ESTIMATIVA DO TEOR DE CARBONO ORGÂNICO DO SOLO USANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Ana Paula Sousa Santos[1], Maria Clara das Neves Palhete dos Santos[1], Rodrigo Nogueira Martins[2]

[1]Instituto Federal do Norte de Minas Gerais (IFNMG) – Campus Araçuaí, Estrada São José do Caraí, nº 385 - CEP 39.606-270 - Bairro Universitários, Araçuaí – MG, Brasil, apss3@aluno.ifnmg.edu.br, mcdnpds@aluno.ifnmg.edu.br.

[2]Instituto Federal do Norte de Minas Gerais (IFNMG) – Campus Salinas, Fazenda Varginha, Km 02, Rod. Salinas/Taiobeiras, Zona Rural — CEP: 39560-000, Salinas – MG, Brasil, rodrigo.nogueira@ifnmg.edu.br.

Resumo

O carbono orgânico do solo (SOC) é um importante indicador da qualidade e do funcionamento dos ecossistemas agrícolas. Nesse sentido, este estudo objetivou estimar o SOC em um Latossolo Vermelho Distrófico, na camada de 0–20 cm, integrando variáveis físico-químicas, espectrais e topográficas. Foram avaliadas argila, pH, condutividade elétrica aparente (CEa), índice de vegetação por diferença normalizada (NDVI) e altitude. O SOC apresentou média de 0,83% e coeficiente de variação de 31,33%. Houve correlação positiva forte entre SOC e argila ($r = 0,86$; $p < 0,01$) e correlações positivas moderadas com CEa ($r = 0,56$; $p < 0,01$) e NDVI ($r = 0,67$; $p < 0,01$). A altitude apresentou correlação negativa ($r = -0,37$; $p < 0,01$), enquanto o pH não apresentou correlação significativa. A regressão linear múltipla e a máquina de vetores de suporte (SVM) com kernel linear apresentaram os melhores desempenhos ($R^2 = 0,73$; RMSE = 0,11%). Os resultados evidenciam o potencial da integração dessas variáveis para estimar o SOC e apoiar estratégias de manejo e conservação do solo na agricultura de precisão.

Palavras-chave: Agricultura digital. SOC. *Support vector machine*. Inteligência artificial.

Área do Conhecimento: Ciências agrárias – Engenharia agrícola.

Introdução

O carbono orgânico do solo (SOC) desempenha papel fundamental no funcionamento dos ecossistemas terrestres, estando diretamente relacionado à fertilidade, à estrutura física, à disponibilidade de nutrientes e à produtividade agrícola (Liptzin et al., 2022). Sua distribuição e dinâmica espacial são influenciadas por diferentes fatores ambientais e de manejo, incluindo clima, vegetação, classe de solo, topografia, uso da terra e propriedades físico-químicas do solo (Mahmoudzadeh et al., 2020).

O monitoramento do SOC em áreas agrícolas é relevante tanto para a manutenção da produtividade quanto para o fortalecimento de estratégias de conservação dos recursos naturais. A quantificação e o acompanhamento de seus estoques podem contribuir para a identificação de áreas com diferentes condições de qualidade do solo, além de subsidiar práticas de manejo voltadas à conservação da matéria orgânica e ao armazenamento de carbono (Zhang et al., 2025).

Nesse contexto, a agricultura de precisão e a agricultura digital têm ampliado as possibilidades de caracterização da variabilidade espacial dos atributos do solo. O emprego de dados provenientes de sensores proximais e orbitais, associado a técnicas de análise espacial, possibilita a obtenção de informações em diferentes escalas e pode auxiliar na identificação de padrões de variabilidade dentro das áreas agrícolas. Entre essas ferramentas, destacam-se os índices de vegetação, como o *Normalized Difference Vegetation Index* (NDVI), e a condutividade elétrica aparente do solo (CEa), que têm sido empregados na caracterização da variabilidade espacial do solo e da cobertura vegetal (Serrano et al., 2016; Esteves et al., 2022).

A integração dessas informações permite a delimitação de zonas de manejo, entendidas como subdivisões da área agrícola que apresentam características relativamente homogêneas. Essa abordagem pode contribuir para o planejamento da amostragem do solo, possibilitando maior

representatividade espacial com potencial redução do número de amostras e, consequentemente, dos custos associados às análises laboratoriais.

Apesar dessas vantagens, a caracterização detalhada dos atributos de fertilidade e qualidade do solo ainda depende, em grande medida, da coleta de amostras e de análises laboratoriais. Embora o aumento da densidade amostral possa melhorar a caracterização da variabilidade espacial, essa estratégia também implica maior demanda de recursos financeiros e operacionais (Silva et al., 2021). Dessa forma, torna-se relevante o desenvolvimento de métodos que conciliem precisão, capacidade preditiva e viabilidade operacional.

Nesse cenário, técnicas de aprendizado de máquina vêm sendo utilizadas de forma crescente na agricultura de precisão e no mapeamento digital de solos. Entre os métodos empregados destacam-se as redes neurais artificiais (ANN), as máquinas de vetores de suporte (SVM), os modelos de regressão linear, o random forest (RF) e as árvores de decisão (DT), os quais apresentam diferentes capacidades de representar relações entre variáveis ambientais e atributos do solo (Breiman, 2001; Szatmári et al., 2021). Estudos comparativos também demonstram o potencial de algoritmos como SVM e RF na predição do carbono orgânico do solo (Were et al., 2015; Meliho et al., 2023). Diante desse contexto, este estudo teve como objetivo integrar informações físico-químicas, espectrais e topográficas para estimar o teor de carbono orgânico do solo por meio de diferentes algoritmos de aprendizado de máquina, buscando identificar as variáveis com maior associação ao SOC e avaliar o desempenho de modelos de diferentes níveis de complexidade.

Metodologia

O estudo foi realizado em uma área comercial de produção de grãos de 106 ha, localizada em Chapadão do Céu, Goiás, Brasil (18°20′10″ S e 52°37′12″ W), cujo solo é classificado como Latossolo Vermelho Distrófico (EMBRAPA, 2013). Foram utilizados dados do banco disponibilizado por Schmidinger et al. (2025), composto por 112 amostras coletadas na profundidade de 0–20 cm, juntamente com informações topográficas e espectrais dos respectivos pontos. As variáveis analisadas foram carbono orgânico do solo (SOC), teor de argila, pH, altitude, condutividade elétrica aparente (CEa) e índice de vegetação por diferença normalizada (NDVI).

O SOC foi determinado pelo método de Walkley e Black (1934), a argila conforme Gee e Bauder (1986) e o pH segundo a Embrapa (2017). A altitude foi obtida de modelo digital de elevação derivado da missão *Shuttle Radar Topography Mission* (SRTM), enquanto a CEa foi determinada com a plataforma Veris U3, utilizando discos condutivos e profundidade de 0–30 cm. O NDVI, proposto por Rouse et al. (1973), foi calculado a partir de imagem Sentinel-2 de maio de 2022.

Inicialmente, realizou-se estatística descritiva, considerando valores mínimo, médio, máximo, desvio padrão e coeficiente de variação (CV), classificado conforme Warrick e Nielsen (1980) em baixa (CV < 12%), moderada (12% ≤ CV < 60%) e alta (CV ≥ 60%). As associações entre o SOC e as demais variáveis foram avaliadas pelo coeficiente de correlação de *Pearson*. Para a modelagem, os dados foram divididos aleatoriamente em 79 amostras para treinamento (70%) e 33 para teste (30%). Foram avaliados seis métodos: regressão linear simples, regressão linear múltipla, *random forest* (RF), máquina de vetores de suporte com kernel linear (SVM), árvore de decisão (DT) e redes neurais artificiais (ANN). Nas regressões, utilizou-se o método *best subset regression* para seleção das variáveis, com o pacote *olsrr* (HEBBALI, 2017). Os modelos foram implementados no R 3.6.3, utilizando os pacotes *randomForest*, *e1071*, *nnet*, *caret* e *olsrr* (Breiman, 2001; Kuhn, 2008; Meyer et al., 2019; Venables; Ripley, 2002). O desempenho preditivo foi avaliado no conjunto de teste por meio do coeficiente de determinação (R^2) e da raiz do erro quadrático médio (RMSE).

Resultados

A análise estatística descritiva dos atributos físico-químicos, espectrais e topográficos evidenciou variabilidade entre os pontos amostrados (Tabela 1). O teor de SOC variou de 0,27 a 1,62%, apresentando média de 0,83%, desvio padrão de 0,26% e CV de 31,33%, indicando variabilidade moderada. O teor de argila apresentou valores entre 16,30 e 62,00%, com média de 35,65% e CV de 33,42%, também caracterizando variabilidade moderada. O pH apresentou média de 6,11 e baixa dispersão, com CV de 4,55%. A altitude variou de 413,80 a 439,87 m, com média de 428,82 m e CV de 1,44%, indicando baixa variabilidade espacial. A CEa apresentou média de 2,67 dS m^{-1}, variando

de 1,17 a 4,90 dS m^{-1}, com CV de 27,83%. O NDVI apresentou média de 0,55, com valores entre 0,40 e 0,62 e CV de 9,79%, indicando baixa variabilidade.

Tabela 1 - Estatística descritiva dos atributos físico-químicos e topográficos do solo.

Variável	n	Média	Valor Mínimo	Valor Máximo	DP	CV (%)
SOC (%)	112	0,83	0,27	1,62	0,26	31,33
Argila (%)	112	35,65	16,30	62,00	11,91	33,42
pH	112	6,11	5,76	7,78	0,28	4,55
Altitude (m)	112	428,82	413,80	439,87	6,15	1,44
CEa (dS/m)	112	2,67	1,17	4,90	0,74	27,83
NDVI	112	0,55	0,40	0,62	0,05	9,79

n, número de amostras; DP, desvio padrão; CV, coeficiente de variação; SOC, Carbono orgânico do solo; CEa, condutividade elétrica aparente. Fonte: elaborado pelos autores (2026).

A análise de correlação de *Pearson* revelou associações distintas entre o SOC e as variáveis físico-químicas, espectrais e topográficas (Tabela 2). O SOC apresentou forte correlação positiva com o teor de argila ($r = 0,86$; $p < 0,01$), evidenciando uma associação expressiva entre essas variáveis. Também foram observadas correlações positivas e moderadas entre o SOC e a CEa ($r = 0,56$; $p < 0,01$) e entre o SOC e o NDVI ($r = 0,67$; $p < 0,01$). A altitude apresentou correlação negativa e significativa com o SOC ($r = -0,37$; $p < 0,01$), indicando tendência de redução dos teores de carbono orgânico nas áreas de maior altitude. O pH, por sua vez, não apresentou correlação estatisticamente significativa com o SOC.

Tabela 2 - Correlação de *Pearson* dos atributos físico-químicos e topográficos do solo..

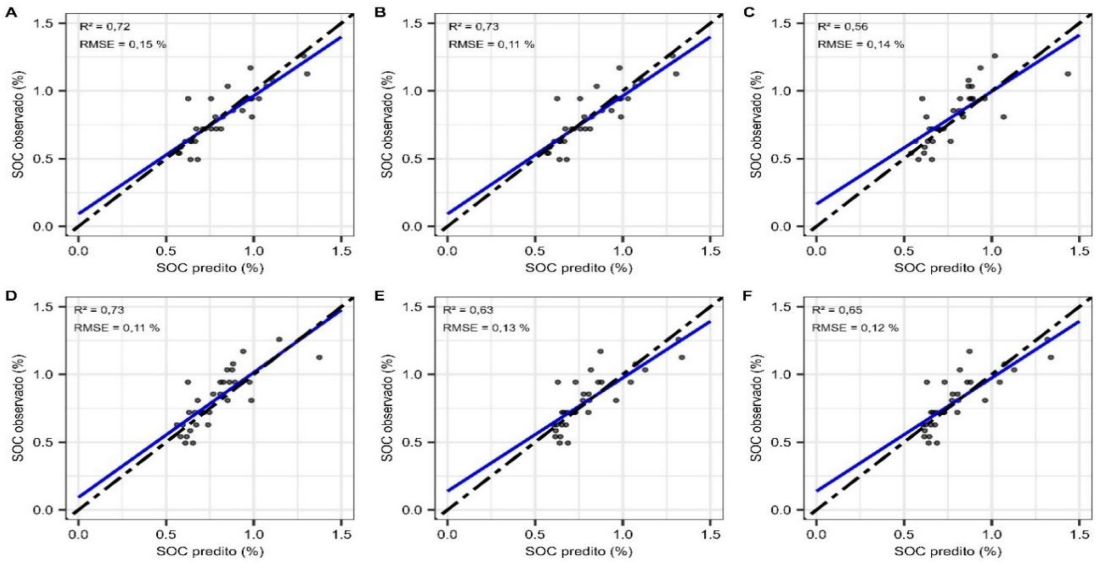
	SOC	Argila	pH	Altitude	CEa	NDVI
SOC	-	-	-	-	-	-
Argila	0.86**	-	-	-	-	-
pH	0.15^{ns}	0.16^{ns}	-	-	-	-
Altitude	-0.37**	-0.45**	0.18^{ns}	-	-	-
CEa	0.56**	0.79**	0.33**	-0.30*	-	-
NDVI	0.67**	0.76**	0.11^{ns}	-0.45**	0.65**	-

SOC, Carbono orgânico do solo; CEa, Condutividade elétrica aparente; ** e * Correlações significativas pelo teste *t* ao nível de 1% e 5% ($p < 0,01$; $p < 0,05$); ^{ns}, Correlação não significativa. Fonte: elaborado pelos autores (2026).

A avaliação dos modelos de estimativa do SOC demonstrou diferenças no desempenho dos algoritmos empregados (Figura 1). A regressão linear múltipla e o SVM com kernel linear apresentaram os melhores resultados, ambos com $R^2 = 0,73$ e RMSE = 0,11%. A regressão linear simples também apresentou desempenho satisfatório, com $R^2 = 0,72$ e RMSE = 0,15%, demonstrando que um modelo de menor complexidade foi capaz de explicar parcela expressiva da variabilidade observada no SOC.

Entre os métodos não lineares, a árvore de decisão (DT) apresentou desempenho intermediário, com $R^2 = 0,63$ e RMSE = 0,13%. O RF e a ANN apresentaram os menores valores de R^2, correspondentes a 0,56 e 0,55, respectivamente. De modo geral, os resultados demonstraram que os modelos lineares, particularmente a regressão linear múltipla e o SVM com kernel linear, apresentaram maior consistência na estimativa do SOC. Esse comportamento sugere que as relações entre as variáveis físico-químicas, espectrais e topográficas consideradas no estudo apresentam, para esse conjunto de dados, comportamento predominantemente linear.

Figura 1 - Gráficos de dispersão dos valores previstos versus observados de carbono orgânico do solo (SOC), obtidos utilizando os modelos de regressão linear simples (A), regressão linear múltipla (B), *random forest* (C), máquina de vetores de suporte com kernel linear (D), *decision tree* (E) e redes neurais artificiais (F).



Fonte: elaborado pelos autores (2026).

Discussão

Os resultados demonstram o potencial da integração de diferentes fontes de informação para a estimativa do carbono orgânico do solo no contexto da agricultura digital. A aplicação de técnicas de aprendizado de máquina e a utilização conjunta de variáveis físico-químicas, espectrais e topográficas constituem estratégias relevantes para o mapeamento digital de atributos do solo (Sakhaee et al., 2022; Zhu et al., 2022). O teor médio de SOC observado foi de 0,83%, com CV de 31,33%, caracterizando variabilidade moderada. A presença de variabilidade espacial do SOC é esperada em áreas agrícolas, uma vez que sua distribuição está relacionada a fatores associados ao material de origem, propriedades físicas do solo, relevo, vegetação e práticas de manejo (Mahmoudzadeh et al., 2020).

Entre os atributos analisados, o teor de argila apresentou a maior correlação com o SOC (r = 0,86). Esse resultado reforça a importância da fração argila na estabilização e proteção da matéria orgânica do solo. As partículas de menor tamanho apresentam elevada superfície específica e podem favorecer a associação da matéria orgânica com a fração mineral, contribuindo para sua proteção contra processos de decomposição (Carvalho et al., 2023). A elevada associação entre argila e SOC também indica que a argila apresenta potencial como variável preditora em modelos de estimativa do SOC. A utilização desse atributo pode ser particularmente relevante em abordagens de mapeamento digital, nas quais variáveis com forte relação com o atributo de interesse podem contribuir para melhorar a capacidade preditiva dos modelos (Zhang et al., 2025).

O NDVI apresentou correlação positiva e moderada com o SOC (r = 0,67), indicando associação entre o vigor da cobertura vegetal e os teores de carbono orgânico. Esse resultado é compatível com estudos que demonstram o potencial de variáveis derivadas do sensoriamento remoto na caracterização e predição de atributos relacionados ao carbono do solo (Wang et al., 2021). A maior atividade e cobertura vegetal podem estar associadas a maior aporte de resíduos orgânicos ao solo, estabelecendo uma relação entre a condição da vegetação e a dinâmica do carbono.

A CEa também apresentou correlação positiva e moderada com o SOC (r = 0,56). A CEa é influenciada por diferentes propriedades do solo, incluindo textura, umidade e concentração de íons. Assim, sua utilização como variável auxiliar pode contribuir para a identificação da variabilidade espacial de atributos do solo e para a delimitação de ZMs (Moral et al., 2010). A correlação positiva observada entre CEa e argila (r = 0,79) reforça a possibilidade de que parte da relação entre CEa e SOC esteja associada à influência da textura do solo. A altitude apresentou correlação negativa com o SOC (r = −0,37), sugerindo tendência de menores teores de SOC nas áreas mais elevadas. A influência da topografia sobre a distribuição do SOC pode estar relacionada a processos de redistribuição de partículas e matéria orgânica, erosão, deposição e diferenças nas condições de desenvolvimento da vegetação (Mahmoudzadeh et al., 2020).

Em relação ao desempenho dos modelos, a regressão linear múltipla e o SVM com kernel linear apresentaram os melhores resultados. Isso demonstra que modelos de menor complexidade podem

apresentar desempenho equivalente ou superior ao de algoritmos mais complexos quando as relações entre as variáveis são predominantemente lineares. Resultados semelhantes são discutidos na literatura de mapeamento digital de solos, na qual a escolha do algoritmo deve considerar não apenas sua complexidade, mas também as características e a estrutura do conjunto de dados (Wadoux et al., 2020). A regressão linear simples também apresentou desempenho próximo aos melhores modelos, com R^2 = 0,72 e RMSE = 0,15%. Esse resultado é particularmente relevante porque demonstra que uma abordagem mais simples pode apresentar capacidade preditiva satisfatória quando existe uma variável explicativa com forte associação com o atributo de interesse. Por outro lado, os modelos RF e ANN apresentaram os menores valores de R^2, respectivamente 0,56 e 0,55. Esse desempenho inferior pode estar relacionado às características do conjunto de dados e ao número relativamente limitado de amostras disponíveis para treinamento. Algoritmos mais complexos podem demandar maior quantidade e diversidade de dados para representar adequadamente relações não lineares e reduzir o risco de sobreajuste (Li et al., 2019).

Os resultados reforçam, portanto, que a maior complexidade de um algoritmo não implica necessariamente maior capacidade preditiva. A seleção do método deve considerar a estrutura dos dados, o número de observações, a relação entre as variáveis explicativas e a variável resposta e os objetivos específicos da modelagem. Nesse sentido, a integração de atributos físico-químicos, espectrais e topográficos constitui uma estratégia promissora para o desenvolvimento de ferramentas de apoio à tomada de decisão na agricultura de precisão.

Conclusão

A integração de variáveis físico-químicas, espectrais e topográficas mostrou-se eficaz para estimar o SOC. O teor de argila foi a variável mais correlacionada com o SOC, seguido do NDVI e da CEa, enquanto a altitude apresentou associação negativa e o pH não foi significativo. A regressão linear múltipla e o SVM com kernel linear obtiveram os melhores desempenhos (R^2 = 0,73; RMSE = 0,11%), superando os métodos não lineares testados. Isso evidencia que, para o conjunto de dados analisado, modelos lineares foram suficientes. Conclui-se que a integração de variáveis é uma ferramenta viável para o mapeamento espacial do SOC e para o planejamento de estratégias de manejo e conservação do solo.

Referências

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. Disponível em: https://doi.org/10.1023/A:1010933404324. Acesso em: 2 ago. 2025.

CARVALHO, M. L. et al. Stabilization of organic matter in soils: drivers, mechanisms, and analytical tools – a literature review. **Revista Brasileira de Ciência do Solo**, v. 47, e0230130, 2023.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Sistema brasileiro de classificação de solos**. 3. ed. rev. e ampl. Brasília, DF: Embrapa, 2013. 353 p.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA. **Manual de métodos de análise de solo**. 2. ed. rev. atual. Rio de Janeiro: Embrapa Solos, 2017. 574 p.

ESTEVES, C. et al. Assessing the contribution of ECa and NDVI in the delineation of management zones in a vineyard. **Agronomy**, v. 12, n. 6, p. 1331, 2022.

GEE, G. W.; BAUDER, J. W. Particle-size analysis. **In**: KLUTE, A. (ed.). Methods of soil analysis: Part 1 – Physical and mineralogical methods. 2. ed. Madison: American Society of Agronomy; Soil Science Society of America, 1986. p. 383–411.

HEBBALI, A. **Package 'olsrr'. Version 0.5-3**. 2017. Disponível em: https://cloud.r-project.org/web/packages/olsrr/olsrr.pdf. Acesso em: 2 ago. 2025.

KUHN, M. Caret Package. **Journal of Statistical Software**, v. 28, n. 5, 2008.

LI, J. et al. Application of machine learning methods to spatial interpolation of environmental variables. Environmental Modelling & Software, v. 119, p. 285–301, 2019.

LIPTZIN, D. et al. An evaluation of carbon indicators of soil health in long-term agricultural experiments. **Soil Biology and Biochemistry**, v. 172, 108708, 2022.

MAHMOUDZADEH, H. et al. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. **Geoderma Regional**, v. 21, e00260, 2020.

MELIHO, M. et al. Spatial prediction of soil organic carbon stock in the Moroccan High Atlas using machine learning. **Remote Sensing**, v. 15, n. 10, 2494, 2023.

MEYER, D. et al. Package "e1071". **R News**, 2019. Disponível em: https://cran.r-project.org/package=e1071. Acesso em: 2 ago. 2025.

MORAL, F. J. et al. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. **Soil and Tillage Research**, v. 106, n. 2, p. 335–343, 2010.

R CORE TEAM. R: **A Language and Environment for Statistical Computing**. Vienna: R Foundation for Statistical Computing, 2020.

ROUSE, J. W. et al. Monitoring vegetation systems in the Great Plains with ERTS (Earth Resources Technology Satellite). **In**: Earth Resources Technology Satellite-1 Symposium, 3., 1973, Washington, DC. Proceedings... Washington, DC: NASA, Goddard Space Flight Center, 1973.

SAKHAEE, E. et al. Spatial prediction of organic carbon in German agricultural topsoils using machine learning algorithms. **Soil**, v. 8, n. 2, p. 587–604, 2022.

SERRANO, J. M. et al. Monitoring of soil organic carbon over 10 years in a Mediterranean silvo-pastoral system: potential evaluation for differential management. **Precision Agriculture**, v. 17, n. 3, p. 274–295, 2016.

SILVA, S. A. et al. Artificial neural networks in the prediction of soil chemical attributes using apparent electrical conductivity. **Spanish Journal of Agricultural Research**, v. 19, n. 3, c0208, 2021.

SZATMÁRI, G.; PÁSZTOR, L.; HEUVELINK, G. B. Estimating soil organic carbon stock change at multiple scales using machine learning and multivariate geostatistics. **Geoderma**, v. 403, 115356, 2021.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. 4. ed. New York: Springer-Verlag, 2002. Disponível em: http://www.stats.ox.ac.uk/pub/MASS4. Acesso em: 2 ago. 2025.

WADOUX, A. M. C. et al. Machine learning for digital soil mapping: applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, 103359, 2020.

WANG, S. et al. Prediction potential of remote sensing-related variables in the topsoil organic carbon density of Liaohekou coastal wetlands, northeast China. **Remote Sensing**, v. 13, n. 20, 4106, 2021.

WARRICK, A. W.; NIELSEN, D. R. **Spatial variability of soil physical properties in the field**. In: HILLEL, D. (ed.). Applications of soil physics. New York: Academic Press, 1980. p. 319–344.

WERE, K. et al. Machine learning for mapping soil organic carbon: a comparison of random forest, support vector regression and artificial neural networks. **Environmental Modelling & Software**, v. 70, p. 241–250, 2015.

ZHANG, W. et al. Enhancing spectral estimation accuracy of soil organic carbon by using geographic region and clay content as covariates. **Journal of Environmental Management**, v. 391, 126571, 2025.

ZHU, C. et al. Digital mapping of soil organic carbon based on machine learning and regression kriging. **Sensors**, v. 22, n. 22, 8997, 2022.

Agradecimentos

Os autores agradecem ao Instituto Federal do Norte de Minas Gerais (IFNMG) – Campi Araçuaí e Salinas pelo apoio à pesquisa.