# Comparing Machine Learning Methods for Predicting Seismic P-wave Velocity on Global Scale

Yousef Razeghi[1,2] [0000−0002−0007−630X]
Wilhelm Hasselbring[2] [0000−0001−6625−4335]
Christian Berndt[1] [0000−0001−5055−0180]
Ines Dumke[1] [0000−0003−0323−8578]

[1] GEOMAR, Helmholtz Centre for Ocean Research Kiel, Germany
[2] Software Engineering Group, Kiel University, Germany

**Abstract.** A huge amount of ocean observation data is available. A purpose of interpreting that data in marine–geophysical applications is to find, for instance, anomalies which are the signs of reservoirs in earth layers beneath the ocean floor. In this position paper, we compare different machine learning methods to predict the overall trend of seismic P-wave velocity as a function of depth for any marine location. Our study is based on a dataset consisting of data from 333 boreholes and 38 geological and spatial predictors. Our preliminary results indicate that random forests provide best results on this dataset, but also suggest to apply data augmentation for improved results with other methods.

**Keywords:** ocean observation · machine learning · anomaly detection.

## 1 Introduction

Following the recent advancements in sensor technology and increasing use of various sensors in different environments, huge amounts of ocean observation data are available today. Machine learning excelled in many fields (e.g. market prediction, cancer diagnosis, object recognition, and etc.) and promises more accurate results in comparison to methods which were used in past. Similar to other fields of science, the field of marine sciences is challenged with big data, which is gathered through (expensive) expeditions and observations via numerous different sensors.

The amount and diversity in the form of data requires specific methods such as machine learning models to process and analyze the data. The goal of this position paper is to compare machine learning methods on the data related to geophysical and geological properties of rocks beneath the sea floor. We aim to make predictions on average properties of these sub-surface materials. Our predictions will help to detect the anomalies in data which are the signs of reservoirs or effects of natural phenomena on earth layers beneath the ocean.

## 2 Prediction of Seismic P-wave Velocities

P-wave velocity is one of the physical properties of sub-surface rocks which can help to predict the materials beneath the ocean floor. Based on measurements done during different drilling projects Dumke et al. [5] gathered the information of 333 borehole logs to form a dataset which is depicted in Figure 1. They have compared the results of prediction using Random Decision Forests (RDF) [2] with hamiltonian functions [6], which were used in the past as a conventional method to compute the average P-wave velocity.

### 2.1 Characteristics of the Dataset

Our data have been gathered during different drilling campaigns performed on a global scale, see Figure 1. The dataset is divided into 10 folds to perform cross validation. The 10 folds are separated from each other based on geographic location. This is an important feature of the dataset to prevent overfitting. Location-wise separation implies that the samples which have been used in training sessions belong to different locations than the samples used for the prediction.

We compared the results of our predictions on this dataset. Some key characteristics of this dataset are that,

1. the number of categories for independent variables which are usually referred to as features is large,
2. the amount data with respect to the number of categories on each sample is scarce for each borehole,
3. the trend in data differs in each correlation of data pairs, and
4. there are many outliers and noise in the dataset.

Figure 2 depicts the correlation of data points in a training set between 4 arbitrary data categories. We chose longitude, latitude, water depth, and sediment

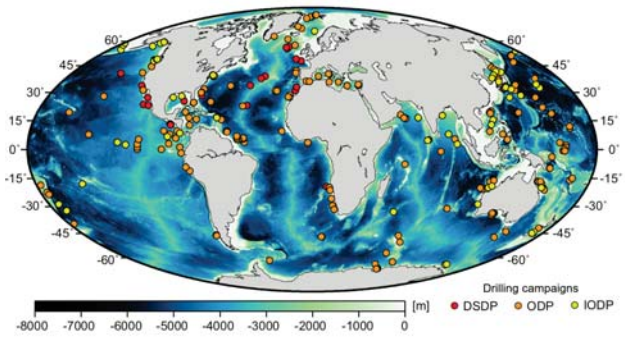


Fig. 1: Distribution of the 333 boreholes [5]. DSDP: Deep Sea Drilling Project, ODP: Ocean Drilling Program, IODP: International Ocean Discovery Program. The Bathymetry is from the GEBCO_2014 grid (http://www.gebco.net)

thickness in this pair plot. Many outliers can be detected. In addition, in each pair of the variables there is clear change in the trend of data distribution.

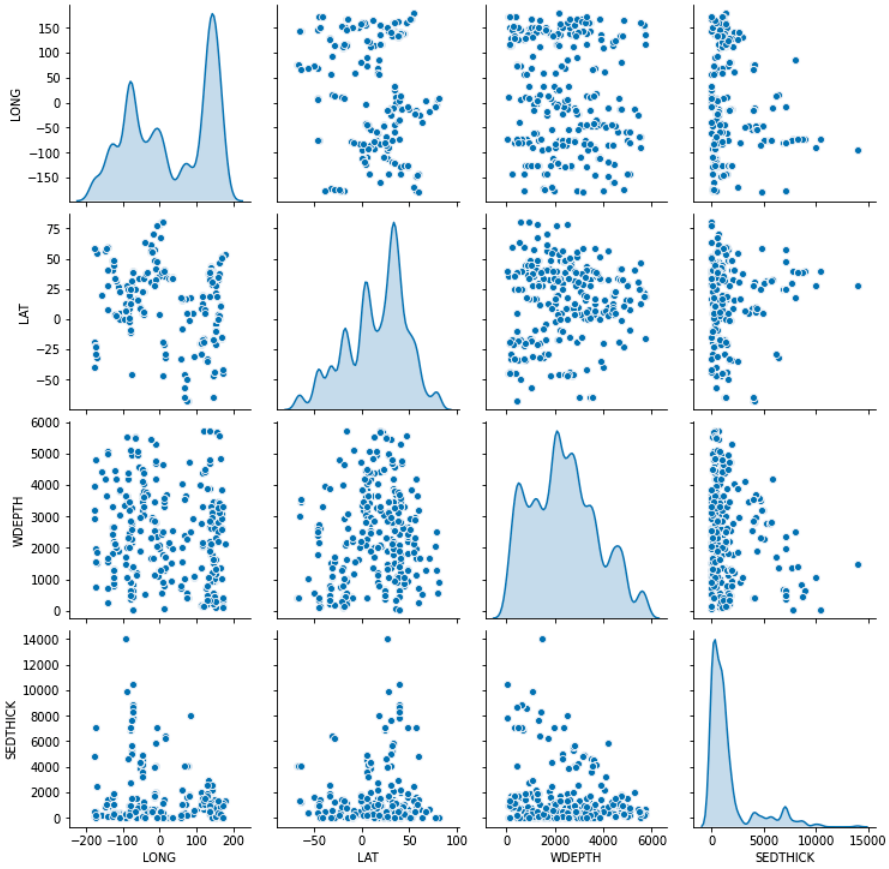


Fig. 2: Dataset pair plot

### 2.2 Investigated Machine Learning Methods

To achieve improved results, we have done prediction on seismic p-wave velocity using different machine learning methods to assess the changes in prediction accuracy. We have used scikit-learn [9], and keras [1] to implement 3 different machine learning methods:

– Support vector regression (SVR): The kernel used in SVR [4] is a radial basis function [11] of degree 8.

- Polynomial regression: In polynomial regression, the degree of the polynomial is set to 2.
- Neural networks (NN): We applied a 2-layer feed forward neural network, each layer consists of 32 neurons. As activation function we used the relu function, which is a widely used activation function for neural networks. In each training iteration, a batch of 50 data samples goes through the model. The training session contained 400 epochs and the learning rate was set to 0.001. We used the RMSProp [3] optimizer for our model.

Dumke et al [5] used a Random Forest Regressor using scikit-learn [9]. They used 1000 decision trees in their study. The number of the predictor variables was 38. Based on their study, using the most important 16 features results in better predictions.

### 2.3 Prediction Results

We have selected the most important 16 features of the dataset as done before [5]. The importance of the features are calculated using RDF. Feature importance can be defined as a score assigned to each category of independent variables with respect to their influence on the performance of the prediction model. Thus, the performance of the prediction is divided into 4 categories based on the previous work by Dumke et al [5]. Each category specifies the level of performance compared to conventional methods (i.e., Hamiltonian curves [6]) that is used for calculating the seismic P-wave velocity. Scores assigned to each prediction are based on comparison of the error metrics RMSE (Root Mean Square Error), Mean Absolute Error, and $R^2$ error metrics. In case of each prediction, the error metrics of the machine learning algorithm is compared to the error metrics of the conventional Hamiltonian functions [6]. Figure 3 illustrates the difference in performance of the machine-learning methods in 4 score categories:

**3:** All 3 error metrics indicate a better fit than the Hamiltonian functions
**2:** 2 of 3 error metrics indicate a better fit than the Hamiltonian functions
**1:** 1 of 3 error metrics indicate a better fit than the Hamiltonian functions
**0:** No error metrics indicate a better fit than the Hamiltonian functions

Categories 0 and 1 are representing the bad predictions, while 2 and 3 represent the good predictions. The RDF method could drop the number of bad predictions and increase good predictions to almost the half comparing to SVR, polynomial regression, and neural network. These data characteristics (see Section 2.1) makes fitting a regression model using SVR, polynomial and neural network a challenge.

## 3 Summary and Future Work

In the field of ocean observation, the data can be in various forms (e.g. borehole logs, seismic data, images, and etc.). All different data types refer to geographical

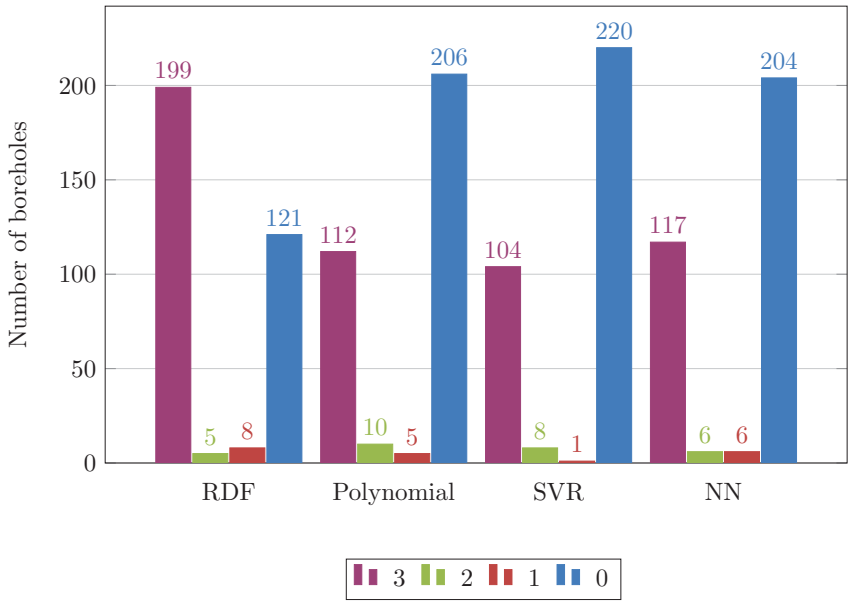


Fig. 3: The number of predictions in each score category (0 – 3) for each machine learning method RDF, Polynomial, SVR and NN.

locations. To have a better understanding of materials in specific locations and interpret the data, integration of different data types is necessary. Our aim is to find appropriate methods to integrate and adapt different forms and types of data to form datasets, and increase the precision and accuracy of the predictions to enhance our understanding.

Another challenge is to select the best predictor variables to get the highest accuracy and reduce the dimensionality of data. Previously, ensemble methods like Random Decision Forest (RDF) [2] is used in several studies. The ability of RDF to deal with unscaled and high-dimensional data makes it popular among the scientific community dealing with ocean observation data.

The major challenge is to make accurate predictions on global scale. This means to predict the properties of materials in unobserved locations. Global scale predictions would reduce the cost in terms of energy and time. Precise predictions would reduce the need for explorations and expensive expeditions.

Data augmentation [12] is a usual performance enhancement method for machine learning models which is used for different data types. Considering the essence of some data types like images, modifications on existing data points (e.g. performing transformations, adding noise, manipulating color tones, etc.) can result in having more new data points. To increase the accuracy in our future

attempts we are up to investigate methods for data augmentation in our field of marine data science.

For reproducibility and reusability, we publish our software open source [10] and the data together with the analytics service OceanTEA [7], as we did in the past [8].

## Acknowledgment

The first author is funded through the Helmholtz School for Marine Data Science (MarDATA, *https* : *//www.mardata.de/*), Grant No. HIDSS-0005.

## References

1. François Chollet et al. Keras: The Python deep learning library. *ascl*, pages ascl–1806, 2018.
2. Antonio Criminisi, Jamie Shotton, and Ender Konukoglu. Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning. *Microsoft Research Cambridge, Tech. Rep. MSRTR-2011-114*, 5(6):12, 2011.
3. Yann N. Dauphin, Harm de Vries, and Yoshua Bengio. RMSProp and equilibrated adaptive learning rates for non-convex optimization. *arXiv abs/1502.04390*, 2015.
4. Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. Support vector regression machines. In *Advances in neural information processing systems*, pages 155–161, 1997.
5. Ines Dumke and Christian Berndt. Prediction of seismic P-wave velocity using machine learning. *Solid Earth*, 10(6), 2019. doi:10.5194/se-2019-58.
6. Edwin L Hamilton. Elastic properties of marine sediments. *Journal of geophysical research*, 76(2):579–604, 1971.
7. Arne Johanson, Sascha Flögel, Christian Dullo, and Wilhelm Hasselbring. OceanTEA: Exploring ocean-derived climate data using microservices. In *Proceedings of the Sixth International Workshop on Climate Informatics (CI 2016)*, pages 25–28, 2016. NCAR Technical Note NCAR/TN-529+PROC. doi:10.5065/D6K072N6.
8. Arne Johanson, Sascha Flögel, Christian Dullo, Peter Linke, and Wilhelm Hasselbring. Modeling polyp activity of Paragorgia arborea using supervised learning. *Ecological Informatics*, 39:109–118, 2017. doi:10.1016/j.ecoinf.2017.02.007.
9. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.
10. Yousef Razeghi, Wilhelm Hasselbring, Christian Berndt, and Ines Dumke. Data and software for: Comparing machine learning methods for predicting seismic p-wave velocity on global scale. doi:10.24433/CO.3341619.v1.
11. Bernhard Scholkopf, Kah-Kay Sung, Christopher JC Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765, 1997.
12. Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, December 2019. doi:10.1186/s40537-019-0197-0.