Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-22-AC2, 2020

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

Markus Diesing

markus.diesing@ngu.no

Received and published: 9 July 2020

Referee 2 attested overall good quality, but identified several issues that could be further addressed:

1. Move the section 3.5 Environmental Space before section 3.3 Random Forest classification model.

Agreed, the section was moved accordingly.

2. Similar, the results of this analysis can be presented before the model results.

Agreed, the section was moved accordingly.

3. 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.

The general strategy for predictor variable selection was to initially find all predictors that are potentially relevant, then reduce redundancy by limiting the set of predictors to those that are uncorrelated. Such an approach has also been advocated by Millard and Richardson (2015). Finding important predictors is achieved with the Boruta algorithm, an "all-relevant" feature selection method (Nilsson et al., 2007), which identifies all predictors that might be relevant for classification (Kursa and Rudnicki, 2010). Limiting redundancy is subsequently achieved by a correlation analysis. To do so, it is necessary to define a critical value of the correlation coefficient r. This is arguably subjective, and the choice of r will influence the number of predictor variables that are selected. To investigate the influence of r on the OOB error of a Random Forest model (using default parameter values), several values of r between 0.1 and 1 (step 0.01) were now trialled. When the OOB error is plotted over r, it is apparent that initially (with small values of r) the OOB error drops off quickly but stabilises at values of r approximately 0.4 – 0.5. As many predictors potentially lead to overfitting, increase processing time, and decrease interpretability of the model, it was decided to select a small value of 0.5. This led to eight predictors being selected.

For clarity, the section on predictor variable selection was updated.

4. 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.

Dutkiewicz et al. (2016) highlighted that iron concentration in surface waters might be an important predictor, as phytoplankton blooms (e.g. diatoms) are enhanced by iron fertilisation. However, they also point out that the Southern Ocean where diatom oozes are abundant, receives very little iron. In fact, productivity is not the only factor determining the composition of deep-sea sediments, dissolution during sinking and dilution with other materials also have to be considered.

Iron concentration was initially included in this study. However, iron concentration was not included in the final model because it was correlated (above the selected threshold) with sea-floor minimum temperature. Box plots of iron concentration versus seafloor lithology did not reveal high discriminatory power of this predictor (see attached figures). Because of that and because it was intended to keep the variable selection process as “automated” as possible, it was decided not to intervene and force the inclusion of iron concentration as a finally selected predictor.

5. 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.

The Carbonate Compensation Depth (CCD) is potentially another important predictor that was not included in the model. This was due to the absence of relevant datasets in the utilised databases. A literature search did not yield any publications with associated datasets that could be used. However, it is assumed that this missing information is (at least partly) provided by other predictors, such as water depth and bottom water temperature.

6. 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.

Agreed. The text has been updated accordingly and now reads:

RF also provides a relative estimate of predictor variable importance. The importance() function of the randomForest package allows to assess variable importance as the mean decrease in either accuracy or node purity. However, the latter approach might be biased when predictor variables vary in their scale of measurement or their num-

ber of categories (Strobl et al., 2007) and was not used here. Variable importance is therefore measured as the mean decrease in accuracy associated with each variable when it is assigned random but realistic values and the rest of the variables are left unchanged. The worse a model performs when a predictor is randomised, the more important that predictor is in predicting the response variable. The mean decrease in accuracy was left unscaled as recommended by Strobl and Zeileis (2008), and is reported as a fraction ranging from 0 to 1.

7. 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.

Agreed. The predictor variables have now been scaled and the subsampling of predictor variables is performed without replacement. The manuscript was updated accordingly.

8. 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.

Agreed. I have now implemented a spatial leave one out cross validation scheme to test the accuracy of the model in a more robust way. The spatial autocorrelation distance is determined with the spatialAutoRange function of the blockCV package. This value is utilised to determine the buffer size around observations which serve as a test point. Details can be found in a new section “3.5 Spatial cross-validation”. However, this meant that model tuning would have become very complex and even more time-consuming.

As model tuning gave a very limited gain in performance, it was therefore decided to run the Random Forest with default parameter values.

As a result of this more robust estimation of map accuracy, the accuracy of the model is now lower (or rather less inflated), but still significantly larger than the no information rate.

9. 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.

The problem of class imbalances is now addressed by utilising a balanced version of Random Forest. This is achieved by using the strata and sampsize arguments of the

randomForest function. Here, we stratify by lithology class. The sampsize is then set to the same number for every class. This means that the class with the lowest frequency (Radiolarian ooze) determines the number of observations used to fit individual trees. However, each sample is still drawn from all available observations and hence this approach is likely more effective then downsampling the whole dataset prior to model building.

In addition to the overall accuracy, I have now also included the balanced error rate, which is more appropriate for datasets with unbalanced class frequencies.

As a result, the final map has a different appearance in some areas of the global ocean. It now approximates hand-drawn maps of the distribution of deep-sea sediments in much more detail. For example, equatorial patches of radiolarian ooze in the Indian Ocean are visible now. A near-equatorial band of radiolarian ooze in the eastern Pacific is now visible, too.

10. 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 was not the intention of this contribution, and might go beyond the scope of a data description paper, to compare the final map with that of Dutkiewicz et al. (2015). I would prefer to leave such an analysis to whoever is interested in it. The map products of both publications are readily available.

11. 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.

I agree that the choice of the colour scheme should be inclusive, but when consulting colorbrewer2.org, I could not find a colour-blind safe option for qualitative data with 5 classes. (NB: The classification has been reduced to the five classes used by Dutkiewicz et al. (2016), as suggested by reviewer 1.) The purple colour has, nevertheless, now been removed, as the respective class is no longer included. I also removed the hillshade bathymetry to make the map clearer.

References

Dutkiewicz, A., Müller, R. D., O'Callaghan, S. and Jónasson, H.: Census of seafloor sediments in the world's ocean, Geology, 43(9), 795–798, doi:10.1130/G36883.1, 2015.

Dutkiewicz, A., O'Callaghan, S. and Müller, R. D.: Controls on the distribution of deep-sea sediments, Geochemistry, Geophys. Geosystems, 17(8), 3075–3098, doi:10.1002/2016GC006428, 2016.

Kursa, M. and Rudnicki, W.: Feature selection with the Boruta Package, J. Stat. Softw., 36(11), 1–11 [online] Available from: http://www.jstatsoft.org/v36/i11/paper/, 2010.

Millard, K. and Richardson, M.: On the importance of training data sample selection in random forest image classification: A case study in peatland ecosystem mapping, Remote Sens., 7(7), 8489–8515, doi:10.3390/rs70708489, 2015.

Nilsson, R., Peña, J. M., Björkegren, J. and Tegnér, J.: Consistent feature selection for pattern recognition in polynomial time, J. Mach. Learn. Res., 8, 589–612, 2007.

Strobl, C. and Zeileis, A.: Danger: High Power! – Exploring the Statistical Properties of a Test for Random Forest Variable Importance, Munich. [online] Available from: https://epub.ub.uni-muenchen.de/2111/1/techreport.pdf, 2008.

Strobl, C., Boulesteix, A.-L., Zeileis, A. and Hothorn, T.: Bias in random forest variable

importance measures: Illustrations, sources and a solution, BMC Bioinformatics, 8(1), 25, doi:10.1186/1471-2105-8-25, 2007.

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2020-22/essd-2020-22-AC2-supplement.zip

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

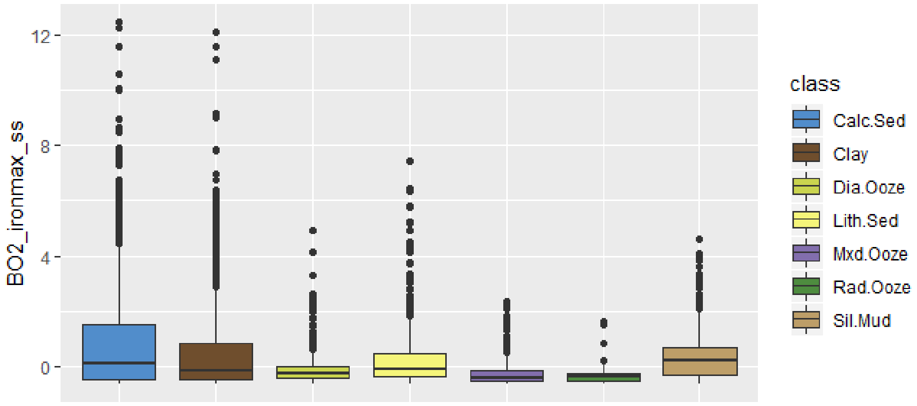

Fig. 1. Iron concentration (max)

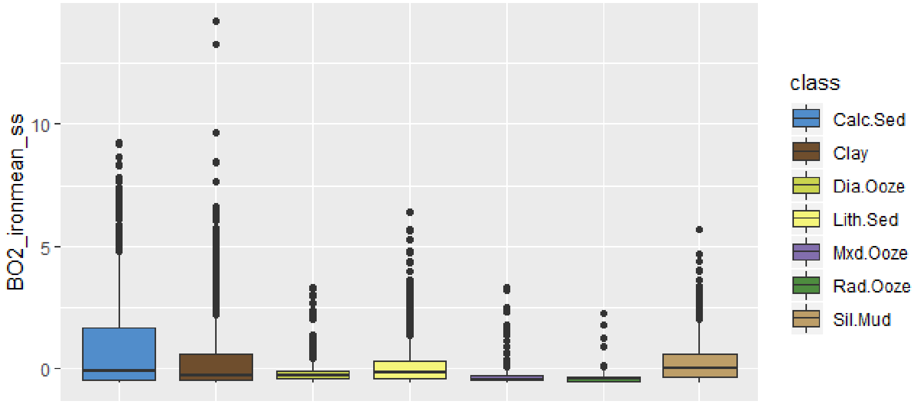

Fig. 2. Iron concentration (mean)

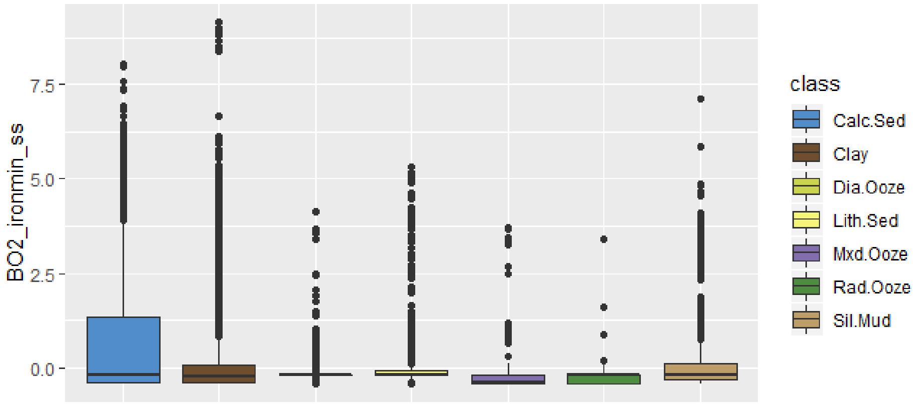

Fig. 3. Iron concentration (min)

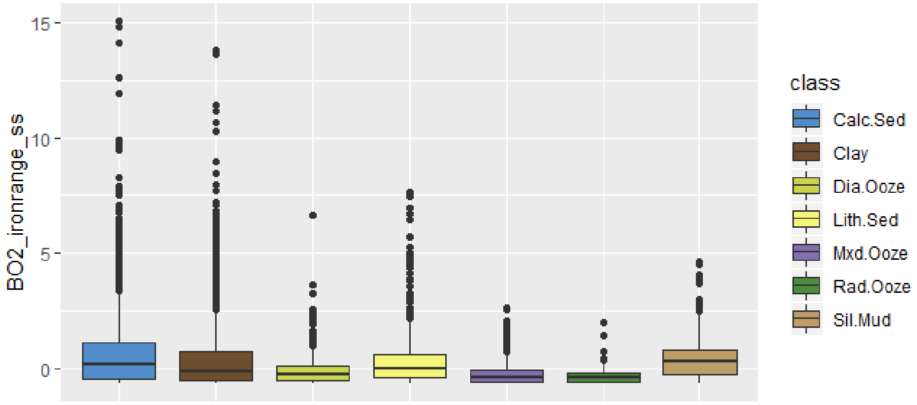

Fig. 4. Iron concentration (range)