2010-04-20

Cluster and Classification Analysis of Fossil Invertebrates within the Bird Spring Formation, Arrow Canyon, Nevada: Implications for Relative Rise and Fall of Sea-Level

Scott L. Morris
Brigham Young University - Provo

Cluster and Classification Analysis of Fossil Invertebrates within the Bird Spring Formation, Arrow Canyon, Nevada: *Implications for Relative Rise and Fall of Sea-Level*

Scott Lee Morris

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Dr. William F. Christensen, Chair
Dr. Scott M. Ritter
Dr. W. Evan Johnson

Department of Statistics

Brigham Young University

August 2010

ABSTRACT

Cluster and Classification Analysis of Fossil Invertebrates
within the Bird Spring Formation,
Arrow Canyon, Nevada:
Implications for Relative Rise and Fall of Sea-Level

Scott Lee Morris

Department of Statistics

Master of Science

Carbonate strata preserve indicators of local marine environments through time. Such indicators often include microfossils that have relatively unique conditions under which they can survive, including light, nutrients, salinity, and especially water temperature. As such, microfossils are environmental proxies. When these microfossils are preserved in the rock record, they constitute key components of depositional facies. Spence et al. (2004, 2007) has proposed several approaches for determining the facies of a given stratigraphic succession based upon these proxies. Cluster analysis can be used to determine microfossil groups that represent specific environmental conditions. Identifying which microfossil groups exist through time can indicate local environmental change. When new observations (microfossils) are found, classification analysis can be used to predict group membership.

Kristen Briggs (2005) identified the microfossils present in sedimentary strata within a specific time interval (Morrowan) of Pennsylvanian-age rocks. In this study we expand analysis to overlying Atokan and Desmoinesian strata. The Bird Spring Formation in Arrow Canyon, Nevada records cycles of environmental change as evidenced by changes in microfossils. Our research investigates cluster and classification analyses as tools for determining the marine facies succession. Light, nutrients, salinity, and water temperature are very dependent on water depth; therefore, our analyses essentially indicate the relative rise and fall of sea-level during Early to Middle Pennsylvanian time.

Keywords: classification, geostatistics

CONTENTS

CHAPTER 1

INTRODUCTION

The surface of the earth changes continuously over geologic time. Nevada's Arrow Canyon, currently situated hundreds of miles from any massive oceanic body, was once submerged by ocean water. The area fluctuated during that time between shallow and deep marine environments as a function of glacioeustatic sea-level change. The purpose of this project is to identify the geologic structure in petrographic data from Pennsylvanian-age depositional cycles from the Morrowan through Desmoinesian strata of Arrow Canyon's Bird Spring formation to enhance recognition of sea-level change.

Facies refers to the environment where deposition of both organic and inorganic material occurs, or conversely, to the composition and texture of sedimentary rocks deposited

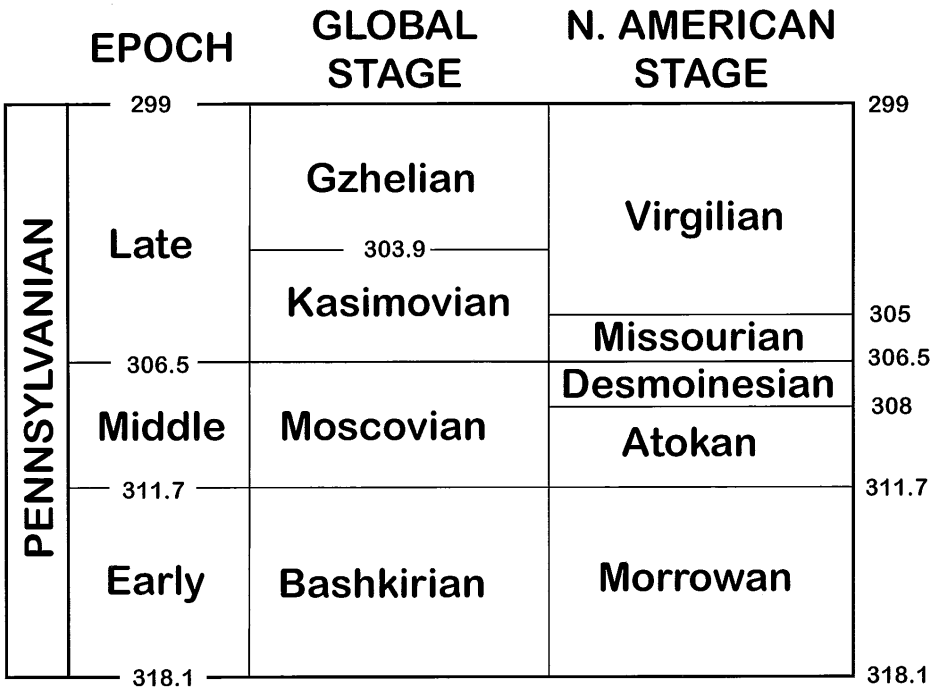

Figure 1.1: Data analyzed were deposited during Pennsylvanian geologic time.

in a given environment. This paper proposes several methods for identifying the facies for each of the 809 samples of sedimentary rock from Arrow Canyon. Cluster and classification analyses were used to cluster samples based on 29 attributes found in various abundances within the samples, and to create a rule for assigning new observations into one of the defined clusters.

1.1 Description of Data

The data matrix is composed of 809 samples, each containing measurements based on 29 attributes. Samples were taken from a successive stratigraphic column covering the Early and Middle Pennsylvanian Epoch, ranging from approximately 306 million to 318 million years ago.

The 29 attributes measured are limestone characteristics that indicate both high amplitude and high frequency climate cyclicity over time (Briggs, 2005). Specifically, the data includes measurements of many types of invertebrate microfossils as well as foraminifera (ocean-dwelling organisms) and several types of algae. Geologists can identify shallow to deep marine environments based on the composition of samples; that is, a sample high in algae and low in deeper-ocean fossils indicate shallow marine environments (James, 1997).

Heterozoan organisms are light-independent fossils; in other words, they thrive in deep, cold, and dark environments. Photozoan organisms are light dependent and require light and warmer water (present in shallow marine environments) to survive. Thus, the samples will generally exhibit tendencies to have either heterozoan or photozoan organisms. Quantifying these samples can determine the type of marine environment (facies) in which each sample was deposited.

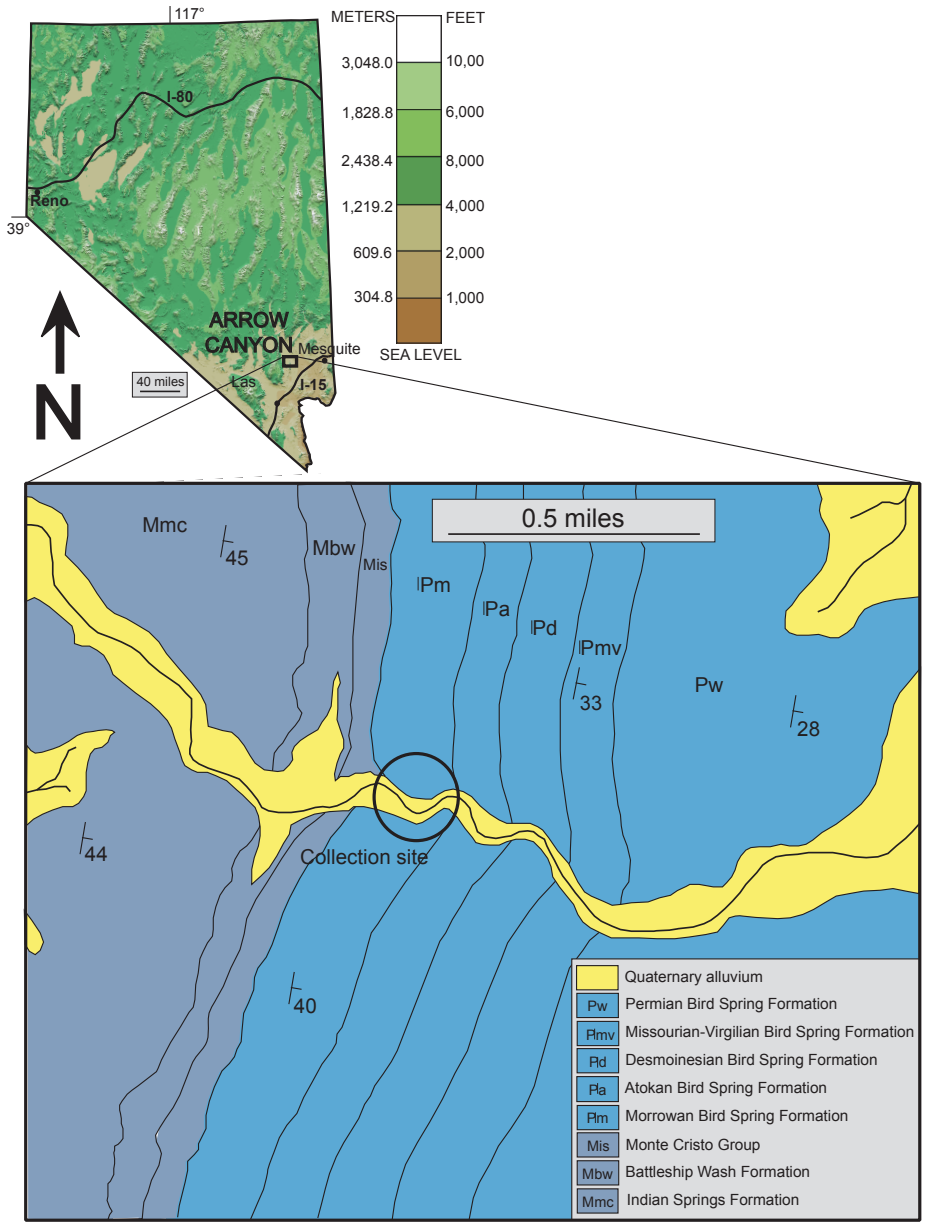

Figure 1.2: Location of Arrow Canyon, Nevada.

1.2 Literature Review

Briggs (2005) researches the Arrow Canyon area with a sedimentologically applied focus. She uses approximately 400 observed samples with 21 measured variables to estimate the type and order of marine facies cycles in Morrowan strata.

Briggs's research focuses on determining the periodicity and duration of climate shifts as evidenced by changes in sea level (and the corresponding marine environments evident from the fossils in the Arrow Canyon area). Her approach incorporates many nonstatistical measures of association, and her actual statistical methodology is not described in any depth. To determine her clusters, she applies hierarchical agglomerative clustering, using average linkage as her choice of distance metric. She concludes that 12 clusters adequately describes the differences in the samples. However, she gives no measure of the clusters' significance nor any idea of how well the chosen clusters would classify a new observation.

To geologists and nongeologists alike, identifying a facies based on the attributes found in a given cluster is nontrivial. James (1997) has studied the characteristics of many of the attributes used in our study, and we will use his work to aid in the identification of facies based on the attributes found within a sample.

Spence and Tucker (2007) give several possible ways to reconstruct marine facies given numerous attributes, and Spence et al. (2004) propose heavily computerized methods of accomplishing similar tasks. We will use these researchers' results to inform our own statistical results.

CHAPTER 2

STATISTICAL METHODOLOGY

Using a relatively richer data set and more rigorously defined statistical methods, we cluster and classify 809 samples to determine which number of clusters is most informative for the data. Furthermore, we determine classification rules for identifying the group membership of a new observation.

2.1 Cluster Analysis

We used the complete linkage distance metric to determine the initial clusters to be used in a k-means clustering algorithm. K-means clustering uses k previously established data clusters, then assigns each object to a cluster with the closest mean, according to Euclidean distance. These reassignments are repeatedly made until every observation remains in the cluster to which it was assigned before being compared with all other cluster means.

We initially believed that the data would naturally form between 10 and 20 clusters (Ritter, 2010). Using the k-means algorithm, 8–15 clusters were formed. Since the interpretation of more than 15 clusters is not practically feasible and using less than eight clusters might omit pertinent information, we evaluate the goodness-of-fit for these different clusterings that used distinct values for the number of clusters.

A MANOVA analysis of the clusters shows that each clustering is statistically significant, according to Wilk's Λ test. Since a large part of this analysis, and future work, requires good classification rates, we used cross-validation to determine which clustering strategy results in the lowest misclassification rates. Table 2.1 shows the misclassification rates using a k-means clustering algorithm with linear, quadratic, and 10-nearest neighbor cross-validated classification rules. A simple pilot study showed that using 10 instead of 5 as the number of nearest neighbors results in better clustering ability. Since we are simply

interested in using the method that gives the lowest misclassification rates, we choose $k = 9$ clusters, using a linear classification scheme.

Table 2.1: Given in the table are # of clusters, Wilk's Λ Statistics, P-values, and linear, quadratic, and 10-nearest neighbor misclassification rates.

# of Clusters	$\Lambda \times 10^4$	P-value	Linear	Quadratic	10-N.Neighbor
8	0.8181	<.0001	0.0729	0.1545	0.1446
9	0.3379	<.0001	0.0581	0.1125	0.1248
10	0.1068	<.0001	0.0618	0.1928	0.1409
11	0.0427	<.0001	0.0724	0.0989	0.1520
12	0.0255	<.0001	0.0816	0.0939	0.1632
13	0.0143	<.0001	0.0828	0.1508	0.1656
14	0.0075	<.0001	0.0816	0.1014	0.1434
15	0.0036	<.0001	0.0717	0.3028	0.1520

Figure 2.1 gives a plot of the misclassification rates among linear, quadratic, and 10-nearest neighbor classification schemes, as a function of the number of clusters. It is apparent that nine clusters relatively minimizes misclassification rates universally and uniquely minimizes the misclassification rates for linear classification.

2.2 Classification and Discriminant Analysis

Discriminant function analysis allows us to significantly reduce the number of dimensions needed to describe the differences between clusters. Using k=9 clusters, there are at most eight dimensions with which to describe differences between the clusters. Table 2.2 gives the standardized discriminant function coefficient vectors ($\times 10,000$), the corresponding eigenvalues, percentage of the difference in cluster means explained by each dimension, and the cumulative sum of these percentages.

The eigenvalues and eigenvectors from Table 2.2 indicate that the cluster means are relatively diffuse in eight-space. The diffuseness in clusters may be due to the relatively low resolution of the data set. Currently, the data are given a score between zero to five, where zero represents the attribute never occurring in a sample and five represents the attribute

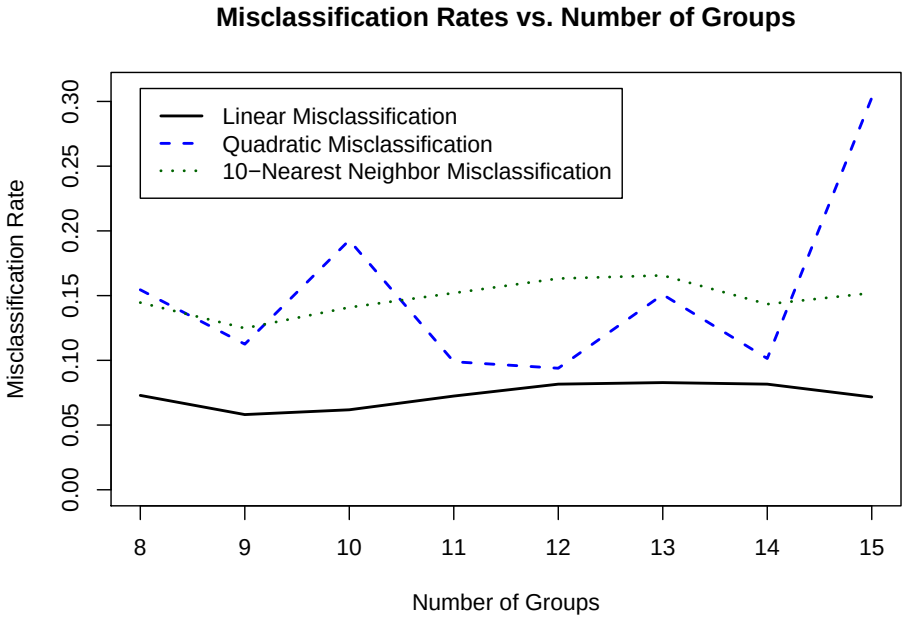

Figure 2.1: Misclassification rates of linear, quadratic, and 10-nearest neighbor classification schemes, as functions of number of clusters.

occurring more than 100 times in the sample. A higher resolution data set may reduce the dimensionality in the decomposition of $\mathbf{E}^{-1}\mathbf{H}$.

However, we still draw many useful conclusions from our data. We choose to interpret the first four dimensions on the grounds that nearly $\frac{3}{4}$ of the variability between clusters is accounted for in those first four, and most scientists would be hard-pressed to make meaningful interpretations of any information beyond four dimensions.

Table 2.2: Shown in the table are the Attribute Names, $10,000\times$ the standardized discriminant coefficient vectors, and the corresponding eigenvalues with their contributions to group separation.

Attribute Name	$\mathbf{a}_1^\star$	$\mathbf{a}_2^\star$	$\mathbf{a}_3^\star$	$\mathbf{a}_4^\star$	$\mathbf{a}_5^\star$	$\mathbf{a}_6^\star$	$\mathbf{a}_7^\star$	$\mathbf{a}_8^\star$
Spongespicules	-0.87	-0.04	-0.60	-0.55	-0.65	-1.38	-0.28	0.13
Phosphaticgrains	0.64	-0.34	0.05	0.48	0.20	2.10	0.28	-0.97
Articulatebrachiopods	1.46	0.48	0.45	2.25	-1.75	3.89	1.21	-0.47
Bryozoaencrusting	0.98	-0.20	-0.39	-1.55	-0.41	2.03	0.59	-1.40
Bryozoafenestrateramose	0.43	0.23	0.11	0.45	-1.79	4.89	1.72	-1.29
Crinoidea	1.66	0.45	0.45	1.27	0.35	5.26	0.22	-0.20
Echinoids	-0.10	-0.49	1.00	1.59	0.22	1.84	0.69	-1.93
Ostracoda	0.98	0.33	0.98	0.22	-0.15	1.40	-0.01	-1.81
Trilobita	0.57	0.06	0.11	1.35	-1.09	2.84	0.76	-0.80
Bivalvia	-0.75	-1.53	3.17	0.87	3.72	-0.87	-0.38	-1.98
Gastropoda	-0.08	-0.48	3.37	0.35	3.50	-0.98	0.67	-2.41
Osagialiths	-1.91	-0.34	1.23	0.89	1.71	-1.19	-1.70	-0.16
Girvanellaliths	0.20	-2.27	3.47	-0.01	5.35	-1.05	-0.21	-3.94
Encrustingalgae	0.58	-1.01	0.77	-0.35	0.25	1.66	-0.58	-1.28
Erectalgae	-0.02	0.28	-0.31	2.18	1.01	3.18	-10.79	4.24
Tuberatinids	0.00	1.39	-0.09	0.69	-1.21	0.03	-1.95	-0.22
Staffellids	1.63	1.37	1.61	-1.84	-1.08	1.53	2.45	-1.01
Millerellids	2.42	0.25	9.29	-4.69	-6.53	-2.34	-0.57	1.23
Endothyrids	1.89	0.26	-0.49	0.14	-0.42	1.41	0.83	-0.76
Pseudoendothyrids	0.25	12.81	-1.12	-2.92	0.94	0.39	0.01	-0.30
Biseriaminids	-0.40	-0.61	0.47	1.52	1.61	0.00	-4.91	-5.04
Bradyinids	0.15	1.45	0.64	5.31	-2.26	-2.69	2.40	-1.07
Tetrataxids	-0.56	0.60	0.45	4.81	-1.41	-2.03	-0.69	-0.05
Paleotextulariids	0.36	0.02	1.67	2.55	-0.71	-0.21	-0.34	-0.17
Lasiodiscids	-0.01	0.48	0.28	0.60	-0.40	-1.44	-1.41	-1.47
Archeodiscids	12.39	-0.18	-2.01	0.09	1.47	-2.67	0.01	0.42
Schubertellids	-0.50	1.31	1.09	7.73	-3.03	-4.24	2.10	-0.71
Irregularencrusters	0.58	0.37	4.19	1.09	4.67	2.65	5.46	11.85
Fusiformfusulinids	0.10	1.13	1.12	4.54	-2.31	-0.32	1.27	-0.22
Eigenvalues	**6.48**	**4.48**	**3.33**	**2.76**	**2.28**	**1.63**	**1.30**	**1.23**
% Explained	**27.57**	**19.06**	**14.19**	**11.77**	**9.69**	**6.96**	**5.52**	**5.25**
$\sum$(% Explained)	**27.57**	**46.63**	**60.82**	**72.59**	**82.28**	**89.24**	**94.76**	**100**

CHAPTER 3

INTERPRETATION OF ATTRIBUTES-BASED CLUSTERS

In addition to studying the discriminant function coefficient vectors, a graphical exploration of the data was implemented. Discriminant scores for each of the 809 variables were plotted, and Figure 3.1 shows the two-dimensional plots of the most important distinctions between groups in the first four dimensions. For example, a plot of discriminant scores between dimensions two and four shows that most of the observations sit together in one large indistinguishable cluster. However, a plot of dimensions 3 and 4 reveals some useful cluster distinctions. This is the reason why regions one and two, two and three, and three and four were plotted against each other. The meaning of the dimension interactions represented in Figure 3.1 is summarized as follows.

Dimensions 1 and 2

Clusters nine and four are most separated from the others in this perspective. Cluster nine contains much higher values in dimension one, whereas cluster four has high dimension two values. Closer inspection clearly reveals first dimension distinctions between clusters three, two, and seven. Second dimension separations are evident between clusters eight, two/three, and one.

Dimensions 2 and 3

This perspective introduces a separation between clusters seven (dimension three) and four (dimension two) from the remaining clusters. Cluster five, which only has two observations, also emerges in the third dimension. A deeper look does not profit much; clusters one, two,

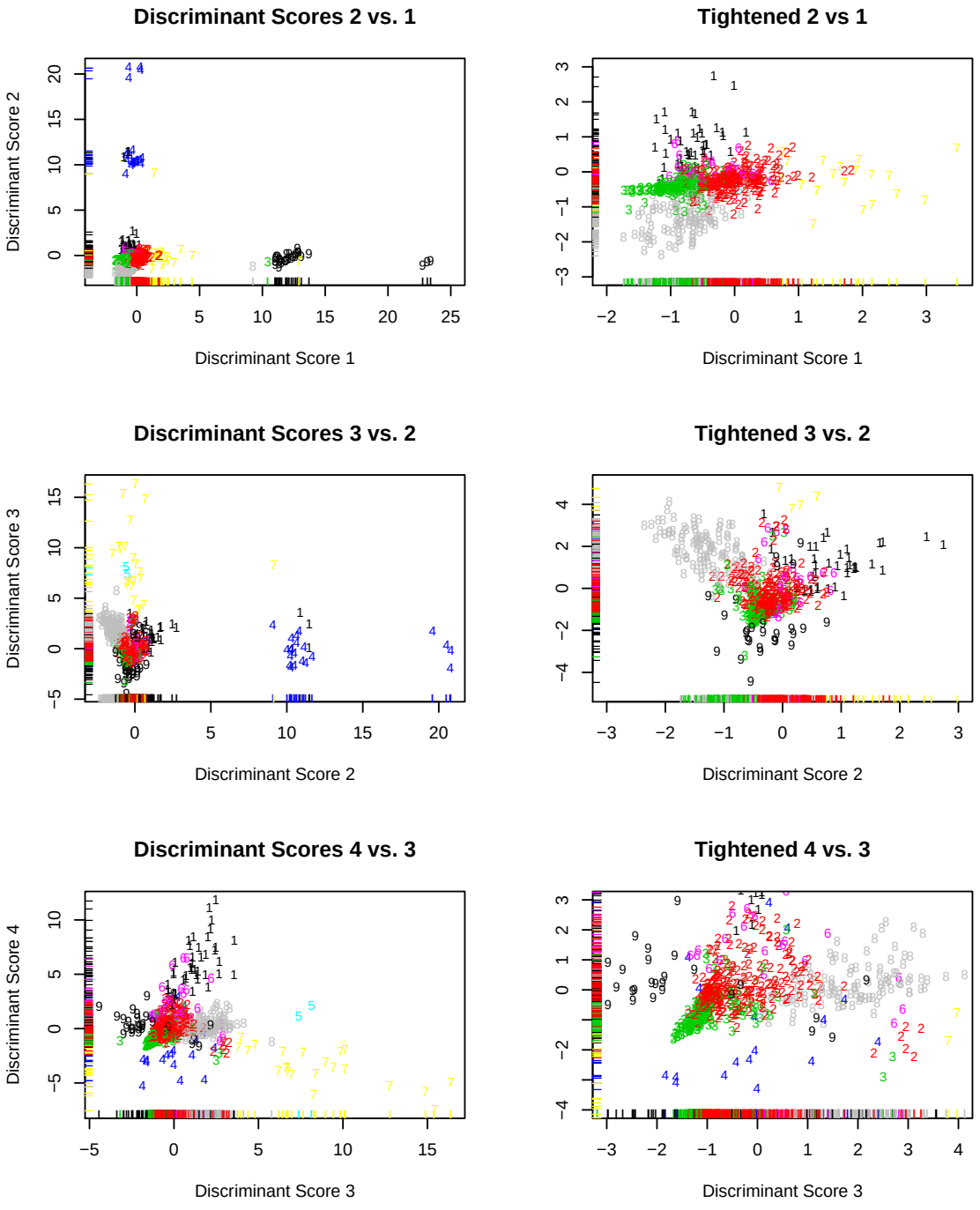

Figure 3.1: Discriminant scores plotted as 1^{st} vs. 2^{nd}, 2^{nd} vs. 3^{rd}, and 3^{rd} vs. 4^{th} dimensions plotted against each other (left side plots). The plots to the right are enlarged shots of the condensed clusterings.

three, six, and nine are nearly indistinguishable from each other, and cluster eight is slightly higher in the third dimension and lower in the second dimension.

Dimensions 3 and 4

Cluster five stands alone with substantial values in both dimensions. On an easily observed scale, clusters seven, five, four, and one are separated from the others. On a smaller scale, clusters two, three, six, eight, and nine are clearly separable as shown.

Interpretation

The paper by James (1997) helps in the interpretation of what the cluster memberships actually mean in context. We now briefly discuss the interpretation of the four dimensions represented in Figure 3.1.

- Dimension 1: Heterozoan/Photozoan Gradient. Samples with attributes that need light (shallower environments) are given higher scores. The dominant attributes found in this dimension are Articulate Brachiopods, Crinoidea, Osagialiths, Staffelids, Millerids, Endothyrids, and Archeodiscids.

- Dimension 2: Cool/Warm Water Gradient. Similar to the first dimension, samples with warmer water attributes are given higher scores. The dominant attributes in the second dimension are Bivalvia, Girvanellaliths, Encrusting Algae, Tuberatinids, Staffelids, Pseudoendothyrids, Bradyinids, Schubertellids, and Fusiform Fusulinids.

- Dimension 3: Nutrient Availability. Warmer and/or shallower water necessitates more nutrients for organisms to survive. The dominant attributes in the third dimension are Bivalvia, Gastropoda, Girvanellaliths, Millerellids, and Irregular Encrusters.

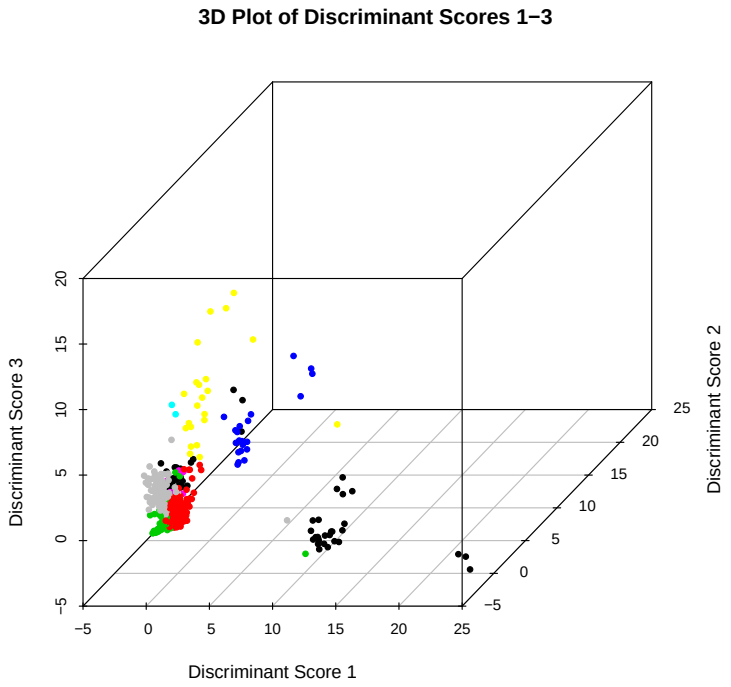

Figure 3.2: 3-D plot of first three dimensions.

- Dimension 4: Composite of many characteristics. The interpretation of this dimension becomes difficult, but it is believed that a combination of sea salinity with the previous dimensions' characteristics encompass this dimension. The dominant attributes in the fourth dimension are Millerellids, Bradyinids, Tetrataxids, Schubertellids, and Fusiformfusilinids.

Figure 3.2 shows a three-dimensional plot of the first three dimensions, which are the most important of the four dimensions we interpreted, where separation of clusters occurs. This figure clearly illustrates that clusters four (blue), seven (yellow), and nine (black) are most separated from each other and the other clusters.

The knowledge gained from a basic understanding of these dimensions is a springboard to a series of questions that can be addressed in the future.

CHAPTER 4

INCORPORATING STRATIGRAPHIC DISTANCE INFORMATION IN CLUSTER FORMATION

4.1 Outline of Methods

An understanding of the associations between clusters is merely the first step toward the much larger picture. Having established clusters and knowing the characteristics that determine the placement of samples into clusters, the next step is to look at the stratigraphic column and piece together a sequencing of warm/shallow to cold/deep oceanic climates. This sequencing of the observed stacking pattern may answer questions about climate and environmental variables during the time of deposition.

Up to this point, the distance metric used to compare distances between observations is purely a statistical function (in our case, Euclidean). This metric is represented by a 809×809 matrix denoted $\boldsymbol{D_b}$. Our research, as well as the studies referenced, relies on the premise that a sample's position within the stratigraphic column reveals information about the facies of the time period. Therefore, we hope to incorporate this information in our analyses. We will do this by constructing another 809×809 distance matrix denoted $\boldsymbol{D_d}$, which simply contains the stratigraphic column distances between observations.

A major component of our research is to determine the optimal combination of statistical (i.e., Euclidean) distances and position within the stratigraphic column. This allows us to reconstruct a sequencing of facies from bottom to top of the samples in this data set. This combination can be represented by

$$\tilde{\boldsymbol{D}} = a_b \boldsymbol{D_b} + a_d \boldsymbol{D_d}, \quad \text{where } a_b + a_d \equiv 1.$$

When $a_b = 1$, $\tilde{\boldsymbol{D}}$ is simply equal to $\boldsymbol{D_b}$, and as a_b decreases toward 0, $\tilde{\boldsymbol{D}}$ becomes increasingly a function of $\boldsymbol{D_d}$. In essence, we propose an "ideal" distance matrix with which to cluster the marine facies. We use the information about samples' assignments into a given cluster to reconstruct the relative rise and fall of sea-level during the Morrowan through Desmoinesian time period in the Arrow Canyon area.

4.2 Choice of Linkage Metric

From earlier results, we have determined to cluster our 809 observations into nine clusters. We compared the performance of several linkage metrics in forming clusters: average linkage, Ward's method, and complete linkage. In this composite ($\tilde{\mathbf{D}}$) situation, the k-means algorithm will not perform properly. K-means requires raw data values to recalculate clusters, but when we add $\mathbf{D_b}$ and $\mathbf{D_d}$, the raw information in the samples is incorporated into $\mathbf{D_b}$.

The use of average linkage resulted in one very large cluster with many other clusters containing only one observation. The benefit of average linkage is that misclassification rates are very low relative to other linkage metrics. However, the purpose of creating $\tilde{\mathbf{D}}$ is nullified since the resulting succession of clusters is even less interpretable than the $\mathbf{D_b}$ distance matrix when $\mathbf{a_d} = 1$.

While Ward's method produced a succession of clusters that appeared more desirable, the misclassification rates were far too high. The major improvements in the ability to interpret clusters was heavily offset by the significant increase in misclassification rates. Although the clusters made intuitive sense, more so than $\mathbf{D_b}$ by itself, the 30%+ misclassification rates indicate that the clusters were not definitively formed.

Complete linkage satisfied the motivation for $\tilde{\mathbf{D}}$. That is, analyses with complete linkage created a stratigraphic succession that was more smooth than the analysis of $\mathbf{D}_b$. Using suitable values of a_b and a_b also resulted in clusters with low misclassification rates. The choice of a_b and a_d is further discussed in Chapter 5.

CHAPTER 5

INTERPRETATION OF CLUSTERS BASED ON ATTRIBUTES AND STRATIGRAPHIC DISTANCE

We created 11 $\tilde{\mathbf{D}}$ matrices by sequentially increasing a_b by 0.1 increments, between zero and one. By default, this implies that a_d values were a sequence from one to zero by 0.1 increments. To determine the optimal values for a_b and a_d, we observed the misclassification rates for the corresponding $\tilde{\mathbf{D}}$ matrices. Figure 5.1 shows that a_b values between 0.8 and 1 give the largest drop in misclassification rates.

Considering a_b values between 0.8 and 0.9, we conclude that $a_b = 0.86$ seems to reduce misclassification rates the most, as shown in Figure 5.2. Consider that lower values

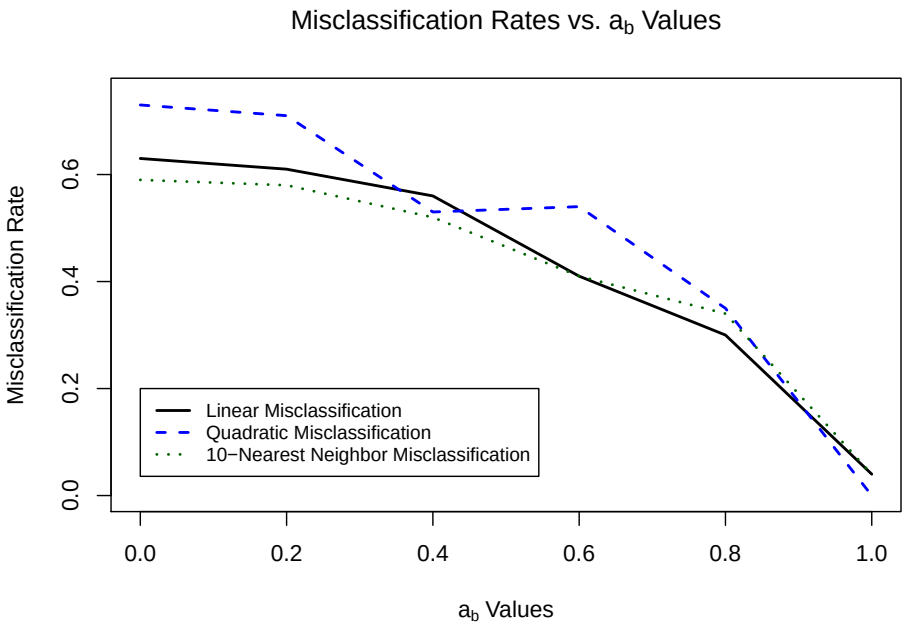

Figure 5.1: For a_b values between 0 and 1, a_b values between 0.8 and 1 give the greatest decrease in misclassification rates.

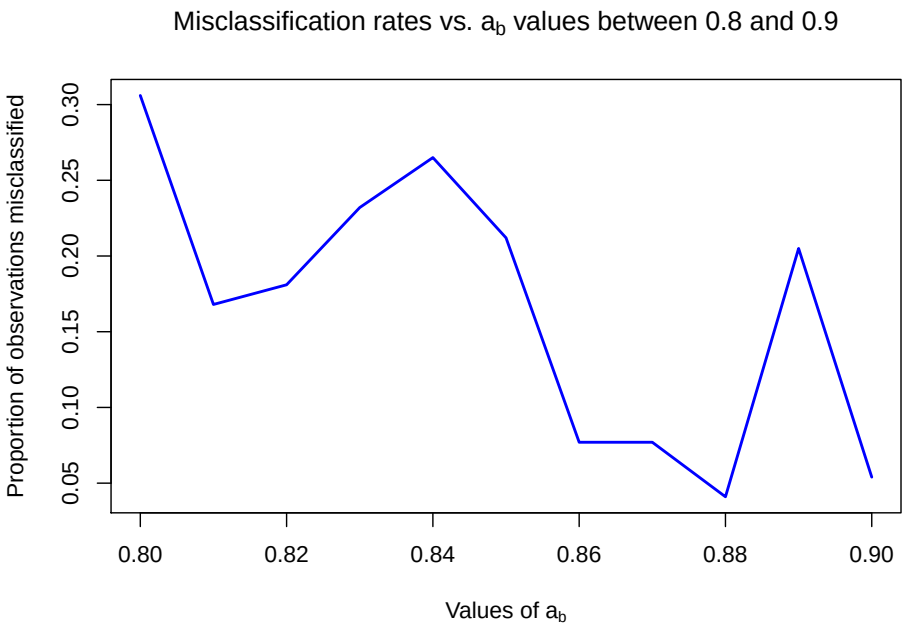

Figure 5.2: For a_b values between 0.8 and 0.9, $a_b = 0.86$ reduces misclassification rates to 7.7%. $\tilde{\mathbf{D}}$ as used for constructing sea-level sequencing is formed with $a_b = 0.86$.

of a_b offer greater "smoothness." Ideally, we want as low values of a_b as possible while still maintaining sufficient misclassification rates. Figure 5.3 shows the cluster memberships of the 809 samples when $a_b = 0$. Figure 5.4 shows the cluster memberships when $a_b = 1$.

Figures 5.3 and 5.4 illustrate how $a_b = 0$ creates a smooth sequencing of observations while $a_b = 1$ creates a choppy, sporadic sequencing. Figure 5.5 shows the sequencing when $a_b = 0.86$. The sequencing does not appear markedly different than when $a_b = 1$. However, since Figure 5.5 incorporates stratigraphic distance, this sequencing should be more intuitive and practical than when $a_b = 1$.

Although somewhat higher misclassification rates accompany smaller a_b values than $a_b = 0.86$, it is possible to see features in the data that may be glanced over when $a_b = 0.86$ is used. Figure 5.5 shows the succession of clusters when $a_b = 0.81$. Although the misclassification rate is approximately 10% higher, this plot shows a more intuitive sequencing with more than one large cluster. Even with carefully chosen values for a_b, it is still difficult to determine if, and how many, cycles occurred in these samples.

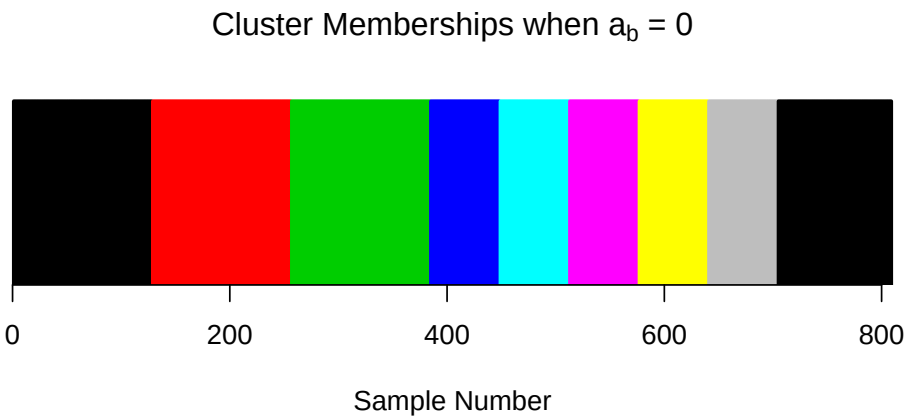

Figure 5.3: Cluster memberships of the 809 samples when $a_b = 0$. Each of the nine colors shown represents a unique cluster.

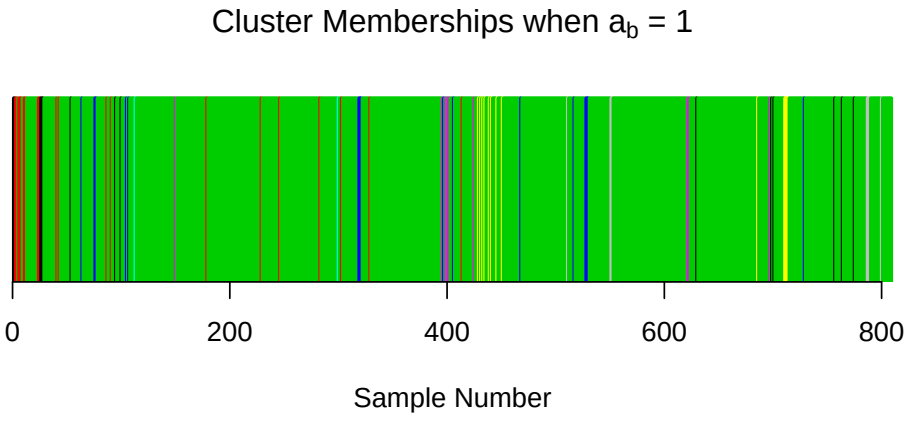

Figure 5.4: Cluster memberships of the 809 samples when $a_b = 1$.

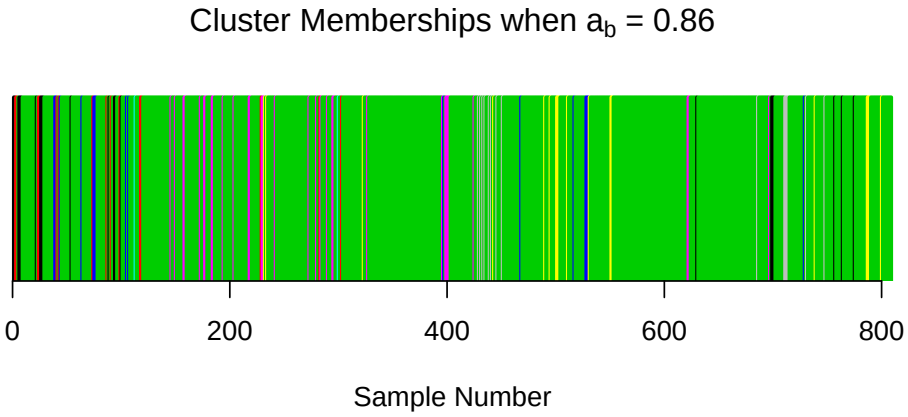

Figure 5.5: Cluster memberships of the 809 samples using optimal $\tilde{\mathbf{D}}$, where $a_b = 0.86$.

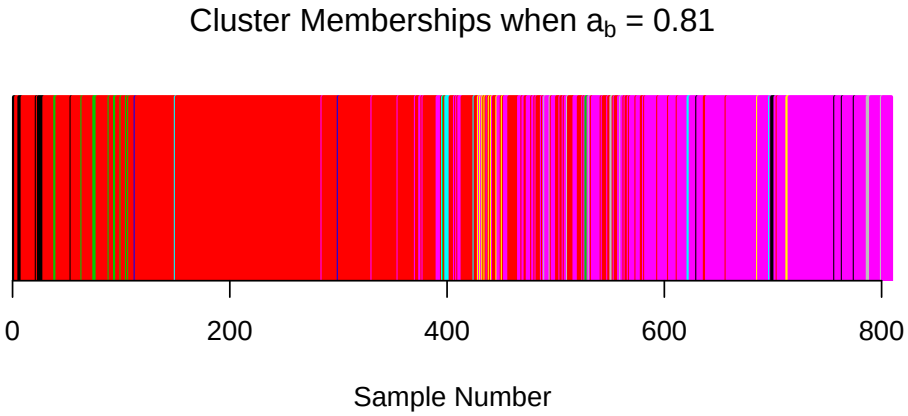

Figure 5.6: Cluster memberships of the 809 samples using $\tilde{\mathbf{D}}$, where $a_b = 0.81$.

CHAPTER 6

DISCUSSION OF RESULTS

Having established the distance matrix $\tilde{\mathbf{D}}$ where $a_b = 0.86$ and $a_d = 0.81$, each sample was assigned membership into one of the nine clusters. The interpretation of our results in large part relies on human assessment of the relationship between and among clusters (Ritter, 2010). The results from the analyses using $\tilde{\mathbf{D}}$ were not interpreted using discriminant scores and discriminant functions.

Table 6.1 was used to draw conclusions about cluster relationships (Ritter, 2010). The nine columns represent the nine clusters into which all observations were clustered. The 29 rows show all 29 fossil attributes. The values in the body of the table give the mean of all microfossil attributes within each cluster. To illustrate this, note that the mean of all Sponge Spicules in samples belonging to cluster three is 0.17.

Table 6.1 conveys that if a certain cluster's samples have high means in a deep-water attribute and low means in a shallow-water attribute, then that cluster represents deeper-dwelling organisms. We analyzed Table 6.1 and assigned relative depths to each of the nine clusters. Table 6.2 shows each of the nine clusters with their associated relative depths.

Having assigned each cluster a relative depth, each of the 809 samples was assigned a relative depth according to its cluster membership. Figure 6.1 was then produced. The vertical gaps are present simply because the only "response" variables in this scenario were integers one through eight.

A kernel smoother was used to draw the curve shown in Figure 6.1. A Normal distribution kernel with a bandwidth of 50 samples gives the most intuitive and meaningful sequencing of relative sea-level. By definition, the Normal distribution kernel is scaled such

Table 6.1: Rows give attribute names, columns give cluster number, and values in the table are mean values of a given (row) attribute within the (column) cluster.

	1	2	3	4	5	6	7	8	9
spongespicules	0.00	0.00	0.17	0.06	0.00	0.00	0.00	0.27	0.29
phosphaticgrains	0.00	0.20	0.30	0.06	0.00	0.16	0.17	0.33	0.29
articulatebrachiopods	1.07	1.27	1.46	1.12	1.50	1.47	1.39	1.73	2.14
bryozoaencrusting	0.00	0.07	0.04	0.00	0.00	0.00	0.22	0.00	0.00
bryozoafenestrateramose	0.43	0.67	0.73	0.62	0.00	1.03	0.89	0.67	2.29
crinoidea	1.57	1.73	1.69	1.56	1.50	1.75	1.89	2.13	3.00
echinoids	0.43	0.73	0.44	0.38	0.50	0.44	0.56	1.00	1.00
ostracoda	1.00	0.67	0.51	0.69	0.50	0.72	0.56	0.80	0.86
trilobita	0.21	0.27	0.31	0.44	0.50	0.34	0.83	0.67	0.29
bivalvia	0.36	0.53	0.19	0.31	0.50	0.06	0.44	0.60	0.43
gastropoda	0.29	0.40	0.17	0.38	0.50	0.06	1.06	0.60	0.14
osagialiths	0.00	0.07	0.03	0.06	0.50	0.03	1.33	0.20	0.00
girvanellaliths	0.71	0.47	0.43	1.00	2.50	0.03	2.33	1.53	0.00
encrustingalgae	0.14	0.27	0.14	0.19	0.00	0.25	0.28	0.33	0.14
erectalgae	0.21	0.13	0.08	0.25	0.00	1.50	0.00	0.40	1.14
tuberatinids	1.71	1.40	0.20	0.56	0.00	1.06	0.11	0.07	0.29
staffellids	0.71	0.40	0.41	0.94	0.00	0.09	0.56	2.13	1.57
millerellids	0.00	0.00	0.03	0.00	0.00	0.06	0.00	2.40	0.29
endothyrids	0.14	0.20	0.23	0.19	0.00	0.28	0.28	0.40	2.57
pseudoendothyrids	0.29	0.00	0.01	1.25	0.00	0.00	0.00	0.07	0.00
biseriaminids	1.79	1.33	0.35	0.94	0.50	1.12	0.50	0.80	1.00
bradyinids	0.36	1.13	0.02	0.19	0.00	0.06	0.00	0.00	0.00
tetrataxids	0.50	0.47	0.02	0.06	0.00	0.72	0.00	0.13	0.29
paleotextulariids	0.21	0.80	0.11	0.25	0.00	0.44	0.00	0.13	0.29
lasiodiscids	0.07	0.07	0.07	0.12	0.50	0.12	0.00	0.07	0.29
archeodiscids	0.00	0.00	0.03	0.00	0.00	0.00	0.17	0.00	1.43
schubertellids	2.36	0.73	0.02	0.19	0.00	0.19	0.00	0.00	0.00
irregularencrusters	2.43	1.27	0.48	1.38	2.00	1.16	0.89	1.40	0.43
fusiformfusulinids	1.00	1.67	0.17	0.44	0.00	1.00	0.00	0.07	0.00

that its quantiles are 12.5 ($\pm 50 \times 0.25$) locations above and below the location being calculated.

As Figure 6.1 clearly shows, most of the observations receive relative depth scores of five. Since most samples have been given the same depth scores, it is difficult to discern differences in sea-level without the aid of a smoother. This fact indicates that the data, as currently constituted, are quite noisy.

Table 6.2: Each of the nine clusters was assigned a relative depth. Lower relative depths correspond to shallower (or closer to shore) organisms. That is, depth 1 refers to shallowest relative depth, and depth 8 represents deepest relative depth.

	Cluster Number								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Relative Depths	4	3	5	5	1	8	2	6	7

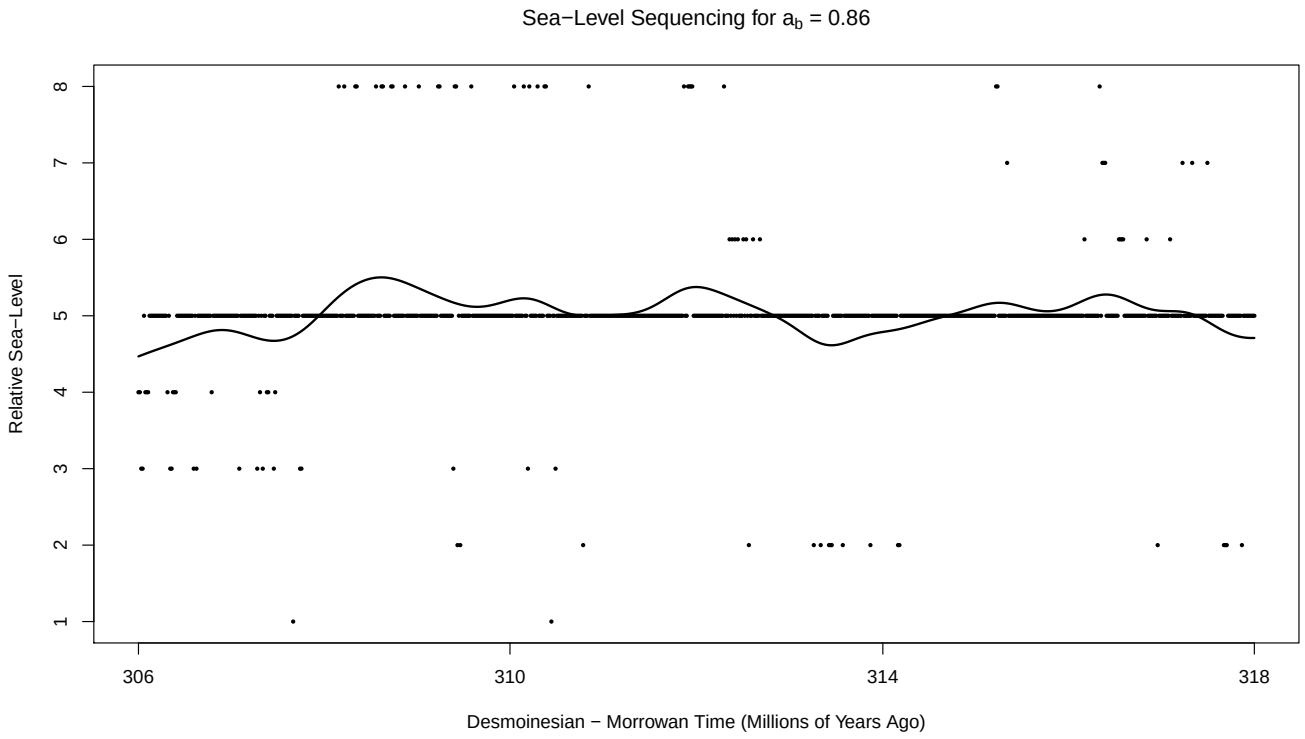

Figure 6.1: Relative sea-level sequencing of Arrow Canyon, NV, between 318–306 Mya for $a_b = 0.86$. Higher relative depth values correspond to higher sea-level, and lower values correspond to lower sea-level. The X-axis shows time before present (Ma), where Sample 1 occurs approximately 306 Ma and Sample 809 occurs 318 Ma.

The approximate eight million earlier years have less variability than the more recent four million years. The Permian period (approximately 290 to 250 Mya), which followed the Pennsylvanian, is known by geologists as the "Permian Icehouse." Icehouse conditions occur when ocean water is largely stored on continents as ice. Icehouse time periods are marked by large variability in sea-levels. Figure 6.1 and Figure 6.2 indicate that in addition to sea-level

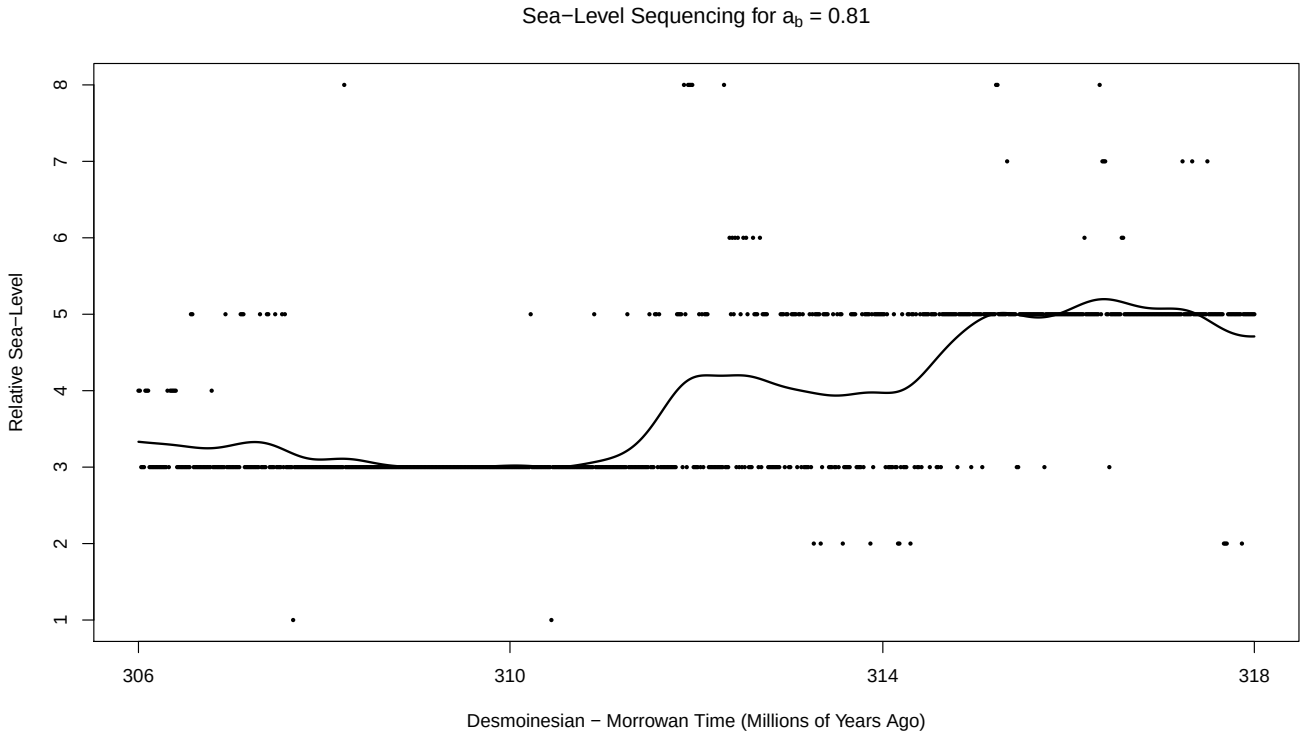

Figure 6.2: Relative sea-level sequencing of Arrow Canyon, NV, between 318–306 Mya for $a_b = 0.81$. Higher relative depth values correspond to higher sea-level, and lower values correspond to lower sea-level.

decreasing with time, the variability in these levels increases. The increasing variability in these figures indicates the onset of such an Icehouse state.

If a smaller value of a_b is used, imposing more smoothness in the distance matrix, the sea-level sequence can change dramatically. Figure 6.2 shows one such dramatic change. Earlier years are marked by higher sea-levels, whereas more recent years have lower sea-levels overall. A major shift in sea-level occurs between approximately 315 and 312 million years ago.

CHAPTER 7

CONCLUSION

Sea-level changed several times during the Morrowan through Desmoinesian time period in the Arrow Canyon, Nevada area. During those twelve million years, clear differences in both sea-level *and* the variability of sea-level changes are evident.

Constructing the sea-level sequencing for this particular data set is a nontrivial task for two primary reasons. First, this is not a problem easily solved by geologist or statistician working alone. Cross-disciplinary efforts are necessary in order to extract the most information possible from the data. Second, the interpretation of attribute associations and assigning accurate relative depths to the clusters formed is a highly subjective procedure. Judgmental errors in cluster associations and the assignment of relative depths may greatly influence the validity of our results.

Despite the potential shortcomings of this study, we have successfully integrated statistical tools and geological science to answer questions about the Arrow Canyon, Nevada area during Morrowan through Desmoinesian time. As time progressed, increasing variations in sea-level changes indicate the onset of the Permian Icehouse. Further considerations, such as incorporating grain-size information in analyses, may lead to additional useful results.

CHAPTER 8

FUTURE WORK

We have discussed cluster analysis, MANOVA, classification analysis, and discriminant analysis as tools for constructing sea-level sequences. Numerous other approaches may be used to accomplish this task.

Up to this point in our analyses, we have used biological and stratigraphic distances between samples to create clusters. In addition to these distance characteristics, the textures associated with each sample may be used to further define distinctions between samples. The texture of a sample refers to the type of stone (or texture) predominantly found in each sample. Such textures range from very muddy, fine samples (mudstones) to coarse, more grainy samples (grainstones). We believe that if textures can be effectively incorporated into the distance matrices used to create clusters, the signal-to-noise ratio in the data can be increased. This may improve the quality of a final analysis.

As shown in Figure 5.2, misclassification rates vary substantially with very small changes in a_b. We are currently unsure if this erratic behavior is due to important features in the data, or whether the clustering and classification algorithms are computationally unstable. We suggest sensitivity-type analyses, wherein values of $a_b = 0.808, 0.809, 0.810, 0.811, 0.812$ and $a_b = 0.858, 0.859, 0.060, 0.861, 0.862$ are run, to observe the behavior of misclassification rates. We need to further understand the volatility of the solutions in the small neighborhoods surrounding the potential values for a_b.

Our current approach relies on constructing sea-level as a function of eight unique depths. The results we currently use to plot sea-level treat each sample within a cluster as identical.

However, we note that in the larger clusters, there is actually a wide variety of attribute profiles, indicating that clustering may be over-simplifying the sea-level depth assessments. Although assessing the approximate sea-level at each of 809 samples is intractable, a promising approach is the use of the cluster-specific sea-level associated with each observation as only a preliminary estimate of depth. Regressing the preliminary sea-level on the matrix of attributes would yield a vector of partial slopes that would relate the attributes to sea-level. Using these slopes and the matrix of attributes, a predicted sea-level is then obtained.

Note that in this regression-based approach, the clusters now serve merely as a way to reduce the dimensionality of the attribute matrix to a small collection of profiles in the data. Although any small collection of attribute profiles (observed or constructed) could be used to elicit the attribute-to-sea-level relationships from the geologist, the regression analysis based on the cluster centroids (means) is optimal because the cluster centroids essentially span the data-inhabited subspace of the 29-dimensional attribute space. Thus, each sample will have a unique depth instead of having identical depths as other samples within the same cluster.

Finally, our analyses have only examined the relationships between clusters of *samples*. Some valuable insights may be gained from also examining the relationships between *attributes*. There may be important relationships between clusters of samples and clusters of attributes that are important, yet currently unknown. In addition to identifying these relationships, should they exist, this method may also aid in the creation and interpretation of sea-level during Pennsylvanian geologic time.

BIBLIOGRAPHY

Briggs, K. P. (2005), *Establishing a High-Frequency Standard Reference Sequence Stratigraphy, Sea-Level Curve, and Biostratigraphy for Morrowan Strata of the Lower Absaroka I Time Slice Based upon the Bird Spring Formation, Arrow Canyon, Nevada*, Brigham Young University.

James, N. P. (1997), "The Cool-Water Carbonate Depositional Realm," *Society for Sedimentary Geology*, 56, 1–22.

Ritter, S. M. (2010), Personal Communication.

Spence, G. H., Arnaud-Vanneau, A., Arnaud, H., and Tucker, M. E. (2004), "Unravelling the Microfacies Signatures of Parasequences Using Computer-Optimized Similarity Matrices," *Sedimentology*, 51, 1243–1271.

Spence, G. H. and Tucker, M. E. (2007), "A Proposed Integrated Multi-Signature Model for Peritidal Cycles in Carbonates," *Journal of Sedimentary Research*, 77, 797–808.

APPENDICES

APPENDIX

A.1 R Code

```
# Read in SPEC data
library(xtable)
data <- read.table("Project_data.txt",header=T,sep="")
n <- dim(data)[1]
p <- dim(data)[2]
X <- as.matrix(data)
# There are erroneous values that need to be fixed
data[413,16] <- 0 # instead of 7
data[113,28] <- 2 # instead of 23
data[300,28] <- 2 # instead of 20

##################
### Clustering ###
##################
# No need to standardize data - all on same scale, but scaled nonetheless
X.std <- apply(X,2,scale)

# Distance matrices for data (bio) and distance (dist) are standardized to
# make each matrix have mean of 1

# Create distance matrix based solely on data
```

```
#dist.bio <- as.matrix(dist(X.std))
dist.bio <- as.matrix(dist(X.std)) / mean(as.matrix(dist(X.std)))

# Create distance matrix (distance apart in stratigraphic column)
#dist.dist <- as.matrix(dist(t(t(1:809))))
dist.dist <- as.matrix(dist(t(t(1:809)))) / mean(as.matrix(dist(t(t(1:809)))))

# Now create distance matrix as combination of two dist. matrices
abio <- .86
adist <- 1-abio
distance <- as.dist(abio*dist.bio + adist*dist.dist)

# Now do clustering, classification analysis on "distance"
complink <- hclust(distance,method='complete')

# At this point, do not use kmeans since kmeans requires data in order
# to properly function. From the results before (666 project and prospectus),
# we know that 9 groups is optimal in terms of using only data, in terms
# of misclassification rates.
ab10 <- cutree(complink,9)

# Plot vertical succession of clusters
#par(mfrow=c(11,1))
#par(mar=c(0,0,0,0))
#plot(rep(1,809),1:809,col=ad1,pch=95,lwd=1)
plot(1:809,rep(1,809),type='h',col=ab10,lwd=2,ylim=c(.5,1),axes=FALSE)
```

```
# We study the misclassification rates of the many 'distance' matrices in SAS.
# We will choose both the combination of distance matrix which looks believable
# heuristically, as well as that which minimizes misclassification.

# Write to file to analyze with SAS
data.out <- cbind(X,ab0,ab1,ab2,ab3,ab4,ab5,ab6,ab7,ab8,ab9,ab10)
write.table(data.out,file='AbData.txt',row.names=F,col.names=F)

#####################################
### See SAS Code, then code below ###
#####################################

#############################################################################
### Get Discriminant Score info. from SAS output (for discriminant approach ###
#############################################################################
lam.vals <- c(.8181,.3379,.1068,.0427,.0255,.0143,.0075,.0036)
pval <- rep("<.0001",8)
table1 <- cbind(lam.vals,pval,lmis,qmis,nmis)
xtable(table1)

# We choose k=9 groups because it has lowest overall misclassification,
# and lowest of the linear misclassifications

# There 8 nonzero eigenvalues => hard to interpret. 6 gives about 90%,
# and we interpret the first two dimensions. This information comes out
# in Table 2 of the document
```

```
evals <- c(6.47704464,4.47859203,3.33291477,2.76456368,
   2.27630179,1.63466139,1.29720735,1.23376702)

evec1 <- c(-0.00408253,0.00356323,0.00461044,0.01426292,0.00155748,
0.00510302,-0.00055257,0.00542694,0.00334446,-0.00431526,-0.00067382,
-0.01832886,0.00069469,0.00402495,-0.00015885,0.00001513,0.00538865,
0.03296040,0.01084000,0.00738912,-0.00162020,0.00218393,-0.00683798,
0.00281649,-0.00007732,0.40813672,-0.00448798,0.00161305,0.00050805)

evec2 <- c(-0.00020110,-0.00191883,0.00150949,-0.00293755,0.00082593,
0.00139271,-0.00282061,0.00182771,0.00033219,-0.00884332,-0.00403782,
-0.00324842,-0.00784091,-0.00698241,0.00215624,0.00577848,0.00455009,
0.00335024,0.00151987,0.37783746,-0.00248846,0.02086287,0.00738100,
0.00011869,0.00433625,-0.00577112,0.01173105,0.00103771,0.00574358)

evec3 <- c(-0.00282530,0.00027749,0.00140362,-0.00561666,0.00040257,
0.00138048,0.00568779,0.00539246,0.00066767,0.01826271,0.02829649,
0.01179981,0.01196854,0.00534503,-0.00238235,-0.00037703,0.00535092,
0.12673382,-0.00283622,-0.03290954,0.00191474,0.00926392,0.00549231,
0.01311749,0.00253836,-0.06628833,0.00973708,0.01167058,0.00571604)

evec4 <- c(-0.00257674,0.00267555,0.00708951,-0.02249448,0.00164732,
0.00390444,0.00905756,0.00121311,0.00797951,0.00502745,0.00294850,
0.00853737,-0.00002779,-0.00245294,0.01680206,0.00287565,-0.00610017,
-0.06391467,0.00078708,-0.08616510,0.00620333,0.07631979,0.05898440,
0.02004980,0.00545155,0.00290456,0.06897862,0.00302788,0.02311582)
```

```
evec5 <- c(-0.00303194,0.00112718,-0.00550908,-0.00597627,-0.00654615,
0.00107593,0.00123808,-0.00080553,-0.00645779,0.02147909,0.02944741,
0.01645087,0.01846178,0.00173306,0.00774493,-0.00503522,-0.00358482,
-0.08906186,-0.00240523,0.02775912,0.00657862,-0.03250500,-0.01733076,
-0.00561248,-0.00367348,0.04831343,-0.02708111,0.01299630,-0.01179337)

evec6 <- c(-0.00646814,0.01168835,0.01225496,0.02939249,0.01785008,
0.01617235,0.01049490,0.00771885,0.01673458,-0.00499205,-0.00821348,
-0.01140464,-0.00362495,0.01148456,0.02443749,0.00013511,0.00506656,
-0.03190773,0.00808299,0.01147120,0.00001817,-0.03869384,-0.02485477,
-0.00167988,-0.01304322,-0.08797051,-0.03780286,0.00738605,-0.00162205)

evec7 <- c(-0.00133048,0.00156407,0.00380155,0.00861617,0.00628716,
0.00066246,0.00392860,-0.00004977,0.00446100,-0.00222002,0.00559336,
-0.01629610,-0.00072266,-0.00403524,-0.08294829,-0.00811006,0.00810538,
-0.00778824,0.00474131,0.00018696,-0.02000035,0.03444071,-0.00846098,
-0.00270545,-0.01280839,0.00046359,0.01871532,0.01519687,0.00647006)

evec8 <- c(0.00061901,-0.00537356,-0.00146703,-0.02035594,-0.00471283,
-0.00062819,-0.01104918,-0.00996409,-0.00470915,-0.01143445,-0.02028007,
-0.00153337,-0.01357417,-0.00883001,0.03259597,-0.00092549,-0.00335755,
0.01682444,-0.00436800,-0.00894192,-0.02053146,-0.01540914,-0.00057150,
-0.00134125,-0.01332032,0.01368329,-0.00634824,0.03299657,-0.00110480)

# Sdiags refers to the estimated Variance (E matrix from SAS)
# The following standardizes the vectors to give "astar"
Sdiags <- c(277.24588394,196.57001458,611.44791963,28.979940652,
```

```
456.04276528,643.82108223,186.36276687,199.90335459,174.62459062,
182.87327033,86.175327844,65.939278098,511.74491405,126.98328743,
102.88214483,351.27350157,553.73002798,32.722630261,184.52913226,
6.9926406926,366.52133896,29.442044911,40.538448868,98.717864481,
73.715573151,5.6086816134,76.389707355,784.90347579,234.42079897)

d1 <- (sqrt(Sdiags)/(n-p))*evec1
d2 <- (sqrt(Sdiags)/(n-p))*evec2
d3 <- (sqrt(Sdiags)/(n-p))*evec3
d4 <- (sqrt(Sdiags)/(n-p))*evec4
d5 <- (sqrt(Sdiags)/(n-p))*evec5
d6 <- (sqrt(Sdiags)/(n-p))*evec6
d7 <- (sqrt(Sdiags)/(n-p))*evec7
d8 <- (sqrt(Sdiags)/(n-p))*evec8

discrim.coefs <- rbind(d1,d2,d3,d4,d5,d6,d7,d8)
colnames(discrim.coefs) <- names(data)

easy.discrim <- round(discrim.coefs*10000,digits=4)

xtable(rbind(t(easy.discrim),evals))

################################################
### Plot Discriminant Scores - Earlier Work ###
################################################
# Plot comparisons of first four - that accounts for nearly 3/4 variability
par(mar=c(5,4,4,2)+.1)
```

```
dscores <- read.table("dscores.txt")
names(dscores) <- paste("DS",1:8,sep="")

library(scatterplot3d)
pdf(file="disc3d.pdf")
scatterplot3d(dscores$DS1,dscores$DS2,dscores$DS3,color=km9,pch=20,angle=45,
  xlab="Discriminant Score 1",ylab="Discriminant Score 2",
  zlab="Discriminant Score 3",main="3D Plot of Discriminant Scores 1-3")
dev.off()

# Plot 1 vs 2, whole thing, then zoomed in on unclear area
pdf(file="ds_plot.pdf",width=6.5,height=8)
par(mfrow=c(3,2))

plot(dscores$DS1,dscores$DS2,col=km9,pch=(48+km9),cex=.75,
xlab="Discriminant Score 1",ylab="Discriminant Score 2",
main="Discriminant Scores 2 vs. 1",xlim=c(-3,25))
points(dscores$DS1,rep(-3.3,809),col=km9,pch=124)
points(rep(-3.9,809),dscores$DS2,col=km9,pch=95)

plot(dscores$DS1,dscores$DS2,col=km9,pch=(48+km9),cex=.75,
xlim=c(-2,3.5),ylim=c(-3,3),xlab="Discriminant Score 1",
ylab="",main="Tightened 2 vs 1")
points(dscores$DS1,rep(-3.25,809),col=km9,pch=124)
points(rep(-2.2,809),dscores$DS2,col=km9,pch=95)

# Plot 2 vs 3
```

```
plot(dscores$DS2,dscores$DS3,col=km9,pch=(48+km9),cex=.75,
xlab="Discriminant Score 2",ylab="Discriminant Score 3",
main="Discriminant Scores 3 vs. 2")
points(dscores$DS2,rep(-5.3,809),col=km9,pch=124)
points(rep(-3.1,809),dscores$DS3,col=km9,pch=95)

plot(dscores$DS2,dscores$DS3,col=km9,pch=(48+km9),cex=.75,
xlim=c(-3,3),ylim=c(-5,5),xlab="Discriminant Score 2",
ylab="",main="Tightened 3 vs. 2")
points(dscores$DS1,rep(-5.45,809),col=km9,pch=124)
points(rep(-3.2,809),dscores$DS3,col=km9,pch=95)

# Plot 3 vs 4
plot(dscores$DS3,dscores$DS4,col=km9,pch=(48+km9),cex=.75,
xlab="Discriminant Score 3",ylab="Discriminant Score 4",
main="Discriminant Scores 4 vs. 3")
points(dscores$DS3,rep(-8.2,809),col=km9,pch=124)
points(rep(-5.1,809),dscores$DS4,col=km9,pch=95)

plot(dscores$DS3,dscores$DS4,col=km9,pch=(48+km9),cex=.75,xlim=c(-3,4),
ylim=c(-4,3),xlab="Discriminant Score 3",ylab="",main="Tightened 4 vs. 3")
points(dscores$DS3,rep(-4.25,809),col=km9,pch=124)
points(rep(-3.2,809),dscores$DS4,col=km9,pch=95)

dev.off()

##############################
```

```
### Plots for Presentation ###
##############################
# Create "Hypothetical Plot" of Kristen's Results
ts.sim <- arima.sim(list(order = c(1,1,0), ar = 0.7), n = 200)
pdf(file="pplot2.pdf",height=5)
ts.plot(ts.sim,col="blue",lwd=2,ylab="Ocean Depth",main="Ocean Depth vs. Time")
dev.off()

pdf(file="onetwo.pdf",width=10,height=5)
par(mfrow=c(1,2))
plot(dscores$DS1,dscores$DS2,col=km9,pch=(48+km9),cex=.75,xlab=
"Discriminant Score 1",ylab="Discriminant Score 2",main=
"Discriminant Scores 2 vs. 1",xlim=c(-3,25))
points(dscores$DS1,rep(-3.3,809),col=km9,pch=124)
points(rep(-3.9,809),dscores$DS2,col=km9,pch=95)

plot(dscores$DS1,dscores$DS2,col=km9,pch=(48+km9),cex=.75,xlim=
c(-2,3.5),ylim=c(-3,3),xlab="Discriminant Score 1",ylab="",
main="Tightened 2 vs 1")
points(dscores$DS1,rep(-3.25,809),col=km9,pch=124)
points(rep(-2.2,809),dscores$DS2,col=km9,pch=95)
dev.off()

pdf(file="threefour.pdf",width=10,height=5)
par(mfrow=c(1,2))
plot(dscores$DS3,dscores$DS4,col=km9,pch=(48+km9),cex=.75,
xlab="Discriminant Score 3",ylab="Discriminant Score 4",main=
```

```
"Discriminant Scores 4 vs. 3")
points(dscores$DS3,rep(-8.2,809),col=km9,pch=124)
points(rep(-5.1,809),dscores$DS4,col=km9,pch=95)

plot(dscores$DS3,dscores$DS4,col=km9,pch=(48+km9),cex=.75,xlim=c(-3,4),
ylim=c(-4,3),xlab="Discriminant Score 3",ylab="",main="Tightened 4 vs. 3")
points(dscores$DS3,rep(-4.25,809),col=km9,pch=124)
points(rep(-3.2,809),dscores$DS4,col=km9,pch=95)
dev.off()

################################################
### Misclassification Stuff for Dtilde matrix ###
################################################
# Plot of misclassification rates vs. Ab values
lmis <- c(.63,.61,.56,.41,.3,.04)
qmis <- c(.73,.71,.53,.54,.35,0)
nmis <- c(.59,.58,.52,.41,.34,.04)
par(mar=c(5,4,4,2)+.1)
#pdf(file="MisSpecPlot.pdf",height=5)
plot(c(0,.2,.4,.6,.8,1),lmis,type="l",lwd=2,ylim=c(0,.75),xlab=expression(
paste(a[b]," Values",sep="")),ylab="Misclassification Rate",main=expression(
paste("Misclassification Rates vs. ",a[b]," Values",sep="")))
lines(c(0,.2,.4,.6,.8,1),qmis,type="l",lty=2,lwd=2,col="blue")
lines(c(0,.2,.4,.6,.8,1),nmis,type="l",lty=3,lwd=2,col="darkgreen")
legend(0,.2,col=c("black","blue","darkgreen"),lwd=c(2,2,2),lty=c(1,2,3),cex=.8,
c("Linear Misclassification","Quadratic Misclassification","10-Nearest Neighbor
Misclassification"))
```

```
#dev.off()

# Linear misclassification rates on finer grid scale
misclass.vals <- c(.63,.62,.61,.56,.56,.51,.41,.43,.31,.05,.05)
plot(seq(0,1,by=.1),misclass.vals,type="l")
points(seq(0,1,by=.1),misclass.vals,col='red',lwd=5)

#########################################################
### Now, hone in on values between .8 and .9 - see ###
### where misclassification really drops off       ###
#########################################################
abio <- 0
adist <- 1-abio
distance <- as.dist(abio*dist.bio + adist*dist.dist)
complink <- hclust(distance,method='complete')
ab0 <- cutree(complink,9)
pdf(file="ab0.pdf",height=3)
plot(1:809,rep(1,809),type='h',col=ab86,lwd=2,ylim=c(.5,1),axes=FALSE,
 xlab="Sample Number",ylab="",main=expression(paste(
 "Cluster Memberships when ",a[b]," = 0",sep="")))
axis(1)
dev.off()

abio <- 1
adist <- 1-abio
distance <- as.dist(abio*dist.bio + adist*dist.dist)
complink <- hclust(distance,method='complete')
```

```
ab1 <- cutree(complink,9)
pdf(file="ab1.pdf",height=3)
plot(1:809,rep(1,809),type='h',col=ab86,lwd=2,ylim=c(.5,1),axes=FALSE,
 xlab="Sample Number",ylab="",main=expression(paste(
 "Cluster Memberships when ",a[b]," = 1",sep="")))
axis(1)
dev.off()

abio <- .81
adist <- 1-abio
distance <- as.dist(abio*dist.bio + adist*dist.dist)
complink <- hclust(distance,method='complete')
ab81 <- cutree(complink,9)
pdf(file="ab81.pdf",height=3)
plot(1:809,rep(1,809),type='h',col=ab81,lwd=2,ylim=c(.5,1),axes=FALSE,
 xlab="Sample Number",ylab="",main=expression(paste(
 "Cluster Memberships when ",a[b]," = 0.81",sep="")))
axis(1)
dev.off()

abio <- .86
adist <- 1-abio
distance <- as.dist(abio*dist.bio + adist*dist.dist)
complink <- hclust(distance,method='complete')
ab86 <- cutree(complink,9)
pdf(file="ab86.pdf",height=3)
plot(1:809,rep(1,809),type='h',col=ab86,lwd=2,ylim=c(.5,1),axes=FALSE,
```

```
 xlab="Sample Number",ylab="",main=expression(paste(
 "Cluster Memberships when ",a[b]," = 0.86",sep="")))
axis(1)
dev.off()

ab86.data <- cbind(X,ab86)
write.table(ab86.data,file='ab86data.txt',row.names=F,col.names=F)

# Looks like .86 gives 7.7% misclassification - use this value since
# we want to "smooth" as much as is reasonable
zoom.vals <- c(.306,.168,.181,.232,.265,.212,.077,.077,.041,.205,.054)

pdf(file="zoomed.pdf",height=5)
plot(seq(.80,.90,by=.01),zoom.vals,type="l",lwd=2,col="blue",ylab=
"Proportion of observations misclassified",xlab=expression(
paste("Values of ",a[b],sep="")),main=expression(paste(
 "Misclassification rates vs. ",a[b]," values between .8 and .9",sep="")))
dev.off()

#########################################################################
### For Dr. Ritter, get mean of each attribute value within a cluster ###
### for him to determine which clusters represent which sea-level     ###
#########################################################################
names(ab86) <- NULL
sum(ab86==1) # 15
c1 <- data[ab86==1,]
mc1 <- as.matrix(apply(c1,2,mean))
```

```
sum(ab86==2) # 690
c2 <- data[ab86==2,]
mc2 <- as.matrix(apply(c2,2,mean))

sum(ab86==3) # 16
c3 <- data[ab86==3,]
mc3 <- as.matrix(apply(c3,2,mean))

sum(ab86==4) # 2
c4 <- data[ab86==4,]
mc4 <- as.matrix(apply(c4,2,mean))

sum(ab86==5) # 2
c5 <- data[ab86==5,]
mc5 <- as.matrix(apply(c5,2,mean))

sum(ab86==6) # 32
c6 <- data[ab86==6,]
mc6 <- as.matrix(apply(c6,2,mean))

sum(ab86==7) # 18
c7 <- data[ab86==7,]
mc7 <- as.matrix(apply(c7,2,mean))

sum(ab86==8) # 15
c8 <- data[ab86==8,]
```

```
mc8 <- as.matrix(apply(c8,2,mean))


sum(ab86==9) # 7
c9 <- data[ab86==9,]
mc9 <- as.matrix(apply(c9,2,mean))


observations <- 1:809
length(observations[ab86==1])
length(observations[ab86==2])
length(observations[ab86==3])
length(observations[ab86==4])
length(observations[ab86==5])
length(observations[ab86==6])
length(observations[ab86==7])
length(observations[ab86==8])
length(observations[ab86==9])


observations[ab86==1]
observations[ab86==2]
observations[ab86==3]
observations[ab86==4]
observations[ab86==5]
observations[ab86==6]
observations[ab86==7]
observations[ab86==8]
observations[ab86==9]
```

```
results.mat <- cbind(mc1,mc2,mc3,mc4,mc5,mc6,mc7,mc8,mc9)
xtable(results.mat)

# Ran through above algorithm with ab86 (as ab81). Once results.mat is
# formed, save it below as ab81.mat. Then run through and do same analysis
# with actual ab86. Then we compute correlations to find which of the new
# ab81.mat values are deepest, to shallowest, by the definitions of which are
# deepest to shallowest in ab86 results.mat
ab81.mat <- results.mat

# Cluster 1 stays as cluster 1
# Cluster 2 stays as cluster 2
# Cluster 4 became cluster 3
# Cluster 5 became cluster 4
# Cluster 6 became cluster 5
# Cluster 3 became cluster 6
# Cluster 8 became cluster 7
# Cluster 7 became cluster 8
# Cluster 9 stays as cluster 9

corrs <- NULL
for(i in 1:9){
corrs[i] <- cor(ab81.mat[,6],results.mat[,i])
}

# Get Sea-Level Heights - for ab81
heights81 <- function(Clusters){
```

```
Depth <- NULL
for(i in 1:809){
if(Clusters[i]==1){Depth[i] <- 4}
if(Clusters[i]==2){Depth[i] <- 3}
if(Clusters[i]==3){Depth[i] <- 5}
if(Clusters[i]==4){Depth[i] <- 1}
if(Clusters[i]==5){Depth[i] <- 8}
if(Clusters[i]==6){Depth[i] <- 5}
if(Clusters[i]==7){Depth[i] <- 6}
if(Clusters[i]==8){Depth[i] <- 2}
if(Clusters[i]==9){Depth[i] <- 7}
}
return(Depth)
}

# Get Sea-Level Heights - for ab86
heights86 <- function(Clusters){
Depth <- NULL
for(i in 1:809){
if(Clusters[i]==1){Depth[i] <- 4}
if(Clusters[i]==2){Depth[i] <- 3}
if(Clusters[i]==3){Depth[i] <- 5}
if(Clusters[i]==4){Depth[i] <- 5}
if(Clusters[i]==5){Depth[i] <- 1}
if(Clusters[i]==6){Depth[i] <- 8}
if(Clusters[i]==7){Depth[i] <- 2}
if(Clusters[i]==8){Depth[i] <- 6}
```

```
if(Clusters[i]==9){Depth[i] <- 7}
}
return(Depth)
}


# Kernel Smoother and plot
depths <- heights81(ab81)
time <- 1:809
kmod50 <- ksmooth(time,depths,kernel="normal",bandwidth=50)
plot(time,depths,pch=20,cex=.25)
lines(time,kmod50$y)

#################################################################
### Create sea-level sequencing plots - for ab81 and ab86 ###
#################################################################
pdf(file="sequencing81.pdf",width=12)
plot(time,depths,pch=20,lwd=2,cex=.35,xlab="Desmoinesian - Morrowan Time
 (Millions of Years Ago)",ylab="Relative Sea-Level",main=expression(paste(
 "Sea-Level Sequencing for ",a[b]," = .81",sep="")),axes=F)
lines(time,kmod50$y,lwd=2)
box()
axis(2)
axis(1,at=c(1,270,540,809),labels=c('306','310','314','318'))
dev.off()

pdf(file="sequencing86.pdf",width=12)
```

```
plot(time,depths,pch=20,lwd=2,cex=.35,xlab="Desmoinesian - Morrowan Time
 (Millions of Years Ago)",ylab="Relative Sea-Level",main=expression(paste(
 "Sea-Level Sequencing for ",a[b]," = .86",sep="")),axes=F)
lines(time,kmod50$y,lwd=2)
box()
axis(2)
axis(1,at=c(1,270,540,809),labels=c('306','310','314','318'))
dev.off()
```

A.2 SAS Code

```
OPTIONS ls=256 formdlim="#";
FILENAME specdata 'C:\Documents and Settings\Administrator\
    Desktop\Master_Project\AbData.txt';

* ############################################################;
* ### s1 stands for "Specimen 1," and so forth              ###;
* ### This code used for misclassification rate comparisons ###;
* ############################################################;
DATA spec;
  INFILE specdata;
  INPUT s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15
        s16 s17 s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29
        ab0 ab1 ab2 ab3 ab4 ab5 ab6 ab7 ab8 ab9 ab10;
RUN;

* ############################################################;
```

```
* ### The following chunk of code helps create R plot that  ###;
* ### determines we will use linear classification function ###;
* ############################################################;

TITLE 'Determine to use 9 clusters';
PROC DISCRIM data=spec method=normal pool=yes crosslisterr manova;
CLASS ab6;
VAR s1-s29;
PRIORS proportional;
RUN;

PROC DISCRIM data=spec method=normal pool=no crosslisterr manova;
CLASS ab6;
VAR s1-s29;
PRIORS proportional;
RUN;

PROC DISCRIM data=spec method=npar k=10 crosslisterr manova;
CLASS ab6;
VAR s1-s29;
PRIORS proportional;
RUN;

* ############################################################;
* ### Once we determined that lowest misclassification      ###;
* ### occurs between ab=.8,.9, we feed in new zoomed data   ###;
* ### and determine which value in between .8 and .9 to use ###;
```

```
* ################################################################;
FILENAME zoomdata 'C:\Documents and Settings\Administrator\
    Desktop\Master_Project\AbZoomData.txt';
DATA zoomspec;
  INFILE zoomdata;
  INPUT s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15
        s16 s17 s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29
        ab80 ab81 ab82 ab83 ab84 ab85 ab86 ab87 ab88 ab89 ab90;
RUN;

PROC DISCRIM data=zoomspec method=normal pool=yes crosslisterr manova;
CLASS ab86;
VAR s1-s29;
PRIORS proportional;
RUN;

* ################################################################;
* ### Following code outputs discriminant scores for plotting ###;
* ################################################################;
FILENAME ab86data 'C:\Documents and Settings\Administrator\
    Desktop\Master_Project\ab86data.txt';
DATA ab86data;
  INFILE ab86data;
  INPUT s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15
        s16 s17 s18 s19 s20 s21 s22 s23 s24 s25 s26 s27 s28 s29
        ab86;
RUN;
```

```
TITLE 'Supervised Classification Analysis: Ab=.86';
PROC GLM data=zoomspec;
CLASS ab86;
MODEL s1-s29 = ab86;
MANOVA H=ab86/PRINTE PRINTH;
RUN;
QUIT;

PROC CANDISC data=spec anova out=outputdata;
  CLASS km9;
  VAR s1-s29;
RUN;

DATA _NULL_;
  SET outputdata;
  FILE  'C:\Documents and Settings\Administrator\Desktop\dscores.txt';
  PUT can1 can2 can3 can4 can5 can6 can7 can8;
RUN;
```