Streamlining Principal Component and Cluster Analysis to Support Conceptual Site Models at Complex Remediation Sites

Tori Ward

May 14, 2025

Introduction

Woodard & Curran

Menti Question (www.menti.com):

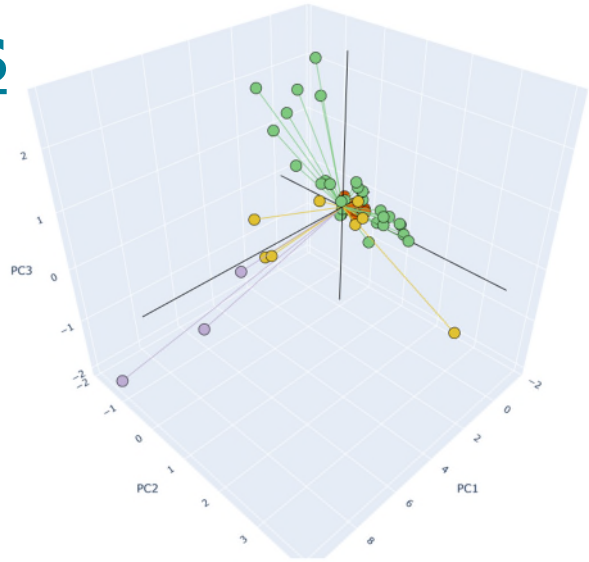
How familiar are you with PCA and Cluster analysis?

6927 2321

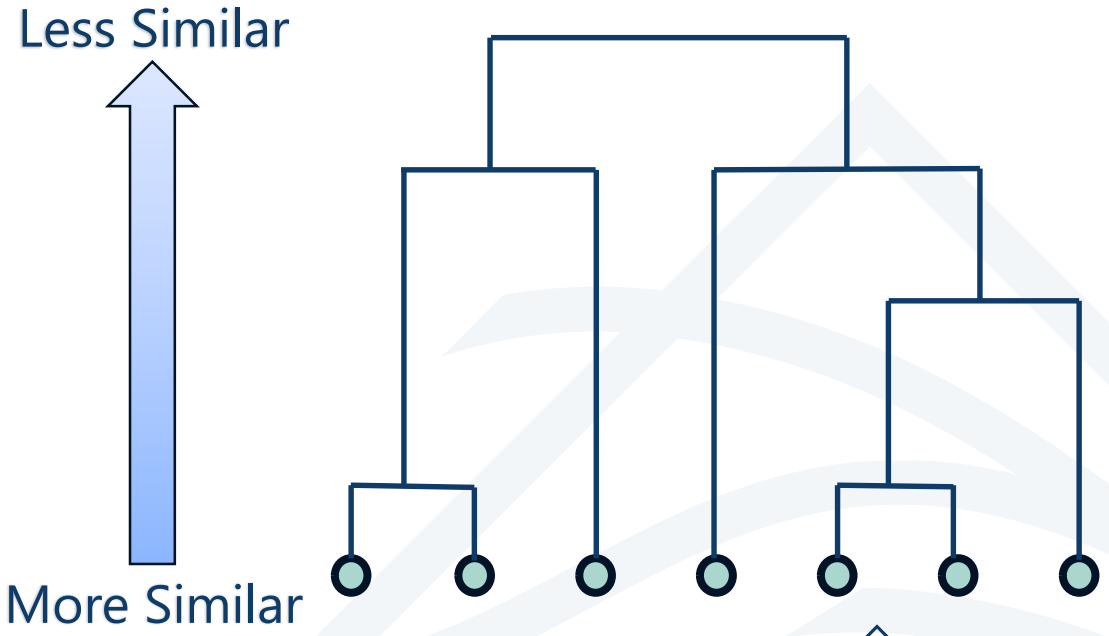
Mentimeter

Intro to PCA and HCA

Unsupervised Machine Learning Methods



- **Principal Component Analysis (PCA)**
  - Reduces many variables to a few components that explain most of the variability in the data
  - **Components** are linear combinations of variables, ordered by decreasing importance



- **Hierarchical Cluster Analysis (HCA)***
  - Starts by placing each observation in its own **cluster** (Bottom-up approach)
  - Applies a distance measure to merge clusters that are most similar

- ***HCA is just one example of a cluster analysis***

Woodard & Curran

Why use PCA and Cluster Analysis?

Complex Geology

Multiple Suspected Sources
not a real Site

Woodard & Curran

Menti Question (www.menti.com):
What are other ways you are aware of PCA and cluster analysis being used in our industry?

6927 2321

Mentimeter

Methodology and Workflow

Woodard & Curran

python™ Tools for PCA and HCA

plotly | Graphing Libraries

scikit learn

SciPy

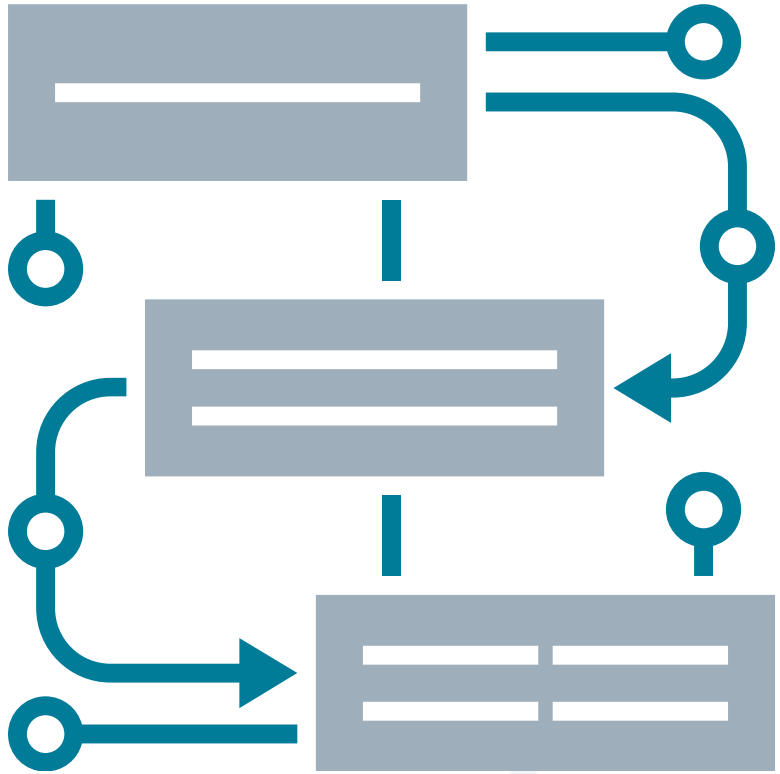
NumPy

pandas

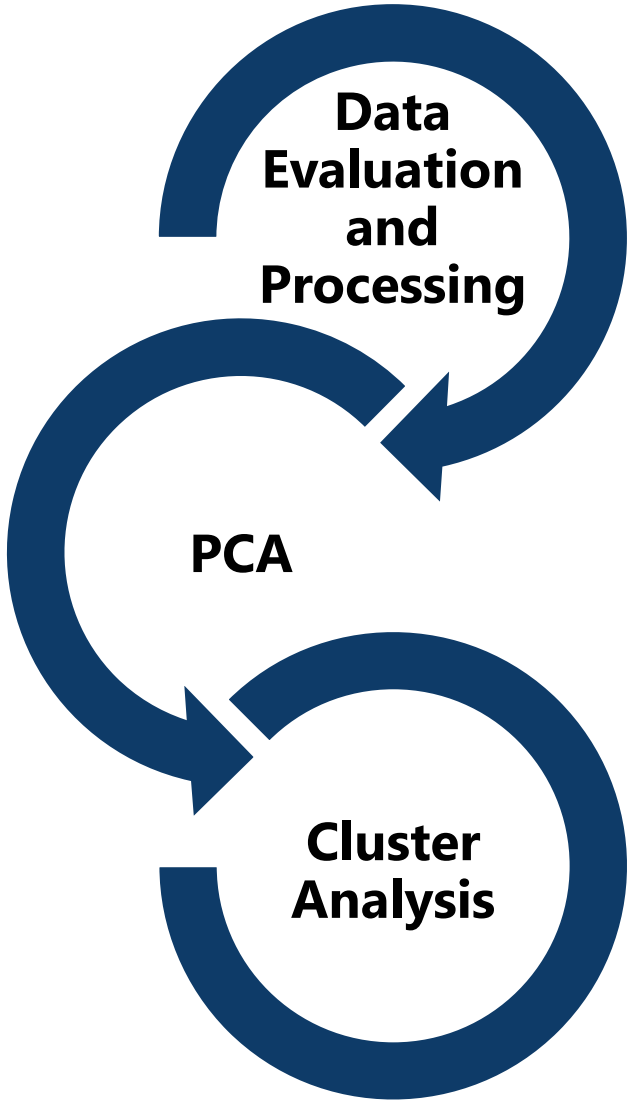
Woodard & Curran

Standardized Methodology

- Basic statistics
- Correlation matrix
- PCA
  - Data transformations
  - Review components, explained variance, and loadings
- Visualize PCA
  - 3D charts of PCA factor loadings
- Clustering
  - Selecting an appropriate number of clusters
    - Dendrograms
    - Elbow and Silhouette Scores
    - Professional Judgement when reviewing output
  - Applied a color-scale based on cluster labels to 2D and 3D PCA charts to assess corroboration between PCA and agglomerative clustering results



Streamlined Workflow



Project-specific customizations

Generalized functions across all 3 scripts are stored in a utilities file

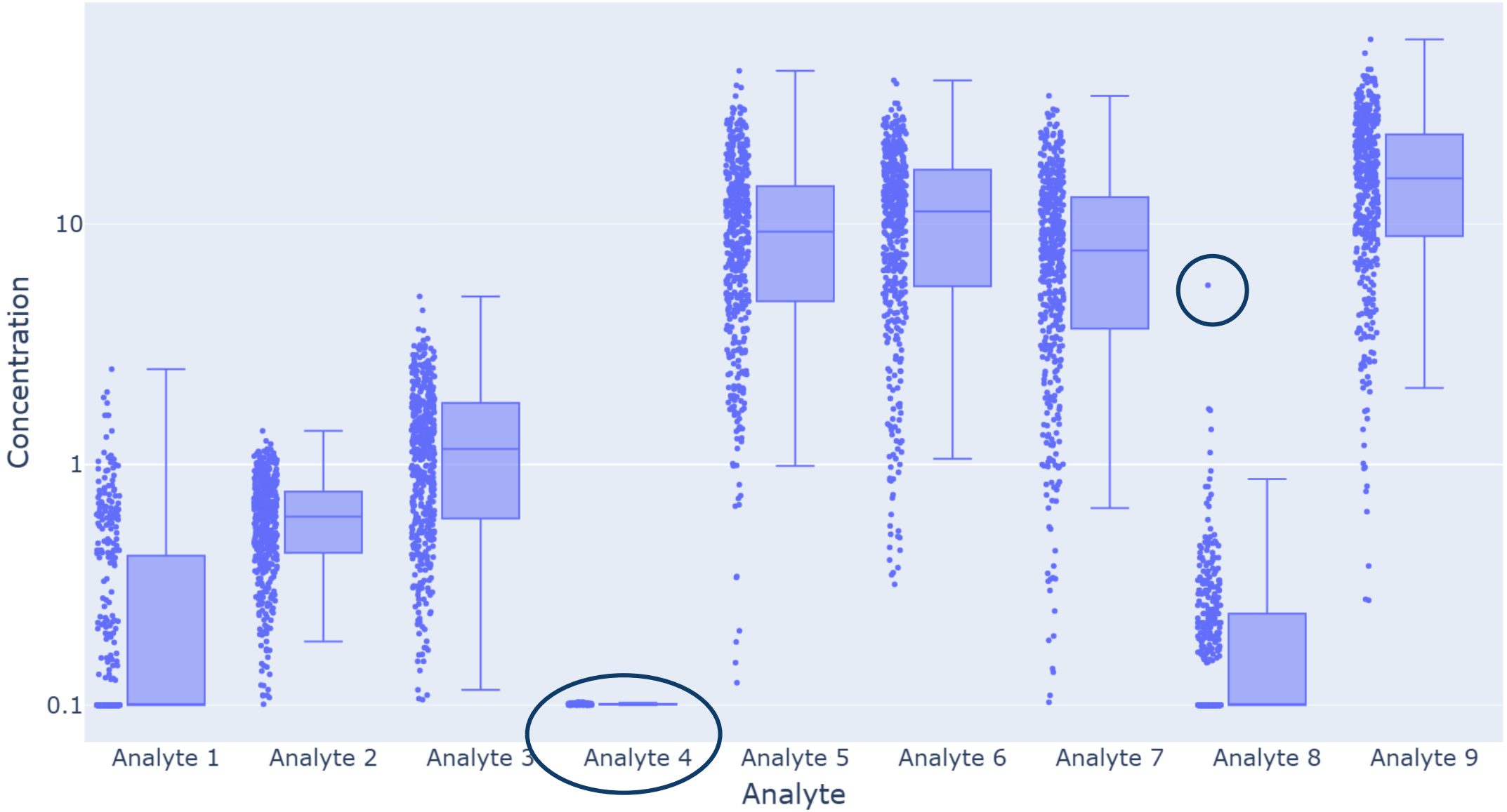
Code maintained on GitHub

Example Output Visualizations

Graphics and tables presented do not summarize project data to preserve client confidentiality

Woodard & Curran

Methodology – Basic Stats



Woodard & Curran

Methodology – Basic Stats

Analyte	Number of Detections	Number of Analyses	Frequency of Detection
Analyte 6	404	442	91.40%
Analyte 7	350	442	79.19%
Analyte 5	311	442	70.36%
Analyte 3	281	436	64.45%
Analyte 2	236	442	53.39%
Analyte 9	211	415	50.84%
Analyte 8	185	410	45.12%
Analyte 4	163	408	39.95%
Analyte 1	155	415	37.35%

Menti Question (www.menti.com):
Which methods have you used to impute non-detect results (multiple selections possible)?

6927 2321

Mentimeter

Methodology – Correlation Matrix

Analyte	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5	Analyte 6
Analyte 1	1.00					
Analyte 2	0.33	1.00				
Analyte 3	**0.70**	0.39	1.00			
Analyte 4	0.34	0.33	0.40	1.00		
Analyte 5	**0.78**	0.66	**0.73**	0.27	1.00	
Analyte 6	0.42	0.39	0.64	0.21	0.48	1.00

Key Point: High correlation between certain analytes suggests they can be aggregated into a smaller number of components

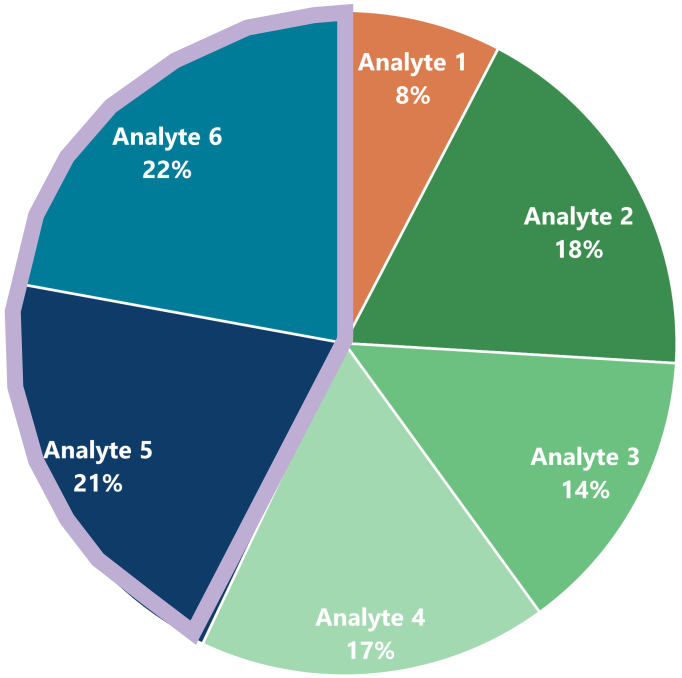
Methodology – PCA Component Weights

	Analyte 1	Analyte 2	Analyte 3	Analyte 4	Analyte 5	Analyte 6	Explained Variance
PC1	0.18	0.43	0.33	0.40	0.49	0.52	58.0%
PC2	0.95	-0.11	-0.13	0.14	-0.19	-0.10	15.6%
PC3	0.12	-0.41	0.85	-0.32	0.04	-0.04	13.3%
PC4	0.20	0.52	-0.09	-0.77	0.29	-0.12	9.4%
PC5	-0.05	0.61	0.38	0.15	-0.64	-0.24	2.9%
PC6	0.02	-0.05	-0.09	-0.33	-0.48	0.80	0.9%

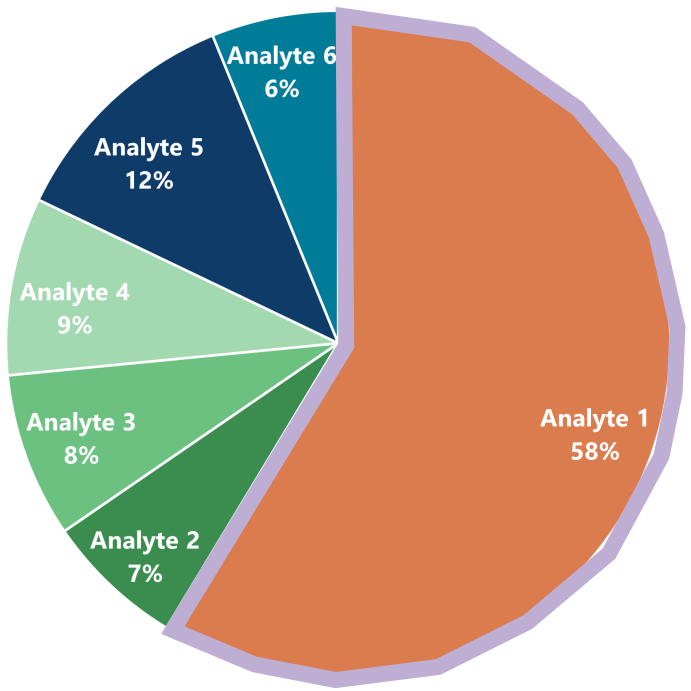
Key Point: Analytes 1, 3, 5, and 6 explain most of the variability in the original data set.

Methodology – PCA Component Weights Visualization

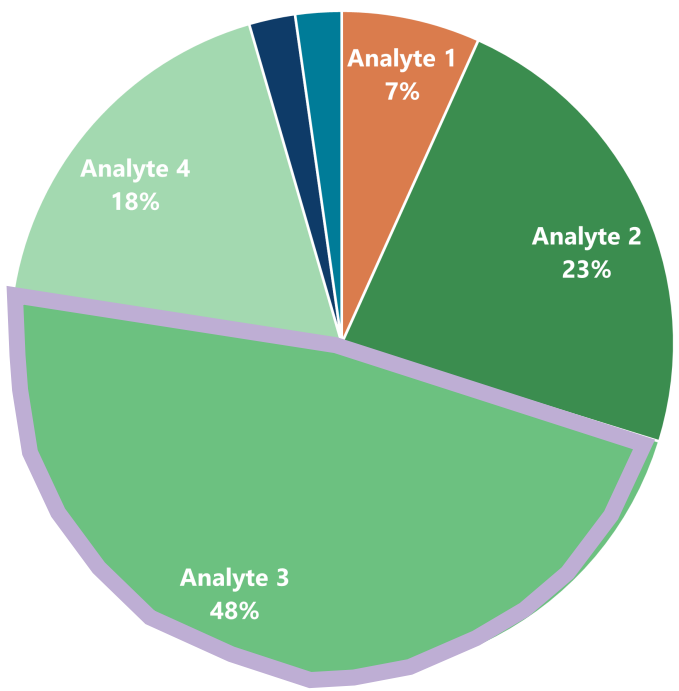
Principal Component 1



Principal Component 2



Principal Component 3

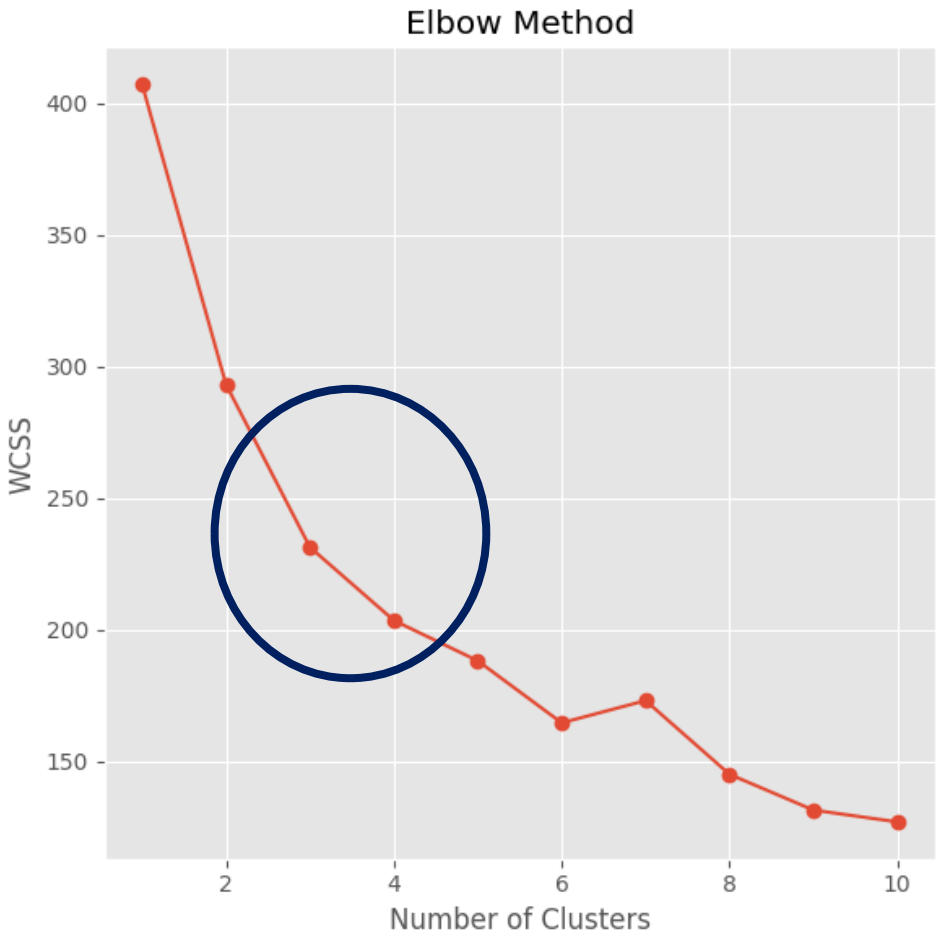


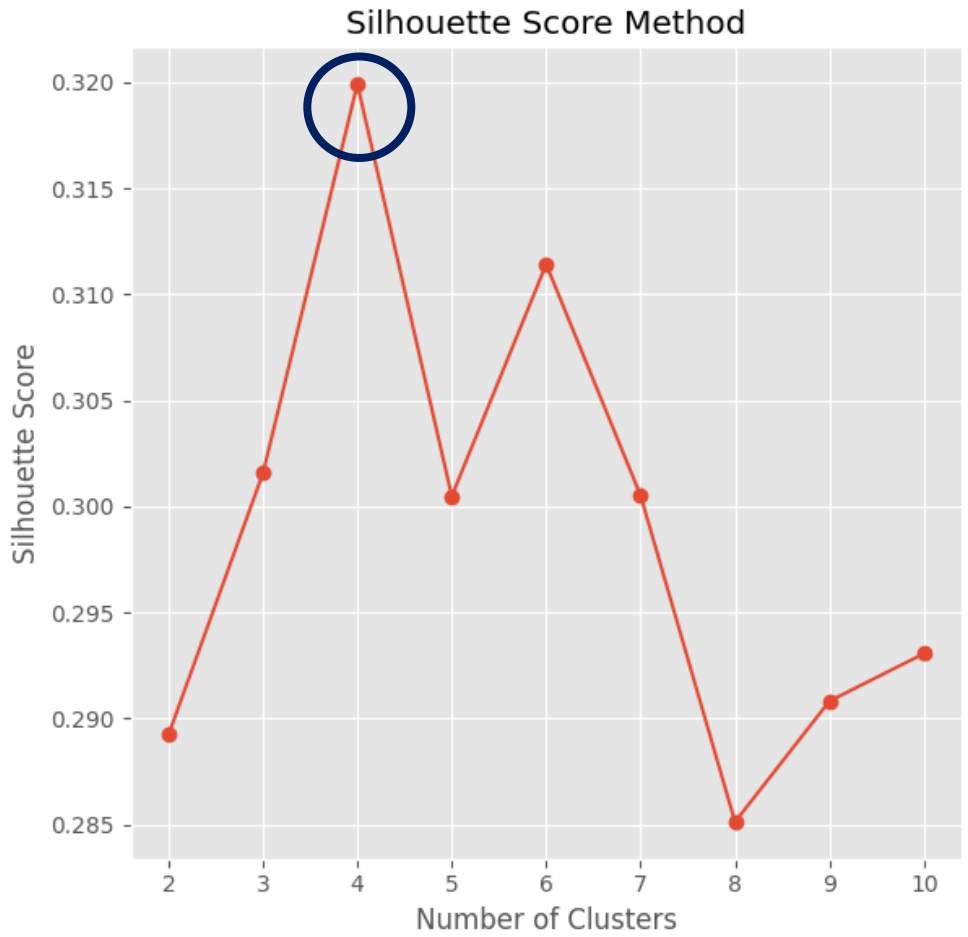
Key Points:

- Principal Component 1 is strongly tied to Analytes 5 and 6
- Principal Component 2 is strongly related to Analyte 1
- Principal Component 3 is strongly related to Analyte 3

Woodard & Curran

Results – Cluster Analysis Elbow and Silhouette Charts

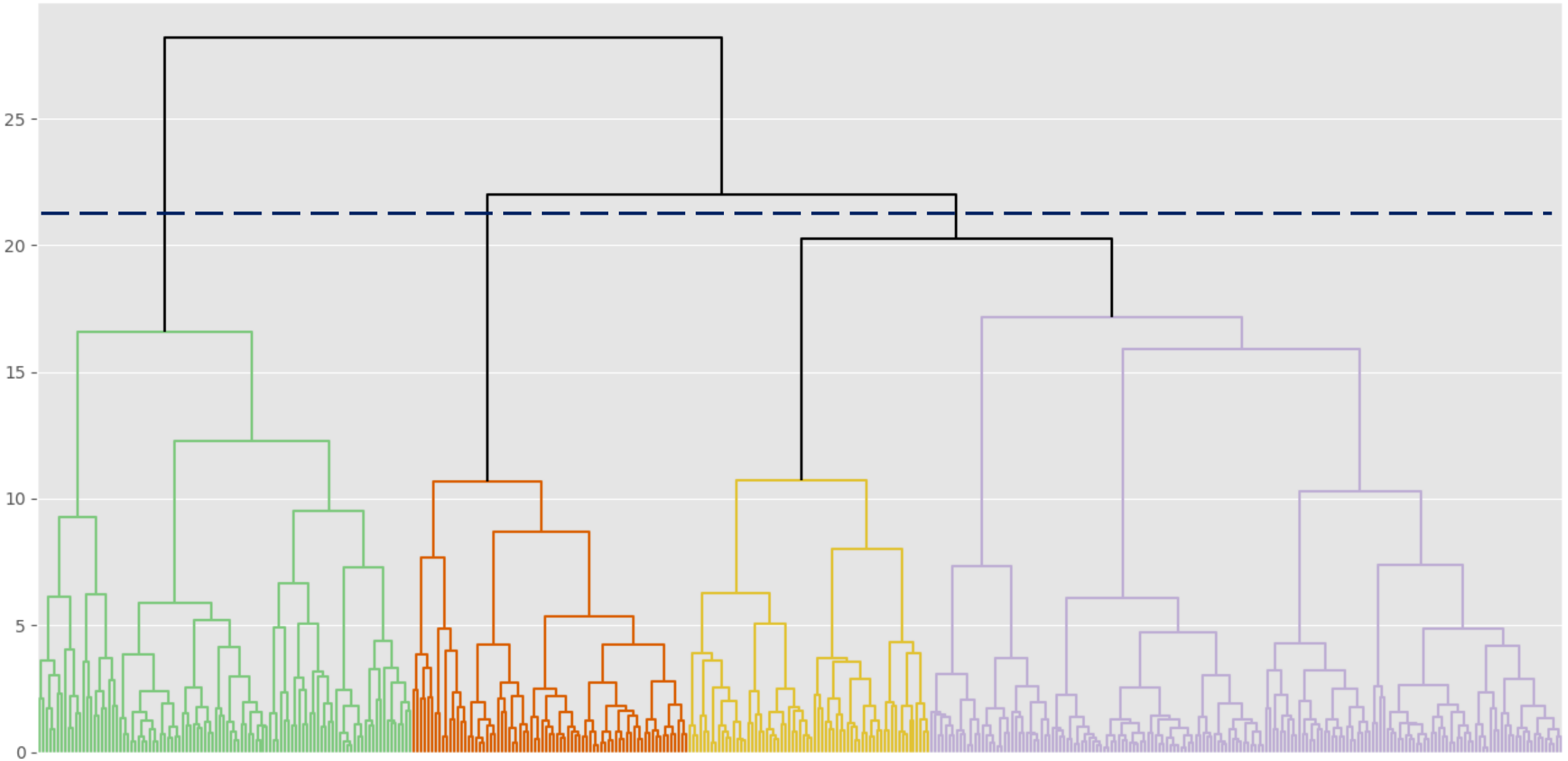




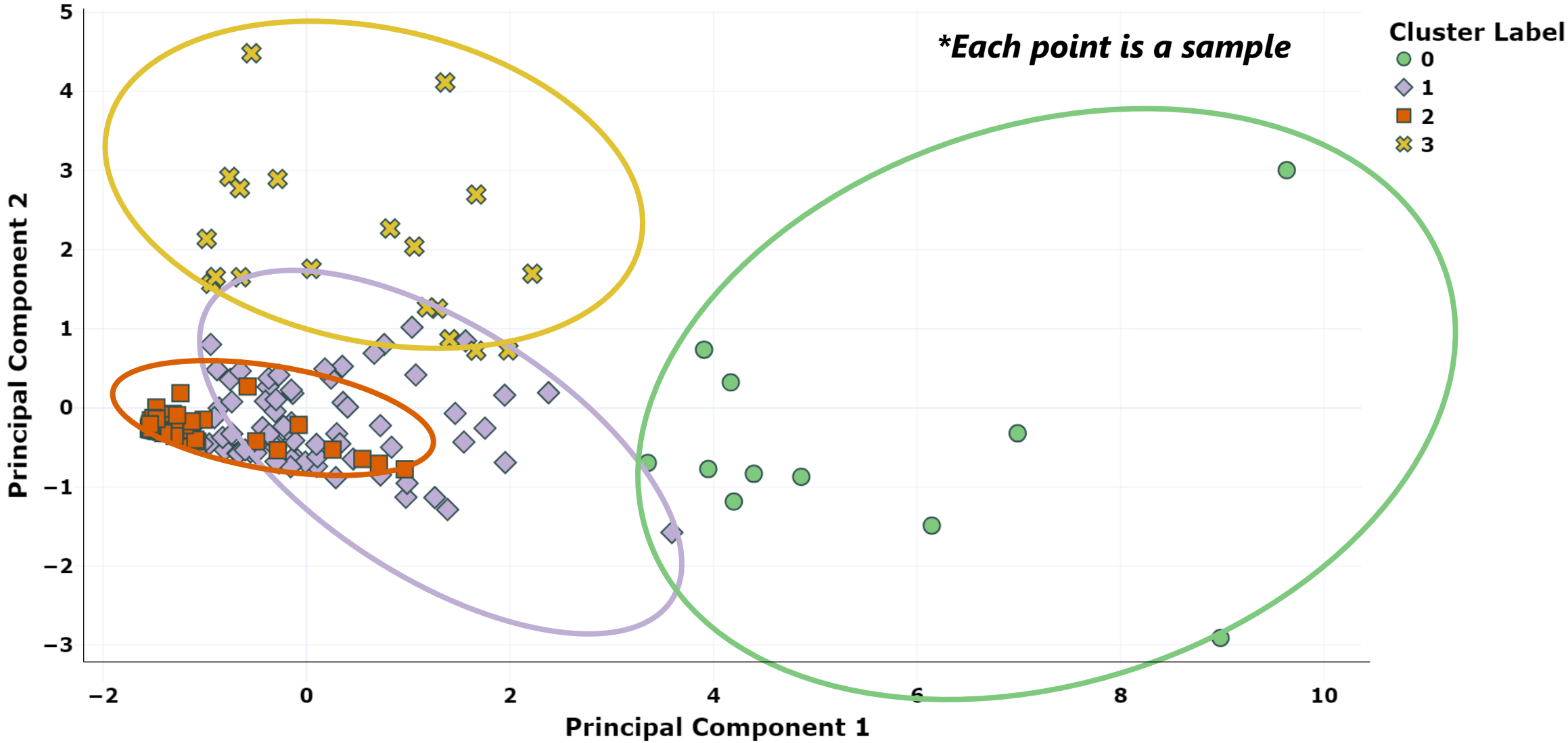
Key Point: 3 to 4 clusters is reasonable

Woodard & Curran

Results - HCA Dendrogram



Results – PCA Scatter with Cluster Labels

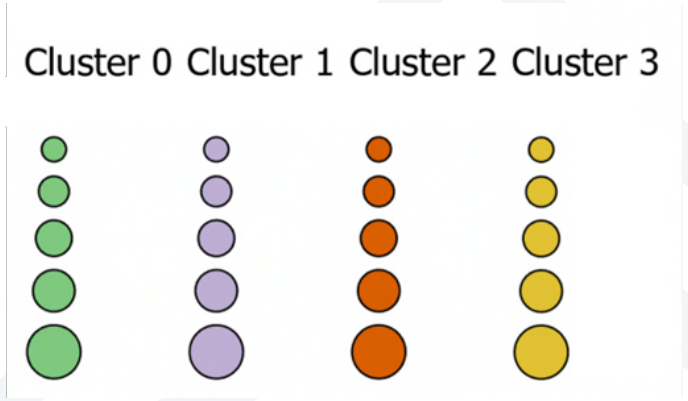


Woodard & Curran

Results – Geospatial Evaluation of Cluster Labels

Key Points:

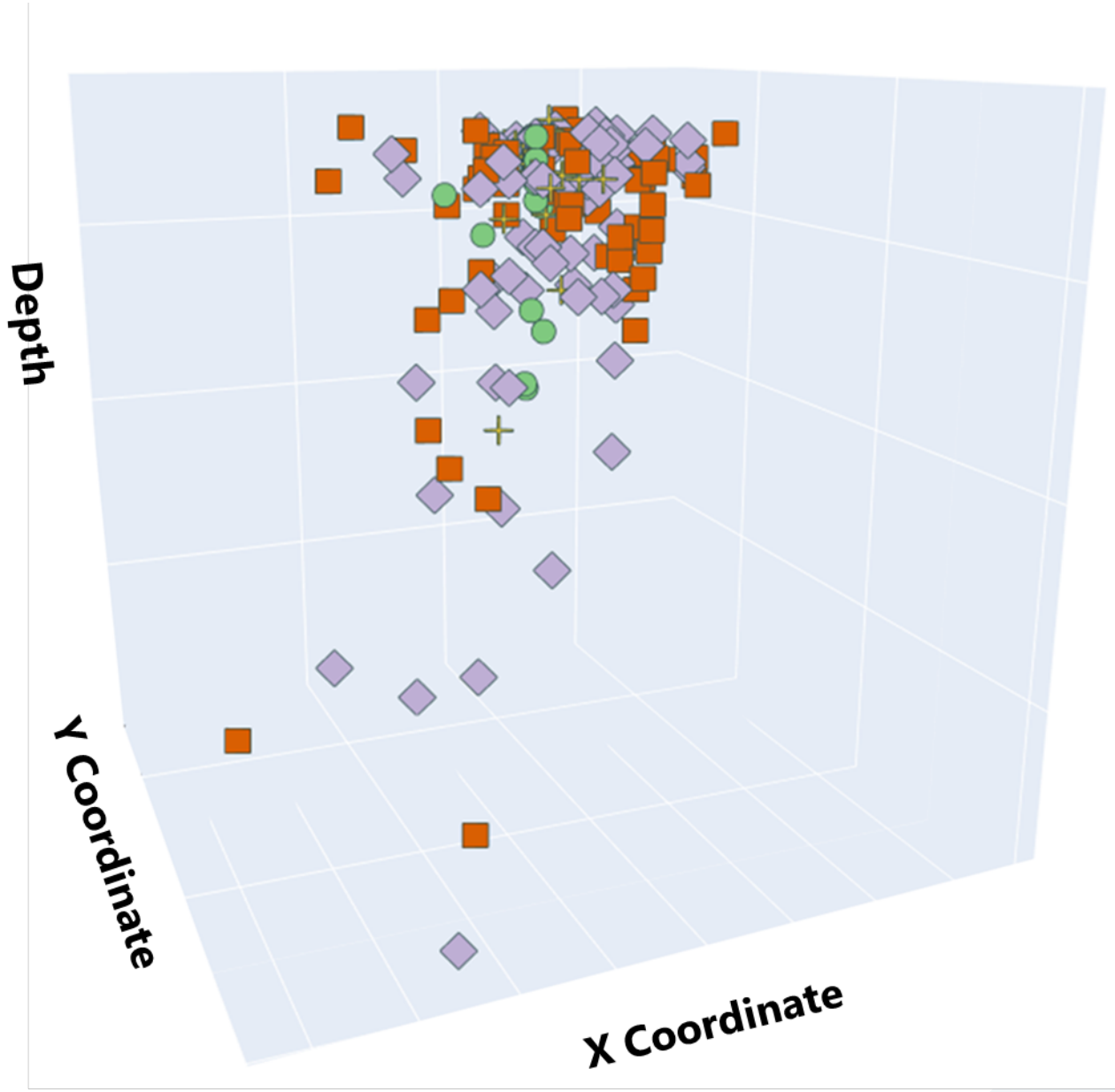
- Locations associated with clusters 0, 1, and 2 are generally located near each of the source areas
- Contaminant concentrations are generally highest in Cluster 1 and lowest in Cluster 0



**** Not a real Site***

Woodard & Curran

X,Y,Z-Coordinates with Cluster Labels



Each point is a sample

Woodard & Curran

Relevance

- Focus on a few key variables in a large data set
- Objective, quantitative line of evidence
- Can use results to target site remediation and planning

Woodard & Curran

Tori Ward

vward@woodardcurran.com

Questions?