AWWA DC Office Internal Report 2023-12

American Water Works Association
Dedicated to the World's Most Vital Resource

SDWIS Arsenic Distributions from 2006-2011

Sean Franco

Data

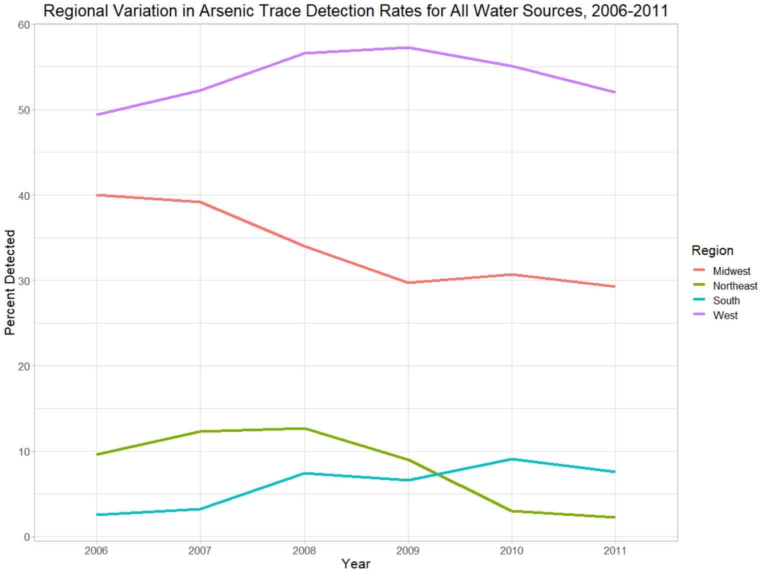

Figure 1: Regional percent detection (1) of arsenic.

Data was collected from the Environmental Protection Agency (EPA) SWDIS database website [1], [2]. The data contains 31 variables, and over 200,000 observations from 2006 to 2011. Relevant information include the state for each water utility, if arsenic testing has occurred, the trace arsenic value, and region and division information. The following report shows distributions of states and regional arsenic detection occurrence. Some states were exempt from reporting, and were mainly in the South census region, and the east central south census division.

Year	Occurrences of 1	Occurrences of 0	Total	Occurrence 1 Percentage
2006	9,456	21,417	30,873	30.6
2007	12,889	22,451	35,340	36.4
2008	13,557	21,803	35,360	38.3
2009	13,429	21,284	34,713	38.6
2010	14,555	22,270	36,825	39.5
2011	13,986	22,822	36,808	37.9

Table 1: Occurrences of Detection level by year.

Machine learning was used to determine the grouping criteria that best showcases arsenic detection imputation. Guiding questions are state, region, or division the best grouping criteria? Or is it arsenic detection best explained thru source water type? A logistical regression was created to compare the odds of Region, State, and Source Water Type independent variables on a positive detection imputation. A logit regression was performed for each year, and was restricted to the select year's data. Each year's total observations were between 31,000 to 37,000 which was deemed reasonable to compare among each year as in Table 1. The data was then split on a 80% training, and 20% testing and ran each year. Model performance evaluation metrics are observed in Figure 2. The results were plotted in Figure 2 from the corresponding year's coefficients in odds-ratio form. Over the time frame, region incurred an odds ratio of 20% to 54% above other variables as predicting detection status. States had no meaningful odds ratio compared to the other variables. This is likely due to constrained observations for each state resulting in noisy data. Source Water Type did not perform well at predicting detection status.

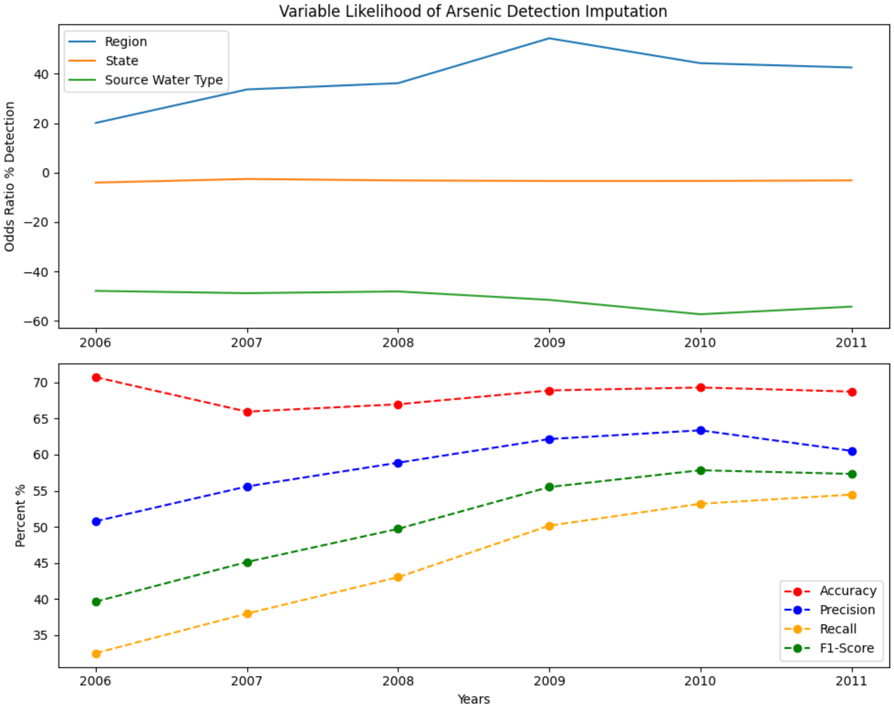

Figure 2: Logistical odds ratio results (top), Accuracy, Precision, Recall, and F1-Score by year (bottom).

The model had an accuracy (predicting true positives and negatives) detection level between 66% to 71%. However the ratio of positives to negatives for each year's data set is ranges from 30% to 37% for positive detection. This means that the accuracy score is inflated from the model correctly predicting the higher occurrences of negatives. To address dataset imbalances, precision (false positives), and recall (false negatives) evaluation scores provide better metrics for the model. The model's precision of detecting false positives ranges from 50% to 60%. The model's recall of detecting predicting false negatives ranges from 32% to 54%. The balance between Precision and Recall evaluations is represented thru F1-Score. Generally, starting from between 2008 and 2009 data sets, the model evaluation criteria report higher values for each performance metric. Given that the model test-training data split, and the comparable dataset size for each year, the model is robust. Regions offer the grouping to best reflect true positive and true negative detect v. non-detect status of Arsenic. The arsenic detection concentration amounts will be analyzed at the regional scale.

Methods

Trace arsenic percentages were calculated across all water sources by grouping census regions through time and analyzing the ratio of detected (1) and non-detected (0) as observed in Figure 1 from 2006 to 2011. Arsenic concentration distributions were grouped by region, and time for all source water types in Figure 4. Figures 4, 5, 3, 6, have upper and lower lines represent the 75th and 25th percentile respectively. Additionally, a middle line or

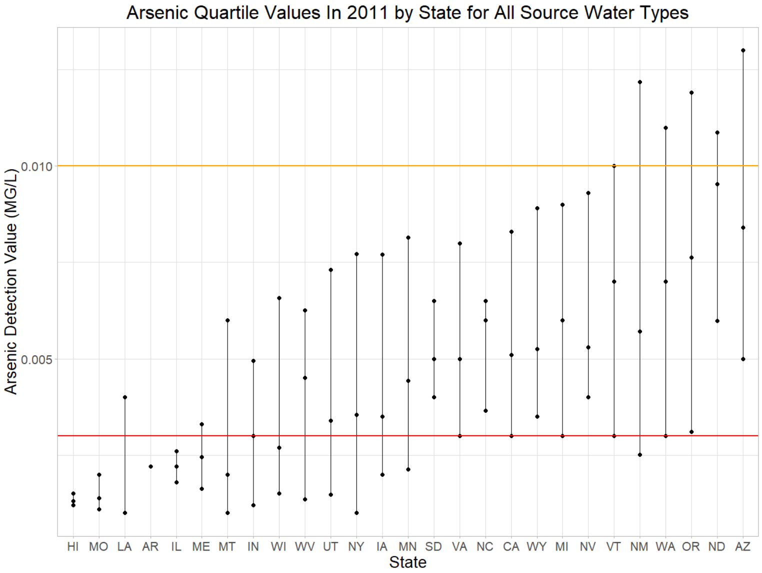

Figure 3: 2011 state quartile ranges of arsenic concentrations.

point represents for the median. Policy thresholds are included in these figures as horizontal orange (0.01 MG/L) and red (0.003 MG/L) lines [3]. Showcasing the regulatory thresholds and regional or state Arsenic distributions estimate benchmarks of arsenic occurrence in finished H_2O through time.

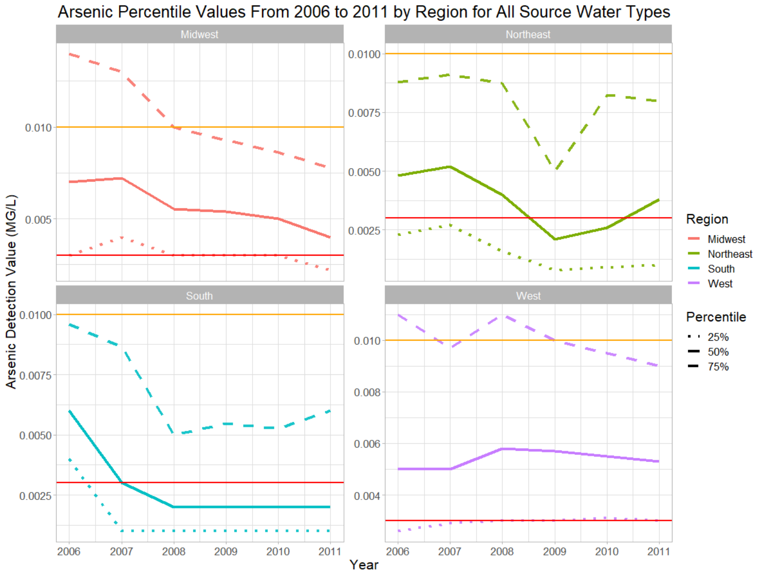

Figure 4: Regional arsenic concentration quartile ranges.

In Figure 1 the West had a higher rate of positive detection cases followed by the Midwest than other regions. From 2006 to 2011, the Midwest decreased their ratio of detected cases by 10%. Similarly, the Northeast decreased their ration of detected cases by 10% from 2008 to 2011. From 2006 to 2011, in Figure 4 all region have their 25^{th} percentile around or below the 0.003 MG/L threshold. From 2006 to after 2008 the Midwest experiences the 75^{th} percentile above or at the current policy threshold (0.01 MG/L). The West also experiences peaks where the 75^{th} percentile is above the current policy threshold in 2006, and from the middle of 2007 to 2009. The Northeast and South regions have the majority of their arsenic occurrences in finished H_2O below the 0.01 MG/L threshold.

Additionally, regions exhibit differing influences by water source as observed in Figure 5. Groundwater (GW), groundwater under direct influence of surface water (GU), and surface water(SW) had different concentrations of Arsenic MG/L in time, and by region. The Midwest experienced the GW 75% percentile above the current thresholds, which abated after 2008. The West experienced the GW 75% percentile above the current threshold

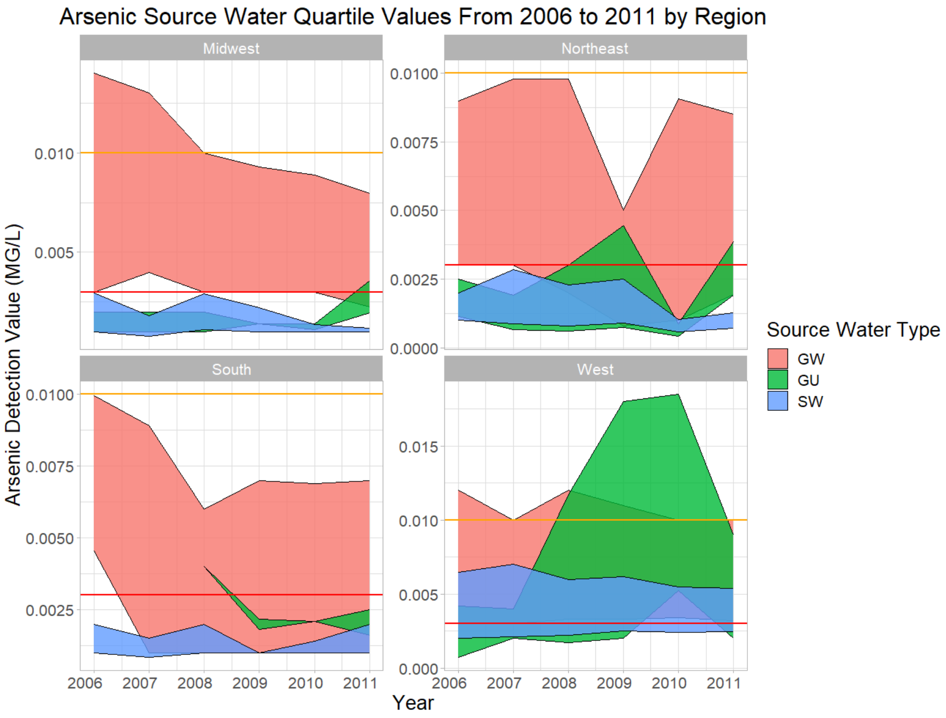

Figure 5: Regional source water arsenic quartile ranges.

from 2006 to 2010 as well as a spike of the GU 75% percentile above the threshold in the middle of 2007.

Big Picture

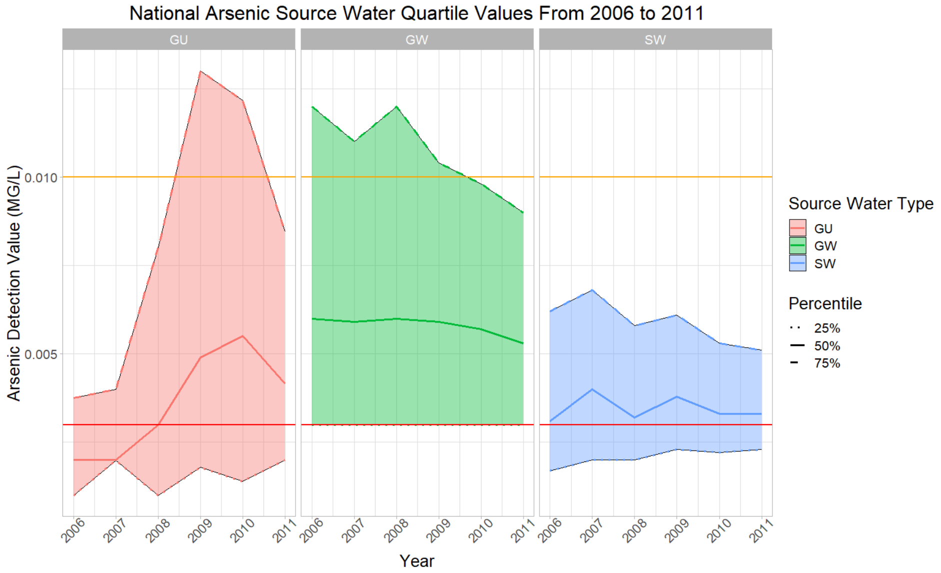

Figure 6: National water source arsenic concentration quartile ranges.

The West has higher detection arsenic detection rates as in Figure 1, overall higher distribution of arsenic in GW, GU, and SW water sources. The South has overall the lowest arsenic detection rates than the other regions. This is misleading because the South also has exceptions to reporting with multiple NA values in the data set. In 2011, across all water sources, most reported states maintain a median below current thresholds, but not below the upcoming threshold as observed in Figure 3. Observed in Figure 6, on a national level, SW and GW Arsenic concentrations are generally decreasing from 2006 to 2011, but GU Arsenic concentrations spiked around

2009. The biggest take-away is that these distributions are not expressing the majority of their variance below the upcoming threshold of 0.003 MG/L. In the absence of updated data, and routine testing, regionally and by extension utility-level arsenic occurrence in finished H_2O distributions have large portions of their distribution above policy thresholds.

References

[1] Office of Water. Sdwis federal reports advanced search.

[2] Office of Water. Six-Year Review 3 Compliance Monitoring Data (2006-2011).

[3] Office of Water (2001). Arsenic and clarifications to compliance and new source monitoring rule: A quick reference guide.

Funding: This work was supported through an internship with the American Water Works Association funded through AWWA's Water Industry Technical Action Fund (WITAF).