Geophysical Research Abstracts
Vol. 20, EGU2018-9760, 2018
EGU General Assembly 2018

Practical quantification of uncertainty in seabed property prediction using geospatial KNN machine learning

Warren Wood (1), Taylor Lee (1), and Jeffrey Obelcz (2)

(1) U. S. Naval Research Laboratory, Geology and Geophysics, United States, (2) NRC Post Doc, U. S. Naval Research Laboratory, Geology and Geophysics, United States

The method of K-Nearest Neighbor (KNN) is arguably the most data-driven form of machine learning regression (i.e. there is no pre-supposed model form). The entire set of observations serves as the model. As such, there is no explicit training step, and the operational parameters that one may refine for any given prediction are minimal. Where little is known about the relationship between predictors and desired quantity (predictand), this form of machine learning has advantages over more formal probabilistic forms of machine learning (e.g. Bayesian methods) which require prior assumptions about the probability distribution of the model and data. However, the lack of formal model structure leaves the KNN prediction estimate without a formal probabilistic uncertainty. Because estimates without some form of uncertainty (either stated or implied) are effectively useless, and because the KNN technique is inherently non-formal, we require a practical approach to address the issue of KNN uncertainty.

We present here possible proxies for uncertainty (posterior probability distribution function) as applied to a geospatial KNN prediction of seafloor properties. Our approach is general, but we focus on seafloor porosity and total organic carbon. Our analyses were performed first on synthetic data so that the true values and errors could be more extensively compared to the to the value and error estimates resulting from the prediction. Specifically, as part of the multi-fold validation process, we compare at each observation the average value of the K-nearest neighbors (which is the estimate of the predictand) to the true value. We used small values for K, ranging from K=3 to k=9. We also compute statistics of both the values and distances in parameter space of these K-nearest neighbors. Using several numerical simulations, we have found that although the process is somewhat problem-dependent, the standard deviation in the values of the K-nearest neighbors comes the closest to approximating the true error in prediction. When applied to actual marine observations, the standard deviation in nearest neighbor values also came the closest to approximating the error.