CS 6463
Project Proposal

Kyle Bailey

Project Task and Importance

Project 2: Forecast PFAS Data

Goal: The goal of this project is to train a model to forecast future levels of PFAS contamination in different regions using historical PFAS data. The objective is to explore how well models can predict future growth or changes in PFAS levels based on historical trends.

What is PFAS and Why is Forecasting Important?

PFAS (Per- and Polyfluoroalkyl Substances) are harmful man-made chemicals that are highly persistent in the environment and human body. Forecasting future PFAS levels can help authorities plan for future contamination risks, implement protective measures, and inform the public about potential exposure hazards.

Data Description

The Unregulated Contaminant Monitoring Rule (UCMR) serves as a data profiling mechanism for the EPA to gather information on suspected drinking water contaminants not yet regulated by the Safe Drinking Water Act (SDWA). This profiling process involves:

1. Data collection: Gathering occurrence data for potential contaminants in drinking water systems.
2. Representative sampling: Ensuring the collected data provides a nationally representative picture of contaminant presence.

This data profiling approach allows for a comprehensive understanding of emerging contaminants in drinking water, forming a basis for informed decision-making regarding water quality and public health.

For this analysis, we will be using available UCMR data from the EPA [1] , for UCMR 1 2001-2005, UCMR 2 2008-2010, UCMR 3 2013-2015, UCMR 4 2018-2020, and UCMR 5 2023-2024. (Data was last refreshed July 11, 2024.) Because this analysis is for predicting PFAS levels for 2024, any 2024 data will be used for testing our model through the last refresh date. The model will then predict PFAS levels by region for the remainder of 2024. The UCMR 1-5 occurrence data is labeled, and the text files all have the same data fields.

Each dataset includes:

- Public Water System (PWS) information
- Contaminant levels for various PFAS substances
- Water system type (e.g., groundwater, surface water)
- Collection dates for each sample
- Geolocation information for each public water system

Expected Outcomes

Deliverables:

Trained forecasting models and code: The model will be trained using PFAS historical data from 2001-2023 to predict future levels in 2024. We hope to use machine learning algorithms like random forest and deep learning approaches like LSTM.

Research Report: A detailed report on the methodologies used, prediction output, and discussion of our findings. We will also add metrics on Error Analysis showing where the model generates the largest errors.

Final Poster: A poster visually summarizing the above.

Evalution Metrics:

Prediction Accuracy: We will use metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared to evaluate how closely the model's forecasts match actual PFAS levels.

Generalization Ability: Evaluate how well the model generalizes to new regions not included in the training data using out-of-sample testing.

Anticipated Challenges

- There is over twenty years of data provided, broken up by region, state, and method. We must gain a high-level understanding of water contaminants to evaluate which datasets and features are relevant.
- Next, we will be challenged with determining what type of model is best suited to the task. We need to evaluate various models, ranging from classical machine learning algorithms to neural networks, to find the best fit.
- We will need to determine how we can measure accuracy since most of the 2024 data that we need to predict will not be available.
- Once we know how to measure accuracy, it will be challenging to find ways to increase accuracy in the time allotted.

[1] https://www.epa.gov/dwucmr/occurrence-data-unregulated-contaminant-monitoring-rule