Week 2 Lab Report

Nguyen Tien Anh Quach

2024-02-01

Introduction

The objective of this document is to answer whether or not shale-gas extraction negatively impacts the groundwater resources in PA. Data analyses utilized methane data from Molofsky et al. (2013). Data were collected from 1701 drinking water wells in Susquenhanna County, PA to examine the prevalence and distribution of methane in groundwater. In the dataset, variables collected include (1) methane concentration, (2) proximity of a well from a fracking site, which is classified into near (<1 km) and far (>1 km), (3) location of a well, which is divided into valley and upland categories, and (4) whether or not a concentration value is above detection limit.

Results

Summary statistics of methane data

The total number of wells is 1701; however, they are predominantly far wells (Table 1). Although the number of upland and valley wells is relatively equal, far wells in each location category still dominate numerically (Table 2). Methane concentrations varied widely among all wells (0.05, 4.3 x 10^4), wells in upland (0.05, 3.2 x 10^4) and valley (0.08, 4.3 x 10^4), and wells near (0.08, 4.3 x 10^4) to and far (0.05, 3.9 x 10^4) from fracking sites. Far wells in the valley have relatively similar methane concentration (Table 3), whereas far wells in the upland have higher methane concentration, standard deviation, and as a result, skewness (Table 4).

The methane data collected are highly right-skewed for all wells, well locations, and well proximities (Fig. 1, 2, & 3). The authors did not give any explanation about the censored observations and how they play a role in the analyses. Therefore, the analyses in this lab did not take into account of the censored observation.

Statistical tests

Near vs. Far wells for all observations

Due to the nature of the data, I expected non-parametric tests are needed to test for the mean difference in methane concentrations (Fig. 1). Indeed, a Shapiro-Wilk normality test yielded a p-value of 0, which indicated that the methane data are highly non-normal. Log-transforming did not help generate a normal distribution, as the p-value of the Shapiro-Wilk test of the logged data returned a p-value of 0. Outliers cannot be removed because it is unknown how a well would have a distinctively high methane concentration, which could be due to (1) equipment error, (2) data entry error, or (3) methane is naturally concentrated in that well.

My hypothesis is that far wells would have lower mean methane concentration, as they are far away from fracking sites. Therefore, a one-sided Wilcoxon rank-sum test was conducted. Results showed that the measured mean concentration of methane at far wells was statistically and significantly lower than that at near wells ($W = 1.71369 \times 10^5$, $p = 8.21 \times 10^{-11}$, $d = 0.154974$).

Interestingly, the authors instead conducted a two-sided Mann-Whitney U-test. Their hypothesis was far wells could have higher or lower concentrations of methane. However, that would not answer their question, as they were asking if fracking has any impacts on methane concentration. In addition, their test result was non-significant ($p = 0.503$). My analysis result has shown that the test was likely done incorrectly to generate the results they were expecting.

Upland vs. Valley wells

As data for this analysis are similar to those used to compare near vs. far wells, it was not necessary to test for the normality of the original and transformed data. My hypothesis was that the mean concentration of methane in upland wells is less than that in valley wells. A one-sided Wilcoxon rank-sum test was conducted. Results showed that the measured mean concentration of methane at upland wells was statistically and significantly lower than that at valley wells ($W = 2.84413 \times 10^5$, $p = 1.18 \times 10^{-14}$, $d = 0.1849923$). This was the only test that the results of my analysis and the authors' analysis were similar.

Near vs. Far wells for valley

Methane data collected from valley wells are highly right-skewed (Fig. 2). Indeed, a Shapiro-Wilk normality test of methane data yielded a p-value of 0, which indicated that the methane data are highly non-normal. Log-transforming did not help generate a normal distribution, as the p-value of the Shapiro-Wilk test of the logged data returned a p-value of 0. Outliers were not removed due to similar reasons (see section above). Therefore, a one-sided Wilcoxon rank-sum test was conducted. Results showed that the measured mean concentration of methane at far wells was statistically and significantly lower than that at near wells ($W = 5.55645 \times 10^4$, $p = 7.27 \times 10^{-4}$, $d = 0.1082596$).

Interestingly, the authors also conducted a one-sided Mann-Whitney U-test. They also obtained a statistically significant p-value of 0.007. However, my p-value is about 10 times smaller,which demonstrated that that the test was likely done incorrectly or data were not cherrypicked to generate the results they were expecting.

Near vs. Far wells for upland

Methane data collected from upland wells are also highly right-skewed (Fig. 3). A Shapiro-Wilk normality test of methane data yielded a p-value of 0, which indicated that the methane data are highly non-normal. Log-transforming did not help generate a normal distribution because the p-value of the Shapiro-Wilk test of the log-transformed data returned a p-value of 0. Therefore, a one-sided Wilcoxon rank-sum test was performed. The test indicated that the measured mean concentration of methane at far wells was statistically and significantly lower than that at near wells ($W = 3.2805 \times 10^4$, $p = 5.27 \times 10^{-7}$, $d = 0.1688332$).

Interestingly, the authors again conducted a two-sided Mann-Whitney U-test, which had an insignificant result ($p = 0.154$). However, my analysis demonstrated that there was actually a significant difference in mean values of methane concentration between far and near wells for upland observations.

Discussion

In general, my results only aligned with the authors' results in the analysis of upland vs. valley wells, where we both found that mean methane concentration at upland wells was significantly lower than that at valley sites. For the other analyses, I found that far wells had significantly lower mean methane concentration. Meanwhile, Molofsky et al. (2013) found non-significant results for those analyses, showing that far and near wells had no significant difference in mean methane concentrations.

What the authors were lacking was a clear alternative hypothesis. As a result, their test choice was not correct (i.e., one-sided instead of two-sided) and their results were ambiguous. Perhaps their ultimate goal was to be vague and to show readers that fracking indeed poses no harm to groundwater resources through their poor selection of statistical tests. My results seemed to be more intuititive, as wells near the fracking sites were shown to have higher mean methane concentration. In addition, our similar result from the upland vs. valley test may have shown that fracking is having a negative impacts on groundwater and perhaps surface water. Groundwater travels down the hydraulic gradient, carrying methane. As valley is within 1000 ft of a major NHD flowline and 500 ft of minor tributaries, the impact of fracking on the methane concentration at valley wells is quite apparent.

Many limitations exist in this study. First, it is questionable why there is a censored variable. Second, there are so many categorical variables, which can be turned into continuous variables. For example, distance to the nearest stream could replace the location variable. Distance to fracking sites could replace the proximity variable. Third, there are many wells and many of them are far from the sites. Near wells are needed to further determine the impacts of fracking on groundwater. Fourth, more water chemistry could be collected, other than methane and metal concentrations. Last, this paper presented a huge conflict of interest. A big question could come up from the readers after reading toward the end and encountering the Acknowledgement section: Why would someone report that fracking is harmful when they are funded by a fracking company to do research and collect data?

Appendix

Table 1. Summary statistics of methane concentration at near and far wells for all observations.

Proximity Category	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	1379	684.2574	0.6	3132.928	15.830	6.230358
Near	322	795.0171	5.9	4086.957	25.575	6.467009

Table 2. Summary statistics of methane concentration at upland and valley wells.

Location	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Upland	836	198.2077	0.47	1644.111	3.565	13.326158
Valley	865	1195.2426	1.80	4331.548	25.780	4.790541

Table 3. Summary statistics of methane concentration at far and near wells in the valley.

Proximity Category	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	670	1186.406	1.3	4058.772	25.8075	4.510468
Near	195	1225.604	19.0	5172.061	25.4800	5.020043

Table 4. Summary statistics of methane concentration at far and near wells in the upland.

Proximity Category	Number of Sites	Mean Methane Conc.	Median Methane Conc.	Standard Deviation of Methane Conc.	IQR of Methane Conc.	Skewness of Methane Conc.
Far	709	209.7305	0.4	1753.1023	2.25	12.798759
Near	127	133.8798	1.4	799.4312	25.69	8.959691

Figure 1. The distribution of methane concentration at near and far wells for all observations.

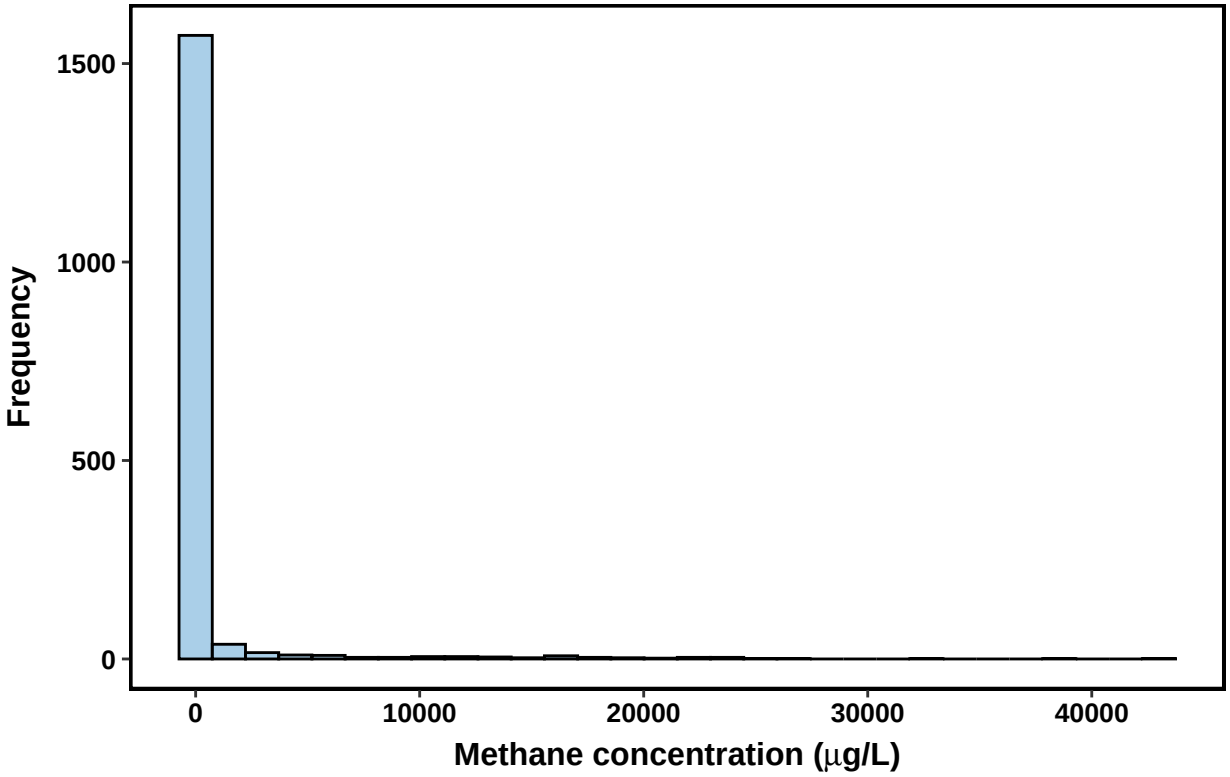

Figure 2. The distribution of methane concentration at far and near wells in the valley.

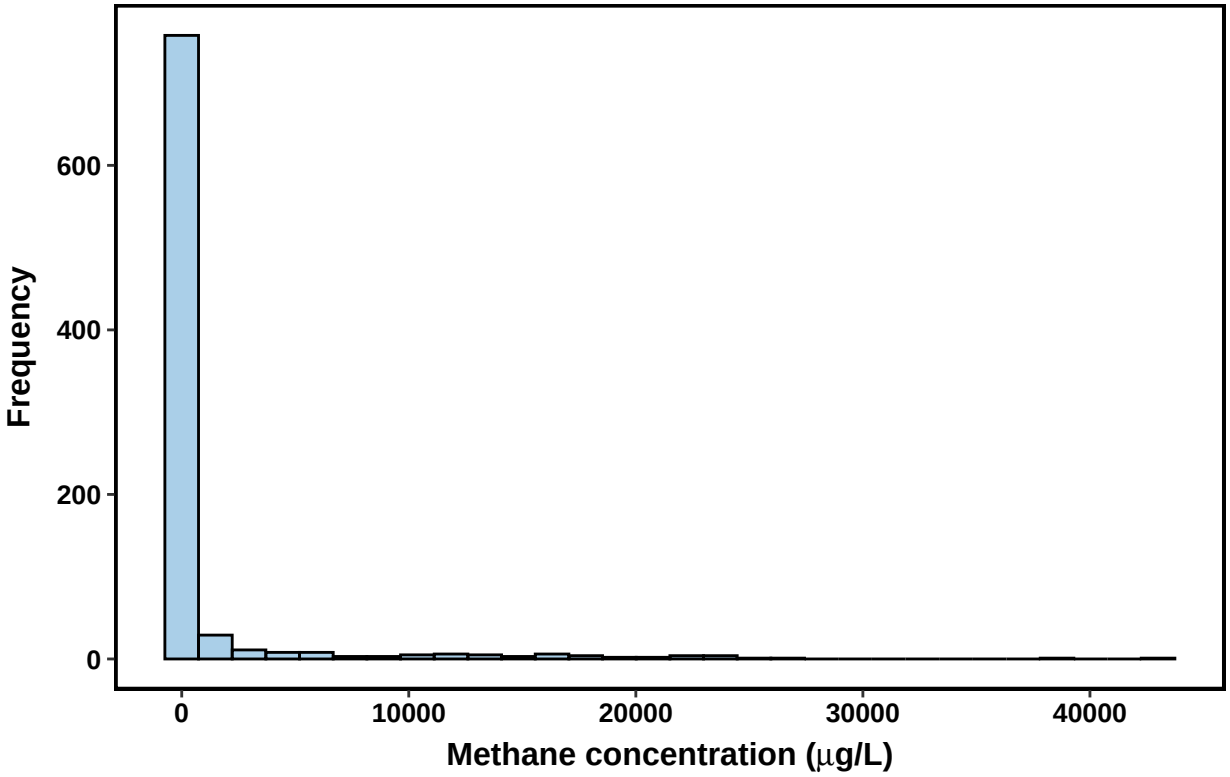

Figure 3. The distribution of methane concentration at far and near wells in the upland.