# Analysis of drinking water water contaminant occurrence in the northeastern and southeastern United States

https://github.com/rachel-gonsenhauser/Final_Project_Environmental_Data_Analytics

Rachel Gonsenhauser

## Abstract

add text here once finnndinngs clear!!!!

# Contents

# List of Tables

# List of Figures

# 1 Rationale and Research Questions

While the EPA establishes standards for 90 drinking water contaminants by means of the federal Safe Drinking Water Act (SDWA) and its regulations, public water systems still often struggle to remain in compliance with such policies (USEPA, 2020). This issue of compliance with the SDWA can stem from myriad causes, for instance financial capacity of the water system. This is especially concerning in areas where geologic conditions and/or anthropogenic activities frequently introduce contaminants into drinking water supplies. Additionally, some known contaminants still have yet to be regulated by the EPA, such as poly- and perfluoroaklyl substances (PFAS), which introduces even more complexity to the issue of water quality monitoring of drinking water sources.

This analysis seeks to investigate the co-occurrence of water quality indicators including arsenic, trihalomethane, uranium, and PFAS, which originate from both geogenic and anthropogenic sources. Additionally, given pervasive questions related to environmental justice and how socioeconomic factors may be related to water quality indicators, this analysis seeks to examine relationships between water quality indicators and county-level median household income (MHI) and size of the population served by a given community water system (CWS), which is often a proxy for how rural an area is and the financial capacity of a CWS. Additionally, questions regarding how contaminant occurrence differs across time and between states are explored.

To narrow the scope of this project, most analyses are targeted to southeastern region states and northeastern region states. These regions were chosen given their differences in geology and socioeconomic makeup. Additionally, individual case studies of Massachusetts and North Carolina are explored in further depth. As arsenic is present in much of the underlying geology in New England and other northeastern states, arsenic data is used is many of the analyses performed. Due to issues of PFAS data limitations, discussed in more detail in the subsequent section, analyses using PFAS data are limited. Specifically, the following questions are explored in this analysis: Question 1: Do arsenic concentrations vary significantly from state to state in northeastern and southeastern states? Question 2: Do socioeconomic factors or the presence of other contaminants predict arsenic concentrations in Massachusetts and North Carolina? Question 3: Do PFAS concentrations vary signficantly across time and from state to state in the United States? Are socioeconomic factors significant predictors of PFAS concentrations?

***EDIT THIS LATER TO PUUT EMPHASIS ON ARSENIC AND PFAS AND EXPLAIN WHY!!!!

## 2 Dataset Information

Data used for this analysis was downloaded from the Centers for Disease Control and Prevention (CDC)'s National Environmental Public Health Tracking Network at Centers for Disease Control and Prevention (CDC)'s National Environmental Public Health Tracking Network https://ephtracking.cdc.gov/DataExplorer/#/. Output from this online tool containing geographic and CWS data associated with individual variables was combined into the final processed dataset used for this analysis.

The wrangling process entailed taking individual datasets containing data for arsenic, PFAS, uranium, trihalomethane, and MHI and joining them into the final processed dataset. Each of these variables had accompanying data inncluding the year, state, county, and CWS in which the data was collected for each parameter. As unique county Federal Information Processing Standards (FIPS) codes were standard across all individual datasets, this variable was used to join datasets into the final processed dataset.

Figure 1: Summary information for processed dataset

| **Parameter** | **Summary** |
|---|---|
| Number of states | 28 |
| Number of CWSs | 25,583 |
| Water quality indicators | Arsenic, trihalomethane, uranium, PFAS |
| Socioeconomic variables | Population served by CWS, MHI |
| Data collection time span | 1999-2018 |

Figure 2: Description of Variables Used in Analyses

| **Column Heading** | **Variable Description** | **Data Range** |
|---|---|---|
| stateFIPS | Federal Information Processing Standard state code | N/A |
| State | state measurement was taken in | N/A |
| countyFIPS | Federal Information Processing Standard county code | N/A |
| County | county measurement was taken in | N/A |
| Year | year measurement was taken in | N/A |
| Arsenic_ugL | mean arsenic concentration (micrograms per liter) | 1-2,422 micrograms/liter |
| PWS.ID | Public Water System Identification Number | N/A |
| CWS.Name | Community Water System Name | N/A |
| Population.Served | number of people served by CWS | 0-8,271,000 people |
| MHI | median household income ($) | $16,435-$113,336 |
| PFAS_ppt | PFAS conncentration (parts per trillion) | 1-60 ppt |
| TTHM_ugl | mean trihalomethane concentration (micrograms per liter) | 0-219.20 micrograms/liter |
| Uranium_ugL | mean uranium concentration (micrograms per liter) | 0-379.10 micrograms/liter |
| MCL_TTM | whether MCL for trihalomethanes is exceeded | N/A |

| Column Heading | Variable Description | Data Range |
| --- | --- | --- |
| MCL_Uranium | whether MCL for uranium is exceeded | N/A |
| MCL_Arsenic | whether MCL for arsenic is exceeded | N/A |

Figure 1 provides a high level summary of the data provided in the processed dataset. It should be noted that PFAS data was only available for 2013-2015. For ease of analysis, this date range was changed to 2014 during the raw dataset wrangling process to create a common annual unit of analysis for all variables. Figure 2 provides descriptions of all variables included in the processed dataset with data ranges provided for continuous variables.

# 3 Exploratory Analysis

While the distributions of all variables in the dataset were explored during data wrangling and exploration, this section highlights key results from the data exploration process. As research questions focus on arsenic and PFAS concentrations, this exploratory analysis will primarily explore data for these two variables.

## 3.1 Data Exploration for Southeastern States

Table 3: Summary Statistics for Southeastern State Variables

| **Parameter** | **Mean** | **Data Range** |
|---|---|---|
| MHI | $40,848 | $16,435-$92,097 |
| Population served by CWS | 14,893 people | 0-2,300,000 people |
| Arsenic | 410.6 micrograms/liter | 1.0-2,395.0 micrograms/liter |
| Uranium | 0.681 micrograms/liter | 0-23.84 micrograms/liter |
| PFAS | 31.71 ppt | 7.0-59.00 ppt |
| Trihalomethane | 17.47 micrograms/liter | 0-80.00 micrograms/liter |

Southeastern states examined include counties with a large range of income levels and water system sizes; additionally, arsenic concentrations vary more than any other contaminant examined (Table 3).

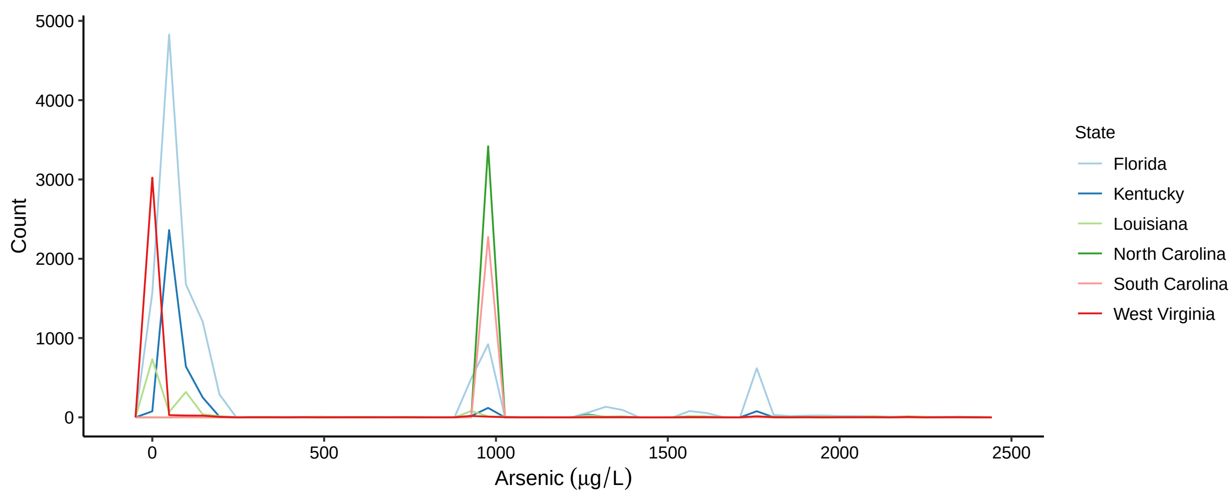


Figure 1: Frequency of Arsenic Concentration Data in Southeastern states.

West Virginia, Florida, Louisiana, and Kentucky all have relatively high counts of arsenic observations at low concentration levels (Figure 1). At the same time, North Carolina, South Carolina, and Florida have relatively high counts of arsenic observations at higher concentrations (around 1,000 micrograms/liter). This exploration of arsenic data in southeastern states justifies an examination of how arsenic concentrations vary among southeastern states and which explanatory variables might predict arsenic concentrations.

```
## Warning: Removed 26290 rows containing non-finite values (stat_bin).
```

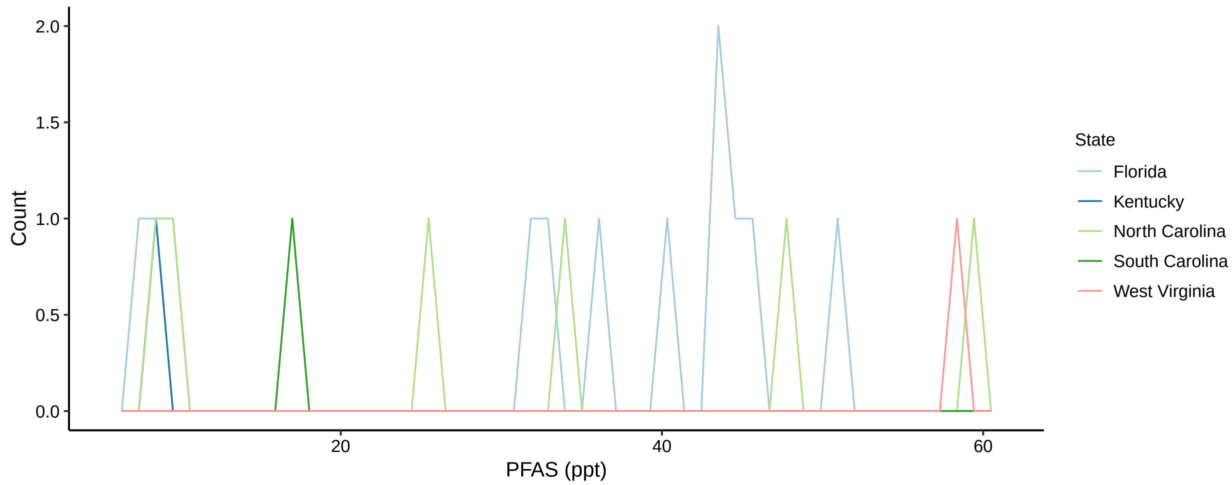


Figure 2: Frequency of PFAS Concentration Data in Southeastern states.

Compared to arsenic data, PFAS concentration data is far more limited in southeastern states (Figure 2). The largest frequency of counts occurs in Florida at around 45 ppt with only two counts available. This limitation of data on PFAS concentrations justifies a separate analysis of PFAS data at a nationwide scale as examining PFAS data regionally severely limits the availability of data.

## 3.2 Data Exploration for Northeastern States

Table 3: Summary Statistics for Northeastern State Variables

| Parameter | Mean | Data Range |
|---|---|---|
| MHI | $54,513 | $26,323-$113,336 |
| Population served by CWS | 13,791 people | 0-8,271,000 people |
| Arsenic | 359.2 micrograms/liter | 1.0-2,2,422.0 micrograms/liter |
| Uranium | 1.75 micrograms/liter | 0-43.00 micrograms/liter |
| PFAS | 29.76 ppt | 3-60.00 ppt |
| Trihalomethane | 10.37 micrograms/liter | 0-134.10 micrograms/liter |

Summary statistics for the northeastern United States indicate the the median household income is higher than that in the southeast and that while the average water system size is similar across regions, the northeast has systems to serve upwards of 8 million people, as compared to a maxiumum size of 2 million people in the southeast (Tables 3 and 4). Ranges and average values for water quality indicators were relatively similar in both regions (Tables 3 and 4).

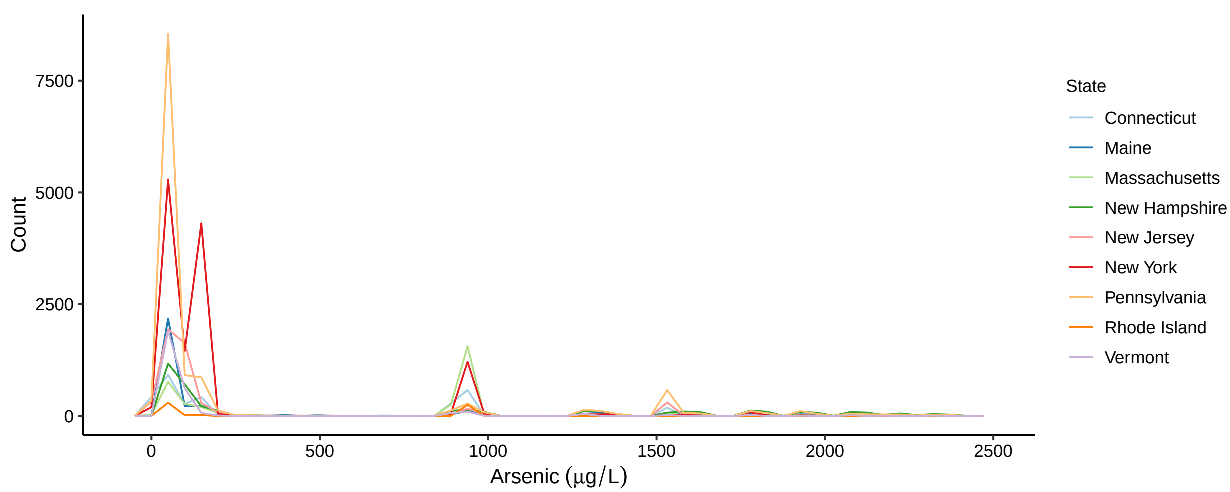


Figure 3: Frequency of Arsenic Concentration Data in Noutheastern states.

All northeastn states appear to have relatively high counts of arsenic observations at low concentration levels, with New York and Pennsylvania maintaining the highest counts (Figure 3). Additionally, Massachusetts, New York, and Connecticut, among other states also have many observations at elevated arsenic concentrations (around 1,000 micrograms/liter). Similarly to southeastern states, the exploration of arenic data in northeastern states justifies a comparison of arsenic concentrations across northeastern states. Additionally, understanding which explanatory variables predict arsenic concentrations, particularly in states with elevated levels, will be explored in the subsequent analysis performed.

```
## Warning: Removed 49146 rows containing non-finite values (stat_bin).
```

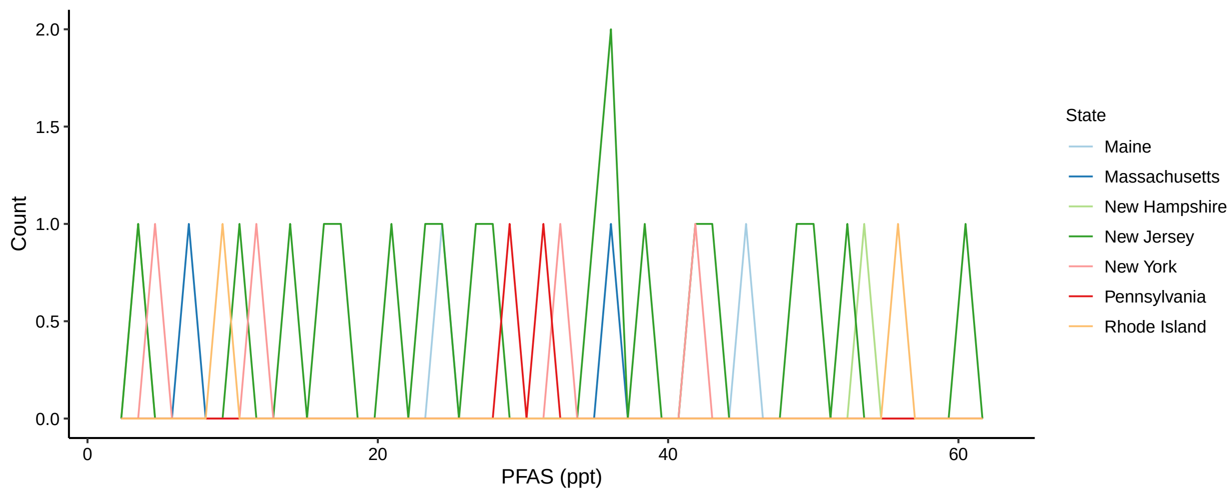


Figure 4: Frequency of PFAS Concentration Data in Noutheastern states.

Much like in southeastern states, PFAS concentration data in northeastern states is very limited (Figure 4). The largest frequency of counts occurs in New Jersey at around 38 ppt with only two counts available. Again, this data limitation provides rationale for examining PFAS data on a nationwide scale, rather than regionally, which will be conducted in the subsequent analysis.

## 3.3 Case Studies: Data Exploration for North Carolina and Massachusetts

North Carolina and Massachusetts were selected as individual case studies due to the prevalence of arsenic data at high concentrations revealed earlier in the exploratory analysis. To gain insight into how arsenic concentrations vary with respect to the variation of other contaminants, arsenic, uranium, and trihalomethane data are provided below across the data collection time span (1999-2018).

```
## Warning in as_grob.default(plot): Cannot convert object of class data.frame into
## a grob.

## Warning: Removed 2865 rows containing non-finite values (stat_smooth).

## Warning: Removed 2865 rows containing missing values (geom_point).

## Warning: Removed 2810 rows containing non-finite values (stat_smooth).

## Warning: Removed 2810 rows containing missing values (geom_point).

## Warning: Graphs cannot be horizontally aligned unless the axis parameter is set.
## Placing graphs unaligned.
```

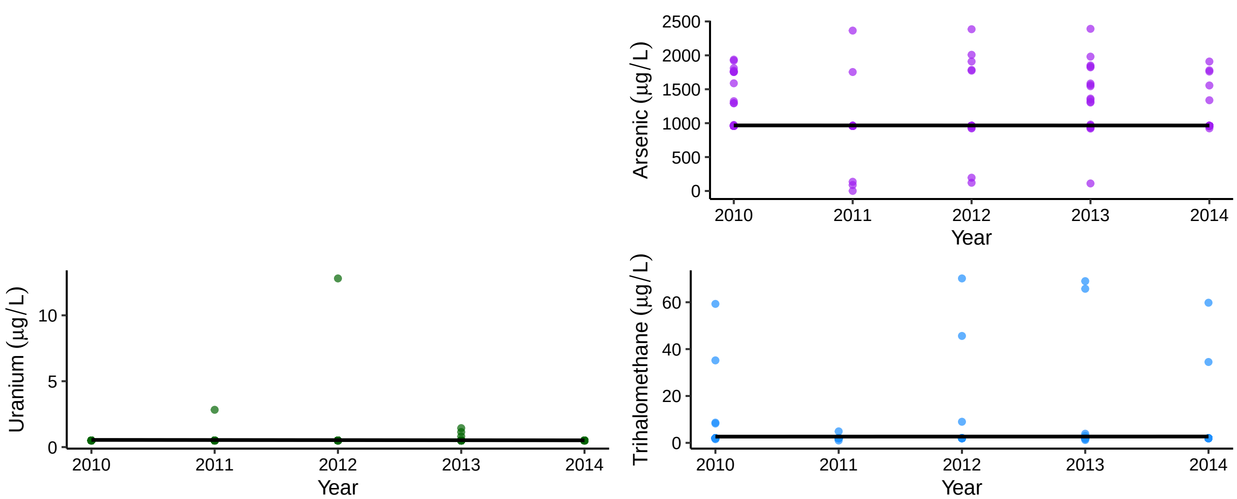


Figure 5: Water contaminant concentrations over time in North Carolina.

Over the period of 1999-2018 examined, North Carolina experiences consistent levels of arsenic, uranium, and trihalomethane (Figure 5). The maximum cotaminant levels (MCL) for these three contaminants, set by the US EPA, are 10 micrograms/liter, 30 micograms/liter, and 80 micrograms/liter respectively. As such, uranium and trihalomethane concenntratinos in North Carolina over this period of time remain safely below the MCL standard. Arsenic concentrationns, on the other hand, appear to occur well above the MCL for the entire period of time examined.

```
## Warning in as_grob.default(plot): Cannot convert object of class data.frame into
## a grob.

## Warning: Removed 3509 rows containing non-finite values (stat_smooth).

## Warning: Removed 3509 rows containing missing values (geom_point).

## Warning: Removed 3321 rows containing non-finite values (stat_smooth).

## Warning: Removed 3321 rows containing missing values (geom_point).

## Warning: Graphs cannot be horizontally aligned unless the axis parameter is set.
## Placing graphs unaligned.
```

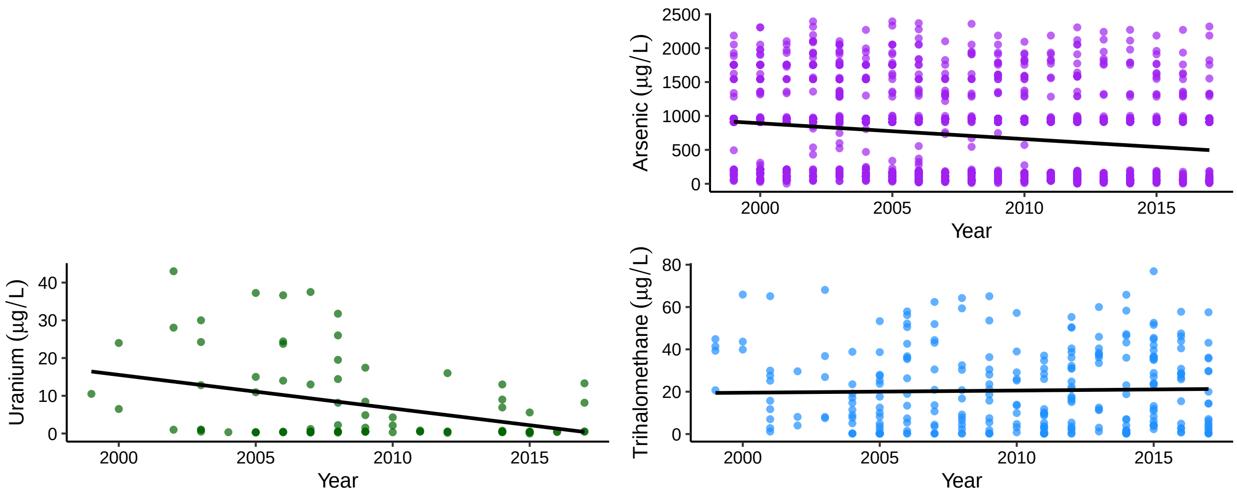


Figure 6: Water contaminant concentrations over time in Massachusetts.

Over the same period of time, Massachusetts experiences slightly declining arsenic concentrations, substantially declining uranium concentrations, and relatively constant trihalomethaen concentrations (Figure 6). Trihalomethane concentrations are safely below the MCL standard, as are uranium concentrations from about 2010-2018. Conversely, arsenic concentrations remain high over the entire 19-year time span examined.
This exploratory analysis has highlighted that arsenic concentrations likely vary among states in the southeast and northeast and that elevated arsenic levels occur in both North Carolina and Massachusetts. As such, the subsequent analysis will focus primarily on examining arsenic concentrations. Additionally, limited PFAS data highlighted during this data exploration process motivates a separate analysis of PFAS data during the analysis phase.

# 4 Analysis

The following analysis seeks to answer the three questions stated at the onset of this report. Key results of statistical analyses performed will be provided below.

## 4.1 Question 1: Do arsenic concentrations vary significantly from state to state in northeastern and southeastern states?

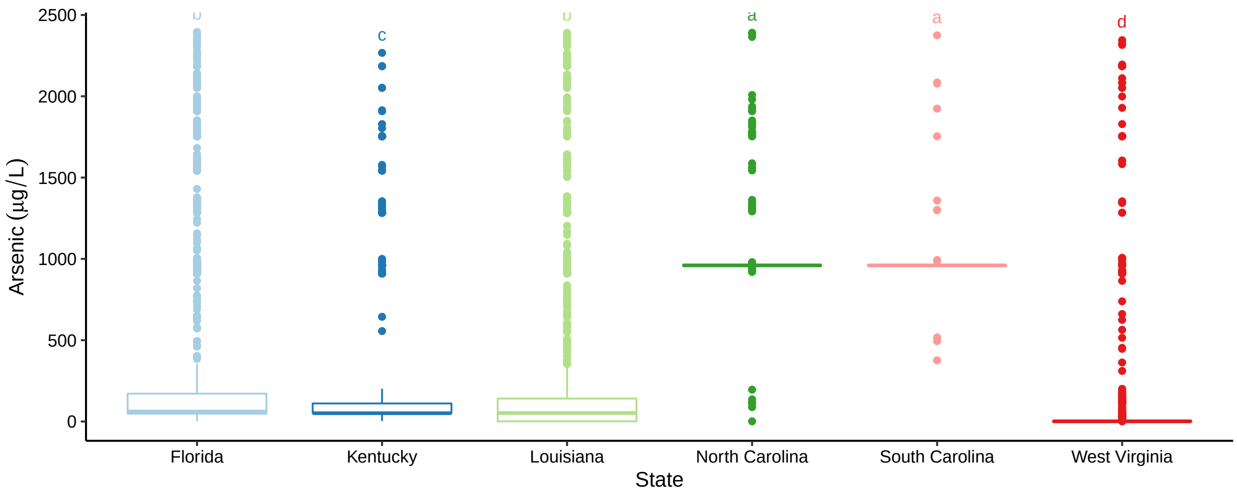


Figure 7: State by state comparison of arsenic values for southeastern states.

Mean annual arsenic concentrations differ significantly between states in the southeast (ANOVA; df=5, F=2872, p <0.0001). Mean arsenic concentrations in West Virginia were significantly lower than in other states and those in North Carolina and South Carolina were significantly higher than those in other states (Post-hoc Tukey test; Figure 7).
Meaan annual arsenic concentrations also differ significantly between states in the northeast (ANOVA; df=8, F=630.9, p<0.0001). Mean arsenic concentrations in New York and Vermont were significantly lower than in other states and those in New Hampshire and Massachusetts were significantly higher than those in other states (Post-hoc Tukey test; Figure 8).

## 4.2 Question 2: Do socioeconomic factors or the presence of other contaminants predict arsenic concentrations in Massachusetts and North Carolina?

For both North Carolina and Massachusetts, uranium was explored as an explanatory variable but was ultimately removed from both models for improved parsimony. For both states, final variables included to explain

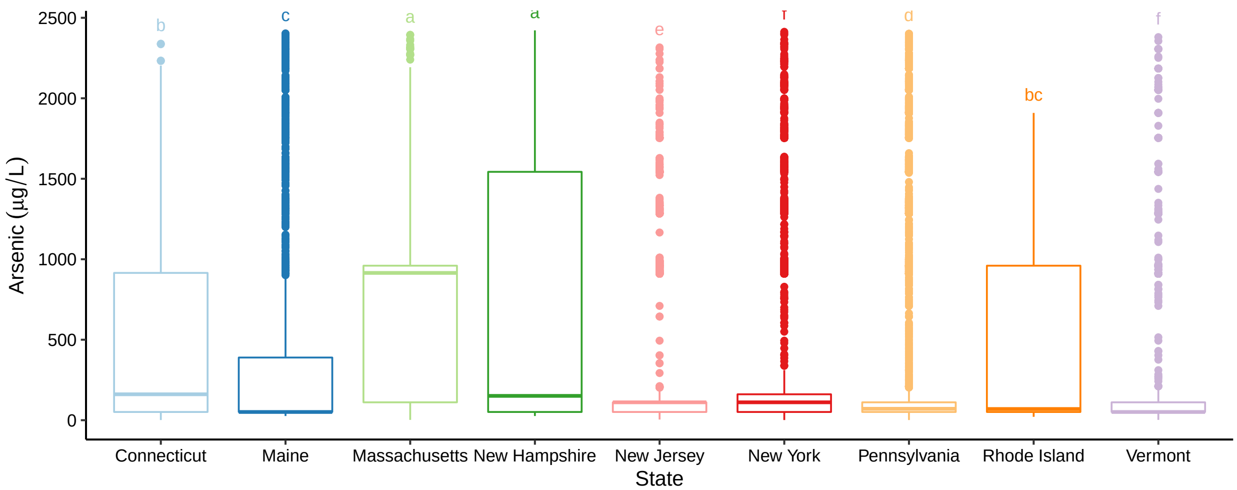


Figure 8: State by state comparison of arsenic values for northeastern states.

variation in arsenic concentration are trihalomethane concentration, MHI, and population served by the CWS.

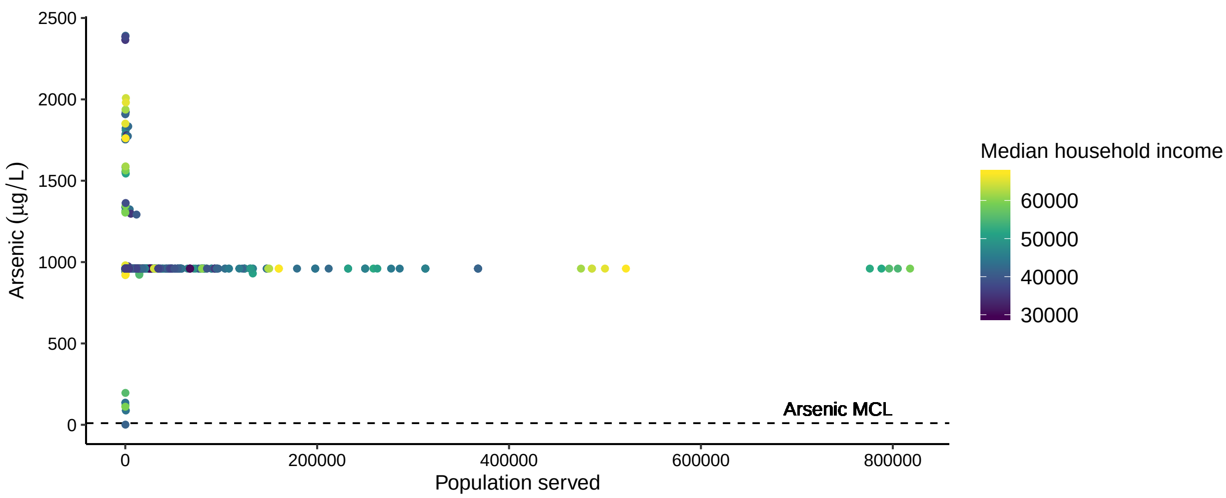


Figure 9: North Carolina Arsenic Concentrations by Population Served by CWS Across Income Levels.

In North Carolina, population served by a CWS and trihalomethane concentrations significantly predict arsenic concentrations, whereas MHI is not a significant predictor of arsenic concentration (Multiple linear regression; df=3 and 654, F=34.74, p<0.0001). Inreasing arsenic concentration is associated with increasing trihalomethane concentration and with decreasing population size served by a CWS. Additionally, there is no discernible relationship between arsenic concentration and MHI (Figure 9).

```
## Warning: Removed 2810 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2810 rows containing missing values (geom_point).
```

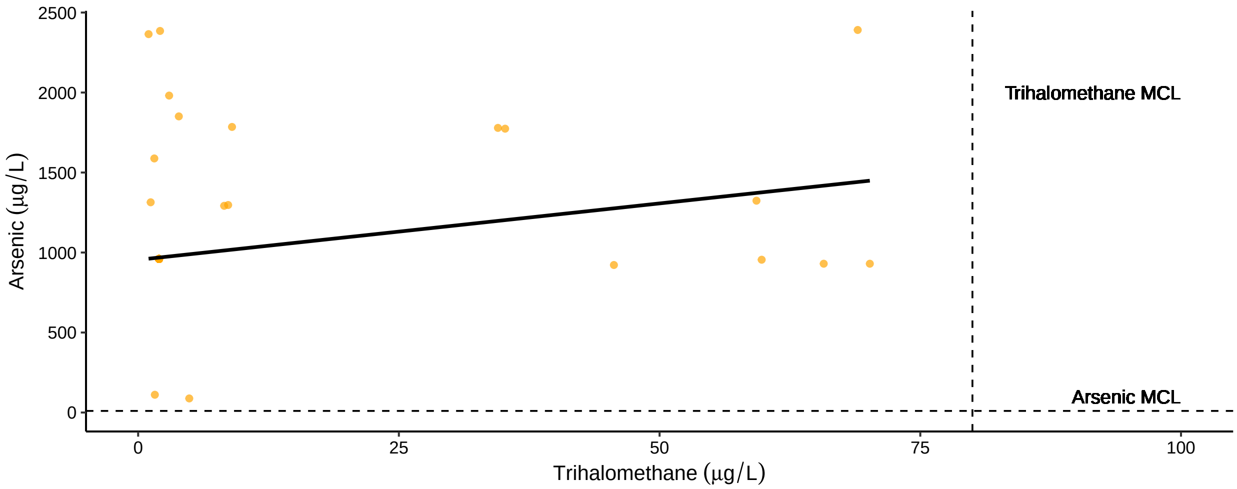


Figure 10: North Carolina Arsenic Concentrations by Trihalomethane Concentration.

```
## Warning: Removed 18 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 18 rows containing missing values (geom_point).
```

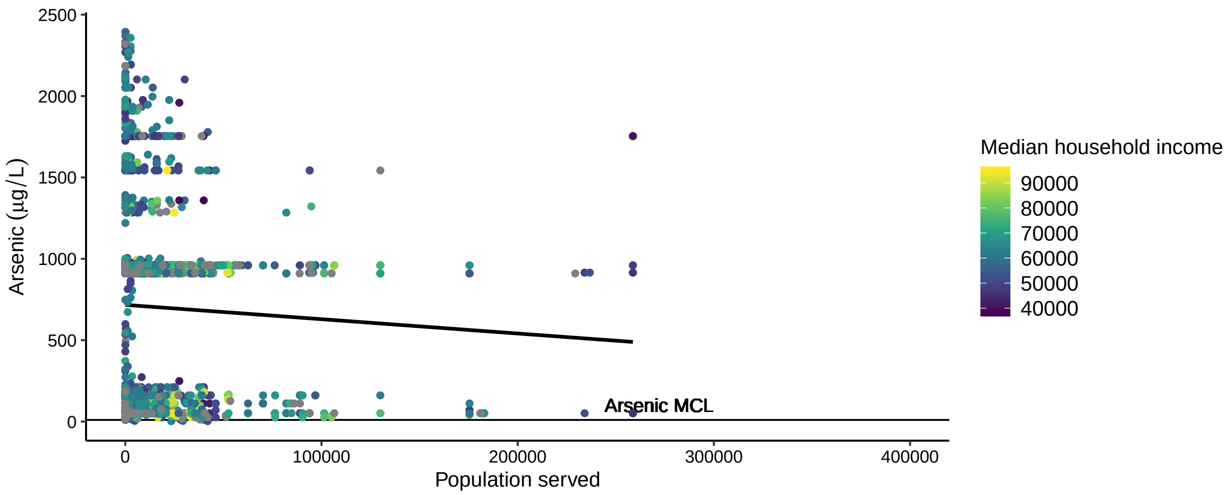


Figure 11: Massachusetts Arsenic Concentrations by Population Served by CWS Across Income Levels.

In Massachusetts, neither population served by a CWS, trihalomethane concentrations, nor MHI significantly predict arsenic conncentrations (Multiple linear regression; df=3 annd 242, F=1.664, p=0.1753). A large range of arsenic cooncentratinos is experienced both across different population sizes served by CWSs annd across MHI levels (Figure 10).

```
## Warning: Removed 3321 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3321 rows containing missing values (geom_point).
```

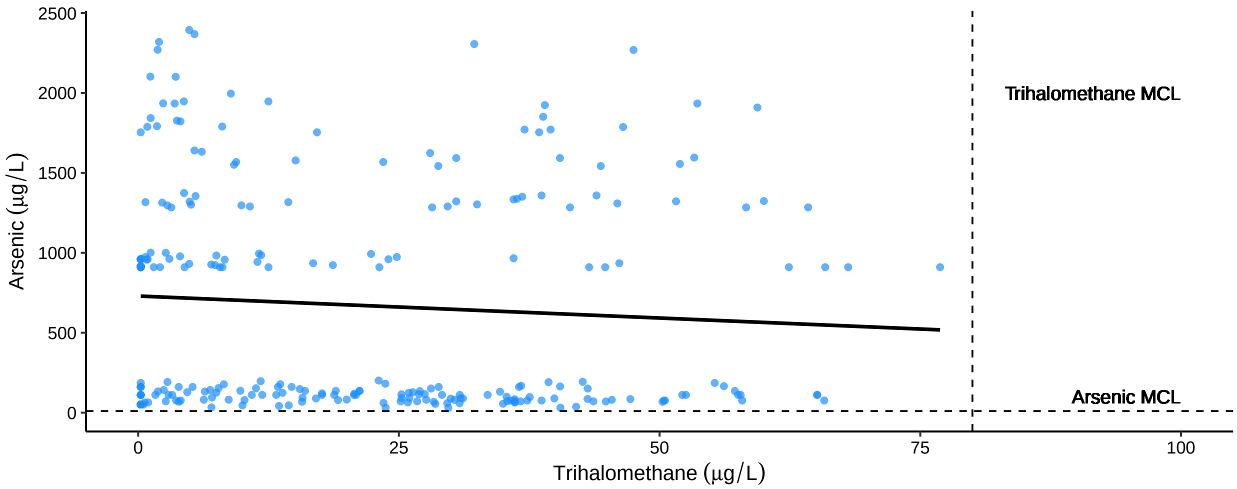


Figure 12: North Carolina v. Massachusetts: Arsenic Concentrations by Trihalomethane Concentration.

Multiple linear regression results discussed previously indicated that trihalomethane concentrations significantly predict arsenic concentrations in North Carolina but not in Massachusetts (Figure 11).

## 4.3 Question 3. Do PFAS concentrations vary signficantly across time and from state to state in the United States? Are socioeconomic factors significant predictors of PFAS concentrations?

Neither MHI nor population served by a CWS significantly predict PFAS concentrations (Multiple linear regression; F=2 and 86, F=0.5912, p=0.5559). An ANOVA was also run in preliminary analyses conducted for PFAS data, but results are not included here due to issues with missingness (25,938 observations were deleted when ANOVA was run). TAKEAWAYS!

```
## Warning: Removed 25946 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 25946 rows containing missing values (geom_point).
```

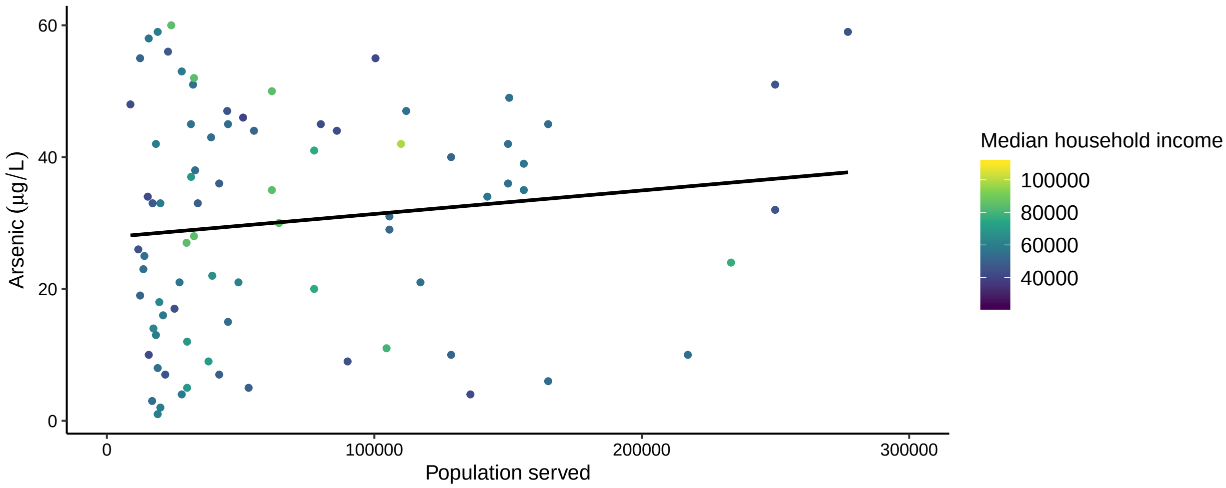


Figure 13: PFAS Concentration by Population Served by CWS Across Income Levels.

# 5 Summary and Conclusions

talk about conclusions and implications of findings. . . and supposition as to why these relationships exist Omitted variable bias: R2s for all models were very low, talk about this Future: given scope of project, couldn't possibly look in depth at each conntamiannt, so i nthe future would be interested in exploring relationships for uranium and TTHM such as those explored here for arsenic. . . it's possible that since arsenic is so ubiquitous inn geology that other contaminants could have diff relationships with income for example. . .

# 6 References

United States Environmental Protection Agency (USEPA). 2020. Safe Drinking Water Act (SDWA). Retrieved from: https://www.epa.gov/sdwa.