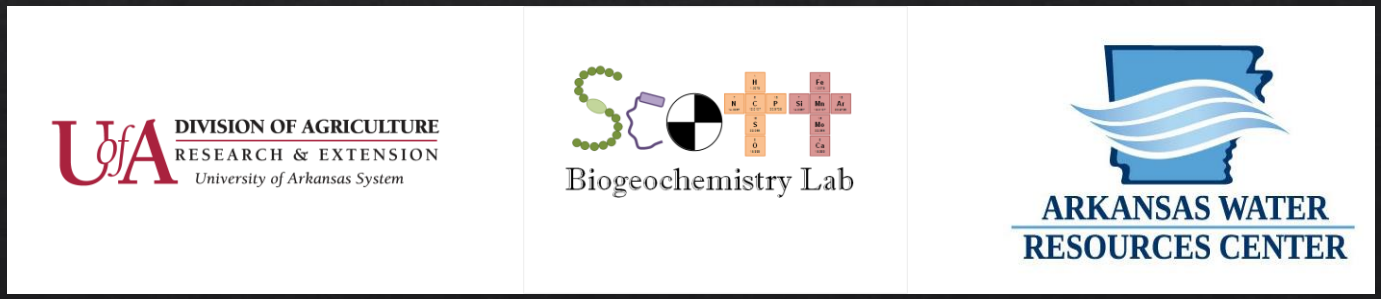
Censored data may obscure low-range nutrient thresholds in reservoirs

Erin M. Grantz, Brian E. Haggard, and J. Thad Scott



Introduction

- Clean Water Action Plan (1998) & (2009) executive directive:
  - Adopt USEPA recommended criteria, or
  - Develop scientifically defensible numeric nutrient criteria at the state level

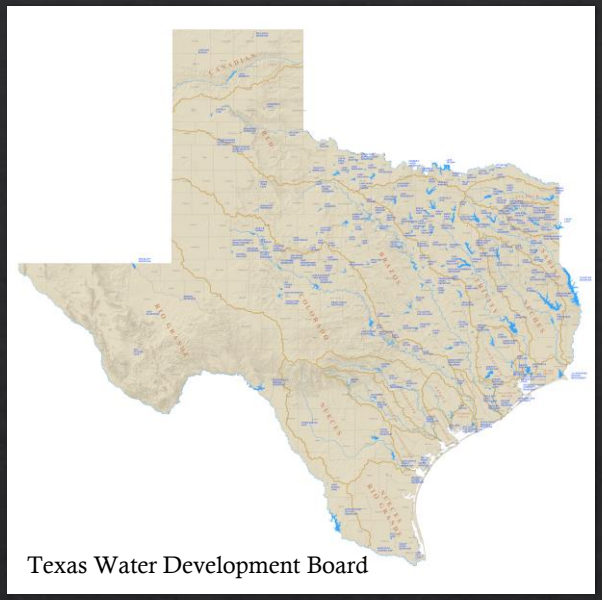
- Obstacles to developing scientifically defensible numeric nutrient criteria:
  - Political, economic, social
  - Data limitations

What are censored data?

- The value of the observation is unknown, except that it falls within a range of possible values
- For concentrations of environmental contaminants, this range is typically between 0 and a quantification limit (QL)
- Quantification limits
  - Detection limit
  - Reporting limit
  - In this study = Minimum concentration meeting desired confidence levels

Dataset attributes

- Texas Commission on Environmental Quality statewide reservoir water quality database
- 764 stations, ~ 100 reservoirs
- Parameters include:
  - Chlorophyll-a (chl-a)
  - Total phosphorus (TP)
  - Secchi transparency
- Common QL's:
  - Chl-a – 10 µg/L
  - TP – 0.060 or 0.050 mg/L
- % Stations with >50% censored data=
  - 40% for TP
  - 22% for chl-a


Texas Water Development Board

Trophic class	Chl-a (ug/L)	TP (mg/L)
Oligotrophic	<2.6	<0.012
Mesotrophic	2.6 – 20	0.012 – 0.024
Eutrophic	20 – 56	0.024 – 0.096
Hypereutrophic	56 – 155+	0.096 – 0.384+

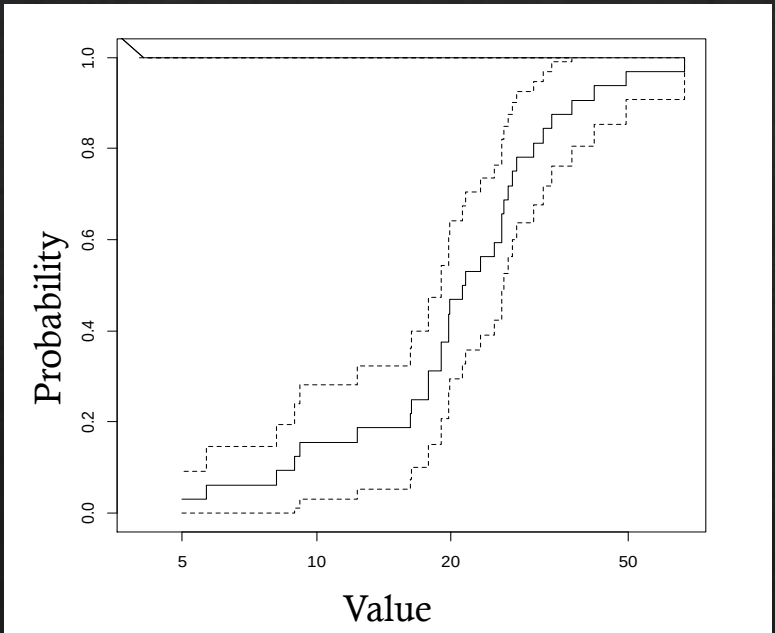
What to do with censored data?: Non-statistical assumptions

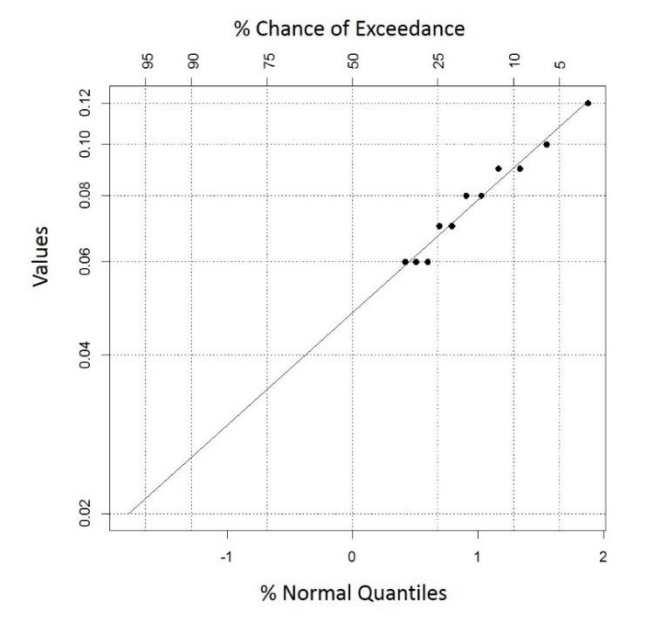
- Delete censored data

- Substitute a value
  - Substitute the QL
  - Substitute ½ the value of the QL
  - Substitute 0

What to do with censored data?: Statistical assumptions

- Kaplan-Meier survival analysis (KM) – non parametric, uses ranking
- Maximum likelihood estimate (MLE) – parametric, uses data distribution
- Regression order statistics (ROS) – parametric, uses data distribution





Statistics for Censored Environmental Data
Helsel (2012)

Do these assumptions affect analytical outcomes?

Objective 1: Compare TP and chl-a reservoir station medians ($n \geq 12$) calculated with different assumptions about censored data

Medians were calculated

1. After substituting the QL for censored observations (Med_{sub}), and
2. By applying statistically-based methodologies (Med_{cen}), using R

Table 2.3.1. Summary of the conditions under which each method for calculating summary statistics in datasets with censored observations is preferred. Adapted from Helsel (2012).

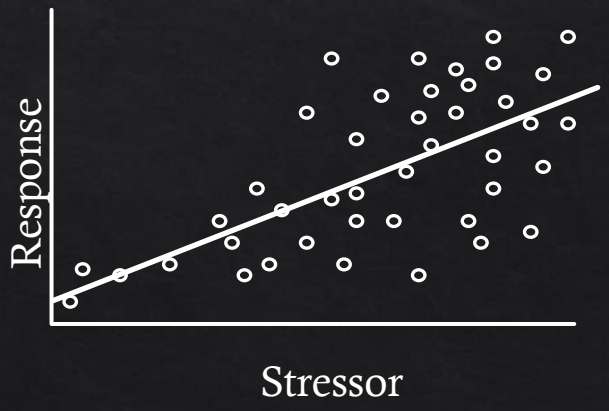
	Amount of Available Data	
Percent Censored	<50 Observations	>50 Observations
< 50% censored	Kaplan-Meier	Kaplan-Meier
50-80% censored	Regression order statistics	Maximum likelihood estimate
>80% censored	Not recommended	Not recommended

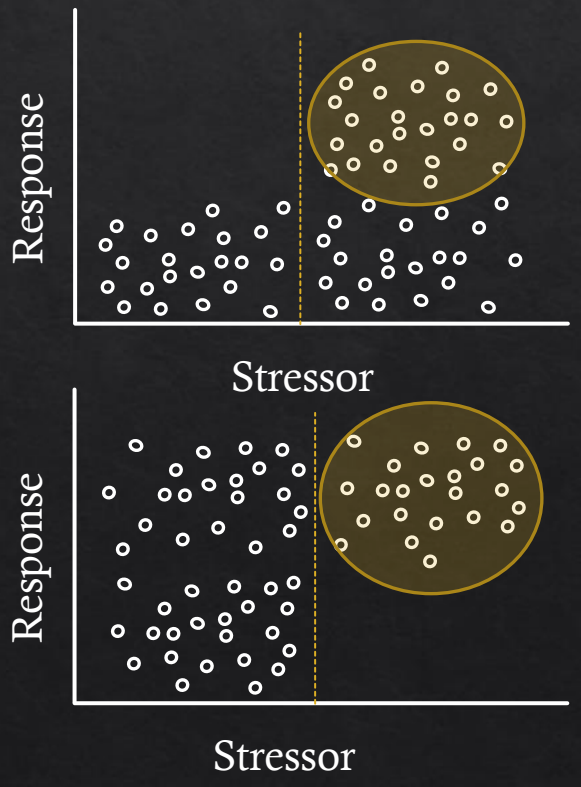
Comparing medians

- Median comparison metric = % difference between station medians

$$\%Diff = \frac{(Med_{sub} - Med_{cen})}{Med_{sub}} \times 100\%$$

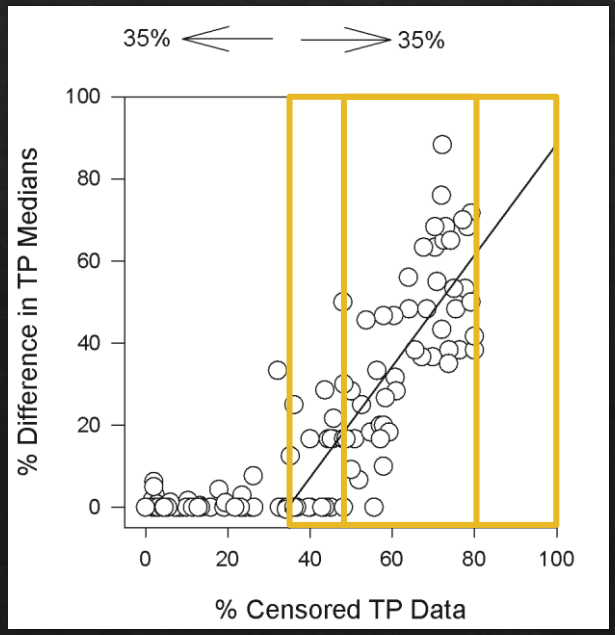
- %Diff vs. Censored data frequency:
  1. Non-parametric changepoint analysis (King & Richardson 2003)
  2. Linear regression analysis

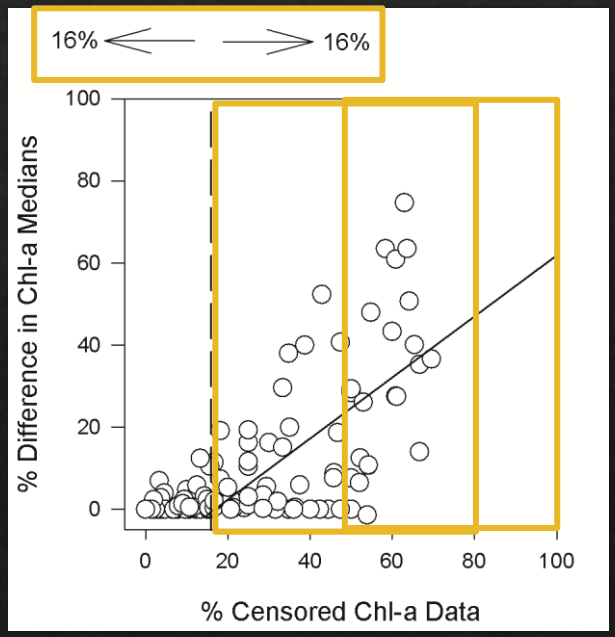




Different assumptions = different medians

- Difference seen as low as 16% censored data
- Most common effects, biggest differences at > 50% censored data
- Linear increase in %Diff between threshold & 80% censored data





Do assumptions about censored data affect analytical outcomes?

Objective 2: Calculate chl-a, TP, and Secchi transparency station medians calculated using four approaches to handling censored data & compare TP thresholds identified using changepoint analysis

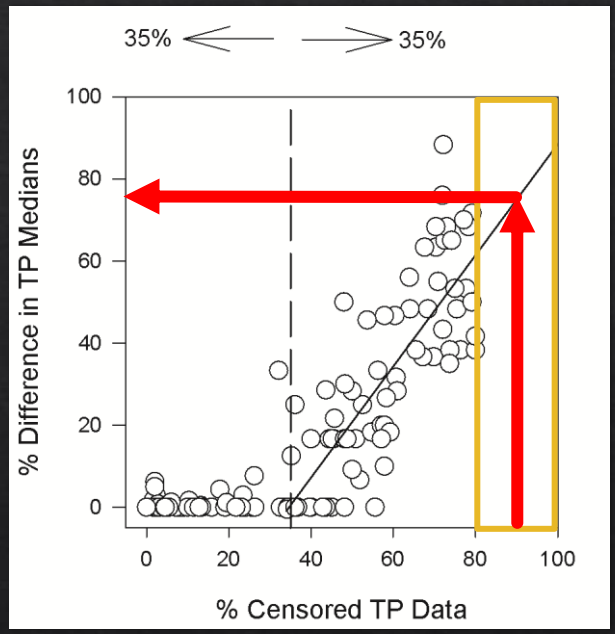
1. Substitute QL
2. Substitute ½QL
3. Statistical methodologies to estimate measures of central tendency (0-80% censored data)
4. Hybrid method (statistical methodologies 0-80% & substitute values from linear regression model for >80% censored data).

Estimating medians for stations with >80% censored data

1. Regression model to project a median difference

2. Use projection + station Med_{sub} to estimate a median with censored data correction

$$\%Diff = m \times \%Censored + b$$

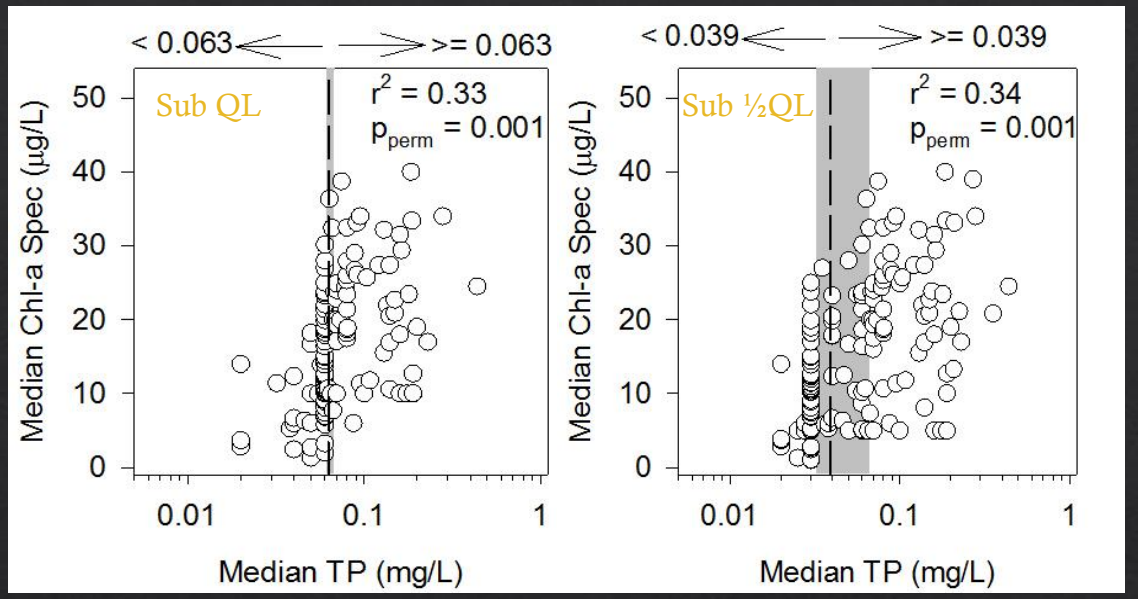


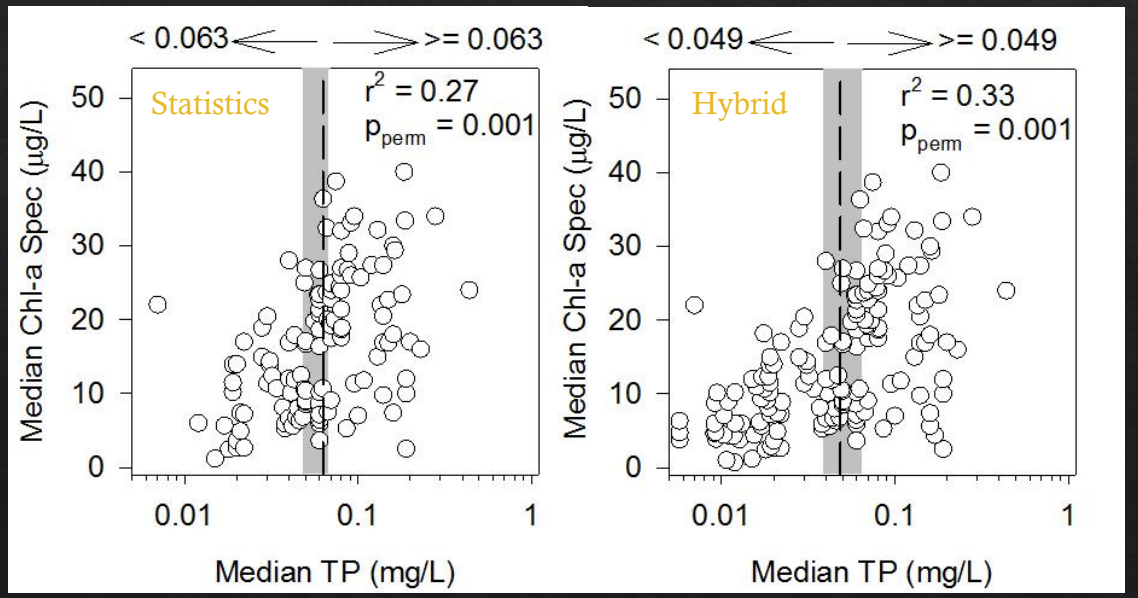
$$\%Diff = \frac{(Med_{sub} - Med_{cen})}{Med_{sub}} \times 100\%$$

$$Med_{cenR} = Med_{sub} - \frac{(Med_{sub} \times \%Diff)}{100\%}$$

Results: Chl-a vs. TP

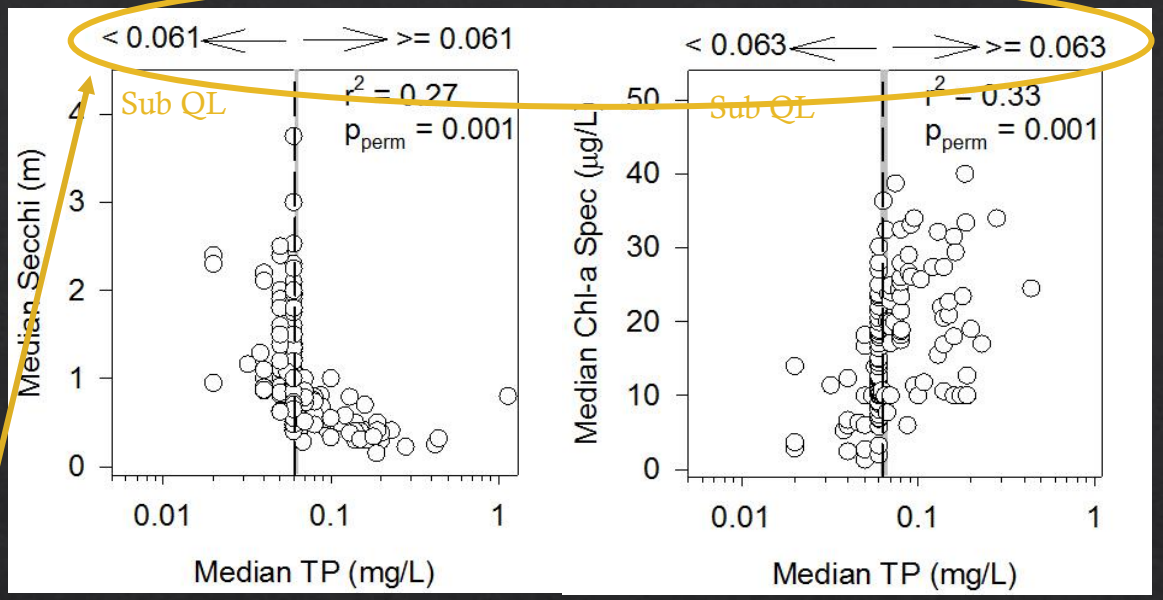
- TP thresholds = 0.039 – 0.063 mg/L
- Thresholds differed with subbed values
- Statistical dataset threshold in agreement with subbed QL dataset
- Hybrid dataset threshold not equal to statistical dataset, but close

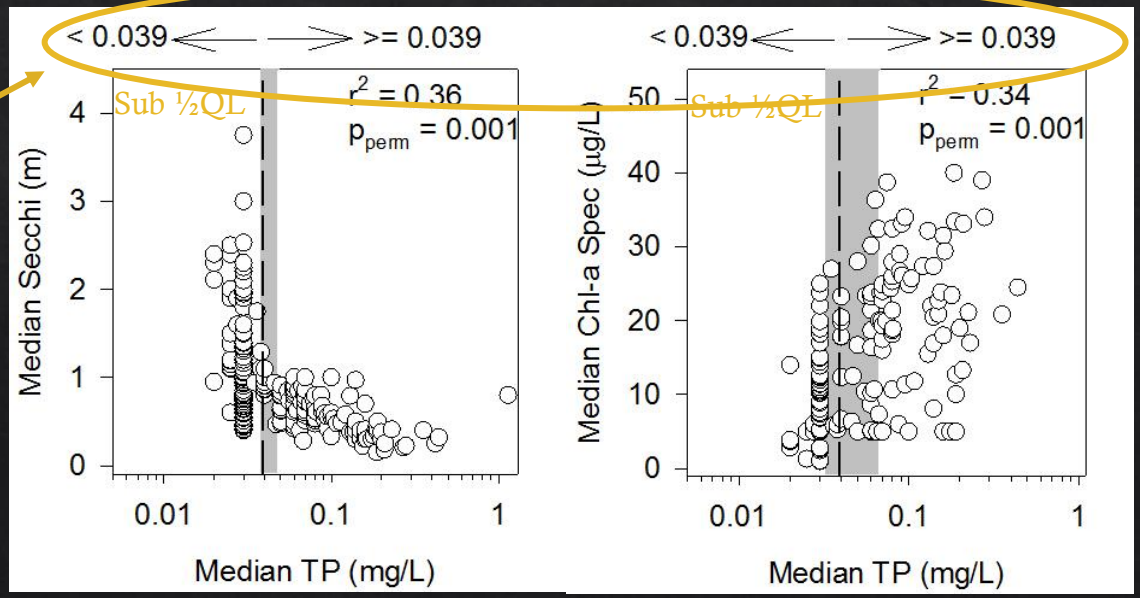




Results: Secchi vs. TP

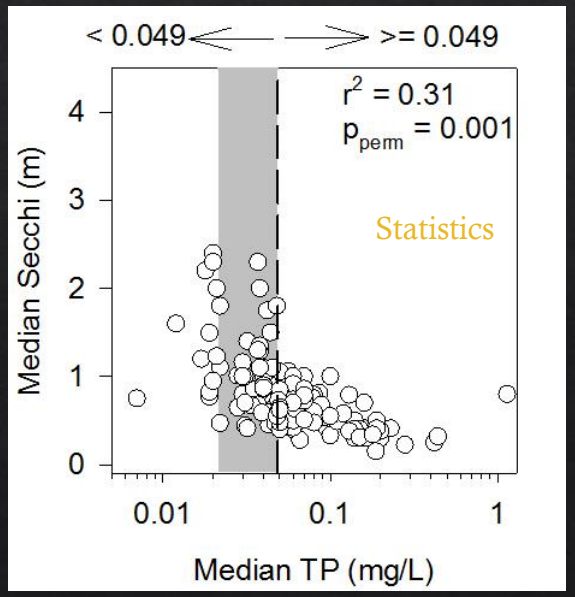
- TP thresholds = 0.025 – 0.061 mg/L
- Thresholds differed with subbed values
- Sub dataset thresholds for Secchi almost identical to those for Chl-a….

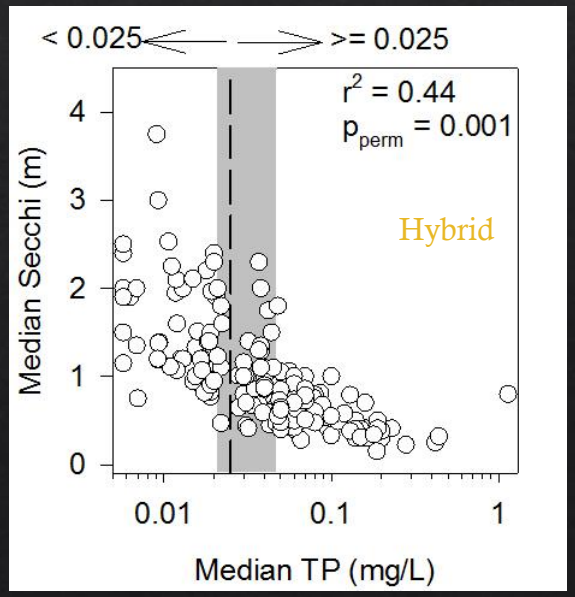




Results: Secchi vs. TP

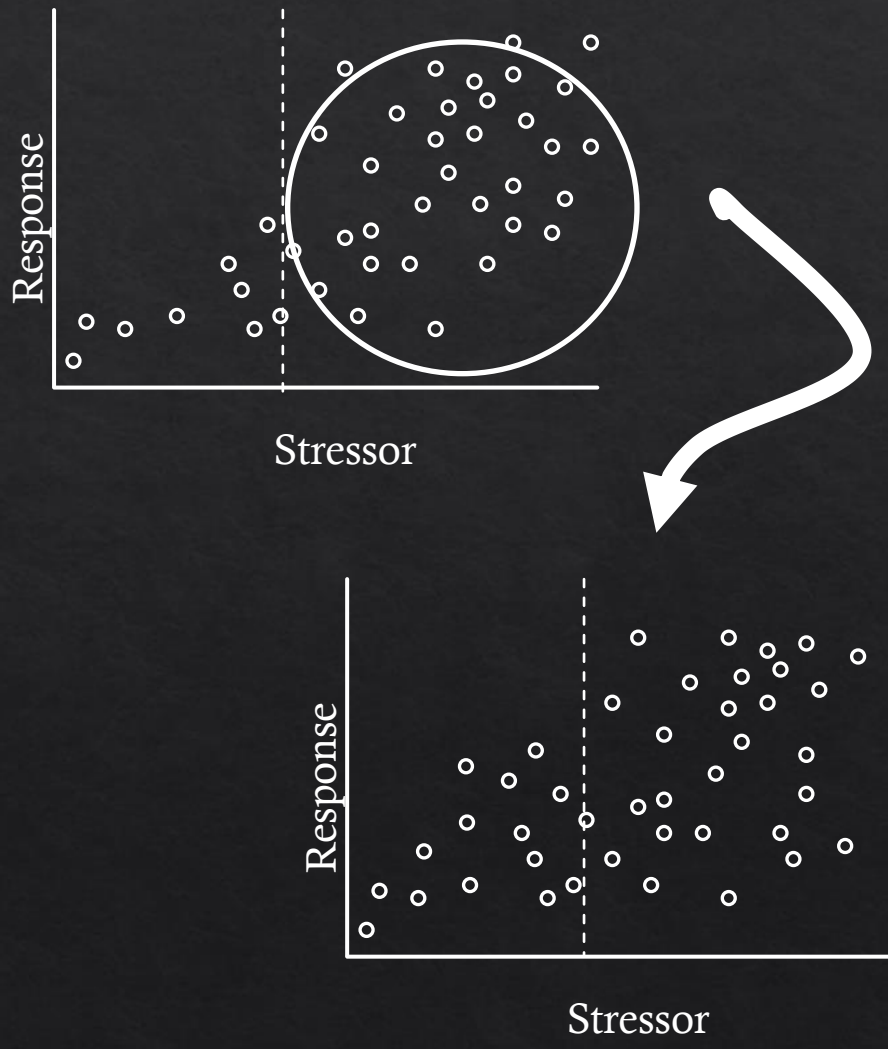
- Mid-range TP threshold for statistical medians dataset
- BUT, much lower threshold identified in hybrid dataset…





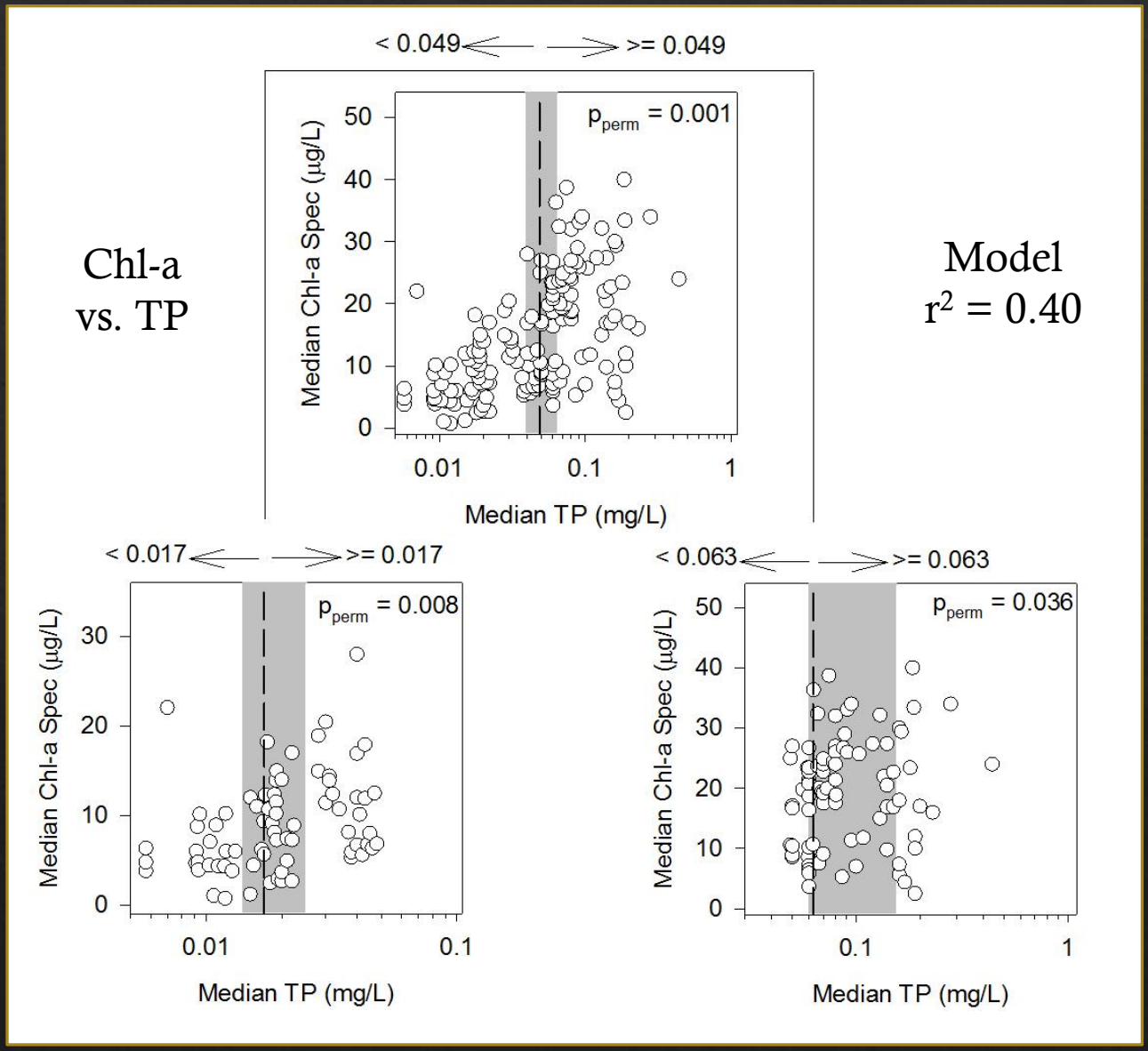
Further analysis of hybrid method dataset

- Threshold relationships may also be hierarchical
- Classification and regression tree analysis (CART; De'ath and Fabricius 2000)



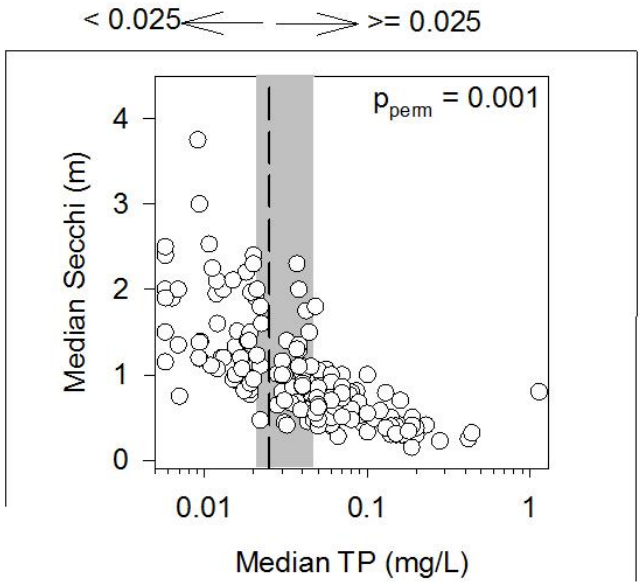
Added complexity in hybrid data models

- Multiple TP thresholds found for Chl-a response
- Low threshold = 0.017 mg/L
- High threshold = 0.063 mg/L



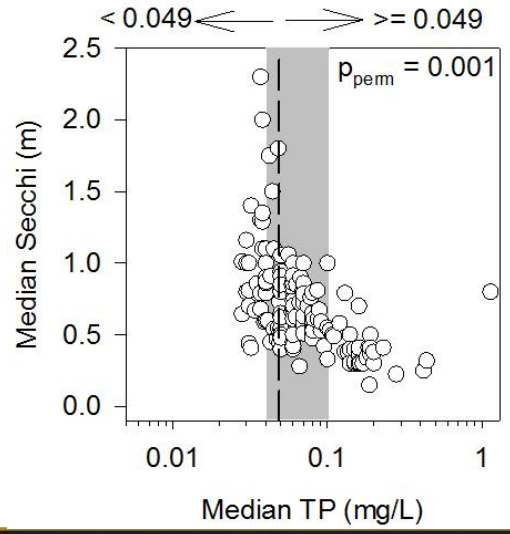
Added complexity in hybrid data models

- Multiple TP thresholds for Secchi response
- High threshold = 0.049 mg/L



Secchi vs. TP

Model $r^2 = 0.50$

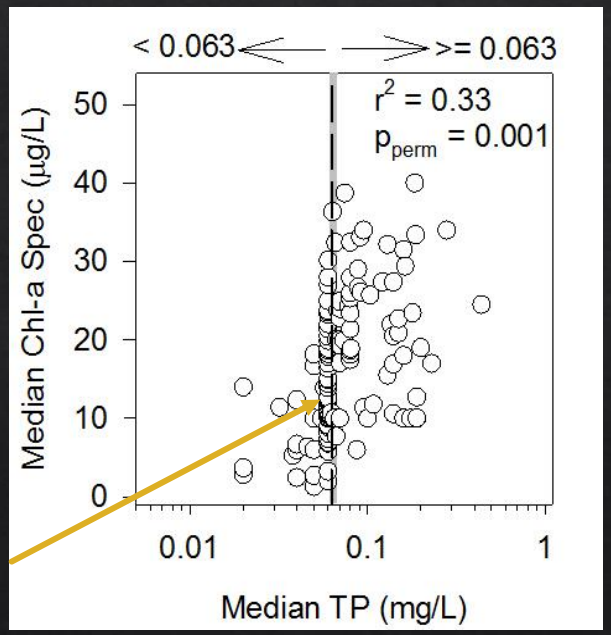


Discussion

1. The assumptions we make about censored observations affect analytical outcomes

Substitution can introduce spurious trends

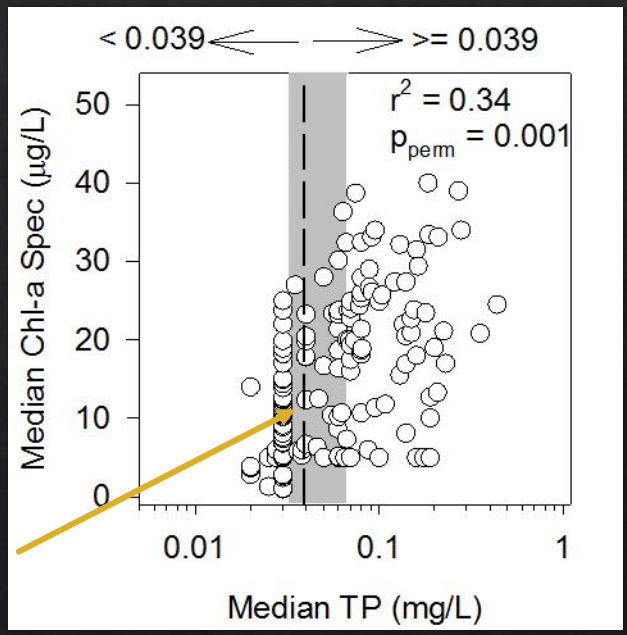
- These data are identical, except for assumptions about censored data!
- Inserting a single value for a large number of observations problematic



0.060 mg/L

Substitution with QL

VS.

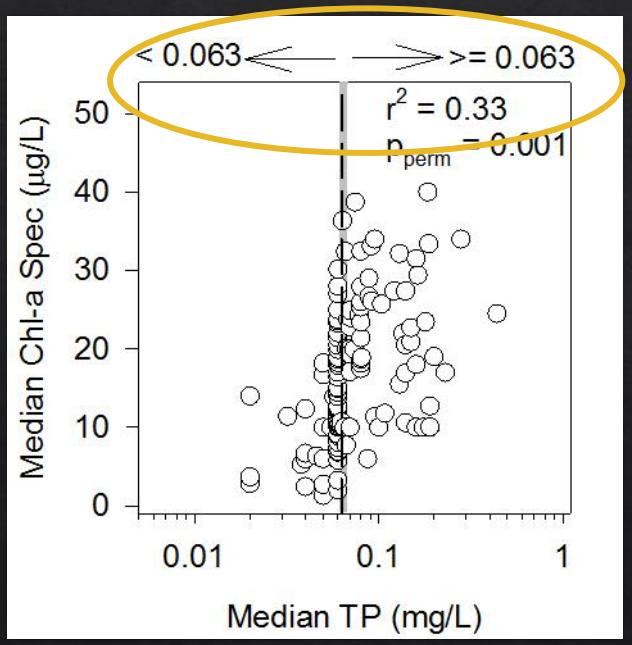


0.030 mg/L

Substitution with ½ QL

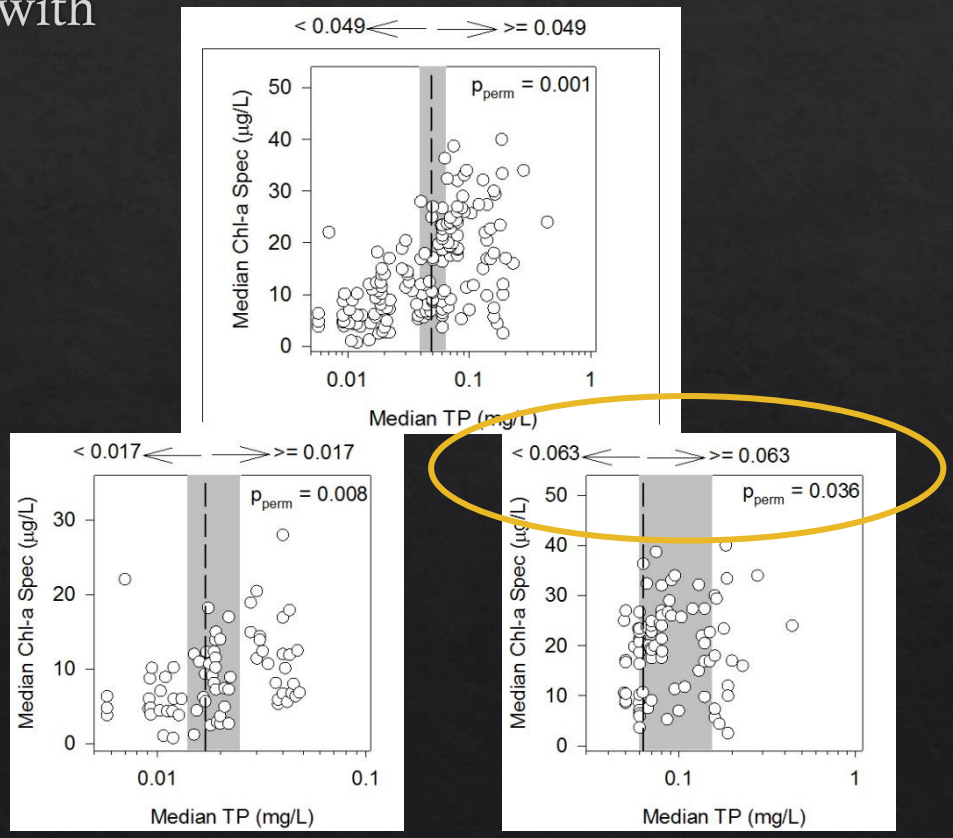
Substitution can inflate weak trends

- TP threshold = 0.063 mg/L identified in multiple median datasets
- But not a primary threshold & with much lower explanatory power



Substitution with QL

VS.



Hybrid data

Discussion

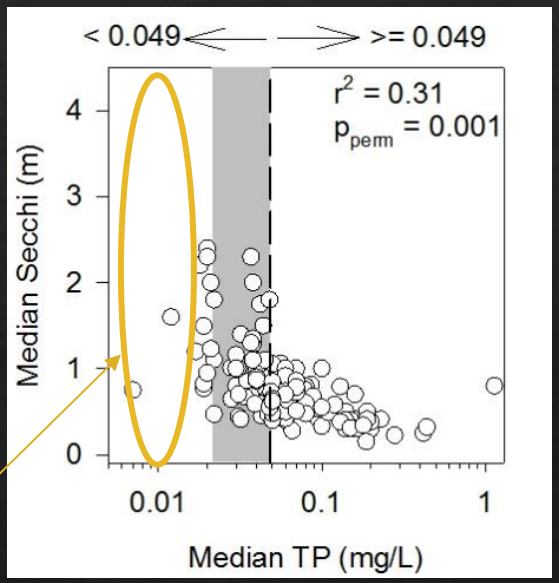
1. The assumptions we make about censored observations affect analytical outcomes

2. Highly censored datasets with high QL's limit utility of even best practice methods for censored data analysis

Low-range & hierarchical thresholds obscured

- Max correction for censoring yielded lowest thresholds & hierarchy
- Not detectable when
  1. spurious trends are introduced with substitution, or
  2. information from sites with >80% censoring was excluded

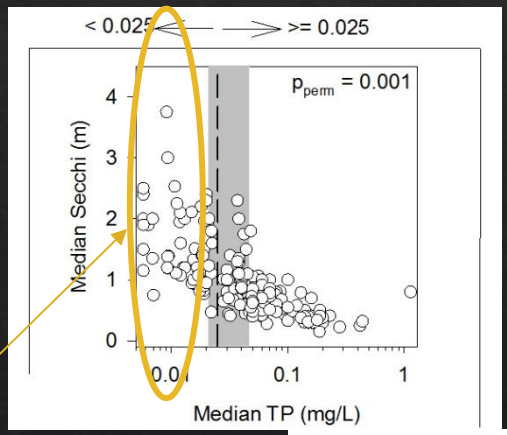
Statistical methods



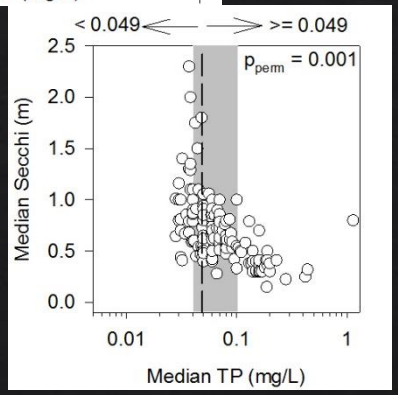
No data!

VS.

Data!



Hybrid data



We need better data!

- Multiple answer to the same question = uncertainty for lawmakers, regulators, & stakeholders

- We can achieve this goal…
  - Best practices should be used in collecting, analyzing, and documenting water quality data
  - Concentrations already identified as environmentally relevant should be considered in selecting analytical methods with relevant QL's

Acknowledgements

Texas Commission on Environmental Quality – funding and data source

TCEQ Project Team, Julie Mcentire & Jill Csekitz

Arkansas Water Resources Center support staff

Scott Biogeochemistry Lab

My coauthors, Thad Scott & Brian Haggard