Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-22-AC1, 2020

Open Access Earth System Science Data Discussions

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 1 was generally positive about the submitted manuscript, but pointed out the following issues that should be addressed:

1. 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 predictor variables/features is a two-step process. The first step (Boruta algorithm) narrows the selection to those variables that are relevant. In this case, all variables were identified as relevant/important. It is important to note that the Boruta algorithm is an "all-relevant" feature selection method (Nilsson et al., 2007), which identifies all predictors that might be relevant for classification (Kursa and Rud-

nicki, 2010). It does not address the question of redundancy in the predictor variable data. To limit redundancy, the second step seeks to identify predictor variables that are correlated with other predictors of higher importance. 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 a large number of predictors potentially leads to overfitting, increases processing time and decreases interpretability of the model, it was decided to select a small value of 0.5. Referee 1 argues that the Random Forest algorithm allows for non-linear input relationships and hence the elimination of correlated input variables was unnecessary. This claim has indeed frequently been made; however, some studies recommend reducing high-dimension datasets to uncorrelated important variables (Millard and Richardson, 2015). In this case, the number of predictor variables has limited influence on prediction performance but marked influence on processing time and interpretability of the results. Therefore, a low number of predictors was selected. This selection of eight variables is also not in contradiction to the results of the Boruta analysis, as the Boruta algorithm identifies relevant features (all-relevant problem, Nilsson et al., 2007), while for model building a low number of variables that achieve comparatively high accuracy was the goal (minimal-optimal problem, Nilsson et al., 2007). I agree that the reasoning for carrying out a two-step variable selection process was not as clear as it could be. Therefore, changes were made to the relevant section in the manuscript.

2. 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.

Following the advice from reviewer 1, it was decided to use the five classes of Dutkiewicz et al. (2016). These are broadly in line with lithology classes depicted in textbook maps and comprise Calcareous sediment, Clay, Diatom ooze, Radiolarian ooze and Lithogenous sediment. This choice still introduced large imbalances in class frequencies, as pointed out by both reviewers. To deal with the issue, a balanced version of Random Forest has now been utilised. This is achieved by using the strata and sampsize arguments of the randomForest function. In this case, 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.

3. 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.

I believe that overall accuracy as a metric for the performance of the model still has some meaning. Basically, if a map has an accuracy of say 60%, then 60 out of 100 randomly placed points in the map will likely be classified correctly. That information is still of importance. However, I agree that this metric is less well suited to give information on rare classes. Such information was conveiyed by the class-specific metrics error of commission and error of omission. To further address this shortcoming, I am now also providing the Balanced Error Rate (BER). The BER treats all mapped classes equally, as it is the mean error of all mapped classes, regardless of how often a class is contained in the test set. Together, both global metrics give a more comprehensive picture to assess the accuracy of the map.

References

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.

Please also note the supplement to this comment:
https://essd.copernicus.org/preprints/essd-2020-22/essd-2020-22-AC1-supplement.zip

---

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