Title
Application of Clustering Techniques to Study Environmental Characteristics of Microbialite-Bearing Aquatic Systems

Permalink
https://escholarship.org/uc/item/48q8632d

Author
Dalinina, Ruslana

Publication Date
2013
Application of Clustering Techniques to Study Environmental Characteristics of Microbialite-Bearing Aquatic Systems

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Statistics

by

Ruslana Dalinina

2013
Abstract of the Thesis

Application of Clustering Techniques to Study Environmental Characteristics of Microbialite-Bearing Aquatic Systems

by

Ruslana Dalinina
Master of Science in Statistics
University of California, Los Angeles, 2013
Professor Rick Schoenberg, Chair

Microbialites are a product of trapping and binding of sediment by microbial communities, and are considered to be some of the most ancient records of life on Earth. It is widely held that microbialites are limited to extreme, hypersaline environments. However, literature also shows evidence of their occurrence in a wider range of environments. The goal of this thesis is to explore geochemical properties of aquatic environments in which microbialites have been found. We apply statistical techniques to distinguish any common traits in microbialite environments. These techniques ultimately could be used to address the question are microbialites restricted to hypersaline environments, other environments with specific characteristics, or are they more broadly distributed? A dataset containing hydrographic characteristics of several microbialite sites with data on pH, conductivity, alkalinity, and concentrations of several major anions and cations was constructed from previously published studies. Qualitative inspection of these data show that microbialites are not restricted to hypersaline environments, as they are present at Pavilion Lake, a freshwater body. In order to group the water samples by their natural similarities and differences, a clustering approach was chosen for analysis. K-means clustering with partial distances was applied to the dataset with missing values, and separated the data into 2 distinct but geochemically similar clusters. One of the clusters is formed by samples from atoll Kiritimati, and the second
cluster contains all other observations. Then, the missing values were imputed by $k$-nearest neighbor method, producing a complete dataset that can be used for further multivariate analysis. We find that microbialites occur within environments spanning a range of salinities (as indicated by conductivity), pH values, and ionic compositions. Furthermore, pH seems to define geochemical profile of this dataset and its clustering. Clustering and imputation procedure outlined here can be applied to an expanded dataset on microbialite characteristics in order to determine if unique properties are associated with microbialite-containing environments, and can also serve as an outline for analysis of such datasets with missing data.
The thesis of Ruslana Dalinina is approved.

Aradhna Tripati
Amy Braverman
Janet Sinsheimer
Hongquan Xu
Rick Schoenberg, Committee Chair

University of California, Los Angeles
2013
To my friends and my parents, for their encouragement and support
Table of Contents

1 Introduction .................................................................................. 1

2 Analysis .................................................................................... 4
   2.1 Exploratory Analysis ............................................................... 4
   2.2 Clustering ........................................................................... 6
      2.2.1 Finding the Number of Clusters ....................................... 7
      2.2.2 Hypothesis Test on Number of Clusters ......................... 9
      2.2.3 K-means Clustering ....................................................... 11
      2.2.4 Clustering a subset of data ........................................... 12
   2.3 Missing At Random and Missing Completely at Random .......... 14
   2.4 Imputation and Validation ................................................... 20
   2.5 Environmental characteristics limiting microbialite distribution .. 23

3 Conclusions and Future Work ..................................................... 26

A Frequency Tables for Missingness Patterns ............................... 28

References ...................................................................................... 30
List of Figures

2.1 Parallel Coordinates Plot ........................................... 5
2.2 Hierarchical clustering of all observations ........................ 6
2.3 Silhouette Plot for Finding the Number of Clusters .................. 8
2.4 Gap Statistic .......................................................... 9
2.5 CH Index ............................................................... 10
2.6 Dendrogram of 6 most complete variables .......................... 14
2.7 Mean By Missing Data Patterns ........................................ 18
2.8 Average Variability of Clustering and Imputation Procedures .... 22
LIST OF TABLES

2.1 Set of variables producing 2 original clusters ........................................ 13
2.2 Missing Data Patterns ........................................................................... 16
2.3 Ranges by Site ...................................................................................... 23
2.4 Ranges by Site, continued ..................................................................... 24
2.5 Full ranges for all variables ................................................................. 24

A.1 Counts of fully and partially observed rows for each variable. 1 = partially
observed, 0 = otherwise ........................................................................ 29
ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my thesis advisors, Amy Braverman, Aradhna Tripati, Janet Sinsheimer, Rick Schoenberg, and Hongquan Xu for their support and guidance throughout my graduate studies. I also would like to thank Gernot Arp, Jozef Kazmierczak, Dale Andersen, and Jerry Kaster for sharing their data and making this research project possible. Special thanks go to Darlene Lim and Victoria Petryshyn for their invaluable input and advice. Finally, I extend my appreciation to American Chemical Society for providing funding for this research.
CHAPTER 1

Introduction

Microbialites are organo-sedimentary deposits formed by the trapping and binding of sediment by benthic microbial communities [9]. Studying microbialites can provide insights into early life on Earth, as well as give clues for the search of life elsewhere [12]. Although they are perceived by some as being limited to harsh environments (i.e. salinity and/or high alkalinity), which may exclude certain types of metazoa, a survey of the literature shows they also frequently occur in aquatic systems that do not exhibit extreme geochemical properties [21],[12]. It is of interest to study the breadth of these environments in order to determine whether there is a characteristic set of geochemical similarities/differences common to sites where microbialites are forming.

In this work, we have created a dataset from published studies on bodies of water containing microbialites. Each study focused on a specific set of geochemical or physical characteristics of an aquatic environment, such as the concentration of certain chemical species in the water. Combining data from each publication led to the overall dataset, with variables being geochemical measurements and observations being samples from various bodies of water (hereafter referred to as sites). The resulting dataset has blocks of missing data due to the little overlap between the variables measured across different studies.

The research question posed is to distinguish any distinct characteristics of microbialite environments through comparison of sites and their features. For preliminary analysis, sites with the most complete observations were selected (Pavilion Lake region, atoll Kiritimati, and 4 German streams). Pavilion Lake is a ground-water fed lake that has not been identified as an extreme environment [21], samples from atoll Kiritimati were taken from a number of lakes
in the region ranging from hypersaline to fresh [3], and German streams are characterized as alkaline karst water [5]. The variable set was restricted to 12 most commonly measured features: concentrations of $Ca^{2+}, Mg^{2+}, SO_4^{2-}, K^+, Na^+, Si, Cl^-, Sr^{2+}, Ba^{2+}$, as well as pH, conductivity, and alkalinity.

Disregarding incomplete observations from analysis would ignore valuable information; a more preferable strategy is to impute the missing values. We therefore focus on clustering methods, which can reveal any natural structure or similarities between the sites and can also be used to impute missing values. When clusters are created, missing items can be filled in based on the characteristics of the cluster to which they belong.

Traditional clustering methods are designed for datasets with no missing values, and therefore cannot be directly applied here. For example, model-based techniques described by C.Fraley and A.E. Raftery, where each cluster is modeled by a probability distribution, assume that data does not contain missing values. K. Wagstaff [26] introduces a version of k-means clustering based on soft constraints. Given a dataset divided onto a set of completely observed variables $F_o$ and a set of partially observed variables $F_m$, as of constraints between items $d_i, d_j$ is defined as

$$s = \sqrt{-\sum_{f \in F_m} (d_{if} - d_{jf})^2}. \quad (1.1)$$

This quantity is then used in clustering over the observed variables $F_o$ with constraints based on $F_m$. The underlying idea in this algorithm is borrowing strength, or information, from the observed features to cluster only partially observed items. However, our dataset on microbialite distribution contains no completely observed features, and therefore this method cannot be applied.

A different approach was suggested by L. Himmelspach and S. Conrad [14]. Authors modify k-means algorithm to use a partial distance instead of traditional distance measure (such as Euclidean). This requires technique little modification to the k-means algorithm, and utilizes all available information from complete and partially observed items. Therefore, we will apply k-means (or rather k-medoids) clustering based on partial distance to our
dataset. This technique has not yet been used on biological/chemical systems, as far as we are aware.

Clustering was chosen because it will provide insights into answering the original question of identifying any common traits among the microbialite sites, as well as the basis for imputation. Then, missing values are imputed by their corresponding cluster’s mean. Stability of clustering classification is checked by randomly deleting some complete observations and performing clustering analysis on this new dataset.

The rest of this paper is structured as follows. The k-means clustering algorithm and a method for finding the number of clusters are applied to the dataset in sections 1.1 -1.2. Section 1.3 explores the relationship between data and missingness. Section 1.4 describes imputation procedure used and validation of the results by random deletion of some values and clustering the resulting dataset. Finally, a short discussion of the statistical analysis results and their applications is presented.
CHAPTER 2

Analysis

2.1 Exploratory Analysis

First, to visualize any obvious patterns in the data, some exploratory plots are made. All variables were standardized for comparison. Tables containing ranges for each variable by site can be found in the Appendix.

To visualize patterns within sites for each variable, a parallel coordinates plot was used (Figure 2.1). The horizontal axis contains all 12 variables, each variable with its own vertical axis. A couple of things can be observed. First, Kiritimati appears to be different from the rest of the sites. Within Kiritimati, sample measurements vary greatly, almost creating two separate clusters - one with high pH, and high concentrations of Mg, Ca, a high alkalinity, and one with low values for these parameters.

The rest of the sites tend to follow more or less the same pattern, with the exception of Pavilion lake having much lower alkalinity. Also, in Pavilion lake most of the variance can be concentrated in two variables - in Si and Ba concentrations (Figure 2.1).

A different way to assess similarity between observations is a dendrogram, a visualization of hierarchical agglomerative clustering (Figure 2.2). Each observation starts out as its own cluster, and then is joined to its closest neighbor based on some distance measure.

From Figure 2.2, it is obvious that Pavilion Lake tends to separate from other sites by forming a cluster onto itself; German streams combine into another group, along with samples from atoll Kiritimati. Even though Kiritimati stands out as a separate cluster, it is joined with German streams sooner than with Pavilion Lake, suggesting more similarities
with the streams than with the samples from the Pavilion region. Thus, the two plots above suggest that samples from atoll Kiritimati are much different from the rest. We now proceed to apply k-means clustering in order to separate the data into $k$ distinct clusters. This separation will give further insights into differences and similarities between observations and will be later used for imputation of missing values.

Figure 2.1: Parallel Coordinates Plot
2.2 Clustering

In this section, we apply k-means with partial distances to divide the dataset into $k$ distinct clusters. K-means is a divisive clustering algorithm that separates the data into $k$ groups by minimizing the distance between observations and the cluster’s center (within sum of squares)[15]. The algorithm requires $k$ to be known in advance, which is rarely the case in practice. A number of techniques have been developed to estimating $k$; here we use a simple silhouette plot as well as a pseudo-F statistic as described in sections 1.2.1 and 1.2.2. Section 1.2.3 describes a modified k-means algorithm used and summarizes clustering results. Section 1.2.4 discusses effects of narrowing the set of variables considered in clustering in order to asses which data features contain most information.

Figure 2.2: Hierarchical clustering of all observations
2.2.1 Finding the Number of Clusters

To search for the best number of clusters, one can compare some measure of goodness of observations’ classification across different values of $k$. One such measure, suggested by Rousseeuw (1987), is a cluster silhouette. It is a simple visual way to determine the optimal number of clusters based on minimizing distance between observations in the same cluster while maximizing separation between the clusters.

Let observation $i$ be classified as a member of cluster $A$. The silhouette value $s(i)$ is defined as

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$  \hspace{1cm} (2.1)

where $a(i)$ is the average distance of $i$ to all other observations in $A$ and $b(i)$ is defined as minimum average distance of $i$ to all observations in clusters other than $A$:

$$b(i) = \min_{C \neq A} d(i, C) = \min_{C \neq A} \frac{1}{|C|} \sum_{j \in C} d(i, j)$$  \hspace{1cm} (2.2)

Since $s(i)$ always lies between -1 and 1, its value may be interpreted as follows:

- $s(i)$ close to 1: object $i$ is well classified (in $A$);
- $s(i)$ close to 0: object $i$ lies intermediate between two clusters ($A$ and $B$);
- $s(i)$ close to -1: object $i$ is badly classified (closer to $B$ than to $A$).

After $s(i)$ is calculated for various number of clusters $k$, a simple plot can be constructed to visualize average silhouette values. The number of clusters that maximize average silhouette can be used as the $k$ in k-means algorithm.

For our data, silhouette was calculated for the number of clusters ranging from 2 to 10 (Figure 2.3). Based on the silhouette plot, it appears that the optimal number of clusters is 2. The average silhouette width for $k = 2$ is 0.947, indicating a well-classified set of observations.
A different approach for determining number of clusters was suggested by R. Tibshirani [24], who defines a gap statistic as follows, with $k$ being the number of clusters:

$$Gap_n(k) = E[\log(W_k)] - \log(W_k)$$

where $W_k$ is pooled within cluster sum of squares. The gap statistic represents the deviation of total within cluster sum of squares from its expected value. A sharp decrease in $\log(W_k)$ indicates that adding the $k^{th}$ cluster to the data is beneficial. In other words, the value of $k$ which maximizes the gap statistic can be taken as the optimal number of clusters for the data. The gap curve plot below was constructed based on the complete observations only (Figure 2.4). Figure 2.4 suggests that 2 clusters are a big improvement in the gap statistic.
from just one cluster, which confirms previous results from hypothesis test. After $k = 4$, the gap statistic curve keeps increasing monotonically with no defined global maximum, suggesting that there might be subclusters defined within clusters.

### 2.2.2 Hypothesis Test on Number of Clusters

In addition to determining the optimal number of clusters it is of interest to ask whether any clustering is beneficial at all or if the data should be treated as one pool of observations ($k = 1$). This is equivalent to testing the following hypothesis:

$$H_0 : k = 1$$

$$H_a : k > 1.$$  

In particular, we first consider the alternative hypothesis with $k = 2$. 

![Gap Statistic Curve, Standardized Data](image)
A number of techniques for such a test have been suggested in literature. For example, a pseudo F-test statistic can be defined. Let WSS and BSS be within and between-cluster sum of squares, respectively. Let n be the number of points being clustered. Then, Calinski and Harabazs [10] define

\[
F = \frac{BSS}{k-1} / \frac{WSS}{n-k}
\]

and suggest using it as an informal test statistic (also referred to as CH index). Cavalli-Sforza (1965) have used this criteria in a context of multivariate cluster analysis as well. The pseudo F ratio above for \( k = 2 \) was found to be 227.94, corresponding to p-value of less than 0.0001. Thus, there is evidence to reject the null hypothesis. Separating data into 2 clusters appears to be more beneficial than considering all observations as a uniform pool.

Naturally, the next question arises if \( k = 3, 4 \) or more clusters would better fit the data than \( k = 2 \). Calinski and Harabazs suggest choosing \( k \) for which the CH index reaches a global or a local maximum, or at least has a rapid increase. A plot of CH index as a function of the number of the clusters reveals that 2 clusters appear to be a significantly better choice than any other number up to 10 (Figure 4).

![Figure 2.5: CH Index](image)
Thus, so far silhouette plot, a hypothesis test, and gap statistic confirm that 2 clusters is the optimal fit for the data. We know proceed to applying a modified version of k-means algorithm with \( k = 2 \).

### 2.2.3 K-means Clustering

Since traditional k-means is based on Euclidean distance and cannot handle data with missing values, we modify the algorithm to use partial distance instead, and therefore utilize all available information [14]. Note that a more robust version of k-means, based on clustering around medoids is used here (Kaufman and Rosseuw) [16]. A medoid is an object representative of a cluster such that total dissimilarity of all objects to their nearest medoid is minimal. An example of medoid is a cluster mean.

Partial distance \( d_{\text{part}}(\mathbf{x}, \mathbf{y}) \) between observations \( \mathbf{x} = (x_1, \ldots, x_i, \ldots, x_p) \) and \( \mathbf{y} = (y_1, \ldots y_i, \ldots, y_p) \) in a \( p \)-dimensional space is computed as follows [14]:

\[
d_{\text{part}}(\mathbf{x}, \mathbf{y}) = \frac{p}{p - \sum_{i=1}^{p} b_i \sum_{\forall i: b_i = 0} (x_i - y_i)^2} (2.5)
\]

where

\[
b_i = \begin{cases} 
0, & \text{if } x_i, y_i \text{ are observed} \\
1, & \text{otherwise} 
\end{cases} (2.6)
\]

and \( i = 1 \ldots p \).

Let \( k \) be the pre-determined number of clusters and \( n \) be the number of observations. The clustering algorithm proceeds as follows:

1. Construct initial medoids \( m_1 \ldots m_k \) (i.e. cluster centers) to minimize the sum of all distances between observations in a cluster and cluster’s center. This is equivalent to minimizing the objective function \( \sum_{i=1}^{n} \min_{t=1 \ldots k} d_{\text{part}}(i, m_t) \) as follows:

   - \( m_1 \) minimizes \( \sum_{i=1}^{n} d_{\text{part}}(i, m_1) \)
• each $m_2 \ldots m_k$ maximize the decrease in the objective function.

Assign observation $i$ to cluster $k$ based on the minimum $d_{\text{part}}(i, m_k)$.

2. Let $i$ be an observation belonging to cluster $C$, $j$ observation not in $C$. Swap $i$ and $j$ cluster membership if doing so will decrease the objective function. Repeat until convergence.

After applying this algorithm to the data, two clusters emerge. One cluster is composed of 6 observations from Kiritimati atoll, and the second cluster consists of all other observations. In particular, the 6 observations from Kiritimati are characterized by high concentration of Ca and Mg and consist of samples collected from 4 different lakes in the basin, a trench/crab burrow in a lake, and a sample from Artemia ponds.

We next explore which data features (i.e. variables) might be responsible for the achieved clustering.

2.2.4 Clustering a subset of data

In order to find out what variables contain the most information useful for clustering, we apply k-means to the set of 6 most complete variables and compare the resulting cluster assignment to that of the whole data. Then, we narrow down this set to see what how much we can reduce the set of variables in consideration while still retaining original clustering assignment.

We start with the most complete variables (pH, Ca, Mg, Conductivity, K, Na), delete incomplete observations (resulting number of rows=144) and apply clustering to this dataset. Resulting clustering on these six variables are identical to that on whole dataset with partial distances (two clusters, six observations from Kiritimati forming its own cluster). This suggests that variables that were left out of this clustering analysis did not add any new information to the structure.

Next, we try to see if any of the six variables can also be omitted. In fact, clustering
<table>
<thead>
<tr>
<th>Number of Variables Clustered</th>
<th>Set of variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Ca, Conductivity, Mg, K, Na</td>
</tr>
<tr>
<td>3</td>
<td>Ca, Conductivity, Mg, K, Na</td>
</tr>
<tr>
<td>4</td>
<td>Any of the six (pH, Ca, Conductivity, Mg, K, Na)</td>
</tr>
<tr>
<td>5</td>
<td>Any of the six</td>
</tr>
<tr>
<td>6</td>
<td>Any of the six</td>
</tr>
</tbody>
</table>

Table 2.1: Set of variables producing 2 original clusters

each combination of two out of the six variables (but excluding pH) lead to the clustering assignment identical to that of the whole data. In other words, restricting the dataset to contain information only on any two of the variables in Table 1 will produce 2 clusters with one of the groups being the 6 observations from Kiritimati.

The same holds for combinations of three variables: any combination of three variables, except for those combinations containing pH, produces the same clustering assignment. Combinations containing pH produce 3 clusters: 2 original clusters and a third cluster containing 26 observations from all sites.

When clustering any 4 variables at a time, resulting cluster assignment is identical to that of clustering six variables (or the whole dataset), even for combinations containing pH. So any 3rd variable combined with pH and another variable (or with any 2 variables) will produce 2 clusters as when clustering the whole dataset.

Using correlation between the six most complete variables as a dissimilarity matrix, we can construct a dendrogram that shows similarity between the variables. pH once again appears separate from the other 5 variables. Thus, we can conclude that pH seems to play a crucial role in separating the data into two groups.

This concludes the clustering part of the analysis. We have determined the optimal number of clusters to be 2, separating the data into one group of 6 observations from atoll Kiritimati and another containing the rest of observations. We then explored the variables used in clustering and concluded that pH plays an important role in separating data into 2 groups.
In the remainder of this paper, we use the clustering information to impute the missing values. First, we explore the missingness mechanism and search for any dependencies between the data and missingness. Then, we impute values by the appropriate cluster mean. Finally, as a way to assess sensitivity of the clustering assignment, we randomly delete some observations, perform clustering analysis again, and compare results to the original groups.

2.3 Missing At Random and Missing Completely at Random

Next, we explore the relationship between missingness and observed values in the dataset. It is important to understand the mechanism that led to missing data and address it in the analysis. If data is missing at random (or completely at random), the mechanism can be ignorable (Little and Rubin 1987). However, if missingness of a particular variable seem to depend on the value of that variable, this fact will need to be taken into account in
imputation.

In general, let $R = [r_{i,j}]_{m \times n}$ be an indicator matrix, with $r_{ij} = 1$ if a value is observed and 0 otherwise. Let $\psi$ be some unknown parameter. Then, missing data mechanism can be classified as one of the following:

1. Data is missing at random (MAR) and observed at random (OAR), in other words missing completely at random - MCAR. Observed values are a random subsample of the whole sample. Probability of a data point being observed is independent of its value:

$$f(R|Y_{obs}, Y_{miss}, \psi) = f(R|\psi)$$

(2.7)

2. Data are missing at random (MAR). Observed values of variable $Y$ do not necessarily form a random subsample of the whole sample, but they are a random sample within subclasses defined by variable an observed variable $X$. Probability of $Y$ being observed may depend on values of $X$, but not on $Y$:

$$f(R|Y_{obs}, Y_{miss}, \psi) = f(R|Y_{obs}, \psi).$$

(2.8)

3. The probability of $Y$ being observed depends on the value of $Y$ and possibly $X$ as well. In this case the data is neither MAR nor OAR. The missing data mechanism should be accounted for in the analysis (i.e. non-ignorable).

Little (1988) proposed a formal test statistic to test if data are MCAR. Notation is as follows. If $p$ is the number of variables in the data and $n$ is the number of observations, let $y_{i(1 \times p)}, i = 1 \ldots n$ be a vector of observations, $r_{i(1 \times p)}$ be a vector of missing data indicators (1 if a value is observed, 0 otherwise). Each unique $r_{i}$ is a pattern of missing data. Let $J$ be the number of distinct missing data patterns in the dataset. Let $\hat{\mu}$ and $\hat{\Sigma}$ be MLE estimates of population mean and variance, and $m_{j}$ the number of cases pattern $j$ occurs.
Then, assuming that $\Sigma$ is unknown, a test statistics is defined as

$$d^2 = \sum_{j=1}^{J} m_j (\bar{y}_{obs,j} - \hat{\mu}_{obs,j}) \hat{\Sigma}^{-1}_{obs,j} (\bar{y}_{obs,j} - \hat{\mu}_{obs,j})^T.$$  \hfill (2.9)

R. Little proposes that if data is MCAR then $d^2$ has a chi-squared distribution with $f = \sum p_j - p$ degrees of freedom.

We first test the MCAR assumption using Little’s test. A matrix $R$ of missing values indicators was created, and 8 distinct missingness patterns were found. Table 2 below shows the patterns and number of observations per each pattern.

<table>
<thead>
<tr>
<th>Pattern</th>
<th># of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>95</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
</tbody>
</table>

Table 2.2: Missing Data Patterns

Test statistic $d^2$ was found to be 68.66, which corresponds to a p-value of less than 0.001. Thus, we conclude that the data at hand is not MCAR.
To further explore the Missing At Random (MAR) assumption, plots of variable means for each pattern were made. Although this visualization is by no means a formal test, it can give us an idea about the distribution of the mean of a particular variable marginalized by the missingness pattern. For instance, if sample mean varies dramatically across the patterns, it may be a sign of some relationship between missing and observed values. These plots can provide an insight about $f(Y_{\text{obs}}|R)$. 

![Plot of pH](image1)

![Plot of Ca](image2)

![Plot of Mg](image3)

![Plot of COND](image4)
Figure 2.7: Mean By Missing Data Patterns
To explore this relationship more formally, each variable was categorized by its quantiles (top two and bottom two quantiles), and further by missingness. In particular, if all values are observed completely, the category is coded as 1. If some of the values are not observed, then category is coded as 0. This produces a set of 12 2x2 tables (see Appendix A), each table corresponding to a variable. A chi-square test for association is applied to each table. For all 12 tables, the test detects some association at a 5% significance level, which could imply there may be some significance to association. However, it is important to note that this association might not be the product of a true dependency between missingness and observed values, but rather a consequence of the missingness patterns themselves. For example, pattern 4 (with all variables observed except for the last one, Ba) is unique to observations taken from atoll Kiritimati. Since observations taken from this site were found to be different from the rest, so are the means of all variables for pattern 4.

Information from the Appendix A and Figure 2.7 can be used to provide some intuition about the missing at random assumption. Since means for most patterns (with the exception of pattern 4) are similar for majority of variables, it does not appear that the means are depended upon missingness. Although this procedure is not a substitution for a formal test of MAR assumption (as defined on p.15), due to the lack of such a test we proceed to make that assumption.

It is important to understand the consequences of wrongly assuming that the data are MAR. For instance, a researcher might not consider values of a variable in geochemical analysis if concentration is below a certain threshold. In this case missingness of the variable is dependent on its value, and therefore the data are not missing at random. If that is the case, the missing data mechanism is what Little and Rubin call non-ignorable. Little and Rubin describe some imputation methods for data that is not MAR, largely likelihood-based theory for estimating parameters of the missing data distribution [23].

However, our dataset seem to satisfy the MAR assumption. Since the data is a combination of measurement from different studies, it is reasonable to say that a particular variable was not measured in a study due to the research question of interest, and not because of the
variable’s value.

## 2.4 Imputation and Validation

Now that the data are separated into two distinct groups, we can use cluster information to impute the missing values. A number of imputation procedures have been developed and used extensively in literature. One simple way to impute missing values is to use cluster mean (imputation by unconditional means, [23]). Specifically, we can use the basic idea that observations in a given cluster are similar to each other, and therefore missing values can be imputed by a simple cluster mean of that variable. For variable $Y$ we impute missing values $y_{ij}$ with cluster mean of observed values $\bar{y}_j$. We also keep track of cluster variance for each variable. However, this procedure is sensitive to the MCAR assumptions, which is not satisfied in our case.

Another possible procedure is imputation by conditional means (Bucks’ Method)[23]. Missing values are filled in with means that are conditioned on recorded variables in an incomplete case. In other words, missing values are regressed on observed variables. This procedure rests on normality assumption. Instead, we would like to use a non-parametric method that is not dependent on MCAR assumption.

We choose k-nearest neighbor imputation (KNN) described by Dixon J.K. in 1979, Troyanskaya O. in 2001 [11]. This method relaxes assumption of MCAR (requiring only a weaker MAR) and does not make any parametric assumptions.

KNN imputation borrows from the simple idea of similarity between the features based on some metric, such as Euclidean distance. If a value in column is missing, then $K$ nearest neighbors are found based on Euclidian distance and the average of these neighbors is used to fill in the missing value. More specifically, the algorithm proceeds as follows:

1. Divide the dataset $X$ into $X_{obs}$, $X_m$, where $X_m$ contains all observations with at least one missing values, and $X_{obs}$ contains all observations with no missing values.
2. For each observation $x$ in $X_m$,

   (a) Separate $x$ into missing and observed items, $x_m, x_{obs}$, respectively.

   (b) Calculate Euclidean distances between $x_{obs}$ and all observations in $X_{obs}$ using only those items (columns) that are observed in vector $x_{obs}$.

   (c) Use $K$ nearest (based on the Euclidean distance) neighbors of $x_{obs}$ to calculate an average. Replace missing items $x_m$ with the corresponding average.

This procedure requires $K$, number of neighbors to use, as an input. Troyanskaya O. shows that the algorithm is relatively insensitive to the choice of $K$ [25]. Here we calculate $K$ using R package *imputation*, which calculates $K$ based on cross-validation. Taking the dataset as the input, the R function randomly erases some data and fills in missing values using KNN imputation with various values of $K$. Mean absolute error is then used as the criteria for choosing the best $K$.

In our case, observations separated into clusters are already grouped by their similarity to each other. Further applying KNN imputation to each individual cluster will impute the missing values by the average of $K$ nearest neighbors in the cluster. Proceeding in such fashion, we arrive at a complete dataset.

In order to validate this analysis, randomly delete missing values and repeat clustering and imputation procedures, as follows:

1. Randomly delete 8% (the original proportion of missing values) of values.

2. Cluster the resulting dataset using k-means with partial distances, as described in section 2.2.

3. Perform KNN imputation using results from clustering

4. Repeat steps 1-3 100 times

5. Calculate average distance between each imputed dataset and the original.
After this simulation is complete, a histogram of average distances from step 5 was made (Figure 2.7). This histogram was constructed by calculating pairwise Euclidean distances between observations from original and simulated datasets, and then taking an average. This produced a measure of variability in our procedure of imputation based on clustering. Resulting histogram has a shape resembling normal distribution. This provides an approximate validation to our results, since it looks like the average distance is stable, producing a bell curve. Since most distances fall into the fairly small interval of (1.66, 1.73), it implies that clustering and imputation procedure consistently producing datasets similar to each other, with no big variations. Thus, this procedure might be applied to an expanded dataset containing information on more microbialite-containing sites.

Figure 2.8: Average Variability of Clustering and Imputation Procedures
2.5 Environmental characteristics limiting microbialite distribution

It was found that all variables in the dataset vary widely across the aquatic environments by pH, salinity (indicated by variable conductivity), and ionic composition. Tables 2.3-2.5 summarize the range of each variable by microbialite site. German karst water streams names were abbreviated as follows: Westerhofer - WB, Deinschwanger - DB, Reinsgraben - Rein, Steinerne Rinne - Stein. Samples from atoll Kiritimati vary the most across all variables, as previously shown on Figure 2.1, while somewhat separating into two distinct groups (one with low concentration of Ca, Mg, and salinity and one with high). Values for Ca, Mg, K, Cl, SO\textsubscript{4} and conductivity have a particularly large range (Table 2.3) for samples from atoll Kiritimati and Pavilion Lake. Such a wide range may be explained by local conditions of the aquatic environment from which the samples were taken. In addition to the biggest variation within the site, observations from Kiritimati also appear to be different from the others, forming a cluster on their own. Thus, the data at hand shows a wide range in values of geochemical features for given microbialite environments.

<table>
<thead>
<tr>
<th></th>
<th>Pavilion Lake</th>
<th>Kiritimati</th>
<th>WB</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.12-9.13</td>
<td>7.16-9.64</td>
<td>7.32-8.3</td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td>0.92-10.63</td>
<td>0.22-38.6</td>
<td>3.58-3.95</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Mg</td>
<td>0.22-255.09</td>
<td>0.1-279.7</td>
<td>1.67-1.72</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Conductivity</td>
<td>0.27-26.6</td>
<td>0.2-156.3</td>
<td>0.90-1.04</td>
<td>mS/cm</td>
</tr>
<tr>
<td>Alkalinity</td>
<td>0-0.018</td>
<td>0.19-14.2</td>
<td>4.8-5.4</td>
<td>mequiv/L</td>
</tr>
<tr>
<td>SO\textsubscript{4}</td>
<td>0.075-321.66</td>
<td>0.05-139.17</td>
<td>2.81-2.95</td>
<td>mmol/L</td>
</tr>
<tr>
<td>K</td>
<td>0.018-22.30</td>
<td>0-48</td>
<td>0.052-0.055</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Na</td>
<td>0.061-211.83</td>
<td>1-2334</td>
<td>0.33-0.34</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Si</td>
<td>0.0018-0.584</td>
<td>0.002-0.1924</td>
<td>0.157-0.159</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Cl</td>
<td>0.0169-17.347</td>
<td>0-2637</td>
<td>0.291-0.299</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Sr</td>
<td>0.0009-0.0944</td>
<td>0.0013-0.3237</td>
<td>0.0175-0.019</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Ba</td>
<td>7.28E-05-0.0012</td>
<td>n/a</td>
<td>0.00021-0.00025</td>
<td>mmol/L</td>
</tr>
</tbody>
</table>

Table 2.3: Ranges by Site

Furthermore, it was found that pH is a defining feature of the geochemical profile in
<table>
<thead>
<tr>
<th></th>
<th>DB</th>
<th>Rein</th>
<th>Stein</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.36-8.5</td>
<td>7.31-8.28</td>
<td>7.1-8.03</td>
</tr>
<tr>
<td>Ca</td>
<td>1.87-2.28</td>
<td>4.89-5.2</td>
<td>2.86-3.52</td>
</tr>
<tr>
<td>Mg</td>
<td>0.86-1.27</td>
<td>1.11-1.13</td>
<td>0.13-0.14</td>
</tr>
<tr>
<td>Conductivity</td>
<td>0.58-0.64</td>
<td>1.14-1.23</td>
<td>0.52-0.65</td>
</tr>
<tr>
<td>Alkalinity</td>
<td>4.68-5.22</td>
<td>4.63-5.2</td>
<td>4.42-6.1</td>
</tr>
<tr>
<td>SO4</td>
<td>0.17-0.19</td>
<td>3.63-3.8</td>
<td>0.27-0.27</td>
</tr>
<tr>
<td>K</td>
<td>0.028-0.058</td>
<td>0.043-0.046</td>
<td>0.014-0.022</td>
</tr>
<tr>
<td>Na</td>
<td>0.31-0.35</td>
<td>0.606-0.623</td>
<td>0.13-0.20</td>
</tr>
<tr>
<td>Si</td>
<td>0.097-0.102</td>
<td>0.148-0.155</td>
<td>0.084-0.084</td>
</tr>
<tr>
<td>Cl</td>
<td>0.55-0.65</td>
<td>0.546-0.596</td>
<td>0.14-0.14</td>
</tr>
<tr>
<td>Sr</td>
<td>0.00032-0.00055</td>
<td>n/a</td>
<td>0.00151-0.00156</td>
</tr>
<tr>
<td>Ba</td>
<td>0.00011-0.00013</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>


Table 2.4: Ranges by Site, continued

<table>
<thead>
<tr>
<th></th>
<th>Range</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>7.10 - 9.64</td>
<td></td>
</tr>
<tr>
<td>Ca</td>
<td>0.22 - 38.60</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Mg</td>
<td>0.10 - 279.70</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Conductivity</td>
<td>0.20 - 156.30</td>
<td>mS/cm</td>
</tr>
<tr>
<td>Alkalinity</td>
<td>0.0 - 14.20</td>
<td>mequiv/L</td>
</tr>
<tr>
<td>SO4</td>
<td>0.05 - 139.17</td>
<td>mmol/L</td>
</tr>
<tr>
<td>K</td>
<td>0.0 - 48.0</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Na</td>
<td>0.061 - 2334.0</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Si</td>
<td>0.00200 - 0.52696</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Cl</td>
<td>0.0 - 2637.0</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Sr</td>
<td>0.00032 - 0.32370</td>
<td>mmol/L</td>
</tr>
<tr>
<td>Ba</td>
<td>0.00011 - 0.001</td>
<td>mmol/L</td>
</tr>
</tbody>
</table>

Table 2.5: Full ranges for all variables
this dataset, and determines the clustering of the observations. This implies that certain microbial taxa may be restricted to particular types of environments, which could be tested in the future using genetic or other types of geobiological data.

Thus, results from this work suggest that microbialites are broadly distributed across the environments with a wide spectrum of geochemical characteristics. This has implications for the conditions for the occurrence of microbialites on the Early Earth and other planetary bodies. However, statistical analysis conducted here was restricted to the set of 6 microbialite sites. In the future, k-means clustering with partial distance can be easily applied to a bigger dataset, possibly containing hydrographic data from a different planetary body. If the new dataset will contain missing values, KNN imputation can be used to fill in the missing values, conduct further statistical analysis, and to help draw conclusions on conditions conducive to microbialite development.
In this work, we have explored properties of geochemical characteristics of 6 microbialite environments. The initial dataset with missing values was clustered via k-means algorithms using partial distances. Then, the dataset was narrowed to the 6 most complete variables (pH, Ca, Mg, conductivity, K, Na) and analysis was repeated to determine whether there is a subset of variables that produces the same clustering results as with the whole dataset.

This resulted in pH being separated out as particularly different from the rest of the variables, with any combination of the 3 variables resulting in the same clustering as originally. This suggests that pH can be used as the defining feature of the geochemical profile in this dataset.

Clustering techniques also distinguished samples from atoll Kiritimati to be particularly distinct from the rest. Observations from this site vary greatly in values of all variables, and seem to create two sub clusters - one with high pH, and high concentrations of Mg, Ca, and alkalinity, and one with low.

Another feature found is low alkalinity of Pavilion Lake in comparison to other observations. Also, most of the variance in Pavilion Lake samples seems to be mostly explained by only two variables, Si and Ba.

Then, observations that were omitted originally were clustered using partial distances as the measure of similarity. The resulting clustering assignment was used to impute missing values using k-nearest neighbors procedure. This imputation and clustering procedure was repeated 100 times to asses variability of this method. Plotting a histogram of these average distances resulted in a bell-shaped curve, with fairly small variance. This suggests cluster-
ing and imputation results achieved here are fairly stable, since comparable datasets were reproduced even from different simulated missingness patterns.

This paper can be used as a general outline of methods that could be applied to an expanded multivariate dataset with missing values. Potential areas for future work include considering more sites that contain microbialites, and conducting a different multivariate statistical analysis on the expanded and imputed dataset. Also, different imputation techniques could be applied and compared against the one presented here.
APPENDIX A

Frequency Tables for Missingness Patterns

Each table contains counts of observations in corresponding category. Columns of each table are the missingness categories (1 means observation is complete, 0 means otherwise). Rows are top and bottom 50% quantiles for each variable.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Observed</th>
<th>Partially Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(7.1, 8.25]</td>
<td>33  40</td>
</tr>
<tr>
<td></td>
<td>(8.25, 9.64]</td>
<td>19  55</td>
</tr>
<tr>
<td>Ca</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.22, 1.95]</td>
<td>8   65</td>
</tr>
<tr>
<td></td>
<td>(1.95, 38.6]</td>
<td>43  30</td>
</tr>
<tr>
<td>Mg</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.1, 0.86]</td>
<td>9   65</td>
</tr>
<tr>
<td></td>
<td>(0.86, 280]</td>
<td>42  30</td>
</tr>
<tr>
<td>Conductivity</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.2, 0.602]</td>
<td>4   69</td>
</tr>
<tr>
<td></td>
<td>(0.602, 156]</td>
<td>47  26</td>
</tr>
<tr>
<td>Alkalinity</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0, 0.0026]</td>
<td>1   70</td>
</tr>
<tr>
<td></td>
<td>(0.0026, 14.2]</td>
<td>48  24</td>
</tr>
<tr>
<td>SO4</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.05, 0.541]</td>
<td>6   64</td>
</tr>
<tr>
<td></td>
<td>(0.541, 139]</td>
<td>38  31</td>
</tr>
<tr>
<td>K</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0, 0.0582]</td>
<td>25  48</td>
</tr>
<tr>
<td></td>
<td>(0.0582, 48]</td>
<td>26  47</td>
</tr>
<tr>
<td>Na</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.0609, 0.339]</td>
<td>9   64</td>
</tr>
<tr>
<td></td>
<td>(0.339, 2.33e+03]</td>
<td>43  30</td>
</tr>
<tr>
<td>Si</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.002, 0.174]</td>
<td>39  30</td>
</tr>
<tr>
<td></td>
<td>(0.174, 0.527]</td>
<td>4   65</td>
</tr>
<tr>
<td>Cl</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.017]</td>
<td>6   63</td>
</tr>
<tr>
<td></td>
<td>(0.17, 2.64e+03]</td>
<td>38  32</td>
</tr>
<tr>
<td>Sr</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.00032, 0.00288]</td>
<td>10  54</td>
</tr>
<tr>
<td></td>
<td>(0.00288, 0.324]</td>
<td>25  40</td>
</tr>
<tr>
<td>Ba</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>(0.00011, 0.000328]</td>
<td>0   47</td>
</tr>
<tr>
<td></td>
<td>(0.000328, 0.00111]</td>
<td>4   44</td>
</tr>
</tbody>
</table>

Table A.1: Counts of fully and partially observed rows for each variable. 1 = partially observed, 0 = otherwise
REFERENCES


