Solid Earth Discuss.,
https://doi.org/10.5194/se-2019-58-AC2, 2019

Solid Earth
Discussions

Interactive comment on "Prediction of seismic p-wave velocity using machine learning" *by* Ines Dumke and Christian Berndt

Ines Dumke and Christian Berndt

cberndt@geomar.de

Received and published: 9 August 2019

The manuscript, " Prediction of seismic p-wave velocity using machine learning", is a well-written description of a machine learning method -Random Forests - to predict seismic p-wave velocity as a function of depth for any a generic marine location. This manuscript is suitable for Copernicus, but the manuscript needs to be revised before it can be accepted. I have some suggestions here.

1 Introduction.

Page 2:

L24: You make the statement that the most widely used machine learning methods are ANNs, SVMs, and RFs. It is hard to convince people that these three algorithms are

the most widely used. For specific problems, some algorithms may be more common than the other algorithms. You may say that the most widely used machine learning includes ANNs, SVMs, and RFs.

AC: We agree with the reviewer that this is probably problem-dependent and should not be generalized here.

CM: We replaced "are" by "include". (now line 26)

L31: You mentioned that RF has been repeatedly found superior to other machine learning methods. You need to specify the particular problems that RF has been found superior to "other machine learning methods" in the text. And what other machine learning methods do you mean here? Please specify in the text.

AC: We have given more details on the particular studies and algorithms tested.

CM: We added 2 sentences after the first sentence of this paragraph (now p.3 lines 1-5): "For example, Li et al. (2011) tested 23 machine learning algorithms - including RF, SVM, and kriging methods - to predict mud content in marine sediments, and found that RF, along with RF combined with ordinary kriging or inverse distance squared, provided the best prediction results. Cracknell and Reading (2014) applied five machine learning methods to lithology classification of multispectral satellite data and reported higher classification accuracy for RF than for Naive Bayes, SVM, ANN, and k-Nearest Neighbors."

2 Methods

Page 3, section 2.1.2:

L25: How do you come up with these 38 predictors? Could you specify the reason why you choose these 38 predictors in this section?

AC: We agree that further information would be helpful in this section.

CM: We shortened the first sentence to "A total of 38 geological and spatial variables

were included as predictors (Table 1)." and added the following passage: "These predictors were parameters that were assumed to influence p-wave velocity. However, only predictors that could be obtained for each of the 333 borehole locations were used. Predictors such as latitude (lat), longitude (long), and water depth (wdepth) were taken from the borehole's metadata, whereas other predictors were extracted from freely available global datasets and grids (Table 1). In addition, predictors describing the borehole's geological setting were determined from the site descriptions given in the proceedings of each drilling campaign. Some parameters known to influence seismic velocity - e.g. porosity, density, or pressure - had to be left out as suitable datasets were not available. Although some of these parameters had been measured in DSDP, ODP and IODP boreholes, they had not necessarily been logged at the same locations at which vp data had been measured, and therefore could not be obtained at all of the 333 boreholes used."

Page 4, section 2.2:

L14: How do you define "performance"? I saw you mentioned performance in the later section 2.3. But it is better to define that when you first mention that.

AC: We agree that it is not always clear what we mean by the term "performance", and we also used it inconsistently to refer to the standard error metrics, the proportion of well predicted boreholes, or both - this is obviously confusing. By performance, we mean both the error metrics and the proportion of well predicted boreholes. We now explain this in paragraph 3 of section 2.3 (p. 5 lines 20-25) and removed/replaced the term in the previous paragraphs.

CM: Paragraph 3 was changed to read "Performance of the RF model was evaluated in two ways: (1) by standard error metrics and (2) by the proportion of boreholes with predicted vp(z) superior to that of empirical functions. The standard error metrics root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R2) were calculated based on the comparison of the predicted and true vp(z)

curves for each borehole in the test fold. RMSE, MAE and R2 of all test folds were then averaged to give final performance values." Throughout the manuscript, we replaced the term "performance" where necessary.

In addition, why do you choose 1000 trees? what is the maximum depth of each tree? How does the number of trees and depth affect the bias and variance of the prediction?

AC: Our study of relevant literature showed that most studies used either 500 or 1000 trees. In an early version of our prediction model, we ran RF repeatedly for numbers of trees between 2 and 1500 and evaluated model performance based on the OOB (out-of-bag) score. As the performance still varied after 500 trees but stabilized around 1000 trees, we chose 1000 trees. We did not repeat this procedure with our final prediction model (which no longer used the OOB approach), so it is possible that a lower number of trees might already have been sufficient. In that case, however, a higher number would not have decreased model performance. The depth of the trees was not defined and therefore not varied in the final prediction model. Early tests showed that performance was generally worse when maximum tree depths were specified (e.g. for max_depth = 5).

CM: no changes made in the manuscript

3 Results

Page 6, section 3.1:

The performance of an algorithm should be shown by both bias and variance. I only see the comparison of errors and percentage of boreholes with scores 2 and 3 in Fig. 3 and 4. How does the number of predictors and data smoothing affect the variance of the prediction?

AC: There is no strict rule that algorithm performance should always be evaluated by bias and variance. Many studies applying machine learning methods use other means to validate their results. We chose to evaluate performance by MAE, RMSE

and R^2, which have been used as performance measures by several other studies that predicted environmental parameters (e.g. Gasch et al. (2015), Ließ et al. (2016), Meyer et al. (2015, 2016)). Our own borehole percentage value serves as an additional measure. We think that our performance evaluation approach is now well described in section 2.3. The effects of varying numbers of predictors and data smoothing in terms of prediction performance are already described in the text.

CM: see above for changes regarding clarification of prediction performance

Since you only have 333 boreholes, 2% change due to different model runs only change scores of 7 boreholes. I am curious about the location distributions of those boreholes which changed their scores, and why their scores changed by changing the number of predictors or data smoothing.

AC: Unfortunately, our applied prediction method does not allow determining which boreholes changed their scores across different model runs. We agree that this would be an interesting aspect to look into, but in this case our model cannot easily be adapted accordingly, so this would likely require setting up a completely new model. This is beyond the scope of these revisions.

CM: no changes made in the manuscript

4 Discussion

Page 10:

L1-5: You made a strong statement about performance of RF. As I suggested in your introduction section, the performance of a machine learning algorithm really depends on situations.

AC: We agree that this can also be misunderstood to mean that RF is always the perfect choice, which is of course not the case. What we actually meant to say was that due to the issues with our dataset (spatially inhomogeneous, varying depth ranges, etc), it is much more likely that the cases of poor performance are due to the dataset itself, and

not due to the choice of machine learning algorithm.

CM: We rewrote the last sentence (now lines 31-33) to clarify this: "However, given the present dataset and its spatial inhomogeneity, we doubt that a different algorithm would lead to a significantly improved prediction performance for vp."

5 Conclusion

Page 10:

L15: RF is hard to extrapolate to data outside the range they have been seen. I doubted that RF can be used for geophysical modeling in areas lacking v_p(z) from boreholes or seismic data.

AC: This is why we recommend more data to be added - to increase the data ranges within the RF model and the likelihood that when the RF model is applied to new data, these data are within the ranges known to RF. We agree that at present, this is not always the case, which likely explains some of the lower-performing locations. However, we also point out that our RF model is not meant as a replacement for other sources of vp(z) data. It is meant only as an aid when no other means are available. We do not expect RF to ever replace or be superior (or even very close) to actual vp measurements or vp from seismic data (nor do we claim this in the manuscript). Our approach is only meant to provide an alternative to using an (unrealistic) constant velocity or empirically-derived vp(z) profiles, which are, as we show, often of lower quality than our predicted vp(z) profiles.

CM: no changes made in the manuscript

Interactive comment on Solid Earth Discuss., https://doi.org/10.5194/se-2019-58, 2019.