Classification of Soil Contamination

Aleo Mok
aleo.mok@gmail.com
CS229 Final Project Report

ABSTRACT: Soil pollution is can be considered to be an imbalance of chemicals in the soil at a particular site. Such unnatural contamination must be addressed to avoid hazard to the environment and inhabitants of a polluted site. However, it is important to first be able to identify whether a site is contaminated before determining a solution. This paper explores the classification of soil samples at a particular site (McConnell Air Force Base) to investigate the natural or unnatural contamination of soil. The samples are addressed using k-means clustering, k-fold cross validation, and Gaussian discriminant analysis. From these evaluations, contamination at the site of interest can be considered.

1. Introduction

Soil contamination is characterized by solid or liquid hazardous substances mixed with naturally occurring soil. Soil pollution can arise from a number of sources, which could be both naturally-occurring in soil and man-made. In other words, the ratio of chemicals in the soil of a given site may be attributed to both natural and unnatural accumulation or production of compounds due to specific environmental conditions. These contaminants can adversely impact the health of plants, animals, and humans when directly or indirectly coming into contact with contaminated soil. Due to the detrimental nature of contamination and the multiple methods to address soil pollution, it is of key interest to be able to determine whether specific sites have contaminated soil.

Soil contaminants may vary in both location and type. In this report, the scope of the soil contamination is limited to soil samples taken from the McConnell Air Force Base. The components investigated in the soil samples are fixed as non-biological concentrations. The goal to be accomplished is to be able to discern between naturally-contaminated soil samples and unnaturally-contaminated soil samples. If the samples can be consistently classified, a geographical mapping of the unnatural (i.e. man-made) contamination locations may help to determine the source(s) of contamination.

2. Data Set

The samples taken from McConnell Air Force Base were tasted in a laboratory for concentrations of eighteen metals: Aluminum, Arsenic, Barium, Beryllium, Calcium, Total Chromium, Cobalt, Copper, Iron, Lead, Magnesium, Manganese, Mercury, Nickel, Potassium, Sodium, Vanadium, and Zinc. The 437 samples were also mapped to geographic locations around the site, allowing for postliminary contamination assignments.

3. Features and Preprocessing

Due the wide range of various concentrations in the dataset, the log-distributions of the concentrations were analyzed to scale and compare soil components. The features were assumed to be Gaussian. The log-normal distributions are provided in Figure 1.

For subsequent analysis, the means and standard deviations of the metal components are present in Figure 2. Furthermore, the geographic locations of each soil sample on the site have been provided in Figure 3.

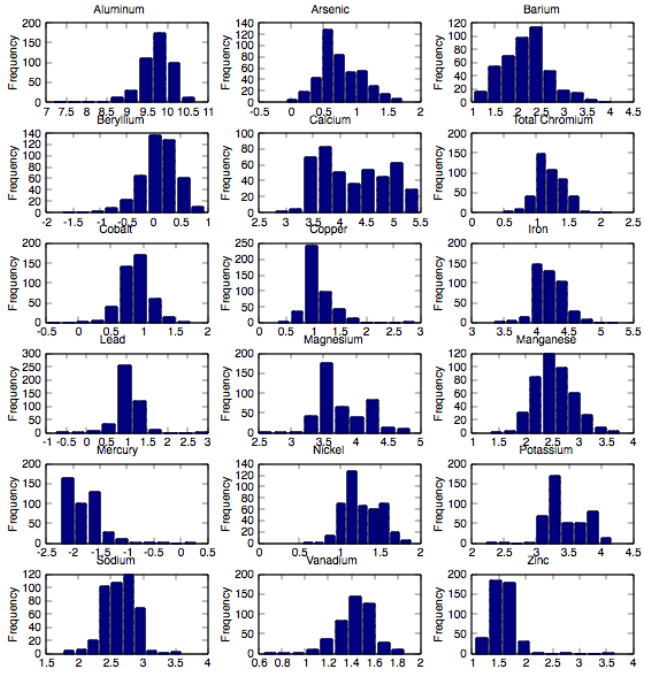

Figure 1: Feature Distribution of Log-Normal Metal Concentrations

Metal	Mean	Standard Deviation
Aluminum	9.729	0.393
Arsenic	0.765	0.325
Barium	2.166	0.496
Beryllium	0.088	0.344
Calcium	4.250	0.622
Total Chromium	1.207	0.212
Cobalt	0.887	0.236
Copper	1.089	0.253
Iron	4.201	0.219
Lead	1.014	0.342
Magnesium	3.786	0.371
Manganese	2.524	0.353
Mercury	-1.783	0.304
Nickel	1.264	0.217
Potassium	3.451	0.296
Sodium	2.638	0.245
Vanadium	1.435	0.155
Zinc	1.581	0.237

Figure 2: Log-normal Feature Distribution Information

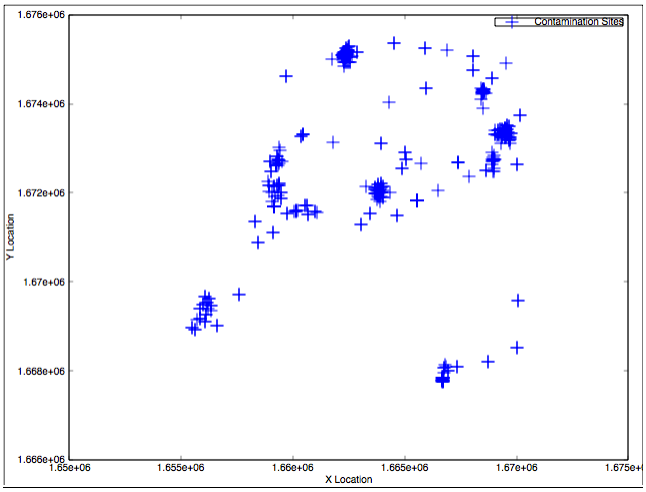

Figure 3: Geographical Mapping of Soil Samples

4. Models and Strategies

To analyze the samples, three approaches were investigated: K-means clustering, K-fold cross validation, and Gaussian discriminant analysis.

4.1 K-means Clustering

The dataset was first grouped assuming that there were two clusters and analyzed for consistency. These two clusters would represent "naturally contaminated" soil samples and "unnaturally contaminated" soil samples. The datasets were subsequently grouped with the assumption that three or more clusters were present. These cluster-centroids would also represent different types of contamination and potentially different sources of contamination.

The features were investigated to determine a hierarchy of significance. The importance of a metal with respect to its impact on classification was evaluated by comparing the standard deviation of the lognormal concentration spread to the "distance" between the centroids at that feature. The closer the concentrations at the centroids, the less "important" it was deemed to be.

$$Centroid\ Closeness_X = \frac{|C_{X,1} - C_{X,2}|}{\sigma_X}$$

$$C_{X,1} = Log\ Concentration\ of\ Feature\ X\ at\ Centroid\ 1$$

$$C_{X,2} = Log\ Concentration\ of\ Feature\ X\ at\ Centroid\ 2$$

$$\sigma_X = Standard\ Deviation\ of\ Feature\ X$$

By determining the order of importance in the features, it would be possible to establish key features that influence the contamination classification of a given soil sample. From a different perspective, it opens the possibility that certain features are not necessary. Removing non-important features greatly simplifies computation for future analyses.

4.2 K-Fold Cross Validation

Samples were randomly grouped into 19 subsets of 23 samples. This method allows for contamination predictions on the given dataset without requiring additional samples from the field. The randomness also provides variation in the training sets for the subsequent Gaussian discriminant analysis.

An extreme version of K-fold cross validation was also explored: the jackknife resampling method. Using this method, an algorithm is trained on the entire dataset, save for a single sample. The algorithm is subsequently tested on the sample to determine effectiveness of training.

4.3 Gaussian Discriminant Analysis

Using the assignments from K-means clustering, linear discriminant analysis was used to generate a predictor for potential future samples to be taken from the site. The performance for linear discriminant analysis (LDA) was compared to the performance for quadratic discriminant analysis (QDA).

For a given cross validation set, predictions were made both using the assumption that the variances between "naturally" and "unnaturally" contaminated samples are equal. The accuracy of the resulting indication is then compared to the accuracy of the prediction made without assuming equal variance.

5. Results and Analysis

5.1 K-means Clustering

Usage of K-means clustering to identify two distinct clusters always produced the same number of "naturally" and "unnaturally" contaminated samples (i.e. 267 "naturally" contaminated samples and 170 "unnaturally" contaminated samples). Furthermore, the centroids were consistently generated. This regularity in results strongly suggests the existence of two classes of contamination samples.

However, when using K-means clustering to identify three or four distinct clusters, the resulting classifications were not unique. Specifically, K-means clustering was not able to converge to the same set of clusters given random centroid initialization. It is unlikely that more than two types of contamination exist within the given dataset; there is a strong implication that there is a single source of contamination on the site. Given the lack of uniformity in classification assignments, gaussian discriminant analysis was not reasonable and was not further explored for more than two classifications.

Cluster #	Natural	Unnatural
Cluster Population	267	170
Aluminum	9.572	9.976
Arsenic	0.582	1.052
Barium	2.179	2.147
Beryllium	-0.005	0.234
Calcium	3.943	4.731
Total Chromium	1.083	1.402
Cobalt	0.801	1.022
Copper	0.960	1.291
Iron	4.071	4.406
Lead	1.040	0.973
Magnesium	3.534	4.182
Manganese	2.371	2.764
Mercury	-1.798	-1.759
Nickel	1.129	1.476
Potassium	3.252	3.765
Sodium	2.532	2.806
Vanadium	1.391	1.505
Zinc	1.501	1.707

Figure 4: Two-Cluster Centroids

In determining a hierarchy of significant features, centroid closeness with respect to feature standard deviation was examined. The following order of feature importance was generated:

	Feature Number	Feature	Centroid Closeness
Least Significant	11	Magnesium	0.064
	15	Potassium	0.127
	14	Nickel	0.198
	9	Iron	0.697
	6	Total Chromium	0.738
	2	Arsenic	0.870
	8	Copper	0.936
	5	Calcium	1.029
	16	Sodium	1.113
	12	Manganese	1.120
	1	Aluminum	1.267
	7	Cobalt	1.305
	18	Zinc	1.445
	17	Vanadium	1.507
	4	Beryllium	1.535
	10	Lead	1.597
Most Significant	13	Mercury	1.737
	3	Barium	1.747

Figure 5: Hierarchy of Significance

From this data, it is apparent that the unnaturally contaminated soil samples have significantly higher concentrations of metals such as Barium, Mercury, and Lead.

To explore the impact of the features on the classifications, K-means clustering was performed on the samples with reduced features. Features were removed in order of increasing importance until the "most significant" feature was left to classify the dataset. In each case, the same centroids were repeatedly generated.

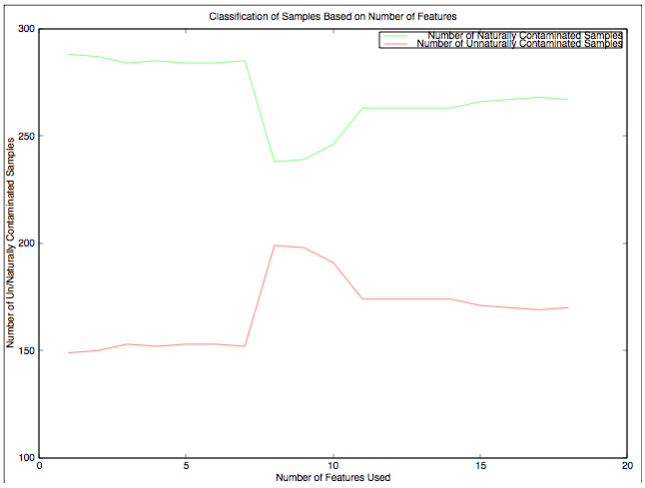

Figure 6: Number of Features vs. Cluster Population

The trend seems to imply that the addition of certain features do not significantly aid in the classification process of the dataset (i.e. where the lines are relatively flat). However, running K-means on the set without those features does not provide an accurate representation of the classifications.

5.2 K-Fold Cross Validation and Gaussian Discriminant Analysis

The accuracy of the Gaussian discriminant analyses across the two types of cross validations are depicted below (Note: 'U' represents unnatural contamination and 'N' represents natural contamination:

		LDA		QDA	
		U	N	U	N
K-Means vs. K-Fold (k=19)	U	0.327	0.062	0.368	0.021
	N	0.002	0.609	0.009	0.602
K-Means vs. Jackknife	U	0.341	0.048	0.362	0.027
	N	0.002	0.609	0.009	0.602

Figure 7: Accuracy of Gaussian Discriminant Analysis Across K-Fold Cross Validations

It is interesting to note that, for this particular dataset, the jackknife resampling method boosted the accuracy of linear discriminant analysis predictions and decreased the accuracy of quadratic discriminant analysis predictions. Furthermore, usage of the jackknife resampling method did not appear to influence positive or negative predictions of naturally contaminated soil samples; it only affected predictions of unnaturally contaminated soil samples.

Although the jackknife resampling method provides small improvements to the linear and quadratic discriminant analyses, generating the predictions also took considerably longer. For significantly larger datasets, the tradeoff may be less favorable. Likewise, for smaller datasets, the jackknife resampling method could be a worthwhile investment.

Regardless of the sampling for training and testing, quadratic discriminant analysis consistently performed better than linear discriminant analysis. This outcome is expected because the quadratic discriminant analysis operates on fewer assumptions (i.e. shared variance) than linear discriminant analysis.

5.3 Geographical Plot of Contamination

After establishing the classification of the data set, the contamination sites were mapped. K-means was executed on both the original concentrations (Figure 9) and the log-normal (Figure 8) concentrations.

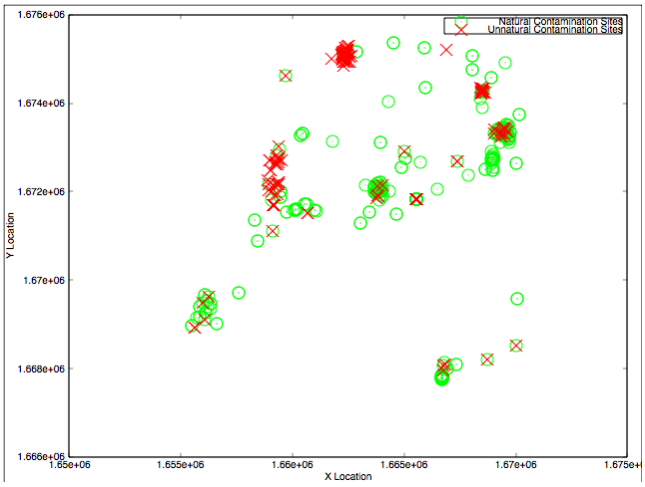

Figure 8: Geographical Mapping of Contamination Sites (Log-Normal)

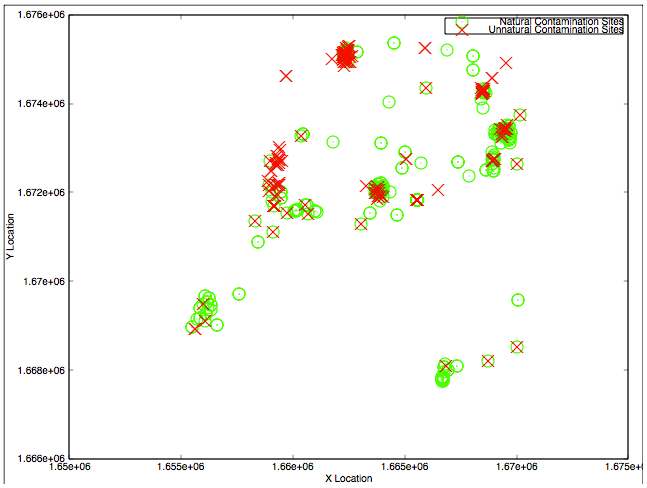

Figure 9: Geographical Mapping of Contamination Sites (Normal)

As seen in Figures 8 and 9, there is no clear geographical basis for the clustering of natural or unnatural contamination. Therefore, the source of the contamination cannot be determined from the provided data.

6. Conclusion

From the analyses of the data with respect to classification, it can be stated with high confidence that the McConnell Air Force Base has apparent soil contamination. This contamination is characterized by especially high concentrations of Barium, Mercury, and Lead. The source of the contamination is yet unknown, given the provided data, but may be related to depth of soil sample or specific site operations.

7. Future Endeavors

Classification of soil sample contamination is one that is constantly undergoing change. Most available data uses hierarchical classification to determine clusters of samples, along with principal component analysis.

In future work, I would like to investigate the accuracy of hierarchical classifications, using either principal component analysis (PCA) or independent component analysis (ICA). These reductions primarily differ in the assumption of Gaussian features, or lack thereof. I would like to compare the classifications from those strategies and models to those achieved from k-means, as done in this report.

8. References

1. Hubert, Lawrence, Hans-Friedrich Köhn, and Douglas Steinley. "Cluster analysis: a toolbox for MATLAB." Handbook of quantitative methods in psychology (2009): 444-512.
2. Fraley, Chris, and Adrian E. Raftery. "Model-based clustering, discriminant analysis, and density estimation." Journal of the American Statistical Association 97.458 (2002): 611-631.
3. Yongming, Han, et al. "Multivariate analysis of heavy metal contamination in urban dusts of Xi'an, Central China." Science of the Total Environment 355.1 (2006): 176-186.
4. Astel AM, Chepanova L, Simeonov V. Soil Contamination Interpretation by the Use of Monitoring Data Analysis. Water, Air, and Soil Pollution 2011;216(1-4):375-390. doi:10.1007/s11270-010-0539-1.