A Machine Learning Approach to Modeling Hydromodification

Ashmita Sengupta

SCCWRP, July 17th

Roadmap for Today’s Presentation

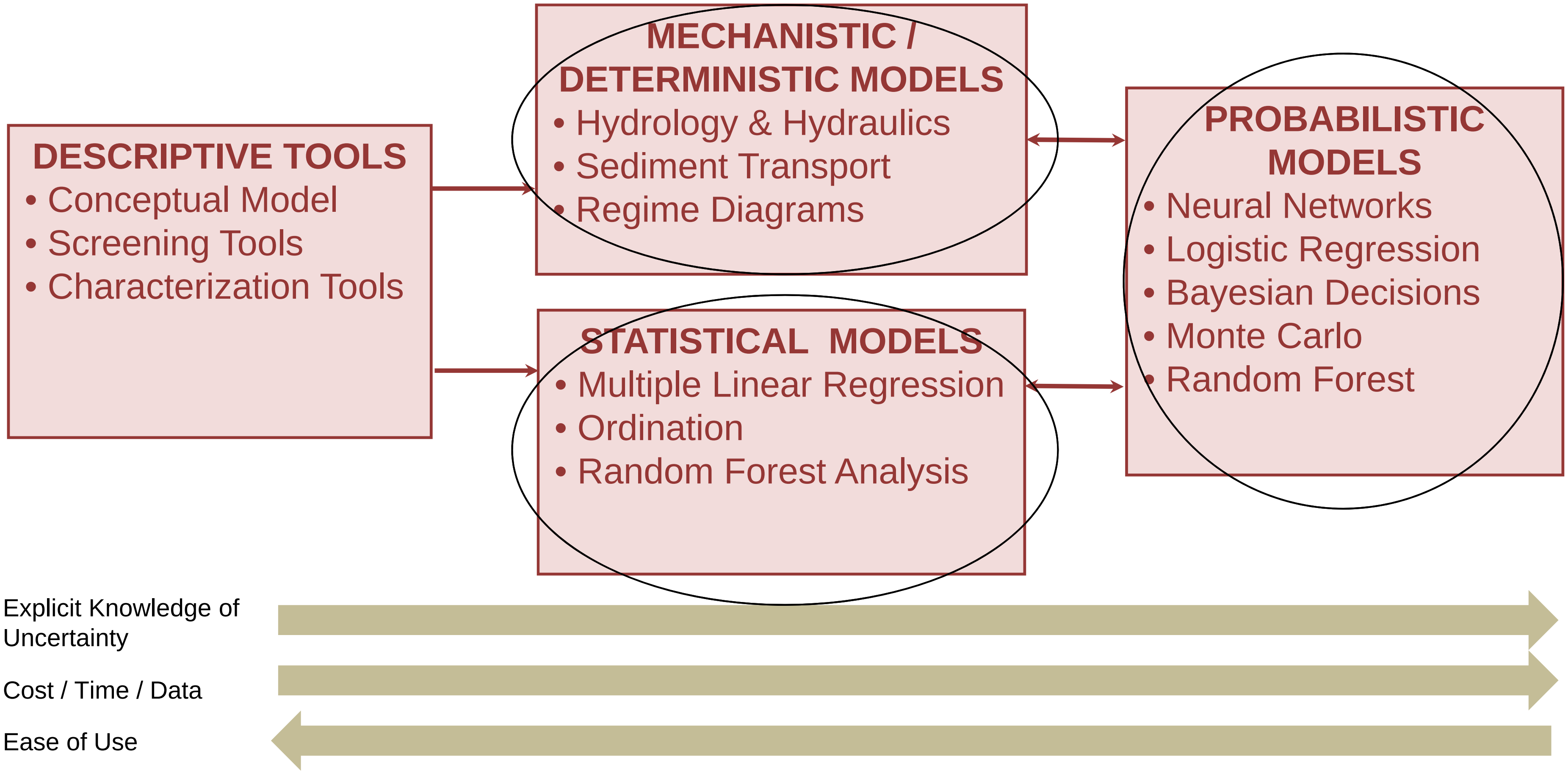
- Background
- Deterministic models: Pros and Cons
- Machine Learning Approaches
- Case Study

Background

- Hydromodification is a serious concern in southern California
- Responses are unpredictable

Hydromodification

Modeling Framework for Hydromodification



DESCRIPTIVE TOOLS

- Conceptual Model
- Screening Tools
- Characterization Tools

MECHANISTIC / DETERMINISTIC MODELS

- Hydrology & Hydraulics
- Sediment Transport
- Regime Diagrams

STATISTICAL MODELS

- Multiple Linear Regression
- Ordination
- Random Forest Analysis

PROBABILISTIC MODELS

- Neural Networks
- Logistic Regression
- Bayesian Decisions
- Monte Carlo
- Random Forest

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

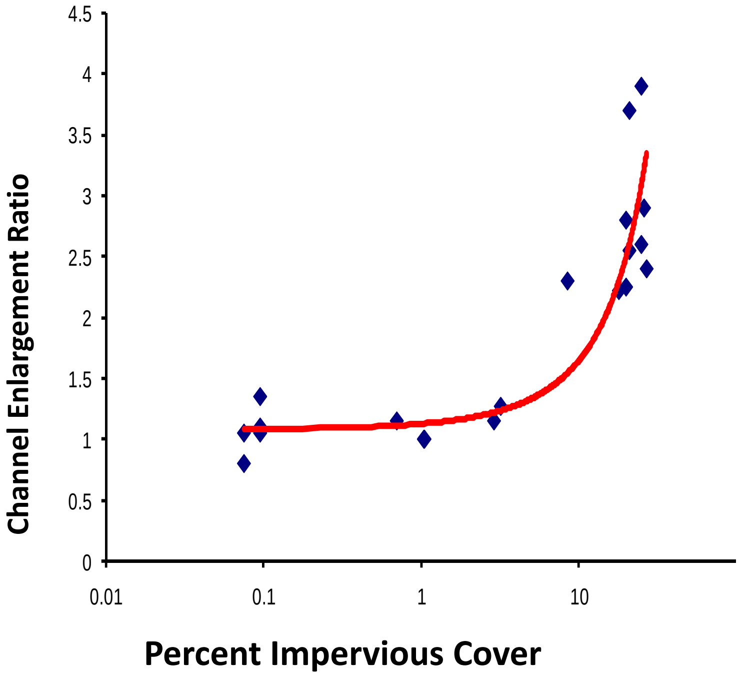
Pros

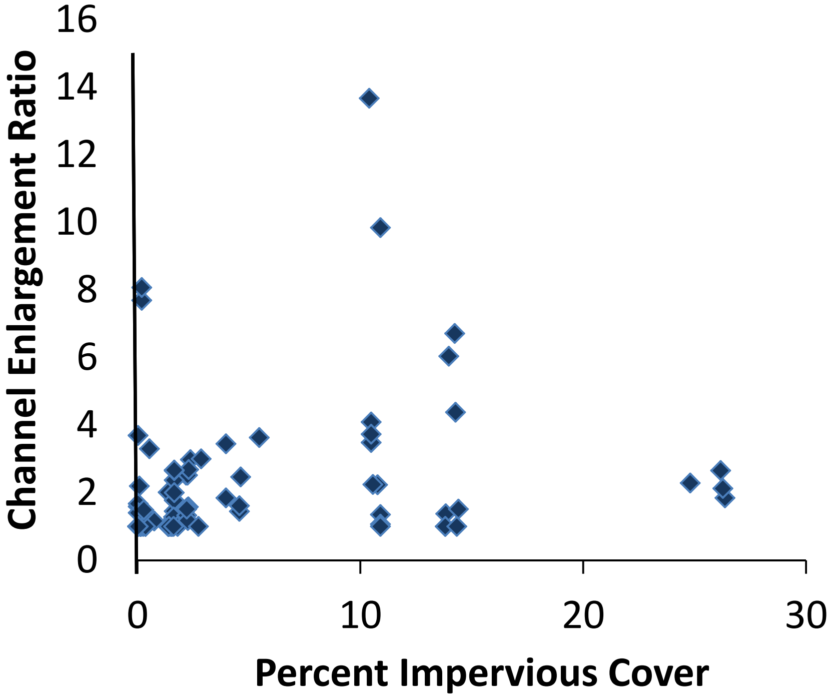
- Addresses questions of basic condition, susceptibility, etc.
- Relatively rapid and easy to apply
- Answers are generally qualitative or semi-quantitative
- Appropriate for screening-level decisions

Cons

- Difficult to model due to uncertain responses
- Cumulative Error

Non-linear responses



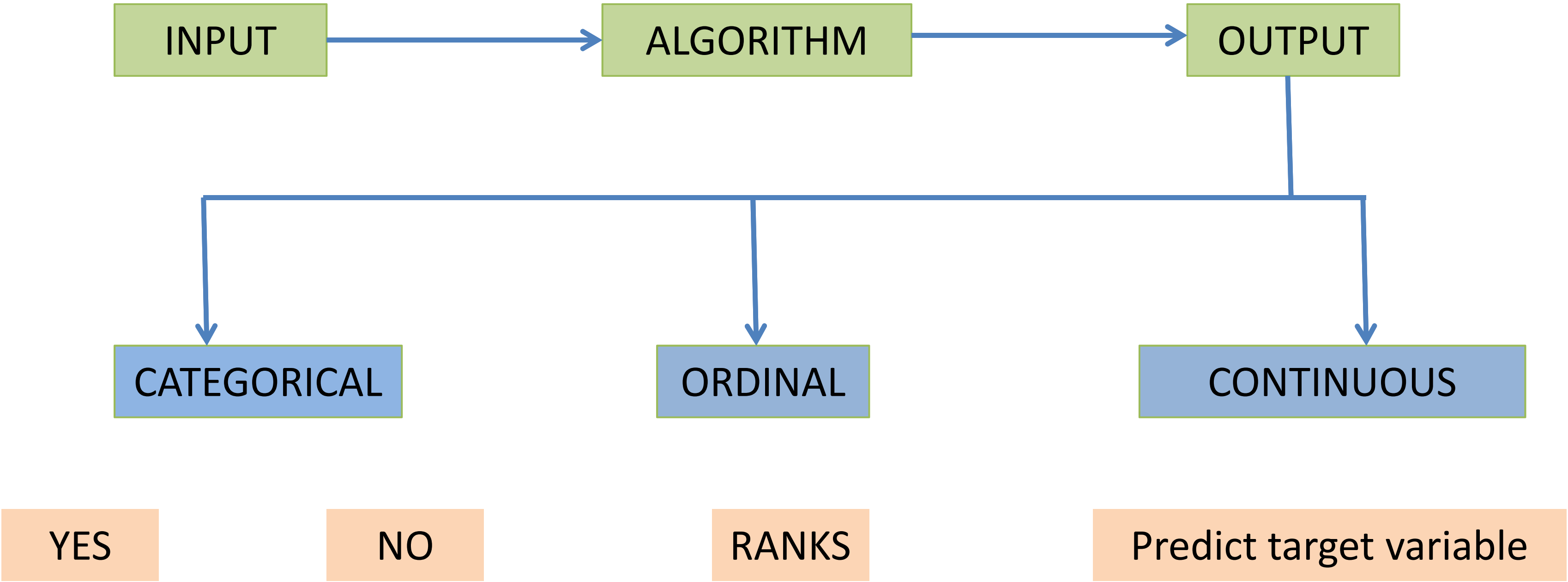


Modeling Tools

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

ORDINAL

CONTINUOUS

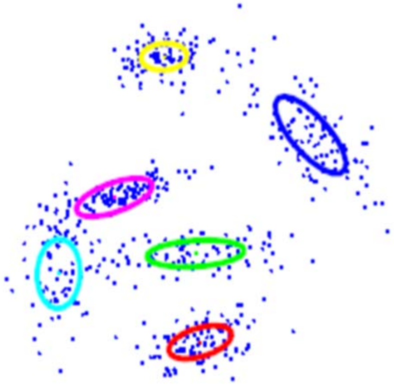
YES

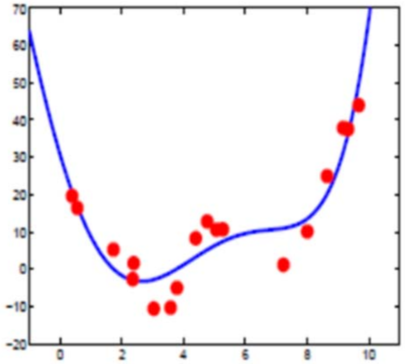
NO

RANKS

Predict target variable

Machine Learning and Approaches

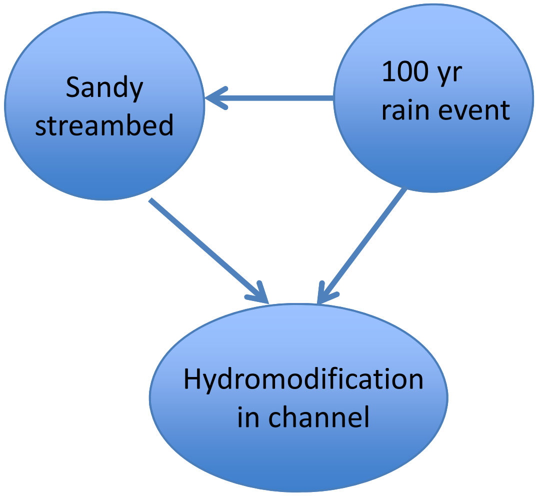




Probabilistic Graphical Models/Bayesian Networks

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

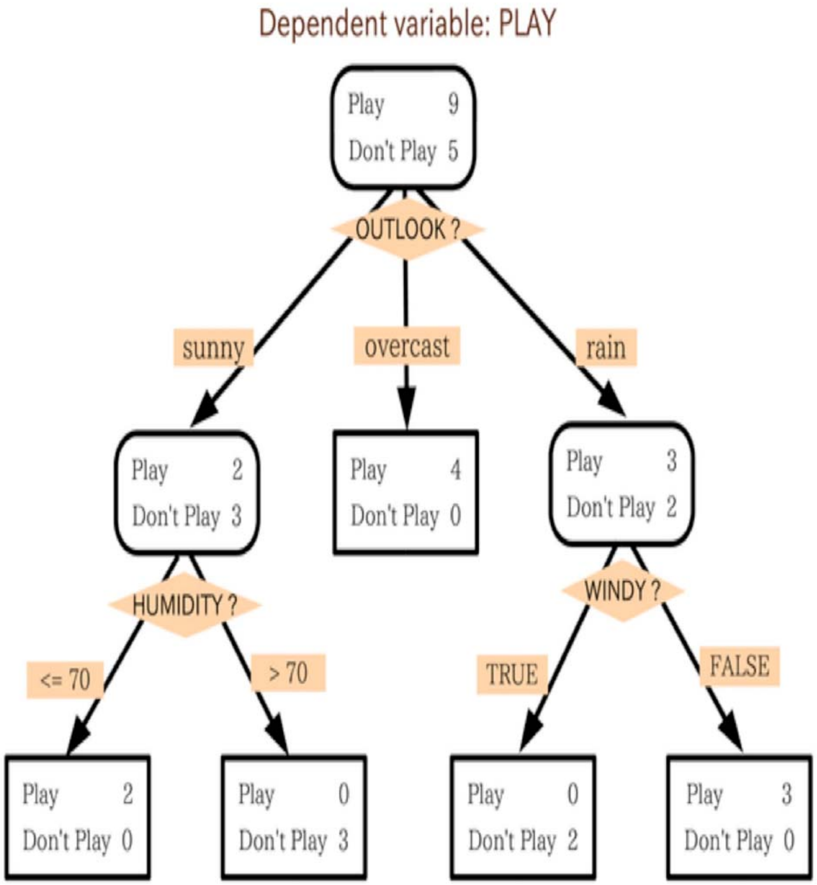
- Model encodes dependencies among variables, accounts for missing data easily
- Learns causal relationships, can be used to gain understanding about a problem domain and to predict the consequences of intervention.
- Model has both causal and probabilistic semantics, it is an ideal representation for 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
- 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.

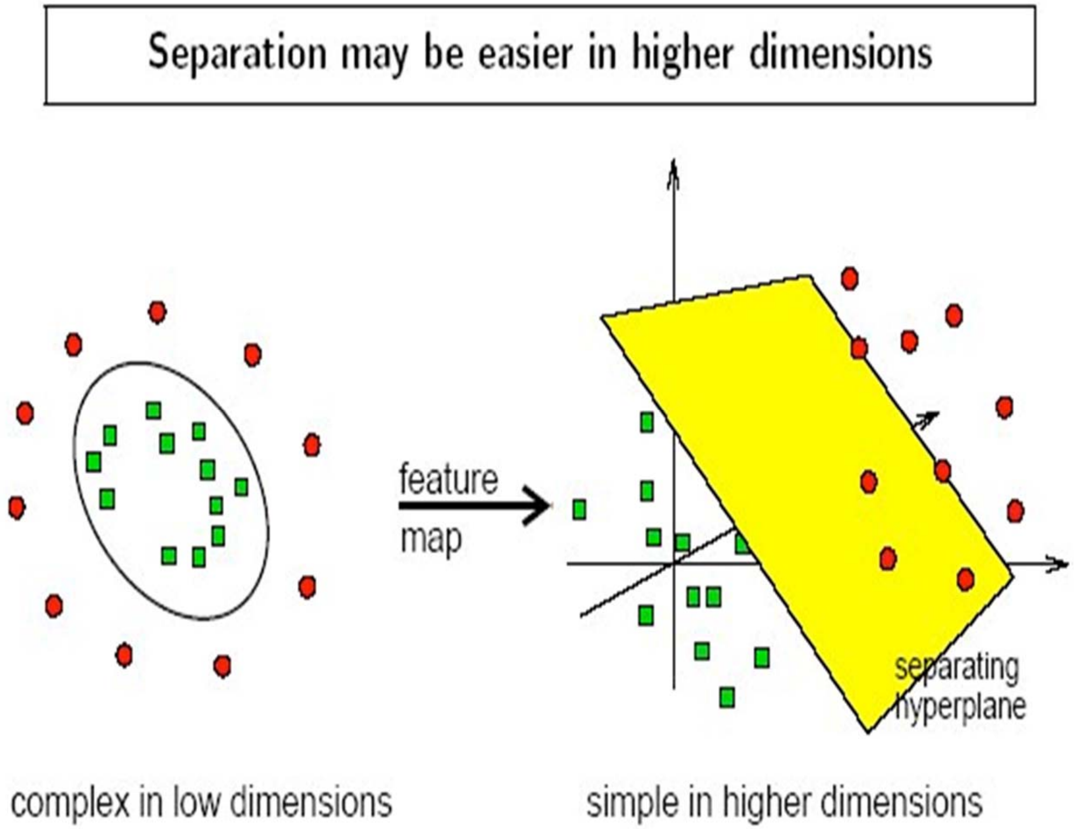


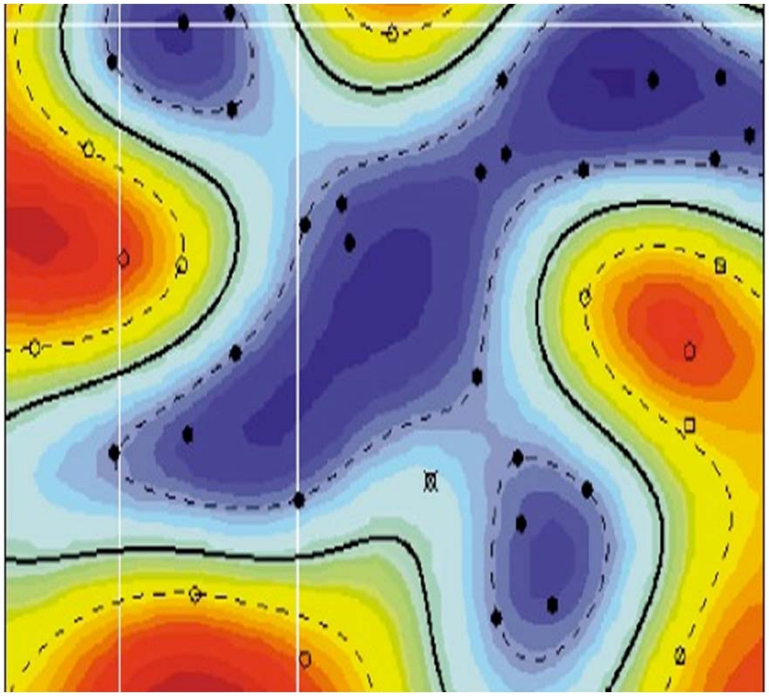
Support Vector Machine

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
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions



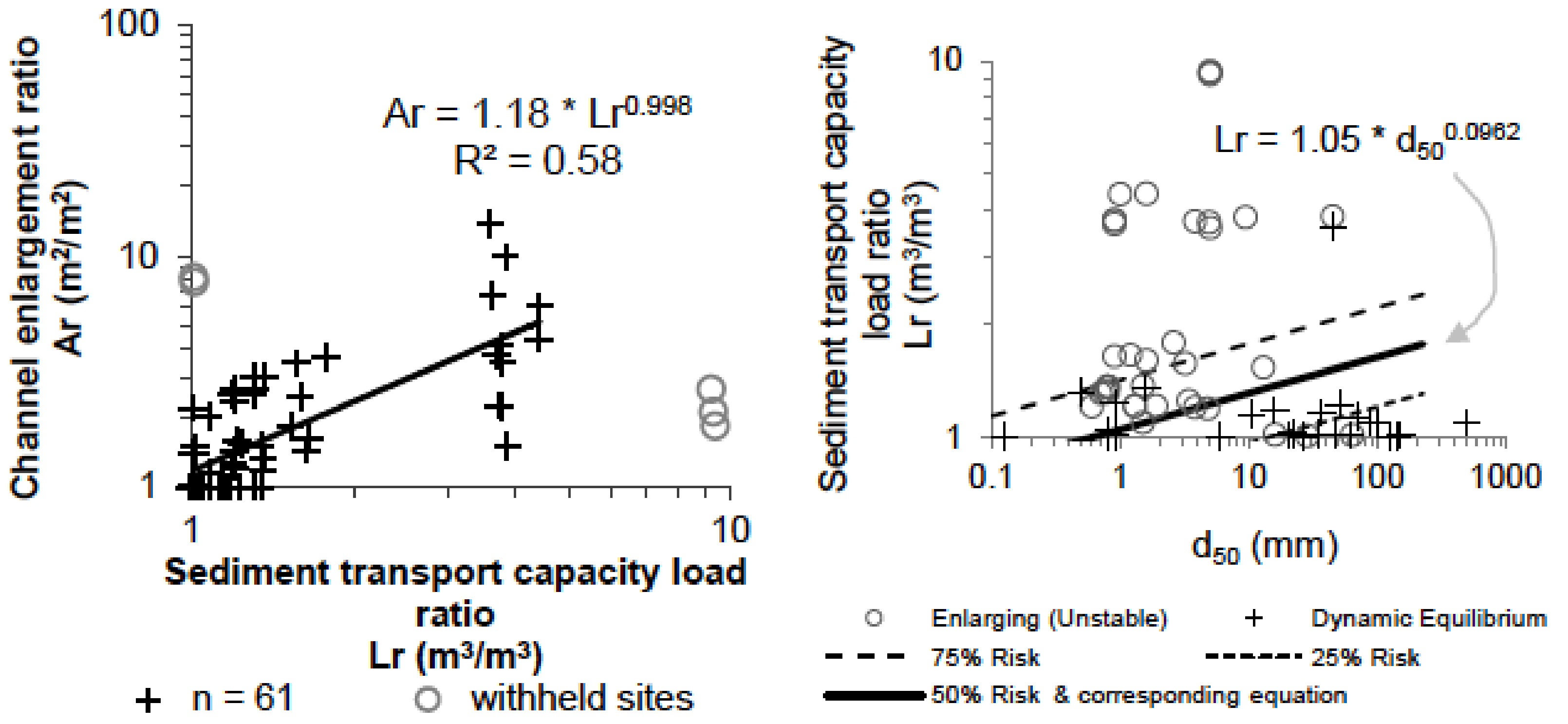
Logistic Regression

- 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



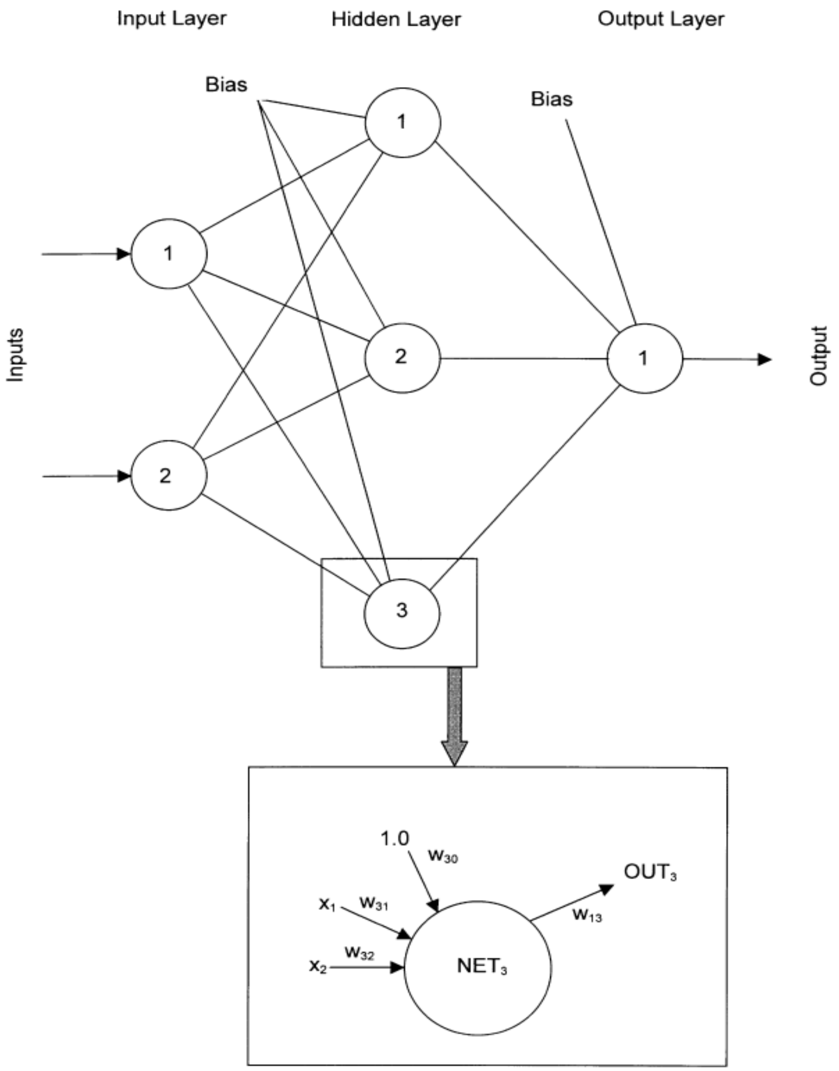
(a) enlargement vs. erosion potential

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

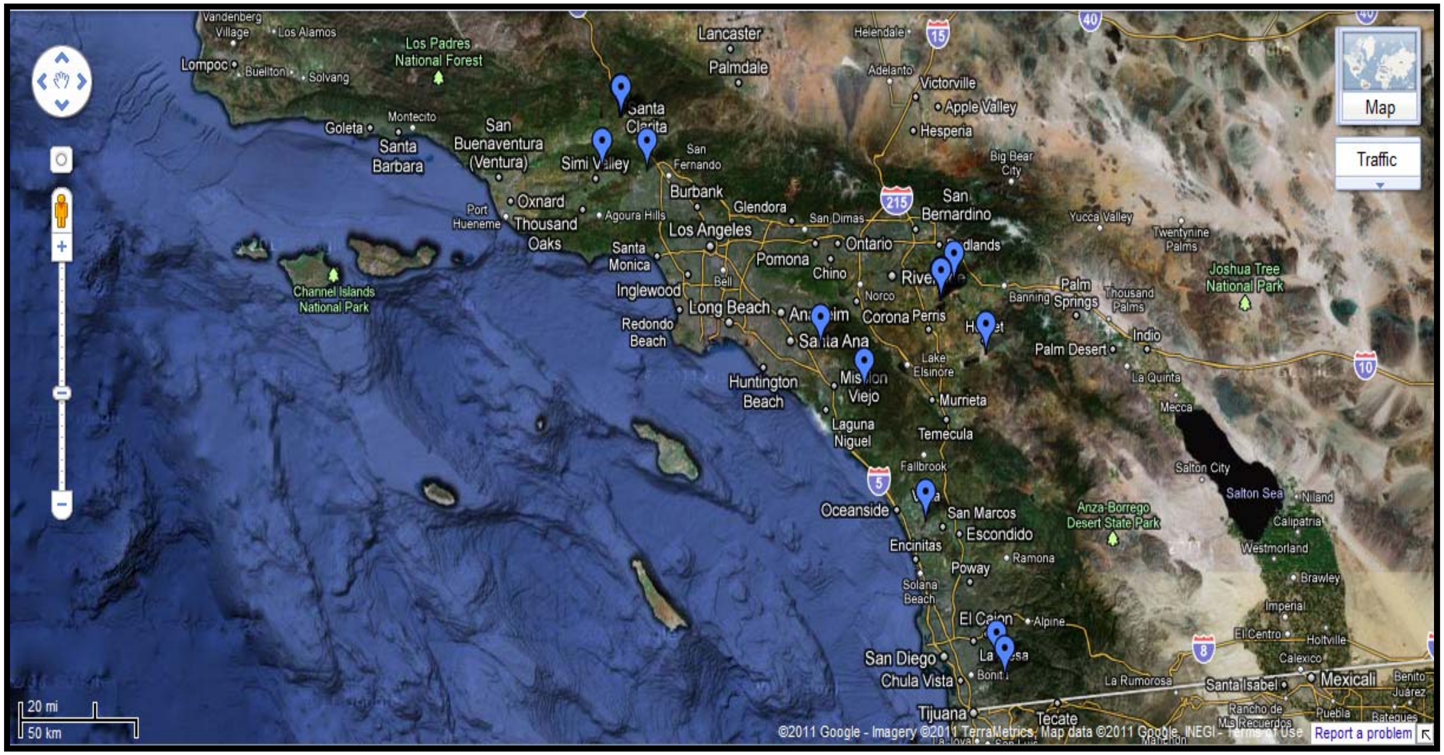
[$] based on the withholding of two stream reaches where enlargement was primarily driven by historic channelization (San Antonio) or kept artificially low due to dense vegetation (Agua Hedionda); both factors were poorly distributed in our dataset

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

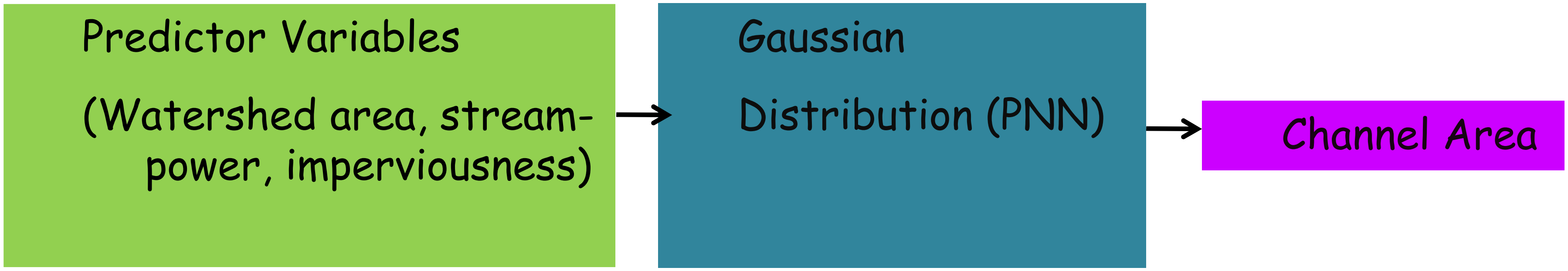


Case Study: General Regression Neural Networks



- 25 different locations
- 85 transects

Neural Network Setup



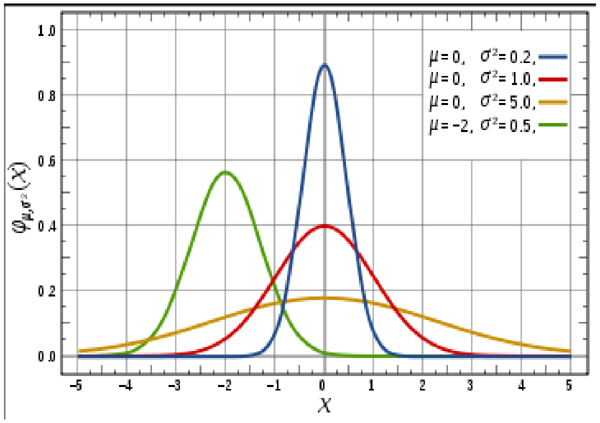
Predictor Variables

(Watershed area, stream-power, imperviousness)

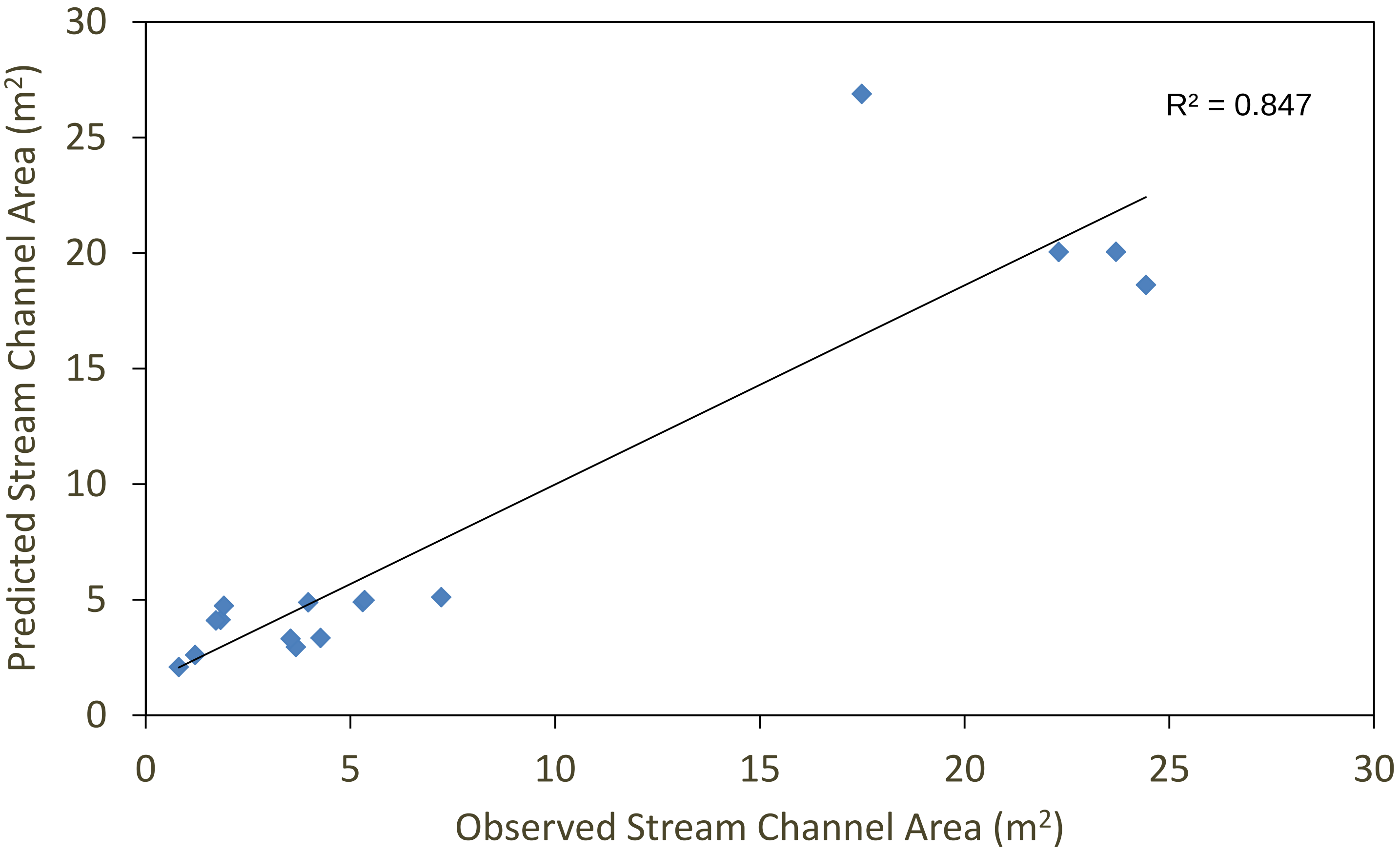
Gaussian

Distribution (PNN)

Channel Area



For Most Validation Scenarios: Channel Area Predicted Well



Comparison with Multivariate

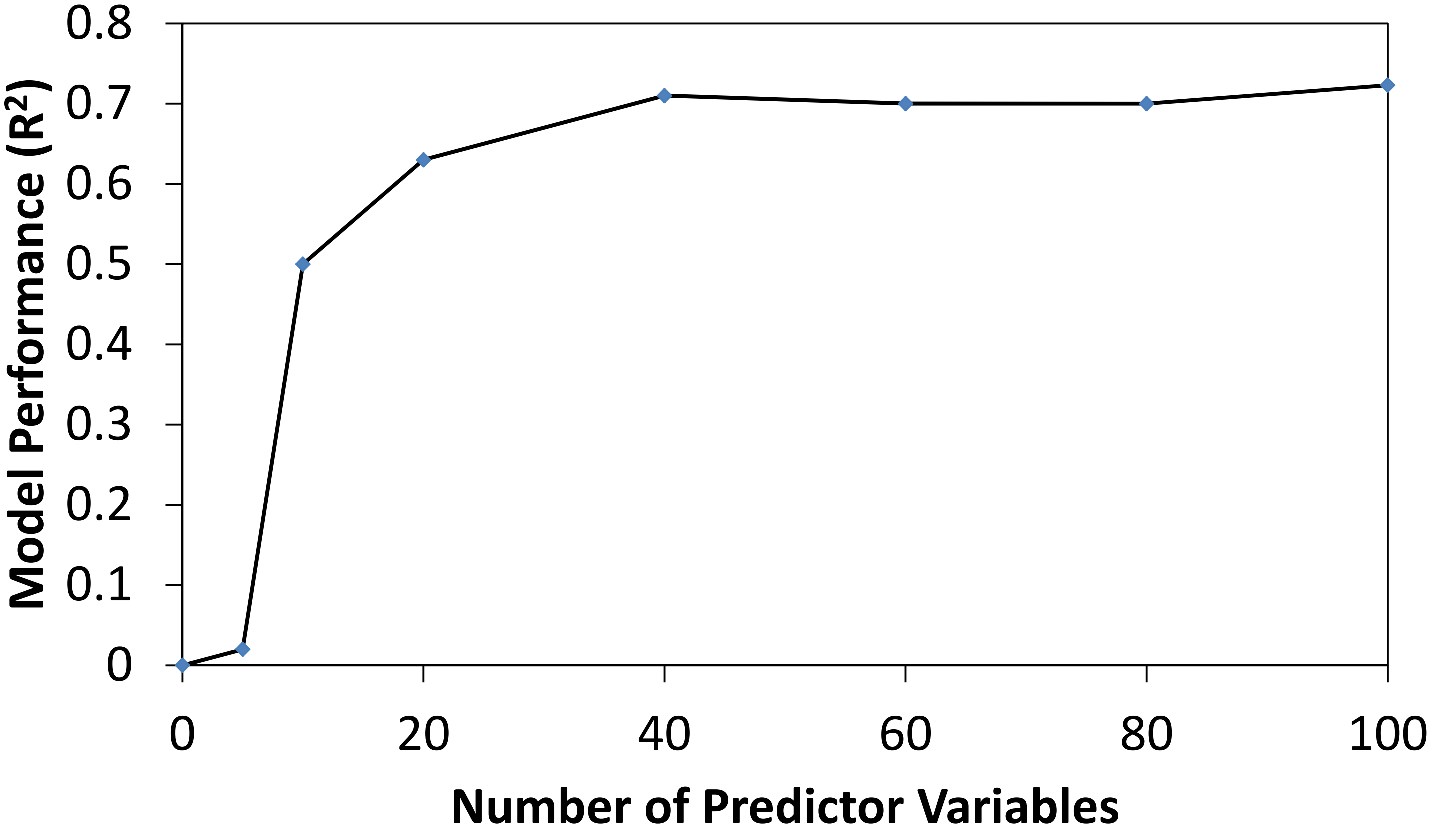
Cross Validation: 10%

	Neural Networks		Multivariate Regression	
Flow	Calibration	Validation	Calibration	Validation
Q2	0.99	0.77	0.74	0.61
Q10	0.99	0.79	0.68	0.53
Q50	0.99	0.85	0.8	0.37
Q100	0.99	0.82	0.73	0.57

Random Holding: 20%

	Neural Networks		Multivariate Regression	
Flow	Calibration	Validation	Calibration	Validation
Q2	0.99	0.66	0.89	0.1
Q10	0.96	0.63	0.82	0.26
Q50	0.99	0.55	0.84	0
Q100	0.99	0.65	0.82	0.4

Sigmoidal Decrease in Model Performance



Certain Variables Were Consistently Ranked Higher

Predictor Variable	Q2	Q10	Q50	Q100
Calculated Flow	1	3	9	0
Bedload Capability	2	5	5	7
Geotechnical Stability of Cross-section	3	3	3	4
Total Impervious Area	4	9	15	0
Stream Power	6	6	NA	NA
Bed material	8	7	10	5
Distance to Hardpoint	0	15	7	3

Interesting Observations