

Ocean Sci. Discuss., referee comment RC2  
<https://doi.org/10.5194/os-2021-83-RC2>, 2021  
© Author(s) 2021. This work is distributed under  
the Creative Commons Attribution 4.0 License.

## **Comment on os-2021-83**

Anonymous Referee #2

---

Referee 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-RC2>, 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.

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?

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.

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

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?

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.?

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?