Predicting suspended sediment particle size distributions using machine learning

Galen Egan
gegan@stanford.edu

Background & Motivation

Marine sediment (i.e., mud) particles cover approximately 70% of the Earth's surface, and play a critical role in a host of environmental engineering problems. Settling rates of marine particles control global carbon sequestration rates, and coastline erosion predictions rely largely on knowledge of sediment properties. Global warming and associated sea level rise have added new urgency to accurately predicting these processes, a task which is generally attempted with numerical sediment transport models. One of the most critical parameters in these models is the mean particle size, which is generally unknown. As of yet, there is no general model (i.e., a model not tuned to specific field observations) available which can predict suspended sediment particle size with sufficient accuracy.

Dataset

- 1648 particle size distributions (PSDs, e.g. figure 1) measured in South San Francisco Bay during three one-month-long field deployments
- Co-located (in time and space) measurements of water salinity, temperature, biological properties, wave velocities, tidal velocities
- **Objective: predict d_{50} and σ^2 based on local hydrodynamics, water chemistry, and biology**
  - These features are much cheaper to measure/model

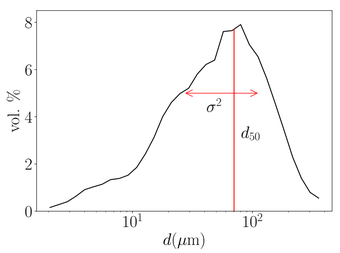


Figure 1: Example PSD with median particle diameter, d_{50} and variance, σ^2 denoted by red lines.

Methods

- Random forest (RF) for feature selection
- Tune and test on RF and support vector regression (SVR) models
- RF implemented with `sklearn.ensemble.RandomForestRegressor`
- SVR implemented with `sklearn.svm.SVR`

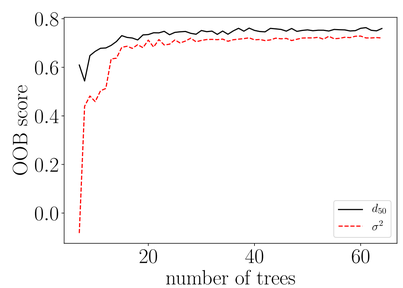
Feature Selection

- Feature Importance (FI) calculated by default in RF algorithm
- Mean of 1000 RF runs predicting σ^2 and d_{50} w/ 10 estimators, 70/30 training/test split
- Top 4 features selected
  - Salinity, S
  - Wave velocity, u_b
  - Particle index of refraction, n_p
  - Temperature, T

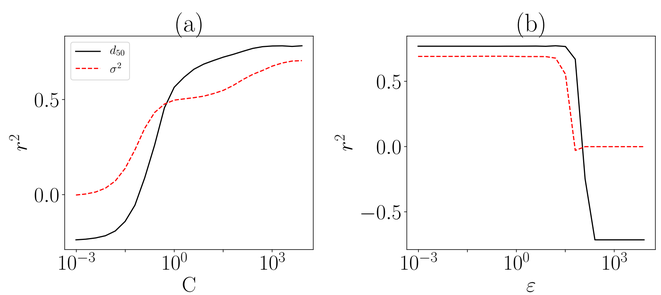
	FI_σ	FI_d
S	0.555	0.722
u_b	0.153	0.058
n_p	0.076	0.051
T	0.049	0.051
a_{676}/a_{650}	0.051	0.037
chl-a	0.050	0.031
a_{450}/a_{676}	0.037	0.025
$\overline{u}$	0.030	0.026

Tuning

- RF tuned to optimal number of trees using out-of-bag (OOB) score

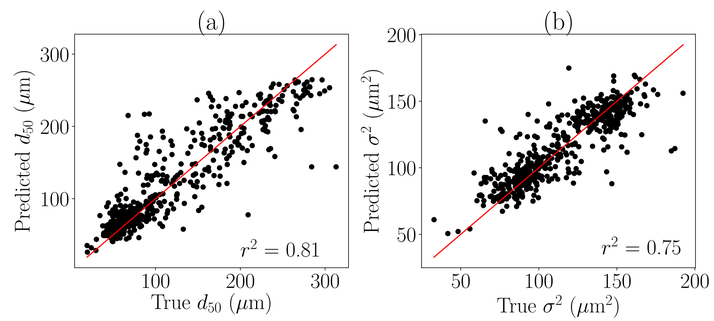


- SVR tuned for C (regularization) and ε (slack variable)
- Optimized for r^2 on 15% cross-validation set

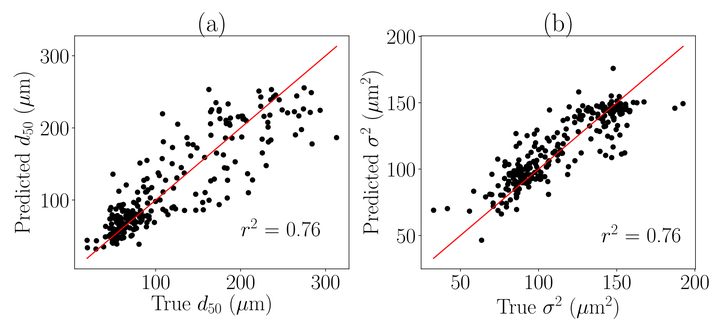


Results on Test Set

- RF results: relatively high accuracy on test set for both d_{50} (panel a) and σ^2 (panel b)



- SVR: tested on smaller (15%) test set
- Results: similar to RF, with impressive prediction on both d_{50} and σ^2.



Conclusions & Future Work

- Both algorithms show promise in predicting d_{50} and σ^2 from features
- Either SVR or RF could be incorporated into San Francisco Bay sediment transport model
- More data needed for testing and training to determine if results are generalizable to other regions

poster video presentation link:

https://stanford.box.com/s/wmouehw8k0de2gxkfyeryibc861clvav