A Machine Learning Approach for Detecting Groundwater Runoff Connectivity

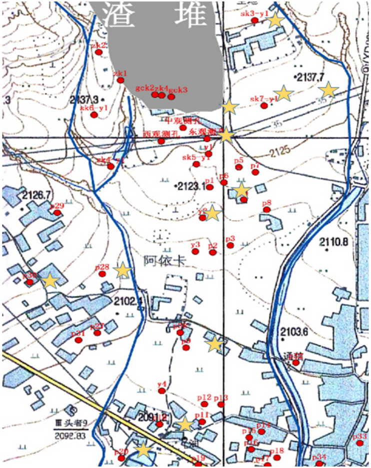
Xiaojun Kang, Xuguang Zhao, Caixia Guo, Junhua Ding

University of North Texas

September 25, 2018

Research Background

Water samples were collected from 12 underground wells in Zhehai mining area, Yunan Province, China, and three samples were collected from each well in different time. Bacteria DNA in the water samples were sequenced.



sampling point	sample numbers
sk3	a
sk7	b
sk12	c
east observation point	d
leachate	e
y2	f
p6	g
p9	h
p10	i
p20	j
p28	k
p30	l

Research Purposes

- Experiment an affordable approach for detecting groundwater runoff connectivity,

Research Purposes

- Experiment an affordable approach for detecting groundwater runoff connectivity,
- Investigate the performance improvement of the cluster analysis.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.
- **Features** are the set of bacterial types, and feature values are the quantity of each type of bacterial in a sample.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.
- **Features** are the set of bacterial types, and feature values are the quantity of each type of bacterial in a sample.
- **Cluster Analysis** of the feature matrix using different clustering algorithms.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.
- **Features** are the set of bacterial types, and feature values are the quantity of each type of bacterial in a sample.
- **Cluster Analysis** of the feature matrix using different clustering algorithms.
- **Cross check** of cluster analysis results.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.
- **Features** are the set of bacterial types, and feature values are the quantity of each type of bacterial in a sample.
- **Cluster Analysis** of the feature matrix using different clustering algorithms.
- **Cross check** of cluster analysis results.
- **Ordination** of the selected clustering results using PCA(principal component analysis) and NMDS (non-metric multidimensional scaling) to improve analysis results.

Approach

- **Analyze** bacterial types and **Calculate** their quantities based on DNA sequences in each sample.
- Each isolated environment should contain unique bacterial community, which is defined by the combination of bacterial types and their quantities.
- **Features** are the set of bacterial types, and feature values are the quantity of each type of bacterial in a sample.
- **Cluster Analysis** of the feature matrix using different clustering algorithms.
- **Cross check** of cluster analysis results.
- **Ordination** of the selected clustering results using PCA(principal component analysis) and NMDS (non-metric multidimensional scaling) to improve analysis results.
- **Refine** the clustering results referring to environmental parameters.

Cluster Analysis

- Hierarchical clustering with four different distances: min, max, average, Ward Minimum Variance

Cluster Analysis

- Hierarchical clustering with four different distances: min, max, average, Ward Minimum Variance
- Partition clustering using K-mean and K-medoids (K-medoids is also called PAM - partitioning around medoids)

Hierachical Clustering Results

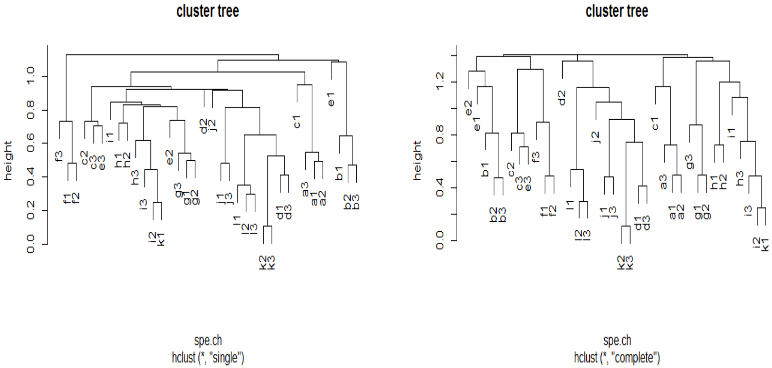


Figure 2: The cluster trees of min cluster (left) and max cluster (right)

Hierachical Clustering Results

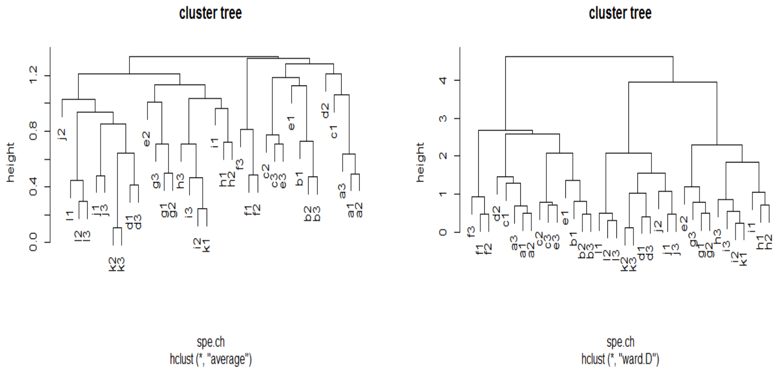


Figure 3: The cluster trees of mean cluster (left) and ward minimal variance cluster (right)

Evaluation of the Results

Calculate Pearson correlation coefficient, represented in Shepard. The larger, the better.

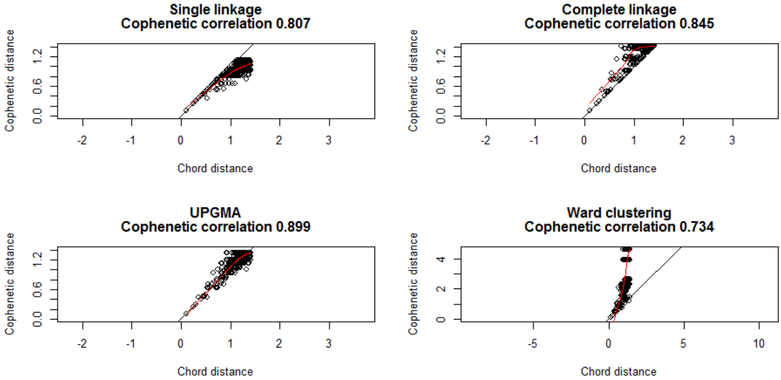


Figure 4: Shepard diagrams for each hierarchical analysis

Clustering Results

Calculate level of integration values to cut the tree, each subtree is a cluster.

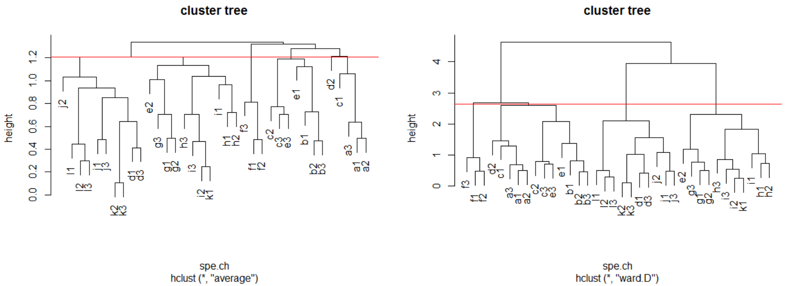


Figure 5: Clustering result, average clustering (left), and ward clustering (right)

Clustering Results

Each row represent one hierarchical analysis (min, max, average, and ward), and each column represents a cluster.

f	b	a	c	k、l、d、j	g、h、i
f、c	b、e	k、l、d、j	g、h、i、a	N/A	N/A
f	a、b、c、e	k、l、d、j	g、h、i	N/A	N/A
f	a、b、c、e	k、l、d、j	g、h、i	N/A	N/A

Figure 6: Clustering result

Partition Clustering Results

K-mean and PAM (or K-medodis). Iteratively experiment different k from 2 to 10 and evaluate the optimal K with SSI values.
PAM: {{a}, {b}, {k, l, j}, {g, h, i, f}}
K-mean (K = 4): {{a}, {b, c, d, e, f}, {k, l, j}, {g, h, i}}
K-mean (K = 6): {{a}, {b, c, e}, {h,i}, {g}, {f}, {k, l, j, d}}

Refine Results with Environmental Parameters

Understand how the environment such as heavy metal ion impact to the clustering. If two wells are connected, their heavy metal ion should impact the both of them.

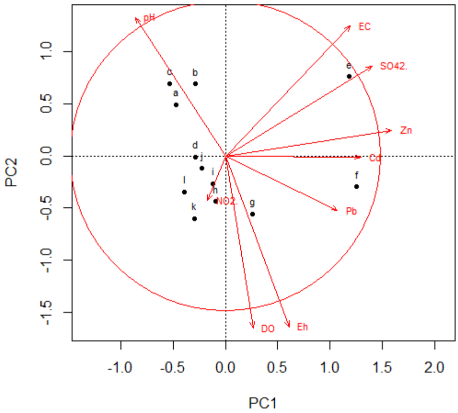


Figure 7: PCA on environmental parameters to clustering

Ordination with NMDS

Final clustering result: {{a, b, c, d, e}, {f}, {k, l, j}, {g, h, i}}

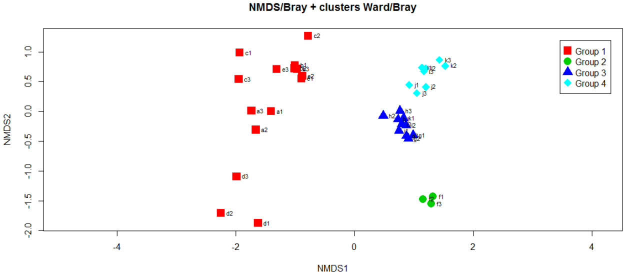


Figure 8: NMDS + Ward cluster.

Validation

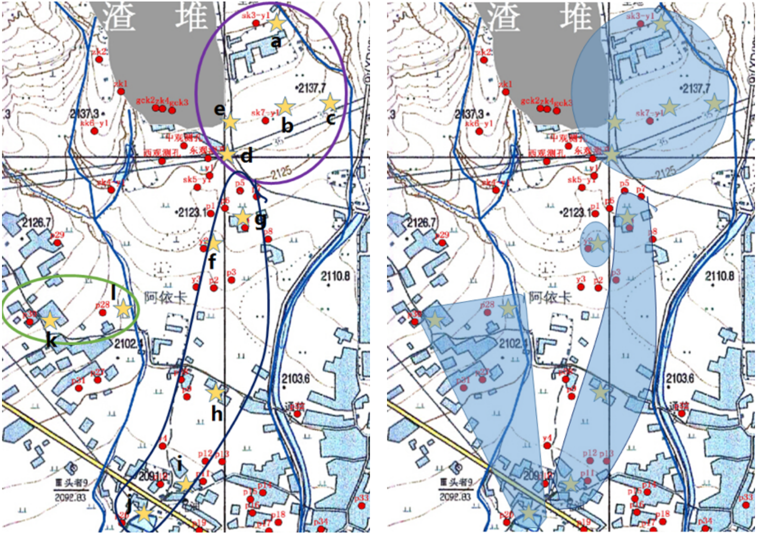


Figure 9: The distribution of groundwater derived from (a) physical and chemical experiments, and (b) ordination + clustering method.

Conclusion

- Preliminary results show clustering of DNA sequences has the great potential for detecting the groundwater runoff connectivity.

Conclusion

- Preliminary results show clustering of DNA sequences has the great potential for detecting the groundwater runoff connectivity.
- More samples, more scenarios should be investigated.