados. Nesta área, os atributos que se destacaram com maior frequência foram os teores de Fósforo disponível (P) e de Cálcio (Ca) no solo, sendo esses atributos químicos consistentemente escolhidos por todos os métodos de aprendizado em todos os dias de avaliação. Os atributos químicos e físicos do solo estão intimamente relacionados à produção e transporte de gases do solo para a atmosfera, conforme mencionado por Farhate et al. (2018a) e Silva et al. (2019). A falta de relação entre FCO_2 e EstC em alguns dias pode indicar a respiração das plantas e raízes como sendo a principal componente de FCO_2 na área de Eucalipto, nesses dias e horários, como a principal fonte de CO_2 para a atmosfera.

Um outro aspecto a se destacar é quanto à temperatura (Ts) e à umidade do solo (Us), variáveis que, assim como o FCO_2, exibem variações nas escalas espaciais e temporais. Tais variáveis, surpreendentemente não demonstraram uma seleção consistente pelos modelos. Observe que em SI (Figura 12), a Ts foi escolhida por

todos os modelos em três dos cinco dias de estudo, enquanto na área EU (Figura 13), Ts foi selecionada em quatro dos cinco dias avaliados. Já para a Us, sua seleção ocorreu em todos os modelos, apenas em um dia para SI e em dois dias para EU.

A emissão de CO_2 do solo é uma das maiores fontes globais de CO_2 para a atmosfera, indicando o nível de atividade microbiana no solo. Contudo, a emissão de CO_2 do solo deve ser entendida como uma componente resultante da soma de duas outras, associadas a processos respiratórios: respiração autotrófica (relacionada às raízes das plantas) e respiração heterotrófica (atividade microbiana e demais organismos presentes no solo) (Hanson et al., 2000). Os resultados indicam que um desses processos pode estar mais relacionado a FCO_2 na área silvopastoril e não necessariamente o mesmo processo estaria contribuindo mais para FCO_2 na área de Eucalipto. A temperatura e a umidade do solo são os principais fatores da componente heterotrófica, afetando a decomposição microbiana em escalas desde milímetros até quilômetros (Rodrigo et al., 1997; Kim et al., 2012; Zhang et al., 2022). Estimativas sugerem que um aumento anual na temperatura global de apenas 0,03 °C poderia levar a uma liberação adicional de aproximadamente 10 Pg de carbono na atmosfera a cada década, devido à respiração do solo (Xu; Qi, 2001). Isso significa que o aumento da temperatura global levaria a uma decomposição mais rápida do carbono do solo e, consequentemente, a uma maior emissão de CO_2 pelos solos (Xu; Qi, 2001; Grunwald, 2022). No entanto, apesar de sua importância, a resposta do carbono do solo ao aquecimento global ainda representa uma das maiores incertezas no contexto do ciclo global do carbono (Kirschbaum, 2006; Haaf; Six; Doetterl, 2021).

Tabela 3. Avaliação de desempenho dos algoritmos de aprendizado de máquina, Árvore de Decisão (DT), *Random Forest* (RF) e *Extreme Gradient Boosting* (XGBoost) aplicados aos usos do solo SI e EU para abordagem espacial. Somente foram considerado o grupo de variáveis Solo para essa modelagem.

Métricas	DT	RF	XGB	DT	RF	XGB	DT	RF	XGB	DT	RF	XGB	DT	RF	XGB
	Silvipastoril														
	22/02/2017			17/03/2017			03/06/2017			10/06/2017			17/06/2017		
r	0,92	0,99	**0,99**	0,97	**0,99**	0,99	0,92	0,98	**0,99**	0,95	**0,99**	0,99	0,95	**0,99**	0,99
R^2	0,85	0,98	**0,98**	0,93	**0,99**	0,98	0,85	0,97	**0,98**	0,90	**0,98**	0,98	0,90	**0,98**	0,97
MSE	0,034	0,007	**0,005**	0,005	**0,001**	0,001	0,022	0,005	**0,004**	0,006	**0,001**	0,001	0,005	**0,001**	0,001
RMSE	0,18	0,08	**0,07**	0,07	**0,03**	0,04	0,15	0,07	**0,06**	0,08	**0,03**	0,04	0,07	**0,03**	0,04
MAE	0,12	0,06	**0,05**	0,04	**0,02**	0,03	0,09	0,04	**0,04**	0,04	**0,02**	0,03	0,04	**0,02**	0,03
MAPE	2,73	1,26	**1,21**	1,00	**0,50**	0,62	2,62	1,21	**1,21**	1,56	**0,81**	1,09	1,50	**0,68**	0,92
	Eucalipto														
	17/02/2017			15/03/2017			03/06/2017			10/06/2017			17/06/2017		
r	0,94	0,99	**0,99**	0,98	0,99	**0,99**	0,97	0,99	**0,99**	0,97	**0,99**	0,99	0,98	0,99	**0,99**
R^2	0,88	0,98	**0,99**	0,96	0,99	**0,99**	0,93	0,99	**0,99**	0,94	**0,99**	0,99	0,96	0,99	**0,99**
MSE	0,31	0,06	**0,039**	0,13	0,03	**0,020**	0,03	0,01	**0,004**	0,03	**0,005**	0,00	0,01	0,00	**0,003**
RMSE	0,55	0,23	**0,20**	0,36	0,18	**0,14**	0,18	0,08	**0,06**	0,17	**0,07**	0,07	0,11	0,05	**0,05**
MAE	0,36	0,16	**0,14**	0,25	0,13	**0,10**	0,12	0,06	**0,05**	0,11	**0,05**	0,05	0,08	0,04	**0,04**
MAPE	5,37	2,42	**2,17**	6,58	3,39	**2,70**	3,16	1,48	**1,25**	3,43	**1,51**	1,58	2,04	1,03	**1,02**

r= coeficiente de correlação linear; R^2= coeficiente de determinação; MSE = erro quadrático médio (μmol m^{-2} s^{-2})^2; RMSE raiz do erro quadrático médio (μmol m^{-2} s^{-2}); MAE erro médio absoluto (μmol m^{-2} s^{-2}) e MAPE erro médio absoluto percentual (%). Os valores destacados em negrito e sublinhado representam os melhores resultados obtidos para cada métrica de desempenho.

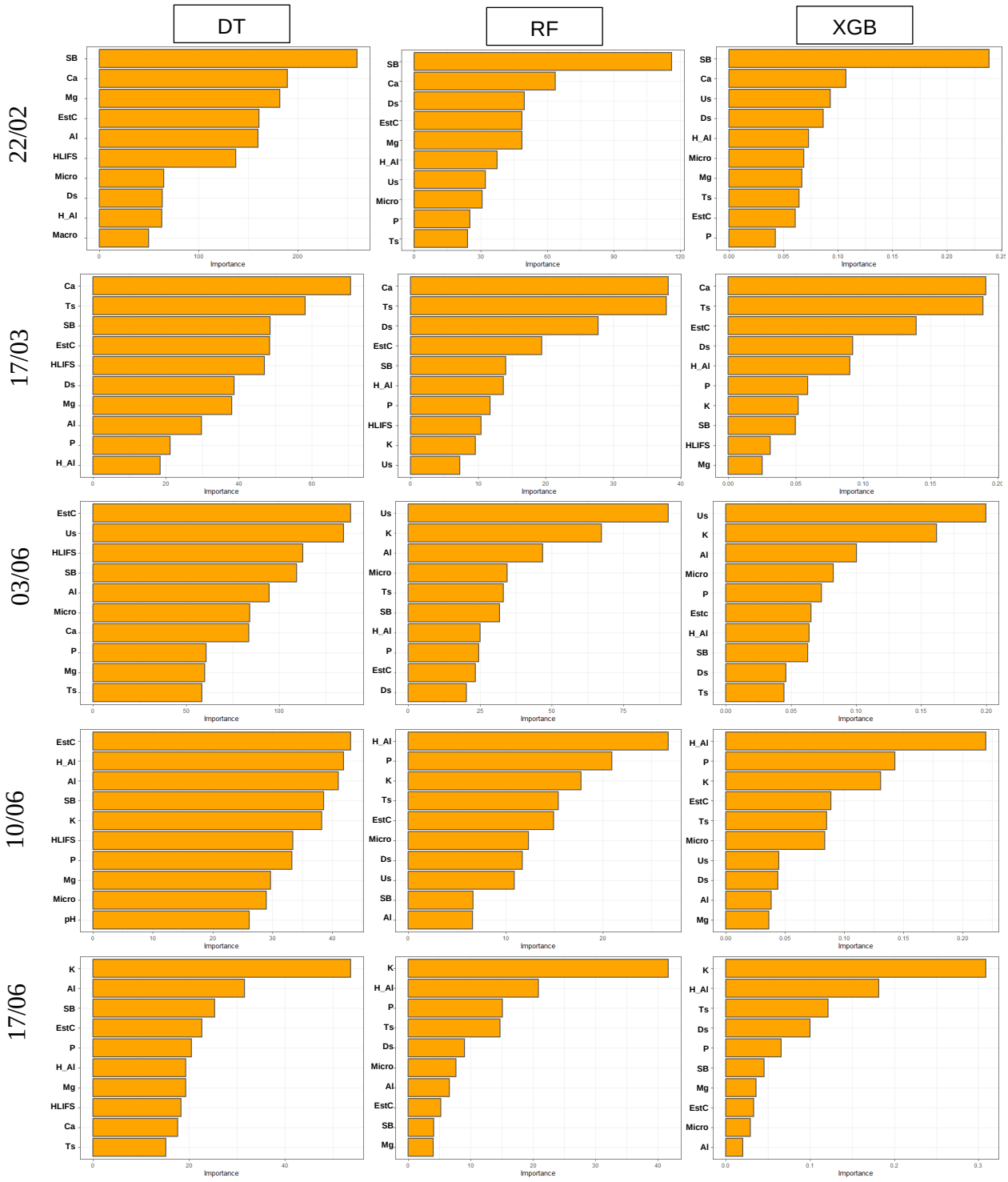

Figura 12. Importância de variáveis dos melhores modelos ajustados em cada algoritmo de aprendizado testado, para o conjunto de covariáveis Solo (n= 3.791), uso do solo Silvipastoril (SI), para o estudo conduzido ano de 2017, Fazenda Experimental da UNESP de Ilha Solteira – Selvíria-MS.

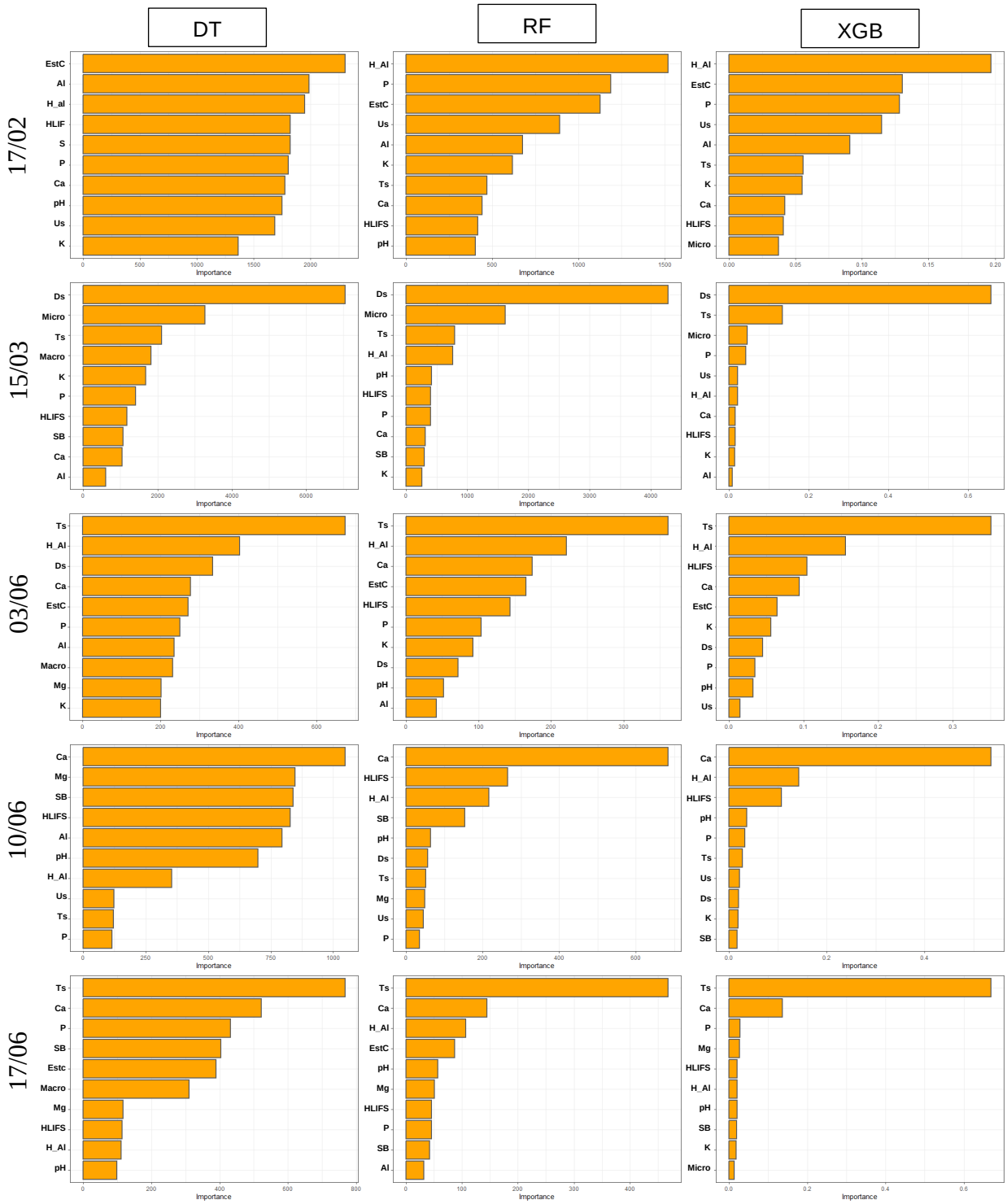

Figura 13. Importância de variáveis dos melhores modelos ajustados em cada algoritmo de aprendizado testado, para o conjunto de covariáveis Solo (n= 4.531), uso do solo Eucalipto (EU), para o estudo conduzido ano de 2017, Fazenda Experimental da UNESP de Ilha Solteira – Selvíria-MS.

Nossos resultados, de certa forma, contribuem para elucidar essas incertezas, uma vez que apontam para a temperatura do solo como um fator controlador da variabilidade espacial FCO_2 especialmente em florestas plantadas de eucalipto, ou

seja, sua importância depende dos manejos adotados na área. Deve-se destacar que mesmo quando as avaliações foram conduzidas no mesmo dia e horário (03, 09 e 17 de junho de 2017, conforme Figura 11 e Tabela 03), os modelos em sua maioria selecionaram a temperatura do solo (Ts). No entanto, as maiores importâncias dessa variável foram observadas na área de eucalipto em comparação com a área silvipastoril. Esse efeito se deve principalmente às condições microclimáticas presentes ao redor da planta e da superfície do solo, resultantes de fatores como diferenças no espaçamento de plantio (aproximadamente de 3 a 4 metros entre plantas para os dois cultivos), tipo de crescimento da planta (monopodial no caso do eucalipto e simpodial no caso da aroeira-vermelha) e a cobertura do solo (folhas e galhos secos para eucalipto e gramínea, no sistema silvipastoril). Provavelmente a combinação desses fatores influenciaram nos níveis de radiação solar incidente e nas taxas de evapotranspiração nas áreas, sendo essas menores para a área silvipastoril e, consequentemente, proporcionando melhores condições de umidade do solo para esta área. Tais aspectos podem ser vislumbrados na Figura 14, a qual apresenta as condições das áreas no momento da condução das avaliações.

Para a área SI, houve uma atenuação da influência de Ts na variabilidade espacial de FCO_2, deixando que sua variabilidade fosse controlada por aspectos ligados à degradação da matéria orgânica do solo realizada por microrganismos em boas condição de umidade, relacionados portanto à componente respiração heterotrófica do solo (Vicentini et al., 2023). Em trabalho recente, onde aspectos ligados à variabilidade espaçotemporal das emissões de CO_2 do solo foram monitorados em floresta tropical no Panamá, Rubio; Detto (2017) constataram que a umidade do solo foi responsável por ciclos sazonais, ciclos diurnos, variabilidade intrasazonal e interanual de FCO_2. Em contra partida, a variabilidade espacial da respiração do solo revelou um papel emergente da estrutura florestal e de fatores como a temperatura do solo e a topografia (Rubio; Detto, 2017). Os autores afirmam que o solo pode ser entendido como uma mistura complexa e espacialmente heterogênea de minerais, estoque de carbono, raízes e microorganismos, e fatores climáticos como precipitação e incidência de radiação solar contribuem para a formação de microclimas diferentes que impulsionam a variabilidade espaçotemporal na respiração do solo, de maneiras diferentes. Além disso, estudos em ecossistemas

florestais sugerem que a respiração autotrófica é mais sensível à temperatura do que a respiração heterotrófica (Makiranta et al., 2008; Chen et al., 2017).

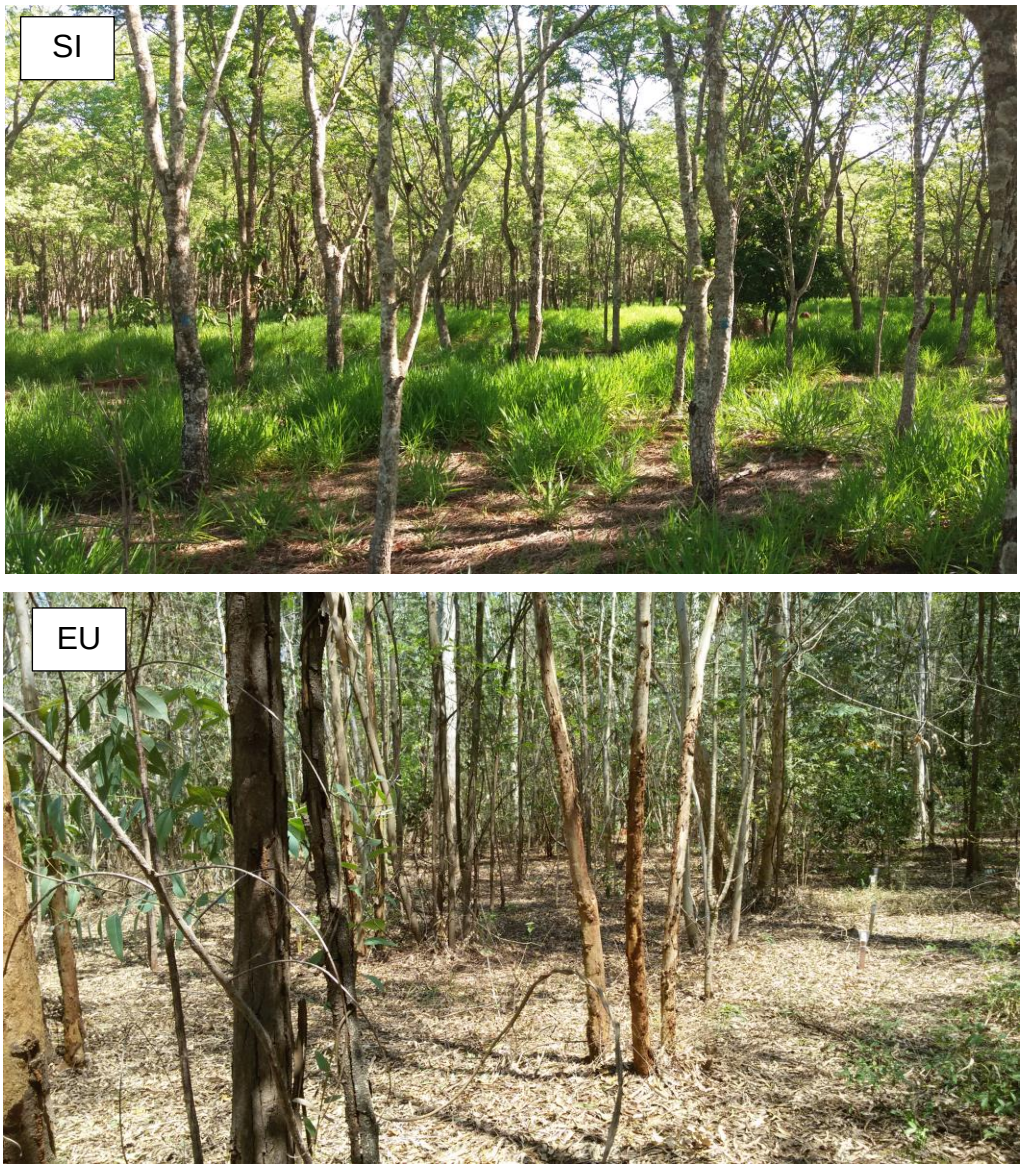

Figura 11. Aspectos gerais das áreas de estudo, silvipastoril (SI) e eucalipto (EU), ano de 2017, Município de Selvíria, Mato Grosso do Sul.

Outro aspecto que corrobora nossa afirmação é a análise da importância e seleção de variáveis para os dias 03, 10 e 17 de junho 2017 (Figuras 12 a 13) aliada à ocorrência de chuvas (precipitações), e consequentemente aumento da umidade do solo durante o período experimental (Figura 11). Observe que para EU a temperatura do solo foi a variável mais importante selecionada por todos os modelos nos dias 03 e 17 de junho, contudo, para o dia 10 de junho, a temperatura do solo perdeu sua importância, dando lugar aos atributos químicos do solo, Ca, Mg e H_Al e ao grau de

humificação da matéria orgânica do solo (HLIFS) que é uma variável atrelada à facilidade com a qual a matéria orgânica é degradada pelos microrganismos do solo (Milori et al., 2006; Panosso et al., 2011). Tal efeito pode ser atribuído a ocorrência de precipitação de leve intensidade, ocorridas nos dias 9, 10 e 11 de junho, inferiores a 10 mm, mas suficiente para alterar os padrões de variabilidade espacial de Ts, as quais foram captadas pelos modelos de aprendizado de máquina estudados.

A comparação da importância das variáveis na abordagem espacial (Figuras 12 e 13) e na abordagem temporal (Figura 8, modelo *Random Forest*) revela que a temperatura e a umidade do solo, juntamente com o tipo de cultura e o manejo das áreas, são as principais variáveis para estimar os fluxos de CO_2 nos diferentes ecossistemas estudados. Por outro lado, quanto à variabilidade espacial, o processo de modelagem deve ser realizado de maneira individual para cada área, onde as variáveis relacionadas aos aspectos químicos do solo assumem uma importância maior do que a temperatura e a umidade do solo. Além disso, é interessante notar que, a despeito da magnitude da variável alvo, a área de silvipastoril apresentou uma maior dependência espacial em comparação com a área de Eucalipto (EU). Isso se traduziu em uma maior continuidade nos padrões espaciais dos mapas e maior homogeneidade desses mapas ao longo dos dias, como observado nas Figuras 15 e 16 (Oliveira, 2018). Em adição a análise das Figuras 15 e 16, referentes aos padrões de variabilidade de FCO_2 para as áreas de SI e EU, respectivamente, nota-se subestimativas dos valores de FCO_2 nos padrões de variabilidade gerados pelos algoritmos de aprendizado de máquina para área de eucalipto. Indicando uma menor acurácia das estimativas para essa área.

Apesar de todos os esforços e do notável poder das técnicas de aprendizado estatístico, como o avançado algoritmo XGBoost, é fundamental destacar que, embora os modelos de Árvores de Decisão (DT) apresentem um desempenho inferior, eles oferecem as vantagens de custos computacionais mais baixos, ajuste simples dos hiperparâmetros e, principalmente, interpretabilidade do modelo. Portanto, eles continuam altamente recomendados para o processo de modelagem espacial da emissão de CO_2 em áreas agrícolas.

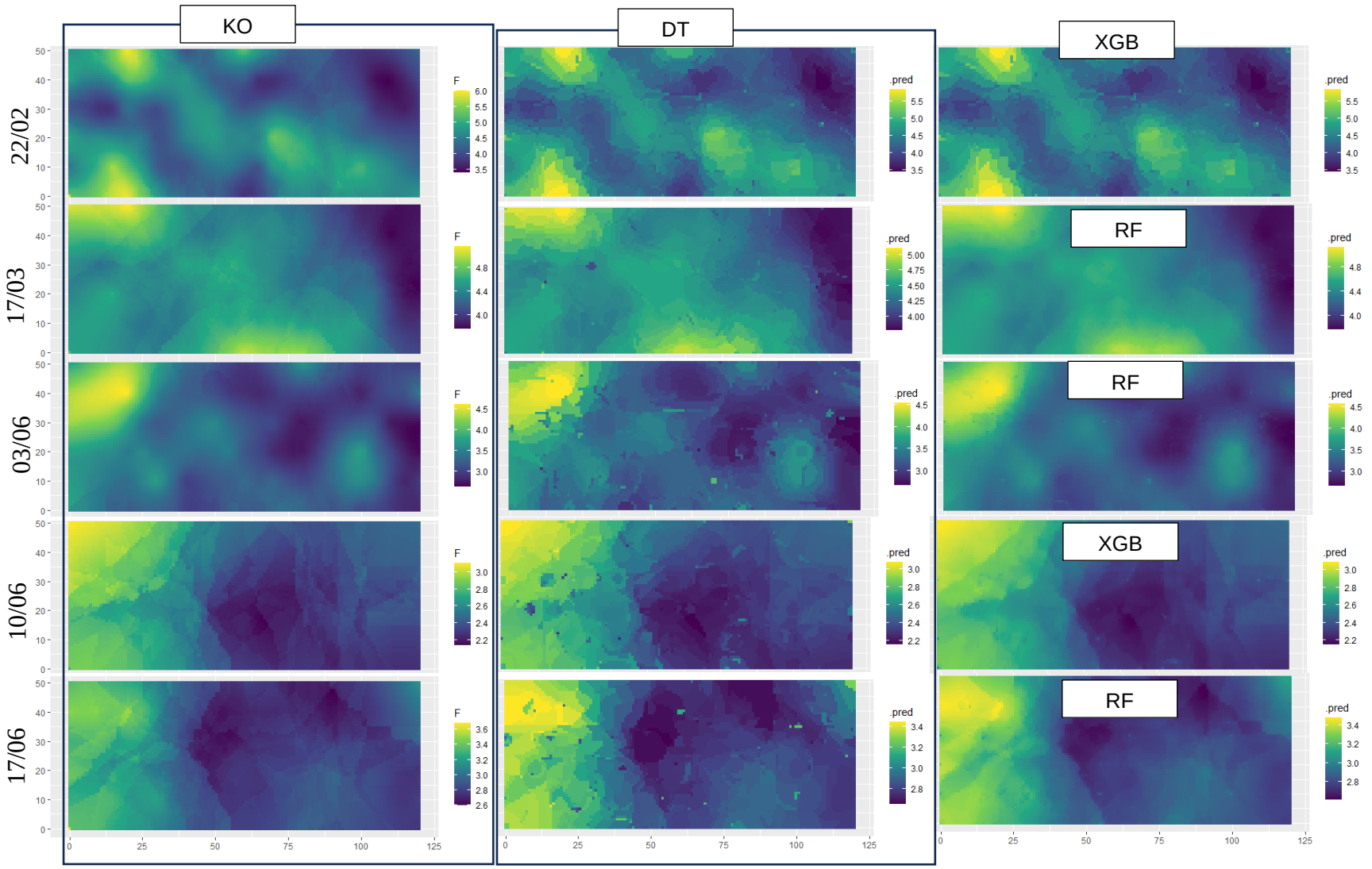

Figura 15. Padrões espaciais de emissão de CO_2 do solo estimados pela Krigagem Ordinária (KO), comparados com os mapas de padrões gerados pelo algoritmo de Árvores de Decisão (DT), bem como o melhor mapa entre os algoritmos RF e XGBoost, para cada dia estudado no uso do solo Silvipastoril.

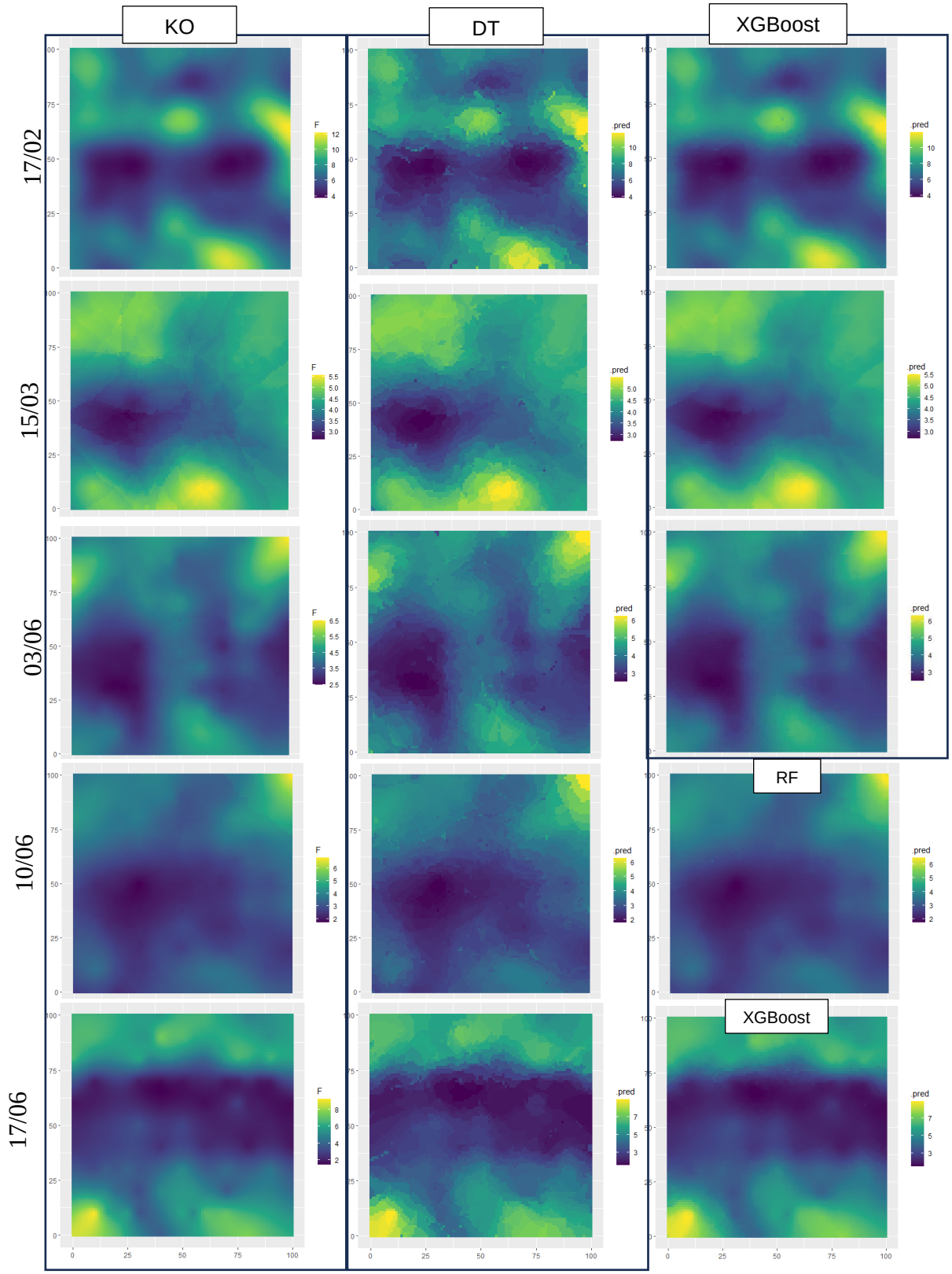

Figura 16. Padrões espaciais de emissão de CO_2 do solo estimados pela Krigagem Ordinária (KO), comparados com os mapas de padrões gerados pelo algoritmo de Árvores de Decisão (DT), bem como o melhor mapa entre os algoritmos RF e XGBoost, para cada dia estudado no uso do solo Eucalipto.

5 CONCLUSÕES

A presente pesquisa investigou a dinâmica espaçotemporal da emissão de CO_2 do solo, considerando-a como um fenômeno de estruturas multidimensionais em constante evolução ao longo do tempo, sendo que tais padrões complexos foram adequadamente descritos e modelados por meio de algoritmos do aprendizado de máquina. As técnicas de aprendizado de máquina, especialmente RF e XGBoost, demonstraram potencial em se tornarem ferramentas chaves para a compreensão e previsão da emissão de CO_2 em solos agrícolas.

Na modelagem temporal o maior valor de R^2 (0,83%) e os menores valores de RMSE e MAPE (0,93 µmol m^{-2} s^{-1} e 19,78%, respectivamente), foram observados para XGBoost. Na modelagem espacial, o estoque de carbono no solo foi a variável que definiu os padrões de variabilidade espacial da emissão de CO_2 do solo no sistema silvipastoril, ou seja, área com melhores condições de umidade do solo e menores incidências de radiação solar e evapotranspiração, associado, portanto, à componente heterotrófica da respiração do solo. Já a temperatura do solo e os atributos químicos do solo definiram os padrões de variabilidade espacial da emissão de CO_2 do solo na área de eucalipto, manejo mais sensível às variações, principalmente de umidade do solo, associado, principalmente, à componente autotrófica da emissão de CO_2, associada à respiração das raízes.

A despeito de todos os esforços, devemos destacar que, apesar das complexidades inerentes ao estudo, a abordagem de árvore de decisão (DT) demonstrou ser uma ferramenta valiosa para o mapeamento digital. Isso devido à sua simplicidade relativa, baixo custo computacional e interpretabilidade do modelo podem ser vantagens significativas na divulgação dos resultados obtidos (Apêndices).

6 CONSIDERAÇÕES FINAIS

Compreender a dinâmica espaçotemporal da emissão de CO_2 do solo é fundamental para a orientação das escolhas de práticas agrícolas que promovam a sustentabilidade ambiental e a mitigação das mudanças climáticas. Nesse contexto, os algoritmos de aprendizado de máquina revelaram-se ferramentas valiosas para modelar e estimar a transferência de carbono do solo, por meio da análise da respiração do solo. Essa abordagem contribuiu para a redução das incertezas associadas a tais projeções.

A capacidade de se identificar a contribuição relativa dos componentes heterotróficos e autotróficos para FCO_2 ao longo do tempo, e como essa dinâmica influencia a sua variabilidade espacial, é um indicador crucial para a tomada de decisões eficazes em curto prazo, necessárias no cenário atual. Portanto, é imperativo explorar novas estratégias com o objetivo de aprimorar a precisão dos modelos, como a incorporação de novos métodos de aprendizado, como redes neurais artificiais, que podem aprofundar ainda mais nossa compreensão e testes sobre as especificidades e generalidades dos modelos implementados.

Aliada à exploração de novas técnicas, variáveis estratégicas devem ser incorporadas, oriundas das mais diversas fontes de dados como: dados meteorológicos mais detalhados; informações sobre propriedades e composição do solo e da planta; biomassa radicular e dados de sensoriamento remoto de alta resolução, podem enriquecer significativamente nosso entendimento das dinâmicas ligadas ao ciclo do carbono. Essas adições tornariam os algoritmos ainda mais sensíveis a pequenas variações, permitindo uma melhor avaliação e previsão das emissões, possivelmente com menos variáveis necessárias.

Por fim, para garantir a longevidade e a eficácia contínua dessas aplicações, é essencial manter um banco de dados bem curado e atualizado. Isso envolve não apenas a coleta de novas observações e atualizações, mas também os processos de manutenção, validação e correção de informações incorretas ou ausentes. A melhoria contínua e o aumento da base de dados, com contribuições de novos experimentos e grupos de pesquisa, são passos críticos para assegurar a confiabilidade das análises futuras garantindo uma contribuição efetiva para a gestão sustentável do solo e de suas emissões de carbono.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLAIRE, S. E.; LANGE, S. F.; LAFOND, J. A.; PELLETIER, B.; CAMBOURIS, A. N.; DUTILLEUL, P. Multiscale spatial variability of CO_2 emissions and correlations with physico-chemical soil properties. **Geoderma**, v. 170, n., p. 251-260, 2012.

ALMEIDA, R. F. D.; TEIXEIRA, D. D. B.; MONTANARI, R.; BOLONHEZI, A. C.; TEIXEIRA, E. B.; MOITINHO, M. R.; PANOSSO, A. R.; SPOKAS, K. A.; JR, N. L. S. Ratio of CO_2 and O_2 as index for categorising soil biological activity in sugarcane areas under contrasting straw management regimes. **Soil Research**, v. 20 april n. 1-9, p., 2018.

AMORIM, W. P.; TETILA, E. C.; PISTORI, H.; PAPA, J. P. Semi-supervised learning with convolutional neural networks for UAV images automatic recognition. **Computers and Electronics in Agriculture**, v. 164, n., p., 2019.

BENGTSSON, H., 2023. future: Unified Parallel and Distributed Processing in R for Everyone. The Comprehensive R Archive Network.

BESALATPOUR, A. A.; AYOUBI, S.; HAJABBASI, M. A.; MOSADDEGHI, M. R.; SCHULIN, R. Estimating wet soil aggregate stability from easily available properties in a highly mountainous watershed. **Catena**, v. 111, n., p. 72-79, 2013.

BICALHO, E. S.; MOITINHO, M. R.; DE BORTOLI TEIXEIRA, D.; PANOSSO, A. R.; SPOKAS, K. A.; LA SCALA, N. Soil Greenhouse Gases: Relations to Soil Attributes in a Sugarcane Production Area. **Soil Science Society of America Journal**, v. 81, n. 5, p. 1168-1178, 2017.

BICALHO, E. S.; PANOSSO, A. R.; TEIXEIRA, D. D. B.; MIRANDA, J. G. V.; PEREIRA, G. T.; LA SCALA, N. Spatial variability structure of soil CO_2 emission and soil attributes in a sugarcane area. **Agriculture Ecosystems & Environment**, v. 189, n., p. 206-215, 2014.

BOCCA, F. F.; RODRIGUES, L. H. A. The effect of tuning, feature engineering, and feature selection in data mining applied to rainfed sugarcane yield modelling. **Computers and Electronics in Agriculture**, v. 128, n., p. 67-76, 2016.

BOCKTING, C. L.; VAN DIS, E. A. M.; BOLLEN, J.; VAN ROOIJ, R.; ZUIDEMA, W. ChatGPT: five priorities for research. **Nature**, v. 614, n. 7947, p. 224-226, 2023.

BREIMAN, L. Statistical modeling: The two cultures. **Statistical Science**, v. 16, n. 3, p. 199-215, 2001.

BREIMAN , L.; CUTLER, A.; LIAW, A.; WIENER, M., 2022. randomForest: Breiman and Cutler's Random Forests for Classification and Regression, The Comprehensive R Archive Network.

BRUCE, P.; BRUCE, A. **Practical Statistics for Data Scientists**. United States of America: O'Reilly Media, Inc., 2017. 562 p.

BRUCE, P.; BRUCE, A.; GEDECK, P. **Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python**. O'Reilly Media, 2020. p.

BURROUGH, P. A.; MCDONNELL, R. A. **Principles of geographical information system**. Oxford: University Press, 1998. 333 p.

CANTERAL, K. F. F.; VICENTINI, M. E.; DE LUCENA, W. B.; DE MORAES, M. L. T.; MONTANARI, R.; FERRAUDO, A. S.; PERUZZI, N. J.; LA SCALA, N.; PANOSSO, A. R. Machine learning for prediction of soil CO_2 emission in tropical forests in the Brazilian Cerrado. **Environmental Science and Pollution Research**, v., n., p., 2023a.

CANTERAL, K. F. F.; VICENTINI, M. E.; DE LUCENA, W. B.; DE MORAES, M. L. T.; MONTANARI, R.; FERRAUDO, A. S.; PERUZZI, N. J.; LA SCALA, N. L.; PANOSSO, A. R. Machine learning for prediction of soil CO2 emission in tropical forests in the Brazilian Cerrado. **Environmental Science and Pollution Research**, v. 30, n. 21, p. 61052-61071, 2023b.

CARVALHO, J. L. N.; CERRI, C. E. P.; FEIGL, B. J.; PICCOLO, M. C.; GODINHO, V. P.; CERRI, C. C. Carbon sequestration in agricultural soils in the Cerrado region of the Brazilian Amazon. **Soil & Tillage Research**, v. 103, n. 2, p. 342-349, 2009.

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. R.; WIRTH, R., 2000. CRISP-DM 1.0: Step-by-step data mining guide.

CHEN, G. Q.; CHEN, Z. M. Carbon emissions and resources use by Chinese economy 2007: A 135-sector inventory and input-output embodiment. **Communications in Nonlinear Science and Numerical Simulation**, v. 15, n. 11, p. 3647-3732, 2010.

CHEN, T.; HE, T.; BENESTY, M.; KHOTILOVICH, V.; TANG, Y.; CHO, H.; CHEN, K.; MITCHELL, R.; CANO, I.; ZHOU, T.; LI, M.; XIE, J.; LIN, M.; GENG, Y.; LI, Y.; YUAN, J.; IMPLEMENTATION), B. X., 2023. xgboost: Extreme Gradient Boosting. The Comprehensive R Archive Network.

CHEN, Y. B.; XU, P.; CHU, Y. Y.; LI, W. L.; WU, Y. T.; NI, L. Z.; BAO, Y.; WANG, K. Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings. **Applied Energy**, v. 195, n., p. 659-670, 2017.

CHIAVEGATTO FILHO, A. D. P.; DOS SANTOS, H. G.; DO NASCIMENTO, C. F.; MASSA, K.; KAWACHI, I. Overachieving Municipalities in Public Health: A Machine-learning Approach. **Epidemiology**, v. 29, n. 6, p. 836-840, 2018.

CHOLLET, F.; ALLAIRE, J. J. **Deep Learning with R**. 2017. 341 p.

CRESSIE, N. THE ORIGINS OF KRIGING. **Mathematical Geology**, v. 22, n. 3, p. 239-252, 1990.

CRISP, D.; FISHER, B. M.; O'DELL, C.; FRANKENBERG, C.; BASILIO, R.; BÖSCH, H.; BROWN, L. R.; CASTANO, R.; CONNOR, B.; DEUTSCHER, N. M.; ELDERING, A.; GRIFFITH, D.; GUNSON, M.; KUZE, A.; MANDRAKE, L.; MCDUFFIE, J.; MESSERSCHMIDT, J.; MILLER, C. E.; MORINO, I.; NATRAJ, V.; NOTHOLT, J.; O'BRIEN, D. M.; OYAFUSO, F.; POLONSKY, I.; ROBINSON, J.; SALAWITCH, R.; SHERLOCK, V.; SMYTH, M.; SUTO, H.; TAYLOR, T. E.; THOMPSON, D. R.; WENNBERG, P. O.; WUNCH, D.; YUNG, Y. L. The ACOS CO_2 retrieval algorithm – Part II: Global X_{CO2} data characterization. **Atmos. Meas. Tech.**, v. 5, n. 4, p. 687-707, 2012.

DA CUNHA, J. M.; CAMPOS, M. C. C.; GAIO, D. C.; DE SOUZA, Z. M.; SOARES, M. D. R.; DA SILVA, D. M. P.; SIMÕES, E. L. Spatial variability of soil respiration in Archaeological Dark Earth areas in the Amazon. **Catena**, v. 162, n., p. 148-156, 2018.

DA SILVA, P. A.; DE LIMA, B. H.; LA SCALA, N.; PERUZZI, N. J.; CHAVARETTE, F. R.; PANOSSO, A. R. Spatial variation of soil carbon stability in sugarcane crops, central-south of Brazil. **Soil & Tillage Research**, v. 202, n., p., 2020.

DAS, B.; DESAI, S.; DARIPA, A.; ANAND, G. C.; KUMAR, U.; KHALKHO, D.; THANGAVEL, V.; KUMAR, N.; REDDY, G. O.; KUMAR, P. Land degradation vulnerability mapping in a west coast river basin of India using analytical hierarchy process combined machine learning models. **Environmental Science and Pollution Research**, v., n., p., 2023.

DE FIGUEIREDO, E. B.; JAYASUNDARA, S.; DE OLIVEIRA BORDONAL, R.; BERCHIELLI, T. T.; REIS, R. A.; WAGNER-RIDDLE, C.; LA SCALA JR, N. Greenhouse gas balance and carbon footprint of beef cattle in three contrasting pasture-management systems in Brazil. **Journal of Cleaner Production**, v. 142, n., p. 420-431, 2017.

DE FIGUEIREDO, E. B.; PANOSSO, A. R.; BORDONAL, R. D.; TEIXEIRA, D. D.; BERCHIELLI, T. T.; LA SCALA, N. Soil CO2-C Emissions and Correlations with Soil Properties in Degraded and Managed Pastures in Southern Brazil. **Land Degradation & Development**, v. 28, n. 4, p. 1263-1273, 2016.

DI MININ, E.; FINK, C.; TENKANEN, H.; HIIPPALA, T. Machine learning for tracking illegal wildlife trade on social media. **Nature Ecology & Evolution**, v. 2, n. 3, p. 406-407, 2018.

DIGGLE, P. J.; RIBEIRO, P. J. **Model-based Geostatistics** New York: Springer 2010. p.

DILUSTRO, J. J.; COLLINS, B.; DUNCAN, L.; CRAWFORD, C. Moisture and soil texture effects on Soil CO_2 efflux components in southeastern mixed pine forests. **Forest Ecology and Management**, v. 204, n. 1, p. 85-95, 2005.

DUNNINGTON, D.; THORNE, B.; HERNANGÓMEZ, D., 2023. ggspatial: Spatial Data Framework for ggplot2, The Comprehensive R Archive Network.

DUTT, S.; CHANDRAMOULI, S.; DAS, A. K. **Machine Learning**. Uttar Pradesh: Pearson India Education Services Pvt. Ltd, 2018. 740 p.

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.

EMBRAPA. **Manual de métodos de análise de solo**. 2 ed. Brasília: Ministério da Agricultura e do Abastecimento / EMBRAPA-CNPS, 1997. 212 p.

FARHATE, C. V. V.; DE SOUZA, Z. M.; OLIVEIRA, S. R. D.; TAVARES, R. L. M.; CARVALHO, J. L. N. Use of data mining techniques to classify soil CO2 emission induced by crop management in sugarcane field. **Plos One**, v. 13, n. 3, p., 2018a.

FARHATE, C. V. V.; SOUZA, Z. M. D.; OLIVEIRA, S. R. D. M.; CARVALHO, J. L. N.; SCALA JÚNIOR, N. L.; SANTOS, A. P. G. Classification of soil respiration in areas of sugarcane renewal using decision tree. **Scientia Agricola**, v. 75, n., p. 216-224, 2018b.

FERRACIOLLI, M. A.; BOCCA, F. F.; RODRIGUES, L. H. A. Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models. **Computers and Electronics in Agriculture**, v. 161, n., p. 233-240, 2019.

FERREIRA, E. C.; ANZANO, J. M.; MILORI, D.; FERREIRA, E. J.; LASHERAS, R. J.; BONILLA, B.; MONTULL-IBOR, B.; CASAS, J.; NETO, L. M. Multiple Response Optimization of Laser-Induced Breakdown Spectroscopy Parameters for Multi-element Analysis of Soil Samples. **Applied Spectroscopy**, v. 63, n. 9, p. 1081-1088, 2009.

FREITAS, L. P. S.; LOPES, M. L. M.; CARVALHO, L. B.; PANOSSO, A. R.; LA SCALA JÚNIOR, N.; FREITAS, R. L. B.; MINUSSI, C. R.; LOTUFO, A. D. P. Forecasting the spatiotemporal variability of soil CO2 emissions in sugarcane areas in southeastern Brazil using artificial neural networks. **Environmental Monitoring and Assessment**, v. 190, n. 12, p. 741, 2018.

GE, Z.; SONG, Z.; DING, S. X.; HUANG, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. **IEEE Access**, v. 5, n., p. 20590-20616, 2017.

GRAF, A.; HERBST, M.; WEIHERMULLER, L.; HUISMAN, J. A.; PROLINGHEUER, N.; BORNEMANN, L.; VEREECKEN, H. Analyzing spatiotemporal variability of heterotrophic soil respiration at the field scale using orthogonal functions. **Geoderma**, v. 181, n., p. 91-101, 2012.

GREENWELL, B. M.; BOEHMKE, B., 2023. vip: Variable Importance Plots. The Comprehensive R Archive Network.

GRUNWALD, S. Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils. **Artificial intelligence and soil carbon modeling demystified: power, potentials, and perils**, v. 1, n. 1, p. 5, 2022.

HAAF, D.; SIX, J.; DOETTERL, S. Global patterns of geo-ecological controls on the response of soil respiration to warming. **Nature Climate Change**, v. 11, n. 7, p. 623-+, 2021.

HAMNER, B.; FRASCO, M.; LEDELL, E., 2018. Metrics: Evaluation Metrics for Machine Learning. The Comprehensive R Archive Network.

HANSON, B.; SUGDEN, A.; ALBERTS, B. Making Data Maximally Available. **Science**, v. 331, n. 6018, p. 649-649, 2011.

HANSON, P. J.; EDWARDS, N. T.; GARTEN, C. T.; ANDREWS, J. A. Separating root and soil microbial contributions to soil respiration: A review of methods and observations. **Biogeochemistry**, v. 48, n. 1, p. 115-146, 2000.

HENGL, T.; DE JESUS, J. M.; HEUVELINK, G. B. M.; GONZALEZ, M. R.; KILIBARDA, M.; BLAGOTIC, A.; SHANGGUAN, W.; WRIGHT, M. N.; GENG, X. Y.; BAUER-MARSCHALLINGER, B.; GUEVARA, M. A.; VARGAS, R.; MACMILLAN, R. A.; BATJES, N. H.; LEENAARS, J. G. B.; RIBEIRO, E.; WHEELER, I.; MANTEL, S.; KEMPEN, B. SoilGrids250m: Global gridded soil information based on machine learning. **Plos One**, v. 12, n. 2, p., 2017.

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

HEUVELINK, G. B. M.; ANGELINI, M. E.; POGGIO, L.; BAI, Z. G.; BATJES, N. H.; VAN DEN BOSCH, R.; BOSSIO, D.; ESTELLA, S.; LEHMANN, J.; OLMEDO, G. F.; SANDERMAN, J. Machine learning in space and time for modelling soil organic carbon change. **European Journal of Soil Science**, v., n., p., 2020.

HUANG, N.; WANG, L.; SONG, X. P.; BLACK, T. A.; JASSAL, R. S.; MYNENI, R. B.; WU, C. Y.; WANG, L.; SONG, W. J.; JI, D. B.; YU, S. S.; NIU, Z. Spatial and temporal variations in global soil respiration and their relationships with climate and land cover. **Science Advances**, v. 6, n. 41, p., 2020.

INCE, D. C.; HATTON, L.; GRAHAM-CUMMING, J. The case for open computer programs. **Nature**, v. 482, n. 7386, p. 485-488, 2012.

ISAAKS, E. H.; SRIVASTAVA, R. M. **Applied geostatistics**. Nova York: Oxford University Press, 1989. 561 p.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. São Carlos, SP. http://www.rizbicki.ufscar.br/ame/.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning -- with Applications in R**. New York: Springer, 2013. p.

JIANG, H. H.; ZHANG, C. Y.; QIAO, Y. L.; ZHANG, Z.; ZHANG, W. J.; SONG, C. Q. CNN feature based graph convolutional network for weed and crop recognition in smart farming. **Computers and Electronics in Agriculture**, v. 174, n., p., 2020.

JIANG, Y. P.; CHEN, S. F.; BIAN, B.; LI, Y. H.; SUN, Y.; WANG, X. C. Discrimination of Tomato Maturity Using Hyperspectral Imaging Combined with Graph-Based Semi-supervised Method Considering Class Probability Information. **Food Analytical Methods**, v. 14, n. 5, p. 968-983, 2021.

KAELBLING, L. P.; LITTMAN, M. L.; MOORE, A. W. Reinforcement learning: A survey. **Journal of Artificial Intelligence Research**, v. 4, n., p. 237-285, 1996.

KHAN, M. Z.; KHAN, M. F. Application of ANFIS, ANN and fuzzy time series models to CO2 emission from the energy sector and global temperature increase. **International Journal of Climate Change Strategies and Management**, v. 11, n. 5, p. 622-642, 2019.

KHAN, S.; TUFAIL, M.; KHAN, M. T.; KHAN, Z. A.; IQBAL, J.; ALAM, M. A novel semi-supervised framework for UAV based crop/weed classification. **Plos One**, v. 16, n. 5, p., 2021.

KIM, D. G.; VARGAS, R.; BOND-LAMBERTY, B.; TURETSKY, M. R. Effects of soil rewetting and thawing on soil gas fluxes: a review of current literature and suggestions for future research. **Biogeosciences**, v. 9, n. 7, p. 2459-2483, 2012.

KIRSCHBAUM, M. U. F. The temperature dependence of organic-matter decomposition - still a topic of debate. **Soil Biology & Biochemistry**, v. 38, n. 9, p. 2510-2518, 2006.

KROESE, D. P.; BOTEV, Z. I.; TAIMRE, T.; VAISMAN, R. **Data Science and Machine Learning: Mathematical and Statistical Methods**. Boca Raton: Chapman and Hall/CRC, 2019. p.

KUHN, M.; JOHNSON, K., 2013. Applied predictive modeling.

KUHN, M.; VAUGHAN, D.; HVITFELDT, E.; POSIT SOFTWARE; PBC, 2023a. parsnip: A Common API to Modeling and Analysis Functions. The Comprehensive R Archive Network.

KUHN, M.; WICKHAM, H., 2020. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles, In: tidymodels (Ed.), The Comprehensive R Archive Network.

KUHN, M.; WING, J.; WESTON, S.; WILLIAMS, A.; KEEFER, C.; ENGELHARDT, A.; COOPER, T.; MAYER, Z.; KENKEL, B.; R CORE TEAM; BENESTY, M.; LESCARBEAU, R.; ZIEM, A.; SCRUCCA, L.; TANG, Y.; CANDAN, C.; HUNT, T., 2023b. caret: Classification and Regression Training. The Comprehensive R Archive Network.

LA SCALA JR, N.; PANOSSO, A. R.; PEREIRA, G. T. Variabilidade espacial e temporal da emissão de CO_2 num latossolo desprovido de vegetação. **Engenharia Agricola**, v. 23, n. 1, p. 88-95, 2003.

LA SCALA JR., N.; PANOSSO, A. R.; PEREIRA, G. T. Modelling short-term temporal changes of bare soil CO_2 emissions in a tropical agrosystem by using meteorological data. **Applied Soil Ecology**, v. 24, n. 1, p. 113-116, 2003.

LA SCALA, N.; BOLONHEZI, D.; PEREIRA, G. T. Short-term soil CO_2 emission after conventional and reduced tillage of a no-till sugar cane area in southern Brazil. **Soil & Tillage Research**, v. 91, n. 1-2, p. 244-248, 2006.

LA SCALA, N.; LOPES, A.; PANOSSO, A. R.; CAMARA, F. T.; PEREIRA, G. T. Soil CO_2 efflux following rotary tillage of a tropical soil. **Soil & Tillage Research**, v. 84, n. 2, p. 222-225, 2005.

LA SCALA, N.; LOPES, A.; SPOKAS, K.; BOLONHEZI, D.; ARCHER, D. W.; REICOSKY, D. C. Short-term temporal changes of soil carbon losses after tillage described by a first-order decay model. **Soil & Tillage Research**, v. 99, n. 1, p. 108-118, 2008.

LA SCALA, N.; MARQUES, J.; PEREIRA, G. T.; CORA, J. E. Short-term temporal changes in the spatial variability model of CO_2 emissions from a Brazilian bare soil. **Soil Biology & Biochemistry**, v. 32, n. 10, p. 1459-1462, 2000.

LA SCALA, N.; PANOSSO, A. R.; PEREIRA, G. T.; GONZALEZ, A. P.; MIRANDA, J. G. V. Fractal dimension and anisotropy of soil CO_2 emission in an agricultural field during fallow. **International Agrophysics**, v. 23, n. 4, p. 353-358, 2009.

LAGANIERE, J.; ANGERS, D. A.; PARE, D. Carbon accumulation in agricultural soils after afforestation: a meta-analysis. **Global Change Biology**, v. 16, n. 1, p. 439-453, 2010.

LARMARANGE, J., 2023. ggstats: Extension to 'ggplot2' for Plotting Stats, In: Network, T. C. R. A. (Ed.).

LEON, E.; VARGAS, R.; BULLOCK, S.; LOPEZ, E.; PANOSSO, A. R.; LA SCALA JR, N. Hot spots, hot moments, and spatio-temporal controls on soil CO2 efflux in a water-limited ecosystem. **Soil Biology and Biochemistry**, v. 77, n. 0, p. 12-21, 2014.

LIAKOS, K. G.; BUSATO, P.; MOSHOU, D.; PEARSON, S.; BOCHTIS, D. Machine Learning in Agriculture: A Review. **Sensors**, v. 18, n. 8, p., 2018.

LINN, D. M.; DORAN, J. W. Effect of water-filled pore space on carbon dioxide and nitrous oxide production in tilled and non-tilled soils. **Soil Science Society of American Journal**, v. 48, n. 6, p. 1267-1272, 1984.

LU, H. B.; LI, S. H.; MA, M. N.; BASTRIKOV, V.; CHEN, X. Z.; CIAIS, P.; DAI, Y. J.; ITO, A.; JU, W. M.; LIENERT, S.; LOMBARDOZZI, D.; LU, X. J.; MAIGNAN, F.; NAKHAVALI, M.; QUINE, T.; SCHINDLBACHER, A.; WANG, J.; WANG, Y. P.; WARLIND, D.; ZHANG, S. P.; YUAN, W. P. Comparing machine learning-derived global estimates of soil respiration and its components with those from terrestrial ecosystem models. **Environmental Research Letters**, v. 16, n. 5, p., 2021.

LU, R. H.; ZHANG, P.; FU, Z. P.; JIANG, J.; WU, J. C.; CAO, Q.; TIAN, Y. C.; ZHU, Y.; CAO, W. X.; LIU, X. J. Improving the spatial and temporal estimation of ecosystem respiration using multi-source data and machine learning methods in a rainfed winter wheat cropland. **Science of the Total Environment**, v. 871, n., p., 2023.

MAHMOUDZADEH, H.; MATINFAR, H. R.; TAGHIZADEH-MEHRJARDI, R.; KERRY, R. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. **Geoderma Regional**, v. 21, n., p., 2020.

MAKIRANTA, P.; MINKKINEN, K.; HYTONEN, J.; LAINE, J. Factors causing temporal and spatial variation in heterotrophic and rhizospheric components of soil respiration in afforested organic soil croplands in Finland. **Soil Biology & Biochemistry**, v. 40, n. 7, p. 1592-1600, 2008.

MALAVOLTA, E.; VITTI, G. C.; OLIVEIRA, A. S. **Avaliação do estado nutricional das plantas: princípios e aplicações**. Piracicaba: Potafós, 1997. p.

MANTOVANELLI, B. C.; CAMPOS, M. C. C.; ALHO, L. C.; SILVA, P. C.; SILVA, D. A. P.; CUNHA, J. M.; SILVA, D. M. P.; SOARES, M. D. R. Distribuição espacial da emissão de CO_2 e atributos do solo sob campo nativo na região de Humaitá, Amazonas. **Sociedade & Natureza**, v. 28, n. 2, p., 2016.

MARTINEZ, G.; WELTZ, M.; PIERSON, F. B.; SPAETH, K. E.; PACHEPSKY, Y. Scale effects on runoff and soil erosion in rangelands: estimations with predictors of different availability Observations and. **Catena**, v. 151, n., p. 161-173, 2017.

MCBRATNEY, A.; DE GRUIJTER, J.; BRYCE, A. Pedometrics timeline. **Geoderma**, v. 338, n., p. 568-575, 2019.

MCBRATNEY, A. B.; ODEH, I. O. A.; BISHOP, T. F. A.; DUNBAR, M. S.; SHATAR, T. M. An overview of pedometric techniques for use in soil survey. **Geoderma**, v. 97, n. 3-4, p. 293-327, 2000.

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

MILORI, D.; GALETI, H. V. A.; MARTIN-NETO, L.; DIECKOW, J.; GONZALEZ-PEREZ, M.; BAYER, C.; SALTON, J. Organic matter study of whole soil samples using laser-induced fluorescence spectroscopy. **Soil Science Society of America Journal**, v. 70, n. 1, p. 57-63, 2006.

MOHAMMED, G. H.; COLOMBO, R.; MIDDLETON, E. M.; RASCHER, U.; VAN DER TOL, C.; NEDBAL, L.; GOULAS, Y.; PÉREZ-PRIEGO, O.; DAMM, A.; MERONI, M.; JOINER, J.; COGLIATI, S.; VERHOEF, W.; MALENOVSKÝ, Z.; GASTELLU-

ETCHEGORRY, J.-P.; MILLER, J. R.; GUANTER, L.; MORENO, J.; MOYA, I.; BERRY, J. A.; FRANKENBERG, C.; ZARCO-TEJADA, P. J. Remote sensing of solar-induced chlorophyll fluorescence (SIF) in vegetation: 50 years of progress. **Remote Sensing of Environment**, v. 231, n., p. 111177, 2019.

MOITINHO, M. R.; PADOVAN, M. P.; PANOSSO, A. R.; TEIXEIRA, D. D. B.; FERRAUDO, A. S.; LA SCALA, N., JR. On the spatial and temporal dependence of CO2 emission on soil properties in sugarcane (Saccharum spp.) production. **Soil & Tillage Research**, v. 148, n., p. 127-132, 2015a.

MOITINHO, M. R.; PANDOVAN, M. P.; PANOSSO, A. R.; LA SCALA JR, N. Efeitos do preparo do solo e resíduos da colheita da cna-de-açúcar sobre a emissão de CO2. **Revista Brasileira de Ciência do Solo**, v. 37, n., p. 1720-1728, 2013.

MOITINHO, M. R.; PANDOVAN, M. P.; PANOSSO, A. R.; TEIXEIRA, D. B.; FERRAUDO, A. S.; LA SCALA JR, N. On the spatial and temporal dependence of CO_2 emission on soil properties in sugarcane (Saccharum spp.) production. **Soil & Tillage Research**, v. 148, n., p. 127–132, 2015b.

MORAES, J. F. L.; VOLKOFF, B.; CERRI, C. C.; BERNOUX, M. Soil properties under Amazon forest and changes due to pasture installation in Rondonia, Brazil. **Geoderma**, v. 70, n. 1, p. 63-81, 1996.

NG, A. Y.-T. **Machine Learning Yearning**. deeplearning.ai, https://github.com/ajaymache/machine-learning-yearning/tree/master.

O'DELL, C. W.; CONNOR, B.; BÖSCH, H.; O'BRIEN, D.; FRANKENBERG, C.; CASTANO, R.; CHRISTI, M.; ELDERING, D.; FISHER, B.; GUNSON, M.; MCDUFFIE, J.; MILLER, C. E.; NATRAJ, V.; OYAFUSO, F.; POLONSKY, I.; SMYTH, M.; TAYLOR, T.; TOON, G. C.; WENNBERG, P. O.; WUNCH, D. The ACOS CO_2 retrieval algorithm – Part 1: Description and validation against synthetic observations. **Atmos. Meas. Tech.**, v. 5, n. 1, p. 99-121, 2012.

OLIVEIRA, C. F. **Variabilidade espacial da emissão de CO2 e estoque de carbono do solo em áreas de eucalipto e sistema silvipastoril**. Ilha Solteira - SP, Universidade Estadual Paulista, 2018. 158p.

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.

PANOSSO, A. R.; CAMARA, F. T.; LOPES, A.; PEREIRA, G. T.; LA SCALA JR, N. Emissão de CO_2 em um Latossolo após preparo convencional e reduzido em períodos seco e chuvoso. **Científica**, v. 34, n. 2, p. 257-262, 2006.

PANOSSO, A. R.; MARQUES, J.; MILORI, D.; FERRAUDO, A. S.; BARBIERI, D. M.; PEREIRA, G. T.; LA SCALA, N. Soil CO_2 emission and its relation to soil properties in sugarcane areas under Slash-and-burn and Green harvest. **Soil & Tillage Research**, v. 111, n. 2, p. 190-196, 2011.

PANOSSO, A. R.; MARQUES, J.; PEREIRA, G. T.; LA SCALA, N. Spatial and temporal variability of soil CO_2 emission in a sugarcane area under green and slash-and-burn managements. **Soil & Tillage Research**, v. 105, n. 2, p. 275-282, 2009a.

PANOSSO, A. R.; PEREIRA, G. T.; MARQUES, J.; LA SCALA, N. Spatial Variability of CO_2 emission on oxisol soils cultivated with sugar cane under different management practices. **Engenharia Agricola**, v. 28, n. 2, p. 227-236, 2008.

PANOSSO, A. R.; PERILLO, L. I.; FERRAUDO, A. S.; PEREIRA, G. T.; MIRANDA, J. G. V.; LA SCALA JR, N. Fractal dimension and anisotropy of soil CO_2 emission in a mechanically harvested sugarcane production area. **Soil & Tillage Research**, v. 124, n., p. 8-16, 2012.

PANOSSO, A. R.; RIBEIRO, C. E. R.; ZANINI, J. R.; PAVANI, L. C.; PEREIRA, G. T.; LA SCALA, N. Spatial variability of CO_2 emission, temperature and moisture of a bare oxisol submitted to different wetting levels. **Semina-Ciencias Agrarias**, v. 30, n., p. 1017-1033, 2009b.

PEBESMA, E.; BIVAND, R. Classes and methods for spatial data in R. **R News**, v. 5, n. 3, p. 9–13, 2015.

PEDERSEN, T. L., 2023. patchwork: The Composer of Plots. The Comprehensive R Archive Network.

PEREIRA, R. H. M.; GONCALVES, C. N.; ARAUJO, P. H. F. D.; CARVALHO, G. D.; ARRUDA, R. A. D.; NASCIMENTO, I.; DA COSTA, B. S. P.; CAVEDO, W. S.; ANDRADE, P. R.; DA SILVA, A.; BRAGA, C. K. V.; SCHMERTMANN, C.; SAMUEL-ROSA, A.; FERREIRA, D.; SARAIVA, M., 2023. Easy access to official spatial data sets of Brazil as 'sf' objects in R. The package includes a wide range of geospatial data available at various geographic scales and for various years with harmonized attributes, projection and fixed topology., In: Network, T. C. R. A. (Ed.).

PINHEIRO DA SILVA, D. A.; CAMPOS, M. C. C.; MANTOVANELLI, B. C.; SANTOS, L. A. C. D.; SOARES, M. D. R.; CUNHA, J. M. D. Variabilidade espacial da emissão de CO_2, temperatura e umidade do solo em área de pastagem na região Amazônica, Brasil. **Revista de Ciências Agroveterinárias**, v. 18, n. 1, p. 119-126, 2019.

POSIT TEAM. **RStudio: Integrated Development Environment for R.** Posit Software, PBC, Boston, MA. http://www.posit.co/.

R DEVELOPMENT CORE TEAM. **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

RAIJ, B. V. **Análise química para avaliação da fertilidade de solos tropicais**. Campinas: Instituto Agronômico, 2001. 285 p.

RAIJ, B. V.; DE ANDRADE, J. C.; CANTARELLA, H.; QUAGGIO, J. A. **Análise química do solo para fins de fertilidade**. Campinas: Fundação Cargill, 1987. 170 p.

REICHSTEIN, M.; CAMPS-VALLS, G.; STEVENS, B.; JUNG, M.; DENZLER, J.; CARVALHAIS, N.; PRABHAT Deep learning and process understanding for data-driven Earth system science. **Nature**, v. 566, n. 7743, p. 195-204, 2019.

RETH, S.; REICHSTEIN, M.; FALGE, E. The effect of soil water content, soil temperature, soil pH-value and the root mass on soil CO_2 efflux - A modified model. **Plant and Soil**, v. 268, n. 1-2, p. 21-33, 2005.

RODRIGO, A.; RECOUS, S.; NEEL, C.; MARY, B. Modelling temperature and moisture effects on C-N transformations in soils: comparison of nine models. **Ecological Modelling**, v. 102, n. 2-3, p. 325-339, 1997.

RUBIO, V. E.; DETTO, M. Spatiotemporal variability of soil respiration in a seasonal tropical forest. **Ecology and Evolution**, v. 7, n. 17, p. 7104-7116, 2017.

SAIZ, G.; GREEN, C.; BUTTERBACH-BAHL, K.; KIESE, R.; AVITABILE, V.; FARRELL, E. P. Seasonal and spatial variability of soil respiration in four Sitka spruce stands. **Plant and Soil**, v. 287, n. 1-2, p. 161-176, 2006.

SAMUEL, A. L. SOME STUDIES IN MACHINE LEARNING USING THE GAME OF CHECKERS. **Ibm Journal of Research and Development**, v. 3, n. 3, p. 211-&, 1959.

SANCHES, G. M.; GRAZIANO MAGALHÃES, P. S.; JUNQUEIRA FRANCO, H. C. Site-specific assessment of spatial and temporal variability of sugarcane yield related to soil attributes. **Geoderma**, v. 334, n., p. 90-98, 2019.

SANTOS, G. A. D.; MOITINHO, M. R.; SILVA, B. D.; XAVIER, C. V.; TEIXEIRA, D. D.; CORA, J. E.; LA SCALA, N. Effects of long-term no-tillage systems with different succession cropping strategies on the variation of soil CO2 emission. **Science of the Total Environment**, v. 686, n., p. 413-424, 2019a.

SANTOS, H. G. D.; NASCIMENTO, C. F. D.; IZBICKI, R.; DUARTE, Y. A. D. O.; PORTO CHIAVEGATTO FILHO, A. D. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, v. 35, n., p., 2019b.

SARTORI, F.; LAL, R.; EBINGER, M. H.; PARRISH, D. J. Potential soil carbon sequestration and CO_2 offset by dedicated energy crops in the USA. **Critical Reviews in Plant Sciences**, v. 25, n. 5, p. 441-472, 2006.

SEEG. **Análise das emissões brasileiras de gases de efeito estufa e suas implicações para as metas de clima do Brasil 1970-2019**. Observatório do Clima, 2020. 40 p.

SEGNINI, A.; CARVALHO, J. L. N.; BOLONHEZI, D.; MILORI, D.; DA SILVA, W. T. L.; SIMOES, M. L.; CANTARELLA, H.; DE MARIA, I. C.; MARTIN-NETO, L. Carbon stock and humification index of organic matter affected by sugarcane straw and soil management. **Scientia Agricola**, v. 70, n. 5, p. 321-326, 2013.

SHADRIN, D.; PUKALCHIK, M.; KOVALEVA, E.; FEDOROV, M. Artificial intelligence models to predict acute phytotoxicity in petroleum contaminated soils. **Ecotoxicology and Environmental Safety**, v. 194, n., p., 2020.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. **Understanding Machine Learning: From Theory to Algorithms**. New York, NY: Cambridge University Press, 2014. 449 p.

SHOREWALA, S.; ASHFAQUE, A.; SIDHARTH, R.; VERMA, U. Weed Density and Distribution Estimation for Precision Agriculture Using Semi-Supervised Learning. **IEEE Access**, v. 9, n., p. 27971-27986, 2021.

SIERRA, C. A.; TRUMBORE, S. E.; DAVIDSON, E. A.; VICCA, S.; JANSSENS, I. Sensitivity of decomposition rates of soil organic matter with respect to simultaneous changes in temperature and moisture. **Journal of Advances in Modeling Earth Systems**, v. 7, n. 1, p. 335-356, 2015.

SILVA, B. D.; MOITINHO, M. R.; SANTOS, G. A. D.; TEIXEIRA, D. D.; FERNANDES, C.; LA SCALA, N. Soil CO2 emission and short-term soil pore class distribution after tillage operations. **Soil & Tillage Research**, v. 186, n., p. 224-232, 2019.

SMITH, P.; MARTINO, D.; CAI, Z.; GWARY, D.; JANZEN, H.; KUMAR, P.; MCCARL, B.; OGLE, S.; O'MARA, F.; RICE, C.; SCHOLES, B.; SIROTENKO, O.; HOWDEN, M.; MCALLISTER, T.; PAN, G.; ROMANENKOV, V.; SCHNEIDER, U.; TOWPRAYOON, S.; WATTENBACH, M.; SMITH, J. Greenhouse gas mitigation in agriculture. **Philosophical Transactions of the Royal Society B-Biological Sciences**, v. 363, n. 1492, p. 789-813, 2008.

SOE, A. R. B.; BUCHMANN, N. Spatial and temporal variations in soil respiration in relation to stand structure and soil parameters in an unmanaged beech forest. **Tree Physiology**, v. 25, n. 11, p. 1427-1436, 2005.

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.

STOKEL-WALKER, C. CHATGPT LISTED AS AUTHOR ON RESEARCH PAPERS. **Nature**, v. 613, n. 7945, p. 620-U621, 2023.

TANG, X. L.; FAN, S. H.; DU, M. Y.; ZHANG, W. J.; GAO, S. C.; LIU, S. B.; CHEN, G.; YU, Z.; YANG, W. N. Spatial and temporal patterns of global soil heterotrophic respiration in terrestrial ecosystems. **Earth System Science Data**, v. 12, n. 2, p. 1037-1051, 2020.

TAVANTI, R. F. R.; MONTANARI, R.; PANOSSO, A. R.; FREDDI, O. D. S.; PAZ-GONZÁLEZ, A. PEDOTRANSFER FUNCTION TO ESTIMATE THE SOIL STRUCTURAL S INDEX AND SPATIAL VARIABILITY IN AN OXISOL WITHIN A LIVESTOCK FARMING SYSTEM. **Engenharia Agricola**, v. 40, n., p. 34-44, 2020a.

TAVANTI, R. F. R.; MONTANARI, R.; PANOSSO, A. R.; LA SCALA, N.; CHIQUITELLI NETO, M.; FREDDI, O. D. S.; PAZ GONZÁLEZ, A.; DE CARVALHO, M. A. C.; SOARES, M. B.; TAVANTI, T. R.; GALINDO, F. S. What is the impact of pasture reform on organic carbon compartments and CO2 emissions in the Brazilian Cerrado? **Catena**, v. 194, n., p. 104702, 2020b.

TAVARES, R. L. M.; OLIVEIRA, S. R. D.; DE BARROS, F. M. M.; FARHATE, C. V. V.; DE SOUZA, Z. M.; LA SCALA, N. Prediction of soil CO2 flux in sugarcane management systems using the Random Forest approach. **Scientia Agricola**, v. 75, n. 4, p. 281-287, 2018.

TEIXEIRA, D. B.; BICALHO, E. S.; CERRI, C. E. P.; PANOSSO, A. R.; PEREIRA, G. T.; LA SCALA JR, N. Quantification of uncertainties associated with space-time estimates of short-term soil CO_2 emissions in a sugar cane area. **Agriculture, Ecosystems and Environment**, v. 167, n., p. 33-37, 2013a.

TEIXEIRA, D. B.; BICALHO, E. S.; PANOSSO, A. R.; CERRI, C. E. P.; PEREIRA, G. T.; LA SCALA JR, N. Spatial variability of soil CO_2 emission in a sugarcane area characterized by secondary information. **Scientia Agricola**, v. 70, n. 3, p. 195-203, 2013b.

TEIXEIRA, D. B.; PANOSSO, A. R.; CERRI, C. E. P.; PEREIRA, G. T.; LA SCALA, N. Soil CO(2) emission estimated by different interpolation techniques. **Plant and Soil**, v. 345, n. 1-2, p. 187-194, 2011a.

TEIXEIRA, L. G.; FUKUDA, A.; PANOSSO, A. R.; LOPES, A.; LA SCALA, N. SOIL CO(2) EMISSION AS RELATED TO INCORPORATION OF SUGARCANE CROP RESIDUES AND AGGREGATE BREAKING AFTER ROTARY TILLER. **Engenharia Agricola**, v. 31, n. 6, p. 1075-1084, 2011b.

TERÇARIOL, M. C.; BRANCAGLIONI, V. A.; ARTÊNCIO JÚNIOR, J. P.; MONTANARI, R.; TEIXEIRA FILHO, M. C. M.; BOLONHEZI, A. C.; LA SCALA JR, N.; CHAVARETTE, F. R.; PANOSSO, A. R. Spatial variability of soil CO_2 emission in soybean and sugarcane areas in Mato Grosso do Sul Cerrado, Brazil. **Journal of Geospatial Modelling**, v. 2, n., p. 1-7, 2016.

TETILA, E. C.; MACHADO, B. B.; ASTOLFI, G.; BELETE, N. A. D.; AMORIM, W. P.; ROEL, A. R.; PISTORI, H. Detection and classification of soybean pests using deep learning with UAV images. **Computers and Electronics in Agriculture**, v. 179, n., p., 2020.

THERNEAU, T.; ATKINSON, B.; RIPLEY, B., 2022. rpart: Recursive Partitioning and Regression Trees. The Comprehensive R Archive Network.

TIERNEY, N. visdat: Visualising Whole Data Frames. **JOSS**, v. 2, n. 16, p. 355, 2017.

TRAMONTANA, G.; JUNG, M.; SCHWALM, C. R.; ICHII, K.; CAMPS-VALLS, G.; RADULY, B.; REICHSTEIN, M.; ARAIN, M. A.; CESCATTI, A.; KIELY, G.; MERBOLD, L.; SERRANO-ORTIZ, P.; SICKERT, S.; WOLF, S.; PAPALE, D. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. **Biogeosciences**, v. 13, n. 14, p. 4291-4313, 2016.

TRANGMAR, B. B.; YOST, R. S.; UEHARA, G. Application of geostatistics to spatial studies of soil properties. **Advances in Agronomy**, v. 38, n., p. 45-94, 1985.

USSIRI, D. A. N.; LAL, R. Long-term tillage effects on soil carbon storage and carbon dioxide emissions in continuous corn cropping system from an alfisol in Ohio. **Soil & Tillage Research**, v. 104, n. 1, p. 39-47, 2009.

VARGAS, R.; SANCHEZ-CANETE, E.; SERRANO-ORTIZ, P.; YUSTE, J. C.; DOMINGO, F.; LOPEZ-BALLESTEROS, A.; OYONARTE, C. Hot-Moments of Soil CO2 Efflux in a Water-Limited Grassland. **Soil Systems**, v. 2, n. 3, p., 2018.

VAUCLIN, M.; VIEIRA, S. R.; VACHAUD, G.; NIELSEN, D. R. The use of cokriging with limited field soil observations. **Soil Science Society of America Journal**, v. 47, n. 2, p. 175-184, 1983.

VICENTINI, M. E.; DA SILVA, P. A.; CANTERAL, K. F. F.; DE LUCENA, W. B.; DE MORAES, M. L. T.; MONTANARI, R.; FILHO, M. C. M. T.; PERUZZI, N. J.; LA SCALA, N.; DE SOUZA ROLIM, G.; PANOSSO, A. R. Artificial neural networks and adaptive neuro-fuzzy inference systems for prediction of soil respiration in forested areas southern Brazil. **Environmental Monitoring and Assessment**, v. 195, n. 9, p. 1074, 2023.

VICENTINI, M. E.; PINOTTI, C. R.; HIRAI, W. Y.; DE MORAES, M. L. T.; MONTANARI, R.; FILHO, M. C. M. T.; MILORI, D. M. B. P.; JÚNIOR, N. L. S.; PANOSSO, A. R. CO_2

emission and its relation to soil temperature, moisture, and O2 absorption in the reforested areas of Cerrado biome, Central Brazil. **Plant and Soil**, v., n., p., 2019.

VIEIRA, S. R. Geoestatística em estudos de variabilidade espacial do solo, In: Novais, R. F., Alvarez V., V. H.; Schaefer, C. E. (Eds.), **Tópicos em ciência do solo**. Viçosa, MG: Sociedade Brasileira de Ciência do Solo, 2000. p. 1-54.

VIEIRA, S. R.; TILLOTSON, P. M.; BIGGAR, J. W.; NIELSEN, D. R. Scaling of semivariograms and the kriging estimation of field-measured properties. **Revista Brasileira de Ciência do Solo**, v. 21, n., p. 525-533, 1997.

WARING, E.; QUINN, M.; MCNAMARA, A.; DE LA RUBIA, E. A.; ZHU, H.; LOWNDES, J.; SELLIS, H.; MCLEOD, H.; WICKHAM, H.; MÜLLER, K.; RSTUDIO, I.; KIRKPATRICK, C.; SCOTT BRENSTUHL; PATRICK SCHRATZ; LBUSETT; KORPELA, M.; THOMPSON, J.; MCGEHEE, H.; ROEPKE, M.; KENNEDY, P.; POSSENRIEDE, D.; ZIMMERMANN, D.; BUTTS, K.; TORGES, B.; SAPORTA, R.; STEWART, H. M., 2022. skimr: Compact and Flexible Summaries of Data. The Comprehensive R Archive Network.

WARNER, D. L.; BOND-LAMBERTY, B.; JIAN, J.; STELL, E.; VARGAS, R. Spatial Predictions and Associated Uncertainty of Annual Soil Respiration at the Global Scale. **Global Biogeochemical Cycles**, v. 33, n. 12, p. 1733-1745, 2019.

WARNER, D. L.; VARGAS, R.; SEYFFERTH, A.; INAMDAR, S. Transitional slopes act as hotspots of both soil CO2 emission and CH4 uptake in a temperate forest landscape. **Biogeochemistry**, v. 138, n. 2, p. 121-135, 2018.

WEBSTER, R. Quantitative spatial analysis of soil in the field, In: Stewart, B. A. (Ed.), **Advances in Soil Science**. New York: Springer, 1985. p. 1-70.

WEBSTER, R.; OLIVER, M. A. **Statistical methods in soil and land resource survey**. New York: Oxford University Press, 1990. 328 p.

WEI, T.; SIMKO, V., 2021. R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.92). The Comprehensive R Archive Network.

WICKHAM, H.; AVERICK, M.; BRYAN, J.; CHANG, W.; MCGOWAN, L.; FRANÇOIS, R.; GROLEMUND, G.; HAYES, A.; HENRY, L.; HESTER, J.; KUHN, M.; PEDERSEN, T.; MILLER, E.; BACHE, S.; MÜLLER, K.; OOMS, J.; ROBINSON, D.; SEIDEL, D.; SPINU, V.; TAKAHASHI, K.; VAUGHAN, D.; WILKE, C.; WOO, K.; YUTANI, H. Welcome to the tidyverse. **Journal of Open Source Software**, v. 4, n. 43, p. 1686, 2019.

WMO. **WMO Greenhouse Gas Bulletin (GHG Bulletin) - No. 18: The State of Greenhouse Gases in the Atmosphere Based on Global Observations through 2021**. World Meteorological Organization, World Data Centre for Greenhouse Gases, Tokyo, Japan. https://reliefweb.int/attachments/839863f6-d52c-4e9d-bf79-ba72a405748d/GHG_18_en.pdf.

XU, M.; QI, Y. Soil-surface CO_2 efflux and its spatial and temporal variations in a young ponderosa pine plantation in northern California. **Global Change Biology**, v. 7, n. 6, p. 667-677, 2001.

YANG, P.; ZHAO, Q. K.; CAI, X. M. Machine learning based estimation of land productivity in the contiguous US using biophysical predictors. **Environmental Research Letters**, v. 15, n. 7, p., 2020.

YILMAZ, I.; KAYNAR, O. Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. **Expert Systems with Applications**, v. 38, n. 5, p. 5958-5966, 2011.

YU, L.; WEN, J.; CHANG, C. Y.; FRANKENBERG, C.; SUN, Y. High-Resolution Global Contiguous SIF of OCO-2. **Geophysical Research Letters**, v. 46, n. 3, p. 1449-1458, 2019.

ZHANG, L.; YAN, W. D.; LIU, Y. J.; LIANG, X. C.; CHEN, X. Y. Simulation of soil CO2 efflux under different hydrothermal conditions based on general regression neural network. **Agricultural and Forest Meteorology**, v. 316, n., p., 2022.

ZHOU, T.; GENG, Y. J.; CHEN, J.; PAN, J. J.; HAASE, D.; LAUSCH, A. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. **Science of the Total Environment**, v. 729, n., p., 2020.

APÊNDICE

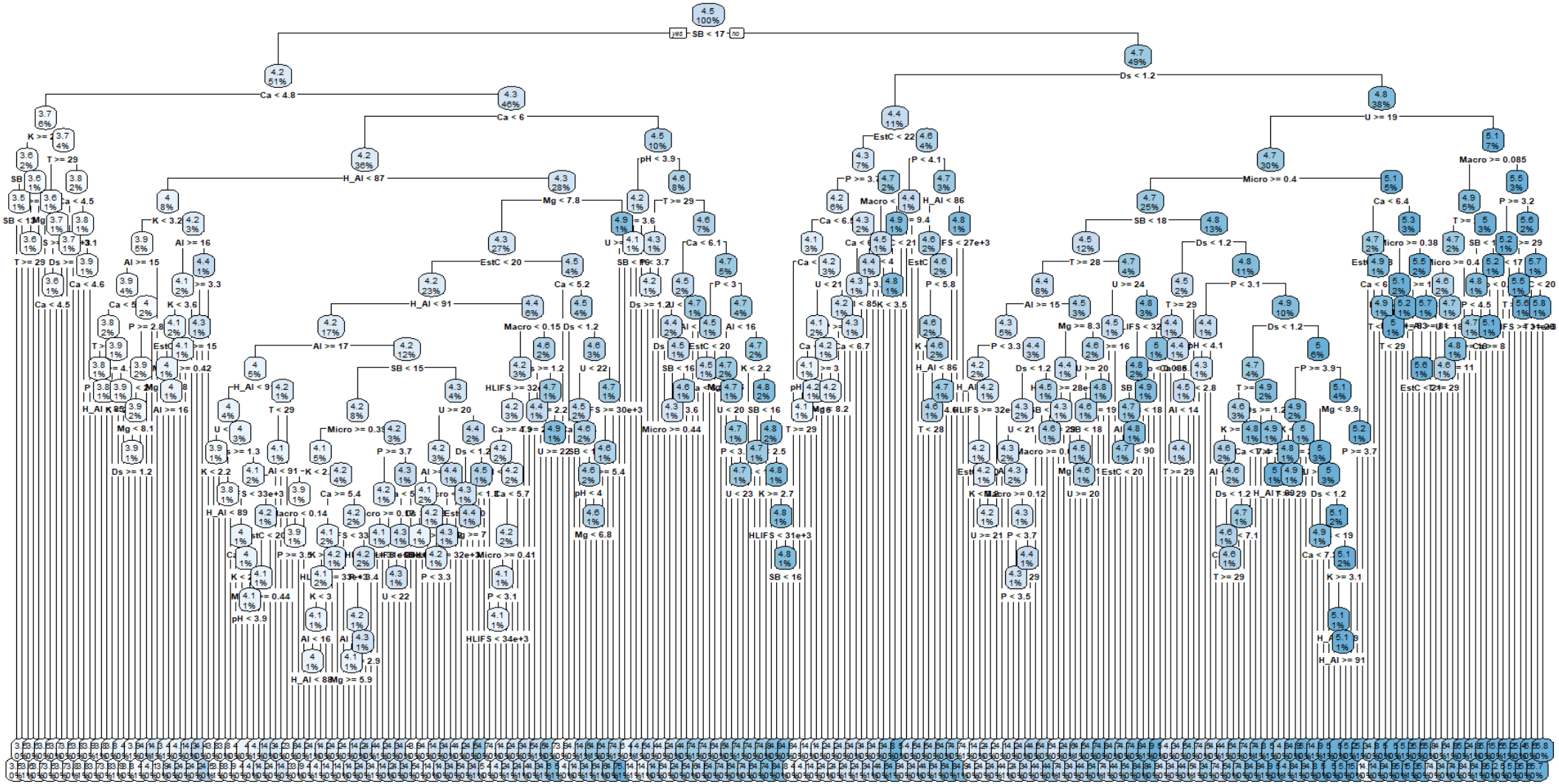

Figura 1A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 22 de fevereiro de 2017 para sistema Silvipastoril. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

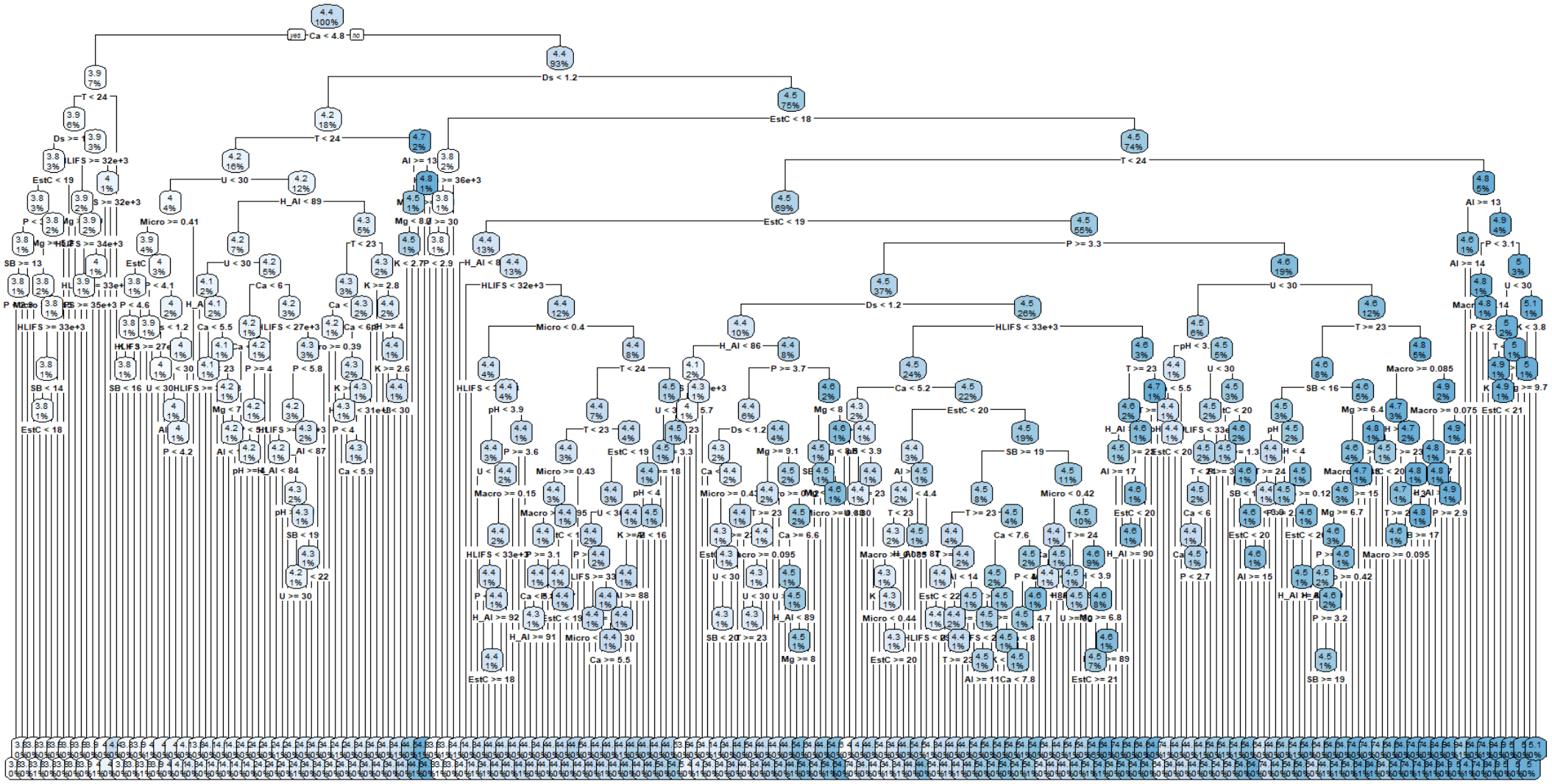

Figura 2A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 17 de março de 2017 para sistema Silvipastoril. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

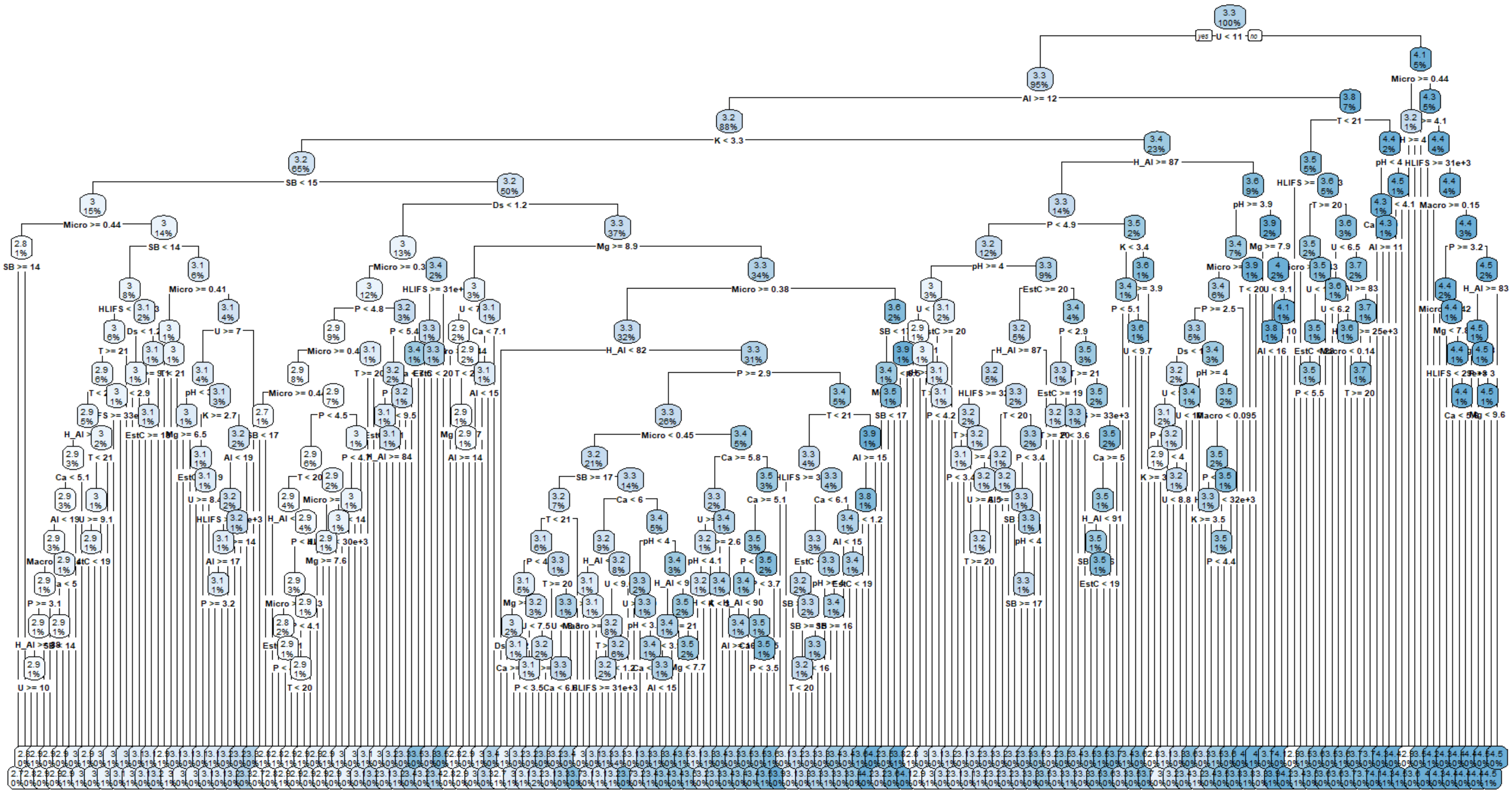

Figura 3A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 03 de junho de 2017 para sistema Silvipastoril. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

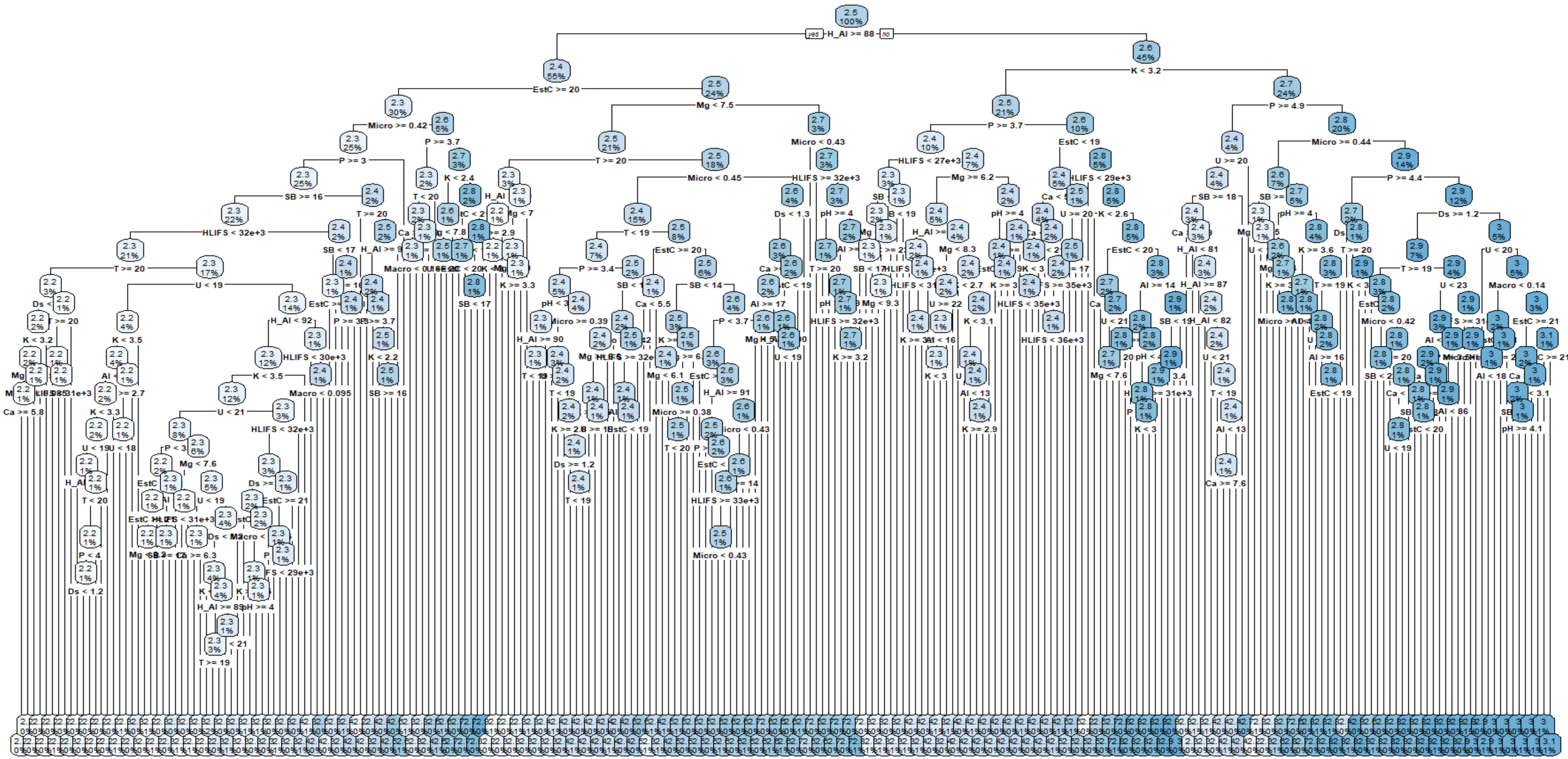

Figura 4A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 10 de junho de 2017 para sistema Silvipastoril. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

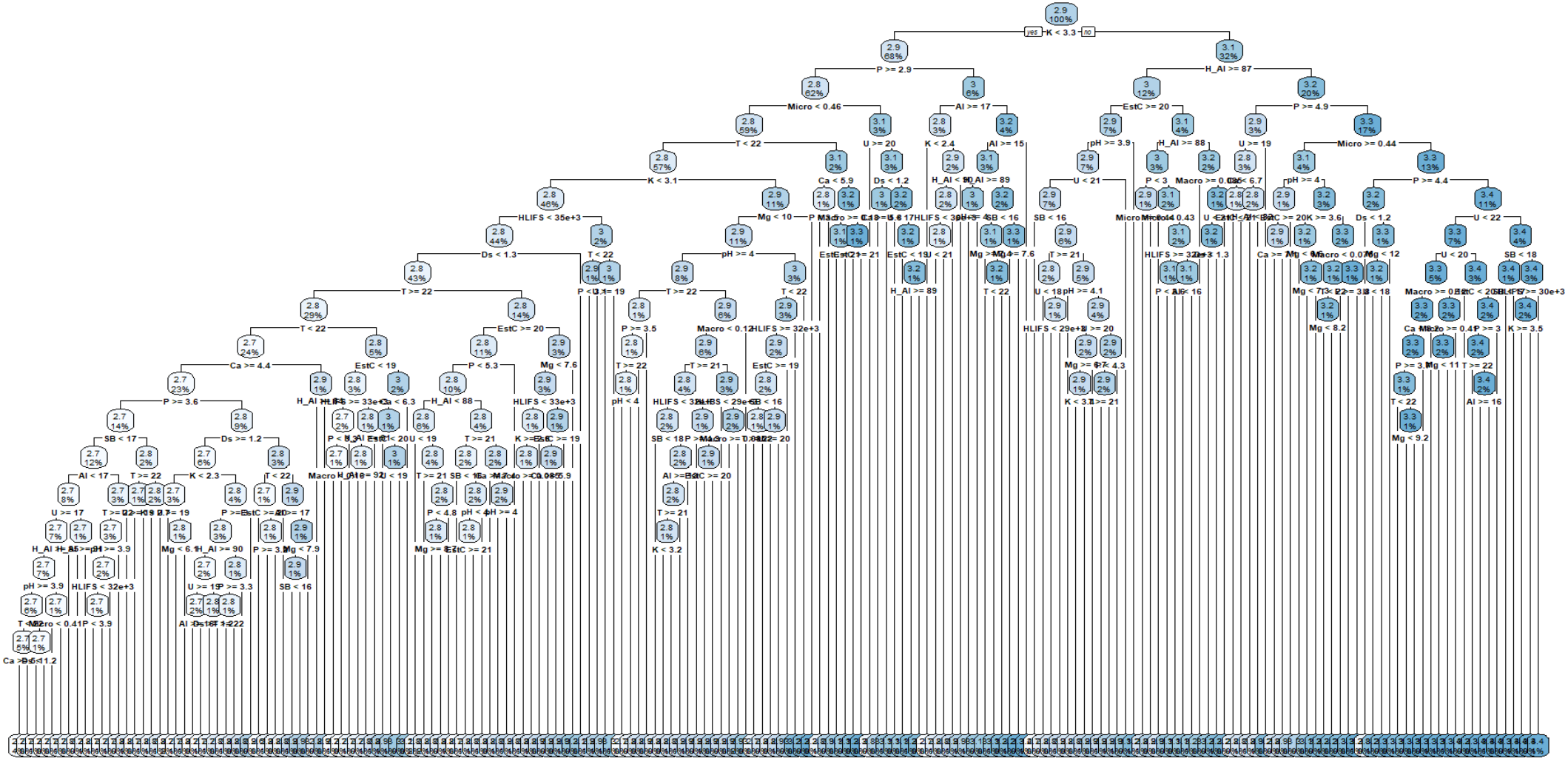

Figura 5A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 17 de junho de 2017 para sistema Silvipastoril. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

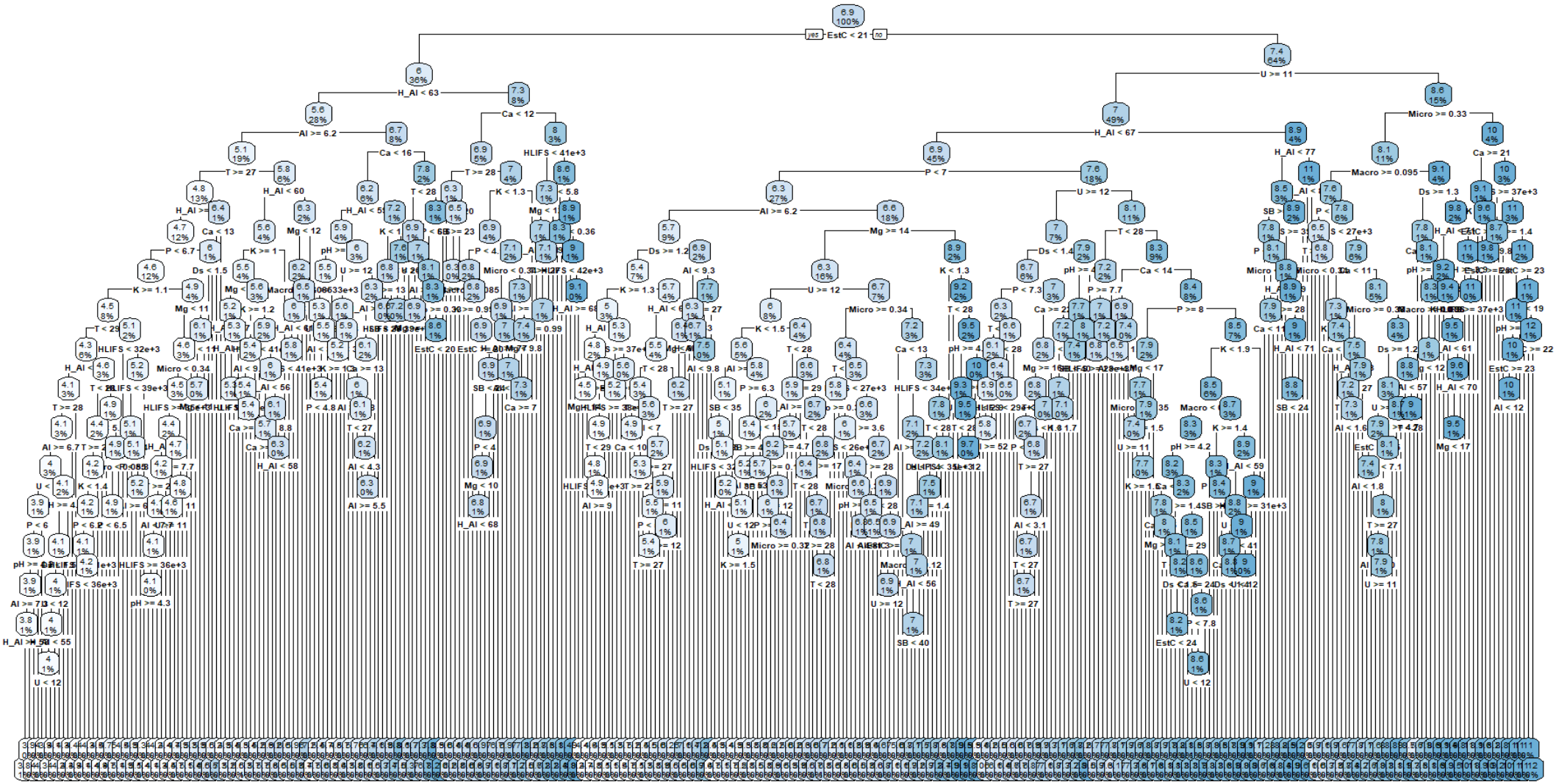

Figura 6A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 17 de fevereiro de 2017 para sistema Eucalipto. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

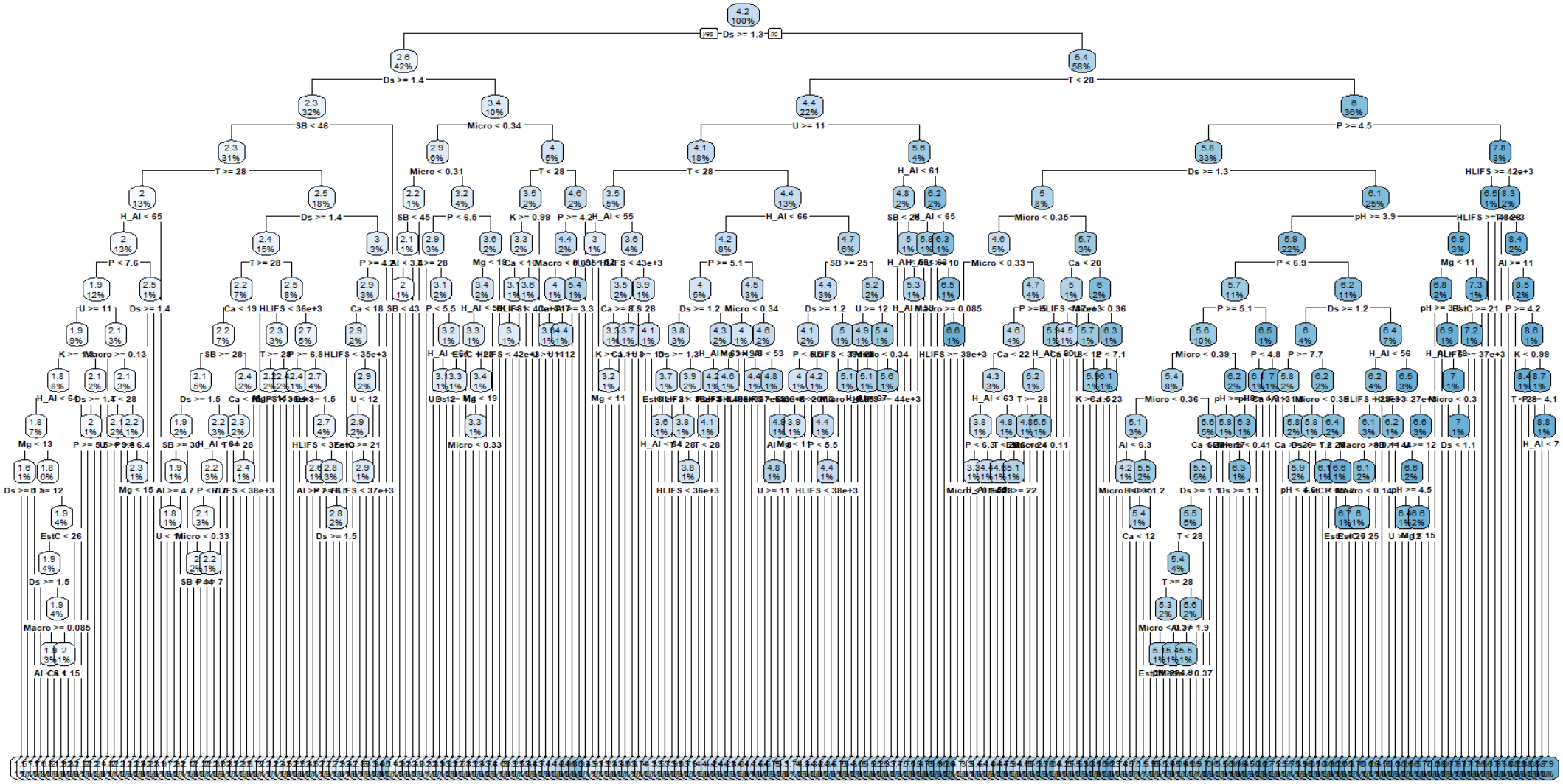

Figura 7A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 15 de março de 2017 para sistema Eucalipto. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

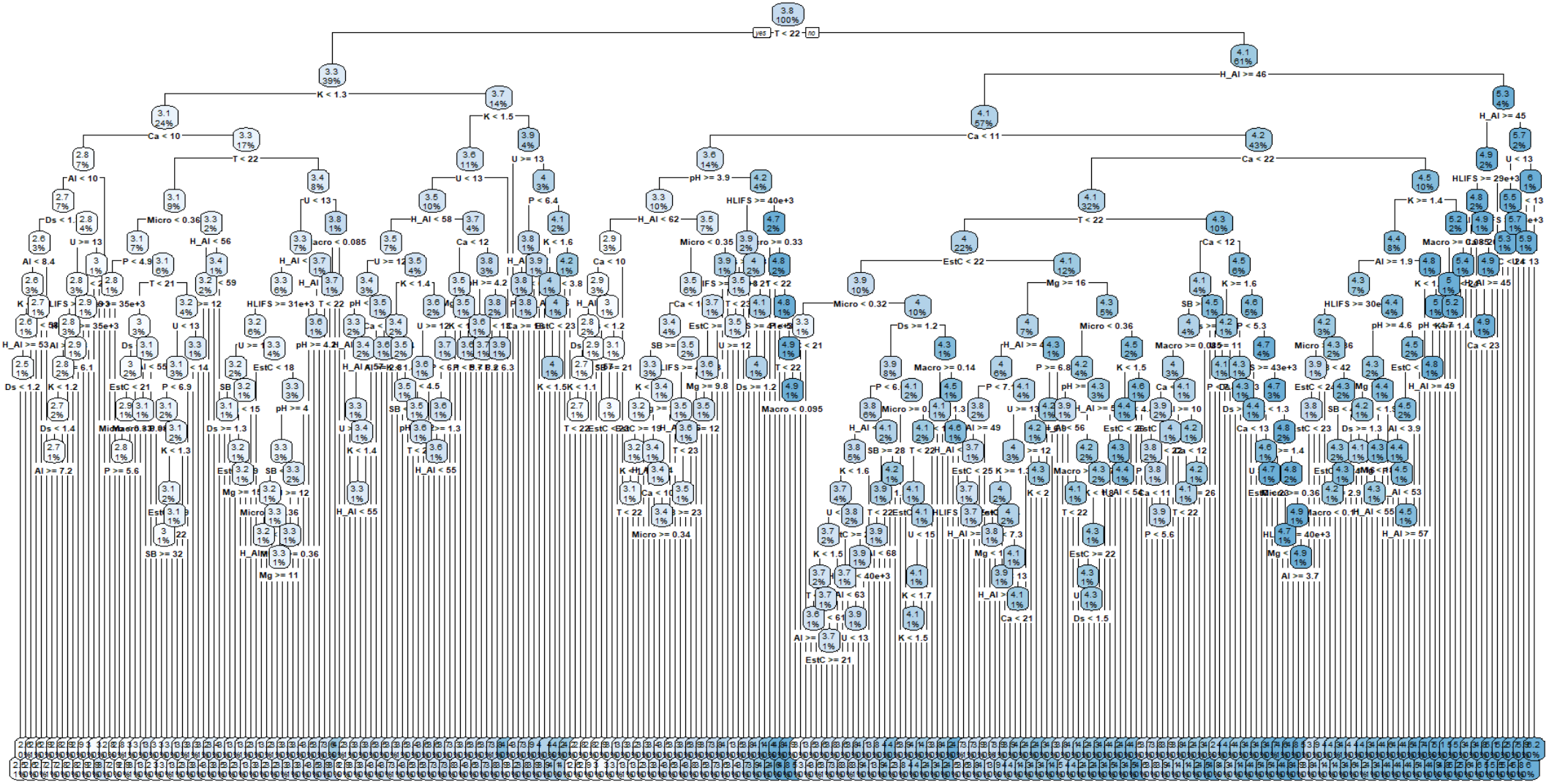

Figura 8A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 03 de junho de 2017 para sistema Eucalipto. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

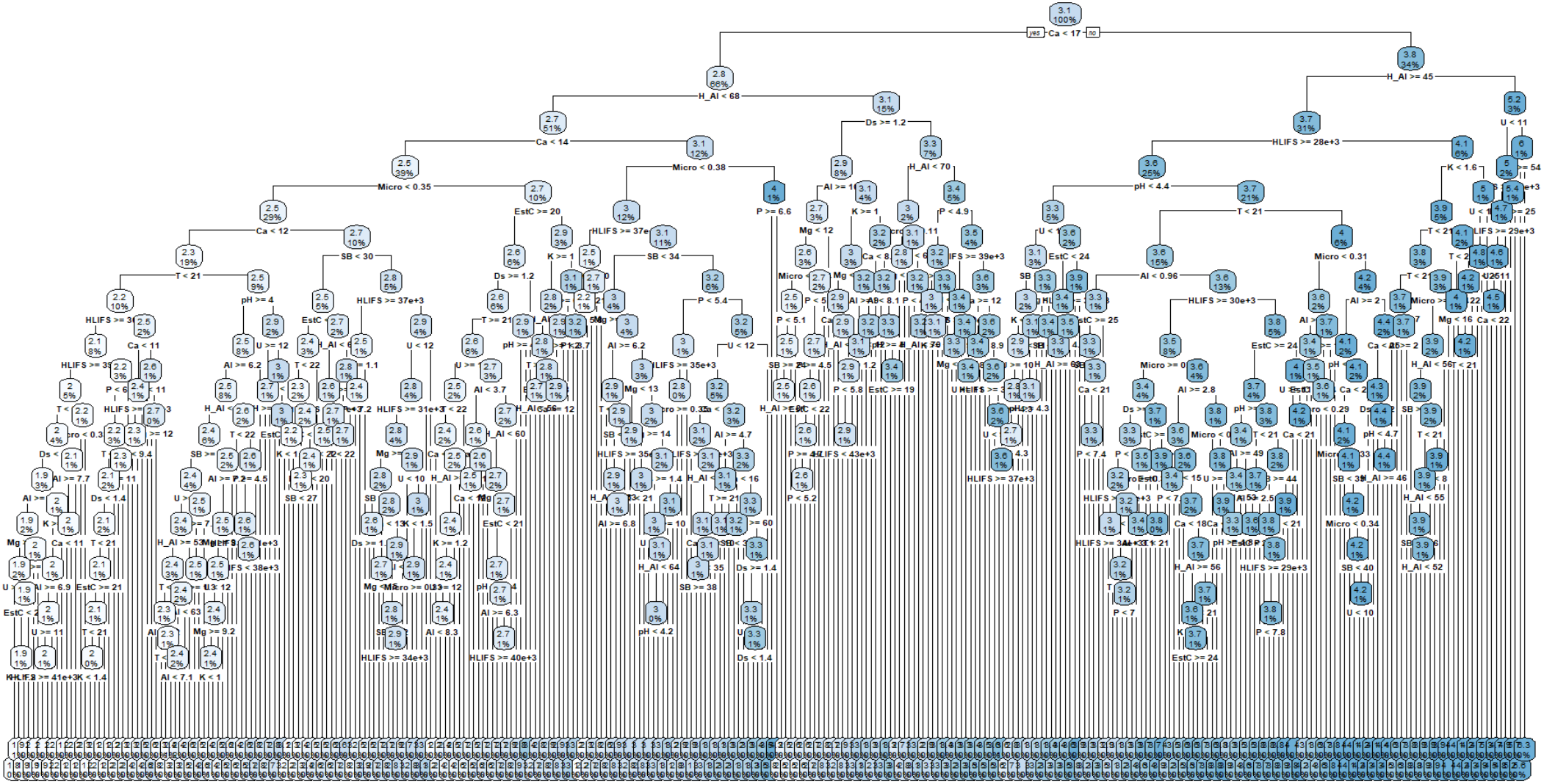

Figura 9A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 10 de junho de 2017 para sistema Eucalipto. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.

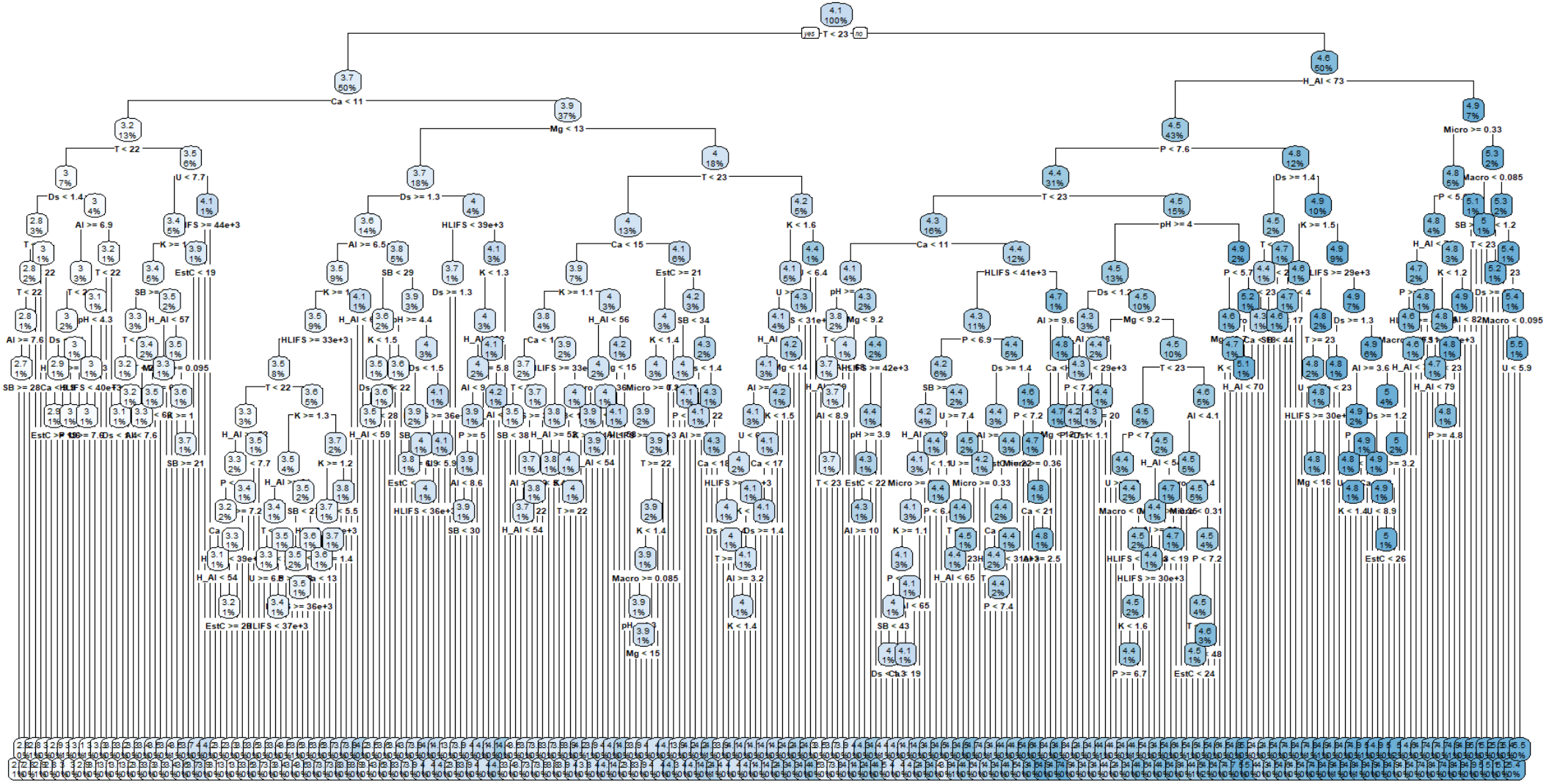

Figura 10A. Representação gráfica da árvore de decisão gerada pelo modelo de aprendizado para o dia 17 de junho de 2017 para sistema Eucalipto. A árvore ilustra as decisões sequenciais tomadas pelo algoritmo para prever as emissões de CO_2 do solo com base nas variáveis selecionadas.