Public Water System Contamination Prediction

Team 125:
Anzar Chowdury, Brandon Gunasti, Ethan Kurtz

Background & Motivation

Background: Managing water quality is difficult to do efficiently as there are many needs to be met by the government agencies from The Safe Drinking Water Act . Allocating the limited resources available efficiently while simultaneously treating for multiple contaminants across vast water system networks is difficult to do effectively.

Motivation: Address the uncertainty of contamination risk in water systems using the list of unmonitored contaminants from the Fifth Unregulated Contaminant Monitoring Rule. Help with the efficiency of this testing.

Problem Statement

Problem: Water quality management faces the challenge of efficiently allocating limited resources for monitoring and treatment across numerous contaminants and vast water system networks. The primary problem is the uncertainty regarding where and when specific contaminants are likely to pose a risk, requiring a data-driven approach to prioritize efforts.

Intended Impact

Resource Optimization: By predicting contamination risks, water authorities can better allocate their resources, focusing monitoring and treatment efforts where they are most needed, thereby improving efficiency. Referenced contaminants are not currently regulated, but are still important to monitor for health standards.

Public Health Protection: Early detection and proactive management of contamination risks can significantly reduce public exposure to harmful contaminants, safeguarding community health.

Regulatory Compliance: The model can aid PWSs in achieving and maintaining compliance with water quality regulations by identifying potential vulnerabilities and guiding corrective actions before regulatory limits are exceeded.

Environmental Insight: Analyzing the factors that contribute to contamination risks can offer insights into environmental impacts on water quality, such as industrial discharge, agricultural runoff, or natural geological features, informing broader environmental protection efforts.

Future Planning: The insights gained from the predictive model can inform long-term planning and investment in water infrastructure and treatment technologies, guiding decisions on upgrades, expansions, or new installations to address identified risks.

Dataset Overview

Data comes from the U.S. Environmental Protection Agency (EPA) collected under the Fifth Unregulated Contaminant Monitoring Rule (UCMR)

The current data is as of January 2024, and will be updated quarterly until 2026. Current data is 24% of the planned total to be collected

Data is updated based on further review by analytical laboratories, public water systems, states, and the EPA

The UCMR program to collect nationally representative data for contaminants (such as Lithium and PFAs) that may be present in drinking water but are not yet subject to regulatory standards set under Safe Drinking Water Act (SDWA)

Data is formatted with each row as a test for a contaminant per water system

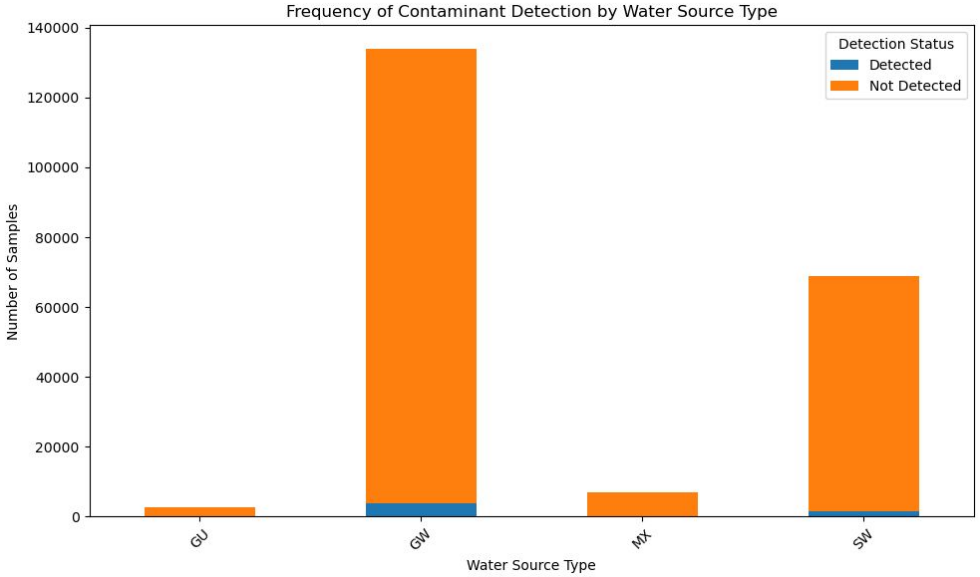
Data Ethical Concerns

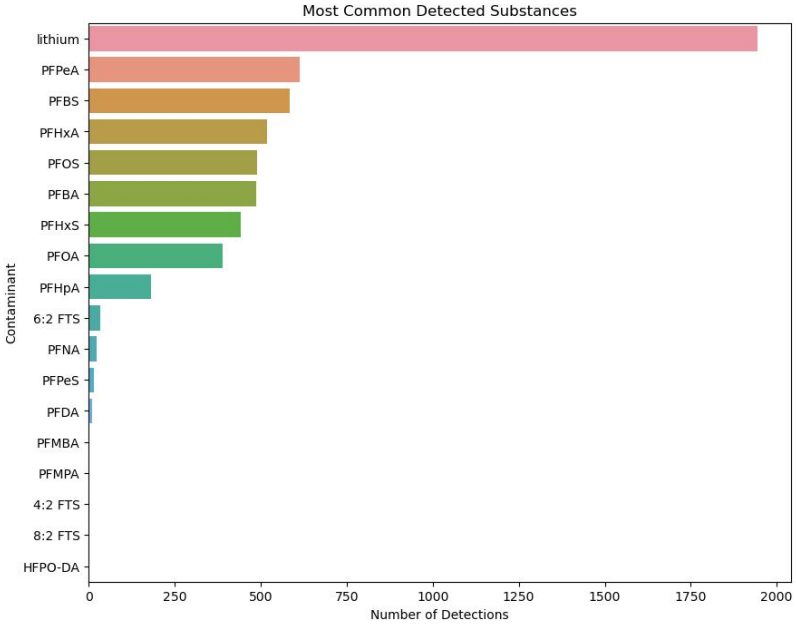
The data appears to have no ethical concerns as to how it was collected. All are in line with government guidelines, and are collected by a government agency for the fulfillment of government acts passed to keep US drinking water clean and safe for consumption.

Methodology

1. Utilize pandas dataframe to parse and format EPA contamination data
2. Removed irrelevant data columns (mostly relating to administrative codes
3. Visualize dataframe to convey frequency of contamination by water source type
4. Visualize most common detected substances within water
5. Calculate number of detections per state and then visualize
6. Create binary column within dataframe that that determines if each specific contaminant was found in water source, replacing initial detection column
7. Create column for collection month, to account for seasonal variables
8. Train a logistic regression model for each contaminant to predict whether it is detected

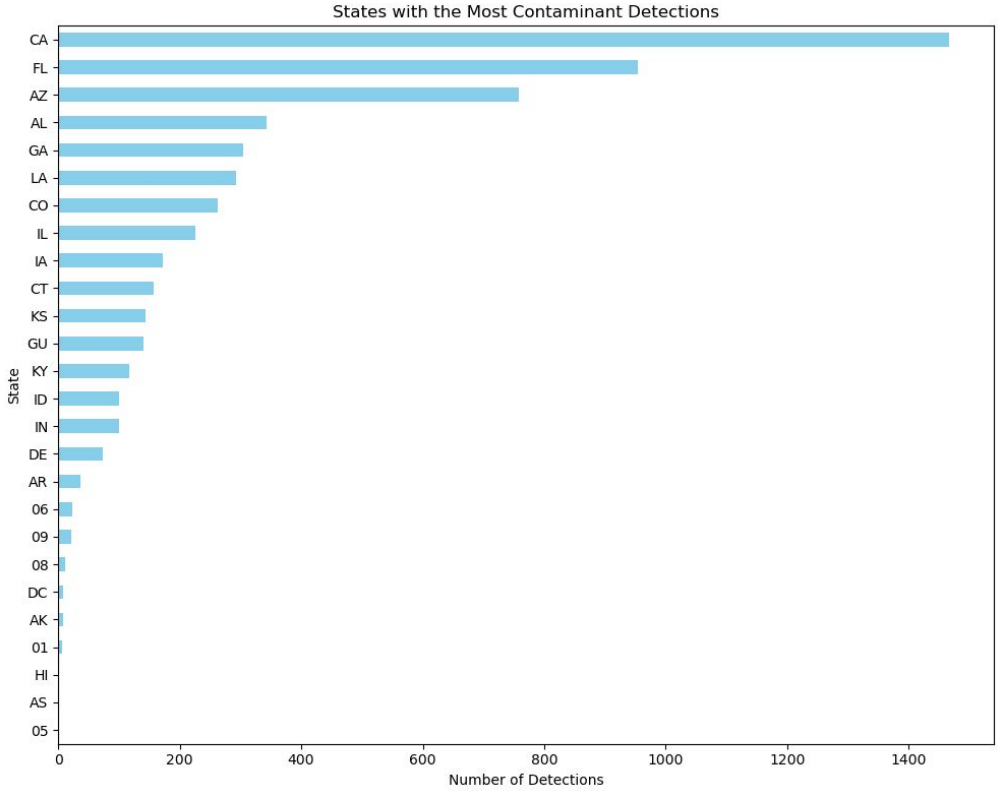
Results





GU: GW Under Influence of SW | **GW**: Ground Water | **MX**: Combination of others | **SW**: Surface Water

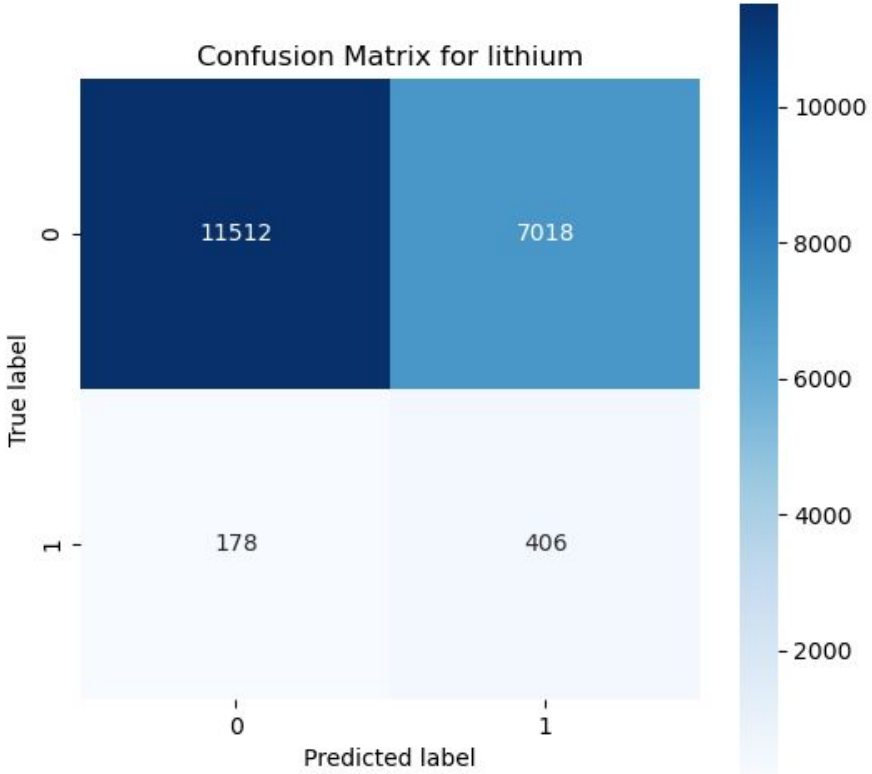
Results (continued)



Results of Logistic Regression

- **Majority Class Recall:** 0.58-0.77, suggesting that the model misses some true negatives
- **Minority Class Recall:** Higher, suggesting the model is better at identifying most of the true positives, but this is likely at the cost of incorrectly classifying many negatives as positives
- **Support Values**: Indicate a significant class imbalance for all contaminants, with the majority class (0, or not detected) having much larger sample sizes compared to the minority class (1, or detected)
- **Majority Class F1-score**: Relatively high for the majority class due to the high precision and moderate recall
- **Minority class F1-score**: Very low across all contaminants, which indicates a poor balance between precision and recall — the model struggles to predict the presence of contaminants accurately
- **Accuracy**: 0.62%-0.77%

Results Continued



Training model for: lithium

	precision	recall	f1-score	support
0	0.98	0.62	0.76	18530
1	0.05	0.70	0.10	584
accuracy			0.62	19114
macro avg	0.52	0.66	0.43	19114
weighted avg	0.96	0.62	0.74	19114

Limitations & Future Work

Limitations

- Models are relatively proficient at identifying samples without contaminants but they struggle to accurately predict when contaminants are present
- The discrepancy in detection rates of contaminants led to a major class imbalance, so the model doesn't predict when contaminants are present as accurately as when there are no contaminants
- The model should not be used in a professional setting unless more balanced datasets with more positive detection cases arise

Future Work

- A more updated version of the data may provide better outcomes as the model would be better trained using a completed data set
- Monitoring changes in concentrations over time as a result of environmental policy

Citations

5th Unregulated Contaminant Monitoring Rule:
https://www.epa.gov/dwucmr/fifth-unregulated-contaminant-monitoring-rule

Data of 5th Summary Monitoring Rule:

https://www.epa.gov/dwucmr/data-summary-fifth-unregulated-contaminant-monitoring-rule

Drinking Water Standards & Regulations:

https://www.cdc.gov/healthywater/drinking/public/regulations.html#:~:text=Under%20the%20SDWA%2C%20EPA%20sets,contaminants%20in%20public%20drinking%20water.

THANK YOU!

Any Questions?