Métodos de avaliação da série temporal do estoque de carbono orgânico do solo brasileiro (1985-2022)

Time series evaluation methods of the organic carbon stock of Brazilian soil (1985-2022)

Marcos Vinícius Souza Cardoso[1], Aline Mari Huf dos Reis[2], Alessandro Samuel-Rosa[3], Taciara Zborowski Horst[4]

RESUMO

O solo desempenha papel crucial no sistema climático global e no bem-estar humano, sendo um importante reservatório de carbono. Para mapear sua distribuição, foram testadas várias abordagens, mas a escassez de dados de solo resultou em produtos estáticos que representam décadas de dados coletados. Recentemente, foram lançados mapas anuais de carbono orgânico no solo no Brasil (coleção beta), representando um avanço no entendimento de sua dinâmica. Entretanto, a qualidade desse produto ainda não foi avaliada devido a restrições no número e distribuição dos dados de solo disponíveis. Este trabalho visa explicitar como a qualidade desses mapas será avaliada utilizando métodos quantitativos e qualitativos como estratégias para superar essas limitações. Dentre os métodos quantitativos, será realizada a validação cruzada espacial e explicitação da incerteza espacial. Também serão explorados métodos de comparação de mapas legados detalhados. Os resultados serão sintetizados no relatório que acompanhará o documento técnico da coleção 1 do MapBiomas Solo, enriquecendo a compreensão da dinâmica do solo brasileiro. O relatório servirá como uma base sólida para analisar as estimativas de COS, tornando a série temporal mais confiável para tomadas de decisões relacionadas ao solo. Impactando positivamente na gestão sustentável de recursos e no desenvolvimento de políticas ambientais mais eficazes.
PALAVRAS-CHAVE: Dados legados; Incerteza; Métricas de avaliação; Validação Cruzada.

ABSTRACT

Soil plays a critical role in the global climate system and human well-being, serving as a significant carbon reservoir. Various approaches have been tested to map its distribution, but the scarcity of soil data has led to the creation of static products representing decades of collected data. Recently, annual maps of soil organic carbon (SOC) in Brazil (beta collection) were released, marking an advancement in understanding its dynamics. However, the quality of this product has not been assessed yet due to limitations in the quantity and distribution of available soil data. This work aims to elucidate how the quality of these maps will be evaluated using quantitative and qualitative methods as strategies to overcome these limitations. Among the quantitative methods, spatial cross-validation and the explicit representation of spatial uncertainty will be performed. Detailed legacy map comparison methods will also be explored. The results will be synthesized in the report accompanying the technical document of MapBiomas Solo Collection 1, enriching our understanding of Brazilian soil dynamics. This report will serve as a solid foundation for analyzing SOC estimates, making the time series more valuable for soil-related decision-making, and positively impacting sustainable resource management and the development of more effective environmental policies.
KEYWORDS: Cross-Validation; Evaluation Metrics; Legacy Data; ; Uncertainty.

INTRODUÇÃO

O solo é um dos maiores reservatórios de carbono do planeta, com importância crítica no contexto do sistema climático global e no bem-estar da humanidade. As

[1] Bolsista Programa Institucional de Iniciação Científica - PIBIC. Universidade Tecnológica Federal do Paraná, Dois Vizinhos, Paraná, Brasil. E-mail: cardoso.mvs@gmail.com. ID Lattes: 8167522617917644.
[2] Pesquisadora de Pós-Doutorado. Universidade Federal de Goiás, Goiânia, Goiás, Brasil. E-mail: huf.aline@gmail.com. ID Lattes: 1570834132484121.
[3] Docente do curso de bacharelado em Agronomia da Universidade Tecnológica Federal do Paraná, Santa Helena, Paraná, Brasil. E-mail: alessandrosamuelrosa@gmail.com. ID Lattes: 1609751519717461.
[4] Docente do curso de bacharelado em Agronomia e Engenharia Florestal da Universidade Tecnológica Federal do Paraná, Dois Vizinhos, Paraná, Brasil. E-mail: tacihorst@gmail.com. ID Lattes: 6763043931071514.

mudanças nos estoques de carbono orgânico do solo (COS) são influenciadas por diversos fatores, incluindo uso da terra e clima (LAL, 2013). Diversas abordagens já foram utilizadas para estimar a distribuição espacial do COS, desde modelos estatísticos lineares simples (MOORE et al., 1993) até modelos geoestatísticos e técnicas avançadas de aprendizado de máquina (LAMICHHANE; KUMAR; WILSON, 2019).

Iniciativas nacionais e internacionais já mapearam os estoques de COS do Brasil (VASQUES, 2017; GOMES et al., 2019; POGGIO, 2021). Essas abordagens visaram identificar tendências em larga escala, priorizando a dimensão espacial e empregando técnicas de mapeamento digital de solos (MDS) (MCBRATNEY; MENDONÇA SANTOS; MINASNY, 2003). No entanto, esses esforços enfrentaram desafios devido à disponibilidade limitada de dados pontuais de solo e covariáveis que representassem os forçantes de mudança do COS, o que resultou na ausência da dimensão temporal. A composição dos dados coletados em diferentes períodos resultou em produtos mapeados que representam um intervalo de tempo, tipicamente abrangendo décadas, para os quais os dados de campo foram coletados. No caso do Brasil, considerando que a maioria dos dados abertos de solo remonta às décadas de 1970 e 1980 (SAMUEL-ROSA et al., 2019), os produtos resultantes tendem a representar com mais precisão esses períodos.

O Brasil, um país altamente dependente do uso da terra, carece de informações detalhadas sobre o solo ao longo do tempo e espaço. A expansão das atividades humanas no território brasileiro afeta diretamente os estoques de COS, mas as estimativas existentes não levam em conta essas mudanças (POGGIO, 2021). De acordo com a coleção 8 de mapas anuais de cobertura e uso da terra produzida pelo MapBiomas, cerca de 33% do território brasileiro está sob uso antrópico (287 Mha), evidenciando uma expansão contínua das atividades humanas (MAPBIOMAS, 2023a). Estes dados confirmam a transformação substancial e rápida que o território do país tem passado, uma evolução que, inevitavelmente, afeta os estoques de COS.

Recentemente, foi lançada uma série temporal de mapas de estoque de COS no Brasil, abrangendo o período de 1985 a 2021, coleção beta, representando um avanço significativo. No entanto, esses mapas têm incertezas inerentes aos dados e modelos utilizados. A série temporal foi desenvolvida utilizando modelos estatísticos que integram informações sobre solo, clima e vegetação, gerando múltiplas possibilidades de uso (produtos). Entretanto, os dados pontuais de COS usados no treinamento dos modelos preditivos apresentam heterogeneidade espacial e temporal d e as restrições do modelo preditivo empregando aprendizado de máquina (MAPBIOMAS, 2023b).

O objetivo deste trabalho foi estabelecer uma estratégia de avaliação da qualidade da série temporal de mapas de estoque de COS fornecida pelo MapBiomas Solo frente a escassez de dados de campo disponíveis para validação direta.

REFERENCIAL TEÓRICO

Apesar da versatilidade da série temporal, os produtos resultantes não estão isentos de limitações. Os modelos estatísticos, embora ferramentas valiosas para compreender um fenômeno, são simplificações da realidade e nunca alcançam uma verdade definitiva, frequentemente apresentando erros e sujeitos a contínuos ajustes (SOUZA, 2021). Os mapas obtidos a partir destes modelos sempre se desviam da realidade, ou do que acreditamos que ela seja. Isso ocorre, porque o MDS em si também contém incertezas (HEUVELINK, 2018).

A escolha da métrica para avaliar o modelo deve ser cuidadosamente considerada dentro do contexto do problema em questão, uma vez que diferentes parâmetros podem levar a interpretações distintas (BRANCO; TORGO; RIBEIRO, 2015). Essas métricas são essenciais para interpretar de maneira apropriada e aplicar eficazmente os resultados obtidos. Para que isso ocorra, é fundamental explicitar a qualidade dos dados, sendo um desafio pois cada metodologia possui sua estruturação própria. Sendo assim, como podemos avaliar a precisão e a incerteza desses mapas de COS em uma série temporal quando não dispomos de dados adicionais para validação e não podemos retroceder no tempo para coletar informações extras?

A avaliação de mapas de solos é desafiadora devido à falta de um design de amostragem específico para mapas temporais, como os baseados em dados legados, especialmente em áreas extensas como o Brasil. Validar mapas de solos por amostragem probabilística, como sugerido por BRUS; KEMPEN; HEUVELINK (2011), é quase impraticável, tornando a validação cruzada uma alternativa viável, especialmente em áreas com amostragem limitada. Além disso, a representação adequada da incerteza, usando estatísticas e distribuições de probabilidade, desempenha um papel essencial em todas as etapas da modelagem de solos, identificando fontes de incerteza e melhorando a qualidade dos mapas de solo, ao fornecer insights sobre como essa incerteza afeta as estimativas finais (HEUVELINK, 2018).

Outra estratégia valiosa é a aplicação de métodos qualitativos de comparação de mapas, que envolvem identificar padrões geográficos conhecidos do solo e avaliar sua representação nos mapas. Isso é alcançado por meio de pontos de referência conhecidos, geralmente obtidos por observações de campo em áreas de teste, que determinam a precisão das previsões sobre o tipo ou propriedade do solo e identificam oportunidades de melhoria nas covariáveis do modelo (ROSSITER et al., 2022). Essas abordagens múltiplas de avaliação são cruciais para aprimorar a qualidade e a utilidade dos mapas de solo em cenários de escassez de dados de campo.

Para alcançar esse objetivo, será utilizado técnicas quantitativas e qualitativas.

MATERIAL E MÉTODOS

CARACTERIZAÇÃO DA SÉRIE TEMPORAL ANALISADA

Os mapas anuais da coleção beta do Mapbiomas Solo foram gerados por meio de modelos de regressão, utilizando algoritmos de aprendizado de máquina. Esses algoritmos estabelecem relações numéricas entre os estoques de COS e variáveis ambientais espaciais, com dados de campo coletados desde a segunda metade do século XX que estão depositados no Repositório Brasileiro de Dados do Solo, conhecido como SoilData (https://soildata.mapbiomas.org). As covariáveis ambientais, que representam fatores de formação do solo, são obtidas de bancos de dados espaciais abertos (MAPBIOMAS, 2023b).

Para esse fim, é utilizado o modelo de Floresta Aleatória, que é um conjunto de modelos de árvore de regressão, onde as previsões de várias árvores de regressão aleatória são combinadas para obter a estimativa final. Os parâmetros desse modelo, como *mtry* e *ntrees*, são ajustados usando validação cruzada *leave-one-out* e o pacote *Caret* em software R. O modelo final de Floresta Aleatória foi desenvolvido com base na

minimização do erro quadrático médio (RMSE) e implementado usando o pacote randomForest em R (LIAW; WIENER, 2002).

VALIDAÇÃO CRUZADA

Será realizada a validação cruzada espacial e a validação cruzada comum para evitar estimativas otimistas de incerteza, especialmente em cenários com amostras de solo agrupadas. A validação será feita usando os mesmos hiperparâmetros dos modelos presentes no Google Earth Engine, utilizando os pacotes *randomForest* e *caret* no R.

Na validação cruzada comum, as amostras serão divididas aleatoriamente em 10 subconjuntos para treinamento e validação do modelo. Na validação cruzada espacial, os dados de treinamento serão agrupados em 30 grupos usando *k-Means* e, em seguida, diferentes grupos de amostras serão selecionados aleatoriamente para treinamento ou validação, também em 10 subconjuntos. Serão explorados diferentes números de clusters (10, 20 e 30) para equilibrar a representação de padrões gerais e locais (BRUS; KEMPEN; HEUVELINK, 2011).

As métricas de ajuste, como Erro Médio (ME) (Eq. 1), Erro Absoluto Médio (MAE) (Eq. 2), Erro Quadrático Médio (MSE) (Eq. 3), Raiz do Erro Quadrático Médio (RSME) (Eq. 4) e Eficiência de Nash-Sutcliffe (NSE) (Eq. 5), serão calculadas em nível nacional e por bioma. Apenas as métricas de validação cruzada comum serão apresentadas em nível de bioma. Isso permitirá avaliar o desempenho do modelo em regiões com um menor número de amostras.

$$= \sum_{i=1}^{n} \frac{yi-xi}{n} \quad (1)$$

$$= \frac{\sum_{i=1}^{n} |yi-xi|}{n} \quad (2)$$

$$= \frac{(yi-xi)^2}{n} \quad (3)$$

$$= \sqrt{\frac{\sum_{i=1}^{n} (yi-xi)^2}{n}} \quad (4)$$

$$= 1 - \frac{MSE}{\frac{(\overline{xi}-xi)^2}{n}} \quad (5)$$

em que: yi é o valor predito; xi é o valor observado; n é o número de observações; $\overline{xi}$ é a média dos valores observados.

INCERTEZA LOCAL E GLOBAL

A incerteza espacial em relação ao estoque de COS é a imprecisão associada às informações geoespaciais. Para estimar a incerteza das previsões em um modelo de Floresta aleatória no contexto dos estoques de COS, existem várias abordagens disponíveis, dependendo das necessidades e recursos disponíveis.

Uma abordagem comum é o método Bootstrap, que envolve reamostrar os dados e treinar várias árvores de decisão em subconjuntos aleatórios do conjunto de treinamento com reposição. Em seguida, você pode calcular a média das previsões dessas árvores, o desvio padrão ou construir um intervalo de confiança com base nessas previsões para avaliar a incerteza local (BREIMAN, 2001). Outra abordagem é usar os dados *out-of-bag* (OOB) gerados pela própria Floresta aleatória. Como cada árvore é treinada em um subconjunto específico dos dados, os pontos de dados que não fazem parte da amostra de treinamento de uma árvore são considerados dados OOB. Esses dados podem ser usados para avaliar a incerteza global das previsões, calculando o erro OOB para cada ponto de dados e examinando a distribuição desses erros (BYLANDER, 2002). O modelo para estimar a incerteza das previsões será desenvolvido na plataforma Google Earth Engine usando linguagem JavaScript.

MAPAS DE REFERÊNCIA

Para levantamento dos mapas de referência, estabeleceremos critérios rigorosos para a seleção de artigos relacionados ao MDS no contexto brasileiro, com foco na comparação desses estudos com os mapas de estoques de COS. Serão incluídos apenas estudos que tenham aplicado o MDS *stricto sensu*, o que implica o uso de covariáveis preditoras, funções matemáticas e um conjunto de dados de treinamento para a classificação numérica de solos. Abordagens *lato sensu* do MDS, que não envolvem essa classificação numérica, serão excluídas desta revisão.

RESULTADOS ESPERADOS

Serão calculadas cinco métricas de validação cruzada e validação cruzada espacial para a série histórica e 38 mapas de incerteza para cada profundidade (0 - 30 e 30 - 100 cm).

Espera-se que a validação cruzada seja uma métrica mais realista para os locais com alta densidade amostral. Nos locais com baixa densidade de amostras, espera-se que a validação cruzada espacial forneça uma métrica mais realista e contextualizada. A ilustração da incerteza espacial poderá complementar essas informações, fornecendo um intervalo de valores a partir do qual se espera que o valor real do estoque de COS esteja, permitindo assim identificar regiões com maior incerteza.

O levantamento irá gerar um banco de dados (mapas) para comparação entre os mapas da série temporal e mapas de referência que permitirá avaliar a concordância e identificar discrepâncias, oferecendo direcionamentos para aprimorar o processo de mapeamento.

Todos esses resultados serão compilados em um relatório de avaliação que será publicado junto ao *Algorithm Theoretical Basis Document* (ATBD) da coleção 1 dos mapas de estoque de COS do MapBiomas Solo, que a partir disso levará em consideração os anos de 1985-2022, contribuindo para uma compreensão mais completa e assertiva do uso dessas informações em diferentes contextos e estruturas.

Agradecimentos

À Universidade Tecnológica Federal do Paraná, MapBiomas, CNPq, Fundação Araucária e Instituto Arapyaú pelo financiamento do projeto. À UTFPR pela concessão da bolsa.

Conflito de interesse

Não há conflito de interesse

REFERÊNCIAS

BRANCO, P.; TORGO, L.; RIBEIRO, R. **A Survey of Predictive Modelling under Imbalanced Distributions**. arXiv, , 13 maio 2015. Disponível em: <http://arxiv.org/abs/1505.01658>. Acesso em: 4 set. 2023

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001.

BRUS, D. J.; KEMPEN, B.; HEUVELINK, G. B. M. Sampling for validation of digital soil maps. **European Journal of Soil Science**, v. 62, n. 3, p. 394–407, 2011.

BYLANDER, T. Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates. p. 287–297, jul. 2002.

GOMES, L. C. et al. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, p. 337–350, 15 abr. 2019.

HEUVELINK, G. B. M. **Soil Organic Carbon Mapping Cookbook**. [s.l: s.n.].

LAL, R. Soil carbon management and climate change. **Carbon Management**, v. 4, n. 4, p. 439–462, 1 ago. 2013.

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

LIAW, A.; WIENER, M. Classification and Regression by randomForest. v. 2, 2002.

MAPBIOMAS. **Annual mapping of soil organic carbon stock in Brazil 1985-2021 (beta collection). Algorithm theoretical basis document and results**. MapBiomas Data, , 2023. Disponível em: <https://data.mapbiomas.org/citation?persistentId=doi:10.58053/MapBiomas/3KXXVV>. Acesso em: 4 set. 2023

MAPBIOMAS. “Projeto MapBiomas – Coleção 8 da Série Anual de Mapas de Cobertura e Uso de Solo do Brasil. MapBiomas Data, 2023a. Disponível em: http://mapbiomas.org.

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

MOORE, I. D. et al. Soil Attribute Prediction Using Terrain Analysis. **Soil Science Society of America Journal - SSSAJ**, v. 57, 1 mar. 1993.

POGGIO, L. **SOIL - SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty**. Disponível em: <https://soil.copernicus.org/articles/7/217/2021/>.

ROSSITER, D. G. et al. How well does digital soil mapping represent soil geography? An investigation from the USA. **SOIL**, v. 8, n. 2, p. 559–586, 5 set. 2022.

SAMUEL-ROSA, A. et al. Open legacy soil survey data in Brazil: geospatial data quality and how to improve it. **Scientia Agricola**, v. 77, p. e20170430, 1 jul. 2019.

SOUZA, J. S. S. DE. Modelagem Matemática e Aprendizagem Significativa: uma Relação Subjacente. **Jornal Internacional de Estudos em Educação Matemática**, v. 14, n. 2, p. 241–247, 27 set. 2021.

VASQUES, G. **Mapa de estoque de carbono orgânico do solo (COS) a 0-30 cm do Brasil. - Portal Embrapa**. Disponível em: <https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1085197/mapa-de-estoque-de-carbono-organico-do-solo-cos-a-0-30-cm-do-brasil>. Acesso em: 1 set. 2023.