Groundwater Contamination of Per- and Polyfluoroalkyl Substances in the United States– Insights from an Ecological Sampling Bias Correction Method

Bumjun Park[a], Hyunseung Kang[a], William Gnesda[b], Christopher Zahasky[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, manufacturing facilities, extent of industrialization, and hydrologic conditions.

In this work, **methods of ecological sampling**[6] are used to account for **both the distribution of PFAS and the sampling biases**. Certain areas are more likely than other areas to be tested for PFAS, which, if unaccounted for, leads to an overrepresentation of PFAS risk in that region. Treating PFAS as an ecological species, the "population distribution" is measured in various geographic points across the contiguous United States. Further geospatial analysis is conducted by interpolating the model's predictions to create a national distribution 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 **Inhomogeneous Poisson Process (IPP) model** is trained on **8308 observations** of PFAS and **20000** randomly generated "potential observation" points across the contiguous United States (shown in **Figure 1**). Each observation point consists of 16 different variables measuring distances to various PFAS sources, geographic characteristics, and most importantly, the PFAS levels (shown in **Table 1**). The model then creates a linear model predicting the "species intensity" of PFAS, which roughly means the expected number of "high PFAS" (greater than 50 ppt) observations we would make. It then adjusts this intensity by accounting for observer bias (shown in **Figure 4**), producing its final prediction (shown in **Figure 3**).

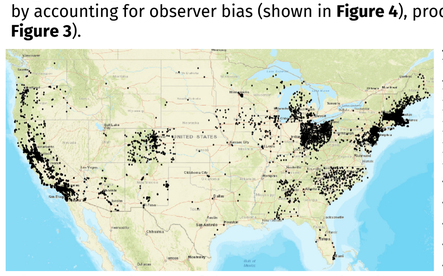

Figure 1: Distribution of data points used in the modelmodel. Each point represents an observation site from which PFAS levels were measured. Data was gathered from multiple different sources, such as the Environmental Working Group[1], California Groundwater Ambient Monitoring Assessment Program[2],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

Variable	Unit
Distance to Airport	log(meters)
Distance to Military Base	log(meters)
Distance to Landfill	log(meters)
Distance to Manufacturers	log(meters)
- Chemical	
- Rubbet	
- Textile	
- Apparel	
- Food and Beverage	
- Furniture	
- Leather	
- Paper	
- Miscellaneous	
Precipitation	mm/year (30 years average)
Population Density	
Median Earnings at Zipcode	Dollars (2020)
PFAS Concentration	ppt

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

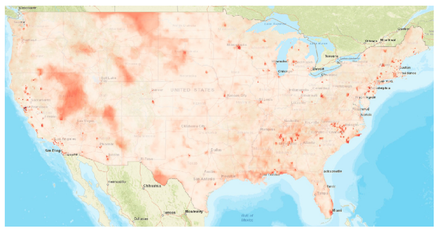

Figure 2: The same PFAS risk map created using a random forest model, one of the more popular machine learning approaches today. The random model **does not account for observer bias**, and thus, compared to **Figure 3**, there are **salient inaccuracies centered around areas where there are insufficient data**. Namely, the desert areas in Nevada and Arizona and the plains of northeastern Montana are arbitrarily highlighted as areas of high PFAS risk.

3. National PFAS Distribution

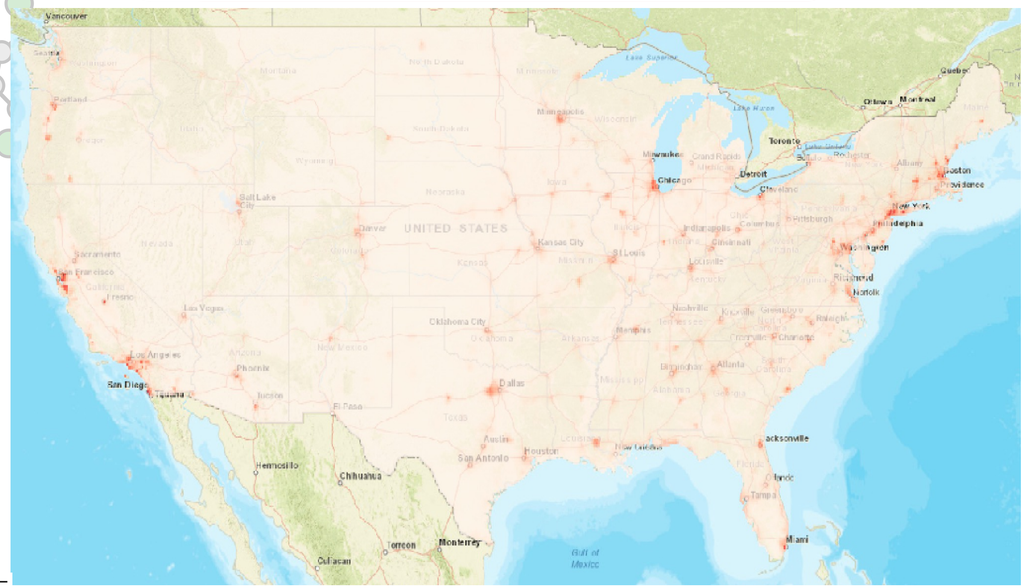

Figure 3: PFAS "**species intensity**" map over contiguous United States. Using the Triangulated Irregular Network (TIN) method, the PFAS risk of every geographic point over the contiguous United States is calculated. The red areas represent the regions with highest amount of expected "high PFAS" observations.

4. National Observer Bias

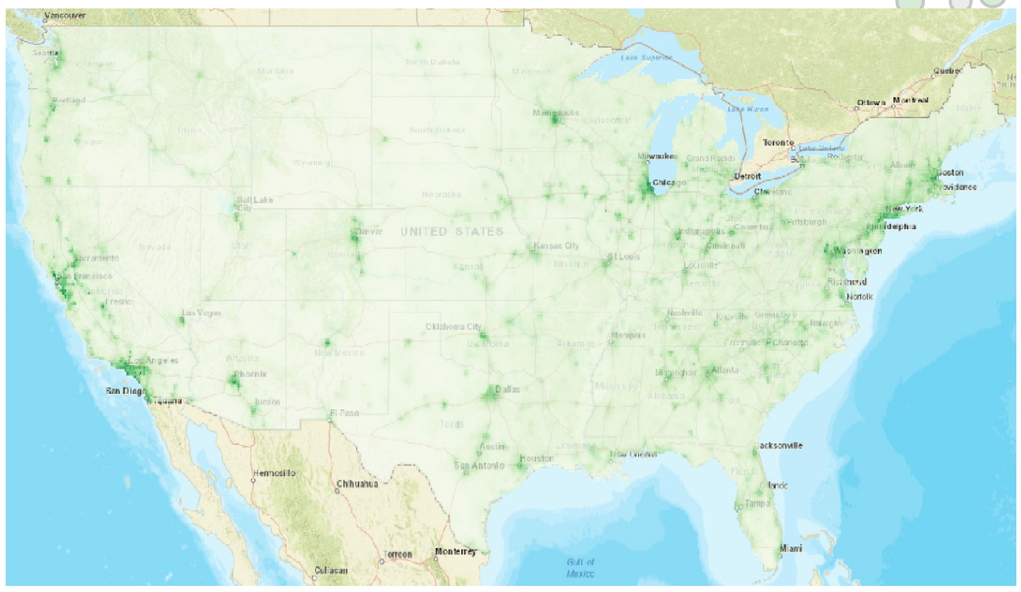

Figure 4: PFAS "**observer bias**" map over contiguous United States. Also using the TIN method, the likelihood of an area being sampled is calculated. The dark-green areas represent the regions which are more likely to be sampled compared to other regions.

5. Model Interpretation

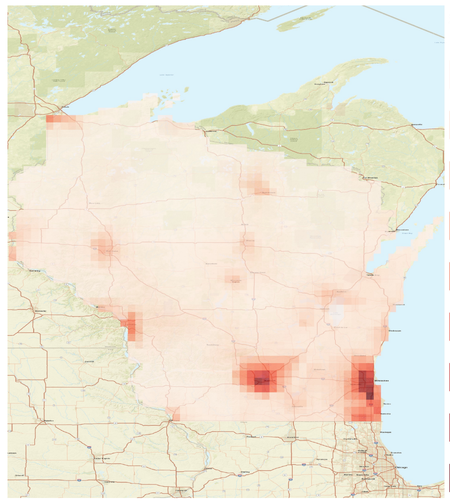

Figure 5: A closer look into the state of Wisconsin, with the colors rescaled to fit the state only .

The darkest pixel in Wisconsin has an intensity of 0.116, meaning that **if we sample groundwater in that pixel 100 times, we would expect to observe PFAS 11.6 times at a concentration above 50 ppt.** In a relative sense, compared to the national scale, a species intensity of 0.116 corresponds to the 99.3 percentile.

6. Summary

- The risk map **points out the regions where we expect to observe the most PFAS in groundwater**, specifically, **the relative risk of observing PFAS greater than 50 ppt at a given point**. Environmental agencies may consult this map for future PFAS testing and policy designing.
- The model accounts for the fact that certain regions, such as **areas with higher populations or areas closer to airports and military bases**, are more likely to be tested and adjusts accordingly.
- The model has room for improvement as it does not fully account for **different testing and reporting standards** of different states. Future models may improve and adjust the model by accounting for various localized features.

7. 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] Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2014). "Bias correction in species distribution models: Pooling survey and collection data for multiple species". *Methods in Ecology and Evolution*, *6*(4), 424–438. https://doi.org/10.1111/2041-210x.12242