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

The introduction should provide an overview of the work you set out to do and provide structure for the remainder of the document.

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.

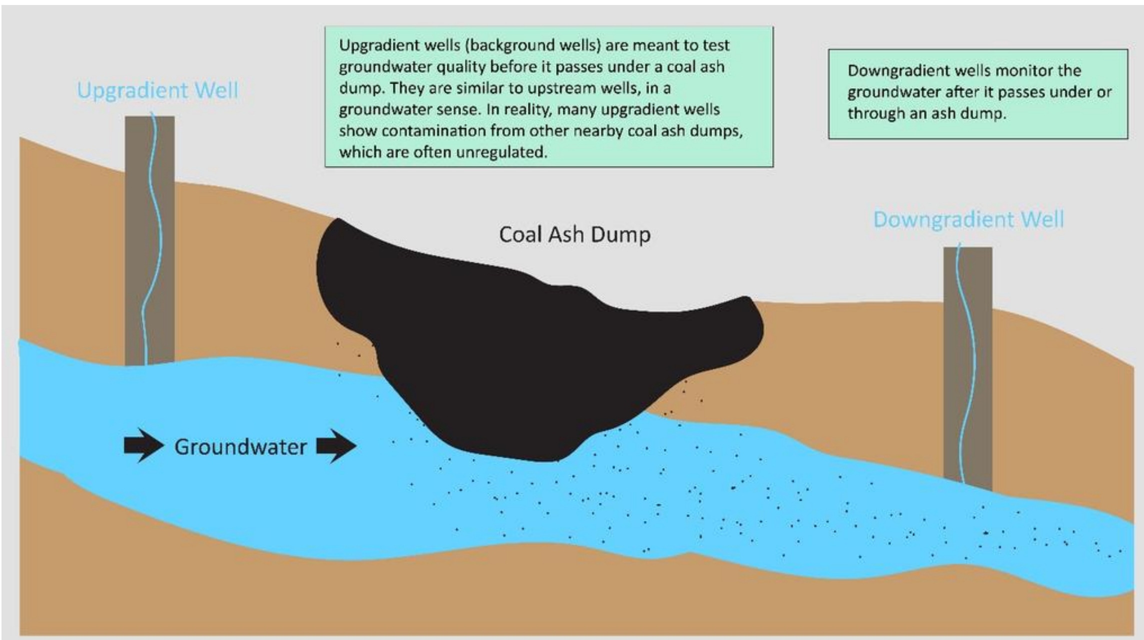

Figure 1.1: Difference Between Upgradient and Downgradient Wells

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.

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 (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 wranged 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 (Environmental Integrity Project, 2020)

- they 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)

1.2.3 Variables

- a coal ash site consists of multiple disposal areas
- within these disposal areas lie multiple wells
- each observation represents a well
- wells are split into 2 different types - upgradient and downgradient wells
- variables consist of information regard chemical contaminant concentrations and specifics regarding the 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
- we are currently using information from plants within illinois but there is data for all the states in the US

Chapter 2 Methodology

2.1 Plan of Action

- we wanted to identify these contaminated upgradient wells and then "correct" these measurements

- we will use manual code to flag contaminated vs noncontaminated wells (filter) using threshold values (this table is from coal ash pdf, but could make manual one)

*INSERT TABLE OF THRESHOLD VALUES HERE (working on it atm! sorry :())

- firstly, we used agglomerative hierarchical clustering to identify contaminated upgradient wells in our 'illinois' dataset (thoughts, maybe we want to expand/use a bigger dataset) using Ward's Method

- then, we separated our data into two parts – one dataset containing these contaminated upgradient wells and another dataset containing UNcontaminated upgradient wells

- then, we randomly sampled (with replacement) (500) times from the measurements of the chemical from non-contaminated upgradient wells to create an empirical distribution of naturally occurring chemical levels. this serves as the set of imputed "corrected" measurements of the chemical for each contaminated upgradient well

- then, we identify the specific 'disposal_area' that the contaminated wells belong to and FILTERED to have a dataset contain only the downgradient wells that corresponded to the upgradient wells – calculating the average of the downgradient wells (for the illinois dataset, we only had contaminated upgradient wells from TWO disposal areas)

- finally, we subtracted each of the (500) imputed upgradient measurements from the average downgradient measure. This creates a distribution of (500) values of the contaminant concentrations caused by the disposal area.

- we can then take the median of these (500) values as the estimate of the contamination caused by the disposal area (for the given chemical) and then use the 2.5 percentile and 97.5 percentile of the distribution as a bootstrap-type confidence interval.

- we found that the first disposal area didn't have any obvious contamination b/c the difference that we calculated (upgrad - downgradient) was mostly 0, while for the second disposal area the different was much greater than 0

2.2 Clustering

- unsupervised ml task whose goal is to divide the data in to clusters without knowing what the groups will look like beforehand (Lantz, 2013)

- used mainly for knowledge discovery rather than prediction (Lantz, 2013)

- many different ways to go about conducting a clustering based investigation, k-means clustering is the method used to try to find relationships between the wells

- our reasons to using this is to see whether if we can identify contaminated wells from uncontaminated wells (we don't anticipate it working due to the messed-up data, but MAYBE we would want to do some sort of study where we 1. run clustering with the messed up data and compare it to 2. run clustering with the corrected data (whatever that might be))

2.2.1 K-Means Clustering

- very popular and widely used clustering algorithm even since its inception decades ago (Lantz, 2013)

- STRENGTHS: uses simple ideas to identify clsuters that can be explained in non-statistical terms, is flexible and has lots of parameters which can be adjusted to address its issues, and it is efficient (Lantz, 2013)

- WEAKNESSES: not as sophisticated than some recent clustering techniques which have arisen recently, since it uses randomness within it, the clusters which it finds is not guaranteed to be optimal, requires a guess as to how many clusters may naturally exist in the data in order for the algorithm to run (Lantz, 2013)

- HOW IT WORKS: (add in later, if relevant?)

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

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

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.

Lantz, B. (2013). *Machine Learning with R*. Birmingham: Packt Publishing.