Methodology

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

Clustering

- unsupervised ml task whose goal is to divide the data in to clusters without knowing what the groups will look like beforehand [@Lantz2013]
- used mainly for knowledge discovery rather than prediction [@Lantz2013]
- 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))

K-Means Clustering

- very popular and widely used clustering algorithm even since its inception decades ago [@Lantz2013]
- 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 [@Lantz2013]

- 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 [@Lantz2013]
- HOW IT WORKS: (add in later, if relevant?)