USING TIME SERIES AND MULTIVARIATE TECHNIQUES TO ANALYZE CHANGES IN THE OCEAN NITROGEN CYCLE THROUGHOUT HISTORY

by

David Fay

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
August 2014

Table of Contents

List of Figures

Abstract

Nitrogen is a limiting resource in marine ecosystems that directly impacts the productivity of marine life. Values of $\delta^{15}N$ extracted from down core sediments are used as a proxy measure of nitrogen in past marine environments. We analyzed historic $\delta^{15}N$ records from around the world, covering time periods that ranged 125,000 to 5000 years ago. The Kalman smoother was used to extract the true signal of $\delta^{15}N$ from the noisy observations. Applying multivariate techniques, we found both global and regional signals of $\delta^{15}N$. From the principal components analysis we found global signals characterized by sharp increases in $\delta^{15}N$ values that began 60,000 and 20,000 years ago. Using k-means clustering, we identified cores with statistically similar $\delta^{15}N$ signals that were in close geographical proximity. These findings suggest that there may be both global and regional forcing of the marine nitrogen cycle.

List of Abbreviations and Symbols Used

Notation	Description
^{14}N	Nitrogen-14 isotope
^{15}N	Nitrogen-15 isotope
CO_2	carbon dioxide
$\delta^{15}N$	expression of the nitrogen stable isotope ($^{14}N/^{15}N$) ratio
e	eigenvector
IG1	interglacial 1
IG2	interglacial 2
K	Kalman gain
km	kilometres
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimation
N	sample size
N_2	nitrogen gas
N_2O	nitrous oxide
O_2	oxygen
PC1	principal component 1
PC2	principal component 2
PC3	principal component 3
PC4	principal component 4
PPT	parts per thousand / parts per mill
Q	process error variance
R	observation error variance
t	time
θ	active parameter set
θ^*	proposal parameter set

Notation	Description
λ	eigenvalue
π	prior

Acknowledgements

I would like to express my gratitude and most sincere appreciation to my supervisors, Drs. Michael Dowd and Markus Kienast, for their support and guidance in the preparation of this Master's thesis.

Dr. Dowd has mentored me in both my undergraduate and graduate academic studies and is responsible for much of my knowledge and interest in time series analysis and state space models. I'd like to thank him for his guidance on appropriate analysis techniques and for insights that helped me put my findings into perspective.

Dr. Kienast provided the data that was the subject of my analysis and always made time, despite having so little of it, to tutor me in Nitrogen 101! I especially appreciated his assistance with the writing and interpretation of Chapters 1 and 4 of my thesis.

I would like to acknowledge the many professors I have studied under at Dalhousie University: in particular, Dr. David Hamilton, who encouraged me to pursue graduate studies in statistics and who was always available to discuss course material, career plans and life in general; as well as Dr. Ammar Sarhan, who was one of my first professors at Dalhousie and from whom I have taken many courses and learned much.

I would also like to thank Greg Britten, a fellow graduate student, for helping me brainstorm ideas around this analysis and especially for his help in debugging the code I used to execute it.

Finally, I would like to acknowledge my immediate and extended family, who have supported me in all of my pursuits, especially my education, and who have inspired me by their example.

Chapter 1

Introduction

Nitrogen is the seventh element on the periodic table and is a key element in ocean biogeochemistry. Fixed forms of nitrogen, when in short supply, are known to be a limiting factor in marine productivity [1]. It is not surprising, then, that changes in nitrogen levels in the ocean over time are an important area of study for oceanographers. By understanding how the processes that affect nitrogen levels have changed through geological time (*e.g.*, the past 125,000 years), oceanographers can get a better sense of how marine ecosystems faired over that time and what future anthropological changes could mean for our planet.

1.1 The Marine Nitrogen Cycle

All organisms in the ocean require nitrogen as well as carbon and other elements as nutrients for food and energy. The vast majority of nitrogen is in the form of nitrogen gas (N_2), which most living beings are not capable of using [1]. These organisms require other organisms to convert the N_2 gas into a form that they can use. There are also organisms that will reconvert these fixed forms of nitrogen into N_2 gas. The transformation of nitrogen between these different states is what is known as the nitrogen cycle.

Overall, there are two processes that predominantly determine the amount of fixed nitrogen the ocean. These processes are nitrogen fixation and denitrification. Nitrogen fixation is the process by which N_2 gas is eventually converted into nitrate or ammonium that can be used by phytoplankton to make organic matter [1]. Phytoplankton are food for more complex organisms, and form the base of the marine food chain. Phytoplankton use this "fixed nitrogen" to make organic tissue from carbon

and other elements. This means that the nitrogen cycle has an influence on, and is influenced by the carbon cycle. Denitrification is the reverse of nitrogen fixation in that it is the process by which nitrogen is converted from biologically accessible nitrogen species back into N_2 gas, and is lost from the ocean to the atmosphere. The rates of nitrogen fixation and denitrification determine how much nitrogen is available in the ocean at any given time and, as a result, ultimately regulate marine productivity and thus, marine carbon dioxide (CO_2) fixation.

The nitrogen cycle in the ocean has therefore been of interest to many oceanographers and marine biochemists. A particular area of interest has been to measure the impact of humans on the marine nitrogen cycle. Humans have started to artificially fix nitrogen for use in soils, which gives plants on farms and gardens more nutrients to grow. The amount of fixation done by humans has rivaled that of the fixation of nitrogen done by marine organisms [1, 2]. A second reason is the impact that the nitrogen cycle has on climate change. Nitrous oxide (N_2O), which is created in both nitrification and denitrification, is a potent greenhouse gas, and when it is created some of it can escape the ocean and goes into the atmosphere [1]. Having more N_2O in the atmosphere would speed up the effects of climate change.

1.2 Studying Nitrogen in Past Environments

Commonly, researchers who investigate how nitrogen behaved in the past, look at the behaviour of nitrogen in specific periods in time: glacial (ice ages) and interglacial (the transition from the glacial to the next glacial period). One reason for looking at these periods in particular, is that they are easier to date and analyze as they are well recorded in sediments on the ocean floor [1]. A second reason is that they offer researchers a real life or "natural experiment" setting in which observations range from a stable equilibrium state in a glacial time period, then transition to a state of rapid change in an interglacial time period, before returning again to a stable state in the next glacial period.

There is no direct record of the historic levels of nitrogen in the ocean [1]. Instead,

there is a proxy measurement that is based on the two stable isotopes of nitrogen. These two stable isotopes are ^{15}N, which is very rare, and ^{14}N, which accounts for 99.63% of all nitrogen. Using these two stable isotopes of nitrogen, an isotopic ratio is calculated and is expressed in terms of ($\delta^{15}N$) as follows:

$$\delta^{15}N = (\frac{^{15}N/^{14}N_{Sample}}{^{15}N/^{14}N_{Air}} - 1) * 1000PPT \qquad (1)$$

where the ratio in the air is used to standardize the results, and PPT stands for parts per thousand. The isotope ratios in the sample are determined on dry ocean sediments sampled from the ocean floor.

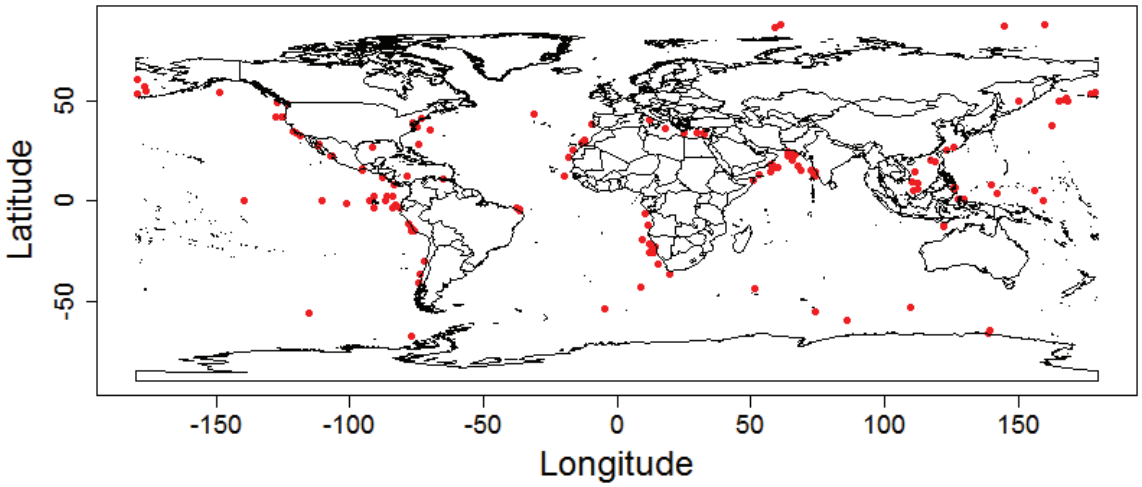
The nitrogen isotope ratio does not give a measurement of the level of nitrogen in the ocean *per se*. Rather, it indicates which process was more active in a particular region. For example, in areas where denitrification is the driving force, $\delta^{15}N$ values tend to be higher. This is because the process of denitrification uses the ^{14}N preferentially over the ^{15}N. When this process is the driving force, it leaves behind higher than normal levels of ^{15}N, which increase $\delta^{15}N$. Conversely, in areas that have very high rates of fixation, $\delta^{15}N$ tends to be much lower. The average $\delta^{15}N$ of N_2 gas is 0 PPT, whereas the mean $\delta^{15}N$ of nitrate in the ocean is 5 PPT. No isotope fractionation occurs during nitrogen fixation. Therefore, nitrogen fixation is adding nitrogen with relatively lower nitrogen isotope ratio to the ocean, overall lowering the $\delta^{15}N$ of mean oceanic nitrate. This means that in an area where there are higher levels of $\delta^{15}N$, the denitrification rate might have been larger than the nitrogen fixation rate. In much of the ocean, however, the $\delta^{15}N$ signal is determined by different utilization of nitrogen species, a process that fractionates isotopes as well.

1.3 Previous Research

As discussed above, when studying the marine nitrogen cycle in the past, researchers use $\delta^{15}N$ as a proxy measurement. In their 2013 paper, Tesdal *et al.* [3] compiled 173 downcore records of ocean sediments, most of which included measurements of

$\delta^{15}N$ (153). The geological locations of these core samples can be seen in Figure 1.1. (Note that the world map graphic used in this and all subsequent figures where locations are indicatted, were generated using the maptools package in R. [4]) The individual records were collected by a number of different researchers and represent different periods of time.

Figure 1.1: Location of all cores with $\delta^{15}N$ measurements (n=153).



The ages associated with each $\delta^{15}N$ measurement were retained from the original sources, unless an age measurement was not provided. It should be noted that there is a sampling bias in this data set as most of the samples were collected in areas of interest to the original researchers, who selected areas where nitrogen cycle processes of interest were known to be present. It should also be noted that the farther back in time the measurements are the less accurate they are believed to be (*i.e.*, $\delta^{15}N$ measured 5000 years ago are considered more accurate than $\delta^{15}N$ measured 100,000 years ago).

In addition to compiling this data set, Tesdal *et al.* [3] conducted a clustering analysis to group samples of $\delta^{15}N$ into regions. They looked at cores within 100 km of defined reference cores, and looked at their similarities based on their means over a defined time period. In order to do these comparisons, they had to place

the observations on a constant time line that was the same across all cores being investigated. They used interpolation to estinate $\delta^{15}N$ where the values were missing. For cores within the 100 km radius, they found that differences in the $\delta^{15}N$ were relatively small.

When looking at any time series sample there is an amount of error associated with the observations that are made. It is impossible to make 100% accurate observations as the methods of measurement and the calibrations made to collection devices can add in some error. This holds true for the observations of $\delta^{15}N$ discussed by Tesdal *et al*. [3].

In another research paper, Galbraith *et al*. [5], grouped the cores in this data set into 16 clusters based on the known oceanographic biological provinces and common $\delta^{15}N$ signal. To summarize a few of these regions, there were three on the west coast of the Americas (one in South America, Central America and North America).

1.4 Thesis Goals

The main goal of this thesis is to try to expand on the research of Tesdal *et al*. [3] and Galbraith *et al*. [5], by approaching the data set from a purely statistical point of interpretation, without prior assumptions, in an attempt to extract the true signal of $\delta^{15}N$ values from the noisy observations over three time periods. The noise in this case is the error in the observations that is due to the measurement of $\delta^{15}N$ and not the natural variability of $\delta^{15}N$ itself. The difference between these two errors will be discussed more in Chapter 2 when we talk about state space models. By doing this analysis, we aim to get a better understanding of changes in the nitrogen cycle (mainly fixation and denitrification rates) through time.

The first time period of interest, named interglacial 1 (IG1), covered the time span of 30,000 - 5,000 years ago. This time period looked at the time from the last glacial maximum to almost the present, *i.e*, including a glacial-interglacial transition. Our second time period was called the glacial time period, which covered the time span of 70,000 - 30,000 years ago. The glacial time period represents the time when the Earth

was in the most recent glacial cycle. Finally, we looked at the time period we called interglacial 2 (IG2), which spanned from 125,000 to 5,000 years ago. This time period covered both the previous two time periods as well as the preceding interglacial cycle. These three time periods were chosen because they covered the glacial and interglacial cycles that the majority of oceanographers are interested in.

To achieve our main goal and make accurate comparisons based on these data, we needed to extract the signal of $\delta^{15}N$ and conduct a multivariate analysis to find dominant signals over the defined time periods and in certain regions over those time periods. As a first step, we aim to ensure that any differences observed in $\delta^{15}N$ values represent a change in state rather than a change associated with unequal time step intervals or missing values. Here, we are specifically concerned with the mixture of two error terms: measurement (or observation) error and process error. Measurement error is the error associated with the measurement of the observation; that is, the error that can be attributed to the techniques for sample collection and the determination of the $\delta^{15}N$ values at each time point. The process error is the error associated with the statistical process itself, that is how much variability is there in the state from one time step to the next. It is important to separate out the measurement error because it could affect the analysis by providing inaccurate measurements of the true state (underlying statistical process). This could lead to results that do not reflect the true signal of $\delta^{15}N$ but instead reflect the methods used to measure the data, depending on the amount of measurement error in the observations. In Chapter 2 we will discuss techniques used to separate these two errors, as well as the methods for establishing a constant time line and dealing with missing data.

Our second step will be to look for dominant signals over the defined time periods and in certain regions over those time periods using multivariate analysis techniques. We will use principal component analysis to determine whether there is a global dominant $\delta^{15}N$ signal - *i.e.*, one that is not distinct to a particular region. Using this technique, we aim to account for the majority of the variance in the signals of $\delta^{15}N$ over a given time period in a small number of principal components (signals). That

is, we hope to take all of the cores that cover a respective period (e.g., 60 cores) and reduce them to just those signals that represent the time period as a whole (e.g., 3 signals). To see if there are any distinct signals in particular regions, we used the k means algorithm. This algorithm will allow us to compare, between cores, the extracted values of $\delta^{15}N$ at each time point in order to create defined clusters of cores that share similar signals. We will be looking to see if these clusters of cores with similar signals are also located in similar geographic regions. The multivariate techniques will be described in greater detail in Chapter 3.

Chapter 2

Univariate Analysis

The goal of this chapter is to extract the true signal of $\delta^{15}N$ from the noisy observations. This will allow us to get accurate measurements of $\delta^{15}N$ signals for all of the cores that cover their respective time periods and will set up the data to be used in a multivariate cluster analysis (Chapter 3). To accomplish this we first constructed a constant time line for each distinct time period. This was necessary because each core sample provides a unique time line of observations, varying both in the number of observations and the time intervals between observations (the latter even within a single core). To make meaningful comparisons between cores, the comparisons must be made at the same points in time. Establishing a constant time line with regular intervals could help to make sure that any changes in $\delta^{15}N$ found in sequential intervals are not influenced by variations in the size of the intervals. However, by imposing a fixed time line, there will be a number of time points with missing a $\delta^{15}N$ value for a given core. Allowing missing values could, in effect, be the same as allowing unequal intervals between observations. Therefore, our second task was to fill in missing data on the constant time line. Finally, and most importantly, we extracted the true signal of $\delta^{15}N$ from the noisy observations. Each observation of $\delta^{15}N$ has a certain amount of error associated with it: observation error (error due to the measurement processes itself), which could lead to inaccurate or false results, and process error (the error associated with the natural changes in $\delta^{15}N$ through time), which is part of the process we want to measure. It is important to our analysis that we separate these two errors.

All three of these issues need to be addressed before we move on to the multivariate analysis and compare changes in the $\delta^{15}N$ ratio between cores within the three defined time periods (Chapter 3). We want to ensure that any differences found are based

on actual differences in the $\delta^{15}N$ and not based on the varying time steps, missing values, or error in observations due to the collection and measurement process. By solving the problems described above (and in more detail below), we are increasing the likelihood that differences found between cores in the multivariate analysis will be based on differences in the true signals of $\delta^{15}N$ values alone.

It should also be noted that, due to the large number of data sets and the fact that there are three time periods of interest, all of the described techniques were automated; that is, instead of looking at each core individually over the three time periods, the techniques were set up to run over all data sets in a one-size-fits-all model. However, as a first step, we ran some basic statistical analysis and data visualization on the raw data to see if any patterns were evident.

2.1 Data Visualization

We first looked at the mean and slope of $\delta^{15}N$ values over each time period. The mean $\delta^{15}N$ as well as the slope of $\delta^{15}N$ were calculated for each core that covered the given time period. These values were then plotted on a global map to see if cores that were geographically closer together shared similar values for these two statistics.

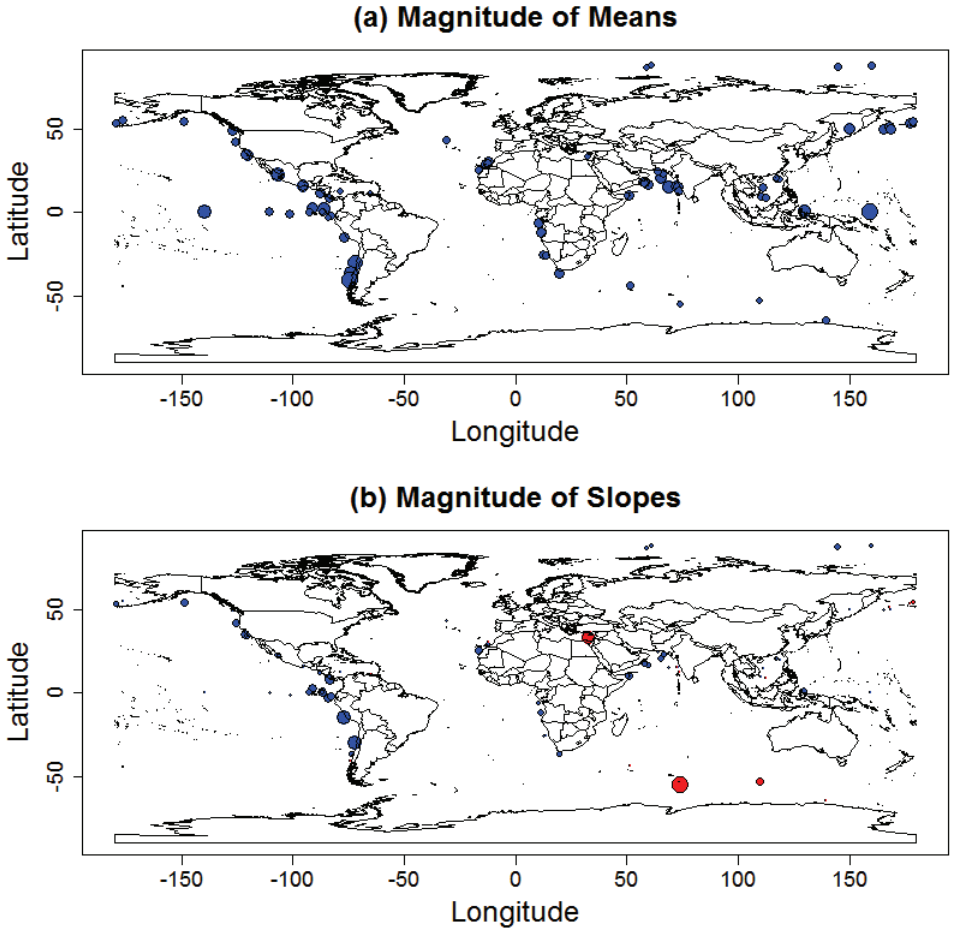
2.1.1 Inter Glacial 1 Period - 30,000 to 5,000 Years Ago

In total, 78 cores covered this time period. As shown in Figure 2.1, we can see that the mean of $\delta^{15}N$ values in cores that are close together geographically are similar. For example, the three cores on the southwest coast of South America have similar means (9.72, 10.77 and 8.82), which are also quite large when compared to the rest of the cores. The cores on the southwest coast of Africa also have similar means (6.65, 6.79, 5.40 and 5.80), but are much smaller than those on the South American Coast. This is consistent with the findings of Tesdal et al. [3] in their clustering analysis, which showed that cores within 100 km circles were not very different and that the greatest variability in mean values occurred in the Arabian Sea, off the northwest coast of South America near the equator, and on the northern coast of Africa in the

Atlantic Ocean.

Looking at the slopes of $\delta^{15}N$ values (Figure 2.1), based on ordinary least squares regression, it can be seen that the vast majority of cores (63 of 78) have a positive slope (depicted with a blue circle), indicating that the majority of the $\delta^{15}N$ values increase as we go from 30,000 to 5,000 years ago. This again is consistent with the finding in Tesdal et. al. [3] that older observations of $\delta^{15}N$ have values that were smaller than more recent observations.

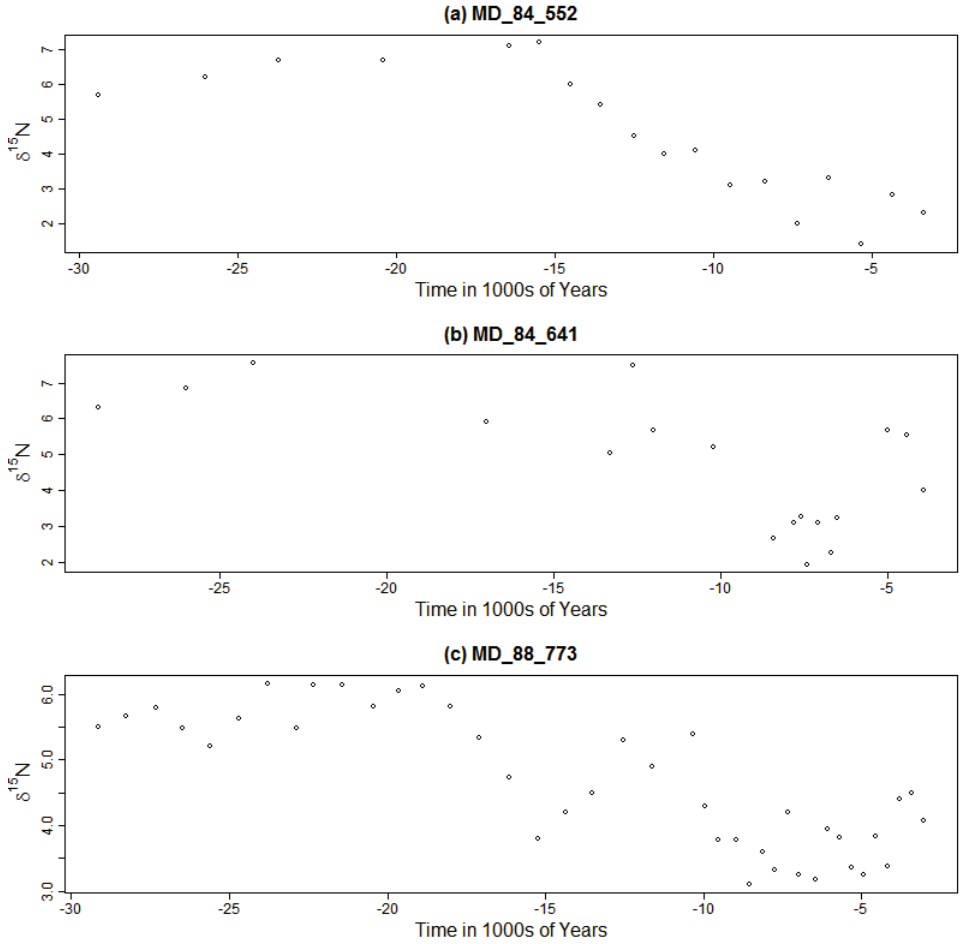
Figure 2.1: a) Plot of the magnitudes of the means of the cores that span the interglacial 1 time period. b) Plot of the magnitude of the slopes that span the interglacial 1 time period. Red circles indicate a negative slope (decreasing trend), while blue circles indicate a positive slope (increasing trend) in $\delta^{15}N$ as time goes from 30,000 years ago to 5,000 years ago.



A smaller set of cores (15 of 78) have decreasing $\delta^{15}N$ values. However, it was

noted that most cores with negative slopes (12 of 15) were very close to 0 (< 0.06). Three cores, two in the Southern Ocean, stood out with unusually high negative slopes (magnitude over 0.1).

Figure 2.2: a) Plot of the $\delta^{15}N$ values vs time for the MD 84-552 core. b) Plot of the $\delta^{15}N$ values vs time for the MD 84-641 core. c) Plot of the $\delta^{15}N$ values vs time for the MD 88-773 core.



In Figure 2.2, we take a closer look at these three cores. The plots show a general overall pattern in each core. Between 30,000 and 20,000 years ago the $\delta^{15}N$ values tend to be relatively stable (around 5.5 to 7 PPT). Sometime between 20,000 and 15,000 years ago the $\delta^{15}N$ values began to decrease, reaching values of about 3 PPT between 10,000 and 5,0000 years ago. Based on this visual analysis, one could speculate that these three cores may have shared a common historic event that caused the $\delta^{15}N$ values to drop rapidly about 15,000 years ago.

2.1.2 Glacial Period - 70,000 to 30,000 Years Ago

Looking at the glacial means in Figure 2.3, just as they were in IG1, the cores that are closer together appear similar in magnitude. This is best demonstrated by the two cores with large means on the west coast of South America and the grouping of cores with smaller means around the northeast of Africa in the Indian Ocean. Again, staying consistent with Tesdal *et al.* [3] , as with IG1, we see that the means for the cores on the west coast of South America are larger than most of the other cores. However, the cores located in the Arabian Sea, off the northwest coast of South America near the equator, and on the northern coast of Africa, in the Atlantic Ocean, still take on a wide range of mean values, as they did in IG1 and in Tesdal *et al.* [3] .

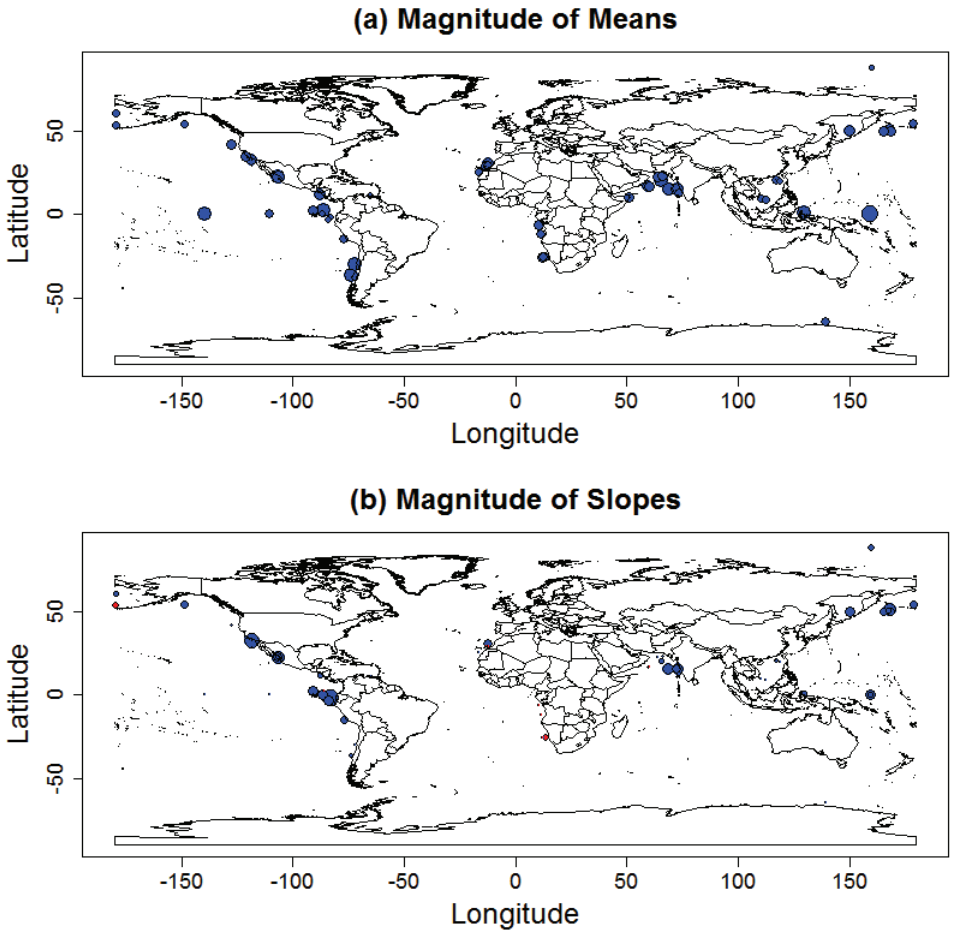
Several cores had negative slopes, but none were as large as those observed in the IG1 time period. Most cores had a positive slope, which is again consistent with Tesdal *et al.* [3], but the magnitudes of the slopes seem to have decreased from those in the IG1 time period.

2.1.3 Inter Glacial 2 Period - 125,000 to 5,000 Years Ago

In the last time period of interest (which encompasses the previous two time periods), the same trends continue and are even more evident. Cores that are closer together have similar means. For example, the cores in the Arabian Sea and off the northern coast of South America now have roughly equivalent means, whereas in the previous time periods, there were more cores in these regions with more variation (Figure 2.4). Generally, over this time period, the slopes of the $\delta^{15}N$ tend to be positive, indicating that $\delta^{15}N$ values tend to get higher as we get closer to the present.

The visual analysis of the raw data suggest that, generally, $\delta^{15}N$ values trend upwards over the three time periods and that cores that are geographically closer together appear similar. This was in agreement with the analysis done by Tesdal et al. [3]. It also became clear from the data in Figure 2.2 that some cores have observations in shorter intervals of time than others (for example, see cores MD 84-552 and MD 88-773), and that some cores have large gaps between observations (for

Figure 2.3: a) Plot of the magnitudes of the means of the cores that span the glacial time period. b) Plot of the magnitudes of the slopes that span the glacial time period. Red circles indicate a negative slope (decreasing trend) while blue circles indicate a positive slope (increasing trend) in $\delta^{15}N$ as time goes from 70,000 years ago to 30,000 years ago.

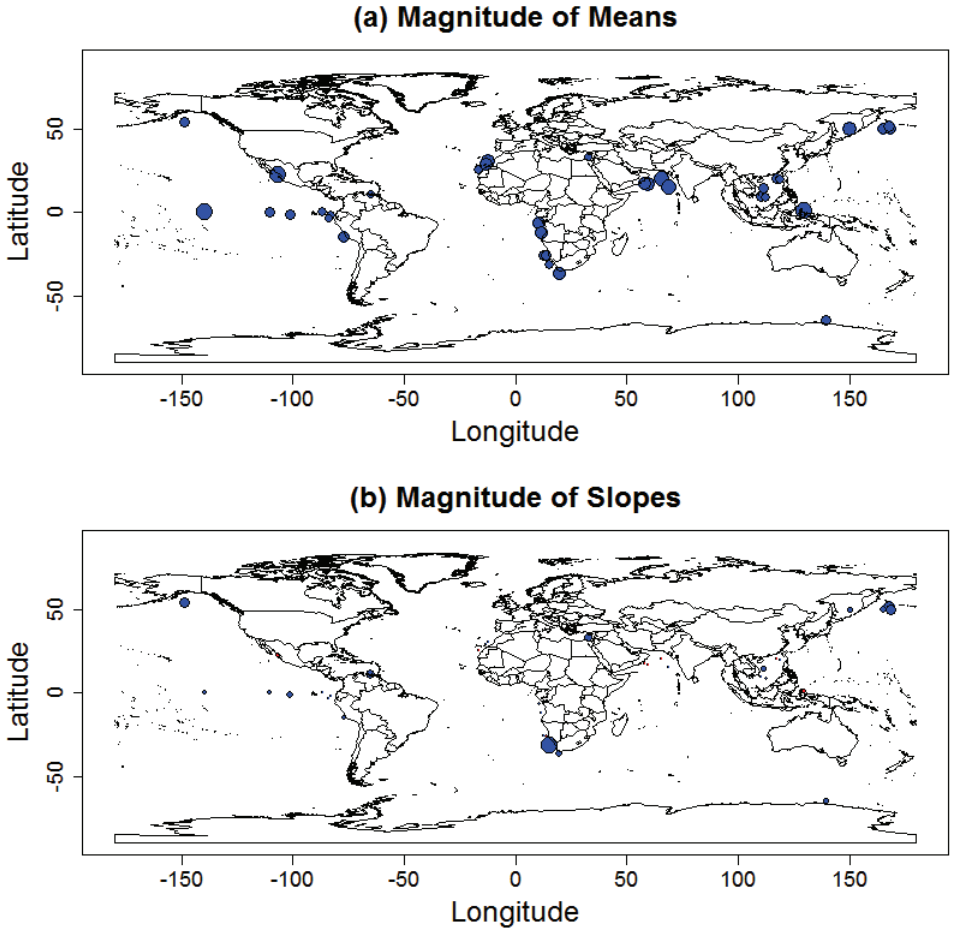


example, in MD 84-641 there is a 7000 year gap between observations around 20,000 years ago).

2.2 State Space Models

In this section we plan to use a state space model to extract the true underlying signal of the $\delta^{15}N$ values. To do this we will separate the error in the observations from the error in the statistical process itself, while at the same time, filling in the gaps of missing $\delta^{15}N$ values on our constant time line. State space models aim to

Figure 2.4: a) Plot of the magnitudes of the means of the cores that span the interglacial 2 time period. b) Plot of the magnitudes of the slopes that span the interglacial 2 time period. Red circles indicate a negative slope (decreasing trend) while blue circles indicate a positive slope (increasing trend) in $\delta^{15}N$ as time goes from 125,000 years ago to 5,000 years ago.



obtain an accurate measurement of the state at a given time point t, (x_t), based on a weighted average between an observation of $\delta^{15}N$ at that time, (y_t), and an estimate from the stochastic process at that time, $(x_{t|t-1})$. This weighted average is based on the variances of the two components, where the component with the smaller variance has the most weight. State space models can be used to separate the observation error, which we want to remove, from the process error (hence, signal extraction). The algorithms and equations used in this section were sourced from Shumway and Stoffer [6].

For this analysis the stochastic process that was used was the random walk. This is the simplest model that could be used to model the underlying statistical process ($\delta^{15}N$ signal). The set of equations that will be used for the state space model are as follows:

$$x_t = x_{t-1} + v_t \quad v_t \sim N(0, Q) \qquad (2)$$

$$y_t = x_t + w_t \quad w_t \sim N(0, R) \qquad (3)$$

Here, v_t represents the process error. The process error is associated with natural changes in $\delta^{15}N$, such as changes in nitrogen fixation and denitrification rates. This is the error we are interested in studying as it represents how much nitrogen levels change in the series. Measurement error is represented as w_t. This is the error that is associated with the data collection process and has nothing to do with changes in the $\delta^{15}N$ values over time. Rather it is the error added into the observation by the process used to collect and measure the data. When looking at these two equations, the only two unknowns that need to be estimated are the variances Q and R for the state and observations respectively.

Using a state space model we aimed to do two things: (i) create a likelihood of Q and R based on the observations of $\delta^{15}N$; and (ii) obtain estimates of the true $\delta^{15}N$ signal. For each core in each time period we will obtain both a likelihood and state estimate from our state space model.

A state space model was our method of choice because of its ability to separate observation from process error. Other methods such as interpolation, splines and nearest neighbour search (which will be described below), would be able to place our value on a constant time line, but, would not be capable of separating out the observation and process errors. When it comes to using splines we would need to set the number of knots (break points in the data) which would affect how "smooth" our state estimates are. Using state space models we will let the data and the parameters determine the amount of smoothing instead of a pre-decided number of break points.

2.2.1 Observation Error (R) and Process Error (Q) Distributions

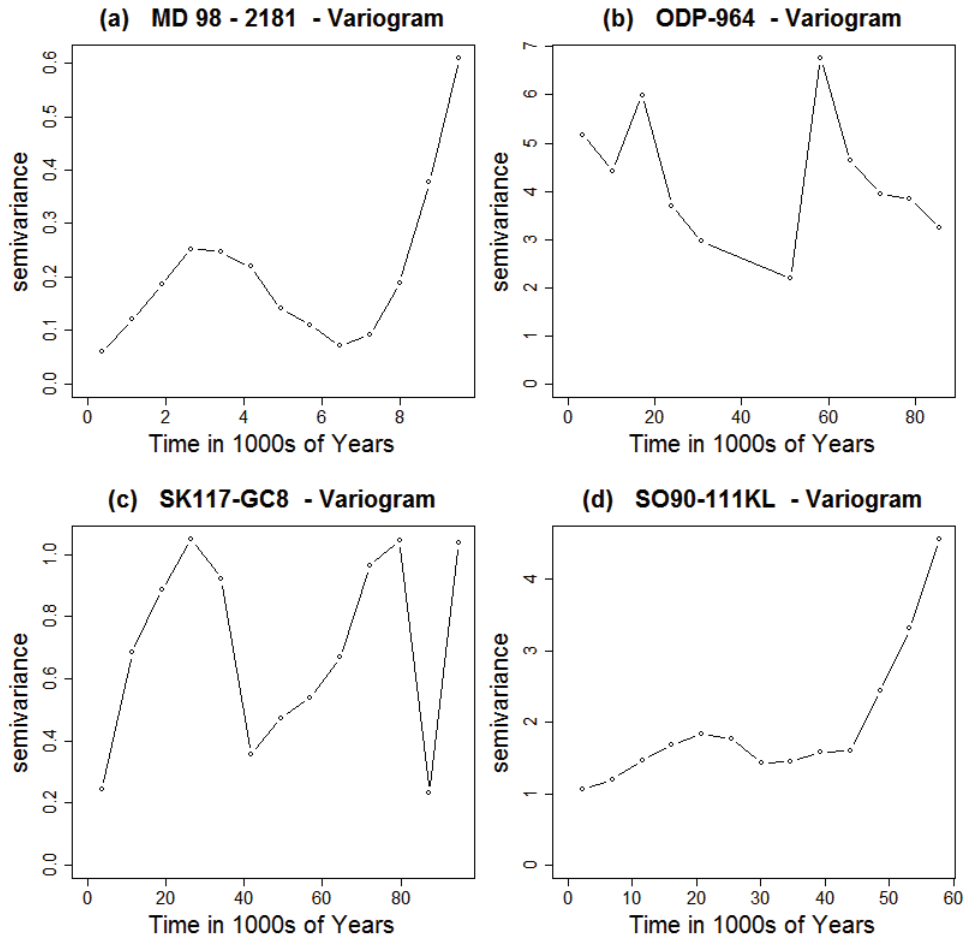
Before moving on with the state space model estimation of Q and R, we first decided to get a rough idea of the distributions of these two parameters based on all of the cores we had available. This would give us a reasonable understanding of the range of values that could be expected from any estimation we would make.

Observation Error Distribution

In order to get an idea of the distribution of the observation error, a variogram was used. Variograms, used for irregular sampling intervals, are used in spatial statistics and are equivalent to the auto-covariance function in time series. A variogram looks at closely related points within certain distances and computes the covariance between those points. That is, it looks at the covariance of the observations based on the distance (in time) they are from each other. To do this we use the variog() function in the geoR package for the statistical program R [7,8]. When fitting a variogram, the estimate of the observation error variance is known as the nugget. The nugget is the value on the variogram at which the distance is zero (the y-intercept). By getting an estimate of the nugget for each of the cores we will be able to generate a distribution of observation error variances.

Even with the variograms values calculated, they were very challenging to fit to a variogram model. This is illustrated in Figure 2.5 where it can be seen that the variograms take on many different shapes. As we are dealing with a large number of data sets, it would be very difficult to find a single variogram model that fits all of the data sets. Therefore, instead of fitting a variogram model, a regression line was fit to the first four points (the four smallest distances) of the variogram. These four smallest distances varied between cores. The nugget was then estimated as the intercept of this regression line and was the estimate of R for that core. This approach was acceptable as, at this point, we were only trying to get a rough idea of what the observation error was.

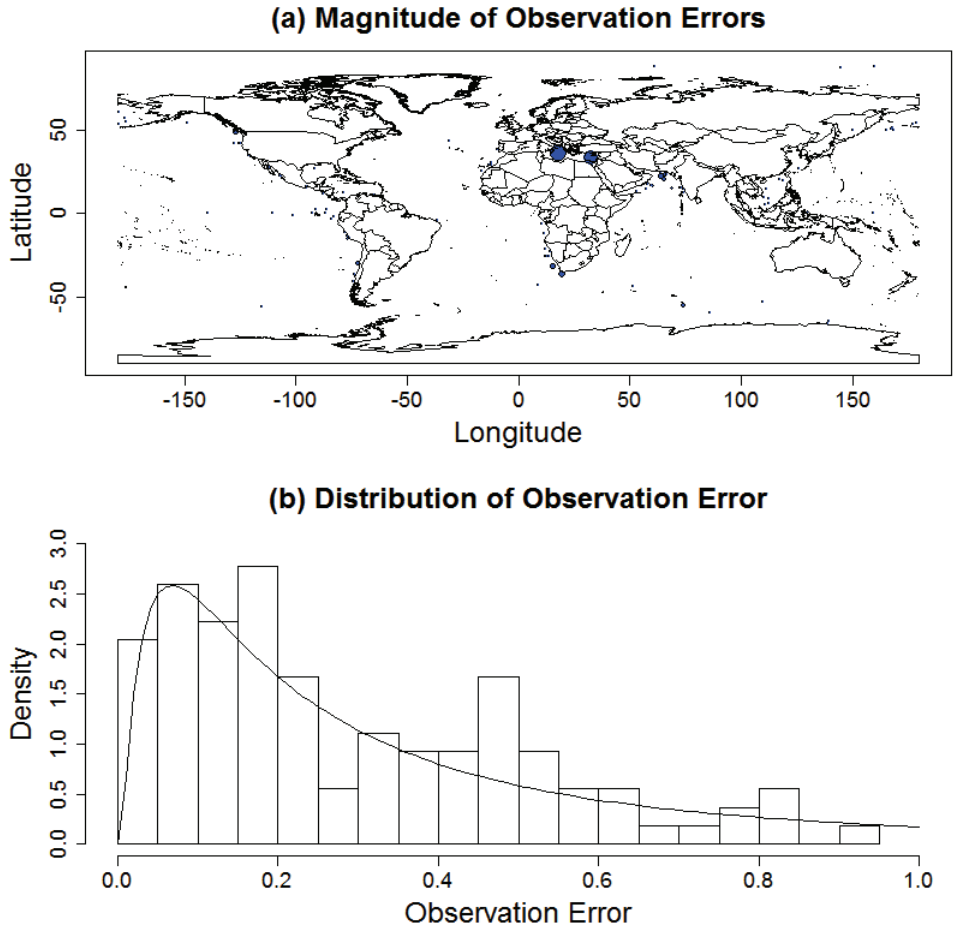
Figure 2.5: Variograms for four cores.



Once all cores had an estimate of the observation error (R), we generated a histogram of the results. The histogram showed a skewed distribution and was fitted with a log-normal distribution with a mean of 0.10 and a standard deviation of 0.07 (see Figure 2.6). These results were lower than expected as it was believed that the observation error variance would be 0.2 at a minimum (M. Kienast personal communication).

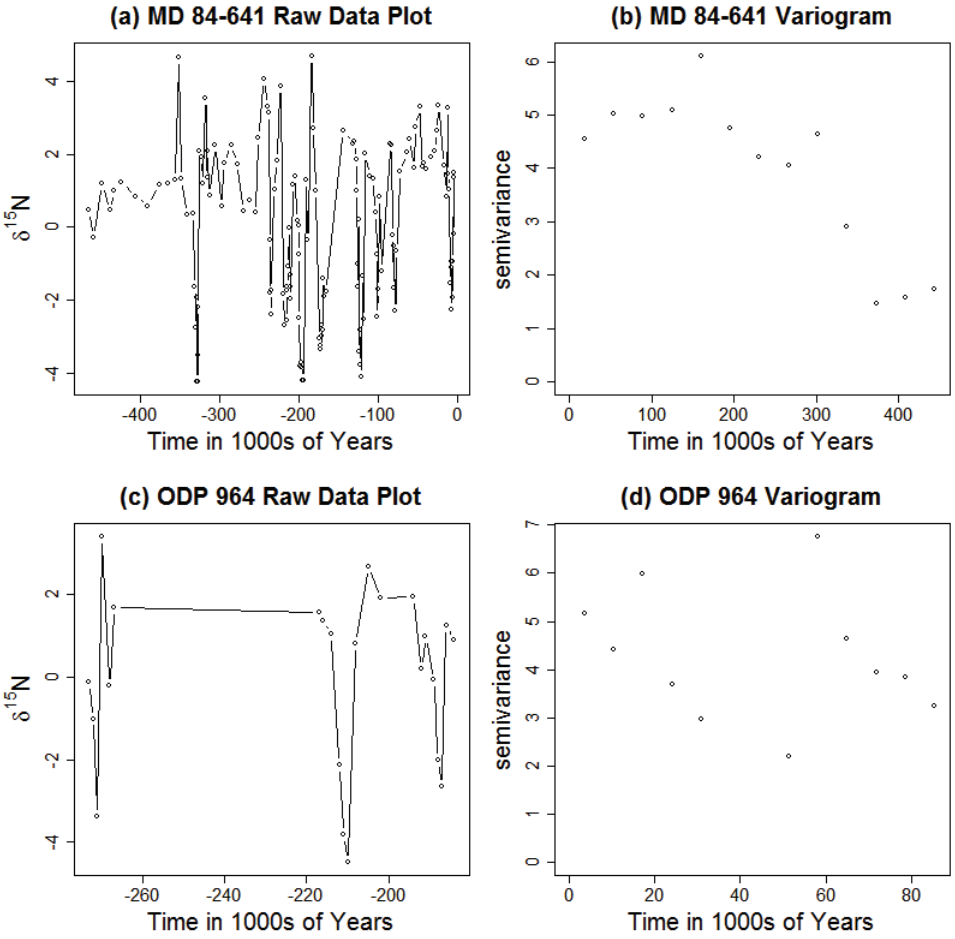
In Figure 2.6, we see two cores (shown as large blue circles) with large observation error estimates in the Mediterranean Sea. The other estimates are so small compared to these two that they do not even show up visibly on the map, which means the estimated observation error was much larger in these cores than in the rest. Looking just at these two series (see Figure 2.7), we see that core MD 84-641 had periodic

Figure 2.6: a) The magnitude plot of the observation error estimates (nuggets) obtained from the variograms. b) Histogram with overlaying log normal distribution with mean = 0.1, and standard deviation = 0.07.



behavior which could cause periodic variograms. In the core ODP 964's variogram, that the first four data points created a negative slope instead of a positive slope. This means that our regression line will be increasing instead of decreasing as it moves towards zero. Also, looking at both cores, we saw that there were not many observations at short distances (i.e., distances less than 1000 years apart). This is a good example of how the observation error cannot be estimated the same way for all cores. In the cases described above, the large observation error estimate likely had more to do with the method used to make the estimation than with actual observation error.

Figure 2.7: Raw data plots and variograms showing cores with unusually high estimates of the observation error. a) The $\delta^{15}N$ values plotted against time for MD 84-641. b) The variogram plot for MD 84-641. c) The $\delta^{15}N$ values plotted against time for ODP 964. d) The variogram plot for ODP 964.



The Process Error

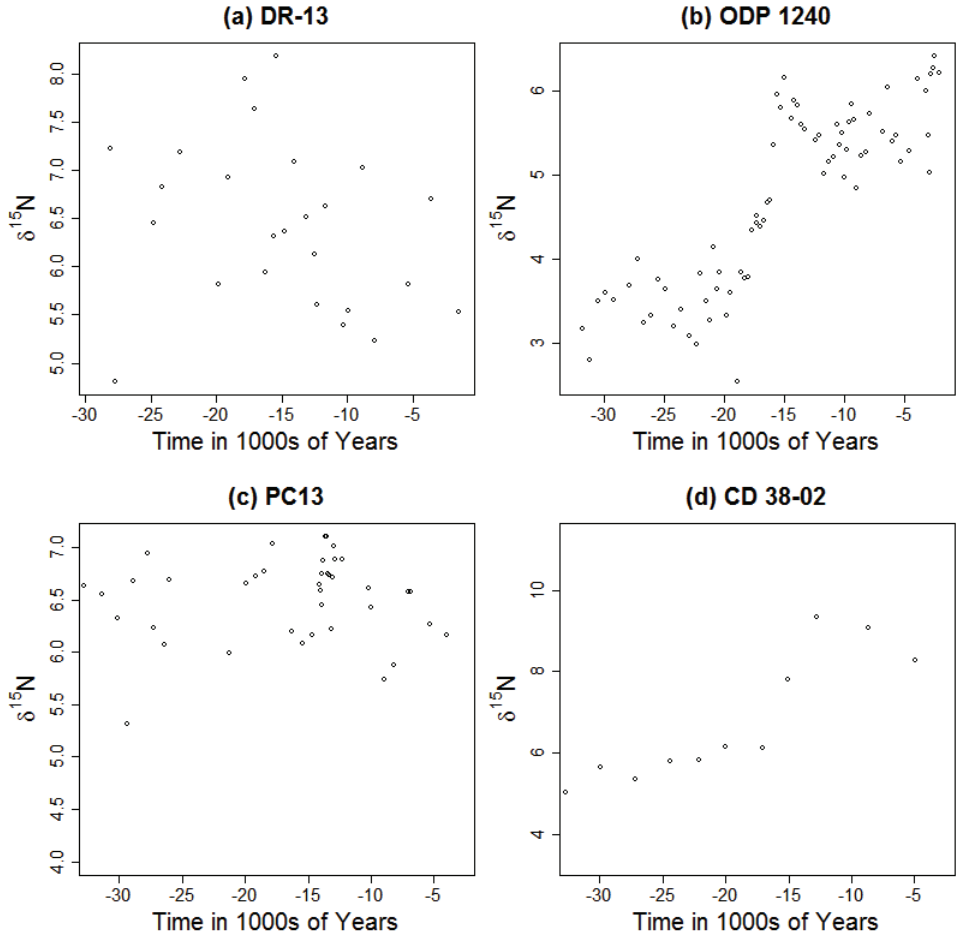
In this section we discuss the techniques used to get a get a rough estimate of the magnitude of the process error. As discussed above, the process error is an estimate of how much the state can vary (increase or decrease) in one time step. The time steps within and between series are not all of the same length. This could lead to conflicting results if we compute the difference in $\delta^{15}N$ on an interval of 1000 years and compare it, with equal weight, to a difference in $\delta^{15}N$ based on 100 years. Before calculating the difference in $\delta^{15}N$ from one time step to the next, we first had to deal with the irregular sampling intervals in the data sets.

One of the problems encountered with this data set was that the observations were not separated equally on the time axes within and between the cores. Within a single series, the time steps (jumps) between observations did not occur at equally spaced intervals. As a result, when measuring how much $\delta^{15}N$ changes from one time point to another, any differences observed could be a matter of different time gaps between intervals and not changes in actual $\delta^{15}N$ values. Similar problems were noted between cores, which did not capture observations at matching time points. Again, this is problematic because to compare changes in $\delta^{15}N$ values between cores, those comparisons need to be made over the same specific period of time (*e.g.*, between 10,000 and 9,000 years ago). If not addressed, this could lead to findings that suggest changes in the nitrogen fixation and denitrification rates occur more rapidly in one core than in another when, in fact, the changes were simply measured on differing time scales.

As can be seen in Figure 2.8, the observations of $\delta^{15}N$ are not uniformly distributed on the time axis. If we compare the cores ODP 1240 and CD 38-02 (panels b and d), it can be seen that between 20,000 and 15,000 years ago, both series of $\delta^{15}N$ values increase by about 2 PPT. However, we can also see that over that 5000 year interval, ODP 1240 had almost four times as many observations as CD 38-02. Looking only from one observation to the next, it may seem that $\delta^{15}N$ increases more rapidly in core CD 38-02 than in core ODP 1240, when in fact the time intervals in the latter are shorter. This can be particularly problematic when trying to compare process errors. Since the process error is a measurement of how much the $\delta^{15}N$ signal can vary from one time step to the next, if different cores don't follow the same time intervals then comparisons of the process error will be based on the time gaps and not on the $\delta^{15}N$ signal. To solve this problem we put the data on a regular interval time line so that the spacing between $\delta^{15}N$ measurements was not only consistent within the core but also between cores, allowing meaningful comparisons to be made.

For this analysis we put the data on a constant time line with a time step of 1000 years. The unit of 1000 years was deemed the most logical choice as it was the base

Figure 2.8: Sample plots of the $\delta^{15}N$ observations from cores covering the interglacial 1 time period.

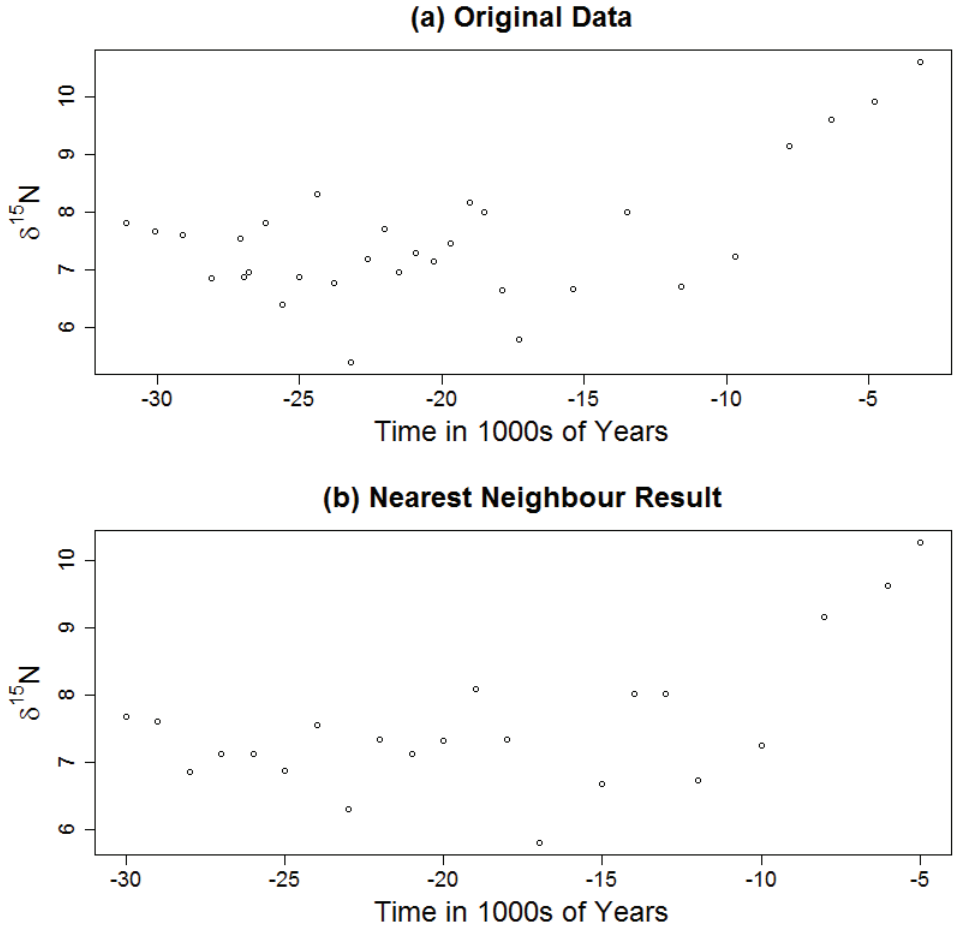


unit of measurement. Still, the observations of $\delta^{15}N$ values do not fall precisely on this time line at the designated time points. (This was not a problem based on the interval selected but on the non-synchronous nature of the raw observations.) For example, there are no observations at exactly 8000 years ago. A nearest neighbor search was used to put all of the $\delta^{15}N$ values onto the created constant time line. We chose nearest neighbor because it is a relatively simple algorithm for solving problems like this. It is also worth noting that the 1000 year time step continues to have missing values, but when we tested the time line with shorter time steps the number of missing values increased.

Each observation of $\delta^{15}N$ was grouped to the closest time point on the constant

time line. In the event that an observation was equally distant from two time points, the observation was placed at the highest time point. For example, an observation at 8,200 years ago would be placed at 8,000 years ago on the constant time line, while observatins at 8,500 years ago and 8,700 years ago would be placed at 9,000 years ago on the constant time line. If two or more observations within the same core were closer to the same time point, the mean value was used for the observation at the designated time point. An example of the outcome of the nearest neighbour search can be seen in Figure 2.9.

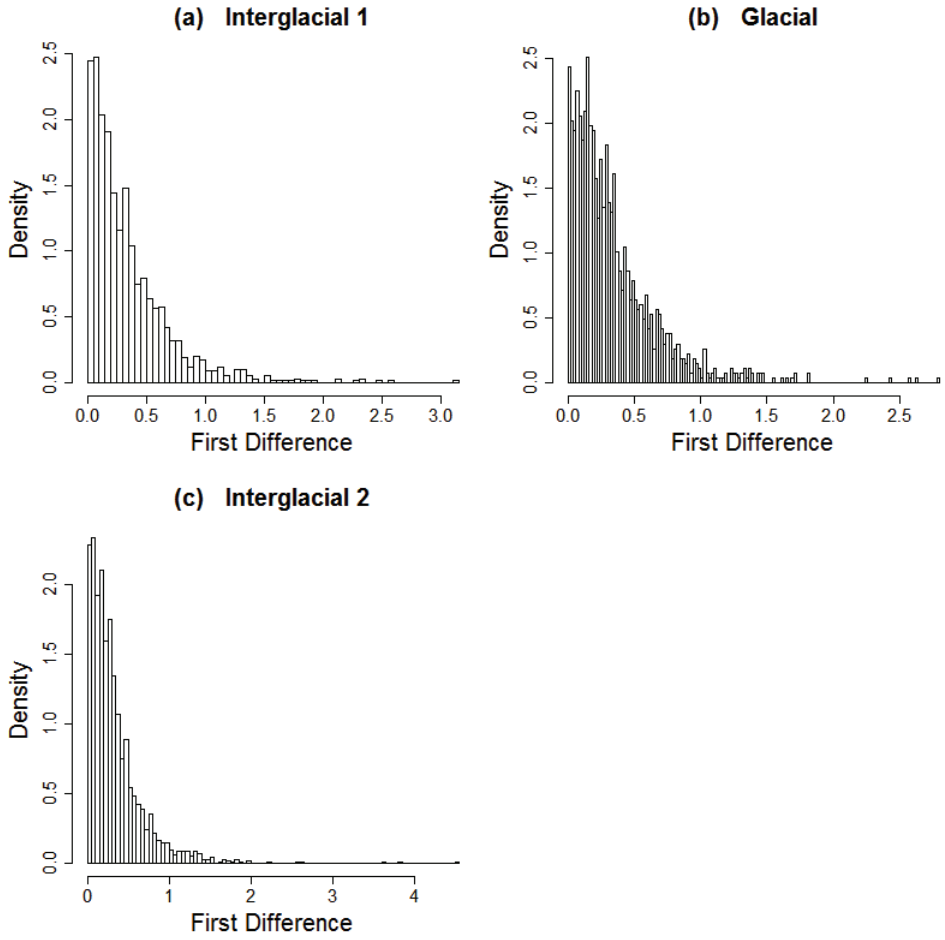
Figure 2.9: Plots showing the original (a) data for core ME33-NAST and the same data after the application of the Nearest Neighbour Search (b).



The nearest neighbor search was completed for each of the three time periods. In order to be included in the analysis for a given time period, the core data set had to

have at least one observation within ±1,000 years of both the start and end point of the time period (i.e., for inclusion in the IG1 period, a core needed to have at least 1 observation between both 4000-6000 years ago and 29,000-31,000 years ago). This was done to ensure that, if a core was selected for analysis for a given time period, it had data that covered the entire time period. After putting all of the cores onto the constant time line, it was found that 78 of the cores spanned the IG1 time period, 55 spanned the glacial time period and 35 spanned the IG2 time period. There are other methods that can map data to a constant time line (such as interpolation), but we were satisfied with the results of the nearest neighbour search.

Figure 2.10: Process error distributions showing the first difference absolute values for each of the three time periods: (a) interglacial 1, (b) glacial, and (c) interglacial 2.



Having established a constant time line, the next step was to get a benchmark for what the process error should be by calculating all of the first differences for the respective time periods. This would give us an idea of how much $\delta^{15}N$ can vary from one time step to the next. However, despite being grouped on the constant time line, the majority of cores still had missing values of $\delta^{15}N$ at various time points (i.e., no observation near one or more of the regular time intervals, even after the nearest neighbor search). Consequently, the problem of unequal time gaps between $\delta^{15}N$ values remained. As stated above, comparisons between cores that do not have measurements at the same time points can make meaningful comparisons between the cores difficult. Therefore, for the calculation of first differences, only $\delta^{15}N$ values that were 1,000 years apart were used (i.e., there were no comparisons between sequential observations made between 7000 years ago and 5000 years ago). Once the first difference values were stored, a histogram was created, as displayed in Figure 2.10. Here we see that in all 3 time periods the majority of first differences are below 0.5. This indicates that we should not see fluctuations of $\delta^{15}N$ of more than 1 PPT over 1000 year intervals when we do our parameter estimations. It is important to note that this is a rough estimate of the process error, since we have not yet separated the observation and process error this estimate has some observation error in it as well.

2.2.2 State Estimation

By conducting the previous analysis on the observation error and process error, we now have a rough idea of what to expect from the state estimation. The Kalman filter method was chosen to estimate these two parameters. These estimations will be based on maximum likelihood estimation through the use of the residuals from the Kalman filter state estimates. The Kalman filter is one of the simpler methods for analysis of state space models, which is why it was chosen for this first attempt at analyzing the cores. If the Kalman filter gives estimates of Q that are larger than 1 on a consistent basis it could indicate that the Kalman filter estimation is not working

properly.

The Kalman Filter

The Kalman filter is an algorithm that is used to get an estimate of the state at a given time. It works by using the state estimate from the Kalman filter at time t-1 and the observed value of $\delta^{15}N$ (if present) to get a weighted average (the estimate of the state), and an uncertainty estimate.

In order to get an estimate of the state at time t, there needs to be an estimate of the state from the Kalman filter at time t-1. The estimate of the state at time t-1 will be denoted as $\hat{x}_{t-1|t-1}$. This estimation is done in two steps: the forecast step and the measurement step. The estimate of the state at time t-1 has a variance of P_{t-1}, which will be defined later. In the forecast step an estimate of the state is calculated using the estimate of the state at the previous time only. This is done by using equation 1. Next the variance of $\hat{x}_{t|t-1}$ needs to be calculated. The forecast variance will be represented as M_t and is calculated using the following equation:

$$M_t = P_{t-1} + Q \qquad (4)$$

$$where \quad P_{t-1} = [R^{-1} + M_{t-1}^{-1}]^{-1} \qquad (5)$$

In the second step, the measurement step, the estimate of the state at time t, $\hat{x}_{t|t}$, is calculated using the equation:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K(y_t - D\hat{x}_{t|t-1}) \qquad (6)$$

$$where \quad K = PR^{-1} \qquad (7)$$

The multiplier K is known as the Kalman gain. The Kalman gain is what applies the weighting between the forecast mean and the observation at time t. As can be seen from the equation for K above, the Kalman gain is a ratio between the observation error variance and the forecast variance. Here, it is important to note

that we are dealing with a scalar case; there is no multivariate component to the estimation process. If the forecast variance is much larger than the observation error variance then the Kalman filter estimates will be closer to the observation than the forecast estimate. If the situation is reversed, the Kalman filter estimates will follow the forecasted estimates more closely. Using this algorithm we can get estimates of the state at all time points from t=2,...,N.

One thing to note about the Kalman filter is that it only takes into account the past value. It does not use any information about the future observations (i.e., x_{t+1} has no effect on x_t, but, x_t is used to calculate x_{t+1}). Since we actually have all of the observations, we were also able to apply the Kalman Smoother, which uses all of the data, past and future, to get an estimate of the state.

Kalman Smoother

The Kalman smoother is an extension of the Kalman filter, in which an estimate of the state is made based on both past and future observations that were collected. This is the estimate of the state that we want as it takes into account all of the information that we have available. This new estimate will be represented as $x_{t|N}$. By using both the past and future observations at time t, our intention is to get a more informative estimate of the state at time t, as all of the information that was available in the data was used.

The algorithm for the Kalman smoother is as follows:

1. Run the Kalman filter and have estimates of $x_{t|t}$, P_t and M_t for all t = 0,..,N.

2. Starting with the estimates at time t=N, run the Kalman filter backwards through the data (N-1 to 0), with the following adjustments to the equations

   $K_t^* = P_t D P_{t+1|N} \sim$ the Kalman Gain

   $P_{t|N} = P_t + K_t^*(P_{t+1|N} - M_{t+1})K_t^{*T} \sim$ Smoother Variance

   $\hat{x}_{t|N} = \hat{x}_t + K_t^*(\hat{x}_{t+1|N} - D\hat{x}_t) \sim$ Smoother Estimate of the State

These estimates of the state from the Kalman smoother will be the values used for the analysis of the cores. This algorithm is simply running the Kalman filter in reverse time and using the state estimates from the Kalman filter as the "observations". As a result of this procedure, $\delta^{15}N$ values were estimated for every time point on the constant time line such that there is no longer any missing data. Comparisons of $\delta^{15}N$ values (between and within cores) will now be possible as all series are now populated with values on identical time lines with no gaps in the data. With techniques now in place to separate the process and observation error, all that remains is to get estimates of these two parameters (R and Q) using the Kalman filter likelihood.

2.2.3 Parameter Estimation

The first attempt to get estimates for R and Q was done using maximum likelihood estimation (MLE) using the Kalman filter. The log likelihood of the Kalman filter is as follows:

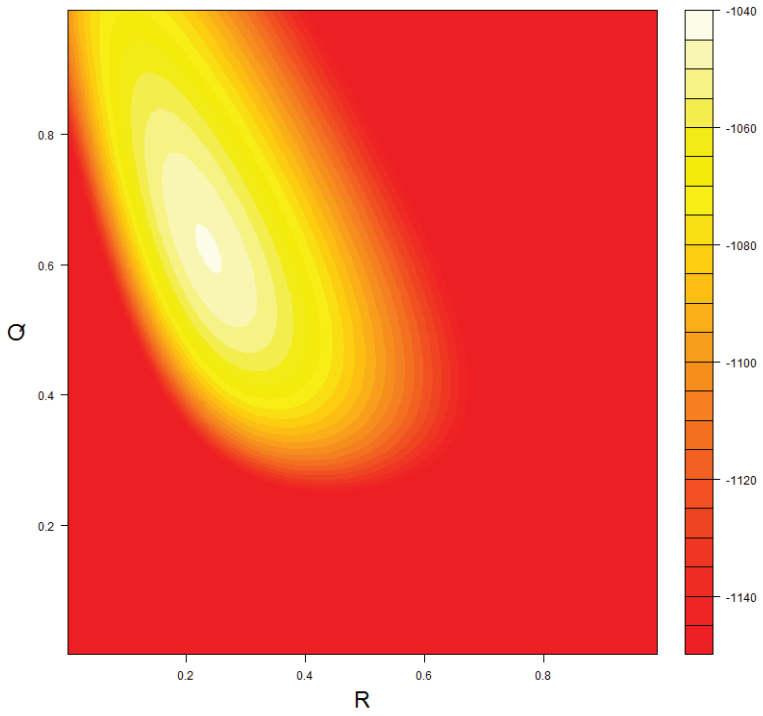
$$l = \sum_{t=1}^{N} ln(M_t + R) - (y_t - x_{t|t-1})^2(M_t + R) \qquad (8)$$

(Recall that M_t comes from equation 4.) For each core, the goal will be to find the values of R and Q that maximize the log likelihood function. To make sure that the Kalman filter likelihood was working properly, data were simulated under controlled conditions. The data were simulated using the following values: N=1,000, R=0.28 and Q=0.6. After the data were simulated, pairings of R and Q values were sent through the likelihood. There were 100 values for both R and Q, where both started at 0.01 and went as high as 1 by increments of 0.01. In total there were 10,000 R and Q pairs that were tested. After running through the 10,000 pairings the MLEs came out to be 0.241 and 0.621 for R and Q respectively. This is illustrated in Figure 2.11. Based on these results, as well as other numerical experiments with parameter recovery, it was concluded that the Kalman filter likelihood was working properly.

2.2.4 Joint Estimation of Error Terms

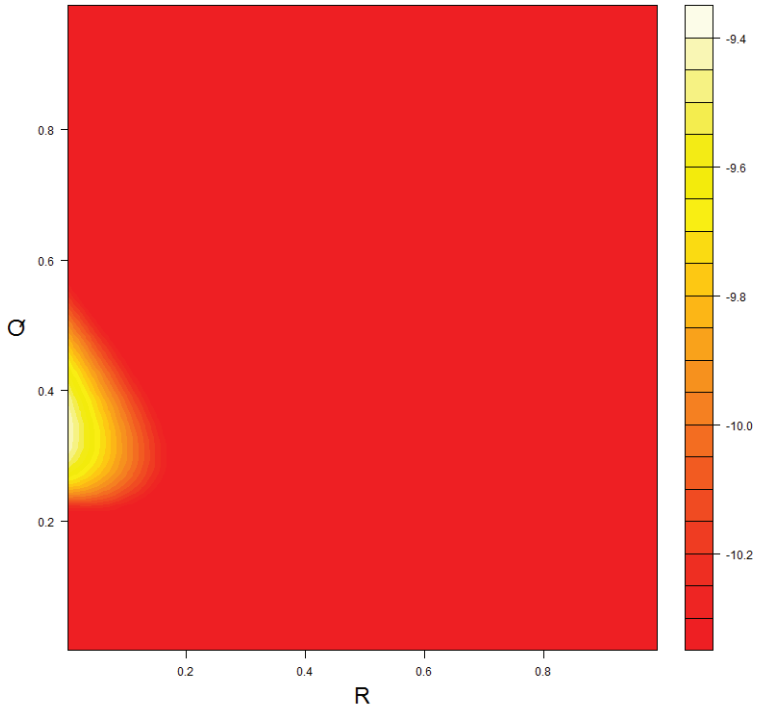
Knowing that the Kalman filter likelihood was working properly, it was time to run the $\delta^{15}N$ through the MLE estimation process using the R function optim as the optimizer. When this analysis was completed, we became aware of a troublesome problem relating to the parameter identifiability. In almost all cases, for all three time periods, the observation error gravitated towards zero. This caused the Kalman smoother to essentially connect the dots between $\delta^{15}N$ values because the MLE estimate for the observation error variance was so small. In a number of other cases the reverse happened where the process error Q was so low that the Kalman smoother almost completely ignored the observations altogether. Figure 2.12, which shows the joint likelihood for core CD 38-02, provides an example of this.

Figure 2.11: Simulated Joint Likelihood.



In this figure, we can see that the observation error (R) is on the zero bound. Even when bounds were placed on the observation and process error, in many cases the MLEs came out to be values very close to one of the bounds. Looking at the

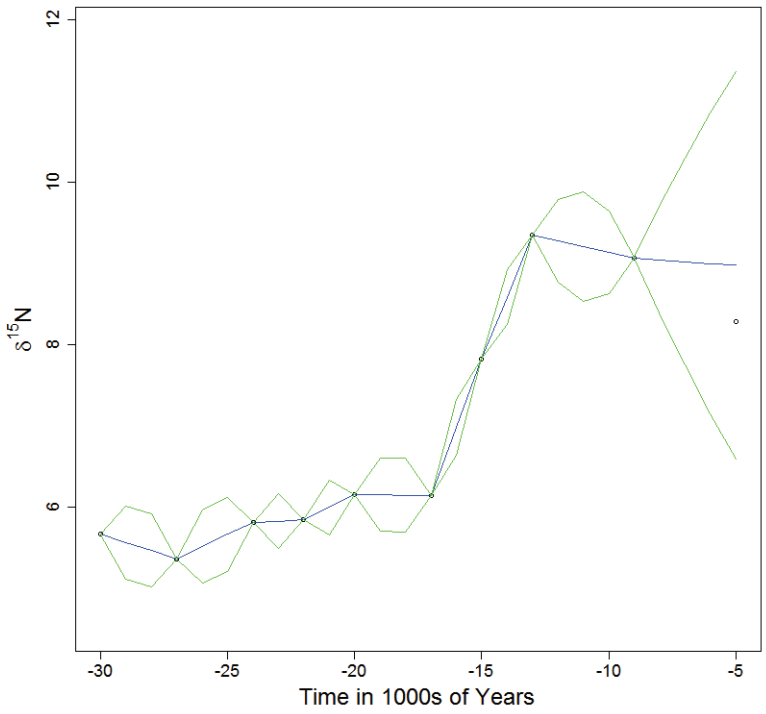
Figure 2.12: CD 38-02 Joint Likelihood.



Kalman smoother fit, based on these estimates in Figure 2.13, it is shown that when the observation error is under-estimated, the Kalman smoother just connects the observations. This is telling us that, in this particular core, there is no observation error and all of the error comes from the statistical process - which is very unlikely.

To test whether this was simply due to the smaller sample sizes, more simulations were conducted. When originally testing the Kalman filter likelihood, we used a sample size of 1000 observations and were able to get back our original parameters. Ten additional simulations were carried out with $Q = 0.6$ and $R = 0.28$ using six different values for N, including N = 26 (sample size for the IG1 time period), 41 (sample size for the Glacial time period), 60, 80, 100 and 121 (sample size for the IG2 time period). The sample sizes of 60, 80 and 100 were used to bridge the gap between the glacial sample size and the IG2 sample size. We found that the Kalman filter was not nearly as accurate as it was when N = 1,000, but the errors did not tend to go to the extremes like in the real data. As Figure 2.14 shows, Q and R took

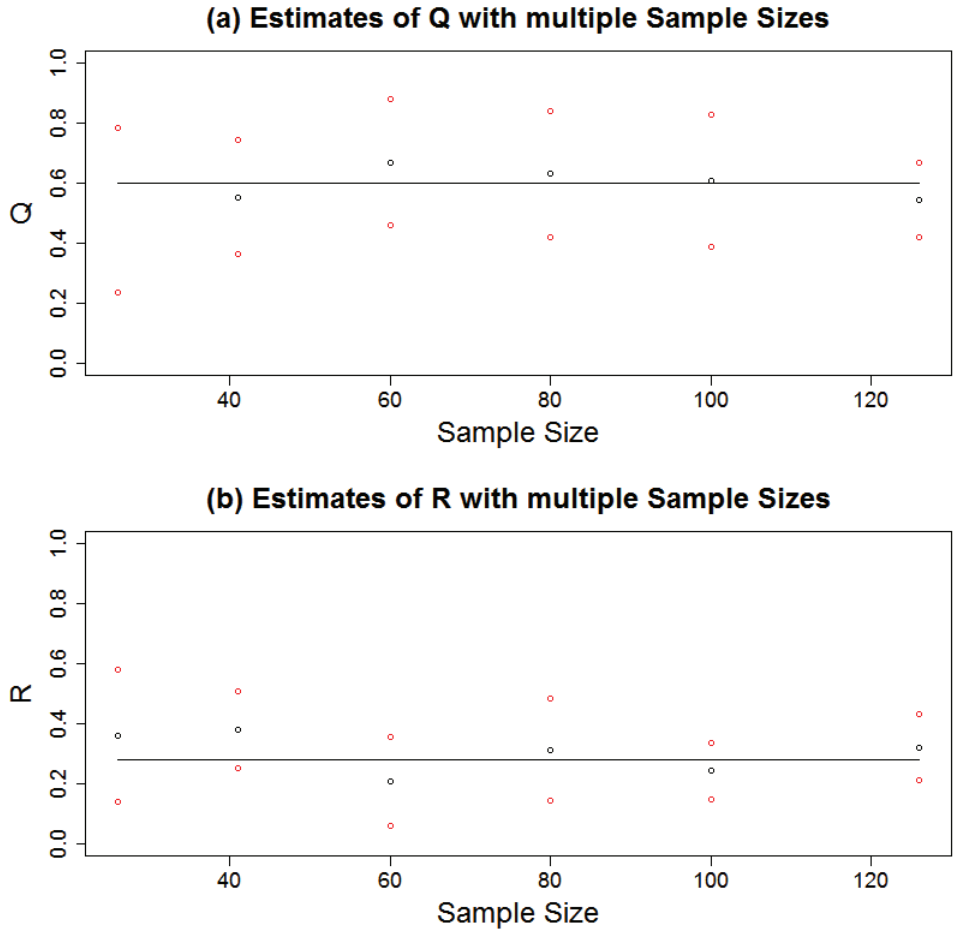
Figure 2.13: Kalman smoother estimates for CD 38-02 with confidence bounds based on the results of the joint estimation of Q and R. The black circles are the $\delta^{15}N$ values generated from the nearest neighbour search. The blue line is the Kalman smoother estimate, while the green lines are the credible regions for the Kalman smoother estimates.



on a wider range of values when N = 26 compared to when N = 121. This means that the smaller the sample size the less accurate the Kalman filter estimates will be. It is also important to note that these simulations were completed using no missing data. For an actual series of $\delta^{15}N$ values, it would rarely be the case that a full set of observations would be available for the core (*e.g.*, 26 observations for a core in the IG1 time line). This further complicates the analysis as it lowers the sample size even more than in the simulations, impacting the accuracy of the estimates.

We found that under completely controlled situations, where all of the statistical assumptions (such as normal errors) were met, the MLE estimation worked very well, even in smaller sample sizes. However, when moving away from simulated data, the estimates obtained were very low and were not consistent with the belief that the observation error was a minimum of 0.2. Possible explanations for these results

Figure 2.14: Plots of the results of the simulations of different sample sizes. The black horizontal line represents the true values of Q (a) and R (b). The black circles represent the means of MLE estimates of the 10 simulations at each sample size. The red circles are the mean +1 standard deviation and the mean -1 standard deviation based on the same created from the 10 MLE estimates at each time point.



are that, in the Kalman filter, the R and Q trade off with each other, or that the errors might not be normally distributed, as was assumed. As we felt we had a better understanding of the observation error, we determined to fix its value and only estimate the value of Q, the characteristics of which were of more interest to us.

2.2.5 Process Error Estimation

Recall that when looking at the histogram for the observation error and the log-normal distribution that was fit to it (see Figure 2.6), the mean was 0.10. This was

considerably lower than was expected (about half of the minimum value). However, our estimates were based on a one-size-fits-all approach, so it was possible that the calculated mean might be off. Rather than use the mean of the log-normal distribution, a simple average (0.28) of the observation error estimates calculated in Section 2.2.1 was used. This value seemed reasonable based on the prior belief that 0.2 was a minimum threshold. This results in the following equations for our state space model:

$$x_t = x_{t-1} + v_t, \quad v_t \sim N(0, Q)$$

$$y_t = x_t + w_t, \quad w_t \sim N(0, 0.28)$$

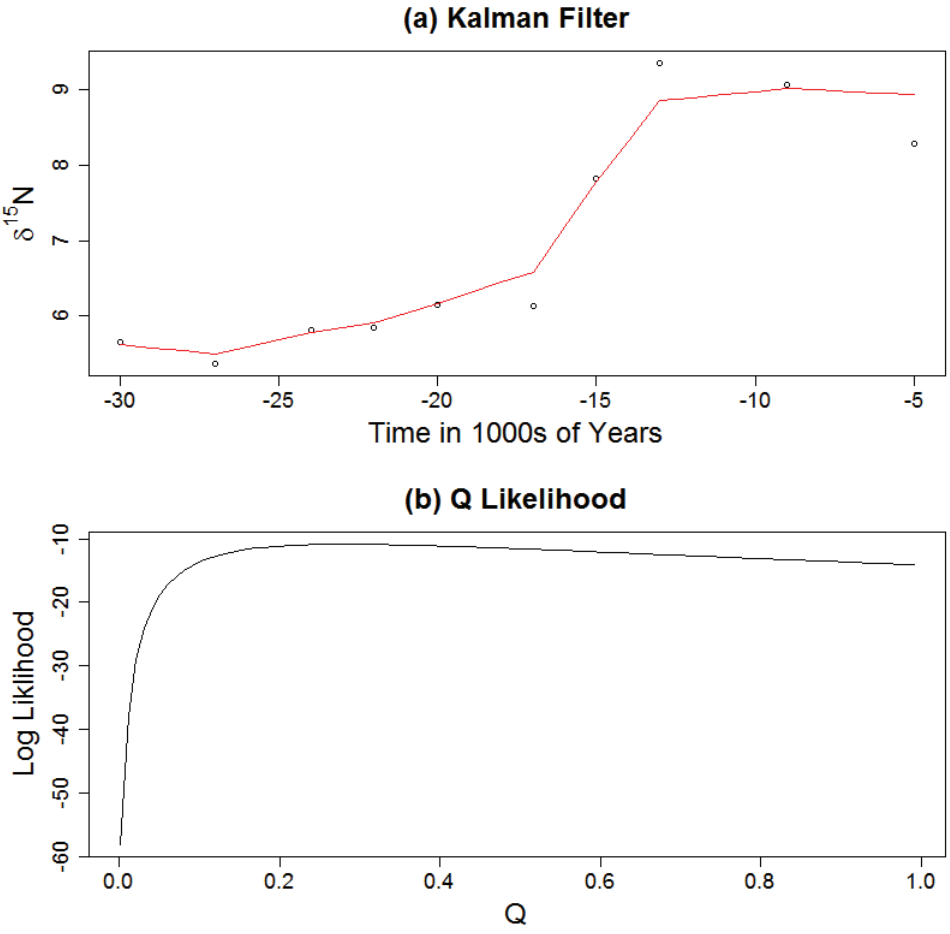
.

Having fixed the observation error, the only random component of the equations left to be estimated was Q. The analysis described above was repeated, but this time only trying to estimate Q. This resulted in an improvement in the cores where the Kalman smoother had followed the data too closely. Building on the previous example, when re-plotting the results for core CD 38-02, as seen in Figure 2.15, the Kalman smoother estimate no longer connects the dots but rather follows a slowly varying trend. However, fixing the observation error, did not solve the problem when the process error was underestimated. In cases where the process error had been underestimated (value near 0), those cores still had process error estimates that were very low and continued to ignore the observations. In these cores the Kalman smoother fitted a relatively straight lines that did not fit the data appropriately. In many cases where this problem persisted, the cores were removed from the analysis.

2.3 Metropolis Hastings

Even though using MLE while fixing the observation error gave reasonable results, it would still be more ideal to jointly estimate both the Q and R values. To try and get a good estimate of both, we decided to take a Bayesian approach. Bayesian techniques are driven by using prior information known about a parameter to assist

Figure 2.15: a) Kalman smoother estimates for CD 38-02 with confidence bounds based on the results of the estimation of Q while R is fixed at 0.28. The black circles are the $\delta^{15}N$ values generated from the nearest neighbour search. The red line is the Kalman smoother estimate. b) Plot of the log likelihood of Q for the core CD 38-02.



in the estimation process [9]. To this end, we put prior distributions on both Q and R, so the likelihood of the parameters were based both on the data (Kalman filter likelihood) and the parameters prior likelihood. This would help us investigate whether adding a prior to the Q and R values would break the dependence structure found in the maximum likelihood estimation. This was our last attempt to jointly estimate the Q and R values.

The Bayesian technique used for this analysis was a Markov Chain Monte Carlo (MCMC) algorithm, specifically the Metropolis Hastings. The Bayesian method requires a prior (π) distribution of the parameters being estimated to weight how likely

the values of Q and R are based on the prior information we know about them. In this study we used a relatively uninformative prior, Uniform(0,1), for both Q and R. Briefly, at each point in the Markov chain, we compared the likelihood of our current parameters, $\theta = (Q, R)$, to the likelihood of a proposal set of parameters, $\theta^* = (Q*, R*)$, and randomly chose one with probabilities based on the ratio of their likelihood. This procedure was done 6500 times, with the first 2500 dropped from the data set as a burn in. The Q and R values at each of the remaining 4000 iterations were stored. Distributions for both Q and R were generated from these stored values, and the median was used as the estimate for the parameters (Q, R). A detailed algorithm for the MCMC procedure can be found in the Appendix B.

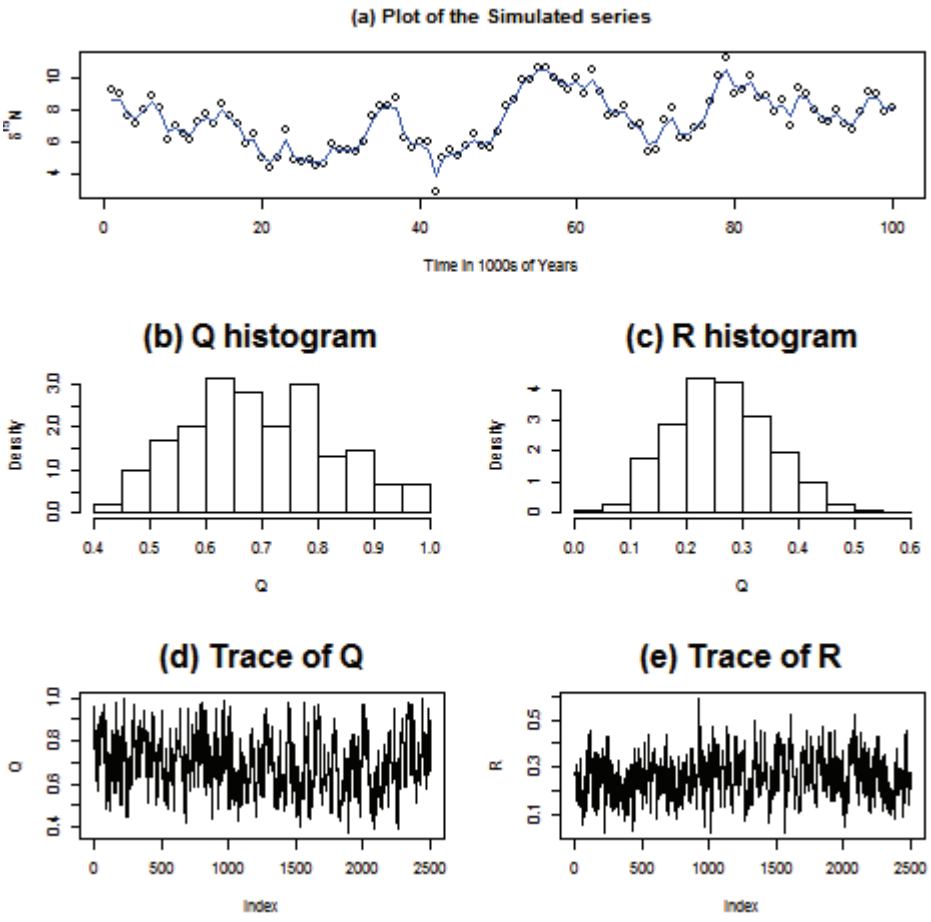
Two problems arose when using this MCMC method. The first was that the chains of Q and R were each highly correlated with themselves. This was due to θ^* being so close to θ. To decrease the amount of correlation between adjacent iterations, we looked at the auto correlation function to find the lag at which the correlation dropped below 0.2 (we called this lag value t^*). Once that lag value was found, we sampled out every t^{*th} value of the chain for Q and R separately. This left us with uncorrelated realizations.

The second problem was that the acceptance ratio was either too low or too high in many cases. The acceptance ratio is the ratio of the number of times you accept θ^* over the total number of iterations of the chain. This was a byproduct of trying to automate the analysis due to the number of cores we had. We set the proposal to be the same for all cores, $\theta^* = \theta + N(0, 0.05)$. However, for many cores, this increase/decrease jump was either too small or too large. Because of this, an adaptive MCMC was created where the jump size could increase or decrease depending on how far the acceptance ratio was from the ideal acceptance range (0.15 - 0.4). A more detailed algorithm is located in Appendix B, as well as a plot that shows the difference in the traces (i.e., of how Q and R change through progressing iterations) between the two versions.

The algorithm that was used was created by Hartig [10, 11] and altered by the

author. The original algorithm was modified in three ways: (i) an adaptive model was created; (ii) it was based on the Kalman filter likelihood instead of a regression likelihood; and (iii) the priors were changed to uniform distributions between 0 and 1 (the prior distributions for Q and R were independent of one another).

Figure 2.16: Simulated MCMC Results. a) Plot of the smoother estimate of the $\delta^{15}N$. b) Histogram of the thinned Q chain from the MCMC analysis. c) Histogram of the thinned R chain from the MCMC analysis. d) Trace of the Q chain from the MCMC analysis. c) Trace of the R chain from the MCMC analysis.



To confirm that the MCMC algorithm was working properly, a chain was run on simulated data of length 100 using the same Q and R values from previous simulations: Q=0.6, R=0.28. As shown in Figure 2.16, we were able to obtain a distribution around our preset values, which indicated the MCMC was working as intended and that we were able to use the MCMC to jointly estimate Q and R.

2.3.1 Joint Estimation of Error Terms

When trying to jointly estimate the Q and R parameters using MCMC, once again the same problem arose where Q and R were being underestimated, as can be seen in Figure 2.17. When looking at the $\delta^{15}N$ series plot itself, we can see that the blue line indicating the Kalman smoother estimate is very reactive to the data points. It is not completely connecting the dots but it still jumps up in large spikes when a data point is present. When looking at the histogram it is shown that both Q and R have very small errors associated with them (under 0.1). This is showing the same problem as with the MLE estimation, where we were getting estimates of the measurement and process error which were too low to be believable.

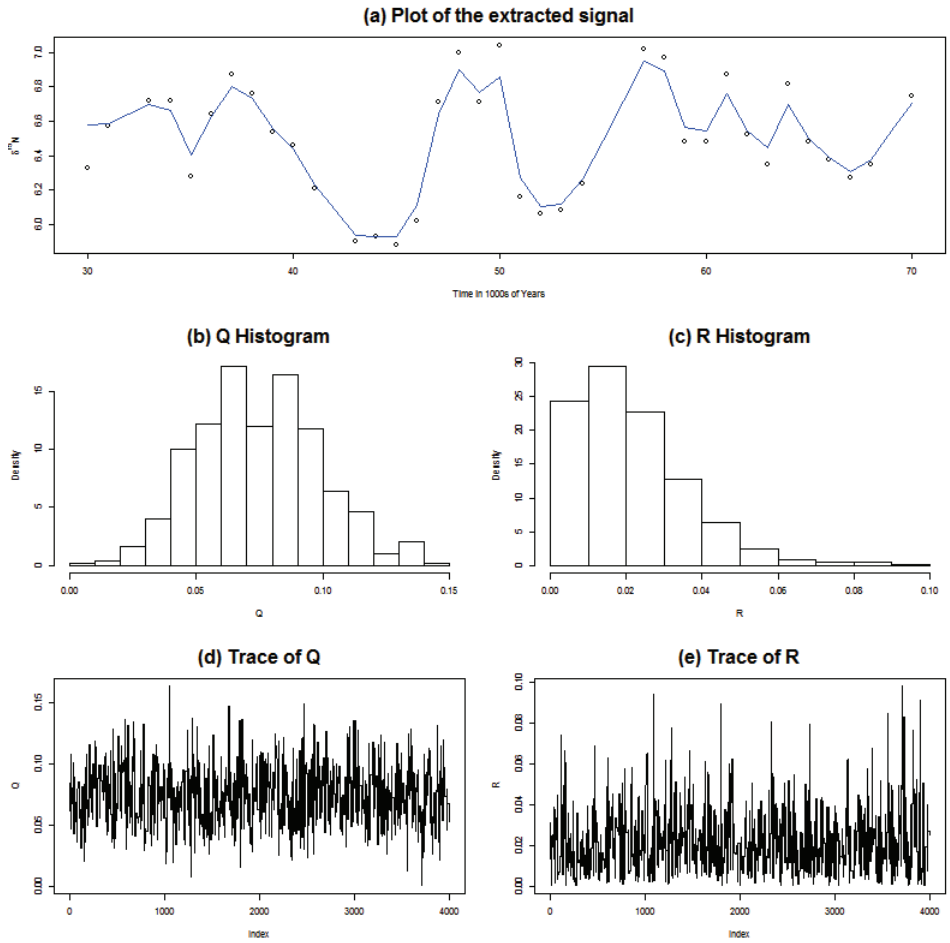
2.3.2 Process Error Estimation

Given that we could not break the dependence structure using the MCMC method, we decided to see if using this method for estimating only Q would fix some of the instances where maximum likelihood underestimated its value. To do this, observation error variance was again set to 0.28 and the MCMC chains were run again. The results were practically identical to the MLE estimates of the state. Consequently, we decided to use the MCMC estimate of the state as the de-noised $\delta^{15}N$ signals for the multivariate analysis.

2.4 Acknowledgment of Unresolved Issues

The goal of this chapter was to extract the true signal of $\delta^{15}N$ from the noisy observations in the core samples. We outlined the methods that were used to extract the signals. However, a number of the problems we encountered were not fully resolved. First, a number of cores in each time period still had underestimated process errors forcing the smoother to almost ignore the observations. Second, by doing the nearest neighbour search it should be noted that we lost some information from the observations by grouping them together (averaging). Finally, by fixing the measurement error at 0.28 we added in a small amount of bias. Since all cores were not measured

Figure 2.17: Real data MCMC Results. a) Plot of the smoother estimate of the $\delta^{15}N$. b) Histogram of the thinned Q chain from the MCMC analysis. c) Histogram of the thinned R chain from the MCMC analysis. d) Trace of the Q chain from the MCMC analysis. c) Trace of the R chain from the MCMC analysis.



the same way and were collected by different researchers, it is unlikely the cores all share the same observation error. Still, we made this assumption so that we could break the dependence structure and carry on with the modeling. Acknowledging that these problems persist, we decided to move on to the multivariate analysis.

As a result of the univariate analysis described in this chapter, we had a processed data set that was ready for the multivariate analysis. That is, all of the time series in the data set: represent the extracted signal of $\delta^{15}N$; were on equivalent time scales (within the designated time periods); and no longer had any missing values of $\delta^{15}N$.

Chapter 3

Multivariate Analysis

In this chapter, we aimed to get a better understanding of the changes in the nitrogen cycle (mainly fixation and denitrification rates) over three time periods using multivariate analysis to compare and analyze the $\delta^{15}N$ signals that were estimated in Chapter 2. Such an analysis is useful because if we are able to find distinct signals in the $\delta^{15}N$ values, either globally or in specific regions, it might be an indicator that an historic event occurred that impacted the nitrogen cycle in the ocean. This, in turn, could lead to future research on why certain events had a larger impact on some regions compared to others (*i.e.*, some regions are less drastically effected by ice ages).

We looked at all of the cores in each time period to characterize variability of $\delta^{15}N$ on the regular interval time scale. This was done to determine whether there was a signal present in all or most of the cores, regardless of their geographical location, which might be an indicator of global change in the marine nitrogen cycle. We also wanted to group cores based on the smoothed $\delta^{15}N$ values obtained in Chapter 2. This could indicate regional variations in the marine nitrogen cycle that might not have been present on a global scale. In their paper, Tesdal *et al.* [3] compared the means of cores in 100 km predefined circles and found that the cores in those circles had similar means of $\delta^{15}N$. We used the entire series of $\delta^{15}N$ values on our constant time line, rather than means, to group the cores with no distance limitation. This would result in findings based on statistical characteristics in the full series (*e.g.*, rapid fluctuations instead of slow changes in $\delta^{15}N$), rather than groups within a defined distance.

3.1 Methods

3.1.1 Correlations vs. Distance

Before we started with our cluster analysis, we wanted to look at the similarity of cores that were close in geographical distance. The geographical distance was calculated using the statistical package fossil in R [12]. If cores in closer regions have the same $\delta^{15}N$ variation, it would be a strong indication that the variation was related to a change in the regional marine nitrogen cycle. To determine whether cores that were geographically close had similar characteristics, the correlations between two cores (at a time) were plotted against the global distance between them. One of the shortcomings of using global distance is that it does not take into account whether (or not) a land mass is crossed. For example, two cores could be really close together, but separated by a land mass and, as a result, be in two completely different oceanographic basins. In such cases, we might not expect the cores to be similar, despite their geographic proximity. For this analysis, the correlations were based on the smoothed estimates of the state from the MCMC analysis at each time point on our constant time line (as described in Chapter 2). It is likely that the smoothing process, which removes many of the small fluctuations in $\delta^{15}N$, will increase the correlations between the series. Since this analysis is concerned with cores that are close together, we only looked at cores that were 2,000 km or less apart from each other. By picking such a short distance we hoped to limit the number of comparisons that crossed land masses.

3.1.2 Principal Components Analysis

Principal components analysis was used to determine whether any distinctive patterns in $\delta^{15}N$ were present in the cores over the time periods. In this analysis we were looking for global signals that could define the respective time periods in their entirety, not necessarily by geographical regions. The equations and algorithms in this section came from the book "Applied Multivariate Analysis" by Johnson and Wichern [13].

Principal components analysis is a statistical technique used to reduce the number

of variables in the model, in our case cores, to a smaller number of variables that are an orthogonal linear combination of the cores. By doing this, we hoped to be able to explain the majority of the variance, in all of the cores, in just a few variables. That is, if there was a dominant trend, such as the decreasing $\delta^{15}N$ as you go back farther in time found by Tesdal *et al.* [3], we might be able to account for the majority of variances in all of the cores with that one trend.

For the principal components analysis, we first found the eigenvalues λ_i and eigenvectors (e_i) of the correlation matrix, ρ, obtained for the p cores in a given time period. The correlation matrix was used instead of the covariance matrix because we wanted the analysis to be completed on the standardized variables. These eigenvectors are orthogonal linear combinations of the p cores that make up the principal components, as shown in the equation below:

$$Y_i = e_i X = e_{i,1}X_1 + e_{i,2}X_2 + ... + e_{i,p}X_p \quad (9)$$

where;

X = is a n x p data matrix of the $\delta^{15}N$ series for the cores in that time period;

Y_i = the i^{th} principal component in a n x 1 vector;

e_i = the i^{th} eigenvector (n x 1).

The proportion of the population variance accounted for by the i^{th} eigenvector is equal to the i^{th} eigenvalue divided by p. In order to find out which principal components should be used, we ordered them based on their eigenvalues, from largest to smallest. That is, the principal component with the largest eigenvalue is principal component 1 (PC1) and the principal component with the smallest eigenvalue is principal component p. We used the k principal components that accounted for the majority of the total variance.

Principal components analysis gives us a number of interpretable results that could assist us in understanding how the nitrogen cycle has changed in the past. The first of these is the loadings which are located in the eigenvectors. The loading for the i^{th} principal component are the coefficients that designate the weight to which each core

contributes to the score of the i^{th} principal component. The score is the weighted average, at a given time point, of the cores that cover the time period of interest. By getting the score values for the i^{th} principal at all time points we can identify the signal of $\delta^{15}N$ that the principal component represents. As we are using the correlation matrix to calculate the eigenvalues and the loadings, our principal component analysis is being done on the standardized variables. As a result, the loading values indicate the correlation or importance of the core to the i^{th} principal component. This means that cores with a large positive loading for principal component 1, are highly correlated with principal component 1.

3.1.3 K-means Clustering

We used the k-means clustering algorithm to define groups of cores that had similar $\delta^{15}N$ signals within each time period. Our goal was to create core clusters that could characterize changes in the marine nitrogen cycle distinct to particular regions. The k-means clustering algorithm is a simple technique used to group items (cores) into distinct clusters by minimizing the Euclidean distance between the cores and the mean of the cluster they are in. We grouped the p cores that covered a given time period into k clusters based on the smoothed estimates of $\delta^{15}N$ values at each time point on the constant time line. For example, in IG1 there were 26 time points on the constant time line: 30,000, 29,000,..., 5000 years ago. We grouped the cores into clusters that minimized the sum of the squared Euclidean distance of $\delta^{15}N$ from the mean $\delta^{15}N$ at each of the 26 time points.

The algorithm for this is as follows:

1. Assign the p cores into k clusters randomly, this will be our starting point.

2. Select the first core (this core will be called the "active core"). Place the active core in each cluster in successive iterations and calculate the summed distance of $\delta^{15}N$ from the mean $\delta^{15}N$ value. After all of the squared differences have been calculated for all of the iterations, assign the active core to the cluster in which the summed distance was the smallest for the active core.

3. If the active core was moved, recalculate the mean of the $\delta^{15}N$ value in the active core's new cluster and recalculate the mean of the $\delta^{15}N$ value in the active core's old cluster. If the active core was not moved, recalculation is not necessary.

4. Repeat procedure with the next core.

5. Upon reaching core p, go back to core 1 and start the loop all over again.

The algorithm stops when a full loop is completed, from 1 to p, with no further reassignment of cores among the clusters.

To determine the optimal number of clusters that should be used for a given time period, we tried to minimize the sum of squares within the clusters. The smaller the sum of squares, the more alike the cores within the clusters are. A scree plot was created to plot the sum of squares within clusters against the number of clusters. In each scree plot we looked at where the bend or elbow in the plot occurred (the point at which the sum of squares within clusters starts to level off), and used that value as the optimal number of clusters. The equations and algorithms in this section came from the book "Applied Multivariate Analysis" chapter 8, by Johnson and Wichern [13].

3.2 Results: Interglacial 1 (IG1)

In this section we describe the results of analyses using the three multivariate techniques described above on the IG1 time period. We wanted to see if there was a dominant trend that characterized the global marine nitrogen cycle over this time period, whether neighboring cores had similar signals of $\delta^{15}N$, and to determine the optimal set of clusters for the cores in this time period. Figure 3.1 shows the locations of the 67 cores included in the analysis for this time period.

3.2.1 Correlation vs. Distance

Looking at the correlation vs distance plot for IG1 (Figure 3.2), it shows that even cores that are 500 km apart are highly correlated with each other. For this correlation

Figure 3.1: Plot of the locations of the 67 cores that were used in the multivariate analysis of the $\delta^{15}N$ values in the interglacial 1 time period.

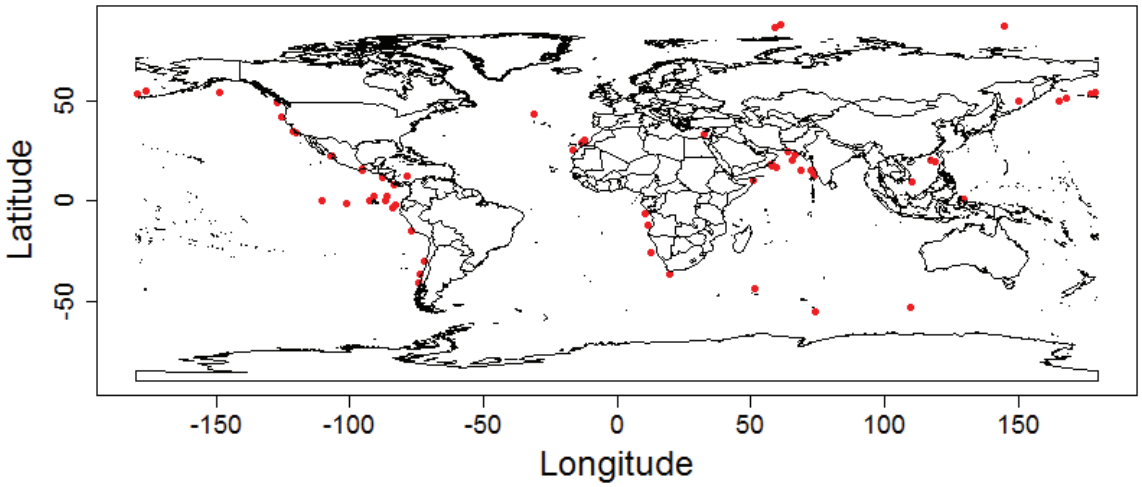
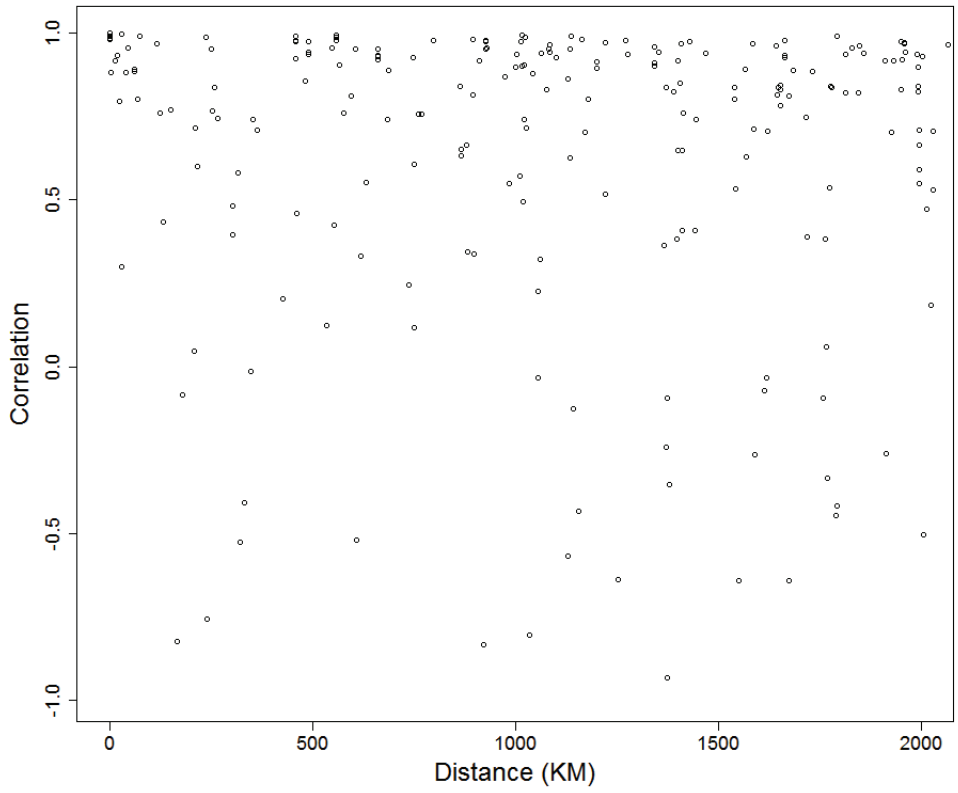


Figure 3.2: Plot of the correlations between all pairing of cores that are within 2,000 km of each other in the interglacial 1 time period. On the X-axis is the distance between the two cores and on the Y-axis is the correlation based on the 26 smoother estimated observations of $\delta^{15}N$.



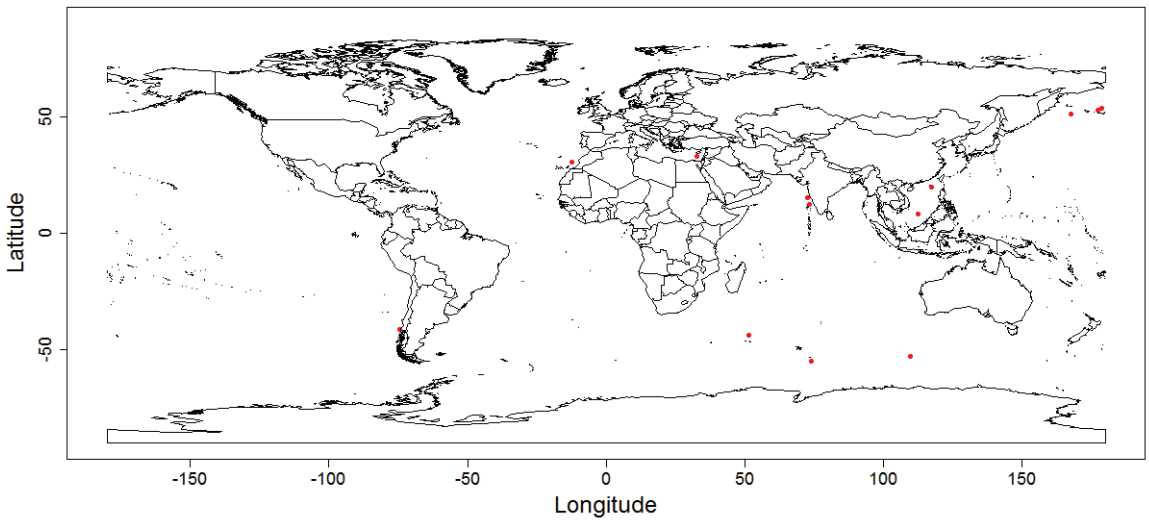
analysis we did not standardize the series before calculating the correlation because we were interested to see if overall trends were similar between cores. As Tesdal *et al.* [3] found, the $\delta^{15}N$ values increase over time, which could mean that the high correlations are a product of this increase and not a product of short term fluctuations in the series. As the distance between cores gets larger, the correlation values start to spread out more and more, so that the majority of the cores within 100 km of each other have positive correlations above 0.7. Once you move past the 100 km mark, the correlations take on a wider range of values, but still maintain some highly correlated pairings. If we expand the distance beyond 500 km, some high correlations remained, but there was still a wide spread in correlations overall. Comparison of distances this large would cross land masses and geographical regions. Since the goal here was to see if cores that were closer together were highly correlated, these large distance comparisons turned out to be of less interest, despite the fact that, in some cases, even cores that were over 5,000 km apart had correlations values over 0.9. It is important to recall that global distance is not the same as water distance, and we did not try to distinguish between cores that were 1,500 km apart in the same ocean versus those that were 1,500 km apart but which crossed land boundaries. Our findings support the belief that high correlations between cores that are substantially distant were based on the increasing trend in $\delta^{15}N$ values rather than the short term fluctuations of the series.

Looking at Figure 3.2, there are a surprisingly large number of very strong negative correlations. Digging deeper to investigate the potential cause, we saw that in 13 of the cores the strong negative correlations were a result of decreasing $\delta^{15}N$ values as time moved towards the present. This is consistent with the results found in the data visualization completed in Chapter 2. In that analysis it was found that a number of cores had a negative (decreasing) slope with respect to time and $\delta^{15}N$. There was one pairing of cores, located in the South China Sea, that had a negative correlation under -0.8 (see Figure 3.2). These two core series were ODP 1144 and SO95 GIK17924-3. Checking on these two cores it was found that these two cores

had a very sharp increase over the time period (ODP 1144) and a very small increase over the same time period (SO95 GIK17924-3). All of the $\delta^{15}N$ values did not differ by more than 0.28 so we cannot draw distinct conclusions from this finding. In fact, all of the negative correlations shown in Figure 3.2 were associated with a core with a negative trend over this time period.

Over the IG1 time period, the majority of cores have an increasing trend where $\delta^{15}N$ is increasing as time moves forward, resulting in large positive correlations between most cores. However, cores like ODP 1144, that have decreasing $\delta^{15}N$ as time passes, are the cause of the negative correlations. It is notable that these cores do not seem to be randomly placed throughout the oceans. Plotting the cores with a negative slope, as in Figure 3.3, it can be seen that these negative trends (slopes) can be found in the Southern Ocean and a number of smaller negative trends can be found off the east coast of Russia. Even though the cores with this negative trend

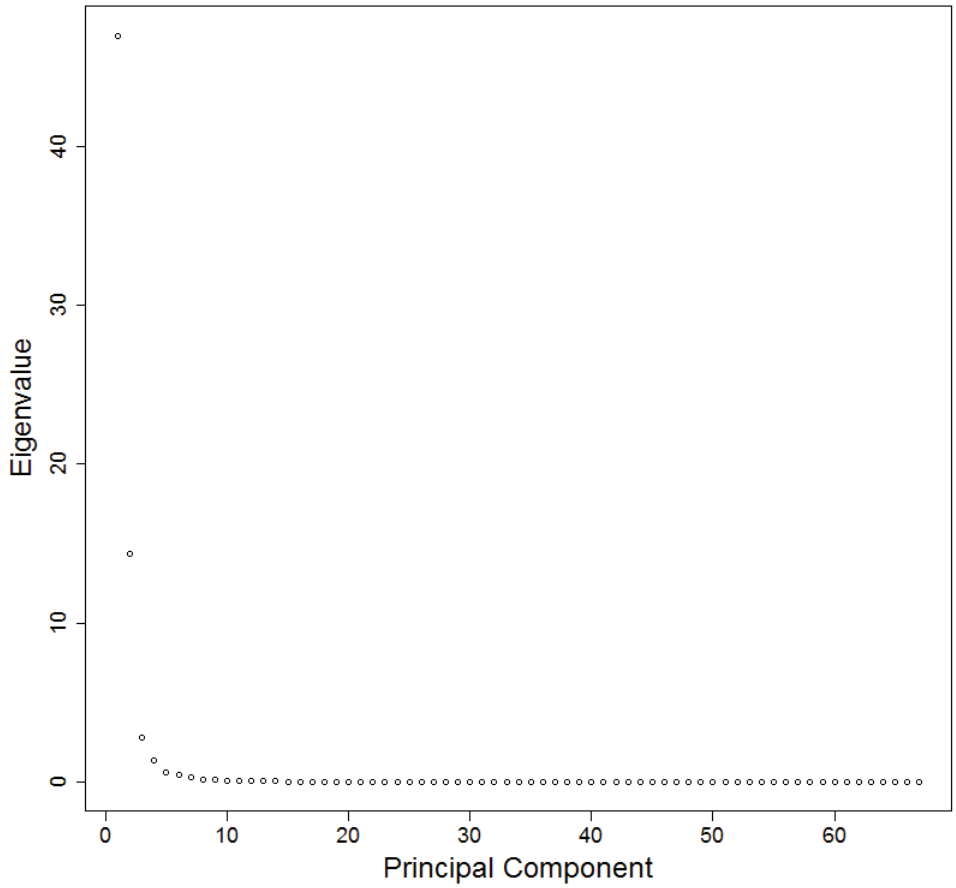
Figure 3.3: Locations of the cores with negative trends (decreasing $\delta^{15}N$ values as time moves forward) in the interglacial 1 time period.



are not completely grouped together in one location, they do tend to appear within close proximity to another core sharing the same feature. This could indicate that there is a natural process at work in these select locations that might be causing this

abnormality.

Figure 3.4: Scree plot of the eigenvalues from the principal component analysis for the interglacial 1 time period, plotting the eigenvalues from largest to smallest.

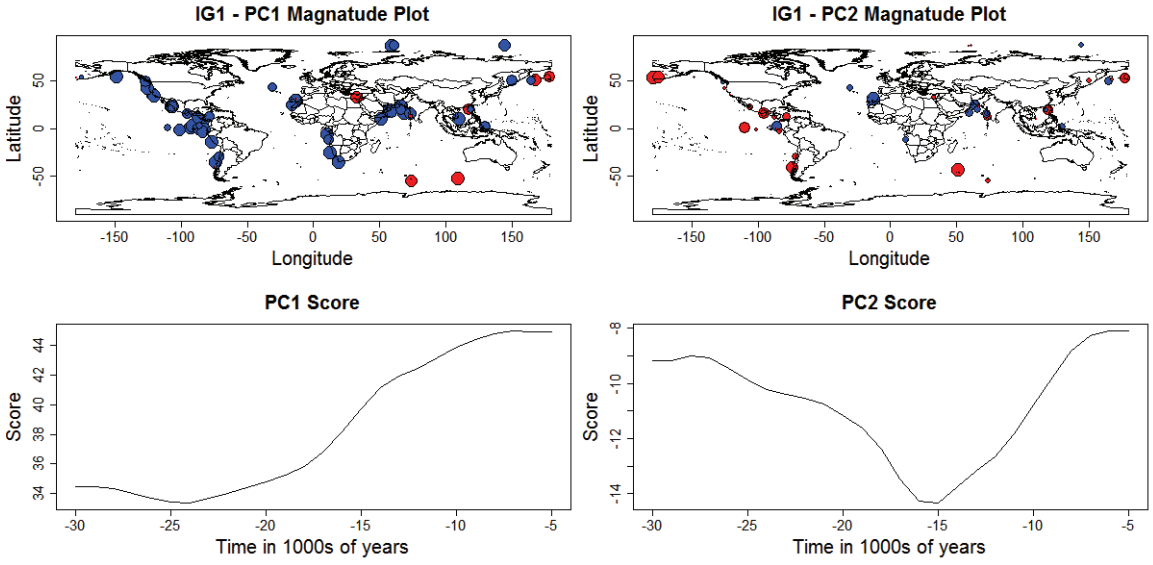


3.2.2 Principal Components

In order to determine how many principal components are required to account for the majority of the covariance structure, a scree plot was created. Looking at the eigenvalues for the IG1 time period, it was found that only the first two principal components seemed to account for the majority of the population variance. This can be seen in Figure 3.4, where there is a significant drop off after the second eigenvalue. In total, the first two principal components accounted for 91.5% of the variability, with 70% coming from the first principal component and 21% coming from the second.

When looking at the score plot of the first principal component, it shows that the dominant signal in the IG1 time period is a large increase in $\delta^{15}N$ starting about 25,000 years ago (Figure 3.5, panel c). When looking at the plot of the magnitudes of the principal components, the predominance of large positive loading values (represented as blue circles) indicates that the majority of the cores share this characteristic (Figure 3.5, panel b). This is in agreement with the data visualization conducted in Chapter 2 and with the finding by Tesdal et al. [3] that $\delta^{15}N$ decreases in value the farther back in time you go. All of the cores shown to have a negative slope in the Chapter 2 data visualization, were highly negatively correlated with the score plot for principle component 1. Looking at the locations of the cores negatively correlated with the first principal component (PC1) we can see that they are grouped in the Southern Ocean and off the east coast of Asia.

Figure 3.5: Plots of magnitudes of the loadings for the first (a) and second (b) principal component respectively for the interglacial 1 time period. The size of the circle represents the magnitude (larger circle = larger coefficient) and the colour represents the sign of the magnitude (blue = positive, red = negative). Score plots for the first (c) and second (d) principal components.



The second principal component has adistinct "V" shape (Figure 3.5, panel d). It decreases from 30,000 years ago, reaching a minimum at 16,000 years ago, then

begins to increase from 15,000 years ago, reaching its maximum at 5,000 years ago. This principal component accounts for 21% of the variability in the core, excluding the first principal component. Based on the principal component analysis from IG1 it can be said that the most dominant signal was a spike in $\delta^{15}N$ values at around 25,000 years ago, which was present in 55 of the 67 cores.

3.2.3 K-means Clustering

The principal component analysis gave us a good understanding of the dominant signal in the interglacial 1 time period; one of an increasing $\delta^{15}N$ value starting around 25,000 years ago, with a few cores having a delay or head start on this increase. With this knowledge, we attempted to group the cores for the IG1 time period based on their smoothed $\delta^{15}N$ values.

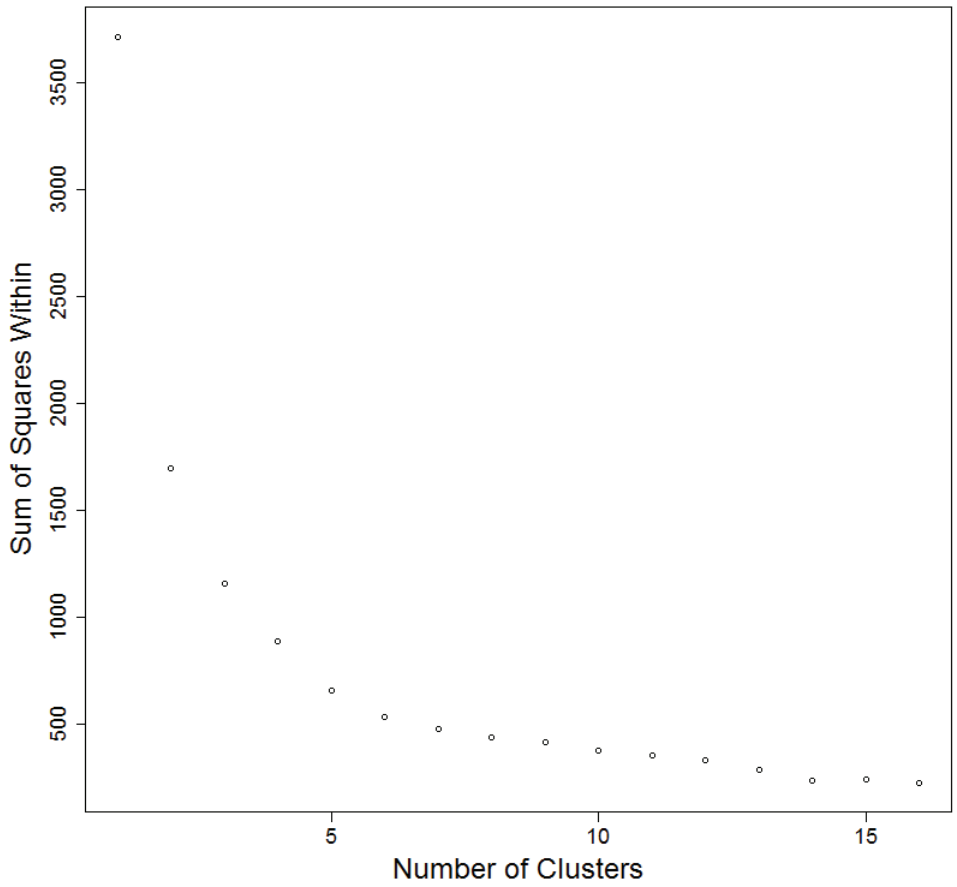
$\delta^{15}N$ Magnitude Analysis

As a first attempt we simply tried to group the cores based solely on the smoother estimates. Before doing any analysis based on the clusters, a scree plot of the sum of squares within groups was done to determine the optimal number of clusters. Looking at Figure 3.6, we see that the "elbow" occurs at five clusters. Therefore, we used five clusters for the analysis.

Figure 3.7 shows the locations of the cores that are within each cluster. The only cluster that appears to be distinct to a geographic region is cluster one, which is on the southwest coast of South America. In order to determine why this particular cluster stood out we plotted the series within each cluster.

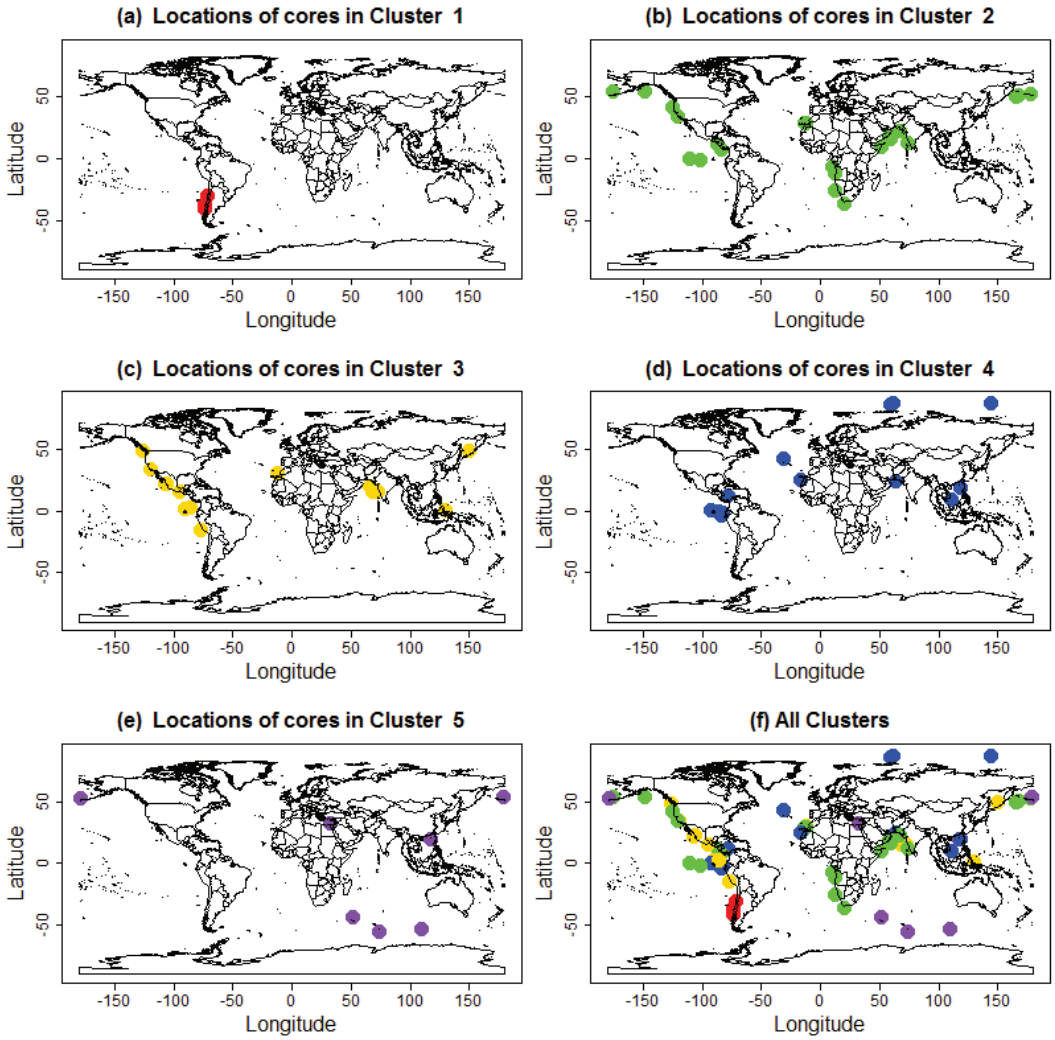
Applying the K-means algorithm to these series, cores with similar magnitudes of $\delta^{15}N$ values were grouped rather than cores with similar statistical characteristics (*i.e.*, trends). This can be seen easily in cluster 1 where there are only three cores (Figure 3.8). When comparing the red and black series, we see that the increase in $\delta^{15}N$ starts 25,000 years ago for the black series and about 20,000 years ago for the red series. Also, when comparing the green series to the red and black series, it

Figure 3.6: Plot of the sum of squares within cores by numbers of clusters for the interglacial 1 time period.



can be seen that while the red and black decrease over the last 5,000 years the green series does not. From this we can conclude that the cores in cluster 1 were grouped based on the sheer magnitude of the measurements of the $\delta^{15}N$ values - *i.e.*, they all have larger $\delta^{15}N$ values than the other cores shown in Figure 3.8, and they are the only 3 series with $\delta^{15}N$ values consistently above 8. Cluster 5, however, does seem to show some similarity in statistical characteristics as it grouped all of the cores that had the negative trend. Since the K-means algorithm only grouped the cores based on magnitude, it was decided to standardize all of the series by subtracting the mean and dividing by the standard deviation.

Figure 3.7: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the interglacial 1 time period.

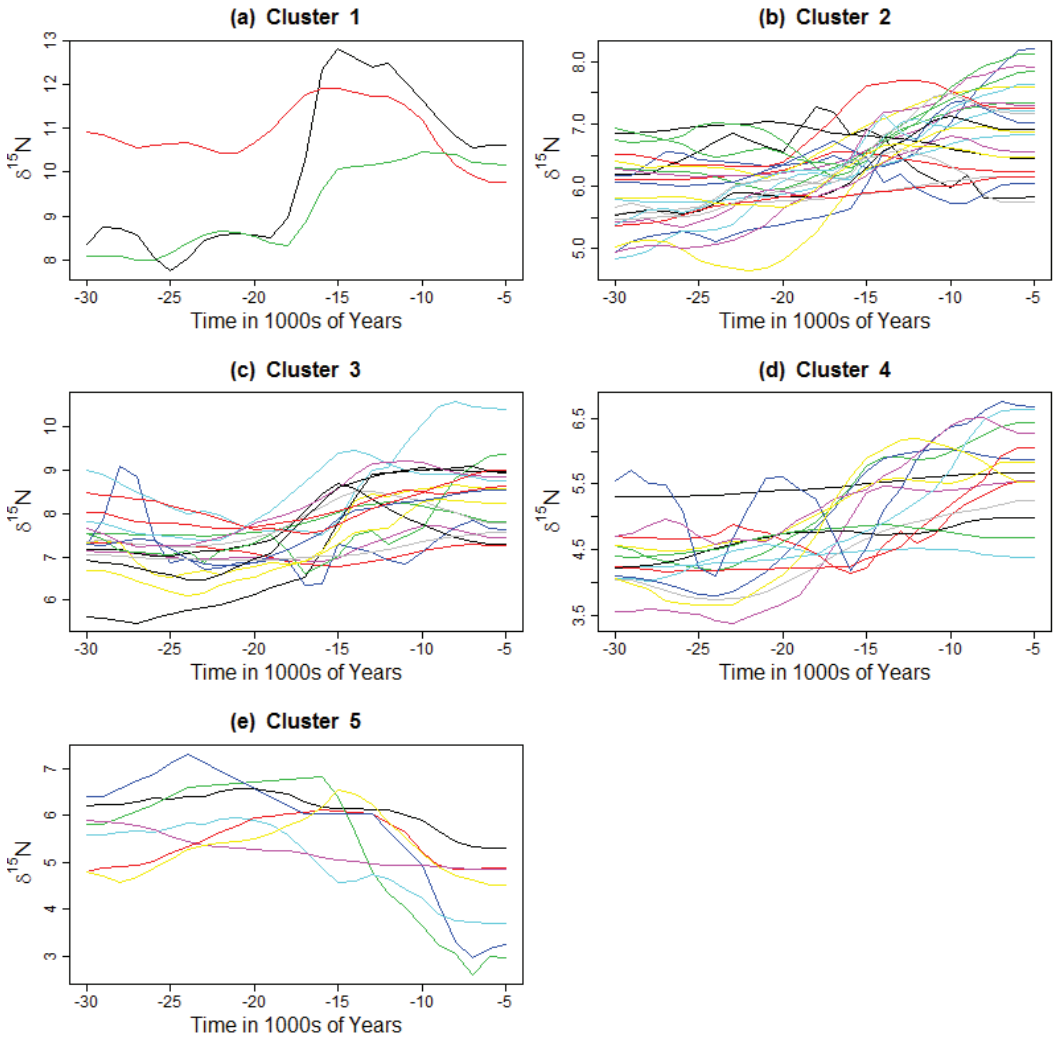


$\delta^{15}N$ Trend Analysis

When the standardized series were used in the k-means clustering analysis, the optimal number of clusters also turned out to be five (see Figure 3.9). When looking into the series plots within each cluster, the k-means algorithm clustered the cores together based on their shape and not by their magnitude.

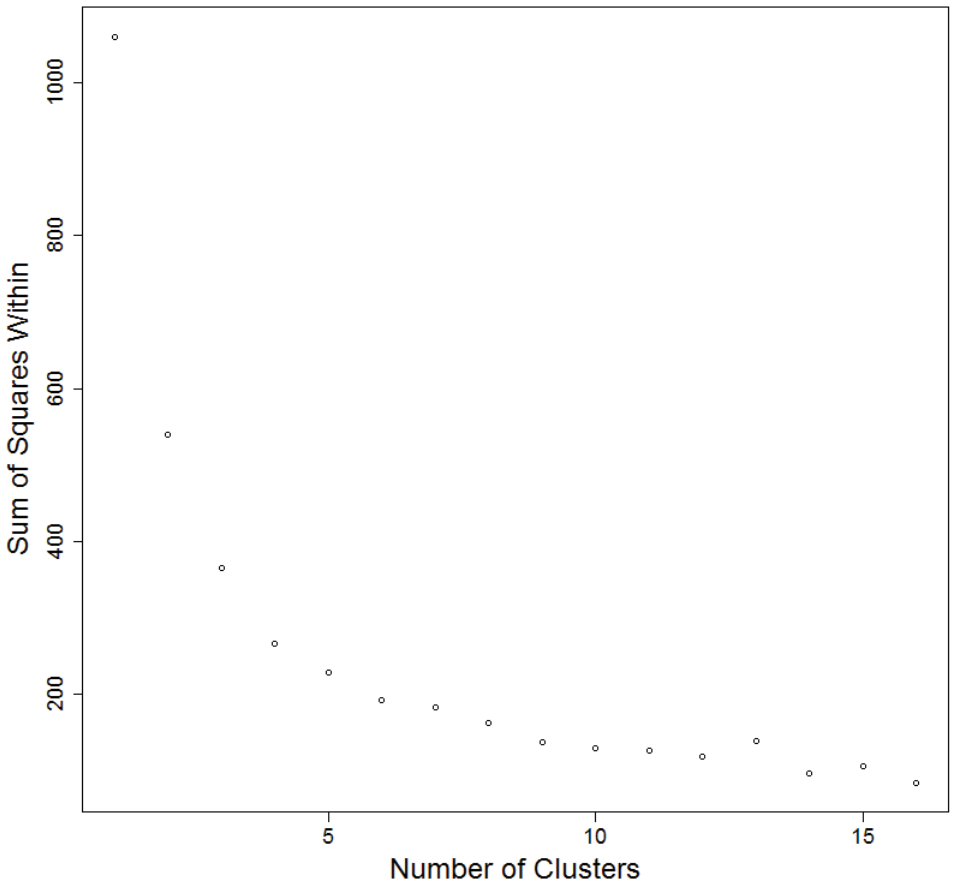
The cores in cluster 1 are all located in the Pacific Ocean, with a high concentration off the west coast of the Americas (Figure 3.10). This group of 11 cores started to spike at around 20,000 - 17,000 years ago (Figure 3.11). Another similar characteristic

Figure 3.8: Plots of the $\delta^{15}N$ series that fall within each of the five defined clusters in the interglacial 1 time period.



in this group is that the $\delta^{15}N$ values start to decrease at around 10,000 years ago. Cluster 2 includes 5 cores that seem odd in that they do not tend to increase or decrease and they are spread out all over the world. These clusters do not fluctuate much and they do not have the strong spike that is prevalent in many of the cores in this time period. The third cluster had a total of 13 cores. The signals of $\delta^{15}N$ were similar to those in cluster 1 in that a spike was seen in the majority of the cores. However, in this cluster the spike occurred at around 15,000 years ago and, unlike cluster 1, the $\delta^{15}N$ values stabilize and do not start to decrease at 10,000 years ago. This signal appears in cores all over the world but appears to have a

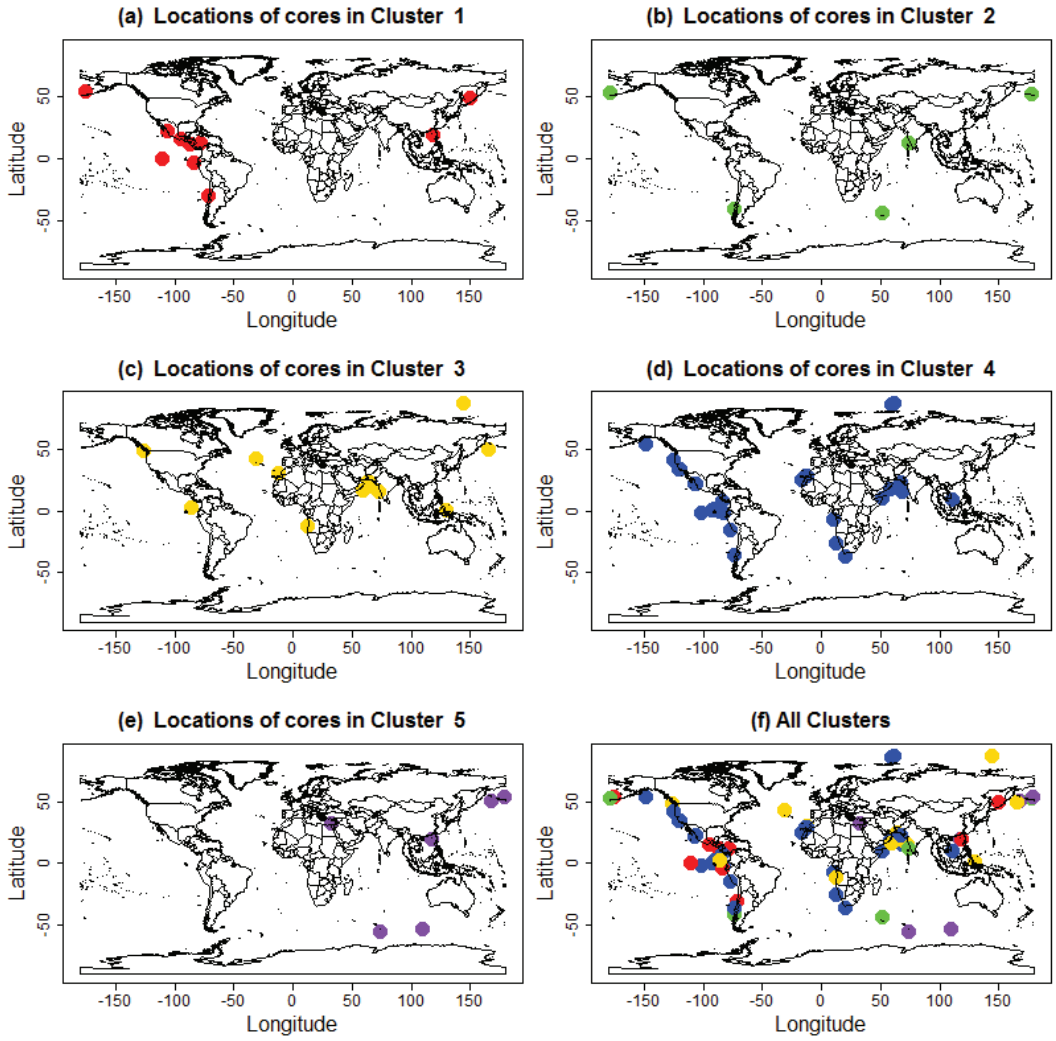
Figure 3.9: Plot of the sum of squares within by the numbers of clusters for the standardized series in the interglacial 1 time period.



large concentration in the Arabian Sea. The cores in cluster 4 have a more gradual increase of $\delta^{15}N$ values. This looks like the first principal component and these cores are located all over the globe. This is not surprising as this was the strongest signal found in the principal component analysis, so it make sense that a large number (32) of cores around the world would have this signal. Finally, the 6 cores in cluster 5 had the strongest decreasing trend. Three of these cores were located in the Southern Ocean and they all seem to fall into the eastern hemisphere.

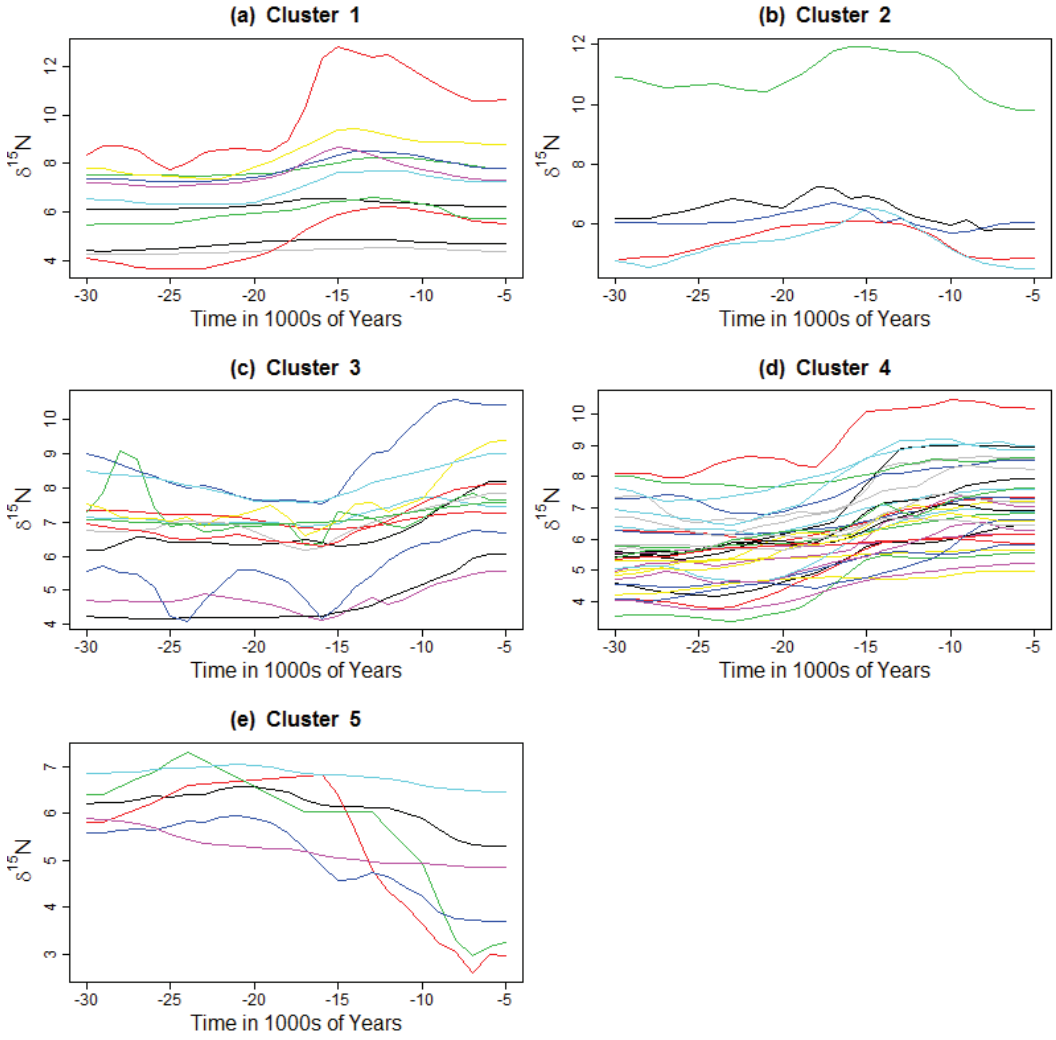
Based on the cluster analysis, cores located off the west coast of the Americas (cluster 1) show an increase in $\delta^{15}N$ values about 5,000 years before the majority of the cores in the Mediterranean Sea (cluster 3). Also the cores on the west coast

Figure 3.10: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the interglacial 1 time period for the standardized series.



of the Americas have a drop off in $\delta^{15}N$ values around 10,000 years ago that is not present in the cores in the Arabian Sea or any other cluster or region. The cluster that stands out the most is the one in the eastern hemisphere (cluster 5), where a steep decreasing $\delta^{15}N$ signal was present (and is not present in other regions of the world). This decrease happens at the same time the cores in cluster 1 start to increase. Overall there was only one real cluster (cluster 1) that stood out geographically in this time period. This may be a limitation of our smoother estimates; as was stated in Chapter 2, the Kalman smoother did not perform well with the small sample sizes.

Figure 3.11: Plots of the $\delta^{15}N$ series that fall within each of the five defined clusters in the interglacial 1 time period for the standardized series.



3.3 Results: Glacial

In this section we repeat the multivariate techniques for the glacial time period (70,000 to 30,000 years ago). When looking at the correlations for the glacial period, it is clear that when we go farther back in time there are fewer series that cover this time period. The locations of the 36 cores in this time period are shown in Figure 3.12

Figure 3.12: Plot of the locations of the 36 series used in the multivariate analysis for the glacial time period.

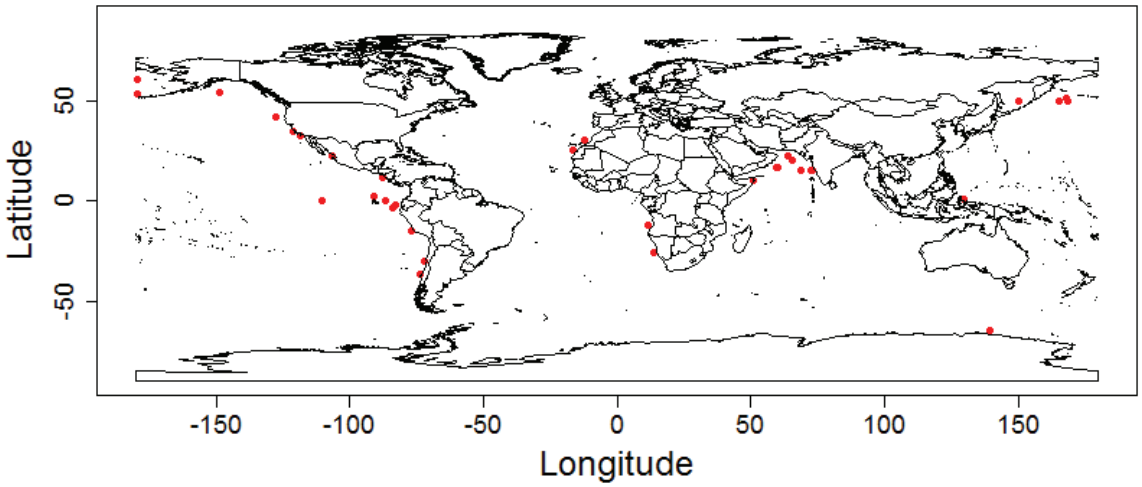
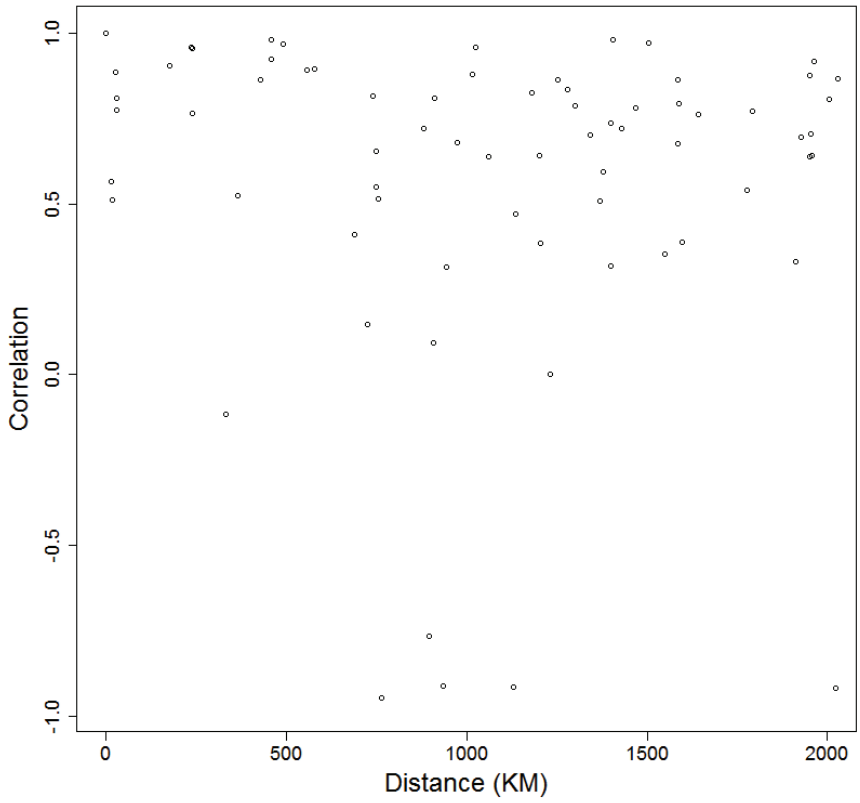


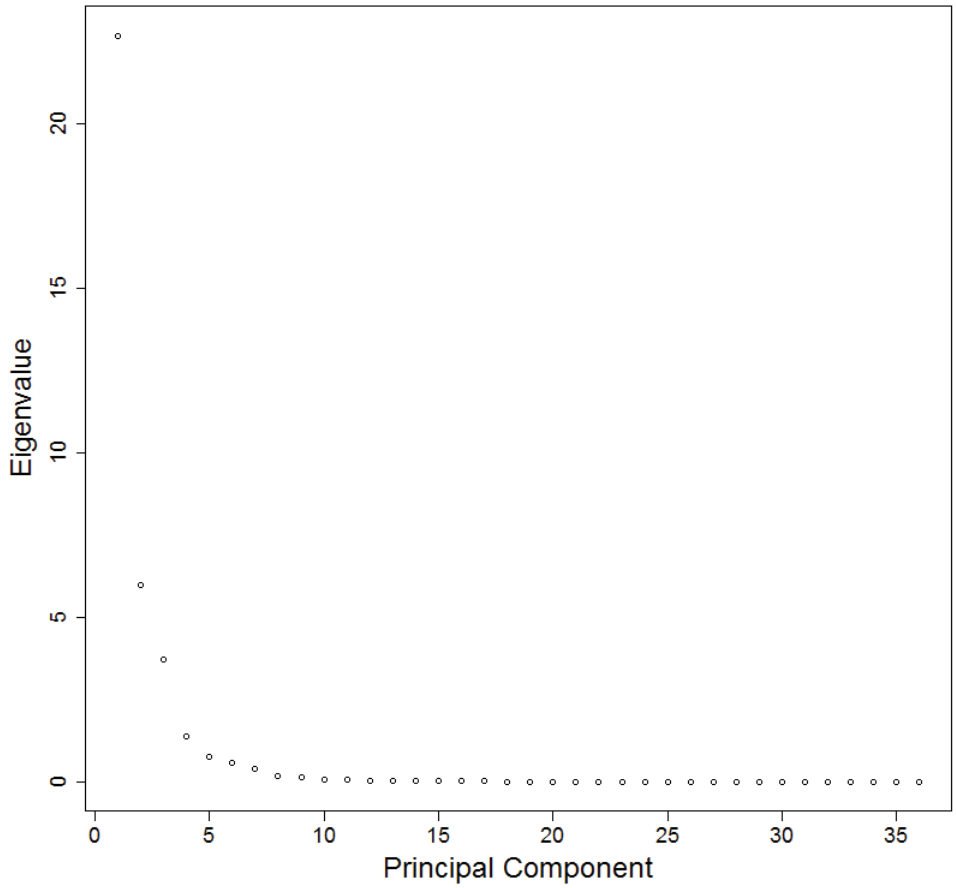
Figure 3.13: Correlation vs. distance plot for the glacial time period.



3.3.1 Correlation vs. Distance

The scatter plot shows that there are only 15 pairings of cores that are located within 500 km of each other (Figure 3.13). Still, we see a similar pattern among these cores, as was seen in the IG1 time period; the correlation of cores within 500 km tend to be highly positively correlated. The exception is that there is only one slightly negative correlation and the remainder are above 0.5. From this we can say that cores that are close together are highly correlated, so we should be able to find similarities in these cores.

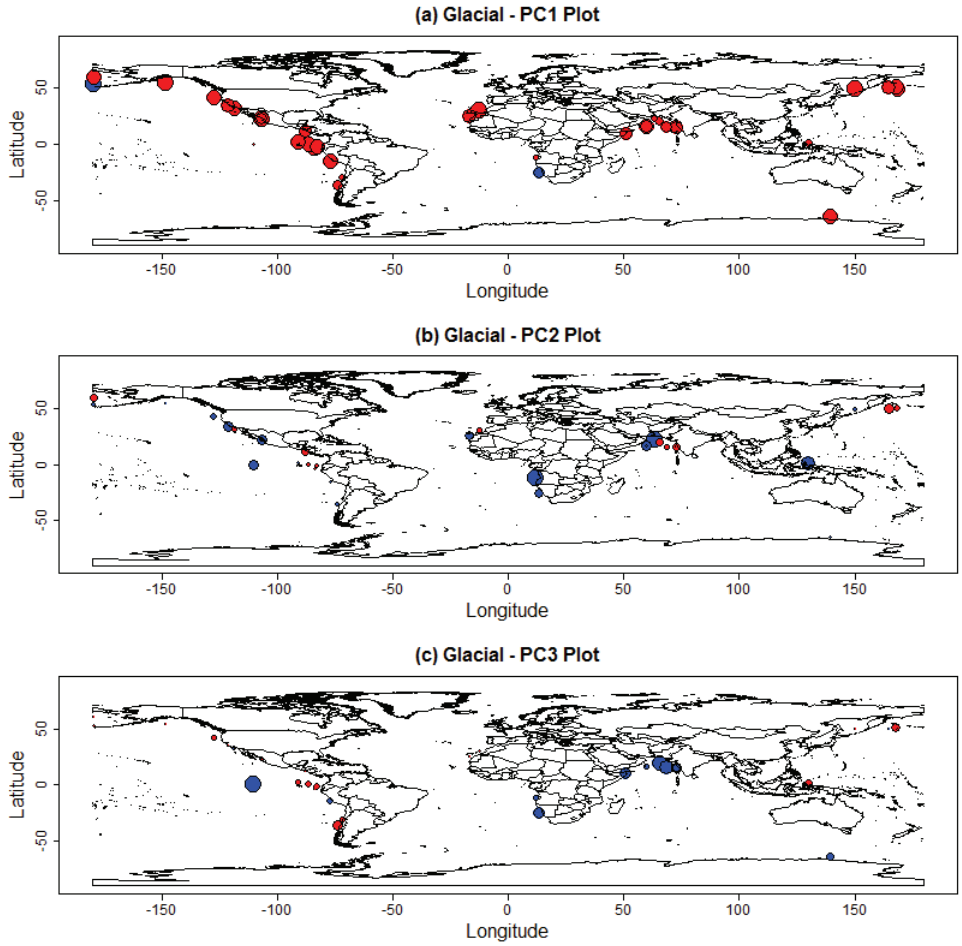
Figure 3.14: Scree plot of the eigenvalues from largest to smallest for the glacial time period.



3.3.2 Principal Component Analysis

Figure 3.14 shows that the elbow in the plot happens at the third principal component. Collectively, principal components 1, 2 and 3 account for a total of 90% of the variance (63%, 17% and 10% , respectively). As with IG1, we see that the first principal component is the strongest signal by a wide margin.

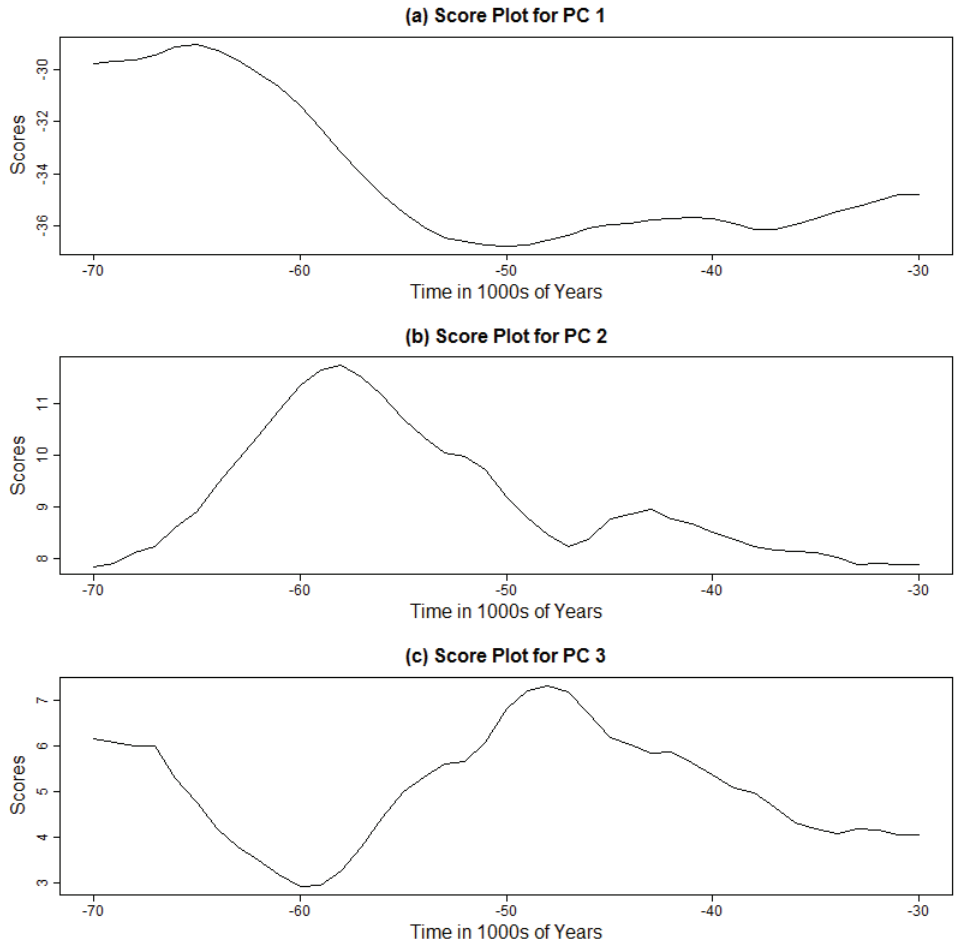
Figure 3.15: Plots of magnitudes of the coefficients for the first (a), second (b) and third (c) principal components in the glacial time period. The size of the circle represents the magnitude (larger circle = larger coefficient) and the colour represents the sign of the magnitude (blue = positive, red = negative).



When looking at the first principal component we can see that it is characterized by a large decrease in $\delta^{15}N$ values starting at around 60,000 years ago (Figure 3.16). However, the magnitude plot shows that the majority of the cores are negatively

correlated (Figure 3.15), which means that they have a sharp increase in $\delta^{15}N$ values at that time.

Figure 3.16: Score plots for the first (a), second (b), and third (c) principal components in the glacial time period.

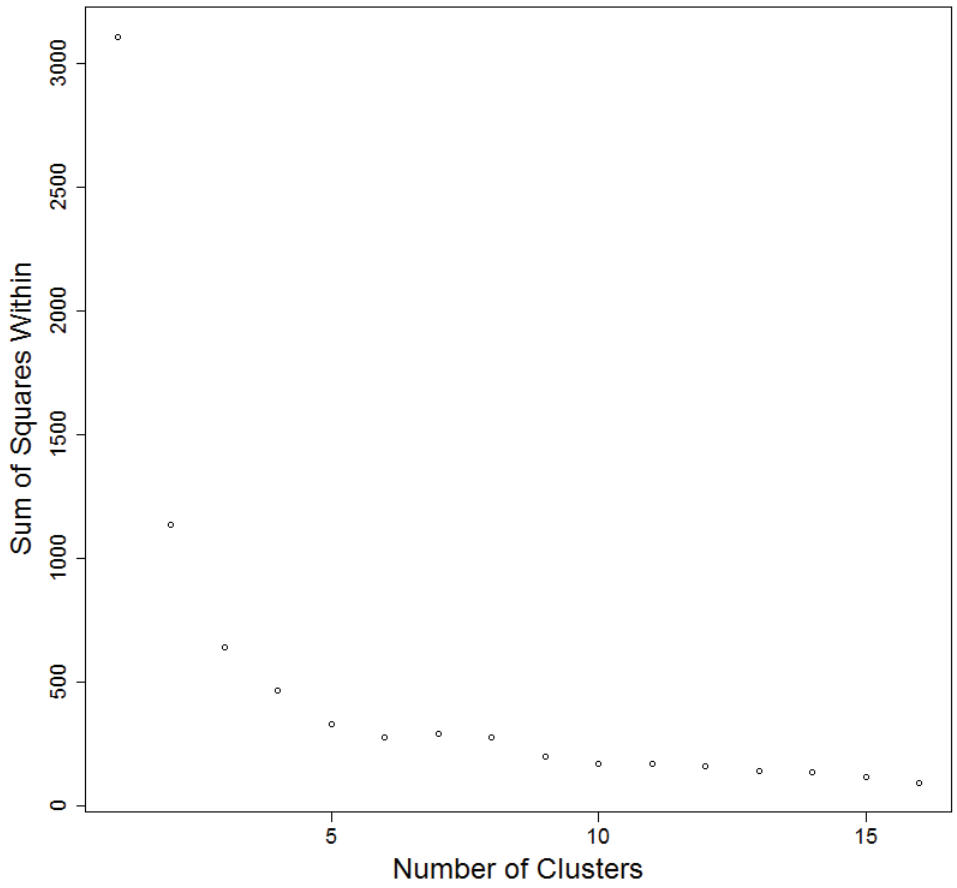


This means the most dominant signal in the glacial time period is a sharp increase at about 60,000 years ago. After the sharp increase, the $\delta^{15}N$ values begin to level off. The second principal component shows that there is a large, almost immediate increasing spike in $\delta^{15}N$ values near the start of this time period (Figure 3.16). After this sharp increase there is a sharp decrease that brings the $\delta^{15}N$ values back down to the initial levels. The signal from the third principal component has a sharp decrease between 70,000 and 60,000 years ago, followed by an increase between 60,000 and 50,000 years ago, at which point it starts to decline again (Figure 3.16). This looks

like the exact reverse of the second principal component.

Overall, the most dominant signal in the glacial time period is a steep increase in $\delta^{15}N$ around 60,000 years ago, which was present in some form in 33 of the 36 series. After about 10,000 years of an increasing trend, the $\delta^{15}N$ values tend to level off and then start to decrease. This signal of $\delta^{15}N$ accounts for 65% of the total population variance in the cores.

Figure 3.17: Scree plot of the sum of squares within by the number clusters used in the glacial time period.

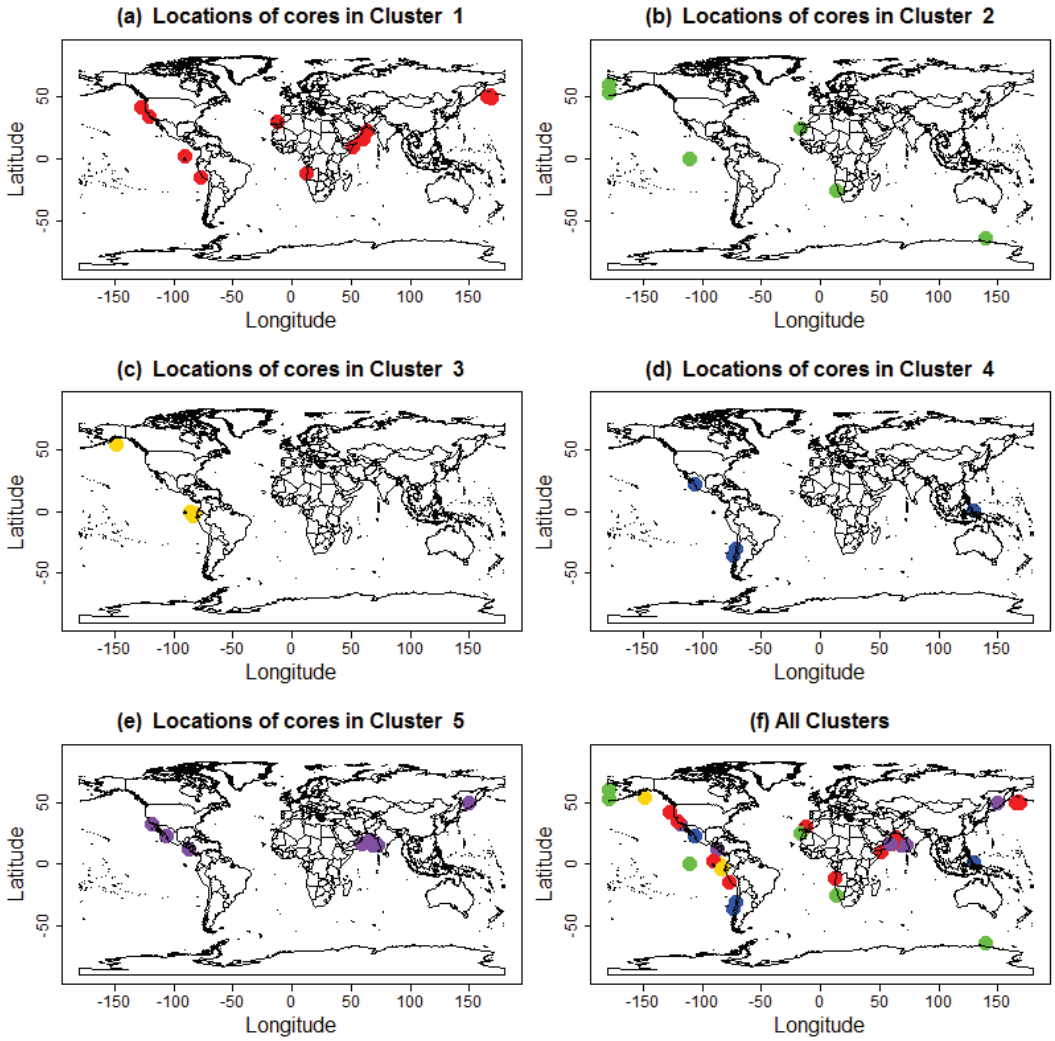


3.3.3 K-means Clustering

$\delta^{15}N$ Magnitude Analysis

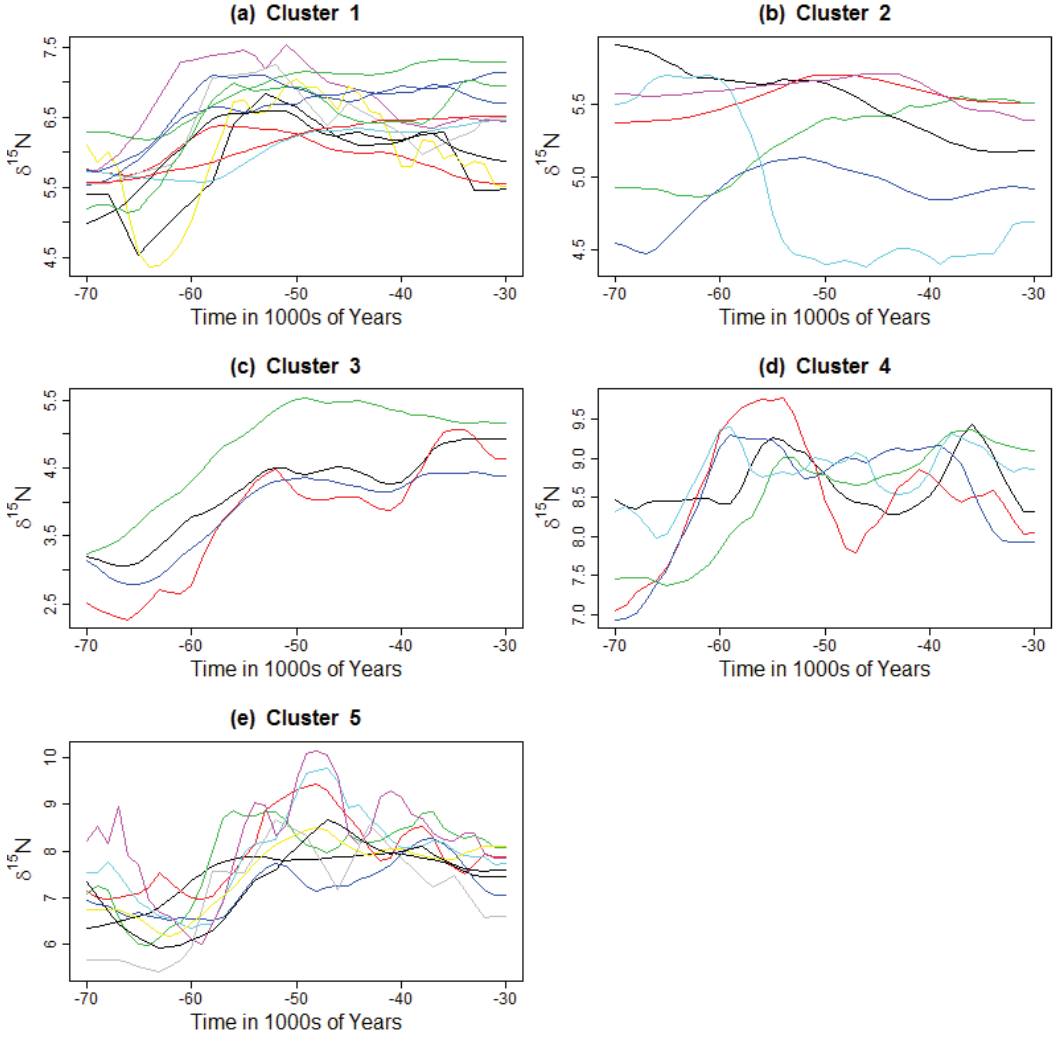
When clustering the cores based on the smoothed values, the result was similar to that found in the IG1 time period; the algorithm simply grouped based on the magnitude of the $\delta^{15}N$ values. This is evident in cluster 2 as the light blue core shares similar values of $\delta^{15}N$ in terms of magnitude with the 5 other cores in the cluster, but it has decreasing $\delta^{15}N$ signal while the other 5 do not. It was determined, based on the scree plot, that 5 was once again the ideal number of clusters (Figure 3.17).

Figure 3.18: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the glacial time period.



Interestingly, even over this time period, the two cores off the southwest coast of South America still had some of the largest values of $\delta^{15}N$. The smallest values of $\delta^{15}N$ were clustered around the northwest coast of South America and the west coast of North America, as shown in Figures 3.18 and 3.19. This could be an indication of two natural processes colliding causing the $\delta^{15}N$ values in the northwest coast of the Americas to be lower than the $\delta^{15}N$ values in the southern hemisphere. This area might be of particular interest as it has both the highest and lowest levels of $\delta^{15}N$ in the same general region.

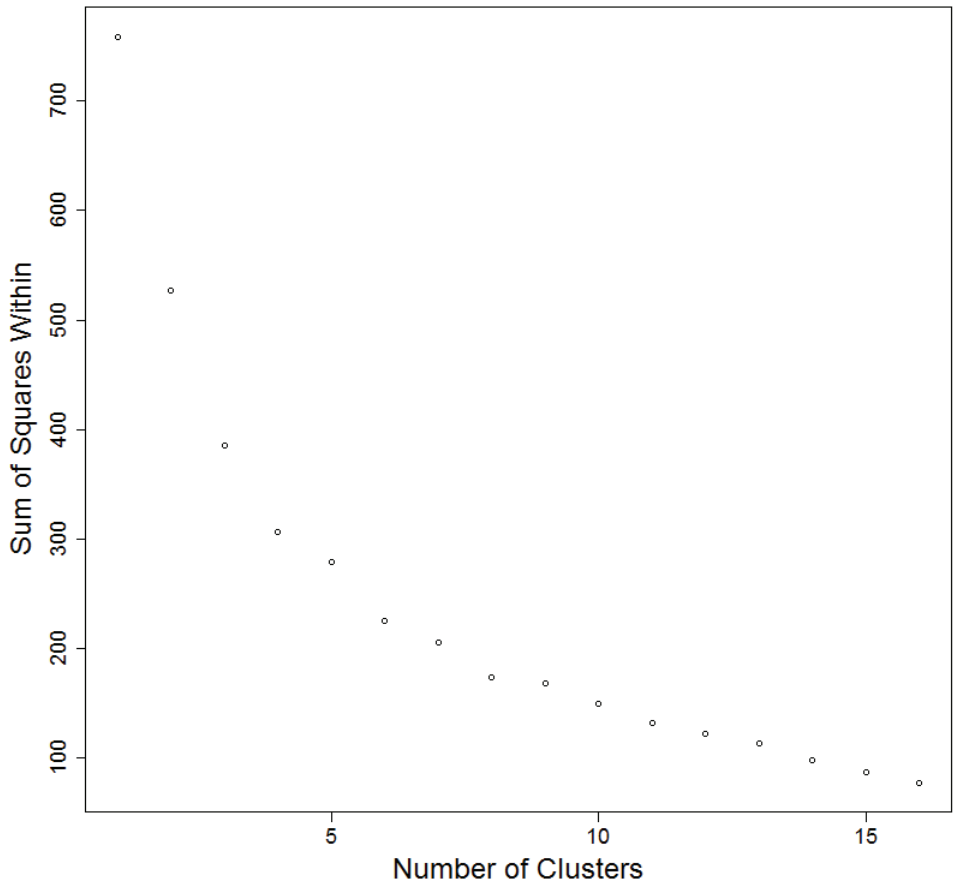
Figure 3.19: Plots of the $\delta^{15}N$ series that fall within each of the five defined clusters in the glacial time period.



$\delta^{15}N$ Trend Analysis

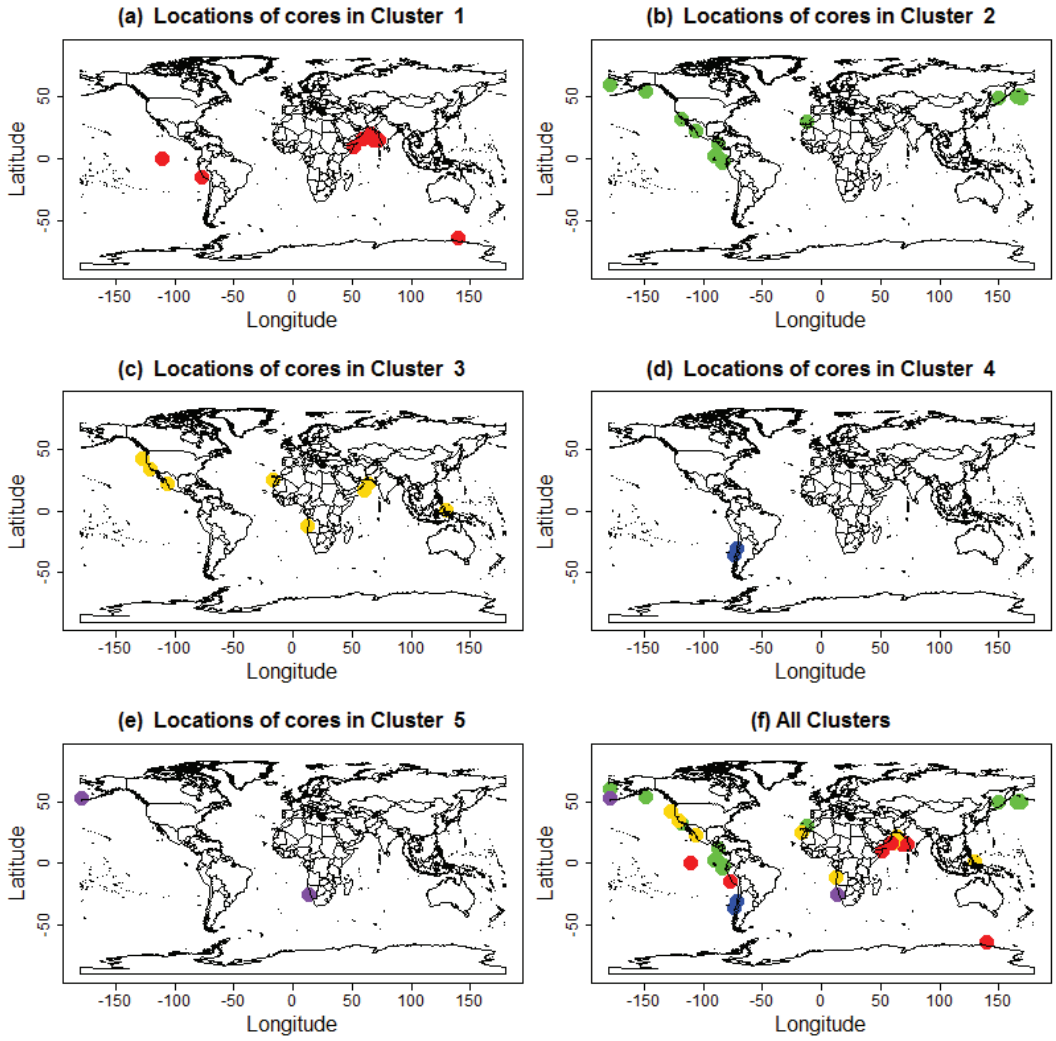
To determine if there were any signals that might be distinct to a given region, instead of just the magnitude of $\delta^{15}N$, the series were standardized. This analysis showed that 5 was still the ideal number of clusters. There is no clear bend in the scree plot shown in Figure 3.20. Despite this, we reasoned that it was appropriate to use 5 clusters in this period based on this scree plot and the previous clustering analysis.

Figure 3.20: Scree plot of the sum of squares within by the number clusters used in the glacial time period based on the standardized series.



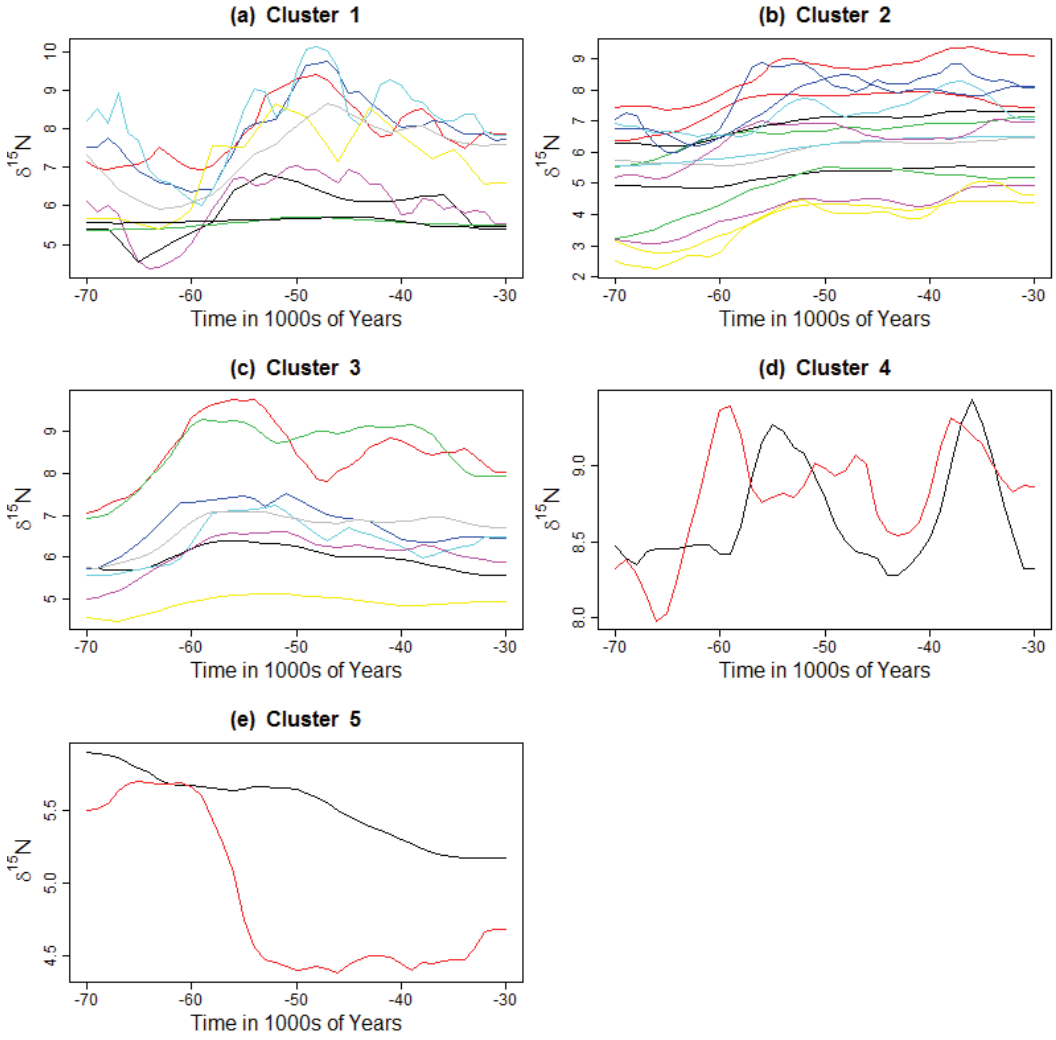
Looking at cluster 1, we immediately notice that the majority of the cores appear in the Arabian Sea. In fact, five of the seven cores in the Arabian Sea are located in this cluster (Figure 3.21). The signal itself appears to be characterized by a decrease in $\delta^{15}N$ values between 70,000 and 60,000 years ago, followed by a sharp increase

Figure 3.21: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the glacial time period for the standardized series.



in $\delta^{15}N$ values between 65,000 and 55,000 years ago (Figure 3.22). That increase was followed by a decrease in $\delta^{15}N$ values over the rest of the time period that was about half as large in magnitude as the increase. This looks fairly similar to results of the first principal component. The cores in cluster 2 did not have decreasing $\delta^{15}N$ values between 70,000 and 60,000 years ago, but did have the same increase in $\delta^{15}N$ values between 65,000 and 55,000 years ago. Most of the cores in cluster 2 hit their maximum peak closer to 50,000 years ago instead of 55,000 years ago (Figure 3.22). The majority of the cores in this cluster did not have a decrease in $\delta^{15}N$ after they

Figure 3.22: These are plots of the $\delta^{15}N$ series that fall within their respective clusters for the standardized series in the glacial time period.



hit their maximum. Instead, they stayed around the maximum level of $\delta^{15}N$ for the remainder of the time period. Most of the cores in cluster 2 are located off the west coast of North America and the northwest coast of South America. The cores in cluster 3 were fairly spread out across the globe. The distinct signal in cluster 3 is that the cores peaked at their maximum about 5000 years before cluster 1. However, the cores in cluster 3 do share the same decrease in $\delta^{15}N$ values after they hit their peak. Cluster 4 has only two members and they are both located on the southwest coast of South America. These cores are unique in that they are the only two cores that have two distinct peaks. The interesting thing about these cores is that the black

core seems to lag behind the red core by about 5,000 years, even though the cores are very close geographically. Cluster 5 is also comprised of only two cores. This cluster grabbed all of the cores that had a strong negative trend, which is unlike the majority of the cores in this time period. Unlike the other clusters, the $\delta^{15}N$ values in cluster 5 start to decrease between 65,000 and 55,000 years ago instead of increasing.

As with IG1, we were able to obtain a cluster that was predominantly made up of the cores located in the North Pacific, off the west coast of the Americas (cluster 2). This core was defined by the $\delta^{15}N$ values not decreasing after they hit their maximum peak. This indicates that the $\delta^{15}N$ values were more stable on the west coast of the Americas. A second distinct group appeared in the Arabian Sea (cluster 1). This cluster was defined by a steep drop off of $\delta^{15}N$ values after hitting its maximum. The cores in cluster 3 were located all over the globe and were very similar to the cores in cluster 1, but preceded the trends found in cluster 1 by about 5,000 years. There could be natural process that caused this delay in cluster 1 or it could just be a byproduct of the aging of the samples. Cluster 4 consisted of only two cores. This cluster was defined as the only two series that had two distinct maximum peaks and should be flagged for further investigation as the lag seen between the cores could be indicative of a natural process or it could be based on the methods used for calculating the age. Cluster 5 included the two cores that had an overall decreasing trend over the glacial time period.

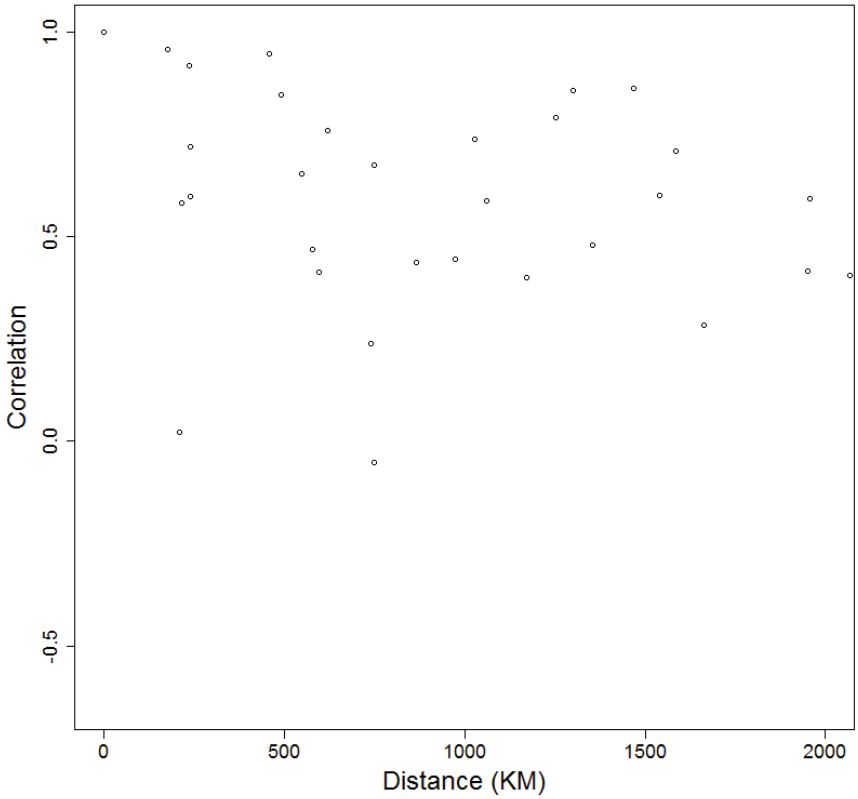
3.4 Results: Interglacial 2 (IG2)

In this section we repeated the multivariate techniques for the interglacial 2 (IG2) time period (125,000 to 5,000 years ago) to identify any characterizing trends in $\delta^{15}N$ values, compare $\delta^{15}N$ values in neighboring cores and determine the optimal set of clusters. The IG2 time period includes 29 cores that cover both of the previous time periods. When analyzing this time period, we were interested to see if the predominant increases from the previous two time periods would continue to be present.

3.4.1 Correlations vs. Distance

The correlation plots for the IG2 time period show that there are very few cores that are less than 500 km apart (Figure 3.23). This was to be expected as not many of the cores in the data set had data points as far back as 125,000 years ago. In total there were eight pairings of cores that were separated by 500 km or less. None of the correlations are negative, and all but one is above 0.5. This means that 7 of the 8 pairings have fairly high correlations, which is consistent with findings in the glacial time period.

Figure 3.23: Plot for the correlation vs. distance plot for the interglacial 2 time period.

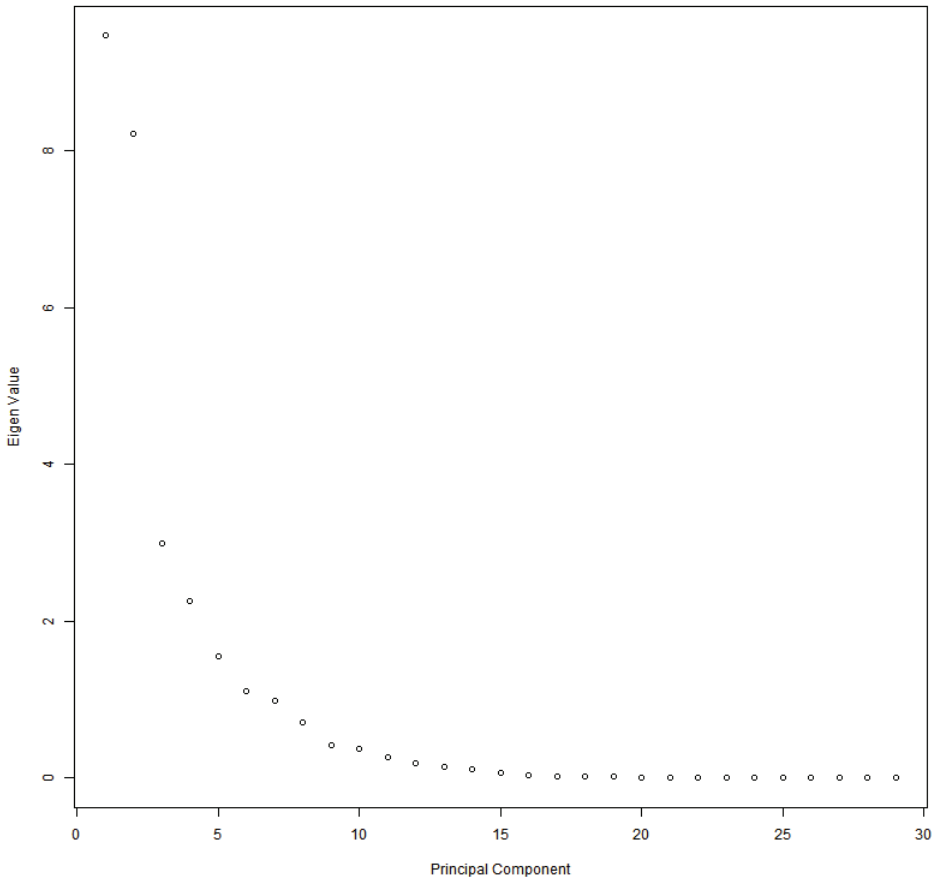


3.4.2 Principal Component Analysis

It was determined that 4 principle components accounted for 79% of the variability in the cores. When looking at the scree plot of the eigenvalues (see Figure 3.24), the first

two principal components were relatively close. This is different from the other two time periods where there was a dominant signal that accounted for a majority (over 50%) of the total variability. In this time period, the first two principal components (combined) account for 51% of the variability, which suggests that there is no single signal that defines this time period.

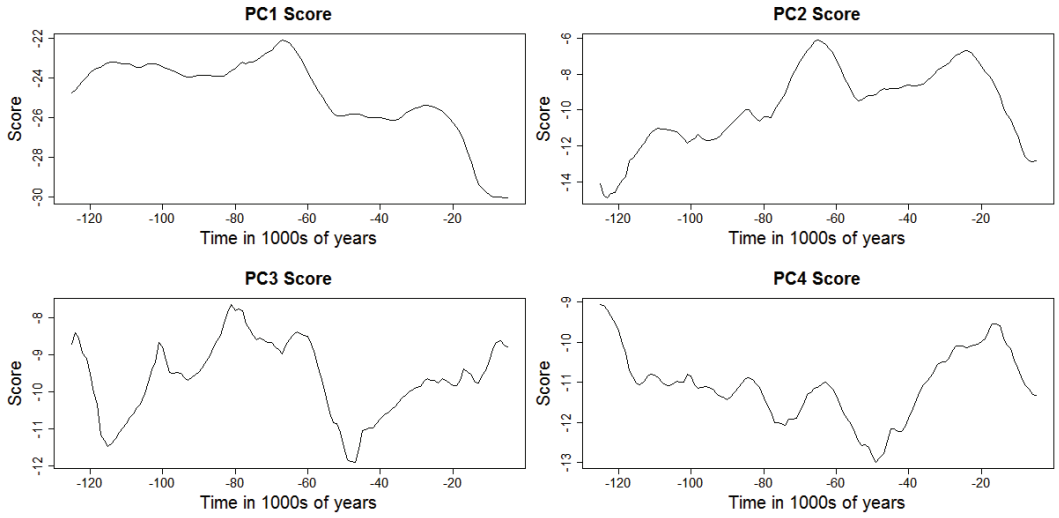
Figure 3.24: Scree plot of the eigenvalues from largest to smallest for the interglacial 2 time period.



Looking at the first principal component (Figure 3.25), which accounted for 33% of the variability, there was a significant drop approximately 60,000 years ago, followed by a second significant drop at approximately 30,000 years ago. The magnitude plot (Figure 3.26) for this principal component is full of red circles indicating a negative coefficient for 27 of the 29 cores in this time period - *i.e.*, they are negatively correlated

with this principal component. The negative coefficient in these 27 cores means that the two drops in $\delta^{15}N$ values are actually increases. The increases appear at approximately the same time as the dominant increases on both the IG1 and glacial time periods. This principal component seems to be connecting the first principal component for the previous two analyses. Once past the 70,000 years ago mark, the score increased slightly followed by a moderate decrease from approximately 115,000 years ago to 125,000 years ago.

Figure 3.25: Score plots for the 4 principal components for the interglacial 2 time period.

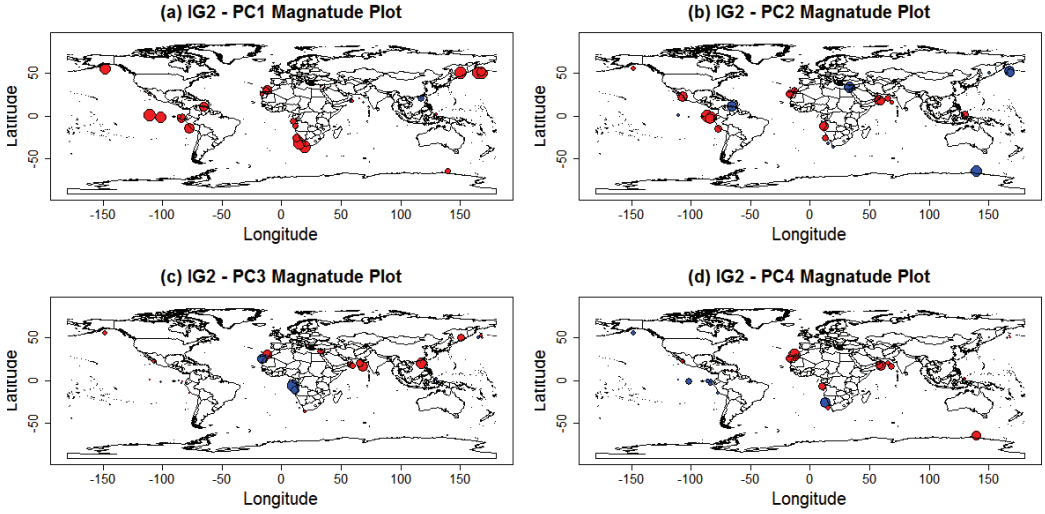


The second principal component (accounting for 28% of the variability) correlates negatively with 15 cores and positively with 14 cores. This principal component is signaling an increase in $\delta^{15}N$ over the majority of the time period. It is characterized by two large increases at 120,000 years ago and about 75,000 years ago, before a steep decrease around 25,000 years ago. Looking at the magnitude plot for this principal component, the most negatively correlated cores tend to be on the west coast of the Americas.

The third principal component (accounting for 10% of the variability) starts off with a sharp decrease followed by two distinct increases at around 115,000 years ago and 95,000 years ago. After the second increase, the $\delta^{15}N$ signal begins to decrease slowly at 80,000 years ago, before decreasing rapidly at 60,000 years ago. For the

remainder of this time period, the $\delta^{15}N$ signal shows an increase, ultimately reaching the same approximate level of $\delta^{15}N$ that it started at. The magnitude plot shows that this signal does not have a lot of weight off the west coast of the Americas. This signal seems to be highly positively correlated with the cores on the west coast of Africa and highly negatively correlated with the cores in the Arabian sea.

Figure 3.26: Magnitude of the coefficients for the principal components for the 29 cores for the interglacial 2 time period.



For the fourth and final principal component (accounting for 8% of the variability), the overall trend appears to be one of a decreasing $\delta^{15}N$ signal over the time period. It starts with a steep increase over the first 10,000 years in the time period then slowly decreases (with a few minor increases) over the next 60,000 years. After that long period of a decreasing trend, the $\delta^{15}N$ signal then starts to increase rapidly over the next 20,000 years before decreasing again by about half the magnitude of the previous increase. The cores in the Arabian sea and off the northwestern coast of Africa have very strong negative correlations with this signal. In fact, only one core has a highly positive correlation with this signal, located off the southwest coast of Africa. Once again the magnitudes of the loading values are very small near the Americas, so this signal does not have much weight in that region.

3.4.3 K-means Clustering

$\delta^{15}N$ Magnitude Analysis

The scree plot for the unstandardized series show that the cores in the IG2 period were again grouped based only on magnitude. There were no geographically distinct clusters that emerged, so we moved on to the standardized plots. The plots for the k-means clustering can be found in Appendix A1.

$\delta^{15}N$ Trend Analysis

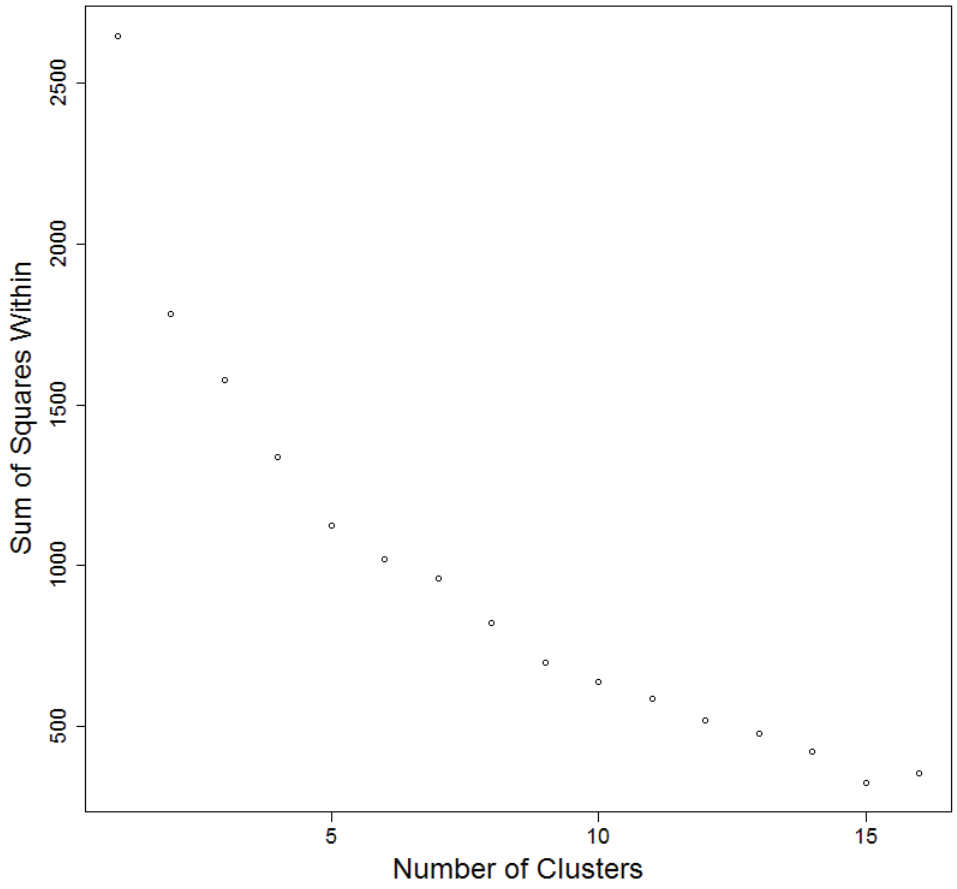
Looking at the standardized k-mean scree plot there was again no real elbow in the plot (Figure 3.27). So once again we used 5 clusters as it looked reasonable based on the plot and that it would stay consistent with the other two time periods.

Looking at cluster 1, the 5 cores in this cluster had similar signals to those found in the other two time periods (Figure 3.28). They were defined by a large spike in $\delta^{15}N$ values at both 15,000 years ago and 65,000 years ago. A steep decrease occurred in each immediately before the glacial time period started. The cores in this time period look stationary; the mean and variance, outside of the large dip at 60,000 years ago, stays fairly constant across the time periods. Four of the 5 cores are located just off the west coast of Africa (Figure 3.29).

Cluster 2, which contained 7 cores, was characterized by an upward trend over the time period. Looking at the short term fluctuations, four of the cores (teal, yellow, purple, and blue) all have sharp increases from 125,000 years ago until approximately 80,000 years ago (Figure 3.28). Following this they share the similar decrease at 80,000 years ago and the two increases at 60,000 and 25,000 years ago were seen with cluster 1. However the red, black and green cores look different in that they do not have the steep increase from 125,000 years ago to 80,000 years ago. Also after their increase at around 65,000 years they level off and do not increase at the same rate as the other 4 cores. These cores, which were located all over the globe (Figure 3.29), were most likely grouped together based on this upward trend.

Cluster 3 has five cores, four of which, do not have much fluctuation in $\delta^{15}N$ in

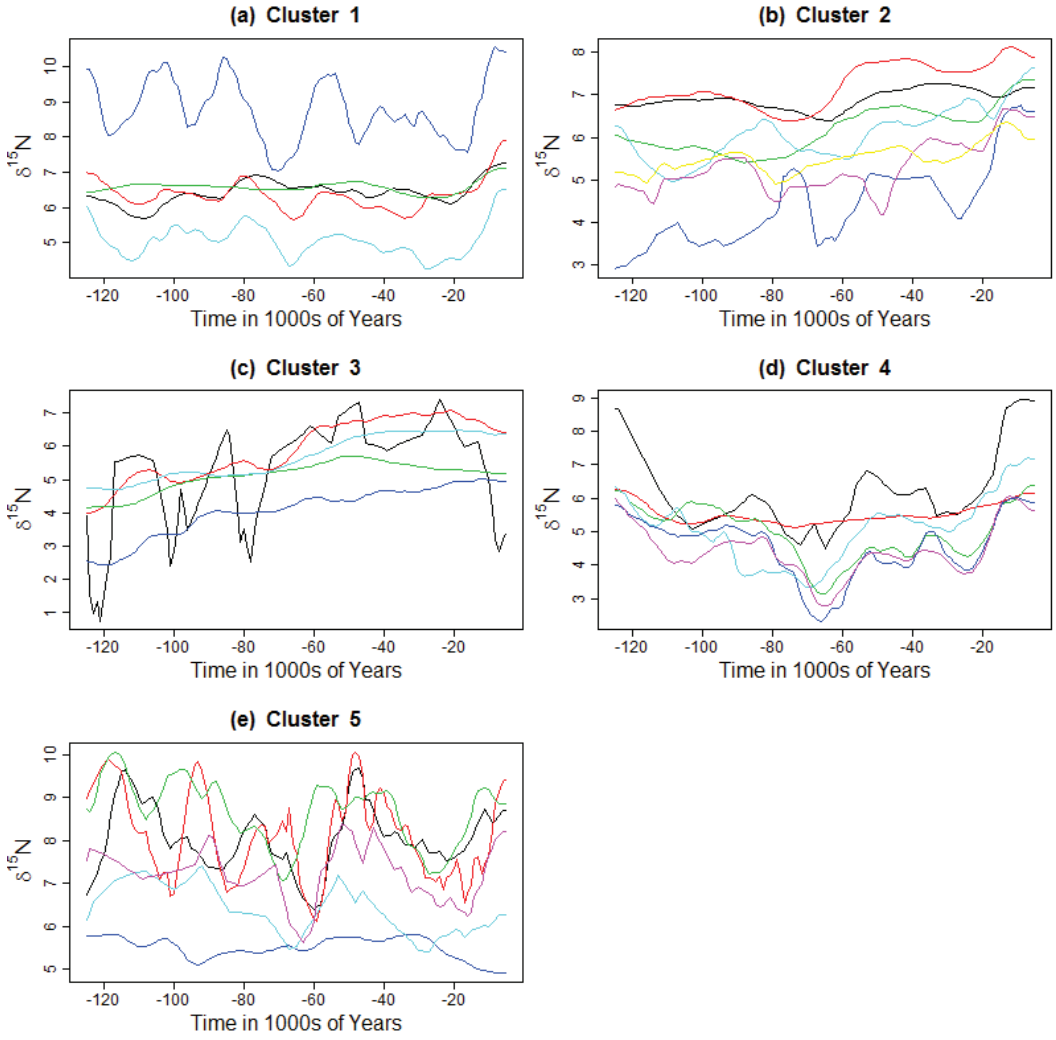
Figure 3.27: Scree plot of the K-mean analysis for the interglacial 2 time period.



short intervals. The signals (excluding the black signal) are very smooth and have a gradual upward trend over the time period (Figure 3.28). The black signal looks very similar to the cores in cluster 1, with the exception that at approximately 15,000 years ago it had a sharp decrease instead of a sharp increase. The cluster 3 cores were scattered all across the globe (Figure 3.29).

The 6 cores in cluster 4 share one common feature that is not present in the other clusters. These cores have a steady decrease in $\delta^{15}N$ values from 120,000 to 100,000 years ago, and tend to stay at this low level until they near the end of the time period when spike back up (Figure 3.28). The cores in clusters 1 and 2 also show a dip in $\delta^{15}N$ values at the beginning of the time period, but reach values similar to the levels at which they started frequently throughout the time period. The cores in cluster 4

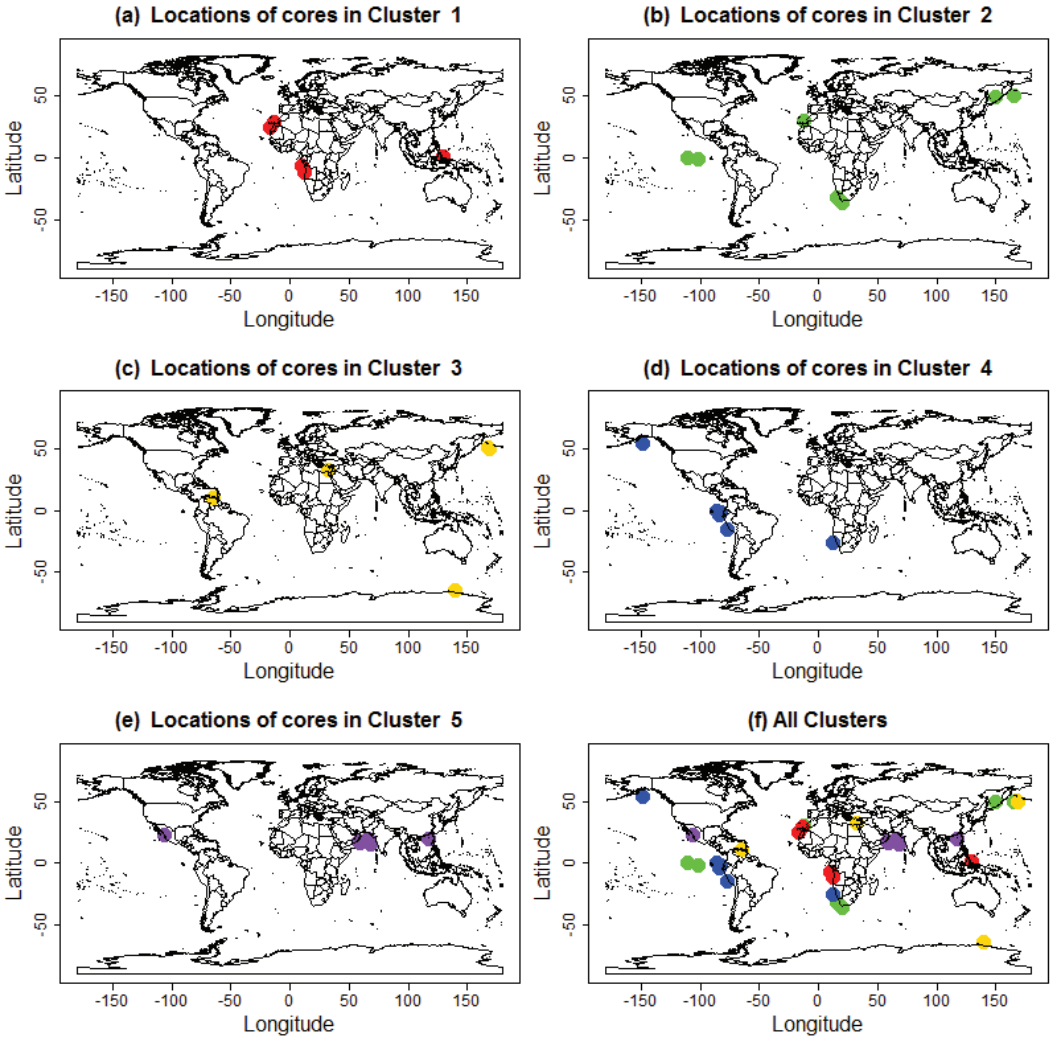
Figure 3.28: These are plots of the $\delta^{15}N$ series that fall within their respective clusters for the standardized series in the interglacial 2 time period.



have the same dips and increase over the short term (*i.e.* increases in $\delta^{15}N$ values at 60,000 and 20,000 years ago), but they do not get back to the high levels of $\delta^{15}N$ at which they started until the end of the time line. All but one of these cores are located on the west coast of the Americas (Figure 3.29). It would be of interest to look into why, over the middle section of this time time period the cores on the west coast of the Americas could not reach the high levels of $\delta^{15}N$ they had at the start and end of the time period, while other regions were able to.

Cluster 5, which contains 6 cores, looks almost identical to cluster 1. The only difference in the signals in the cores in cluster 5 is that they appear to lag behind the

Figure 3.29: Locations of the cores in each cluster for the interglacial 2 time period.



cores in cluster 1 by about 15,000 - 20,000 years (Figure 3.28). Again, there could be a natural cause to create this lag or it could simply be the way in which the cores were aged. These cores were found mainly in the Arabian Sea (Figure 3.29). Further investigation may be warranted to determine why cores in the Arabian Sea seem to lag behind the cores off the west coast of Africa.

Chapter 4

Discussion and Conclusions

Oceanographers and marine biochemists are keenly interested in quantifying changes in the fixed marine nitrogen inventory, and the processes that determine changes in this inventory through geological time. Unfortunately, there is no actual, historical record of the marine nitrogen inventory itself. However, researchers can measure the processes that are known to increase (nitrogen fixation) or decrease (denitrification) the levels of nitrogen in the ocean. $\delta^{15}N$ is a ratio of the two stable isotopes of nitrogen: ^{14}N, which is very common, and ^{15}N, which is rare. The ratio of these two isotopes is expressed as $\delta^{15}N$ (see equation 1). High levels of $\delta^{15}N$ tend to represent areas that have higher rates of denitrification (versus fixation). The process of denitrification uses the more common ^{14}N isotope preferentially, leaving higher concentrations of the ^{15}N isotope in the ocean. Lower levels of $\delta^{15}N$ can represent areas that have higher rates of nitrogen fixation (over denitrification). The process of fixation does not have a preference as to which isotope is used so such areas produce $\delta^{15}N$ values lower than areas with high denitrification rates.

In a previous analysis of a database of 173 downcore records, a team of researchers conducted a clustering analysis to group samples of $\delta^{15}N$ into geographic regions and found that cores within a 100 km radius of a defined reference core had similar mean values of $\delta^{15}N$ [3]. The multivariate analysis reported in this thesis, conducted on the same database was designed from a purely statistical point of interpretation to characterize the $\delta^{15}N$ signals extracted from the noisy observations, without geographical boundaries.

We conducted a univariate analysis that separated the process error (natural changes in $\delta^{15}N$) from the observation error (error from the measurement of $\delta^{15}N$), which was fixed at 0.28, to extract the true underlying signal of $\delta^{15}N$ from all of

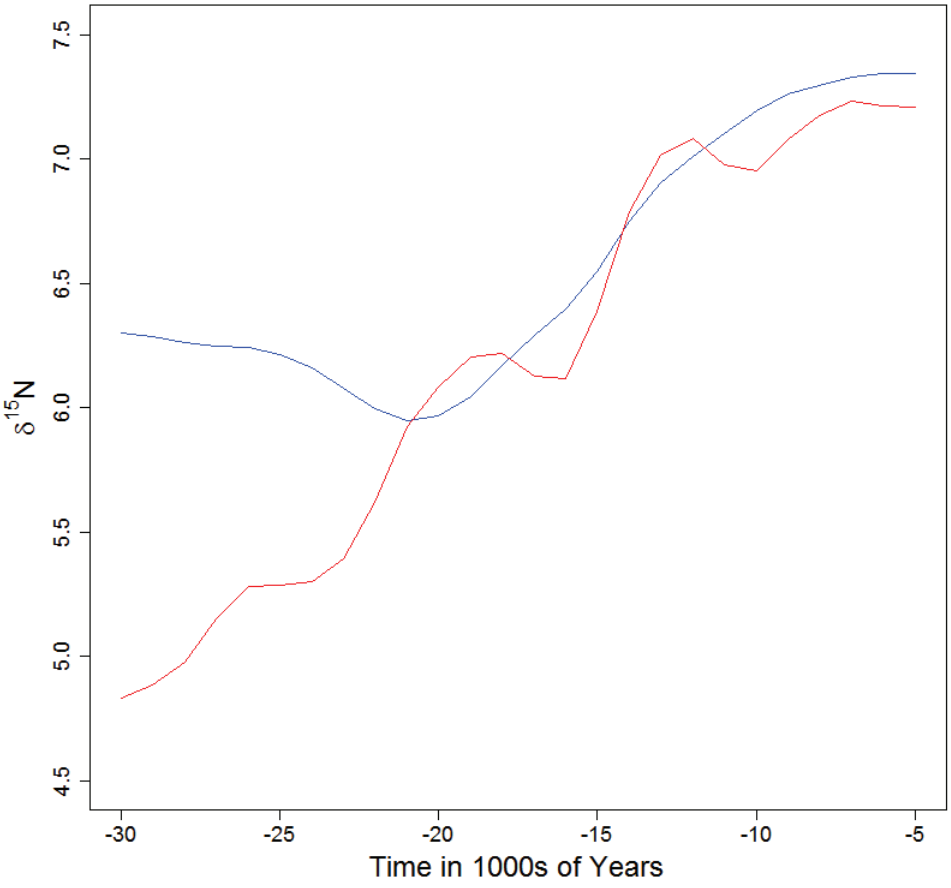
the cores. As a byproduct of signal extraction, using the Kalman smoother, we were able to place the data on a fixed interval time line across all cores, completing the preparation for the multivariate analysis.

The multivariate analysis found that cores that were close in proximity were highly correlated in all three time periods. Using principal component analysis, we found that the two most dominant $\delta^{15}N$ signals, present in most cores and in most regions, were an increase in $\delta^{15}N$ values around 20,000 years ago during the interglacial 1 period (IG1), and around 60,000 years ago during the glacial period. During the interglacial 2 (IG2) time period, there was no single dominant signal; however, the signals that were dominant in the IG1 and glacial time periods were both evident in the IG2 time period.

The most dominant signal seen in any of the principal component analyses was the sharp increase in $\delta^{15}N$ seen in IG1 (accounting for 70% of the variance in that time period). Since there were only two principle components (PC1 and PC2) in that time period, we decided to look at whether cores that had similar loading values were more similar than cores that had differing loading values. In the IG1 time period, PC2 has a "V" type shape. Cores that were positively correlated with both PC1 and PC2 seem to have a delay in the increase that was seen in PC1 alone. Cores that have a positive correlation with PC1 and a negative correlation with PC2 tend to have a steeper increase in $\delta^{15}N$ than seen in PC1 alone. To demonstrate, we plotted a core with a highly positive correlation with both PC1 and PC2 (Figure 4.1). Here we can see that this delays the increase, so the $\delta^{15}N$ values start to increase at around 16,000 years ago instead of 25,000 years ago. If we look at a core that has a high positive correlation with PC1 and a high negative correlation with PC2, we see that it increases the amplitude of the increase - *i.e.*, the $\delta^{15}N$ values increase at around 25,000 years ago, but at a much faster rate.

In terms of clustering, the Americas were the only region that had a fairly distinct grouping of cores across all three time periods. The Arabian Sea started to define itself as a cluster in the glacial and IG2 time periods, while the cores on the coast of

Figure 4.1: Plot of two series with different relationships to the first two principal components. GeoB 1008 (blue line), located off the the coast of Africa has a positive correlation with both principal components while ODP 887 (red line), located off the coast of North America has a positive correlation with the first principal component but a negative correlation with the second principal component.



Africa started to distinguish themselves in IG2.

4.1 Post Hoc Analyses 1 - Comparisons Between Time Periods

In Chapter 3, we clustered the cores based on their $\delta^{15}N$ signals in the three different time periods (IG1, glacial and IG2), but did not compare the findings between time periods. In this section we will discuss and compare the grouping patterns between the three time periods. In total there will be four comparisons: IG1 vs. Glacial, IG1 vs. IG2, Glacial vs. IG2 and IG1 vs. Glacial vs. IG2. This post hoc analysis

was conducted to determine whether cores that covered multiple time periods stayed grouped together and if similar regions were defined in all of the time periods clusters.

4.1.1 Interglacial 1 vs. Glacial

For the comparisons made here we will be using the clusters in the glacial period as our point of reference.

Cluster 1 in the glacial period had 9 total cores, 8 of which were also represented in the IG1 time period. In this cluster there were two distinct groups that were also clustered together in the IG1 period. The first group (group 1) consisted of four cores in the Arabian Sea (MD 76-131, ME33 NAST, RC27-61, and SK117-GC8) that were all in cluster 3 in the IG1 time period. The second group (group 2) consisted of three cores that were present in cluster four of IG1. Of these three cores two were located in the Arabian Sea (ME33 EAST and NIOP 38-02) and one was from the Americas (CD 38-02).

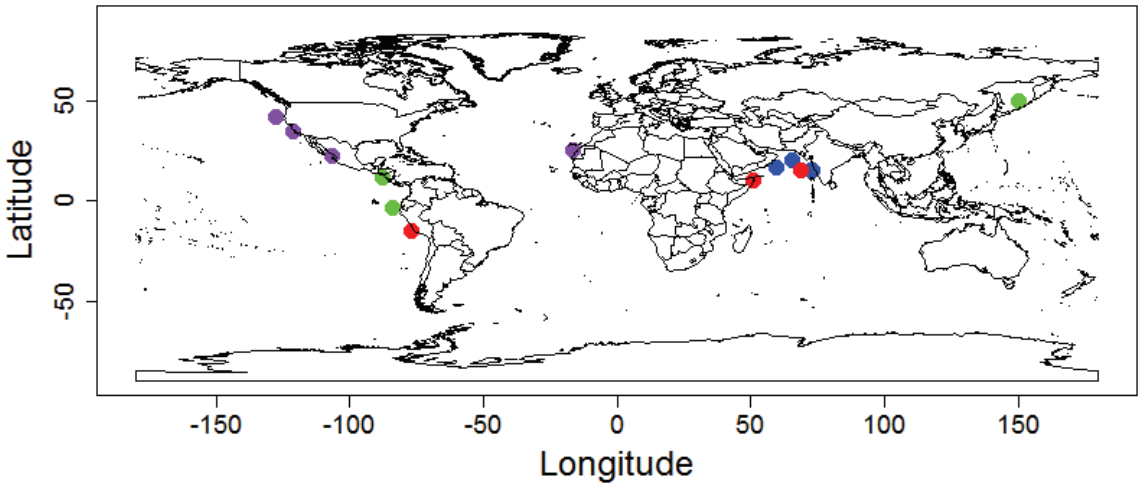
In cluster 2 there were 15 cores in the glacial time period, but only 5 of them were present in the IG1 time period. Of the 10 cores that were not in IG1, eight were located by the Americas and two were located on the west coast of Africa. In this cluster however there were only three cores that were clustered together in cluster 1 of IG1 (Group 3). Two of these cores were located by the Americas (MD 02-2524 and TR163-31) and one by Asia (GGC27).

Cluster 3 had eight total cores, six of which were also in the IG1 time period. Four of the six (group 4) were clustered together in cluster 4 of IG1. Three of these were located in the Americas (NH22P, ODP 1017, W8709-8 PC) and one by Africa (SU94 20bK).

In glacial time period, clusters 4 and 5 consisted of 2 cores each - in both cases one core was present in the IG1 time period and one was not. As a result, comparisons were not made for these two clusters.

In summary there were four distinct groups that were clustered together in the IG1 and glacial time periods. Two of these groups were predominantly made up of cores

Figure 4.2: Plot showing the locations of the four groups of cores that were in the same cluster in both the interglacial 1 and glacial time periods: group 1 (MD 76-131, ME33 NAST, RC27-61, and SK117-GC8) is represented by the blue circles; group 2 (CD 38-02, ME33 EAST and NIOP 38-02) by the red circles; group 3 (GGC27, MD 02-2524 and TR163-31) by the green circles; and group 4 (NH22P, ODP 1017, W8709-8 PC and SU94 20bK) by the purple circles.



from the Arabian Sea (group 1 and group 2) while the other two were predominantly made up of cores in the Americas (groups 3 and 4). This suggests that these core groups behave similarly in the two time periods, which may indicate patterns of change in the marine nitrogen cycles that are specific to those regions over those time periods. Plots of the groups can be seen in Figure 4.2

4.1.2 Interglacial 1 vs. Interglacial 2

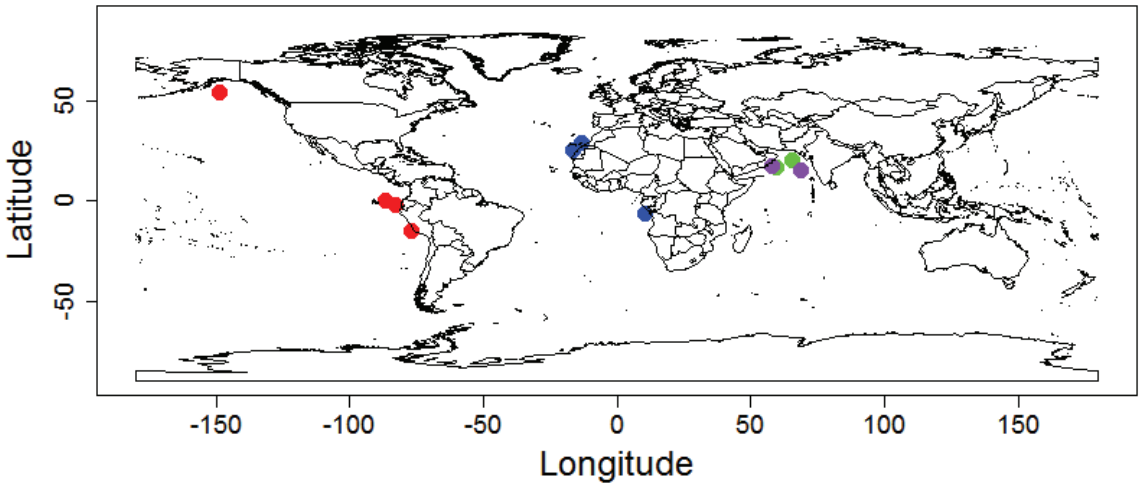
The comparisons in this section use the clusters in IG2 as the point of reference. (Note that it is possible for a core to be represented in IG2, but not be represented in IG1. The selection method for the IG1 group required cores to have an observation of $\delta^{15}N$ within 1000 years of 30,000 years ago. This was not a constraint on the IG2 time period.)

Cluster 1 in the IG2 time period consisted of 5 cores, all of which were included

in the IG1 time period. In IG2, the cores were primarily located on the west coast of Africa. Three of the five cores (group 1) in this cluster, all located on the west cost of Africa, were also grouped in cluster 4 in the IG1 time period (GeoB 1008, GeoB 4240 and Su94-20bK). Another core located near Africa (GeoB 1016) from cluster 1 in the IG1 time period, also joined this group.

In Clusters 2 and 3 in the IG2 time period there were no groupings that were close geographically. Of the seven cores in cluster 2, six were also in the IG1 time period. There were two cores each from clusters 1, 3 and 4 in the IG1 time period, but none of the pairing were close geographically. In cluster 3 there were two cores that shared the same cluster in IG1 but again they were not close geographically (one from the Mediterranean and one from Asia).

Figure 4.3: Plot showing the locations of the four groups of cores that were in the same cluster in both the interglacial 1 and interglacial 2 time periods: group 1 (GeoB 1008, GeoB 4240 and Su94-20bK) is represented by the blue circles; group 2 (CD 38-02, ME0005A 24JC, ME0005A 27JC and OSP 887) by the red circles; group 3 (ME33 NAST and RC27 61) is represented by the green circles; and group 4 (ME33 EAST and RC27 24) by the purple circles.



Cluster 4 was comprised of six cores, all of which were in the IG1 time period as well. Five of the six cores were in cluster 4 in the IG1 time period, four of which were located on the west coast of the Americas (CD 38-02, ME0005A 24JC, ME0005A

27JC and OSP 887) (group 2). A fifth core from the Americas (from cluster 1 in IG1) joined the four above in the IG2 time period.

Cluster 5 was comprised of mainly of cores in the Arabian Sea (4 of 6), all of which were in the IG1 time period. Two of the cores (group 3) from the Arabian Sea (ME33 NAST and RC27 61) were in cluster 3 in IG1 and two others (ME33 EAST and RC27 24) came from cluster 4 (group 4).

Again, there were four distinct groups that clustered together across the IG1 and IG2 time periods. As with the previous comparison, the groupings were located in the Arabian Sea and the Americas, with the west coast of Africa emerging as a new group. This reinforces the idea that there may be oceanic processes at work in these regions that respond to a common forcing. The location of the cores in these four groups can be seen in Figure 4.3.

4.1.3 Glacial vs Interglacial 2

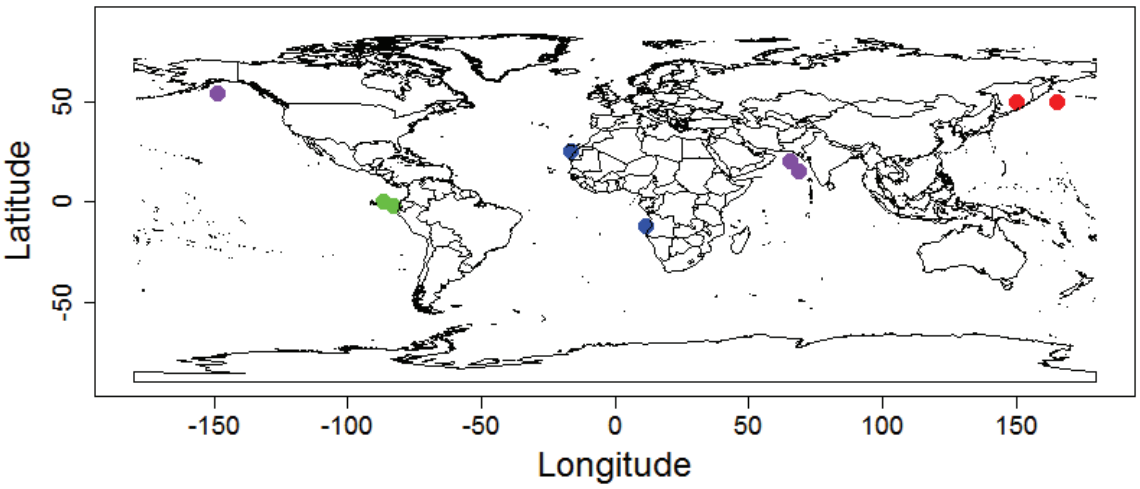
This comparison is the most interesting of the comparisons as it uses the two time periods for which we had the most confidence in our estimates of the $\delta^{15}N$ signal. (Note again that it is possible for a core to cover the IG2 but not the glacial period for reasons similar to those stated in Section 4.1.2.) Cluster 1 of the IG2 time period included 5 cores. Three cores were in both time periods: two from cluster 3 in IG1 on the west coast of Africa (group 1; GeoB 1016 and SU94-20bK) and one from Asia (MD 01-2386). Two remaining cores, neither of which were represented in the glacial time period (GeoB 1008 and GeoB 4240), were also near Africa.

The seven cores in cluster 2 in IG2 were located all over the world. Of those, four cores were present in the glacial time period, including two cores off the coast of Asia (group 2; GGC27 and MR 98-05-3) in cluster 2, along with MD 01-2386 near Africa. There were no other pairings in this cluster.

Cluster 3 only had one core that was in both time periods so no comparisons are possible.

Cluster 4, which had six cores mostly located near the Americas, had a group of

Figure 4.4: Plot showing the locations of the four groups of cores that were in the same cluster in both the glacial and interglacial 2 time periods: group 1 (GeoB 1016 and SU94-20bK) is represented by the blue circles; group 2 (GGC27 and MR 98-05-3) by the red circles; group 3 (ME 0005A 24JC, ME 0005A 27JC, and ODP 887) by the green circles; and group 4 (ME33 NAST, ME33 EAST and RC27 61) the purple circles.



four cores that came from cluster 2 in the glacial time period. Three of these cores were located on the west coast of the Americas (group 3; ME 0005A 24JC, ME 0005A 27JC, and ODP 887).

In cluster 5, which had six cores, there were 3 cores that were in cluster 1 of the glacial time period. All three of these cores were in the Arabian Sea (group 4; ME33 NAST, ME33 EAST and RC27 61). A fourth core from cluster 5, that was not in the glacial time period (RC27 24), was also located in the Arabian Sea.

Four distinct groups clustered together across the IG2 and glacial time periods. The groupings were located in the Arabian Sea, the Americas, the west coast of Africa and near Asia. Again, this suggests that there may be natural processes at work in these regions that allow these cores to continuously be grouped together independent of the time periods. The location of the cores in these four groups can be seen in Figure 4.4.

4.1.4 Comparisons Across All Time Periods

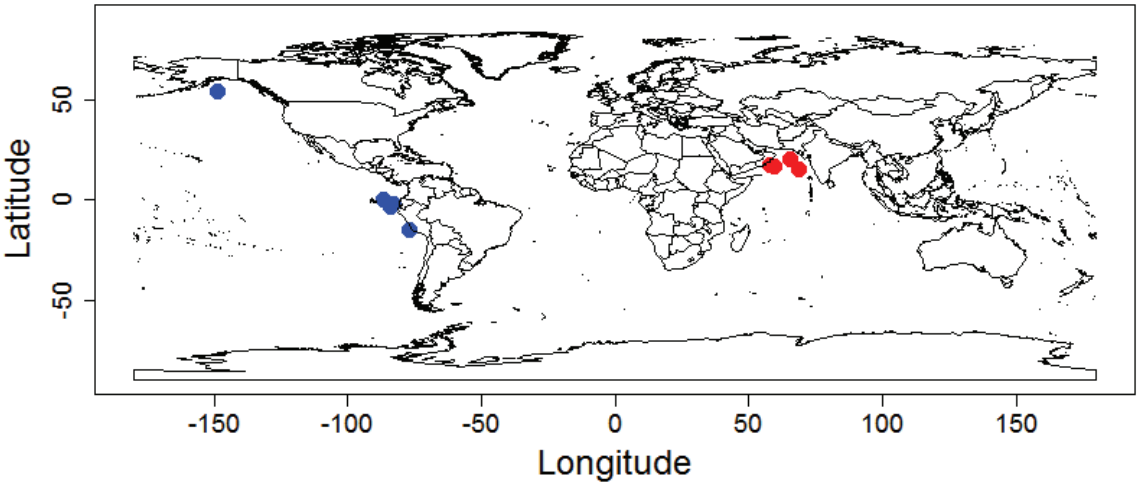
Looking across all three time periods, there are two groups of cores that stood out. The first consists of 5 cores on the west coast of the Americas: ME0005A 24JC, ME0005A 27JC, ODP 887, CD 38-02 and TR163-31 (locations of the cores can be seen in Figure 4.5). In this group, ME0005A 24JC, ME0005A 27JC and ODP 887 are clustered together in every time period. In the IG1 and IG2 time period CD 38-02 is also clustered with the three cores identified above, but it is removed from the cluster in the glacial time period. Core TR163-31 is clustered with the three cores in the glacial and IG2 time periods, but not in the IG1 time period. Looking into why TR163-31 was not grouped together with these cores in the IG1 time period it was found that it had a decrease in $\delta^{15}N$ during approximately the last 10,000 years of the IG1 time period while the other cores in this group did not. This is the only real substantial difference from the other cores. This 10,000 year time period accounts for a third of the total time span of IG1, so it is not surprising that the TR163-31 core was not grouped with the others in IG1. The $\delta^{15}N$ signal plots (Figure: 4.6) show that these five cores do in fact have similar signals overall - all have a distinctive decrease over the first half of the period and an increase starting at around 60,000 years ago that continues for the rest of the period.

The second distinct group was found in the Arabian Sea and consisted of four cores: ME33 NAST, RC27 61, ME33 EAST and RC27 24 (locations of the cores can be seen in Figure 4.5). ME33 NAST and RC27 61 were clustered together in every time period. While RC27 24 and ME33 EAST were paired either with those two cores or with each other in every time period that they covered. Looking at the signal plots for these cores, it can be seen that they too are also all very similar to each other over the three time periods in question (Figure: 4.7) .

4.2 Comparing our Clusters to Previous Groupings

As mentioned before, Galbraith *et al.* [5] clustered the samples in a previous study based on the biological "provinces" and common $\delta^{15}N$ signal in surface sediments.

Figure 4.5: Plot showing the locations of two groups of cores that were clustered together across all time periods investigated: group 1 (ME0005A 24JC, ME0005A 27JC, ODP 887, CD 38-02 and TR163-31) is represented by the blue circles; and group 2 (ME33 NAST, RC27 61, ME33 EAST and RC27 24) by the red circles.

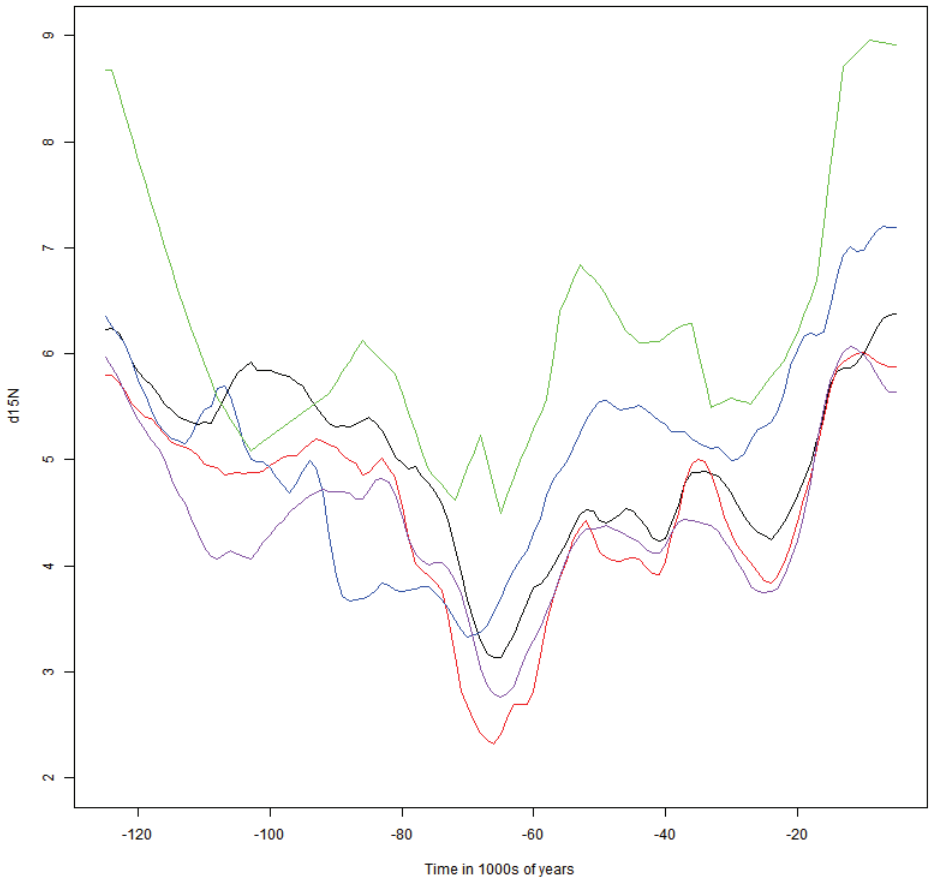


In total they had sixteen clusters and they only looked at the IG1 time period. The groupings were based on the location of the records only, and not on statistical similarity. Here we will compare the results found in Galbraith *et al*. [5] to our results. The first big difference in our results was the number of clusters. In our analysis we only used five clusters in all three time periods while Galbraith *et al*. [5] used sixteen.

In the IG1 time period we were only able to distinguish 1 geographical region in the west coast of the Americas. This one region encompassed cores that were located in three clusters defined by Galbraith *et al*. [5]. Galbraith *et al*. had three clusters one for the west coast of the Americas, for the west coast of Central America and one for the west coast of North America. Our one cluster covered all three of these regions so we did not find a difference in the cores located there.

When we looked into the clustering of the glacial time period we now had three defined regions. One of which was located in the Arabian Sea. When comparing this to the results in Galbraith *et al*. [5], they defined three clusters in the Arabian sea compared to our one. Two clusters appeared in the Americas in this time period,

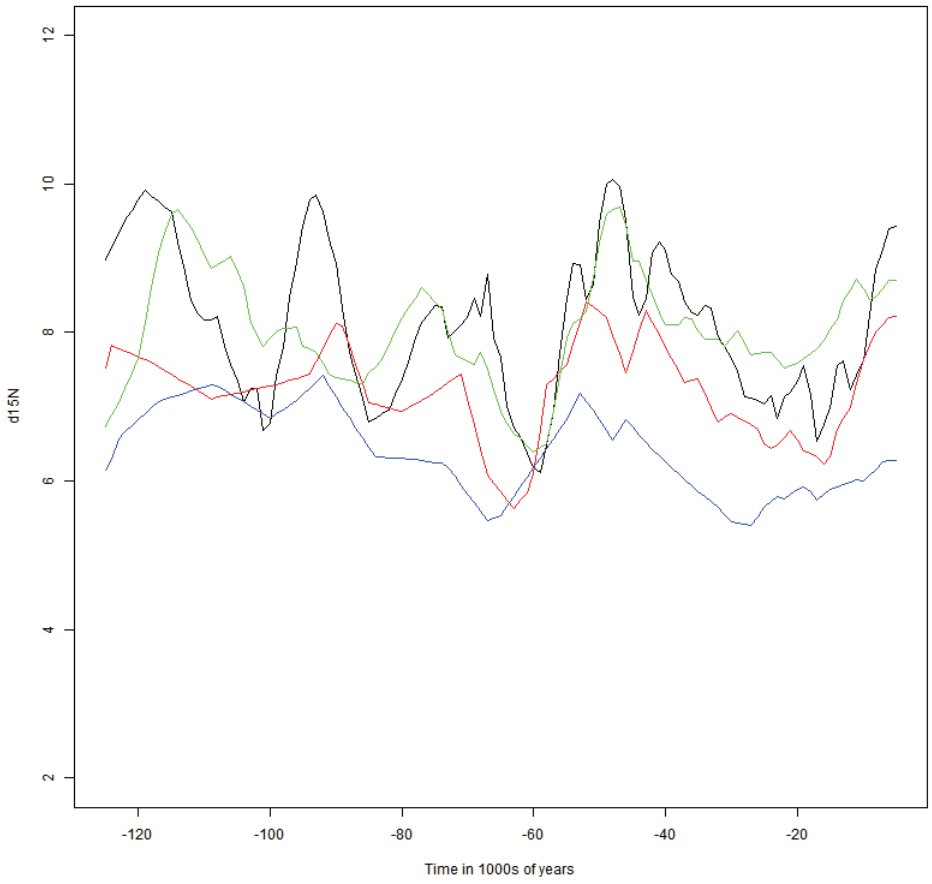
Figure 4.6: Plot of the $\delta^{15}N$ series of the five cores from the Americas: ME0005A 24JC (black); ME0005A 27JC (red); ODP 887 (blue); CD 38-02 (green); and TR163-31 (purple).



one core consisted of two cores and was located on the west coast of South America. This is in agreement with one of the Galbraith *et al.* [5] clusters. The second cluster in the Americas consisted of cores on the west coast of Central and North America. This again is a combination two clusters created Galbraith *et al.* [5].

Finally when looking into the IG2 time period we were able to distinguish three geographical clusters. The first was one consisting of cores on North and Central America. Again this was a combination of two clusters created by Galbraith *et al.* [5]. The second cluster was one in the Arabian Sea, again this cluster in the Arabian sea was a combination of three clusters created by Galbraith *et al.* [5]. The final cluster was one on the west coast of Africa. Again this was a combination of three clusters

Figure 4.7: Plot of the $\delta^{15}N$ series of the four cores from the Arabian Sea: ME33 NAST (black);, ME33 EAST (green); RC27 24 (blue); and RC27 61(red).



defined by Galbraith *et al.* [5].

Overall we were able to separate out similar regions to Galbraith *et al.* [5]. Galbraith *et al.* [5] defined the groups more precisely and narrower (geographically) than our clusters. The clusters that we were able to obtain tended to cover large geographical regions (i.e. the entire west coast of the Americas) instead of a small portion of the west coast. The reason for this is most likely due to the fact that our groupings were based on the standardized series and we did not take the magnitude of change or the absolute value of $\delta^{15}N$ into account. Figure 4.6 shows that CD 38-02 (green line) has much larger $\delta^{15}N$ values than the other four cores. In their groupings, Galbraith *et al.* [5] would have placed this core in a different cluster because of this magnitude,

while we were just interested in the trends in the signals.

Overall our analysis found similar results to that of Galbraith *et al*. [5]. The main difference the approaches taken were that Galbraith *et al*. forced cores into groups based on their geographical regions, while we grouped only based on statistical characteristics. Our analysis would be better suited to determining if there were any distinct signals of $\delta^{15}N$ in a given region because we did not use the core location as part of the clustering criteria. However, if one was interested in determining a regional mean for an oceanographic province, then the approach made by Galbraith *et al*. [5] might be better suited. A future study could look at combining these two clustering methods, that is redo the analysis that was conducted in this thesis on the extracted signal of $\delta^{15}N$, taking into account both the magnitude and geographical location of the $\delta^{15}N$ signals.

4.3 Post Hoc Analysis 2 - Interglacial 2 Subset Cluster Analysis

While running the analyses described in Chapters 2 and 3, we noted that there were a number of cores that we were able to fit accurately, but which did not show much fluctuation in $\delta^{15}N$ values. These cores were essentially a trend that did not fluctuate much outside of the 0.28 observation error threshold - that is, all observations were less than 0.28 apart. We decided to remove these cores as a second post hoc analysis and run the cluster analysis on the IG2 time period again. We reasoned that this might produce more defined regions. We chose to do this on the IG2 time period because it was the only time period that appeared to start forming three distinct groups (west coast of Africa, west coast of the Americas and the Arabian Sea) and it was the time period we were the most confident in our estimates of the state of $\delta^{15}N$.

From this post hoc analysis, we obtained new clusters that included only the cores off the Americas as well as one that included only cores on the west coast of Africa. The four cores in the Arabian Sea were still grouped together, with a single core in the Americas rounding out the group. This analysis showed that the cores that didn't have a wide range of $\delta^{15}N$ were just being grouped with other cores that had a similar

upward trends. They were not very similar and they may have been confounding the clustering results. By removing these cores we were able to get more distinct groups than we had with the original analysis. All figures for this result can be seen in appendix A2.

4.4 Limitations

The univariate analysis was not accomplished without some complications. First, our Kalman smoother did not perform well when there were a small number of data points (under 40). As a result, we might not have as accurate measurements of the $\delta^{15}N$ signal in the IG1 time period. Second, we had a lot of missing data that we needed to get estimates for on our constant time line. By condensing the data points on the constant time line we lost some of the information, which could have impacted the accuracy of the Kalman smoother estimates. Third, when we did the multivariate analysis, we assumed that the process error was constant over the entire time period. However when looking at some of the plots we can see that $\delta^{15}N$ values increase and decrease more rapidly over equal time intervals, which could indicate that the process error is not constant over the entire time period (Figure 2.2). If this is the case it would affect the accuracy of our Kalman smoother estimates. Another limitation of this analysis is that we assumed steps to be constant and known. However, the time line for each core is actually an interpolation between age fixed points (such as carbon 14 dating); so it is, in a sense, a random variable. Therefore, the age differences for each sample has some uncertainty in time that we did not account for. Finally, and perhaps most importantly, is that a dependence structure was uncovered in the process and observation error, which violated one of the assumptions we made in order to use the Kalman filter. These two errors are not independent of each other and this made estimation of the parameters (Q and R) based on MLE very challenging. It also forced us to fix the observation error to break the dependence structure. Even when moving to a Bayesian framework we were still unable to break the dependence structure. If we had more information on the distributions of the

process and observation errors we might have been able to break the dependence structure through the use of priors. However, since we did not have this information we were forced to fix the observation error at 0.28 to continue with the analysis.

Despite the problems summarized above, we obtained a reasonable estimate of the true signal of $\delta^{15}N$. The estimates in the longer time periods are probably more accurate than those in the shorter time periods, as it was shown that the more data points we had the better we could recover the true signal. Even with the problems stated above we were able to conduct a multivariate analysis that found numerous interesting trends in the $\delta^{15}N$ values as they changed through time.

4.5 Potential Areas for Future Investigation

Future studies on the signal extraction could try to remedy the problems stated in the preceding paragraph. One approach would be to create an adaptive Kalman smoother to account for the possible changes in the process error on the time line inside a given core. This could lead to more accurate results as it would not force a single process error on all points in the time line. Another line of investigation could try to fit a model where it allowed the time steps to be treated as random or uncertain. Since the time points we had were estimated from the data and not designated as the sample was being collected, there is an amount of uncertainty surrounding them. Finally, if we had more information on the process error and the observation error it might be possible to make more informative priors that could help break the dependence structure that we encountered throughout the parameter estimation process. As we did not know much about the process and observation error distributions, we used a relatively uninformative prior (a uniform distribution) which did help slightly with the estimation process. If we could make a more informative prior we might be able to break this dependence and jointly estimate the process error and observation error variances instead of being forced to fix one.

4.6 Conclusion

In conclusion, we were able to distinguish three distinct geographical clusters, one off the west coast of the Americas, one in the Arabian Sea and one off the west coast of Africa. These clusters became more distinct when the number of cores in the analysis decreased (via some cores not covering longer time periods), and when we analyzed the longer time periods (*i.e.*. glacial and IG2) where the Kalman smoother was shown to be more accurate. In all three time periods it was found that the optimal number of clusters was five, much fewer than the 16 used by Galbraith *et al*. [5]. This could indicate that, in given regions, certain forcings caused similar changes in $\delta^{15}N$ values, regardless of magnitude of $\delta^{15}N$, that might not have been present in other geographical regions. The cores that were grouped together off the Americas during most of the time periods shared similar $\delta^{15}N$ signals (see Figure 4.6). Similarly, the cores that were grouped together in the Arabian Sea over most of the time periods shared similar $\delta^{15}N$ signals (see Figure 4.7). However, the signals from the Arabian Sea were vastly different than the signals observed off the Americas. This provides additional evidence that there are similarities between the $\delta^{15}N$ records that are in close geographical proximity - as shown by both Tesdal *et al*. [3] and Galbraith *et al*. [5] - but also demonstrates that cores in different geographical regions have distinct signals of $\delta^{15}N$. If, in the future, we find more accurate ways to estimate and denoise the true signal of $\delta^{15}N$ from the noisy observations (like the examples stated earlier in this chapter) it could lead to more definitive results about the signals of $\delta^{15}N$ and, in turn, the changes in the nitrogen cycle that are unique to specific regions. Still, while we were able to find distinct signals in two regions over most of the time periods, the two most dominant signals observed were the two sharp increases in $\delta^{15}N$ at 60,000 and 20,000 years ago. These signals are not confined to specific regions and were found all over the globe, clearly indicating that the marine nitrogen cycling responded to a common forcing at these times.

Bibliography

[1] Douglas G Capone, Deborah A Bronk, Margaret R Mulholland, and Edward J Carpenter. *Nitrogen in the marine environment*. Academic Press, 2008.

[2] James N Galloway, Frank J Dentener, Douglas G Capone, Elisabeth W Boyer, Robert W Howarth, Sybil P Seitzinger, Gregory P Asner, CC Cleveland, PA Green, EA Holland, et al. Nitrogen cycles: past, present, and future. *Biogeochemistry*, 70(2):153–226, 2004.

[3] J-E Tesdal, ED Galbraith, and M Kienast. Nitrogen isotopes in bulk marine sediment: linking seafloor observations with subseafloor records. *Biogeosciences*, 10:101–118, 2013.

[4] Roger Bivand and Nicholas Lewin-Koh. *maptools: Tools for reading and handling spatial objects*, 2014. R package version 0.8-29.

[5] Eric D Galbraith, Markus Kienast, et al. The acceleration of oceanic denitrification during deglacial warming. *Nature Geoscience*, 6(7):579–584, 2013.

[6] Robert H Shumway and David S Stoffer. *Time series analysis and its applications: with R examples*. Springer, 2010.

[7] Paulo J. Ribeiro Jr and Peter J. Diggle. geoR: a package for geostatistical analysis. *R-NEWS*, 1(2):14–18, June 2001. ISSN 1609-3631.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[9] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[10] Florian Hartig. A simple metropolis-hastings mcmc in r, 2010. http://theoreticalecology.wordpress.com/2010/09/17/metropolis-hastings-mcmc-in-r/. accessed July 14, 2013.

[11] Florian Hartig, Justin M Calabrese, Björn Reineking, Thorsten Wiegand, and Andreas Huth. Statistical inference for stochastic simulation models–theory and application. *Ecology Letters*, 14(8):816–827, 2011.

[12] Matthew J. Vavrek. fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, 14(1):1T, 2011. R package version 0.3.0.

[13] Richard A Johnson and Dean W Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education, Inc, 2007.

Appendix A

Plots

A.1 Plots for the IG2 non standardized K means Analysis

Figure A.1: A Plot of the sum of squares within cores by numbers of clusters.

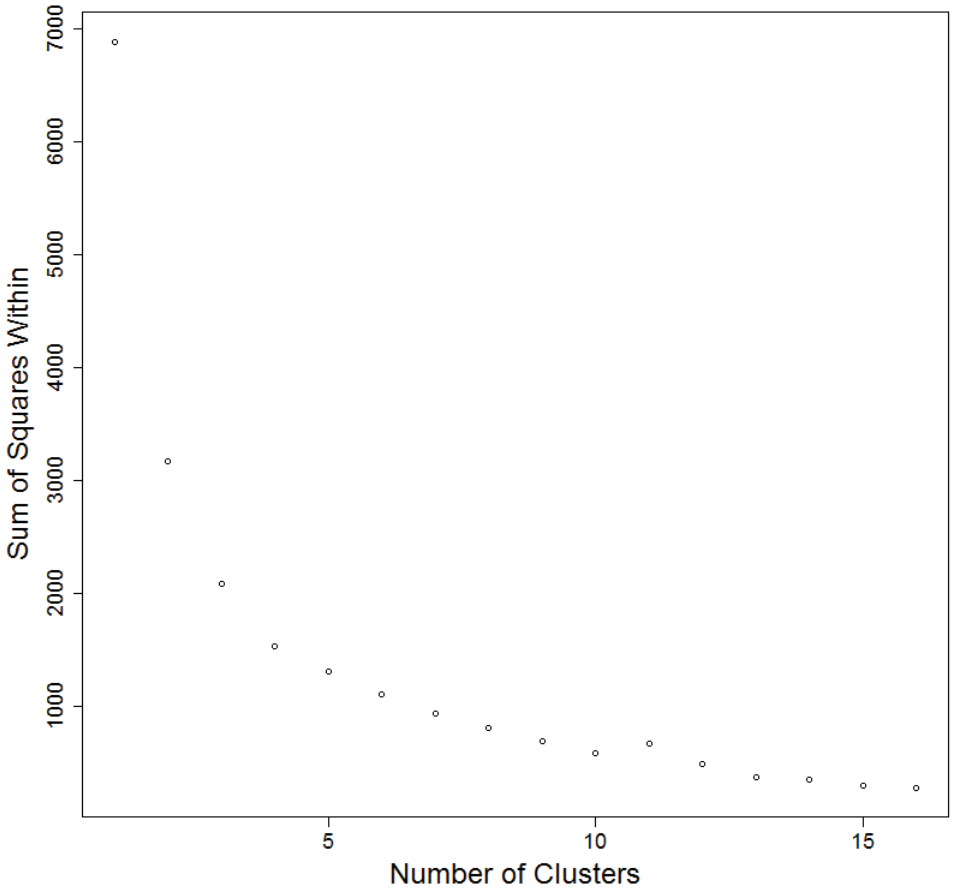


Figure A.2: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the interglacial 2 time period for the non-standardized series.

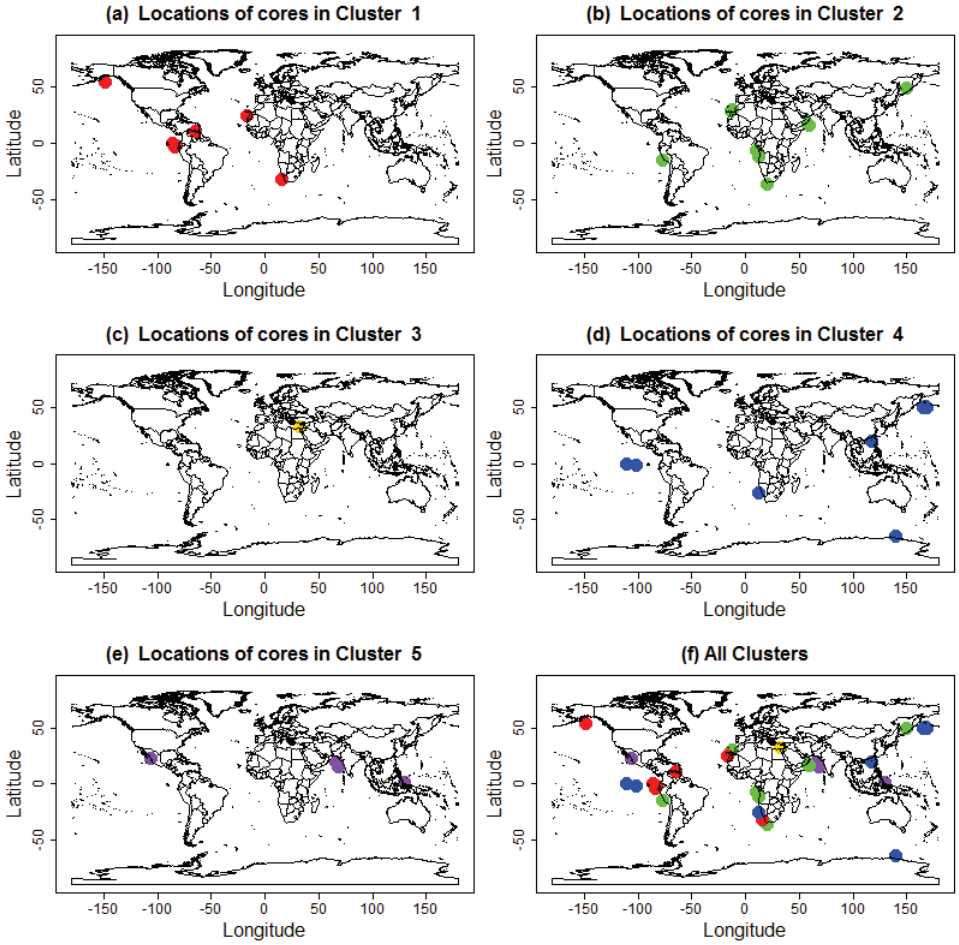
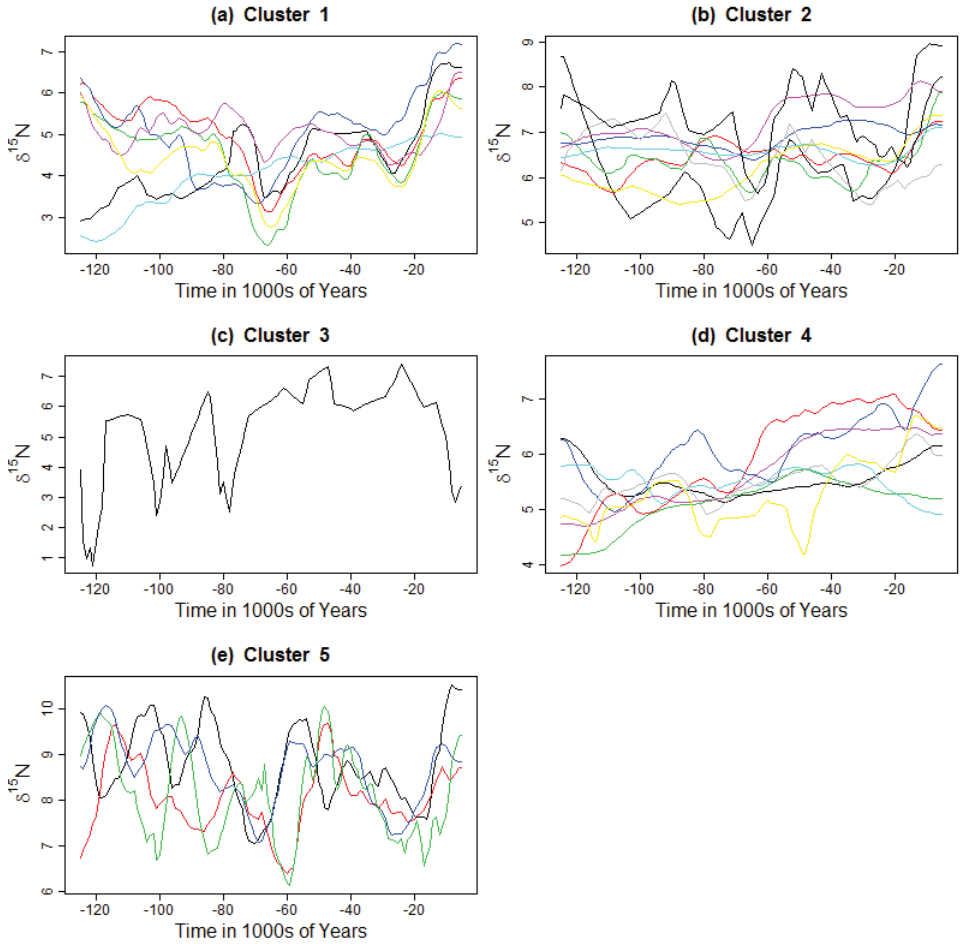


Figure A.3: Plots of the $\delta^{15}N$ series that fall within each of the five defined clusters in the interglacial 2 time period for the non-standardized series.



A.2 Plots for the interglacial 2 Standardized with selected cores removed from the analysis.

Figure A.4: A Plot of the sum of squares within cores by numbers of clusters.

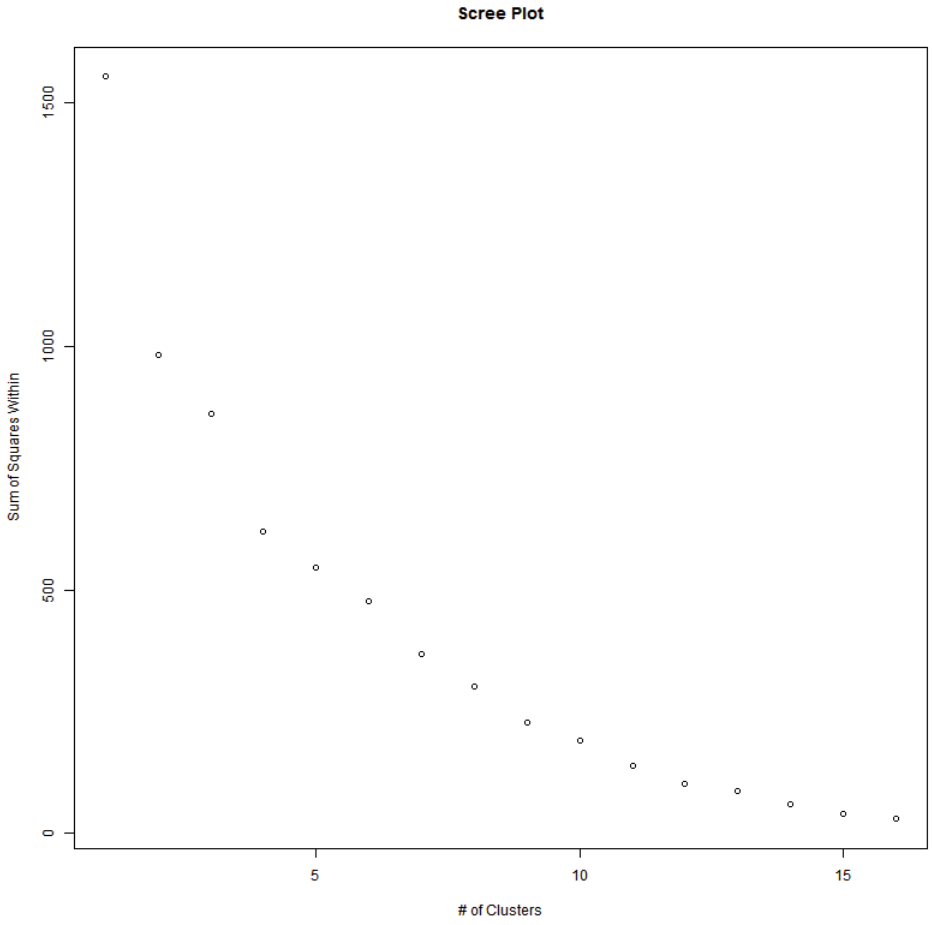


Figure A.5: Plots showing the locations of the cores within their individual clusters (a-e) and with all clusters combined (f) in the interglacial 2 time period for the standardized series.

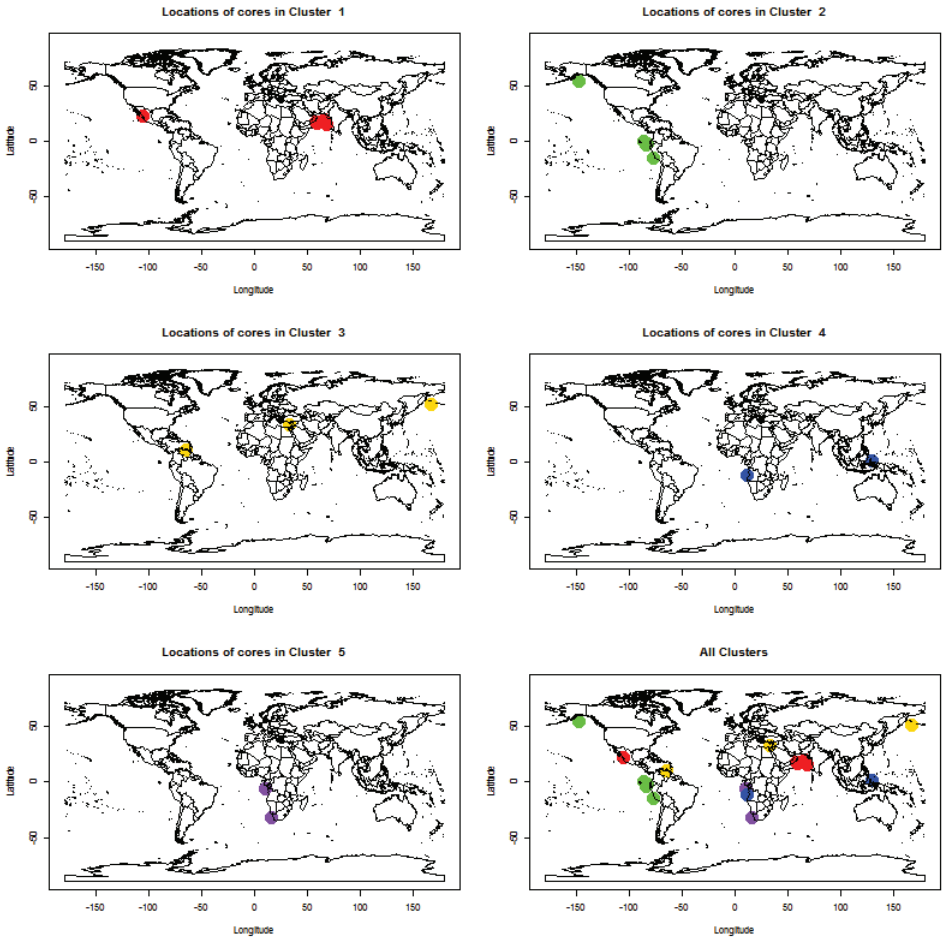
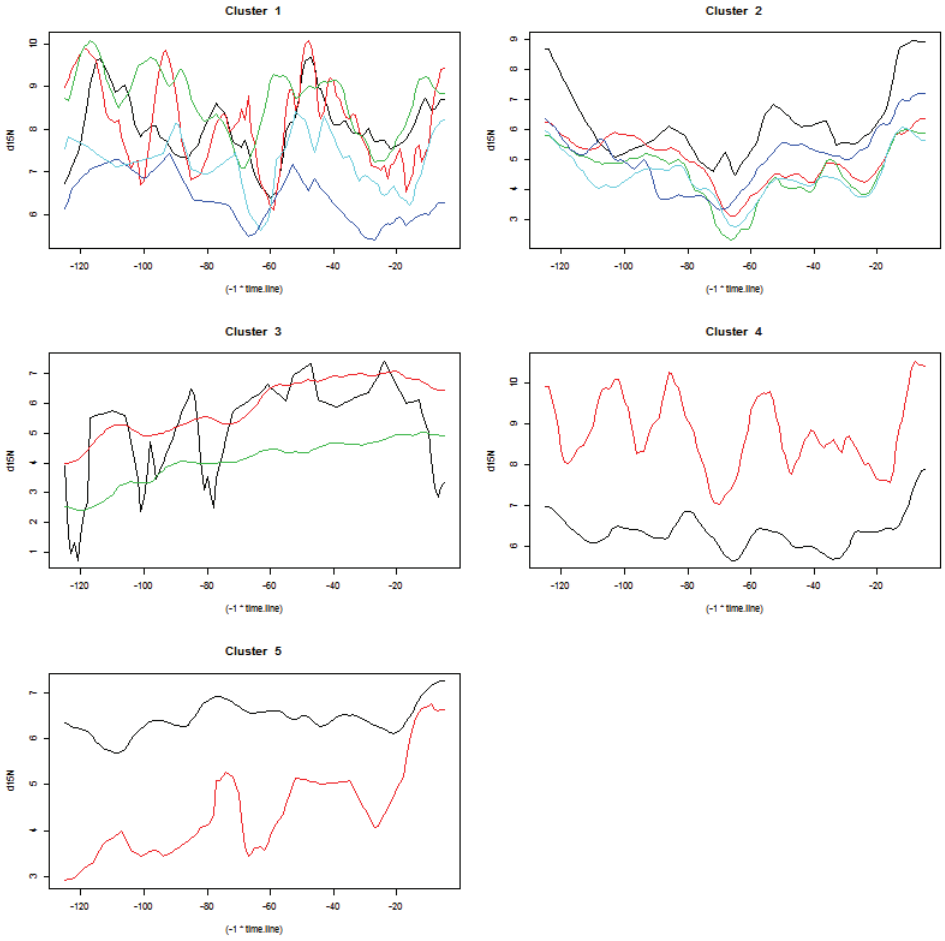


Figure A.6: Plots of the $\delta^{15}N$ series that fall within each of the five defined clusters in the interglacial 2 time period for the standardized series.



Appendix B

Markov Chain Monte Carlo Algorithm

How our MCMC worked was that you start off with initial values for the parameter, θ, where $\theta = (Q, R)$ equals the vector (Q,R). Next you calculate a proposal set of parameters $\theta^* = (Q^*, R^*)$ which is equal to the initial parameters plus a little bit of noise ($\theta^* = \theta + N(0, 0.05)$. For θ^* you calculated the log likelihood based on the Kalman filter log likelihood as well as the log likelihood of both Q and R based on their priors and sum them up. This is known as the posterior distribution:

$$Posterior \simeq e^{l(\theta)+l(\pi(\theta))}$$

Next the acceptance criteria is set by creating a ratio of the posterior likelihood of θ^* to θ, this ratio will be denoted by α.

$$\alpha = \frac{e^{l(\theta^*)+l(\pi(\theta^*))}}{e^{l(\theta)+l(\pi(\theta))}}$$

At this point a decision had to be made whether to keep θ or θ^* before moving to the next iteration of the chain. To do this we use the α value as a biased coined flip to determining which we keep based on the following conditions:

1. If $\alpha \geq 1$ then $\theta = \theta$
2. If $\alpha \leq 1$ then $\begin{cases} \theta = \theta^* & with \quad probability \quad 1-\alpha \\ \theta = \theta & with \quad probability \quad \alpha \end{cases}$

This means that the algorithm at each step keeps θ with a probability of α. This procedure was done 6500 times and the first 2500 were dropped from the data set as a burn in. The Q and R values at each of the 4000 iterations was stored and "chains" of parameters were used to get the estimates of Q and R.

The adaptive algorithm followed the same procedure as above with one exception. At every 100th iteration the MCMC calculated the acceptance ratio of the past 100 iterations (*i.e.*, if we are one iteration 1000 then it looked at iterations 901-1000 and computed the acceptance ratio of those 100 iterations). Based on the acceptance ratio one of three steps was taken.

1. If acceptance >0.4 then we increase the proposal function jump to $\theta^* = \theta + N(0, \sigma) * 1.1$.

2. If acceptance <0.15 then we decrease the proposal function jump to $\theta^* = \theta + N(0, \sigma) * 0.1$.

3. otherwise $\theta^* = \theta + N(0, \sigma)$.