Water Contaminants in the U.S.

Kaleigh Benesch

Western Carolina University

MATH 678

August 1^{st}, 2023

Abstract

This paper examines the levels of certain contaminants found in water supplies across the United States. We will talk about "contaminants" as substances found in water that make it impure, whether they be naturally occurring, added intentionally for health reasons, or accidentally caused by humans. By incorporating evidence from different styles of visualizations and various statistical methods, we reveal several patterns that appear in the data. Supporting analysis will show the importance of location when investigating water suppliers that exceed health and legal limits of contaminants allowed in water. Location will also play a large role when analyzing where contaminants are more commonly found throughout the United States. Further, the data will validate correlation patterns between certain contaminants and show which ones are more or less likely to be found together. These areas of interest will shed more light on what is found in everyday drinking water sources used across the country.

1 Introduction of Data

Results in this paper are based on a data set that contains about 367,000 observations and 11 attributes. The data set "contaminant.csv" was posted to the open data catalog, data.world, in 2017 and was originally collected from the Environmental Working Group's (EWG) National Drinking Water Database. This data set was eventually merged with an updated set of drinking water standards from the EWG in 2021 for comparison. Variables that are included in "contaminant.csv" are the name of each contaminant, details of each location where a water sample was taken, the name of each supplier, and the number of people each sample serves. The water samples were taken from about 42,600 unique suppliers across the country, but not all suppliers' water was tested for the same number of contaminants. This may be due to different legislation between states that require only specific contaminants for documentation. For example, in Figure 1, we can see that "Selma Water Works & Sewer Board" does not have a recorded measurement of aluminum, but "Northern Dallas Water

Authority" does.

contaminant <chr>	average_result <chr>	max_result <chr>	supplier_name <chr>
Nitrate	0.14 ppm	0.17 ppm	Selma Water Works & Sewer Board
Total haloacetic acids (HAAs)	0.33 ppb	0.66 ppb	Selma Water Works & Sewer Board
Barium (total)	185.5 ppb	185.5 ppb	Selma Water Works & Sewer Board
Manganese	32 ppb	32 ppb	Selma Water Works & Sewer Board
Nitrate & nitrite	0.11 ppm	0.11 ppm	Selma Water Works & Sewer Board
Aluminum	245.33 ppb	245.33 ppb	North Dallas Water Authority
Nitrate	0.1 ppm	0.2 ppm	North Dallas Water Authority

Figure 1: Example of columns in the data set.

The data set also contains the average and maximum measurements of contaminants taken from a sample with corresponding units, as shown above. Finally, there are two columns that give the health and legal limits of the corresponding contaminant, also with units, and whether or not that sample has exceeded the limit allowed. It is important to note why a health limit is also included, instead of only a federal legal limit. After studying health effects, the U.S. Environmental Protection Agency (EPA) states a maximum contaminant level goal (MCLG) which is a non-enforceable health goal that is set at a level "at which no known or anticipated adverse effect on the health of persons occurs and which allows an adequate margin of safety"[2]. MCLGs consider only public health and not the limits of detection and treatment technology effectiveness. They sometimes are set at levels which water systems cannot meet because of technological limitations. Updated drinking water standards from the EWG in 2021 also include contaminant names and their corresponding health and legal limits with units.

2 Data Preparation

To begin making the data more usable, the "locations_served" column was duplicated and everything was removed before the separating comma to create a "state" column and everything after to create a different column for the "city." After creating distinct columns for city and state, the duplicated columns were renamed accordingly to be used later for analysis and visualizations. Before moving on, the maximum length of characters in the "state" column

was checked to make sure it was only of length two and consistent with postal abbreviations. The next stage of data cleaning considered contaminant results and whether or not health and legal limits were exceeded. For example, an observation in the "health_limit_exceeded" column would contain "No10 ppm" indicating that the health limit is ten parts per million and was not exceeded by the sample's results. The `mutate_at()` function was used to find the character sequences "No" and "Yes" in the limit columns and replaced these parts of the string with nothing, represented by "". Next, the "legal_limit_exceeded" variable was split to create a new column that contained only each observations corresponding unit of measurement. With a stand-alone units column, I then removed additional units from the "average_result", "max_result", and "Health_limit" columns. To do this, the function `gsub()` was used which finds the first space in a character string and replaces everything following it with a new character. In this case, all characters after the first space were replaced with nothing, similar to previous cleaning methods. With only numeric values left in these character columns, they were converted to have numeric properties using the `as.numeric()` function. Another problem encountered was that some legal limits stated that the contaminant was "legal at any level," which created about 30,000 observations with NA values in this column. Since these specific contaminants cannot exceed a legal limit that does not exist, the NAs were replaced in this column with a large number (999,999) to suggest they would never be exceeded. Similarly, NAs also appeared when there were values in the "Health_limit" column concerning MCLGs, meaning there is no known limit of some contaminants that will not affect human health. In this case, the entries also stated that the health limit should be set at zero, so these NA values were replaced with "0." The purpose of replacing the aforementioned values with a real number is to create new binary columns that make it easier to analyze when a water sample has exceeded a health or legal limit.

After importing the data set with updated drinking water standards, the health standard was separated from its attached unit and made into their own unique columns. As mentioned for the first data set, with a stand-alone units column, units were removed from the

"Federal_Legal_Limit" column using `gsub()`. The nonexistent legal limits were also replaced with a large number for evaluation purposes. After converting the updated health and legal limit columns to have numeric properties, they were renamed to include the year "2021," as to not create confusion when joined with the first data set. Once both data sets were cleaned individually, the `merge()` function was used to join them by the column with contaminant names. Since there were 99 unique contaminants in the updated standards data set and 315 in the first data set, there were about 75,000 missing observations in the health and legal limit columns for 2021. When the 2021 limit was not specified in the updated standards, the limit took on an "NA" value. To facilitate later analysis, all samples were converted to measurements of parts per billion (ppb) to parts per million (ppm). By using a combination of the `mutate()` and `if_else()` functions to search across multiple columns where a limit is taken in ppb, we can transform the corresponding limit so that it is divided by 1,000 to yield the desired result. Now, the merged data set is ready to be used to investigate what insights may lie within the data.

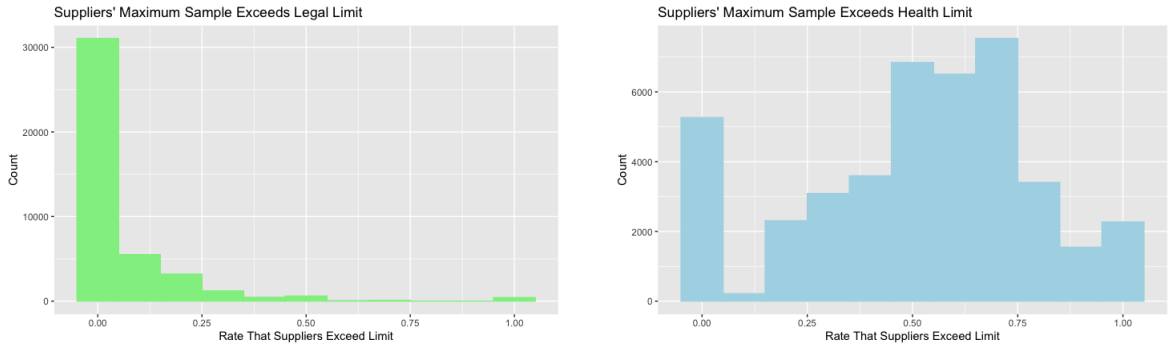
3 Analysis

The first area of interest investigated was how often suppliers across the country exceeded the legal limits of contaminants in their drinking water. To begin, a new data frame was created consisting of the binary variables created previously to determine whether or not a health or legal limit was exceeded in a given water sample. From this, a table was created to show how many times each supplier's name occurred and merged it with another data frame containing the sums of how many times each supplier's maximum and average result exceeded a health or legal limit, as shown in Figure 2 below.

	supplier_name	total_samples	avg_health_exceed	max_health_exceed	avg_legal_exceed	max_legal_exceed
1	@fcsa #40/Del Rio	5	5	5	3	3
2	10 4 Water System	11	3	3	0	0
3	100 Mountain View Road Park	8	7	7	0	1
4	101 Plaza	4	0	0	0	0
5	1133 Taconic Llc	6	3	3	1	1

Figure 2: Supplier Occurrences and Number of Times Limits Are Exceeded.

With this layout of information, we can then make additional columns that divide a supplier's number of times exceeded in a category by the total number of sample taken from them to compute a proportion of times that each limit was exceeded. The histograms in Figure 3 show how often suppliers exceed various limits.



(a) Rate suppliers exceed federal legal limit.

(b) Rate suppliers exceed health limit

Figure 3: Rates at which suppliers exceed limits.

From the left graph, most suppliers do not exceed the federal legal limits of contaminants allowed in drinking water. While there is more of a normal distribution on the right-hand graph, there is still a significant portion of suppliers who do not even exceed the stated health limits. On the other hand, after further investigation, there are 248 unique suppliers whose contaminant levels exceeded both health and legal limits every time. However, all but one of these suppliers only had either one or two contaminants recorded, thus making it much more likely to exceed all limits. For example, Bellavista Estates in Virginia had a water sample that only tested for aluminum, which exceeded the heal and legal limit for this contaminant. Since suggested health limits cannot be enforced, it is understandable why we see the data take this general shape that we do not see when looking at rates of legal limits

being exceeded.

Another subject investigated was whether or not certain contaminants are correlated with each other. In other words, if a specific contaminant has a large concentration in a water sample, then it may be more or less likely that another specific contaminant is also present in large concentration. Since there are 315 different contaminants in the data set, only eight most frequently occurring contaminants were considered for relevancy purposes. We will choose this number because after the eighth most frequently appearing contaminant, the number of observations seems to drop off at a higher rate. With the top eight most common contaminants as variables, two data frames were created; one data frame recorded the average concentration for each contaminant by about 15,600 unique suppliers, and the second data frame considered the eight contaminants for each of the 50 states. We will exclude the data frame with states because there is much more averaging that takes place when we confine all contaminant results down to 50 rows. The former data frame had 162,000 observations out of 367,000. The data frame initially holds each contaminant name in a single column, so the `pivot_wider()` function was used to make each of the eight contaminants their own column to show each average result by state or supplier name. For example, if we averaged each contaminant result by state, we would have the following data frame.

	State	Barium	Bromodichloromethane	Chloroform	Copper	Dibromochloromethane	Nitrate	Total haloacetic acids	Total trihalomethanes
1	AK	0.06777120	0.005683548	0.021141667	0.13975500	0.000861250	0.6740000	0.020532917	0.018935600
2	AL	0.17886714	0.002745833	0.015348636	0.08199867	0.001200870	0.6744681	*NA*	0.016735278
3	AR	0.03444016	0.005755921	0.017035316	0.13102394	0.003250119	0.3357143	0.018050990	0.026322581
4	AZ	0.06027741	0.001853125	0.005826667	0.07537707	0.004955357	1.9313415	0.006153333	0.006776765
5	CA	0.09552172	*NA*	0.015373474	0.03349079	*NA*	*NA*	0.017568148	*NA*

Figure 4: Average Results of Contaminants by State

Likewise, with this type of data frame averaged by supplier, the correlations among all eight contaminants can be determined and is provided below in Figure 5.

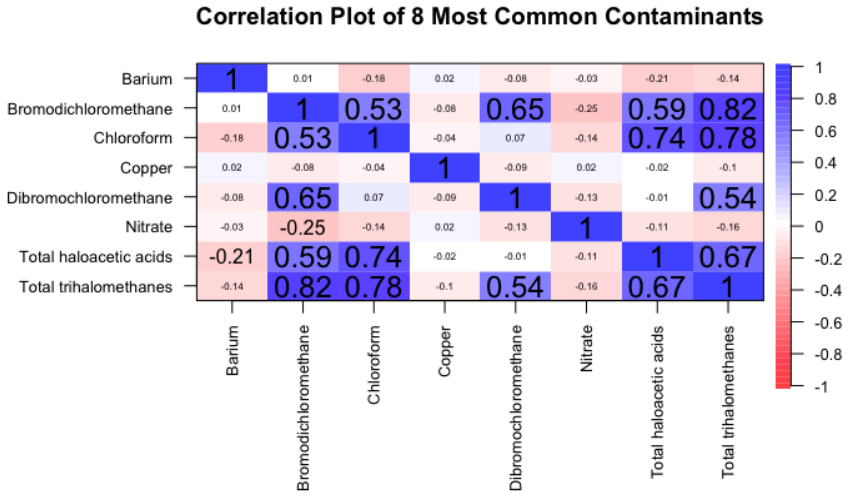


Figure 5: Correlation Plot of Contaminants by Supplier

When we explore the average results of these contaminants by supplier, we can see that more pairs of contaminants show higher correlation coefficients than in the data frame organized by state. In this data frame, bromodichloromethane and chloroform are most correlated with total trihalomethanes. These two have a positive correlation coefficients of 0.82 and 0.78, respectively, and might be more likely to be found together in similar amounts when testing water supplies. According to the U.S. Environmental Protection Agency, bromodichloromethane, chloroform, and trihalomethanes are all disinfection byproducts that form when water disinfectants, such as chlorine, react with other naturally occurring contaminants in the water[1]. With this in mind, it is understandable why these contaminants are more likely to be found together.

Finally, the locations in the United States in which some contaminants are commonly found was considered. Because there are 315 contaminants, maps were created considering only the five most common contaminants in the data set, as to keep individual analysis focused rather than flooded with maps. The five contaminants examined were nitrate, total trihalomethanes, barium, copper, and chloroform. A map of the U.S. was created that shows the average amounts of each contaminant by state. For this topic, the package "`usmap`" was used to create a map of the United States. After individually analyzing the five most common

contaminants, the only one that showed a significant geographical pattern was copper.

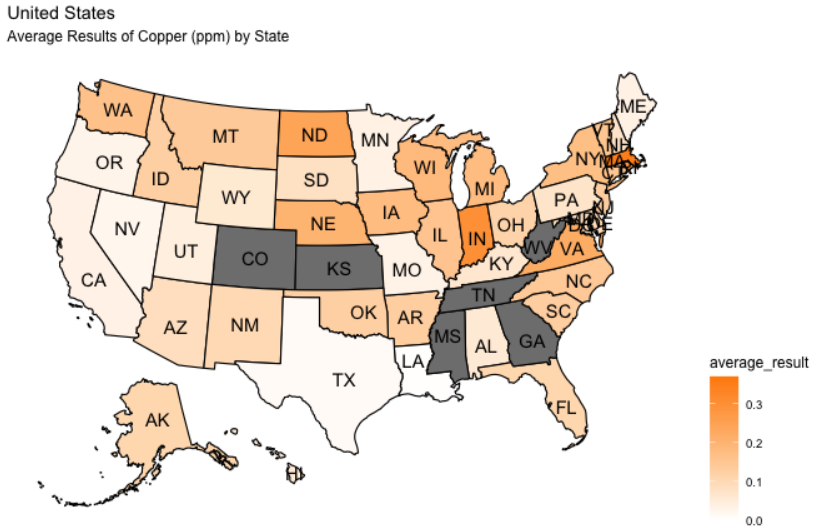


Figure 6: Average Level of Copper Contaminant by State

According to the map results in Figure 6, copper mostly appears in higher amounts in northeastern parts of the United States and in smaller amounts in west/southwestern states. Even though the state of Massachusetts has the highest average result (0.353 ppm) of copper in drinking water, this is not reason for concern. The recommended health limit of copper contaminant allowed in drinking water is 0.300 ppm and the stated federal legal limit is 1.000 ppm, so the darker areas on the map should not always be construed as "bad" results. With some knowledge of general U.S. history, we might be able to assume this pattern of copper being more common in the northeast is because of the amount of industrialization that has taken place in this area of the country. A combination of historical industrial pollution and residential copper plumbing leeching into water could be a reasonable factor for higher levels of copper in the samples taken.

4 Conclusion

As expressed through further analysis, several insightful patterns about drinking water were uncovered from the data set. Histogram representations showed that a majority of suppliers

do not exceed, or much less often exceed, the legal limits of contaminants allowed in water. These visuals also show that the frequency of suppliers exceeding the recommended health limits are more normally distributed because these limits cannot be enforced. Next, contaminants with similar purposes or causes, such as byproducts of chlorination, are more likely to be found together in water samples. Finally, maps were created that easily show whether or not contaminants follow geographical patterns throughout the United States.

Going forward, we may want to utilize the updated standards from 2021 to compare how certain health and legal limits have changed over time and how often suppliers would have exceeded these limits instead of the older standards. If I were to do this project again, I would have pulled more relevant data sets that were posted on the same page as the original "contaminant.csv." These files were also collected from the Environmental Working Group's National Drinking Water Database at the same time, and they include more information on specific locations of samples taken. With this we would be able to create different mapping visuals to display new locational patterns of certain contaminants or suppliers of drinking water.

References

[1] U.S. Environmental Protection Agency, *Disinfection byproducts: A reference resource*, 2016. `https://archive.epa.gov/enviro/html/icr/web/html/gloss_dbp.html`.

[2] ———, *How epa regulates drinking water contaminants*, 2022. `https://www.epa.gov/sdwa/how-epa-regulates-drinking-water-contaminants`.