Valores de Shapley como ferramenta explicativa para modelos complexos de mapeamento digital de solos

Shapley values as an explanatory tool for complex digital soil mapping models

MOQUEDACE, Cássio Marques[1]; BALDI, Clara Glória Oliveira[1]; SIQUEIRA, Rafael Gomes[1], PEREIRA, Luís Flávio[1]; GOMES, Lucas Carvalho[2]; FERNANDES-FILHO, Elpídio Inácio[1]

[1]Laboratório de Geoprocessamento e Pedometria (LabGeo) da Universidade Federal de Viçosa - UFV, cassiomoquedace@gmail.com, clara.gloria.oliveira@gmail.com, luis.flavio@ufv.br, elpidio@ufv.br; [2]Departamento de Agroecologia da Universidade de Aarhus - Dinamarca, lucas.gomes@agro.au.dk

RESUMO EXPANDIDO TÉCNICO CIENTÍFICO

Eixo Temático: Crise ecológica e mudança climática: resistências e impactos na agricultura, nas águas e nos bens

Resumo: O solo é um bem capaz de garantir os serviços ecossistêmicos necessários para alcançar os objetivos de desenvolvimento sustentável. Objetivou-se avaliar de forma espacial a importância de proxies ambientais em modelo de aprendizado de máquina como preditores do teor de argila do solo. A área de estudo foi Minas Gerais. Utilizou-se quatro diferentes grupos de informações ambientais como controladores dos teores de argila do solo. Todo o processamento foi realizado em ambiente R. Calculou-se os valores de Shapley espacialmente, distribuindo o peso para as covariáveis de acordo com sua participação relativa no resultado. Os valores de Shapley evidenciaram a ampla variabilidade espacial da importância de diferentes variáveis ambientais que controlam a distribuição de argila no estado de Minas Gerais. Nas regiões onde os solos são mais arenosos o clima foi a variável que impulsionou as estimativas. Os valores de Shapley proporcionaram maior explicabilidade do modelo.
Palavras-chave: MDS; aprendizado de máquina; linguagem R; quantile random forest.

Introdução

O solo tem papel central para garantir os serviços ecossistêmicos necessários para alcançar os objetivos de desenvolvimento sustentável estabelecidos pela ONU em 2015 (BOUMA et al., 2019; MCBRATNEY et al., 2014). Produzir informações espaciais em alta resolução de atributos do solo é necessário para apoiar pesquisas científicas e subsidiar tomadas de decisão. O avanço computacional tem proporcionado a construção de modelos com boa acurácia no mapeamento digital de solos (MDS), superando métodos convencionais e produzindo mapas mais plausíveis com a realidade de campo (GOMES et al., 2019).

No entanto, diante da elevada complexibilidade destes modelos, se tornam em sua maioria, modelos preditivos, carecendo de explicabilidade ou interpretabilidade, sendo até classificados como “modelos caixa-preta” (DIKSHIT et al., 2021). De forma geral os autores utilizam métricas estatísticas como a importância relativa de uma variável ambiental utilizada como preditor do modelo para avaliar como o atributo do solo modelado se relaciona com os proxies utilizados (SIQUEIRA et al., 2023).

Embora esta seja uma forma válida de avaliar a contribuição das variáveis ambientais, este método implica em avaliar a redução na média da precisão do modelo por permutação. Isto relata somente a importância isolada da variável para o modelo ajustado. Além disso, a importância destes fatores ambientais covariam no espaço.

Alguns métodos para melhorar a compreensão dos fatores que impulsionam o atributo do solo modelado vem sendo explorado, sendo o valor Shapley um deles. É um método baseado na teoria dos jogos de coalizão que foi desenvolvido como um meio de distribuir o ganho entre diferentes jogadores de acordo com sua participação relativa em um jogo. Este método é promissor para descrever a relação funcional entre um atributo do solo e as covariáveis ambientais, inclusive no espaço (WADOUX; MOLNAR, 2022).

Sob a perspectiva da agroecologia, compreender os fatores ambientais que determinam os teores de argila do solo e como eles se relacionam com o ambiente é potencialmente capaz de construir e/ou remodelar agroecossistemas mais adaptados localmente a crise ecológica e as mudanças climáticas. Objetivou-se avaliar de forma espacial a importância de proxies ambientais em modelo de aprendizado de máquina como preditores do teor de argila do solo.

Metodologia

Como área de estudo utilizou-se o estado brasileiro de Minas Gerais, localizado na região Sudeste do Brasil. O clima varia entre As, Aw, Cfa, Cfb, Cwa e Cwb (ALVARES et al., 2013). A região detém grande diversidade ambiental, inclusive geomorfológica. A paisagem é dominada por Latossolos e Argissolos profundos. Para o ajuste do modelo utilizou-se o banco de dados de Guevara et al. (2018) e Souza et al. (2015) com 685 amostras de solo na profundidade de 0-20 cm (Figura 1).

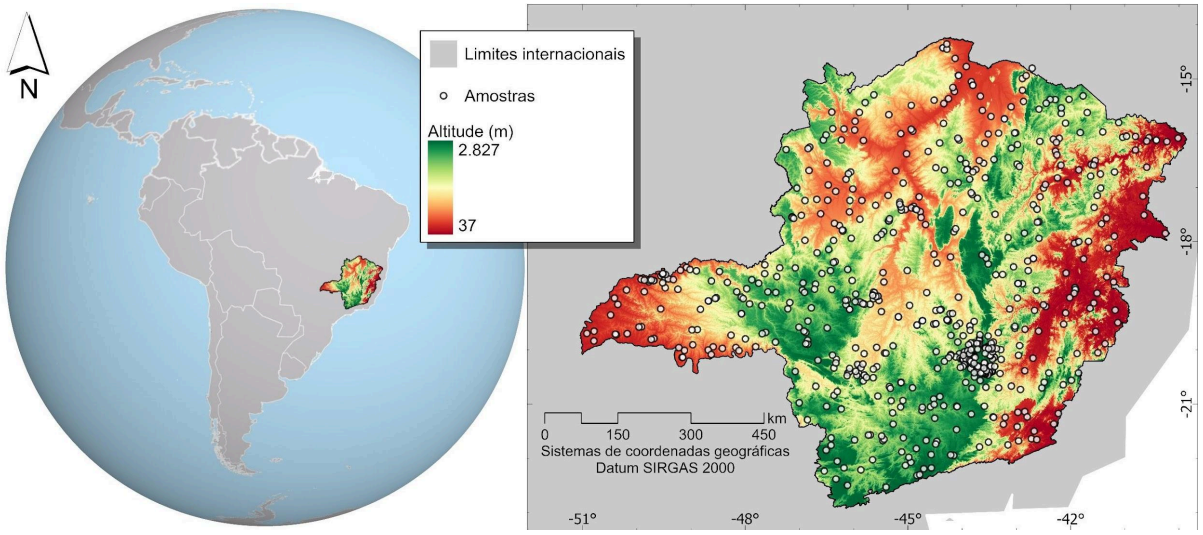

Figura 1. Mapa de localização das amostras distribuídas no relevo do estado de Minas Gerais, Brasil.

Utilizou-se quatro diferentes grupos de informações ambientais como controladores dos teores de argila do solo (Tabela 1). Os preditores foram harmonizados na resolução de 30m. O modelo de aprendizado de máquina utilizado foi o quantile random forest (QRF) e o processo de ajuste se deu sob a seguinte estrutura repetida 100 vezes: i. separação em treino com validação cruzada 10 folds (80% dos dados de argila) e teste (20%); ii. ajuste do modelo e hiperparâmetros e; iii. avaliação de performance. As métricas de performance utilizadas foram o erro médio absoluto (MAE), a raiz quadrada do erro médio (RMSE) e o coeficiente de correlação de concordância (ρ_c).

Tabela 1. Preditores ambientais utilizados no ajuste do quantile random forest para os teores de argila do solo no estado de Minas Gerais.

Clima	NDVI	**Solo**	K40
TMA	NPP	CT	DZ1
PMA	**Relevo**	eTh	IMT
Organismos	MDT	eU	US

Em que: TMA = temperatura média anual; PMA = precipitação média anual; NDVI = índice de vegetação por diferença normalizada; NPP = produtividade primária líquida; MDT = modelo digital de terreno; CT = contagem total da radiação gama; eTh = equivalente de tório; eU = equivalente de urânio; K40 = equivalente de potássio; DZ1 = 1ª derivada vertical da anomalia magnetométrica; IMT = magnometria – intensidade magnética total e; US = Umidade do solo.

Todo o processamento foi realizado em ambiente R (R CORE TEAM, 2023) e para o cálculo dos valores de Shapley foi utilizado o pacote fastshap (GREENWELL, 2023). Os valores de Shapley distribuem o peso para as covariáveis de acordo com sua participação relativa no resultado. O peso é calculado pela previsão para uma determinada unidade menos a previsão média.

Para a construção dos mapas a espacialização dos valores de Shapley calculou-se o valor médio por pixel dos grupos de covariáveis, resultando em quatro mapas (Clima, Organismos, Relevo e Solo). Além disso, foi gerado um mapa único de

contribuição que consistiu em atribuir a classe dominante com maior valor Shapley por pixel.

Resultados e Discussão

O modelo ajustado alcançou ρ_c = 0,58 ± 0,05, MAE = 15,2 ± 2,53 % e RMSE = 18,5 ± 2,28 %. Estes valores denotaram um ajuste satisfatório do QRF para os teores de argila do solo comparado a outros autores que ao modelar atributos do solo com modelos de árvore de decisão alcançaram métricas mais modestas (GOMES et al., 2019; SIQUEIRA et al., 2023).

Os valores de Shapley evidenciaram a ampla variabilidade espacial da importância de diferentes variáveis ambientais que controlam a distribuição de argila no estado de Minas Gerais (Figura 2). O grupo de covariáveis com as maiores importâncias globais foram o Solo e o Clima, respectivamente.

O solo aqui representado pelos dados de gamaespectrometria e pela umidade do solo tem ampla fundamentação na literatura tangente a textura do solo. Observa-se que a importância das covariáveis de solo aumentou nas regiões onde há maior concentração de argila. Isso pode estar associado tanto a granulometria fina do solo que geralmente retém maiores quantidades de água (no caso do preditor de umidade do solo), quanto nas relações mais fortes das assinaturas espectrais dos raios gama em solos com textura mais fina (TAYLOR et al., 2002).

Espacialmente, observa-se que esta importância foi mais acentuada na porção central do estado (Central mineira e parte do Triângulo Mineiro) e uma faixa que margeia da região leste até a região sul (Vale do Rio Doce, Zona da Mata e parte do Sul de Minas).

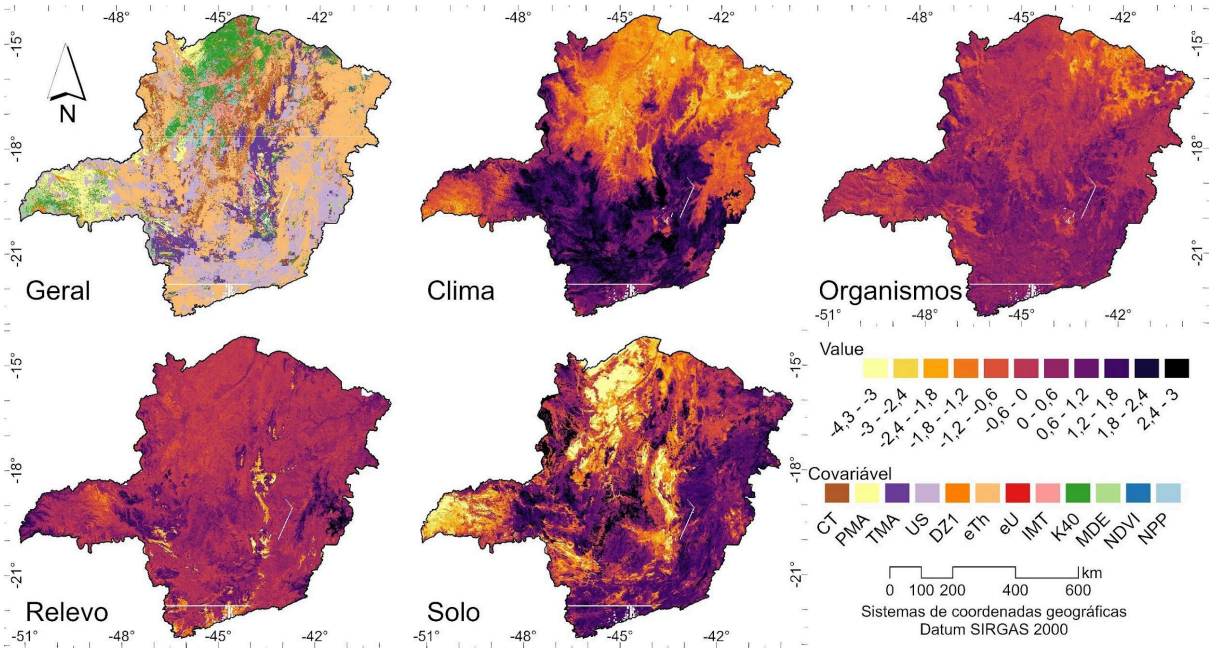

Figura 2. Mapa da covariável de localização mais importante que contribuiu para a predição de argila e mapas do padrão espacial dos valores de Shapley médios para quatro grupos de covariáveis utilizadas na predição de argila com o quantile random forest para o estado de Minas Gerais, Brasil. Em que: As abreviações utilizadas seguem descritas na Tabela 1.

O clima destaca-se entre a parte central (Mesorregião Metropolitana de Belo Horizonte e Serra do Espinhaço) e leste do estado, e na parte noroeste mineira. Verifica-se que nas regiões central e leste o clima úmido favorece o desenvolvimento de solos argilosos, enquanto na região noroeste, a baixa precipitação se destaca como o principal fator controlador dos valores reduzidos de argila naquela parte do Estado.

Na Serra do Espinhaço, as menores temperaturas associadas ao clima serrano foram o fator mais importante no modelo, não havendo necessariamente uma relação direta de causa-efeito com os baixos valores de argila desta região, mas sim com a geologia quartzítica que sustenta as grandes altitudes da Serra do Espinhaço. Na região noroeste, o efeito do clima captado pelo modelo também é integrado com o efeito da geologia, em particular das rochas areníticas que predominam nesta parte.

Na importância geral, observa-se dominância do radionuclídeo eTh nas mais variadas regiões. O que explica a elevada importância do grupo de covariáveis solos na predição de argila. Nas regiões onde os solos são mais arenosos (principalmente no norte do estado) o clima, principalmente a PMA foi a variável que impulsionou as estimativas, sugerindo que os raios gama se correlacionam de forma mais fraca em solos arenosos.

A dependência parcial eTh reafirma tal resultado (Figura 3). Os valores de eTh aumentam de forma expressiva ao passo que os teores de argila são aumentados, até alcançar uma assíntota (em torno de 50 ppm de eTh).

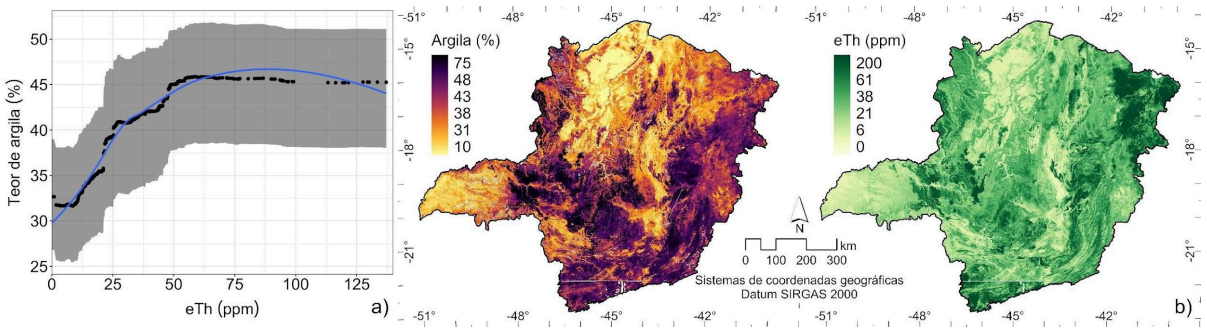

Figura 3. Dependência parcial (a), distribuição dos teores de argila do solo e distribuição do equivalente de tório – eTh (b) para o estado de Minas Gerais, Brasil.

Ao observar a distribuição espacial da argila e do eTh nota-se uma equivalência de maiores teores de argila e maior intensidade do radionuclídeo. Isto reafirma a importância da gamaespectrometria para o mapeamento digital de solos, sobretudo para a textura do solo.

Conclusões

Os valores de Shapley foram eficientes na avaliação dos impulsionadores da predição espacial da argila no estado de Minas Gerais. Os valores de Shapleys proporcionaram maior explicabilidade do modelo nos principais proxies que controlam os teores de argila. Os produtos gerados neste trabalho são importantes ferramentas para subsidiar as tomadas de decisão tangentes à resiliência dos ecossistemas mineiros.

Referências bibliográficas

BOUMA, Johan et al. The challenge for the soil science community to contribute to the implementation of the UN Sustainable Development Goals. **Soil Use and Management**, v. 35, n. 4, p. 538–546, 2019. DOI: 10.1111/sum.12518. Disponível em: https://onlinelibrary.wiley.com/doi/10.1111/sum.12518.

DIKSHIT, Abhirup et al. Pathways and challenges of the application of artificial intelligence to geohazards modelling. **Gondwana Research**, v. 100, p. 290–301, 2021. DOI: 10.1016/J.GR.2020.08.007. Acesso em: 14 jul. 2023.

GOMES, Lucas C. et al. Modelling and mapping soil organic carbon stocks in Brazil. **Geoderma**, v. 340, n. January, p. 337–350, 2019. DOI: 10.1016/j.geoderma.2019.01.007. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0016706117320669.

GREENWELL, Brandon. **fastshap: Fast Approximate Shapley Values**. 2023. Disponível em: https://cran.r-project.org/package=fastshap.

GUEVARA, Yang. Z. C. et al. Reference values of soil quality for the Rio Doce Basin. **Revista Brasileira de Ciência do Solo**, v. 42, 2018. DOI:

10.1590/18069657rbcs20170231. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-06832018000100515&lng=en&tlng=en.

MCBRATNEY, Alex et al. The dimensions of soil security. **Geoderma**, v. 213, p. 203–213, 2014. DOI: 10.1016/j.geoderma.2013.08.013. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0016706113002954.

R CORE TEAM. R: **A Language and Environment for Statistical Computing**. R Foundation for Statistical ComputingVienna, Austria, 2023. Disponível em: http://www.r-project.org.

SIQUEIRA, Rafael G. et al. Machine learning applied for Antarctic soil mapping: Spatial prediction of soil texture for Maritime Antarctica and Northern Antarctic Peninsula. **Geoderma**, v. 432, p. 116405, 2023. DOI: 10.1016/j.geoderma.2023.116405. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S0016706123000824. Acesso em: 21 jun. 2023.

SOUZA, José. J. L. L. et al. Geochemistry and spatial variability of metal(loid) concentrations in soils of the state of Minas Gerais, Brazil. **Science of the Total Environment**, v. 505, p. 338–349, 2015. DOI: 10.1016/j.scitotenv.2014.09.098. Disponível em: https://linkinghub.elsevier.com/retrieve/pii/S004896971401420X.

TAYLOR, Miranda. J. et al. Relationships between soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia. **Exploration Geophysics**, v. 33, n. 2, p. 95–102, 2002. DOI: 10.1071/EG02095. Disponível em: https://www.tandfonline.com/doi/full/10.1071/EG02095.

WADOUX, Alexandre M. J. C.; MOLNAR, C. Beyond prediction: methods for interpreting complex models of soil variation. **Geoderma**, v. 422, p. 115953, 2022. DOI: 10.1016/J.GEODERMA.2022.115953. Acesso em: 14 jul. 2023.