Predicting the Probability of Attaining Water Quality Classes

Tom Danielson, Ph.D.
Jeanne DiFranco
Beth Connors
Leon Tsomides

MAINE DEPARTMENT OF ENVIRONMENTAL PROTECTION
Protecting Maine's Air, Land and Water

Overview

- Conceptual Introduction to Discriminant Analysis
  - Example with iris data
- Case study related to water quality
  - Wetland bioassessments

Discriminant Analysis

- Predicts the probability of a sample belonging to a pre-defined group

Ronald Fisher, 1936

"The greatest biologist since Darwin"
Richard Dawkins

Edgar Anderson's Iris Data (1935)

Beachhead Iris
Iris setosa

L. Bernler

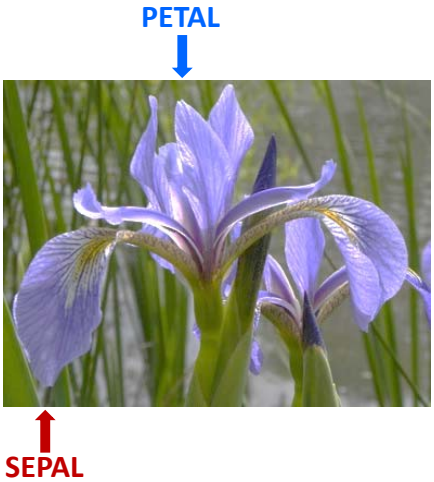
Northern Blue Flag
Iris versicolor

Bob Gutowski

Southern Blue Flag
Iris virginica

C. Pierce

Iris Measurements

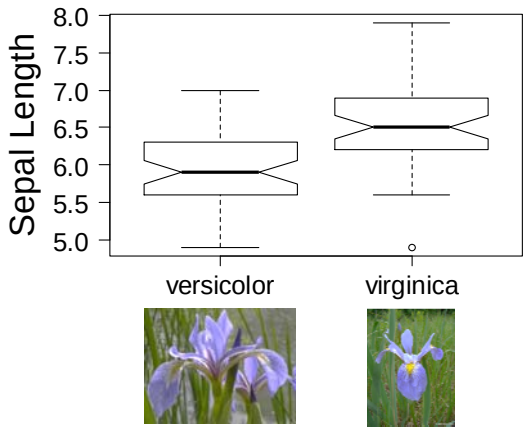


50 flowers of each species

- Petal length (cm)
- Petal width (cm)
- Sepal length (cm)
- Sepal width (cm)

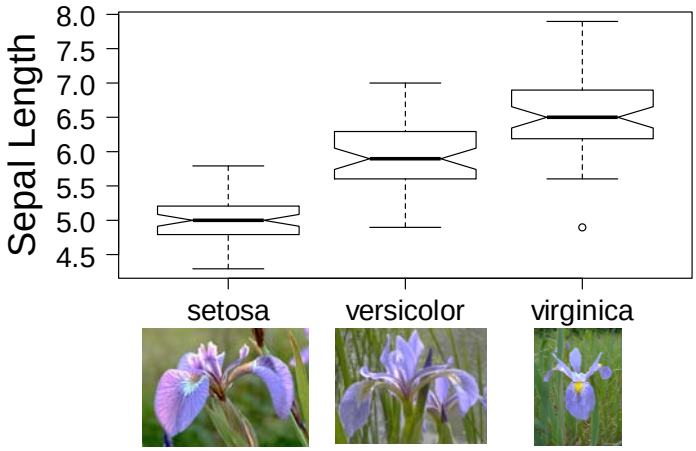
Comparing One Variable with Two Species

T-test ($p < 0.001$, means are different)

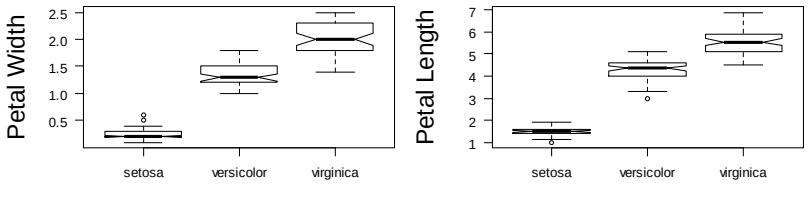


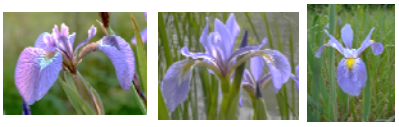
Comparing One Variable with Three Species

ANOVA ($p < 0.001$, at least one mean is different)



Same For All Four Variables





Discriminant Analysis

- Classifies samples based on patterns in the variables
  - 2 or more groups
  - Groups must be already defined
  - Multiple variables that show some differences among groups (metrics)
  - Transform to approximate normal distribution

Data Format

n = 150, randomly selected 100 for model building and 50 for validating model

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
…	…	…	…	…

2-Variable Model

- **Sepal Length**
- **Sepal Width**

		Model		
		setosa	versicolor	virginica
Actual	setosa	**33**	1	0
	versicolor	1	**21**	9
	virginica	0	9	**26**



- Model Building (n=100)

  80% correct

2-Variable Model

- **Sepal Length**
- **Sepal Width**

		Model		
		setosa	versicolor	virginica
Actual	setosa	**33**	1	0
	versicolor	1	**21**	9
	virginica	0	9	**26**

- Model Building (n=100)

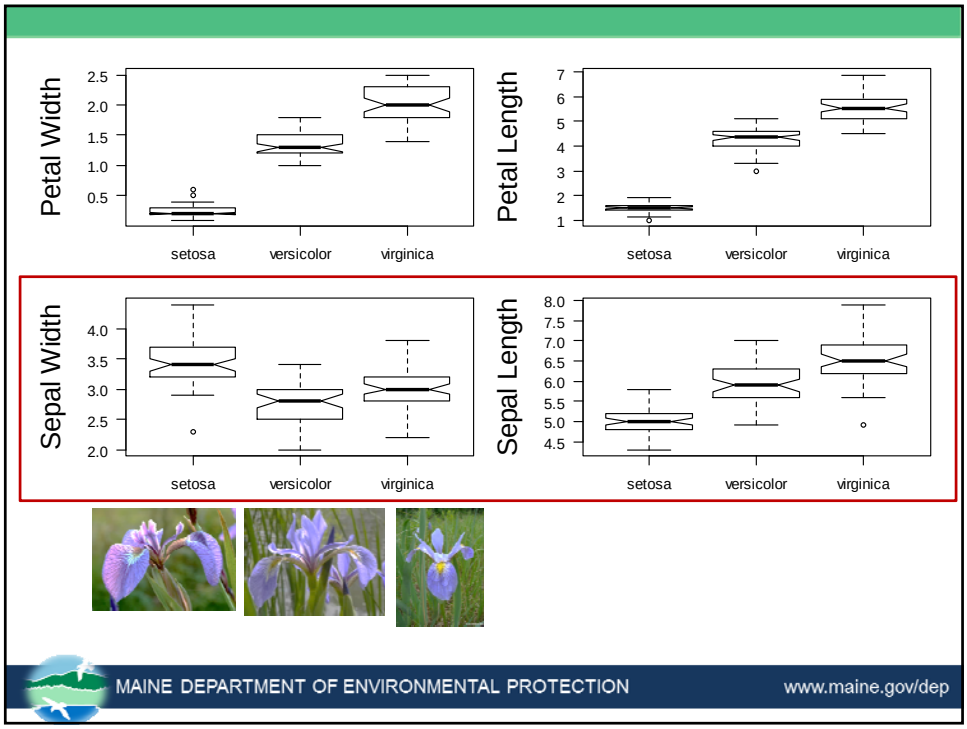
  80% correct

		Model		
		setosa	versicolor	virginica
Actual	setosa	**16**	0	0
	versicolor	0	**14**	5
	virginica	0	6	**9**

- Validation (n=50)

  88% correct

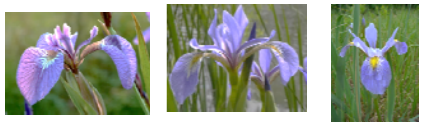
MAINE DEPARTMENT OF ENVIRONMENTAL PROTECTION www.maine.gov/dep



3-Variable Model

- **Petal Length**
  Sepal Length
  Sepal Width

		Model		
		setosa	versicolor	virginica
Actual	setosa	**34**	0	0
	versicolor	0	**29**	2
	virginica	0	2	**33**



- Model Building
  98% correct

  100% correct with validation data

MAINE DEPARTMENT OF ENVIRONMENTAL PROTECTION
www.maine.gov/dep

Variable Coefficients

	Constant	Sepal Length (V_1)	Sepal Width (V_2)	Petal Length (V_3)
setosa	-77.7	24.6	17.3	-20.3
versicolor	-73.4	14.8	5.3	9.6
virginica	**-101.3**	**10.5**	**5.0**	**20.9**

$$Z_{virginica} = \alpha_0 + \alpha_1 V_1 + \alpha_2 V_2 + \alpha_3 V_3$$

What species is this new flower?

Virginia Botanical Associates

Iris setosa

Iris versicolor

Iris virginica

	constant	Sepal Length	Sepal Width	Petal Length
setosa	-77.7	24.6	17.3	-20.3
versicolor	-73.4	14.8	5.3	9.6
virginica	**-101.3**	**10.5**	**5.0**	**20.9**

$$Z_{virginica} = -101.3 + 10.5\ (6.0) + 5.0\ (3.0) + 20.9\ (5.0)$$

Sepal Length	Sepal Width	Petal Length
6.0	**3.0**	**5.0**

	constant	Sepal Length	Sepal Width	Petal Length
setosa	-77.7	24.6	17.3	-20.3
versicolor	**-73.4**	**14.8**	**5.3**	**9.6**
virginica	-101.3	10.5	5.0	20.9

$$Z_{versicolor} = -73.4 + 14.8\ (6.0) + 5.3\ (3.0) + 9.6\ (5.0)$$

Sepal Length	Sepal Width	Petal Length
6.0	**3.0**	**5.0**

Calculate Z Scores



	setosa	versicolor	virginica
Z	20	80	82

Calculate Probabilities



	setosa	versicolor	virginica	Sum
Z	20	80	82	--
e^Z	7.1 E+08	3.4 E+34	2.8 E+35	**3.2 E+35**

Calculate Probabilities



	setosa	versicolor	virginica	Sum
Z	20	80	82	--
e^Z	7.1 E+08	3.4 E+34	2.8 E+35	3.2 E+35

$$Probability_{virginica} = \frac{e^{Z virginica}}{e^{Z setosa} + e^{Z versicolor} + e^{Z virginica}}$$

Result



	setosa	versicolor	virginica
Z	20	80	82
e^Z	7.1 E+08	3.4 E+34	2.8 E+35
Probability	0.00	0.11	0.89

Result



	setosa	versicolor	virginica
z	20	80	82
e^z	7.1 E+08	3.4 E+34	2.8 E+35
Probability	0.00	0.11	**0.89**

CONCLUSION:
The new flower is *Iris virginica*.

Case Study with Wetland Data

Narrative Aquatic Life Criteria

Class	Narrative Aquatic Life Criteria
GPA (Lakes)	Habitat natural. Stable or decreasing trophic state free of culturally induced algal blooms.
AA	Habitat natural and free flowing. Aquatic life as naturally occurs.
A	Habitat natural. Aquatic life as naturally occurs.
B	Habitat unimpaired. Must support all indigenous aquatic species. No detrimental changes to resident biological community.
C	Must support all indigenous fish species and maintain structure and function of resident biological community.

Narrative Aquatic Life Criteria

Class	Narrative Aquatic Life Criteria
GPA (Lakes)	Habitat natural. Stable or decreasing trophic state free of culturally induced algal blooms.
AA	Habitat natural and free flowing. Aquatic life as naturally occurs.
A	Habitat natural. Aquatic life as naturally occurs.
B	Habitat unimpaired. Must support all indigenous aquatic species. No detrimental changes to resident biological community.
C	Must support all indigenous fish species and maintain structure and function of resident biological community.

Study Sites

- **203 sites ranging from minimally disturbed to severely degraded**
- **Isolated wetlands or on fringe of lakes or low gradient streams**

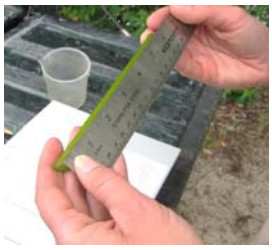
Sample Methods

- Clip submerged stems of aquatic plants
  - Yellow water lily
  - Pickerelweed
- 3 locations in wetland
- 5 stems per location

Sample Methods

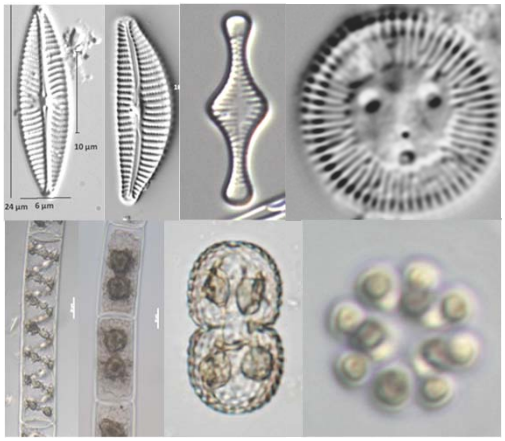
Rub **Rinse** **Measure**





Types of Algae

- >1,300 taxa
- Diatoms
- Green
- Red
- Cyanobacteria
- Dinoflagellates
- several other groups



Assigning Water Quality Groups

Step	Comment
4 biologists independently evaluated samples	• "Blind" – biologists only had algal species abundances and community metrics • Looked at patterns in metrics and key components of algal community • Compared samples to known ranges of values observed in minimally disturbed reference sites

Assigning Water Quality Groups

Step	Comment
4 biologists independently evaluated samples	• "Blind" – biologists only had algal species abundances and community metrics • Looked at patterns in metrics and key components of algal community • Compared samples to known ranges of values observed in minimally disturbed reference sites
4 biologists independently assigned water quality groups	• Assigned each sample to a group based on similarity or departure from reference conditions • Groups were A, B, C, or N • Also assigned BCG Tier

MAINE DEPARTMENT OF ENVIRONMENTAL PROTECTION www.maine.gov/dep

Assigning Water Quality Groups

Step	Comment
4 biologists independently evaluated samples	• "Blind" – biologists only had algal species abundances and community metrics • Looked at patterns in metrics and key components of algal community • Compared samples to known ranges of values observed in minimally disturbed reference sites
4 biologists independently assigned water quality groups	• Assigned each sample to a group based on similarity or departure from reference conditions • Groups were A, B, C, or N • Also assigned BCG Tier
Team made final decisions	• Discussed what variables were most important • Referred back to narrative aquatic life criteria and BCG • Compared each sample to reference conditions • Assigned samples to groups (A, B, C, N) using a consensus approach

Discriminant Model

- Build model to predict water quality groups
  - **A** - Class GPA/AA/A
  - **B** - Class B
  - **C** - Class C
  - **N** - Non-attainment (doesn't even attain C)
- 203 samples (163 training, 40 validating)

Potential Variables

- 57 candidate metrics, such as…
  - Relative abundance of species that like a lot of nutrients
  - Richness of species that like acidic water
  - Relative richness of clean water species
  - Index of organic enrichment

Stepwise Selection of Variables

- Program adds variables one at a time
- Stops when adding another variable doesn't help much

Stepwise Selection of Variables

- Program adds variables one at a time
- Stops when adding another variable doesn't help much

- Model with 7 variables
  - 92% correct
  - 82% correct with validation data

Our Approach

- Wrote an R code that tries all combinations of metrics
  - models based on 5 or 6 metrics
  - code identified the very best models

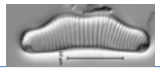
Selecting Model

- Biologists evaluated candidate models
  - Low error rate
  - Low error rate with validation data
  - Balance of misclassifications
  - "Good mix" of variables

"Good Mix" of Variables

- Metrics that relate back to biological criteria
- Metrics that are ecologically meaningful
- No redundant variables
  - Biologically redundant
  - Highly correlated
- Variety
  - Different measures, such as relative abundance, richness, index, etc.
  - Different components of algal community

Model Variables

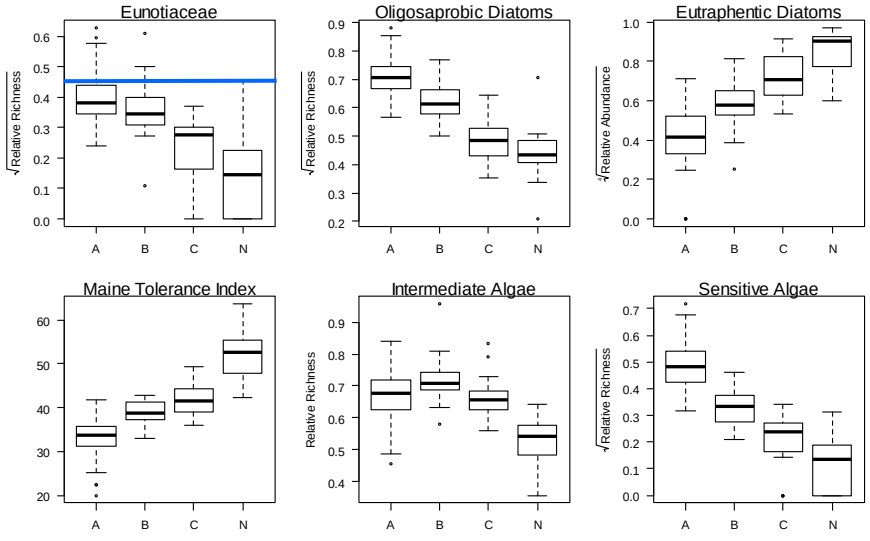
Metric Description	Environmental Significance
Eunotiaceae relative richness 	Prefer oligotrophic to mesotrophic, somewhat acidic conditions
Eutrophentic diatoms relative abundance	Prefer eutrophic conditions
Oligosaprobic diatoms relative richness	Prefer low organic enrichment
Maine Tolerance Index	Weighted-average community index
Sensitive taxa relative richness	Require clean water
Intermediate taxa relative richness	Unimodal response to disturbance

Variable based on only diatoms or all algae

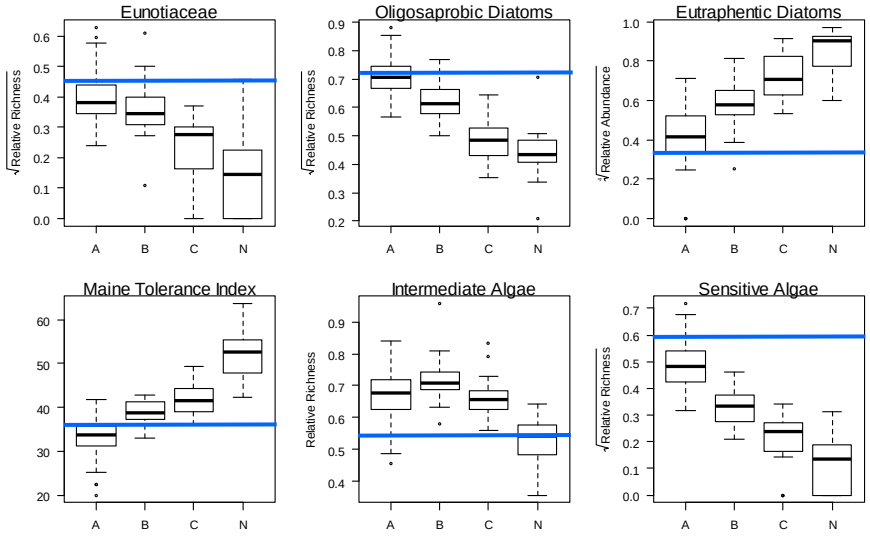
Model Performance

		Model				
		Class A	Class B	Class C	NA	% correct
Actual	Class A	94	1	0	0	99
	Class B	4	21	1	0	80
	Class C	0	1	26	0	96
	NA	0	0	1	18	95
					Overall % Correct	95%
					% Correct with Validation Data	85%

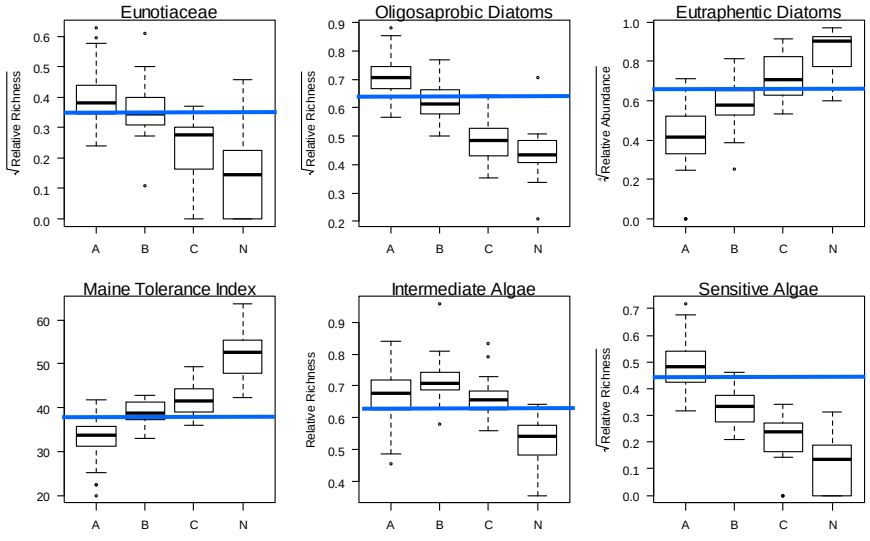
Great Works Pond



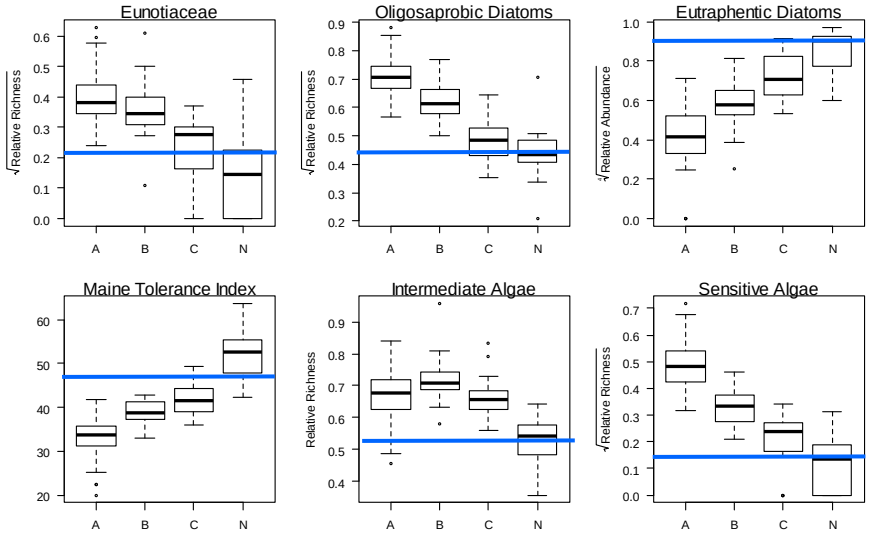
Great Works Pond (Probability A = 1.00)



Boyden Lake (Probability A = 0.50)



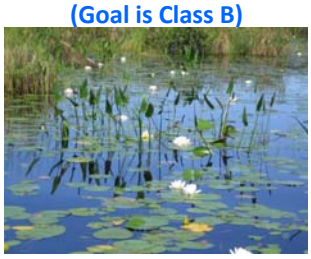
Capisic Pond (Probability NA = 0.98)



Application of Model

- Computes probability of attaining biological criteria based on the 6 metrics

New Wetland Sample (Goal is Class B)



Class	Probability
AA/A	0.20
B	**0.80**
C	0.00
Non-attainment	0.00

Contact:

Tom Danielson
thomas.j.danielson@maine.gov
207-441-7430

www.maine.gov/dep