Goldschmidt 2021 Abstract
https://doi.org/10.7185/gold2021.3423

Applying machine learning models to predict basin of origin for produced waters using major ion chemistry

JENNA SHELTON[1], AARON M. JUBB[1], SAMUEL SAXE[2], EMIL ATTANASI[2], ALEXEI MILKOV[3], MARK ENGLE[4], PHILIP FREEMAN[2], CHRISTOPHER SHAFFER[5] AND MADALYN BLONDES[1]

[1]U.S. Geological Survey
[2]US Geological Survey
[3]Colorado School of Mines
[4]The University of Texas at El Paso
[5]US Fish and Wildlife Service

Presenting Author: jlshelton@usgs.gov

Understanding the geochemistry of waters produced during the extraction of oil and gas resources is essential to informing the best treatment and reuse options, potentially optimized for a given geologic basin. In this study, we used the U.S. Geological Survey's National Produced Waters Geochemical Database (USGS PWGD) to determine if major ion chemistry could be used to accurately classify a produced water sample to a given geologic basin using machine learning techniques. Two datasets were derived from the USGS PWGD and split into training and test datasets: one with fewer features (n = 7) but more samples (n = 58,541) named PWGD7, and another with more features (n = 9) but fewer samples (n = 33,271) named PWGD9. Three supervised machine learning algorithms (Random Forest, Naïve Bayes, and k-Nearest Neighbors) were used to develop multi-class classification models to predict basin of origin for produced waters using major ion chemistry. After training, the models were tested on three different datasets: one based on PWGD7, one based on PWGD9, and one based on data absent from the USGS PWGD.

Overall prediction accuracies across the various models developed ranged from 23.5% to 73.5% when tested on the two PWGD-based datasets. A model using the Random Forest algorithm predicted most accurately compared to all other models tested. The models generally predicted basin of origin more accurately on the PWGD7-based dataset than the PWGD9-based dataset, suggesting that either a larger sample size and/or fewer features lead to a more accurate model. Individual balanced accuracies for each basin within PWGD7 ranged from 50.6% (Anadarko) to 100% (Raton), and from 44.5% (Gulf Coast) to 99.8% (Sedgwick) for PWGD9. Results from testing the most accurate Random Forest model on recently published data outside of the USGS PWGD suggest that some provinces may lack information regarding geochemical diversity while others included in this dataset are well described. A compelling result of this work is that basin of origin for produced waters can usually be determined using major ion composition alone and therefore, deep basinal fluid compositions may have more inter-basin variability than intra-basin variability.