Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-22-RC1, 2020

Open Access Earth System Science Data Discussions

Interactive comment on "Deep-sea sediments of the global ocean" *by* Markus Diesing

Everardo González Ávalos (Referee)

egonzalez@geomar.de

Received and published: 26 May 2020

A lithology map for the seafloor below 500m depth was created with the use of the Random Forests technique. The algorithm was trained using a set of 8 global predictor variables and around 10000 measurements of seven lithology classes as response variables and achieved an accuracy of 69.5%. Two byproducts of the analysis were the probability maps for individual sediment classes and an assessment of predictor variable importances.

Since I am not an earth scientist by formation, I do not consider myself qualified to make an assessment of novelty and usefulness from a geological point of view and will not do so. I will instead comment on the general content and the technical side of the article, focusing on the materials and methods.

The article was structured clearly with a meaningful division in the following sections:

introduction, data, methods, results, limitations of the approach, potential usage, data availability, and conclusion. These sections are streamlined towards the understanding of the algorithmic implementation and its results; they retain completeness while remaining pleasantly concise, "Limitations of the approach" being the only exception to this. All accompanying figures and tables are clear and understandable, both, in digital form and in paper.

The software was tested for reproducibility using the ERC tool under https://o2r.uni-muenster.de/#/erc/GWME2voTDb5oeaQFuTWMCEMveKS1MiXm, and performed positively in this aspect. Upon closer examination, the discrepancies that led to it being flagged with failed reproducibility multiple times, appear to be minor formatting changes. The data products found under https://doi.pangaea.de/10.1594/PANGAEA.911692 are accessible, complete, and use standard file types.

For the most part, the methodology was clearly explained, with enough references to the sources for the used techniques as well as a clear specification of the software implementations used. In contrast to this, the following considerations regarding the selection of input and output variables for the Random Forest algorithm did not seem properly addressed: The importance of eliminating correlated input variables was not discussed. Considering that the Random Forests algorithm allows for non linear input relationships, a justification for this does not seem obvious, particularly since its results contradict those of the Boruta algorithm, which deemed all 38 predictors as important. The selection of 7 lithology classes as response variables instead of 5 or 13 seemed arbitrary, but the repercussions of this choice are big: the great disparity of class contribution to the test dataset (44.8% for the most represented class v.s. 0.9% for the least represented class) render model accuracy less suitable as a performance metric. The reason for this becomes evident when looking at the high errors of commission and omission of the underrepresented classes.

The above considerations do not in anyway discredit the relevance of the results, in fact the contrary is true: the prediction accuracy for the two most common sediment classes, calcareous sediment and clay, is above the overall accuracy. In the predictions, these two classes combined make up for almost 90% of the total seabed. In contrast to this, the 0.4% share of mixed calcareous-sciuliseous ooze with an error of commission of 75% make for a statistically insignificant portion of the results; however this also makes its inclusion in this analysis questionable. The same argument can be done for the radiolarian ooze and lithogenous sediment classes.

The discussion of the predictor variable importance and the inclusion of class probability maps in the results made for a great addition to the analysis. Interpretability in machine learning is an important subject currently gaining much deserved attention and it makes an argument for the use of the Random Forests algorithm instead of other techniques such as neural networks, the latter being generally regarded as more sophisticated but requiring bigger efforts to achieve interpretability.

The overall quality of the article was satisfactory to me. It showed good understanding of the machine learning methods used and displayed outstanding transparency in their software implementation. In addition to this, the analysis of predictor variable importance and the individual probability maps in the results made for two strong points. A more detailed discussion of both, the reasoning behind the selection of predictor variable, and its effect on the results and variable predictor importance would be desirable. Since the most underrepresented response variable classes have extremely low impact on the result meaningfulness, their exclusion might be advisable. Finally, if the disparity between response variable classes should remain as large as it currently is, replacing accuracy with another performance metrics such as Intersection over Union (averaging the IoU for all individual classes) would account for a more fair result interpretation and allow for better performance comparisons in future works.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-22, 2020.