VARIÁVEIS GEOMORFOMÉTRICAS E MACHINE LEARNING NA PREDIÇÃO DE SÍTIOS ARQUEOLÓGICOS EM SERRANÓPOLIS - GOIÁS - BRASIL

Alessandra Cristina Pereira[1], Édipo Henrique Cremon[2], Rosiclér Theodoro da Silva[3], Júlio Cezar Rubin de Rubin[4] e Nilton Ricetti Xavier de Nazareno[5]

[1]Instituto Federal de Goiás (IFG) – Campus Goiânia, R. 75, 46 - Centro, Goiânia - GO, 74055-110, alepereira.amb@gmail.com, [2]Instituto Federal de Goiás (IFG) – Campus Goiânia, R. 75, 46 - Centro, Goiânia - GO, 74055-110, edipo.cremon@ifg.edu.br, [3]Pontifícia Universidade Católica de Goiás, Av, Universitária, 1.440, Setor Universitário 74605-010 - Goiânia, GO, silva.rosicler@gmail.com, [4]Pontifícia Universidade Católica de Goiás, Av, Universitária, 1.440, Setor Universitário 74605-010 - Goiânia, GO, rubin@pucgoias.edu.br e [5]Instituto Federal de Goiás (IFG) – Campus Goiânia, R. 75, 46 - Centro, Goiânia - GO, 74055-110, nilton.nazareno@ifg.edu.br

RESUMO

O objetivo deste trabalho foi predizer locais de sítios arqueológicos utilizando variáveis geomorfométricas e algoritmos de *machine learning* (aprendizado de máquina) na região de Serranópolis, Goiás, por possuir um dos mais importantes complexos arqueológicos do Brasil. Foram utilizados pontos amostrais de presença de sítios arqueológicos e amostras aleatórias de pseudoausências de sítios, além de 18 covariáveis geomorfométricas derivadas diretamente de modelo digital de elevação. A predição foi realizada por dois algoritmos de classificação supervisionada baseados em árvore de decisão: Random Forest (RF) e Extreme Gradient Boosting (XGBoost). O RF apresentou melhor habilidade preditiva para esse estudo. As covariáveis geomorfométricas mais importantes para a predição de sítios arqueológicos foram associadas a rugosidade, fator de vista para o céu, depressão topográfica e amplitude do relevo.

Palavras-chave – Inteligência Artificial, Random Forest, Geoarqueologia, Dado geográfico, MDE.

ABSTRACT

The aim of this work was to predict archaeological site locations based on geomorphometric covariates and machine learning algorithms in the region of Serranópolis, Goiás, as it has one of the most important archaeological complexes in Brazil. Sample points of presence of archaeological sites and random samples of pseudo-absence of sites were used, in addition to 18 geomorphometric covariates derived directly from digital elevation model. Prediction was performed by two supervised classification algorithms based on decision tree: Random Forest (RF) and Extreme Gradient Boosting (XGBoost). RF showed better predictive ability for this study. The most important topographic covariates for archaeological site prediction were associated with Vector Ruggedness Measure, Sky View Factor, Topographic Depression, and Local Relief.

Key words – *Artificial intelligence, Random Forest, Archaeology, Geographic data, DEM.*

1. INTRODUÇÃO

A associação entre os processos de ocupação humana do passado e o ambiente físico tem permitido a modelagem de sítios arqueológicos, no sentido da predição de potenciais locais de sítios em áreas que nunca foram estudadas com base nos padrões de localização de sítios conhecidos e relação com variáveis do meio ambiente [1]. Os modelos preditivos arqueológicos produzidos até o momento, no geral, utilizam-se de uma abordagem metodológica baseada em técnicas estatísticas multivariadas, considerando dados arqueológicos e variáveis ambientais.

Diante disso, com o avanço da inteligência artificial, em especial de algoritmos de *machine learning* (aprendizado de máquina), modelos preditivos a partir de dados geográficos para fins arqueológicos têm sido beneficiados [2–4]. Não obstante, ainda há poucos trabalhos que exploram tais ferramentas para a predição de sítios arqueológicos no Brasil. Neste sentido, o município de Serranópolis, no estado de Goiás, consiste em um local ideal para a aplicação de técnicas de *machine learning* utilizando dados geográficos para a predição de sítios arqueológicos. Pesquisas arqueológicas nesta região tem apresentado dados importantes, com significância histórico-cultural para a compreensão dos processos ocupacionais do Planalto Central Brasileiro [5].

Este trabalho parte do pressuposto que variáveis geomorfométricas associadas ao meio físico possam estabelecer um modelo preditivo que possibilite inferir áreas potenciais de sítios arqueológicos na região de Serranópolis, empregando inteligência artificial, em especial o *machine learning*, de forma que possa contribuir para as pesquisas geoarqueológicas.

O objetivo deste trabalho foi predizer locais de sítios arqueológicos utilizando *machine learning* e variáveis geomorfométricas na região de Serranópolis, Goiás, analisando a importância das covariáveis geomorfométricas no contexto geoarqueológico.

2. MATERIAL E MÉTODOS

A área de estudo corresponde ao município de Serranópolis, situada na região sudoeste do estado de Goiás, com 5.526,4 km² e distante cerca de 450 km da capital estadual, Goiânia.

Os procedimentos metodológicos para o trabalho, iniciaram com a aquisição e processamento das variáveis

geomorfométricas, usadas como covariáveis independentes. A aquisição dos pontos dos sítios arqueológicos e criação de pontos de pseudoausência de sítios arqueológicos, usados como covariável dependente. A predição foi realizada pelos algoritmos de classificação supervisionada baseados em árvore de decisão Random Forest (RF) e Extreme Gradient Boosting (XGBoost). Estes algoritmos foram aplicados seguindo a validação cruzada do tipo *k-fold* para o treinamento e calibração do modelo (70% das amostras). Para aferir a confiabilidade dos resultados da classificação, 30% das amostras foram utilizadas para teste (validação indepentente) e as métricas de performance AUC (*area under the ROC curve*), sensibilidade, especificidade e F-score foram analisados.

O estudo utilizou um conjunto de 18 variáveis geomorfométricas (Tabela 1) para inferências de sítios arqueológicos, extraídos diretamente do MDE COP-30 (*Copernicus DEM*). Esse MDE está disponível em coordenadas geográficas com datum horizontal WGS-84 e tamanho de pixel de 1 segundo de arco e consiste no melhor dado altimétrico gratuito disponível para o contexto do estado de Goiás [6]. Este dado foi reprojetada para coordenadas métricas no sistema de coordenadas Universal Transversa de Mercator (UTM) fuso 22 Sul, datum horizontal SIRGAS2000, tamanho de pixel de 30 m com o método de convolução cúbica pelo *software* QGIS [7]. Em seguida, foi realizado o processo de correção do MDE para eliminação dos vazios (fossos), tornando o MDE hidrologicamente consistente, utilizando o *software* TerraViewHidro. Por fim, a extração das variáveis geomorfométricas foi realizada por meio de algoritmos disponíveis nos *softwares* SAGA-GIS, QGIS e Whitebox. Neste trabalho, optou-se por adotar o nome das variáveis geomorfométricas em inglês, assim como está disponível na maioria dos *softwares* e literatura (Tabela 1).

Variáveis geomorfométricas		
Slope	*Dissection*	*Multi-Resolution Index of Valley Bottom Flatness* (MRVBF)
Profile curvature	*Local relief*	*Topographic height*
Plan curvature	*Relative Topographic Position* (RTP)	*Vector Ruggedness Measure* (VRM)
Aspect	*Vertical Distance to Channel Network* (VDCM)	*Positive openness*
Circular variance of aspect	*SAGA Wetness Index* (TWI)	*Negative openness*
Depression	*Sky View Factor* (SVF)	*Multi-Resolution Index of the Ridge Top Flatness* (MRRTF)

Tabela 1: Variáveis geomorfométricas utilizadas neste estudo para predição de sítios arqueológicos.

Os registros de sítios arqueológicos usados neste estudo foram levantados da base de dados do Cadastro Nacional de Sítios Arqueológicos (CNSA), além de outros pontos de localização e informações cedidos pelos pesquisadores do IGPA/PUC Goiás. Estas fontes permitiram reunir 54 registros de sítios arqueológicos usados para amostras de treinamento e teste do modelo preditivo (Figura 1).

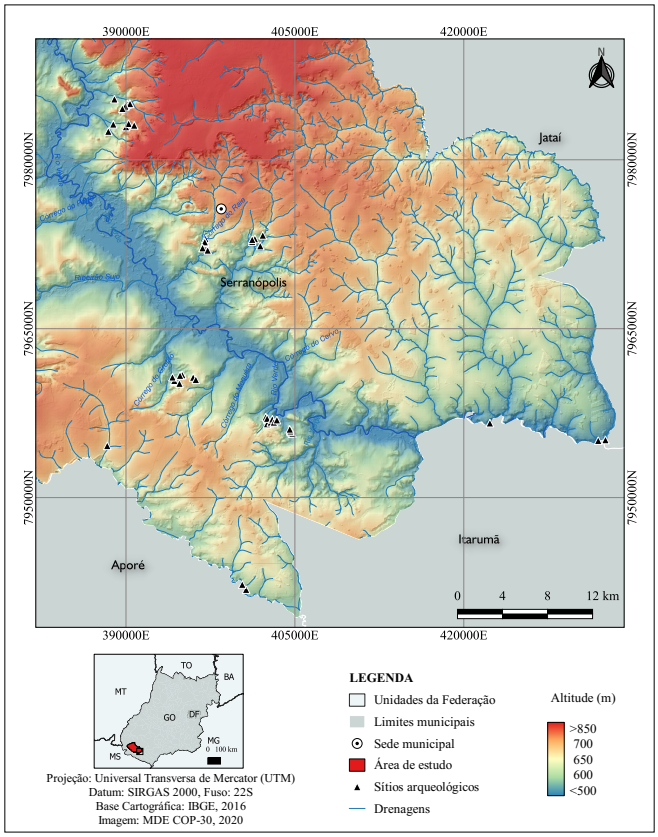


Figura 1: Localização geográfica dos sítios arqueológicos..

A respeito das principais características dos sítios selecionados, dos 54 sítios, 37 são classificados como abrigo sob rocha e 17 a céu aberto; 50 registros foram classificados como sítios pré-colonial e quatro como sítios históricos; e quanto a relevância científica dos sítios, 30 são classificados com alta relevância científica, 3 como média e 21 foram classificados como baixa relevância.

Em função dos pontos de presença de sítios arqueológicos, foram definidos os pontos de pseudoausência de sítio e assumidos como não sítio, ou seja, correspondem às alternativas para os reais pontos de ausência. A definição de pontos de pseudoausência é comum em modelagens onde se conhece a presença espacial de um fenômeno ou alvo a ser predito, mas a ausência é desconhecida, sendo bastante utilizada em modelagens de distribuição de espécies na ecologia [8]. A partir de um raio de 500 metros dos pontos de ocorrência de sítios, foram gerados, aleatoriamente, um conjunto de 60 pontos de pseudoausência com distância mínima de cinco quilômetro entre eles.

Para a análse de quais variáveis geomorfométricas mais importantes na predição de locais de ocorrência de sítios

arqueológicos, foi usado o pacote *caret* do *software* R [9] no qual permite rankear a importância dos algoritmos em valores normalizados entre 0 e 100.

3. RESULTADOS

Na comparação entre os algoritmos utilizados neste trabalho (Tabela 2), o algoritmo RF apresentou uma AUC de 0,932 e o XGBoost de 0,879. Com a métrica de sensibilidade, observa-se que o algoritmo RF apresentou o maior valor (0,850), enquanto o algoritmo XGBoost apresentou um desempenho pouco inferior (0,810). Esses valores representam a proporção de verdadeiros positivos, ou seja, quanto que o modelo acertou a presença de sítios arqueológicos. Para a especificidade, associada a quanto que o modelo foi capaz de predizer corretamente a ausência de uma determinada condição, observa-se que o algoritmo RF e XGBoost tiveram valores muito próximos. Por fim, em relação ao F-score observa-se que o algoritmo RF apresentou o melhor desempenho (Tabela 2). A predição do algoritmo RF pode ser visualizada por um webmapa no seguinte endereço: https://geoifg.github.io/serranopolis_sbsr/.

Algoritmo	RF	XGBoost
AUC	0,932	0,879
Sensibilidade	0,850	0,810
Especificidade	0,929	0,923
F-score	0,866	0,831

Tabela 2: Métricas de performance dos algoritmos RF e XGBoost.

No ranqueamento de importância das covariáveis geomorfométricas (Figura 2), as covariáveis mais importantes para o classificador RF destacam-se as covariávies VRM, SVF, *Depression*, MRVBR e *Local relief*, como as cinco mais preditoras. Já com o classificador XGBoost, as cinco principais covariáveis mais preditoras foram VRM, *Depression*, SVF, RTP e *Dissection*.

4. DISCUSSÃO

De modo geral, ambos classificadores avaliados tiveram bons desempenhos, mostrando que a estratégia de *ensemble* dos algoritmos baseados em árvore de decisão representa uma importante ferramenta para predição [10–12]. O melhor desempenho do classificador RF, baseado em *bagging*, em relação ao classificador XGBoost, baseados em *boosting*, não é uma novidade na literatura. O mesmo ocorreu em estudo efetuado por 12, entretanto isso não é uma regra. Há estudos no qual o desempenho de algoritmos baseados em *boosting* possuem desempenho superior do que os de bagging para classificação binária [13]. Logo, sempre é importante avaliar mais casos, mas no momento o RF tem se mostrado uma poderosa ferramenta preditiva para sítios arqueológicos como demonstrado neste trabalho e em outros [2, 3].

No que se refere as covariáveis geomorfométricas mais importantes para predizer locais de sítios arqueológicos, neste estudo, destaca-se a VRM, apontadas pelos modelos de *machine learning*, como a mais preditiva. Outras também se destacaram: SVF, *Depression*, *Local relief* e MRVBF. O

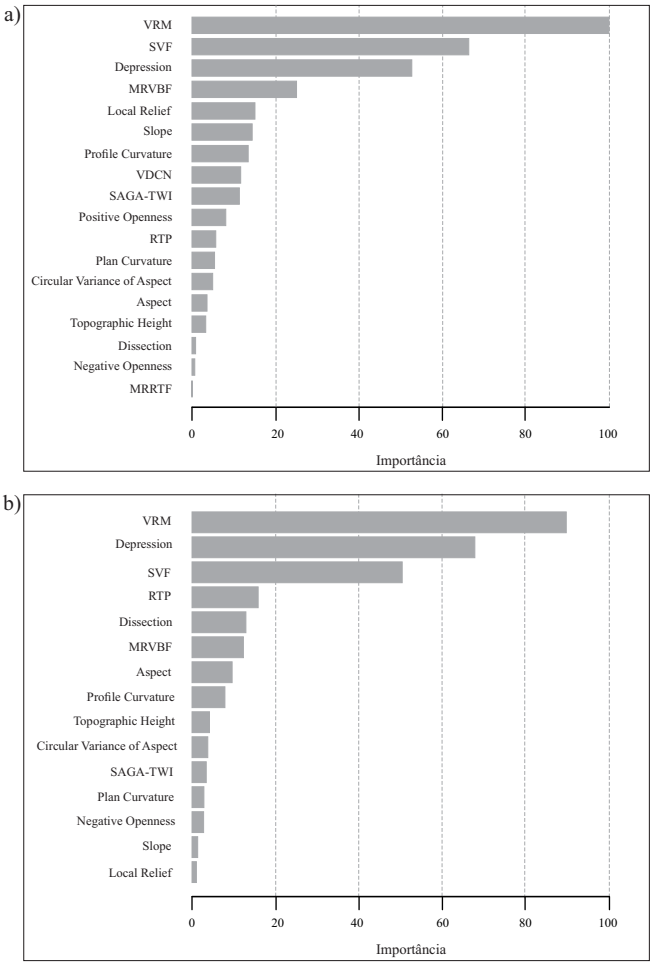


Figura 2: Importâncias das covariáveis geomorfométricas para a predição de sítios arqueológicas usando os classificadores a) RF; e b) XGBoost.

VRM quantifica a rugosidade do terreno, onde uma menor rugosidade representa nenhuma variação do terreno, ao passo que rugosidade máxima representa uma maior variação do terreno.

O SVF calcula o fator de visibilidade do céu, baseada na iluminação difusa influenciada pelo relevo. Esta variável consiste em uma ferramenta muito útil em arqueologia, pois melhora o reconhecimento das características do relevo, bem como é um fator importante para estimativa da radiação solar [14]. Terrenos mais abertos, como em feições expostas (planos e picos) são mais visíveis e, portanto, recebem maior quantidade de luminosidade provenientes do céu claro, enquanto terrenos com relevos mais profundos e partes mais baixas de vales, por exemplo, quase nenhum céu é visível, logo, recebem pouca radiação solar. Uma variável que pode ser relacionada ao SVF é o MRRTF, pois esse índice identifica área planas em grandes altitudes.

As covariáveis geomorfométricas *Depression* e *Dissection* destacam a unidade de relevo de baixa altitudes. *Local relief* gera um modelo do relevo local, acentuando depressões, independentemente do ângulo de iluminação. O uso da variável *Local relief* vem sendo empregada para prospecção arqueológica onde há variações de relevo positivo e/ou negativo [15], a exemplo também das covariável *Positive openness* e *Negative openness* utilizadas nesse trabalho.

Slope (declividade) representa a inclinação da superfície, enquanto *Aspect* (orientação de vertente) mostra sua direção de inclinação. Variáveis como *Slope* e *Aspect* são utilizadas em muitos modelos preditivos arqueológicos, uma vez que estudos comprovam que estas feições topográficas foram usadas como fatores determinantes para escolha de local para assentamento [16], mas neste estudo, para Serranólis, tais variáveis não se mostraram as mais preditivas.

É importante salientar que a região de Serranópolis se encontra inserida em chapadões, resultantes dos processos erosivos de arenitos da Formação Botucatu e áreas de aplainamento caracterizado por morros tabulares de bordas escarpadas, vales encaixados e drenagem de média densidade. No contexto arqueológico, sobretudo os sítios do tipo abrigo sob rocha, se beneficiam do recuo das escarpas pela erosão diferencial. Logo podemos aferir que os sítios arqueológicos estão localizados em regiões onde o relevo é bem delimitado, o que justifica a indicação destas covariáveis como as mais importantes.

5. CONCLUSÕES

O algoritmo RF apresentou o melhor resultado de predição de sítios arqueológicos com base em variáveis geomorfimétricas neste estudo. Contudo, a abordagem de modelagem com o maior poder preditivo (RF) não é uma resposta definitiva, mas pode indicar um roteiro para análise exploratória e preditiva. É uma ferramenta que pode ser utilizada para auxiliar os pesquisadores nas prospecção de estudos arqueológicos combinado com as variáveis geomorfométricas permitindo obter uma quantidade melhor de informações quanto a potenciais regiões com sítios arqueológicos.

Importante frisar quanto as informações dos sítios, que o total utilizado para este estudo (54) não esgota a possibilidade da existência de mais ocorrência na região, visto que o a extensa área territorial do município de Serranópolis ainda é objeto de estudos, o que faz que este sistema possa ser retroalimentado a cada campanha de pesquisa, em virtude do seu potencial geoarqueológico.

Tendo em vista a significância cultural para a compreensão dos processos ocupacionais do planalto Central Brasileiro, almeja-se que este modelo preditivo (RF), aqui apresentado, possa contribuir para as pesquisas nos âmbitos acadêmicos, de arqueologia preventiva, na gestão ambiental e patrimonial de áreas arqueológicas, reforçando assim a importância de pesquisas interdisciplinares.

6. REFERÊNCIAS

[1] Philip Verhagen and Thomas G Whitley. Integrating archaeological theory and predictive modeling: a live report from the scene. *Journal of Archaeological Method and Theory*, 19(1):49–100, 2012.

[2] Maria Elena Castiello and Marj Tonini. An explorative application of random forest algorithm for archaeological predictive modeling. a swiss case study. *Journal of Computer Applications in Archaeology*, 4(1), 2021.

[3] Peter M Yaworsky, Kenneth B Vernon, Jerry D Spangler, Simon C Brewer, and Brian F Codding. Advancing predictive modeling in archaeology: An evaluation of regression and machine learning methods on the Grand Staircase-Escalante National Monument. *PloS one*, 15(10):e0239424, 2020.

[4] Lisa Stubert, Antoni Martín i Oliveras, Michael Märker, Harald Schernthanner, and Sebastian Vogel. Viticulture in the Laetanian region (Spain) during the Roman period: predictive modelling and geomatic analysis. *Geosciences*, 10(6):206, 2020.

[5] Julio Cezar Rubin de Rubin, Sibeli Aparecida Viana, Rosiclér Theodoro da Silva, Maira Barberi, Fernanda Elisa Costa Paulino, Joanne Ester Ribeiro-Freitas, Mariana Garcia de Souza, Eloah Vargas Ribeiro, et al. Cazadores-recolectores y el paisaje en Serranópolis, goiás, brasil. *Boletín de Arqueología PUCP*, (29):129–158, 2020.

[6] Édipo Henrique Cremon, Giovana Maranhão Bettiol, João Paulo Magna Junior, Fabio Campos Macedo, and Max Well de Oliveira Rabelo. Avaliação da altimetria do MDE COP-30 no Centro-Oeste do Brasil. *Revista Brasileira de Cartografia*, 74(3):536–546, set. 2022.

[7] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2022.

[8] Damaris Zurell, Janet Franklin, Christian König, Phil J Bouchet, Carsten F Dormann, Jane Elith, Guillermo Fandos, Xiao Feng, Gurutzeta Guillera-Arroita, Antoine Guisan, et al. A standard protocol for reporting species distribution models. *Ecography*, 43(9):1261–1277, 2020.

[9] Max Kuhn. *The caret Package*, 2022. R package version 6.0-93.

[10] Bardan Ghimire, John Rogan, Víctor Rodríguez Galiano, Prajjwal Panday, and Neeti Neeti. An evaluation of bagging, boosting, and random forests for land-cover classification in Cape Cod, Massachusetts, USA. *GIScience & Remote Sensing*, 49(5):623–643, 2012.

[11] Li Wen and Michael Hughes. Coastal wetland mapping using ensemble learning algorithms: A comparative study of bagging, boosting and stacking techniques. *Remote Sensing*, 12(10):1683, 2020.

[12] Amirhosein Mosavi, Farzaneh Sajedi Hosseini, Bahram Choubin, Massoud Goodarzi, Adrienn A Dineva, and Elham Rafiei Sardooi. Ensemble boosting and bagging based machine learning models for groundwater potential prediction. *Water Resources Management*, 35(1):23–37, 2021.

[13] Yash Singhal, Ayushi Jain, Shrey Batra, Yash Varshney, and Megha Rathi. Review of bagging and boosting classification performance on unbalanced binary classification. In *2018 IEEE 8th International Advance Computing Conference (IACC)*, pages 338–343. IEEE, 2018.

[14] Klemen Zakšek, Kristof Oštir, and Žiga Kokalj. Sky-view factor as a relief visualization technique. *Remote sensing*, 3(2):398–415, 2011.

[15] Ralf Hesse. Lidar-derived local relief models – a new tool for archaeological prospection. *Archaeological Prospection*, 17(2):67–72, 2010.

[16] Minrui Zheng, Wenwu Tang, Akinwumi Ogundiran, and Jianxin Yang. Spatial simulation modeling of settlement distribution driven by random forest: consideration of landscape visibility. *Sustainability*, 12(11):4748, 2020.