Introduction

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

Draft of Introduction

- (coal ash report - car) coal one of the most dangerous combustible fossil fuels is comprised of a long list of dangerous chemicals – including substances such as arsenic, radium, other carcinogens, metals that can impair developing children's brains, toxins dangerous to aquatic life, etc.
- power plants produce 100mil tons of coal ash every year, which is dumped into landfills and waste ponds
- only recently (2015) have complaints and lawsuit arisen in which certain ecological organizations have attempted to sue the EPA to regulate disposal of coal ash
- this coal ash rule has forced power companies to make publicly available data regarding chemical concentrations in 265 coal plants containing ponds and landfills (about 3/4 of all coal power plants across the US)
- environmental agencies have concluded that the groundwater under basically all coal plants are contaminated
- HOWEVER this might be overstated? we wanted to investigate whether or not if this was true.
- wells are split into 2 different types - upgradient and downgradient

- upgradient wells (background wells) measures groundwater chemical levels BEFORE passing through a coal ash dump while downgradient wells monitor the groundwater AFTER it passes through an ash dump
- we have reason to believe that many chemicals are NATURALLY OCCURING and as such, the statement made my environmental agencies regarding all groundwater being contaminated may be overstated
- typically, we would estimate the amount of chemical contamination, for this example – arsenic – caused by a coal ash dump with the equation: downgradient arsenic concentration minus upgradient arsenic concentration
- however, because there may be retired/unregulated upgradient wells that are occasionally contaminated already, this might be inaccurate
- we wanted to identify these contaminated upgradient wells and then "correct" these measurements
- 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

---

*everything below is just ideas on where to go from here

*something that needs to be mentioned is that the chemical concentration values for some of the wells are strange in that the instruments that were used to measure concentrations had a certain threshold value that they were able to detect. for example, the tool they used can only detect antimony at concentrations above 0.0010000 and if the actual concentration were lower, it reports it as 0.0010000. this leads to a LOT of data points which have these meaningless threshold values as measurements, which may have drastically affected our analysis…

*this could potentially be something interesting to explore, and try to find a way that missing data imputation techniques work and explore how it might be relevant to this dataset?

Resources for new(er) users

If you are new to bookdown, the package this thesis template is built on, I highly recommend bookmarking Chapter 2 of Yihui Xie's book, "bookdown: Authoring Books and Technical Documents with R Markdown" as a reference for **R** and LaTeX components useful to writing your thesis, including

- *R Markdown* syntax
- math expressions
- numbering and referencing equations
- special chunks for *theorems*, *definitions*, *proofs*, etc.
- captioning and referencing figures
- captioning and referencing tables
- in-text citations and bibliographies

What to expect

While Yihui Xie's books is the best resource, the remainder of this document will include examples of commonly used formatting. The amherst thesis github repo also contains a guide on dealing with all the different files in this thesis directory.