Training Datasets for Modeling with AI across the Deep-Ocean Seafloor

Chris Jenkins, INSTAAR CU Boulder
PI for ‘dbSEABED’
Machine Learning AI&ML Group in CSDMS 2020

(Chris Jenkins, Daniel Buscombe, Evan Goldstein, Kelly Kochanski, Jeffrey Obelcz)

- Decided: best strategy to help CSDMS members to pick up & use ML is for CSDMS to provide training data

- At the same time it’s been observed that the deep-sea tends to be neglected as an activity in CSDMS
Background

• dbSEABED Project objective – maps and database for the **materials of the seabed**, everywhere maximizing data and optimizing the math / algorithm methods; resolve facies on 10km scales worldwide

• Research **Heterogeneous Data methods** – leveraging the massive amounts of data that have been collected by oceanographers over the decades; difficult, but with many advantages

• **Applied to research and ocean management** – fisheries, biogeochemistry, habitats, contamination monitoring, safe navigation, marine conservation, sonar prediction, mine-countermeasures, stratigraphy, deposition/erosion, seafloor stability, paleoceanography; over 100 cases

• A persistent problem: How to **create the best gridded data from sparse point data**? Avoid pitfalls! Realistic spatial predictions. THAT is why we look to MACHINE LEARNING
Overview of ML in our field

• “learn from already labeled data how to predict the class of unlabeled data”; “machine memory, not machine learning”

• Informative example papers (see refs) ...
  • Dutkiewicz Müller O’Callaghan & Jónasson 2015. Census of seafloor sediments in the world’s ocean
  • Restreppo Wood Phrampus 2020. Oceanic sediment accumulation rates predicted via machine learning algorithm: towards sediment characterization on a global scale

• The range of Al methods includes Random Forest, Neural Network, Support Vector Machine, K-Nearest Neighbors; SciKit-Learn is our starting package

• Training data – absolutely critical, and difficult to compile on large scales

• Important separations: Supervised and Unsupervised, Regression and Classification
**Primer on Training Data**

- Column-wise data, ‘input vectors’, with headers
- ‘Features’ – each parameter, attribute
- ‘Labelled Data’, ‘Target Values’
- ‘Labels’ – the desired output values on the training data
- ‘Target’ the attribute to predict
- ‘Dimensionality’ – here 4

<table>
<thead>
<tr>
<th>Sample index</th>
<th>sepal length</th>
<th>sepal width</th>
<th>petal length</th>
<th>petal width</th>
<th>target</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>1</td>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>2</td>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>3</td>
<td>4.6</td>
<td>3.1</td>
<td>1.5</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
<tr>
<td>4</td>
<td>5.0</td>
<td>3.6</td>
<td>1.4</td>
<td>0.2</td>
<td>Iris-setosa</td>
</tr>
</tbody>
</table>
Common Training Data issues

- Missing attribute values: fix by imputation, completing (e.g., with feature medians), or deletion
- Standardizing – usually centered on MEAN and scaled by the STDEV
- Not too many features (Random Forest)
- Imbalanced data (fix by down- or up-sampling, or reduction)
- Incomplete parameter values coverage
- Feature selection / reduction
- Overfitting

- The modeler is not a specialist on the data (e.g., geologist using physical oceanography data)
Specifically in dbSEABED ...

- Apply to Test data (Seafloor Mapping, Research Result)
- Performance / Validation Tools
- Choice (or Suite) of AI Algorithm/s
  = Scikit-Learn
- Training Data (organize, select labelled samples, feature selection, ...)
- Outside data bases (eg dbSEABED, WOA, MODIS)

The processing flow for ML-Mapping of Seafloor Properties

- Software package “Contributed Model”
- Digital stack of the Terrain, Geo and Enviro data grids, sampled by the Training and Test data; skewering at Labelled data locations
Specifically in dbSEABED (cont)

- Assemble griddings of the Training parameters
- Normalize and step the grid-values
- Create uniques-classified map area of the grid-stack; assign an index to each unique area, and associate the training parameter values
- Assemble labeled Training samples with Target-parameter values and locations; pare down to a set of uniques with no blanks
- Attach Training parameter values from the stack to the Labelled locations using their locations
- Run the ML algorithm/s, collect performance metrics
- Transfer the outputs to
- Assess the outputs; adjust and re-run
Correlation matrix

Here, ‘Token’ = Organic Carbon content

Pearson correlation coefficient $R$, $p$
$p \leq 0.05$: \( p_{stars} = **\)
$p \leq 0.01$: \( p_{stars} = ***\)
$p \leq 0.001$: \( p_{stars} = ****\)

This data release from dbSEABED

**Gulf of Mexico** 0.02deg (~2km), but also **Global**
(0.25 deg to match World Ocean Atlas)

- **Terrain** – bathymetry, slope, aspect, geomorphic provinces, hessian largest eigenvalue (ridged-ness)
- **Geologic*** – gravel content (% wt), sand content (% wt), mud content (% wt), clay content (% wt), central grainsize (phi), sorting (phi), rock presence (%), color (rgb), red/green (/1), grain component memberships (%), feature memberships (%)
- **Compositional*** – carbonate content (% wt, dry), organic carbon content (%wt, dry),
- **Geophysical*** – porosity (%), critical shear stress (kPa), shear strength (kPa), sound velocity (m/s)
- **Environment** – bottom & surface water temperatures, turbidity, surface chlorophyll-a, bottom dissolved oxygen (μmol/kg), bottom oxygen lows (μmol/kg), ...
This data release from dbSEABED (cont)

Seabed textures
This data release from dbSEABED (cont)

Compositions & Color
Figure 43. Dead bivalves at the seafloor after an oxygen starvation event, still with soft tissues visible. From: Norkko & others (2013). Baltic Sea; width of view is about one metre.

Source: World Ocean Atlas 2018
1dg averages, statistical mean & stddvn

https://www.ncei.noaa.gov/access/world-ocean-atlas-2018/bin/woa18oxnu.pl
Temperature (Surface & Bottom)

Turbidity, productivity

Source: MODIS Aqua

1dg averages, time averaged, visible spectral proxies

https://modis.gsfc.nasa.gov/data/dataprod/kd_490.php
Current flows

https://www.hycom.org/dataserver
(short sample of)
For the clinic:  

What seabed applications do YOU and your institution want / do you imagine for ML? 

• Habitat Suitability models  
• Inter-parameter relations  
• Substrate mappings  
• ... add more (with links?)  

• Get the data from “HERE” and enter it into your GIS (QGIS or ESRI)
HSM – *Macoma balthica*

Habitat suitability model
- standard management tool
- here, physical parameters
- usually a logistic equation

What are the chief correlates for the living *M. balthica*?
- dbSEABED _Enviro layers
- GBIF (OBIS) occurrence data

From FLIKR - https://www.flickr.com/photos/gridarendal/31636259221
Clay/silt ratio
Important for seabed uptake and sequestration of radionucleides, e.g., from the Fukushima releases

Method:
• Collect all conceivable parameter inputs, with a rationale for their use
• Collect all the silt/clay or clay/mud data from dbSEABED, separately for analysis and description data
• Associate the silt/clay & parameter locations and water depths.
• Test the correlations
Wrap-up

• What do researchers want in global gridded data-layers?
• What are the questions they want to answer?
• How will we obtain/build those layers?
• What exactly is the parameter / statistic that we want in each case?
• How important is having a process-model rationale for each parameter?
Wrap-up (cont)

dbSEABED Mud contents (%)

HYCOM Bottom Currents (Vs, m/s)


Extra for questions ...
<table>
<thead>
<tr>
<th>Target Feature</th>
<th>Predictive Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbonate</td>
<td>lgwd, prox, btemp, stemp, kd490, chlora</td>
</tr>
<tr>
<td>Rock exposure</td>
<td>lgwd, slp, hsn</td>
</tr>
<tr>
<td>Mud content</td>
<td>lgwd, slp, kd490, chlora</td>
</tr>
<tr>
<td>Sand content</td>
<td>lgwd, slp, prx</td>
</tr>
<tr>
<td>Gravel content</td>
<td>lgwd, slp, prox</td>
</tr>
<tr>
<td>Organic carbon</td>
<td>lgwd, btemp, stemp, kd490, chlora, doxMN, doxLO</td>
</tr>
<tr>
<td>Porosity</td>
<td>lgwd, btemp, stemp, kd490</td>
</tr>
</tbody>
</table>