

US EPA ARCHIVE DOCUMENT

Conditional Probability Analysis and Changepoints: Computational Sessions I & II

Questions that this session will address:

What is conditional probability analysis (CPA)?

What are changepoints?

**How are these tools being used in environmental
analyses?**

**What computational tools are available for
exploratory data analysis?**

Conditional Probability Analysis and Changepoints: Computational Sessions I & II

1:00 John Paul (USEPA, RTP NC) - Conditional probability analysis and change-point analysis: Background and Introduction.

1:30 Neil Kamman (VT, IPA to EPA) Use of conditional probability in development of nutrient criteria in VT lakes.

2:00 Ben Jessup (TetraTech) - Setting sediment benchmarks using multiple analytical methods.

3:00 John Kiddon (USEPA, AED) - Finding the Story in the Data: Using Excel to Visualize and Analyze Lake and Estuarine Monitoring Data.

3:00 break

3:30: Jeff Hollister (USEPA, AED) - Using CProb in R and Excel to conduct conditional probability analysis.

4:30 Jeffrey White (TetraTech) - EDAS2, Environmental Data Analysis System

4:00 open question / discussion period



Conditional probability analysis and changepoint analysis: Background and Introduction.

John F. Paul

**National Health and Environmental Effects Research
Laboratory**

U.S. Environmental Protection Agency

Research Triangle Park, North Carolina 27711



What is conditional probability?

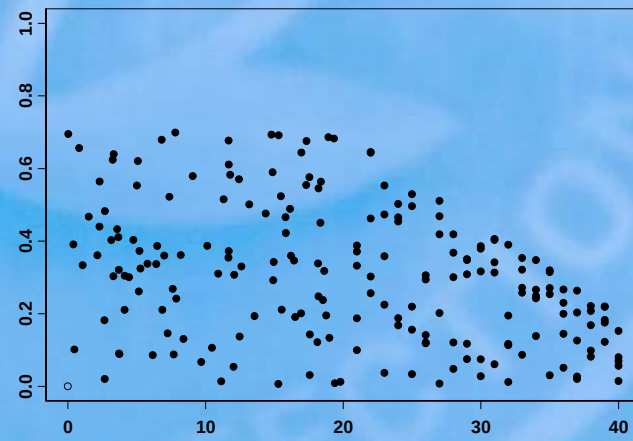
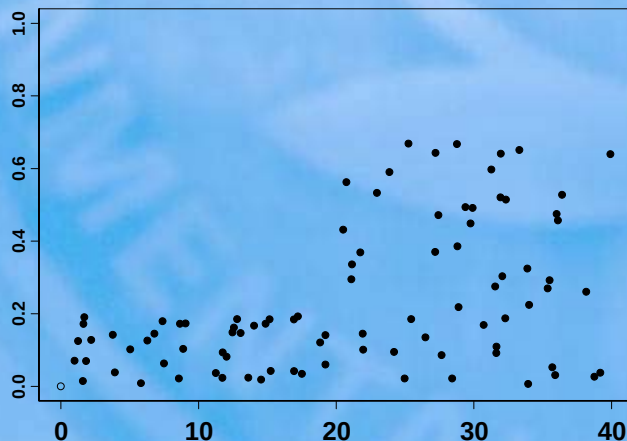
The probability that an event occurs if another event has occurred.

e.g., probability of a stream being impaired if sustained DO < 2 mg/L

What is a changepoint?

Value that separates response variable into 2 groups, each with distinct characteristics.

Corresponding value of environmental variable is called a threshold.



Outline of Presentation

- History
- Background
- Simple Examples
- Wrap-up

History of Conditional Probability Analysis

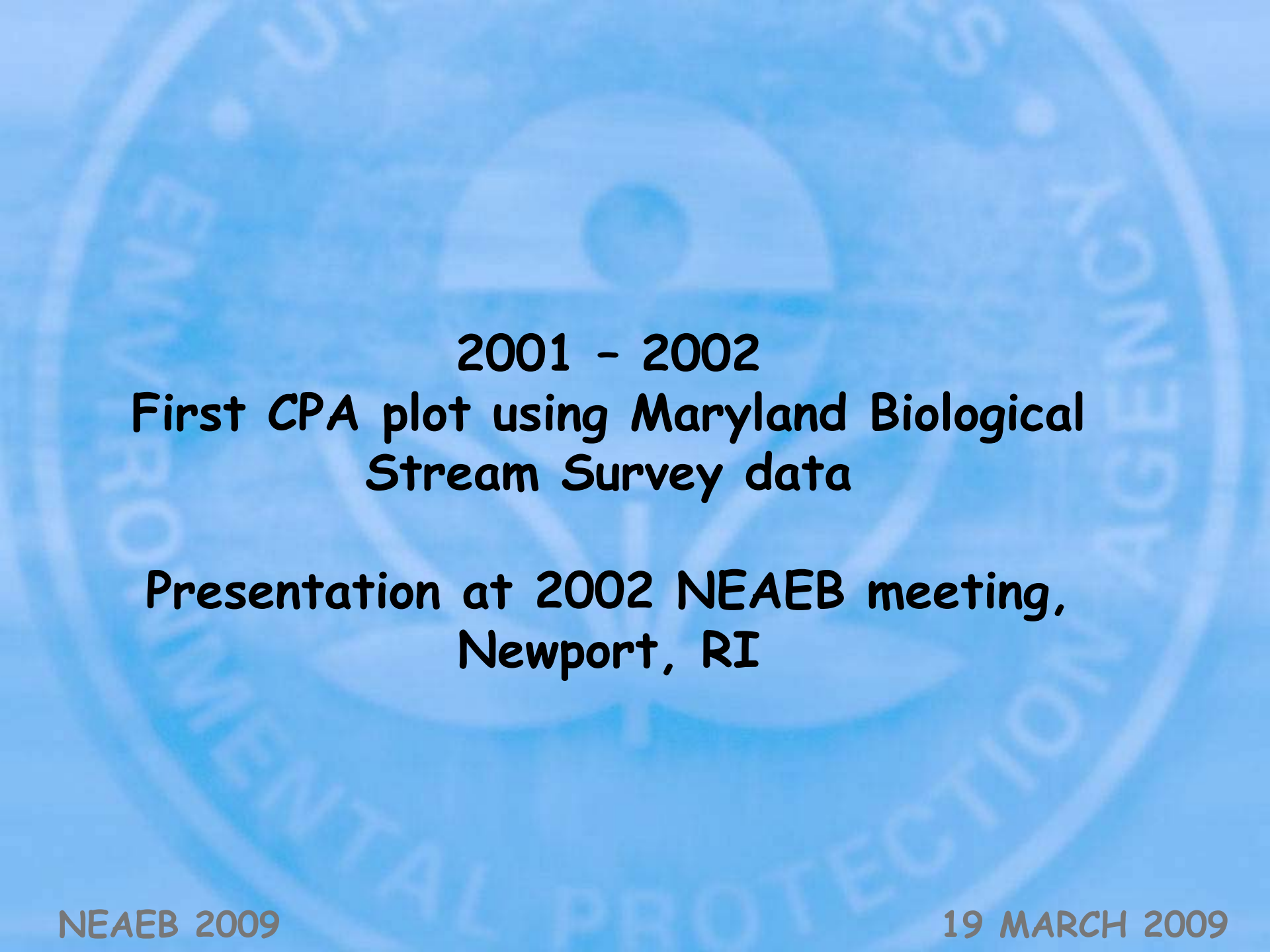
(as applied to ecological assessments)

2000 - Critical Realization: results of analysis using probability data can be expressed as probability of occurrence

40% of streams impaired

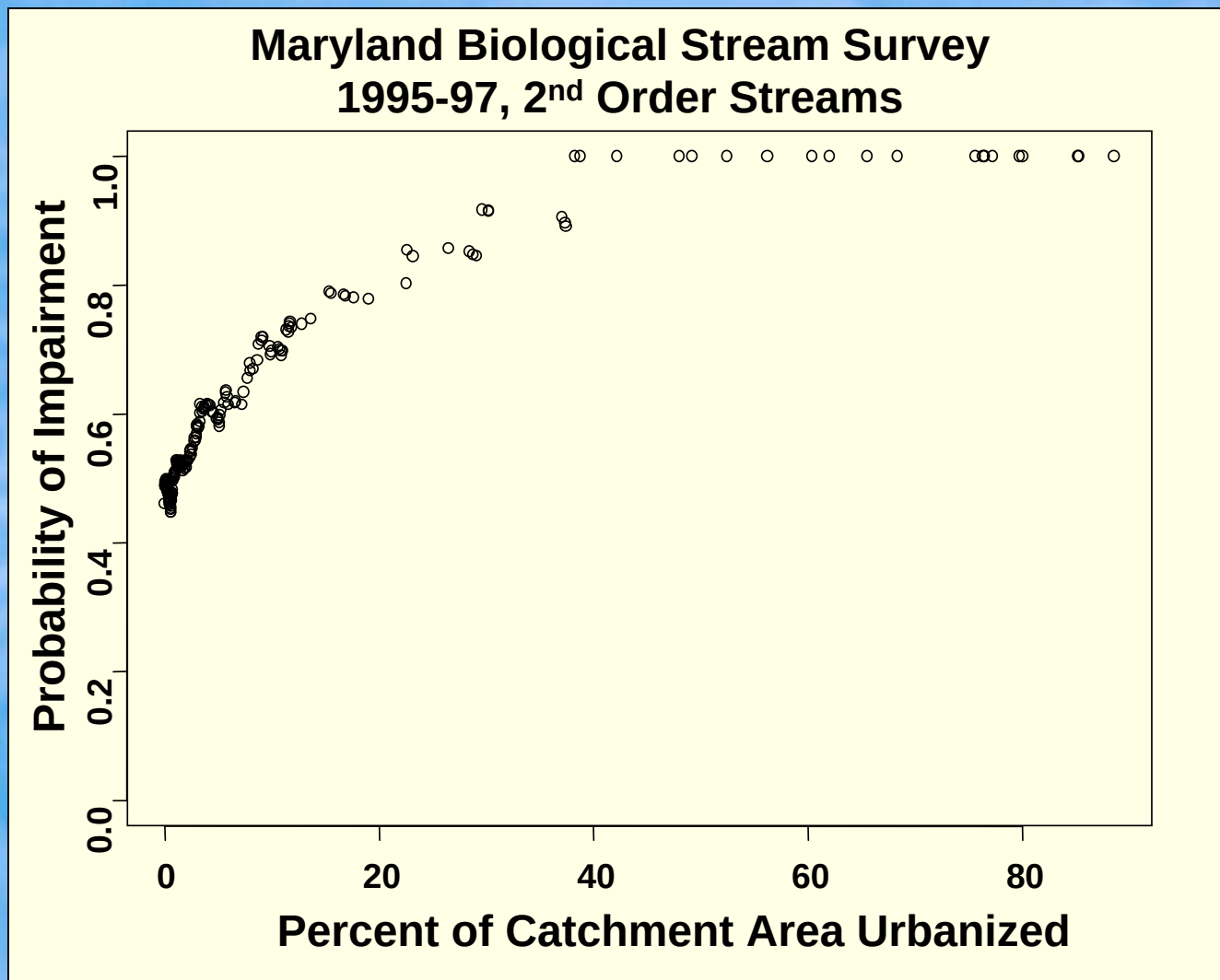


probability of having observed impaired streams is 0.40



2001 - 2002
**First CPA plot using Maryland Biological
Stream Survey data**

**Presentation at 2002 NEAEB meeting,
Newport, RI**

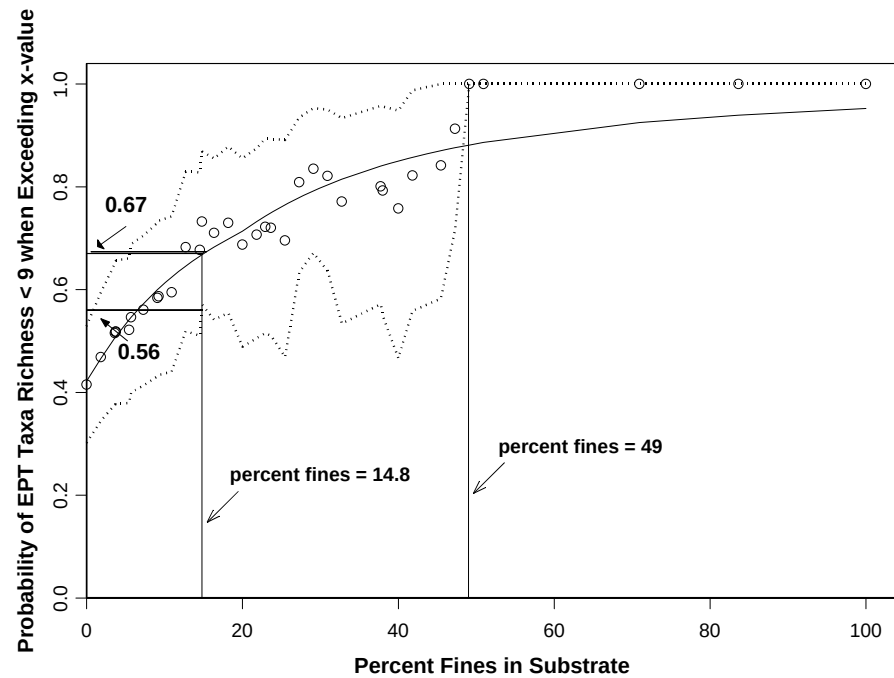
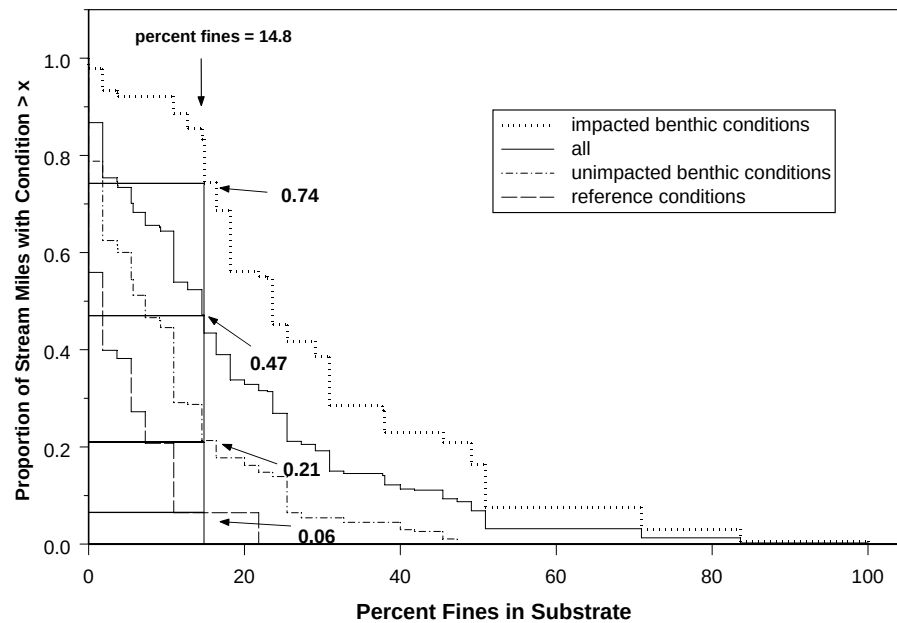


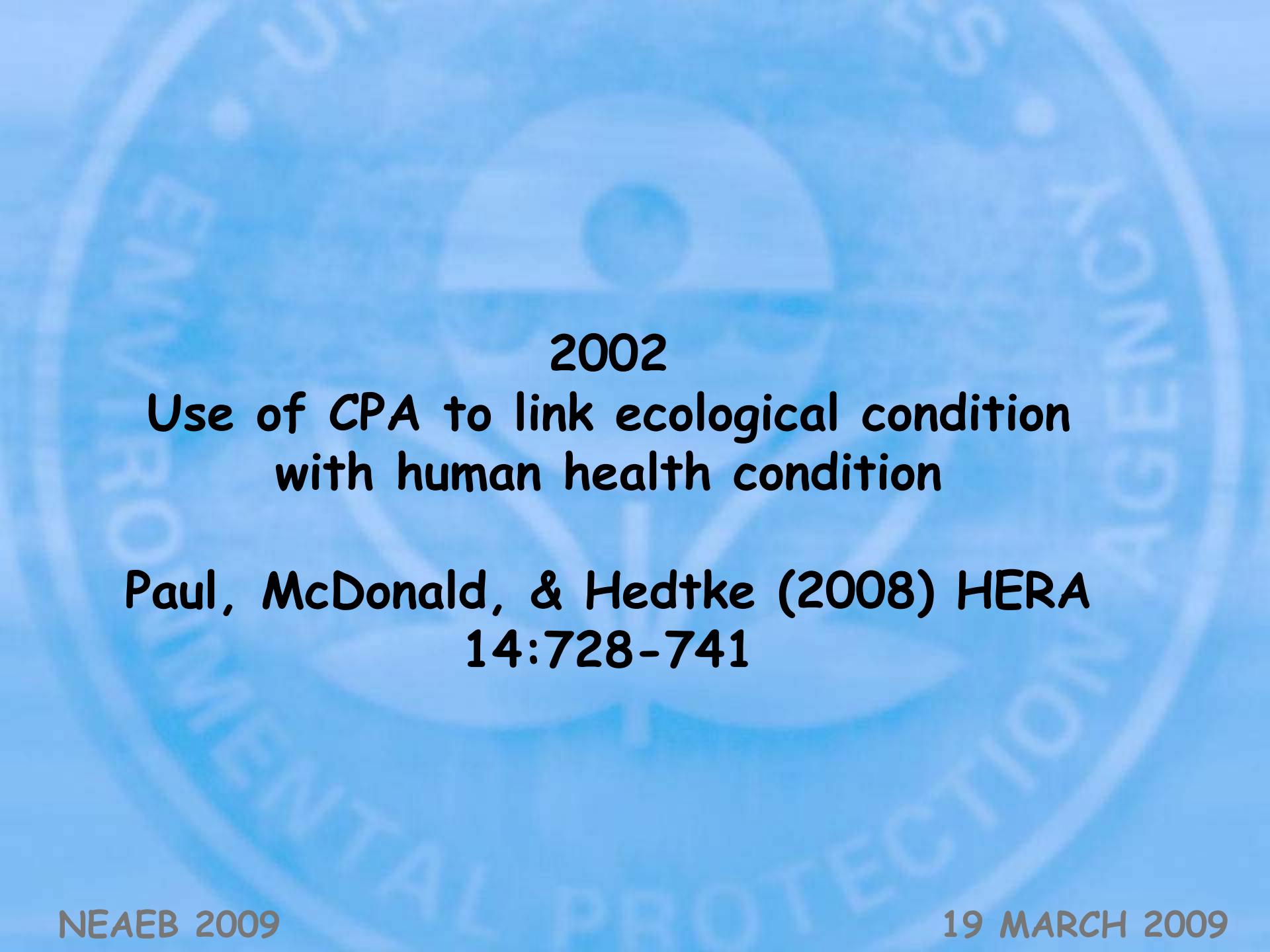
2002

Use of CPA to identify candidate criteria

**Workshop on Suspended Sediments and Solids,
Cincinnati, 11-12 July 2002**

**Paul, J. F. and M. E. McDonald (2005)
JAWRA 41(5): 1211-1223.**

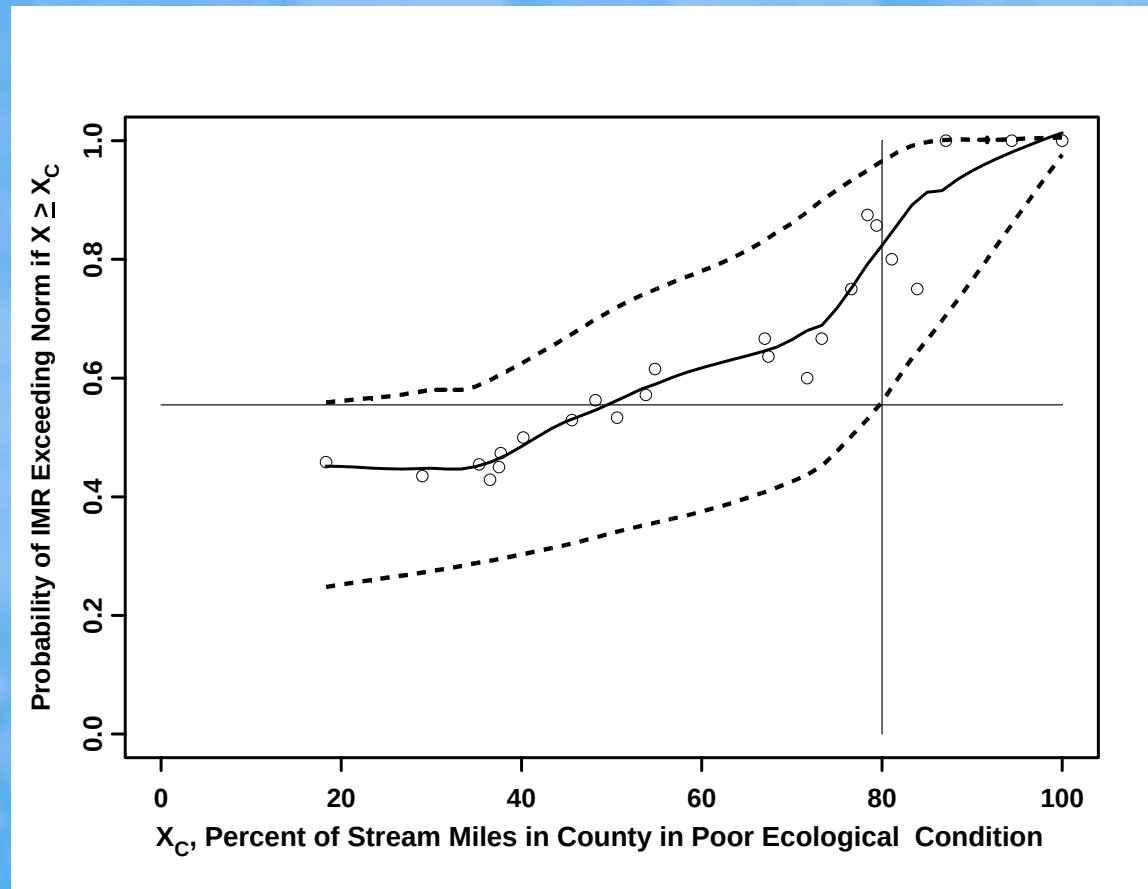




2002

**Use of CPA to link ecological condition
with human health condition**

**Paul, McDonald, & Hedtke (2008) HERA
14:728-741**



2005

SABS Framework published by OW & ORD

**U.S. EPA. (2006) Framework for Developing
Suspended and Bedded Sediments (SABS)
Water Quality Criteria EPA-822-R-06-001.
Washington DC: U.S. EPA.**



United States Environmental Protection Agency
Office of Water
Office of Research and Development

EPA-822-R-06-001

March 2006

**FRAMEWORK FOR DEVELOPING
SUSPENDED AND BEDDED SEDIMENTS (SABS)
WATER QUALITY CRITERIA**

Rationale for Using with Field Data

Conditional probability is a logical way to support identification of thresholds in ecological analyses:

- we protect against biological impact by specifying a threshold value not to be exceeded
- if we exceed the threshold value, then we can estimate the probability of aquatic impact
- incorporates uncertainty and natural variability into analysis

Simple Example to Illustrate State-Wide Wadable Stream Survey

Site ID	Stressor	Fish IBI	Benthic IBI
XX-1	3	2.98	4.01
XX-2	6	3.27	3.27
XX-3	2	1.78	1.78
XX-4	9	3.63	3.63
XX-5	1	3.84	3.51
XX-6	4	1.68	2.42
.....			

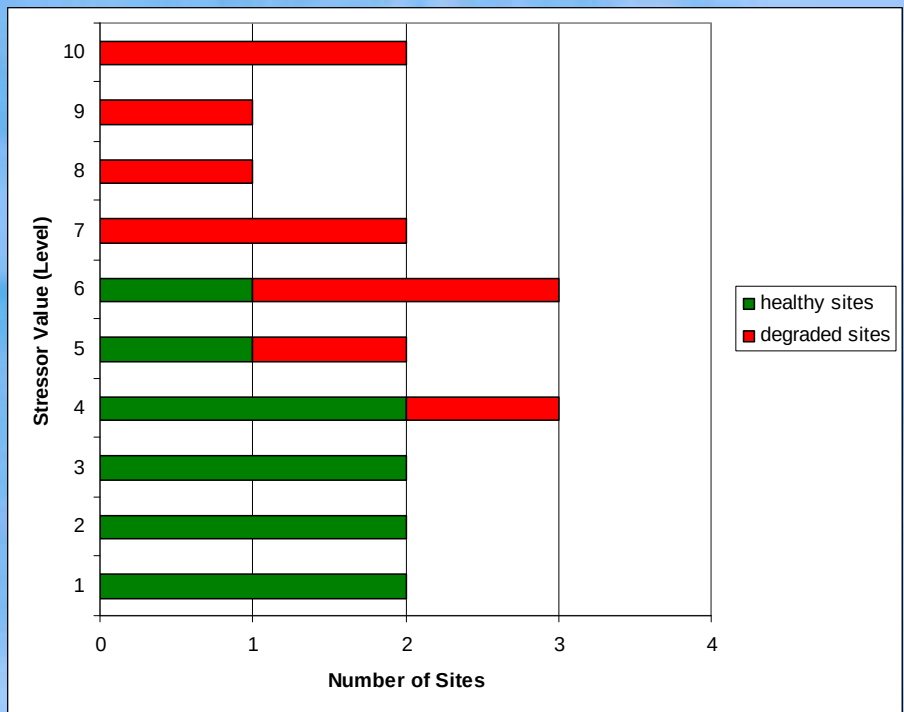
Simple data example: Simple plot of Data

Response variable as dichotomous variable

Degraded site - unacceptable biological conditions based on biotic index

Healthy site - acceptable biological conditions based on biotic index

Stressor - exposure variable, biological conditions become more unacceptable as value increases

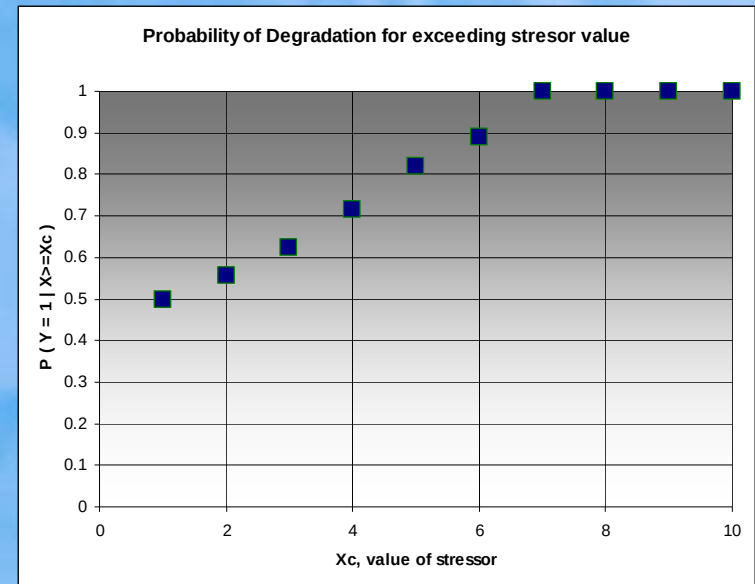


Stressor value	Number of Healthy Sites	Number of Degraded Sites
1	2	0
2	2	0
3	2	0
4	2	1
5	1	1
6	1	2
7	0	2
8	0	1
9	0	1
10	0	2

Simple data example:

CPA for stressor value $\geq X_c$

Probability that a site would be degraded if we observe exceedance of a specific stressor value (X_c)

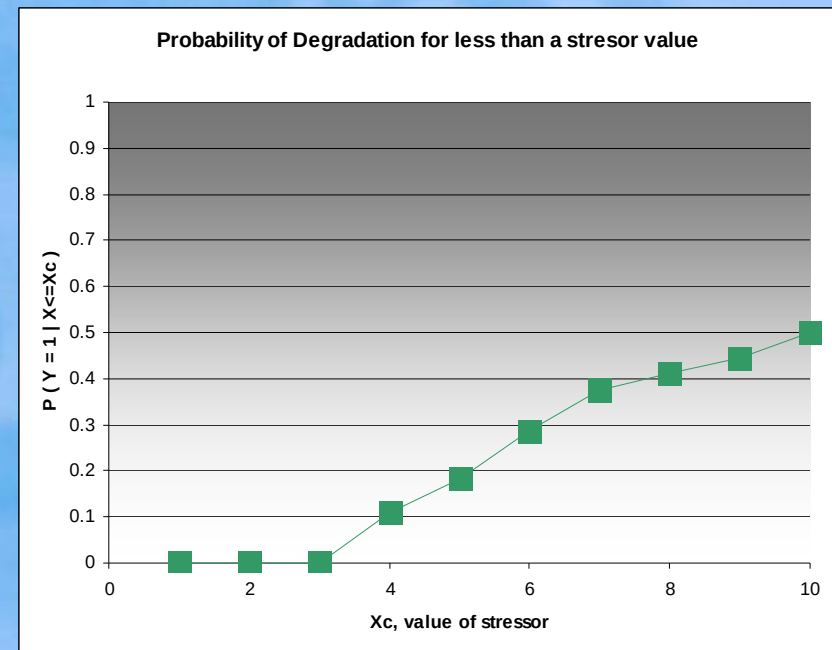


Stressor value (X_c)	Number of healthy sites	Number of degraded sites	Number of degraded sites for $X \leq X_c$	Total number of sites for $X \leq X_c$	$P(Y=1 X \geq X_c)$
1	2	0	10	20	0.5
2	2	0	10	18	0.555556
3	2	0	10	16	0.625
4	2	1	10	14	0.714286
5	1	1	9	11	0.818182
6	1	2	8	9	0.888889
7	0	2	6	6	1
8	0	1	4	4	1
9	0	1	3	3	1
10	0	2	2	2	1

Simple data example:

CPA for stressor value $\leq X_c$

Probability that a site would be degraded if we observe less than a specific stressor value (X_c)



Stressor value (X_c)	Number of healthy sites	Number of degraded sites	Number of degraded sites for $X \leq X_c$	Total number of sites for $X \leq X_c$	$P(Y=1 X \leq X_c)$
1	2	0	0	2	0
2	2	0	0	4	0
3	2	0	0	6	0
4	2	1	1	9	0.111111
5	1	1	2	11	0.181818
6	1	2	4	14	0.285714
7	0	2	6	16	0.375
8	0	1	7	17	0.41185
9	0	1	8	18	0.444444
10	0	2	10	10	0.5

Example with Real Data:

Eutrophication in Northeast Lakes

EMAP sampling of lakes in northeast in 1991-94

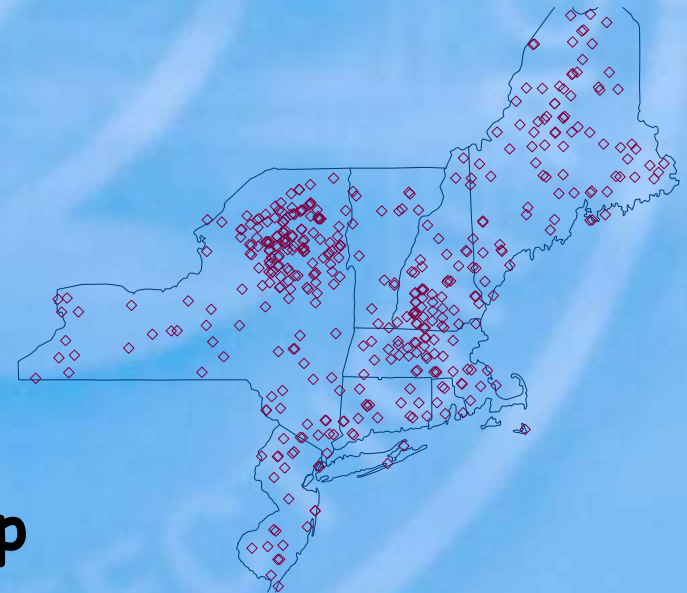
Response variable - chlorophyll a

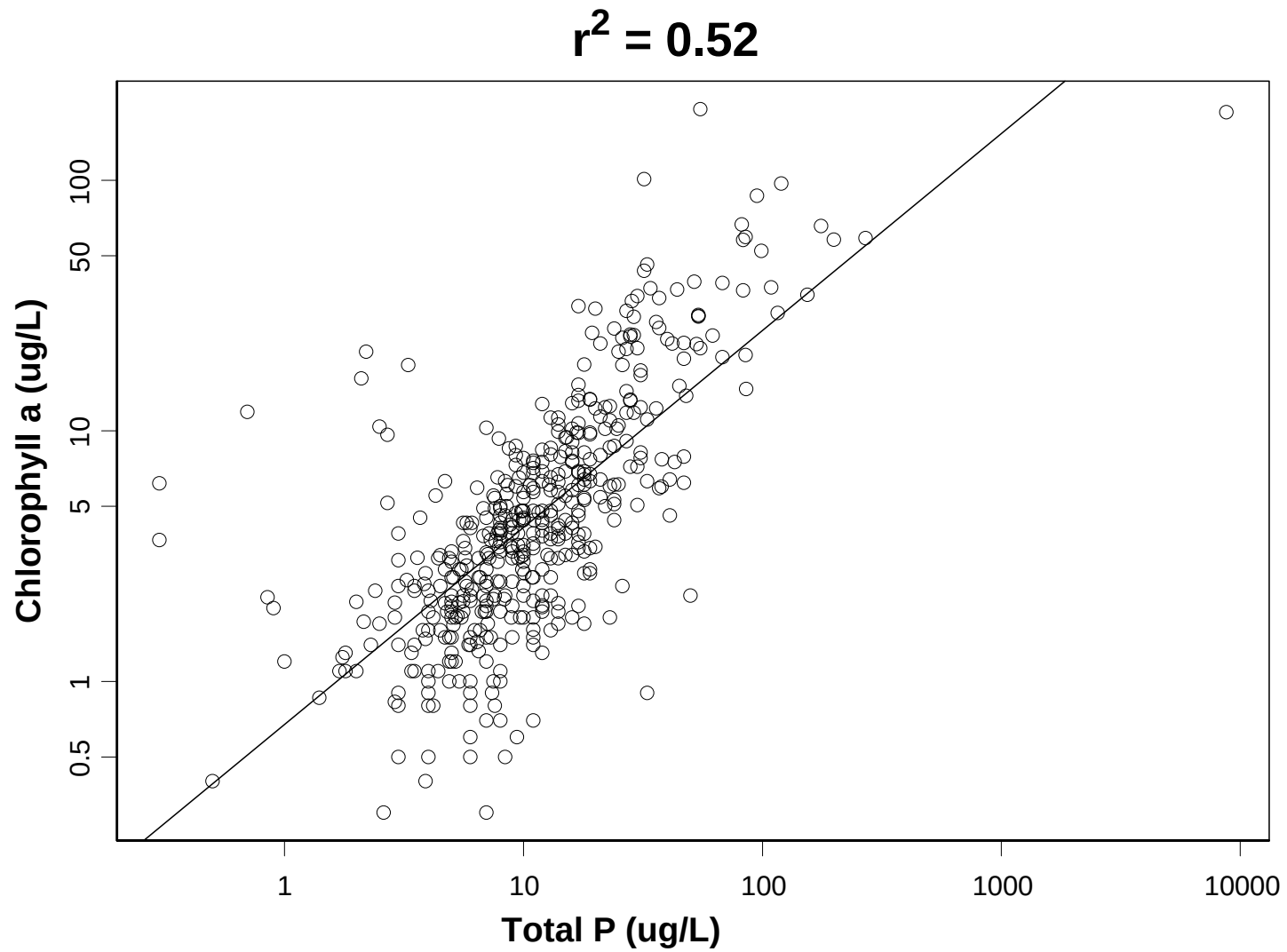
Define eutrophy as $chl a > 30 \text{ ug/L}$

Threshold parameter - phosphorus

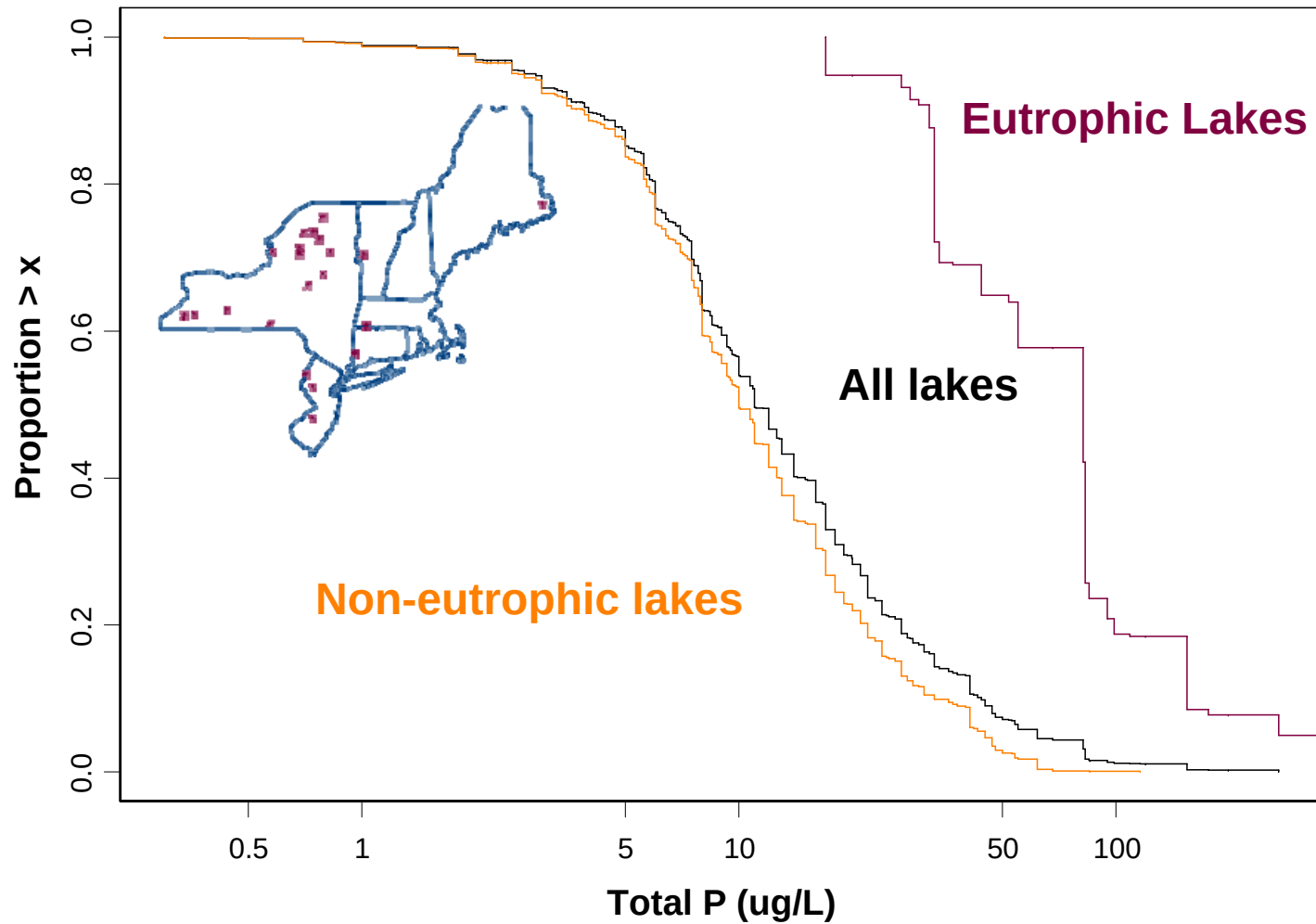
333 lakes sampled

Data available at www.epa.gov/emap

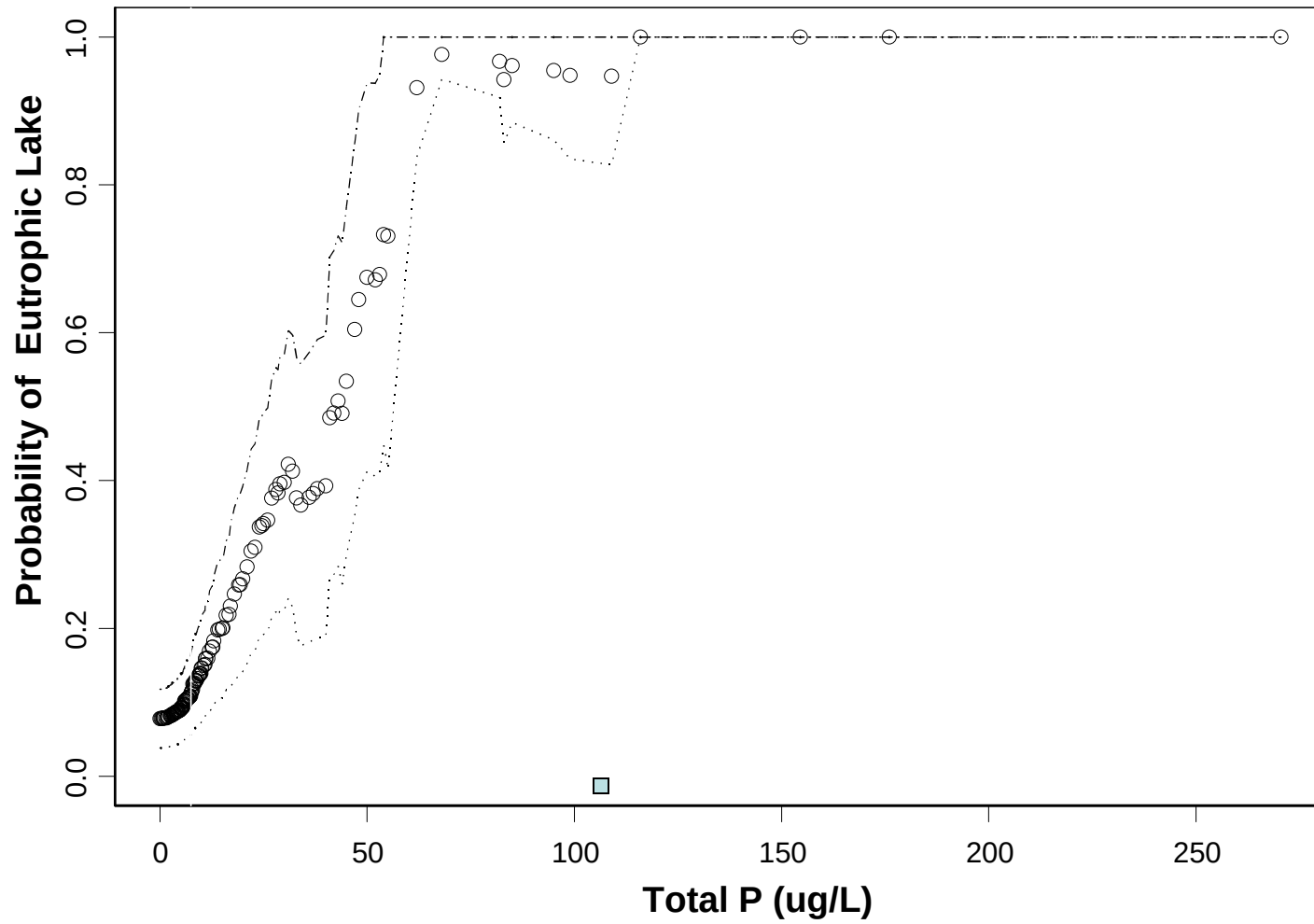




Reverse CDF and CCDFs for Total P



Conditional Probability for Exceeding Total Phosphorus Level



How do we pick off a candidate criterion value from a CPA curve?

The literature calls identifying a change in the curve as changepoint analysis

Changepoint analysis is just the initial step, it is just a statistical test - must check on biological significance of this value

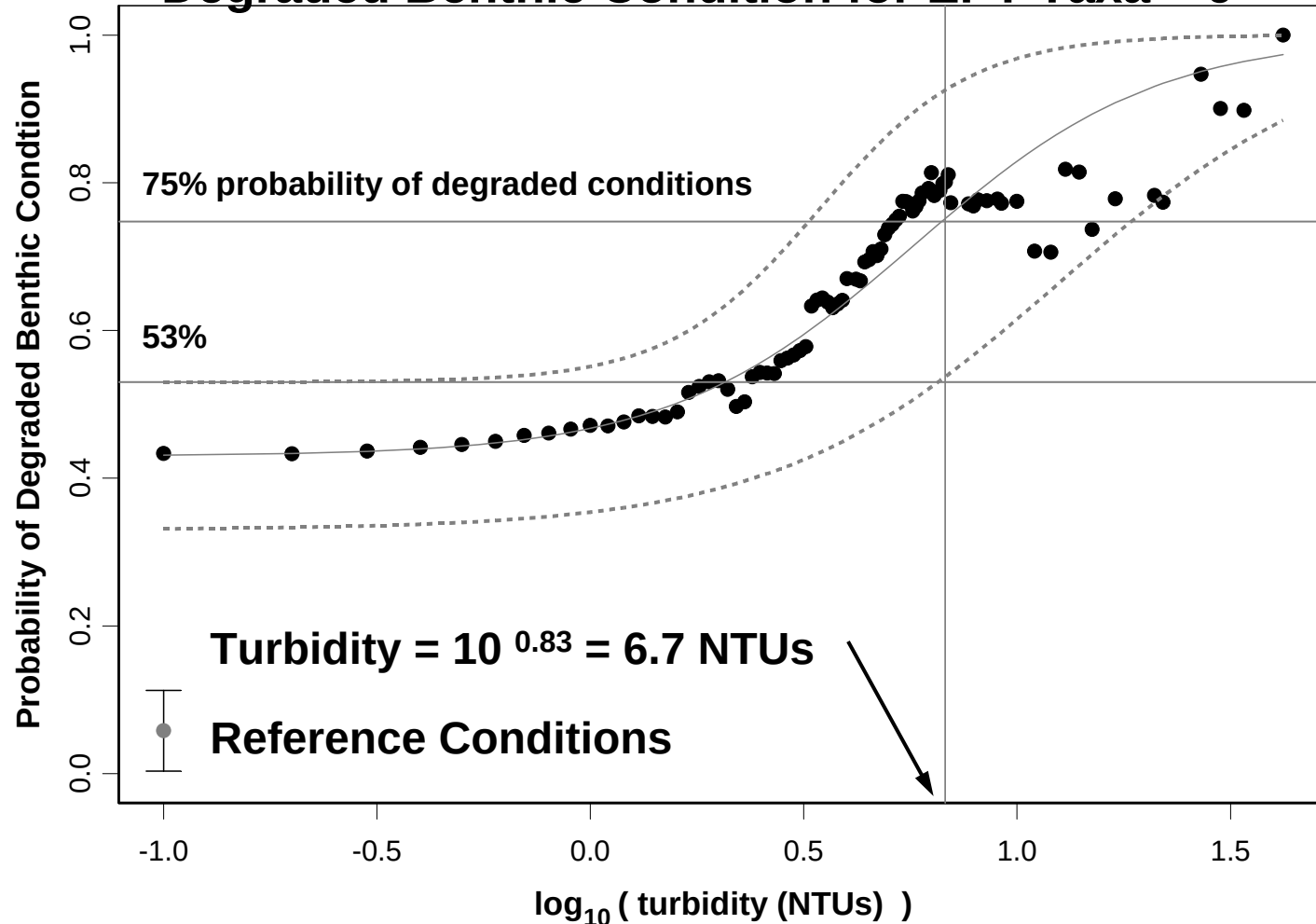
Approaches for Identifying Change point

- non-overlapping confidence intervals
- visual inspection
- nonparametric deviance reduction
- EC_{50} equivalent
- change in curvature of fitted curve

$$P(Y = 1|X > X_C) = \begin{cases} 1 + (D_0 - 1) / (1 + \exp(B_0(X_C - X_0))), & \text{for } X_C > X_0 \\ 1 + (D_0 - 1) / (1 + \exp(B_1(X_C - X_0))), & \text{for } X_C \leq X_0 \end{cases}$$

Conditional Probability Analysis for Turbidity

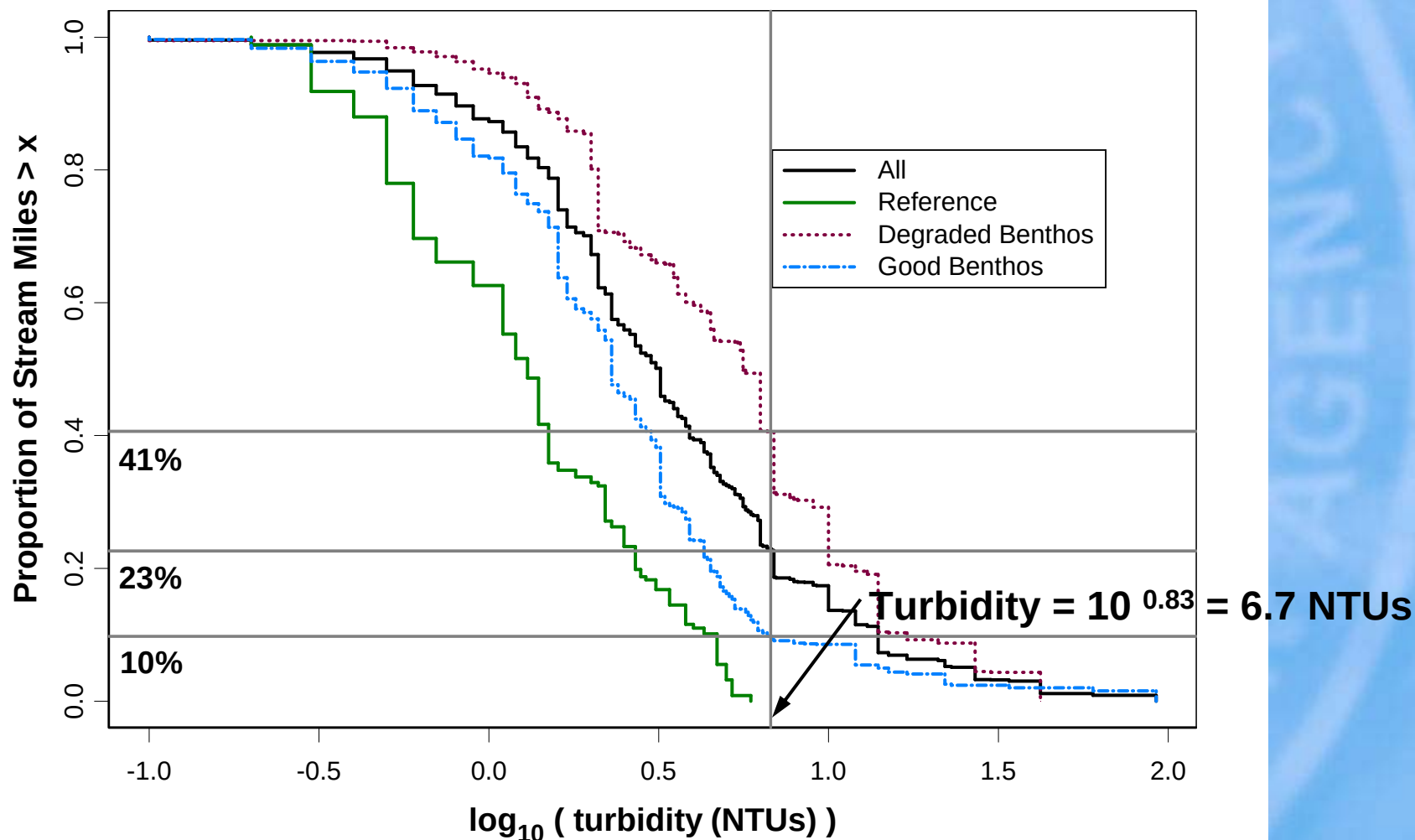
Degraded Benthic Condition for EPT Taxa < 9



For this geographic area, if turbidity levels in streams exceed 6.7 NTUs, probability of observing degraded benthic conditions in those streams is 75%

Distribution of Turbidity Values per Stream Mile

Turbidity Threshold of 6.7 NTUs



What are Minimal Conditions to Apply CPA to a Data Set

- 1) A set of matching data with a stressor metric paired with a response metric
- 2) Application of CPA currently based on a dichotomous response variable, so must identify threshold level for response metric that defines unacceptable conditions.

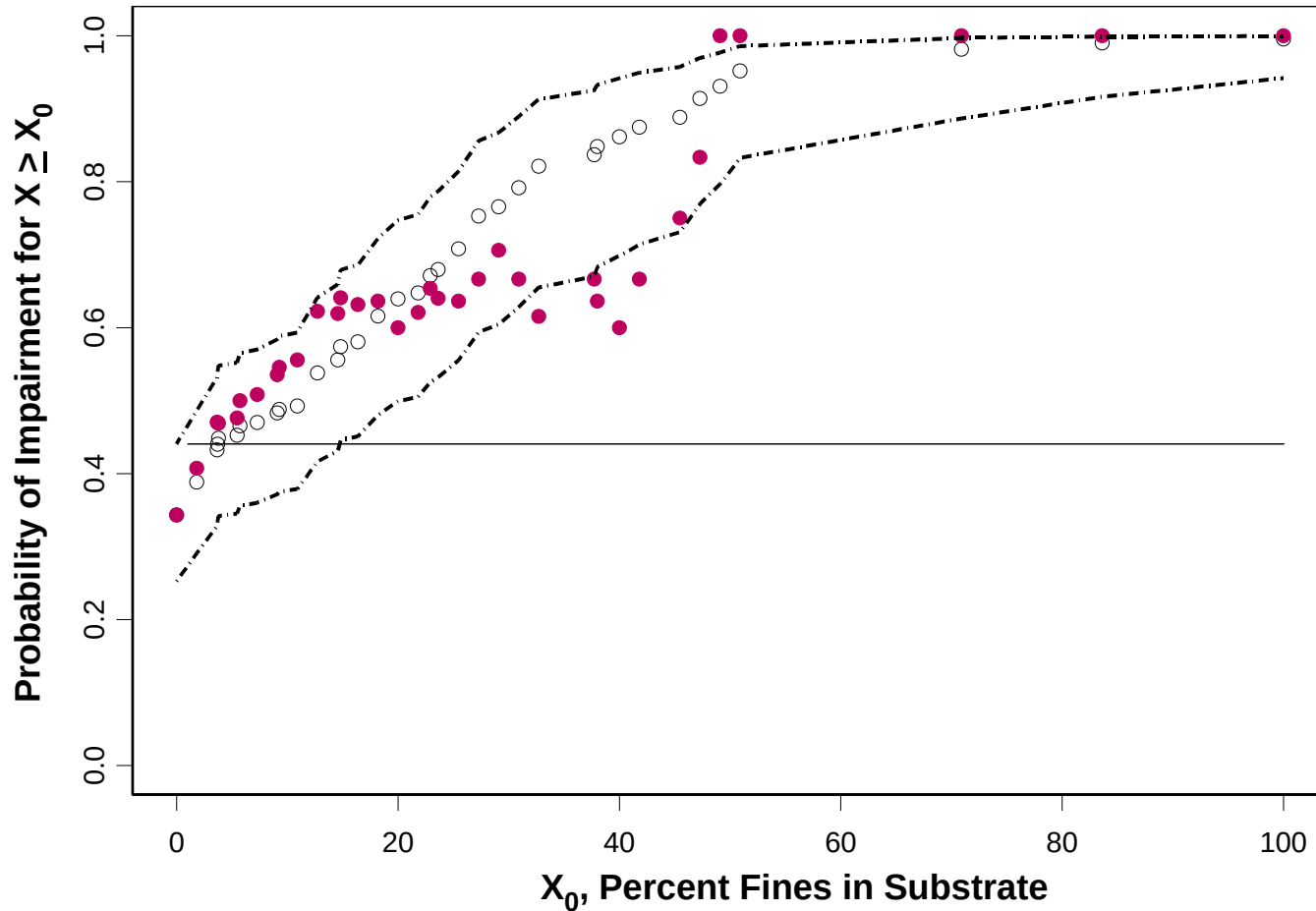
New Ideas in CPA

Traditional approach computes the empirical CPA curve using paired observed stressor and response data

New approach starts with stressor-response relationship, then compute using response values from relationship instead of observed response values

Stressor-response relationship can be generated from other study (e.g., Jay Fields, NOAA, work) or developed from paired data

Analysis of Multiple Stressors?



Important Point to Remember

Conditional probability analysis and changepoint analysis are just statistical tools that can be used to extract very specific information from a data set.

Before applying, it is imperative that extensive exploratory data analysis (EDA) be conducted

EDA is a form of detective work, primarily using graphical depictions of various renditions of the data

Typical Graphical Depictions

Scatter plots (response vs. stressor, stressor vs stressor)

Cumulative Distributions (CDFs &CCDFs)

Data plots on maps (colors, symbols)

Scatter plots with regressions (linear, quantile)

Conditional Probability Plots

ROC curves

For further information:

John F. Paul

paul.john@epa.gov

919-541-3160

USEPA (B343-06)

RTP, NC 27711

Session Questions & Discussion

How does one establish the ecological significance of an identified threshold (changepoint)?

We would like to hear about your experience with conditional probability analysis and changepoint