Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-22-RC2, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Open Access Earth System Science Data Discussions

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

Anonymous Referee #2

Received and published: 29 May 2020

General comments: The manuscript submitted by Markus Diesing presents a global scale modelling approach of the deep-sea sediments spatial distribution. This important topic for marine geology and related fields, came to the fore decades ago (the 1970s), and this paper shows how newer available global data combined with machine learning, can contribute to this direction. However, a similar approach with data from the same dataset has been presented by Dutkiewicz et al., 2015. Also, the controls on the distribution of deep-sea sediments (based on the analysis of this dataset) have been extensively discussed by Dutkiewicz et al., 2016. Both studies are mentioned in the manuscript. Nevertheless, the different applied algorithm, workflow, and the inclusion of the sea-surface iron concentration differentiate this effort. The author made a point of being very transparent about his approach, providing a thorough workflow in a well-written manuscript. Moreover, the data and code availability shows the path for similar studies promoting scientific research and knowledge.

Specific comments: Although the overall good quality of the manuscript, there are issues that could be addressed further. These are divided into different sections, as follows:

Manuscript structure: It is a well-constructed manuscript; however, it would be better to move the section 3.5 Environmental Space before section 3.3 Random Forest classification model. The quality characteristics (i.e. representatives) of the training samples have to be examined and assessed before any modelling approach. Similar, the results of this analysis can be presented before the model results.

Selection of predictor variables: a) The author, performed a detailed feature selection, trying to include all the relevant and not redundant predictors to this problem by applying a two-stage process: Firstly, the Boruta algorithm, and secondly a correlation filtering. While there is a detailed analysis of the Boruta algorithm, the correlation filtering is not analysed. The same is also pointed from the first Reviewer. It is not clear which correlation coefficient has been used, and the significance of these correlations. The reasons for excluding high-correlated predictor variables could be analysed further, mainly focused on the RF performance and potential bias in cases of high-correlated predictors, as in this study the variable importance & interpretation has increased value.

b) Dutkiewicz et al., 2015 and Dutkiewicz et al., 2016, referred to the absence of iron concentration as potential predictor variable form their model and correlation analysis, respectively. In this study, the author initially includes iron concentration as a potential predictor variable. However, in the final model, the iron concentration is not used. Based on the results from Boruta algorithm, the iron concentration is in a relatively high position (ranked 17th/38). Thus, further analysis of this exclusion is needed. To the same direction, it would be interesting to include a boxplot analysis between the iron concentration and the various lithology classes, enlighten further the understanding of the causal mechanisms. Such analysis is provided only for the final selected predictors inside the script.

c) The Carbonate Compensation Depth is considered as an essential factor for the distribution of the deep-sea sediments. This parameter is captured by seafloor lithology in the analysis. Nevertheless, it would be interesting to be discussed the presence of available global CCD models (and its limitations) that could be used or not as a predictor.

Random Forest (RF) modelling: The use of the RF, seems appropriate for this kind of analysis, due to interpretability that offers. The author followed an analytical workflow, including data pre-processing, stratified splitting between training & test datasets & model parameter tuning through 10-fold cross-validation. However, there are methodological issues that could be addressed further:

1a) The author uses the unscaled mean decrease in accuracy as a variable importance measure. Although this is the recommended approach (e.g. Strobl et al., 2007; Strobl et al., 2008), it would be better if the reasons behind this choice are stated inside the text.

1b) In the manuscript, it is not stated if the RF sub-sampling of predictor variables is performed with or without replacement. In the provided script seems to be with replacement (as the default option in the used package). However, studies have shown that this approach can be biased when predictor variables vary in their scale and/or in their number of categories (e.g. Strobl et al., 2007). Also, it is not also clear if any type of feature scaling has been applied in the predictor variables before modelling. In case that the author would like to continue without feature scaling for the predictor variables, packages like party seem to provide more unbiased results. In any case, it would be good to be provided with a more comprehensive explanation, as the model interpretability is one of the main targets in this work.

2a) Studies have shown that the traditional cross-validation can result in overoptimistic errors when applied in spatial data (e.g. Roberts et al., 2016). Consequently, a spatial model should also include a spatial cross-validation analysis. Moreover, in cases of

spatially unbalanced class distribution, stratified cross-validation can be applied (e.g. Lawson et al., 2017). Here, train & test samples were split with stratification, but the training was conducted only with cross-validation. Recently published spatial versions of RF could alternatively help to this direction (e.g. Hengl et al, 2018, Georganos et al, 2019). The existing dataset (despite the tremendous effort of Dutkiewicz et al., 2015) is spatially imbalanced, with areas have experienced heavier sampling efforts than other, making even more important the concept of spatial cv or/and spatial RF. The author addresses this issue by setting a minimum distance among the training points and by removing duplicates.

2b) Despite the selected class simplification, the two main classes still count for the 77.6% of the training and test sample, resulting in relatively high overall accuracy, but with limited accuracy in the rare classes. Considering the availability of methods and algorithms that try to overcome these class imbalances, (e.g. weights, penalty costs, over/under-sampling. SMOTE, Isolation Forests) it would be interesting to see how further can be improved the performance compared to the presented (baseline) model. The overall accuracy as a performance metric is not the best option in such situations (it is also mentioned from the first Reviewer) However, the no information rate is provided, showing that there is still gain.

Results The results show good overall agreement with the above mentioned previous mapping efforts. However, the comparison demands parallel examinations of the maps from the two papers. Given the availability of the results from Dutkiewicz et al., 2015, it would be interesting to include a direct categorical map comparison between the two approaches (after the proper modifications due to the different number of lithological classes) showing clearer the areas with the highest agreement and disagreement. It is helpful that the author included the probability prediction of each class, strengthen the interpretation analysis of the results.

Technical corrections In Figures 1 & 6a, the use of purple colour for the Mixed Ooze is not ideal, as it has limited contrast with the background map and its surrounded

classes. An essential part of this study and its results is related to map creation and interpretation. Consequently, the use of colourblind-friendly palette is recommended, making the manuscript more comfortable on a broader audience.

Hope this is helpful!

Please also note the supplement to this comment:
https://www.earth-syst-sci-data-discuss.net/essd-2020-22/essd-2020-22-RC2-supplement.pdf

---

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