My amazing title

Tony Ni
April DD, 20YY

Submitted to the Department of
Mathematics and Statistics
of Amherst College in partial fulfillment
of the requirements for the degree of
Bachelor of Arts with honors.

Advisor:
Brittney Bailey

Abstract

The abstract should be a short summary of your thesis work. A paragraph is usually sufficient here.

Acknowledgments

Use this space to thank those who have helped you in the thesis process (professors, staff, friends, family, etc.). If you had special funding to conduct your thesis work, that should be acknowledged here as well.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Coal is one of the most dangerous combustible fuels which is being burned in all across the world as one of the largest methods of obtaining energy. Yet, although it is a fossil fuel which is naturally abundant and easy to utilize, it is comprised of a long list of dangerous chemicals including – but not limited to: arsenic, radium, boron, and a large list of other chemicals which prove to be dangerous to humans and animals alike. (Kelderman et al., 2019)

Power plants produce electricity by burning this coal, and as a result of how prevalent it is within the US - over 100 million tons of coal ash are produced every year. This side-product as a result of the coal combustion is often disposed by directly being dumped into landfills and waste ponds. (Kelderman et al., 2019)

Only recently have these complaints and lawsuits regarding the disposing practices made by non-profit environmental organizations been heard. Due to the onslaught of pressure put on the Environmental Protection Agency – the Coal Ash Rule was bor in 2015. (Kelderman et al., 2019)

This rule has forced over 265 coal power plants – about 3/4 of all coal power plants in the US - to make data regarding chemical concentrations publicly available to the general population. (Kelderman et al., 2019)

In their analysis using this data, the Environmental Integrity Project – a non-profit

organization dedicated to issues involving environmental justice have concluded that essentially all groundwater under coal plants are contaminated. (Kelderman et al., 2019)

However, is this really the case? There are many naturally occurring chemicals existing in groundwater as as such, perhaps their claims are overstated.

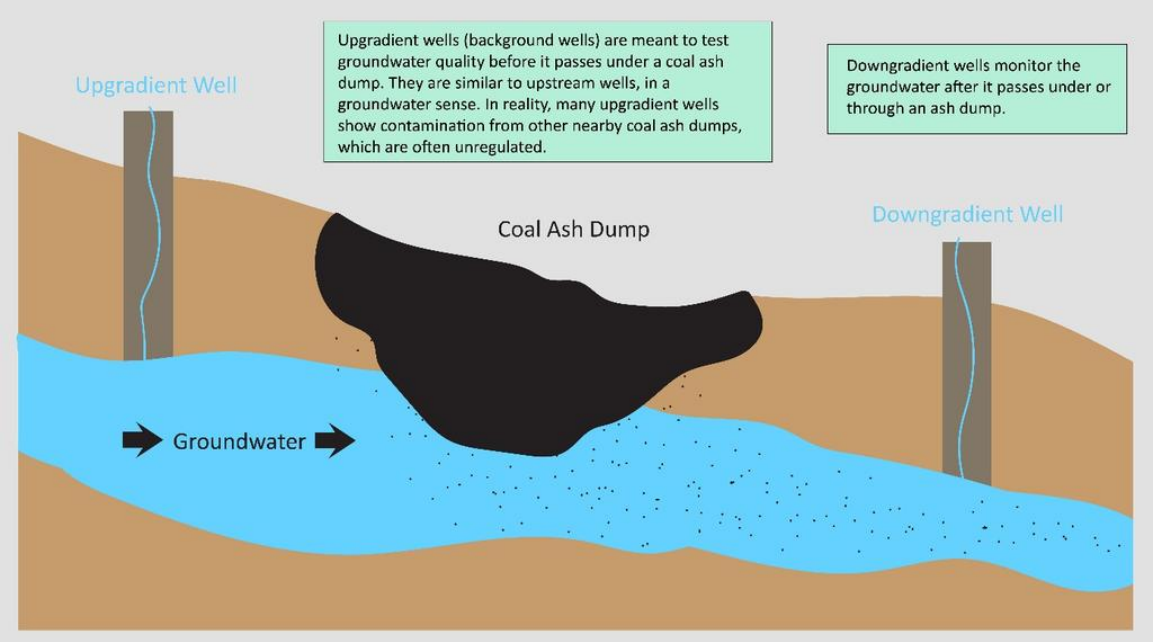

Figure 1.1: Difference Between Upgradient and Downgradient Wells

Typically in a coal ash plant, there exists two types of wells: upgradient wells and downgradient wells. These wells are essential to measure the amount of contamination being caused by coal ash. Upgradient wells, also known as background wells, measures the concentrations of chemicals in groundwater before it passes through an coal ash dump. Conversely, downgradient wells measure the concentrations of chemicals in groundwater after it passes through a coal ash dump.

- "80% of the US population is served by 14% of the utilities," so if something were to get into the water distribution system, it can easily spread amongst the

US population which is why contamination in water services is so important. (Byer & Carlson, 2019)

With this information, typically – one estimates the amount of chemical contamination caused by a coal as dump by subtracting the upgradient concentration from the downgradient concentration of a chemical (downgradient concentration - upgradient concentration).

However, due to the lack of proper reporting guidelines prior to the enactment of the Coal Ash Rule, we believe that there may be retired or even unregulated upgradient wells which can cause the concentrations of chemicals being recorded from these upgradient wells to be inaccurate or even completely wrong.

Our end goal remains the same as the EIP: to identify contaminated groundwater in coal plants – but to attempt to find a way to effectively correct the improper/inaccurate values resulting from LOD errors and other factors which the EIP may not have considered.

The limit of detection problem stems from the measuring devices' inability to obtain chemical concentrations smaller than a certain threshold amount, thus affecting the measurements recorded.

Our plan is to utilize bootstrapping and imputation techniques to correct for these measurements by accounting for the innate contamination which may be caused by factors such as retired and unregulated wells that were mentioned before.

1.2 Data

1.2.1 Coal Ash Rule

A large coal ash spill at the Tennessee Valley Authority (TVA) which occured on December 22, 2008 in Kingston, TN – prompted the Environmental Protection Agency

(EPA) to propose a set of standardized regulations and procedures to address the concerns regarding coal ash plants nationwide in the US. (Environmental Protection Agency, 2020)

This was known as the Coal Ash Rule, passed on December 19, 2014. (Environmental Protection Agency, 2020)

Changes were made to the Coal Ash Rule over the years in the form of 'amendments,' one of which made required facility information and data to be made publicaly available to the public (April 15, 2015 rule change) (Environmental Protection Agency, 2020)

1.2.2 Source of Data

The data used in the study are from the results published in "Annual Groundwater Monitoring and Corrective Action Reports" which were made available to the public in March 2018 as a result of the Coal Ash Rule. (Environmental Integrity Project, 2020)

These reports are in PDF format and are thousands of pages long, which makes it difficult for individuals to look through the data in a meaningful way. (Environmental Integrity Project, 2020)

The EIP obtained the data from an online, publicly available database containing groundwater monitoring results from the first "Annual Groundwater Monitoring and Corrective Action Reports" in 2018 which was collected from coal plants and coal ash dumps under the Coal Ash Rule (Environmental Integrity Project, 2020)

They wrangled the data into a more accessible machine-readable format which contains information from over 443 annual groundwater monitoring reports posted by 265 coal ash plants, which is downloadable from the EIP's website. (Environmental Integrity Project, 2020)

1.2.3 Variables

The dataset contains information regarding chemical concentrations at coal plants. A coal plant consists of multiple disposal areas for the coal ash that it produces. At each disposal area, there are specific locations that groundwater is being measured, known as wells which represent an observation in the dataset. There are two types of wells – upgradient and downgradient wells. The variables consist of information regarding the specific chemical concentrations of each well. From the 19 different contaminants (antimony, arsenic, boron, etc.) a major problem is that some wells only have measurements for certain chemicals and don't have them for others.

1.2.4 Plan of Action

Within the report, the EIP mentions certain restrictions within the data that have caused their data to potentially be inaccurate (specifically, with limit of detection problems, and a large amount of missing chemical data). The limit of detection problem comes when measuring devices used to measure chemical concentrations are unable to detect below a certain threshold, causing large numbers of observations to have duplicate, wrong values – which can cause for misguided analysis.The other issue is less guided/formed, but for brevity, we think that a lot of the issues in the data comes from the potential possibility of contamination during data collection from investigators from non coal-ash sources. This may include things like: retired/unregulated wells which are old and have chemicals leaking into the groundwater, mismanagement in measuring, etc.My project hopes to work with methods on handling this missing data – alongside investing potential uses of bootstrapping and other resampling methods (potentially?) in order to try to come up with a more statistically accurate and sound result by looking to assuage the problems that the EIP faced in their analysis. Specifically, to

find a way to split up the data into ”uncontaminated” and ”contaminated” wells in order to find the natural distribution of chemicals in each – and doing to so in the face of data corrupted by LOD problems and inaccuracies. I'm hoping to apply and compare different ways of altering the data to account for these myriad of issues in order to look for more salient findings that the EIP might have missed or if not, to see if improvements can be made regarding the way that contaminated coal ash wells are being identified.

Chapter 2 Methodology

2.1 Overview

- There are many different ways that researchers are dealing with detection limits, which are ubiquitous in the scientific realm. Substitution, nonparametric methods, and maximum likelihood methods are all ways to combat this problem. (Lafleur et al., 2011)
- Detection limits are constantly changing: as technology improves, so too does our ability to accurately measure substances (Elias & Goodman, 1999)
- Reporting limits can often be used be misleading as non-statisticians (practices in environmental law) may often interpret ND (non-detect) as nullity, when in fact it only means that the measurement falls below a certain limit (Elias & Goodman, 1999)
- Substitution is the worst way, nonparametric ways do better, maximum likelihood methods are the best (Lafleur et al., 2011)
- The general form of the equation used to determine LOD is of the form:

$$LOD = k * s_{zero}$$

where "k is the constant for defining LOD" and s_{zero} is the standard deviation

of the zero/blank" (Akard, Tsurumi, Oestergaard, & Inoue, 2002)

- 2 or 3 SDs is often what is used, refer to picture (I think it's neat) (Akard et al., 2002)

- Throwing away/Discarding values that are below of the LOD gets rid of tons of useful information (Berthouex, 2020)

- LOD entries still contain information that a lot of people don't realize – specifically information that the values is between 0 and the LOD. (Chen et al., 2011)

- The way that chemists/researchers report low concentrations are varied and not standardized. Some may report: "ported. They may report the datum to the data analyst as (1) trace, (2) the letters ND (not detected), (3) the numerical value of MDL itself, (4) a"less than" value, that is, the numerical value of the MDL preceded by a "<" sign, (5) zero, (6) some value between zero and the MDL, for example, one-half the MDL, (7) the actual measured concentration even if it is below the MDL (that is, whether the value is positive or negative), (8) the actual measured value followed by the MDL in parenthesis, or (9) the actual measured value with a statement of its precision (for example, 2 ± 4 Mg/L, where the $\pm$ value indicates the precision of the estimate). The last three methods are the best." (Berthouex, 2020)

2.1.1 Substitution Approach

- Substitution methods are easy to implement, but are biased (common values: LOD/2, LOD/sqrt(2), LOD) but are discouraged b/c results in estimates of parameters being biased (Chen et al., 2011)

- First method they tried was replacing all values with LOD/2, they claim these method was recommended for datasets where lots of data is below LOD or when data is highly skewed with a geometric sd of 3 or more. Some people also use LOD/sqrt(2) and is recommended to be used when few data us below LOD or when data is not highly skewed Through their study, they found that there really isn't a difference between these two methods (Glass & Gray, 2001)

2.1.2 Kaplan-Meier Estimate Approach (#Kaplan-Meier)

- Suggests using the reverse Kaplan-Meier (KM) estimator to estimate the distribution function and population percentiles for data where there is "left-censored data" (data point is below a certain value but known by how much) (Gillespie et al., 2010)

- After their study, they found that even though THEORETICALLY the reverse KM is for left-censored data (just like KM is for right-censored data), it is still limited in its usage since all it does it estimate a distribution function (it's really just an exploratory thing, they said) (Gillespie et al., 2010)

2.1.3 Distribution-Based Multiple Imputation Approach (#DBML)

- Study exploring different options to handle LOD laboratory data – specifically with regards to multiple imputation methods for left-censored data. They concluded that "the distribution-based MI method" worked well for bivariate data where the values were < LOD. (Chen et al., 2011)

- What this study used was distribution-based multiple imputation methods – they used MLEs to estimate distribution parameters based on all datas (< LOD and those not). They repeatedly imput the values to create multiple complete

sets of data, and then analyzed each one individually (Chen et al., 2011)

- Mathematically, they created a log-likelihood function with all the data, then derived MLEs of each parameters on multiple bootstrapped datasets. Each bootstrap data gives different estimates for the mean, sd, etc. (refer to article for math) (Chen et al., 2011)

2.1.4 Write about other approaches here

- Another method is "Cohen's Method" where one extrapolates the left hand side of distribution based on the distribution of the uncensored data and then calculate the MLE estimate of the arithmetic mean – found to be unreliable with data with outliers, this method can ONLY be used with data where there is a single LOD. From their study they found that this method gave high, unlikely results of the mean (Glass & Gray, 2001)

- Imputing from a uniform is useful when you don't know the distribution of the data (Canales, 2018)

2.1.5 Comparison between different methods

- Compared 5 methods: found that in terms of performance: imputation method using MLE to estimate distribution parameters and then imputing censored data points with values from this distribution below the LOD > imputation from a uniform distribution > other 3 methods (substitution, log-normal MLE to estimate mean and SD, and kaplan-meier estimate) (Canales, 2018)

- KM method is better than MLE for data where there are TONS of missing data or if data is highly skewed (distribution not assumed in KM method) (Canales, 2018)

- Uses RMSE (root mean squared error) to see how close the estimated values are to the true values (lower RMSE means closer estimation to known values) (Canales, 2018)

2.1.6 Packages

- MLE and KM methods were implemented using the NADA package (https://cran.r-project.org/web/packages/NADA/NADA.pdf) in R (`cenmle` and `cenfit` functions) where data is labeled as censored or uncensored, for censored values, LOD is used as a placeholder, since these methods aren't imputation methods – these censored values weren't replaced. Instead, summary statistics were generated with the entire data set (including the censored data) (Canales, 2018)

- The imputations methods used mostly followed the general ideas: we have to assume the entire data set follows a particular distribution. Then we use this distribution to impute in values for the censored data. The MLE imputation method uses MLE methods to estimate the parameters of a distribution to fit the dataset, then values lower than the LOD are imputed FROM this dataset for all censored values (they used the function `fitdistcens` from the R package `fitdistRplus`). The second uniform imputation method assumes a uniform distribution with minimum 0, maximum LOD – for all values less than the LOD, then the left-censored values are replaced with a number randomly selected from this uniform distribution. (Canales, 2018)

Corrections

A list of corrections after submission to department.

Corrections may be made to the body of the thesis, but every such correction will be acknowledged in a list under the heading "Corrections," along with the statement "When originally submitted, this honors thesis contained some errors which have been corrected in the current version. Here is a list of the errors that were corrected." This list will be given on a sheet or sheets to be appended to the thesis. Corrections to spelling, grammar, or typography may be acknowledged by a general statement such as "30 spellings were corrected in various places in the thesis, and the notation for definite integral was changed in approximately 10 places." However, any correction that affects the meaning of a sentence or paragraph should be described in careful detail. The files samplethesis.tex and samplethesis.pdf show what the "Corrections" section should look like. Questions about what should appear in the "Corrections" should be directed to the Chair.

References

Akard, M., Tsurumi, K., Oestergaard, K., & Inoue, K. (2002). Why the Limit of Detection (LOD) Value is Not an Appropriate Specification for Automotive Emissions Analyzers. *SAE Technical Papers*, *111*(2002), 1321–1328. http://doi.org/10.4271/2002-01-2711

Berthouex, P. (2020). A Study of the Precision of Lead Measurements at Concentrations Near the Method Limit of Detection, *65*(5), 620–629.

Byer, D., & Carlson, K. H. (2019). Real-time detection of intentional chemical contamination in the distributional system, *97*(7), 130–133.

Canales, R. (2018). Methods for Handling Left-Censored Data in Quantitative Microbial Risk Assessment, *84*(20), 1–10.

Chen, H., Quandt, S. A., Grzywacz, J. G., Arcury, T. A., Environmental, S., Perspectives, H., ... Arcury, T. A. (2011). A Distribution-Based Multiple Imputation Method for Handling Bivariate Pesticide Data with Values below the Limit of Detection, *119*(3), 351–356. http://doi.org/10.1289/ehp.1002124

Elias, D., & Goodman, R. C. (1999). When Nothing Is Something : Understanding Detection Limits, *13*(4), 519–521.

Environmental Integrity Project. (2020). Coal Ash Groundwater Contamination: Documenting Coal Ash Pollution. Retrieved from https://environmentalintegrity.org/coal-ash-groundwater-contamination/

Environmental Protection Agency. (2020). Disposal of Coal Combustion Residuals from Electric Utilities Rulemakings. Retrieved from https://www.epa.gov/coalash/coal-ash-rule

Gillespie, B. W., Chen, Q., Reichert, H., Franzblau, A., Lepkowski, J., Adriaens, P., ... Garabrant, D. H. (2010). Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. *Epidemiology*, *21*. http://doi.org/10.1097/EDE.0b013e3181ce9fD8

Glass, D. C., & Gray, C. N. (2001). Estimating mean exposures from censored data: Exposure to benzene in the Australian petroleum industry. *Annals of Occupational Hygiene*, *45*(4), 275–282. http://doi.org/10.1016/S0003-4878(01)00022-9

Kelderman, K., Kunstman, B., Roy, H., Sivakumar, N., Mccormick, S., & Bernhardt, C. (2019). Coal's Poisonous Legacy: Groundwater Contaminated by Coal Ash Across the U.S.

Lafleur, B., Lee, W., Billhiemer, D., Lockhart, C., Liu, J., & Merchant, N. (2011). Statistical methods for assays with limits of detection: Serum bile acid as a differentiator between patients with normal colons, adenomas, and colorectal cancer. *Journal of Carcinogenesis*, *10*, 1–8. http://doi.org/10.4103/1477-3163.79681