Social Vulnerability and Health Exposure to Well Nitrate Contamination in Iowa: An Interpretable Machine Learning Approach

Jinyi Cai, Caglar Koylu, Eric Tate, David Cwiertny

ABSTRACT: Under the higher risk of water pollution issues posed by extreme weather, this research proposes an approach to investigate the adverse health effects of environmental hazards on socially vulnerable populations. The Shapley additive explanation (SHAP) is adopted to interpret the Extreme Gradient Boosting (XGBoost) model to discern the contributions of socio-demographic variables to colorectal disease risks within elevated nitrate water pollution areas in Iowa. The comparison between the XGBoost model and the Multi-scale Geographically Weighted Regression (MGWR) shows a better model fit with XGBoost. The feature importance results suggest that populations with housing cost burden and lower levels of education attainment are vulnerable to colorectal cancer risk associated with nitrate pollution. In the southwest areas of Iowa, this effect is more evident for the population facing housing cost burdens.

KEYWORDS: *water quality, private wells, social vulnerability, environmental health, interpretable machine learning*

Introduction

Climate change is amplifying the duration and frequency of extreme weather events, such as droughts and floods. In agricultural areas, these events increase nitrate concentration in hydrological systems, posing a threat to private wells. Well users are responsible for the stewardship to maintain their water quality while their socio-economic status influences their ability to stewardship. Meanwhile, epidemiologic studies found strong evidence for the association between nitrate exposure, especially above 5mg/L, and certain diseases (Schullehner et al., 2018).

Previous research has explored the relationship between socio-demographic characteristics and drinking water pollution using areal unit data, like the American Community Survey (ACS), and household surveys (Andrew George et al., 2023; Soriano Jr. et al., 2023). However, the absence of areal dataset on stewardship behaviors impedes the analysis of whether people residing in nitrate-polluted areas are vulnerable in health. Household surveys offer insights into social disparities in stewardship but are limited by sample sizes and recall bias. Moreover, complex factors like water table level and well depth complicate the examination of vulnerability correlations through linear regression.

Using the conceptual framework in Figure 1, this study aims to address the questions: Who are the populations vulnerable to the health risks posed by nitrate pollution in drinking well water? Where are they? To answer these questions, the study integrates data on well water nitrate levels and conducts regression analysis with an interpretable machine learning approach between socio-demographic factors and colorectal cancer risk rates.

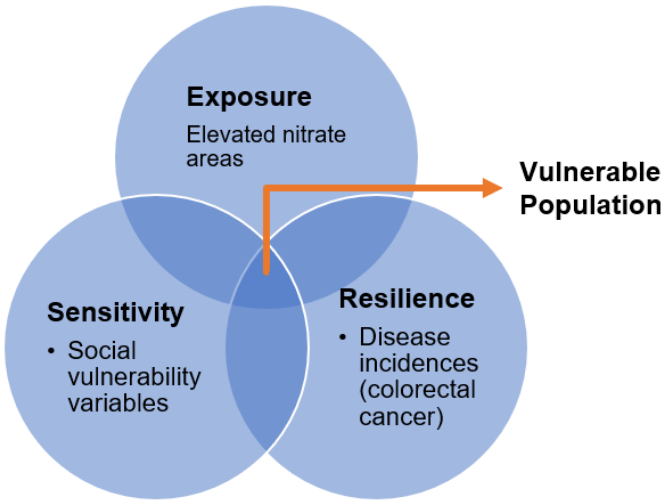

Figure 1 Conceptual Framework

Data

Nitrate test results from 8,532 shallow wells (<100 feet) in Iowa (2014-2018) were analyzed using data from the Private Well Tracking System by the Iowa DNR. Tracts of elevated (>=5 mg/L) and not-elevated (<5 mg/L) nitrate levels were categorized, as shown in Figure 2, through interpolation and zonal statistics.

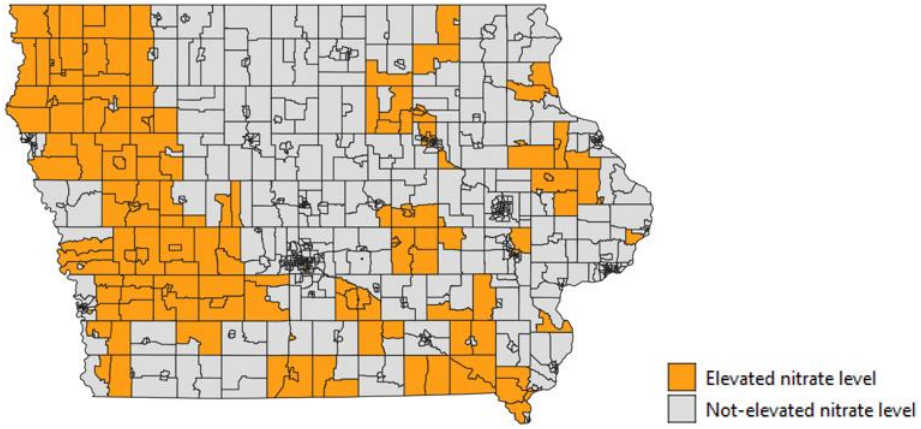

Figure 2 Elevated nitrate pollution tracts in Iowa

Seven socio-demographic indicators from ACS 2016-2020 5-year estimates were analyzed at the census tract level in Iowa as measures of social vulnerability. These include: The percentage of African American (EP_AFAM), Hispanic (EP_HISP), high housing cost burden (EP_HBURD), adults over 25 without a high school diploma (EP_NOHSDP), individuals below 150% poverty line (EP_POV150), people aged 65+ (EP_AGE65), and population density (POPDEN).

The colorectal cancer incidence rates (2014 - 2018) from the Iowa Cancer Registry website (*Iowa Cancer Registry*, 2024) were used to evaluate the health impact of nitrate exposure. This measure compares the cancer incidence probability in each unit against statewide rates (Ward et al., 2019). Major city areas were excluded to minimize influence of public water systems, as illustrated in Figure 3.

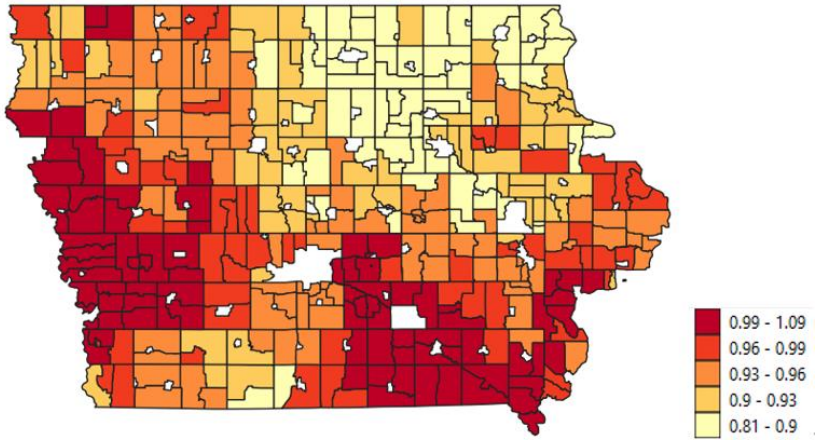

Figure 3 Relative risk of colorectal cancer in Iowa, 2014-2018

Method

The Extreme Gradient Boosting (XGBoost) model is used for regression analysis with the relative risk of colorectal cancer incidence as the dependent variable and the seven social vulnerability indicators as the features. It is unaffected by multicollinearity, which improves the investigation of the relationship among the influential while correlated variables.

The SHapely Addictive explanations (SHAP) analysis is used to interpret the contribution of the explanatory variables to the prediction result (Lundberg & Lee, 2017). SHAP is based on game theory, aiming to measure the importance of each feature in the model. The Shapley value for feature i in a model f is calculated by:

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

Here, n is the total number of features, $N\setminus\{i\}$ is all the combinations of features excluding feature i, S one of the subsets of $N\setminus\{i\}$, $f(S)$ is the model prediction with feature values in S, $f(S \cup \{i\})$ is the model prediction with feature values in S and feature value of i. $f(S \cup \{i\}) - f(S)$ calculate the difference between prediction results of whether adding the feature i. The SHAP value reflects a feature's contribution to a prediction: positive for enhancing and negative for diminishing the prediction. Thus, it could be used to interpret

the regression results to reveal the impact of socio-demographic factors on predicting colorectal cancer risk.

A recent finding indicates that SHAP-explained XGBoost can better identify spatial and non-spatial effects than multiscale geographically weighted regression (MGWR) (Li, 2022). To further verify, the MGWR model was also fitted to compare the performance.

Results

The regression analysis was performed on 128 elevated nitrate tracts in Iowa. The dataset for 277 not-elevated nitrate tracts was trained for comparison. XGBoost has the R2 value of 0.93 on elevated tracts and 0.92 on not elevated tracts. While MGWR has R2 value of 0.71 on elevated tracts and 0.81 on not elevated tracts. Figure 4 displays the SHAP summary plot with the top 15 contributing feature effects and interaction effects. The high contribution of X and Y coordinates and their interactions indicates a notable spatial effect. The following socio-demographic characteristics are Hispanic, high housing cost burden, people above 65, population density and persons with no high school diploma. Moreover, the SHAP identifies some interaction effects between location and socio-demographic characteristics.

Table 1 compares the mean SHAP values of each feature between areas with elevated and not-elevated nitrate pollution. A high ratio for certain features indicates its different effect on colorectal disease risk in areas of different nitrate levels. Characteristics of African Americans, high housing cost burden, persons with no high school diploma, Hispanic and persons aged 65 and older show noticeable differences. Among these features, the positive mean SHAP of housing cost burden and low education attainment suggests a correlation with a high risk of colorectal disease. Conversely, negative mean SHAP values for racial minority groups, including African American and Hispanic, highlight their association with a lower risk of colorectal disease. Additionally, while spatial effect contributes significantly to prediction results, their differences between the high and low nitrate level areas are not as pronounced as those observed in socio-demographic features.

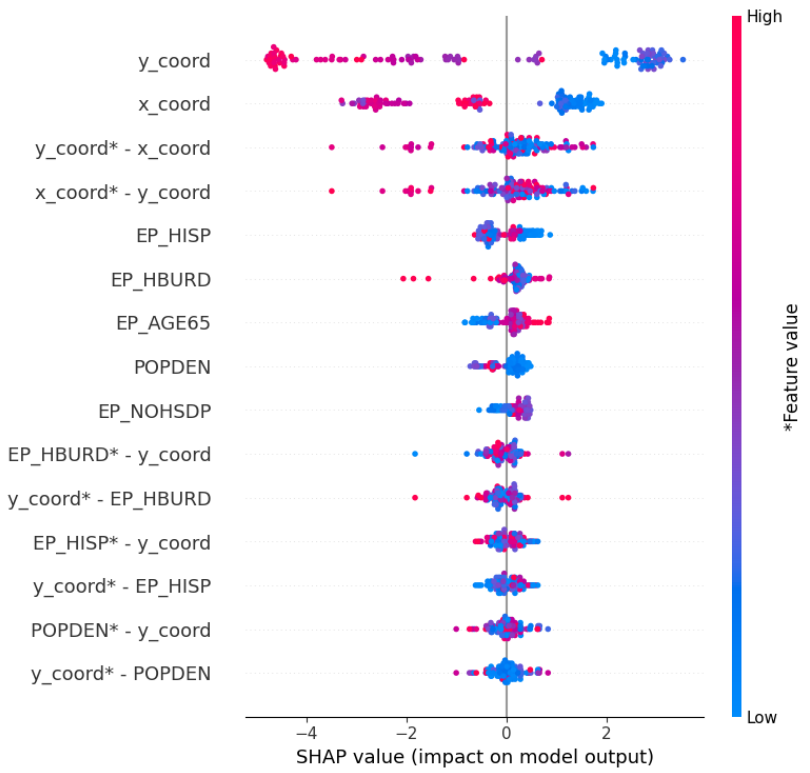

Figure 4 SHAP summary plot for the elevated nitrate model

Table 1 Comparison of feature importance at elevated and not elevated nitrate areas

Characteristics	*Elevated (> 5 mg/L)*	*Not elevated (< 5 mg/L)*	*Ratio*
% of Black/African American	-0.0138	-0.0008	16.7710
% of high housing cost burden	0.0986	0.0091	10.8498
% of persons with no high school diploma (age 25+)	0.0897	0.0085	10.5196
Y coordinate	-0.2678	-0.0834	3.2113
X coordinate	-0.1131	-0.2431	0.4652
% of persons below 150% poverty	-0.0068	0.0683	-0.0990
Population density	0.0265	-0.0366	-0.7239
% of persons aged 65 and older	0.0108	-0.0031	-3.5139
% of Hispanic	-0.0855	0.0030	-28.3155

Figure 6 displays the spatial distribution of SHAP values for housing cost burden. Red tracts indicate a positive contribution to colorectal disease risk, and blue tracts indicate a negative contribution. A cluster in the southwest for housing cost burden shows its positive influence on increasing disease risk.

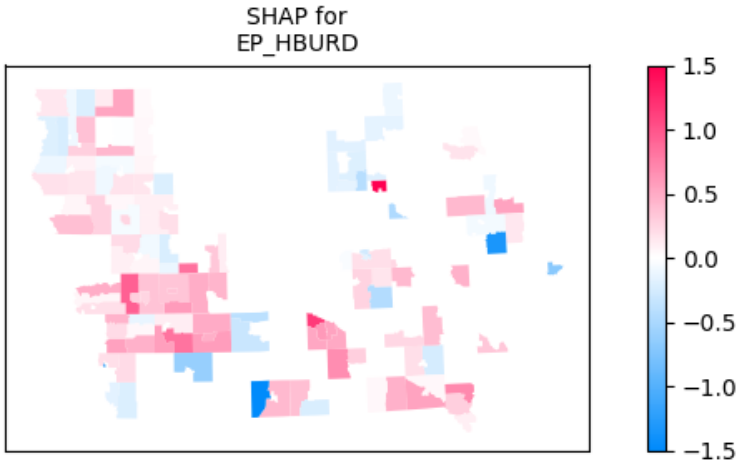

Figure 5 Spatial pattern of the SHAP values for housing cost burden by census tracts

Conclusions

This study investigates the relationship among nitrate pollution, socio-demographic characteristics, and colorectal cancer risk. We utilized SHAP-explained XGBoost to conduct vulnerability analysis by identifying feature contributions, spatial effects, and interaction effects. The findings highlight the increased vulnerability of populations with housing cost burdens and lower education levels, which suggests that stewardship behaviors among private well users relate more to economic burdens and literacy, particularly in Iowa's southwest.

Future work will further address current limitations. First, areas with a higher risk of colorectal cancer may be affected by nitrate pollution from urban community water services. Removing tracts with major city areas leads to decreased data volume and may cause bias. So further analysis could benefit from disease incidences at a finer scale. Second, considering the time lag effect in the exposure to nitrate and increased risk of colorectal disease, the interpretable results can be improved by incorporating older nitrate level data.

This study offers valuable insights into the socio-demographic and spatial disparities among well users in agricultural areas, which underlines the importance of targeted interventions for water quality and private well stewardship in vulnerable communities. This becomes increasingly vital with the heightened risk from extreme weather events affecting water quality.

References

Andrew George, Kathleen Gray, Wait, K., Gallagher, D., Edwards, M., Currie, J., Hogan, J., Kwasikpui, A. W., & Pieper, K. J. (2023). Drinking Water Disparities in North Carolina Communities Served by Private Wells. *Environmental Justice*, env.2022.0100. Q2. https://doi.org/10.1089/env.2022.0100

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Iowa Cancer Registry. (2024). https://shri.public-health.uiowa.edu/cancer-data/interactive-iowa-data-tools/iowa-cancer-maps/colorectal-cancer-incidence/

Li, Z. (2022). Extracting spatial effects from machine learning model using local interpretation method: An example of SHAP and XGBoost. *Computers, Environment and Urban Systems*, *96*, 101845. https://doi.org/10.1016/j.compenvurbsys.2022.101845

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Schullehner, J., Hansen, B., Thygesen, M., Pedersen, C. B., & Sigsgaard, T. (2018). Nitrate in drinking water and colorectal cancer risk: A nationwide population-based cohort study. *International Journal of Cancer*, *143*(1), 73–79. Q1. https://doi.org/10.1002/ijc.31306

Soriano Jr., M. A., Warren, J. L., Clark, C. J., Johnson, N. P., Siegel, H. G., Deziel, N. C., & Saiers, J. E. (2023). Social Vulnerability and Groundwater Vulnerability to Contamination From Unconventional Hydrocarbon Extraction in the Appalachian Basin. *GeoHealth*, *7*(4), e2022GH000758. https://doi.org/10.1029/2022GH000758

Turner, B. L., Kasperson, R. E., Matson, P. A., McCarthy, J. J., Corell, R. W., Christensen, L., Eckley, N., Kasperson, J. X., Luers, A., Martello, M. L., Polsky, C., Pulsipher, A., & Schiller, A. (2003). A framework for vulnerability analysis in sustainability science. *Proceedings of the National Academy of Sciences*, *100*(14), 8074–8079. https://doi.org/10.1073/pnas.1231335100

Ward, C., Oleson, J., Jones, K., & Charlton, M. (2019). Showcasing Cancer Incidence and Mortality in Rural ZCTAs Using Risk Probabilities via Spatio-Temporal Bayesian Disease Mapping. *Applied Spatial Analysis and Policy*, *12*(4), 907–921. Q4. https://doi.org/10.1007/s12061-018-9276-4

Jinyi Cai, Ph.D. Student, Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA 52240

Caglar Koylu, Associate Professor, Department of Geographical and Sustainability Sciences, University of Iowa, Iowa City, IA 52240

Eric Tate, Professor, School of Public and International Affairs, Princeton University, Princeton, NJ 08544

David Cwiertny, Professor, Department of Civil and Environmental Engineering, University of Iowa, Iowa City, IA 52242