VARIATION IN LOW-FREQUENCY UNDERWATER AMBIENT SOUND LEVEL ESTIMATES BASED ON DIFFERENT TEMPORAL UNITS OF ANALYSIS

A Thesis in

Acoustics

by

Russell S. Hawkins

© 2013 Russell S. Hawkins

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

August 2013
The thesis of Russell Stewart Hawkins was reviewed and approved* by the following:

Jennifer L. Miksis-Olds  
Senior Research Associate  
Associate Professor of Acoustics  
Thesis Advisor

R. Lee Culver  
Senior Research Associate  
Associate Professor of Acoustics

David L. Bradley  
Senior Scientist  
Professor of Acoustics

Victor W. Sparrow  
Professor of Acoustics  
Head of the Department of Acoustics

*Signatures are on file in the Graduate School
Abstract

Underwater ambient sound is by nature a nonstationary random process because of the variable duration of the contributing sound sources. A nonstationary label only identifies an absence of static properties and does not define the nature of nonstationarity. A universally-accepted methodology for analysis of underwater ambient sound does not exist. Because the distribution of underwater ambient sound does not always follow a Gaussian structure and is nonstationary in time, analysis over an extended period is required to accurately characterize the data. In examining long time series, large fluctuations in the ambient sound level caused by the variable duration of transient sounds create a need to determine the optimal sampling window based on a robust statistical analysis of variability. Failure to do so can lead to inaccurate representations of the actual ambient sound level.

Utilizing data from the Comprehensive Nuclear Test Ban-Treaty International Monitoring System (CTBT IMS) at Diego Garcia in the Indian Ocean, the temporal variation in ambient sound was examined over multiple time scales and sampling intervals. Periodicity and geographical region of sources were also examined. The data were analyzed using various window lengths and subsampling, or using data at specifically defined intervals, to determine the dependence of the characterization on sampling protocol. Detection of periodicity with subsampling was also examined.

Results indicated significant differences for both the varying window lengths and varying subsampling intervals in the 1-110 Hz range. The maximum difference between sound level estimates for the 10-30 Hz band due to subsampling was on average 2 dB re 1 μPa²/Hz and as high as 4 dB re 1 μPa²/Hz. The maximum difference for the full band was approximately 1 dB re 1 μPa²/Hz but as high as 6 dB re 1 μPa²/Hz. Window length analyses indicated significant differences of approximately 1 dB re 1 μPa²/Hz. Differences between the 60s and 200s subsample analyses indicated that the larger the window, the more susceptible the analysis is to variation due to subsampling. Detection of periodicity did not decrease due to subsampling. These results were used to recommend a general procedure for determining underwater ambient sound level estimate variation based on sampling protocols.
# Table of Contents

List of Figures ........................................................................................................ vi
List of Tables .......................................................................................................... ix

Acknowledgments .................................................................................................. x

Chapter 1

Introduction .............................................................................................................. 1
1.1 Sound in the Ocean .......................................................................................... 1
1.2 Nonstationary Random Data .......................................................................... 2
1.3 Sound Sources .................................................................................................. 3
1.4 Long-term Studies in Ambient Sound .............................................................. 7
    1.4.1 Segment Size ............................................................................................ 8
1.5 Periodicity ........................................................................................................ 9
1.6 Test Site Details ............................................................................................. 11
1.7 Goals, Objectives, and Questions .................................................................. 14

Chapter 2

Methodology and Procedures .............................................................................. 15
2.1 Introduction ..................................................................................................... 15
2.2 Data Collection ............................................................................................... 15
2.3 Signal Processing ............................................................................................ 16
2.4 Spectral Average Subsampling ....................................................................... 19
2.5 Distribution ................................................................................................... 21
2.6 Kruskal Wallis Test ....................................................................................... 23
2.7 Periodicity at the Northern and Southern Hydrophones .............................. 25
    2.7.1 Autocorrelation ....................................................................................... 26
Chapter 3

Results ........................................................................................................... 28

3.1 Introduction ............................................................................................ 28
3.2 Window Length Analysis ........................................................................ 28
3.3 Subsample Analysis ................................................................................ 33
  3.3.1 Using 60 second windows ................................................................. 33
  3.3.2 Using 200s Window Lengths ............................................................. 36
3.4 Summary of Subsample Results .............................................................. 40
3.5 Periodicity Results .................................................................................. 44

Chapter 4

Conclusion ........................................................................................................ 48

4.1 Ambient Sound Study Comparison ......................................................... 48
4.2 Frequency band discussion .................................................................... 49
4.3 Direction of Trend ................................................................................... 51
4.4 Differences between northern and southern hydrophones .................... 51
4.5 Periodicity discussion ............................................................................. 52
4.6 Unit of analysis recommendation ............................................................ 53
4.7 Procedure for selecting optimal unit of analysis .................................... 56
4.8 Recommendation for Future Work ........................................................ 59

References ...................................................................................................... 60

Appendix .......................................................................................................... 67
List of Figures

**Figure 1.1** Hydrophone Sensor being deployed and attached cable leading inland (www.ctbto.org)........................................................................................................................................... 12

**Figure 1.2** Organization of Hydrophones around Diego Garcia in the Indian Ocean (NOAA, 2012)......... 13

**Figure 2.1** Timeline of data used in analysis with red portions indicating missing days of data.
(Note: image is not to scale and does not include missing hours of data)........................................ 15

**Figure 2.2** Algorithm Schematic of steps in data analysis........................................................................ 18

**Figure 2.3** Distribution of the 1000 1 minute samples from Diego Garcia N1 starting 1-1-2005 and location of mode, median, and mean ........................................................................................................ 22

**Figure 2.4** Large variation due to self-calibration sequence that occurred on September 18th, 2004 at 8:20 p.m. ........................................................................................................................................ 26

**Figure 2.5** Daily averages over the full spectrum indicating the change due to the large transient event ........................................................................................................................................ 27

**Figure 3.1** 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference) .................................................. 29

**Figure 3.2** 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference) .................................................. 30

**Figure 3.3** 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. ........................................................................................................................................ 31

**Figure 3.4** 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference) .................................................. 32

**Figure 3.5** 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 33
Figure 3.6 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 34

Figure 3.7 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrow indicates location of significant difference) .................................................. 35

Figure 3.8 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrow indicates location of significant difference) .................................................. 36

Figure 3.9 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 37

Figure 3.10 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 38

Figure 3.11 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 39

Figure 3.12 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference) .................................................. 40

Figure 3.13 Significant/non-significant difference results from 100 randomized Kruskal-Wallis Analyses by frequency band. The bar graph is organized into groups by frequency band. Each group displays both hydrophones using both window sizes (60s and 200s). .................................................. 41

Figure 3.14 Distribution of p-values for 4 frequency bands. The red dotted line indicates the adjusted alpha value used for significant difference ($\alpha = 0.01$). Any peaks that fell above the red dotted line would be considered to have non-significant differences. The top figures are scaled from 0 to 0.015 alpha value to display the values which fell at approximately 0. The 40-60 Hz figure (lower left) indicates that all distribution peaks fell below 0.01 alpha value apart from the 60s window southern hydrophone which had a peak above 0.01 (the single peak is indicated with a green triangle). The 85-105 Hz band cumulative distribution (lower right) indicated two tests that had non-significant differences, which was the north and south hydrophone for the 60s window (peaks for this band are indicated with the blue and green triangles). .................................................. 42
**Figure 3.15** Median differences between highest and lowest subsample rates. In this figure, the median is designated by the red line with the 25th and 75th percentiles bounding the blue box. Red (+) symbols indicate outliers and are shown to demonstrate overall variation, while the whiskers indicate all other data points not considered outliers. ................................................................. 43

**Figure 3.16** 10-Year Spectrogram of Southern Hydrophone H08S2. The black arrows indicate signals that were present throughout the entire decade. The two lower black arrows point to the 10-30 Hz signal which seems to be very periodic in nature. ......................................................................................... 44

**Figure 3.17** 10-Year Spectral Autocorrelation of Southern Hydrophone H08S2 ........................................ 45

**Figure 3.18** SDL (Sound Density Level) over 10 Years of Data H08S2 ............................................................. 46

**Figure 3.19** Autocorrelation 10-30 Hz for H08S2 (2002-2012)............................................................................ 47

**Figure 4.1** Comparison Sound Pressure Level Plots for the Indian Ocean. Figure A depicts the sound pressure level found at Diego Garcia North for a six week time period using continuous 60s spectral averages, beginning on Jan. 1, 2005. Figure B is from the Gulf of Oman in the Indian Ocean. It consists of data from 5 different sites recorded during single months in 1977 (Wagstaff and Aitkenhead, 2005). Both figures are based on median sound pressure levels in the Indian Ocean. .................................................................................................................................................. 49

**Figure 4.2** Smith and Sandwell database with a ray stop criterion of the sound channel axis depth equaling that of the bathymetry (Pulli and Upton, 2001). This figure shows the bathymetry-based ray paths leading from both the northern hydrophones H08N (B) and the southern hydrophones H08S (A).................................................................................................................................................. 52

**Figure 4.3** The subsampling effect on the sound level from 10-30 Hz for hydrophone H08N1 on January 1, 2005 8:00am-4:00pm. The continuous 1 minute averages picks up the transients while the 60 minute subsample, which only has 1 data point every hour, excludes the high-intensity transient signals. ........................................................................................................................................... 54
List of Tables

Table 1.1: Low frequency marine mammal vocalizations relevant to Diego Garcia........................................... 5
Table 2.1 Subsampling Method in dB re 1 μPa²/Hz .......................................................................................... 18
Table 2.2 Averaging Method for calculating vectors for subsample analysis ..................................................... 20
Table 4.1 Calculation of Variables Parameters for Kruskal-Wallis test. In the following table, $N_{gris}$ is the number of groups, $S_{max}$ is the maximum subsample, $N$ is the FFT size, $f_s$ is the sampling rate, $S_{min}$ is the minimum subsample, $L$ is the length of time of the data set, $R_{max}$ is the maximum test range, $R_{min}$ is the minimum test range, and $\alpha$ is the alpha value ....................................................... 58
I would like to thank the Office of Naval Research for providing the opportunity and funding for my research. I would like to thank the CTBTO for providing the data which has been crucial to this study. I would like to thank the PSU’s Graduate Program in Acoustics for allowing me the opportunity to learn and grow in my understanding of acoustics.

I would like to acknowledge my advisor, Dr. Jennifer Miksis-Olds, who managed to coach me through the writing process and who spared no expense of painstakingly leading me through the thesis process. I would also like to thank Sam Denes and Chad Smith for always being willing to answer my coding questions and for being there when I needed help along the way. I would like to thank Dr. Lee Culver for the literature he provided which has helped me to complete this.

Mostly, I would like to thank my wife for all of her support and for her uplifting spirit. Without her hard work next to mine, I could not have completed this endeavor.
Introduction

1.1 Sound in the Ocean

Because sound in the ocean propagates tremendously well, many sounds can travel for long periods of time and spread over the entire ocean (Medwin et al., 2005). Therefore, the ambient sound level can be extremely high, stemming from the contributions of marine mammals, anthropogenic activities, and nature. An understanding of what is occurring in ocean dynamics can be gained simply by listening to the ambient sound. The purpose of this research was to quantify the variation of low-frequency ambient sound (1-110 Hz) based on its temporal variation in the Indian Ocean and determine what interval of analysis is most appropriate for that area.

Underwater ambient sound is the sound field against which a target signal must be detected. High ambient sound levels sometimes make it difficult to detect transient, or temporally short spanned, signals and other continuous signals (Richardson et al., 1995). If the signal-to-noise ratio (SNR) is too low, it may be impossible for a receiver to detect and subsequently identify any signal or variation in the sound level. Ambient sound is impacted by the number of sources as well as the efficiency of propagation of the contributing sound sources. An understanding of how the ambient sound level varies in the ocean will elucidate a broader description of the overall acoustic environment.
1.2 Nonstationary Random Data

Historically, beginning as early as the 1930’s, ambient sound measurements have focused on the mean square of the pressure (Carey and Evans, 2011). This is found from Parseval’s Theorem, which states that the integral of the spectral density multiplied over all time is equal to the mean square of the pressure (Bendat and Piersol, 2000).

\[ \int_{-\infty}^{\infty} |p(t)|^2 \, dt = \int_{-\infty}^{\infty} P(f) \ast P(f)^* \, df \] (1.1)

In equation (1.1), \( P(f) \) is the Fourier Transform of the pressure. However, this representation of the pressure squared is inconvenient for signal processing as it is impossible to apply an infinite integral on any amount of real data. Therefore, the equation can be rewritten in terms of a sampled signal version of Parseval’s Theorem.

\[ \sum_{n=0}^{N-1} p_n^2 \ast \Delta t = \sum_{m=0}^{N-1} |X_m|^2 \ast \Delta f = \sum_{m=0}^{N/2} G_{xx} \ast \Delta f \] (1.2)

In equation (1.2), \( X_m \) is the linear spectrum, \( \Delta f \) is the frequency width, \( N \) is the sample, and \( G_{xx} \) is the single sided spectral density. Determining generalities about the average mean square of the pressure for an ambient sound field is complex because the mean of the squared pressure can only be determined for a bounded, nonperiodic, and stationary pressure. Underwater ambient sound is by nature a nonstationary random process because of the variable duration of the sound sources contributing to it. If the spectrum level is analyzed in specific frequency bands, it is easy to identify the nonstationarity of the random data. Consequently, defining the mean of ambient sound over time can be more complicated, because it is time-varying. Creating a stationary conclusion for nonstationary data can lead to inaccurate representations of sound level, and a nonstationary conclusion only identifies an absence of static properties and does not define the nature of nonstationarity. Because of this, a universally-accepted methodology for analysis of underwater ambient sound does not exist.

A good visual representation of ambient sound over time is a spectrogram, which shows the single sided spectral density of ambient sound over time. At a single moment in time, a value over a specific frequency band from the spectrogram gives the spectrum level.
In this equation, \( G_{xx} \) is the single-sided spectral density at a specific sampled time \( t \), \( \Delta f \) is the window length, and \( f_1 \) and \( f_2 \) are the bounds of the frequency band being examined. However, averaging spectrum levels over time from the spectrogram only gives reliable estimates of the actual variation of the sound level if the distribution of the frequency specific spectrum levels is Gaussian. Because the distribution of underwater ambient sound does not follow a Gaussian structure, analysis over an extended period is required to accurately characterize the data (Bendat and Piersol, 2000). As the duration of the analysis sample increases, the number of transient events in that sample also increases. In low-frequency analysis, these transient events can occur quite often. A better understanding of what is occurring in ambient sound can be gained by understanding some of the contributing sources.

1.3 Sound Sources

To better understand how different sound sources contribute to the overall ambient sound level, we must understand the sound level variation. Since low frequency signals propagate energy extremely efficiently in water, it is difficult to predict how the low-frequency ambient sound will vary in a particular location (Medwin et al., 2005). The efficiency of propagation can also make it difficult to ascribe a specific sound to a localized source. The major variations of low frequency sound come from seismic activity, explosions, wind, marine animal vocalizations, and industrial activity (National Research Council, 2003). Each of these source groups may contribute to the ambient sound level over different frequency ranges, rates, and time scales.

Maximum levels from natural seismic sources occur between 2 and 20 Hz (Richardson et al., 2005). Seismic sounds can be 30-40 dB above the ambient sound level and can last from seconds to several minutes while the reverberations can continue for even longer (Schreiner et al., 1995). A tsunami near Thailand occurred on December 26, 2004, which caused an increase in the sound level of approximately 30 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \) between 1 and 100 Hz at Diego Garcia for more than 24 hours. It would have been nearly impossible to detect any other transient sources that day below 40 Hz. In locations near fault lines of frequent seismic activity, sound has been detected for up to 3
weeks continuously (Fox et al., 1995). A low level interminable background sound from the varying seismic motions of the earth has also been observed. Most often, geophysical activity in ambient sound is detected as a relatively short transient, or a series of short-duration transients (Wenz, 1962). Ambient sound analyses must account for the duration of the short signals caused by geophysical activity as well as the duration of the longer signals.

Surface waves caused by varying wind speeds, turbulent sea states, swell patterns, and the impact of raindrops are the dominant contributor of natural sound to the ambient sound level (Franz, 1959; Hildebrand, 2009). Locally, surface waves caused by the wind are the dominant contributor to the ambient sound field. Wave agitation caused by the wind contributes over frequencies from 1 Hz to 100 kHz. At locations where the ambient sound level is high enough (above 60 dB re 1 μPa²/Hz), the relationship between wind speed and ambient sound level is difficult to determine (McDonald et al., 2008). At low wind speeds (below 5 m/sec) ambient sound is distinctly independent of wind speed in the low frequency range (1-100 Hz). As the wind speed and direction can change instantaneously, the sound caused by breaking waves is difficult to categorize. It can be very continuous or extremely brief. However, there is evidence that there is a continual contribution from sea surface agitation to the ambient sound level although it may be masked by the higher level ambient sound sources, such as shipping (Hildebrand, 2009).

Although a single marine animal's vocalizations occur over time spans as little as a few seconds, single animals or groups can create sound that can continue for days or even weeks on end, which has been shown to dominate the soundscape in certain regions at specific times of the year (Watkins et al., 1987). Some blue and fin whales in the North Pacific have been shown to dominate the ambient sound field between 15-22 Hz up to 25 dB above the sound floor for large portions of the year (Curtis et al., 1999). In the southern Hemisphere, blue and sperm whales followed by fin whales are the most common whale vocalizations contributing to the ambient sound level (Sirovic et al., 2009; Gedamke and Robinson, 2010). However, since sperm whales are not known to vocalize in the low frequency range targeted in this study (1-100 Hz), the soundscape would most likely be dominated by blue and fin whales only, although there are other marine mammals that contribute to it (Table 1.1). Because each species vocalizes at different rates, it makes the decision of how to sample data much more important if the analysis seeks to include all of their contributions.
<table>
<thead>
<tr>
<th>Species</th>
<th>Signal Type</th>
<th>Frequency Range (Hz)</th>
<th>Dominant Frequencies (Hz)</th>
<th>Source Level (dB re 1 μPa at 1m)</th>
<th>Duration single call (s)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antarctic Blue</td>
<td>moans</td>
<td>12-390</td>
<td>16-25</td>
<td>188</td>
<td>6.0-18.0</td>
<td>Cummings and Thompson, 1971a</td>
</tr>
<tr>
<td></td>
<td>A downsweep phrase</td>
<td>17-18</td>
<td>17-18</td>
<td>188</td>
<td>11</td>
<td>Nieukirk et al., 2012</td>
</tr>
<tr>
<td></td>
<td>B downsweep phrase</td>
<td>16-18</td>
<td>16-18</td>
<td>188</td>
<td>9</td>
<td>Sirovic et al., 2007</td>
</tr>
<tr>
<td></td>
<td>moan</td>
<td>25-29</td>
<td>25-29</td>
<td>189</td>
<td>8.0-18.0</td>
<td></td>
</tr>
<tr>
<td>Sri-Lanka Pygmy Blue</td>
<td>Pulsive Call</td>
<td>20-44</td>
<td>25-35</td>
<td>-</td>
<td>18</td>
<td>Stafford et al., 2011</td>
</tr>
<tr>
<td></td>
<td>FM upsweep</td>
<td>56-72</td>
<td>56-72</td>
<td>-</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Long Tone</td>
<td>105-108</td>
<td>105-108</td>
<td>-</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>Madagascar Pygmy Blue</td>
<td>Six-pulse call</td>
<td>34-36</td>
<td>35</td>
<td>-</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FM downsweep</td>
<td>22-30</td>
<td>25</td>
<td>-</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>Fin</td>
<td>moans, downsweeps</td>
<td>14-118</td>
<td>20</td>
<td>160-186</td>
<td>1</td>
<td>Watkins et al., 1987</td>
</tr>
<tr>
<td></td>
<td>constant call</td>
<td>20-40</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tones, upsweeps</td>
<td>30-750</td>
<td>-</td>
<td>155-165</td>
<td>6.0-26.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rumble</td>
<td>10-30.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>downsweeps</td>
<td>15-35</td>
<td>15-30</td>
<td>183</td>
<td>0.9</td>
<td>Nieukirk et al., 2012</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15-28</td>
<td>15-28</td>
<td>189</td>
<td>1</td>
<td>Sirovic et al., 2007</td>
</tr>
<tr>
<td>Gray</td>
<td>moans</td>
<td>20-200</td>
<td>20-200</td>
<td>185</td>
<td>1.5</td>
<td>Cummings et al., 1968</td>
</tr>
<tr>
<td></td>
<td>blow sounds</td>
<td>15-175</td>
<td>15-100</td>
<td>-</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td>bubble sounds</td>
<td>15-305</td>
<td>-</td>
<td>-</td>
<td>0.7</td>
<td>Crane and Lashkari, 1996</td>
</tr>
<tr>
<td></td>
<td>moans</td>
<td>12-550</td>
<td>40-132</td>
<td>-</td>
<td>-</td>
<td>Stafford et al., 2007</td>
</tr>
<tr>
<td></td>
<td>pulse</td>
<td>35-150</td>
<td>35-150</td>
<td>-</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Humpback</td>
<td>moans</td>
<td>20-1800</td>
<td>35-360</td>
<td>175</td>
<td>0.4-0.9</td>
<td>Thompson et al., 1986</td>
</tr>
<tr>
<td></td>
<td>grunts</td>
<td>25-1900</td>
<td>25-400</td>
<td>190</td>
<td>0.1-0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pulse trains</td>
<td>25-1250</td>
<td>25-80</td>
<td>179-181</td>
<td>0.3-0.4</td>
<td></td>
</tr>
<tr>
<td>Minke</td>
<td>downsweeps</td>
<td>60-130</td>
<td>60-130</td>
<td>165</td>
<td>0.2-0.3</td>
<td>Edds et al., 1997</td>
</tr>
<tr>
<td>South Right</td>
<td>tonal</td>
<td>30-1250</td>
<td>235</td>
<td>-</td>
<td>0.2-4.1</td>
<td>Cummings et al., 1972</td>
</tr>
<tr>
<td></td>
<td>simple moans</td>
<td>70-320</td>
<td>160</td>
<td>-</td>
<td>0.6-1.6</td>
<td></td>
</tr>
<tr>
<td>Sei</td>
<td>downsweeps</td>
<td>34-82</td>
<td>34-82</td>
<td>156</td>
<td>1.4</td>
<td>Baumgartner et al., 2008</td>
</tr>
<tr>
<td></td>
<td>double sweep</td>
<td>40-120</td>
<td>40-80</td>
<td>179</td>
<td>1</td>
<td>Newhall et al., 2012</td>
</tr>
<tr>
<td>Bryde</td>
<td>downsweeps</td>
<td>57-232</td>
<td>75-208</td>
<td>152-174</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>double sweep</td>
<td>20-24</td>
<td>21</td>
<td>152-174</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>tonal call</td>
<td>36-38</td>
<td>37</td>
<td>152-174</td>
<td>2.7</td>
<td>Oleson et al., 2003</td>
</tr>
<tr>
<td></td>
<td>non-harmonic call</td>
<td>59-60</td>
<td>60</td>
<td>152-174</td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rapid call</td>
<td>41-43</td>
<td>42</td>
<td>152-174</td>
<td>1.3</td>
<td></td>
</tr>
</tbody>
</table>
Anthropogenic sound contributions are extremely varied in the interval over which they produce sound and frequencies they span. Explosions, seismic exploration, sonar, and shipping are just a few of the anthropogenic activities that produce large acoustic signals ranging from 0.1 Hz to over 100 kHz, and their durations are similarly varied, ranging from milliseconds to continuous (Blackman et al., 2004). Air gun arrays used in seismic exploration are a major topic of research in long-term ambient sound (Hildebrand, 2009; Blackman et al., 2004; Madsen et al., 2006). Air guns sounds are particularly broadband. A single air gun can have source levels up to 216-232 dB re 1 μPa at 1 m, but arrays of up to 70 air guns can emit signals at 261 dB re 1 μPa at 1 m. The rate at which air guns produce sound during active surveys is 1-6 sounds per minute over a time span of days to weeks (Nieukirk et al., 2012). The sounds from these airgun arrays can dominate the entire sound field for large spans of time. Nieukirk et al. (2012) found that in the mid-Atlantic ocean, the number of days in a month that had airgun sounds dominating the sound field exceeded 80% routinely. The peak energy in a spectrum for the signal from an airgun or airgun array can range anywhere between 1-100 Hz, though Blackman et al. (2004) calculated a high signal-to-noise ratio at 30-60 Hz in the Indian Ocean. Many analyses of airgun sound spectrum find most of the energy between 1-100 Hz (Nieukirk et al., 2012, Blackman et al., 2004).

The most heavily studied anthropogenic sound source in long-term ambient sound analysis is shipping. Most of the sound resulting from shipping has been shown to be a byproduct of propulsion systems (Ross, 1976). Recent studies of modern container ships report peak frequency of 40 Hz, while the faster container ships emit sound with a peak frequency of 100 Hz. These ships have source levels up to 190 dB re 1 μPa²/Hz at 1 m, with an average source level of 180 dB re 1 μPa²/Hz at 1 m. The signal from a passing ship can dominate much of the low-frequency spectrum in a single location for over 30 minutes. Because of the variety of ship noise spectra, determining an exact link between any single variable, such as size, will lead to incorrect assumptions about source level (McKenna et al., 2012). However, for the purposes of identification of sources, size and speed of ships may be instrumental in recognizing them. In the Indian Ocean, there is a relatively dense distribution of ships southeast of the Gulf of Oman, which is northwest of Diego Garcia (Wagstaff and Aitkenhead, 2005). This distribution of ships contributes to the ambient noise level because of the shipping lane from the Gulf of Oman to Southern Australia which passes just south of Diego Garcia. This is the closest shipping lane to Diego Garcia. This study would suggest that shipping should be considered as one of the primary sources to the ambient sound level in the Indian Ocean. Therefore,
analysis of ambient ocean sound in the Indian Ocean should consider contributions from local shipping.

### 1.4 Long-term Studies in Ambient Sound

Over the past 50 years, there have been numerous studies that have documented the increase in the ambient sound level over time. Most of the studies have concluded that since the 1950's up to around 1980, there has been an overall increase in the low frequency ocean noise budget by about 3 dB per decade in the Pacific Ocean (Andrew et al., 2002; Chapman and Price, 2011; National Research Council, 2003). If this pattern were to continue, it would have an effect on marine animals’ communication as well as affect the ability of humans to detect signals. Although the low frequency noise between 1960 and 1980 has increased steadily at 3 dB per decade, the rise in ambient sound level since 1980 has increased more gradually. Since 1980, the rate of increase has fallen to 0.25 dB per year or 2.5 dB per decade (Chapman and Price, 2011). While the rise in the ocean ambient sound level has been recognized by several studies, many of these studies obtain the results based on a limited amount of data and focused on particular areas, such as the Pacific Ocean (McDonald et al., 2006). The significance of trends in long-term ambient sound analyses could be impacted by variations in sampling protocol. Also, long-term studies of ambient sound have not been conducted in the Indian Ocean. Therefore, this project will provide another data point on the global scale effort.

Some long-term time series studies focus on the ambient sound without the transients (Merchant et al., 2012). This is referred to as the sound floor or background noise. Excluding transient signals is a method of sampling the ambient sound level although it should be defined differently than the ambient sound level. Basing an ambient sound measurement on only the low level interminable background sound gives no information on the sources contributing to the ambient sound level. Another method that has been used to measure the ambient sound level is subsampling, which is taking samples on a regular interval. This can be done with or without transients. However, subsampling could possibly lead to an inaccurate representation of the sound level if the sources contributing to it produce signals shorter than the amount of time between subsamples. Everything apart from the designated acoustic signal being analyzed is considered ambient sound and that includes other transient signals. A complete picture of ambient sound in the ocean can only be obtained through the inclusion of transient signals, and such an analysis requires careful selection of the subsampling rate.
Some transient signals are as short as a few milliseconds. An examination of the ambient sound level that omits periods of data can miss these transient signals and therefore the representation of the ambient sound can be skewed. However, many of the transient signals that contribute to the ambient sound are recurrent, such as marine mammal calls and rain, and therefore some of these could be detected using an irregular sampling strategy (Au and Hastings, 2008). The purpose of this research was to determine at what point subsampling of the ambient sound level causes a significant deviation from the continuous sound level. Then, it can be determined whether it is necessary to use continuous data or to exclude some data to reduce processing time and resources.

Long-term examinations of the ambient sound level require particular attention in choosing the unit of analysis because each source in the ocean contributes to the ambient sound level on a different time scale. These long-term analyses attempt to include all the necessary information while excluding redundant data (Curtis et al., 1999). The amount of data that has been used in previous studies of ambient sound fluctuates. Some studies of ambient noise have been limited in time and space. Wenz (1969) performed an analysis on the increase in ambient noise from 1963-1966, using 200s averages of the sound level three times per hour. Andrew et al. (2002) performed a similar experiment in the Pacific collecting three minutes of data every three to six minutes. Other recent research has followed the same pattern of selecting data on a limited basis, though few of these analyses offer any explanation as to the choice of analysis interval (Chapman and Price, 2011; Curtis et al., 1999). Yet other analyses use continuous data with no overlap between spectral averages to ensure accurate ambient sound representation (McDonald et al., 2006, 2008). It remains statistically unknown whether utilizing continuous data is necessary to accurately represent the sound level.

1.4.1 Segment Size

Wenz (1969) was one of the first to conduct an experiment on long-term trends in ambient sound. He, most likely due to processing and computational constraints at the time, chose to use 200 seconds of data 3 times every hour for each spectral average (McDonald et al., 2008). Therefore, since part of determining long-term trends in ambient sound is comparison with previous results, subsequent studies on ambient sound have used 200 s spectral averages. These analyses, however, use continuous averages with no overlap as opposed to Wenz’s 3 samples per hour. McDonald et al. (2008) made
a comparison to Wenz’s data by taking 179 days of continuous 200 s spectral averages to determine the overall ambient sound level in the Pacific Ocean. A three year study on noise from sea ice located in the Chukchi Sea was performed using the same parameters by Roth et al. (2012). This and many other examinations of long-term ambient sound have been established using a component of Wenz’ original choice of processing methodology, which was to have a 200 s window length.

In examining long time series, large fluctuations in the ambient sound level caused by the variable duration of transient sounds create a need to determine the optimal sampling window based on a robust statistical analysis of variability. Failure to do so can lead to inaccurate representations of the actual ambient sound level and impact differences between 2 different time periods. The degree of shift between mean sound levels with different window sizes, even using non-overlapping continuous data, could be significant. Therefore, determination of the ideal window size is a principal concern. Also, in the absence of a continuous data set, the degree of uncertainty should be accounted for in the representation of the sound level.

1.5 Periodicity

Acoustic periodicity is another important factor in selection of interval of analysis (Wenz, 1961). At times, the acoustic periodicity can be very difficult to detect amidst the background noise. This difficulty can be amplified if the resources used in the analysis are limited and more particularly if the analysis is based upon non-continuous data. It’s possible that the detectability of periodic signals decreases with subsampled data. In an autocorrelation algorithm to detect periodicities, the amplitude of the peak of the periodic event is based on the magnitude of the signal. However, if the ambient sound analysis is based on subsampled data, the magnitude of that signal could be reduced significantly in the correlation analysis. Subsequent analyses may not detect these periodic variations in such a situation.

Conversely, the time scales used in subsampling are vastly different than the time scales of the periodicities being examined. Marine mammals, for instance, have seasonal periodicities. Because each of these source groups can dominate the soundscape at certain times of the year, the periodicity of these sources should still be detectable, although with possibly a decreased amplitude. One consideration is whether subsampling
the data makes the autocorrelation amplitude decrease to a point that makes it impossible to detect against background sound levels.

Many marine mammals cause periodic variations based on their migration patterns as well as nocturnal activity. In the Pacific Ocean near Hawaii, the diurnal vocalization pattern of humpback whales has been shown to reverse depending on the season, and the number of calls also increases between February and April (Au and Hastings, 2000). Blue whales vocalize most often in the summer and autumn and the number of calls significantly decreases in the winter and early spring in the North Pacific (Stafford et al., 2001). Also in the North Pacific, the number of humpback whale vocalizations increases in sufficient quantity to increase the average noise level in January and February (McDonald et al., 2008). In the southern hemisphere, Antarctic blue whales migrate to the Indian Ocean in the winter months, specifically April through September (Branch et al., 2007; Stafford et al., 2003). Fin whales also migrate from the polar waters surrounding Antarctica to the tropical waters of the Indian Ocean in the winter months (Mizroch et al., 1984).

While distant shipping contributes very continuously to the sound level, local shipping lanes may have increased traffic at certain times causing periodic variations. Between May and July, shipping will increase globally and have a minimum between January and April (Schreier et al., 2007). However, while it is clear that there are seasonal increases in shipping activity throughout the year, no numerical evidence has yet been published to account for its contribution to seasonal variance in the underwater ambient sound level. Although seismic exploration occurs all year long, it increases during the summer time, from the late spring to the fall (Nieukirk et al., 2012). The peak energy detected from seismic exploration is 30-60 Hz, and during the summer months, the increase in activity can raise the daily median at that frequency up to 20 dB re 1 μPa²/Hz (Blackman et al., 2004; Klinck et al., 2012).

It has been shown that natural seismic activity occurs on a periodic basis as well. It increases at midnight, which is caused by microseisms brought about by the earth's diurnal thermal activity (Sidorin, 2010). Natural seismic activity is not only periodic on a diurnal scale but also a seasonal scale. During the winter, the activity of microseisms increases remarkably (Schimmel et al., 2011). In the Indian Ocean, because of the western fault line of the Indo-Australian tectonic plate, contribution to the ambient sound level stemming from nearby microseisms is possible. The periodic sound patterns exhibited by this fault line should be accounted for in an ambient sound analysis in the Indian Ocean.
Seasonal monsoons and changes in wind speed can also cause fluctuations in the low-frequency ambient sound level. Pigott et al. (1964) found an increase in the noise caused by wind-driven waves in the winter. He also found that this seasonal variation was independent of the frequency. This could be caused by the change in the thermal structure of the water column (Pigott et al., 1964).

An ambient sound analysis that bases its findings on limited datasets can omit acoustic contributions from any of these sources. The periodicities evident in these contributing sources make it important to determine how the subsampling rate affects the periodicity of sound. Consequently, in examining long time-series data sets for the variation in ambient noise, an investigation of the acoustic periodicity is advantageous in depicting the actual sound level and for interpreting sources and source contribution.

1.6 Test Site Details

Since 1990, the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) Preparatory Commission has been monitoring nuclear activity via seismic, hydroacoustic, infrasonic, and radionuclieic stations all over the world through the International Monitoring System (IMS). These monitoring stations are continually active and are therefore ideal for use in scientific endeavors such as underwater acoustics. The hydrophones utilized by the CTBT IMS are located in the Sound Fixing and Ranging (SOFAR) channel at an approximate underwater depth of 1200 meters, depending on location. They are suspended from sub-surface floats and connected to ocean-bottom anchors so that they can maintain the requisite depth. Individually protected, non-repeatered fiber optic cables extend up to 100 km from the hydrophones to the island station where the data is transmitted (Lawrence et al., 2004). At times, to guard the fiber optic cables against breakage, they are buried beneath the seabed to avoid the rough surf zones (Figure 1.1).
The first passive hydroacoustic station, which began acquiring data in 2002, was at Diego Garcia in the Chagos Archipelago Islands. The hydrophones are separated into triads, with three hydrophones located northwest of Diego Garcia and three located south of it (CTBTO Preparatory Commission, 2008) (Figure 1.2). The hydrophones are designated by the station number for Diego Garcia, H08, a letter indicating the geographical location (N for north, or S for south), and the number of the hydrophone (1, 2, or 3). The triangular configuration was designed to localize nuclear activity and therefore it can also be used to obtain directional information on the sources contributing to the ambient sound level. Because direct identification of individual sources will not be integral to the ambient sound level estimate analysis, the results are based on a single hydrophone in the north and a single hydrophone in the south (H08N1 and H08S2). At times the other hydrophones were also analyzed for corroboration of results from these two. The targeted bandwidth for these hydrophones is from 1 to 120 Hz; however the reliability of the data records decreases significantly above 110 Hz due to a steep roll-off from the anti-aliasing filter. The data set has been calibrated and corrected using hydrophone response curves.
The hydrophones at Diego Garcia were certified by the International Monitoring System (IMS) on the 18th of December, 2000. The data are continuously transferred via satellite to the US National Data Center and the International Data Center (IDC). These hydrophones undergo yearly inspections to verify they are working correctly. (CTBTO Preparatory Commission, 2008). The raw data collected by these hydrophones were formatted into two hour sections in a digital format applied by the International Data Centre (IDC). The IDC is a central element of the CTBT verification system. The IDC collects, analyzes and stores all of the CTBT IMS data.

The hydrophones located around Diego Garcia are separated by the Chagos Laccadive Plateau, which is an underwater sea mount. This bathymetry helps in the localization of the geographical region of sound sources, circumventing the need for complex calculation of exact location of a specific source using the triangular

![Organization of Hydrophones around Diego Garcia in the Indian Ocean](image)

**Figure 1.2** Organization of Hydrophones around Diego Garcia in the Indian Ocean (NOAA, 2012)
configuration. The land mass between the northern and southern phones acts like a barrier, separating the northern phones from some southern sound sources and vice versa. This is beneficial in determining the geographical region of the sources contributing to variation.

1.7 Goals, Objectives, and Questions

The goal of this research was to gain a better understanding of how the low frequency sound level and sources in the Indian Ocean change over time to determine the most appropriate unit of analysis for long-term ambient sound research.

Goal: Determine the temporal ambient noise variability near Diego Garcia.

Objectives: A) Determine the overall ambient sound level variation based on subsampling rate and window length at Diego Garcia.

Null Hypothesis – The subsampling rate and window length have no effect on the sound level estimates of the mean value of ambient noise in the Indian Ocean.

B) Determine how the sound level varies between the northern and southern hydrophones of the island.

Null Hypothesis – The ambient sound field detected by the southern hydrophones is not significantly different ambient sound field detected by the northern hydrophones.

C) Determine if the detectability of periodicity is impacted by subsampling.

Null Hypothesis – Peak amplitudes of periodic variations in the ambient sound do not decrease with subsampled data.

D) Based on the results of objectives A, B, and C, develop a recommendation for how to determine the most appropriate unit of analysis in long-term ambient noise trend analysis for any ocean region.
Methodology and Procedures

2.1 Introduction

This chapter presents the processing, methods, and analysis used to determine significant differences in ambient sound means based on different FFT-sizes and subsampling rates. The first section highlights the collection parameters and procedures. This is followed by an outline of the algorithm for calculating the mean sound level estimate, method of selecting the subsampled data, and the method of vector averaging. A discussion of the distribution and its use in the analysis is then presented. The Kruskal-Wallis test, a statistical test used to determine significant differences between non-normally distributed data, is then presented and the final section details the periodicity analysis.

2.2 Data Collection

Data from the CTBT IMS at Diego Garcia were used from Jan. 21, 2002 to Sep. 1, 2012. Although the data set spans 1 decade, there are certain portions where data are missing. Most of the time, the missing data only span a few hours, but there is one extended time period in particular that spans almost 1 month. That spans from July 13th, 2004 to August 18th, 2004 (Figure 2.1).

![Figure 2.1 Timeline of data used in analysis with red portions indicating missing days of data.](image)
>Note: image is not to scale and does not include missing hours of data)
2.3 Signal Processing

Two hydrophones were used in the analysis: one from the northern triad (H08N1) at Diego Garcia and one from the southern triad (H08S2). The selection of the second southern hydrophone was due to a sensitivity shift of hydrophone H08S1 during the decade.

Four bandwidths were selected for analysis of sound level and variation: 10-30 Hz, 40-60 Hz, 85-105 Hz, and 1-110Hz. One bandwidth was selected in consideration of the most common low-frequency marine mammal vocalizations (10-30 Hz). These vocalizations come from blue and fin whales (Sirovic et al., 2009). While seismic survey sounds are broadband and can be detected at any frequency in the 1-120 Hz band, one analysis by Blackman et al. (2004) indicated that the peak frequency range for these signals in the Indian Ocean is 5-60 Hz with highest SNR between 30-60 Hz. Therefore, the 40-60 Hz range was selected to best depict the contribution of airgun sounds that would be encountered in the Indian Ocean. An 85-105 Hz bandwidth was selected to be in the range of bulk shipping which has a distinct peak in the acoustic signature at approximately 100 Hz (McKenna et al., 2012). An analysis was also performed over the full spectrum (1-110 Hz). Many other sounds may overlap in these frequency ranges, so the analysis is simply a measurement of the degree of variation in each of the bands. If sources are contributing to the ambient sound level over different frequency bands, then one subsampling rate may not be universally constant in long-term ambient sound analysis. This analysis considers contributions from marine mammals, shipping sources, and seismic activity to determine an optimal unit of analysis for calculating an accurate sound level estimate.

The data set was recorded with a sampling rate of 250 Hz. It was converted to a spectral density plot using the following FFT sizes: 2500, 3750, 7500, 15000, and 50000, which correspond to the following window lengths: 10s, 15s, 30s, 60s, 200s. A Hanning window weighting with no overlap was employed to be consistent with previous ambient sound analyses (Andrew et al., 2002; McDonald et al., 2008; Wenz, 1961). The data were in partitions of two-hour intervals. So, the first step was to combine the data to be able to examine much longer time intervals regardless of the choice of unit of analysis. The data set is analyzed through various “for” loops in Matlab using a Matlab algorithm developed by Chad Smith (ARL, Penn State) which allows it to run through multiple 2-hour data sets, while simultaneously checking for data fragmentation, applying calibration parameters, and converting to a power spectral density.
The data set was imported into the Matlab algorithm as A/D counts and then converted into volts using an analog to digital counts conversion factor. A discrete window function was applied to the data and it was converted into a linear spectrum using a Fast Fourier Transform (FFT) to transform it to the frequency domain.

\[ FFT(\text{win} \ast \text{data}) \ast dt = X_m \]  

(2.1)

In equation (2.1), \( X_m \) represents the linear spectrum of the data, and \( dt \) is \( 1/f_s \). \( f_s \) is the sampling frequency. Correction and calibration factors were then applied and data were converted into a spectral density. Calibration was performed with individual hydrophone response curves. Due to the sensitive nature of the information contained the response curve, details cannot be provided. There was a decrease in sensitivity (1.5 dB) since installation on one hydrophone (H08S1), and for this purpose the second southern hydrophone (H08S2) was used instead. The resulting output is the spectral density over the selected time period. When applying the calibration factors, the data were also converted into pressure (\( \text{Pa}^2/\text{Hz} \)).

\[ A_w \ast \frac{2}{T} |X_m|^2 \ast \text{cal}_m^2 = G_{xx} \]  

(2.2)

In equation (2.2), \( G_{xx} \) is the single sided spectra, \( T \) is the time length, \( A_w \) is the corrected window function, and \( \text{cal} \) is the frequency dependent (m) calibration factor provided by the response curves. The sound density level was then calculated for each interval of time over the desired frequency bands or full spectrum.

\[ 10 \ast \log_{10} \left( \frac{1}{N} \sum_{n=m_1}^{m_2} G_{xx_n} \right) = SDL_{dB} \]  

(2.3)

In equation (2.3), \( N \) is the number of spectral averages. \( m_1 \) and \( m_2 \) represent the limits of the frequency band which may or may not fall directly on the specified value depending on the bin length given by the FFT size. The limits round to whatever bin value fell closest to the specified frequency band. The resulting sound density level units were dB re 1 \( \mu \text{Pa}^2/\text{Hz} \). This analysis focused on the mean level of the spectral density across a specific frequency band instead of the more commonly used sound pressure level found by energy summation over that window (Bendat and Piersol, 2000). This was necessary to have the results of the band levels be comparable to corresponding spectrograms. By averaging, the sound pressure level values were more representative of the values that would be seen on a spectrogram. Figure 2.2 shows the order that the algorithm performs the necessary calculations to find the spectrum level over each bandwidth.
The spectral density level values were saved for each sample of time (example: 1 minute). This created a table of sound pressure values over each frequency band upon which the statistical analysis was performed. Subsampled data were calculated the same way with the exception of spectral density levels being saved only at specifically chosen intervals (example: 1 minute every 4 minutes). The following table shows an example of the method of subsampling:

**Table 2.1** Results of Subsampling Method in dB re 1 μPa²/Hz

<table>
<thead>
<tr>
<th>Minute</th>
<th>Data Taken Every Minute</th>
<th>2 Minute Subsample</th>
<th>4 Minute Subsample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>99</td>
<td>99</td>
<td>99</td>
</tr>
<tr>
<td>2</td>
<td>98.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>98.2</td>
<td>98.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>98.4</td>
<td>99.3</td>
<td>99.3</td>
</tr>
<tr>
<td>5</td>
<td>99.3</td>
<td>99.3</td>
<td>99.3</td>
</tr>
<tr>
<td>6</td>
<td>100.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>100.5</td>
<td>100.5</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>100.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.4 Spectral Average Subsampling

To account for variation in vector sizes with subsampling, the spectral densities selected most often were averaged to equal the vector size of the minimum subsampling rate. For example, a comparison between a 1 minute subsample rate (non-overlapping continuous 1 minute segments of data) and a 60 minute subsample rate (1 minute of data every 60 minutes) would yield vastly different vector sizes. For every 60 points of the 1 minute subsampled data, there would be only 1 point of the 60 minute subsampled data. The 60 1-min values were averaged to get a single value so the vector lengths would be equal. The following formula was used to calculate each of the subsampled spectral averages.

\[ S_1 = 10 \log_{10} \left( \frac{N_1}{N_2} \sum_{k=1}^{N_2} Gxx_k \right) \]  

In equation (2.4), \( Gxx_k \) is the value of the spectral density over the specified frequency band for a single subsample \( k \). \( N_1 \) is the length in time of the subsampled data (i.e. 1 minute) and \( N_2 \) is the length in time of the comparison subsample (i.e. 60 minutes) and \( S_1 \) gives the combined average value for the matrix. The most appropriate test for non-Gaussian data is the Kruskal Wallis, which tests mean ranks of the groups instead of means. Averaging after subsampling results in equal sample sizes among groups giving equal confidence intervals in the analysis. More about the Kruskal-Wallis is discussed in Section 2.6. Because the FFT size dictates the smallest interval of subsampling, each of the subsampling rates were multiples of the original interval size. The following tables show examples of how the averages were computed. Data were averaged in intensity and converted back to decibels.
### Table 2.2 Averaging Method for calculating vectors for subsample analysis

<table>
<thead>
<tr>
<th>Subsample Rate (dB)</th>
<th>1 Min</th>
<th>2 Min</th>
<th>4 Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>99.0</td>
<td>99.0</td>
<td>99.0</td>
<td></td>
</tr>
<tr>
<td>98.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>98.2</td>
<td>98.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>98.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99.3</td>
<td>99.3</td>
<td>99.3</td>
<td></td>
</tr>
<tr>
<td>100.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.5</td>
<td>100.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Averages (dB)</th>
<th>1 Min</th>
<th>2 Min</th>
<th>4 Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.6</td>
<td>98.6</td>
<td>99.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100.2</td>
<td>100.0</td>
<td>99.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[10^{\frac{s}{10}} = I\]

\[10 \log_{10}(I_A) = s_A\]

**Intensity (Pa^2/Hz)**

<table>
<thead>
<tr>
<th>Intensity (Pa^2/Hz)</th>
<th>1 Min</th>
<th>2 Min</th>
<th>4 Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.0E+09</td>
<td>8.0E+09</td>
<td>8.0E+09</td>
<td></td>
</tr>
<tr>
<td>7.7E+09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.7E+09</td>
<td>6.7E+09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.9E+09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.6E+09</td>
<td>8.6E+09</td>
<td>8.6E+09</td>
<td></td>
</tr>
<tr>
<td>1.0E+10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1E+10</td>
<td>1.1E+10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.2E+10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[\frac{1}{N} \sum I = I_A\]

**Averaged Intensity**

<table>
<thead>
<tr>
<th>Averaged Intensity</th>
<th>1 Min</th>
<th>2 Min</th>
<th>4 Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3E+09</td>
<td>7.3E+09</td>
<td>8.0E+09</td>
<td></td>
</tr>
<tr>
<td>1.0E+10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.9E+09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.6E+09</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The size of FFT dictates the window length of the spectral density. Part of selecting an optimal unit of analysis is the selection of FFT size for analyzing the data. This may be different for different frequency bands. While one would expect the average sound level to be approximately the same regardless of the window length if the data set is continuous and non-overlapping, part of the analysis was to determine if this is true. The Kruskal-Wallis test, which can be used on data with distributions that are not normal, was used for determining significant differences between mean ranks based upon varying window lengths and subsample rates.
2.5 Distribution

The distribution of random data can be used to interpret many of its properties (Bendat and Piersol, 2000). For example, a Gaussian distribution follows very specific patterns in analysis that won't necessarily be true for non-Gaussian distributed data. More specifically, the most likely value in a distribution of data is not calculated by the mean in non-Gaussian distributed data. The probability structure of nonstationary data is not invariant in time but is a function of the time. So, the mean, median, and mode values of the data can also change with time. However, taking a large enough sample, the data can be treated as a stationary process to examine the distribution (Bendat and Piersol, 2000).

Preliminary analyses in Matlab indicated that the Indian Ocean ambient sound data estimates were not normally distributed, but followed most closely a lognormal distribution (Figure 2.3). The smoothing curve was applied using a Matlab kernel smoothing function called "ksdensity". Three points are highlighted in Figure 2.2. First, the mode was 77.0 dB re 1 μPa²/Hz. Second, the median value was 77.3 dB re 1 μPa²/Hz. Third, the mean value was 77.8 dB re 1 μPa²/Hz. Because the data set has a lognormal distribution, higher rates of transient events shift the mean sound level up above the median value.
There has been much debate in underwater ambient noise analysis about which value in a distribution best represents the actual sound level (Merchant et al., 2012; Parks et al., 2009; Welch, 1976). Many factors influence the selection of “best” representation such as source contribution, location, and even marine mammal hearing sensitivity.
Merchant et al. (2012) performed a study to determine the differences between the mean, median, and mode of the ambient sound level and suggested that the mean of the sound pressure level be used due to its relation to the sound energy level, which is the following,

\[ SEL = SPL_{\text{lin}} + 10 \log_{10} T \]  

(2.5)

In equation 2.5, \( T \) is the exposure period in seconds, \( SEL \) is the sound energy level, and \( SPL_{\text{lin}} \) is the linear sound pressure level. The disadvantage of this method of calculating the mean is that it is susceptible to overestimates. However, the robustness in variable averaging times indicates that this can be an accurate representation over long periods of time (Merchant et al., 2012). This research focused on how the mean rank of the data changes with subsampling; therefore, the distribution of data was crucial to understanding the ambient sound level over time. Because subsampling did not account for the periodicity of sources in ambient sound, subsequent subsampled distributions would not necessarily follow the same distribution as the parent. If a source periodicity coincides with the rate of subsampling, then the subsequent distribution should be shifted to accommodate that source's consistent reoccurrence. However, the subsampled distributions should also be narrower than the parent distribution due to samples being based on a much smaller group of data.

### 2.6 Kruskal Wallis Test

An analysis of variation (ANOVA) is often used to determine whether groups of data have significant variation between means. However, ANOVA is an analysis specified for normally Gaussian distributed data, and may not be appropriate for non-Gaussian distributions. If the variable is not normally distributed, the chance of false positives, or calculating variations as significant when they are not, increase (Dixon, 2008). The recommended test for non-uniformly distributed data is the Kruskal-Wallis test, which is a nonparametric generalization of the ANOVA (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996).

The Kruskal-Wallis test was used to determine if the shape and mean rank of the probability distributions were similar across different temporal windows. The Kruskal-Wallis analysis is a comparison between the mean ranks of the groups. Each of the data points is organized from the smallest to the largest and then given a value based on where it falls in the vector (Glass et al., 1972). Using the mean rank of each group
enables the Kruskal-Wallis to test significant variations among distributions that are non-Gaussian. The comparison p-value that was used to determine if differences were significant was based on a Bonferroni adjusted alpha value of 0.01. The Bonferroni adjustment is a statistical correction implemented to counteract the problem of family-wise error rate. As the number of tests increase, the probability of finding at least one significant result increases due to chance. This increased error rate is controlled by dividing the significance level by the number of groups being tested. In Matlab, simultaneous confidence intervals for the multiple comparison test are calculated based on the specified alpha value, sample sizes, and population standard deviation. Therefore, the confidence intervals would be the same size due to the vectors being the same length. A significant difference was found only when the confidence intervals between groups did not overlap.

Alpha values reflect statistical probabilities, which relate to the number of tests which are performed and the sample size. A 99% confidence interval is constructed to have a coverage probability of 0.99 based on the entire collection of confidence intervals for all values of x. Having equal sample sizes from the same distribution based on a probability alpha value 0.01 give equal confidence intervals. If a p-value is below the specified alpha value (0.01), then the test indicates the there is a significant variation between groups, rejecting the null hypothesis.

A single Kruskal-Wallis test only gives information about the specific time period being analyzed and doesn’t necessarily give information on the universal characteristics of the ambient sound field. Therefore, a bootstrapping method of 100 different Kruskal-Wallis analyses was implemented to determine the average result for each frequency band. This method departed from the standard procedure for a bootstrapping analysis because the bootstrapping analysis typically consists of a test over randomized points. The investigation of subsampling constrained the Kruskal-Wallis test to be run over consecutive times. So, they were not randomized. This test was run 100 different times over randomly selected 6 week time periods throughout the decade.

The significance of results from the bootstrapping analysis was determined by a normalized cumulative probability density plot of the Kruskal-Wallis p-values. Significance was indicated if the peak of the distribution was less than the alpha value of 0.01.
2.7 Periodicity at the Northern and Southern Hydrophones

The periodicity of seasonal sources was examined in an autocorrelation analysis to determine if the peak amplitudes of the periodicities of the ambient sound level changed with subsampling. An autocorrelation function was used to determine periodicities in the variation of ambient sound. The autocorrelation of a time series was calculated by reversing the series in time and then shifting with respect to the initial time series, multiplying, and then integrating.

\[
R_{xx}(\tau) = \frac{1}{T} \int_{0}^{T} x(t)x(\tau + t)dt
\]

Anytime the ambient sound level followed a similar shape and level, the autocorrelation function indicated a periodicity with a peak. Because the autocorrelation function of a time series with itself is one dimensional, it was adjusted to run only in the time direction along the entire spectrum.

Many periodic sources are seasonal. So, examination of the PSD over large windows (greater than two years) was expected to reveal seasonal periodicities in the variation of the ambient sound level (Wenz, 1961). Migration patterns of blue or fin whales in the Indian Ocean should be evident in an extended analysis at these hydrophones. Because of the large amount of data used, it was unrealistic to display every second of data for the entire decade. For this reason, the autocorrelation was based on daily averages of the spectral density. It should be expected that the periodicities from seasonal sources will still be visible from these daily averages because of the amount of the calls during these time periods. However, it is possible that an autocorrelation analysis based on incomplete data sets, or subsampled data, would make it impossible to differentiate the periodic events from the ambient noise in a long-term analysis. The periodicity analysis was performed over frequency ranges where clear repetition was visible in the spectrogram. Based on the results of the Kruskal-Wallis and post-hoc multiple comparison tests, which indicated if the amplitude of the peak decreased or remained the same, the most appropriate unit of analysis for long-term time series was recommended. The results of the autocorrelation function in coordination with the statistical tests helped to interpret the results based on knowledge of the contributions of ships, weather, and marine mammals.
2.7.1 Autocorrelation

To detect periodicities over the full spectrum, the following autocorrelation formula was used over each frequency bin (n).

\[
R(\tau) = \frac{E[(V(n)_t - \mu) - (V(n)_{t-\tau} - \mu)]}{\sigma^2}
\]

(2.8)

In equation 2.8, \(E\) is the expected value operator, \(\sigma\) is the variance, and \(\mu\) is the mean. \(V(n)\) is the variation from the mean at the frequency bin (n) (Box et al., 1994). This equation made it possible to view a spectrum level autocorrelation and to determine at what frequencies periodicities occurred.

2.7.2 Calibration Signals

Initial analysis of the sound level at Diego Garcia indicated some high intensity broadband transient events that occurred approximately monthly. These variations occurred simultaneously across all six hydrophones and were the same duration. After consulting the CTBTO Preparatory Commission, it was determined that these transient sounds were caused by self-calibration sequences of the hydrophones. Long-term quality monitoring is undertaken on these hydrophones to ensure high standard performance and data acquisition (CTBTO Preparatory Commission, 2008).

Figure 2.4 Large variation due to self-calibration sequence that occurred on September 18th, 2004 at 8:20 p.m.
These sequences are initiated by the command station to test whether the digitizer and associated electronics are functioning correctly for the electronics module. The input of each sensor channel is connected to a calibration generator instead of the outputs of the hydrophone. The sequence proceeds as follows: a four second no input is recorded by the generator, followed by random noise sequence (RBC), and then 10 discrete variable frequency sine waves, finally concluding in 38 seconds of no input. Even though these transient self-calibration signals are relatively short, lasting only 164 seconds, the overall level of the daily averages for these days, which are used for the periodicity analysis, are shifted by about 50 dB re 1 μPa²/Hz.

![Mean Sound Level for the week of September 16, 2006](image)

**Figure 2.5** Daily averages over the full spectrum indicating the change due to the large transient event.

Therefore, to remove any bias in the data, days in which self-calibrations occurred were eliminated from the final analysis.
Chapter 3

Results

3.1 Introduction

This chapter presents the results from the variable window length and subsample analyses as well as the periodicity analysis. It begins with a discussion of the window length analysis parameters followed by a presentation of the results. This is followed by a discussion of the subsample analysis parameters, 60s subsample results, and 200s subsample results. The results of the subsample analysis provides visualizations of the collective outcomes from subsampling, and the final section presents the spectrograms of the seasonal periodicity and periodicity subsampling results. Each multiple comparison figure displays the results of the Kruskal-Wallis analysis, accompanied by a probability density figure indicating the overall shape of the data distribution.

3.2 Window Length Analysis

The window length analysis was performed over 10 randomized 1-week time periods on both H08N1 and H08S2 throughout the decade spanning 2002 to 2012. The following results in this section provide an example from the 10-week analysis and are a selection from a single 1-week time period for each frequency band. The sample results presented were taken from 1 week of data starting October 1st, 2008 and ending October 8th, 2008. The window length analysis was calculated for an FFT size equaling the following: 2500, 3750, 7500, 15000, and 50000. With a sampling frequency equal to 250 Hz and 0 overlap, these values correspond to window lengths of 10, 15, 30, 60, and 200 seconds respectively. The average result of the window length analysis from the 10 randomized time periods is then summarized.

For the north hydrophone, the window length analysis over the 1-110 Hz frequency band often presented unexpected results. It was assumed the sound level estimate would
increase with a larger window length. As you increase the length of time for each spectral average, you also increase the number of potential transient signals. So, the sound level was expected to be higher. However, the 1-110 Hz frequency band showed that as the window length decreased the sound level estimate increased (Figure 3.1, left). The probability density curve indicated two peaks in the distribution of the data. For the selected time period on the northern hydrophone, there was a significant difference (H = 62.7, 4 d.f., p < 0.001). The results for the southern hydrophone were very similar (Figure 3.1, right). The overall sound level was about 2 dB re 1 μPa²/Hz higher than the northern hydrophone but the shape was comparable. There was also significant differences here, for this time period (H = 47.9, 4 d.f., p < 0.001). The multiple comparison for the north showed that the 10s window length had a greater mean rank than the 15s, 30s, and 60s window lengths. The south was the same except that there was no significant difference between the 10s and 15s window lengths.

**Figure 3.1** 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference)
The analysis of the 10-30 Hz bandwidth for H08N1 indicated that as the size of the window length increased, the mean value also increased, which was opposite of the full spectrum (Figure 3.2, left). There was a significant difference between the mean ranks based on the confidence intervals for the selected time period (H = 171.5, 4 d.f., p < 0.001). There was also a significant difference on the southern (H = 119.2, 4 d.f., p < 0.001) (Figure 3.2, right). The only difference between the northern and southern hydrophones was in the overall mean of the distribution which was about 7 dB re 1 μPa^2/Hz higher for the southern hydrophone. The significant differences are indicated with the arrows on Figure 3.2.

![Figure 3.2](image_url)

**Figure 3.2** 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference)

While the 10-30 Hz range showed a significant difference between window lengths over the example time period, the 40-60 Hz range failed to produce similar results (Figure 3.3). The mean rank of the sound level increased as the window length increased but not to a point where it would be considered significant at the 0.01 adjusted level (H
The overall increase in mean level between the largest and smallest window lengths was no more than 0.10 dB re 1 μPa²/Hz. The probability densities of the distribution over this time period show the similarities over the separate window lengths. The southern hydrophone, H08S2, over the same time period showed similar results. There was no significant difference (H = 10.1, 4 d.f., 𝑝 = 0.039).

![Figure 3.3](image.png)

**Figure 3.3** 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008.

The 85-105 Hz band, much like the 40-60 Hz band, failed to show significant differences between window lengths for the northern hydrophone for this time period (H = 12.5, 4 d.f., 𝑝 = 0.14) (Figure 3.4, left). The difference between sound level estimate means was less than 0.1 dB re 1 μPa²/Hz between the largest and smallest window sizes. The probability density showed very little difference between any of the curves indicating that in the 85-105 Hz band, there was approximately no difference in sound level estimates across any of the window lengths between 10s and 200s. The southern hydrophone showed similar shape in the multiple comparison (Figure 3.4, right).
However, unlike the northern hydrophone for this time period, the results from the southern hydrophone showed a significant difference between the groups \( (H = 33.4, \text{4 d.f., } p < 0.001) \). The significant difference results were isolated to this time period as tests over other time periods indicated no significant differences, which follows in the summary.

![Graphs showing multiple comparisons and probability densities for northern and southern hydrophones.](image)

**Figure 3.4** 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different window lengths (10s, 15s, 30s, 60s, 200s) from 10-1-2008 to 10-7-2008. (Arrows indicate location of significant difference)

The results of the 9 other randomized tests were very similar to the selected example. 100% of the tests over the 10-30 Hz band showed significant difference and 80% of the tests over the 1-110 Hz band showed significant differences. The 40-60 Hz band and the 85-105 Hz band showed non-significant differences the majority of the time. 90% of the tests over the 85-105 Hz band showed no significant difference, and 70% of the tests over the 40-60 Hz band showed no significant difference.


3.3 Subsample Analysis

3.3.1 Using 60 second windows

The subsample analysis was performed over 100 randomized 6-week periods throughout the decade. Significance was tabulated via a normalized cumulative probability density function. The following results are an example selection from a single time period for each frequency band. The results are taken from 6 weeks of data (1000 points) starting Jan 1st, 2005. The subsample analysis was calculated over 5 subsample rates (1, 5, 15, 30, and 60 min).

Although the difference in mean sound level estimates for the 1-110 Hz band was approximately 1 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \), the Kruskal-Wallis test indicated no significant difference at the northern hydrophone for the specified time period (\( H = 10.9, 4 \text{ d.f.}, p = 0.027 \)). A significant difference was found for the southern hydrophone (\( H = 498.2414, 4 \text{ d.f.}, p < 0.001 \)).

![Figure 3.5](image)

**Figure 3.5** 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)
The multiple comparison test for the 10-30 Hz band showed an increase in sound level as the subsample rate was increased (Figure 3.6). For the northern hydrophone, there was a significant difference ($H = 232.1$, 4 d.f., $p < 0.001$). On average, the difference between mean estimates of the largest subsampling rate and the smallest was approximately 2 dB re $1 \mu Pa^2/Hz$ but at times reached 4 dB re $1 \mu Pa^2/Hz$. The southern hydrophones showed similar results ($H = 572.3$, 4 d.f., $p < 0.001$) (Figure 3.6, right). The difference between the means of highest and lowest subsampled rates was approximately 2 dB re $1 \mu Pa^2/Hz$.

![Figure 3.6](image)

**Figure 3.6** 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)

In the 40-60 Hz range, the results of the subsample analysis were very different from the 10-30 Hz range (Figure 3.7). For the 40-60 Hz band, at the northern hydrophone, the analysis over the 6-week time period beginning on Jan. 1, 2005 showed no significant difference in sound level as the subsample rate was decreased ($H = 12.6$, 4 d.f., $p = 0.013$). The results for the southern hydrophone were similar, with the continuous sound...
level estimate being slightly above the subsampled sound level estimate. However, there was a significant difference for the southern hydrophone due the mean rank of the continuous sampling being higher than that of the 60 minute subsample ($H = 23.0567$, 4 d.f., $p < 0.001$).

Figure 3.7 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrow indicates location of significant difference)

Subsampling the data had very little effect on the 85-105 Hz band on the northern hydrophone (Figure 3.8). In this band, taking only 1 minute of data every hour gave the same results as using continuous data. For the example time period at the northern hydrophone, there was no significant difference ($H = 2.0$, 4 d.f., $p = 0.73$). There was, however, a significant difference for this time period at the southern hydrophone ($H = 20.2$, 4 d.f., $p < 0.001$).
Figure 3.8 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (1 min, 5 min, 10 min, 30 min, 60 min) from 1-1-2005 to 2-12-2005. (Arrow indicates location of significant difference)

3.3.2 Using 200s Window Lengths

The 200s subsample analysis was performed in the same way as the previous analysis. The results were taken from 6 weeks of data (1000 points) starting Jan 1st, 2005. The subsample analysis was calculated over 5 subsample rates (3.3, 10, 16.6, 33.3, and 50 min).

The 1-110 Hz multiple comparison test on the northern hydrophone for the selected time period showed a clear division between subsampling every 16.6 minutes and less frequent subsampling rates (H = 58.5, 4 d.f., p < 0.001) (Figure 3.9). The results for the southern hydrophone were similar, showing a significant difference between subsampling rates (H = 516.3, 4 d.f., p < 0.001).
Figure 3.9 1-110 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)

For the 10-30 Hz band over the 6 week period starting on Jan 1, 2005, there was a significant difference between subsampled sound level estimates for the northern hydrophone (H = 286.2, 4 d.f., \( p < 0.01 \)) (Figure 3.10, left). The difference between the highest and lowest subsampling rates means was approximately 3 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \). The southern hydrophone H08S2 showed similar results (Figure 3.10, right). For example time period, there was a significant difference between subsampled sound level estimates (H = 393.8, 4 d.f., \( p < 0.01 \)). The difference between the highest and lowest subsampling rates means was approximately 3.5 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \).
Figure 3.10 10-30 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)

For the northern hydrophone over the 40-60 Hz band, the variation in sound level estimate only differed by approximately 0.5 dB re 1 μPa²/Hz. However, this was a significant difference between the groups (H = 22.7, 4 d.f., p < 0.001). For the southern hydrophone, a significant difference was also found (H = 40.0, 4 d.f., p < 0.001).
Figure 3.11 40-60 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)

For the 85-105 Hz band at the northern hydrophone, for the example time period, there was no significant difference between the groups \( (H = 4.8, 4 \text{ d.f.}, p = 0.3) \) (Figure 3.12, left). For the southern hydrophone however, there was a significant difference between the groups \( (H = 40.0, 4 \text{ d.f.}, p < 0.001) \). This was due to a division between the three lowest subsample rates (3.3 min, 10 min, 16.6 min) and the two highest (33.3 min, 50 min) (Figure 3.12, right).
3.4 Summary of Subsample Results

In the following summary of results, the first figure shows the percentages of results that indicated significant differences (Figure 3.13). The second figure shows the normalized cumulative distribution of those p-values for each of the separate frequency bands for both hydrophones with both window lengths (Figure 3.14). In summary, the 40-60 Hz band and the 85-105 Hz band produced results indicating that approximately 50-90% of the time, subsampling had no significant effect on sound level estimates. The 10-30 Hz band and the full spectrum tests both indicated significant differences the majority of the time.

In general, subsampling the 60s window data set produced less of an effect than subsampling the 200s window data set, with the exception of the 40-60 Hz band at the northern hydrophone H08N1. Also, the majority of the time, the southern hydrophone

Figure 3.12 85-105 Hz multiple comparison and probability density of northern hydrophone H08N1 (left) and of H08S2 (right) over 5 different subsample rates (3.3 min, 10 min, 16.6 min, 33.3 min, 50 min) from 1-1-2005 to 2-12-2005. (Arrows indicate location of significant difference)
analysis indicated significant differences more often than the northern hydrophone analysis, with the exception of the 40-60 Hz band.

**Figure 3.13** Significant/non-significant difference results from 100 randomized Kruskal-Wallis Analyses by frequency band. The bar graph is organized into groups by frequency band. Each group displays both hydrophones using both window sizes (60s and 200s).

The subsampling of the data has a much larger effect on the sound level estimates in the 10-30 Hz band and the 1-110 Hz band. Each of these 100 randomized tests showed the sound level estimates increasing with increased subsample rate, but for the 85-105 Hz band (North and South) and 40-60 Hz band (South), this increase was not significant across subsample groups. Using the normalized cumulative distribution of the p-values, the null hypothesis can be either accepted or rejected. If the peak of the distribution falls below the alpha value line, the mean value of the sound level estimates between subsampling rates is significantly different (Figure 3.14).
Figure 3.14 Distribution of p-values for 4 frequency bands. The red dotted line indicates the adjusted alpha value used for significant difference ($\alpha = 0.01$). Any peaks that fell above the red dotted line would be considered to have non-significant differences. The top figures are scaled from 0 to 0.015 alpha value to display the values which fell at approximately 0. The 40-60 Hz figure (lower left) indicates that all distribution peaks fell below 0.01 alpha value apart from the 60s window southern hydrophone which had a peak above 0.01 (the single peak is indicated with a green triangle). The 85-105 Hz band cumulative distribution (lower right) indicated two tests that had non-significant differences, which was the north and south hydrophone for the 60s window (peaks for this band are indicated with the blue and green triangles).
For most of the frequency bands, the peak of the distribution function for the p-values fell below the designated alpha value of 0.01. For both the 10-30 Hz band and the full spectrum (1-110 Hz) the number of tests that indicated significant differences was very large. On average, this difference was approximately 2 dB re 1 μPa²/Hz for the 10-30 Hz band with outliers reaching 4 dB re 1 μPa²/Hz (Figure 3.15). For the 1-110 Hz band the difference between sound level estimates due to subsampling was 1 dB re 1 μPa²/Hz with outliers reaching 6 re 1 μPa²/Hz. Based on these results, it is assumed that the optimal unit of analysis for analyzing long-term trends in ambient sound is frequency dependent. For the full spectrum and 10-30 Hz band, any analysis that is not continuous can create significant uncertainty in the final sound level estimate. However, for the 85-105 Hz band and 40-60 Hz band, a 60 min subsample would give an equivalent sound level estimate to continuous sampling.

**Figure 3.15** Median differences between highest and lowest subsample rates. In this figure, the median is designated by the red line with the 25th and 75th percentiles bounding the blue box. Red (+) symbols indicate outliers and are shown to demonstrate overall variation, while the whiskers indicate all other data points not considered outliers.
The differences between the 200s subsample analysis and the 60s subsample analysis were that the 200s analysis revealed significant differences more often. Also, the shape of the probability density was less consistent in the 200s analysis. Most of the time the 200s window results indicated a significant difference approximately twice as often as the analysis over the 60 second window. This result implies that the larger the window length, the greater effect subsampling has on the mean sound level estimate. Because it is a larger window length, the ambient sound level is composed of more transient sounds per window length. Therefore, it is logical that a subsampling process would have a larger effect on the overall mean sound level estimate.

3.5 Periodicity Results

Preliminary analysis of autocorrelations over windows shorter than 1 year failed to return any easily discernible acoustic periodicities. On performing the autocorrelation on daily averages over one decade, clear periodic signals became evident (Figure 3.16). To determine whether subsampling had an effect on the detection of periodicity, seasonal signals were identified.

![Spectrogram H08S2 (2002-2012)](image.png)

**Figure 3.16** 10-Year Spectrogram of Southern Hydrophone H08S2. The black arrows indicate signals that were present throughout the entire decade. The two lower black arrows point to the 10-30 Hz signal which seems to be very periodic in nature.
Daily spectral averages with an FFT window length of 60 seconds over 1 decade revealed seasonal periodicities on both the northern and southern hydrophones. The periodicities in the 10-30 Hz range at the southern hydrophone were particularly clear. The spectral autocorrelation helped to identify periodicities (Figure 3.17). Periodic sources were increasing the ambient sound level in the 10-30 Hz frequency band seasonally. This signal is most likely from fin whales whose most common vocalization is a downsweep between 10-30 Hz and from Antarctic blue whales who often emit vocalizations around 30 Hz. Their periodic contribution to the ambient sound level could be due to the migration patterns of these whales in the Indian Ocean.

![10 Year Spectrum level Autocorrelation of H08S2](image)

**Figure 3.17** 10-Year Spectral Autocorrelation of Southern Hydrophone H08S2

Creating a line plot over the 10-30 Hz band for the decade revealed the periodic pattern even more. In Figure 3.18, the sound level for the 10-30 Hz band is shown from Jan 22, 2002 to September 1, 2012 using 5 different subsample rates. This data has been
smoothed with a 100 point moving average. The difference between the lowest subsampling rate (60 min) and the continuous was no more than 3 dB re 1 μPa^2/Hz.

Although the level of the subsampled data is decreased, the overall shape of the data appears to be unchanged. Most of the time, the sound pressure level increased approximately the same amount during each season, but toward the beginning of 2005, the overall level at this band dramatically increases.

![10-30 Hz Band for Station H08S2 with 100 Point Moving Average](image)

**Figure 3.18** SDL (Sound Density Level) over 10 Years of Data H08S2

Figure 3.19 shows the autocorrelation of the unsmoothed data which was then smoothed with a 100 point moving average function. The subsampling of the data had no effect on the detectability of the periodic sources. It is interesting to note that the correlation values actually increased for the subsampled data.
Figure 3.19 Autocorrelation 10-30 Hz for H08S2 (2002-2012)
Conclusion

4.1 Ambient Sound Study Comparison

Very few studies in low-frequency ambient noise have been done in the Indian Ocean. The sound level estimates calculated in this analysis were consistent with other ambient noise studies in the Indian Ocean. A pilot study of ambient noise prior to the official start of data collection at CTBT IMS Diego Garcia hydrophones was performed in 2001. It shows similar levels to those found in this study between 10-100 Hz (75 dB +/- 5 dB re 1 μPa^2/Hz) (Hanson, 2001). Spectral densities of the southern hydrophone were similar as well. In the Southeast Indian Ocean, levels have been calculated that are much lower than what was calculated here. At 100 Hz, the average spectrum level off the western coast of Australia was only 62 +/- 7 dB re 1 μPa (Cato, 1976). This is lower than the mean value found in this study (approximately 70 dB re 1 μPa^2/Hz at 100 Hz).

Wagstaff and Aitkenhead et al. (2005), performed an ambient noise measurement study in the Gulf of Oman in the northeast Indian Ocean. A sound pressure level comparison between this study and the current shows that median levels are slightly different between the two data sets. The overall level at Diego Garcia is slightly lower than the levels found off of the Gulf of Oman (Figure 4.2). Part of the reason for this difference could be due to the regularity and population of ships in each region. The Gulf of Oman has a very high distribution of ships. Therefore, a higher ambient sound level around the Gulf of Oman is logical (McKenna et al., 2012).
This research focused on the differences in ambient sound level estimates based on different units of analysis and quantifies the level of uncertainty introduced by different sampling protocols. Variation in the sound field gives an indication of the number of individual transient signals or of the transient component of discrete bands. Because sound propagates extremely well at low frequencies, sources can travel from particularly distant regions and contribute to the ambient sound field. The lower the frequency, the more efficiently the sounds can propagate. This may explain why the 10-30 Hz band had so many more transient events than the other bands.

For the 1-110 Hz band, the average difference between individual subsampling rates was approximately 1 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \) with outliers reaching 6 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \). Therefore, to accurately represent the mean ambient sound level, using anything other than continuous data increases the uncertainty of the final result by an average of 1 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \). This is an important result for comparative studies in ambient sound level trends. If different sampling protocols are used to compare one time period to another, then the mean ambient sound level can have an uncertainty between 1 and 6 dB re 1 \( \mu \text{Pa}^2/\text{Hz} \). For the 2-3 dB increase per decade which multiple studies have found, if that result was found via a comparative study using different sampling protocols, it is
probable that the value is inaccurate by a significant amount (Andrew et al., 2001; Chapman and Price, 2011; Wenz, 1969)

The frequencies between 10-30 Hz are dominated by both natural and anthropogenic seismic events and by marine mammal sounds, particularly blue whales and fin whales. The number of transient sounds encountered in this frequency band is larger due to the efficiency of propagation of low-frequency sounds. The window length results showed that the difference between mean sound level estimates were the largest over this frequency band. Between the largest and smallest window size, on average, there was a 2 dB re 1 μPa²/Hz between means for the northern hydrophone. There was less of a difference on the southern hydrophone H08S2, though still significant. The subsample results in the 10-30 Hz band were the most consistent. 100% of the time for all but one analysis indicated significant differences in the mean sound level estimate. At times the difference between subsampled sound level estimates reached 4 dB re 1 μPa²/Hz. The frequent contribution of sound sources in this range coupled with the efficiency of propagation illustrates the reason for the difference in sound level estimates in this region. In the 10-30 Hz range, there are a far greater number of high-intensity transient events than some of the higher frequency bands. Because the ambient sound field in the 10-30 Hz range is so variant due to transient events, any analysis of the ambient sound level in the Indian Ocean that used a subsampling procedure would find a significant decrease in the estimation of the mean SPL.

In the 40-60 Hz frequency range, contributions from marine mammal calls are not as numerous, though still present. Seismic airguns have spectral peaks in this frequency range, and nearby and distance ships also contribute to the ambient sound level here. The results of the window length analysis and the subsampling analyses over both window lengths show that the ambient sound level in this range is far more stable than the 10-30 Hz range. However, the distribution of the p-values for the bootstrapping analysis indicated rejection of the null-hypothesis, which was that the means across subsampling rates were equal. Subsampling this frequency range consistently can result in slightly inaccurate mean sound level estimates. However, the 5 minute subsample sound level estimate more frequently showed no significant difference between any other subsampling mean sound level estimates. Therefore, the 5 minute subsample would be optimal for this range to minimize computational resources while conserving precision.

The 85-105 Hz band most consistently returned results that indicated no significant difference in variation (approximately 70% of 400 analyses). Distance and local shipping, marine mammals, and seismic signals are among the sources that would be expected to
contribute to the ambient sound level in this frequency band. A large majority of the time, results indicating non-significant differences between mean sound level estimates were observed. Because this band showed no significant difference between subsampling rates most of the time, an extremely infrequent sampling method (as low as 1 minute every hour) will return accurate representations for the mean sound level estimate.

4.3 Direction of Trend

In almost all cases, the subsample analysis indicated that as the amount of data decreased, the ambient sound level estimate also decreased. This trend was always evident even if the probability test indicated that it was not significant. A similar trend was observed in the window length analysis as well. That is, that the sound level estimate decreased with continuous data. These trends are logical considering the nonstationary nature of ambient sound. If the data set could be treated as stationary, we would not only expect approximately equal sound level estimates, but also similar distributions (Bendat and Piersol, 2000). The trend in sound level estimates would not have increased with continuous data. The nonstationarity of the data drives these results. The transient component of the ambient sound level is what drives this nonstationarity. However, the degree of the nonstationarity varies depending on what frequency band is being examined. For the 10-30 Hz band and the 1-110 Hz band, the trend shown in the multiple comparison tests was much more pronounced, indicating that the transient component and therefore the nonstationarity is much larger than for either the 40-60 Hz band or the 85-105 Hz band.

4.4 Differences between northern and southern hydrophones

While results followed similar shape and agreed across frequency bands, there were some poignant differences between the northern and southern hydrophones ambient sound fields. Even though the magnitude of difference calculated on the northern hydrophone was larger, the southern hydrophone returned more results indicating significant differences between means due to differences in subsampling rates and window lengths. This leads to the conclusion that the ambient sound field recorded by the south hydrophones has a larger transient component than that recorded by the northern hydrophone. This could be a result of a greater number of sources, closer sources, or increased rates of source contribution detected by the southern hydrophone. This is
reasonable as the southern hydrophones have larger bathymetry-based coverage area considering propagation paths (Pulli and Upton, 2001). The Smith and Sandwell database shows bathymetry based ray stop criterion at Diego Garcia, which indicates how the coverage area is larger at the southern hydrophone (Figure 4.1). Based on this, some sounds detected from the south by the southern hydrophone will be less detectable by the northern hydrophone due to the bathymetry interfering with the SOFAR channel propagation.

![Figure 4.2](image_url)

**Figure 4.2** Smith and Sandwell database with a ray stop criterion of the sound channel axis depth equaling that of the bathymetry (Pulli and Upton, 2001). This figure shows the bathymetry-based ray paths leading from both the northern hydrophones H08N (B) and the southern hydrophones H08S (A).

### 4.5 Periodicity discussion

Subsampling the data set over the selected time periods (1 min to 60 min) did not affect the detection of periodicity. Subsampling the data as infrequently as 1 minute every 2 hours returned equally detectable seasonal periodicities in the 10-30 Hz band. In fact, the subsampled data usually returned higher amplitude peaks in the autocorrelation analysis. This is due to the differences in time scales which were being examined. The subsampling of the data set only scaled 2 hours at the most, and the periodicities being examined in the long-term analysis were seasonal. If the subsampling were done on a seasonal basis (selecting only a few months of data), then the periodicities would be more difficult to detect. The same result would follow for examining hourly or daily periodicities with hourly subsampling. Though the current analysis was only
investigating long-term time scales (multiple years), if the same analysis was to be done on short term scales, the hourly subsampling could have a much larger effect on the detection of daily or hourly periodicities. However, this analysis showed that the subsampling of the data set on hourly time scales does not decrease the peak level of the periodicity in the autocorrelation analysis.

4.6 Unit of analysis recommendation

The recommendation for unit of analysis in long-term ambient sound level trends in the Indian Ocean is both frequency-dependent and location-dependent. Any type of analysis focusing on this low-frequency range is going to be affected by the lower frequency end of the spectrum. In this study, the results from the transient component of the 10-30 Hz band affect the selection of unit of analysis for the whole spectrum. The differences between the 60s and 200s analysis indicated that the larger the window, the more susceptible the analysis is to variation due to subsampling. Therefore, the smallest window possible within computational time constraints would be recommended for the analysis. Since the smallest window examined here was 1 min, it is recommended to use this for the window length at Diego Garcia. Subsampling the data is also not recommended as the uncertainty in the final mean value can be up to 6 dB re 1 μPa^2/Hz.

The 10-30 Hz window constrains the selection of sampling protocol to be continuous. In other words, no subsampling would be advisable for the 1-110 Hz spectrum. The number of transient signals present in the 10-30 Hz bandwidth makes the entire spectrum susceptible to bias if any data are absent. Figure 4.2 shows an illustration of the same quantitative results demonstrating that the subsampling of the sound level will omit enough of the transient signals to have a significant effect on the mean sound level estimate.
Figure 4.3 The subsampling effect on the sound level from 10-30 Hz for hydrophone H08N1 on January 1, 2005 8:00am-4:00pm. The continuous 1 minute averages picks up the transients while the 60 minute subsample, which only has 1 data point every hour, excludes the high-intensity transient signals.

The degree of shift of the ambient sound level estimate due to subsampling for the 10-30 Hz band was as high as 4 dB re 1 μPa²/Hz and on average 2 dB re 1 μPa²/Hz. This degree of uncertainty is hugely significant in examination of trends in ambient sound level over time.

The recommendation for continuous sampling of the 1 minute window is based on the complete spectrum. If a study were to focus on a particular frequency band, such as 85-105 Hz, the recommendation would be different. Because the 85-105 Hz band is composed of far fewer transient sounds than other frequency bands, the window length and subsampling of the data set have a much smaller effect on the mean sound level estimate. So, an analysis using a continuous data set would give approximately the same
sound level as an analysis using data subsampled 1 minute every 2 hours. Therefore, if
an analysis was focused on only the 85-105 Hz band, the recommended unit of analysis
and subsample rate would be 60 seconds every 2 hours to drastically decrease
computational time while preserving accuracy in ambient sound level estimate.

This study provides a method of calculating the uncertainty in mean sound level
estimates over long time periods based on subsampling. Many different studies have
documented the rise in ambient sound level over time and have found an increase in the
ambient sound level of approximately 2-3 dB per decade (Andrew et al., 2001; Chapman
and Price, 2011; Wenz, 1969). A mean sound level estimate that varies even 1.5 dB re 1
μPa²/Hz would be very significant in such a study as it is 50% of the increase. A uniform
selection of sampling would result in accurate comparative results even if subsampling
the data set was utilized. However, differences in the sampling method between studies
can cause a deviation in results.

A comparative study between a data set collected by Wenz (1969) from 1958-1959
and another data set collected from 2005-2006 by McDonald et al. (2008) was performed.
McDonald et al. (2008) states that spectral averages collected by Wenz (1969) were
based on three 200 s averages per hour, while McDonald et al. (2008) used spectra based
on 200 s averages from a continuous data set. The power spectra compared in this study,
for the lower frequency bands (i.e. below 60 Hz) could carry an uncertainty of about 1-2
dB re 1 μPa²/Hz. However, in both of these studies there was a significant difference in
the analysis that is of note here. Wenz (1969) and McDonald et al. (2008) removed
transient signals that were greater than 3 dB above the background and also removed all
contributions due to shipping. Therefore, these studies are more representative of the
sound floor or background noise than the ambient sound level. Because the differences in
sound level estimates found in this study are driven by the transient component of the
data, a similar analysis on the background sound level would likely return different
results. Therefore, there may be no significant difference between sound level estimates
due to the subsampling artifacts in the studies performed by Wenz (1969) and McDonald
et al. (2008). However, because no universally accepted procedure exists for selecting
window length or sampling method, if the results had been based on the entire ambient
data set including transients, uncertainty in results could be substantial.

This study has shown the degree of uncertainty in ambient sound level estimates
based on different units of analysis. The shift in the sound level estimates between
comparative studies can be substantial if signal processing parameters are not accounted
for. So, in the field of low-frequency underwater ambient sound in the ocean, either
guidelines must be established to specify signal processing parameters for comparative studies by the scientific community, or the degree of variation caused by dissimilar processing parameters must be considered statistically to confirm interpretation of results and observed trends.

Basing an ambient sound measurement on only the low level interminable background sound gives little information on the sources contributing to the ambient sound level. Therefore, the inclusion of all transient signals is essential to the characterization of the ambient sound level. Selection of sampling protocol excluding transients is likely extraneous, because less information about sources is included in the background noise. A complete study of ambient sound in the ocean, which includes transients, entails careful selection of sampling protocols.

4.7 Procedure for selecting optimal unit of analysis

The recommendation of a continuous data set with a 1 min window is isolated to the Indian Ocean hydrophones at Diego Garcia. The ambient sound field depends on location, bathymetry, attenuation, source contribution, and many other factors. The steps used in determining the optimal unit of analysis can be applied to any acoustic data set.

**STEPS IN DETERMINATION OF SUBSAMPLING PARAMETER**

1. *Interval Selection for the Subsample Analysis:*

   The number of groups to test should be between 2 and 5, since the Kruskal-Wallis test cannot test more than 5 groups. The maximum subsample rate should always be continuous which is what the selection of unit of analysis will be based on. The minimum subsample rate must be greater than the length of time of the data set over 2 (> L/2).

   The recommended number of groups to test for the FFT-analysis is 5. The maximum and minimum window sizes depend on the sampling rate and length of the data set. The recommended minimum FFT-size would give a frequency resolution of 1 Hz.
2. *Duration of time series*

The duration of the time series for the analysis should be no smaller than your window length times 2 and can be no larger than the length of the entire data set.

3. *Kruskal-Wallis Alpha*

The alpha value should be 0.05 divided by the number of groups. This is the Bonferroni correction which is used to control a family-wise error rate.

4. *Interpreting Multiple Comparison*

Overlapping confidence intervals indicate no significant difference between groups, and selection of subsampling rate should reflect the group that is not significantly different than the continuous method.

5. *Probability Density Comparison*

Individual tests benefit from a side-by-side probability density comparison.

6. *Degree of Uncertainty or Variation*

Because the Kruskal-Wallis tests the ranks of groups of data, it cannot be used to quantify the variation in mean. Those must be calculated separately. The mean of each of the separate subsampled data sets should be calculated on tests which indicate significant differences.
Table 4.1 Calculation of Variables Parameters for Kruskal-Wallis test. In the following table, $N_g$ is the number of groups, $S_{\text{max}}$ is the maximum subsample, $N$ is the FFT size, $f_s$ is the sampling rate, $S_{\text{min}}$ is the minimum subsample, $L$ is the length of time of the data set, $R_{\text{max}}$ is the maximum test range, $R_{\text{min}}$ is the minimum test range, and $\alpha$ is the alpha value.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of groups</td>
<td>$N_g = 2 - 5$</td>
</tr>
<tr>
<td>Maximum subsample rate</td>
<td>$S_{\text{max}} = \frac{N}{f_s}$</td>
</tr>
<tr>
<td>Minimum subsample rate</td>
<td>$S_{\text{min}} = \frac{L}{2}$</td>
</tr>
<tr>
<td>Maximum test range</td>
<td>$R_{\text{max}} = L$</td>
</tr>
<tr>
<td>Minimum test range</td>
<td>$R_{\text{min}} = \frac{2N}{f_s}$</td>
</tr>
<tr>
<td>Alpha value</td>
<td>$\alpha = \frac{0.05}{N_g}$</td>
</tr>
</tbody>
</table>

The selection for window length should be based on whatever comparison study is being made. The Kruskal-Wallis and post-hoc multiple comparison tests are crucial for the selection of unit of analysis and subsampling method. The Kruskal-Wallis test is the ideal test for acoustic sound level data because it doesn’t require normally distributed data. The post-hoc multiple comparison test is useful for determining the location of variation in the ambient sound level estimate. The probability density function gives sound level estimate distribution which can be useful in interpreting the results of the multiple comparison test. If multiple Kruskal-Wallis tests are employed in the analysis, then the resulting p-values should be displayed on a normalized cumulative distribution function, and significance difference is determined if the peak of the distribution falls below the alpha value.
4.8 Recommendation for Future Work

One area that was not investigated here that would be useful in future analysis is a study on computational time required in an analysis. It was assumed that a larger window size requires less time in computational analysis of ambient sound level. However quantifying the exact time period required in each of these analysis was not part of this thesis. It would be a valuable investigation to determine the length of time required for each type of analysis.

A specific analysis on directivity of individual sources was not included in this type of analysis, but such an exploration could illuminate many of the properties of the ambient sound field in an area, and help to define the transient component of that field.

Examining much higher frequency bands would also be a useful analysis. The transient component of the ambient sound field due to low-frequency propagation is what caused the significant differences between subsampling rates and window lengths. Therefore, high frequency analyses could possibly produce different results.

Comparative ambient sound level studies that used different sampling protocol across analyses, such as McDonald et al. 2008, should also be re-investigated. While the results of these tests are likely still valid because transients were not included, the degree of uncertainty in the estimates should be calculated to ensure accuracy.

This study was based on the mean of the ambient sound level across frequency and time. Because the mean tends to bias the sound level, an examination of total energy by $\Sigma Gxx \cdot df$ as well as the median of the sound level would be beneficial to determine if similar results are observed.
References


Appendix

Matlab Codes

The following are codes used in Matlab to calculate the statistics and figures used in this study. Most codes are written so that they can either be run as a function with given inputs or so they can be run using the command window or the function window. Each algorithm specifies which inputs are required to run.

Table of Contents

1. Bootstrapping analysis

This algorithm provides the Matlab code used in the bootstrapping analysis. It performs a given number of Kruskal-Wallis and post-hoc multiple comparison analyses specified by run, on the specified hydrophone, station, at the site location. It calculates the p-values and the percentage of tests returning significant differences. It saves the table of p-values for each of the bands under a specified location.

2. Kruskal Wallis and PDF

This algorithm performs the Kruskal-Wallis and post-hoc multiple comparison test on a single time period for the specified hydrophone over the 5 subsample rates specified. It plots the multiple comparison as well as probability density functions for each subsample rate.

3. FFT Comparison analysis

This algorithm performs the Kruskal-Wallis and post-hoc multiple comparison on a single time period for a specified hydrophone over the 5 specified window lengths. It also plots a multiple comparison as well as a probability density function.
4. **Decadal Spectrogram**

This code calculates the daily average power spectral density and plots a spectrogram of those daily PSD over an input time period. It simultaneously edits out any PSD's that are unrealistically high due to calibration.

5. **Spectrum Level Autocorrelation**

This algorithm requires a matrix input calculated using the code Decadal Spectrogram. The required input matrix can be saved as GXX and input into the workspace from which this code performs the autocorrelation. The spectral autocorrelation figure title must be adjusted manually, as well as the timeline, specified by days. It plots the PSD from 1-120 Hz and it should be noted there is a significant roll-off above 110 Hz.

6. **Read Raw Data File**

This algorithm was developed by Chad Smith to select data from the CTBTO hydrophone, which is in wfdisc format. It converts the data into absolute time vs. A/D counts. The algorithm was adjusted to be able to continue running through missing portions of data. Another adjustment also accounts for the shift in formatting that occurs toward the middle of 2011.

7. **Calibration and Spectrum Level**

This algorithm was also developed by Chad Smith to apply the calibration curves to the data and convert it into a spectral density for a specified time period and hydrophone.

8. **Subsampled Periodicity Analysis**

This final code is used to determine the daily averages over a specific frequency band, and calculates those daily averages based on subsampled data as well. It plots the daily averages and the autocorrelation figure to show the difference in correlation values with subsampling.

9. **P-value Distribution**

This code plots the normalized cumulative distribution of p-values for the bootstrapping analysis. It requires inputs of each of the p-value matrices under specific names shown below.
10. Window Length Bootstrap

Performs the same bootstrapping analysis specified in the bootstrapping analysis for the window lengths instead of subsample rates.

**Bootstrapping Analysis**

```matlab
%% rerun
% function
rerun(runlength, location, station, fftsize, overlap, multiple, Datalength)

% Run as Matlab Script; Compares through ANOVA and Post-hoc Multiple
% rerun tests the means and variation of the sound level over 5
% different subsampled rates. (i.e. 1 min, 5 min, 60 min) It does this for
% a specified number of times.
% Also, it performs the Kruskal-Wallis test to determine whether the
% shape of the distribution is similar or significantly different
% and lastly, it plots the probability density function over the two
% frequency bands selected.
% To view data segments at 30 seconds a piece use a 7500 point fft
% (7500/250 = 30 seconds). To view data segments at 1 minute a piece use
% a 15000 pt fft (15000/250 = 60 seconds). ETC.
%
% INPUTS(7):
% runlength - Number of runs over which to do bootstrap analysis
% location - The site location of the CTBTO dataset (i.e 'HA08', etc.)
% station - The Hydrophone Station number (i.e. 'H08N1', 'H11S2', etc.)
% fftsize - Size of fft over which it will be run
% overlap - The overlap of the samples for the FFT (i.e. 0)
% multiple - The subsampling rate which are multiples of the initial Data Segment size. Input = 4 (i.e. [2 3 4 5])
% Datalength - The number of data points in the analysis

% OUTPUTS(1):
% Diary text file under Specified location with the P-values and other information from the tests
% Displays percents of results from Kruskal-Wallis

%% Create Date Matrix from which to pull the starting dates
```
clear all;close all;clc;
for s = 731237:735071; % Based on datenum
datematrix(s-731237+1,:) = s;
end

% Remove starting days with no data -42 to run through 100 points
datematrix(datematrix==732040,:)=[];
datematrix(datematrix==733971,:)=[];
datematrix(datematrix==734098,:)=[];
for ss = 732096:732178;
    datematrix(datematrix==ss,:)=[];
end

%% RUN AS MATLAB SCRIPT
runlength = 100;
location = 'HA08';
station = 'H08S2';
fftsize = 50000;
overlap = 0.0;
initial = fftsize/250;
multiple = [3 5 10 15];
Datalength = 1000;

Averagingrates =
    [initial/60,initial*multiple(1)/60,initial*multiple(2)/60,...
     initial*multiple(3)/60,initial*multiple(4)/60];
% Converts the averaging rates to seconds

%% Preallocate Space
gxxaverage10to30 = nan(ceil(Datalength*max(Averagingrates)),5);
gxxaverage40to60 = nan(ceil(Datalength*max(Averagingrates)),5);%Preallocate space

gxxaverage85to105 = nan(ceil(Datalength*max(Averagingrates)),5);
gxxaverageSpectrum = nan(ceil(Datalength*max(Averagingrates)),5);
gxx10to30 = nan(Datalength,5);
gxx40to60 = nan(Datalength,5);
gxx85to105 = nan(Datalength,5);
gxxSpectrum = nan(Datalength,5);
Soundlevel10to30 = nan(Datalength,5);
Soundlevel140to60 = nan(Datalength,5);
Soundlevel185to105 = nan(Datalength,5);
SoundlevelSpectrum = nan(Datalength,5);
prob10to30 = nan(runlength,4);
prob40to60 = nan(runlength,4);
prob85to105 = nan(runlength,4);
probSpectrum = nan(runlength,4);
means = nan(runlength,4);
medians = nan(runlength,4);
addpath read_e1_files -begin

%% Run through Bootstrap
for z = 1:runlength;
date = randsample(datematrix,1);

Start_date = datestr(date);

%% Start Diary
diary(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\'...
    'Bootstrap' station '-' num2str(fftsize) '-fft.txt']);

disp('____________________________________')
disp(['Run Number ' num2str(z) ' -->'])
disp(['Start Date:' Start_date]);
disp(['Station: ' station]);
disp(['Location:' location]);
disp(['Averaging Rates = ' num2str(Averagingrates(1)) ',
    num2str(Averagingrates(2)) ',
    num2str(Averagingrates(3)) ',
    num2str(Averagingrates(4)) ',
    num2str(Averagingrates(5))]);
disp(['Data Length = ' num2str(Datalength)]);
disp(['Overlap = ' num2str(overlap)]);
disp('____________________________________')

Start_day = datenum(Start_date,'dd-mmm-yyyy');
End_day = datenum(Start_date,'dd-mmm-yyyy') +
    (ceil(max(Averagingrates)*Datalength/1440)-1);
Number_of_days = End_day - Start_day + 1;
End_date = datestr(End_day);

for d = Start_day:End_day     % increment through multiple days
    for hour = 0:2:22; % increment through day
        % Site info and data/time
        start_file=datestr2DOY(datestr(d,23)); %get starting directory name
        plot_flag = 0;
        %% Read 2 Hours of Raw Data
        try
            [data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =...
                read_raw_data_v3r0(location,station,start_file,hour,plot_flag);
        catch MException % Checks if there is data missing and skip sectors
            if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
                data(1,1:1800000) = nan;
                continue
            end
        end
        if exist('data','var')==0;  % This helps to account for the missing
          data sections
            if overlap == 0;
                Gxx = nan(N/2+1,(2*N/250)/(1-overlap));
                eval(['Gxx' num2str(hour) ' = Gxx;']);
            continue
            else
            end
        end
    end
end
Gxx = nan(N/2+1,(2*N/250)/(1-overlap)-1);  
eval({'Gxx' num2str(hour) '=Gxx;' });  
continue  
end  
end  
% This portion of the code checks the file to make sure it is the  
% correct  
% size, and assumes any extra rows are superfluous.  
if length(data)>2*60*60 % assume we can chop anything more than 2 hours  
data=data(1:2*250*60*60); % (bad assumption?)  
end  
if length(data)<2*60*60 % assume we can chop anything more than 2 hours  
error('*** Incorrect data length!!! ***')  
end  
%% Apply FFT to Data  
N = fftsize; % parse length  
fs = 250; %Hz Sampling rate  
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station,  
t_ser(1),location,N,overlap);  
Gxx_ave=mean(Gxx,2); % [uPa^2/Hz]  
% Finds the average across the entire window giving a single PSD  
Gxx_ave_dB=10*log10(Gxx_ave);  
% Makes it possible to view the PSD in dB re 1 uPa^2/Hz  
%% Frequency Band Indeces  
% Find the mean at each of the subsample rates  
[~, ind0] = min(abs(freqs-10.0)); %10Hz bin  
[~, ind1] = min(abs(freqs-30.0));  
[~, ind2] = min(abs(freqs-40.0));  
[~, ind3] = min(abs(freqs-60.0));  
[~, ind4] = min(abs(freqs-85.0));  
[~, ind5] = min(abs(freqs-105.0));  
%% Subsampling and Accounting for Overlap  
% Loop Through each of the subsample rates  
for a = 1:5;  
Averagingrate = Averagingrates(a);  
if overlap < 0.50; % This compensates for the overlap by adjusting the  
Gxx.  
    points = ceil((60*2)/((fftsize/fs)/60));  
    Gxxsubsampled = Gxx(:,1:Averagingrate/min(Averagingrates):points);  
    %Gxx subsampled at subsample rate  
else  
    Gxxnew = nan(length(Gxx(:,1)),(fftsize/250)*2);  
    points = ceil(((60*2)/(1-overlap))/((N/fs)/60))/(1/(1-overlap));  
end
for \( w = 1:floor(1/(1-overlap)):length(Gxx(1,:))-1; \)
\[
Gxxnew(:,ceil(w/(1/(1-overlap)))) = mean(Gxx(:,w:w+(floor(1/(1-overlap))-1)),2);
\]
\[
Gxxnew(:,ceil(length(Gxx(1,:))/(1/(1-overlap)))) = \]
\[
Gxx(:,length(Gxx(1,:)));
\]
end
Gxxsubsampled =
Gxxnew(:,1:Averagingrate/min(Averagingrates):points);
% Accounted for overlap by averaging
end
section = (d-Start_day)*length(Gxxsubsampled(1,:))*12 ...
+ length(Gxxsubsampled(1,:))*hour/2 + 1;
fs = 250.0;
dt = 1/fs;
df = 1/(dt*N);

for m = 1:length(Gxxsubsampled(1,:));
Gxxtemp = Gxxsubsampled(:,m);
gxxaverage10to30(section+m-1,a) = mean(Gxxtemp(ind0:ind1));
gxxaverage40to60(section+m-1,a) = mean(Gxxtemp(ind2:ind3));
gxxaverage85to105(section+m-1,a) = mean(Gxxtemp(ind4:ind5));
gxxaverageSpectrum(section+m-1,a) = mean(Gxxtemp);
end
end % end multiple averaging rate
end % end increment through hour
end % end increment through multiple days

%% Averaging the smaller intervals to match data length size over the
same time period
for a = 1:5;
for t = 1:Datalength;

Step = floor(max(Averagingrates)/Averagingrates(a));

gxx10to30(t,a) = nanmean(gxxaverage10to30((Step*(t-1)+1):t*Step,a));
Soundlevel10to30(t,a) = 10.*log10(gxx10to30(t,a));

gxx40to60(t,a) = nanmean(gxxaverage40to60((Step*(t-1)+1):t*Step,a));
Soundlevel40to60(t,a) = 10.*log10(gxx40to60(t,a));

gxx85to105(t,a) = nanmean(gxxaverage85to105((Step*(t-1)+1):t*Step,a));
Soundlevel85to105(t,a) = 10.*log10(gxx85to105(t,a));

gxxSpectrum(t,a) = nanmean(gxxaverageSpectrum((Step*(t-1)+1):t*Step,a));
SoundlevelSpectrum(t,a) = 10.*log10(gxxSpectrum(t,a));

end % end multiple averaging rate
end % end increment through hour
end % end increment through multiple days
end
% This portion takes the different vector sizes and averages them to make
% equal vectors for the analysis. This makes the confidence intervals the
% same size

Soundlevel10to30(~isfinite(Soundlevel10to30))=mean(median(Soundlevel10to30,2));

Soundlevel40to60(~isfinite(Soundlevel40to60))=mean(median(Soundlevel40to60,2));

Soundlevel85to105(~isfinite(Soundlevel85to105))=mean(median(Soundlevel85to105,2));

SoundlevelSpectrum(~isfinite(SoundlevelSpectrum))=mean(median(SoundlevelSpectrum,2));
% Remove infinite values from the matrix

end % multiple averaging rates

%% Determine mean difference of two farthest subsample rates (i.e. 1 and 60)
means(z,1) = abs(nanmean(Soundlevel10to30(:,1))...-
-nanmean(Soundlevel10to30(:,5)));
means(z,2) = abs(nanmean(Soundlevel140to60(:,1))...-
-nanmean(Soundlevel140to60(:,5)));
means(z,3) = abs(nanmean(Soundlevel85to105(:,1))...-
-nanmean(Soundlevel85to105(:,5)));
means(z,4) = abs(nanmean(SoundlevelSpectrum(:,1))...-
-nanmean(SoundlevelSpectrum(:,5)));

medians(z,1) = abs(nanmedian(Soundlevel10to30(:,1))...-
-nanmedian(Soundlevel10to30(:,5)));
medians(z,2) = abs(nanmedian(Soundlevel40to60(:,1))...-
-nanmedian(Soundlevel40to60(:,5)));
medians(z,3) = abs(nanmedian(Soundlevel85to105(:,1))...-
-nanmedian(Soundlevel85to105(:,5)));
medians(z,4) = abs(nanmedian(SoundlevelSpectrum(:,1))...-
-nanmedian(SoundlevelSpectrum(:,5)));

%% Perform ANOVA and post-hoc multiple comparison on points
windows = {num2str(Averagingrates(1)),num2str(Averagingrates(2)),... num2str(Averagingrates(3)),num2str(Averagingrates(4)),... num2str(Averagingrates(5))};
% Specify (name) windows for the analysis

[p1,table1,stat1] = anova1(Soundlevel10to30,windows,'off');
[p2,table2,stat2] = anova1(Soundlevel40to60,windows,'off');
[p3,table3,stat3] = anova1(Soundlevel85to105,windows,'off');
[p4,table4,stat4] = anova1(SoundlevelSpectrum,windows,'off');
% Runs the anova analysis without plotting
%% Kruskal-Wallis Test
[p5,anovatab5,stat5] = kruskalwallis(Soundlevel10to30,windows,'off');
[p6,anovatab6,stat6] = kruskalwallis(Soundlevel40to60,windows,'off');
[p7,anovatab7,stat7] = kruskalwallis(Soundlevel85to105,windows,'off');
[p8,anovatab8,stat8] = kruskalwallis(SoundlevelSpectrum,windows,'off');
% Runs the Kruskal-Wallis analysis without plotting

%% Save analysis details to .txt file
disp('____________________________________')
disp('ANOVA Test P-Value for...')
disp(['    ...10-30 Hz = ' num2str(p1)])
disp(['    ...40-60 Hz = ' num2str(p2)])
disp(['    ...85-105 Hz = ' num2str(p3)])
disp(['    ...1-125 Hz = ' num2str(p4)])
disp('____________________________________')
disp('ANOVA Test F-Statistic for...')
disp(['    ...10-30 Hz = ' num2str(cell2mat(table1(2,5)))]
disp(['    ...40-60 Hz = ' num2str(cell2mat(table2(2,5)))]
disp(['    ...85-105 Hz = ' num2str(cell2mat(table3(2,5)))]
disp(['    ...1-125 Hz = ' num2str(cell2mat(table4(2,5)))]
disp('____________________________________')
disp('Kruskal Wallis Test P-Value for...')
disp(['    ...10-30 Hz = ' num2str(p5)])
disp(['    ...40-60 Hz = ' num2str(p6)])
disp(['    ...85-105 Hz = ' num2str(p7)])
disp(['    ...1-125 Hz = ' num2str(p8)])
disp('Kruskal Wallis Test H-Statistic for...')
disp(['    ...10-30 Hz = ' num2str(cell2mat(anovatab5(2,5)))]
disp(['    ...40-60 Hz = ' num2str(cell2mat(anovatab6(2,5)))]
disp(['    ...85-105 Hz = ' num2str(cell2mat(anovatab7(2,5)))]
disp(['    ...1-125 Hz = ' num2str(cell2mat(anovatab8(2,5)))]
disp('____________________________________')
diary off
% Displays the p values for each day and saves them in a log file

%% Save p-values for each 42 day period

% 1-ANOVA p-value, 2-ANOVA f-statistic, 3-Kruskal-Wallis p-value, 4-H-Statistic
prob10to30(z,1:4)=[p1,cell2mat(table1(2,5)),p5,cell2mat(anovatab5(2,5))];
prob40to60(z,1:4)=[p2,cell2mat(table2(2,5)),p6,cell2mat(anovatab6(2,5))];
prob85to105(z,1:4)=[p3,cell2mat(table3(2,5)),p7,cell2mat(anovatab7(2,5))];
probSpectrum(z,1:4)=[p4,cell2mat(table4(2,5)),p8,cell2mat(anovatab8(2,5))];
clearvars -except prob10to30 prob40to60 prob85to105 probSpectrum
runlength...
    location station fftsize overlap initial multiple Datalength...
    Averagingrates datematrix means medians
clc;
% clears the variables and the command window
end

%% Calculate Statistics
pass10to30 = sum(prob10to30(:,1:2:4)>0.01)/length(prob10to30(:,1));
pass40to60 = sum(prob40to60(:,1:2:4)>0.01)/length(prob40to60(:,1));
pass85to105 = sum(prob85to105(:,1:2:4)>0.01)/length(prob85to105(:,1));
passSpectrum = sum(probSpectrum(:,1:2:4)>0.01)/length(probSpectrum(:,1));

disp('____________________________________')
disp('Percent no significant change ANOVA:')
disp(['           ...10-30 Hz:' num2str(pass10to30(1)*100) ' %'])
disp(['           ...40-60 Hz:' num2str(pass40to60(1)*100) ' %'])
disp(['           ...85-105 Hz:' num2str(pass85to105(1)*100) ' %'])
disp(['           ...1-120 Hz:' num2str(passSpectrum(1)*100) ' %'])
disp('____________________________________')

%% Calculate Statistics
averagemeandifference = nanmean(means);
averagemediandifference = nanmean(medians);

disp('____________________________________')
disp('Average Difference between means for:')
disp(['           ...10-30 Hz:' num2str(averagemeandifference(1)) 'dB'])
disp(['           ...40-60 Hz:' num2str(averagemeandifference(2)) 'dB'])
disp(['           ...85-105 Hz:' num2str(averagemeandifference(3)) 'dB'])
disp(['           ...1-120 Hz:' num2str(averagemeandifference(4)) 'dB'])
disp('____________________________________')

disp('Average Difference between medians for')
disp(['           ...10-30 Hz:' num2str(averagemediandifference(1)) 'dB'])
disp(['           ...40-60 Hz:' num2str(averagemediandifference(2)) 'dB'])
disp(['           ...85-105 Hz:' num2str(averagemediandifference(3)) 'dB'])
disp(['...1-120 Hz: ' num2str(averagemediandifference(4)) 'dB'])
disp('____________________________________')
disp('***DONE***')
% end

Kruskal-Wallis, and PDF Analysis

% clear all;close all;clc;
% function
anova_data(Start_date, location, station, multiple, Datalength, fftsize, overlap)

%% Values for ANOVA test in Excel
% Sample - anova_data('1/1/2005','HA08','H08N1',[5 15 30 60],1000,15000,0.0)
% Run as Matlab Script; Compares through ANOVA and Post-hoc Multiple
% Comparison tests the means and variation of the sound level over 5
% different subsampled rates. (i.e. 1 min, 5 min, 60 min)
% Also, it performs the Kruskal-Wallis test to determine wether the
% shape of the distribution is similar or significantly different
% and lastly, it plots the probability density function over the two
% frequency bands selected.

% To view data segments at 30 seconds a piece use a 7500 point fft
% (7500/250 = 30 seconds). To view data segments at 1 minute a piece use
% a
% 15000 pt fft (15000/250 = 60 seconds). ETC.

% % INPUTS(5):
% Start_date - The starting date form 'mm/dd/yyyy' (i.e.
% 12/25/1986')
% location - The site location of the CTBTO dataset (i.e
% 'HA08',etc.)
% station - The Hydrophone Station number (i.e.
% 'H08N1','H11S2',etc.)
% multiple - The subsampling rate which are multiples of the
% initial
% Datalength - The number of data points in the analysis
% fftsize - Size of fft over which it will be run
% overlap - The percent overlap in the fft
% (USE 0.0 TO AVOID UNNECESSARY AVERAGING)

% OUTPUTS(1):
% Diary text file under Specified location with the P-values and
% other
% information from the tests

% PLOTS(15):
% ANOVA test P-values for all frequency bands
% ANOVA plots for all frequency bands
% Kruskal Wallis test P-values for all frequency bands
% Kruskal Wallis plots for all frequency bands on Ranks
% Post-hoc multiple comparisons for all frequency bands
% Probability Density Functions for all frequency bands over all
% subsampling rates
%
% RUN AS MATLAB SCRIPT (Give Inputs!)
% clear all;close all;clc;
% Start_date = input('Start Date:','s');
% location = input('Location:','s');
% station = input('Station:','s');
% fftsize = input('FFT Size:');
% overlap = input('Overlap:');
% initial = fftsize/250;
% disp(['Segment size = ' num2str(initial/60) ' minutes'])
% multiple = input('Averaging Rates multiples of initial i.e.[2 3 4 5]');
% Datalength = input('Data Length (number of data points):');

% RUN AS MATLAB SCRIPT (SET INPUTS)
clear all;close all;clc;
Start_date = '1/1/2005';
location = 'HA08';
station = 'H08N1';
fftsize = 15000;
overlap = 0;
initial = fftsize/250;
disp(['Segment size = ' num2str(initial/60) ' minutes'])
multiple = [5,10,30,60];
Datalength = 1000;

Averagingrates =
[initial/60,initial*multiple(1)/60,initial*multiple(2)/60,...
 initial*multiple(3)/60,initial*multiple(4)/60];

% Display Parameters
disp('____________________________________')
disp(['Start Date: ' Start_date]);
disp(['Location: ' location]);
disp(['Station: ' station]);
disp(['Initial segment size based on fft size --> fft_size/250 = '...
 num2str(initial/60) ' minutes'])
disp(['Averaging Rates = ' num2str(Averagingrates(1)) ','...
 num2str(Averagingrates(2)) ',' num2str(Averagingrates(3)) ',' ...
 num2str(Averagingrates(4)) ',' num2str(Averagingrates(5)) ' minutes']);
disp(['Data Length = ' num2str(Datalength)]);
disp(['FFT Size = ' num2str(fftsize)]);
disp(['Overlap = ' num2str(overlap)]);
disp(['Segment Size = ' num2str((fftsize/250)/60) ' minutes or ' num2str((fftsize/250)) ' seconds']);
disp('_________________________')
%% Preallocate Space
qxxaverage10to30 = nan(ceil(Datalength*max(Averagingrates)),5);
qxxaverage40to60 = nan(ceil(Datalength*max(Averagingrates)),5);
qxxaverage85to105 = nan(ceil(Datalength*max(Averagingrates)),5);
gxxaverageSpectrum = nan(ceil(Datalength*max(Averagingrates)),5);
gxx10to30 = nan(Datalength,5);
gxx40to60 = nan(Datalength,5);
gxx85to105 = nan(Datalength,5);
gxxSpectrum = nan(Datalength,5);
Soundlevel10to30 = nan(Datalength,5);
Soundlevel40to60 = nan(Datalength,5);
Soundlevel85to105 = nan(Datalength,5);
SoundlevelSpectrum = nan(Datalength,5);
addpath read_e1_files -begin

Start_day = datenum(Start_date,'dd/mm/yyyy');
End_day = datenum(Start_date,'dd/mm/yyyy') + (ceil(max(Averagingrates)*Datalength/1440) - 1);
Number_of_days = End_day - Start_day + 1;
End_date = datestr(End_day);
for d = Start_day:End_day   % increment through multiple days
    disp(['*** PROCESSING DAY ' datestr(d,23) '... ***'])
    for hour = 0:2:22; % increment through day
        % Site info and data/time
        start_file=datestr2DOY(datestr(d,23)); %get starting directory name
        plot_flag = 0;
        %% Read 2 Hours of Raw Data
        try
            [data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =...
            read_raw_data_v3r0(location,station,start_file,hour,plot_flag);
        catch MException % Checks if there is data missing and skip sectors
            if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
                data(1,1:1800000) = nan;
                continue
            end
        end
        if exist('data','var')==0;  % This helps to account for the missing
            data sections
            if overlap == 0;
                Gxx = nan(N/2+1,(2*N/250)/(1-overlap));
                eval([''Gxx' num2str(hour) '='Gxx; ' ]);
            continue
            else
                Gxx = nan(N/2+1,(2*N/250)/(1-overlap)-1);
                eval([''Gxx' num2str(hour) '='Gxx; ' ]);
            continue
            end
        end
end
if length(data)>2*60*60 % assume we can chop anything more than 2 hours
data=data(1:2*250*60*60); % (bad assumption?)
end
if length(data)<2*60*60 % assume we can chop anything more than 2 hours
    error('*** Incorrect data length!!! ***')
end

% This portion of the code checks the file to make sure it is the correct
% size, and assumes any extra rows are superfluous.

%% Apply FFT to Data
N = fftsize; % parse length
fs = 250; %Hz Sampling rate
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station, t_ser(l), location, N, overlap);

Gxx_ave=mean(Gxx,2); % [uPa^2/Hz]
Gxx_ave_dB=10*log10(Gxx_ave);

%% Frequency Band Indeces
% Find the mean at each of the subsample rates
[~, ind0] = min(abs(freqs-10.0)); %10Hz bin
[~, ind1] = min(abs(freqs-30.0));
[~, ind2] = min(abs(freqs-40.0));
[~, ind3] = min(abs(freqs-60.0));
[~, ind4] = min(abs(freqs-85.0));
[~, ind5] = min(abs(freqs-105.0));

%% Subsampling and Accounting for Overlap

%% Loop Through each of the subsample rates
for a = 1:5;
    Averagingrate = Averagingrates(a);
    if overlap < 0.50; % This compensates for the overlap by adjusting the Gxx.
        points = ceil((60*2)/((fftsize/fs)/60)); % Number of steps in Gxx
        % points are the number of points in each 2 hours based on the size
        % of the fft
        % and the percent overlap. (with 0 percent overlap and an fft size
        % of
        % 7500 we would have sections of 30 seconds, so there would be 240
        % points
        % less than 50 percent overlap would give an uneven index, so the
        % points must be treated as having no overlap.
        Gxxsubsampled = Gxx(:,1:Averagingrate/min(Averagingrates):points);
        %Gxx subsampled at subsample rate
    else
        Gxxnew = nan(length(Gxx(:,1)),(fftsize/250)*2);
        points = ceil(((60*2)/(1-overlap))/((N/fs)/60)/(1/(1-overlap)));
        %
for \( w = 1: \text{floor}(1/(1-\text{overlap})): \text{length}(Gxx(:,1))-1; \)

\[
Gxxnew(:, \text{ceil}(w/(1/(1-\text{overlap})))) = \text{mean}(Gxx(:,w:w+(\text{floor}(1/(1-\text{overlap}))-1)),2);
\]

\[
Gxxnew(:, \text{ceil}(\text{length}(Gxx(:,1))/(1/(1-\text{overlap})))) = Gxx(:, \text{length}(Gxx(:,1)));\]

end

Gxxsubsampled =
Gxxnew(:,1:\text{Averagingrate}/\text{min(Averagingrates)}:\text{points});
% Accounted for overlap by averaging
end

% Overlap used in the fft changes how you subsample (with 50 percent overlap,
% and an fft size of 15000, each segment would be 30 seconds long, so a
% 1 minutes subsample rate would select only every other 30s because of
% the overlap.) A 50 percent overlap would increase the number of
% samples
to double it. So this averages to make it equal to the actual sample size

section = (\text{d}-\text{Start}_\text{day})*\text{length}(Gxxsubsampled(:,1)))*12 ...
+ \text{length}(Gxxsubsampled(:,1))*\text{hour}/2 + 1;

fs = 250.0;
dt = 1/fs;
df = 1/(dt*N);
for m = 1:1:length(Gxxsubsampled(:,1));
    Gxxtemp = Gxxsubsampled(:,m);
gxxaverage10to30(section+m-1,a) = \text{mean}(Gxxtemp(ind0:ind1));
gxxaverage40to60(section+m-1,a) = \text{mean}(Gxxtemp(ind2:ind3));
gxxaverage85to105(section+m-1,a) = \text{mean}(Gxxtemp(ind4:ind5));
gxxaverageSpectrum(section+m-1,a) = \text{mean}(Gxxtemp);
end
end % end multiple averaging rate
end % end increment through hour
end % end increment through multiple days

% Averaging the smaller intervals to match data length size over the
same time period
for a = 1:5;
for t = 1:\text{Data}\text{length};
    \text{Step} = \text{floor}(\text{max(Averagingrates)}/\text{Averagingrates}(a));
    gxx10to30(t,a) = \text{nanmean}(gxxaverage10to30((\text{Step}*(t-1)+1):t*\text{Step},a));
    \text{Soundlevel10to30}(t,a) = 10.*\log10(\text{gxx10to30}(t,a));
    gxx40to60(t,a) = \text{nanmean}(gxxaverage40to60((\text{Step}*(t-1)+1):t*\text{Step},a));
    \text{Soundlevel40to60}(t,a) = 10.*\log10(\text{gxx40to60}(t,a));
end
gxx85to105(t,a) = nanmean(gxxaverage85to105((Step*(t-1)+1):t*Step,a));
Soundlevel85to105(t,a) = 10.*log10(gxx85to105(t,a));

gxxSpectrum(t,a) = nanmean(gxxaverageSpectrum((Step*(t-1)+1):t*Step,a));
SoundlevelSpectrum(t,a) = 10.*log10(gxxSpectrum(t,a));

end
% This portion takes the different vector sizes and averages them to make
% equal vectors for the analysis. This makes the confidence intervals the
% same size

Soundlevel10to30(~isfinite(Soundlevel10to30))=mean(median(Soundlevel10to30,2));
Soundlevel40to60(~isfinite(Soundlevel40to60))=mean(median(Soundlevel40to60,2));
Soundlevel85to105(~isfinite(Soundlevel85to105))=mean(median(Soundlevel85to105,2));
SoundlevelSpectrum(~isfinite(SoundlevelSpectrum))=mean(median(SoundlevelSpectrum,2));
% Remove infinite values from the matrix

end % multiple averaging rates

%% Perform ANOVA and post-hoc multiple comparison on points
windows = {num2str(Averagingrates(1)),num2str(Averagingrates(2)),...
          num2str(Averagingrates(3)),num2str(Averagingrates(4)),...
          num2str(Averagingrates(5))};
% Specify (name) windows for the analysis

[p1,table1,stat1] = anova1(Soundlevel10to30,windows,'off');
[p2,table2,stat2] = anova1(Soundlevel40to60,windows,'off');
[p3,table3,stat3] = anova1(Soundlevel85to105,windows,'off');
[p4,table4,stat4] = anova1(SoundlevelSpectrum,windows,'off');
% Run the analysis of variation without plotting

%% Display probability density Estimates of each of the selected Units
[f1,xi1,u1] = ksdensity(Soundlevel10to30(:,1));
[f2,xi2,u2] = ksdensity(Soundlevel10to30(:,2));
[f3,xi3,u3] = ksdensity(Soundlevel10to30(:,3));
[f4,xi4,u4] = ksdensity(Soundlevel10to30(:,4));
[f5,xi5,u5] = ksdensity(Soundlevel10to30(:,5));
[f401,xi401,u401] = ksdensity(Soundlevel40to60(:,1));
[f402,xi402,u402] = ksdensity(Soundlevel40to60(:,2));
\[ \begin{align*} [f403, xi403, u403] &= \text{ksdensity}(\text{Soundlevel140to60}(:,3)); \\
[f404, xi404, u404] &= \text{ksdensity}(\text{Soundlevel140to60}(:,4)); \\
[f405, xi405, u405] &= \text{ksdensity}(\text{Soundlevel140to60}(:,5)); \\
[f11, xi11, u11] &= \text{ksdensity}(\text{Soundlevel85to105}(:,1)); \\
[f22, xi22, u22] &= \text{ksdensity}(\text{Soundlevel85to105}(:,2)); \\
[f33, xi33, u33] &= \text{ksdensity}(\text{Soundlevel85to105}(:,3)); \\
[f44, xi44, u44] &= \text{ksdensity}(\text{Soundlevel85to105}(:,4)); \\
[f55, xi55, u55] &= \text{ksdensity}(\text{Soundlevel85to105}(:,5)); \\
[f111, xi111, u111] &= \text{ksdensity}(\text{SoundlevelSpectrum}(:,1)); \\
[f222, xi222, u222] &= \text{ksdensity}(\text{SoundlevelSpectrum}(:,2)); \\
[f333, xi333, u333] &= \text{ksdensity}(\text{SoundlevelSpectrum}(:,3)); \\
[f444, xi444, u444] &= \text{ksdensity}(\text{SoundlevelSpectrum}(:,4)); \\
[f555, xi555, u555] &= \text{ksdensity}(\text{SoundlevelSpectrum}(:,5)); \\
\end{align*} \]

\text{type} = 'Subsample';

\text{pdf1} = \text{figure}(1);
\text{plot}(xi1,f1,'-',xi2,f2,'-',xi3,f3,'-',xi4,f4,'-',xi5,f5,'-',
', 'Linewidth',1.5, ... 
'MarkerEdgeColor','k','MarkerFaceColor',[.49 1 .63],'MarkerSize',7);
\text{title('Probability densities:10-30 Hz Band');}
\text{xlabel('Sound Level (dB re \text{1\u00b5Pa}^2/Hz');}
\text{ylabel('Density');}
\text{plotdetails}
\text{legend([num2str(Averagingrates(1)) ' min ' type],[num2str(Averagingrates(2)) ' min ' type],...}
\text{[num2str(Averagingrates(3)) ' min ' type],[num2str(Averagingrates(4)) ' min ' type],...}
\text{[num2str(Averagingrates(5)) ' min ' type]);}

\text{pdf2} = \text{figure}(2);
\text{plot}(xi401,f401,'-',xi402,f402,'-',xi403,f403,'-',xi404,f404,'-',
',xi405,f405,'-', 'Linewidth',1.5, ... 
'MarkerEdgeColor','k','MarkerFaceColor',[.49 1 .63],'MarkerSize',7);
\text{title('Probability densities:40-60 Hz Band');}
\text{xlabel('Sound Level (dB re \text{1\u00b5Pa}^2/Hz');}
\text{ylabel('Density');}
\text{plotdetails}
\text{legend([num2str(Averagingrates(1)) ' min ' type],[num2str(Averagingrates(2)) ' min ' type],...}
\text{[num2str(Averagingrates(3)) ' min ' type],[num2str(Averagingrates(4)) ' min ' type],...}
\text{[num2str(Averagingrates(5)) ' min ' type]);}

\text{pdf3} = \text{figure}(3);
\text{plot}(xi11,f11,'-',xi12,f22,'-',xi13,f33,'-',xi14,f44,'-',xi15,f55,'-',
', 'Linewidth',1.5, ... 
'MarkerEdgeColor','k','MarkerFaceColor',[1 .95 0],'MarkerSize',7);
\text{title('Probability densities:85-105 Hz Band');}
\text{xlabel('Sound Level (dB re \text{1\u00b5Pa}^2/Hz');}
\text{ylabel('Density');}
\text{plotdetails}
legend([num2str(Averagingrates(1)) ' min ' type], [num2str(Averagingrates(2)) ' min ' type],... [num2str(Averagingrates(3)) ' min ' type], [num2str(Averagingrates(4)) ' min ' type],... [num2str(Averagingrates(5)) ' min ' type]);

pdf4 = figure(4);
plot(xi111,f111,'-',xi222,f222,'-',xi333,f333,'-',xi444,f444,'-',xi555,f555,'-', 'Linewidth',1.5,'MarkerEdgeColor','k','MarkerFaceColor',[.5 0 1],'MarkerSize',7);
title('Probability densities:Full Spectrum');
xlabel('Sound Level (dB re 1\mu Pa^2/Hz)');
ylabel('Density');

%% Kruskal-Wallis Test
MC1 = figure(5);
[p5,anovatab5,stat5] = kruskalwallis(Soundlevel10to30,windows,'off');
multcompare(stat5,'alpha',0.01,'ctype','bonferroni');
title('Multiple Comparison Test 10-30 Hz');
xlabel('Mean Rank');
ylabel('Subsample rate (Minutes)');
plotdetails

MC2 = figure(6);
[p6,anovatab6,stat6] = kruskalwallis(Soundlevel40to60,windows,'off');
multcompare(stat6,'alpha',0.01,'ctype','bonferroni');
title('Multiple Comparison Test 40-60 Hz');
xlabel('Mean Rank');
ylabel('Subsample rate (Minutes)');
plotdetails

MC3 = figure(7);
[p7,anovatab7,stat7] = kruskalwallis(Soundlevel85to105,windows,'off');
multcompare(stat7,'alpha',0.01,'ctype','bonferroni');
title('Multiple Comparison Test 85-105 Hz');
xlabel('Mean Rank');
ylabel('Subsample rate (Minutes)');
plotdetails

MC4 = figure(8);
[p8,anovatab8,stat8] = kruskalwallis(SoundlevelSpectrum,windows,'off');
multcompare(stat8,'alpha',0.01,'ctype','bonferroni');
title('Multiple Comparison Test Full Spectrum');
xlabel('Mean Rank');
ylabel('Subsample rate (Minutes)');
plotdetails
%% Kruskal-Wallis Test on Different sized matrices
[p9,anovatab9,stat9] = kruskalwallis(10.*log10(gxxaverage10to30),windows,'off');
[p10,anovatab10,stat10] = kruskalwallis(10.*log10(gxxaverage40to60),windows,'off');
[p11,anovatab11,stat11] = kruskalwallis(10.*log10(gxxaverage85to105),windows,'off');
[p12,anovatab12,stat12] = kruskalwallis(10.*log10(gxxaverageSpectrum),windows,'off');

%% Save analysis details to .txt file
diary(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\...
   datestr(Start_day,1) '-' station '-' num2str(fftsize) '-fft.txt']);
disp('____________________________________')
disp(['Start Date:' Start_date]);
disp(['Location:' location]);
disp(['Station:' station]);
disp(['Averaging Rates = ' num2str(Averagingrates(1)) ',
   num2str(Averagingrates(2)) ',
   num2str(Averagingrates(3)) ',
   num2str(Averagingrates(4)) ',
   num2str(Averagingrates(5))]);
disp(['Data Length = ' num2str(Datalength)]);
disp(['FFT Size = ' num2str(fftsize)]);
disp(['Overlap = ' num2str(overlap)]);
disp(['Segment Size = ' num2str((fftsize/250)/60) ' minutes or ' ...
   num2str(fftsize/250) ' seconds']);
disp('____________________________________')
disp('ANOVA Test P-Value for...')
disp(['...10-30 Hz = ' num2str(p1)])
disp(['...40-60 Hz = ' num2str(p2)])
disp(['...85-105 Hz = ' num2str(p3)])
disp(['...1-125 Hz = ' num2str(p4)])
disp('ANOVA Test F-Statistic for...')
disp(['...10-30 Hz = ' num2str(cell2mat(table1(2,5)))]
disp(['...40-60 Hz = ' num2str(cell2mat(table2(2,5)))]
disp(['...85-105 Hz = ' num2str(cell2mat(table3(2,5)))]
disp(['...1-125 Hz = ' num2str(cell2mat(table4(2,5)))]
disp('Kruskal Wallis Test P-Value for...')
disp(['...10-30 Hz = ' num2str(p5)])
disp(['...40-60 Hz = ' num2str(p6)])
disp(['...85-105 Hz = ' num2str(p7)])
disp(['...1-125 Hz = ' num2str(p8)])
disp('Kruskal Wallis Test H-Statistic for...')
disp(['...10-30 Hz = ' num2str(cell2mat(anovatab5(2,5)))]
disp(['...40-60 Hz = ' num2str(cell2mat(anovatab6(2,5)))]
disp(['...85-105 Hz = ' num2str(cell2mat(anovatab7(2,5)))]
num2str(cell2mat(anovatab7(2,5))))
disp('...1-125 Hz = ' num2str(cell2mat(anovatab8(2,5))))

disp('____________________________________')
disp('***DONE***')
diary off

%% Save figures
mkdir(['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station],datestr(Start_day,'dd.mm.yyyy'))
saveas(MC1,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'MC 10-30 Hz.fig'])
saveas(MC2,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'MC 40-60 Hz.fig'])
saveas(MC3,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'MC 85-105 Hz.fig'])
saveas(MC4,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'MC 1-110 Hz.fig'])
saveas(pdf1,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'PDF 10-30 Hz.fig'])
saveas(pdf2,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'PDF 40-60 Hz.fig'])
saveas(pdf3,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'PDF 85-105 Hz.fig'])
saveas(pdf4,['C:\Users\rsh935\Documents\Research\Matlab Figures\' 
station ' \' datestr(Start_day,'dd.mm.yyyy') ' \'FFT' num2str(fftsize) ' \'PDF 1-110 Hz.fig'])

% end

% FFT Comparison Analysis

% FFT Comparison

% function
fftcomparison(Start_date,End_date,location,station,fftsize,overlap,band1,band2,band3,band4)
% Sample - fftcomparison('6/1/2008','8/1/2008','HA08','H08S2',[2500 3750 7500 15000 50000],0.0,[10 30],[40 60],[85 105],[1 110])
% Run as Matlab Script; Compares through ANOVA and Post-hoc Multiple
% Comparison tests the means and variation of the sound level over 5 different fft sizes. (i.e. N = 15000,7500 etc.)
% Also, it performs the Kruskal-Wallis test to determine wether the shape of the distribution is similar or significantly different
and lastly, it plots the probability density function over the
frequency bands selected.

To view data segments at 30 seconds a piece use a 7500 point fft
(7500/250 = 30 seconds). To view data segments at 1 minute a piece us
a 15000 pt fft (15000/250 = 60 seconds). ETC.

INPUTS(5):
Start_date - The starting date form 'mm/dd/yyyy' (i.e.
'12/25/1986')
End_date - The End data of the analysis
location - The site location of the CTBTO dataset (i.e
'HA08', etc.)
station - The Hydrophone Station number (i.e.
'H08N1', 'H11S2', etc.)
fftsizes - Size of fft over which it will be run
overlap - The percent overlap in the fft
(USE 0.0 TO AVOID UNNECESSARY AVERAGING)

OUTPUTS(1):
Diary text file under Specified location with the P-values and
other information from the tests

PLOTS(15):
ANOVA test P-values for all frequency bands
ANOVA plots for all frequency bands
Kruskal Wallis test P-values for all frequency bands
Kruskal Wallis plots for all frequency bands on Ranks
Post-hoc multiple comparisons for all frequency bands
Probability Density Functions for all frequency bands over all
subsampling rates

RUN AS MATLAB SCRIPT
clear all; close all; clc;
Start_date = '10/1/2008';
End_date = '10/8/2008';
location = 'HA08';
station = 'H08S2';
fftsizes = [2500 3750 7500 15000 50000];
overlap = 0;
band1 = [10,30];
band2 = [40,60];
band3 = [85,105];
band4 = [1,110];

%% Display Parameters
disp('_______________________________')
disp(['Start Date:' Start_date]);
disp(['Location:' location]);
disp(['Station:' station]);
disp(['FFT Sizes = ' num2str(fftsizes)]);
disp(['Overlap = ' num2str(overlap)]);
disp('Segment Sizes: ');
disp(['... num2str((fftsizes/250)) ' seconds']);
disp('Frequency Bands: ');
disp(['... num2str(band1(1)) '-' num2str(band1(2)) ' Hz given by variable number 1']);
disp(['... num2str(band2(1)) '-' num2str(band2(2)) ' Hz given by variable number 2']);
disp(['... num2str(band3(1)) '-' num2str(band3(2)) ' Hz given by variable number 3']);
disp(['... num2str(band4(1)) '-' num2str(band4(2)) ' Hz given by variable number 4']);
disp('____________________________________
addpath read_e1_files -begin
Start_day = datenum(Start_date,'mm/dd/yyyy');
End_day = datenum(End_date,'mm/dd/yyyy');
Number_of_days = End_day - Start_day + 1;

%% Preallocate Space
SPLband1 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband2 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband3 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband4 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);

%% Begin loop for fft_length
for a = 1:5;
    for d = Start_day:End_day % increment through multiple days
        disp(['*** PROCESSING DAY ' datestr(d,23) ' with N = ' ...
              num2str(fftsizes(a)) '... ***'])
        for hour = 0:2:22; % increment through day
            % Site info and data/time
            start_file=datestr2DOY(datestr(d,23)); %get starting directory name
            plot_flag = 0;
            try
                [data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =... read_raw_data_v3r0(location,station,start_file,hour,plot_flag);
            catch MException
                if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
                    data(1,1:1800000) = nan;
            end
        end
    end
end

```
continue
end

if exist('data','var')==0; % This helps to account for the missing data sections
    if overlap == 0;
        Gxx = nan(N/2+1,(2*N/250)/(1-overlap));
        eval(['Gxx num2str(hour) '=Gxx; ' ]);
        continue
    else
        Gxx = nan(N/2+1,(2*N/250)/(1-overlap)-1);
        eval(['Gxx num2str(hour) '=Gxx; ' ]);
        continue
    end
end

if length(data)>2*60*60 % assume we can chop anything more than 2 hours
    data=data(1:2*250*60*60); % (bad assumption?)
end
if length(data)<2*60*60 % assume we can chop anything more than 2 hours
    error('*** Incorrect data length!!! ***')
end

%% Apply FFT to Data
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station, t_ser(1), location, N, overlap);

Gxx_ave=mean(Gxx,2); % [uPa^2/Hz]
Gxx_ave_dB=10*log10(Gxx_ave);

%% Frequency Band Indeces (Bins)
% Find the mean at each of the subsample rates
[~, ind0] = min(abs(freqs-band1(1)));
[~, ind1] = min(abs(freqs-band1(2)));
[~, ind2] = min(abs(freqs-band2(1)));
[~, ind3] = min(abs(freqs-band2(2)));
[~, ind4] = min(abs(freqs-band3(1)));
[~, ind5] = min(abs(freqs-band3(2)));
[~, ind6] = min(abs(freqs-band4(1)));
[~, ind7] = min(abs(freqs-band4(2)));

%% Subsampling and Accounting for Overlap
section = (d-Start_day)*length(Gxx(1,:))*12 ...
    + length(Gxx(1,:))*hour/2 + 1;

for m = 1:1:length(Gxx(1,:));
    Gxxtmp = Gxx(:,m);
    SPLband1(section+m-1,a) = 10.*log10(mean(Gxxtmp(ind0:ind1)));
    SPLband2(section+m-1,a) = 10.*log10(mean(Gxxtmp(ind2:ind3)));
SPLband3(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind4:ind5)));  
SPLband4(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind6:ind7)));  
end  
end % end increment through hour  
end % end increment through multiple days  
end % end multiple averaging rate  

%% Get rid of zeros and invalid values  
SPLband1(isinf(SPLband1)) = nan;  
SPLband2(isinf(SPLband2)) = nan;  
SPLband3(isinf(SPLband3)) = nan;  
SPLband4(isinf(SPLband4)) = nan;  

%% Perform ANOVA and post-hoc multiple comparison on points  
windows = {num2str(fftsizes(1)/250),num2str(fftsizes(2)/250),...  
          num2str(fftsizes(3)/250),num2str(fftsizes(4)/250),...  
          num2str(fftsizes(5)/250)};  

[p1,~,stat] = anova1(SPLband1,windows,'off');  
[p2,~,stat2] = anova1(SPLband2,windows,'off');  
[p3,~,stat3] = anova1(SPLband3,windows,'off');  
[p4,~,stat4] = anova1(SPLband4,windows,'off');  

%% Display probability density Estimates of each of the selected Units  
[f1,x1,u1] = ksdensity(SPLband1(:,1));  
[f2,x2,u2] = ksdensity(SPLband1(:,2));  
[f3,x3,u3] = ksdensity(SPLband1(:,3));  
[f4,x4,u4] = ksdensity(SPLband1(:,4));  
[f5,x5,u5] = ksdensity(SPLband1(:,5));  
[f11,x11,u11] = ksdensity(SPLband2(:,1));  
[f22,x22,u22] = ksdensity(SPLband2(:,2));  
[f33,x33,u33] = ksdensity(SPLband2(:,3));  
[f44,x44,u44] = ksdensity(SPLband2(:,4));  
[f55,x55,u55] = ksdensity(SPLband2(:,5));  
[f111,x111,u111] = ksdensity(SPLband3(:,1));  
[f222,x222,u222] = ksdensity(SPLband3(:,2));  
[f333,x333,u333] = ksdensity(SPLband3(:,3));  
[f444,x444,u444] = ksdensity(SPLband3(:,4));  
[f555,x555,u555] = ksdensity(SPLband3(:,5));  
[f1111,x1111,u1111] = ksdensity(SPLband4(:,1));  
[f2222,x2222,u2222] = ksdensity(SPLband4(:,2));  
[f3333,x3333,u3333] = ksdensity(SPLband4(:,3));  
[f4444,x4444,u4444] = ksdensity(SPLband4(:,4));  
[f5555,x5555,u5555] = ksdensity(SPLband4(:,5));  

type = 'Subsample';  
PDF1 = figure(1);  
plot(x1,f1,'-',x2,f2,'-',x3,f3,'-',x4,f4,'-',x5,f5,'-  
', 'Linewidth', 1.5,...
legend(['Window = ' num2str(fftsizes(1)/250) 's'],...['Window = ' num2str(fftsizes(2)/250) 's'],...['Window = ' num2str(fftsizes(3)/250) 's'],...['Window = ' num2str(fftsizes(4)/250) 's'],...['Window = ' num2str(fftsizes(5)/250) 's']);

%% Kruskal-Wallis Test
MC1 = figure(5); [p5,anovatab5,stat5] = kruskalwallis(SPLband1,windows,'off'); multcompare(stat); title(['Multiple Comparison Test ' num2str(band1(1)) '-' num2str(band1(2)) 'Hz']); xlabel('Mean Sound Level (dB re 1\mu Pa^2/Hz)'); ylabel('Window length (s)') plotdetails

MC2 = figure(6); [p6,anovatab6,stat6] = kruskalwallis(SPLband2,windows,'off'); multcompare(stat2); title(['Multiple Comparison Test ' num2str(band2(1)) '-' num2str(band2(2)) 'Hz']); xlabel('Mean Sound Level (dB re 1\mu Pa^2/Hz)'); ylabel('Window length (s)') plotdetails

MC3 = figure(7); [p7,anovatab7,stat7] = kruskalwallis(SPLband3,windows,'off'); multcompare(stat3); title(['Multiple Comparison Test ' num2str(band3(1)) '-' num2str(band3(2)) 'Hz']); xlabel('Mean Sound Level (dB re 1\mu Pa^2/Hz)'); ylabel('Window length (s)') plotdetails

MC4 = figure(8); [p8,anovatab8,stat8] = kruskalwallis(SPLband4,windows,'off'); multcompare(stat4); title(['Multiple Comparison Test ' num2str(band4(1)) '-' num2str(band4(2)) 'Hz']); xlabel('Mean Sound Level (dB re 1\mu Pa^2/Hz)'); ylabel('Window length (s)') plotdetails

%% Save analysis details to .txt file
warning off MATLAB:MKDIR:DirectoryExists
mkdir(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\' station],datestr(Start_day,'dd.mm.yyyy'))
diary(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\' station ' \' datestr(Start_day,'dd.mm.yyyy') \' fft-analysis.txt']);
disp('_____________________________________________')
disp(['Start Date:' Start_date]);
disp(['End Date:' End_date]);
disp(['Location:' location]);
disp(['Station:' station]);
disp(['fft sizes = ' num2str(fftsizes)]);
disp(['Overlap = ' num2str(overlap)]);
Decadal Spectrogram

% function spectrogramyear(Start_date,End_date,location,station)
% example - spectrogramyear('4/26/2007','12/31/2011','HA08','H08N1')
% The following code creates a spectrogram of any number of days taking
% power spectral density of the ambient noise each day. If a more
detailed
% analysis is required, the code must be adjusted to run on hours
instead
% of days.
%
% INPUTS:
% Start_date - The starting date form 'mm/dd/yyyy' (i.e.
'12/25/1986')
% End_date - The ending date form 'mm/dd/yyyy' (i.e.
'12/26/1987')
% location - The site location of the CTBTO dataset (i.e
'HA08',etc.)
% station - The Hydrophone Station number (i.e.
'H08N1','H11S2',etc.)
%
% PLOTS:
% A plot of the spectrogram for between Start_date and End_date

% Run as Matlab Script
clear all;close all;clc;
Start_date = input('Start Date:','s');
End_date = input('End Date:','s');
location = input('Location:','s');
station = input('Station:','s');

disp('____________________________________')
disp(['Start Date:' Start_date]);
disp(['End Date:' End_date]);
disp(['Location:' location]);
disp(['Station:' station]);
disp('____________________________________')

Start_day = datenum(Start_date, 'mm/dd/yyyy');
End_day = datenum(End_date, 'mm/dd/yyyy');

Number_of_days = datenum(End_date, 'mm/dd/yyyy') -
datenum(Start_date, 'mm/dd/yyyy') + 1;
dB = nan(7501, Number_of_days);

%% Set loops
for n = Start_day:End_day; % increment through multiple days
    disp(['*** PROCESSING DAY ' datestr(n, 23) '... ***'])
clearvars -except Number_of_days Start_day End_day Start_date
end

%% site info and data/time
start_file = datestr2DOY(datestr(n, 23)); % get starting directory name
plot_flag = 0;

%% Check if folder exists
Existence_check = exist(['F:\' location '\' location '. ' start_file]);
if Existence_check == 7;
end

%% Loop Through the Day
for hour=0:2:22 % increment through day
    %% LOAD 2 HOURS
    N = 15000; % parse length
    overlap = 0.50;

    try
        data_t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =...
            read_raw_data_v3r0(location, station, start_file, hour, plot_flag);
catch MException
    if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
        data(1, 1:1800000) = nan;
        continue
    end
end

if exist('data', 'var')==0; % This helps to account for the missing
    data_sections
    if overlap == 0;
        Gxx = nan(N/2+1, (2*N/250)/(1-overlap));
        eval(['Gxx' num2str(hour) ' = Gxx;']);
        continue
    end
end
else
    Gxx = nan(N/2+1,(2*N/2)/(1-overlap)-1);
    eval(['Gxx' num2str(hour) '=Gxx; ']);
    continue
end

if length(data)>2*60*60 % assume we can chop anything more than 2 hours
    data=data(1:2*250*60*60); % (bad assumption?)
end
if length(data)<2*60*60 % assume we can chop anything more than 2 hours
    error('*** Incorrect data length!!! ***')
end

%% SPECTRUM
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station,
t_ser(1),location,N,overlap);

% % Unit of Analysis Test
% Subsample_unit = 5;     % Unit of analysis for subsampling in minutes
% Gxx = Gxx(:,1:Subsample_unit*2:239);

% Find the mean of each day
Gxx_ave=mean(Gxx,2); % [uPa^2/Hz]
Gxx_ave_db=10*log10(Gxx_ave);

% Renaming Gxx and Gxx_ave per hour
eval(['Gxx' num2str(hour) '=Gxx; ']);
eval(['Gxx_ave' num2str(hour) '=Gxx_ave; ']);
end % End increment through hours
else
    disp('FOLDER DOES NOT EXIST!')
    continue
end

%% Combine PSD files
% Daily Averages
Gxx_tot = nanmean({Gxx0,Gxx2,Gxx4,Gxx6,Gxx8,Gxx10,Gxx12,Gxx14,...
    Gxx16,Gxx18,Gxx20,Gxx22},2);

section = (n-Start_day)+1;
dB(:,section) = 10*log10(Gxx_tot);

% % 4 Averages per day
% Gxx_tot = [mean({Gxx0,Gxx2,Gxx4},2),mean({Gxx6,Gxx8,Gxx10},2),...
%     mean({Gxx12,Gxx14,Gxx16},2),mean({Gxx18,Gxx20,Gxx22},2)];
% section = (n-Start_day)*4+1;
% GXX(:,section:section+3) = 10.*log10(Gxx_tot);
end

%% Other Essentials

dB = dB(301:6901,:);  % Don't include the invalid values
time = Start_day:End_day;

%% Spectrogram
figure(1)
imagesc(time,freqs,dB)  % spectrogram plot
datetickzoom('x',2,'keepticks')
plotdetails
xlim([Start_day End_day])
colorbar
colormap('jet')
axis xy
title([{'Spectrogram for station ' station ' from ' datestr(Start_day,22)... ' to ' datestr(End_day,22)]},'fontsize',20,'fontweight','b','interpreter','none ');
xlabel({'Date (colorbar in dB re 1\mu Pa^2/Hz)'},'fontsize',20,'fontweight','b');
ylabel({'Frequency [Hz]'},'fontsize',20,'fontweight','b');

%% Continue Editing Spectrogram
proceed = input('Continue to EDIT Spectrogram? (i.e. y or n):','s');
if strcmp(proceed,'y') == 1;

dB_new = nan(length(dB(:,1))+1,length(dB(1,:)));
dB_new(2:length(dB(:,1))+1,:) = dB(1:length(dB(:,1)),:);
dB_new(1,:) = time;
for n = 1:length(dB_new(1,:));
    if mean(dB_new(2:end,n)) >= 110;
        dB_new(:,n) = nan;
    end
end

%% Edit out the inaccuracies
dB_new(:,all(isnan(dB_new(2:end,:)),1))=[];
dB_new(:,all(isinf(dB_new(2:end,:)),1))=[];

days = dB_new(1,:);

%% Spectrogram
clf
figure(1)
imagesc(days,freqs,dB_new(2:end,:))  % spectrogram plot
datetickzoom('x',2,'keepticks')
plotdetails
Spectrum Level Autocorrelation

%% Spectrum Level Autocorrelation
% Load the GXX data into this to perform a two dimensional
% autocorrelation in the x-direction only along each of the
% frequencies.

GXX_autocorr = nan(length(GXX(:,1)),(length(GXX(1,:))*2-1));
%Preallocate GXX_autocorr

for n = 1:length(GXX(:,1));
    GXX_mean = mean(GXX(n,:));
    GXX_variance = GXX(n,:)-GXX_mean;
    GXX_autocorr(n,:) = xcorr(GXX_variance,'coeff');
end

%% Plot the Autocorrelation
frels = 0:0.018181818181818:120;
days = -10.6:21.2/length(GXX(1,:)):10.6;
imagesc(days,frels,GXX_autocorr)
xlim([0 10])
caxis([-0.1 0.4])
xlabel('Lag (Years)','fontsize',14,'fontweight','b')
ylabel('Frequency (Hz)','fontsize',14,'fontweight','b')
colorbar
axis xy
title('10 Year Spectrum Level Autocorrelation of H08S2','fontsize',14,'fontweight','b')
Read Raw Data File (Created By Chad Smith, Adjusted by Russell Hawkins)

function [data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, error_flag] = read_raw_data_v2r0(location, station, start_file, hour, plot_flag)
% This function will pull a 2 hour section of data from the CTBTO database. It is important to mention that occasionally a 2 hour data section will be returned more than 2 hours' worth of data samples (assuming nominal fs==250Hz this is 1800000 samples). This happens because these samples are actually referenced in the database file (.wfdisc) as having this many samples within this time frame. I assume this is the CTBTO's way of keeping the recording system in absolute real-time with a slightly time dependent fs. However, I will have to check with AFTAC.

% INPUTS:
% location - CTBTO site ('HA08', 'HA10', or 'HA11')
% station - specific element (ex. 'H08N1')
% jdate - Julian date (ex. '2005014')
% hour - starting hour of section (must be 0:2:22)
% plot_flag - ==1 then we plot a time series plot of the data upon return

% OUTPUTS:
% data - 2 hours of data
% t - relative time domain
% t_ser - absolute time domain (data fragments can be skewed in time a little we're trying to keep things on an even time grid within the 2 hours)
% fs_nom - nominal fs used for output data array
% cal_coef - nominal calibration from wfdisc file
% cal_per - nominal calibration periods from wfdisc file
% df_flag - is data contiguous flag (0==NO)

addpath read_e1_files -begin % add required MEX files to path
df_flag = 0; % 0==data section NOT fragmented (contiguous 2 hours)
% 1==data section IS fragmented
error_flag = 0; % 0==data exists
% 1==data nonexistant replace with mean

%% SET DATA AND .wfdisc DIRECTORY
main_dir='F:\'; % main data drive
day=[location '.' start_file];
datenumber = datenum(['1/1/' day(6:9)])+ str2double(day(10:end)) - 1;
%% Check if folder exists to determine type of analysis to run
Existence_check = exist(['F:\' location ' location ' start_file ' day(6:9)]);
if Existence_check == 7;
    data_dir = [main_dir day(1:4) ' day ' day(6:9) ' day(10:end) '];
else
    data_dir = [main_dir day(1:4) ' day ' '];
end

%% Set WF Disc Path
wfdisc_path=[data_dir day '.wfdisc'];
wfdisc=readtostruct_wfdisc(wfdisc_path); % pull database info

%% INDEX SECTION DATA AND CALIBRATION
sec_ind=find(ismember(wfdisc.dir,[./' num2str(hour,'%02d')...
       '-' num2str(hour+2,'%02d')])); % 2 hour section files
% (Finds the current two hour section)
sta_ind=find(ismember(wfdisc.sta(sec_ind),sta)); % station files finds
% current station analyzed
data_ind=sec_ind(sta_ind); % data fragments within 2 hour section

%% DECOMPRESS DATA WITHIN 2 HOURS AND CONCAT IF IT IS FRAGMENTED
% allocate space for 2 hours of data (using NaNs to fill any holes)
data = nan(1,250*2*60*60); %250 points of data per second
% ONLY use this try catch section for long data sets. It determines whether
% the data exists for that two hours of data and fills holes if it doesn't
try
    % allocate time domain for 2 hours
    start_ser=UNIXsecs2MatSer(wfdisc.starttime(data_ind(1))); % Defines the
    % time by converting from UNIX (seconds since a certain date) time to matlab time
catch exception
    if strcmp(exception.identifier,'MATLAB:badsubscript') % Catches any %gaps in the data that would cause the code to stop running
disp(['ERROR: NO DATA AVAILABLE! HOUR:' num2str(hour)])
assignin(WS,'Gxx',nan(15000/2,(2*15000/250)/(1-0.50)));
% Assign the variable Gxx with Nans in the workspace
% MUST BE CHANGED TO RUN A DIFFERENT FFT SIZE AND OVERLAP
error_flag = 1;
return
end
end

% Calculating the time window based on the relative times
fs_nom=(wfdisc.samprate(data_ind(1))); dt=1/fs_nom;

t = double((0:length(data)-1).*dt); % relative times
t_ser = start_ser + t./24/60/60; % MATLAB serials

% fill allocation with frags
for data_frags=1:length(data_ind)

    % PULL DATA FROM .wfdisc
    fs=(wfdisc.samprate(data_ind(data_frags))); dt=1/fs;
    cal_coef=wfdisc.calib(data_ind(data_frags));
    cal_per=wfdisc.calper(data_ind(data_frags));
    nsamps=wfdisc.nsamp(data_ind(data_frags));
    filename=wfdisc.dfile(data_ind(data_frags));
    % sec_name=wfdisc.dir{data_ind(data_frags)};
    byte_offset=wfdisc.foff(data_ind(data_frags));

    % DATA TIMES
    frag_start_ser=UNIXsecs2MatSer(wfdisc.starttime(data_ind(data_frags)));
    frag_stop_ser=UNIXsecs2MatSer(wfdisc.endtime(data_ind(data_frags)));
    % datestring_start=datestr(start_ser,30); % get string for mat
    % filename
    % datestring_stop=datestr(stop_ser,30);

    % READ RAW DATA
    if Existence_check == 7; %This accounts for the different
        formatting
        cur_fl=[data_dir wfdisc.dir{data_ind(data_frags)}(3:end) '\'
            filename];
    else
        cur_fl=[data_dir '\' filename]; %Finds the specific file with
            the data
    end
    data_frag=read_e1_v2(cur_fl,byte_offset,nsamps);
    % data_frag=read_e1_matfunc(cur_fl,nsamps); % read e1 MATLAB
        function

    N=length(data_frag);
    if N~=nsamps % check data length with wfdisc
        error('*** Number of data samples does not match wfdisc file! ')
    end

    % place frag into allocated data space (NaNs fill missing data
        space)
    [t_frag_diff frag_ind] = min(abs(t_ser-frag_start_ser));
    t_frag_diff=t Frag_diff*24*60*60;

    % Check for excessive difference in the absolute sample timing of the
    % data array and fragment array.
if t_frag_diff>dt/5 % check data length with wfdisc
  % warning([" File may have excessive time domain sample skewing!
  t_frag_diff == '"...
  num2str(t_frag_diff) '[s] / '"...
  num2str(100*t_frag_diff/dt) '% of dt'])
end

data(frag_ind:nsamps+frag_ind-1) = data_frag;
end % end loop used to pull data frags

%% CHECK IF DATA IS FRAGMENTED
if max(isnan(data))>0
  % warning([" ' station ' within file ' filename ' is fragmented!''])
  % warning(["File is fragmented! at: Hour ' num2str(hour) ' on '
  datestr(starttime,23)])
  df_flag = 1;
else
  df_flag = 0; % full contiguous 2 hours
end

% place zeros in data instead of nans so they are obvious in plot
data(isnan(data)) = 0.0;

%% PLOTTY STUFF

% plot time domain data if desired
if plot_flag==1
  figure; hold on; %setwindowstate(fig1,'maximize');
  % plot(t_ser,data,'.r','markersize',5);
  plot(t_ser,data,'b');
  title(["Station ' station ' from file ' filename ".'],'fontsize',12,'fontweight','b','...
  'interpreter','none');
  xlabel('Absolute Time [hours]','fontsize',12,'fontweight','b');
  ylabel('Amplitude [A/D counts]','fontsize',12,'fontweight','b');
  grid on
datetick('x',13)
end

Calibration and Spectrum Level (Calibration Values NOT provided)

(Created by Chad Smith)

function [Gxx t_spec freqs] = spectro_mat_data_v0r2(data, element,
  start_ser,location,N,overlap)
% This function calibrates and finds the spectrum level of a data
section.

fs=250.0; %Sample rate indicating 250 samples per second
dt=1/fs;

t=[0:length(data)-1].*dt; % [s] Time in seconds
t_mat_ser=double(t)/60/60/24+start_ser;

%% CALIBRATION DATA (dB re lV/uPa coef)
% load correct calibration file for HA08
switch element

    case {'H08N1'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial001.txt ***')
    cal_data=load('resp_fls\HA08Station001.txt','-ascii'); % HT1 Serial 001
    case {'H08N2'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial002.txt ***')
    cal_data=load('resp_fls\HA08Station002.txt','-ascii'); % HT2 Serial 002
    case {'H08N3'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial006.txt ***')
    cal_data=load('resp_fls\HA08Station003.txt','-ascii'); % ED06 Serial 006
    case {'H08S1'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial003.txt ***')
    cal_data=load('resp_fls\HA08Station004.txt','-ascii'); % HT3 Serial 003
    case {'H08S2'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial004.txt ***')
    cal_data=load('resp_fls\HA08Station005.txt','-ascii'); % ED04 Serial 004
    case {'H08S3'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial007.txt ***')
    cal_data=load('resp_fls\HA08Station006.txt','-ascii'); % ED07 Serial 007

    case {'H10N1'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial001.txt ***')
    cal_data=load('resp_fls\HA10Station001.txt','-ascii'); % HT1 Serial 401
    case {'H10N2'}
%     disp(['*** Element: ' element ' ***'])
%     disp('*** Using calibration file Serial002.txt ***')
cal_data=load('resp_flslHA0Station002.txt','-ascii');   % HT2
Serial 402
    case {'H10N3'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial006.txt ***''])
    end
cal_data=load('resp_flslHA0Station003.txt','-ascii');   % ED07
Serial 407
    case {'H10S1'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial003.txt ***''])
    end
cal_data=load('resp_flslHA0Station004.txt','-ascii');   % HT4
Serial 404
    case {'H10S2'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial004.txt ***''])
    end
cal_data=load('resp_flslHA0Station005.txt','-ascii');   % ED05
Serial 405
    case {'H10S3'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial007.txt ***''])
    end
cal_data=load('resp_flslHA0Station006.txt','-ascii');   % ED06
Serial 406

    case {'H11N1'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial001.txt ***''])
    end
cal_data=load('resp_flslHA11Station001.txt','-ascii');   % HT1
Serial 501
    case {'H11N2'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial002.txt ***''])
    end
cal_data=load('resp_flslHA11Station002.txt','-ascii');   % HT2
Serial 502
    case {'H11N3'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial006.txt ***''])
    end
cal_data=load('resp_flslHA11Station003.txt','-ascii');   % HT3
Serial 503
    case {'H11S1'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial003.txt ***''])
    end
cal_data=load('resp_flslHA11Station004.txt','-ascii');   % HT4
Serial 504
    case {'H11S2'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial004.txt ***''])
    end
cal_data=load('resp_flslHA11Station005.txt','-ascii');   % HT5
Serial 505
    case {'H11S3'}
        disp('[''*** Element: ' element ' ***''])
        disp('[''*** Using calibration file Serial007.txt ***''])
```matlab
% Serial 506

cal_data = load('resp_flsl\HA08Station006.txt', '-ascii');  % HT6

cal_freqs = [0.0; (cal_data(:,1)); 125.0];
cal_resp = 10.0*(cal_data(:,2)./20);  % V/μPa

if strcmp(location,'HA08');
    preampgain = NOT GIVEN;  % Hz from Andrew Forbes email
elseif strcmp(location,'HA10');
    preampgain = NOT GIVEN;  % Hz from Andrew Forbes email
elseif strcmp(location,'HA11');
    preampgain = NOT GIVEN;  % Hz from Andrew Forbes email
end

C2 = 9.0/(NOT GIVEN);  % V/count (from A/D spec sheet) volts per analog to
digital counts
lin_preamp_gain = 10.0^(preampgain/20);
data = data.*C2./lin_preamp_gain;  % convert data to volts because its
originally
% in only counts which is why we multiply by volts per count.

% AMBIENT NOISE AVE

steps = floor((length(data)-ceil(N*overlap))/(N*(1-overlap)));
%Uses the size of the data and the parse length to determine the number of
%steps in the data for the two hour segment. As is would give two per
%minute because of the 50 percent overlap resulting in 240 steps.
advance = (N*(1-overlap));  % rate of advance in samples

df = 1/(dt*N);
T = N*dt-dt;
% t_spec = [0:N-1]*dt;
freqs = [0:N/2]*df;  % Determines the resolution of the frequency axis by
% the sample rate and the Number of Points

% interpolate cal data (i.e. estimates intermediate values of
% calibration
% data for the desired resolution frequency)
cal_dom = abs(interp1(cal_freqs, cal_resp, freqs));
% zero outer band of calibration values
[y ind] = min(abs(freqs-1.0));  % 5Hz
```
cal_dom(1:ind-1) = 0.0;
[y ïnd] = min(abs(freqs-110.0)); %115Hz
cal_dom(ind+1:end) = 0.0;

% % Remove DC Offset
% data = data./2^23; % scaling between +/-1
% data = data - mean(data); % removal of DC offset

win = hanning(N)'; % hanning window weighting
Aw = N/(sum(win.^2)); %applying the hanning window to the parse length
% df_eff=sum(win.^2)/(dt.*sum(win).^2);

Gxx=nan(N/2+1,steps); %preallocate space for the spectrum based on
sample size
t_spec=nan(1,steps); %preallocate space for time span
for n=1:steps
    t_spec(n)=t_mat_ser((n-1)*advance+N/2);
    Xm=fft(win.*data((n-1)*advance+1:(n-1)*advance+N),N)*dt;
    Gxx(:,n)=Aw*sspsd(Xm,T).*cal_dom.^2; % calibrate and weight PSD
end

Subsampled Periodicity Analysis

%% Periodicity Subsample
% This code runs through a specified number of days on a specified
% hydrophone and calculates the daily averages of the sound level over
two
% specific frequency bands (default 10-30 and 30-40 Hz). It subsamples
% the
% data to determine what variation occurs for the correlation values of
% the
% autocorrelation plots.
%
% INPUTS:    Start_date - date of start (i.e. 1/1/2005)
%            End_date - date of end (i.e. 1/2/2005)
%            location - Site (i.e. HA08)
%            station - hydrophone (i.e. H08N1)
%
% OUTPUT FIGURES:     Daily average plot vs time
%                     Autocorrelation plots
%
%% Run as Matlab Script
clear all;close all;clc;
Start_date = input('Start Date:','s');
End_date = input('End Date:','s');
location = input('Location:','s');
station = input('Station:','s');

disp('____________________________________')
disp(['Start Date:' Start_date]);
disp(['End Date:' End_date]);
disp(['Location:' location]);
disp(['Station:' station]);
disp('____________________________________')

Start_day = datenum(Start_date,'mm/dd/yyyy');
End_day = datenum(End_date, 'mm/dd/yyyy');

Number_of_days = datenum(End_date,'mm/dd/yyyy') -
datenum(Start_date,'mm/dd/yyyy') + 1;
dB = nan(7501,Number_of_days);
N = 15000; % parse length
overlap = 0.0;
gxxaverage10to30 = nan((N/250)/(1-overlap)*24,Number_of_days);
gxxaverage30to40 = nan((N/250)/(1-overlap)*24,Number_of_days);

% Set loops
for n = Start_day:End_day; % increment through multiple days
disp(['*** PROCESSING DAY ' datestr(n,23) '... ***'])
clearvars -except Number_of_days Start_day End_day Start_date
End_date...
location station n freqs dB N overlap gxxaverage10to30
gxxaverage30to40 %Clear the large variable data on every round

% site info and data/time
start_file=datestr2DOY(datestr(n,23)); %get starting directory name
plot_flag = 0;

% Check if folder exists
Existence_check = exist(['F:\' location '\' location '. ' start_file]);
if Existence_check == 7;

% Loop Through the Day
for hour=0:2:22 % increment through day

%% LOAD 2 HOURS

try
[data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =...
read_raw_data_v3r0(location,station,start_file,hour,plot_flag);
catch MException
if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
data(1,1:1800000) = nan;
continue
end
end

if exist('data','var')==0; % This helps to account for the missing
data sections
if overlap == 0;
    Gxx = nan(N/2+1,(2*N/250)/(1-overlap));
    eval(['Gxx' num2str(hour) '=Gxx;' ]); continue
else
    Gxx = nan(N/2+1,(2*N/250)/(1-overlap)-1);
    eval(['Gxx' num2str(hour) '=Gxx;' ]); continue
end

if length(data)>2*60*60 % assume we can chop anything more than 2 hours
    data=data(1:2*250*60*60); % (bad assumption?)
end
if length(data)<2*60*60 % assume we can chop anything more than 2 hours
    error('*** Incorrect data length!!! ***')
end

%% SPECTRUM
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station, t_ser(1),location,N,overlap);

[~, ind0] = min(abs(freqs-10.0)); %10Hz bin
[~, ind1] = min(abs(freqs-30.0));
[~, ind2] = min(abs(freqs-40.0));

section = length(Gxx(1,:))*hour/2;

for m = 1:length(Gxx(1,:));
    Gxxtemp = Gxx(:,m);
    gxxaverage10to30(section+m,n-Start_day+1) = mean(Gxxtemp(ind0:ind1));
    gxxaverage30to40(section+m,n-Start_day+1) = mean(Gxxtemp(ind1:ind2));
end
end % End increment through hours
else
    disp('FOLDER DOES NOT EXIST!')
    continue
end

%% Continue Editing Matrix
proceed = input('Continue to EDIT band values? (i.e. y or n)':'s');
if strcmp(proceed,'y') == 1;
    sub-sample1_10to30 = gxxaverage10to30;
    sub-sample5_10to30 = gxxaverage10to30(1:5:end,:);
subsample15_10to30 = gxxaverage10to30(1:15:end,:);  
subsample30_10to30 = gxxaverage10to30(1:30:end,:);  
subsample60_10to30 = gxxaverage10to30(1:60:end,:);  

subsample1_30to40 = gxxaverage30to40;  
subsample5_30to40 = gxxaverage30to40(1:5:end,:);  
subsample15_30to40 = gxxaverage30to40(1:15:end,:);  
subsample30_30to40 = gxxaverage30to40(1:30:end,:);  
subsample60_30to40 = gxxaverage30to40(1:60:end,:);  

%% Take the daily mean of those subsamples  
daily10to30(1,:) = mean(subsample1_10to30);  
daily10to30(2,:) = mean(subsample5_10to30);  
daily10to30(3,:) = mean(subsample15_10to30);  
daily10to30(4,:) = mean(subsample30_10to30);  
daily10to30(5,:) = mean(subsample60_10to30);  

daily30to40(1,:) = mean(subsample1_30to40);  
daily30to40(2,:) = mean(subsample5_30to40);  
daily30to40(3,:) = mean(subsample15_30to40);  
daily30to40(4,:) = mean(subsample30_30to40);  
daily30to40(5,:) = mean(subsample60_30to40);  

%% Other Essentials  
time = Start_day:End_day;  
dB10to30 = 10.*log10(daily10to30);  
dB30to40 = 10.*log10(daily30to40);  

%% Edit out incorrect days and calibration day  
dB10to30(dB10to30>110)=nan;  
dB30to40(dB30to40>110)=nan;  

dB_new10 = nan(length(dB10to30(:,1))+1,length(dB10to30(1,:)));  
dB_new10(2:length(dB10to30(:,1))+1,:) = dB10to30(1:length(dB10to30(:,1)), :);  
dB_new10(1,:) = time;  

dB_new30 = nan(length(dB30to40(:,1))+1,length(dB30to40(1,:)));  
dB_new30(2:length(dB30to40(:,1))+1,:) = dB30to40(1:length(dB30to40(:,1)), :);  
dB_new30(1,:) = time;  

dB_new10(:,any(isnan(dB_new10),1))=[];  
dB_new30(:,any(isnan(dB_new30),1))=[];  

%% Plot both frequency bands  
figure(1)  
plot(dB_new10(1,:),dB_new10(2,:),dB_new10(3,:),dB_new10(4,:),...  
dB_new10(4,:),dB_new10(5,:),dB_new10(6,:))
datetickzoom('x',2,'keeticks')
xlim([Start_day End_day]);
title(['10-30 Hz Band for Station ' station ' from ' Start_date ' to ' End_date'],...
    'Fontweight','b','Fontsize',16)
xlabel('Date','Fontweight','b','Fontsize',16)
ylabel('Sound Level (dB re 1\mu Pa^2/Hz)','Fontweight','b','Fontsize',16)
legend('Continuous','5 Min','15 Min','30 Min','60 Min')
plotdetails

figure(2)
plot(dB_new30(1,:),dB_new30(2,:),dB_new30(1,:),dB_new30(3,:),dB_new30(1,:),...
    dB_new30(4,:),dB_new30(1,:),dB_new30(5,:),dB_new30(1,:),dB_new30(6,:))
datetickzoom('x',2,'keeticks')
xlim([Start_day End_day]);
title(['30-40 Hz Band for Station ' station ' from ' Start_date ' to ' End_date'],...
    'Fontweight','b','Fontsize',16)
xlabel('Date','Fontweight','b','Fontsize',16)
ylabel('Sound Level (dB re 1\mu Pa^2/Hz)','Fontweight','b','Fontsize',16)
legend('Continuous','5 Min','15 Min','30 Min','60 Min')
plotdetails
else
disp('*** DONE ***')
end

%% Perform autocorrelation analysis on Bands
proceed2 = input('Continue to perform correlation? (i.e. y or n):','s');
if strcmp(proceed2,'y') == 1;
    days = -10.6:21.2/length(autocorr10(1,:)):10.6-21.2/length(autocorr10(1,:));

    for n = 1:5;
        A10_mean = mean(dB_new10(n+1,:));
        A10_variance = dB_new10(n+1,:)-A10_mean;
        autocorr10(n,:) = xcorr(A10_variance,'coeff');
    end

    for n = 1:5;
        A30_mean = mean(dB_new30(n+1,:));
        A30_variance = dB_new30(n+1,:)-A30_mean;
        autocorr30(n,:) = xcorr(A30_variance,'coeff');
    end

figure(3)
P-value Distribution

%% pdist

% This code plots the normalized cumulative distribution of p-values for the bootstrapping analysis. It requires inputs of each of the p-value matrices under specific names shown below.

% INPUTS: The following are the matrix names required in this code.
% prob10to30_N60
% prob10to30_S60
% prob10to30_N200
% prob10to30_S200
% prob40to60_N60
% prob40to60_S60
% prob40to60_N200
% prob40to60_S200
% prob85to105_N60
% prob85to105_S60
% prob85to105_N200
% prob85to105_S200
% prob1to110_N60
% prob1to110_S60
% prob1to110_N200
% prob1to110_S200

% OUTPUTS FIGURES (4)
% p-value distribution 10-30 Hz
% p-value distribution 40-60 Hz
% p-value distribution 85-105 Hz
% p-value distribution 1-110 Hz

%% Quad plot of normalized cumulative distributions
pd10to30_N60 = fitdist(prob10to30_N60,'kernel');
pd10to30_S60 = fitdist(prob10to30_S60,'kernel');
pd10to30_N200 = fitdist(prob10to30_N200,'kernel');
pd10to30_S200 = fitdist(prob10to30_S200,'kernel');
pd40to60_N60 = fitdist(prob40to60_N60,'kernel');
pd40to60_S60 = fitdist(prob40to60_S60,'kernel');
pd40to60_N200 = fitdist(prob40to60_N200,'kernel');
pd40to60_S200 = fitdist(prob40to60_S200,'kernel');
pd85to105_N60 = fitdist(prob85to105_N60,'kernel');
pd85to105_S60 = fitdist(prob85to105_S60,'kernel');
pd85to105_N200 = fitdist(prob85to105_N200,'kernel');
pd85to105_S200 = fitdist(prob85to105_S200,'kernel');
pd1to110_N60 = fitdist(prob1to110_N60,'kernel');
pd1to110_S60 = fitdist(prob1to110_S60,'kernel');
pd1to110_N200 = fitdist(prob1to110_N200,'kernel');
pd1to110_S200 = fitdist(prob1to110_S200,'kernel');

%% Determine Probability density
xvalues = linspace(0,0.2,40);
xvalues2 = linspace(0,0.015,10);
xvalues3 = linspace(0,0.001,10);
spread10to30_N60 = pdf(pd10to30_N60,xvalues3);
spread10to30_S60 = pdf(pd10to30_S60,xvalues3);
spread10to30_N200 = pdf(pd10to30_N200,xvalues3);
spread10to30_S200 = pdf(pd10to30_S200,xvalues3);
spread40to60_N60 = pdf(pd40to60_N60,xvalues);
spread40to60_S60 = pdf(pd40to60_S60,xvalues);
spread40to60_N200 = pdf(pd40to60_N200,xvalues);
spread40to60_S200 = pdf(pd40to60_S200,xvalues);
spread85to105_N60 = pdf(pd85to105_N60,xvalues);
spread85to105_S60 = pdf(pd85to105_S60,xvalues);
spread85to105_N200 = pdf(pd85to105_N200,xvalues);
spread85to105_S200 = pdf(pd85to105_S200,xvalues);
spread1to110_N60 = pdf(pd1to110_N60,xvalues2);
spread1to110_S60 = pdf(pd1to110_S60,xvalues2);
spread1to110_N200 = pdf(pd1to110_N200,xvalues2);
spread1to110_S200 = pdf(pd1to110_S200,xvalues2);

%% Normalize Probability Densities
norm10to30_N60 = spread10to30_N60./sum(spread10to30_N60);
norm10to30_S60 = spread10to30_S60./sum(spread10to30_S60);
norm10to30_N200 = spread10to30_N200./sum(spread10to30_N200);
norm10to30_S200 = spread10to30_S200./sum(spread10to30_S200);
norm40to60_N60 = spread40to60_N60./sum(spread40to60_N60);
norm40to60_S60 = spread40to60_S60./sum(spread40to60_S60);
norm40to60_N200 = spread40to60_N200./sum(spread40to60_N200);
norm40to60_S200 = spread40to60_S200./sum(spread40to60_S200);
norm85to105_N60 = spread85to105_N60./sum(spread85to105_N60);
norm85to105_S60 = spread85to105_S60./sum(spread85to105_S60);
norm85to105_N200 = spread85to105_N200./sum(spread85to105_N200);
norm85to105_S200 = spread85to105_S200./sum(spread85to105_S200);
norm1to110_N60 = spread1to110_N60./sum(spread1to110_N60);
norm1to110_S60 = spread1to110_S60./sum(spread1to110_S60);
norm1to110_N200 = spread1to110_N200./sum(spread1to110_N200);
norm1to110_S200 = spread1to110_S200./sum(spread1to110_S200);

%% Plot distributions
subplot(2,2,1)
hold on
```matlab
plot(xvalues3,norm10to30_N60,'b-o',xvalues3,norm10to30_S60,'g-*',...
xvalues3,norm10to30_N200,'m-',...
xvalues3,norm10to30_S200,'k','LineWidth',2.5)
line([0.01 0.01], [0 1],'LineStyle', '--', 'Color', 'r')
xlim([0,0.015])
ylim([0 0.4])
xlabel('p-value')
ylabel('density (10-30 Hz)')
legend('H08N1 - 60s','H08S2 - 60s','H08N1 - 200s','H08S2 - 200s')
plotdetails
hold off

subplot(2,2,3)
hold on
plot(xvalues,norm40to60_N60,'b',xvalues,norm40to60_S60,'g',...
xvalues,norm40to60_N200,'m',xvalues,norm40to60_S200,'k','LineWidth',2.5)
[pks3,locs3] = findpeaks(norm40to60_S60);
plot(xvalues(locs3),pks3,'k^','markerfacecolor','g','markersize',10);
line([0.01 0.01], [0 1],'LineStyle', '--', 'Color', 'r')
xlim([0,0.1])
ylim([0 0.4])
xlabel('p-value')
ylabel('density (40-60 Hz)')
legend('H08N1 - 60s','H08S2 - 60s','H08N1 - 200s','H08S2 - 200s')
plotdetails
hold off

subplot(2,2,4)
hold on
plot(xvalues,norm85to105_N60,'b',xvalues,norm85to105_S60,'g',...
xvalues,norm85to105_N200,'m',xvalues,norm85to105_S200,'k','LineWidth',2.5)
[pks,locs] = findpeaks(norm85to105_N60);
plot(xvalues(locs),pks,'k^','markerfacecolor','b','markersize',10);
[pks2,locs2] = findpeaks(norm85to105_S60);
plot(xvalues(locs2),pks2,'k^','markerfacecolor','g','markersize',10);
line([0.01 0.01], [0 1],'LineStyle', '--', 'Color', 'r')
xlim([0,0.1])
ylim([0 0.4])
xlabel('p-value')
ylabel('density (85-105 Hz)')
legend('H08N1 - 60s','H08S2 - 60s','H08N1 - 200s','H08S2 - 200s')
plotdetails
hold off

subplot(2,2,2)
hold on
plot(xvalues2,norm1to110_N60,'b-o',xvalues2,norm1to110_S60,'g-*',...
xvalues2,norm1to110_N200,'m-',...
xvalues2,norm1to110_S200,'k','LineWidth',2.5)
```
% Window Length Bootstrap

 Window Length Bootstrap

% fftmultiple

fftmultiple(Start_date, End_date, location, station, fftsizes, overlap, band1, band2, band3, band4)
%
Sample - fftmultiple('6/1/2008','8/1/2008','HA08','H08S2',[2500 3750 7500 15000 50000],0.0,[10 30],[40 60],[85 105],[110 110])

% Run as Matlab Script; Compares through ANOVA and Post-hoc Multiple
% Comparison tests the means and variation of the sound level over 5
% different fft sizes. (i.e. N = 15000,7500 etc.)
% Also, it performs the Kruskal-Wallis test to determine whether the
% shape of the distribution is similar or significantly different
% and lastly, it plots the probability density function over the
% frequency bands selected.
% To view data segments at 30 seconds a piece use a 7500 point fft
% (7500/250 = 30 seconds). To view data segments at 1 minute a piece us
% a
% 15000 pt fft (15000/250 = 60 seconds). ETC.
%
% INPUTS(6):
% Runlength - Length of Run
% Start_date - The starting date form 'mm/dd/yyyy' (i.e. '12/25/1986')
% End_date - The End date of the analysis
% location - The site location of the CTBTO dataset (i.e 'HA08', etc.)
% station - The Hydrophone Station number (i.e. 'H08N1', 'H11S2', etc.)
% fftsizes - Size of fft over which it will be run
% overlap - The percent overlap in the fft
% (USE 0.0 TO AVOID UNNECESSARY AVERAGING)
%
% OUTPUTS(1):
% Diary text file under Specified location with the P-values and
% other
% information from the tests
%
% PLOTS(8):
% Post-hoc multiple comparisons for all frequency bands
% Probability Density Functions for all frequency bands over all
% subsampling rates
%% Create Date Matrix from which to pull the starting dates
clear all;close all;clc;
for s = 731237:735071; % Based on datenum
    datematrix(s-731237+1,:) = s;
end

% Remove starting days with no data -42 to run through 100 points
datematrix(datematrix==732040,:)=[];
datematrix(datematrix==733971,:)=[];
datematrix(datematrix==734098,:)=[];
for ss = 732096:732178;
    datematrix(datematrix==ss,:)=[];
end

%% RUN AS MATLAB SCRIPT
runlength = 10;
location = 'HA08';
station = 'H08S2';
fftsizes = [2500 3750 7500 15000 50000];
overlap = 0;
band1 = [10,30];
band2 = [40,60];
band3 = [85,105];
band4 = [1,110];

disp(['Location:' location]);
disp(['Station:' station]);
disp(['FFT Sizes = ' num2str(fftsizes(1)) ',' num2str(fftsizes(2)) '
', ...' num2str(fftsizes(3)) ',' num2str(fftsizes(4)) ','
num2str(fftsizes(5))]);

% Preallocate runlength files
prob10to30 = nan(runlength,4);
prob40to60 = nan(runlength,4);
prob85to105 = nan(runlength,4);
probSpectrum = nan(runlength,4);

% Run through Bootstrap
for z = 1:runlength;
    date = randsample(datematrix,1);
    Start_date = datestr(date);
    Number_of_days = 7;
    Start_day = datenum(Start_date,'dd-mmm-yyyy');
    End_day = datenum(Start_date,'dd-mmm-yyyy') + Number_of_days-1;
    End_date = datestr(End_day);
    addpath read_e1_files -begin
    disp('____________________________________')
disp(['Run Number ' num2str(z) ' -->'])
disp(['Start Date:' Start_date]);
```matlab
disp(['End Date:' End_date]);
disp('____________________________________')

%% Preallocate Space
SPLband1 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband2 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband3 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);
SPLband4 = nan(Number_of_days*(1440/(min(fftsizes)/250/60)),5);

%% Begin loop for fft_length
for a = 1:5;
    fftsize = fftsizes(a); % parse length
    % Start Diary
    diary(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of
    |P-Values\'...
    |'Bootstrap' station '-' num2str(fftsize) '-fft.txt']);

    %% Begin loop for days
    for d = Start_day:End_day   % increment through multiple days
        for hour = 0:2:22; % increment through day
            % Site info and data/time
            start_file=datestr2DOY(datestr(d,23)); %get starting directory name
            plot_flag = 0;
            %% Read 2 Hours of Raw Data
            N = fftsize;
            fs = 250; %Hz Sampling rate
            dt = 1/fs;
            df = 1/(dt*N);
            try
                [data, t, t_ser, fs_nom, cal_coef, cal_per, df_flag, ~] =... 
                read_raw_data_v3r0(location,station,start_file,hour,plot_flag);
            catch MException
                if strcmp(MException.identifier,'MATLAB:unassignedOutputs')
                    data(1,1:1800000) = nan;
                    continue
                end
            end
            if exist('data','var')==0; % This helps to account for the missing
            data sections
                if overlap == 0;
                    Gxx = nan(N/2+1,(2*N/250)/(1-overlap));
                    eval(['Gxx' num2str(hour) '=Gxx;'])
                    continue
                else
                    Gxx = nan(N/2+1,(2*N/250)/(1-overlap)-1);
                    eval(['Gxx' num2str(hour) '=Gxx;'])
                    continue
                end
            end
```

end
end

if length(data)>2*60*60 % assume we can chop anything more than 2 hours
data=data(1:2*250*60*60); % (bad assumption?)
end
if length(data)<2*60*60 % assume we can chop anything more than 2 hours
    error('*** Incorrect data length!!! ***')
end

%% Apply FFT to Data
[Gxx t_spec freqs] = spectro_mat_data_v0r2(data, station, t_ser(l), location, N, overlap);

Gxx_ave=mean(Gxx,2); % [uPa^2/Hz]
Gxx_ave_dB=10*log10(Gxx_ave);

%% Frequency Band Indeces (Bins)
% Find the mean at each of the subsample rates

 [~, ind0] = min(abs(freqs-band1(1)));
 [~, ind1] = min(abs(freqs-band1(2)));
 [~, ind2] = min(abs(freqs-band2(1)));
 [~, ind3] = min(abs(freqs-band2(2)));
 [~, ind4] = min(abs(freqs-band3(1)));
 [~, ind5] = min(abs(freqs-band3(2)));
 [~, ind6] = min(abs(freqs-band4(1)));
 [~, ind7] = min(abs(freqs-band4(2)));

%% Subsampling and Accounting for Overlap
section = (d-Start_day)*length(Gxx(1,:))*12 ...
+ length(Gxx(1,:))*hour/2 + 1;

for m = 1:1:length(Gxx(1,:));
    Gxxtemp = Gxx(:,m);
    SPLband1(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind0:ind1)));
    SPLband2(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind2:ind3)));
    SPLband3(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind4:ind5)));
    SPLband4(section+m-1,a) = 10.*log10(mean(Gxxtemp(ind6:ind7)));
end
end % end increment through hour
end % end increment through multiple days
end % end multiple averaging rate

%% Get rid of zeros and invalid values
SPLband1(isinf(SPLband1)) = nan;
SPLband2(isinf(SPLband2)) = nan;
SPLband3(isinf(SPLband3)) = nan;
SPLband4(isinf(SPLband4)) = nan;

%% Perform ANOVA and post-hoc multiple comparison on points
windows = {num2str(fftsizes(1)/250),num2str(fftsizes(2)/250),...}
    num2str(fftsizes(3)/250),num2str(fftsizes(4)/250),...}
    num2str(fftsizes(5)/250)};

[p1,table1,stat] = anova1(SPLband1,windows,'off');
[p2,table2,stat2] = anova1(SPLband2,windows,'off');
[p3,table3,stat3] = anova1(SPLband3,windows,'off');
[p4,table4,stat4] = anova1(SPLband4,windows,'off');

%% Display probability density Estimates of each of the selected Units
[f1,x11,u11] = ksdensity(SPLband1(:,1));
[f2,x12,u22] = ksdensity(SPLband1(:,2));
[f3,x13,u33] = ksdensity(SPLband1(:,3));
[f4,x14,u44] = ksdensity(SPLband1(:,4));
[f5,x15,u55] = ksdensity(SPLband1(:,5));

[f11,x111,u111] = ksdensity(SPLband2(:,1));
[f22,x122,u222] = ksdensity(SPLband2(:,2));
[f33,x133,u333] = ksdensity(SPLband2(:,3));
[f44,x144,u444] = ksdensity(SPLband2(:,4));
[f55,x155,u555] = ksdensity(SPLband2(:,5));

[f111,x1111,u1111] = ksdensity(SPLband3(:,1));
[f222,x1222,u2222] = ksdensity(SPLband3(:,2));
[f333,x1333,u3333] = ksdensity(SPLband3(:,3));
[f444,x1444,u4444] = ksdensity(SPLband3(:,4));
[f555,x1555,u5555] = ksdensity(SPLband3(:,5));

[f1111,x11111,u11111] = ksdensity(SPLband4(:,1));
[f2222,x12222,u22222] = ksdensity(SPLband4(:,2));
[f3333,x13333,u33333] = ksdensity(SPLband4(:,3));
[f4444,x14444,u44444] = ksdensity(SPLband4(:,4));
[f5555,x15555,u55555] = ksdensity(SPLband4(:,5));

type = 'Subsample';
PDF1 = figure(1);
plot(x11,f1,'-',x12,f2,'-',x13,f3,'-',x14,f4,'-',x15,f5,'-'
    ',Linewidth',1.5,...
    'MarkerEdgeColor','k','MarkerFaceColor',[.49 1 .63],'
    'MarkerSize',7);
    title(['Probability densities:' num2str(band1(1)) ' - ' num2str(band1(2)) '
    ' Hz']);
xlabel('Sound Level (dB re 1\mu Pa^2/Hz)');
ylabel('Density');
plotdetails
legend(['Window = ' num2str(fftsizes(1)/250) ' s'],
    ['Window = ' num2str(fftsizes(2)/250) ' s'],...
    ['Window = ' num2str(fftsizes(3)/250) ' s'],
    ['Window = ' num2str(fftsizes(4)/250) ' s'],
    ['Window = ' num2str(fftsizes(5)/250) ' s']);

PDF2 = figure(2);
plot(x11,f11,'-','x12,f22','-','x13,f33','-','x14,f44','-','x15,f55','-','Linewidth',1.5,...
'MarkerEdgeColor','k','MarkerFaceColor',[1 .95 0],'MarkerSize',7);
title(['Probability densities: ' num2str(band2(1)) ' - ' num2str(band2(2)) ' Hz']);
xlabel('Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Density');
plotdetails
legend([\'Window = ' num2str(fftsizes(1)/250) ' s'],[\'Window = ' num2str(fftsizes(2)/250) ' s'],...
[\'Window = ' num2str(fftsizes(3)/250) ' s'],[\'Window = ' num2str(fftsizes(4)/250) ' s'],...
[\'Window = ' num2str(fftsizes(5)/250) ' s']);

PDF3 = figure(3);
plot(x111,f111,'-','x122,f222','-','x133,f333','-','x144,f444','-','x155,f555',...
'-','Linewidth',1.5,'MarkerEdgeColor','k','MarkerFaceColor',[.5 0 1],'MarkerSize',7);
title(['Probability densities: ' num2str(band3(1)) ' - ' num2str(band3(2)) ' Hz']);
xlabel('Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Density');
plotdetails
legend([\'Window = ' num2str(fftsizes(1)/250) ' s'],[\'Window = ' num2str(fftsizes(2)/250) ' s'],...
[\'Window = ' num2str(fftsizes(3)/250) ' s'],[\'Window = ' num2str(fftsizes(4)/250) ' s'],...
[\'Window = ' num2str(fftsizes(5)/250) ' s']);

PDF4 = figure(4);
plot(x1111,f1111,'-','x1222,f2222','-','x1333,f3333','-','x1444,f4444','-','x1555,f5555',...
'-','Linewidth',1.5,'MarkerEdgeColor','k','MarkerFaceColor',[.5 0 1],'MarkerSize',7);
title(['Probability densities: ' num2str(band4(1)) ' - ' num2str(band4(2)) ' Hz']);
xlabel('Mean Sound Level (dB re 1 \mu Pa^2/Hz)');

%% Kruskal-Wallis Test
MC1 = figure(5);
[p5,anovatab5,stat5] = kruskalwallis(SPLband1,windows,'off');
multcompare(stat);
title(['Multiple Comparison Test ' num2str(band1(1)) ' - ' num2str(band1(2)) ' Hz']);
xlabel('Mean Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Window length (s)')

plotdetails

MC2 = figure(6);
[p6, anovatab6, stat6] = kruskalwallis(SPLband2, windows, 'off');
multcompare(stat2);
title(['Multiple Comparison Test ' num2str(band2(1)) ' - ' num2str(band2(2)) ' Hz']);
xlabel('Mean Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Window length (s)')
plotdetails

MC3 = figure(7);
[p7, anovatab7, stat7] = kruskalwallis(SPLband3, windows, 'off');
multcompare(stat3);
title(['Multiple Comparison Test ' num2str(band3(1)) ' - ' num2str(band3(2)) ' Hz']);
xlabel('Mean Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Window length (s)')
plotdetails

MC4 = figure(8);
[p8, anovatab8, stat8] = kruskalwallis(SPLband4, windows, 'off');
multcompare(stat4);
title(['Multiple Comparison Test ' num2str(band4(1)) ' - ' num2str(band4(2)) ' Hz']);
xlabel('Mean Sound Level (dB re 1 \mu Pa^2/Hz)');
ylabel('Window length (s)')
plotdetails

%% Save analysis details to .txt file
warning off MATLAB:MKDIR:DirectoryExists
mkdir(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\' station], datestr(Start_day, 'dd.mm.yyyy'))
diary(['C:\Users\rsh935\Documents\Research\Matlab Variables\Diaries of P-Values\' station '\ datestr(Start_day, 'dd.mm.yyyy') '\fft-analysis.txt'])

disp('______________________________________')
disp(['Start Date:' Start_date])
disp(['End Date:' End_date])
disp(['Location:' location])
disp(['Station:' station])
disp(['fft sizes = ' num2str(fftsizes)])
disp(['Overlap = ' num2str(overlap)])
disp('______________________________________')

%% Save p-values for each 42 day period

% 1-ANOVA p-value, 2-ANOVA f-statistic, 3-Kruskal-Wallis p-value, 4-H-Statistic
prob10to30(z,1:4)=[p1,cell2mat(table1(2,5)),p5,cell2mat(anovatab5(2,5))];
prob40to60(z,1:4)=[p2,cell2mat(table2(2,5)),p6,cell2mat(anovatab6(2,5))];
prob85to105(z,1:4)=[p3,cell2mat(table3(2,5)),p7,cell2mat(anovatab7(2,5))];
probSpectrum(z,1:4)=[p4,cell2mat(table4(2,5)),p8,cell2mat(anovatab8(2,5))];

disp('ANOVA Test Stats...')
disp([num2str(band1(1)) '-' num2str(band1(2)) 'Hz: p = ' num2str(p1)'
',f = ' num2str(prob10to30(z,2))'])
disp([num2str(band2(1)) '-' num2str(band2(2)) 'Hz: p = ' num2str(p2)'
',f = ' num2str(prob40to60(z,2))'])
disp([num2str(band3(1)) '-' num2str(band3(2)) 'Hz: p = ' num2str(p3)'
',f = ' num2str(prob85to105(z,2))'])
disp([num2str(band4(1)) '-' num2str(band4(2)) 'Hz: p = ' num2str(p4)'
',f = ' num2str(probSpectrum(z,2))'])
disp('____________________________________')

disp('Kruskal Wallis Stats...')
disp([num2str(band1(1)) '-' num2str(band1(2)) 'Hz: p = ' num2str(p5)'
',H = ' num2str(prob10to30(z,4))'])
disp([num2str(band2(1)) '-' num2str(band2(2)) 'Hz: p = ' num2str(p6)'
',H = ' num2str(prob40to60(z,4))'])
disp([num2str(band3(1)) '-' num2str(band3(2)) 'Hz: p = ' num2str(p7)'
',H = ' num2str(prob85to105(z,4))'])
disp([num2str(band4(1)) '-' num2str(band4(2)) 'Hz: p = ' num2str(p8)'
',H = ' num2str(probSpectrum(z,4))'])
disp('____________________________________')
diary off

%% Save figures
mkdir(['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station','datestr(Start_day,'dd.mm.yyyy'))
saveas(MC1,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-MC 10-30 Hz.fig'])
saveas(MC2,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-MC 40-60 Hz.fig'])
saveas(MC3,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-MC 85-105 Hz.fig'])
saveas(MC4,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-MC 1-110 Hz.fig'])
saveas(PDF1,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-PDF 10-30 Hz.fig'])
saveas(PDF2,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-PDF 40-60 Hz.fig'])
saveas(PDF3,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-PDF 85-105 Hz.fig'])
saveas(PDF4,['C:\Users\rsh935\Documents\Research\Matlab Figures\'
station '"
','datestr(Start_day,'dd.mm.yyyy') '"FFT-PDF 1-110 Hz.fig'])
disp('***DONE***')
close all
end

%% Calculate Statistics
pass10to30 = sum(prob10to30(:,1:2:4)>0.01)/length(prob10to30(:,1));
pass40to60 = sum(prob40to60(:,1:2:4)>0.01)/length(prob40to60(:,1));
pass85to105 = sum(prob85to105(:,1:2:4)>0.01)/length(prob85to105(:,1));
passSpectrum = sum(probSpectrum(:,1:2:4)>0.01)/length(probSpectrum(:,1));

disp('____________________________________')
disp('Percent no significant change ANOVA:')
disp(['           ...10-30 Hz:' num2str(pass10to30(1)*100) '%'])
disp(['           ...40-60 Hz:' num2str(pass40to60(1)*100) '%'])
disp(['           ...85-105 Hz:' num2str(pass85to105(1)*100) '%'])
disp(['           ...1-120 Hz:' num2str(passSpectrum(1)*100) '%'])
disp('____________________________________')

disp('Percent no significant change Kruskal-Wallis:')
disp(['           ...10-30 Hz:' num2str(pass10to30(2)*100) '%'])
disp(['           ...40-60 Hz:' num2str(pass40to60(2)*100) '%'])
disp(['           ...85-105 Hz:' num2str(pass85to105(2)*100) '%'])
disp(['           ...1-120 Hz:' num2str(passSpectrum(2)*100) '%'])
disp('____________________________________')

% This portion calculates and displays the percentages from the
% Kruskal-Wallis bootstrapping analysis test

% end