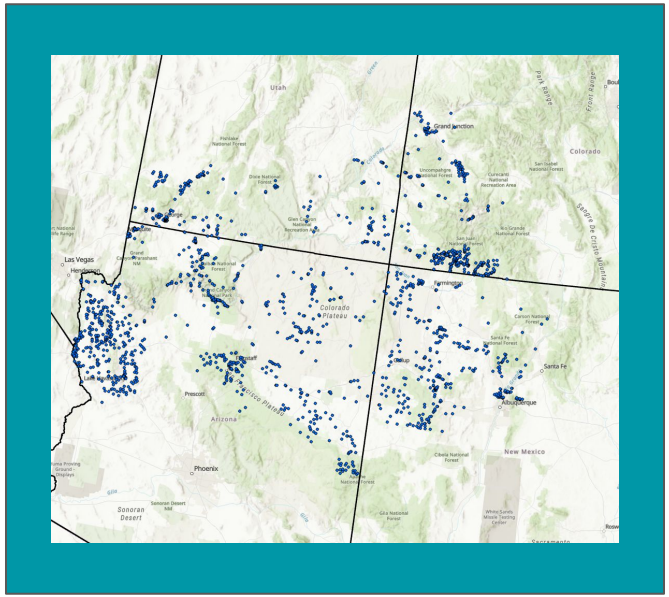
Using Machine Learning and geospatial data to Predict Groundwater Occurrence of Arsenic in the Colorado Plateau



Austin Martinez, Dr. Joseph hoover
University of Arizona, Biosphere 2
Biosphere 2 REU 2024

Arsenic Occurrence in Groundwater

- Arsenic occurrence in groundwater is a global health concern.
- Alluvial aquifer concentrations of arsenic can be high in arid and semi-arid regions like the southwest United states.
- Low hydraulic gradients, low natural recharge rates, minimal groundwater outflow to adjacent basins, long groundwater flow paths, and high potential evapotranspiration.

Health Risks

- Arsenic exposure from wells is a potentially substantial public health concern, but largely hidden.

- Effects of arsenic are:
  - hazardous too pulmonary and cardiovascular systems
  - increase cancer risk
  - affects childhood development.

Why Machine Learning?

- Machine learning methods capture complex patterns and provide better estimates of groundwater quality.
- Geochemical reactions and hydrologic conditions leading to elevated arsenic levels are complex and poorly described by linear models over large areas.
- Linear regression models assume linear relationships and are not well-suited for predictor variable collinearity.

Question

Which machine learning model (XGBoost, Random Forest Classification, and Boosted Regression Tree) most accurately predicts the occurrence of arsenic in groundwater, and which areas within the Colorado Plateau have highest risk of arsenic exposure?

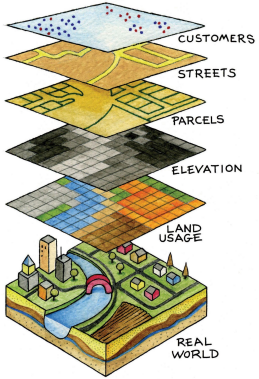
Methods - Data Extraction

1. Extracted regional water quality data for the Colorado Plateau (AZ, UT, NM, CO) from The Water Quality Portal and cleaned for use with GIS.

2. Extracted geospatial data layers (hydrologic, geologic, and soils) for the southwest US and converted to raster surfaces with ArcGIS Pro.

3. Extracted predictor values for well locations and compiled across all layers.

4. Cleaned dataset for machine learning models.





Methods - Model Creation

1. Split cleaned data into training (70%) and testing (30%) sets, balanced by arsenic measurement categories (C1=1ug/L, C2=5ug/L, C3=10ug/L).

2. Applied Random Forest Classification, Boosted Regression Tree and XGBoost models to predict arsenic occurrence by subclass for the Colorado Plateau.

3. Developed models in R(v. 4.4) using randomForest, caret, gbm, and xgboost packages.

4. Tuned hyper-parameters using UArizona Puma Supercomputing Cluster (192 nodes; 94 cores per node) - run using 96 cores on 1 node.

5. Optimized models using the training set and 10-fold cross-validation to maximize accuracy.

	Hyper-parameters							
Model	**Interaction depth**	**Shrinkage rate**	**Number of trees**	**Minimum observation node**	**Mtry**	**Number of rounds**	**Max depth**	**Subsample**
Random Forest Clasification	—	—	500	—	1 - 206	—	—	—
Boosted Regression Tree	2 - 16 by = 2	0.004 - 0.01 by 0.002	100 - 500 by = 10	8	—	—	—	—
XGBoost	—	0.005 - 0.0125 by 0.0025	—	—	—	100-1000 by 100	6 - 14 by 2	0.25 - 0.75 by 0.25

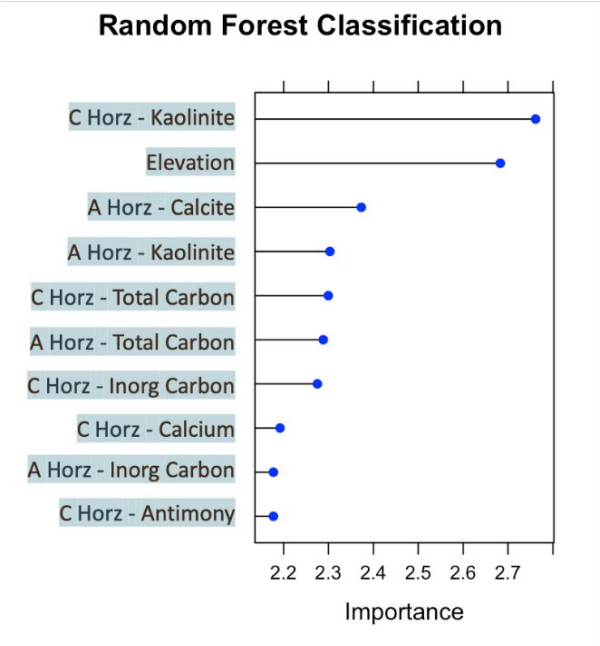
Figure 1: These are the hyper-parameters used to tune all 3 models. (NA = --)

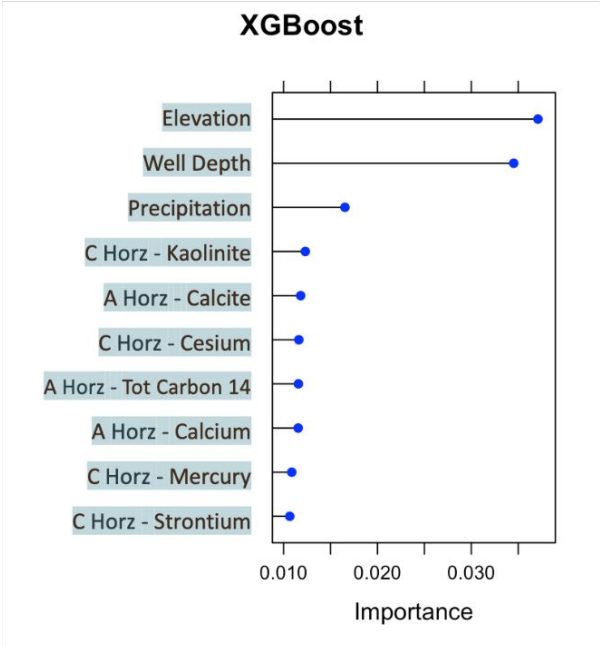
Results - Training and Testing data

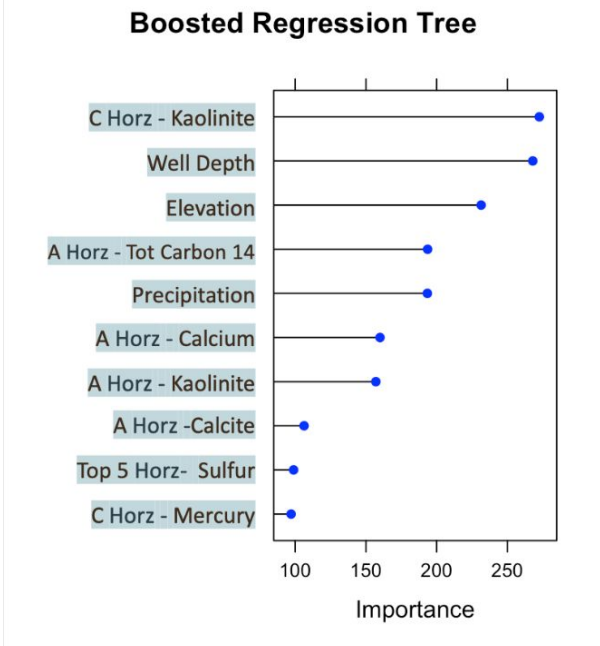
		Testing Data				Training Data		
Model	**Class/Concentration**	**Sensitivity(%)**	**Specificity(%)**	**Kappa**	**Accuracy(%)**	**Model**	**Accuracy(%)**	**Kappa**
Random Forrest Classification	C1 = 1 µg/L	98.71%	81.19%	0.8031	91.32%	**Random Forrest Classification**	77.66937%	0.4132252
	C2 = 5 µg/L	70.21%	98.95%					
	C3 = 10 µg/L	77.78%	97.5%					
Boosted Regression Tree	C1 = 1 µg/L	96.19%	77.63%	0.7383	90.06%	**Boosted Regression Tree**	79.27482%	0.4564251
	C2 = 5 µg/L	53.33%	98.58%					
	C3 = 10 µg/L	82.61%	96.24%					
XGBoost	C1 = 1 µg/L	97.85%	44.55%	0.3987	76.95%	**XGBoost**	82.56228%	0.4369628
	C2 = 5 µg/L	12.77%	97.91%					
	C3 = 10 µg/L	42.59%	94.64%					

Figure 3: (Testing data) The chart compares the performance of all 3 models in predicting arsenic concentrations at thresholds of 1, 5, and 10 µg/L. **(Training data)** The chart compares the accuracy and kappa of all 3 models

Results - Variable Importance Plots

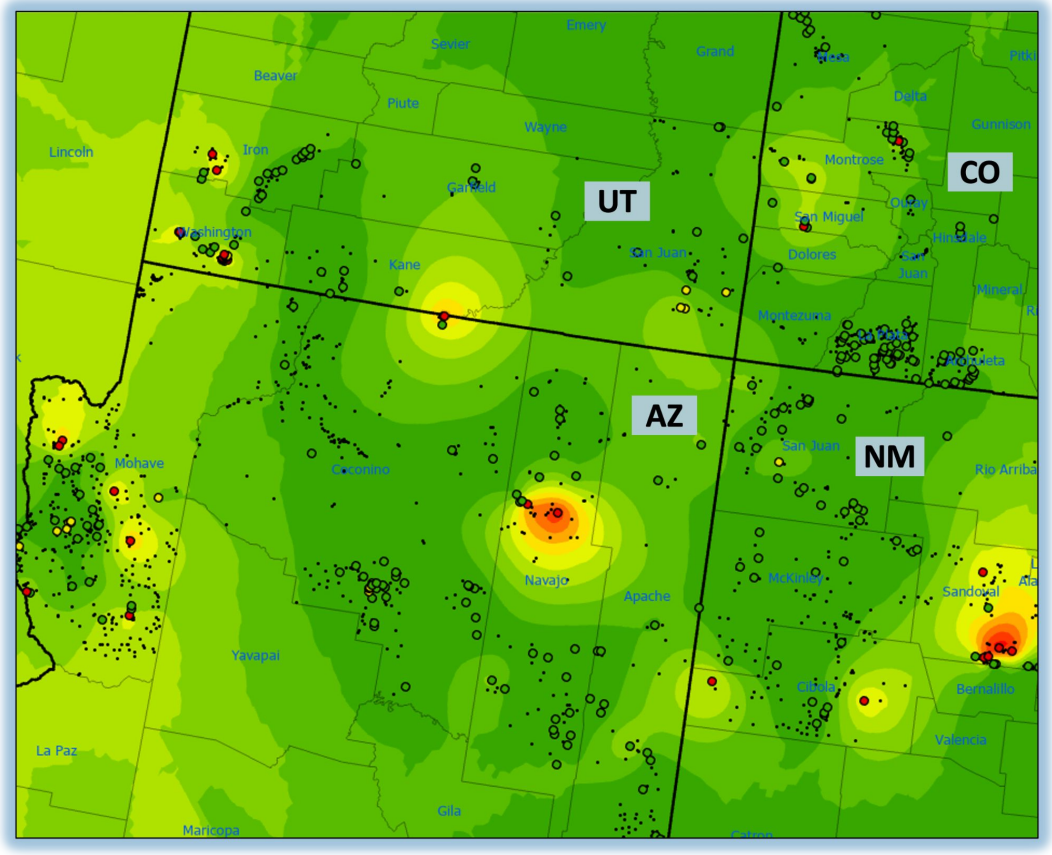






Results - Counties at risk

Map of Interpolated As Concentrations > 10 ug/L



Counties at Risk of Arsenic Exposure:

	At Risk Counties		
Arizona	Navajo	Mohave	Coconino
New Mexico	Sandoval	Cibola	
Utah	Washington	Iron	Kane
Colorado	San Miguel	Montrose	

Figure 4: This chart shows which counties are at risk of arsenic exposure above 10 ug/L.

Conclusions

Which model predicts arsenic occurrence best:

Training data accuracy

- XGBoost showed the highest training data accuracy at 82.56%.

Testing data accuracy

- Random Forest Classification achieved the highest testing data accuracy at 91.31%.

Counties at Risk of Arsenic Exposure:

	At Risk Counties		
Arizona	Navajo	Mohave	Coconino
New Mexico	Sandoval	Cibola	
Utah	Washington	Iron	Kane
Colorado	San Miguel	Montrose	

Figure 4: This chart shows which counties are at risk of arsenic exposure above 10 ug/L.

Future Plans

Uranium in the Colorado Plateau

- This same project is going to be done but with Uranium.
- Model accuracy will be evaluated for all 3 models on predicting Uranium occurrence
- A probability map will be generated to identify at risk communities

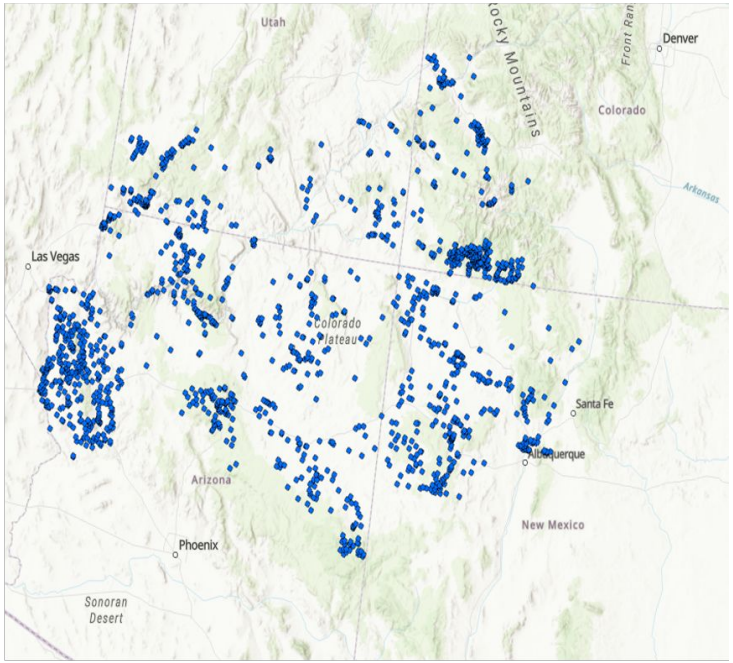


Figure 5: Map of the study area and sampled wells.

Acknowledgments

This research was supported by Biosphere 2 NSF Grant: 2051145 and the Hoover Lab.



THE UNIVERSITY OF ARIZONA
Biosphere 2



National Science Foundation

Any Questions?

