Ocean Science Discussions

Ocean Sci. Discuss., author comment AC2
https://doi.org/10.5194/os-2021-83-AC2, 2021

Reply on RC2

Mikael L. A. Kaandorp et al.

Author comment on "Using machine learning and beach cleanup data to explain litter quantities along the Dutch North Sea coast" by Mikael L. A. Kaandorp et al., Ocean Sci. Discuss., https://doi.org/10.5194/os-2021-83-AC2, 2021

The authors present an interesting research article to identify features that favor marine litter deposition. The article is well laid out, and the subject of this article is very relevant to today's society. The outcomes of the article could be used to further optimize the organization of beach cleanups.

The authors present a well laid out overview of the literature and lead well into the research question.

We would like to thank the reviewer for these very encouraging comments.

In section 3.2.1., the authors describe, that they create a large set of combinations for the explanatory variables. Later in 3.2.2., features that correlate are then assigned to clusters to reduce the dimensionality of the data. I was wondering, if this could have been solved by creating a medium sized set of combinations in the first place, in order to make the clustering part obsolete?

We chose this approach, since we didn't know a priori which set of features would be the most relevant. We didn't know a priori what kind of lead time and what kind of radius of influence is the most dominant in causing litter to beach: does the amount of litter that is found on the beach mainly depend on longer time scales, which might be slowly leading to an increasing standing stock of litter, or is most of the litter arriving on the beaches in the previous couple of days? Since readers who would want to implement similar approaches would also face the same questions, we find it informative to keep the full analysis in the manuscript. We have added some clarification to this reasoning in the manuscript (line 270 track changes):

Some features correlate as these are, for example, derived from the same variable, but for a different radius or lead time. However, we do not know a priori which of these radii and lead times are the most appropriate predictors for the beached litter quantities. For example, litter concentrations might be influenced by long-term processes, slowly increasing the standing stock of litter on the beach, or the concentrations could be better predicted by conditions on the day leading up to the cleanup stage. Since we do not know this, we let the algorithm select the most appropriate variables.

In section 3.2.2. it is not quite clear to me, if the test data is at some point used as training data during the Nested 5-fold cross validation training process. The

authors should describe this in a little more detail.

The k-fold cross validation procedure is also illustrated in Figure D1. One divides the available dataset into a part for training and a part for testing. The test data is never used for training within the outer loop. This means that the model performance is not overpredicted. In fact, it may be slightly underpredicted because the model becomes more accurate as more data becomes available, so one could even say that the real performance of the model could be slightly higher. This procedure is quite standard, we've added a link to the book by Hastie and Tibshirani in which this method is explained more clearly for the interested reader. We have added a clarification to the manuscript (line 280 track changes):

In the inner loop, 80% of the training data (i.e. 64% of the total data) is used to train the model, and 20% of the training data (i.e. 16% of the total data) is used to calculate the importance of the features, also repeated 5 times. Since in the inner loop none of the test data are used to train the model, we do not overpredict the model performance (Hastie et al., 2008)

Could the authors explain more in detail why they chose Random Forest as regression algorithm compared with other regression algorithms?

We have added more clarification to the text (line 259 track changes):

This model allows us to capture non-linear relations between the features and response. It is a non-parametric model, and does not require prior knowledge on the model structure. These are both important reasons to choose the specific algorithm: coastal processes affecting dispersion of marine litter are highly complex (Sebille et al., 2020), so we do not know a priori how the different environmental variables might interact, and how non-linear these interactions might be. The random forest regression model can aid in scientific knowledge discovery (Bortnik and Camporeale, 2021): it gives us Gini importances for all features (Nembrini et al., 2018). This is another reason for choosing this specific algorithm, as it provides us information what processes are important for predicting beached litter concentrations.

In section 4.3., it was not crystal clear to me, why a new regression model with the Top 8 features was trained compared with using the model that was already trained. Could the authors explain the reasons behind this?

We have added clarification for the reasoning in the manuscript (line 421 track changes):

We choose to use a model trained using the top 8 features only for the extrapolations, as increasing the amount of features does not increase the predictive performance (see Figure B5). Furthermore, reducing the amount of features simplifies the computations, as we do not need to compute all 391 variables again for the entire Dutch coastline.

In section 5., the authors give an extrapolation of how much litter is located on the Dutch North Sea coastlines. Can these extrapolations reliably be extrapolated to the whole coastlines of Europe etc.?

No, the model is trained using conditions in the Netherlands, and should therefore not be used to make extrapolations for other regions. Machine learning models generally perform poorly when making extrapolations for conditions not seen before in the training data. Conditions might be very different for other regions (e.g. much stronger currents, more complex coastline geometries, different coastline substrates). We have added a clarification regarding this to the manuscript (line 484 track changes):

The model itself can not directly be used for other geographic regions, since the features used to train the algorithm are specific to the region of interest. The model is likely to perform poorly when making extrapolations for conditions not present in the training data. As an example, the substrate of beaches is likely to have a large impact on litter concentrations (Hardesty et al., 2017), which are relatively uniform in this analysis (all sandy beaches).

I was missing some context of the extrapolations (16T – 31T), i.e., how much waste collect cleanup missions during a whole year. Could the authors include this information?

This is a good point. We have added a table to the supplementary information, where one can see that the total weight gathered over the various years varies from 9,872 to 20,078 kilograms. See table A1

We have added a small discussion on these numbers to the manuscript (line 357 track changes):

The total amount of litter gathered during the cleanup campaigns, and the total amount of kilometers sampled per year is presented in Table A1. The total amount of litter gathered varies from 9,872 to 20,078 kilograms. This is in line with the expected total amount of litter predicted by the model, since the majority of the coastline (222 to 262 kilometers out of 365 kilometers) was cleaned up during the cleanup campaigns.