Exploring the Relationship Between Social Vulnerability and Well Nitrate Concentration in Iowa Through Interpretable Machine Learning

Sumin Bae[1]; Jinyi Cai[2]; Caglar Koylu, PhD[2]

American Community School of Abu Dhabi, United Arab Emirates[1], Department of Geographical and Sustainability Sciences, University of Iowa, IA[2]

Geo-Social

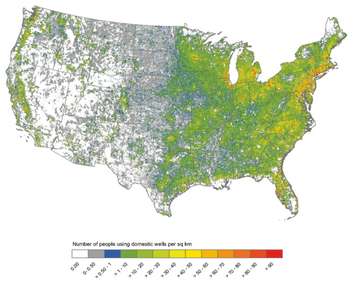
Belin-Blank Center
College of Education
The University of Iowa

Introduction

Background Information

Climate change increases the frequency and severity of **extreme weather events**, such as droughts and floods, leading to higher **nitrate concentrations** in water

An estimated **12-14%** of the US population relies on **unregulated private wells**, primarily in rural areas, putting these populations at greater risk of nitrate contamination

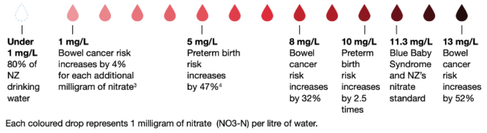


(Source: Johnson TD, etc. 2018)

Problem

Private wells are particularly vulnerable to contamination because they are not covered by the federal **Safe Drinking Water Act**

Nitrate contamination in drinking water has been linked to serious health issues, including **methemoglobinemia** (blue baby syndrome) in infants and various types of **cancers** in adults



(Source: Greenpeace, 2022)

Previous Research

Nationwide studies link **public water contaminants** to **vulnerable populations**, but data on **private wells** is limited

Surveys found no link between **private well contamination** and **population characteristics**

Objectives

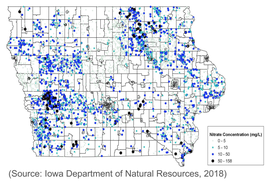
Analyze the spatial distribution of nitrate contamination and socio-demographic factors

Apply an **advanced interpretable machine learning model** to study these relationships

Data and Methodology

Data Collection

Nitrate Test Results:



(Source: Iowa Department of Natural Resources, 2018)

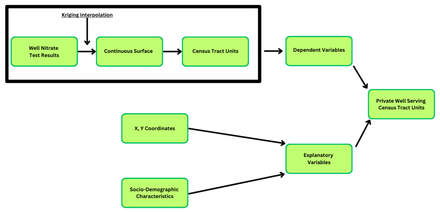
Socio-Demographic Factors:

Socio-demographic data were obtained from the American Community Survey (2014-2018). The variables included:

- **Racial and ethnic minority** percentage
- Percentage of individuals **with less than a college degree**
- **Unemployment rate**

	Mean	Std	Min	Median	Max
% Racial & Ethnic Minority	13.55	14.37	0.00	8.18	88.02
% less than college degree (25+)	41.54	12.50	4.57	42.98	76.92
% unemployed civilian labor force	16.60	5.83	4.80	15.63	40.81

Data Preparation



Regression Analysis Models

Compared several **regression models** to analyze the **relationship** between **socio-demographic characteristics** and **nitrate contamination levels**

	Spatial Lag & Error Model (SLM)	Multi-scale geographically weighted regression (MGWR)	Generalized Additive Models (GAMs)	Extreme Gradient Boosting Model (XGBoost)
Spatial Effects	✓	✓	✓ (Geoadditive Model)	✓ (With Location Variables)
Spatial Non-Stationary		✓		✓ (With Location Variables)
Non-linearity			✓	✓
Interaction Effects				✓
Interpretability	✓	✓	✓ (Less Interpretability with the Smooth Functions)	✓ (With Variable Contribution Analysis (SHAP Value))

Variables Contribution Analysis

SHapley Additive Explanations (SHAP) Values:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n - |S| - 1)!}{n!} \left[f(S \cup \{i\}) - f(S) \right]$$

marginal contribution of feature i

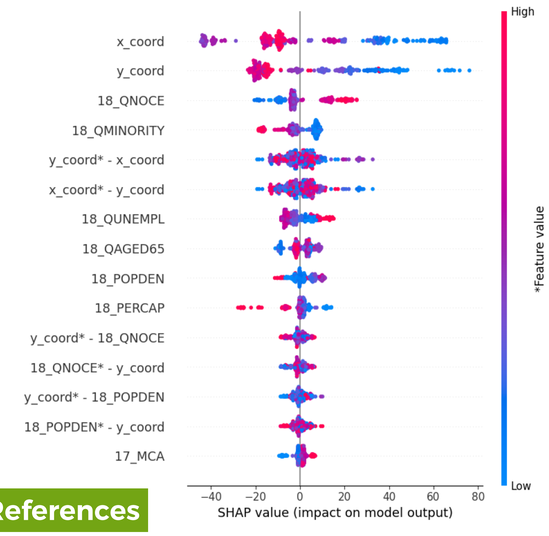
change of predicted value before and after adding the new feature i

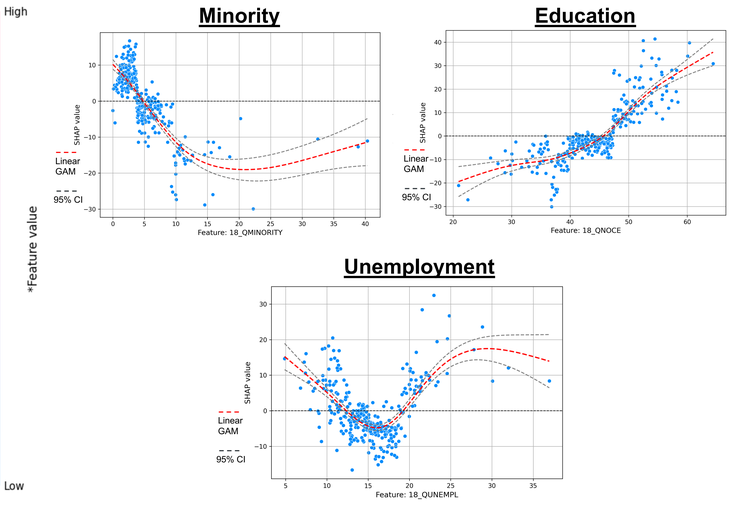
n : total number of features,
$N \setminus \{i\}$: all the combinations of features excluding feature i,
S : one of the subsets of $N \setminus \{i\}$,
$f(S)$: model prediction with feature values in S,
$f(S \cup \{i\})$: model prediction with feature values in S and feature value of i.

Positive SHAP Value: feature i contributes to increasing the predicted outcome value

Negative SHAP Value: feature i contributes to decreasing the predicted outcome value

Results





- **X-Axis:** Percentage of population with no college education / unemployed / minority status
- **Y-Axis:** SHAP value indicating the impact on nitrate levels
- **Red dashed line:** Linear relationship
- **Solid line:** Generalized Additive Model (GAM) fit
- **Dashed black line:** 95% confidence interval

- **Base Value** (nitrate concentration): **4.87 mg/L**
- **X-Axis: Percentage change** in **nitrate concentration** based on **4.87 mg/L**
- **Y-Axis: Analyzed features**

Discussion

Model Evaluation

Use of **SHAP values** enhances the interpretability of the model, providing insights into the relative importance of different features, such as **unemployment rate, educational attainment, and minority status**

Socio-Demographic Factors

Unemployment and **low educational attainment** emerged as significant predictors of risk

- Areas with **high unemployment** (greater than 16%) have a **25% higher risk** of nitrate contamination
- Areas with **low education levels** (less than a college degree) have a **20% higher risk**

SHAP values also highlight the importance of considering **minority status**

- **Hispanic populations** face a **15% higher risk** of nitrate contamination
- **Minority populations** from **0 to 10%** contribute to a **20% decrease** in predicted nitrate levels, indicating less exposure to elevated nitrate well water

Spatial Effects

SHAP values indicate that **location** has the most crucial effect on contributing to nitrate levels prediction, with lower values towards the **west and south areas** relating to an **up to 80% increase** in nitrate concentration

Conclusion

Findings

Understand population disparities in exposure to high nitrate well water pollution

- Communities with **low college degree attainment**, **high number of population with minority status**, **high unemployment**

Underscore **non-linear relationships** in social vulnerability analysis

Future Work

Measure the uncertainty for the SHAP values

- Use **non-parametric inference approaches** to derive confidence intervals

Acknowledgements

I firstly am grateful to God for His love, provision, and grace through Christ. I also want to thank my mentors in the Geo-Social Lab, Dr. Caglar Koylu and Jinyi Cai, for their meticulous and exceptional mentorship and invaluable support, inspiring me to go above and beyond with my research. Moreover, I am thankful for my lab mates, Ariana and Devesh, who supported me throughout this experience. Finally, I would also like to extend my gratitude to my parents and the Belin-Blank Center for this incredible opportunity.

References