Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-353-AC2, 2018

Biogeosciences Discussions

Interactive comment on "Quantitative mapping and predictive modelling of Mn-nodules' distribution from hydroacoustic and optical AUV data linked by Random Forests machine learning" *by* Iason-Zois Gazis et al.

Iason-Zois Gazis et al.

igazis@geomar.de

Received and published: 5 October 2018

Reviewer #2 General comments:

This paper was very interesting to read and clearly demonstrates how combining several state of the art scientific tools can achieve results that were, until recently, difficult to produce. The main idea in the manuscript; using Machine Learning to derive abundance of nodules from predictor variables remotely sensed with an AUV, has been applied by these authors and others datasets and this paper combines data and method-

ologies that have both been featured in other publications (cited in the paper). However, it presents a thorough protocol to make use of these tools, combine and optimize them, account for known caveats in the procedure and demonstrates the applicability of this protocol in a practical situation. The scientific approach is complex but transparently detailed throughout the method, results and appendixes. Thus, this paper is a useful case study and a method that should be applicable to other similar datasets and, as such, is a valuable contribution to the exploration of the Manganese nodules fields in the CCZ. It is well written but could be streamlined and made easier to read. The important findings could be further highlighted in the results section by moving some of the subsections in the appendix (as highlighted by reviewer 1). In addition, I found that several sentences or groups of sentences in the discussion either were confused in their formulation or didn't make a clear point. Furthermore, the discussion could be structured into several paragraphs to help readers perceive the different points made by the authors.

Authors comments:

We welcome all comments of Reviewer #2 and we appreciate the time and effort put to review this manuscript. Below we present our reply for each of the reviewer's points:

Similar to Reviewer #1, Reviewer #2 highlights the transparent, thorough and well-written workflow, and notices also that this methodology has been applied in the past, as well as the need for a different structure of some parts in the manuscript. Both comments have been already answered to Reviewer #1 and are considered in the revised version. In addition, we followed the recommendation of the Reviewer #2 to divide the discussion part into several paragraphs in an effort to state clearly the points of our study.

Reviewer #2 Specific comments:

I also have a couple of specific remarks and suggestion to add to those of reviewer 1: R400: If RF is not good at predicting outside the ranges of the training set, could it

affect the projected map of nodule abundance? Other studies projecting RF models (of species distribution) in space (or time) have used multivariate environmental similarity surfaces (MESS) maps (Elith et al. 2010). This procedure is mapping how dissimilar to known data points the predictors are across the projection area. This could potentially highlight that predictions in deeper and shallower areas than where nodule abundance samples are should be considered with care. This could also help target areas for future sampling. See Elith J, Kearney M, Phillips S (2010) The art of modelling range-shifting species. Methods in Ecology and Evolution 1:330-342

Authors comments:

Doubtlessly, the 'weakness' of the RF method is to predict outside of the range of the training set, this can influence the accuracy of the final abundance map. The need for extrapolation is always given in deep ocean studies by the limited numbers of actual samples. The problem of having not 'entirely' representative samples can only be solved by collecting a great number of sample points (like images in our case) that are well-distributed inside the study area (i.e. data that will include the entire range of the number of Mn-nodules/m2 and they are come from all the different sub-terrains). The comparative use of different machine learning algorithms (Support Vector Machines and Artificial Neural Networks) for the same dataset, which are able to extrapolate beyond the training range (e.g. Balabin and Lomakina, 2011; Martious and Lambert, 2017), can reveal the size of this 'weakness' in RF predictions. Such extrapolated predictions should be treated carefully regarding their accuracy and should always been validated with samples from the outer parts (lower and upper) of the training range. The main difficulty of our approach, is the need for different representative large training data in every different study area. The use of multivariate environmental similarity surfaces (MESS) can contribute to Mn-nodule exploration, by indicating other similar Mn-nodules fields in the wider area, based on the similarity of morphological characteristics of the already studied areas. To our knowledge the combined use of RF and MESS has not been applied yet as. Elith et al. (2010) used Boosted Regression Tree

(BRT) and Maximum Entropy (MaxEnt) machine learnings approaches; an approach interesting for future studies. Another promising, although complex would be the use of the Transfer Learning Approach. This approach can overcome the drawback of traditional machine learning, in which the training predictive algorithms should be trained each time based on previously collected (labelled or unlabeled) data from the study area. By using Transfer Learning, one can take an already trained model and transfer the part of the model that contains the necessary built relationships into a new model (usually smaller) that has to learn only the extra relationships/patterns that may exist in the new study area (e.g. Pan and Yang, 2010; Lu et al, 2015). Thus, the non-linear relationship between the number of Mn-Nodules/m2 and the topographic factors can be transferred and applied to other potential areas, where there is a lack of labelled optic data, and may include slightly different bathymetric range and topographic characteristics.

1. Balabin, R.M. and Romakina, E.I.: Support vector machine regression (LS-SVM) - an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? Phys. Chem. Chem. Phys., 13, 11710–11718, 2011. https://doi.org/10.1039/c1cp00051a

2. Martius, G., Lampert, C.H.: Extrapolation and learning equations. CoRR abs/1610.02995, 2016. http://arxiv.org/abs/1610.02995

3. Elith, J., Kearney, M. and Phillips, S.: The art of modelling range-shifting species. Methods in Ecology and Evolution, 1, 330–342, 2010. https://doi.org/10.1111/j.2041-210X.2010.00036.x

4. Pan, S. J., and Yang, Q.: A Survey on Transfer Learning. IEEE Transactions on knowledge and data engineering, Vol.22, No. 10, 1345-1359, 2010.

5. Lu, J., Behbood, V., Hao, P., Zuo, H., Xue, S., Zhang, G.: Transfer learning using computational intelligence: A survey. Knowledge-Based Systems, 80, 14–23, 2015. http://dx.doi.org/10.1016/j.knosys.2015.01.010

R415: The relevance of the depth as the most important predictor could be discussed further. Is there a geological reason why depth is the main driver of nodules distribution (as it looks unintuitive as to why such small changes in depth could drive nodule distribution)? Is it likely to be a proxy for another driver?

Authors comments:

This question is also asked in the specific comment (7) from Reviewer #1, and it is answered there.

R499: Minor point but Judging by figure 12, the relation between MSR and the different tuning parameters, particularly the number of training samples is not linear and thus, could either increase asymptotically towards a maximum or might continue increase logarithmically. Either way, It is unclear if more data would be a major improvement. Thus, collection of new data should focus on better-distributed data

Authors comments:

Indeed, it is not clear if more data would be a major improvement. The availability of more data and especially if they were better distributed, would most likely reinforce the model to build better and wider relationships between the predictor and response variables. This would allow keeping a larger number of validation data points. The need for more and better-distributed data has been stated in lines 498 – 504, especially when considering the spatial clustering inside the study area. The influence of the number of training data for model performance still remains a discussion point between studies showing an improvement by adding more data (e.g. Bishop, 2006), and other studies presenting stable performance of the model even if more data are added (e.g. Zhu et al, 2012).

1. Bishop, C.M.: Pattern Recognition and Machine Learning. Information Science and Statistics. Springer, Heidelberg, 2006.

2. Zhu, X., Vondrick, C., Ramanan, D., and Fowlkes, C. C.: Do We Need More Training

Data or Better Models for Object Detection? In BMVC, 3, 5, 2012.

R510: Given the rarity of corers data compared to photo data, would it not be better to take all cores where there is photos to strengthen the comparison between the two nodule counting methods? The photos of areas where some of the cores have been taken can still be excluded from the RF model and externally validated afterwards in order to make the best use of available ship time and data.

Authors comments:

In an ideal scenario with a sufficient number of available box-corers (many box corers, in which we are referring in the manuscript), both scenarios should be applied. The greatest number of them should be deployed in areas with photos in order to calculate better the factor between counted Mn-nodules in photos and in box-corers, and the rest in areas without photos in order to estimate the accuracy of the model in areas far away from the optic data but still inside the study area. However, in a realistic scenario the amount box-core samples will always be limited and thus they should be deployed in areas with photos to establish a better relationship between these two quantitative methods.

And a few technical corrections and suggestions -Authors comments:

R56: "data points"? "Data sets"? – Done we use data sets

R180: could you specify what the correction would be? - Done. This correction can be a simple factor that describes the ratio between the number of Mn-nodules seen in the photo and the number of nodules counted in box-corers (considering for the different spatial scales). Kuhn and Rathke (2014) used this approach, but also considered two different nodule size spectra.

1. Kuhn, T., Rathke, M.: Report on visual data acquisition in the field and interpretation for SMnN. Deliverable D1.31 of the EU-Project Blue Mining. BGR Hannover, 34 pp. 2017. www.bluemining.eu/downloads

R474: "resulting in biased results where Mn-nodules are bigger"? – Done (This phrase is in lines 472-473). The sentence has been changed.

R480: This is true, but is it necessary to state that here? Maybe it could be moved to the introduction

R476 - 485: It is hard to follow the authors point here. Do you mean that the observed influence of bathymetric factors on the nodule distribution cannot necessarily be explained? This observation is an interesting fact in itself and may lead to a better understanding of an underlying Mn-nodule formation process?

R490: "as it ignores"? – Done

R490: "To this end, several authors, have included the values of latitude/longitude and even LMI as predictor variables"? – Done (based on Referee's #1 suggestion)

516: "thus, high priority areas (e.g. these with highest commercial interest) can be targeted for sampling based on the results of optic data and RF modelling"? – Done

Please also note the supplement to this comment:
https://www.biogeosciences-discuss.net/bg-2018-353/bg-2018-353-AC2-supplement.pdf

---

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-353, 2018.