UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

ESTIMATIVA DE INDICADORES DE QUALIDADE DO SOLO POR MEIO DE MODELOS DE “MACHINE LEARNING”

Nayane Jaqueline Costa Maia

Engenheira Agrônoma

Ma. em Agronomia (Ciência do Solo)

2022

UNIVERSIDADE ESTADUAL PAULISTA – UNESP

CÂMPUS DE JABOTICABAL

ESTIMATIVA DE INDICADORES DE QUALIDADE DO SOLO POR MEIO DE MODELOS DE "MACHINE LEARNING"

Discente: Nayane Jaqueline Costa Maia

Orientadora: Prof. Dr. Glauco de Souza Rolim

Coorientadora: Dra. Flávia Fernanda Simili

Tese apresentada à Faculdade de Ciências Agrárias e Veterinárias – Unesp, Câmpus de Jaboticabal, como parte das exigências para a obtenção do título de Doutor em Agronomia (Ciência do Solo)

2022

M217e

Maia, Nayane Jaqueline Costa
Estimativa de indicadores de qualidade do solo por meio de modelos de "Machine Learning" / Nayane Jaqueline Costa Maia. -- Jaboticabal, 2022
99 p. : il., tabs., 2 v. + 1 CD-ROM

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências Agrárias e Veterinárias, Jaboticabal
Orientador: Glauco de Souza Rolim
Coorientadora: Flávia Fernanda Simili

1. Ciência do Solo. 2. Modelagem do Solo. 3. Inteligência Artificial. 4. Integração Lavoura-Pecuária. I. Título.

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

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Jaboticabal

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: ESTIMATIVA DE INDICADORES DE QUALIDADE DO SOLO ESTIMADOS POR MEIO DE MODELOS DE MACHINE LEARNING

AUTORA: NAYANE JAQUELINE COSTA MAIA
ORIENTADOR: GLAUCO DE SOUZA ROLIM
COORIENTADORA: FLÁVIA FERNANDA SIMILI

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

Prof. Dr. GLAUCO DE SOUZA ROLIM (Participaçao Presencial)
Departamento de Engenharia e Ciencias Exatas DECEx / FCAV UNESP Jaboticabal

Pesquisadora Dra. MARA REGINA MOITINHO (Participaçao Presencial)
CNPEM / Campinas/SP

Prof. Dr. ROGÉRIO TEIXEIRA DE FARIA (Participaçao Presencial)
Departamento de Engenharia e Ciencias Exatas DECEx / FCAV UNESP Jaboticabal

Taynara Tuany Borges Valeriano
Dra. TAYNARA TUANY BORGES VALERIANO (Participaçao Presencial)
Instiuição Bayer S.A. / Agronomic Predective Modeler at Bayer CropScience

Diego Silva Siqueira
Dr. DIEGO SILVA SIQUEIRA (Participaçao Presencial)
Agrônomo Autônomo - Núcleo de Inovação Tecnológica (NIT) - Supera Parque / Ribeirão Preto/SP

Jaboticabal, 09 de novembro de 2022

DADOS CURRICULARES DA AUTORA

NAYANE JAQUELINE COSTA MAIA – nascida em 23 de novembro de 1993, na cidade de Castanhal – PA. Concluiu o ensino médio integrado ao curso Técnico em Agropecuária (2008-2011), pelo Instituto Federal de Educação, Ciência e Tecnologia (IFPA), Câmpus de Castanhal. É Engenheira Agrônoma formada pelo IFPA-Câmpus Castanhal (2012-2017). Durante a graduação foi integrante do Núcleo de Pesquisa e Estudos Agropecuários (NUPAGRO) e do Núcleo de Pesquisa em Ciência do Solo e Água da Amazônia (NUPECSA). Em 2012 foi bolsista de iniciação científica do Departamento de Zootecnia do IFPA-Câmpus Castanhal, com pesquisas voltadas para ganho de peso animal usando resíduos de frutas da Amazônia. Em 2013 foi bolsista da FAPESPA/CNPq, no projeto em que foram avaliados indicadores da qualidade do solo em áreas sob vegetação natural e cultivos na Amazônia. Em 2014 foi monitora do Laboratório de Solos e Plantas, do Departamento de Solos e Adubos do IFPA-Câmpus Castanhal. Em 2015 foi bolsista de iniciação cientifica do Departamento de Ciência do Solo, com pesquisa sobre indicadores de fertilidade do solo na cultura do açaí (*Euterpe oleraceae*) em áreas de várzea e de terra firme. Nos anos de 2017 e 2018, foi bolsista de mestrado (CAPES) no Programa de Pós-Graduação em Agronomia (Ciência do Solo) da FCAV-Unesp, no departamento de Solos e Adubos e durante esse período participou do projeto da FAPESP “Impacto ambiental, produtividade e viabilidade econômica de sistemas convencional ou integrado de lavoura pecuária” no Instituto de Zootecnia de Sertãozinho-SP. Nos anos de 2019 a 2021 foi bolsista de doutorado (CAPES) no Programa de Pós-Graduação em Agronomia (Ciência do Solo) da FCAV-Unesp, no departamento de Engenharia e Ciências Exatas. No ano de 2021 passou no concurso público para professora substituta de Agronomia no IFMS-Campus Naviraí. Desde 2021 vem trabalhando na área de ciência de dados e modelagem preditiva no setor privado, com otimização da produtividade no campo sucroenergético.

“Nós somos, de fato, aquilo que escolhemos e as consequências que assumimos.”

Mario Sergio Cortella

Aos meus pais Nelma Maia e Jesus Maia pela coragem de mudar de cidade, que mudaria tudo para nós a partir de então. Ao meu querido amigo Washington Pereira – *In memoriam*, por me ensinar que a vida é uma só, e que todo dia ela precisa ser vivida intensamente (*"Abre a porta do teu ser, sinta o vento te soprar").*

DEDICO

AGRADECIMENTOS

À Deus, por proporcionar esta oportunidade e me entregar tanta coragem para chegar até aqui sozinha.

À minha família, em especial aos meus pais, Jesus Maia e Nelma Maia, e aos meus irmãos, especialmente a Nayana Maia, que não mediram esforços para me ajudar; devo a eles tudo o que tenho e o que sou.

Ao professor Dr. Glauco de Souza Rolim, por me apresentar um novo mundo (ciência de dados), por me apoiar na execução desse trabalho e na minha carreira fora da academia.

À Pesquisadora Dra. Flávia Fernanda Simili, por me apoiar com a continuação do projeto e o uso dos dados do seu projeto, e na grande parceria de redação científica que construímos nos últimos anos.

À Professora Dra. Mara Cristina Pessôa da Cruz, pelo carinho e todo aprendizado repassado durante o mestrado no laboratório de fertilidade do solo da FCAV/Unesp.

Ao Conselho do Programa de Pós-Graduação em Agronomia (Ciência do Solo). O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de financiamento 001.

Ao Instituto de Zootecnia de Sertãozinho-SP, pela instalação do projeto. À Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), pelos recursos e financiamento do projeto (Processo Nº 2014/24514-6).

Ao Laboratório de Fertilidade do Solo da FCAV/Unesp – Campus de Jaboticabal onde foram realizadas as análises. E ao Group of Agrometeorological Studies (GAS). A todos os amigos que fiz nesses grupos, por me auxiliarem e compartilharem o aprendizado comigo.

Aos amigos que fiz na cidade de Jaboticabal-SP e São José do Rio Preto-SP, pelo carinho, paciência e toda ajuda.

A todos que contribuíram para a realização desse trabalho.

Muito obrigada!

Sumário

RESUMO ..... **ii**

ABSTRACT ..... **iii**

LISTA DE ABREVIAÇÕES **1**

CAPÍTULO 1 – Considerações gerais **3**

CAPÍTULO 2 - Estimação de carbono e nitrogênio orgânico do solo a partir de variáveis do solo, clima, plantas e animais utilizando modelos de "machine learning" **8**

RESUMO 8

2.1 INTRODUÇÃO 9

2.2 MATERIAL E MÉTODOS 11

2.2.1 Local do experimento 11

2.2.2 Dados climáticos 12

2.2.3 Delineamento experimental e tratamentos 13

2.2.4 Coleta de dados de solo, plantas e animais 14

2.2.5 Processamento de dados 15

2.2.6 Modelos de Machine learning 16

2.2.7 Avaliação do desempenho dos modelos 18

2.3. RESULTADOS E DISCUSSÃO 20

2.3.1 Importância e seleção das variáveis de entrada no modelo 20

2.3.2 Desempenho da estimação de carbono e nitrogênio orgânico do solo 28

2.3.3 Comparação entre os modelos de machine learning para estimar o teor de carbono e nitrogênio orgânico do solo 44

2.4. CONCLUSÕES 46

2.5 REFERÊNCIAS 46

CAPÍTULO 3 - Disponibilidade de nitrogênio potencialmente mineralizável no solo: uma abordagem interpretável de "machine learning" **55**

RESUMO 55

3.1 INTRODUÇÃO 56

3.2 MATERIAL E MÉTODOS 58

3.2.1 Local do experimento e Delineamento experimental 58

3.2.2 Coleta de dados de solo, plantas e animais 59

3.2.3 Processamento de dados 60

3.2.4 Avaliação do desempenho dos modelos 63
3.2.5 Interpretação dos modelos usando o SHAP 63
3.3 RESULTADOS E DISCUSSÃO 64
3.3.1 Importância e Seleção das variáveis estimadoras de nitrogênio potencialmente mineralizável do solo 64
3.3.2 Desempenho dos modelos na estimação de nitrogênio potencialmente mineralizável do solo 67
3.3.3 Interpretação dos modelos usando SHAP 70
3.4 CONCLUSÕES 76
3.5 REFERÊNCIAS 77
APÊNDICES 84

ESTIMATIVA DE INDICADORES DE QUALIDADE DO SOLO POR MEIO DE MODELOS DE "MACHINE LEARNING"

RESUMO – O monitoramento de indicadores de qualidade do solo contribui para mitigar as mudanças climáticas globais sem diminuir a produtividade dos sistemas, além de avaliar a eficiência de sistemas sustentáveis e complexos. O uso de modelos de "machine learning" no atual cenário do aumento dos dados é necessário para que as informações dos sistemas agrícolas sejam mais colaborativas, de maneira que, diminua custos e se traduza em otimização dos sistemas, tornando-os mais sustentáveis. Com este trabalho objetivou-se testar modelos de machine learning para estimar indicadores de qualidade do solo, a fim de melhorar o monitoramento dos solos agrícolas em sistema de monocultivo de milho, de pecuária e de Integração Lavoura-Pecuária, avaliando: 1) as combinações de variáveis para estimar o carbono e nitrogênio orgânico solo; e 2) variáveis que impactam a disponibilidade do nitrogênio potencialmente mineralizável no solo. Os modelos de machine learning testados foram Multilayer Perceptron Regressor (MLP), Random Forest Regressor (RF), K Neighbors Regressor (KNN), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Adaptive Boosting Regressor (AdaBoost) e eXtreme Gradient Boosting Regressor (XGBoost), usando a linguagem de programação Python. Os modelos utilizados se mostraram eficientes para estimar carbono e nitrogênio do solo. As variáveis combinadas de plantas-animais-solo-clima estimaram com boa acurácia e precisão o carbono e nitrogênio orgânico do solo. No sistema de monocultivo de Milho, os modelos AdaBoost e MLP obtiveram alta precisão e MAPE<1%. No sistema monocultivo de Pecuária, os modelos MLP, SVR e Adaboost foram mais precisos e acurados (MAPE<3%). No sistema de integração Lavoura-Pecuária, todos os modelos tiveram elevada acurácia (MAPE<5%), no entanto, o modelo SVR obteve a menor precisão para estimar carbono orgânico do solo (R^2 < 2%), e o SVR e KNN obtiveram a menor precisão para estimar o nitrogênio orgânico do solo (R^2 < 50%). Para estimar o nitrogênio potencialmente mineralizável no solo, o modelo XGBoost foi o mais preciso, com o menor acurácia e viés (R^2 = 0.97, MAPE = 3% e MBE = 0.10 mg kg^{-1}), superando o AdaBoost, RF e MLR, nessa ordem. As diferentes combinações de variáveis estimadoras podem indicar processos importantes que influenciam na liberação de nitrogênio potencialmente mineralizável no solo, por meio de métodos de explicações aditivas de Shapley (SHAP). No geral, esses resultados fornecem uma nova perspectiva com o uso de aplicação de machine learning para estimar importantes nutrientes do solo, aproveitando os mais diversos históricos de dados de sistemas agrícolas, que podem ser úteis para os tomadores de decisão na produção de alimentos.

"Palavras-chave:" combinação de variáveis, carbono orgânico, nitrogênio orgânico, nitrogênio potencialmente mineralizável, inteligência artificial, agricultura 4.0.

ESTIMATION OF SOIL QUALITY INDICATORS WITH MACHINE LEARNING MODELS

ABSTRACT - Monitoring soil quality indicators contribute to mitigating global climate change without reducing the influence of systems, in addition to measuring the efficiency of systems and complexes. The use of "machine learning" models in the current scenario of the data age is necessary for information from agricultural systems to be more collaborative, in a way that reduces costs and translates into the optimization of systems, making them more intelligent. The objective of this work was to test machine learning models to estimate soil quality indicators, to improve the monitoring of agricultural soils in a corn and pasture monoculture system and Crop-Livestock Integration system, evaluating: 1) a combination of variables to estimate carbon and organic only; and 2) variables that impact the availability of mineralizable power in the soil. The machine learning models tested were Multilayer Perceptron Regressor (MLP), Random Forest Regressor (RF), K Neighbors Regressor (KNN), Support Vector Regression (SVR), Multiple Linear Regression (MLR), Adaptive Boosting Regressor (AdaBoost) and eXtreme Gradient Boosting Regressor (XGBoost), using the Python programming language. The models used are efficient to estimate soil carbon and tolerance. The combined plant-animal-soil-climate variables accurately and accurately estimated soil carbon and organics. In the corn monoculture system, the AdaBoost and MLP models achieved high accuracy and MAPE<1%. In the livestock monoculture system, the MLP, SVR, and Adaboost models were more precise and accurate (MAPE<3%). In the Crop-Livestock integration system, all models had high accuracy (MAPE<5%), however, the SVR model obtained a lower accuracy for estimating soil organic carbon (R^2 < 2%), and the SVR and KNN obtained the lowest precision to estimate the soil organic (R^2 < 50%). To estimate the electrically mineralizable in the soil, the XGBoost model was the most accurate, with the lowest accuracy and bias (R^2 = 0.97, MAPE = 3% and MBE = 0.10 mg kg^{-1}), surpassing AdaBoost, RF, and MLR, in that order. The different combinations of estimated variables may indicate important processes that influenced the release of potentially mineralizable material in the soil, using Shapley's induced additive methods (SHAP). Overall, these results provide a new perspective with the use of machine learning applications to estimate important soil nutrients, taking advantage of the most diverse historical data of agricultural systems, which can be useful for decision-makers in food production.

"Keywords:" combination of variables, organic carbon, organic nitrogen, potentially mineralizable nitrogen, artificial inteligence, agriculture 4.0.

LISTA DE ABREVIAÇÕES

AdaBoost: adaptive boosting regressor

B: saturação de bases

BD: densidade do solo

C ou **TOC:** carbono orgânico total

CS: monocultivo de Milho

CMI: índice de manejo do carbono

CEC: capacidade de troca de cátions

C/N-soil: relação carbono e nitrogênio do solo

C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo

d: Índice de Willmott

ICLS: Integração Lavoura-Pecuária

KNN: k neighbors regressor

LS: monocultivos de Pecuária

LC: carbono lábil

LD: liteira depositada

LIC: carbono da liteira

LIN: nitrogênio da liteira

LIC/LIN: relação carbono e nitrogênio da liteira

MA: macroporosidade

MAE: erro absoluto médio

MAPE: erro percentual absoluto médio

MBC: carbono da biomassa microbiana

MBE: víeis de erro médio

MBN: nitrogênio da biomassa microbiana

MI: microporosidade

ML: machine learning

MLP: multilayer perceptron regressor

MLR: multiple linear regression

N ou **TON**: nitrogênio orgânico total

NU: nitrogênio da urina

NF: nitrogênio das fezes

PMN: nitrogênio potencialmente mineralizável

pH: potencial de hidrogênio

Pr: fósforo

R^2: coeficiente de determinação ajustado

RF: random forest regressor

RMSE: raiz quadrada do erro médio

SHAP: shapley additive explanations

SCS: estoque de carbono

SE: erro sistemático

SNS: estoque de nitrogênio

SVR: support vector regression

SWC: potencial de água no solo

TP: porosidade total

VIF: fator de inflação da variância

W: peso animal

XGBoost: extreme gradient boosting regressor

CAPÍTULO 1 – Considerações gerais

A sustentabilidade dos sistemas agrícolas e pecuários se baseia em três importantes pilares: ser economicamente viável, socialmente justo e ambientalmente correto (Otálora et al., 2021). A qualidade do solo é um dos fatores que podem indicar se o ambiente está em equilíbrio e se o manejo do solo vem sendo executado corretamente para garantir o sucesso dos sistemas produtivos. A degradação e a perda de biodiversidade do solo são dois dos problemas globais mais presentes que afetam os ecossistemas terrestres (Muñoz-Rojas, 2018; Sims et al., 2019).

Reduzir a degradação e recuperar os solos degradados são, portanto, ações urgentes e necessárias para manter a função e a produtividade do ecossistema, mitigar as mudanças climáticas, preservar a biodiversidade e garantir a produção de alimentos e o fornecimento de recursos (Bouma e Montanarella, 2016; Kaeesstra et al., 2016). A ONU estabeleceu metas específicas para a restauração de ecossistemas, incluindo a “Agenda 2030 para o Desenvolvimento Sustentável” (adotada pela Assembleia Geral das Nações Unidas - ONU) que define 17 Objetivos de Desenvolvimento Sustentável (ODS). Muitos desses objetivos estão fortemente ligados ao manejo do solo no estabelecimento de estratégias para atingir as metas de recuperação nas próximas décadas (Kaeesstra et al., 2016; Muñoz-Rojas, 2018), reforçando a importância de alcançar um mundo neutro em termos de degradação da terra e recarbonização dos solos agrícolas (Stavi e Lal, 2015; Lal et al., 2021).

A maioria das funções do ecossistema do solo são difíceis de avaliar diretamente e, portanto, são frequentemente inferidas a partir de propriedades mensuráveis, como indicadores de qualidade do solo, que podem abranger uma ampla gama de características físicas, químicas e biológicas do solo (Muñoz-Rojas, 2018). A avaliação da qualidade do solo não pode ser definida medindo características únicas do solo (Maurya et al., 2020), e seria impossível usar todas as suas características para avaliar a qualidade; assim, é necessário um conjunto mínimo de dados que consiste em um conjunto de características, incluindo propriedades físicas, químicas e biológicas do solo que ajudam a monitorar a fertilidade, saúde e qualidade (Yu et al. 2018; Maurya et al., 2020).

O método tradicional de medição dos indicadores de qualidade do solo, é baseado em medição pontual e laboratorial, possui alta precisão, mas grande carga de trabalho, alto custo e alta demanda de reagentes químicos, sendo difícil atender às

reais necessidades de monitoramento de solos agrícolas (Barra et al., 2021; Jiang et al., 2022). As imagens oriundas de satélite têm sido amplamente utilizadas no monitoramento do solo e dos sistemas de cultivo, como por exemplo, a matéria orgânica (Wei et al., 2020; Luo et al., 2022), evapotranspiração da cultura (Niyogi et al., 2020), estresse hídrico (Jamshidi et al., 2020) e monitoramento de produtividade (Skakun et al., 2021).

Existe uma crescente demanda de dados oriundos dos sistemas ou áreas agrícolas (Basso e Antle, 2020), uma vez que essas fontes de dados podem ser transformadas em informações valiosas sobre os diferentes usos dos solos agrícolas, através de modelos de “machine learning” - ML (aprendizado de máquina). O termo aprendizagem de máquina pode ser definido como o processo de descobrir as relações entre variáveis de previsão/estimação e de resposta usando abordagens estatísticas baseadas em algoritmos orientados a dados não-lineares (Heung et al., 2016; Khaledian e Miller, 2020; Wadoux et al., 2020). Os algoritmos de ML também podem lidar com um grande número de covariáveis cruzadas como estimadoras (Wadoux et al., 2020).

A crescente disponibilidade de dados de solo que podem ser adquiridos de forma remota e eficiente, e algoritmos de código aberto disponíveis gratuitamente, levaram a uma adoção acelerada de modelos de ML para analisar dados de solo (Padarian et al., 2020). Várias aplicações de ML bem conhecidas na ciência dos solos incluem a previsão de tipos e propriedades do solo por meio de mapeamento digital do solo e análise de dados espectrais infravermelhos para inferir as propriedades do solo (Barra et al., 2020; Padarian et al., 2020; Wadoux et al., 2020; Goydaragh et al., 2021). Modelos de machine learning também é usado para estudos sobre o controle da distribuição de indicadores de qualidade do solo (Fernandes et al., 2019; John et al., 2020; Diaz-Gonzalez et al., 2022).

Diante desse contexto, os modelos de machine learning se mostram promissoras para estimar indicadores de qualidade do solo. No entanto, a seleção do conjunto de dados deve ser estabelecida de acordo com o tipo de solo, condições ambientais, práticas de manejo e tipo de cultura de cada área (Diaz-Gonzalez et al., 2022). Além disso, as características meteorológicas da área de cultivo devem ser consideradas no processo de estimação dos indicadores do solo para melhor representar as áreas agrícolas (Bünemann et al., 2018; John et al., 2020).

Dessa forma, os modelos de machine learning são ferramentas acuradas, de fácil implementação, e de alta aplicabilidade nos sistemas agrícolas para estimar indicadores de qualidade de solo sob diferentes usos agrícolas. Com este estudo objetiva-se testar modelos de machine learning para estimar indicadores de qualidade do solo, a fim de melhorar o monitoramento de diferentes usos de solos agrícolas. Nos próximos capítulos serão apresentados 1) combinações de diferentes variáveis de solo, planta, clima e animal para estimar o carbono e nitrogênio orgânico do solo; e 2) variáveis que impactam na disponibilidade do nitrogênio potencialmente mineralizável no solo.

REFERÊNCIAS

Barra I, Haefele SM, Sakrabani R, Kebede F (2021) Soil spectroscopy with the use of chemometrics, machine learning and pre-processing techniques in soil diagnosis: Recent advances–A review. **TrAC Trends in Analytical Chemistry** 135: 116166. https://doi.org/10.1016/j.trac.2020.116166

Basso B, Antle J (2020) Digital agriculture to design sustainable agricultural systems. **Nature Sustainability** 3(4): 254-256. https://doi.org/10.1038/s41893-020-0510-0

Bouma J, Montanarella L (2016) Facing policy challenges with inter-and transdisciplinary soil research focused on the UN Sustainable Development Goals. **Soil** 2(2): 135-145. https://doi.org/10.5194/soil-2-135-2016

Bünemann EK et al. (2018) Soil quality–A critical review. **Soil Biology and Biochemistry** 120, 105-125. https://doi.org/10.1016/j.soilbio.2018.01.030

Diaz-Gonzalez FA, Vuelvas J, Correa CA, Vallejo VE, Patino D (2022) Machine learning and remote sensing techniques applied to estimate soil indicators–Review. **Ecological Indicators** 135: 108517. https://doi.org/10.1016/j.ecolind.2021.108517

Fernandes MMH, Coelho AP, Fernandes C, Silva MF, Marta CCD (2019) Estimation of soil organic matter content by modeling with artificial neural networks. **Geoderma** 350: 46-51. https://doi.org/10.1016/j.geoderma.2019.04.044

Goydaragh MG, Taghizadeh-Mehrjardi R, Jafarzadeh AA, Triantafilis J, Lado M (2021) Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. **Catena** 202: 105280. https://doi.org/10.1016/j.catena.2021.105280

Heung B, Ho HC, Zhang J, Knudby A, Bulmer CE, Schmidt MG (2016) An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. **Geoderma** 265: 62-77. https://doi.org/10.1016/j.geoderma.2015.11.014

Jamshidi S, Zand-Parsa S, Niyogi D (2021) Assessing crop water stress index of citrus using in-situ measurements, landsat, and sentinel-2 data. International **Journal of Remote Sensing** 42(5): 1893-1916. https://doi.org/10.1080/01431161.2020.1846224

Jiang X, Luo S, Ye Q, Li X, Jiao W (2022) Hyperspectral Estimates of Soil Moisture Content Incorporating Harmonic Indicators and Machine Learning. **Agriculture** 12(8): 1188. https://doi.org/10.3390/agriculture12081188

John K, Abraham Isong I, Michael Kebonye N, Okon Ayito E, Chapman Agyeman P, Marcus Afu S (2020) Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. **Land** 9(12): 487. https://doi.org/10.3390/land9120487

Khaledian Y, Miller BA (2020) Selecting appropriate machine learning methods for digital soil mapping. **Applied Mathematical Modelling** 81: 401-418. https://doi.org/10.1016/j.apm.2019.12.016

Keesstra SD, Bouma J, Wallinga J, Tittonell P, Smith P, Cerdà A, Montanarella L, Quinton JN, Pachepsky Y, Van der Putten WHV, Bardgett RD, Moolenaar S, Mol G, Jansen B, Fresco LO (2016) The significance of soils and soil science towards realization of the United Nations Sustainable Development Goals. **Soil** 2(2): 111-128. https://doi.org/10.5194/soil-2-111-2016

Lal R, Monger C, Nave L, Smith P (2021) The role of soil in regulation of climate. **Philosophical Transactions of the Royal Society** 376(1834): 20210084. https://doi.org/10.1098/rstb.2021.0084

Luo C, Zhang X, Meng X, Zhu H, Ni C, Chen M, Liu H (2022) Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. **Catena** 209: 105842. https://doi.org/10.1016/j.catena.2021.105842

Maurya S, Abraham JS, Somasundaram S, Toteja R, Gupta R, Makhija S (2020) Indicators for assessment of soil quality: a mini-review. **Environmental Monitoring and Assessment** 192(9): 1-22. https://doi.org/10.1007/s10661-020-08556-z

Muñoz-Rojas M (2018) Soil quality indicators: critical tools in ecosystem restoration. **Current Opinion in Environmental Science & Health** 5: 47-52. https://doi.org/10.1016/j.coesh.2018.04.007

Niyogi D, Jamshidi S, Smith D, Kellner O (2020) Evapotranspiration climatology of indiana using in situ and remotely sensed products. **Journal of Applied Meteorology and Climatology** 59(12): 2093-2111. https://doi.org/10.1175/JAMC-D-20-0024.1

Otálora XD, Del Prado A, Dragoni F, Estellés F, Amon B (2021) Evaluating Three-Pillar Sustainability Modelling Approaches for Dairy Cattle Production Systems. **Sustainability** 13(11): 6332. https://doi.org/10.3390/su13116332

Padarian J, Minasny B, McBratney AB (2020) Machine learning and soil sciences: A review aided by machine learning tools. **Soil** 6(1): 35-52. https://doi.org/10.5194/soil-6-35-2020

Sims NC, England JR, Newnham GJ, Alexander S, Green C, Minelli S, Held A (2019) Developing good practice guidance for estimating land degradation in the context of the United Nations Sustainable Development Goals. **Environmental Science & Policy** 92: 349-355. https://doi.org/10.1016/j.envsci.2018.10.014

Skakun S, Kalecinski NI, Brown MG, Johnson DM, Vermote EF, Roger JC, Franch B (2021) Assessing within-field corn and soybean yield variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 satellite imagery. **Remote Sensing 13**(5): 872. https://doi.org/10.3390/rs13050872

Stavi I, Lal R (2015) Achieving zero net land degradation: challenges and opportunities. **Journal of Arid Environments** 112: 44-51. https://doi.org/10.1016/j.coesh.2018.04.007

Yu P, Liu S, Zhang L, Li Q, Zhou D (2018) Selecting the minimum data set and quantitative soil quality indexing of alkaline soils under different land uses in northeastern China. **Science of The Total Environment.** https://doi.org/10.1016/j.scitotenv.2017.10.301.

Wadoux AMC, Minasny B, McBratney AB (2020) Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews** 210: 103359. https://doi.org/10.1016/j.earscirev.2020.103359

Wei L, Yuan Z, Wang Z, Zhao L, Zhang Y, Lu X, Cao L (2020) Hyperspectral Inversion of Soil Organic Matter Content Based on a Combined Spectral Index Model. **Sensors** 20(10): 2777. https://doi.org/10.3390/s20102777

CAPÍTULO 2 - Estimação de carbono e nitrogênio orgânico do solo a partir de variáveis do solo, clima, plantas e animais utilizando modelos de “machine learning”

RESUMO – O monitoramento de carbono orgânico (TOC) e nitrogênio orgânico (TON) do solo precisam ser ampliados e implementados para mitigar as mudanças climáticas globais. Neste estudo, objetivou-se correlacionar o teor de TOC e TON com variáveis de solos, plantas e animais considerando as condições climáticas e balanço hídrico de sistema de Integração Lavoura-Pecuária (ICLS) e Monocultivos de milho (CS) e sistema de pecuária (LS); para explorar o potencial de seis modelos de regressão em estimar o teor de TOC e TON. Os modelos testados foram Multilayer Perceptron Regressor (MLP), Random Forest Regressor (RF), K Neighbors Regresso;r (KNN), Support Vector Regression (SVR), Multiple Linear Regression (MLR), e Adaptive Boosting Regressor (AdaBoost). As variáveis do solo, plantas e animais foram medidas no local ou analisadas em laboratório durante um período de dois anos, em uma região tropical úmida do estado de São Paulo (Aw, Koppen), no Brasil. No total foram 189 variáveis de entrada: vinte de solos, quatro de plantas, três de animais e nove climáticas, em 18 decêndios. Os modelos de redes neurais MLP e AdaBoost tiveram os melhores desempenhos, comprovado pelo diagrama de Taylor, com o maior coeficiente R^2>0.94 e os menores erros quadráticos médio (RMSE) entre os sistemas avaliados. No sistema CS, os modelos AdaBoost e MLP obtiveram alta precisão e MAPE<1%. No sistema LS, os modelos MLP, SVR e Adaboost foram mais precisos e acurados (MAPE<3%). No sistema ICLS, todos os modelos tiveram alta acurácia (MAPE<5%), no entanto, o modelo SVR obteve a menor precisão para estimar TOC (R^2 < 2%), e o SVR e KNN obtiveram a menor precisão para estimar o TON (R^2<50%). A combinação de variáveis que mais contribuíram na estimativa de TOC no sistema CS, foram: a saturação de bases, umidade relativa, déficit de água no solo, evapotranspiração potencial, precipitação, excedente de água no solo, armazenamento de água no solo. Para a estimativa de TON no sistema CS, foram o nitrogênio potencialmente mineralizável, radiação líquida, excedente de água no solo, evapotranspiração potencial, precipitação, velocidade do vento. Para estimar TOC no sistema LS, foram as variáveis radiação líquida, excedente de água no solo, porosidade total, carbono da liteira, índice de manejo do carbono, e nitrogênio potencialmente mineralizável. Para a estimativa do TON no sistema LS, foram o carbono da liteira, índice de manejo do carbono, excedente de água no solo, saturação por bases, potencial de água no solo, liteira depositada, relação C/N do solo. A combinação de variáveis que mais contribuíram na estimativa de TOC no sistema ICLS, foram a precipitação, excedente de água no solo, macroporosidade, nitrogênio da liteira, relação C/N da biomassa microbiana do solo, nitrogênio das fezes e das urinas. Para estimar TON, foram: carbono orgânico do solo, relação C/N do solo, fósforo, saturação por bases, excedente de água no solo, carbono da liteira, carbono da biomassa microbiana do solo. Com base nos resultados, verificamos a importância do sistema ICLS, que possuem correlações mais densas entre as variáveis plantas-animais-solo-clima que impactam positivamente no ambiente. Além disso, a combinação dessas variáveis foi usada pela primeira vez conjuntamente para estimar indicadores de solos agrícolas como TOC e TON, indicadores que impactam

diretamente na mitigação das mudanças climáticas, e necessitam de análises dispendiosas e demoradas. Esses resultados fornecem uma nova perspectiva para a aplicação de modelagem por aprendizado de máquinas para estimar importantes nutrientes do solo.

“Palavras-chave:” Modelagem do solo, Indicadores de qualidade do solo, Artificial Neural Networks, AdaBoost, Python, inteligência artificial.

2.1 INTRODUÇÃO

A recarbonização dos solos é uma solução global e vem demandando grandes pesquisas sendo de interesse mundial, no que tange o âmbito da agricultura sustentável, qualidade dos solos agrícolas, agricultura regenerativa, tecnologia aliada a agricultura sustentável, e tudo isso junto para diminuir o impacto da agricultura no ambiente, mitigar as mudanças climáticas globais e otimizam a produtividade dos sistemas. A Organização das Nações Unidas para Agricultura e Alimentação (FAO) lançou um programa de recarbonização de solos (RECSOIL), no qual sua mensagem principal identifica o aumento de carbono orgânico no solo (TOC) como uma das opções mais econômicas para a mitigação das mudanças climáticas, bem como para combater a desertificação, a degradação da terra e a insegurança alimentar (Campbell et al.,2018; FAO, 2019; Amelung et al., 2020).

A melhoria do manejo do solo pode resultar em grandes impactos a longo prazo, todos positivamente ligados aos objetivos de desenvolvimento sustentável (ODS). Entre esses manejos do solo, a reutilização de resíduos orgânicos e sua transformação em aditivos orgânicos é uma das estratégias de ODS, resultados positivos dessas estratégias ocorrem com eficiência em sistemas combinados de pecuária e produção de grãos (Oliveira et al., 2022). De acordo com McGuire et al. (2022), o aumento dos níveis de TOC no solo tem o potencial global de sequestro equivalentes a 5 a 10% das emissões globais de gases de efeito estufa. A fim de aproveitar totalmente o potencial do carbono dos solos, que dependendo do sistema de manejo o solo, pode armazenar (sumidouro) ou emitir (fonte) carbono na forma de CO_2 ($C\text{-}CO_2$).

Para avaliar e monitorar a qualidade dos sistemas agrícolas e pecuários, são necessárias análises laboratoriais dispendiosas e demoradas, e que geram grandes

quantidades de resíduos tóxicos no meio ambiente. O método mais consolidado atualmente para a determinação de TOC do solo é o de Walkley e Black (1934), que envolve análise com resíduos ricos em ácidos puros (H_2SO_4, H_3PO_4) e dicromato de potássio ($K_2Cr_2O_7$). Esses resíduos quando descartados incorretamente, podem causar riscos ao meio ambiente (Ontañon et al., 2018) e são cancerígenos aos seres humanos (Wu et al., 2019). Com o intuito de reduzir esses resíduos, alguns pesquisadores desenvolveram estimativas de TOC a partir do uso de ondas magnéticas (Wang et al., 2016). Segundo Mahmoud et al. (2017), a principal desvantagem desse método é a necessidade de vários ajustes de refletância dos raios gama e a compra de equipamentos caros. Para a estimação de nitrogênio orgânico total do solo (TON), é realizada usando sensores do solo, entretanto, é um processo oneroso devido ao alto custo desses sensores. (Zhang et al., 2019; Lin et al., 2020; Xu et al., 2021).

Para superar as restrições e limitações de custo e ambientais envolvidas nas estimações de TOC e TON, uma alternativa viável para a estimação são os algoritmos de machine learning (ML), que podem ser utilizados como uma ferramenta de estimativa ou como um complemento às abordagens químicas. Esse método possibilita simular e prever o teor de TOC e TON no solo a partir da combinação de variáveis oriundas de diferentes fontes em modelos para estimar as propriedades do solo. Modelos de redes neurais artificiais (Multilayer Perceptron-MLP), Multiple Linear Regression (MLR), e também Random Forest (RF) junto com Support Vector Regression (SVR) são os modelos mais utilizados para estimar o teor de TOC e TON do solo (Mahmoud et al., 2017; Emamgholizadeh et al., 2018; Reda et al., 2019; Lin et al., 2020; Mahmoud et al. 2020). Entretanto, o modelo K-Nearest Neighbor (KNN) também já foi utilizado para prever o teor de nitrogênio do solo (Grell et al., 2021). E o algoritmo Adaptive Boosting (AdaBoost) obteve alta acurácia para classificar solos (Pham et al., 2021) e alta precisão (R^2=0.91) para estimar o teor de matéria orgânica do solo (Wei et al., 2020).

Na Alemanha, Wiesmeier et al. (2014) combinou atributos do solo, precipitação e temperatura do ar anual para estimar estoque de carbono no solo, obtendo uma boa precisão usando o modelo RF. Em outro estudo, foram combinados atributos do solo e variáveis climáticas usando índices de vegetação e temperatura do ar anual, e a precisão da estimativa de TOC foi melhorada usando o RF (John et al., 2020). Os

estudos de Wiesmeier et al. (2014) e John et al. (2020) são as únicas tentativas de combinar variáveis ambientais com indicadores de qualidade do solo até agora, e indicam a importância de incluir preditores auxiliares, para representar a importância do clima e propriedades físicas e químicas do solo nos sistemas.

Apesar da aceitabilidade do ML para estimar TOC e TON, poucos estudos consideram a incorporação conjunta de indicadores do solo, análises de plantas, influência dos animais nos sistemas e características climáticas na sua estimação, especialmente comparando monocultivos e sistemas de integração Lavoura-Pecuária (ICLS). Nos sistemas de ICLS os animais são muito importantes, pois grande parte dos nutrientes ingeridos retorna ao solo via fezes e urina, os quais são liberados em um curto intervalo de tempo em formas prontamente disponíveis no solo (Vilela et al., 2011). Essa demanda continuada entre entradas e saídas de nutrientes no sistema pode aumentar a reserva de TOC e TON no solo quando comparados aos monocultivos (Bieluczyk et al., 2017; Liebig et al., 2017).

Diante deste contexto, objetivou-se (1) correlacionar o TOC e TON com variáveis de solos, plantas e animais considerando as condições climáticas e balanço hídrico de cada sistema avaliado, para avaliar e comparar o potencial de diferentes combinações dessas variáveis; visando (2) calibrar e testar seis algoritmos de machine learning (MLP, RF, MLR, SVR, KNN, AdaBoost) para estimar o teor de TOC e TON do solo em sistemas de monocultivos de milho, pecuária e sistema de integração Lavoura-Pecuária.

2.2 MATERIAL E MÉTODOS

2.2.1 Local do experimento

O experimento foi conduzido entre novembro de 2015 e janeiro de 2018 no Centro de Pesquisa de Bovinos de Corte, Instituto de Zootecnia/APTA/SAA, Sertãozinho, São Paulo, Brasil (21°8'16” S e 47°59'25” W), com altitude média de 548 m. O clima da região segundo a classificação de Köppen é Aw (Rolim e Aparecido, 2015), caracterizado como tropical úmido com estação chuvosa no verão e seca no inverno (Figura 1). A avaliação ou a determinação dos atributos do solo, variáveis de plantas e animais foi realizado entre março de 2017 e janeiro de 2018. O solo da área experimental é classificado como Latossolo Vermelho distrófico argiloso (Santos et al.,

2018), equivalente a um Latossolo (Oxisol), de acordo com o sistema de classificação de solos do USDA (Bockheim et al., 2014).

2.2.2 Dados climáticos

Dados climáticos de temperatura máxima, mínima e média do ar (°C dia^{-1}), precipitação (mm dia^{-1}), irradiância global (MJ m^{-2} dia^{-1}), umidade relativa (%), velocidade do vento (m s^{-1}) foram obtidos da plataforma NASA-POWER. A partir dos dados climáticos obtidos, o balanço hídrico e a Evapotranspiração de referência (PET) foram calculados pelo método de Penman-Monteith (Allen et al., 1998), usando a linguagem de programação Python. Os decêndios climáticos foram baseados em seis meses antes de cada coleta do solo, foram 18 decêndios para cada variável climática (9 x 18) (nove variáveis climáticas detalhadas na Tabela 1), totalizando 162 variáveis climáticas de entradas nos modelos, que estão detalhados no Apêndice 3. Esses dados meteorológicos foram obtidos a partir do modelo de assimilação Modern-Era Retrospective Analysis for Research and Applications-2 (MERRA-2) e estão disponíveis como séries temporais diárias. Todos os elementos estavam disponíveis para grades de 0,25° × 0,25° (Stackhouse Jr. et al., 2018).

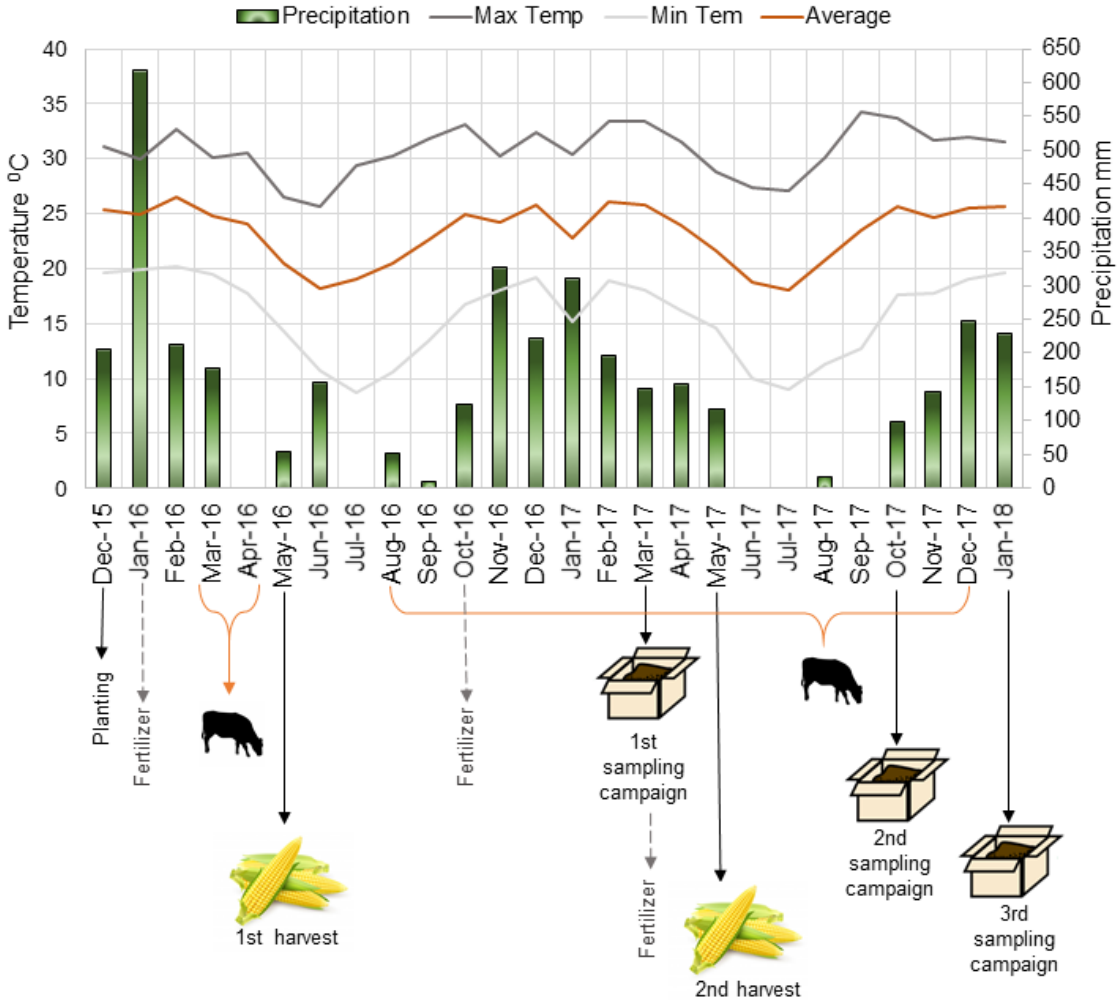


Figura 1. Esquema de atividades e dados climáticos obtidos da estação meteorológica localizada no Centro de Pesquisa de Bovinos de Corte, Instituto de Zootecnia/APTA/SAA, Sertãozinho, São Paulo, Brasil. Fonte: Maia et al. (2021).

2.2.3 Delineamento experimental e tratamentos

O experimento foi conduzido em campo experimental de 16.02 ha dividido em 18 piquetes de 0.89 ha cada, organizados em blocos casualizados com três tratamentos e três repetições. Os tratamentos consistiram na técnica de semeadura consorciada entre milho (*Zea mays* L.) e capim-marandu (*Urochloa brizantha* (Hoechst. Ex A. Rich.) R. D. Webster cultivar marandu), para o estabelecimento do Sistema Integrado Lavoura-Pecuária (ICLS) e duas semeaduras simples (tratamentos testemunha) representando os sistemas convencionais: sistema de cultivo para a produção de grãos de milho (Crop system - CS) e sistema de pecuária para engorda de gado de corte (Livestock system - LS). Todos os tratamentos foram semeados em dezembro de 2015, utilizando semeadora de plantio direto com cinco linhas. Os

detalhes de como foi realizado o manejo desses tratamentos encontram-se em Maia et al. (2021).

No tratamento Pecuária, em que o capim-marandu foi estabelecido solteiro, a pastagem se estabeleceu antes, permitindo que os animais entrassem no sistema em março de 2016. Todos os tratamentos, exceto o tratamento Lavoura, foram submetidos ao pastejo por bovinos de corte da raça Caracu que estavam em fase de recria (média de 14 meses de idade), e a taxa de lotação foi feita segunda a oferta de forragem disponível. Os animais permaneceram em regime de lotação contínua até dezembro de 2017. Foram realizados dois ciclos de lotação contínuos dos animais: o primeiro ciclo foi entre agosto e outubro de 2016 (78 dias), e o segundo ciclo entre novembro de 2016 e dezembro de 2017 (370 dias).

2.2.4 Coleta de dados de solo, plantas e animais

As variáveis de entrada (ou estimadores) incluíram: atributos do solo, planta, animais, e variáveis climáticas (Tabela 1). Os atributos de fertilidade e física do solo foram coletados no início do experimento, novembro de 2015, com profundidade variando entre 10, 40 e 90 cm dependendo do indicador de solo avaliado e metodologia, e durante o decorrer do experimento foram coletadas as informações de plantas e animais. Os atributos microbiológicos e de labilidade de C e N no solo foram coletados durante março de 2017 a janeiro de 2018 (Figura 1). Todas essas variáveis foram analisadas em laboratório seguindo o padrão internacional. Nessa pesquisa foram utilizadas vinte variáveis de solo, quatro variáveis de plantas, três variáveis de animais e nove climáticas para 18 decêndios, totalizando um conjunto total de 189 variáveis de entrada.

Tabela 1. Variáveis de solo, plantas e clima para a estimação do TOC e TON do solo. Médias das variáveis estão detalhadas nos Apêndices 1 e 2.

Descrição (unidade)	Abreviações	Referências
	Planta	
Liteira depositada (g m^{-2})	LD	
Carbono e Nitrogênio da liteira (g kg^{-1})	LIC, LIN	Rezende et al. (1999)
Relação entre Carbono e Nitrogênio da liteira	LIC/LIN	
	Animal	
Nitrogênio da Urina e das Fezes (kg)	NU, NF	Haynes e Williams (1993)
Peso animal (kg)	W	Mendonça et al. (2020)
	Solo	

Potencial de hidrogênio	pH	
Fósforo (mg dm^{-3})	Pr	Raij et al. (2001)
Saturação de bases (%)	B	
Capacidade de troca de cátions ($mmol_c$ dm^{-3})	CEC	
C e N da biomassa microbiana do solo (mg kg^{-1})	MBC, MBN	Monz et al. (1991); Raij et al. (2001); Silva et al. (2007)
Nitrogênio potencialmente mineralizável (mg kg^{-1})	PMN	Roberts et al. (2009)
Carbono lábil (mg kg^{-1})	LC	Weil et al. (2003)
Índice do manejo do carbono	CMI	Blair et al. (1995)
Carbono orgânico total (g kg^{-1})	TOC	Walkley e Black (1934)
Nitrogênio orgânico total (mg kg^{-1})	TON	Raij et al. (2001)
Relação C/N do solo	C/N-soil	
Relação C/N da biomassa microbiana do solo	C/N-SMB	
Estoque de C e N no solo (Ton ha^{-1})	SCS, SNS	EMBRAPA (1997); Veldkamp et al. (1994)
Densidade do solo (g cm^{-3})	BD	
Macroporosidade, Microporosidade (cm^{-3} cm^{-3})	MA, MI	EMBRAPA (1997)
Porosidade total do solo (cm^{-3} cm^{-3})	TP	
Potencial de água no solo (MPa)	SWC	
	Clima	
Temperatura (°C)	T	
Precipitação (mm)	P	
Velocidade do vento (m s^{-1})	U	
Umidade relativa (%)	RH	
Radiação líquida (MJ m^{-2} d^{-1})	RN	
Evapotranspiração potencial (mm)	PET	
Armazenamento de água no solo (mm)	STO	Allen et al. (1998)
Déficit de água no solo (mm mo^{-1})	DEF	
Excedente de água no solo (mm mo-1)	SUR	

2.2.5 Processamento de dados

Os dados utilizados foram as repetições por tratamento, bloco, e período de avaliação ao longo do estudo. Os dados climáticos foram padronizados em escala decendial, iniciando na data de semeadura, e finalizando na data de colheita do milho e da pastagem, totalizando 18 decêndios (Apêndice 3). Antes de iniciar o processo de modelagem, todas as variáveis independentes foram padronizadas por StandardScaler, biblioteca Scikit-learn da linguagem Python.

A modelagem foi realizada em duas abordagens, a primeira foi calcular os coeficientes de correlação de Pearson foram calculados para verificar a influência unitária e diminuir a colinearidade entre as variáveis de entrada e os teores de carbono e nitrogênio orgânico do solo, foram retiradas as variáveis com peso igual a zero ou

correlações menores que 0.30 e maior que 0.90, sendo ela negativa ou positiva. Para melhor representar as influências das variáveis e suas correlações foi usado a biblioteca gráfica Network plot do R Studio.

Na segunda abordagem, o método de Stepwise foi utilizado para selecionar as oito melhores variáveis para estimar o TOC e TON do solo para cada tratamento, dentre as variáveis disponíveis após a limpeza usando a correlação de Pearson. Nessa fase dividimos o banco de dados em duas partes pelo método de CrossValidation (CV= 2) (Figura 2). Em termos simples, o Stepwise ajuda a determinar quais variáveis são importantes e quais não são. O Stepewise por regressão linear múltipla considera que variáveis com o *p-value* alto não contribui significativamente para a precisão do modelo, dessa forma, a eliminação de variáveis é feita através dos métodos forward e backward. Portando, à medida que se insere uma nova variável que é melhor para explicar a variável dependente, as variáveis já incluídas podem se tornar redundantes e são eliminadas pela variável mais explicativa, mais uma etapa importante para evitar a colinearidade. Após essa etapa, as variáveis selecionadas foram representadas graficamente usando o seu coeficiente angular para cada tratamento avaliado.

2.2.6 Modelos de Machine learning

Selecionada as oito variáveis independentes pelo método de Stepwise, foram avaliados seis modelos de ML: Artificial Neural Networks – Multilayer Perceptron Regressor (MLP), Multiple Linear Regression (MLR), Random Forest Regressor (RF), Support Vector Regression (SVR), K Neighbors Regressor (KNN), e Adaptive Boosting Regressor (AdaBoost). O esquema de análises seguiu uma ordem de acordo com a Figura 2.

A rede neural utilizada foi o algoritmo MLP, escolhido devido a possibilidade de modelar problemas não-lineares, apresentar robustez a ruídos e, uma grande capacidade de ser generalizado a resolver qualquer tipo de problema (Heidari et al., 2018; Rana et al., 2018). A principal limitação desse modelo é o seu treinamento inicial, pois é necessário fazer diversas interações aleatórias de treinamentos para escolher o modelo que resulte em menores erros e maior precisão (Pereira et al., 2013).

O modelo RF foi escolhido para criar um conjunto de árvores de decisão e combiná-las, com o objetivo de aumentar o detalhe de informações e o resultado geral, obtendo assim uma predição com maior acurácia (Guo et al., 2004; Genuer et al., 2010). Isso é um ponto positivo, já que uma única árvore de decisão pode ser propensa a ruídos, mas o agregado de muitas árvores de decisão reduz o efeito dos ruídos, gerando resultados mais precisos. Esse modelo também pode lidar com um número muito grande de variáveis de entrada sem ocorrer overfitting (Biau, 2012). A principal desvantagem do RF é a possibilidade de criar árvores tendenciosas. Para evitar esse problema, o conjunto de dados foi padronizado antes de se ajustar à árvore de decisão (Guo et al., 2004).

O modelo MLR foi escolhido por ser um algoritmo simples, além de ser menos complexo em relação aos outros modelos de machine learning. Sua principal desvantagem é assumir que todos os dados são lineares, e quando existem outliers podem ter efeitos negativos na regressão. Ao contrário do MLR, o algoritmo SVR foi escolhido para oferece suporte tanto a regressão linear, quanto a não linear. Outra vantagem é que o SVR performa bem com banco de dados pequenos. Sua principal desvantagem é o tempo de treinamento e sua sensibilidade ao tipo de kernel e C (hyperplane) usado durante a fase de treinamento.

O KNN foi escolhido por ser simples de interpretar, ao mesmo tempo que apresenta uma grande acurácia. Além disso, é bem-sucedido em situações de regressão em que o conjunto de dados é muito irregular (Jadhav e Channe, 2013). Uma desvantagem do modelo KNN é quando contém muitos ruídos quando os dados são incorretamente rotulados, que acabam afetando os resultados (Miloud-Aouidate e Baba-Ali, 2012). O algoritmo AdaBoost foi escolhido para ser usado em pequenos bancos de dados, pois combina várias informações (variáveis) do banco formando uma única variável forte. Mesmo sendo a versão simples do Boosting, esta versão é extremamente rápida em comparação com outros algoritmos, principalmente as redes neurais. No entanto, AdaBoost costuma ter uma baixa performance quando os dados apresentam muitos outliers.

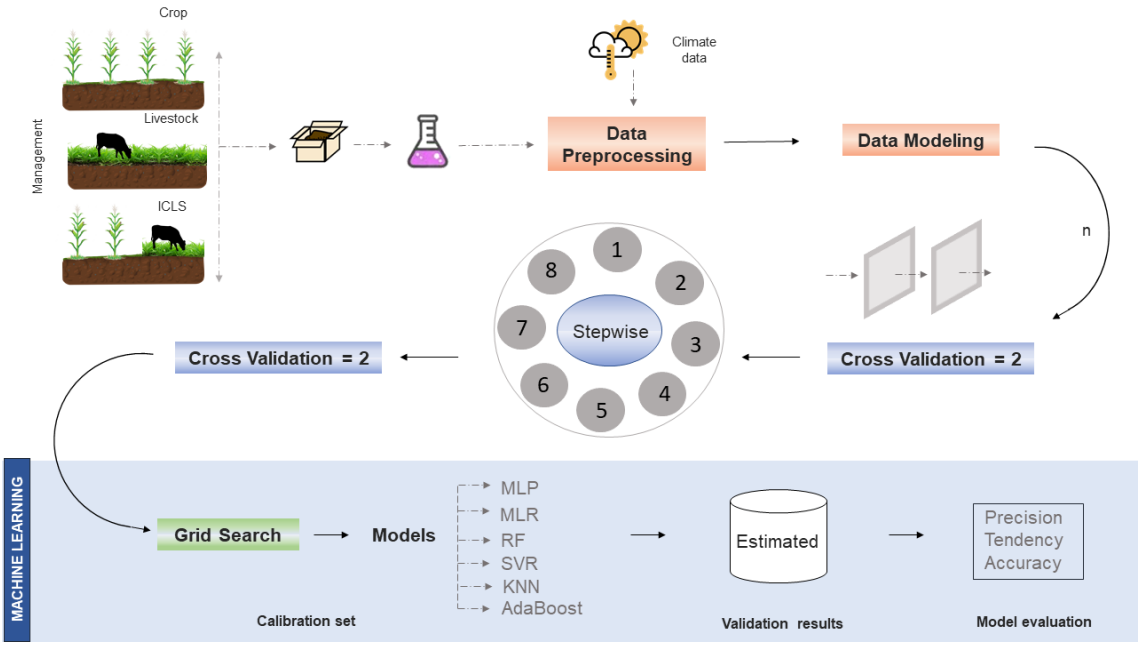


Figura 2. Fluxograma dos processos para a estimação de TOC e TON. MLP: Multilayer Perceptron Regressor, RF: Random Forest Regressor, MLR: Multiple Linear Regression, SVR: Support Vector Regression, KNN: K Neighbors Regressor, e AdaBoost: Adaptive Boosting Regressor.

Em todos os modelos testados, os parâmetros foram ajustados usando o método GridSearch e o banco de dados foi dividido usando o CrossValidation em duas partes (CV= 2) (Figura 2); que funcionou com 2 divisões e uma proporção de 50-50 para a separação dos dados em treinamento e teste. O método Gridsearch permite o teste de diferentes combinações de parâmetros previamente especificados. É um teste poderoso pois permite um mapeamento da melhor combinação que irá minimizar os erros de estimação do modelo em relação aos dados observados. A validação cruzada ou CrossValidation nos permite treinar e testar o modelo para que não capte os padrões dos dados, principalmente em banco de dados pequenos, evitando a variância e generalização do modelo, com isso obtemos um resultado mais robusto. Os parâmetros testados de cada modelo e os melhores valores ajustados para cada tratamento foram apresentados no Apêndice 4.

2.2.7 Avaliação do desempenho dos modelos

Foram utilizadas métricas de precisão (coeficiente de determinação ajustado (R^2 ajustado), Eq. 1), tendência (Erro Sistemático, Eq. 2) e acurácia (Índice de Willmott et al. (1985), Eq. 3), para avaliar o desempenho dos modelos.

$$R^2\ adjusted = 1 - \left[\frac{\left(1-R^2\right)(n-1)}{n-k-1}\right] \quad (1)$$

$$SE = \sqrt{\frac{\sum_{i-1}^{N}(Yobs_i - Ymean)^2}{N}} \quad (2)$$

$$d = 1 - \frac{\sum_{i-1}^{N}(Yobs_i - Yest_i)^2}{\sum_{i-1}^{N}(|Yest_i - Ymean| + |Yobs_i - Ymean|)^2} \quad (3)$$

Para avaliar a acurácia, foram utilizados o MAE (Erro Absoluto Médio) (Eq. 4), MAPE (Erro Percentual Absoluto Médio) (Eq. 5) e RMSE (Raiz Quadrada do Erro Médio) (Eq. 6):

$$MAE = \frac{\sum_{i=1}^{N}\left|Y_{obs_i} - Y_{est_i}\right|}{N} \quad (4)$$

$$MAPE = \frac{\sum_{i=1}^{N}\left(\left|\frac{Yest_i - Yobs_i}{Yobs_i}\right|100\right)}{N} \quad (5)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{obs_i} - Y_{est_i}\right)^2}{N}} \quad (6)$$

em que o R^2 é o coeficiente de determinação ajustado, n é o número de elementos, k é o número de variáveis de regressão, SE é o erro sistemático, d é o Índice de Willmott, Yobs são os dados observados, Yest são os dados estimados, Ymean é a média dos dados.

O Diagrama de Taylor também foi utilizado para avaliação do desempenho dos modelos e destaca os melhores desempenhos dos modelos em comparação com os dados observados. Neste diagrama, os ângulos referem-se à correlação coeficiente entre os dados observados e estimados. A distância radial da origem denota a razão entre o desvio padrão normalizado (SD) da simulação e o da observação (Taylor, 2001). No diagrama de Taylor, se os pontos estimados estão mais próximos do ponto de referência, a precisão do modelo é alta. No entanto, se os pontos estimados

estiverem longe do ponto de referência, a precisão do modelo é baixa (Ahmadalipour et al., 2018; Ghorbani et al. 2018; Taylor, 2001). Além disso, se os pontos estiverem longe do ponto de referência, indica que a correlação entre os parâmetros de entrada e saída é baixa, e o desempenho desses modelos de machine learning não podem estimar corretamente o teor de TOC e TON do solo.

2.3. RESULTADOS E DISCUSSÃO

2.3.1 Importância e seleção das variáveis de entrada no modelo

As variáveis de qualidade do solo (química, física e microbiológica), plantas e animais se correlacionaram fortemente com o TOC e TON do solo (Figuras 3 e 4).

Foram observadas correlações positivas com indicadores de labilidade do carbono e fertilidade do solo, e o teor de TOC e TON no sistema de monocultivo de milho (Figuras 3 A, B). O C/N-Soil foi a variável que apresentou alta correlação negativa com o TON nos tratamentos de monocultivos de milho e pecuária (Figuras 3 B, D). A alta correlação negativa de C/N do solo com o nitrogênio do solo é sustentada pela hipótese de que a relação C/N do solo é um índice que permite avaliar o grau de evolução da matéria orgânica no sistema e sua mineralização (Shen et al., 2019). O C/N do solo variou entre os tratamentos de 14:1 a 15:1 (Apêndice 1), alguns autores afirmaram que a relação C:N menor que 20:1, indica maior mineralização da matéria e acréscimo de nitrogênio disponível para as plantas (Poffenbarger et al., 2018, Sheng et al., 2019).

Os tratamentos de monocultivo (CS e LS) tiveram correlação positiva com variáveis de biomassa microbiana do solo (carbono lábil-LC, Índice do manejo do carbono-CMI e N da biomassa microbiana-MBN) (Figuras 3 A, C). O carbono lábil e o Índice do manejo do carbono estimam um reservatório de C que está mais intimamente associado a funções lábeis do solo e a ciclagem de nutrientes, e são dependentes do C existente na biomassa microbiana do solo (Duval et al. 2018; Bongiorno et al. 2019). Enquanto o teor de N da biomassa microbiana teve um efeito de correlação positiva do pastejo e está associado a urina e fezes dos animais nos sistemas que contém animais (Oliveira et al., 2022), devido a mineralização e disponibilidade de N no solo (Rakkar et al., 2017; Maia et al., 2021).

Além disso, o peso de animais teve uma forte correlação negativa com o teor de TON no tratamento monocultivo de pecuária (Figura 3 D). Essa correlação demonstrou que a diminuição do teor de N ocorreu quando não existiu o consórcio da pastagem e outros sistemas de cultivos, como o que ocorreu nos sistemas integrados, e consequentemente impactou na disponibilidade desse importante nutriente no solo (Teutscherová et al., 2021). Em contraproposta, Dubeux Jr. e Sollemberguer (2020) afirmaram que o fornecimento de nutrientes via produção animal é uma das principais vantagens da pecuária, portanto, não são considerados perdas. No entanto, é necessário o equilíbrio entre as entradas e saídas de nutrientes, para manter balanceadas as quantidades de nutrientes nos sistemas de produção.

Ao contrário dos tratamentos de monocultivos (CS e LS) em que foram obtidas correlações menos densas, os tratamentos ICLS apresentaram diferentes e altas correlações com as variáveis preditoras (Figura 4). Em geral, as correlações mais densas nos sistemas integrados (ICLS) e com maior número de variáveis evidenciam o quanto sistemas integrados podem contribuir com melhores indicadores de qualidade do solo, sendo eficientes quando se avalia os serviços e contribuições ambientais agroecossistemicos (Lemaire et al., 2014). O TOC, TON e a concentração de carbono microbiano do solo (MBC) tiveram uma correlação positiva entre si (0.50 e 0.1, respectivamente), indicando microfaunas mais abundantes no tratamento ICLS (Figura 4). Assim, mudanças ambientais para aumentar a comunidade microbiana do solo podem resultar em aumento de curto prazo dos reservatórios ativos de C e N no solo (Strickland et al., 2019; Maia et al.; 2021).

Outros estudos relataram a boa eficiência de variáveis de solo e ambientais para estimar as concentrações de TOC e TON no solo (Jonh et al., 2020; Chen et al., 2021; Goydaragh et al., 2021), consistente com os achados desse estudo, obtiveram correlações de Pearson variando entre 0.4 a 0.85. No entanto, foi a primeira vez que os resultados das interações de peso animal, urina e fezes foram adicionados nas análises de estimação desses indicadores de qualidade do solo.

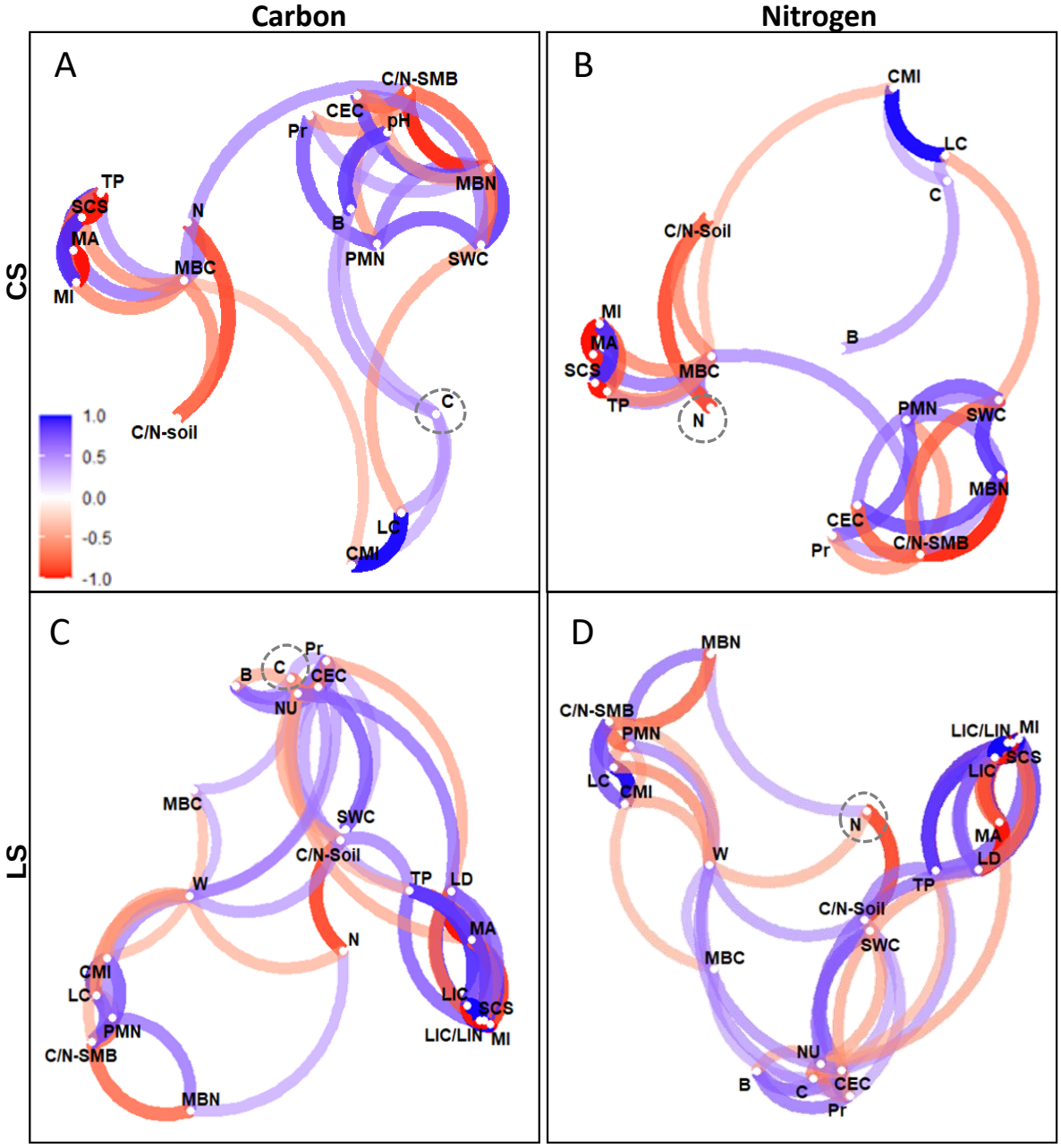


Figura 3. Correlação das variáveis de solo, planta e animal com o carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS. Correlações menores que 0.5 e -0.5 apresentam cores mais opacas. Correlações negativas estão em vermelho. CS: monocultivo de Milho; LS: monocultivos de Pecuária. C: carbono orgânico total; N: nitrogênio orgânico total; MBC: carbono da biomassa microbiana; MBN: nitrogênio da biomassa microbiana; LC: carbono lábil; CMI: índice de manejo do carbono; PMN: nitrogênio potencialmente mineralizável; C/N-soil: relação carbono e nitrogênio do solo; C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo; BD: densidade do solo; MA: macroporosidade; MI: microporosidade; TP: porosidade total; SCS: estoque de carbono; SNS: estoque de nitrogênio; pH: potencial de hidrogênio; Pr: fósforo; CEC: capacidade de troca de cátions; B: saturação de bases; SWC: potencial de água no solo. NU: nitrogênio da urina; NF: nitrogênio das fezes; W: peso animal; LD: liteira

depositada; LIC: carbono da liteira; LIN: nitrogênio da liteira; LIC/LIN: relação carbono e nitrogênio da liteira.

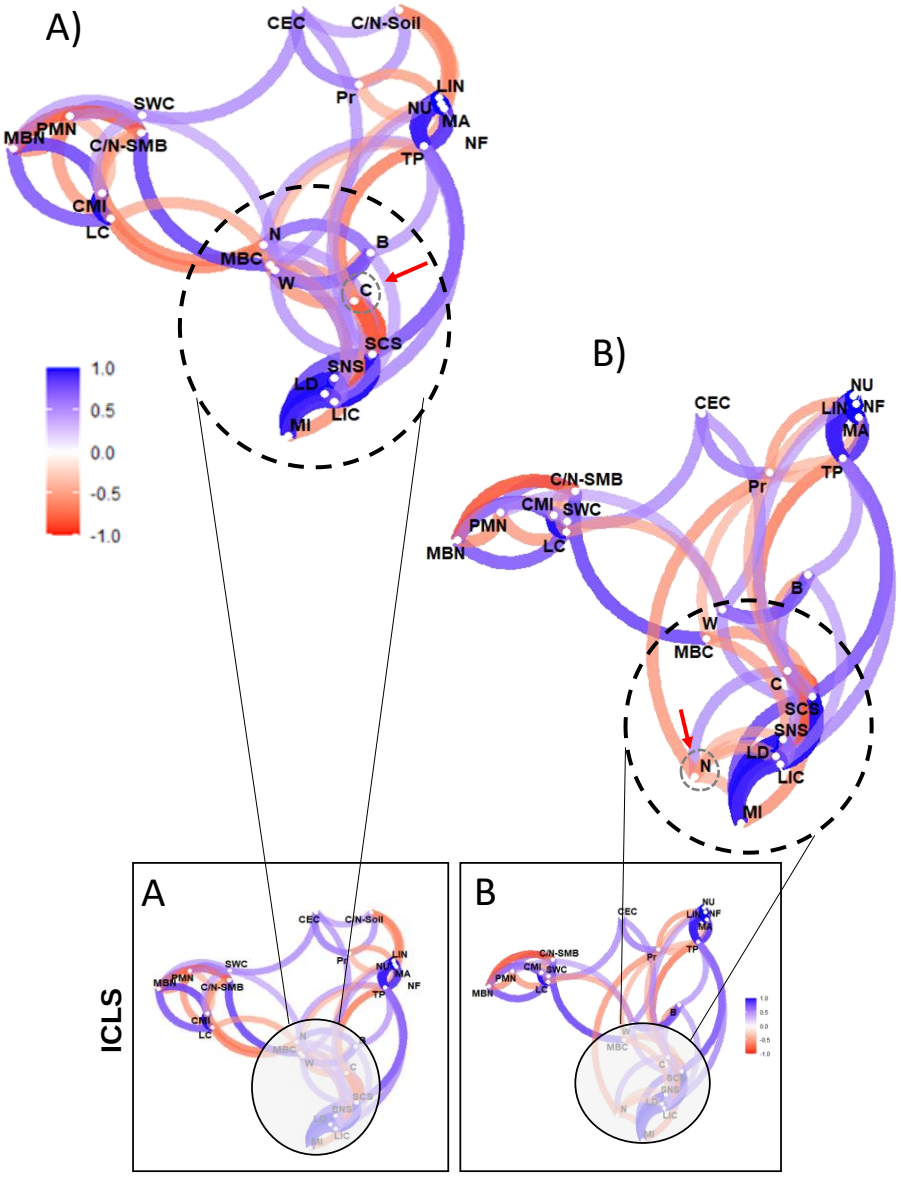


Figura 4. Correlação das variáveis de solo, planta e animal com o carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-ICLS e (**B**) TON-ICLS. Correlações menores que 0.5 e -0.5 apresentam cores mais opacas. Correlações negativas estão em vermelho. ICLS: Integração Lavoura-Pecuária. C: carbono orgânico total; N: nitrogênio orgânico total; MBC: carbono da biomassa microbiana; MBN: nitrogênio da biomassa microbiana; LC: carbono lábil; CMI: índice de manejo do carbono; PMN: nitrogênio potencialmente mineralizável; C/N-soil: relação carbono e nitrogênio do solo; C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo; BD: densidade do solo; MA: macroporosidade; MI: microporosidade; TP: porosidade total; SCS: estoque de carbono; SNS: estoque de nitrogênio; pH: potencial de hidrogênio; Pr: fósforo; CEC: capacidade de troca de cátions; B: saturação de bases; SWC: potencial de água no solo. NU: nitrogênio da urina; NF: nitrogênio das fezes; W: peso

animal; LD: liteira depositada; LIC: carbono da liteira; LIN: nitrogênio da liteira; LIC/LIN: relação carbono e nitrogênio da liteira.

As variáveis climáticas variaram ao longo dos 18 decêndios com correlações positivas e negativas durante o ciclo de cultivo do milho e da pastagem (Figura 5). Em relação ao TOC, os tratamentos LS (Figura 5 C) e ICLS (Figura 5 E) obtiveram as menores correlações com as variáveis climáticas, variando entre 0.4 e -0.4. Em relação ao TON, o tratamento CS foi o único que obteve correlações maiores que 40% (Figura 5 B).

A temperatura (T) teve correlação positiva com os teores de TON e TOC com o aumento da temperatura nos dez primeiros decêndios, e a correlação foi negativa quando a temperatura decresceu no final do ciclo. O mesmo ocorreu com a velocidade do vento (U) e a deficiência de água no solo (DEF), pois durante o inverno e no final do ciclo o solo tende a ser mais seco nessa região. O contrário ocorreu com a precipitação (P), umidade relativa (RH) e a armazenamento de água no solo (STO) durante o ciclo dos cultivos, com uma correlação negativa no início do ciclo e positiva no final, pois houve um aumento de chuvas a partir do decêndio 6, que deixou o solo mais úmido. No entanto, com a chegada do inverno no final do ciclo e a diminuição das chuvas (Figuras 1 e 5), a correlação se tornou negativa. Enquanto a evapotranspiração potencial (PET) e a radiação líquida (RN) tiveram uma grande influência positiva durante todo o ciclo dos cultivos, e não houve excedente hídrico (SUR) durante o ciclo, o que era esperado para a região, que tem em média uma precipitação de 1200 mm distribuídos durante o ano.

De acordo com alguns autores, as condições climáticas influenciaram direta e indiretamente a agricultura, sendo o principal responsável pelas alterações dos teores de TOC e TON no solo (Miller et al., 2004; Omer et al., 2018; Jansson e Hofmockel, 2020; Maia et al., 2021). Além disso, a temperatura do ar e a precipitação estão entre os elementos climáticos que têm grande influência sobre esses indicadores de qualidade do solo (Sun et al., 2013; Wuest, 2014; Iqbal et al., 2015; Engelhardt et al., 2021).

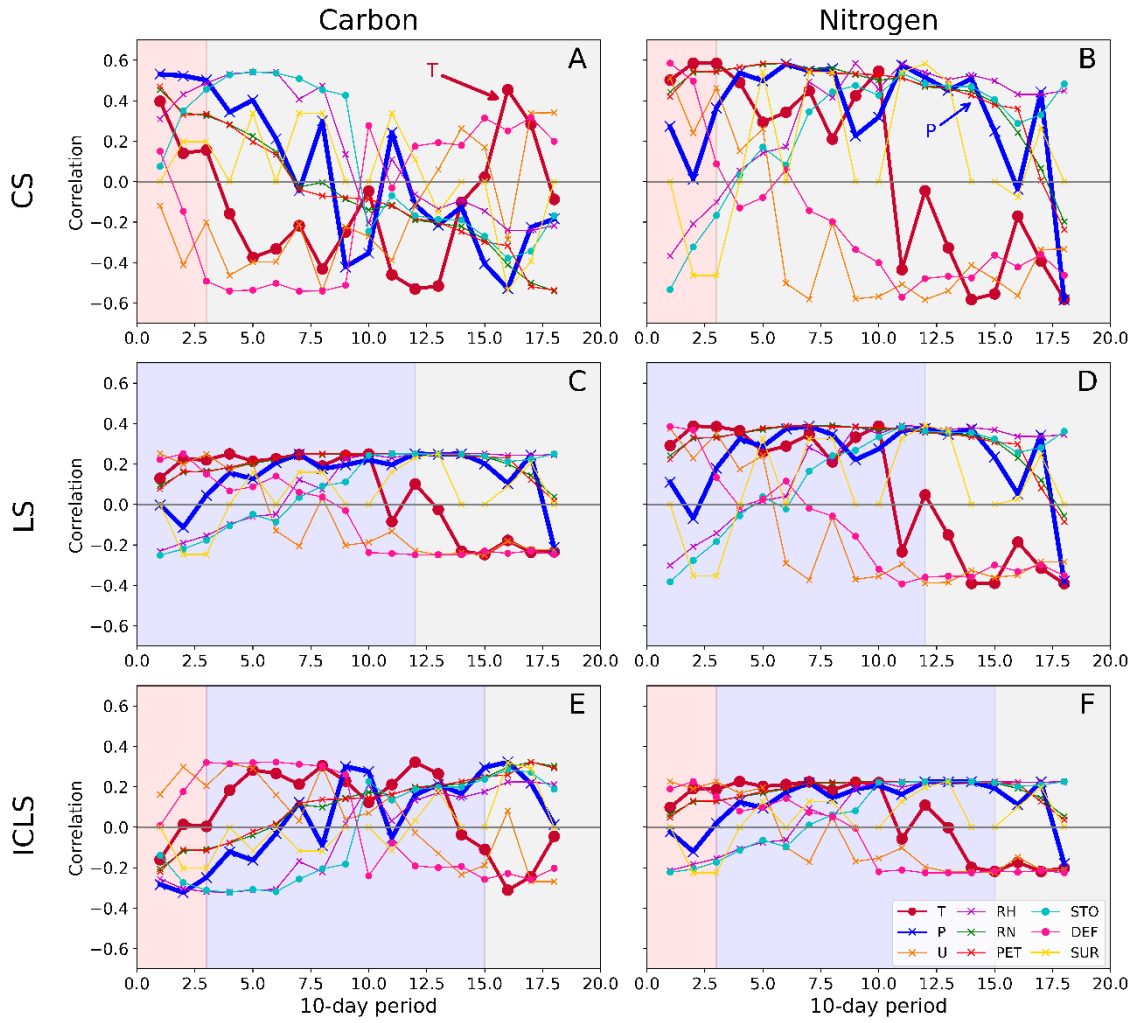


Figura 5. Correlação das variáveis climáticas com o carbono e nitrogênio orgânico do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária. Vermelho: período com o cultivo de milho; Azul: período com o pastejo de animais; Cinza: período de pousio. T: temperatura (°C); P: precipitação (mm); U: velocidade do vento (m s^{-1}); RH: umidade relativa (%); RN: radiação líquida (MJ m^{-2} d^{-1}); PET: evapotranspiração potencial (mm); STO: armazenamento de água no solo (mm); DEF: déficit de água no solo (mm mo^{-1}); SUR: excedente de água no solo (mm mo^{-1}).

A importância relativa das variáveis para a estimativa do conteúdo TOC e TON foi produzida pelo método de Stepwise, após a padronização do banco de dados foram representadas graficamente usando o seu coeficiente angular (Figura 6). No tratamento CS, as variáveis foram classificadas na ordem de (1) solo e (2) clima. A saturação de bases (B) e nitrogênio potencialmente mineralizável (PMN) foram identificados como os estimadores mais importante do conteúdo TOC e TON, seus

coeficientes angulares (R) foram de 0.87 e 0.06, respectivamente (Figuras 6 A e B). A maior proporção de ativos de nitrogênio potencialmente mineralizável para TON sugeriu uma decomposição e mineralização mais rápida da matéria orgânica do solo em CS do que nos outros tratamentos, os resíduos orgânicos de milho podem se traduzir em maior disponibilidade de N a curto prazo durante o ciclo sazonal (Strickland et al., 2019).

No tratamento LS, as variáveis foram classificadas na ordem de (1) clima, (2) solo e (3) plantas (Figuras 6 C e D). O segundo e terceiro decêndio de radiação líquida (RN) apresentaram alta importância relativa para afetar o teor de TOC (coeficiente angular de 30). Enquanto o carbono da liteira (LIC) e índice do manejo do carbono (CMI) foram os fatores mais importante para estimar o conteúdo de nitrogênio do solo no tratamento LS (coeficientes angulares foram de 5 e -5, respectivamente). Os segundo e terceiro decêndio de radiação liquida (RN-2 e RN-3) tiveram uma correlação negativa com o teor de TOC nos primeiros decêndios (Figura 5 C), indicando que a relação inversa se deve ao fato que baixa radiação solar, gera menos produção de biomassa de gramíneas, e consequentemente, podem impactar diretamente no estoque de carbono a longo prazo no solo (Ramírez et al., 2020). Para estimar o teor de TON, os microrganismos do solo continuam tendo impacto no estoque de carbono orgânico do solo. Pois o índice do manejo do carbono do solo é dependente do carbono existente na biomassa microbiana, e em conjunto com o carbono da liteira das gramíneas de pastagens, tem uma presença abundante de raízes mais finas, formando uma rede complexa que afeta a estrutura do solo, interfere na porosidade, essa região da rizosfera proporciona maior abundância de microrganismos, que afetam diretamente na disponibilidade N no solo (Kabiri et al., 2016; Oliveira et al., 2016).

As variáveis do tratamento ICLS foram classificadas como (1) clima, (2) solo e (3) plantas. Variáveis climáticas, solo e plantas apresentaram alta importância relativa para estimar o TOC e TON. Entretanto, o tratamento ICLS foi o único que apresentou variáveis de animais com peso de importância para estimar o carbono do solo (fezes e urina), enquanto o sistema LS (pecuária) não incluiu nenhuma variável de animal com importância. Para estimar o teor de TOC no ICLS (Figura 6 E), as variáveis mais importantes foram as climáticas (quarto decêndio de precipitação-P-4 e água no solo-SUR), seguida de solos, plantas e animais (Urina e Fezes). Para a estimativa de TON

seguiu o fluxo contrário (Figura 6 F), as variáveis mais importantes foram de solos (carbono orgânico, C/N do solo, fósforo e saturação por bases), seguido das climáticas (água no solo-SUR) e de planta (carbono da liteira).

O quarto decêndio de precipitação (P-4) apresentou um grande peso para estimar o teor de TOC no ICLS, com uma correlação negativa (Figura 5 E) indicando diminuição de chuva no início do ciclo, favorecendo a aeração nos poros do solo e a atividade aeróbica da microbiota que estão diretamente ligadas ao estoque de C no solo (Niu et al., 2016; Zhu et al., 2018a; Ma e Chang, 2019).

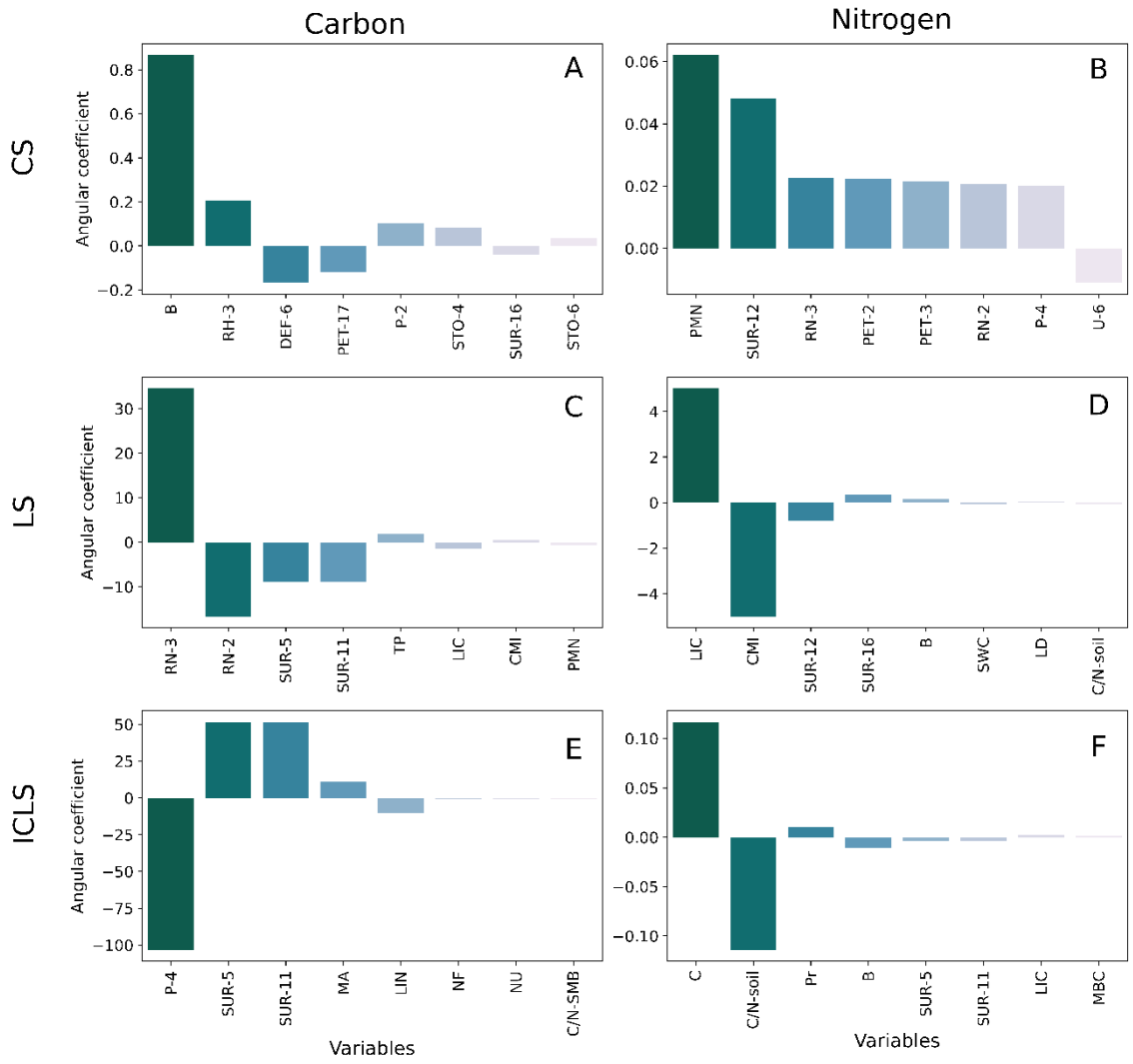


Figura 6. Importância das variáveis pelo método de Stepwise para estimar carbono e nitrogênio orgânico do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária. C: carbono orgânico total; MBC: carbono da biomassa microbiana; CMI: índice de manejo do carbono;

PMN: nitrogênio potencialmente mineralizável; C/N-soil: relação carbono e nitrogênio do solo; C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo; MA: macroporosidade; TP: porosidade total; Pr: fósforo; B: saturação de bases; SWC: potencial de água no solo; NU: nitrogênio da urina; NF: nitrogênio das fezes; LD: liteira depositada; LIC: carbono da liteira; LIN: nitrogênio da liteira. DEF-6: sexto decêndio de déficit de água no solo; P-2: segundo decêndio de precipitação; P-4: quarto decêndio de precipitação; PET-2: segundo decêndio de evapotranspiração potencial; PET-3: terceiro decêndio de evapotranspiração potencial; PET-17: décimo sétimo decêndio de evapotranspiração potencial; RH-3: terceiro decêndio de umidade relativa; RN-2: segundo decêndio de radiação líquida; RN-3: terceiro decêndio de radiação líquida; STO-4: quarto decêndio de armazenamento de água no solo; STO-6: sexto decêndio de armazenamento de água no solo; SUR-5: quinto decêndio de excedente de água no solo; SUR-11: décimo primeiro decêndio de excedente de água no solo; SUR-12: décimo segundo decêndio de excedente de água no solo; SUR-16: décimo sexto decêndio de excedente de água no solo; U-6: sexto decêndio de velocidade do vento.

2.3.2 Desempenho da estimação de carbono e nitrogênio orgânico do solo

Os modelos MLP e AdaBoost indicaram um alto desempenho para estimação de TOC e TON, os modelos MLR, RF, KNN e SVR apresentaram algumas limitações (Tabela 4). De uma forma geral, MLP e o AdaBoost foram os mais precisos e acurados para estimar o TOC e TON, uma vez que obtiveram um erro médio menor que 5% (MAPE) e precisão de R^2= 0.99. Vale ressaltar que um erro de 5% nestes modelos é considerado baixo, pois os tratamentos tiveram uma média de TOC igual a 21 g kg^{-1} e TON igual a 2.5 mg kg^{-1} (Apêndice 1), ou seja, observou-se uma variação média de apenas ±1.05 g kg^{-1} e ±0.125 mg kg^{-1} na sua estimação, respectivamente. O coeficiente de determinação ajustado (R^2) variou de 0.70 a 0.99 entre os modelos mais precisos, indicando um alto ajuste do modelo, onde, 70% a 99% das variações nas estimativas de TOC e TON do solo foram explicadas pelas variáveis selecionadas pelo método de Stepwise.

Tabela 4. Desempenho estatísticos dos algoritmos na estimativa de carbono orgânico e nitrogênio do solo nos tratamentos. MLP: Multilayer Perceptron Regressor, RF: Random Forest Regressor, MLR: Multiple Linear Regression, SVR: Support Vector Regression, KNN: K Neighbors Regressor, e AdaBoost: Adaptive Boosting Regressor.

		Carbono						Nitrogênio					
Tratamentos		**MLP**	**RF**	**MLR**	**SVR**	**KNN**	**AdaBoost**	**MLP**	**RF**	**MLR**	**SVR**	**KNN**	**AdaBoost**
CS	R^2- ajustado	0.99	0.13	0.26	0.01	0.25	0.94	0.99	0.39	0.01	0.05	0.01	0.99
	MAE	7.272	1.084	1.043	1.046	1.046	0.015	0.003	0.108	0.125	0.117	0.138	0.010
	RMSE	0.001	1.2273	1.1368	1.4556	1.4556	0.3253	0.005	0.141	0.181	0.176	0.191	0.018
	MAPE	0.002	3.110	3.005	2.884	3.050	0.449	0.117	4.514	5.252	5.158	5.765	0.437
	d	0.999	0.773	0.822	0.449	0.449	0.989	0.999	0.809	0.743	0.818	0.639	0.998
	SE	1.669	1.669	1.669	1.695	1.679	1.669	0.229	0.229	0.229	0.232	0.229	0.229
LS	R^2- ajustado	0.99	0.14	0.40	0.93	0.01	0.99	0.99	0.25	0.99	0.76	0.56	0.99
	MAE	0.001	0.888	0.836	0.192	1.048	0.028	0.001	0.105	2.319	0.084	0.098	0.001
	RMSE	0.0001	1.1582	0.9642	0.3377	1.3806	0.0766	0.0002	0.1588	2.7255	0.0885	0.1205	0.0030
	MAPE	0.003	2.563	2.395	0.528	3.015	0.077	0.006	4.283	9.881	3.681	4.098	0.068
	d	0.999	0.743	0.874	0.987	0.588	0.999	0.999	0.774	0.999	0.951	0.891	0.999
	SE	1.577	1.577	1.577	1.579	1.580	1.577	0.232	0.232	0.232	0.233	0.232	0.232
ICLS	R^2- ajustado	0.99	0.61	0.84	0.01	0.61	0.99	0.99	0.69	0.99	0.33	0.41	0.99
	MAE	1.761	0.747	0.500	0.943	0.718	0.009	0.001	0.043	0.001	0.064	0.064	0.009
	RMSE	2.2717	0.8834	0.5710	1.5533	0.887	0.0215	0.0002	0.0521	0.0014	0.0762	0.0712	0.0022
	MAPE	4.837	2.032	1.394	2.978	1.930	0.027	0.006	1.831	0.054	2.726	2.640	0.038
	d	0.999	0.918	0.972	0.763	0.925	0.999	0.999	0.938	0.999	0.830	0.876	0.999
	SE	1.789	1.789	1.789	1.807	1.795	1.789	0.118	0.124	0.118	0.124	0.118	0.118

R^2: coeficiente de determinação ajustado; *MAE*: erro médio absoluto; *RMSE*: raiz quadrada do erro médio; *MAPE*: erro percentual absoluto médio; *d*: Índice de Willmott et al. (1985); *SE*: erro sistemático; *CS:* monocultivo de Milho; *LS:* monocultivos de Pecuária; *ICLS:* Integração Lavoura-Pecuária.

De modo geral, as redes neurais artificiais tiveram alta precisão (R^2= 0.99), baixa tendência (e SE ≤2) e alta acurácia (RMSE ≤0.01, d= 0.99) (Figura 7). O uso da rede neural MLP já foi utilizado para a estimar o teor de TOC na Nigéria, e obteve baixa precisão (R^2= 0.36) devido a quantidade baixa de variáveis de entrada na calibração e tendência de multicolinearidade (John et al., 2020). Fernandes et al. (2019) calibrou a MLP utilizando 6 variáveis de fertilidade do solo para estimar o teor de matéria orgânica do solo, testou diversas camadas das redes e obteve R^2= 0.92 de precisão e um erro de 1.82 g kg^{-1}.

Li et al. (2013) propuseram estimar o conteúdo de matéria orgânica com base em 11 variáveis climáticas e de uso da terra na China, usando modelagem por MLP e MLR. Ao final do estudo, os autores observaram que os MLPs foram mais precisos que a MLR na estimação, e que as 11 variáveis de entrada apresentaram boa acurácia na estimativa da matéria orgânica do solo. A alta precisão da estimativa obtida por esses autores se deve à dependência entre fatores climáticos e de uso do solo, como chuva, temperatura, umidade e radiação e fatores do solo, como declividade do terreno e teor de matéria orgânica no solo, atuando diretamente na sua decomposição (Balogh et al., 2011), indicando que, inserir variáveis climáticas na calibração e estimativa de indicadores do solo otimiza o modelo.

Ao estimar o TOC e TON do solo, o algoritmo RF teve baixa precisão para a maioria dos tratamentos avaliados (R^2= 13-69%), apresentando alta tendência na distribuição dos dados (Figura 8, Tabela 4). Apesar da baixa precisão do RF, obteve acurácia moderada para estimar TOC e TON, com MAPE máximo variando de 3-4.5% dos valores verdadeiros. Esses resultados indicam que o modelo RF poderia aumentar a sua precisão aumentando o número de variáveis estimadoras, ou até mesmo, usando o próprio modelo de RF para selecionar as variáveis. Portanto, suportamos a hipótese que o uso do Stepwise baseado em MLR, pode estar ocasionando essa baixa precisão do modelo, mas no geral o modelo é acurado e próximo dos dados observados de TOC e TON (Tabela 4).

Para a previsão espacial das concentrações e estoques de TOC da Ilha Barro Colorado no Panamá, Grimm et al. (2008) também aplicaram RF, mas alcançaram variações explicadas muito baixas de 6-23% (indicadas por MSE variando entre 0,94 a 0,77) para diferentes profundidades do solo. Outro estudo indicou que o modelo RF

apresentou a maior tendência em superestimar do teor de TOC (Were et al., 2015), esses resultados são consistentes com os achados desse estudo, pois os RMSEs dos dados de teste superestimam quando os teores são baixos e subestimam quando os teores de TOC e TON são altos, independente do tratamento (Figura 8). No entanto, outro estudo indicou precisão do RF de 68%, mas com os menores RMSE para estimar TOC (RMSE = 0.20 g kg^{-1}) entre os modelos avaliados, inclusive o SVR (John et al., 2020).

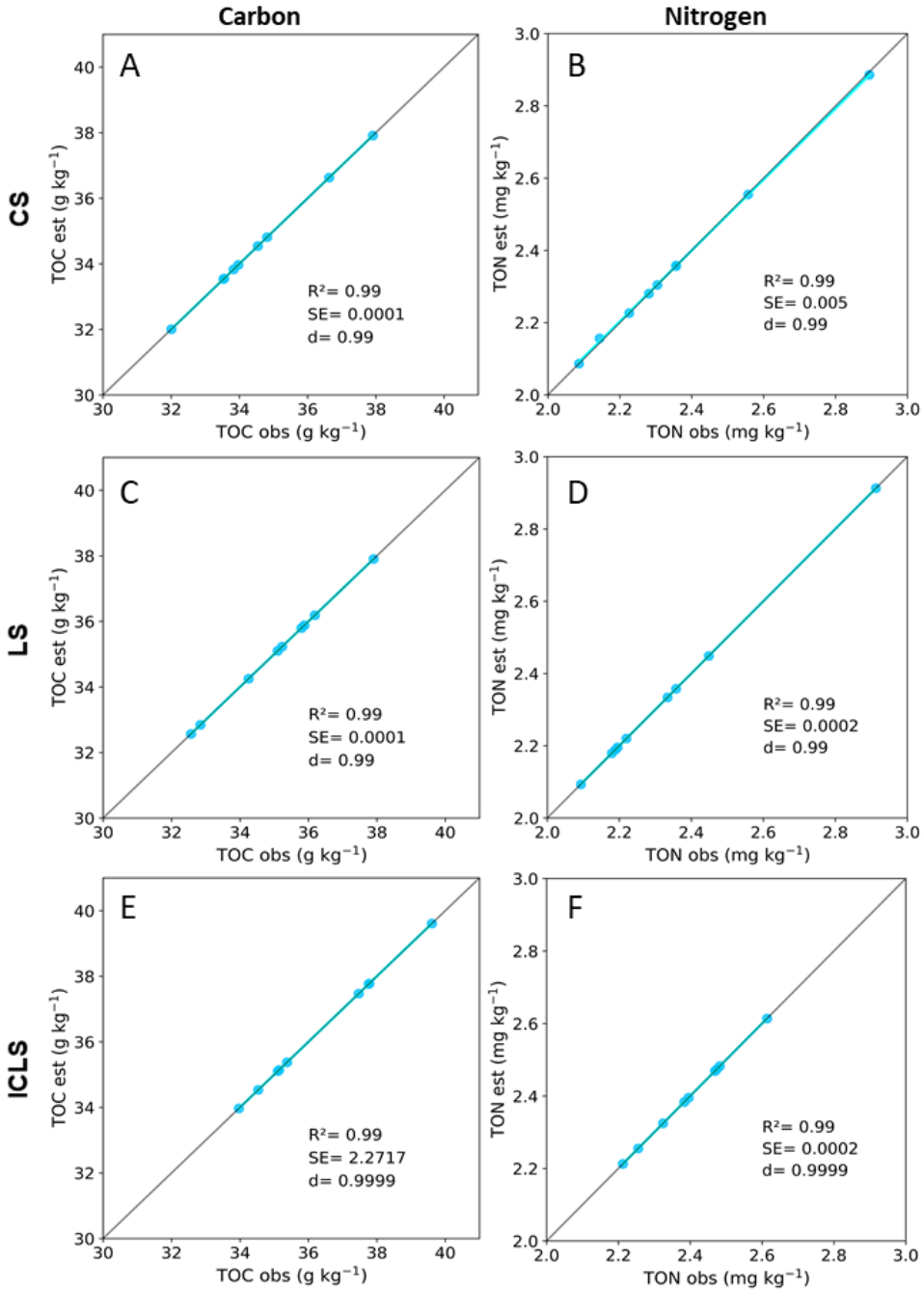


Figura 7. Desempenho do modelo Multilayer Perceptron Regressor (MLP) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

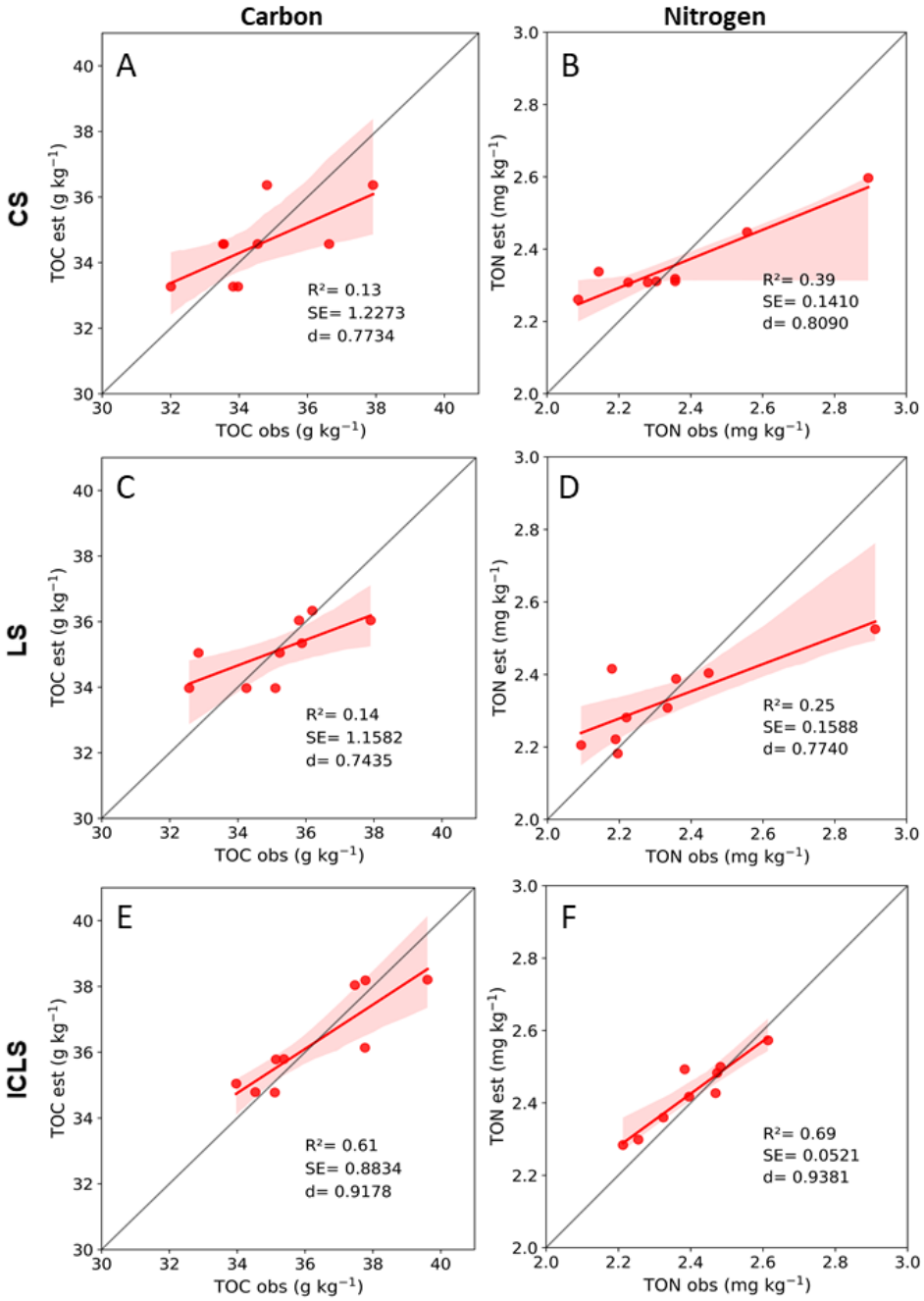


Figura 8. Desempenho do modelo Random Forest Regressor (RF) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

Apesar da menor calibração do algoritmo MLR quando comparado com os outros algoritmos testados neste trabalho, o MLR demonstrou alta precisão, com um

R^2 máximo de 0.99; alta acurácia, com um MAPE máximo de 5% e baixa tendência, com um valor de SE máximo de 1.46 g kg^{-1} para TOC e 2.72 mg kg^{-1} para TON (Figura 9). Apesar da baixa precisão nos tratamentos com monocultivos para estimar TOC (Figuras 9 A e C) e TON (Figura 9 B), ambos tiveram acurácia moderada, com "d" mínimo de 0.82 e 0.74, MAE máximo de 1.04 e 0.12, e MAPE máximo de 3% e 5% para a estimativa de TOC e TON, respectivamente. Com base nesse resultado, foram observados que o desempenho do modelo MLR para estimativa de TOC foi maior em sistema de integração Lavoura-Pecuária (ICLS) (R^2= 0.84 e RMSE = 1.394 g kg^{-1}), em sistemas de monocultivos (LC e CS) o modelo teve tendência em subestimar ou superestimar o teor de TOC no solo (R^2= 0.40). Por fim, o MLR foi preciso para estimar o TON em áreas de LS e ICLS (R^2= 0.99) (Figuras 9 D e F). Isso quer dizer que apesar da baixa precisão desse modelo nos sistemas de monocultivo de pecuária, esse algoritmo apresentou melhoria quando combinado com variáveis climáticas, solo e planta para estimar C e N orgânico do solo, que é o caso dos sistemas integrados.

Esses resultados corroboram com John et al. (2020), que demonstraram a utilidade das variáveis ambientais mais as variáveis relacionadas ao solo para explicar a distribuição do TOC na profundidade do solo, mas somente quando ambas são usadas em conjunto como preditoras de TOC.

Em geral, a capacidade de estimação das redes neurais e o Adaboost foi superior à dos modelos de regressão linear múltipla (Tabela 4). O tratamento de monocultivo de milho CS obteve a menor caracterização pelas variáveis de entrada utilizando o MLR (R^2 = 0.26 para TOC e 0.01 para o TON), que pode estar associado ao menor número de variáveis de entrada diversificadas entre clima-planta-solo-animal (Figuras 6 A e B), e com uma menor densidade de correlações quando comparado aos sistemas integrados (Figura 4). John et al. (2020) indicaram em seu estudo que a calibração do modelo com poucas variáveis e exclusivas de solo, em MLR tende a uma baixa explicabilidade (R^2= 0.17), consistente com os achados nesse estudo. Além do mais, sugeriu inserir variáveis climáticas para aumentar a robustez do modelo.

Assim como o RF, os algoritmos SVR e KNN tiveram baixa precisão para a estimativa do TOC e TON para alguns tratamentos. Para o SVR, a precisão variou de R^2= 0.01 a 0.93 (Figura 10, Tabela 4). Os tratamentos CS e ICLS foram os que

obtiveram a menor precisão (R^2) utilizando o modelo SVR, variando o seu poder de explicação entre 1 a 33% para estimar o TOC e TON do solo (Figuras 10 A, B, E, F). Entre os modelos testados no presente estudo, o algoritmo KNN demonstrou menor precisão para estimar o TOC e TON para a maioria dos tratamentos (Figura 11, Tabela 4). A precisão máxima desse modelo foi de R^2= 0.61 (Figura 11 E), apresentando um valor médio de R^2= 0.31 de precisão entre os tratamentos.

Os modelos SVR e o KNN obtiveram a menor acurácia, com um “d” mínimo de 0.44 (o menor entre os modelos), e um MAPE variando de 3% para TOC e 5% para TON. No entanto, o KNN apresentou um MAPE máximo de aproximadamente 6% (o maior erro percentual entre os modelos). Isso quer dizer que o erro da média do KNN para o TOC é de 0.63 g kg^{-1} e 0.15 mg kg^{-1} para o TON. Apesar da baixa acurácia e precisão do KNN em relação aos outros modelos, esse modelo teve uma tendência na estimativa do TOC similar ao SVR, apresentando um valor máximo de tendência de SE=1.80 g kg^{-1}, em relação aos demais modelos.

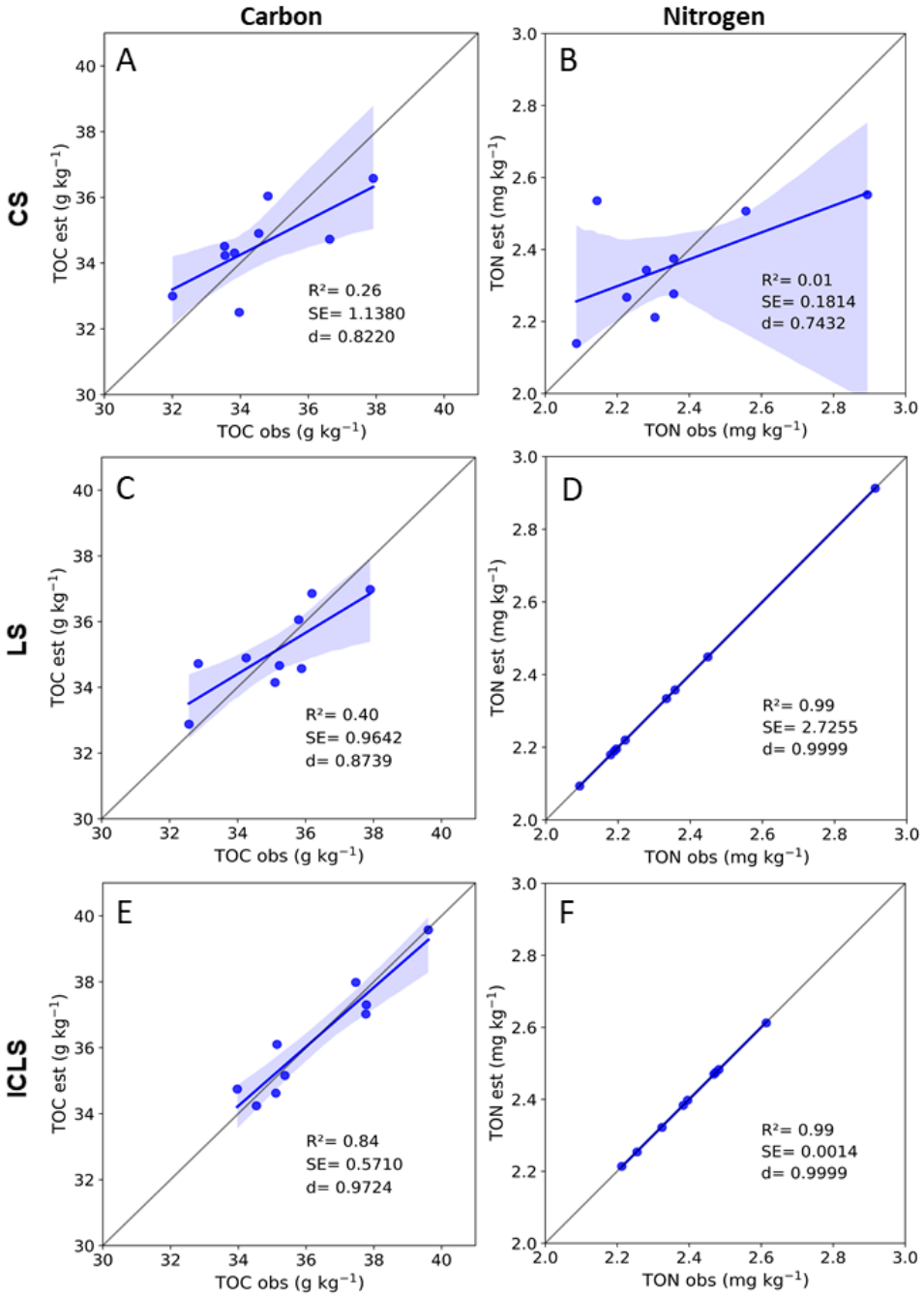


Figura 9. Desempenho do modelo Multiple Linear Regression (MLR) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

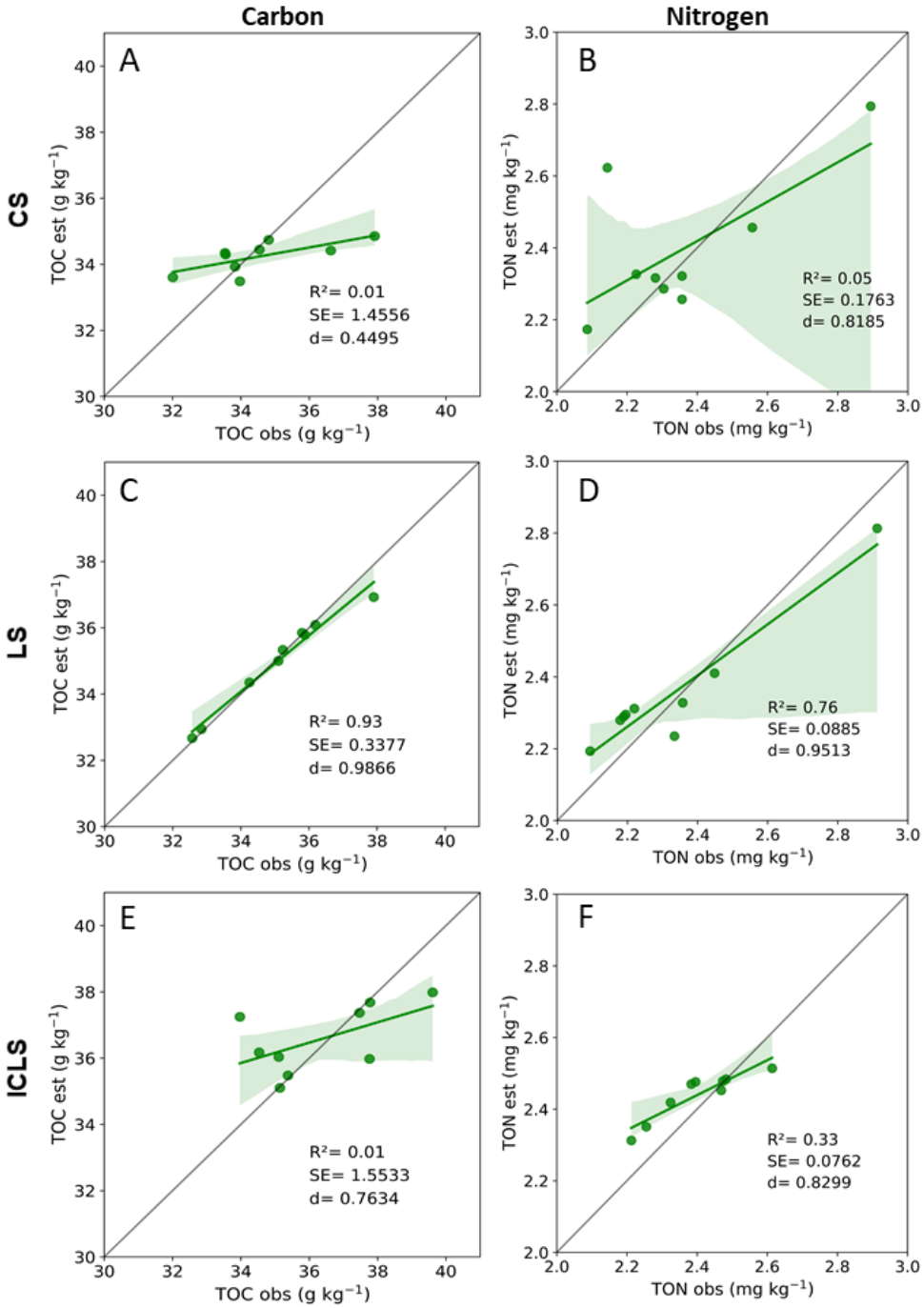


Figura 10. Desempenho do modelo Support Vector Regression (SVR) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

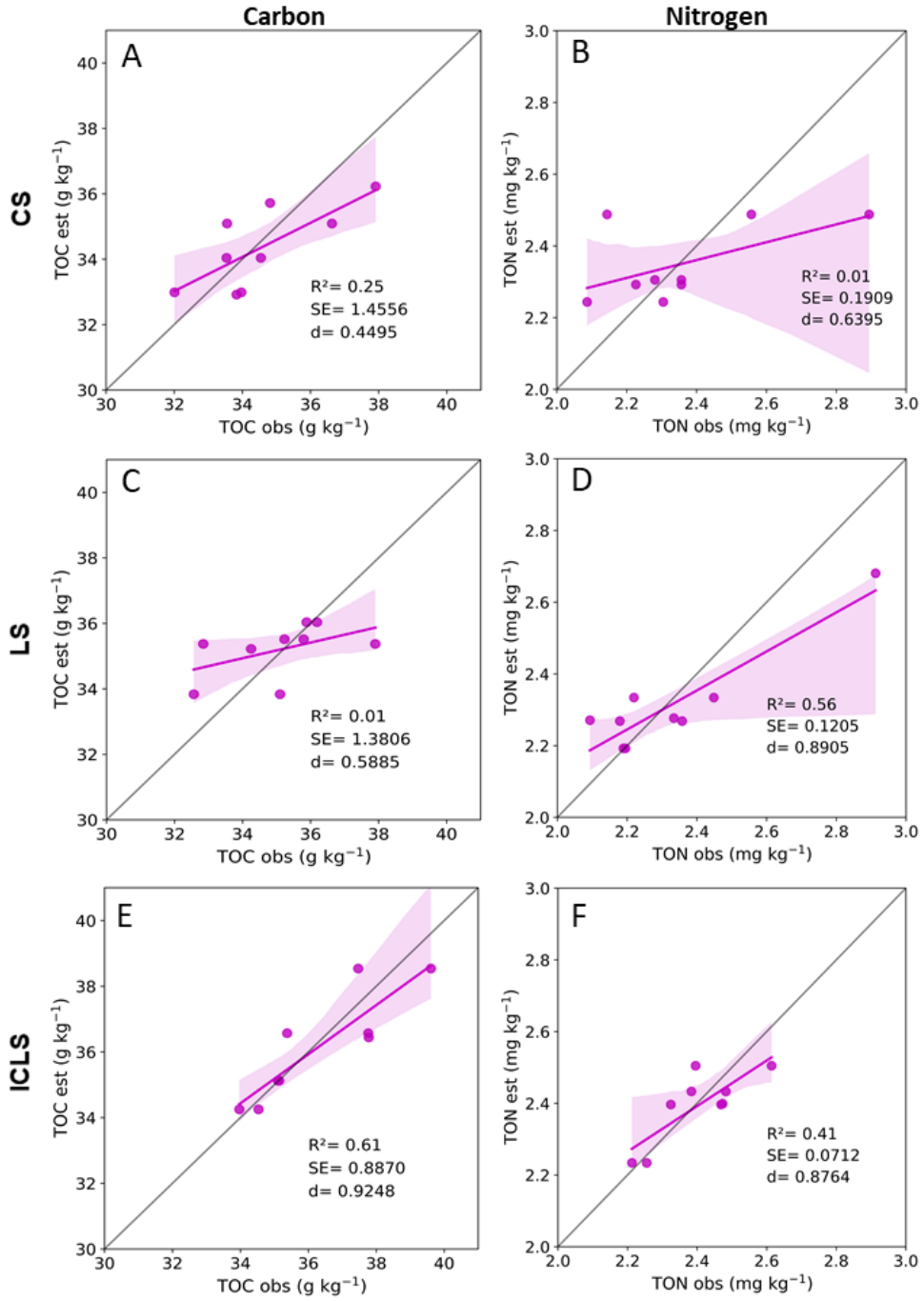


Figura 11. Desempenho do modelo K Neighbors Regressor (KNN) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

O algoritmo AdaBoost teve alto desempenho para a maioria dos tratamentos (Figura 12, Tabela 4), assim como as redes neurais artificiais (Figura 7, Tabela 4). Esse modelo demonstrou alta precisão, com um R^2 máximo de 0.99; alta acurácia, com um MAPE máximo de 0.8% e "d" máximo de 0.99; baixa tendência, com um valor de SE máximo de 1.79 g kg^{-1} para TOC e 0.23 mg kg^{-1} para TON. Essa versão simples do Boosting apresentou menor número de calibrações quando comparado com o MLP (Apêndice 4), assim como também apresentou os menores erros (MAE, RMSE) na estimativa de TOC e TON (Tabela 4).

Um estudo realizado por Pham et al. (2021) mostrou que o modelo Adaboost é menos propenso a erros para classificar solos, em relação a redes neurais artificias, se tornando mais assertivo e preciso. Mishra et al. (2020) em estudos desenvolvidos na Índia, usando diferentes modelos de regressão para prever a produção de arroz em épocas com sensibilidade climática, observaram que o Adaboost obteve os maiores desempenhos (R^2 = 0.98), enquanto o SVR obteve o menor desempenho (R^2 = 0.62), sendo o SVR denominado no estudo como um regressor fraco. Esses resultados corroboram com os achados nesse estudo, com explicação dos dados em média do SVR de 30% e Adaboost em torno de 99%. Entretanto, essa é a primeira vez que o modelo de Adaboost de regressão foi utilizado para estimar indicadores do solo, usado anteriormente apenas para classificar solos (Sirsat et al., 2017; Wei et al., 2020; Pham et al., 2021).

Nossos resultados indicam que as variáveis de plantas, solos e animais quando combinadas com as variáveis climáticas são boas estimadoras de TOC e TON. Entre os sistemas estudados, os sistemas ICLS possuem correlações mais densas, ao contrário dos monocultivos (CS e LS). Isso pode ser explicado devido os sistemas integrados serem mais ricos em aportes de resíduos em cada estação, além de maior aporte de resíduos vegetais durante a variação sazonal (Strickland et al., 2019), no entanto, a condição climática foi a mesma para todos os tratamentos (Figura 1). As variáveis no sistema ICLS: precipitação de chuva, excedente de água no solo, macroporosidade, N da liteira, N das fezes, N da urina, C/N da biomassa microbiana contribuíram para um aumento de RMSE de 2% na estimação de TOC utilizando redes neurais (MLP). Ao passo que, os teores de carbono orgânico, C/N do solo, fósforo, excedente de água no solo, C da liteira e da biomassa microbiana foram as variáveis

mais importantes para estimar o TON no sistema ICLS, e obtiveram RMSE menor que 1% utilizando modelos de MLP e Adaboost.

Além de indicadores do solo, plantas e animais, o clima é um dos cincos elementos básicos que afetam o processo de formação do solo, e seu impacto no carbono e nitrogênio orgânico do solo já foi reportado em outros estudos (Dash et al. 2019; Ma e Chang, 2019). Semelhante aos nossos resultados, Deng et al. (2018) reportaram que a precipitação é uma variável importante e que afeta diretamente a estimação de TOC no leste da china.

Para utilizar os modelos de maiores desempenhos no presente estudo, deve-se empregar as mesmas variáveis de entrada (Figura 6). Vale ressaltar que, embora não seja necessário realizar todo o procedimento de delineamento experimental no item 2.2.3 para aplicar estes modelos de estimativa dos teores de TOC e TON, a capacidade de generalização deve ser avaliada com histórico de amostras analisada em cada laboratório e região.

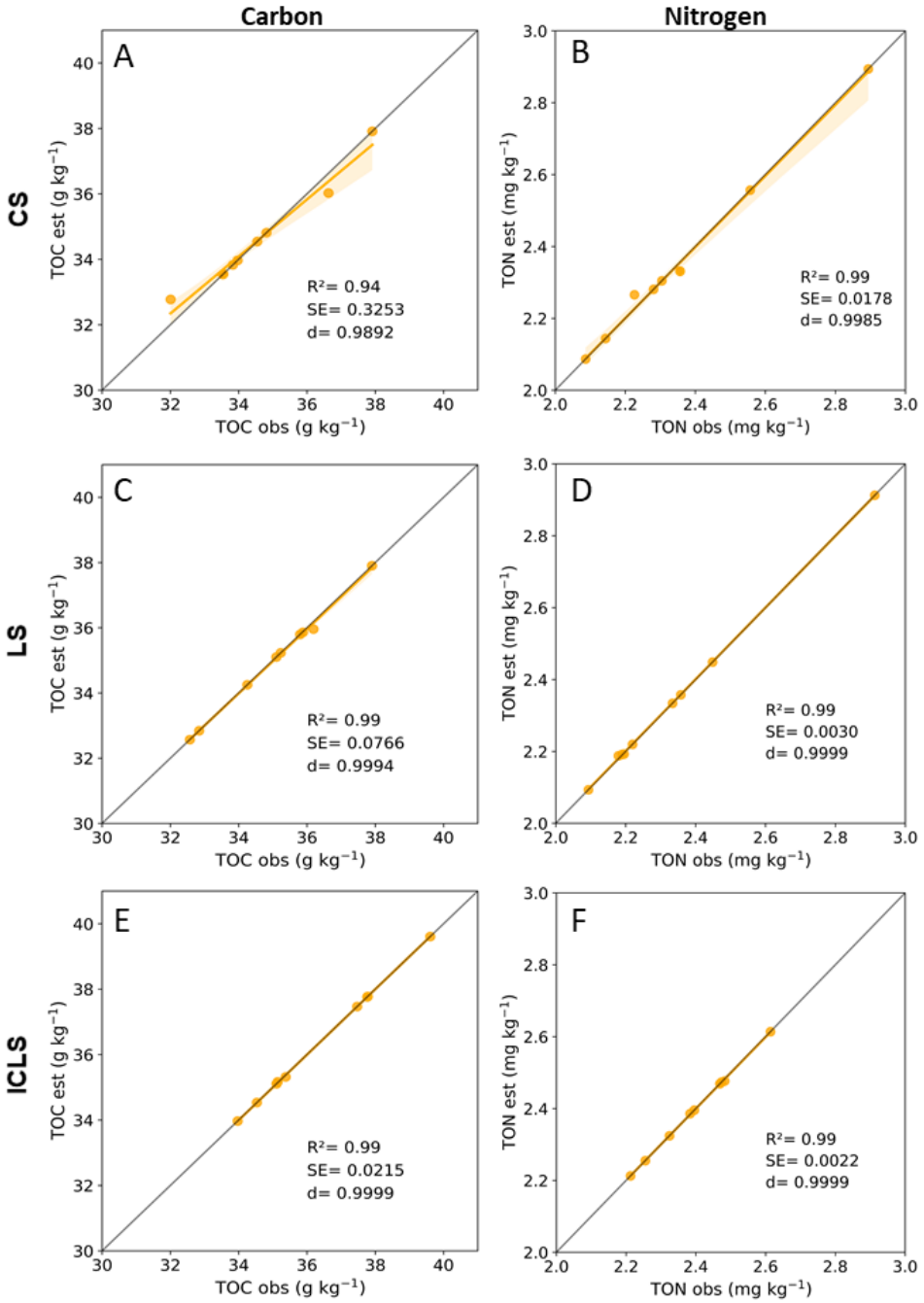


Figura 12. Desempenho do modelo Adaptive Boosting Regressor (AdaBoost) para a estimativa de carbono e nitrogênio orgânico total do solo nos tratamentos (**A**) TOC-CS, (**B**) TON-CS, (**C**) TOC-LS, (**D**) TON-LS, (**E**) TOC-ICLS, (**F**) TON-ICLS. Os pontos são as repetições do tratamento. As linhas completas correspondem a linhas 1:1. R^2: coeficiente de determinação ajustado, d: Índice de Willmott et al. (1985), SE: erro sistemático. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

2.3.3 Comparação entre os modelos de machine learning para estimar o teor de carbono e nitrogênio orgânico do solo

Para identificar os modelos com as maiores acurácias, usamos o diagrama de Taylor que oferece uma visão geral comparativa. Os modelos MLP e AdaBoost se aproximaram do ponto de referência, o que indicou alta precisão desses modelos em comparação aos demais em estudo. O resultado deste diagrama (Figura 13) mostrou que os dados estimados por esses algoritmos foram mais correlacionados com os dados reais do teor de TOC e TON no solo. Além disso, o índice RMSE (< 0.2) que avalia a acurácia, foram menores nos modelos MLP e AdaBoost.

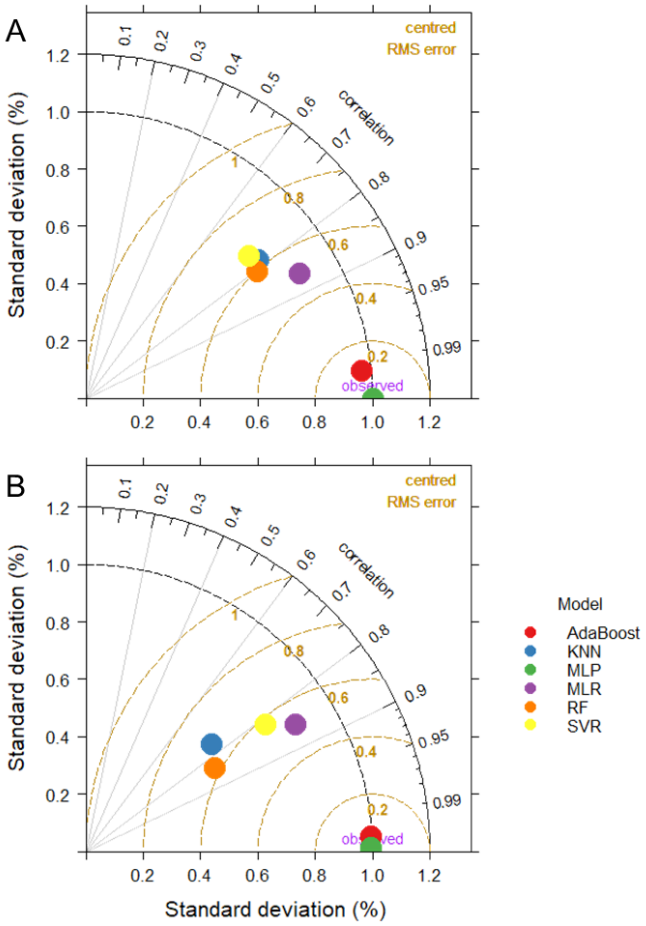


Figura 13. Diagrama de Taylor para avaliar a saída dos dados de teste de seis modelos na estimação dos teores de **A**) TOC e **B**) TON. Os ângulos referem-se à correlação coeficiente entre os dados observados e estimados. A distância radial da origem denota a razão entre o desvio padrão normalizado (SD) da simulação e o da observação. Se os pontos estimados estão mais próximos do ponto de referência, a precisão do modelo e a correlação entre os parâmetros de entrada são altas. Linhas pontilhadas amarelas indicam o RMSE. MLP: Multilayer Perceptron Regressor, RF:

Random Forest Regressor, MLR: Multiple Linear Regression, SVR: Support Vector Regression, KNN: K Neighbors Regressor, e AdaBoost: Adaptive Boosting Regressor.

Para estimar o teor de TOC, os algoritmos RF, SVR e KNN tiveram baixa dispersão entre si, baixa correlação, alto RMSE e maior distância do ponto de referência (Figura 13 A). Apesar disso, o MLR obteve um erro menor (RMSE< 0.6) para estimar o TOC (Figura 13 A), possivelmente causado pela alta precisão para estimar o teor de TOC no tratamento ICLS (R^2=0.84 e RMSE=0.571 g kg^{-1}, Figura 9 E). Para estimar o teor de TON, os algoritmos KNN e RF também tiveram baixa dispersão entre si, alto RMSE e maior distância do ponto de referência (Figura 13 B). Os algoritmos SVR e MLR podem ter se aproximado mais do ponto de referência devido à alta precisão do SVR para estimar o teor de TON no tratamento LS (R^2=0.76 e RMSE=0.0885 mg kg^{-1}, Figura 10 D). E o MLR, devido alcançar uma precisão de 99% para a estimação do teor de TON nos tratamentos LS e ICLS (Figura 9 D, F), apesar do erro de 2.72 mg kg^{-1} para estimar TON no tratamento LS.

Portanto, as concentrações de TOC e TON podem ser estimados com alta precisão usando conjuntamente variáveis de entrada baseados no histórico da área e análises de: solo, plantas, animais e dados climáticos. Isso demonstra uma alta aplicabilidade do presente estudo, e uma nova perspectiva para a aplicação de modelagem por aprendizado de máquinas para estimar importantes nutrientes do solo, e que demandam de análises dispendiosas.

Após uma ampla análise da literatura, não foram encontrados resultados científicos anteriores incluindo características climáticas e urina e fezes de animais combinadas para estimar o teor de carbono e nitrogênio orgânico do solo. A maioria dos estudos se concentram em usar imagens de satélites para prever carbono ou matéria orgânica do solo (Zhu et al. 2018b; Wei et al. 2020; Goydaragh et al. 2021; Zhou et al. 2021), ou usam isoladamente apenas indicadores de qualidade do solo (Wiesmeier et al. 2014; Priya e Ramesh, 2018; Fernandes et al. 2019; John et al. 2020), ou apenas covariáveis ambientais na estimação de TOC e TON (Zeraatpisheh et al., 2022). Além de encontrar novas associações e oportunidades para estimar TOC e TON neste estudo, foi possível identificar variáveis importantes que impactam na recarbonização do solo, que fazem parte de sistemas tão complexos, como o ICLS.

2.4. CONCLUSÕES

A combinação de variáveis de planta, solo, animas e clima contribuíram significativamente na estimativa de TOC e TON nos sistemas avaliados. A partir do estudo, foi identificado que o sistema ICLS possui mais densas correlações entre as variáveis avaliadas, e a seleção mais diversificadas de variáveis de plantas, solos, animais e climas contribuíram para que todos os modelos obtivessem uma alta acurácia na estimativa de TOC e TON do solo.

A estimativa dos teores de TOC e TON do solo usando os modelos de Adaboost e MLP revelou alta precisão (R^2=0.99) e mais próximos dos dados de laboratório (RMSE < 2%). Para cada sistema, os modelos com maior precisão e maior acurácia para estimar TOC e TON foram: no sistema monocultivos de milho (CS), os modelos AdaBoost e MLP obtiveram um MAPE menor que 1%, com alta precisão. No sistema monocultivos de Pecuária (LS), os modelos MLP, SVR e Adaboost foram mais precisos e acurados (MAPE < 3%). No sistema Integração Lavoura-Pecuária (ICLS), todos os modelos tiveram alta acurácia de modo geral (MAPE < 5%), no entanto, o modelo SVR obteve a menor precisão para estimar TOC (R^2 < 2%), e o SVR e KNN obtiveram a menor precisão para estimar o TON (R^2 < 50%).

2.5 REFERÊNCIAS

Ahmadalipour A, Moradkhani, H., Rana, A., 2018. Accounting for downscaling and model uncertainty in fine-resolution seasonal climate projections over the Columbia River Basin. **Climate dynamics** 50(1): 717-733. https://doi.org/10.1007/s00382-017-3639-4

Allen RG, Pereira LS, Raes D, Smith M (1998) **Crop evapotranspiration-Guidelines for computing crop water requirements-FAO**. Irrigation and drainage paper 56. Fao, Rome. 300(9), D05109.

Amelung W et al (2020) Towards a global-scale soil climate mitigation strategy. **Nature communications** 11(1): 1-10. https://doi.org/10.1038/s41467-020-18887-7

Balogh J, Pintér K, Fóti S, Cserhalmi D, Papp M, Nagy Z (2011) Dependence of soil respiration on soil moisture, clay content, soil organic matter, and CO^2 uptake in dry grasslands. **Soil Biology and Biochemistry** 43(5): 1006-1013. https://doi.org/10.1016/j.soilbio.2011.01.017

Biau G (2012) Analysis of a random forests model. Journal of Machine Learning. **Research** 13: 1063-1095.

Bieluczyk W, Pereira MG, Guareschi RF, Bonetti JA, Freó VA, Silva Neto EC (2017) Granulometric and oxidizable carbon fractions of soil organic matter in crop-livestock integration systems. **Semina: Ciências Agrárias** 38: 607-622. https://doi.org/10.5433/1679-0359.2017v38n2p607

Bongiorno G, Bünemann EK, Oguejiofor CU, Meier J, Gort G, Comans R, Mäder P, Brussaard L, Goede R (2019) Sensitivity of labile carbon fractions to tillage and organic matter management and their potential as comprehensive soil quality indicators across pedoclimatic conditions in Europe. **Ecological Indicators** 99: 38-50. https://doi.org/10.1016/j.ecolind.2018.12.008

Blair GJ, Lefroy RDB, Lisle L (1995) Soil carbon fractions based on their degree of oxidation, and the development of a carbon management index for agricultural systems. **Australian Journal of Agricultural Research** 46: 1459-1466. https://doi.org/10.1071/AR9951459

Bockheim JG, Gennadiyev AN, Hartemink AE, Brevik EC (2014) Soil-forming factors and Soil Taxonomy. **Geoderma** 231-237. https://doi.org/10.1016/j.geoderma.2014.02.016

Campbell BM, Hansen J, Rioux J, Stirling CM, Twomlow S (2018) Urgent action to combat climate change and its impacts (SDG 13): transforming agriculture and food systems. **Current opinion in environmental sustainability** 34: 13-20. https://doi.org/10.1016/j.cosust.2018.06.005

Chen J, Zhang H, Fan M, Chen F, Gao C (2021) Machine-learning-based prediction and key factor identification of the organic carbon in riverine floodplain soils with intensive agricultural practices. **Journal of Soils and Sediments** 21(8): 2896-2907. https://doi.org/10.1007/s11368-021-02987-y

Dash PK, Bhattacharyya P, Roy KS, Neogi S, Nayak AK (2019) Environmental constraints' sensitivity of soil organic carbon decomposition to temperature, management practices and climate change. **Ecological Indicators** 107: 105644. https://doi.org/10.1016/j.ecolind.2019.105644

Deng X, Chen X, Ma W, Ren Z, Zhang M, Grieneisen ML, Long W, Ni Z, Zhan Y, Lv X (2018) Baseline map of organic carbon stock in farmland topsoil in East China. **Agriculture, ecosystems & environment** 254: 213-223. https://doi.org/10.1016/j.agee.2017.11.022

Dubeux Jr JCB, Sollenberger L (2020) Nutrient cycling in grazed pastures. In.: Rouquette Jr, M., Aiken, G. (Eds.) **Management strategies for Sustainable Cattle production in Southern Pastures**. Academic, p. 59-75.

Duval ME, Galantini JA, Martínez JM, Limbozzi F (2018) Labile soil organic carbon for assessing soil quality: influence of management practices and edaphic conditions. **Catena** 171: 316-326. https://doi.org/10.1016/j.catena.2018.07.023

Emamgholizadeh S, Esmaeilbeiki F, Babak M, Zarehaghi D, Maroufpoor E, Rezaei H, (2018) Estimation of the organic carbon content by the pattern recognition method. **Communications in Soil Science and Plant Analysis** 49(17): 2143-2154. https://doi.org/10.1080/00103624.2018.1499750

Engelhardt IC, Niklaus PA, Bizouard F, Breuil MC, Rouard N, Deau F, Philippot L, Barnard RL (2021) Precipitation patterns and N availability alter plant-soil microbial C and N dynamics. **Plant and Soil** 466(1): 151-163. https://doi.org/10.1007/s11104-021-05015-7

EMBRAPA (1997) Manual de métodos de análise de solo. Rio de Janeiro: Centro Nacional de Pesquisa de Solos.

FAO (2019) **Recarbonization of global soils: a dynamic response to offset global emissions**. Disponível em: <https://www.fao.org/3/i7235en/I7235EN.pdf>. Acesso em: 28 set. 2022.

Fernandes MMH, Coelho AP, Fernandes C, Silva MF, Marta CCD (2019) Estimation of soil organic matter content by modeling with artificial neural networks. **Geoderma** 350: 46-51. https://doi.org/10.1016/j.geoderma.2019.04.044

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. **Pattern recognition letters** 31(14): 2225-2236. https://doi.org/10.1016/j.patrec.2010.03.014

Ghorbani MA, Deo RC, Yaseen ZM, Kashani M, Mohammadi B (2018) Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. **Theoretical and applied climatology** 133(3): 1119-1131. https://doi.org/10.1007/s00704-017-2244-0

Guo L, Ma Y, Cukic B, Singh H (2004) Robust prediction of fault-proneness by random forests. In: 15th international symposium on software reliability engineering. **IEEE** 417-428. https://doi.org/10.1109/ISSRE.2004.35

Goydaragh MG, Taghizadeh-Mehrjardi R, Jafarzadeh AA, Triantafilis J, Lado M (2021) Using environmental variables and Fourier Transform Infrared Spectroscopy to predict soil organic carbon. **Catena** 202: 105280. https://doi.org/10.1016/j.catena.2021.105280

Grimm R, Behrens T, Märker M, Elsenbeer H (2008) Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. **Geoderma** 146(1-2): 102-113. https://doi.org/10.1016/j.geoderma.2008.05.008

Grell M, Barandun G, Asfour T, Kasimatis M, Collins ASP, Wang J, Güder F (2021) Point-of-use sensors and machine learning enable low-cost determination of soil nitrogen. **Nature Food** 2(12): 981-989. https://doi.org/10.1038/s43016-021-00416-4

Haynes RJ, Williams PH (1993) Nutrient cycling and soil fertility in the grazed pasture ecosystem. **Advances in agronomy** 49: 119-199. https://doi.org/10.1016/S0065-2113(08)60794-4

Heidari E, Sobati MA, Movahedirad S (2016) Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN). **Chemometrics and intelligent laboratory systems** 155: 73-85. https://doi.org/10.1016/j.chemolab.2016.03.031

Jansson JK, Hofmockel KS (2020) Soil microbiomes and climate change. **Nature Reviews Microbiology** 18(1): 35-46. https://doi.org/10.1038/s41579-019-0265-7

Jadhav SD, Channe HP (2016) Comparative study of K-NN, naive Bayes and decision tree classification techniques. **International Journal of Science and Research** 5(1): 1842-1845.

John K, Abraham Isong I, Kebonye, NM, Ayito EO, Agyeman PC, Afu SM (2020) Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. **Land** 9(12): 487. https://doi.org/10.3390/land9120487

Kabiri V, Raiesi F, Ghazavi MA (2016) Tillage effects on soil microbial biomass, SOM mineralization and enzyme activity in a semi-arid Calcixerepts. **Agriculture, Ecosystems & Environment** 232: 73-84. https://doi.org/10.1016/j.agee.2016.07.022

Lemaire G, Franzluebbers A, Carvalho PCF, Dedieu B (2014) Integrated crop–livestock systems: Strategies to achieve synergy between agricultural production and environmental quality. **Agriculture, Ecosystems & Environment** 190: 4-8. https://doi.org/10.1016/j.agee.2013.08.009

Li QQ, Yue TX, Wang CQ, Zhang WJ, Yu Y, Li B, Yang J, Bai GC (2013) Spatially distributed modeling of soil organic matter across China: An application of artificial neural network approach. **Catena** 104: 210-218. https://doi.org/10.1016/j.catena.2012.11.012

Liebig MA, Ryschawy J, Kronberg SL, Archer DW, Scholljegerdes EJ, Hendrickson JR, Tanaka DL (2017) Integrated crop-livestock system effects on soil N, P, and pH in a

semiarid region. **Geoderma** 289: 178-184. https://doi.org/10.1016/j.geoderma.2016.11.036

Lin L, Gao Z, Liu X (2020) Estimation of soil total nitrogen using the synthetic color learning machine (SCLM) method and hyperspectral data. **Geoderma** 380: 114664. https://doi.org/10.1016/j.geoderma.2020.114664

Iqbal A, Aslam S, Alavoine G, Benoit P, Garnier P, Recous S (2015) Rain regime and soil type affect the C and N dynamics in soil columns that are covered with mixed-species mulches. **Plant and soil** 393(1): 319-334. https://doi.org/10.1007/s11104-015-2501-x

Ma M, Chang R (2019) Temperature drive the altitudinal change in soil carbon and nitrogen of montane forests: Implication for global warming. **Catena** 182: 104126. https://doi.org/10.1016/j.catena.2019.104126

McGuire R, Williams PN, Smith P, McGrath SP, Curry D, Donnison I, Emmet B, Scollan N (2022) Potential Co-benefits and trade-offs between improved soil management, climate change mitigation and agri-food productivity. **Food and Energy Security**. https://doi.org/10.1002/fes3.352

Maia NJC, Cruz MCPD, Dubeux Junior JCB, Menegatto LS, Augusto JG, Mendonça GG, Terçariol MC, Oliveira JG, Simili FF (2021) Integrated crop-livestock versus conventional systems: use of soil indicators to detect short-term changes during seasonal variation. **Bragantia** 80. https://doi.org/10.1590/1678-4499.20210127

Mahmoud AA, Elkatatny S, Mahmoud M, Abouelresh M, Abdulraheem A, Ali A (2017) Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. **International Journal of Coal Geology** 179: 72-80. https://doi.org/10.1016/j.coal.2017.05.012

Mahmoud AA, Elkatatny S, Ali A, Abdulraheem A, Abouelresh M (2020) Estimation of the total organic carbon using functional neural networks and support vector machine. In: **International Petroleum Technology Conference**. https://doi.org/10.2523/IPTC-19659-MS

Mendonça GG, Simili FF, Augusto JG, Bonacim PM, Menegatto LS, Gameiro AH, (2020) Economic gains from crop-livestock integration in relation to conventional systems. **Revista Brasileira de Zootecnia** 49: 1-11. https://doi.org/10.37496/rbz4920190029

Miller AJ, Amundson R, Burke IC, Yonker C (2004) The effect of climate and cultivation on soil organic C and N. **Biogeochemistry** 67(1): 57-72. https://doi.org/10.1023/B:BIOG.0000015302.16640.a5

Miloud-Aouidate A, Baba-Ali AR (2012) A hybrid KNN-ant colony optimization algorithm for prototype selection. In: **International Conference on Neural Information Processing** 307-314. https://doi.org/10.1007/978-3-642-34487-9_38

Mishra S, Mishra D, Santra GH (2020) Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: An empirical assessment. **Journal of King Saud University-Computer and Information Sciences** 32(8): 949-964. https://doi.org/10.1016/j.jksuci.2017.12.004

Monz CA, Reuss DE, Elliott ET (1991) Soil microbial biomass carbon and nitrogen estimates using 2450 MHz microwave irradiation or chloroform fumigation followed by direct extraction. **Agriculture, Ecosystems & Environment** 34: 55-63. https://doi.org/10.1016/0167-8809(91)90093-D

Niu S, Classen AT, Dukes JS, Kardol P, Liu L, Luo Y, Rustad L, Sun J, Tang J, Templer PH, Thomas RQ, Tian D, Vicca S, Wang Y, Xia J, Zaehle JS (2016) Global patterns and substrate-based mechanisms of the terrestrial nitrogen cycle. **Ecology Letters** 19: 697-709. https://doi.org/10.1111/ele.12591

Oliveira WRD, Ramos MLG, Carvalho AMD, Coser TR, Silva AMM, Lacerda MM, Souza KW, Marchão RL, Vilela L, Pulrolnik K (2016) Dynamics of soil microbiological attributes under integrated production systems, continuous pasture, and native cerrado. **Pesquisa Agropecuária Brasileira** 51: 1501-1510. https://doi.org/10.1590/S0100-204X2016000900049

Oliveira JG, Santana Jr ML, Maia NJC, Dubeux Junior JCB, Gameiro AH, Kunrath TR, Mendonça GG, Simili FF (2022) Nitrogen balance and efficiency as indicators for monitoring the proper use of fertilizers in agricultural and livestock systems. **Scientific Reports** 12(1): 1-10. https://doi.org/10.1038/s41598-022-15615-7

Omer M, Idowu OJ, Ulery AL, VanLeeuwen D, Guldan SJ (2018) Seasonal changes of soil quality indicators in selected arid cropping systems. **Agriculture** 8: 124. https://doi.org/10.3390/agriculture8080124

Ontañon OM, Fernandez M, Agostini E, González OS (2018) Identification of the main mechanisms involved in the tolerance and bioremediation of Cr (VI) by Bacillus sp. SFC 500-1E. **Environmental Science and Pollution Research** 25(16): 16111-16120. https://doi.org/10.1007/s11356-018-1764-1

Pereira LA, Afonso LC, Papa JP, Vale ZA, Ramos CC, Gastaldello DS, Souza AN (2013) Multilayer perceptron neural networks training through charged system search and its application for non-technical losses detection. In: Conference on Innovative Smart Grid Technologies (ISGT Latin America). **IEEE** 1-6. https://doi.org/10.1109/ISGT-LA.2013.6554383

Pham BT, Nguyen MD, Nguyen-Thoi T, Ho LS, Koopialipoor M, Quoc NK, Armaghani DJ, Van Le H (2021) A novel approach for classification of soils based on laboratory tests using Adaboost, Tree and ANN modeling. **Transportation Geotechnics** 27: 100508. https://doi.org/10.1016/j.trgeo.2020.100508

Poffenbarger HJ, Sawyer JE, Barker DW, Olk DC, Six J, Castellano MJ (2018) Legacy effects of long-term nitrogen fertilizer application on the fate of nitrogen fertilizer inputs in continuous maize. **Agriculture, ecosystems & environment** 265: 544-555. https://doi.org/10.1016/j.agee.2018.07.005

Priya R, Ramesh D (2018) Adaboost.RT based soil npk prediction model for soil and crop specific data: A predictive modelling approach. In: **International Conference on Big Data Analytics** 322-331. https://doi.org/10.1007/978-3-030-04780-1_22

Raij BV, Andrade JC, Cantarella H, Quaggio JA (Eds.) (2001) Análise química para avaliação da fertilidade de solos tropicais. Campinas: Instituto Agronômico, 235p.

Rana A, Rawat AS, Bijalwan A, Bahuguna H (2018) Application of multilayer (Perceptron) artificial neural network in the diagnosis system: a systematic review. **IEEE** 1-6. https://doi.org/10.1109/RICE.2018.8509069

Ramírez PB, Fuentes-Alburquenque S, Díez B, Vargas I, Bonilla CA (2020) Respostas da comunidade microbiana do solo a frações lábeis de carbono orgânico em relação ao tipo de solo e uso da terra ao longo de um gradiente climático. **Soil Biology and Biochemistry** 141: 107692. https://doi.org/10.1016/j.soilbio.2019.107692

Rakkar MK, Blanco-Canqui H, Drijber RA, Drewnoski ME, MacDonald JC, Klopfenstein T (2017) Impacts of cattle grazing of corn residues on soil properties after 16 years. **Soil Science Society of America Journal** 81: 414-424. https://doi.org/10.2136/sssaj2016.07.0227

Reda R, Saffaj T, Ilham B, Saidi O, Issam K, Brahim L (2019) A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. **Chemometrics and Intelligent Laboratory Systems**. 195: 103873. https://doi.org/10.1016/j.chemolab.2019.103873

Rezende CP, Cantarutti RB, Braga JM, Gomide JA, Pereira JM, Ferreira E, Tarré RM, Macedo R, Alves BJR, Urquiaga S, Cadisch G, Giller KE, Boddey RM (1999) Litter deposition and disappearance in Brachiaria pastures in the Atlantic Forest region of the south of Bahia, Brazil. **Nutrient Cycling in Agroecosystems** 54: 99-112.

Roberts TL, Norman RJ, Slaton NA, Wilson CE, Ross WJ, Bushong JT (2009) Direct steam distillation as an alternative to the Illinois soil nitrogen test. **Soil Science Society of America Journal** 73: 1268-1275. https://doi.org/10.2136/sssaj2008.0165

Rolim GS, Aparecido LEO (2015) Camargo, Köppen and Thornthwaite climate classification systems in defining climatical regions of the state of São Paulo, Brazil. **International Journal of Climatology** 36(2): 636-643. https://doi.org/10.1002/joc.4372

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumbreras JF, Coelho MR, Almeida JA, Araujo Filho JC, Oliveira JB, Cunha TJF (Eds.) (2018) Sistema brasileiro de classificação de solos. Brasília: Embrapa.

Silva EE, Azevedo PHS, De-Polli H (2007) Determinação do Nitrogênio da Biomassa Microbiana do Solo (BMS-N). Seropédica – RJ: EMBRAPA (EMBRAPA. **Comunicado técnico, 96).**

Sirsat MS, Cernadas E, Fernández-Delgado M, Khan R (2017) Classification of agricultural soil parameters in India. **Computers and electronics in agriculture** 135: 269-279. https://doi.org/10.1016/j.compag.2017.01.019

Shen F, Wu J, Fan H, Liu W, Guo X, Duan H, Hu L, Lei X, Wei X (2019) Soil N/P and C/P ratio regulate the responses of soil microbial community composition and enzyme activities in a long-term nitrogen loaded Chinese fir forest. **Plant and Soil** 436(1): 91-107. https://doi.org/10.1007/s11104-018-03912-y

Sun S, Liu J, Chang SX (2013) Temperature sensitivity of soil carbon and nitrogen mineralization: impacts of nitrogen species and land use type. **Plant and Soil** 372(1): 597-608. https://doi.org/10.1007/s11104-013-1758-1

Stackhouse Jr PW, Whitlock CH, DiPasquale RC, Brown DE, Chandler WS (2002) Meeting energy-sector needs with NASA climate datasets. **Earth Observation Magazine** 11(8): 6-10.

Strickland MS, Thomason WE, Avera B, Franklin J, Minick K, Yamada S, Badgley BD (2019) Short-Term effects of cover crops on soil microbial characteristics and biogeochemical processes across actively managed farms. **Agrosystems, Geosciences & Environment** 2: 1-9. https://doi.org/10.2134/age2018.12.0064

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. **Journal of Geophysical Research: Atmospheres** 106(7): 7183-7192. https://doi.org/10.1029/2000JD900719

Teutscherová N, Vázquez E, Sotelo M, Villegas D, Velásquez N, Baquero D, Pulleman M, Arango J (2021) Intensive short-duration rotational grazing is associated with improved soil quality within one year after establishment in Colombia. **Applied Soil Ecology** 159: 103835. https://doi.org/10.1016/j.apsoil.2020.103835

Veldkamp E (1994) Organic Carbon Turnover in Three Tropical Soils under Pasture after Deforestation. **Soil Science Society of America Journal** 58: 175-180.

Vilela L, Martha Junior GB, Macedo MCM, Marchão RL, Guimarães Júnior R, Pulrolnik K, Maciel GA (2011) Sistemas de integração lavoura-pecuária na região do cerrado. **Pesquisa Agropecuária Brasileira** 46: 1127–1138.

Xu S, Wang M, Shi X, Yu Q, Zhang Z (2021) Integrating hyperspectral imaging with machine learning techniques for the high-resolution mapping of soil nitrogen fractions in soil profiles. **Science of The Total Environment** 754: 142135. https://doi.org/10.1016/j.scitotenv.2020.142135

Wang P, Chen Z, Pang X, Hu K, Sun M, Chen X (2016) Revised models for determining TOC in shale play: Example from Devonian Duvernay shale, Western Canada sedimentary basin. **Marine and Petroleum Geology** 70: 304-319. https://doi.org/10.1016/j.marpetgeo.2015.11.023

Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic soil titration method. **Soil Science** 37: -38.

Wei L, Yuan Z, Wang Z, Zhao L, Zhang Y, Lu X, Cao L (2020) Hyperspectral Inversion of Soil Organic Matter Content Based on a Combined Spectral Index Model. **Sensors** 20(10): 2777. https://doi.org/10.3390/s20102777

Weil RR, Islam KR, Stine MA, Gruver JB, Samson-Liebig SE (2003) Estimating active carbon for soil quality assessment: a simplified method for laboratory and field use. **America Journal of Alternative Agriculture** 18: 3-17. https://doi.org/10.1079/AJAA200228

Wiesmeier M, Barthold F, Spörlein P, Geuß U, Hangen E, Reischl A, Schilling B, Angst G, Lutzow MV, Kögel-Knabner I (2014) Estimation of total organic carbon storage and its driving factors in soils of Bavaria (southeast Germany). **Geoderma Regional** 1: 67-78. https://doi.org/10.1016/j.geodrs.2014.09.001

Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. **Ecological Indicators** 52: 394-403. https://doi.org/10.1016/j.ecolind.2014.12.028

Willmott CJ (1981) On the validation of models. **Physical geography** 2(2): 184-194.

Wu YH, Lin JC, Wang TY, Lin TJ, Yen MC, Liu YH, Wu P, Chen F, Shih Y, Yeh I (2019) Hexavalent chromium intoxication induces intrinsic and extrinsic apoptosis in human renal cells. **Molecular medicine reports** 21(2): 851-857. https://doi.org/10.3892/mmr.2019.10885

Wuest S (2014) Seasonal variation in soil organic carbon. **Soil Science Society of America Journal** 78: 1442-1447. https://doi.org/10.2136/sssaj2013.10.0447

Zeraatpisheh M, Garosi Y, Owliaie HR, Ayoubi S, Taghizadeh-Mehrjardi R, Scholten T, Xu M (2022) Improving the spatial prediction of soil organic carbon using environmental covariates selection: A comparison of a group of environmental covariates. **Catena** 208: 105723. https://doi.org/10.1016/j.catena.2021.105723

Zhang Y, Li M, Zheng L, Qin Q, Lee WS (2019) Spectral features extraction for estimation of soil total nitrogen content based on modified ant colony optimization algorithm. **Geoderma** 333: 23-34. https://doi.org/10.1016/j.geoderma.2018.07.004

Zhu Q, Castellano MJ, Yang G (2018a) Coupling soil water processes and nitrogen cycle across spatial scales: Potentials, bottlenecks and solutions. **Earth-Science Reviews** 187: 248-258. https://doi.org/10.1016/j.earscirev.2018.10.005

Zhu J, Wu W, Liu HB (2018b) Environmental variables controlling soil organic carbon in top-and sub-soils in karst region of southwestern China. **Ecological indicators** 90: 624-632. https://doi.org/10.1016/j.ecolind.2018.03.073

Zhou T, Geng Y, Ji C, Xu X, Wang H, Pan J, Bumberger J, Haase D, Lausch A (2021) Prediction of soil organic carbon and the C: N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. **Science of The Total Environment** 755: 142661. https://doi.org/10.1016/j.scitotenv.2020.142661

CAPÍTULO 3 - Disponibilidade de nitrogênio potencialmente mineralizável no solo: uma abordagem interpretável de “machine learning”

RESUMO – O nitrogênio mineral (N) é o principal nutriente limitante para o aumento da produtividade das culturas, impactando diretamente na fertilidade do solo e potencial para emissões de N_2O, um importante gás do efeito estufa. O monitoramento de N prontamente disponível no solo (PMN) contribui muito para diminuir a demanda de fertilizantes nitrogenados e impactos nas mudanças climáticas globais. Modelos de machine learning têm demonstrado alta capacidade em simular componentes em sistemas complexos, assim, objetivou-se investigar as variáveis que podem impactar na disponibilidade de PMN, a fim de entender relações complexas de sistemas de integração lavoura-pecuária. Os modelos de machine learning testados foram Multiple Linear Regression (MLR), Random Forest Regressor (RF), Adaptive Boosting Regressor (AdaBoost) e eXtreme Gradient Boosting Regressor (XGBoost). Foram analisadas variáveis do solo, plantas e animais durante um período de dois anos em solos sob sistemas de integração lavoura-pecuária, no Brasil. Os modelos foram ajustados após selecionar cinco variáveis mais importantes. Esta seleção foi realizada por meio da correlação de Spearman e pelo modelo de Random Forest. Para estimar o PMN o modelo XGBoost foi o mais preciso, com o menor erro e viés (R^2 = 0.97, MAPE = 3% e MBE = 0.10 mg kg^{-1}), superando o AdaBoost, RF e MLR, nessa ordem. O método de explicações aditivas de Shapley (SHAP) foi usado para fornecer interpretações globais e locais para os modelos de machine learning. Os resultados demonstraram que as variáveis mais importantes que impactam no PMN do solo foram o carbono lábil, carbono da liteira, porosidade total do solo, capacidade de troca de cátions e estoque de nitrogênio no solo, nessa ordem de importância. Em adição, a diminuição de carbono lábil no solo impactou positivamente na liberação de PMN a curto prazo. Os solos de sistemas integrados liberam mais carbono oriundo dos restos vegetais (carbono da liteira); e o balanceamento entre as estruturas físicas (porosidade total) e de fertilidade (troca catiônica) influenciaram na liberação de PMN do solo. E por fim, observou-se que o estoque de nitrogênio no solo resulta em menor probabilidade de estimar PMN do solo. Um ponto favorável para a estimação de PMN, é que as variáveis de entrada usadas podem ser comumente determinadas em análises de rotinas de solo de áreas agrícolas, portanto, não há necessidade de dispositivos e medições adicionais para a estimação. Este estudo fornece variáveis do solo importantes para uma modelagem mais eficaz da disponibilidade de nitrogênio potencialmente mineralizável do solo, como estratégia de monitoramento do processo de ciclagem desse importante elemento no solo sob sistema de integração lavoura-pecuária.

“Palavras-chave:” Nitrogênio do solo, Indicadores de qualidade do solo, SHAP, XGBoost, inteligência artificial, agricultura 4.0.

3.1 INTRODUÇÃO

Existe uma demanda exponencial por produtos agrícolas à medida que a população mundial cresce (Kastner et al., 2021). O nitrogênio (N) é o principal nutriente limitante da produtividade em sistemas de cultivos (Rütting et al., 2018), e a aplicação de fertilizantes químicos nitrogenados é considerada uma das principais formas de melhorar a produtividade das culturas (Sinclair e Rufty, 2012; Linquist et al., 2013; Abalos et al., 2014). No entanto, a fertilização com N desempenha um papel fundamental na poluição do ar (via emissões de amônia) e nas mudanças climáticas, devido as emissões de N_2O, um dos gases do efeito estufa com maior potencial de aquecimento global (Xuejun et al., 2011; Thakur et al., 2013; Smith et al., 2015; Chatzisymeon et al., 2017). De modo geral, estima-se que as plantas utilizam menos de 50% do fertilizante aplicado, e grande parte deste é perdido no sistema solo-planta-atmosfera. Nesse contexto, o emprego de estratégias para aumentar a eficiência do uso de N fertilizante assumirá mais importância do que a quantidade de N aplicada, tornando-se uma prática fundamental no manejo de solos agrícolas (Abbruzzini et al., 2019). A agricultura é altamente dependente das condições climáticas e, portanto, está sujeita a mudanças e variabilidades, com impactos na segurança alimentar. Além disso, outros cenários podem impactar na demanda desses produtos, como é o caso da guerra entre a Rússia e Ucrânia. A Rússia é o sexto maior exportador mundial de fertilizantes nitrogenados e gás natural, a principal matéria prima para fertilizantes nitrogenados (Hassen et al., 2022). De fato, alguns meses após o início da guerra, as consequências foram claras: as exportações ucranianas e russas cessaram, as colheitas futuras são questionáveis e os preços globais das commodities agrícolas dispararam, ameaçando empurrar milhões para a fome e a pobreza (Hassen et al., 2022; Hebebrand e Laborde, 2022).

Diante desse contexto, a adoção de sistemas sustentáveis para a produção de alimentos é essencial, para que os nutrientes sejam reciclados no solo, diminuindo a aplicação de fertilizantes nitrogenados e os riscos ambientais. Os sistemas integrados lavoura-pecuária (ICLS) atendem a esta característica de ciclagem de nutrientes (Oliveira et al., 2022), trazem como benefícios uma rotatividade de diferentes produtos dentro de uma mesma área (produção de grãos ou pecuária) (Carvalho et al., 2018), e aumentam significativamente a qualidade do solo (Galindo et al., 2020; Maia et al.,

2021; Bansal et al., 2022), além de proporcionar ganhos econômicos (Mendonça et al., 2020).

Para avaliar os processos de ciclagem de nutrientes em ICLS, a qualidade do solo é medida por meio de indicadores importantes, como o nitrogênio potencialmente mineralizável do solo (PMN). O PMN é definido como o N orgânico que pode ser mineralizado durante o crescimento das plantas (Stevenson e Braids, 1968). Logo, o PMN tem sido usado para prever a liberação de N dos solos, pois é capaz de identificar práticas de manejo que aumentam as taxas de ciclagem de N do solo em curto prazo (Maia et al., 2021), a diminuição de demanda de fertilizantes nitrogenados e saúde dos solos (Franzluebbers, 2016; Ghimire et al. 2019).

Visando otimizar custos e processos para a obtenção de indicadores do solo, vários estudos se baseiam na utilização de métodos de “machine learning” (ML, aprendizado de máquina) para estimar propriedades do solo de forma indireta (Reda et al., 2019; John et al., 2020; Lin et al., 2020). Há uma crescente diversidade e disponibilidade de dados dentro de laboratórios e fazendas que podem ser usados para estimar esses indicadores de solo, devido ao volume e a natureza complexa das interações entre as variáveis, tem havido uma tendência de uso de algoritmos para estimar matéria orgânica do solo (Mahmoud et al. 2020; Wei et al., 2020) e nitrogênio do solo (Lin et al., 2020; Grell et al., 2021). No entanto, foi encontrado uma única aplicação de machine learning para o PMN do solo, usando regressão linear múltipla (Osterholz et al., 2017).

Embora os algoritmos de ML possam ser robustos e acurados para estimar propriedades importantes, a interpretabilidade dos modelos continua sendo um desafio devido à dificuldade para discernir como uma estimação ou previsão foi derivada. Apesar de alguns modelos já tenham algumas funcionalidades embutidas para avaliar o nível de importância das variáveis, como é o caso do Random Forest, esses métodos são bastante generalistas e avaliam a importância das variáveis de forma global (Strobl et al., 2008; Filippi et al., 2020). Além disso, nenhum insight é fornecido sobre o que está impulsionando a estimação para um determinado ponto de observação (Jones et al., 2022).

Apesar de encontrar muitas pesquisas usando ML para estimar indicadores de qualidade do solo, poucos são direcionados para interpretar os resultados do modelo,

ou seja, não está claro quais fatores desempenham um papel fundamental nos indicadores de qualidade do solo. No entanto, o surgimento de abordagens de machine learning interpretável parece ter revertido essa situação. O método SHapley Additive exPlanations (SHAP) é uma oportunidade para superar essas limitações, e fornecer uma interpretação mais alinhada ao que realmente pode acontecer, além de, também considerar possíveis efeitos sinérgicos entre as variáveis estudadas (Lundberg e Lee, 2017; Lundberg et al., 2020). Essas vantagens permitiram que o SHAP fosse um importante interpretador de modelos, e ele já foi usado em alguns estudos de qualidade do solo (Pathy et al., 2020; Yang et al., 2021; Zhao et al., 2022) com sucesso.

Diante desse panorama, com o presente estudo, de forma geral, objetivou-se investigar as variáveis que podem impactar na disponibilidade de nitrogênio potencialmente mineralizável do solo, a fim de entender relações complexas de sistemas de integração lavoura-pecuária. Os objetivos específicos foram (1) identificar as variáveis preditoras que mais influenciam o PMN do solo, (2) calibrar um modelo de machine learning que seja acurado para estimar o PMN, ao mesmo tempo que, use o mínimo possível de ajustes, (3) utilizar o SHAP para interpretar os modelos mais acurados, e explicar as características de cada variável que influenciam no PMN do solo a curto prazo.

3.2 MATERIAL E MÉTODOS

3.2.1 Local do experimento e Delineamento experimental

O experimento foi conduzido entre novembro de 2015 e janeiro de 2018 no Centro de Pesquisa de Bovinos de Corte, Instituto de Zootecnia/APTA/SAA, Sertãozinho, São Paulo, Brasil (21°8'16” S e 47°59'25” W), com altitude média de 548 m. O clima da região segundo a classificação de Köppen é Aw (Rolim e Aparecido, 2015), caracterizado como tropical úmido com estação chuvosa no verão e seca no inverno. O solo da área experimental é classificado como Latossolo Vermelho distrófico argiloso (Santos et al., 2018), equivalente a um Latossolo (Oxisol), de acordo com o sistema de classificação de solos do USDA (Bockheim et al., 2014).

O experimento foi conduzido em campo experimental de 16.02 ha, o sistema de manejo consistia na técnica de semeadura consorciada entre milho (*Zea mays* L.) e capim-marandu (*Urochloa brizantha* (Hoechst. Ex A. Rich.) R. D. Webster cultivar marandu), para o estabelecimento do Sistema Integrado Lavoura-Pecuária (ICLS). A semeadura foi realizada em dezembro de 2015, utilizando semeadora de plantio direto com cinco linhas. O manejo do experimento foi realizado de acordo com Maia et al. (2021).

O pastejo foi realizado por bovinos de corte da raça Caracu que estavam em fase de recria (média de 14 meses de idade), e a taxa de lotação foi feita segunda a oferta de forragem disponível. Os animais permaneceram em regime de lotação contínua até dezembro de 2017. Foram realizados dois ciclos de lotação contínuos dos animais: o primeiro ciclo foi entre agosto e outubro de 2016 (78 dias), e o segundo ciclo entre novembro de 2016 e dezembro de 2017 (370 dias).

3.2.2 Coleta de dados de solo, plantas e animais

As variáveis de entrada incluíram: atributos do solo, planta e animais. Os atributos de fertilidade e física do solo foram coletados no início do experimento, novembro de 2015, com profundidade variando entre 10, 40 e 90 cm dependendo do indicador de solo avaliado e metodologia, e durante o decorrer do experimento foram coletadas as informações de plantas e animais. Os atributos microbiológicos e de labilidade de carbono e nitrogênio do solo foram coletados durante março de 2017 a janeiro de 2018 (três amostragens durante esse período). Todas essas variáveis foram analisadas em laboratório seguindo o padrão internacional (Tabela 1). Nessa pesquisa foi utilizado vinte variáveis de solo, quatro variáveis de plantas, três variáveis de animais, totalizando um conjunto total de vinte e sete variáveis de entrada.

Tabela 1. Variáveis de solo, planta e animal para a estimação do PMN do solo. Médias das variáveis estão detalhadas nos Apêndices 1 e 2.

Descrição (unidade)	Símbolos	Referências
	Planta	
Liteira depositada (g m^{-2})	LD	
Carbono e Nitrogênio da liteira (g kg^{-1})	LIC, LIN	Rezende et al. (1999)
Relação entre Carbono e Nitrogênio da liteira	LIC/LIN	
	Animal	
Nitrogênio da Urina e das Fezes (kg)	NU, NF	Haynes e Williams (1993)
Peso animal (kg)	W	Mendonça et al. (2020)
	Solo	
Potencial de hidrogênio	pH	
Fósforo (mg dm^{-3})	Pr	Raij et al. (2001)
Saturação de bases (%)	B	
Capacidade de troca de cátions ($mmol_c$ dm^{-3})	CEC	
C e N da biomassa microbiana do solo (mg kg^{-1})	MBC, MBN	Monz et al. (1991); Raij et al. (2001); Silva et al. (2007)
Nitrogênio potencialmente mineralizável (mg kg^{-1})	PMN	Roberts et al. (2009)
Carbono lábil (mg kg^{-1})	LC	Weil et al. (2003)
Índice do manejo do carbono	CMI	Blair et al. (1995)
Carbono orgânico total (g kg^{-1})	TOC, C	Walkley e Black (1934)
Nitrogênio orgânico total (mg kg^{-1})	TON, N	Raij et al. (2001)
Relação C/N do solo	C/N-soil	
Relação C/N da biomassa microbiana do solo	C/N-SMB	
Estoque de Carbono e Nitrogênio no solo (Ton ha^{-1})	SCS, SNS	EMBRAPA (1997); Veldkamp et al. (1994)
Densidade do solo (g cm^{-3})	BD	
Macroporosidade, Microporosidade (cm^{-3} cm^{-3})	MA, MI	EMBRAPA (1997)
Porosidade total (cm^{-3} cm^{-3})	TP	
Potencial de água no solo (MPa)	SWC	

3.2.3 Processamento de dados

O conjunto de dados mencionado acima foi usado para estimar o PMN do solo, tomando como entrada o valor das variáveis de plantas, solo e animais do sistema de integração Lavoura-Pecuária. Antes de iniciar o processo de modelagem, as variáveis independentes de entrada foram padronizadas por StandardScaler, da biblioteca Scikit-learn da linguagem Python. Na fase de processamento dos dados foi utilizada a técnica de fator de inflação da variância (VIF), tem como objetivo verificar a presença de multicolinearidade ou o relacionamento das variáveis entre si (Taylor et al., 1981). Posteriormente, as cinco melhores variáveis foram selecionadas pelo método

combinado da correlação de Spearman (*p*) e Randon Forest Regressor (Figura 1), que tinha como premissa mais simples a variável se repetir pelo menos três vezes em todos os métodos, para garantir que tivessem alto nível de importância.

Os modelos que usamos para estimar o PMN do solo, foram: Multiple Linear Regression (MLR), Random Forest Regressor (RF), Adaptive Boosting Regressor (AdaBoost) e eXtreme Gradient Boosting Regressor (XGBoost). A técnica de validação cruzada (CrossValidation, cv=2) foi utilizada para comparar o desempenho de ambos os modelos; que funcionou com 2 divisões e uma proporção de 50-50 para a separação dos dados em treinamento e teste.

O ajuste de hiperparâmetros foi usado para cada modelo usando o Cross-Validation Score e Grid Search para selecionar a melhor combinação de hiperparâmetros. O ajuste é o processo de testar iterativamente várias combinações de hiperparâmetros dos modelos para obter os melhores resultados de modelagem em vez de usar valores padrão. Dessa forma, ajustamos a “profundidade máxima” para controlar o overfitting, pois mais profundidade permitirá que o modelo aprenda relações muito específicas para uma determinada amostra. Para evitar que isso ocorra, usamos o Grid Search, e os melhores parâmetros e valores ajustados para cada modelo estão apresentados no Apêndice 5.

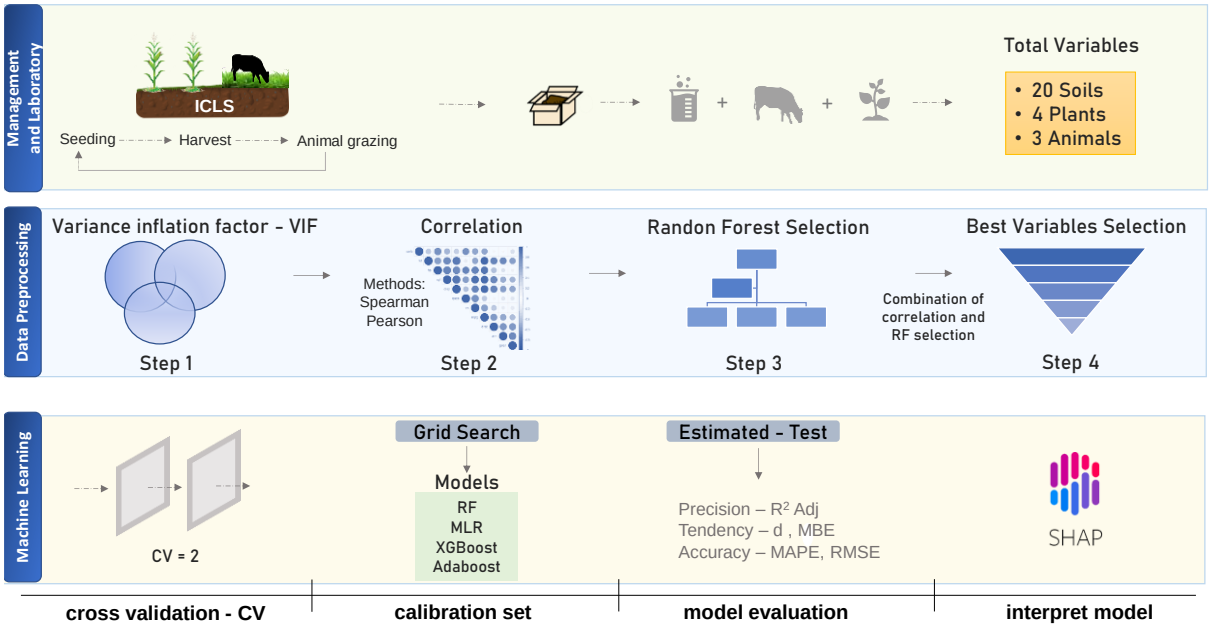


Figura 1. Fluxograma dos processos de experimentação, laboratório e modelagem para a estimação de nitrogênio potencialmente mineralizável do solo. RF: Random

Forest Regressor, MLR: Multiple Linear Regression, AdaBoost: Adaptive Boosting Regressor, XGBoost: eXtreme Gradient Boosting Regressor.

O MLR é a maneira mais simples de estabelecer a dependência da variável de saída em relação às variáveis de entrada. Sua principal desvantagem é assumir que a relação da variável independente com a dependente é linear. Além disso a RLM é sensível a outliers. O modelo RF apresenta uma vantagem por criar um conjunto de árvores de decisão e as combina reduzindo o efeito dos ruídos, gerando resultados mais precisos (Guo et al., 2004; Genuer et al., 2010). Esse modelo também pode ser usado para selecionar variáveis importantes (Strobl et al., 2008). A principal desvantagem do RF é a possibilidade de criar árvores tendenciosas, que pode ser resolvido com a padronização do banco de dados (Guo et al., 2004).

Os algoritmos *boosting* atribui todos os valores estimados de saída de uma coleção de aprendizes “fracos”, para estabelecer um aprendiz “forte” através da implementação de treinamento de aditivos (Tran, 2022). Neste estudo, o modelo de AdaBoost foi escolhido por ter baixo víeis ao trabalhar com bancos de dados pequeno, pois combina várias variáveis do banco de dados para formar uma única variável forte. Mesmo sendo a versão simples do Boosting, esta versão é extremamente rápida em comparação com outros algoritmos, principalmente as redes neurais. A sua principal desvantagem é ter baixa performance quando os dados apresentam muitos outliers.

Outro modelo boosting utilizado foi o XGBoost, este algoritmo foi escolhido devido ser um tipo aprimorado e sofisticado de modelo baseado em árvores, e o desempenho do modelo final é progressivamente melhorado devido ao “aumento de gradiente” (Friedman, 2001). Em outras palavras, um novo preditor é fornecido em cada árvore de decisão, diminuindo o overfitting e a capacidade computacional. Sua principal vantagem é o conjunto de regressores fracos superar até mesmo conjuntos de dados desequilibrados (Ho e Tran, 2022). De fato, XGBoost superou vários outros modelos de machine learning (por exemplo, Gradient Boosting, Artificial Neural Network, Support Vector Regressor, K-nearest Neighbors e Random Forest) para prever ou estimar indicadores de qualidade do solo (Hikouei et al., 2021; Ho e Tran, 2022; Nguyen et al., 2022; Zhou et al., 2022).

3.2.4 Avaliação do desempenho dos modelos

As métricas utilizadas para avaliar o desempenho dos modelos foram precisão, a tendência e acurácia. Para avaliar a precisão, foi usado o coeficiente de determinação ajustado (R^2 ajustado) (Eq. 1), a tendência pelo víeis de Erro Médio (Eq. 2) e a acurácia pelo Índice de Willmott et al. (1985) (Eq. 3):

$$R^2\ adjusted = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad (1)$$

$$MBE = \frac{\sum_{i-1}^{N}(Yobs_i - Yest_i)}{N} \quad (2)$$

$$d = 1 - \frac{\sum_{i-1}^{N}(Yobs_i - Yest_i)^2}{\sum_{i-1}^{N}(|Yest_i - Ymean| + |Yobs_i - Ymean|)^2} \quad (3)$$

Para avaliar a acurácia, também foram utilizados o MAE (Erro Absoluto Médio) (Eq. 4), MAPE (Erro Percentual Absoluto Médio) (Eq. 5) e RMSE (Raiz Quadrada do Erro Médio) (Eq. 6):

$$MAPE = \frac{\sum_{i=1}^{N}\left(\left|\frac{Yest_i - Yobs_i}{Yobs_i}\right|100\right)}{N} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_{obs_i} - Y_{est_i}\right)^2}{N}} \quad (5)$$

em que o R^2 é o coeficiente de determinação ajustado, n é o número de elementos, k é o número de variáveis de regressão, SE é o erro sistemático, d é o Índice de Willmott, Yobs são os dados observados, Yest são os dados estimados, Ymean é a média dos dados.

3.2.5 Interpretação dos modelos usando o SHAP

Para analisar a importância relativa das variáveis dentro do modelo e como elas influenciam no PMN do solo, foi usado a biblioteca Python SHAP (SHapley Additive

exPlanations) (Lundberg et al., 2020). Após o treinamento e validação dos modelos, cada modelo foi treinado novamente utilizando o SHAP com a função *explainer*, mantendo a independência e particularidades de cada modelo de regressão avaliado. Este método calcula o valor Shapley para a variável *i* para estimar PMN e sua contribuição na saída do modelo (Yan et al., 2020). A contribuição de cada variável é calculada através de pesos de shap que podem variar entre negativo ou positivo (Figura 2).

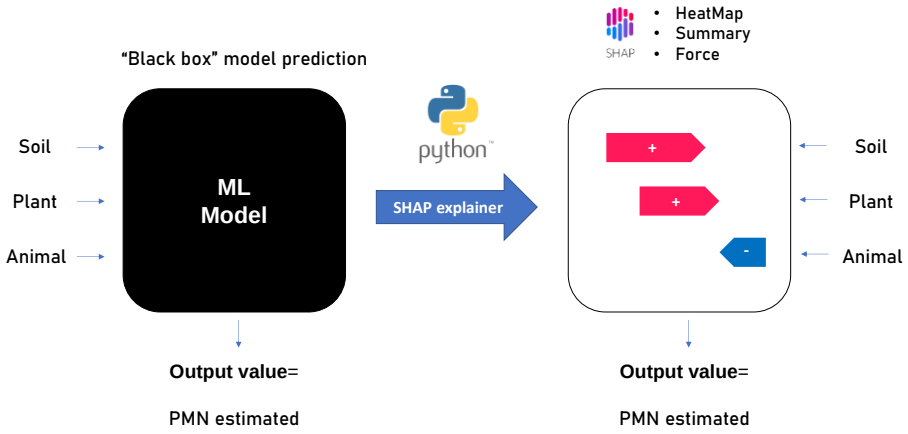


Figura 2. Representação diagramática do SHAP. Figura adaptada de Scavuzzo et al. (2022) e Lundberg et al. (2020). ML: machine learning, PMN: nitrogênio potencialmente mineralizável.

A contribuição das diferentes variáveis e instancias (amostras) foram calculadas por meio dos valores SHAP, e para a explicação de estimação global foram utilizados os gráficos de mapa de calor e resumo absoluto. Para exemplificar a importância local com base em uma única amostra, foi utilizado o gráfico de força, ambos da biblioteca SHAP do Python.

3.3 RESULTADOS E DISCUSSÃO

3.3.1 Importância e Seleção das variáveis estimadoras de nitrogênio potencialmente mineralizável do solo

A matriz de correlação entre as variáveis de planta, solo e animais e o PMN (Figura 3) indicaram uma correlação forte e negativa entre a porosidade total e PMN (p= -0.48, Figura 3 A). Por exemplo, a diminuição na porosidade total do solo aumenta

o teor de PMN no solo, devido a diminuição de fluxo de oxigênio e água nos poros (Schjønning et al.,2003; Chirinda et al., 2010).

Assim, os processos metabólicos dos microrganismos são cessados e os processos de mineralização de N no solo iniciam (Zhang et al., 2019), dando início aos processos de N que podem ser assimilados pelas plantas. Além disso, existe uma correlação negativa entre o PMN versus carbono lábil, índice de manejo do carbono e nitrogênio da liteira (Figura 3). O carbono lábil e índice de manejo do carbono estão associados a funções biológicas do solo, portanto, são dependentes da biomassa microbiana do solo. Logo, a correlação negativa indicou uma diminuição do PMN no solo por conta da imobilização de N pelos microrganismos do solo (Burger e Jackson, 2003). Chirinda et al. (2010) sugeriu que a aplicação de esterco no solo eleva a atividade microbiana em sistemas com grande aporte de resíduos vegetais e esterco, levando a níveis mais altos de processos de transformação de N, ou seja, PMN. Assim, a fração lábil do carbono é importante para a ciclagem de C e N, e a disponibilidade de PMN a longo prazo (Cookson et al., 2005).

As variáveis que mais afetaram positivamente o PMN foi o carbono da liteira (média = 0.38), seguido do estoque de nitrogênio no solo (média = 0.28) e a capacidade de troca de cátions (média = 0.25) (Figura 3), sugerindo que sua principal fonte de N é oriunda de restos vegetais da liteira e do estoque de N no solo. Além de, indicar que sistemas integrados teve maior balanceamento de fertilidade do solo, que também impactou na disponibilidade do N potencialmente mineralizável no solo (como é o caso da capacidade de troca de cátions).

O aumento da fertilidade do solo em sistemas de integração lavoura-pecuária já foi relatado em diversos estudos (Moraes et al., 2014; Rakkar et al., 2017; Galindo et al., 2020; Maia et al., 2021), além da sua influência no nitrogênio mineral do solo (Lantinga et al., 2012; Bansal et al., 2022). No geral, não houve alta correlação entre as variáveis de entrada e o PMN (Spearman < 0.50). Assim, variáveis que obtiveram VIF > 10 foram excluídas e não foram utilizadas na modelagem posterior.

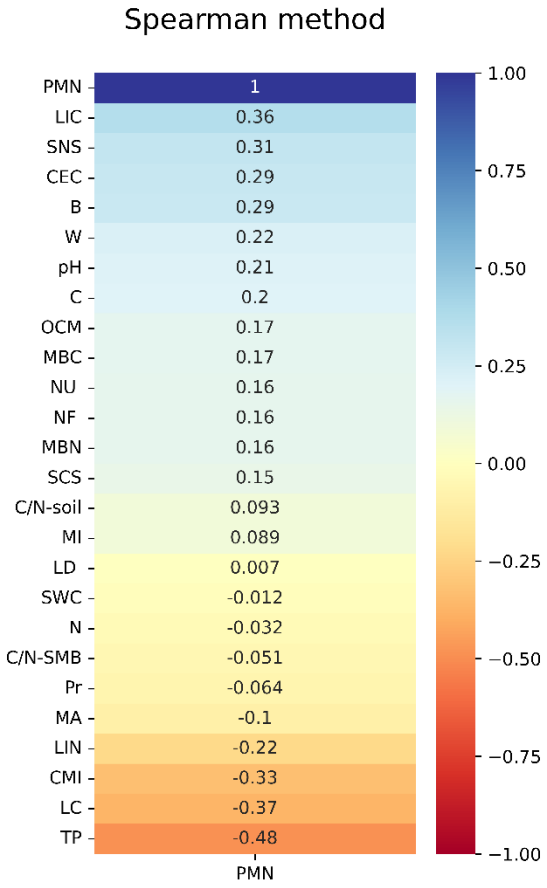


Figura 3. Correlação das variáveis de solo, planta e animal com o PMN do solo pelo método de Spearman. Correlações negativas estão em vermelho. C: carbono orgânico total; N: nitrogênio orgânico total; MBC: carbono da biomassa microbiana; MBN: nitrogênio da biomassa microbiana; LC: carbono lábil; CMI: índice de manejo do carbono; PMN: nitrogênio potencialmente mineralizável; C/N-soil: relação carbono e nitrogênio do solo; C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo; BD: densidade do solo; MA: macroporosidade; MI: microporosidade; TP: porosidade total; SCS: estoque de carbono; SNS: estoque de nitrogênio; pH: potencial de hidrogênio; Pr: fósforo; CEC: capacidade de troca de cátions; B: saturação de bases; SWC: potencial de água no solo. NU: nitrogênio da urina; NF: nitrogênio das fezes; W: peso animal; LD: liteira depositada; LIC: carbono da liteira; LIN: nitrogênio da liteira; LIC/LIN: relação carbono e nitrogênio da liteira.

A importância relativa das cinco variáveis mais importantes usadas para estimar PMN do solo indicaram que o carbono lábil do solo (LC) esteve altamente correlacionado com PMN, com correlação negativa variando entre 34 a 37%, e importância relativa de 33% (Figura 4). As variáveis influentes seguintes foram o carbono da liteira, porosidade total, capacidade de troca de cátions e, estoque de nitrogênio do solo com o menor poder de explicação (4%). As variáveis carbono da leiteira, estoque de nitrogênio e capacidade de troca de cátions foram positivamente

correlacionadas com o PMN do solo, enquanto o carbono lábil e porosidade total foi negativamente correlacionado com PMN do solo (Figura 4). Os ajustes dos modelos foram feitos usando essa seleção de variáveis.

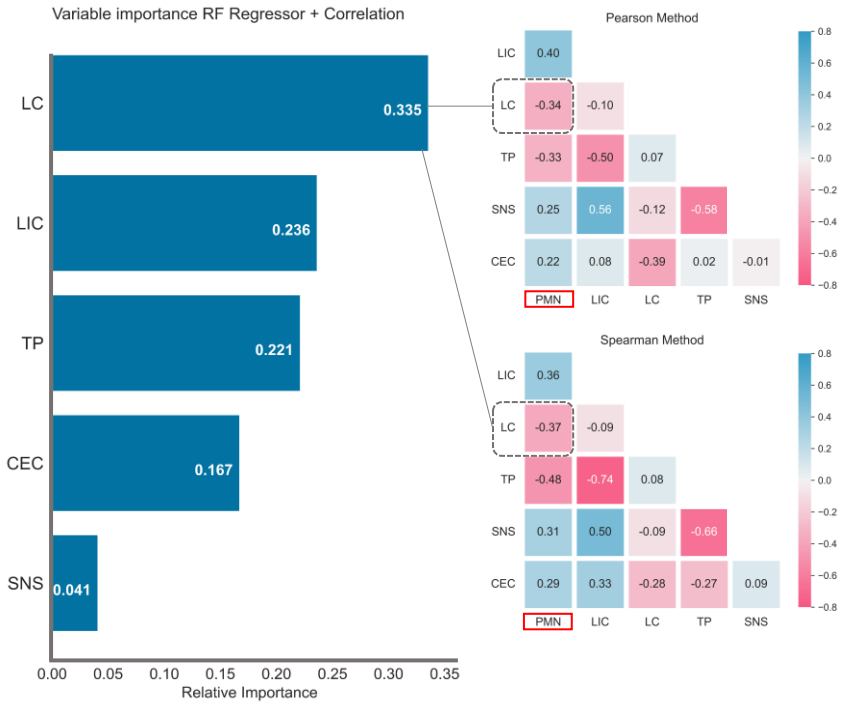


Figura 4. Importância relativa (%) das cinco variáveis mais importantes para estimar PMN do solo, geradas pelo método combinado de correlação e RF Regressor. Correlações negativas estão em vermelho. PMN: nitrogênio potencialmente mineralizável, LC: carbono lábil, TP: porosidade total, SNS: estoque de nitrogênio, CEC: capacidade de troca de cátions, LIC: carbono da liteira.

3.3.2 Desempenho dos modelos na estimação de nitrogênio potencialmente mineralizável do solo

Quanto à precisão, os modelos Adaboost (R^2=0.89) e XGboost (R^2=0.97) apresentaram altos valores de explicabilidade e índice de concordância (d= 0.97 e d= 0.99, respectivamente) (Tabela 3). Considerando o conjunto de dados original, os valores calculados de MAPE foram pequenos, sugerindo que o XGBoost (3%) e AdaBoost (6%) gerou um modelo mais confiável do que o MLR (~14%) e RF (~10%). Portanto, o erro (RMSE) de estimação de PMN variou entre os modelos de 7.22 a

34.82 mg kg^{-1}. Para as condições de estimação, os modelos MLR e RF indicam que 14% e 10% dos erros de víeis apresentam certa tendência decrescente do MLR (-0.47 mg kg^{-1}), e tendência crescente do RF (0.47 mg kg^{-1}). Apesar do AdaBoost ter apresentado alta precisão e baixo erro, indicou alto víeis dos dados estimados, ou seja, 6% do seu erro (MAPE) teve uma tendência decrescente de -0.84 mg kg^{-1}.

Tabela 3. Desempenho estatísticos dos algoritmos na estimativa de PMN do solo. MLR: Multiple Linear Regression, RF: Random Forest Regressor, AdaBoost: Adaptive Boosting Regressor, XGBoost: eXtreme Gradient Boosting Regressor.

Manejo		**MLR**	**RF**	**AdaBoost**	**XGBoost**
	R^2- ajustado	0.21	0.55	0.89	0.97
	RMSE	34.822	26.248	12.739	7.223
ICLS	MAPE	13.822	9.846	5.910	3.004
	MBE	-0.4737	0.4737	-0.8368	0.1049
	d	0.6352	0.8238	0.9733	0.9912

R^2: coeficiente de determinação ajustado; *RMSE*: raiz quadrada do erro médio; *MAPE*: erro percentual absoluto médio; *d*: Índice de Willmott et al. (1985); *MBE*: víeis de erro médio; *ICLS*: Integração Lavoura-Pecuária; PMN: nitrogênio potencialmente mineralizável.

No entanto, como a acurácia deve ser avaliada pela precisão, erro observado e víeis do modelo, o XGBoost apresentou maior acurácia, com o coeficiente de determinação ajustado (R^2) de 97%, RMSE de 7.22 mg kg^{-1} e MBE de 0.10 mg kg^{-1} (Figura 5 D). Isso se deve ao fato do XGBoost ter obtido a menor tendência dos dados estimados (Figura 5), comparados com os outros modelos como o MLR (MBE= -0.47 mg kg^{-1}), RF (MBE= -0.47 mg kg^{-1}) e AdaBoost (MBE= -0.84 mg kg^{-1}).

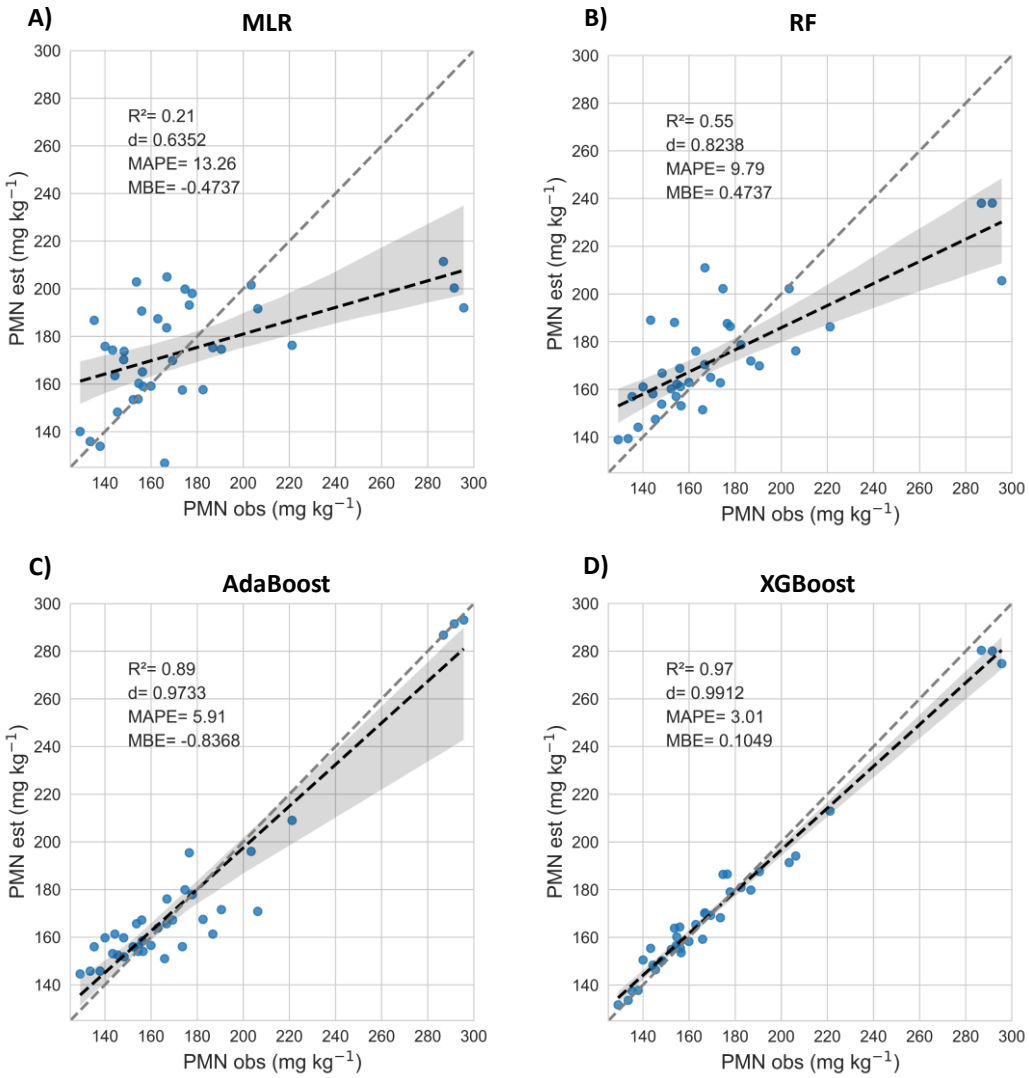


Figura 5. Desempenho dos modelos (**A**) MLR, (**B**) RF, (**C**) AdaBoost, (**D**)XGBoost para estimar o PMN do solo. Os pontos azuis são dados de treinamento. As linhas cinzas correspondem a linhas 1:1. RF: Random Forest Regressor, MLR: Multiple Linear Regression, AdaBoost: Adaptive Boosting Regressor, XGBoost: eXtreme Gradient Boosting Regressor. R^2: coeficiente de determinação ajustado, MAPE: erro percentual absoluto médio, d: Índice de Willmott et al. (1985), MBE: víeis de erro médio. PMN: nitrogênio potencialmente mineralizável.

Nguyen et al. (2022), com o objetivo de estimar carbono orgânico do solo usando índices de vegetação, observaram que o XGBoost (R^2=0.91) também foi superior ao RF (R^2=0.87). Outro estudo indicou precisão do XGBoost de 76% contra 70% do RF para estimar metais pesado do solo (Yang et al., 2020), além disso, segundo os autores, a variável mais importante foi a capacidade de troca de cátions (CEC). Em contraste com nossos resultados, um estudo utilizando o MLR para prever

PMN do solo obteve 80% de precisão (Osterholz et al., 2017), no entanto, utilizou apenas variáveis oriundas de frações da matéria orgânica do solo, análises que precisam de reagentes e equipamentos específicos como o fluorômetro Aqualog. Baseado nisso, um ponto favorável desse estudo, na estimação de PMN, foi que as variáveis de entrada utilizadas podem ser comumente determinadas em análises de rotinas de solo de áreas agrícolas, portanto, não há necessidade de dispositivos e medições adicionais para a estimação. Para entender o peso de cada variável na estimação do PMN, a interpretação dos modelos avaliados é revisada e discutida na seção a seguir.

3.3.3 Interpretação dos modelos usando SHAP

O mapa de calor possibilitou identificar o valor SHAP por trás da mesma previsão em diferentes instâncias (Figura 6). O mapa usou agrupamento hierárquico baseado na similaridade de explicação dos dados estimados. Assim, variáveis com peso de explicabilidade menor teve um valor de shap menor na saída de cada modelo (azul ou branco). Enquanto vermelho indicou que as variáveis contribuíram mais para o teor de PMN do solo. Portanto, as estimativas mais precisas (altos valores de *f(x)* na parte superior) estiveram associadas ao teor de carbono lábil, carbono da liteira e porosidade total do solo nos modelos RF, AdaBoost e XGBoost (Figura 6 B, C, D), mais a capacidade de troca de cátions no AdaBoost (Figura 6 C). Ao contrário dos demais modelos, o MLR variou o seu padrão ao longo das instancias (amostras), além disso, seguindo uma via contrária, a variável que mais contribuiu para a estimativa do PMN foi o carbono da liteira, seguido do carbono lábil e porosidade total do solo (Figura 6 A).

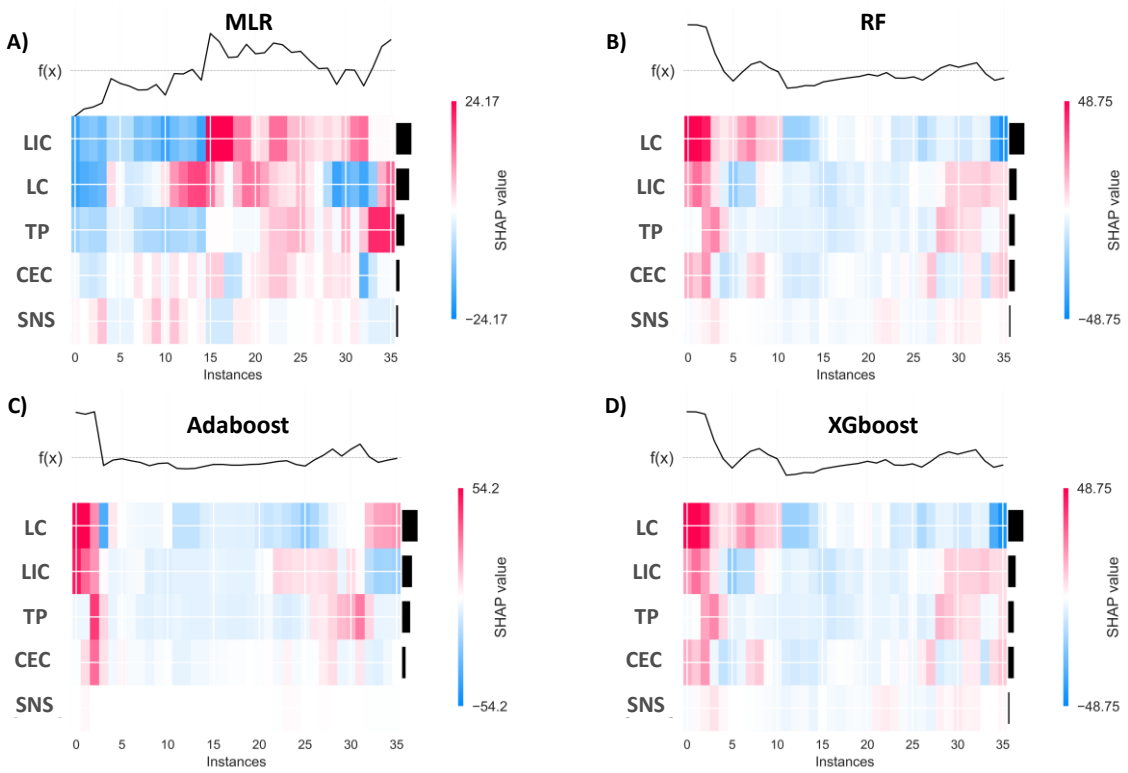


Figura 6. SHAP mapa de calor e impacto de cada registro no valor final estimado de PMN (*f(x)*) dos modelos (**A**) MLR, (**B**) RF, (**C**) AdaBoost, (**D**)XGBoost. O eixo X é a instância entre os registros 0 a 36 (instances). O eixo Y são as variáveis importantes. As cores mostram a magnitude dos valores SHAP, cor vermelha representa alto poder de estimação, cor azul representa baixo pode de estimação. A curva *f(x)* na parte superior da figura são a acurácia das estimativas de PMN do modelo de cada instância. As cores se agrupam porque a classe de mapa de calor de SHAP executa um agrupamento hierárquico nas instancias e, em seguida, ordena as amostras de 1 a 36 no eixo X. RF: Random Forest Regressor, MLR: Multiple Linear Regression, AdaBoost: Adaptive Boosting Regressor, XGBoost: eXtreme Gradient Boosting Regressor. PMN: nitrogênio potencialmente mineralizável, LC: carbono lábil, TP: porosidade total, SNS: estoque de nitrogênio, CEC: capacidade de troca de cátions, LIC: carbono da liteira. As barras em preto indicam o grau de importância geral da variável.

Os resultados acima mostram que os indicadores de solo e planta: carbono lábil, carbono da liteira e porosidade total estiveram entre as características mais importantes na decisão de saída dos algoritmos mais precisos (ou seja, na liberação de PMN no solo). O gráfico de resumo shap (violino) combina a importância das variáveis com os efeitos de cada variável na estimação do PMN (Figura 7 e 8). Neste gráfico, vemos as primeiras indicações da relação entre o valor de um recurso (indivíduo) e o seu impacto na estimação de PMN do solo. No entanto, para ver a

forma exata como cada indivíduo impacta na estimação de PMN, colocamos um exemplo de dados estimados para cada modelo, e através do gráfico de força shap (Figura 7 e 8 A.1 B.2), conseguimos avaliar as explicações para uma única amostra.

Complementando o gráfico de importância das variáveis (Figura 4) e o mapa de calor (Figura 6), através do gráfico de resumo (Figura 7 e 8 A, B), podemos plotar todos os dados estimados e obter interpretações claras das variáveis para cada modelo. Por exemplo, o carbono da liteira (LIC) foi a variável com maior peso de importância no MLR (Figura 7 A), isso nos diz que valores maiores de LIC levarão a maior precisão estimada do PMN do solo. Ao avaliar as figuras 7 e 8, podemos notar que o MLR foi o único que apresentou o LIC como variável estimadora mais importante, indicando que o modelo não acertou os padrões de explicabilidade (Figura 4), portanto, pode ser o caso dessa variável estar levando a maiores erros e tendência no MLR (Figura 5 A, Tabela 3).

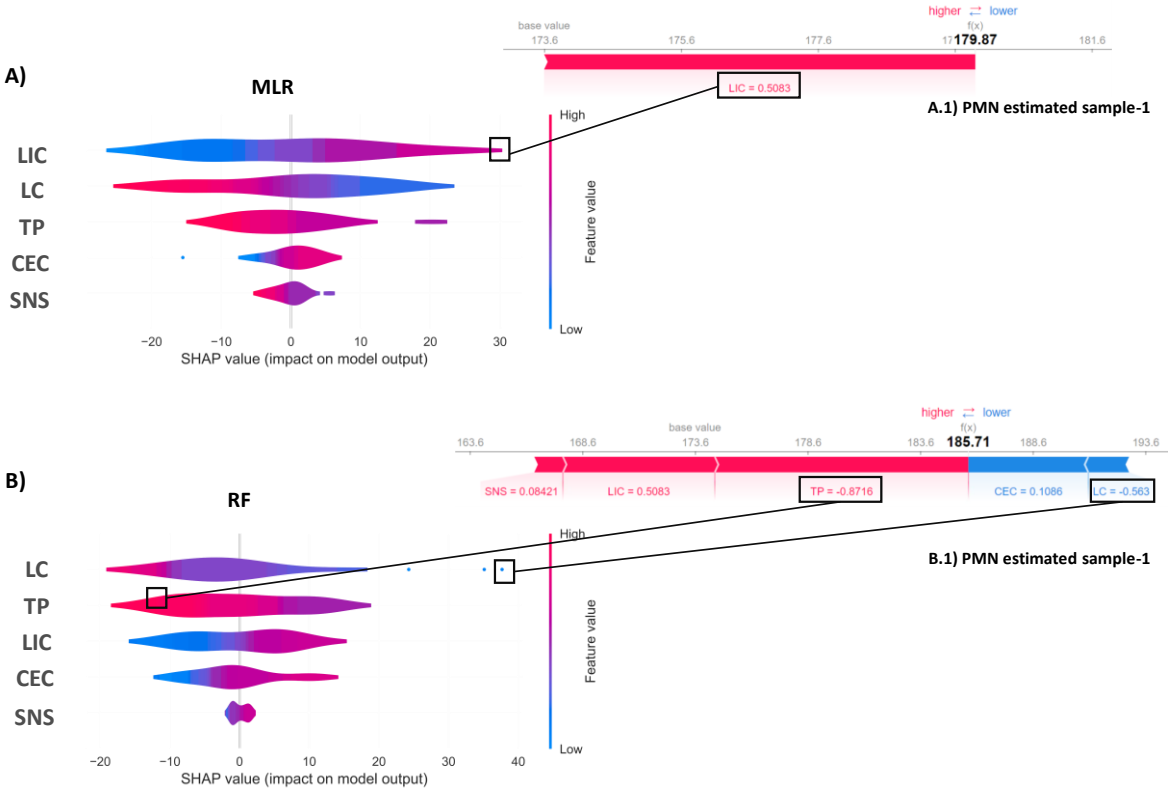


Figura 7. SHAP do resumo de estimação de todos os registros (gráfico de violino) e força de estimação para um único registro (gráfico de barra). (**A)** MLR Regressor, e (**B)** RF Regressor para estimar o PMN do solo. *Gráfico de violino:* As variáveis são classificadas em ordem decrescente de importância. As cores mostram a magnitude

dos valores SHAP, representa se a observação é alta (vermelho) ou baixa (azul) para estimar o PMN do solo. No eixo X o impacto de cada observação pode ter uma correlação positiva ou negativa com a variável alvo. (**A.1)** força de estimação MLR Regressor, e (**B.2)** força de estimação RF Regressor para estimar o PMN do solo. *Gráfico de barra:* O valor de *f(x)* é a estimação final do modelo para essa observação (valor em negrito). O valor base (base value) é a previsão média do modelo sobre o conjunto de treinamento (Yest mean). As variáveis que aumentam a previsão são mostradas em vermelho, e aqueles que forçam a previsão para baixo são mostrados em azul. Valores de variáveis menores que mostrados, indicam que forçam a estimação para a direita ou esquerda. RF: Random Forest Regressor, MLR: Multiple Linear Regression. PMN: nitrogênio potencialmente mineralizável, LC: carbono lábil, TP: porosidade total, SNS: estoque de nitrogênio, CEC: capacidade de troca de cátions, LIC: carbono da liteira.

Nos modelos RF, AdaBoost e XGBoost a variável carbono lábil (LC) foi a variável mais importante, ambos são negativamente correlacionados com o PMN do solo (Figura 7 B, Figura 8 A e B). O SHAP confirmou que o impacto do carbono lábil no PMN do solo é maior em comparação aos efeitos do estoque de nitrogênio nos agregados do solo (SNS). Romanyà et al. (2012) também observaram que em solos organicamente manejados, os estoques de N orgânicos foram menos propensos a se tornarem N potencialmente mineralizável, além do mais, afirmou que solos ricos em matéria orgânica liberam gradativamente e lentamente o PMN no solo, do que solos manejados convencionalmente.

Ao avaliar a magnitude dos dados, observou-se que a diminuição de LC teve um impacto positivo no teor de PMN do solo (Figura 8 B), isso ocorre devido ao carbono lábil do solo representar as funções biológicas do solo, portanto, se a concentração de LC diminui no solo, sugere uma maior rotatividade da matéria orgânica em sistemas integrados, ocasionado pelos processos de mineralização do N. Em outras palavras, a concentração menor de LC no solo está relacionada a maior disponibilidade de nitrogênio a curto prazo no solo (PMN) (Cookson et al., 2005; Chirinda et al., 2010; Maia et al., 2021).

Os valores de SHAP nas Figuras 7 B, 8 A e B mostraram que o carbono oriundo da liteira (LIC) esteve positivamente correlacionado com o PMN do solo, sugerindo que o aumento de resíduos orgânicos no solo também influenciou na concentração de nitrogênio a curto prazo. Liu et al. (2009) encontraram resultado semelhante, mostrando que a entrada de serapilheira no solo impulsiona os ciclos de carbono e nitrogênio no solo. Além disso, o aumento de serapilheira no solo mostrou-se eficaz

no aumento de PMN no solo (Córdova et al., 2018; Zhang et., 2018; Strickland et al., 2019).

A porosidade total (TP) apresentou uma alta correlação negativa com o PMN (Figuras 7 B, 8 A e B), além de ser a terceira variável mais importante nos modelos mais robustos (Figuras 5 C e D), sugeriu que a diminuição na porosidade total do solo aumentou o teor de PMN no solo. Na mesma linha de pesquisa, Souza et al. (2010) concluíram que o sistema de integração Lavoura-Pecuária (ICLS) promoveu maior agregação do solo em comparação com áreas não pastoreadas, devido ao maior crescimento e desenvolvimento das raízes através do consórcio de gramíneas, que minimizam os efeitos da intensidade de pastejo. Portanto, o pequeno efeito de pisoteio sobre os atributos físicos do solo, que por sua vez não limita o crescimento e desenvolvimento das culturas, indica que a qualidade do solo pode ser preservada em ICLS sob condições de pastejo moderado.

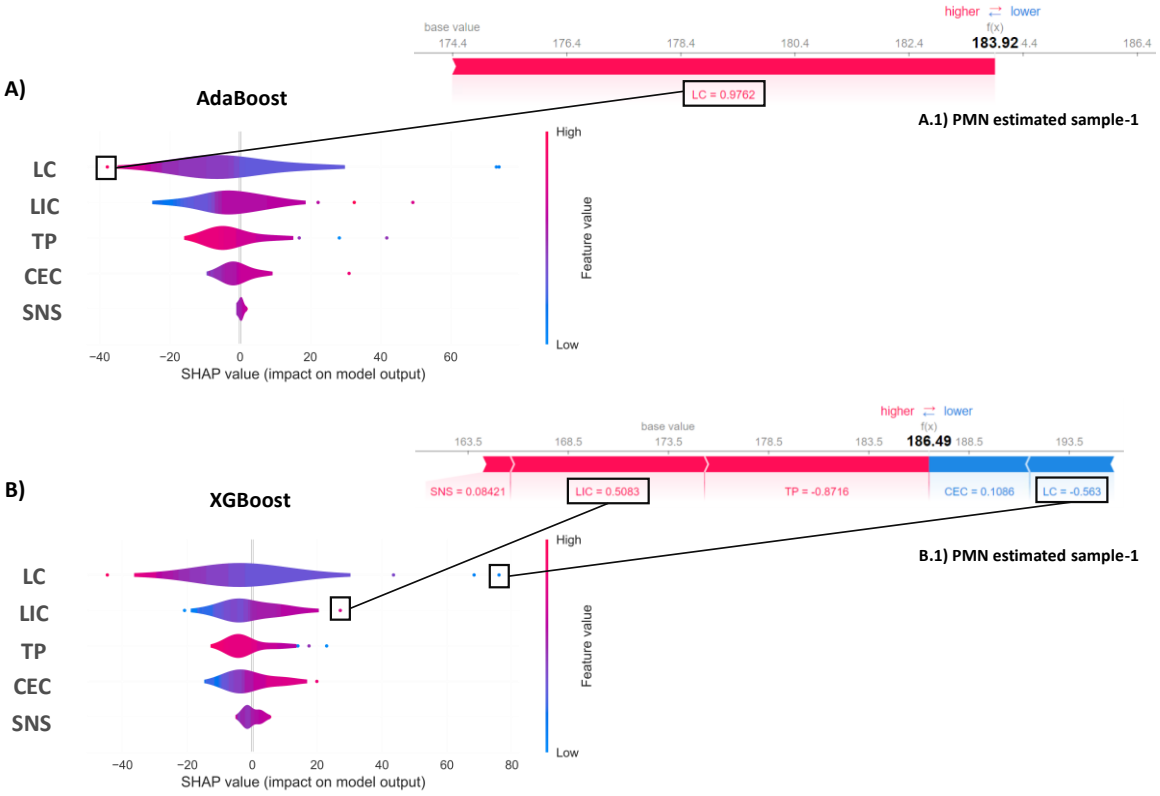


Figura 8. SHAP do resumo de estimação de todos os registros (gráfico de violino) e força de estimação para um único registro (gráfico de barra). (**A)** AdaBoost Regressor, e (**B)** XGBoost Regressor para estimar o PMN do solo. *Gráfico de violino:* As variáveis são classificadas em ordem decrescente de importância. As cores mostram a

magnitude dos valores SHAP, representa se a observação é alta (vermelho) ou baixa (azul) para estimar o PMN do solo. No eixo X o impacto de cada observação pode ter uma correlação positiva ou negativa com a variável alvo. (**A.1)** força de estimação Adaboost Regressor, e (**B.2)** força de estimação XGboost Regressor para estimar o PMN do solo. *Gráfico de barra:* O valor de *f(x)* é a estimação final do modelo para essa observação (valor em negrito). O valor base (base value) é a previsão média do modelo sobre o conjunto de treinamento (Yest mean). As variáveis que aumentam a previsão são mostradas em vermelho, e aqueles que forçam a previsão para baixo são mostrados em azul. Valores de variáveis menores que mostrados, indicam que forçam a estimação para a direita ou esquerda. AdaBoost: Adaptive Boosting Regressor, XGBoost: eXtreme Gradient Boosting Regressor. PMN: nitrogênio potencialmente mineralizável, LC: carbono lábil, TP: porosidade total, SNS: estoque de nitrogênio, CEC: capacidade de troca de cátions, LIC: carbono da liteira.

Os efeitos na saída dos modelos com maior acurácia – Adaboost e XGBoost (Figura 5, Tabela 3) podem ser resumidos como uma maior possibilidade de ter aumentado o PMN no solo (valor SHAP > 0, Figura 8), e uma menor possibilidade de ter aumento de PMN no solo (valor SHAP < 0, Figura 8). Isso sugeriu que solos de sistemas integrados tendem a apresentar valores mais baixos de estoque nitrogênio (SNS) com SHAP < 0 (cor azul), que resultam em menor probabilidade de estimar PMN do solo. E valores mais baixos de carbono lábil (LC) com SHAP > 0 (cor azul), que resultam em maior probabilidade de estimar PMN do solo. Do ponto de vista agronômico, esses resultados apontam que solos de sistemas integrados liberam mais carbono oriundo dos restos vegetais (LIC) em conjunto com as estruturas do solo de porosidade total e o teor de capacidade de troca de cátions, que impactam positivamente na liberação de PMN do solo a curto prazo. Mas para que isso ocorra, as condições entre saídas e entradas de matéria orgânica do solo com a atuação de microrganismos do solo precisam encontrar-se balanceadas (LC).

Os resultados acima levaram em consideração várias amostras simultaneamente, o que foi possível ter uma ideia geral do que o modelo aprendeu, e como isso impactou na qualidade do solo. No entanto, podemos avaliar um único valor SHAP por amostra, através do gráfico de força (Figuras 7 e 8, A.1 e B.1). Os valores médios do conjunto de dados de teste foram indicados por “valor base” em cinza, enquanto a variável de saída da amostra individual foi mostrada em negrito. A amostra do RF e XGboost apresentou o carbono lábil com impacto positivo no PMN, ou seja, valores menores que 0.56 (< 193 mg kg^{-1}) direciona a estimação para a esquerda (Figuras 7 e 8 B.1). Podemos visualizar nesses modelos que o C da liteira e

porosidade total teve impacto maior sobre a estimação dessa amostra, direcionado a estimação para a direita (Figuras 7 e 8 B.1). Além do mais, confirmamos o baixo impacto do estoque de nitrogênio no teor de PMN do solo, uma vez que seu impacto na estimação inicia abaixo do valor inicial de base (< 173.5 mg kg^{-1}). De fato, observamos uma relação entre a entrada de diferentes resíduos no solo, estrutura e fertilidade do solo de sistemas integrados, sugerindo que arranjos integrados tem a capacidade de aumentar o nitrogênio potencialmente mineralizável para as plantas (Liebig et al., 2011; Ryschawy et al., 2017; Ghimire et al., 2018; Osterholz et al., 2018).

3.4 CONCLUSÕES

O carbono lábil, carbono da liteira, porosidade total do solo, capacidade de troca de cátions e estoque de nitrogênio no solo associado ao extreme gradient boosting (XGBoost) é eficiente na estimativa de PMN (R^2=0.97, RMSE de 7.22 mg kg^{-1} e MBE de 0.10 mg kg^{-1}), indicando que pode ser usado para estimar e monitorar indicadores importantes de qualidade do solo.

Ao interpretar os modelos mais acurados com o método SHAP, constatamos que 1) as variáveis mais importantes que impactam no PMN do solo são carbono lábil, carbono da liteira, porosidade total do solo, capacidade de troca de cátions e estoque de nitrogênio no solo, nessa ordem de importância; 2) a diminuição de carbono lábil no solo impacta positivamente na liberação de PMN no solo a curto prazo; 3) solos de sistemas integrados liberam mais carbono oriundo dos restos vegetais (carbono da liteira); 4) o balanceamento entre as estruturas físicas (porosidade total) e de fertilidade do solo (troca catiônica) influenciam na liberação de PMN do solo; 5) estoque de nitrogênio no solo resultam em menor probabilidade de estimar PMN do solo.

Esse estudo demonstra que modelos de machine learning podem ser usados para estimar o nitrogênio potencial mineralizável no solo, e através disso, monitorar a qualidade dos solos. No entanto, limitado pela disponibilidade de dados, nosso estudo não considerou o nitrogênio mineral ou o retorno desse nutriente para os sistemas de cultivo e colheitas. Assim, estudos futuros são necessários para melhorar ainda mais a acurácia dos modelos e auxiliar na interpretação das causas de explicabilidade de liberação de nitrogênio dos solos.

3.5 REFERÊNCIAS

Abalos D, Jeffery S, Sanz-Cobena A, Guardia G, Vallejo A (2014) Meta-analysis of the effect of urease and nitrification inhibitors on crop productivity and nitrogen use efficiency. **Agriculture, Ecosystems & Environment** 189: 136-144. https://doi.org/10.1016/j.agee.2014.03.036

Abbruzzini TF, Davies CA, Toledo FH, Cerri CEP (2019) Dynamic biochar effects on nitrogen use efficiency, crop yield and soil nitrous oxide emissions during a tropical wheat-growing season. **Journal of environmental management** 252: 109638. https://doi.org/10.1016/j.jenvman.2019.109638

Bansal S, Chakraborty P, Kumar S (2022) Crop–livestock integration enhanced soil aggregate-associated carbon and nitrogen, and phospholipid fatty acid. **Scientific Reports** 12(1): 1-13. https://doi.org/10.1038/s41598-022-06560-6

Burger M, Jackson LE (2003) Microbial immobilization of ammonium and nitrate in relation to ammonification and nitrification rates in organic and conventional cropping systems. **Soil Biology and Biochemistry** 35(1): 29-36. https://doi.org/10.1016/S0038-0717(02)00233-X

Blair GJ, Lefroy RDB, Lisle L (1995) Soil carbon fractions based on their degree of oxidation, and the development of a carbon management index for agricultural systems. Australian Journal of Agricultural Research 46: 1459-1466. https://doi.org/10.1071/AR9951459

Bockheim JG, Gennadiyev AN, Hartemink AE, Brevik EC (2014) Soil-forming factors and Soil Taxonomy. Geoderma 231-237. https://doi.org/10.1016/j.geoderma.2014.02.016

Carvalho PCDF, Peterson CA, Nunes PADA, Martins AP, Souza Filho, W, Bertolazi VT, Kunrath TR, Moraes A, Anghinoni I (2018) Animal production and soil characteristics from integrated crop-livestock systems: toward sustainable intensification. **Journal of animal science** 96(8): 3513-3525. https://doi.org/10.1093/jas/sky085

Chatzisymeon E, Foteinis S, Borthwick AGL (2017) Life cycle assessment of the environmental performance of conventional and organic methods of open field pepper cultivation system. **The International Journal of Life Cycle Assessment** 22: 896–908. https://doi.org/10.1007/s11367-016-1204-8

Chirinda N, Olesen JE, Porter JR, Schjønning P (2010). Soil properties, crop production and greenhouse gas emissions from organic and inorganic fertilizer-based arable cropping systems. **Agriculture, Ecosystems & Environment** 139(4): 584-594. https://doi.org/10.1016/j.agee.2010.10.001

Cookson WR, Abaye DA, Marschner P, Murphy DV, Stockdale EA, Goulding KW, (2005) The contribution of soil organic matter fractions to carbon and nitrogen mineralization and microbial community size and structure. **Soil Biology and Biochemistry** 37(9): 1726-1737. https://doi.org/10.1016/j.soilbio.2005.02.007

Córdova SC, Olk DC, Dietzel RN, Mueller KE, Archontouilis SV, Castellano MJ (2018) Plant litter quality affects the accumulation rate, composition, and stability of mineral-associated soil organic matter. **Soil Biology and Biochemistry** 125: 115-124. https://doi.org/10.1016/j.soilbio.2018.07.010

EMBRAPA (1997) Manual de métodos de análise de solo. Rio de Janeiro: Centro Nacional de Pesquisa de Solos.

Filippi P, Whelan BM, Vervoort RW, Bishop TF (2020). Mid-season empirical cotton yield forecasts at fine resolutions using large yield mapping datasets and diverse spatial covariates. **Agricultural Systems** 184: 102894. https://doi.org/10.1016/j.agsy.2020.102894

Franzluebbers AJ (2016) Should soil testing services measure soil biological activity? **Agricultural & Environmental Letters** 1(1). https://doi.org/10.2134/ael2015.11.0009

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. **Annals of statistics** 1189-1232. http://www.jstor.org/stable/2699986.

Galindo FS, Delate K, Heins B, Phillips H, Smith A, Pagliari PH (2020) Cropping System and Rotational Grazing Effects on Soil Fertility and Enzymatic Activity in an Integrated Organic Crop-Livestock System. **Agronomy** 10: 803. https://doi.org/10.3390/agronomy10060803

Genuer R, Poggi JM, Tuleau-Malot C (2010) Variable selection using random forests. **Pattern recognition letters** 31(14): 2225-2236. https://doi.org/10.1016/j.patrec.2010.03.014

Guo L, Ma Y, Cukic B, Singh H (2004) Robust prediction of fault-proneness by random forests. In: 15th international symposium on software reliability engineering. **IEEE** 417-428. https://doi.org/10.1109/ISSRE.2004.35

Ghimire R, Thapa VR, Cano A, Acosta-Martinez V (2019) Soil organic matter and microbial community responses to semiarid croplands and grasslands management. **Applied Soil Ecology** 141: 30-37. https://doi.org/10.1016/j.apsoil.2019.05.002

Grell M, Barandun G, Asfour T, Kasimatis M, Collins ASP, Wang J, Güder F (2021) Point-of-use sensors and machine learning enable low-cost determination of soil nitrogen. **Nature Food** 2(12): 981-989. https://doi.org/10.1038/s43016-021-00416-4

Hassen TB, El Bilali H (2022) Impacts of the Russia-Ukraine War on Global Food Security: Towards More Sustainable and Resilient Food Systems?. **Foods** 11(15): 2301. https://doi.org/10.3390/foods11152301

Hebebrand C, Laborde D (2022) High Fertilizer Prices Contribute to Rising Global Food Security Concerns. Disponível em: <https://www.ifpri.org/blog/high-fertilizer-prices-contribute-rising-global-food-security-concerns>. Acesso em: 22 set. 2022.

Haynes RJ, Williams PH (1993) Nutrient cycling and soil fertility in the grazed pasture ecosystem. Advances in agronomy 49: 119-199. https://doi.org/10.1016/S0065-2113(08)60794-4

Hikouei IS, Kim SS, Mishra DR (2021) Machine-learning classification of soil bulk density in salt marsh environments. **Sensors** 21(13) : 4408. https://doi.org/10.3390/s21134408

Ho LS, Tran VQ (2022) Machine learning approach for predicting and evaluating California bearing ratio of stabilized soil containing industrial waste. **Journal of Cleaner Production** 133587. https://doi.org/10.1016/j.jclepro.2022.133587

Jarray N, Abbes AB, Farah IR (2021) An evaluation of soil moisture retrieval using machine learning methods: Application in arid regions of Tunisia. **IEEE** 6331-6334. https://doi.org/10.1109/IGARSS47720.2021.9554585

Jones EJ, Bishop TF, Malone BP, Hulme PJ, Whelan BM, Filippi P (2022) Identifying causes of crop yield variability with interpretive machine learning. **Computers and Electronics in Agriculture** 192: 106632. https://doi.org/10.1016/j.compag.2021.106632

Kastner T, Chaudhary A, Gingrich S, Marques A, Persson UM, Bidoglio G, Provost GL, Schwarzmüller F (2021) Global agricultural trade and land system sustainability: Implications for ecosystem carbon storage, biodiversity, and human nutrition. **One Earth** 4(10): 1425-1443. https://doi.org/10.1016/j.oneear.2021.09.006

Lantinga EA, Boele E, Rabbinge R (2013) Maximizing the nitrogen efficiency of a prototype mixed crop-livestock farm in The Netherlands. **NJAS: Wageningen Journal of Life** Sciences 66(1): 15-22. https://doi.org/10.1016/j.njas.2013.07.001

Liebig MA, Tanaka DL, Kronberg SL, Scholljegerdes EJ, Karn JF (2012) Integrated crops and livestock in central North Dakota, USA: Agroecosystem management to buffer soil change. **Renewable Agriculture and Food Systems** 27(2): 115-124. https://doi.org/10.1017/S1742170511000172

Lin L, Gao Z, Liu X (2020) Estimation of soil total nitrogen using the synthetic color learning machine (SCLM) method and hyperspectral data. **Geoderma** 380: 114664. https://doi.org/10.1016/j.geoderma.2020.114664

Linquist BA, Liu L, Van Kessel C, Van Groenigen KJ (2013) Enhanced efficiency nitrogen fertilizers for rice systems: Meta-analysis of yield and nitrogen uptake. **Field Crops Research** 154: 246-254. https://doi.org/10.1016/j.fcr.2013.08.014

Liu L, King JS, Booker FL, Giardina CP, Lee Allen H, Hu S (2009) Enhanced litter input rather than changes in litter chemistry drive soil carbon and nitrogen cycles under elevated CO2: a microcosm study. **Global Change Biology** 15(2): 441-453. https://doi.org/10.1111/j.1365-2486.2008.01747.x

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. **Advances in neural information processing systems** 30.

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Katz R, Himmelfarb J, Bansal N, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. **Nature machine intelligence** 2(1): 56-67. https://doi.org/10.1038/s42256-019-0138-9

Maia NJC, Cruz MCPD, Dubeux Jr JCB, Menegatto LS, Augusto JG, Mendonça GG, Terçariol MC, Oliveira JG, Simili FF (2021) Integrated crop-livestock versus conventional systems: use of soil indicators to detect short-term changes during seasonal variation. **Bragantia** 80. https://doi.org/10.1590/1678-4499.20210127

Mahmoud AA, Elkatatny S, Mahmoud M, Abouelresh M, Abdulraheem A, Ali A (2017) Determination of the total organic carbon (TOC) based on conventional well logs using artificial neural network. **International Journal of Coal Geology** 179: 72-80. https://doi.org/10.1016/j.coal.2017.05.012

Mendonça GG, Simili FF, Augusto JG, Bonacim PM, Menegatto LS, Gameiro AH, (2020) Economic gains from crop-livestock integration in relation to conventional systems. Revista Brasileira de Zootecnia 49: 1-11. https://doi.org/10.37496/rbz4920190029

Monz CA, Reuss DE, Elliott ET (1991) Soil microbial biomass carbon and nitrogen estimates using 2450 MHz microwave irradiation or chloroform fumigation followed by direct extraction. Agriculture, Ecosystems & Environment 34: 55-63. https://doi.org/10.1016/0167-8809(91)90093-D

Moraes A, Carvalho PCF, Anghinoni I, Lustosa SBC, Andrade SEVG, Kunrath TR, (2014) Integrated crop–livestock systems in the Brazilian subtropics. **European Journal of Agronomy** 57: 4-9. https://doi.org/10.1016/j.eja.2013.10.004

Nguyen TT, Pham TD, Nguyen CT, Delfos J, Archibald R, Dang KB, Hoang NB, Guo W, Ngo HH (2022) A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion. **Science of the Total Environment** 804: 150187. https://doi.org/10.1016/j.scitotenv.2021.150187

Oliveira JG, Santana Júnior ML, Maia NJC, Dubeux Jr JCB, Gameiro AH, Kunrath TR, Mendonça GG, Simili FF (2022) Nitrogen balance and efficiency as indicators for monitoring the proper use of fertilizers in agricultural and livestock systems. **Scientific Reports** 12(1): 1-10. https://doi.org/10.1038/s41598-022-15615-7

Osterholz WR, Rinot O, Shaviv A, Linker R, Liebman M, Sanford G, Strock J, Castellano MJ (2017) Predicting gross nitrogen mineralization and potentially mineralizable nitrogen using soil organic matter properties. **Soil Science Society of America Journal** 81(5): 1115-1126. https://doi.org/10.2136/sssaj2017.02.0055

Osterholz WR, Liebman M, Castellano MJ (2018) Can soil nitrogen dynamics explain the yield benefit of crop diversification?. **Field crops research** 219: 33-42. https://doi.org/10.1016/j.fcr.2018.01.026

Pathy A, Meher S, Balasubramanian P (2020) Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods. A**lgal Research** 50: 102006. https://doi.org/10.1016/j.algal.2020.102006

Raij BV, Andrade JC, Cantarella H, Quaggio JA (Eds.) (2001) Análise química para avaliação da fertilidade de solos tropicais. Campinas: Instituto Agronômico, 235p.

Rakkar MK, Blanco-Canqui H, Drijber RA, Drewnoski ME, MacDonald JC, Klopfenstein T (2017) Impacts of cattle grazing of corn residues on soil properties after 16 years. **Soil Science Society of America Journal** 81(2): 414-424. https://doi.org/10.2136/sssaj2016.07.0227

Reda R, Saffaj T, Ilham B, Saidi O, Issam K, Brahim L (2019) A comparative study between a new method and other machine learning algorithms for soil organic carbon and total nitrogen prediction using near infrared spectroscopy. Chemometrics and Intelligent Laboratory Systems. 195: 103873. https://doi.org/10.1016/j.chemolab.2019.103873

Rezende CP, Cantarutti RB, Braga JM, Gomide JA, Pereira JM, Ferreira E, Tarré RM, Macedo R, Alves BJR, Urquiaga S, Cadisch G, Giller KE, Boddey RM (1999) Litter deposition and disappearance in Brachiaria pastures in the Atlantic Forest region of the south of Bahia, Brazil. Nutrient Cycling in Agroecosystems 54: 99-112.

Rütting T, Aronsson H, Delin S (2018) Efficient use of nitrogen in agriculture. **Nutrient cycling in Agroecosystems** 110(1): 1-5. https://doi.org/10.1007/s10705-017-9900-8

Roberts TL, Norman RJ, Slaton NA, Wilson CE, Ross WJ, Bushong JT (2009) Direct steam distillation as an alternative to the Illinois soil nitrogen test. Soil Science Society of America Journal 73: 1268-1275. https://doi.org/10.2136/sssaj2008.0165

Romanyà J, Arco N, Solà-Morales I, Armengot L, Sans FX (2012) Carbon and nitrogen stocks and nitrogen mineralization in organically managed soils amended with composted manures. **Journal of Environmental Quality** 41(4): 1337-1347. https://doi.org/10.2134/jeq2011.0456

Ryschawy J, Liebig MA, Kronberg SL, Archer DW, Hendrickson JR (2017) Integrated crop-livestock management effects on soil quality dynamics in a semiarid region: a typology of soil change over time. **Applied and Environmental Soil Science** 2017. https://doi.org/10.1155/2017/3597416

Scavuzzo CM, Scavuzzo JM, Campero MN, Anegagrie M, Aramendia AA, Benito A, Periago V (2022) Feature importance: Opening a soil-transmitted helminth machine learning model via SHAP. **Infectious Disease Modelling** 7(1): 262-276. https://doi.org/10.1016/j.idm.2022.01.004

Silva EE, Azevedo PHS, De-Polli H (2007) Determinação do Nitrogênio da Biomassa Microbiana do Solo (BMS-N). Seropédica – RJ: EMBRAPA (EMBRAPA. Comunicado técnico, 96).

Sinclair TR, Rufty TW (2012) Nitrogen and water resources commonly limit crop yield increases, not necessarily plant genetics. **Global Food Security** 1(2): 94-98. https://doi.org/10.1016/j.gfs.2012.07.001

Smith P, House JI, Bustamante M, Sobocká J, Harper R, Pan G, West PC, Clark JM, Adhya T, Rumpel C, Paustian K, Kuikman P, Cotrufo MF, Elliott JA, McDowell R, Griffiths RI, Asakawa S, Bondeau A, Jain AK, Meersmans J, Pugh TAM (2016) Global change pressures on soils from land use and management. **Global change biology** 22(3): 1008-1028. https://doi.org/10.1111/gcb.13068

Schjønning P, Thomsen IK, Moldrup P, Christensen BT (2003) Linking soil microbial activity to water-and air-phase contents and diffusivities. **Soil Science Society of America Journal** 67(1): 156-165. https://doi.org/10.2136/sssaj2003.1560

Souza EDD, Costa SEVGDA, Anghinoni I, Lima CVSD, Carvalho PCDF, Martins AP (2010) Biomassa microbiana do solo em sistema de integração lavoura-pecuária em plantio direto, submetido a intensidades de pastejo. **Revista Brasileira de Ciência do solo** 34: 79-88. https://doi.org/10.1590/S0100-06832010000100008

Stevenson FJ, Braids OC (1968) Variation in the relative distribution of amino sugar with depth in some soil profiles. **Soil Science Society of America Journal** 32: 590-598. https://doi.org/10.2136/sssaj1968.03615995003200040049x

Strickland MS, Thomason WE, Avera B, Franklin J, Minick K, Yamada S, Badgley BD, (2019) Short-Term effects of cover crops on soil microbial characteristics and biogeochemical processes across actively managed farms. **Agrosystems, Geosciences & Environment** 2, 1-9. https://doi.org/10.2134/age2018.12.0064

Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. **BMC bioinformatics** 9(1): 1-11. https://doi.org/10.1186/1471-2105-9-307

Taylor P, Mansfield ER, Helms BP, Mansfield ER, Helms BP (1981) Detecting Multicollinearity. **The American Statistician** 36: 1–4. https://doi.org/10.1080/00031305.1982.10482818.

Thakur AK, Rath S, Mandal KG (2013) Differential responses of system of rice intensification (SRI) and conventional flooded-rice management methods to applications of nitrogen fertilizer. **Plant Soil** 370: 59–71. https://doi.org/10.1007/s11104-013-1612-5

Tran VQ (2022) Machine learning approach for investigating chloride diffusion coefficient of concrete containing supplementary cementitious materials. **Construction and Building Materials** 328: 127103. https://doi.org/10.1016/j.conbuildmat.2022.127103

Veldkamp E (1994) Organic Carbon Turnover in Three Tropical Soils under Pasture after Deforestation. Soil Science Society of America Journal 58: 175-180.

Xuejun L, Fusuo Z (2011) Nitrogen fertilizer induced greenhouse gas emissions in China. **Current Opinion in Environmental Sustainability** 3(5): 407-413. https://doi.org/10.1016/j.cosust.2011.08.006

Walkley A, Black IA (1934) An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic soil titration method. Soil Science 37: -38.

Wei L, Yuan Z, Wang Z, Zhao L, Zhang Y, Lu X, Cao L (2020) Hyperspectral Inversion of Soil Organic Matter Content Based on a Combined Spectral Index Model. **Sensors** 20(10): 2777. https://doi.org/10.3390/s20102777

Weil RR, Islam KR, Stine MA, Gruver JB, Samson-Liebig SE (2003) Estimating active carbon for soil quality assessment: a simplified method for laboratory and field use. America Journal of Alternative Agriculture 18: 3-17. https://doi.org/10.1079/AJAA200228

Willmott CJ (1981) On the validation of models. Physical geography 2(2): 184-194.

Yan F, Song K, Liu Y, Chen S, Chen J (2020) Predictions and mechanism analyses of the fatigue strength of steel based on machine learning. **Journal of Materials Science** 55(31): 15334-15349. https://doi.org/10.1007/s10853-020-05091-7

Yang H, Huang K, Zhang K, Weng Q, Zhang H, Wang F (2021) Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities. **Environmental Science & Technology** 55(20): 14316-14328. https://doi.org/10.1021/acs.est.1c02479

Zhang T, Luo Y, Chen HY, Ruan H (2018) Responses of litter decomposition and nutrient release to N addition: A meta-analysis of terrestrial ecosystems. **Applied Soil Ecology** 128: 35-42. https://doi.org/10.1016/j.apsoil.2018.04.004

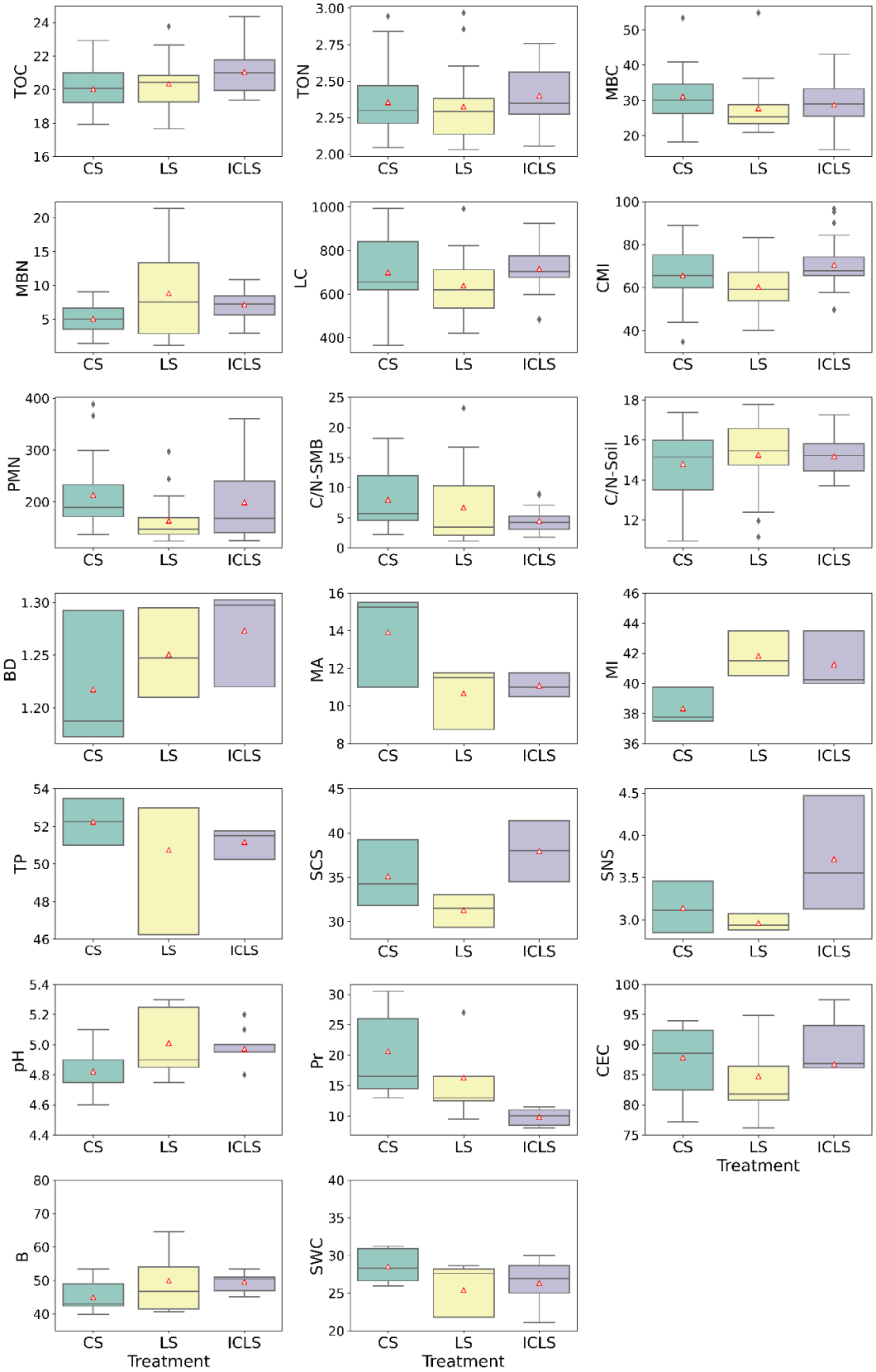
Zhang S, Zheng Q, Noll L, Hu Y, Wanek W (2019) Environmental effects on soil microbial nitrogen use efficiency are controlled by allocation of organic nitrogen to microbial growth and regulate gross N mineralization. **Soil Biology and Biochemistry** 135: 304-315. https://doi.org/10.1016/j.soilbio.2019.05.019

Zhao Y, Gao G, Ding G, Wang L, Chen Y, Zhao Y, Yu M, Zhang Y (2022) Assessing the influencing factors of soil susceptibility to wind erosion: A wind tunnel experiment with a machine learning and model-agnostic interpretation approach. **Catena** 215: 106324. https://doi.org/10.1016/j.catena.2022.106324

Zhou W, Li H, Wen S, Xie L, Wang T, Tian Y, Yu W (2022) Simulation of Soil Organic Carbon Content Based on Laboratory Spectrum in the Three-Rivers Source Region of China. **Remote Sensing** 14(6): 1521. https://doi.org/10.3390/rs14061521

APÊNDICES

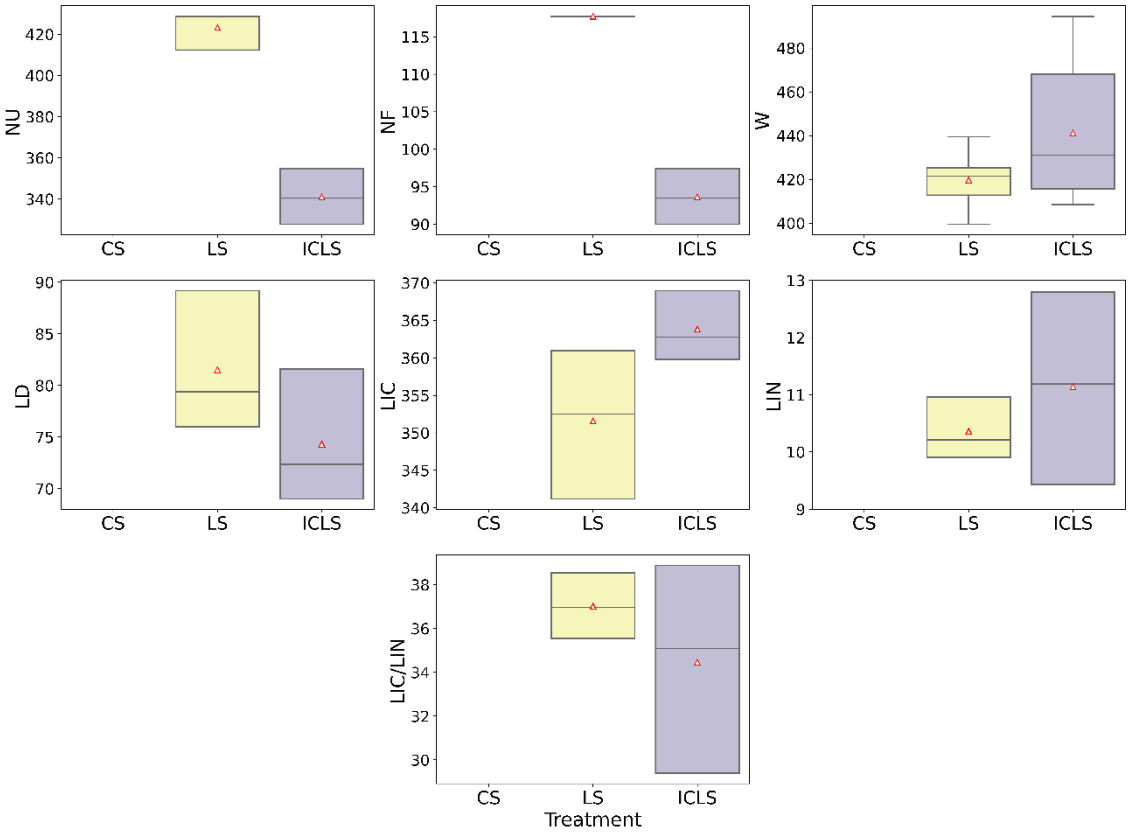
Apêndice 1. Média das variáveis de solo descritas na Tabela 1.



Legenda do gráfico: (- internal) representa o segundo quartil ou mediana; (□) primeiro e terceiro quartil; (⊤ and ⊥) limite inferior e superior; (△) valores médios. CS: monocultivo de Milho; LS: monocultivos de

Pecuária; ICLS: Integração Lavoura-Pecuária. TOC: carbono orgânico total; TON: nitrogênio orgânico total; MBC: carbono da biomassa microbiana; MBN: nitrogênio da biomassa microbiana; LC: carbono lábil; CMI: índice de manejo do carbono; PMN: nitrogênio potencialmente mineralizável; C/N-soil: relação carbono e nitrogênio do solo; C/N-SMB: relação carbono e nitrogênio da biomassa microbiana do solo; BD: densidade do solo; MA: macroporosidade; MI: microporosidade; TP: porosidade total; SCS: estoque de carbono; SNS: estoque de nitrogênio; pH: potencial de hidrogênio; Pr: fósforo; CEC: capacidade de troca de cátions; B: saturação de bases; SWC: potencial de água no solo.

Apêndice 2. Média das variáveis de planta e animal descritas na Tabela 1. O tratamento CS não foi medido porque não havia presença de animais, logo, foi avaliada a liteira somente nos tratamentos onde havia pastagem (LS e ICLS).



Legenda do gráfico: (- internal) representa o segundo quartil ou mediana; (□) primeiro e terceiro quartil; (⊤ and ⊥) limite inferior e superior; (∆) valores médios. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária. NU: nitrogênio da urina; NF: nitrogênio das fezes; W: peso animal; LD: liteira depositada; LIC: carbono da liteira; LIN: nitrogênio da liteira; LIC/LIN: relação carbono e nitrogênio da liteira.

Apêndice 3. Variáveis climáticas e detalhamento por dezoito decêndios. Para cada tratamento as datas dos decêndios variaram de acordo a data de plantio, manejo adotado e as datas de amostragens de solo.

	Variáveis Climáticas em decêndios								
Decêndio	**T**	**P**	**U**	**RH**	**RN**	**PET**	**STO**	**DEF**	**SUR**
1	T-1	P-1	U-1	RH-1	RN-1	PET-1	STO-1	DEF-1	SUR-1
2	T-2	P-2	U-2	RH-2	RN-2	PET-2	STO-2	DEF-2	SUR-2
3	T-3	P-3	U-3	RH-3	RN-3	PET-3	STO-3	DEF-3	SUR-3
4	T-4	P-4	U-4	RH-4	RN-4	PET-4	STO-4	DEF-4	SUR-4
5	T-5	P-5	U-5	RH-5	RN-5	PET-5	STO-5	DEF-5	SUR-5
6	T-6	P-6	U-6	RH-6	RN-6	PET-6	STO-6	DEF-6	SUR-6
7	T-7	P-7	U-7	RH-7	RN-7	PET-7	STO-7	DEF-7	SUR-7
8	T-8	P-8	U-8	RH-8	RN-8	PET-8	STO-8	DEF-8	SUR-8
9	T-9	P-9	U-9	RH-9	RN-9	PET-9	STO-9	DEF-9	SUR-9
10	T-10	P-10	U-10	RH-10	RN-10	PET-10	STO-10	DEF-10	SUR-10
11	T-11	P-11	U-11	RH-11	RN-11	PET-11	STO-11	DEF-11	SUR-11
12	T-12	P-12	U-12	RH-12	RN-12	PET-12	STO-12	DEF-12	SUR-12
13	T-13	P-13	U-13	RH-13	RN-13	PET-13	STO-13	DEF-13	SUR-13
14	T-14	P-14	U-14	RH-14	RN-14	PET-14	STO-14	DEF-14	SUR-14
15	T-15	P-15	U-15	RH-15	RN-15	PET-15	STO-15	DEF-15	SUR-15
16	T-16	P-16	U-16	RH-16	RN-16	PET-16	STO-16	DEF-16	SUR-16
17	T-17	P-17	U-17	RH-17	RN-17	PET-17	STO-17	DEF-17	SUR-17
18	T-18	P-18	U-18	RH-18	RN-18	PET-18	STO-18	DEF-18	SUR-18

T: temperatura (°C), P: precipitação (mm), U: velocidade do vento (m s^{-1}), RH: umidade relativa (%), RN: radiação líquida (MJ m^{-2} d^{-1}), PET: evapotranspiração potencial (mm), STO: armazenamento de água no solo (mm), DEF: déficit de água no solo (mm mo^{-1}), SUR: excedente de água no solo (mm mo^{-1}).

Apêndice 4. Parâmetros com os valores padrão pelo método de GridSearch para cada modelo, e os melhores valores ajustados para cada tratamento. CS: monocultivo de Milho; LS: monocultivos de Pecuária; ICLS: Integração Lavoura-Pecuária.

		Carbono			**Nitrogênio**		
Parâmetros	GridSearch	CS	LS	ICLS	CS	LS	ICLS
Redes Neurais Multilayer Perceptron - MLP							
hidden_layer_sizes	10 to 150	50, 100	10,10,10	10,10,10	50,100,10	10,10,10	10, 20
n_layers	1 to 5	2	3	3	3	3	2
solver	adam, lbfgs	lbfgs	lbfgs	lbfgs	adam	lbfgs	lbfgs
learning_rate	constant, adaptive, invscaling	constant	constant	constant	constant	constant	constant
activation	relu	relu	relu	relu	relu	relu	relu
max_iter	1000	1000	1000	1000	1000	1000	1000
Random Forest Regressor - RF							
max_depth	2 to 10	2	2	2	2	2	3
n_estimators	1 to 300	1	2	110	12	10	10
max_features	sqrt, auto	auto	sqrt	sqrt	auto	sqrt	auto
min_samples_split	2 to 20	5	2	4	2	2	2
min_samples_leaf	1 to 20	2	2	1	1	2	1
bootstrap	True, False	False	False	False	True	False	True
Support Vector Regression - SVR							
C	0.0001 to 50	0.1	16	1	5	0.1	0.11
kernel	rbf, poly, sigmoid, linear	linear	poly	poly	poly	linear	sigmoid
degree	0.0001 to 10	1	2	1	5	0.001	1
gamma	auto, scale	auto	auto	auto	auto	auto	scale
K Neighbors Regressor - KNN							
n_neighbors	1 to 4	2	2	2	4	2	2
leaf_size	1 to 50	1	1	1	1	1	1
p	1 to 5	1	1	1	2	3	2
algorithm	auto, ball_tree, kd_tree, brute	auto	auto	auto	auto	auto	auto
Adaptive Boosting Regressor - AdaBoost							
n_estimators	1 to 100	100	100	100	100	100	30
learning_rate	0.01 to 10	0.01	0.05	0.01	0.05	0.01	0.05
loss	linear, square, exponential	linear	exponential	square	square	square	square

Apêndice 5. Parâmetros com os valores padrão pelo método de GridSearch e os melhores valores ajustados para cada modelo, no ecossistema de programação em Python.

Parâmetros	GridSearch	Ajuste
Random Forest Regressor - RF		
max_depth	2 to 5	4
n_estimators	50 to 100	50
max_features	sqrt, auto	sqrt
criterion	default	squared_error
min_samples_split	1 to 12	8
min_samples_leaf	1 to 5	2
bootstrap	True, False	False
Adaptive Boosting Regressor - Adaboost		
n_estimators	1 to 100	20
learning_rate	0.01 to 10	0.1
loss	linear, square, exponential	exponential
eXtreme Gradient Boosting Regressor - XGBoost		
max_depht	1 to 10	3
n_estimators	100 to 1000	100
learning_rate	0.01 to 0.1	0.1
colsample_bytree	0.3 to 1	0.7