Groundwater Contamination of Per- and Polyfluoroalkyl Substances in the United States–Insights from a Random Forest Model

Bumjun Park[a], William Gnesda[b], Christopher Zahaksy[b]

[a]Department of Statistics, University of Wisconsin–Madison; [b]Department of Geoscience, University of Wisconsin–Madison

Link to my personal website!

1. Introduction

Per- and polyfluoroalkyl substances (PFAS) are a group of synthetic pollutants that have been increasingly found in groundwater in communities across the United States, and thus have been drawing growing interest and concern. The concentration of PFAS in water systems is influenced by a multitudes of factors, namely the proximity to airports, military bases, landfills, or an assortment of manufacturing facilities, as well as geographic conditions such as the climate or population density. In this work, a random forest machine learning model accounting for these factors is applied to assess the likelihood of having hazardous levels of PFAS concentration, greater than 50 parts per trillion(ppt), in various geographic points across the contiguous United States. Based on the model, further geospatial analysis is conducted by interpolating the model's predictions to create a national risk map that highlights the most susceptible areas. The risk map can serve as a guideline for future water sampling investigations into PFAS contamination for different agencies and policymakers.

2. Methodology

The random forest model is trained on 10033 points of PFAS measurements distributed across the contiguous United States (shown in **Figure 1**). Each point consists of 19 different variables measuring distances to various PFAS sources, geographic characteristics, and most importantly, the PFAS levels (shown in **Table 1**). The model then creates 2000 different decision trees, one of which is partially shown in **Figure 2**, ascertaining whether the point has hazardous PFAS levels, greater than 50 ppt, or not. Then, Triangulated Irregular Network, as shown in **Figure 3**, is applied to create a national PFAS risk map.

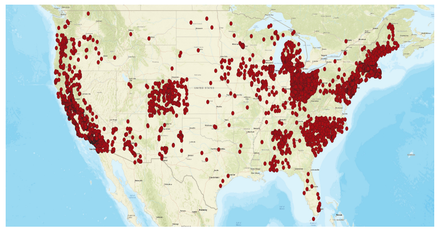

Figure 1: Distribution of data points used in random forest model. Each point represents an observation site from which PFAS levels were measured. Data was gathered from multiple different sources, such as the Environmental Working Group[1], California Groundwater Ambient Monitoring Assessment Program[2], and various state-level agencies.

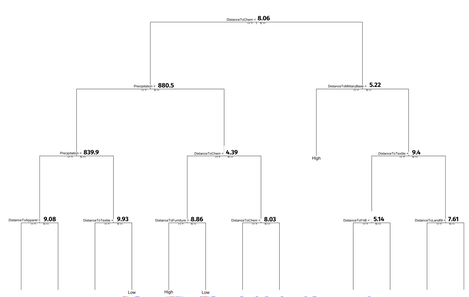

Figure 2: Snippet of a decision tree used in the random forest model. Based on the provided data, the random forest model randomly subsets some of the data, and creates 2000 decision trees. The final prediction is made by compiling the "votes" of these individual decision trees.

Variable	Unit
Distance to Airport	log(meters)
Distance to Military Base	log(meters)
Distance to Landfill	log(meters)
Distance to Manufacturers	log(meters)
- Agricultural	
- Apparel	
- Chemical	
- Food and Beverage	
- Furniture	
- Leather	
- Machinery	
- Miscellaneous	
- Paper	
- Rubber	
- Textile	
- Wood	
Soil Organic Carbon	Tons/HA
Population Density	
Bureau of Land Management	T/F
PFAS Concentration	ppt

Table 1: List of variables considered by the random forest model. Data was gathered from international and federal sources such as the UN[3], NASA[4], and the EPA Facility Registry Service[5].

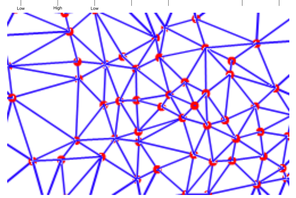

Figure 3[6]: An illustration of Triangulated Irregular network. The random forest model is extrapolated to 20000 randomly generated points in the United States. And based on the extrapolated points, each three adjacent points are "triangulated" to find the PFAS risk in each traingle.

3. National Risk Map

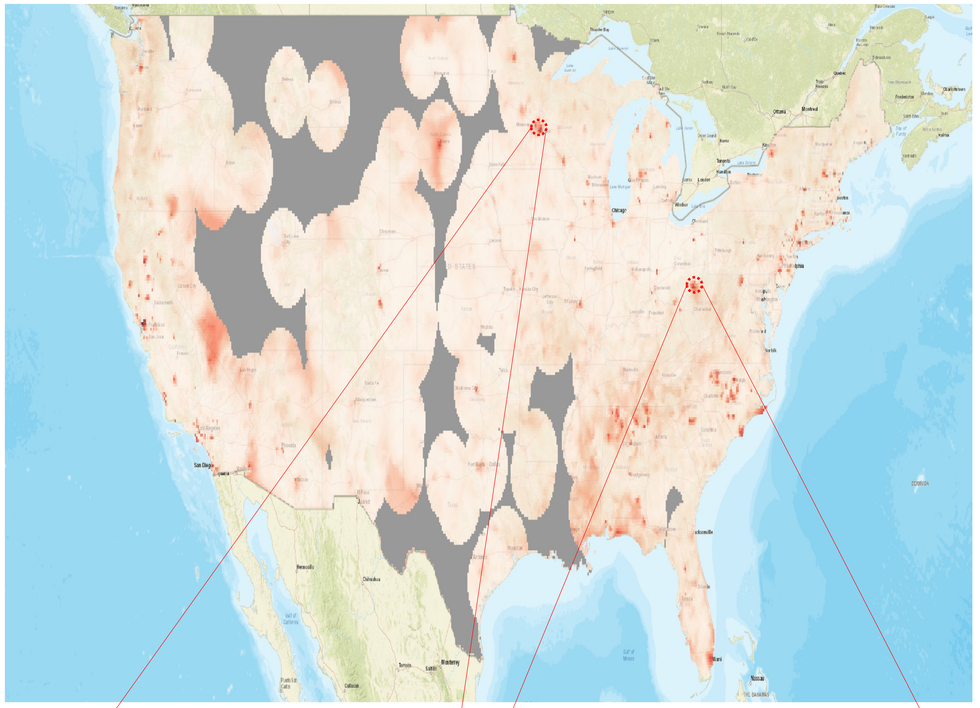

Figure 4: PFAS risk map over contiguous United States. Using the Triangulated Irregular Network method, the PFAS risk of every geographic point over the contiguous United States is calculated. The red areas represent the regions with highest PFAS risk. Areas that did not have sufficient data to begin with have been grayed out, as the model's predictions are unreliable in those areas.

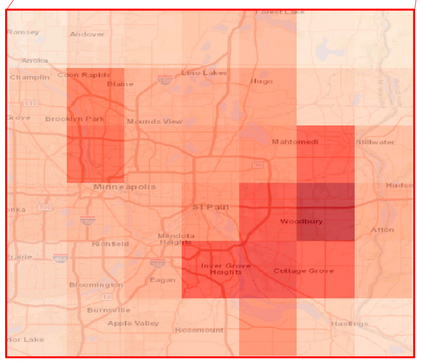

Figure 5: A closer look into Saint Paul, MN, one of the areas in the United States known to have high levels of PFAS

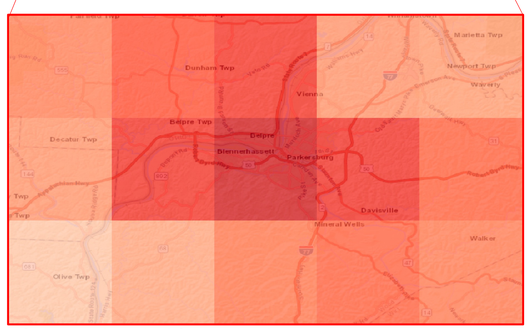

Figure 6: A closer look into Parkersburg, WV, another area known to have high levels of PFAS. This is the site of the film Dark Waters (2019), which deals with water contamination.

4. Model Evaluation

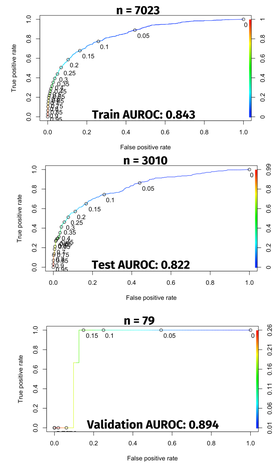

Figure 7: Receiver Operating Characteristic(ROC) curves of the random forest model. The curves display the True Positive Rate, the probability of correctly capturing all points with high PFAS, against the False Positive Rate, the probability of incorrectly diagnosing a low PFAS point as a high PFAS point. A perfect model shows a perfect rectangular shape, while a completely random model shows a 45 degree line. The Area Under the ROC curve(AUROC) is 1 for a perfect model, and 0.5 for a random model. The given data of 10033 points is subsetted into 70% training data, which is actually used to fit the model, and 30% testing data, which is used for purely evaluation purposes. The training ROC curve(top) and the testing ROC curve(middle) both show high AUROC's. Additionally, a separate validation dataset of 79 points from the state of Wisconsin(bottom) also shows a high AUROC.

Table 2: A table ranking in terms of importance. the variables that the random forest model used. When a given variable was excluded from the model, Distance to Textile Mills, Distance to Apparel Manufacturers, and Distance to Leather Manufacturers showed the greatest decrease in model accuracy, suggesting that they are the more important predictors.

	MeanDecreaseAccuracy
DistanceToTextile	2.60E-02
DistanceToApparel	2.53E-02
DistanceToLeather	2.49E-02
DistanceToAirport	2.06E-02
DistanceToMilitaryBase	2.00E-02
PopulationDensity	1.91E-02
DistanceToRubber	1.89E-02
DistanceToMiscellaneous	1.81E-02
DistanceToPaper	1.72E-02
DistanceToChem	1.69E-02
DistanceToFurniture	1.68E-02
DistanceToWood	1.66E-02
Precipitation	1.61E-02
DistanceToMachinery	1.50E-02
DistanceToAg	1.42E-02
DistanceToFnB	1.34E-02
DistanceToLandfill	1.31E-02
SOCTonsPerHA	8.53E-03
BureauLandManagement	1.44E-05

5. Summary

- The risk map **points out the regions more susceptible to PFAS contamination.** Environmental agencies may consult this map for future PFAS testing and policy designing.
- The model points out the more important predictors in gauging PFAS levels, namely **distances to textile mills, apparel manufacturers, and leather manufacturers**.
- The model has room for improvement as the data is extremely **sparsely distributed** and is **spatially autocorrelated**. Future models may expand the data and adjust the model to account for spatially confounding factors.

6. References

[1] Environmental Working Group. (2022). PFAS Contamination in the U.S. map.
[2] "GAMA Groundwater Information System." *GAMA Online Tools.* California Water Boards, Groundwater Ambient Monitoring Assessment Program
[3] Food and Agriculture Organization of the United Nations. 2017. *Global Soil Organic Carbon Map*
[4] Socioeconomic Data and Applications Center (SEDAC). 2018. *Gridded Population of the World (GPW), V4.* Center for International Earth Science Information Network - CIESIN - Columbia University
[5] "Facility Registry Service Geospatial Data Download Service." n.d. *EPA.*
[6] *ArcMap*. What is a TIN surface?-ArcMap Documentation