A Machine Learning Approach to Modeling Hydromodification

Ashmita Sengupta

SCCWRP, July 17th
Roadmap for Today’s Presentation

• Background
• Deterministic models: Pros and Cons
• Machine Learning Approaches
• Case Study
Background

• Hydromodification is a serious concern in southern California

• Responses are unpredictable
Explicit Knowledge of Uncertainty
Cost / Time / Data
Ease of Use

Appropriate tool or combinations of tools based on information needs, desired level of certainty, data availability etc.
Mechanistic/Deterministic Models

• Hydrologic: watershed hydrologic processes-runoff, infiltration, and precipitation
  Hydrologic Engineering Centers (HEC) or HSPF based

• Hydraulic: water-surface profiles, shear stresses, shear stresses, stream power values, and hydraulic characteristic
  Hydrologic Engineering Centers-River Analysis System (HEC-RAS)

• Sediment Transport Models: potential change in channel morphology

• Regime Diagrams
Pros and Cons of Deterministic Models

Pros

• Addresses questions of basic condition, susceptibility, etc.
• Relatively rapid and easy to apply
• Answers are generally qualitative or semi-quantitative
• Appropriate for screening-level decisions

Cons

• Difficult to model due to uncertain responses
• Cumulative Error
Non-linear responses

![Graph showing non-linear responses between channel enlargement ratio and percent impervious cover.](image-url)
Modeling Tools

Modeling tools should:

• Represent uncertainty in model structure and parameters and noise in the data
• Be automated and adaptive
• Exhibit robustness
• Scale well to large data sets
The Anatomy of a Machine Learning Problem

INPUT → ALGORITHM → OUTPUT

- CATEGORICAL
  - YES
  - NO
- ORDINAL
  - RANKS
- CONTINUOUS
  - Predict target variable
Machine Learning and Approaches
Probabilistic Graphical Models/Bayesian Networks

A graphical model that encodes probabilistic relationships among variables of interest.

- Model encodes dependencies among variables, accounts for missing data easily
- Learns causal relationships, can be used to gain understanding about a problem domain and to predict the consequences of intervention.
- Model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge (which often comes in causal form) and data.
- Avoids over-fitting of data.
Random Forests/Decision Trees

Random forest method for classification (and regression)

• Create a model that predicts the value of a target variable based on several input variables.
• The interior node corresponds to one of the input variables.
• Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.
Support Vector Machine

A Support Vector Machine (SVM) performs classification by constructing an $N$-dimensional hyperplane that optimally separates the data into two categories.

- SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors is maximized. In the figure above, the line in the right panel is superior to the line in the left panel.
Support Vector Machine

Separation may be easier in higher dimensions

complex in low dimensions  simple in higher dimensions
Logistic Regression

• Logistic Regression is a type of predictive model that does not involve decision trees and is more akin to nonlinear regression such as fitting a polynomial to a set of data values.

• Logistic regression can be used only with two types of target variables:
  
  a. A categorical target variable that has exactly two categories (i.e., a *binary* or *dichotomous* variable).
  
  b. A continuous target variable that has values in the range 0.0 to 1.0 representing probability values or proportions.
Logistic Regression

(a) enlargement vs. erosion potential

(b) risk of enlargement associated with $d_{50}$ and erosion potential

5 based on the withholding of two stream reaches where enlargement was primarily driven by historic channelization (San Antonio) or kept artificially low due to dense vegetation (Agua Hedionda); both factors were poorly distributed in our dataset.
Case Study: General Regression Neural Network (GRNN)

- Series of iteratively solved equations:
  - Adaptive Learning
  - Ability to model nonlinear relationships
  - Identification of variables that most affect uncertainty in model output
  - Ability to use surrogate variables
  - Easier parameter optimization
Case Study: General Regression Neural Networks

- 25 different locations
- 85 transects
Neural Network Setup

Predictor Variables
(Watershed area, stream-power, imperviousness)

Gaussian Distribution (PNN)

Channel Area
For Most Validation Scenarios: Channel Area Predicted Well

\[ R^2 = 0.847 \]
### Comparison with Multivariate

#### Cross Validation: 10%

<table>
<thead>
<tr>
<th>Flow</th>
<th>Neural Networks Calibration</th>
<th>Validation</th>
<th>Multivariate Regression Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>0.99</td>
<td>0.77</td>
<td>0.74</td>
<td>0.61</td>
</tr>
<tr>
<td>Q10</td>
<td>0.99</td>
<td>0.79</td>
<td>0.68</td>
<td>0.53</td>
</tr>
<tr>
<td>Q50</td>
<td>0.99</td>
<td>0.85</td>
<td>0.8</td>
<td>0.37</td>
</tr>
<tr>
<td>Q100</td>
<td>0.99</td>
<td>0.82</td>
<td>0.73</td>
<td>0.57</td>
</tr>
</tbody>
</table>

#### Random Holding: 20%

<table>
<thead>
<tr>
<th>Flow</th>
<th>Neural Networks Calibration</th>
<th>Validation</th>
<th>Multivariate Regression Calibration</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q2</td>
<td>0.99</td>
<td>0.66</td>
<td>0.89</td>
<td>0.1</td>
</tr>
<tr>
<td>Q10</td>
<td>0.96</td>
<td>0.63</td>
<td>0.82</td>
<td>0.26</td>
</tr>
<tr>
<td>Q50</td>
<td>0.99</td>
<td>0.55</td>
<td>0.84</td>
<td>0</td>
</tr>
<tr>
<td>Q100</td>
<td>0.99</td>
<td>0.65</td>
<td>0.82</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Sigmoidal Decrease in Model Performance
Certain Variables Were Consistently Ranked Higher

<table>
<thead>
<tr>
<th>Predictor Variable</th>
<th>Q2</th>
<th>Q10</th>
<th>Q50</th>
<th>Q100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calculated Flow</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>Bedload Capability</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Geotechnical Stability of Cross-section</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Total Impervious Area</td>
<td>4</td>
<td>9</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>Stream Power</td>
<td>6</td>
<td>6</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Bed material</td>
<td>8</td>
<td>7</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Distance to Hardpoint</td>
<td>0</td>
<td>15</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>
Interesting Observations