Journal of Research High School

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

JRHS Outstanding Research Paper Award

Statistical Evaluation of the Correlations between Socioeconomic Factors and the Amount of Trihalomethane in Drinking Water in the State of NY

Stephanie Lee[1] *

[1]Horace Mann Upper School, Bronx, NY, USA
*Corresponding Author: stephanieyl30@gmail.com

Advisor: Sang Park (Ph.D.), spark1@ggc.edu

Received June 14, 2023; Revised July 13, 2023; Accepted, August 11, 2023

Abstract

Trihalomethanes are common byproducts of disinfection processes in public water systems. The relationships between the total amount of trihalomethanes in public water systems in New York and the corresponding socioeconomic variables were studied. A significant amount of chemical and demographic information representing 81% of the NY state population, was obtained from the NY State Department of Health, the US Environmental Protection Agency, and the US Census Bureau. Statistical tools such as Pearson Coefficients with P-values, Grubbs test, and Mean Comparison using Student's t-test were adopted to evaluate the correlations of total trihalomethanes concentration and various socioeconomic factors. Data analysis found negative correlations between the total amount of trihalomethanes and socioeconomic factors, such as mean household income, Asian percentage, and Hispanic percentage. In addition, the water source and the size of the public water system are considered critical factors. The lowest concentrations of total trihalomethanes were detected in communities served by groundwater with higher median household income and higher Asian populations.

Keywords: Trihalomethanes, Socioeconomic, Drinking Water, Statistical Evaluation, New York

1. Introduction

The recent water contamination crises in Flint, Michigan (because of lead in 2014) and the San Joaquin Valley (caused by nitrates and arsenic in 2007) have highlighted the unequal impact of contaminant exposure and poor water system management on populations of color living in poverty. Numerous studies have linked socioeconomic factors, such as income as well as racial and ethnic diversity, to a decline in water quality (Schaider et al., 2019; Switzer & Teodoro, 2017). According to a countrywide examination of drinking-water quality violations between 1982 and 2015, 8.0% of public water systems (PWS) had at least one health-related violation. Overall, there were over 95,000 breaches throughout the 34, with disinfection byproducts (DBPs) accounting for approximately 25% (Allaire et al., 2018). Although disinfection is generally recognized as a significant public health victory for its potential to inhibit the growth of pathogens in drinking water, trihalomethanes are the most common class of decontamination byproducts (DBPs) that form when natural organic matter and antiseptics like chlorine interact during the treatment of drinking water supplies (DeMarini, 2020). Several epidemiological studies showed associations between rectal, colon, and bladder cancers and chlorinated drinking water. These findings support the proposition that trihalomethanes potentially harbor carcinogenic properties in the human body, as substantiated by the outcomes derived from laboratory trials on animal subjects (Costet et al., 2011; Hildesheim et al., 1998). Since 1979, trihalomethanes have been subject to regulations by the United States Environmental Protection Agency (EPA). Currently, the maximum contamination level (MCL) for total trihalomethanes (TTHM) is set at 80 micrograms per liter (80 ppb). The Environmental Working Group (EWG) is the only group whose TTHM standards are stricter than the ones required by federal and state laws, which

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

the EWG and others consider inadequate. Their maximum allowed concentration is 0.15 ppb (Total Trihalomethanes, n.d.).

In the United States, only a few studies have examined the correlation between trihalomethane levels and sociocultural factors (Harris, 2009; Christman et al., 1983). According to these studies, trihalomethane concentrations in New York were positively correlated with the median family income, racial composition of localities, and community size. However, the previous findings were state-wide normalized information, and could not provide insight into what might be causing the correlations or what other physical factors might be at play.

Considering the previous studies, this research hypothesized that PWSs providing service to communities that are more affluent would have lower trihalomethane levels because of differences in system characteristics and treatment technologies. This study examined the disparities in trihalomethane concentrations among different New York public water systems by identifying the factors that affect these concentrations. Particularly, the association between trihalomethane levels and socioeconomic indicators was analyzed at the city (or town or village) level in the state of New York to determine how the features of water treatment systems may influence these relationships. Multiple statistical approaches were used to accomplish these goals. This is what makes this study unique among others of similar scope.

2. Data Analysis Methods

All communities investigated in this study were identified using the US EPA Drinking Water Information System (SDWIS) to determine the correct PWS IDs before linking the TTHM concentration data with the US Census Bureau data, such as median household income, race percentage, and population. The communities that were not identified on the Census Bureau website were excluded from this study. Most of the excluded towns were small villages with less than 5,000 community members. The total amount of PWS data retrieved was from over 210 systems providing services for approximately 80% of the residents of the state of New York. For towns receiving services from multiple PWS, all TTHM datasets were incorporated.

The correlations between the TTHM concentrations and socioeconomic factors were obtained using the Pearson correlation coefficient (ρ) method with corresponding p-values (Equation 1).

$$\rho = \frac{\sum_{i=1}^{n}(x_i-\underline{x})(y_i-\underline{y})}{(n-1)s_x s_y} \quad (1)$$

where $\underline{x}$, $\underline{y}$, s_x, and s_y indicate the averages and the standard deviations of groups x and y, respectively.

The significance level of the p-value (alpha) for either accepting or rejecting the null hypothesis, "no correlation between TTHM and socioeconomic factors," was set at 5% (0.05). The linear regression tool in the MS Excel data analysis function was used to determine the correlation coefficients and p-values. Additionally, samples were obtained from the NY Department of Health 2019 data for trihalomethanes concentrations and the 2016-2020 Census Bureau data for demographic information.

The outliers of the collected data were identified from each group of data using the Grubbs test, as shown in Equation 2 (Harris, 2009).

$$G = \frac{|questionable\ value - \underline{x}|}{s} \quad (2)$$

where $\underline{x}$ and s indicate the average and standard deviation of the data set, respectively.

A sample was excluded from the calculation when its computed g-value exceeded the critical g-value at the 95% confidence level. If the critical g-value for the specific sample number was unavailable, two pairs of the nearest (one above and one below) critical g-value sample numbers were used to determine the slope, which was then used to calculate a reasonable g-value for the specific sample number adopted for this research.

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

Further analysis was performed to determine whether there are statistically significant differences between the mean values of TTHM of two different socioeconomic factors. The 95% confidence level student's t-test values were compared with the computed t-test based on the mean comparison between the two selected factors. Equation 3 was used to determine the computed t-test between the two groups (Harris, 2009).

$$t = \frac{|\underline{x_1} - \underline{x_2}|}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1+n_2}} \qquad \text{where} \quad s_{pooled} = \sqrt{\frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1 + n_2 - 2}} \tag{3}$$

where s_{pooled} is the pooled standard deviation.

When the t-test value was greater than 95% confidence level student's t value, the compared groups were considered statistically different, rejecting the null hypothesis that there was no significant difference in mean values. Considering the sample sizes for the adopted socioeconomic factors (70 - 208), the tabulated student's t-test values were set between 1.960 and 1.994 at the 95% confidence level, indicating that the two samples were considered significantly different when the computed t-test values exceeded 1.994. All computed t-test values in this research were either higher than 1.994 or lower than 1.960, but none of them fell within that specific range.

The same mean comparison statistical data analysis was conducted after grouping PWS based on the water sources such as groundwater and surface water. The samples within the same socioeconomic variable were divided into two groups: high 50% vs. low 50%. When the total number of samples was odd, the median value was placed into the low 50% group. Based on Equation 3, the differences in specific socioeconomic factors between high and low 50s were reviewed to determine whether they were statistically significant.

3. Results

We examined the relationship between TTHM concentrations and various socioeconomic factors, including the mean household income, racial composition, and population of the serving area. We analyzed 285 PWS, covering 81% of the population of NY.

According to the statistical evaluation, certain socioeconomic factors, such as mean household income, percentage of Asian population, and percentage of Hispanic population, were significantly correlated with the TTHM concentration. Table 1 summarizes the Pearson correlation coefficients and the p-values between TTHM and various socioeconomic factors examined in this research.

Table 1. Pearson correlation coefficients between total trihalomethane (TTHM) concentrations and socioeconomic factors ($\alpha = 0.01$)

	Correlation Coefficient	P-value	Outliers (%)
Income vs. TTHM	-0.430	3.0×10^{-14}	0.96 %
% Asian vs. TTHM	-0.304	1.7×10^{-7}	1.44 %
% Hispanic vs. TTHM	-0.195	9.1×10^{-4}	0.96 %
% White vs. TTHM	0.105	0.076	0.96 %
% Black vs. TTHM	0.011	0.860	1.44 %
Serving size vs. TTHM	0.001	0.991	1.91 %

Table 1 indicates that a relatively small number of samples were identified as outliers and excluded from the statistical analysis. The percentages of the extracted samples for each factor can be found in Table 1. Figures 1 and 2 show the distribution of TTHM concentrations as a function of mean household income and the percentage of Asian population, respectively. According to the data analysis, the TTHM concentration showed a statistically significant relationship with medium income, indicating that an increase in income results in the decrease in the TTHM concentration. Asian and Hispanic percentages illustrated a similar trend, with an inverse relationship with the concentration of TTHM. Particularly, the median household income and the percentage of Asian population revealed a higher negative correlation constant, rejecting the null hypothesis that there is no relationship between TTHM concentration and these variables. However, the percentage of White and Black populations and the total serving size of the water system showed statistically unbiased TTHM concentrations. Figures

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

1 and 2 illustrate the sample distributions for medium household income and the percentage of Asian population adopted for this research.

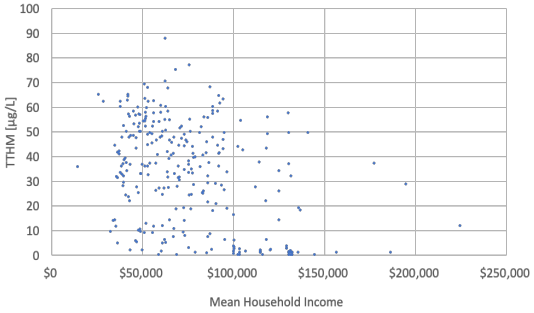

Figure 1. Mean household income plotted against total trihalomethane concentrations at city levels in New York State. Note: TTHM denotes total trihalomethane concentrations

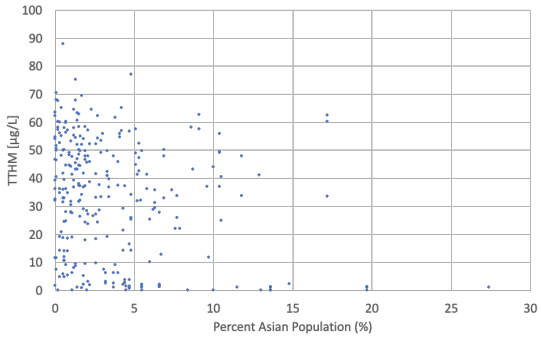

Figure 2. Percentage of Asian population plotted against total trihalomethane concentration at city levels in New York State. Note: TTHM denotes total trihalomethane concentrations.

(mean household income, percentage of Asian population, and community population) were statistically different depending on the water source. As shown in Table 4, no significant difference was observed between the mean TTHM concentrations based on the percentage of Hispanic populations in the two water sources.

Depending on water quality and availability, the State of New York has utilized both surface water and groundwater for drinking water resources. Tables 2 and 3 present the fundamental statistical data calculated separately for the two water resources, focusing on the three statistically significant variables. The average concentration of TTHM in groundwater is about 10.5 µg/L, which is significantly lower than that in surface water (44.0 µg/L). However, the mean values of the three main variables showed higher values in groundwater than those in surface water. According to Table 1, when the correlation coefficients were calculated for each water resource, the groundwater coefficients decreased more than the general values, indicating that the surface water coefficients were neither noticeable nor significant.

Several studies have been conducted to compare the statistical difference between the two mean TTHM concentrations of socioeconomic factors, using the mean student's t-test value comparison method. The computed t-test value of the two sets of TTHM concentrations between the two water sources, including all available socioeconomic factors with the application of Equation 2, was 16.6 (Table 4). This value was significantly greater than the higher cut-off student's t-test value of 1.994, indicating that the two mean TTHM concentrations were statistically different. Table 4summarizes the other results of the comparison between socioeconomic factors. Based on the cut-off range of student's t-test values (1.960-1.994), the mean TTHM values for the three socioeconomic factors

Table 2. Statistical data for TTHM concentration and mean household income with different water sources.

	TTHM (µg/L)			Mean household Income (US $)		
	Mean	STD*	Median	Mean	STD*	Median
Groundwater	10.5	11.9	5.2	99,836	76,903	89,685
Surface water	44.0	14.1	43.6	62,414	49,925	52,493

* Standard deviation

Table 3. Statistical data for Asian and Hispanic Populations with different water sources.

	Asian population (%)			Hispanic population (%)		
	Mean	STD*	Median	Mean	STD*	Median
Groundwater	5.8	5.9	3.9	12.3	9.1	11.6
Surface water	3.0	3.5	1.8	9.1	10.1	7.1

* Standard deviation

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

4. Discussions

While previous studies have indicated a relationship between the THM concentration and various natural variables (12, 13), a detailed systematic evaluation of socioeconomic factors has not yet been performed. Therefore, we examined the amount of THM in drinking water resources as a function of various socioeconomic variables, including median household income, ethnicity, and water source. According to the Pearson coefficients and p-values calculated, the mean household income has a robust reverse relationship with the concentration of TTHM, as shown in Table 1. In addition, the percentages of Asian and Hispanic populations showed negative correlations with the amount of TTHM in the water system. In contrast, the percentages of White and Black populations showed no statistical significance in relation to the amount of TTHM concentration. Since the data analysis indicated that the TTHM concentrations in groundwater were significantly lower than the ones in surface water sources (Table 2), further evaluations were performed to determine whether the water sources influenced the amount of TTHM in the water system. The mean comparison student's t-test method was used to compare the t-test values to the computed t-test values in order to investigate whether the socioeconomic factors in each water source were statistically different. As shown in Table 4, the mean comparison evaluation method confirmed that the difference in the amounts of TTHM between groundwater and surface water was statistically significant. These results suggested that the type of source water plays an essential role in the amount of TTHM in the water system. Additionally, this was confirmed by the computed t-test value of the mean TTHM concentration comparison, which is significantly higher than the student's t-test value at the 95% confidence level. Furthermore, the median household income and the percentage of Asian population were found to be different, supporting the observation that water source impacts the TTHM concentration. The percentages of Black and White populations were confirmed as unbiased variables in the two water sources and had no correlation with the TTHM amount in the water system (Tables 1 and 4). Moreover, the Hispanic population was the only socioeconomic variable that did not show a significant difference between the two water sources, but the amount of TTHM is negatively correlated with the percentage of population. Furthermore, although the average size of the population served by surface water is 43,000 people, which is significantly higher than that for groundwater cases (24,000 people/system), a correlation was not observed between the amount of TTHM and service size.

Table 4. Computed t-test values to determine the statistical difference between groundwater and surface water.

Socioeconomic Factor	Computed t-test value	S_{pooled} value
All socioeconomic samples	16.599	12.447
Mean Household Income	6.151	30,743
% Asian population	3.331	4.247
Population	2.495	56,067
% Black population	1.721	6.749
% White population	0.748	14.043
% Hispanic population	0.723	9.214

The social variables examined in this research were further statistically analyzed within each water source. For both groundwater and surface water, a mean comparison test of TTHM amount was conducted based on each socioeconomic variable. Each group was divided into two groups: upper 50% and lower 50%. The median value was included in the lower 50% when there were fewer samples. Figure 3 shows a summary of the data analysis. The approximate cut-off t-value for statistical significance was 1.99. When the TTHM amount was evaluated without considering the source of water, the differences in the amounts of TTHM were statistically significant between the upper 50% and lower 50% of the median household income and the percentages of Asian and Hispanic populations. These observations were consistent with the general correlation coefficient results, as shown in Table 1. When the same socioeconomic variables were evaluated for surface water systems, no correlation was observed between socioeconomic variables and TTHM. Furthermore, the median household income and the percentage of Asian populations were statistically significant between the upper and lower 50% of the values for communities served with groundwater.

Vol. 2023 (2) 131 – 136
ISSN 2688-3651 [online]

In summary, the water source is a critical factor in controlling the amount of TTHM in water. The amounts are significantly correlated with the median household income and percentage of Asian population when the communities are served with groundwater.

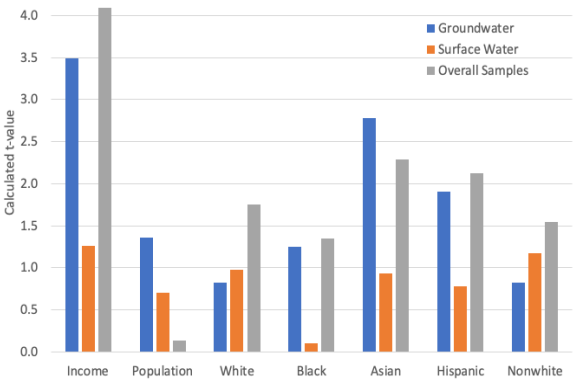

Figure 3. The t-test results for different water sources comparing the upper and lower 50% of each category. When the sample number is odd, the median value was included in the lower 50% of samples.

5. Conclusion

This study examined the relationship between the amount of trihalomethanes, a potential carcinogen, and various environmental and socioeconomic factors, such as income, water sources, and demographic factors. Generally, the amount of TTHM in water is related to the median household income, water source, and some demographic factors, such as composition of Asian and Hispanic populations. The correlation factor increased significantly when the public water system used groundwater. Moreover, high-income families and Asian residents living in communities served by groundwater sources benefited from low TTHM concentration than those served by surface water sources.

References

Allaire, M., et al. (2018). National Trends in Drinking Water Quality Violations. *Proceedings of National Academy of Sciences*, 115(9), 2078–2083. doi:10.1073/pnas.1719805115.

Christman, R., et al. (1983) Identity and yields of major halogenated products of aquatic fulvic acid chlorination. *Environ. Sci. Technol.* 17, 625 - 628.

Costet, N., et al. (2011) Water Disinfection By-products and Bladder Cancer: Is There a European Specificity? A Pooled and Meta-analysis of European Case–Control Studies. *Occupational and Environmental Medicine*, 68(5), 379−85. doi: 10.1136/oem.2010.062703.

DeMarini, D. (2020). A Review on the 40th Anniversary of the First Regulation of Drinking Water Disinfection By-products. *Environmental and Molecular Mutagenesis*, 61(6), 588–601. doi:10.1002/em.22378.

Harris, D. *Exploring Chemical Analysis*. 4th edition, W. H. Freeman and Company, New York, 2009.

Hildesheim, M., et al. (1998). Drinking Water Sources and Chlorination Byproducts II. Risk of Colon and Rectal Cancers. *Epidemiology*, 9(1), 29–35. www.jstor.org/stable/3702610.

Schaider, L., et al. (2019) Environmental Justice and Drinking Water Quality: Are There Socioeconomic Disparities in Nitrate Levels in U.S. Drinking Water? *Environmental Health*, 18(3). doi: 10.1186/s12940-018-0442-6.

Switzer, D., & Teodoro, M. (2017) The Color of Drinking Water: Class, Race, Ethnicity, and Safe Drinking Water Act Compliance. *Journal of American Water Works Association*, 109(9), 40–45. doi:10.5942/jawwa.2017.109.0128.

Total trihalomethanes (TTHMs), Retrieved May 20, 2023 from https://www.ewg.org/tapwater/contaminant.php?contamcode=2950