Machine Learning Aplicado ao Mapeamento Marinho: Caso de Estudo na Costa Oeste da Ilha da Madeira

D. Ceddia Porto Silva (1), L. Constante Reis (2)

(1) RJ/Brasil. diogoceddia@id.uff.br.

(2) Divisão de Geologia Marinha. Instituto Hidrográfico.

Resumo:

O mapeamento marinho é, desde os primórdios da navegação, componente estratégica de diversos povos que utilizaram os oceanos para seus objetivos. Com o avanço tecnológico, o rigor e performance das metodologias empregues para tais fins cresceram exponencialmente, alcançando a era dos computadores e algoritmos de inteligência artificial.

Esse trabalho explora uma metodologia de mapeamento marinho usando aprendizado de máquina (Machine Learning), aplicada à predição do afloramento rochoso junto à costa oeste da Ilha da Madeira, com dados obtidos no âmbito do projeto SEDMAR Madeira. A proposta é alcançada em diversas etapas: (1) tratamento dos dados, (2) aplicação e configuração do modelo, e (3) análise de coerência geológica.

Agora, para além de sua análise crítica e experiente, o especialista possui mais uma ferramenta de auxílio para aumentar sua eficiência: uma superfície do alvo, criada a partir de correlações estatísticas complexas e pormenorizadas, trabalhadas à excelência das ferramentas computacionais disponíveis.

Palavras-chave: Machine Learning, Mapeamento Marinho, Modelagem, Inteligência Artificial, Cartografia Geológica.

1. INTRODUÇÃO

O mapeamento marinho sempre foi de interesse de todos os povos com vínculo ao mar. Da vara de prumar até sistemas batimétricos multifeixe, mapas do fundo marinho foram progressivamente desenvolvidos e aperfeiçoados (OHI, 2005). Nesse âmbito, hoje, com o desenvolvimento computacional, técnicas de inteligência artificial podem ser testadas e desenvolvidas em ambiente doméstico, propiciando uma liberdade de desenvolvimento sem precedentes.

A proposta desse trabalho é o desenvolvimento de uma metodologia de mapeamento marinho utilizando o modelo de *machine learning* XGBoost. O alvo a ser mapeado é um afloramento rochoso, localizado a oeste da Ilha da Madeira. Os dados dispostos são batimetria multifeixe e sísmica monocanal de alta resolução (*sub-bottom profiler*). A batimetria corresponde a uma superfície de profundidades, abrangendo toda área de estudo. A sísmica, por sua vez, corresponde a dados dispostos em linhas de navegação – ou seja, uma distribuição geoespacial limitada. No entanto, é justamente a informação sísmica a referência principal para discernir a ocorrência ou não do afloramento rochoso. Nesse contexto, a aplicação do modelo de *machine learning* possibilita a extrapolação da caracterização do afloramento nas linhas sísmicas para toda extensão da superfície batimétrica.

Não obstante, os resultados serão avaliados segundo a sua coerência geológica, em que serão exploradas as vantagens, desvantagens e particularidades da área de estudo. Em uma visão geral, esse artigo apresenta brevemente uma metodologia e análise crítica para extrapolar qualquer tipo de informação geoespacialmente limitada – seja em função da natureza física do levantamento, dificuldades logísticas ou financeiras – , considerando critérios matemáticos mínimos.

2. ÁREA DE ESTUDO

Foi selecionada a costa oeste da ilha da Madeira como área de estudo, sendo particularmente complexa tanto na morfologia quanto na rugosidade do fundo. Apresenta um relevo irregular, alternando entre afloramentos rochosos e espessuras sedimentares de até 70 metros de profundidade, que delimitam o embasamento.

Todos os dados foram obtidos no âmbito do programa SEDMAR Madeira, em diversas campanhas a bordo do NRP Almirante Gago Coutinho, entre 2013 e 2018. Os dados batimétricos foram obtidos com os sistemas Kongsberg EM120 e EM710, e os dados de sísmica de alta resolução foram obtidos com os sistemas Echoes 3500 da Ixblue e Boomer AA200 da Applied Acoustics (Figura 1).

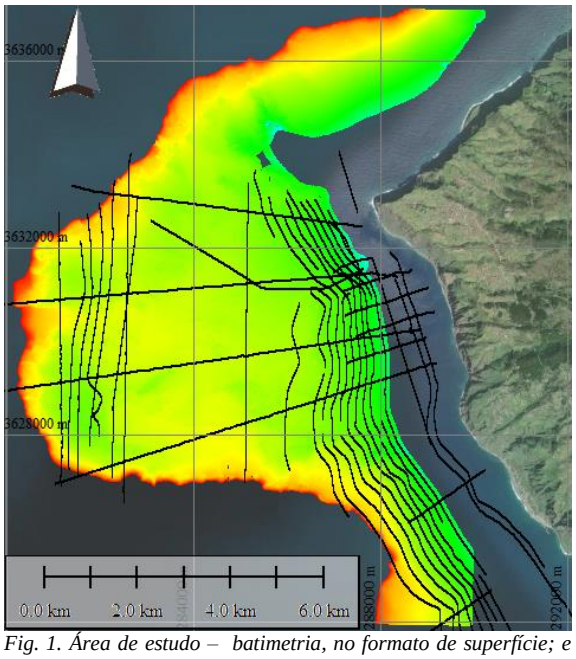

Fig. 1. Área de estudo – batimetria, no formato de superfície; e sísmica, representada em linhas. WGS84 / UTM 28N.

3. MATERIAIS E MÉTODOS

A modelagem utilizando algoritmos de inteligência artificial, incluindo modelos de *machine learning*, possuem diversas etapas fundamentais. Nesse trabalho, o seguinte fluxograma foi abordado: construção dos datasets; aplicação e configuração do modelo; e análise de coerência geológica, presente em todas as etapas deste trabalho para guiar a melhoria de performance do modelo.

3.1. Construção dos Datasets

O processo de construção dos datasets é a etapa que reúne, processa, filtra e condiciona os dados para serem inseridos no modelo desejado. Aqui, o objetivo é fornecer o melhor input possível para o modelo escolhido, com o máximo de informação relevante acerca do alvo de predição, mantendo qualidade, formato e condicionamento adequados (Huang *et al.*, 2015).

Para modelar o afloramento rochoso a oeste da Ilha da Madeira, foram dispostos dados de batimetria, *backscatter* e sísmica monocanal de alta resolução. A partir da batimetria, foram extraídos dados de declividade do fundo e orientação da declividade do fundo. Para aumentar a complexidade do modelo e a qualidade da predição, foi adotada a estratégia de *feature engeneering*. Ela consiste na criação de mais camadas de informação (ou *features*) para alimentar o modelo. Até ao momento são 4 *features*, todas extraídas da batimetria: profundidade, intensidade do *backscatter*, declividade do fundo e orientação da declividade do fundo. Através da técnica de *feature engeneering* foram criadas mais duas *features*: a distância de um dado ponto para a linha de costa e a distância de um dado ponto para o talude. Já a sísmica foi previamente classificada entre rocha e sedimento. Entretanto, durante a classificação e como consequência da resolução dos equipamentos envolvidos, toda espessura sedimentar da superfície do fundo marinho menor que 20 centímetros foi considerada afloramento rochoso. Dessa forma, o *dataset* é constituído de 6 *features* e um alvo (*target*), sendo ele rocha ou sedimento. Todas as *features* são sujeitas à etapa de filtragem, a qual valores discrepantes que não representam a grandeza física são eliminados.

As 6 camadas de informação (ou *features*) correspondem a *datasets* distintos. O objetivo final é ter somente um *dataset*, que contemple um ponto coordenado e seus respectivos atributos (profundidade, intensidade do *backscatter* etc.). Para tal, estabelecemos um ponto coordenado como *pivot* e buscamos seus vizinhos mais próximos das outras *features*. Para busca do vizinho mais próximo foi utilizada uma árvore de busca binária *k-d-tree*, que devido à sua complexidade de algoritmo possui crescimento linear em função do *dataset* e, portanto, apresenta custo computacional baixo (Bentley, 1975). Com os vizinhos mais próximos calculados, um novo *dataset* é construído, o qual associa um ponto coordenado a um valor de cada *feature*. Com o *dataset* montado, procede-se um filtro de natureza geoespacial, que descarta qualquer relação entre as *features* se a distância do vizinho mais próximo for superior a 10 metros. Para isso, foi utilizada a distância euclidiana.

Ao todo, são necessários dois *datasets* para esse trabalho: um para realizar a modelagem (*dataset 1*); e outro para extrapolação (*dataset 2*). O *dataset 1* contém as informações de classe (sedimento/rocha), necessárias para treinar o modelo e calcular métricas de assertividade. É sob o *dataset 1* que são feitas etapas de teste, configuração e ajuste do modelo. O *dataset 2*, por sua vez, corresponde a toda superfície das *features* que não inclui pontos coordenados com informação de classe – portanto, sem informação de validação. O *dataset 2* corresponde à superfície sobre a qual as classes serão extrapoladas/preditas.

3.2. Aplicação e configuração do modelo

Para o presente trabalho, foi aplicado o modelo de *machine learning* XGBoost. Essa escolha foi motivada pelo XGBoost ser considerado um algoritmo "estado-da-arte", tendo vencido diversos desafios no kaggle[1] e em competições[2] (Ariza-Garzon et al., 2020; Chen and Guestrin, 2016; Li et al., 2020). Como o *target* é uma classe (sedimento ou rocha), a abordagem desenvolvida consiste em uma modelagem supervisionada classificatória.

[1] www.kaggle.com/c/datachallenge (em 2015 o XGBoost esteve em 17 das 29 soluções vencedoras)
[2] KDDCup 2015 (XGBoost esteve na solução das 10 equipes vencedoras)

Os modelos foram submetidos a etapas de configuração: *tuning*, análise de *overfit* e filtro estatístico.

O *tuning* consiste em configurar valores para os hiperparâmetros do modelo que maximizem a sua métrica de assertividade (L. Zhang, C. Zhan; 2017). Foi utilizado *cross-validadion*[3] com 5 *folds* para calcular a acurácia balanceada[4], e o modelo foi hiperparametrizado iterativamente[5] até alcançar o melhor conjunto de parâmetros que maximizasse a métrica de desempenho.

A análise de *overfit*[6] consiste em determinar se o modelo consegue abstrair os dados de treinamento para além do domínio treinado. Um modelo em *overfit* é 'viciado' no *dataset* de treinamento, apresenta métricas de desempenho boas e é incapaz de realizar predições coerentes, já que só consegue predizer um universo restrito de *inputs* e não generaliza bem (Wang *et al.*, 2020).

O filtro estatístico foi utilizado com finalidade de 'limpar' falsos positivos do afloramento rochoso. Esse filtro baseou-se na eliminação de qualquer predição da classe rocha que tivesse menos de 90% de probabilidade no output do modelo[7].

*3.3. Análise de coerência geológica*Normalmente, os modelos de IA otimizam funções que aumentam métricas de desempenho. Entretanto, nem sempre os valores de desempenho matematicamente superiores traduzem os resultados com maior coerência geológica. Portanto, durante todas as etapas de ajuste do modelo foi analisado como os hiperparâmetros e filtros impactavam a superfície final, objetivando o melhor *output* sob o ponto de vista do especialista.

4. RESULTADOS

O modelo, de treinamento/validação (*dataset 1*) preliminar sem *tuning*, apresentou acurácia balanceada[8] de 98.876%. Com *tuning*, sua performance subiu para 99.471%, registrando uma melhora de 0.595%. Foi analisado também que o modelo não caiu em *overfit*.

Através do recurso de batimetria com sombra, nota-se que as rochas se expressam majoritariamente como uma rugosidade de fundo. Isso permite a avaliação da performance do modelo, conforme há consistência de coincidência da predição com a rugosidade do fundo. Sob esse aspecto visual, algumas regiões da superfície predita foram analisadas individualmente. Foram elas: uma região com grande densidade e continuidade de registros/linhas sísmicas (Figura 2); e outra com poucos registros sísmicos (Figura 3). Entretanto, como a interpretação sísmica classificou como afloramento rochoso espessuras sedimentares de até 20 centímetros, o modelo classificou como rocha alguns fundos sem rugosidade.

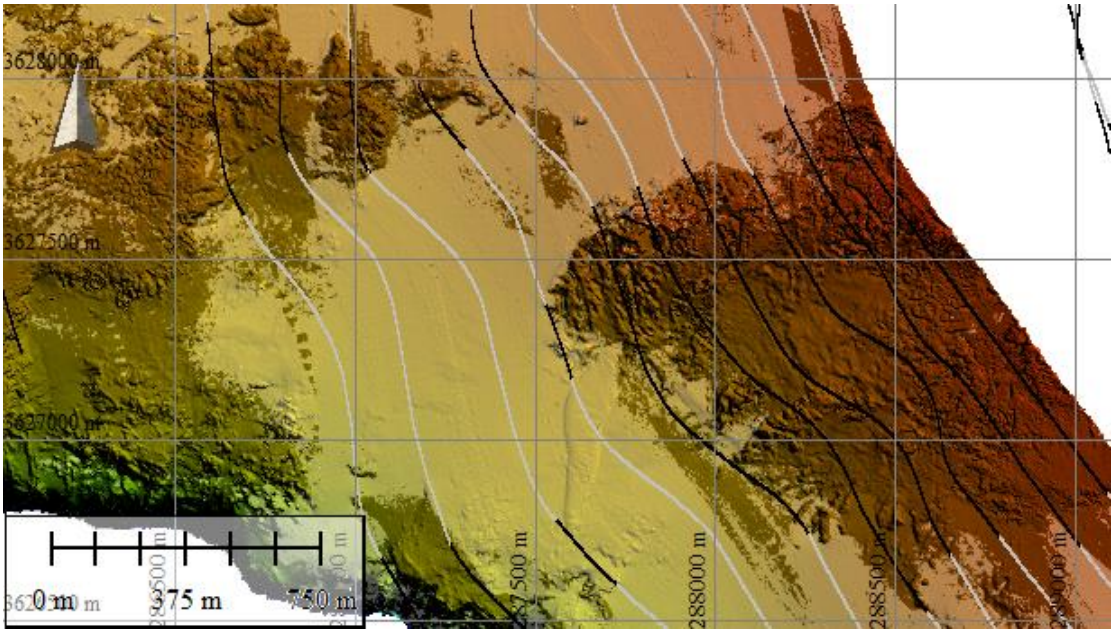

Fig. 2. Detecção do afloramento rochoso através de modelo XGBoost. As linhas correspondem à sismica, classificadas em rocha (cor preta) e sedimento (cor cinza). Ao fundo, superfície batimétrica com sombra, evidenciando rugosidade de fundo. Regiões mais escuras correspondem à predição do modelo do afloramento. Região com alta densidade e continuidade de linhas sísmicas. WGS84 / UTM 28N.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
[4] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html
[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[6] https://xgboost.readthedocs.io/en/latest/python/python_api.html
[7] https://xgboost.readthedocs.io/en/stable/prediction.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

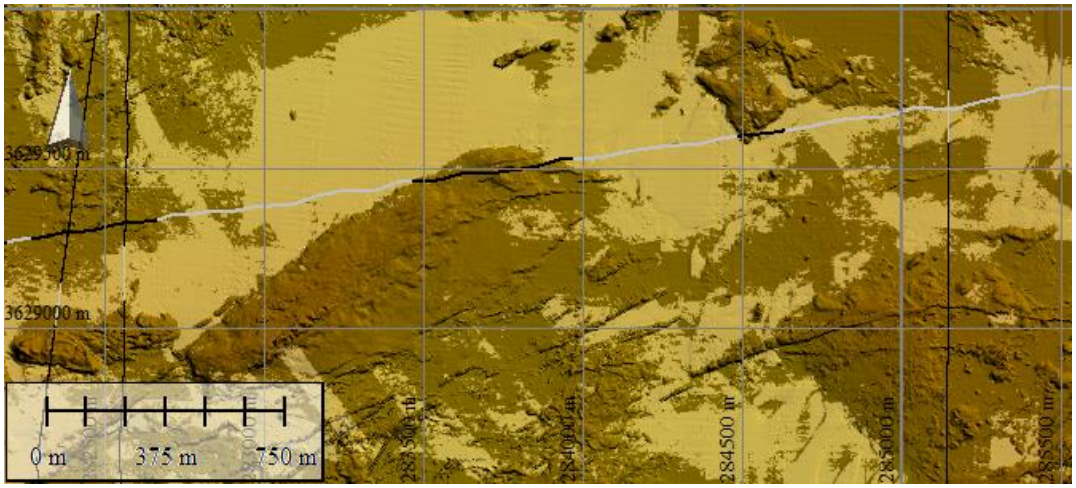

Fig. 3. Detecção do afloramento rochoso através de modelo XGBoost. As linhas correspondem à sismica, classificadas em rocha (cor preta) e sedimento (cor cinza). Ao fundo, superfície batimétrica com sombra, evidenciando rugosidade de fundo. Regiões mais escuras correspondem à predição do modelo do afloramento. Região com baixa densidade e continuidade de linhas sísmicas. WGS84 / UTM 28N.

5. CONCLUSÃO

O modelo XGBoost conseguiu alcançar uma métrica de acurácia balanceada muito satisfatória (99.471% com *cross-validation* de 5 *folds*), sem *overfit* e com devida hiperparametrização. Para extrapolação, o modelo apresentou grande coerência geológica de uma forma geral na área de estudo, se mostrando uma ferramenta promissora e eficiente para a abordagem.

No entanto, a área de estudo não apresenta uma homogeneidade na distribuição das linhas sísmicas. Isso possui dois impactos negativos no modelo: primeiro, o modelo adquire um viés, pois é treinado mais em uma subárea do que em outras; e segundo, que regiões com poucas linhas sísmicas ocasionam em perda de coerência geológica, pois o modelo não consegue predizer bem aquilo que não conheceu na etapa de treinamento. Para que o modelo obtenha o melhor resultado possível, as linhas sísmicas devem estar igualmente distribuídas em toda área de estudo, de forma que explique o máximo da heterogeneidade e complexidade da área. Tal resultado seria alcançado se, durante a aquisição da batimetria multifeixe, dados de sísmica monocanal de alta resolução fossem adquiridos concomitantemente.

Além disso, a superfície do afloramento rochoso feito pelo modelo pode ser exportado em formato ASCII, com pontos coordenados. Dessa forma, pela ótica do especialista que necessita mapear o afloramento, o modelo pode servir ou de referência ou de base para delimitar a fronteira sedimento/afloramento rochoso. Assim, diminui-se significativamente o tempo de trabalho de mapeamento, alcançando maior eficiência e produtividade.

Ao fim e como maior contribuição, esse trabalho sugere que, com uma boa metodologia e dados bem distribuídos geoespacialmente, é possível utilizar modelo de *machine learning* para realizar cartografia sedimentar ou outros mapeamentos expeditos.

Agradecimentos

Os autores agradecem à Doutora Aurora Rodrigues Bizarro pela cedência dos dados.

REFERÊNCIAS

Ariza-Garzon, M.J., Arroyo, J., Caparrini, A., Segovia-Vargas, M.J., 2020. Explainability of a Machine Learning Granting Scoring Model in Peer-to-Peer Lending. https://doi.org/10.1109/ACCESS.2020.2984412.

Bentley, J.L., 1975. Multidimensional Binary Search Trees Used for Associative Searching. Commun. ACM 18, 509–517.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. https://doi.org/10.1145/2939672.2939785.

Jianglin Huang, Yan-Fu Li, Min Xie, (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and Software Technology. https://doi.org/10.1016/j.infsof.2015.07.004.

L. Zhang, C. Zhan (2017). Machine Learning in Rock Facies Classification: An Application of XGBoost. https://doi.org/10.1190/IGC2017-351.

Li, X., Luo, J., Jin, X., He, Q., Niu, Y., 2020. Improving soil thickness estimations based on multiple environmental variables with stacking ensemble methods. Remote Sensing 12, 1–21. https://doi.org/10.3390/rs12213609.

M. Wang, D. Huang, G. Wang, D. Li (2020). SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes. Journal of Geotechnical and Geoenvironmental Engineering. Volume 146. https://doi.org/10.1061/(asce)gt.1943-5606.0002297.