HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Classification/Cluster-Based ML Approaches to Investigate Groundwater Contamination at Coal Ash Dumps

Presenters: Antonella Basso, José Lopez, Tony Ni
Faculty Mentor: Dr. Rachel Nethery
Graduate Student Mentor: Luli Zou

AGNES SCOTT COLLEGE
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Amherst College

Outline

1. Background
2. Data
3. Research Question
4. Exploratory Analysis
5. Classification
6. Clustering
7. Bootstrapping
8. Conclusion

Background

Background

The gravity of the situation:

- The U.S. produces around 100 million tons of coal ash every year.
  - Nearly 130 million tons of coal ash was generated in 2014
- Reckless coal ash disposal in landfills and waste ponds has gone unchecked. Until recently . . . sort of.

Background

Power companies have been reckless in the disposal of coal ash.

- Poor coal ash management has led to major spills time and time again.

- Long term, disposal of coal ash into ponds and landfills resulting in groundwater contamination will have most significant impact.

Background

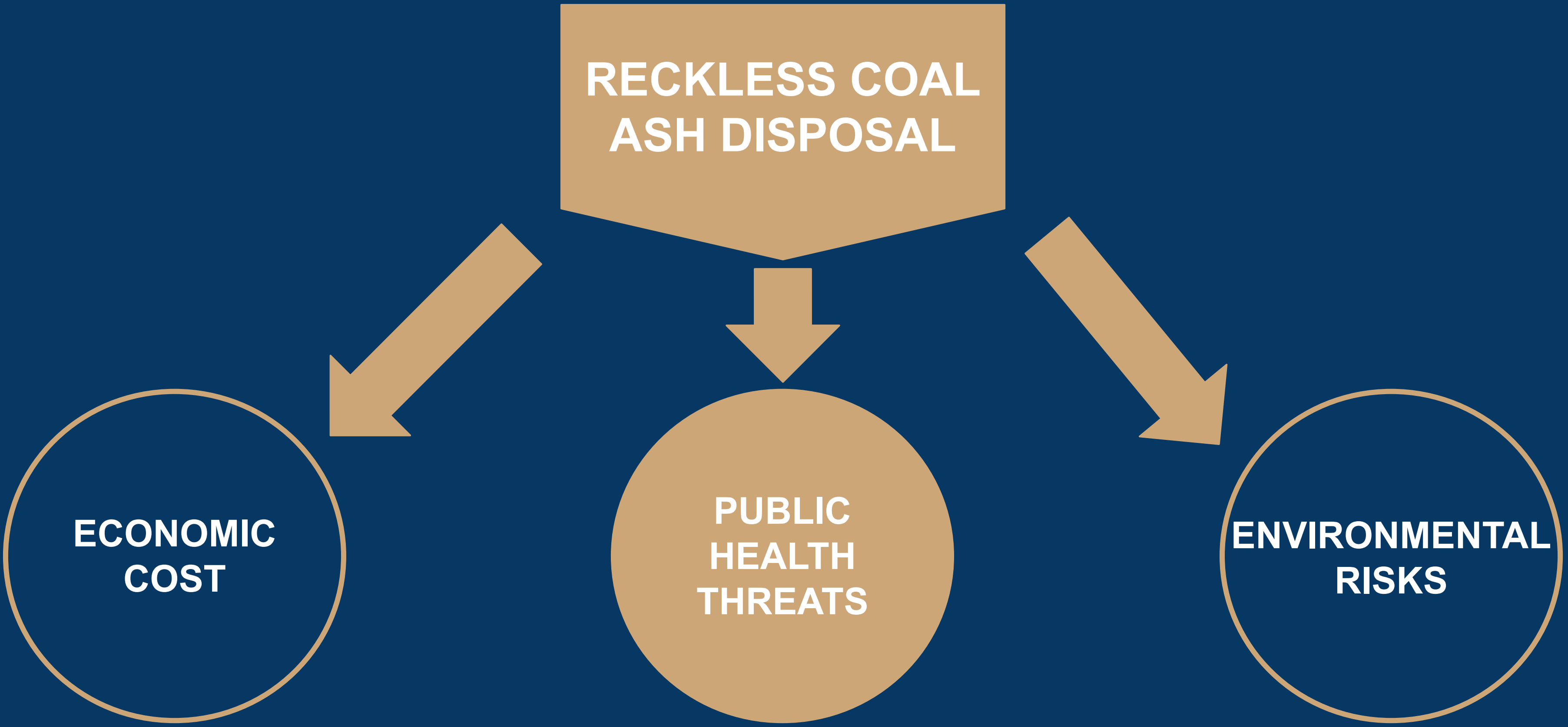
What's wrong with coal ash contaminating groundwater? Well . . . there is a long list of toxic pollutants in high concentrations in coal ash.

- Arsenic
- Boron
- Cadmium
- Cobalt
- Chromium
- Fluoride
- Lead
- Lithium
- Mercury
- Molybdenum
- Radium
- Selenium
- Thallium

Background


RECKLESS COAL ASH DISPOSAL
ECONOMIC COST
PUBLIC HEALTH THREATS
ENVIRONMENTAL RISKS

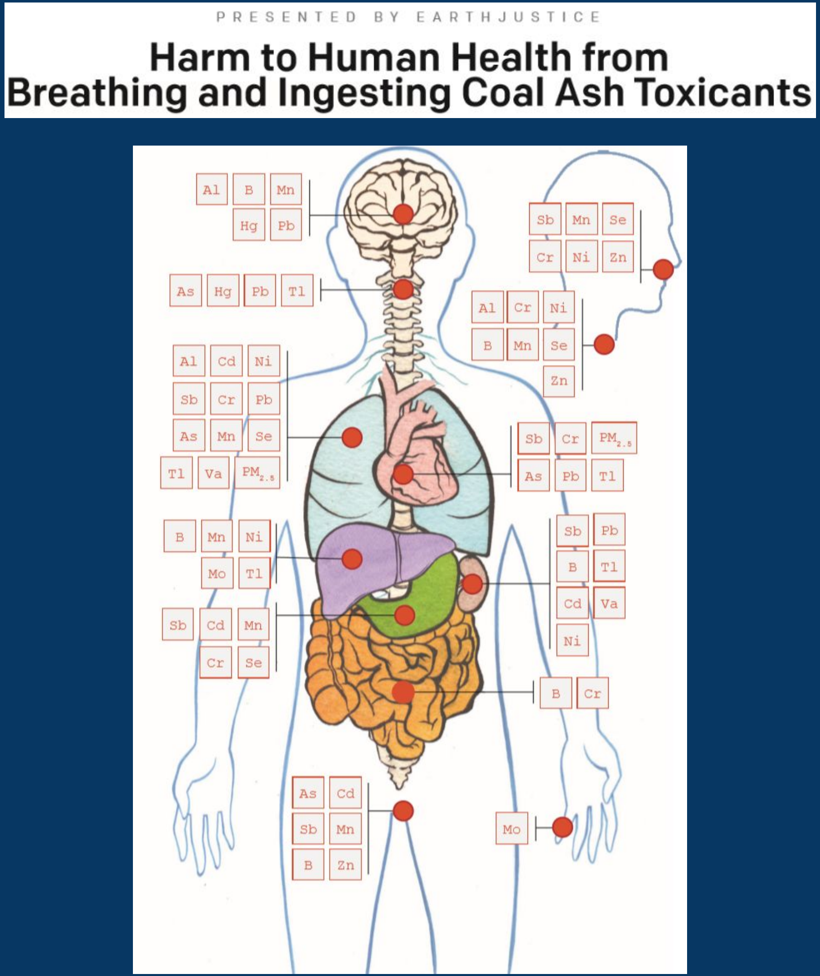
Background


RECKLESS COAL
ASH DISPOSAL
ECONOMIC
COST
PUBLIC
HEALTH
THREATS
ENVIRONMENTAL
RISKS

Background

The general human health risks of coal ash toxic contaminants:

- Cancer
- Heart disease
- Reproductive failure
- Stroke
- Child neurological impairments



Data

Data: Introduction

- EIP's Coal Ash Rule groundwater monitoring results database
- 38,792 groundwater samples
- 198 upgradient/downgradient wells
- 18 sites in Illinois
- 21 contaminants

Data: Format

- Variables:
  - State
  - Site
  - Disposal Area
  - Type
  - Well ID
  - Gradient
  - Sample Date
  - Contaminant
  - Measurement Unit
  - Concentration
- Samples

Before

	Unnamed: 0	state	site	disposal.area	type	well.id	gradient	samp.date	contaminant	measurement.unit	concentration
0	1	IL	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	11/27/17	Total Dissolved Solids	mg/l	2960.0
1	2	IL	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	6/20/17	Total Dissolved Solids	mg/l	2850.0
2	3	IL	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	7/25/17	Total Dissolved Solids	mg/l	2830.0
3	4	IL	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	12/27/16	Total Dissolved Solids	mg/l	2780.0
4	5	IL	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	6/23/16	Total Dissolved Solids	mg/l	2730.0

After

	site	disposal	type	well	gradient	contaminant	unit	concentration
0	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	Total Dissolved Solids	mg/l	2960.0
1	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	Total Dissolved Solids	mg/l	2850.0
2	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	Total Dissolved Solids	mg/l	2830.0
3	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	Total Dissolved Solids	mg/l	2780.0
4	Baldwin Energy Complex	Baldwin Bottom Ash Pond	SI	MW-370	Downgradient	Total Dissolved Solids	mg/l	2730.0

Data Wrangling

contaminant	Antimony	Arsenic	Barium	Beryllium	Boron	Cadmium	Calcium	Chloride	Chromium	Cobalt	...
well											
03R	0.0010	0.001000	0.063700	0.001000	1.318444	0.001113	85.388889	70.555556	0.001075	0.001012	...
05DR	0.0010	0.001000	0.057425	0.001000	1.275556	0.001000	80.033333	77.777778	0.001000	0.001462	...
05R	0.0010	0.001175	0.058175	0.001000	1.433111	0.001313	76.588889	76.222222	0.001113	0.001087	...
08D	0.0010	0.001000	0.168250	0.001000	0.122222	0.001188	205.555556	265.666667	0.001000	0.013663	...
12	0.0010	0.001000	0.048438	0.001000	0.445889	0.001000	72.033333	72.777778	0.001000	0.001000	...
...	...	...	...	...	...	...	...	...	...	...	...

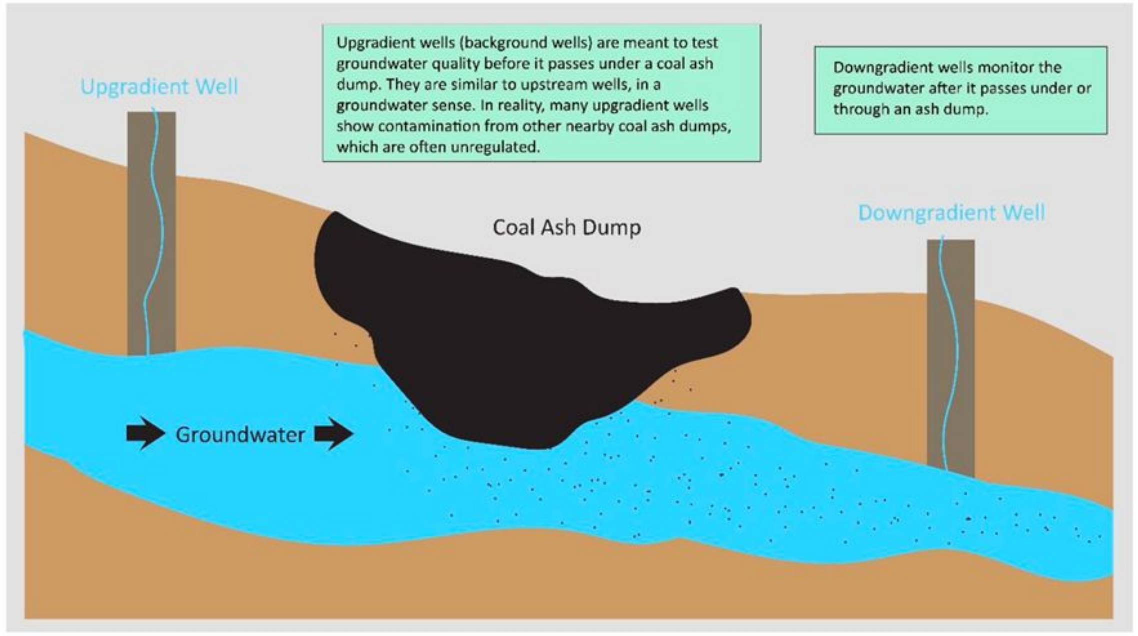
Research Questions

Research Questions

- About how many upgradient wells are contaminated?
- Can we identify contaminated upgradient wells?

Research Questions

- How can we correct contaminant measurements if upgradient wells are contaminated?

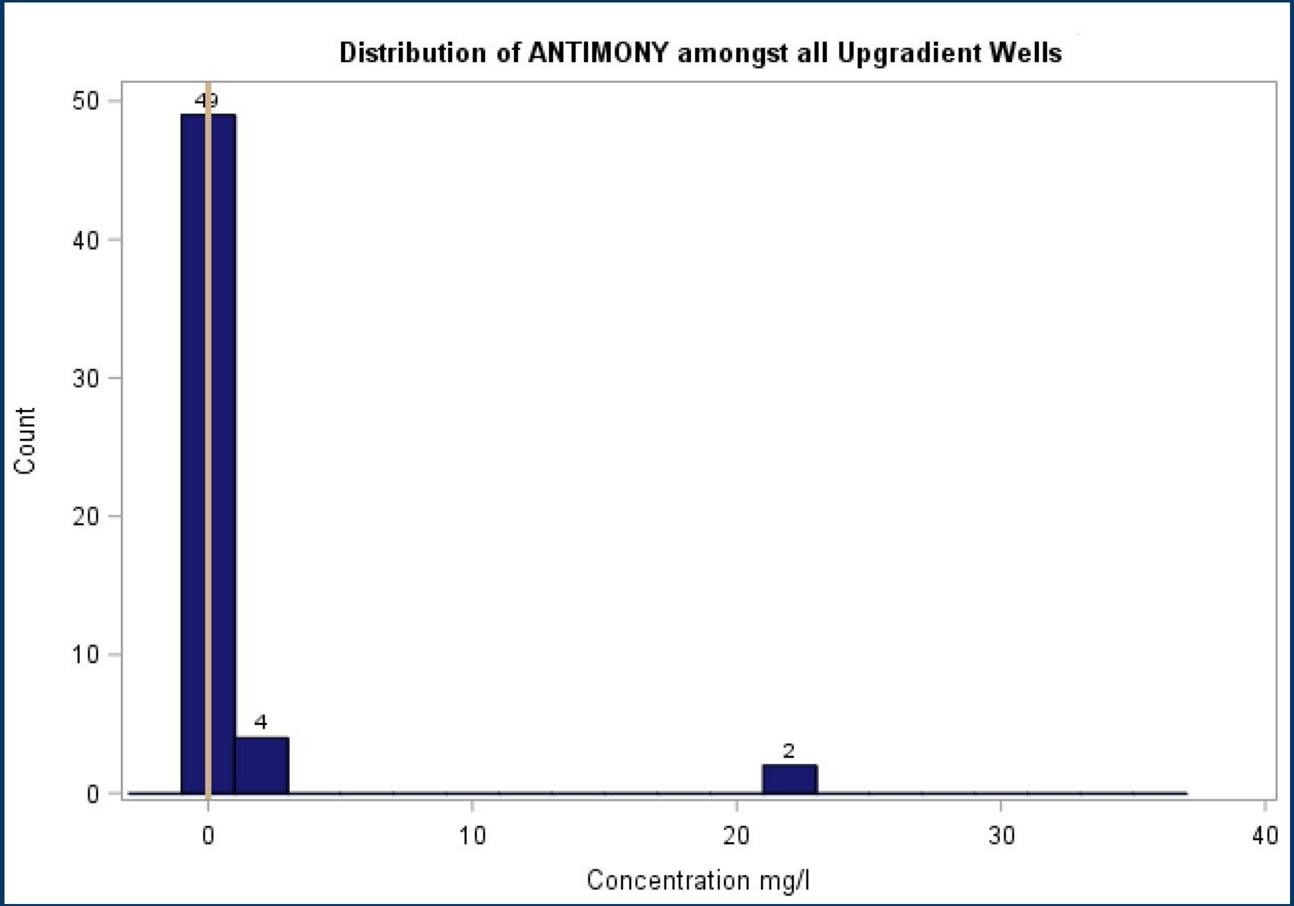


Exploratory Data Analysis

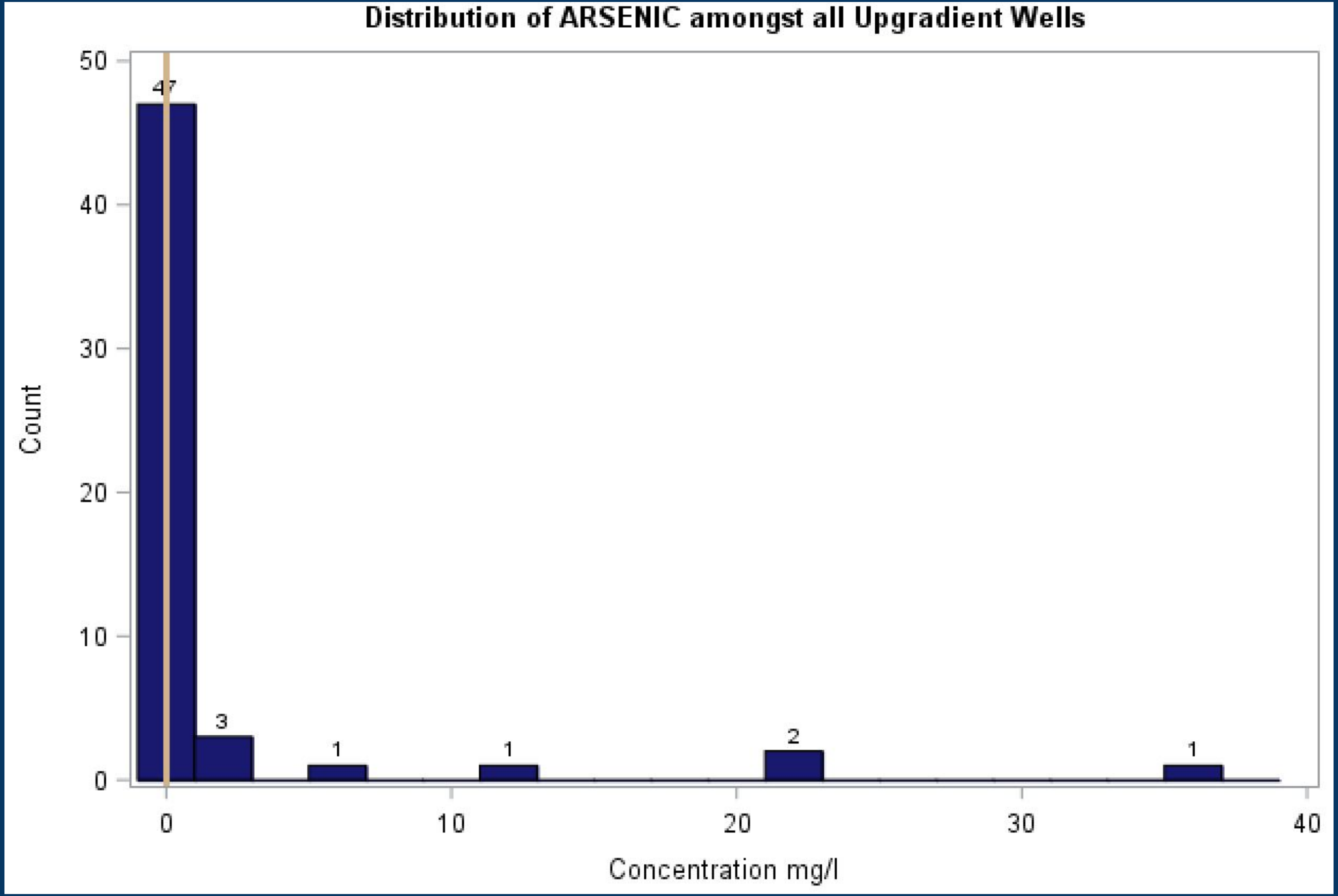
Exploratory Data Analysis

Table B1: Groundwater monitoring pollutants and thresholds used in this report

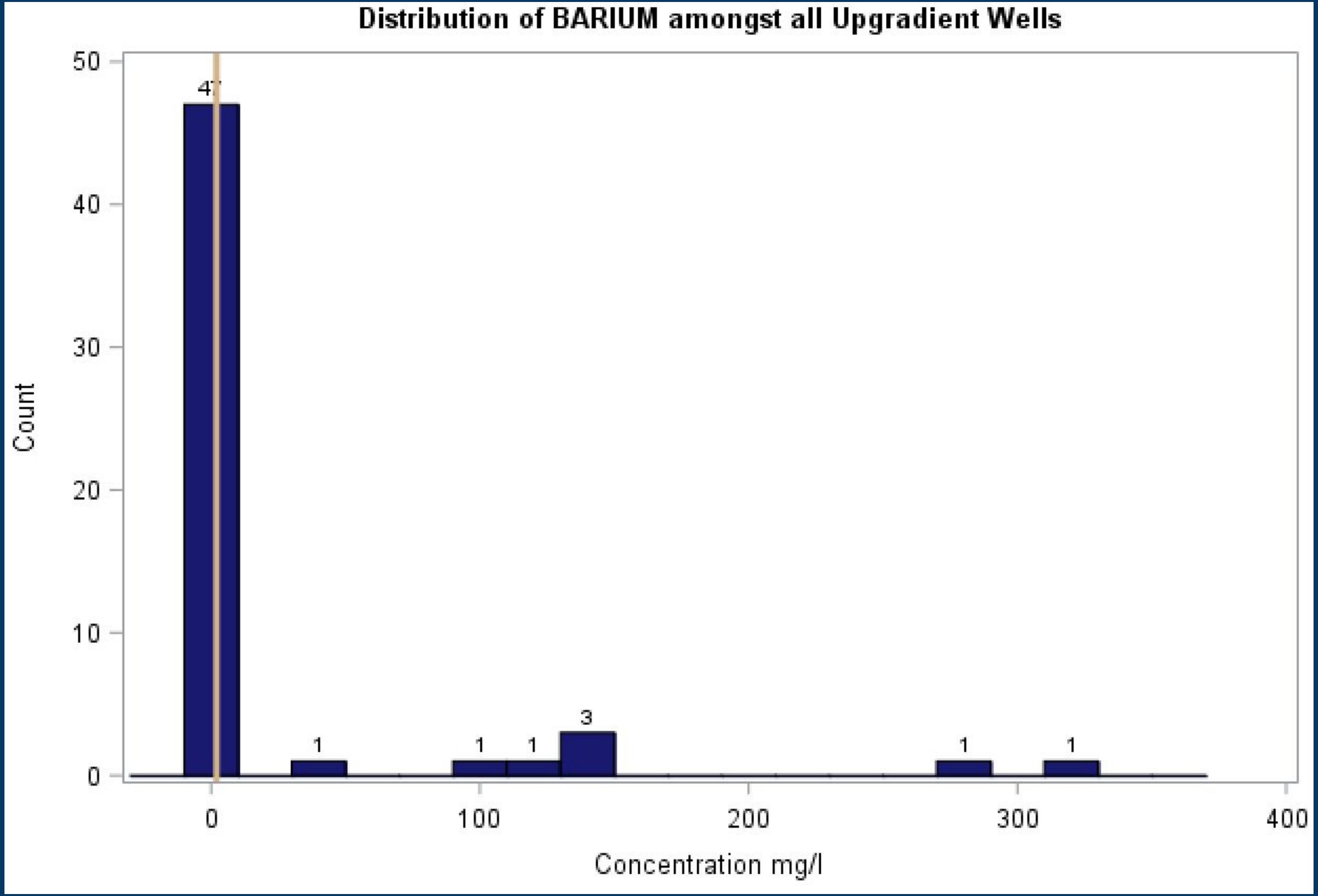
	Health-based threshold	Presumptive groundwater protection standard under CCR rule[163]
Detection monitoring constituents (40 CFR Part 257, Appendix III)		
Boron	3 mg/L[164]	
Calcium		
Chloride		
Fluoride		
pH		
Sulfate	500 mg/L[165]	
Total Dissolved Solids (TDS)		
Assessment monitoring constituents (40 CFR Part 257, Appendix IV)		
Antimony	6 µg/L	6 µg/L
Arsenic	10 µg/L	10 µg/L
Barium	2 mg/L	2 mg/L
Beryllium	4 µg/L	4 µg/L
Cadmium	5 µg/L	5 µg/L
Chromium	100 µg/L	100 µg/L
Cobalt	6 µg/L	6 µg/L
Fluoride	4 mg/L	4 mg/L
Lead	15 µg/L	15 µg/L
Lithium	40 µg/L	40 µg/L
Mercury	2 µg/L	2 µg/L
Molybdenum	40 µg/L[166]	100 µg/L
Selenium	50 µg/L	50 µg/L
Thallium	2 µg/L	2 µg/L
Radium 226 and 228	5 pCi/L	5 pCi/L



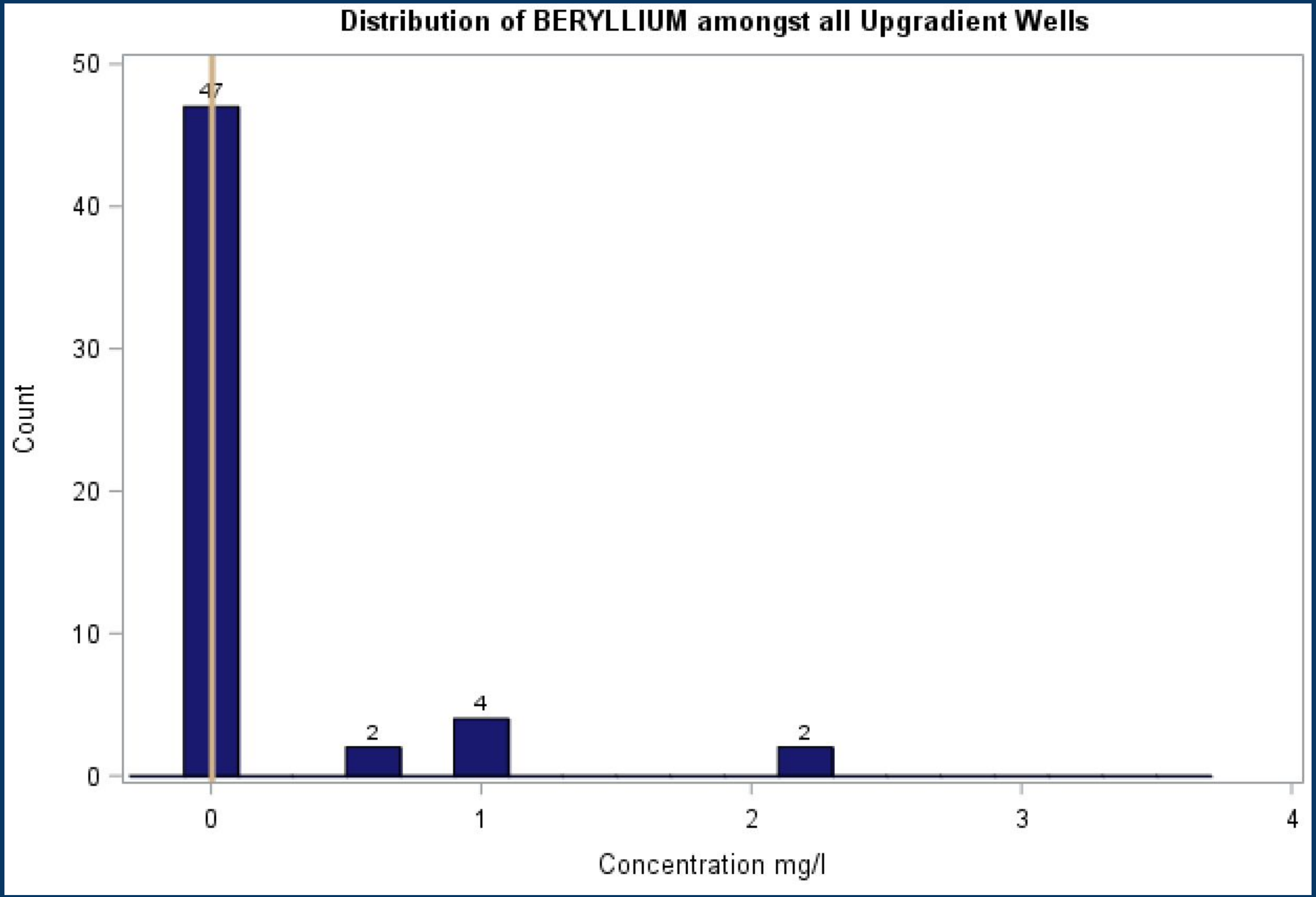
Upgradient well contaminant concentration distribution



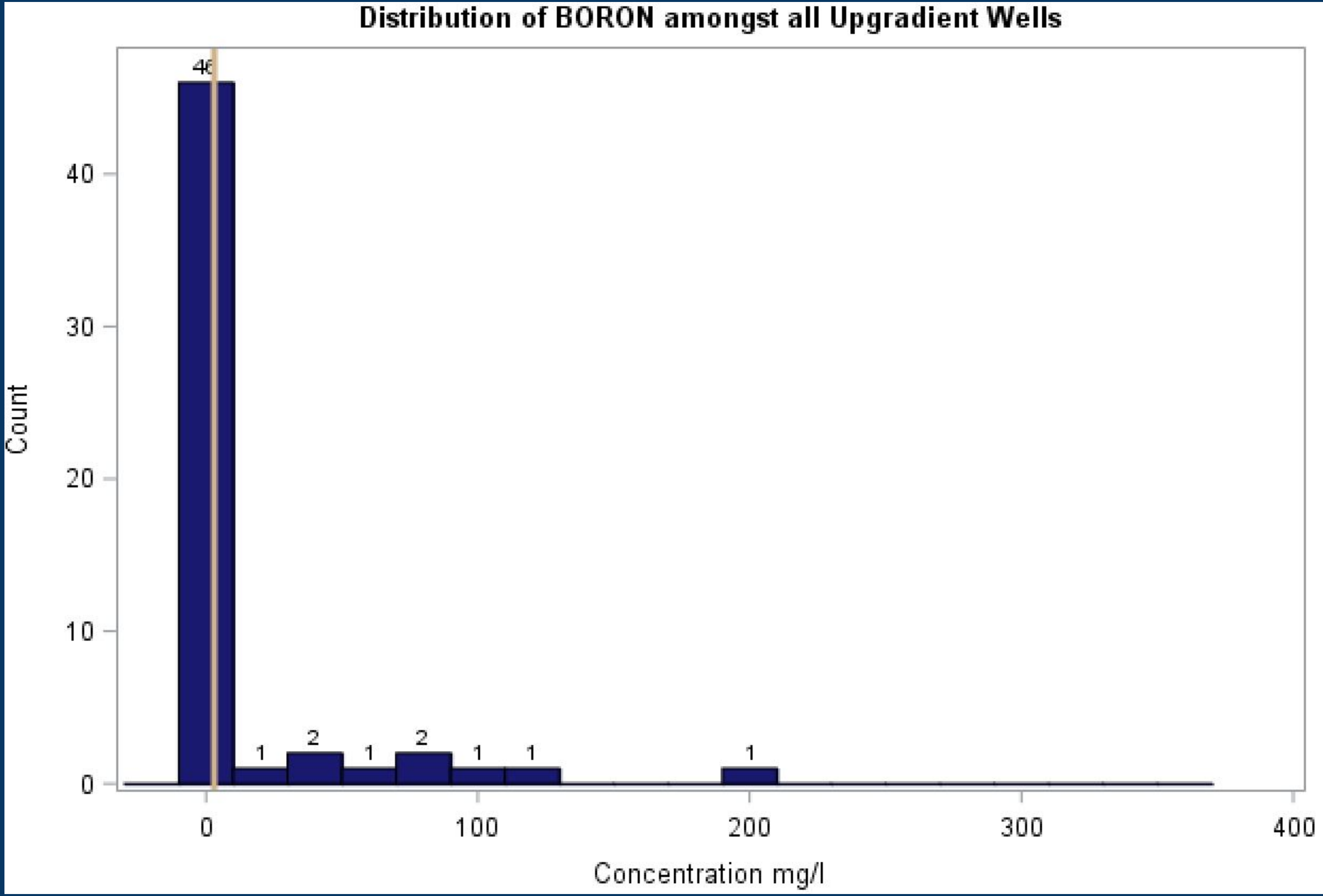
Upgradient well contaminant concentration distribution



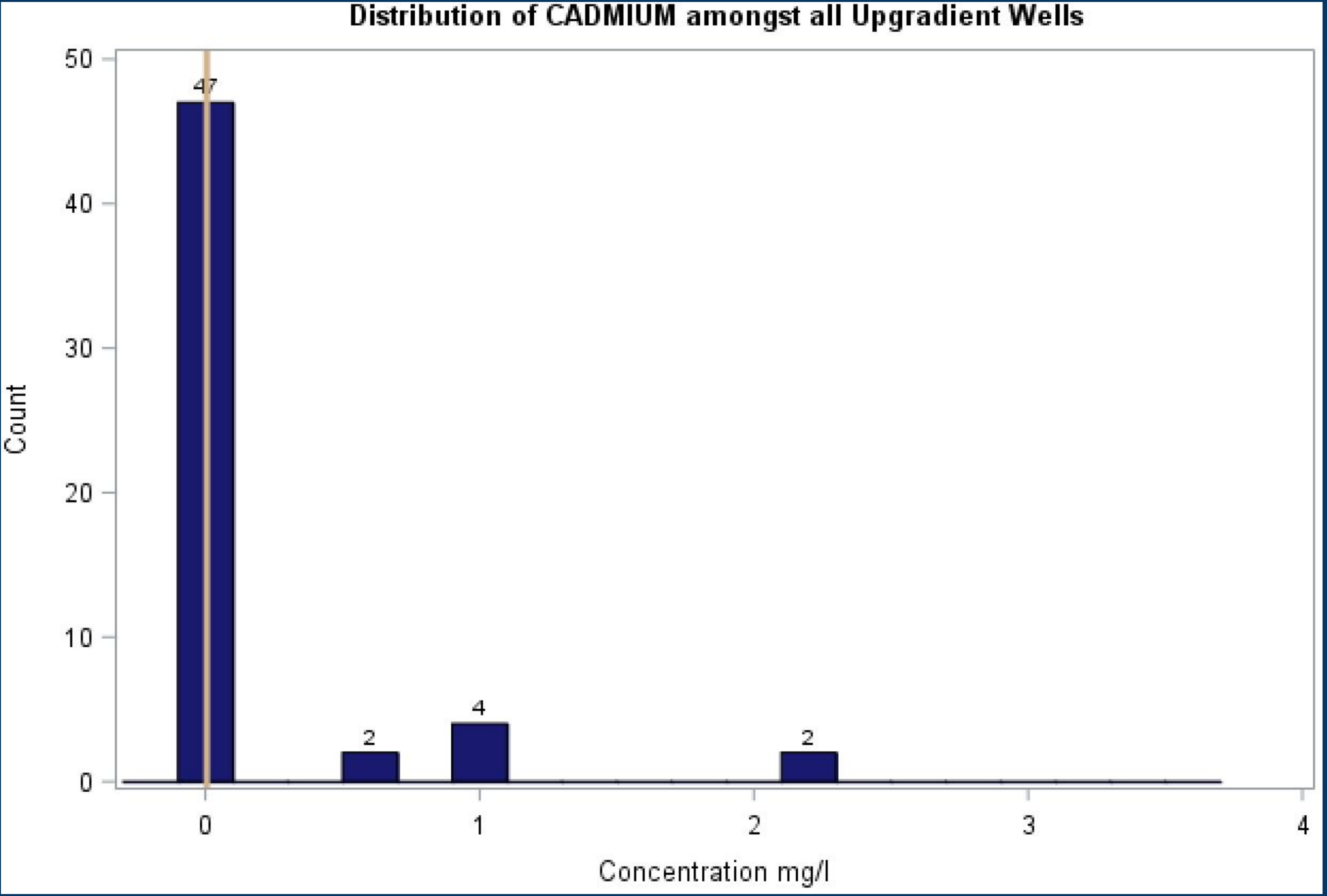
Upgradient well contaminant concentration distribution



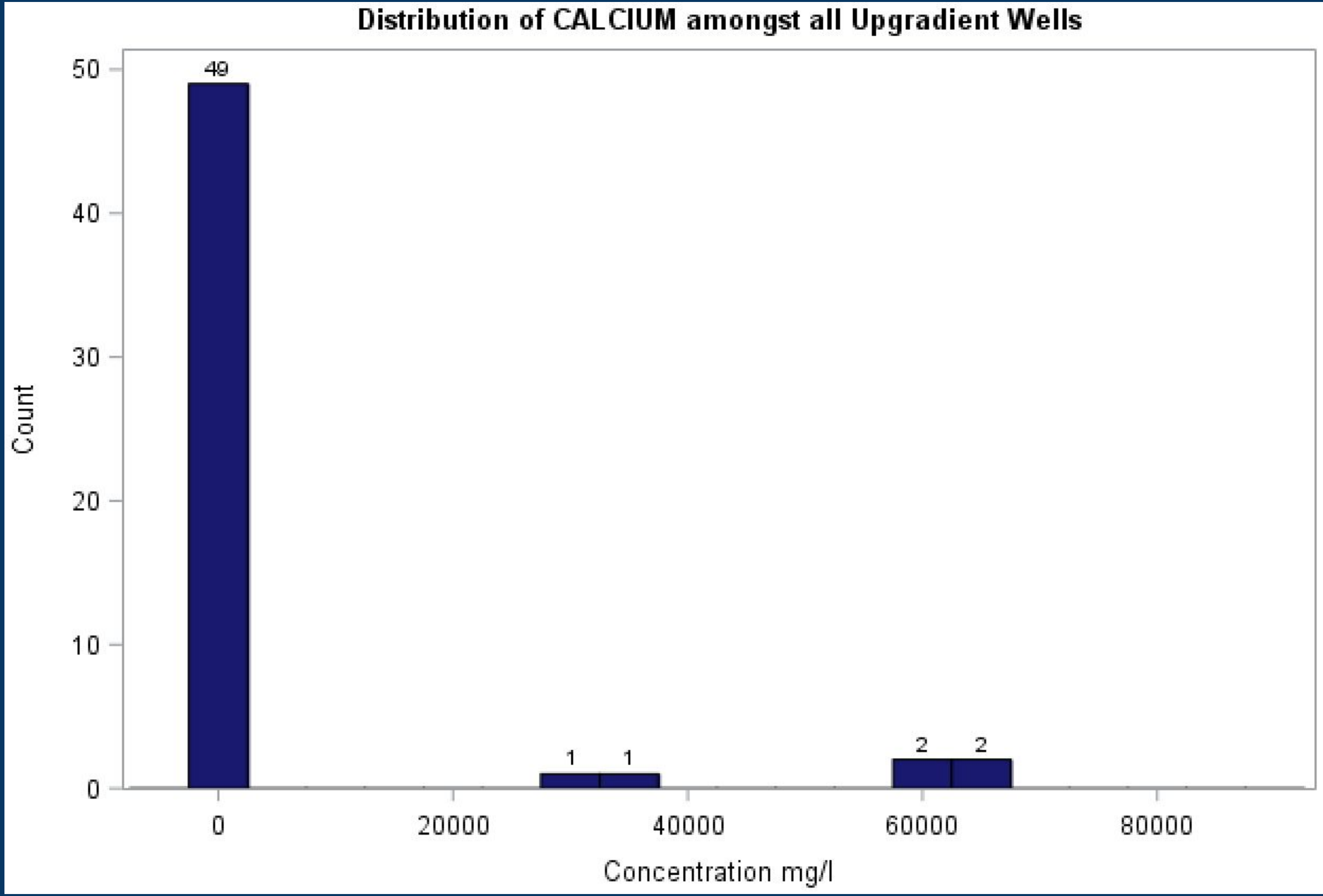
Upgradient well contaminant concentration distribution



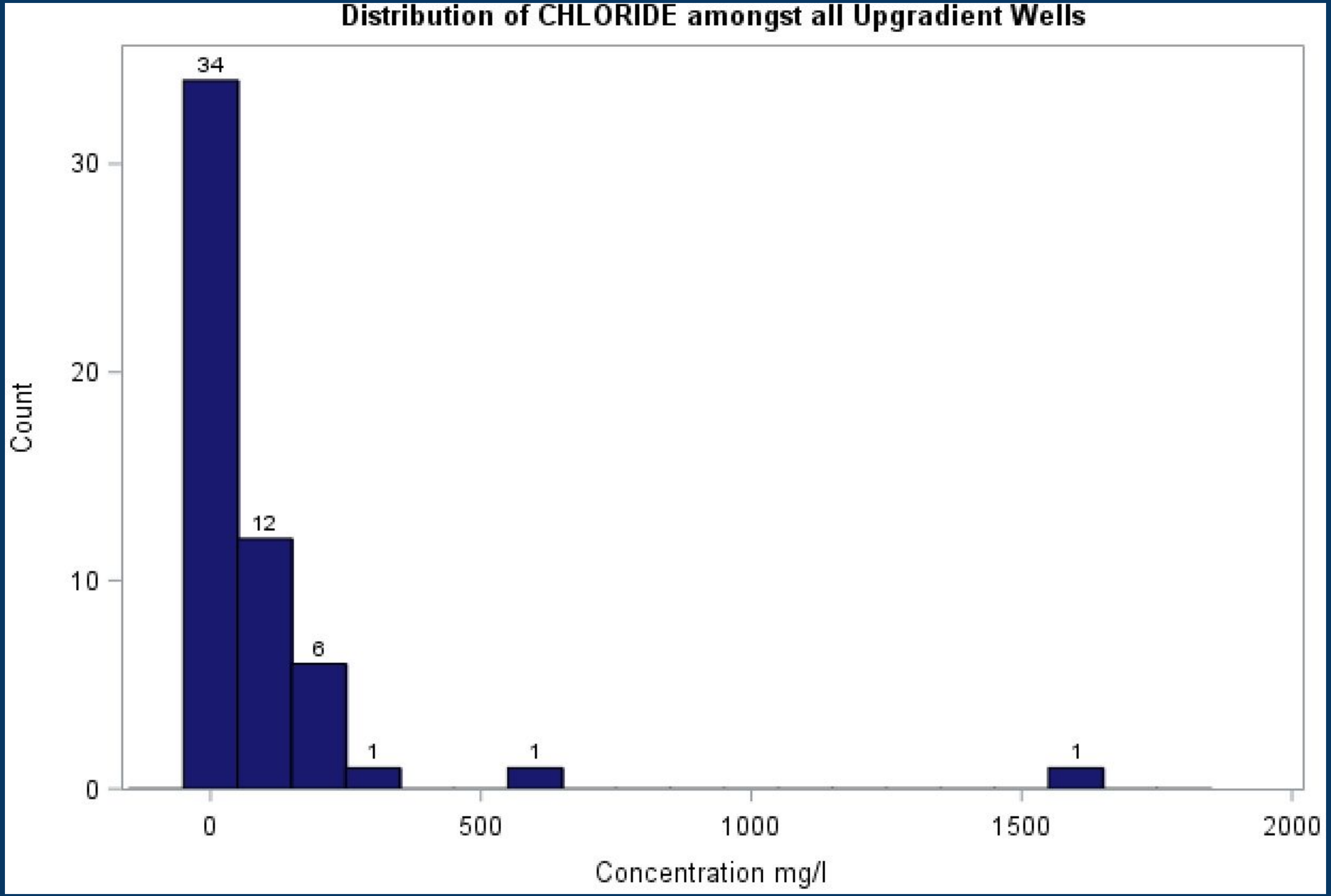
Upgradient well contaminant concentration distribution



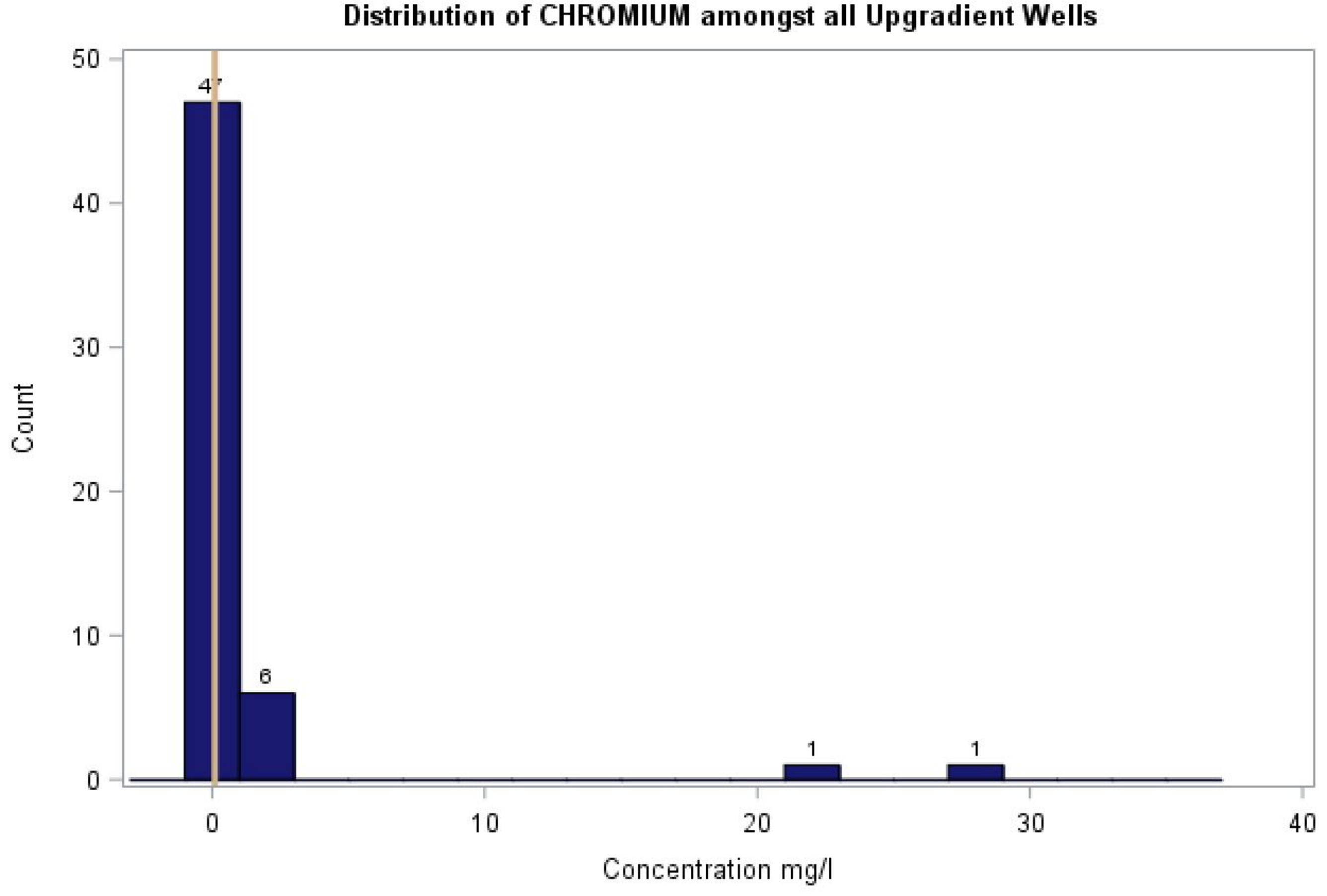
Upgradient well contaminant concentration distribution



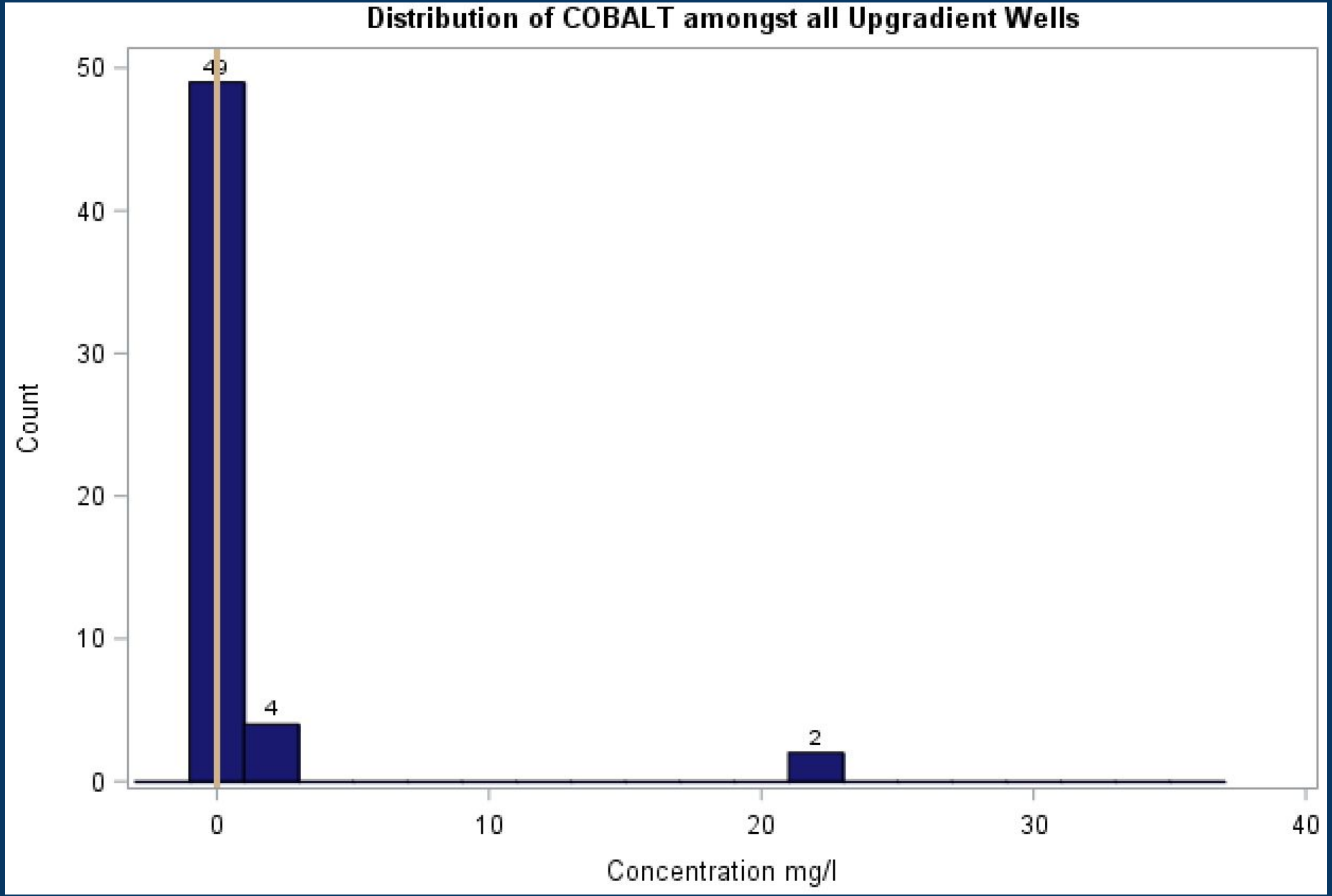
Upgradient well contaminant concentration distribution



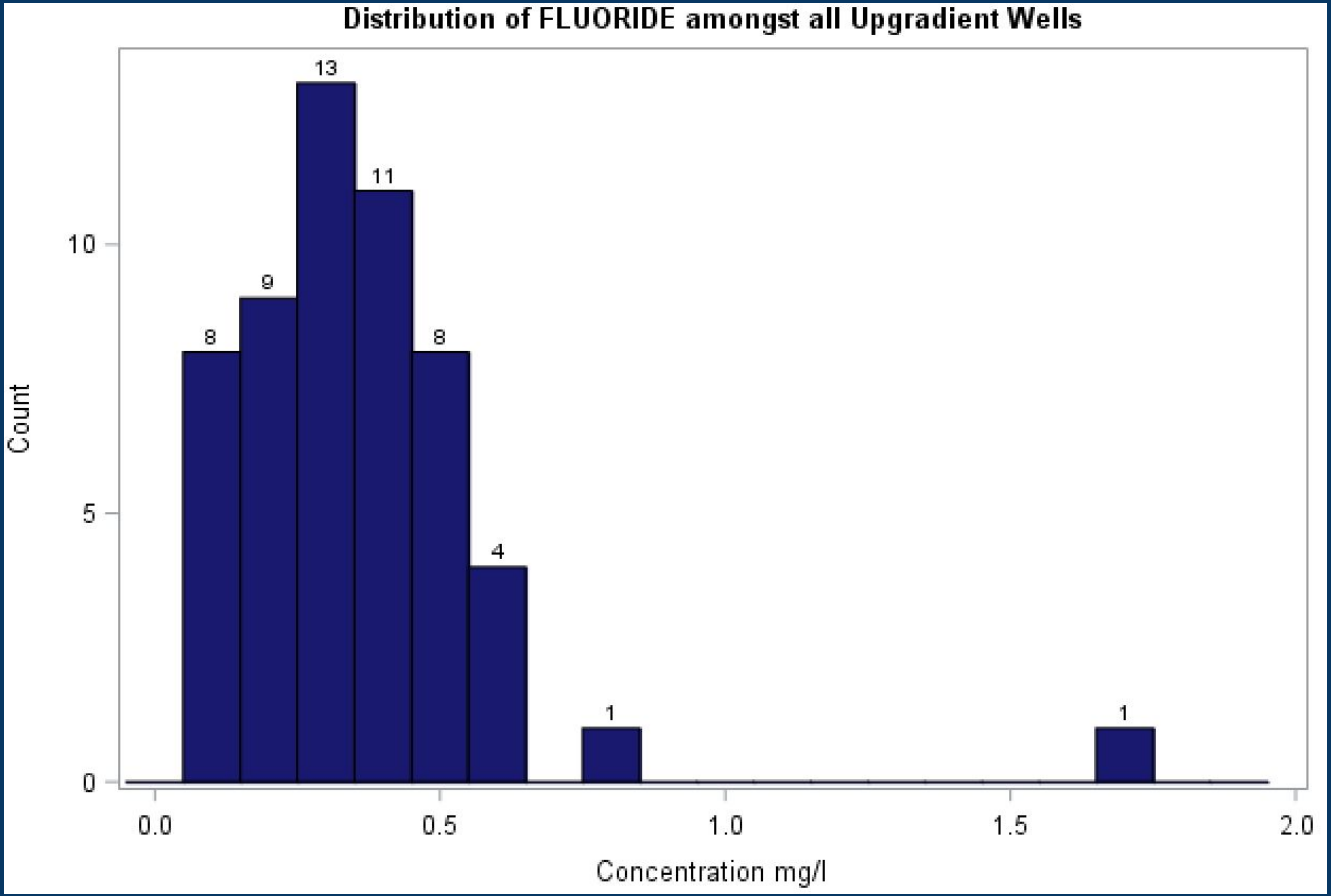
Upgradient well contaminant concentration distribution



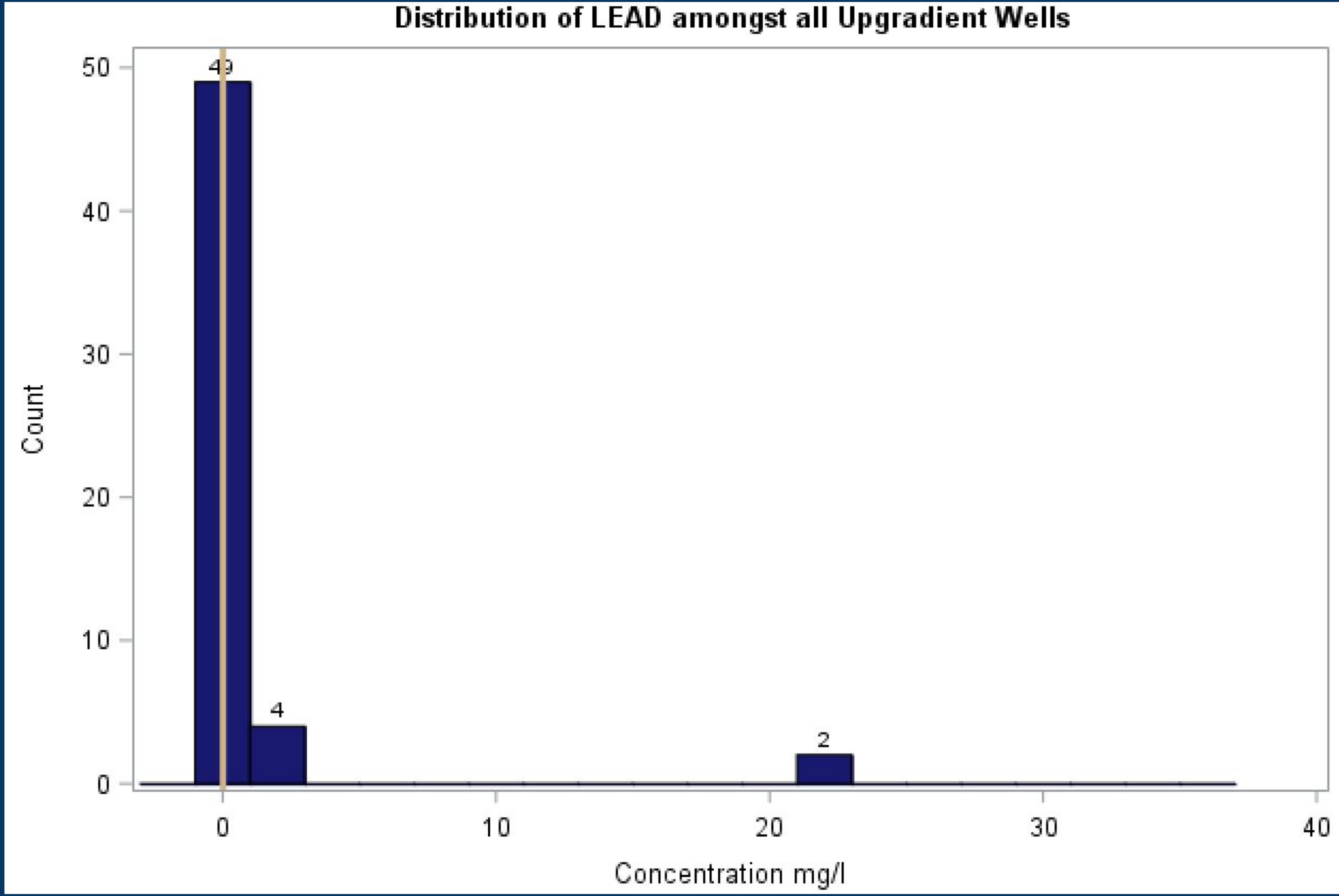
Upgradient well contaminant concentration distribution



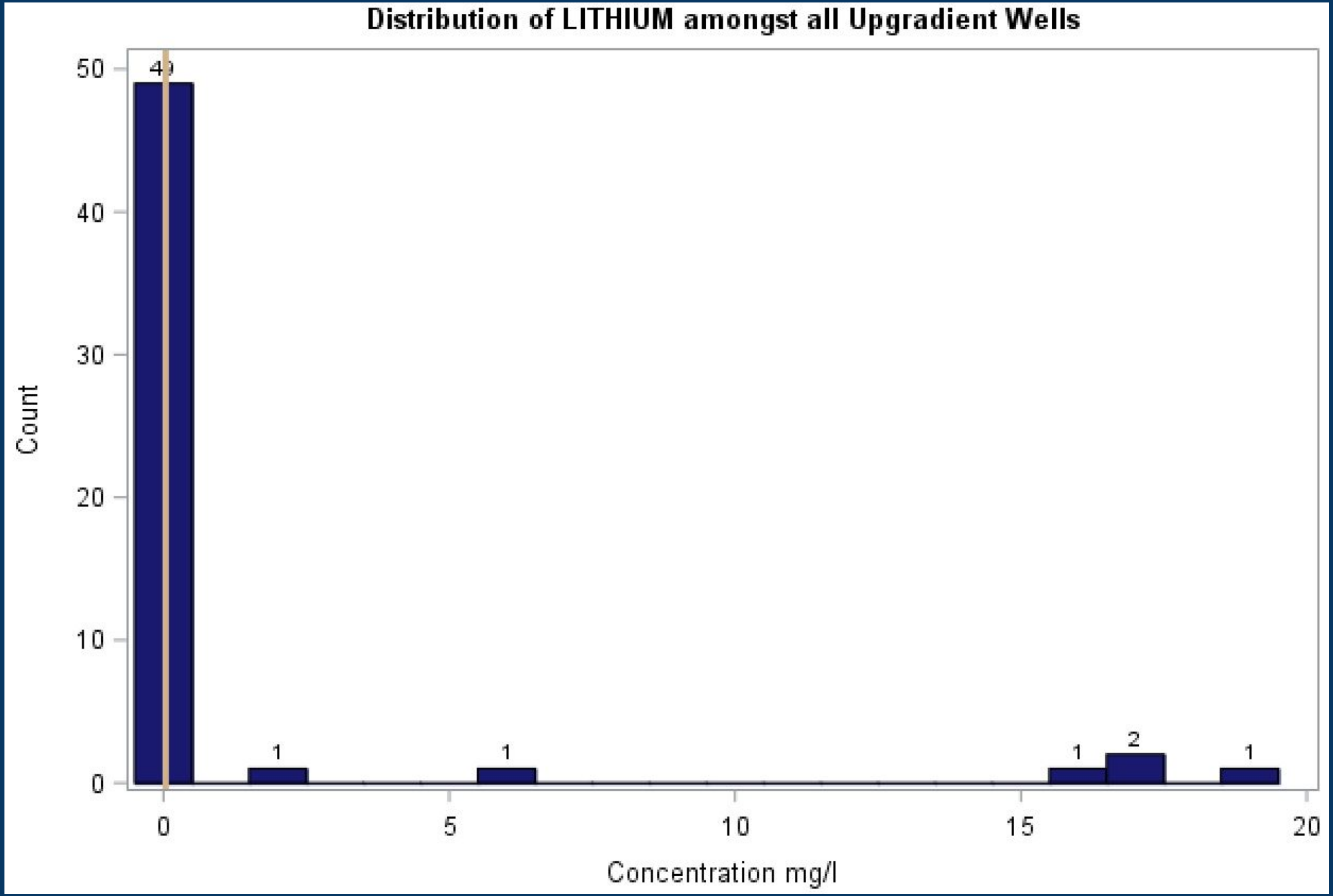
Upgradient well contaminant concentration distribution



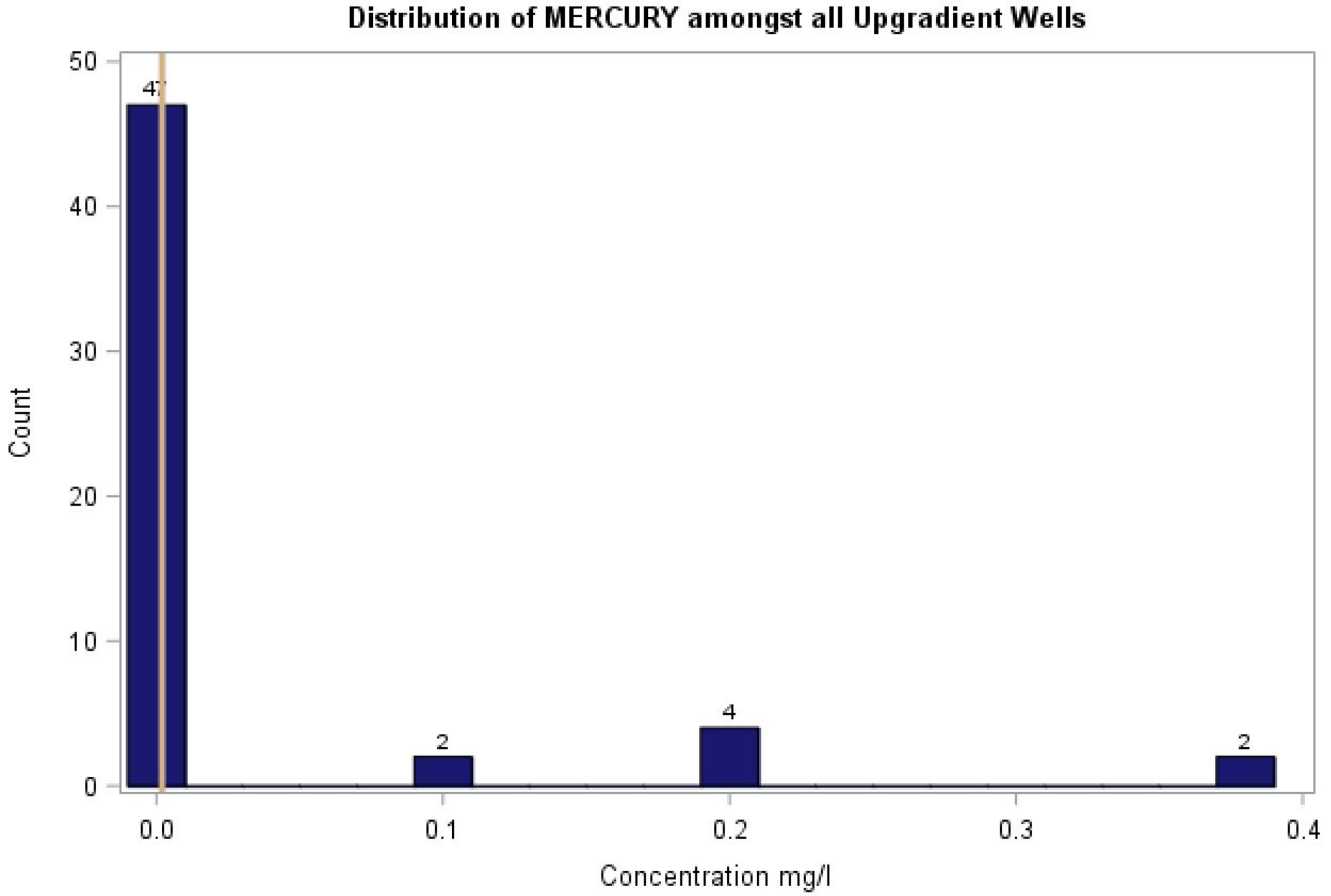
Upgradient well contaminant concentration distribution



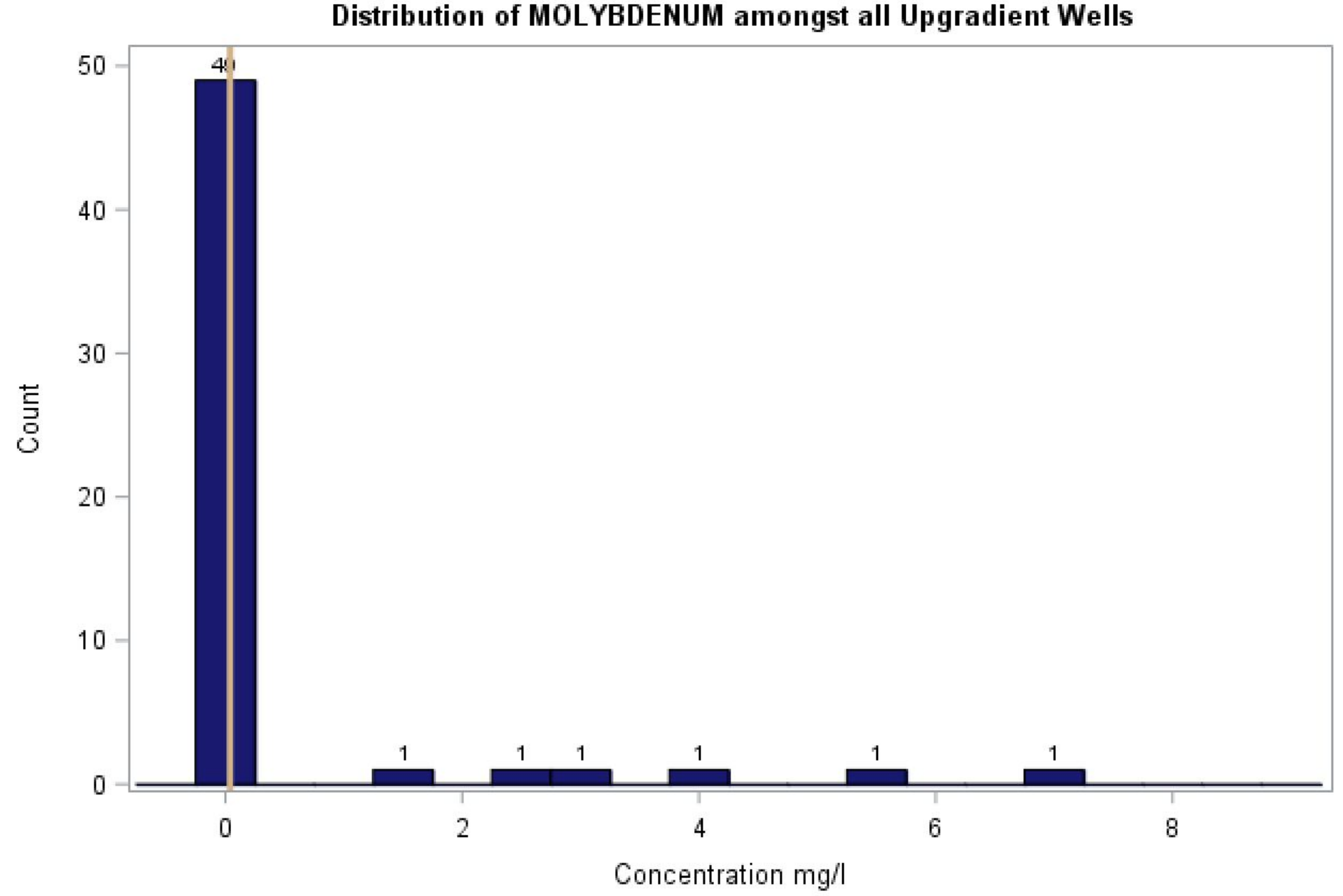
Upgradient well contaminant concentration distribution



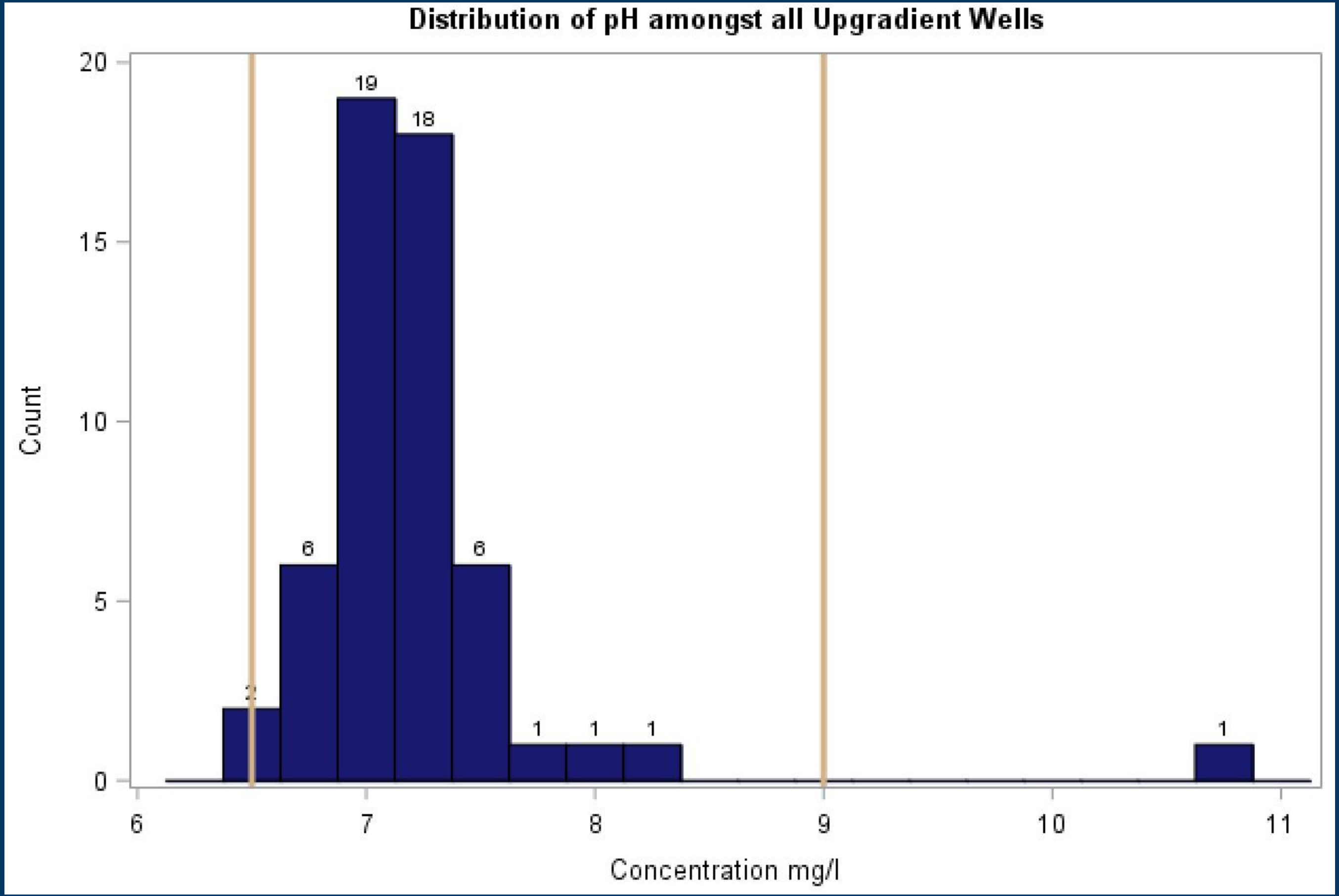
Upgradient well contaminant concentration distribution



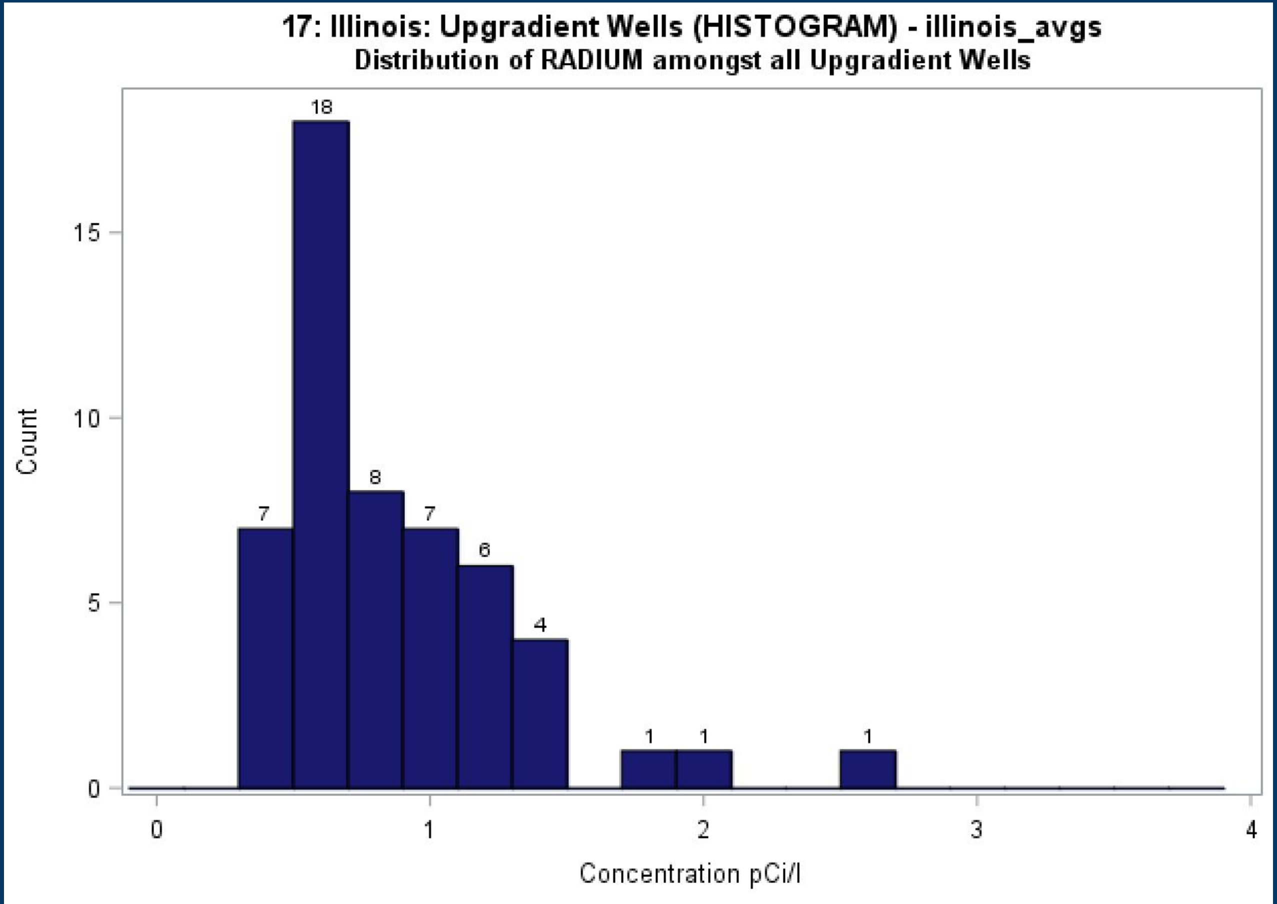
Upgradient well contaminant concentration distribution



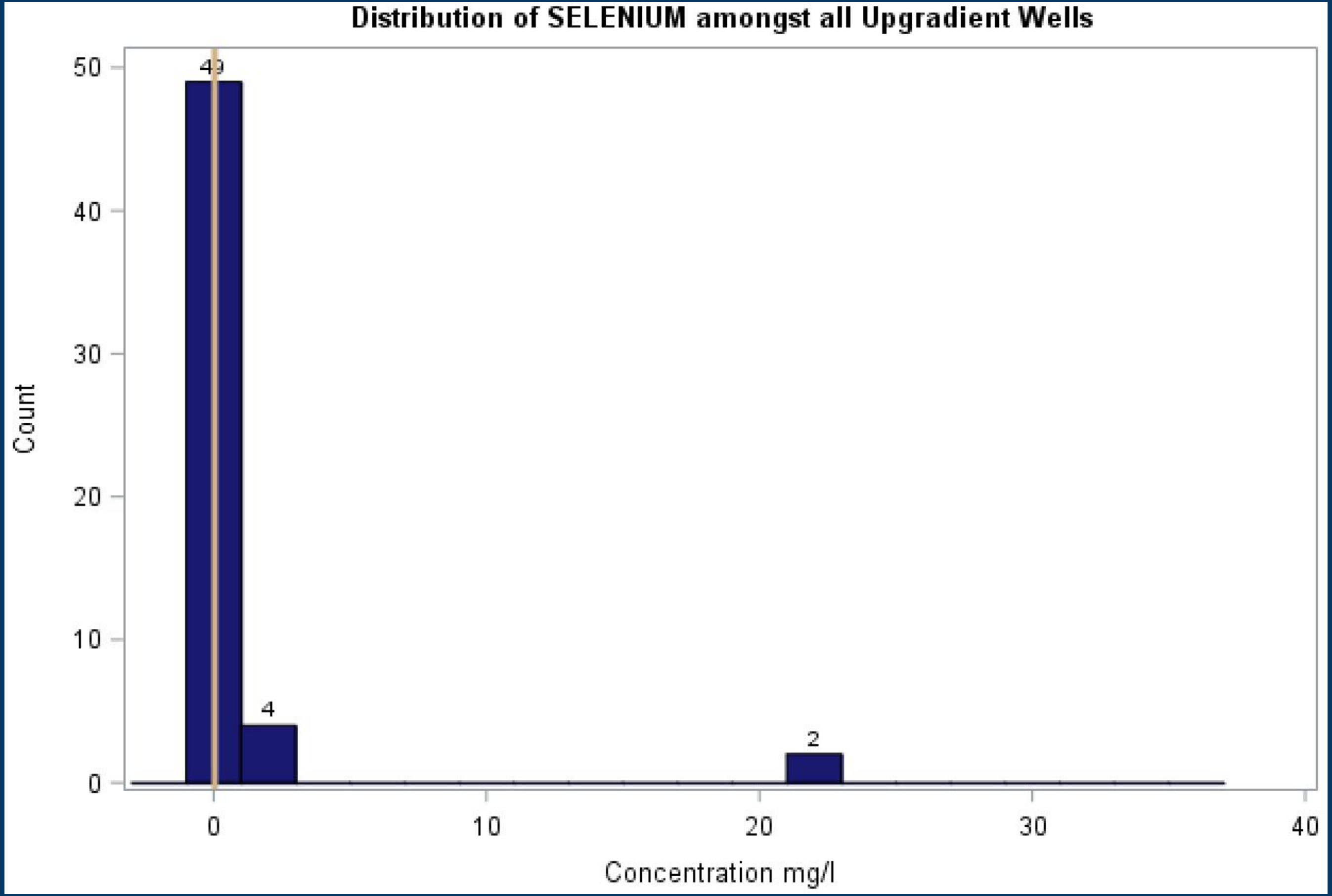
Upgradient well contaminant concentration distribution



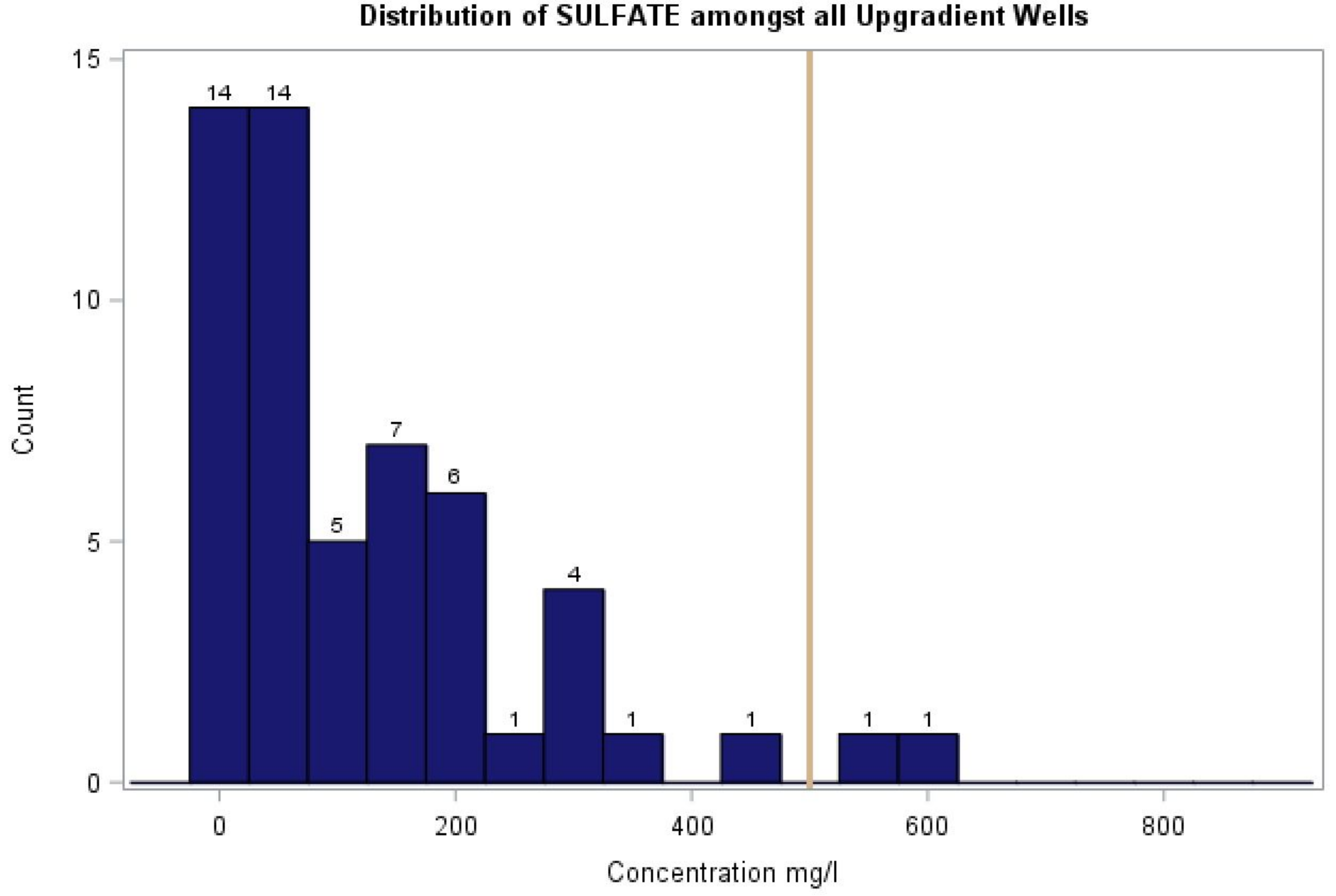
Upgradient well contaminant concentration distribution



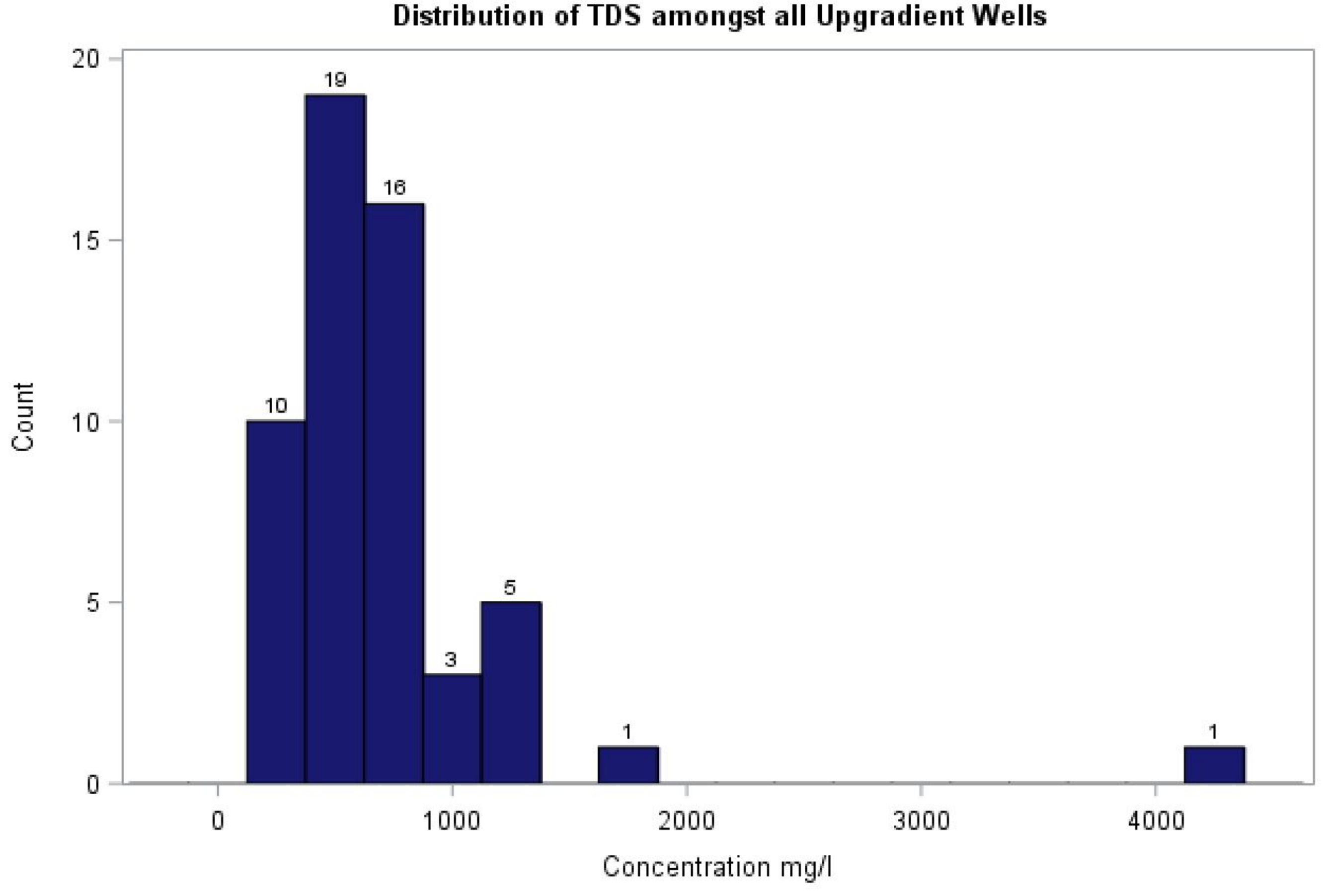
Upgradient well contaminant concentration distribution



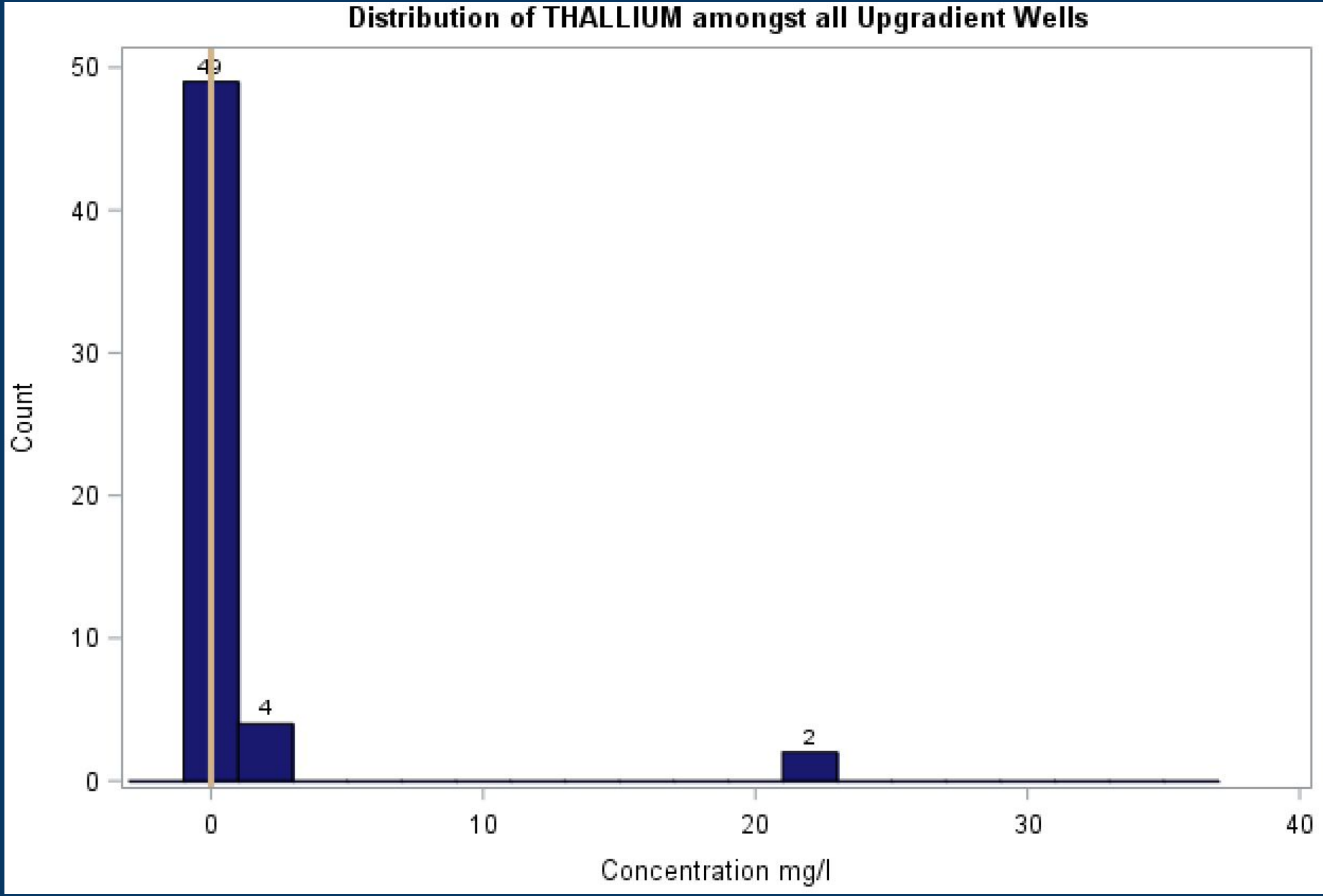
Upgradient well contaminant concentration distribution



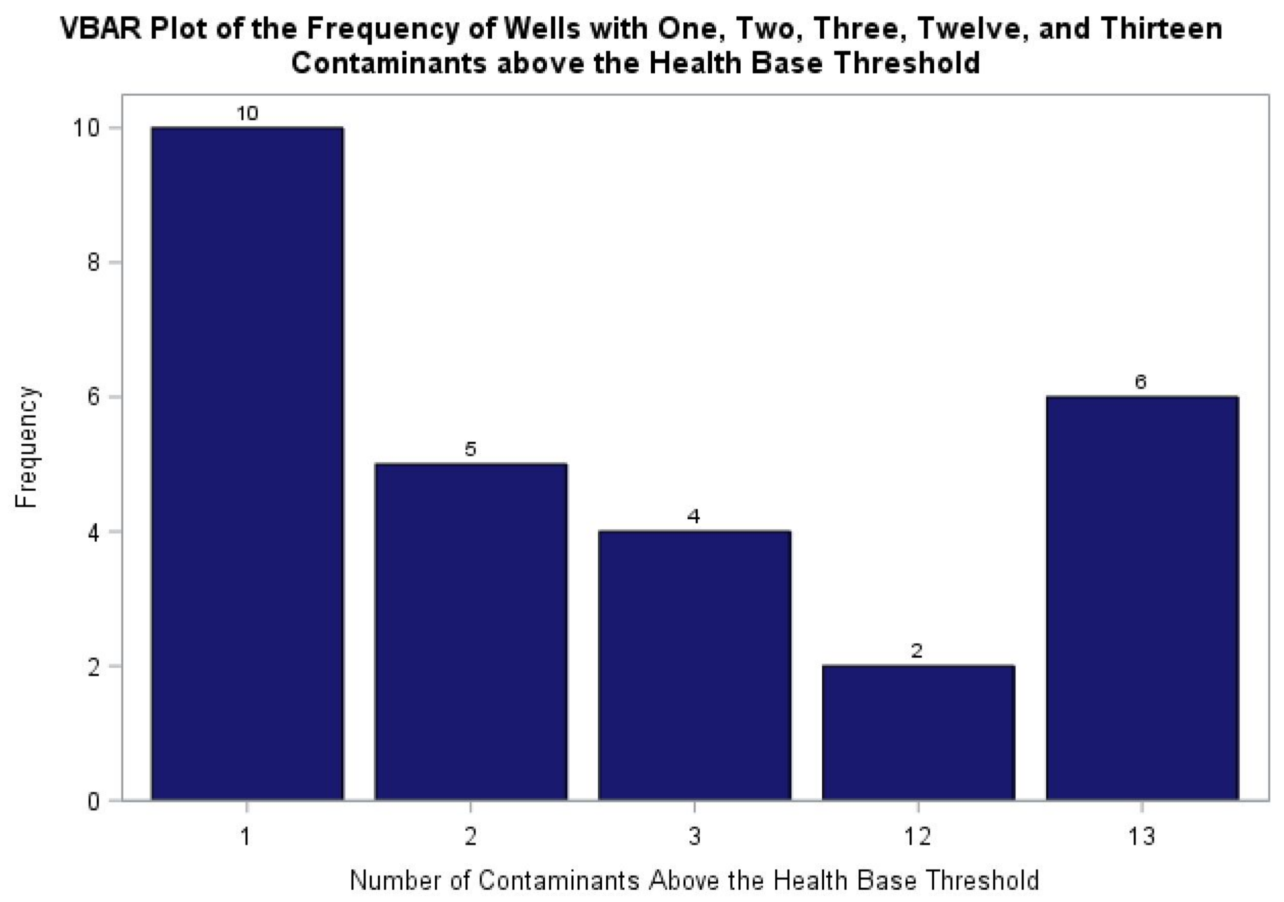
Upgradient well contaminant concentration distribution



Upgradient well contaminant concentration distribution



Upgradient well contaminant concentration distribution



Upgradient well contaminant concentration distribution

Wells identified

Wells with ______ contaminant(s) above the health based threshold				
One	Two	Three	Twelve	Thirteen
08D 31 8 APW5 BA06 G201 G45S MW-01 MW-10 MW-304	25 G48MG MW-05 MW-06 T03S	EBG MW-09 MW-11 MW-14	AP-4 AP-5	G01D G02D G03D MW201 MW203 MW24D

Classification

Classification - Introduction

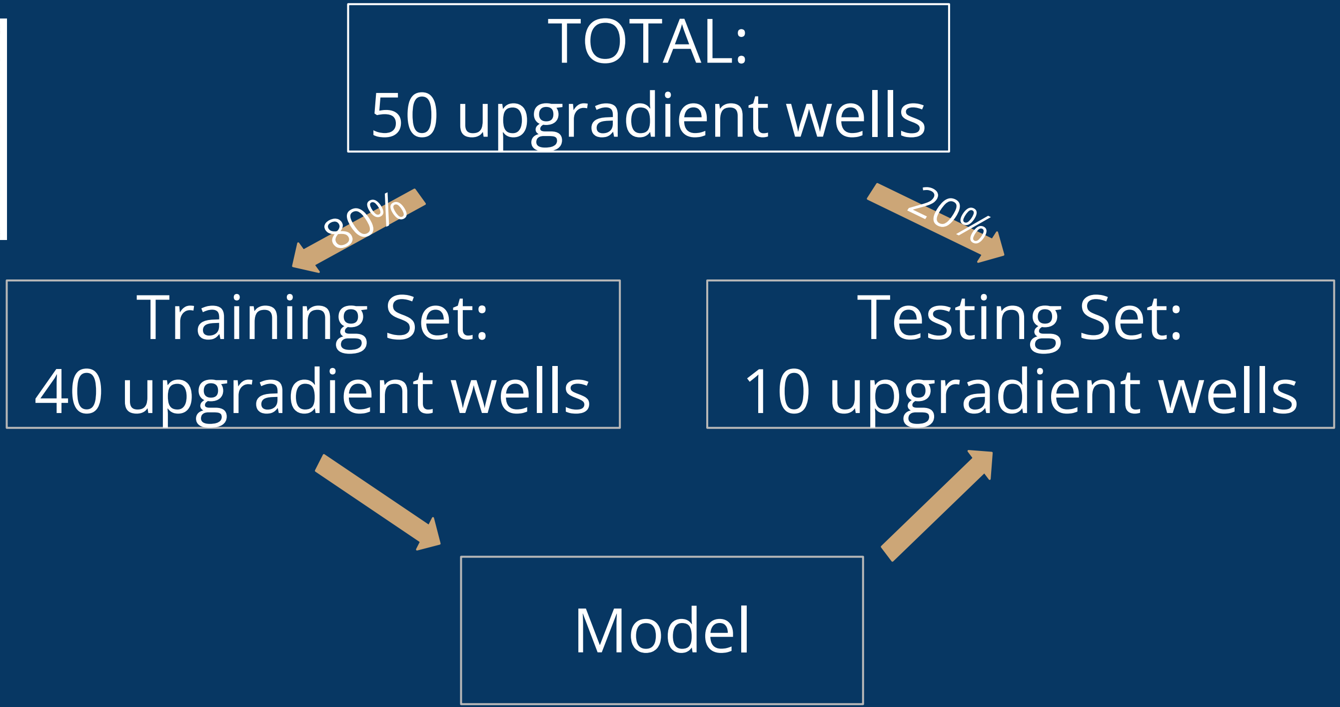
- How does classification work?
- What is the goal of classification?
  - How do we achieve this?
- Assumption that we have to make for our coal ash dataset

Classification - Which method?

- Many different classification methods to choose from, each with their advantages/disadvantages
- We decided to go with kNN (k nearest neighbors)
    - Why?

Classification - Splits

Status	Frequency
dangerous	44
safe	6



Classification - Results

	dangerous	safe
dangerous	8	0
safe	1	1

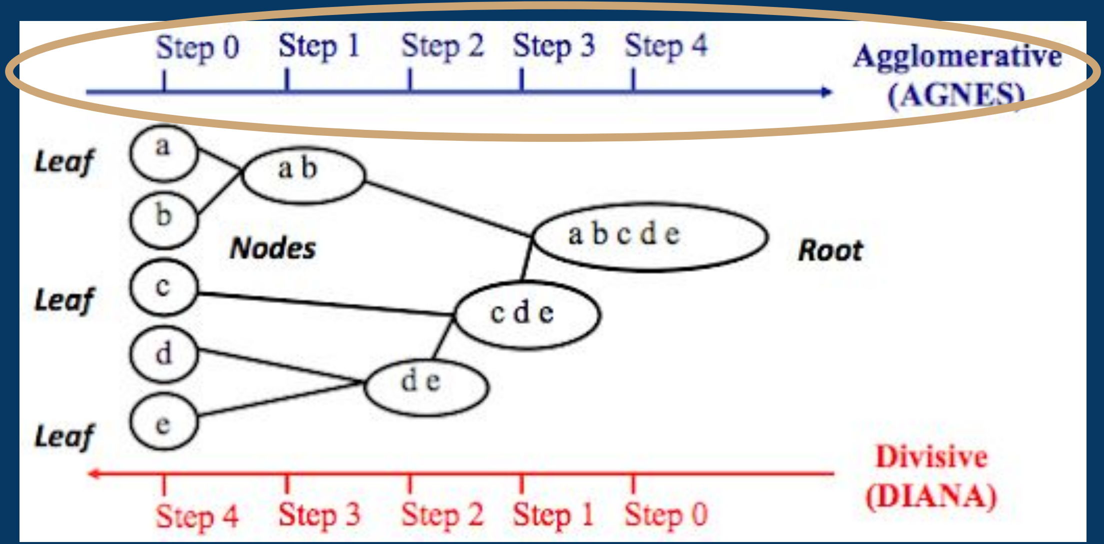
$$AER = \frac{1+0}{8+0+1+1} = 0.1$$

Let’s try to use a clustering-based approach instead!

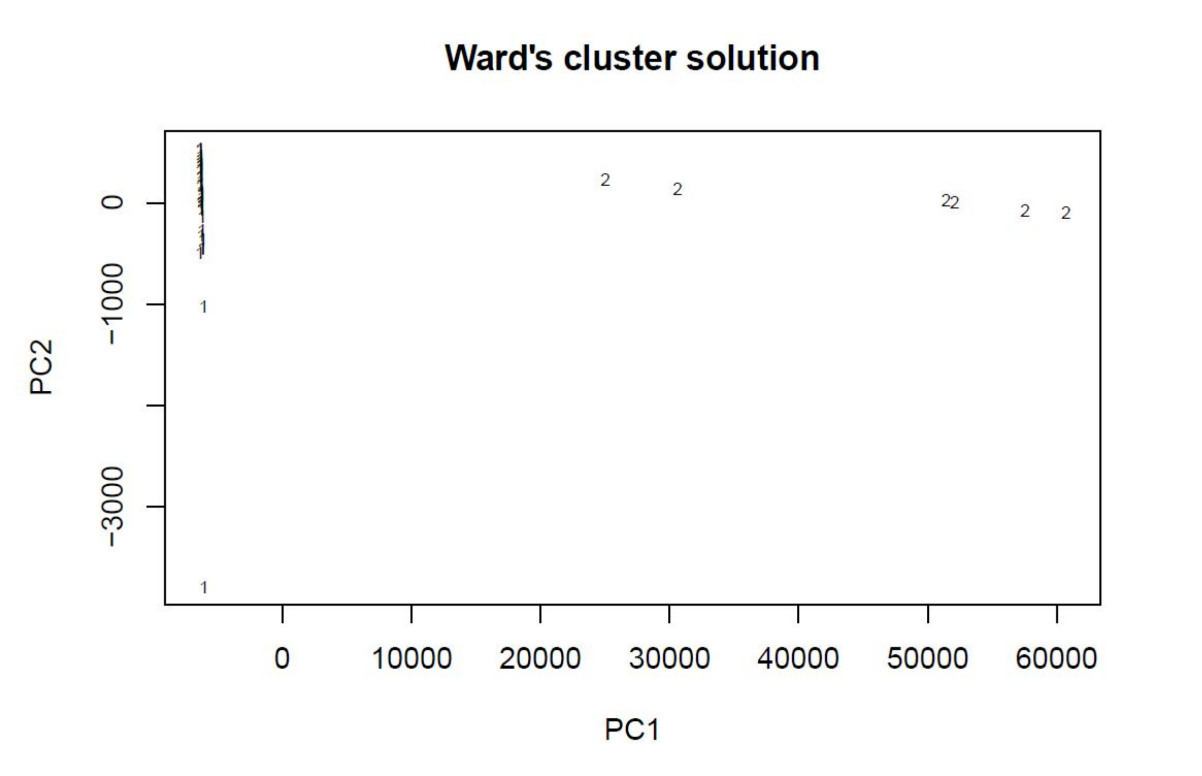
Clustering

Clustering - Hierarchical Clustering

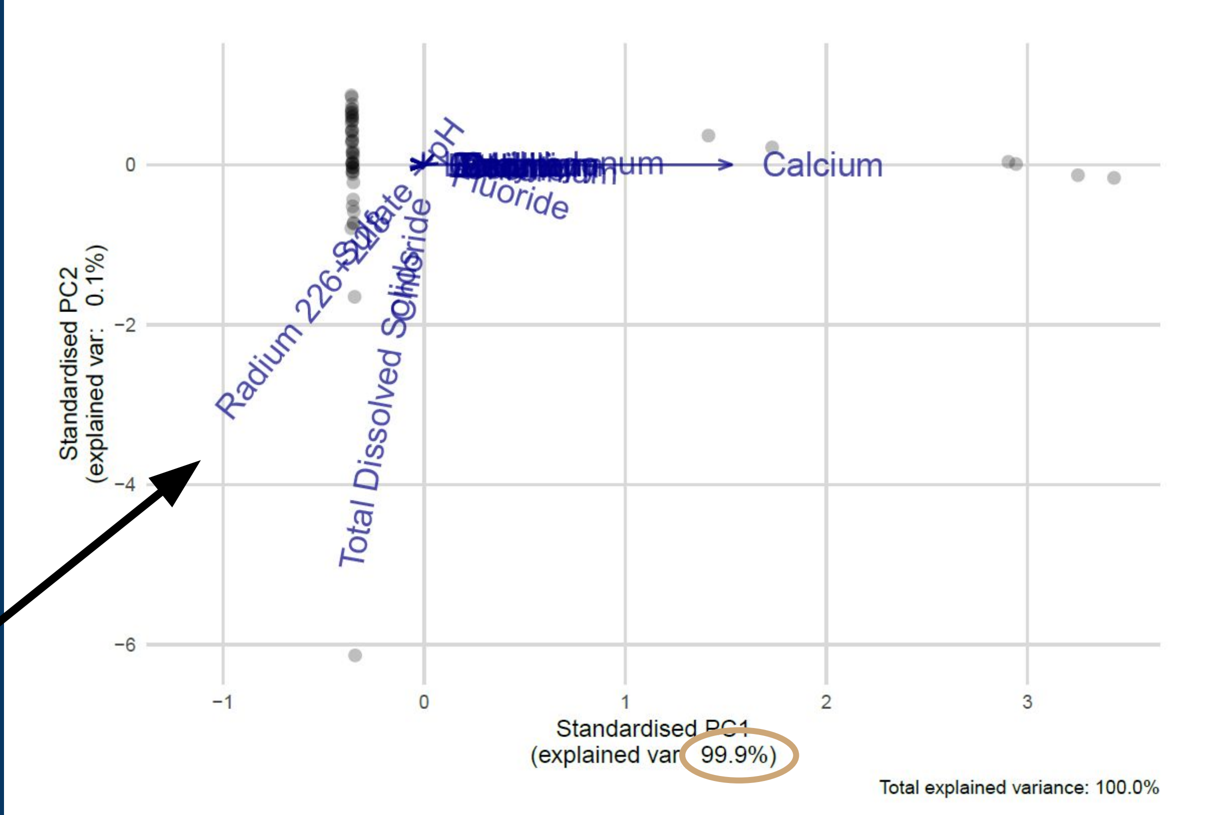
- Ward's Method
  - Agglomerative Hierarchical Clustering
    - Bottom-up approach (small to big)



Clustering (HC) - Ward’s Cluster Solution



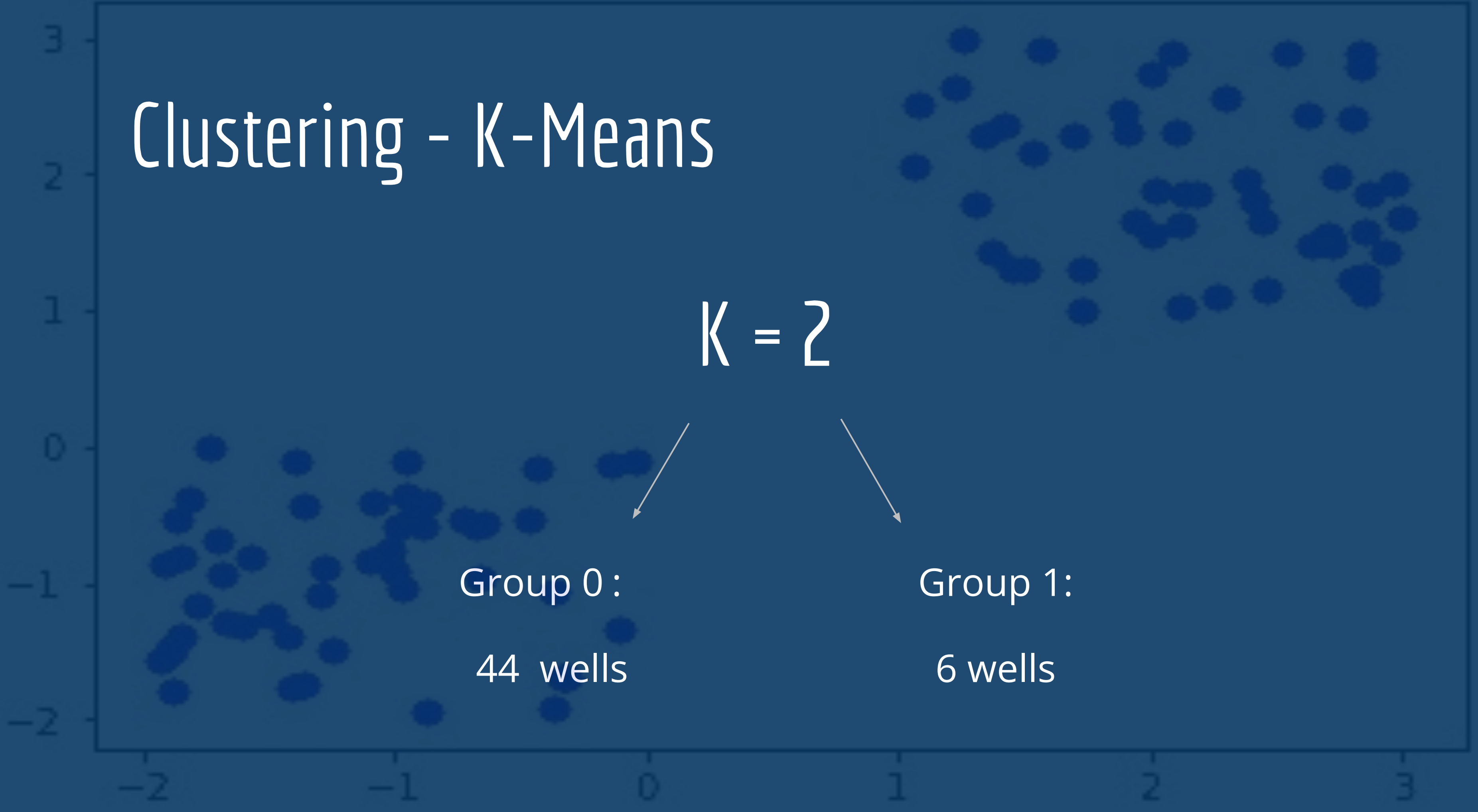
Clustering (HC) - Biplot



- Sulfate
- Radium 226+228
- Chloride
- TDS

K-Means Clustering

Clustering - K-Means



K = 2

Group 0 :

44 wells

Group 1:

6 wells

- Comparing ***mean*** concentrations of each group/cluster

contaminant	Antimony	Arsenic	Barium	Beryllium	Boron	Cadmium	Calcium	Chloride
clusters								
0	0.002301	0.141806	0.129627	0.001088	0.791699	0.001112	111.084382	113.158421
1	0.833500	8.538752	150.552354	0.833500	99.318653	0.833500	52651.684259	10.953704

. . .

	Chloride	Radium 226+228	Sulfate	Total Dissolved Solids
clusters				
0	113.158421	0.914453	141.150590	739.162000
1	10.953704	0.641990	20.490741	395.907407

Radium 226+228 threshold: 5 pCi/L

Sulfate threshold: 500 mg/L

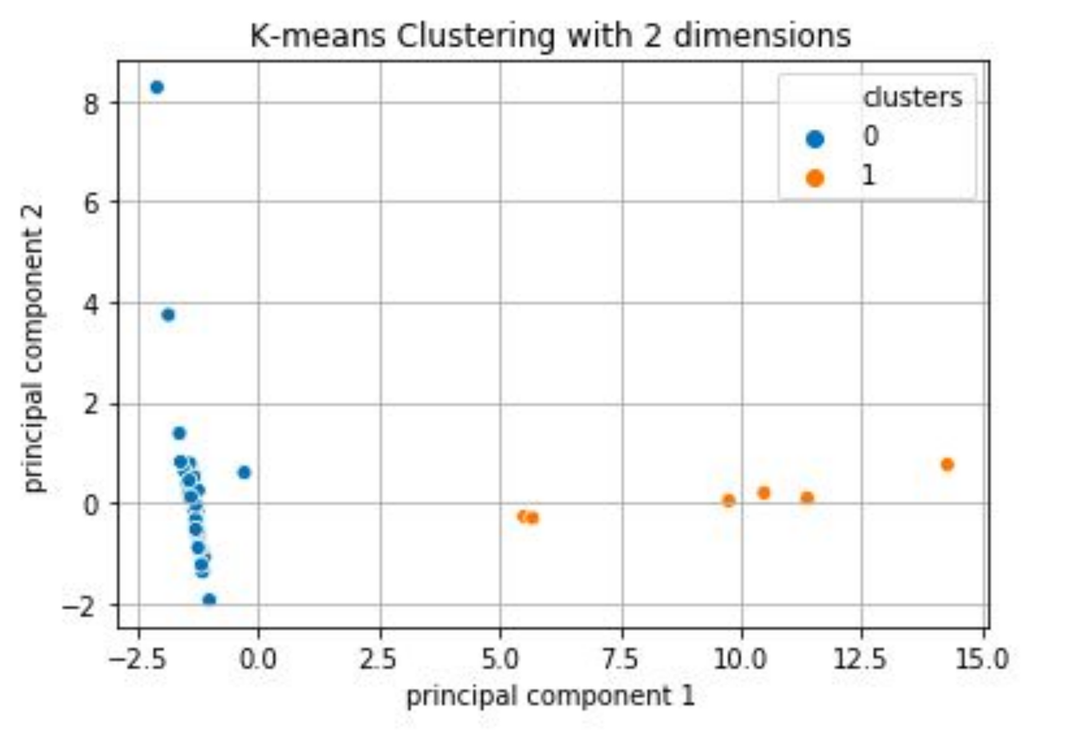
Group 1 wells:

- G01D
- G02D
- G03D
- MW201
- MW203
- MW24D

Wells with 13 threshold violations:

- G01D
- G02D
- G03D
- MW201
- MW203
- MW24D

Visualization - PCA

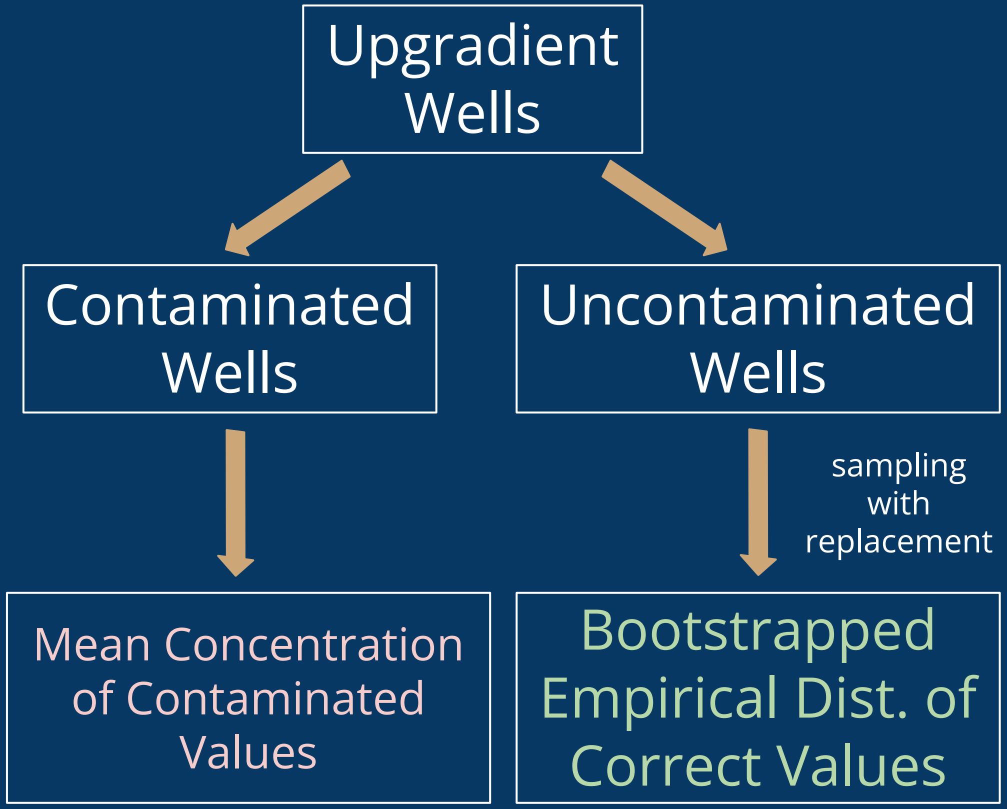


Bootstrapping

Bootstrapping

- Need to correct these values!
    - Contamination from retired/unregulated ponds

- Calculate the actual amount of contamination ONLY from the coal ash with the equation:



Mean Conc. of Contaminated Wells - Bootstrapped Empirical Dist. of Corrected Values

Conclusion

Conclusion

- Did we accomplish our original research goals?
  - Yes, for the most part!
  - Classification based approach was limited
    - Model needs more training data
  - Cluster based approach was more effective
    - Identified two distinct clusters
  - Bootstrapping and imputation methods

Acknowledgements

Thank you so much - Dr. Rachel Nethery and Luli Zou! We appreciate your guidance and support throughout this research project! ❤️

Also a big thank you to Dr. Marcello Pagano and Priti Thareja, and to everyone who made this program possible this summer!