NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

PERFORMANCE OF KERNEL-BASED METHODS IN ANALYZING UNDERWATER NOISE LEVEL STATISTICS

by

Alexandra M. Smith

March 2022

Thesis Advisor: Marko Orescanin
Co-Advisors: Kevin B. Smith
Paul Leary

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** March 2022	**3. REPORT TYPE AND DATES COVERED** Master's thesis

4. TITLE AND SUBTITLE
PERFORMANCE OF KERNEL-BASED METHODS IN ANALYZING UNDERWATER NOISE LEVEL STATISTICS

5. FUNDING NUMBERS

6. AUTHOR(S) Alexandra M. Smith

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT (maximum 200 words)

In this work we present an evaluation of kernel-based methods of estimating spectral density statistics compared to histogram-based methods based on evaluations of fine structures. The data is from the Monterey Accelerated Research System, where month-long windows of data from February 2019 to January 2021 were preprocessed to provide 30-second power averages for analysis. The data was processed in MATLAB where it was analyzed through traditional histogram and kernel-based methods to produce spectral probability densities (SPD). The structure presented in the respective SPDs for each method was compared to determine where the methods converge, what differences appear, and which method performs best. Performance was analyzed through a statistical analysis of physical noise sources, specifically wind speed at the surface of the water. We developed a simple classifier that can categorize ambient noise data as being associated with high wind speed or low wind speed, which demonstrates a distinct difference between days with similar noise level characteristics and different noise level characteristics. This classifier was optimized in the 200-400 Hz range, where the statistics for the dataset for accuracy, precision, recall, and F-value were 0.6302, 0.6317, 0.9936, and 0.7706, respectively, with a windy vs. non-windy threshold value set at 65 dB re 1 μPa/√Hz. We found that for the chosen time duration of analysis window, the classifier had similar performance regardless of method.

14. SUBJECT TERMS
signal analysis, kernel-based methods, underwater noise level statistics, power spectral density, spectrogram, periodogram

15. NUMBER OF PAGES
57

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

PERFORMANCE OF KERNEL-BASED METHODS IN ANALYZING UNDERWATER NOISE LEVEL STATISTICS

Alexandra M. Smith
Ensign, United States Navy
BS, United States Naval Academy, 2020

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN PHYSICS

from the

NAVAL POSTGRADUATE SCHOOL
March 2022

Approved by: Marko Orescanin
Advisor

Kevin B. Smith
Co-Advisor

Paul Leary
Co-Advisor

Joseph P. Hooper
Chair, Department of Physics

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In this work we present an evaluation of kernel-based methods of estimating spectral density statistics compared to histogram-based methods based on evaluations of fine structures. The data is from the Monterey Accelerated Research System, where month-long windows of data from February 2019 to January 2021 were preprocessed to provide 30-second power averages for analysis. The data was processed in MATLAB where it was analyzed through traditional histogram and kernel-based methods to produce spectral probability densities (SPD). The structure presented in the respective SPDs for each method was compared to determine where the methods converge, what differences appear, and which method performs best. Performance was analyzed through a statistical analysis of physical noise sources, specifically wind speed at the surface of the water. We developed a simple classifier that can categorize ambient noise data as being associated with high wind speed or low wind speed, which demonstrates a distinct difference between days with similar noise level characteristics and different noise level characteristics. This classifier was optimized in the 200-400 Hz range, where the statistics for the dataset for accuracy, precision, recall, and F-value were 0.6302, 0.6317, 0.9936, and 0.7706, respectively, with a windy vs. non-windy threshold value set at 65 dB re 1 μPa/√Hz. We found that for the chosen time duration of analysis window, the classifier had similar performance regardless of method.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

DFT	Discrete Fourier Transform
fft	Fast Fourier Transform
KSFE	Kernel Smoothing Function Estimate
MARS	Monterey Accelerated Research System
PAM	Passive Acoustic Monitoring
PDF	Probability Density Function
PSD	Power Spectral Density
SPD	Spectral Probability Density

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Passive acoustic monitoring (PAM) is the technology that allows us to collect and analyze the ambient noise data for this thesis. PAM outputs large quantities of data, which necessitates the use of machine learning techniques to automate processing and analysis data. This thesis seeks to gauge the performance of kernel-based methods compared to conventional histogram methods. This performance is analyzed through the scope of a statistical analysis of wind-induced ambient noise. We also seek to analyze the effect the smoothing aspect of kernel density estimators has on spectral features in the power spectral densities (PSDs) and spectral probability densities (SPDs).

This thesis employs several methods to analyze the data, starting with the conventional method of presenting acoustic data, the PSD. This is followed by the slightly more advanced method of the SPD. From the SPD, we can perform a pairwise distance analysis to compare the SPDs contained in different time segments to each other to establish categorizations of ambient noise based on the effect of wind speed at the surface. From this, we can build a classifier based on the mean values of the SPDs to predict whether a particular segment in time is windy based solely on the ambient data.

The results show that there are no significant differences in the obvious spectral features when comparing kernel-based methods with the histogram method. The pairwise distance analysis establishes a clear relationship between wind speed and ambient noise level as well as establishes an optimal frequency range where the ambient noise is dominated by the wind speed. The simple classifier is able to perform well on average for each month, but it also establishes that there is negligible difference in performance between the kernel density estimator and histogram method.

Future works based on this thesis could involve reducing the high resolution present in the PSDs from the fast fourier transform analysis by using large-scale automated analysis on real-time data in order to fully utilize kernel density estimation. On the other hand, while the efforts for this thesis examine the influence of high wind events on acoustic spectral probability density curves, follow-on efforts could lead to similar classifiers to distinguish vessel types or other signals that can impede a sonar system's performance.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

During the process of writing this thesis, I received much support and guidance. I would first like to thank my department chair, Dr. Kevin Smith, for laying the groundwork in helping me understand the acoustics aspect of my thesis and allowing me to tie my very computationally heavy thesis back to something with physical relevance. I would also like to thank my first co-advisor, Dr. Marko Orescanin, for his guidance in understanding the necessary elements of machine learning that I needed to analyze my results. Finally, I would like to thank my second co-advisor, Dr. Paul Leary, for his support in debugging much of my MATLAB code. I received a great deal of assistance over the course of the past year from these three professors, and they really helped me pull together a working thesis for such a quick turnaround.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

1.1 Motivation

Passive acoustic monitoring (PAM) of the ocean is important because it allows us to measure, monitor, and identify sources of sound [1]. The advances in PAM technology have driven the development of autonomous recording units in the form of gliders, drifting platforms, and cabled ocean observatories [2]. The latter, cabled ocean observatories, are composed of seafloor-mounted hydrophones that connect to shore stations via cables to allow straightforward access to power and data acquisition [2].

Developing the technology to advance PAM systems and the methodology to most effectively use it is relevant to the Navy, particularly for undersea warfare (USW) and anti-submarine warfare (ASW). One of the current SONAR programs for submarines is the Acoustic Rapid Commercial Off-the-Shelf Insertion (A-RCI). This open-architecture SONAR system uses legacy sensors and commercial central processors in order to execute the mission of detecting surface vessels, submarines, mines, and other submerged objects while avoiding counter-detection [3]. A-RCI uses the acoustic arrays (such as the spherical and hull arrays) and the towed arrays in order to process acoustic data [3]. Being able to better utilize passive monitoring systems will contribute to the mission of USW and ASW, especially when it comes to contact detection and identification. Improving the methodologies for PAM systems will allow the Navy to monitor signals that may interfere with SONAR systems and identify the sources of acoustic signals.

With advancements in the PAM systems come the ability to gather large volumes of data, and therefore the need to process these vast quantities of data. Currently, the A-RCI systems struggle to operate effectively in high-density surface ship environments, which prompts the need for some form of automation to assist SONAR operators. The rise of machine learning allows us the ability to automate large-scale streams of passive sonar data.

1.2 Problem Statement

A common acoustic metric is the sound pressure level (SPL), which expresses the RMS acoustic pressure amplitude of a particular time window and frequency range, typically on a decibel scale [2]. Acoustic metrics are often expressed as a dB level relative to a reference pressure, which is 1 μPa for underwater acoustics. A common way to express acoustic data is through a power spectral density (PSD), which shows how sound level varies with frequency. Specifically, PSDs describe the signal power density as a function of discrete frequency bands [2]. The percentiles are also used to present acoustic data. Exceedance levels — the percentile levels across frequency spectra — can be used, but a more comprehensive analysis can be found in the spectral probability density (SPD), which presents the empirical probability density of SPLs in each frequency band [2]. To characterize discrete events, a spectrogram can be used, which is a common time-series plot that shows how sound levels vary with time at each frequency.

Density estimation is commonly used to investigate features such as multimodality and skewness in a dataset [4]. Histograms are the most commonly used density estimator, but depending on the method of use, an alternative density estimator may be necessary. Even in the univariate case, changing the origin — even when bin width is held constant — can alter the presentation of the data that may introduce misinterpretations of features that emerge in the data [4]. Kernel density estimation provides a smoother and more efficient output to the conventional histogram method and therefore could prove to be a valid alternative when analyzing underwater noise level statistics. To establish the effectiveness of kernel density estimation compared to its predecessor, these results need to be processed through a performance metric.

1.3 Research Objectives

The primary research objective is to determine whether kernel-based methods perform better than traditional methods of estimating spectral density statistics. In order to achieve this objective, we focus on analyzing specific case studies with a focus on physical noise sources. These could include shipping noise or wind noise as observed by underwater soundscape monitoring via a passive sonar system.

Specifically, in passive sonar observations there is abundant data available for wind speed near the surface, so we can differentiate between windy and non-windy segments of time by analyzing the statistics of underwater noise levels.

The secondary objective is to determine whether different data features become evident using Kernel-based methods. Since kernel-based methods have a smoothing effect, it is possible that periodogram and spectrogram features will be more obvious but noisier.

1.4 Organization

Chapter II introduces the background of passive acoustic monitoring and the analysis of noise sources. It then covers the background on common presentations of acoustic monitoring and lays the groundwork for comparing histogram and kernel-based methods.

Chapter III discusses overall research methodology.

Chapter IV examines the experiment's outcomes and includes metrics, observations, efficiency, and accuracy of histogram and kernel-based methods in analyzing acoustic data.

Chapter V covers a description of the thesis work's achievements and a list of potential future works.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Background

This chapter provides background and a review of the previous work on passive sonar monitoring and analysis of noise sources. Passive acoustic data can be collected and relayed to shore in near real-time due to advances in low-power instrumentation and computational speed [5]. These measurements provide insight into natural sound sources as well as anthropogenic sound sources, shipping in particular [5]. The study of ambient noise is important to underwater acoustics since ocean acoustic signals are disrupted by the dynamics of ambient noise, potentially leading to a low signal-to-noise ratio and reducing the chances of detection [6].

Studies on shipping noise using PAM technology have demonstrated a steady increase in ambient noise levels in the low-frequency ranges over several decades [7]. Ambient noise has increased in the 10-50 Hz range, which is dominated by shipping noise [5]. This is proportional to the increase in shipping activity over the years, effectively providing insight into anthropogenic noise sources. Similar insight can be sought by looking at natural noise sources.

Understanding natural ambient noise can aid in the measurement of the processes producing the noise as well as assess SONAR performance [8]. Given the prevalence of ambient noise induced by wind in frequency ranges up to 50kHz, it has been shown from previous studies that sound spectra can be predicted based on wind speed at the surface using wind speeds from 2 to 14 m/s [8]. The METOC wind data covers a much wider range of wind speeds, from 0 to over 15 m/s. However, the effect wind speed has on ambient noise decreases as frequency increases, with wind having the greatest effect on ambient noise in mid-range frequencies, which is around 500 Hz [9]. This is likely due to the production of bubbles at high wind speeds, which scatters and absorbs sound [10].

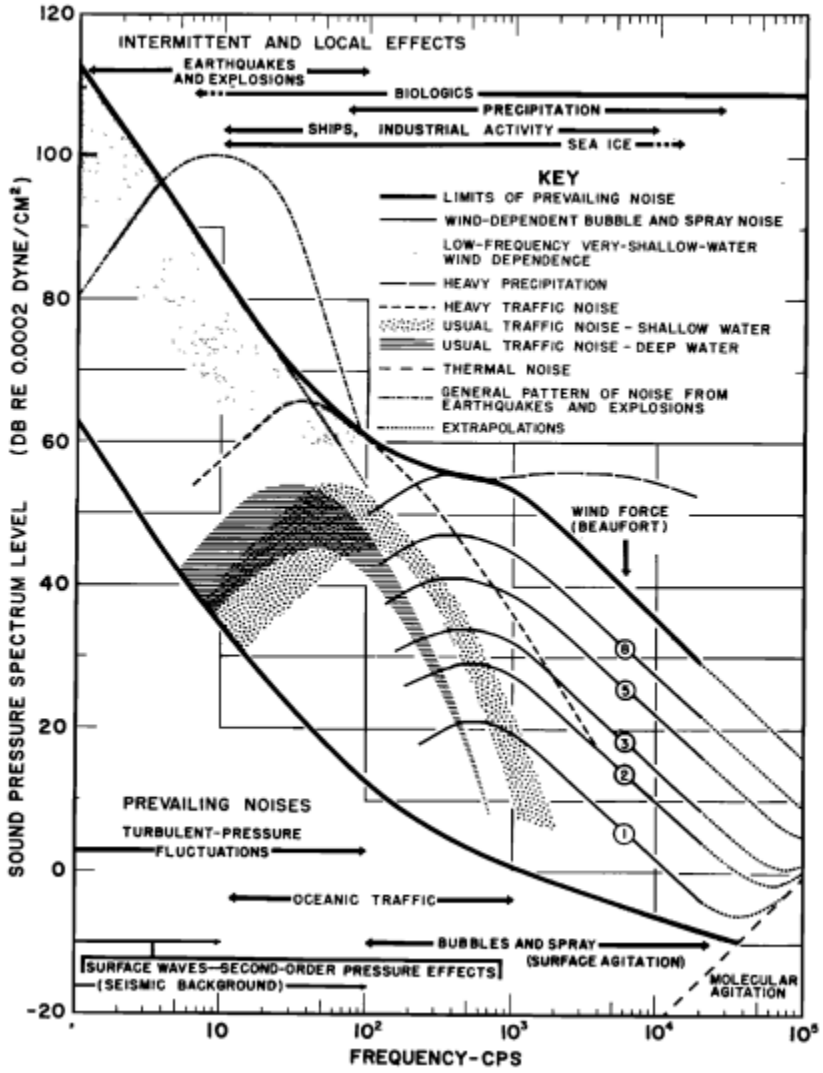


Figure 2.1. From Wenz (1962), this shows how the SPL of ambient noise can be estimated based on knowledge of sea state, shipping activity, wind, etc. Source: [11]

Figure 2.1 shows the Wenz curve, which demonstrates how different noise sources interact to produce particular SPLs. As stated previously, we can see how wind speed correlates with sea state, and that its effect on ambient noise level starts at around 100 Hz and is maximized around 500 Hz, then dropping off as frequency increases. When analyzing ambient noise in the frequency range below 500 Hz, two wind-speed-dependent regions exist: one before the wind speed at which waves begin breaking and one after [12].

For the region after the formation of whitecaps, Kennedy and Goodnow showed that there is a broad maximum at 400 Hz produced by waves breaking due to high winds, and that the generated noise has the characteristic of dipole or doublet radiation [12]. The mentioned methods and experiments in the above section utilized PSDs for evaluations of noises, one of the common presentations of acoustic data. However, PSDs are not the only means of presenting acoustic data.

2.1 Presentation of Acoustic Data

The Power Spectral Density (PSD) is a conventional method used to display acoustic noise when investigating temporal variation, larger datasets from improved passive acoustic monitoring (PAM) systems require more robust analysis techniques [13]. Merchant introduces the idea of spectral probability densities (SPD) in his paper "Spectral Probability Density as a Tool for Ambient Noise Analysis." To retrieve the SPD, a dataset of instantaneous pressure samples is first divided into M segments, where each segment has N samples. Each data segment is given by

$$p^m(n) = \frac{w(n)}{\alpha} p(n + mN) \tag{2.1}$$

where m is the segment, n is the sample, w(n) is a window function, and α is the coherent gain factor of the window function [13]. The discrete Fourier transform (DFT) is then

$$P^m(f) = \sum_{n=0}^{N-1} p^m(n) exp(\frac{-2\pi i f n}{N}) \tag{2.2}$$

where f is the frequency. The DFT is symmetric around the Nyquist frequency for real signals, so the single-sided pressure amplitude spectrum is given by

$$P^m_{ss}(f') = \frac{\sqrt{2}}{N} P^m(f') \tag{2.3}$$

where $1 \leq f' \leq N/2$. The PSD is then

$$PSD^m(f') = \frac{1}{B} |P^m_{ss}(f')|^2 \tag{2.4}$$

where B is the noise power bandwidth of the window function given by

$$B = \frac{1}{N}\sum_{n=0}^{N-1}(\frac{w(n)}{\alpha})^2 \tag{2.5}$$

A periodogram that displays how the power of the initial signal data is distributed with frequency can be constructed as an N/2 by M matrix of the PSDs for each data segment such that

$$PSD(f', m) = 10log_{10}(PSD^m(f')) \tag{2.6}$$

The Spectral Probability Density (SPD) of a frequency bin can be constructed from the histogram of the PSD M-values at frequency f' such that

$$SPD(f') = \frac{1}{Mh}H(PSD(f', m), h) \tag{2.7}$$

where the H function represents the histogram with a bin width h in dB re 1 $\mu Pa^2 Hz^{-1}$ [13].

Merchant shows that SPDs can reveal multimodality and outliers in the data by revealing the underlying noise level distribution. Furthermore, SPD Analysis can preclude the masking of low-amplitude tonals since the full dynamic range is used [13]. The combination of these benefits makes the use of SPDs in statistical analysis preferred when comparing the effectiveness of histogram-based methods and kernel-based methods in this thesis.

2.2 Argument for Kernel Density Estimation

Kernel density estimation provides an alternative technique to the traditional histogram in probability density estimation. A kernel density estimation function is the probability density function (PDF) of a chosen variable. It is defined by

$$\hat{f}_\lambda(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} K(\frac{x - x_i}{\lambda}) \tag{2.8}$$

where $x_1, x_2, ...x_n$ are the samples taken from the given distribution, λ is the bandwidth

(which determines the amount of smoothing for an estimate), and K is a kernel smoothing function, or kernel [14]. A kernel is a wavelength function of a window. Kernel density estimation provides a smoother output compared to histogram methods, and this smoothing effect does not lead to significant information loss [15], [16]. One potential drawback from using kernel density estimation is it will underestimate the tail end of distributions [16]. This thesis then seeks to understand whether the smoother outputs using kernel-based methods outperform the convention of using histogram-based methods.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 3: Methodology

This chapter describes the experiments conducted in this thesis to compare kernel density estimation methods for the evaluation of PSD relative to the classical approach of utilizing histograms. Parameters of kernel methods are introduced. Additionally, dataset and performance metrics for evaluation are described.

3.1 Kernel Parameters

A kernel function needs to satisfy the condition [4]

$$\int_{-\infty}^{\infty} K(x) = 1 \tag{3.1}$$

For this thesis, the normal, box, triangle, and Epanechnikov kernels were used. The normal kernel is expressed by

$$K(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \tag{3.2}$$

where μ is the mean of the distribution and σ^2 is the variance [17]. For the MATLAB function ksdensity, the normal kernel is the default. The box kernel is expressed by

$$K(x) = \begin{cases} 1 & \text{if } -\frac{1}{2} \leq x \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

The triangle kernel is expressed by

$$K(x) = 1 - |x| \tag{3.3}$$

where $|x| \leq 1$ [17]. And the Epanechnikov kernel is expressed as

$$K(x) = \frac{3}{4}(1 - x^2) \tag{3.4}$$

where $|x| \leq 1$ [17]. These kernels are the kernel smoothing functions for the kernel density estimation function referenced in Equation 2.8 from the previous chapter.

3.2 Data Background

The data comes from the Monterey Accelerated Research System (MARS). There are two years of data pre-processed to provide 30-second averages. MARS consists of a 52-kilometer undersea cable that carries data and power to a hub on the seafloor, where a directional hydrophone is attached to one of the nodes at about 891 meters below the surface of Monterey Bay [18]. The project is led by Dr. Kevin B. Smith, chairman and professor of Physics at the Naval Postgraduate School in Monterey, CA. The directional hydrophone samples sound at a rate of 8kHz [18]. An accelerometer-based particle motion sensor is used to determine the direction of fluid current data. For wind data, data was collected from the Station 46042 buoy owned and maintained by the National Data Buoy Center. The buoy is located 27 nautical miles west-northwest of Monterey [19]. The locations of the NOAA buoy and MARS sensor can be seen in Figure 3.1, where the asterisk indicates the location of the NOAA buoy and the "X" marker represents the location of the MARS sensor.

The MARS sensor utilizes a vector sensor system in addition to the pressure sensor system typical of PAM technologies. The output data is organized into four channels, which include one omnidirectional pressure channel and three dipole velocity channels. The data is calibrated through a calibration curve specific to each channel, which limits the operational band of the sensor to 20-1200 Hz. The data is processed in a way that is consistent with Welch's method, which is where a power average is taken over the time window by performing a fast Fourier transform (fft) on 1-second windows to acquire a spectrum with 1 Hz bin width, then taking the power average over the length of time. The spectrograms are produced by taking 1-second samples of the data and applying a 50 percent overlap between samples, which results in spectrogram data in 1 Hz bins every half second.

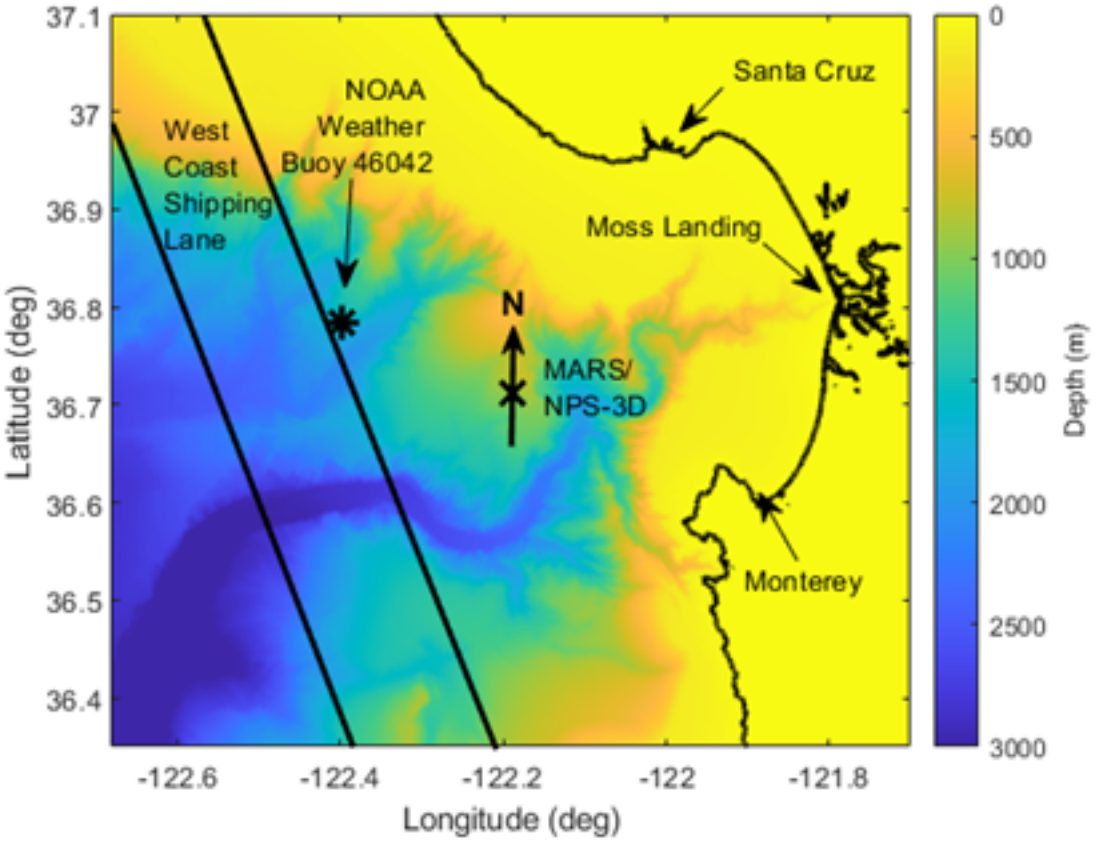


Figure 3.1. Bathymetric Map of the Monterey Bay

3.3 Data Organization and Analysis

The kernel-based method is similar to Welch's method. This uses smoothing to estimate probability density functions on a spectrum. This method samples random 3D vectors (for multivariate data) in kernels (which was previously established to be the wavelength function of a window). The kernel is applied to each frequency bin and analyzed over a time series.

The MATLAB function ksdensity was used to generate the probability density function (PDF) matrices and the spectral probability density (SPD) matrices analyzed here. [kspdf, kscenters] = ksdensity(freqvector, xbins) has the inputs of the RMS pressure data (freqvector) and a specified number of bins (xbins) to output the probability density estimate (kspdf) and the bin centers (kscenters).

3.3.1 Constructing the PDF, SPD, and Percentile Matrices

For the full PDFs and SPDs, a ‘for’ loop created a 2-D matrix for the histogram PDF and each kernel smoothing function PDF for each parameter variation. The ‘for’ loop runs through each frequency vector in the pre-processed pressure data to build each PDF. Within the ‘for’ loop, the code called the ‘ksdensity’ function, which processes the frequency vector from the pressure dataset over an ‘xbins’ argument which spans the range of the pressure dataset. ‘xbins’ is constructed with 761 points, which gives increments of 1/8 dB. The resulting vector from ‘ksdensity’ was then normalized by the bin width. The parameters of the ‘ksdensity’ function were varied to output arrays for the default kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), box kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), triangle kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), and Epanechnikov kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths).

To construct the histogram array, some extra work is needed since the ‘histogram’ function inputs and outputs in terms of bin edges, as opposed to centers (of the bins). Bin edges need to be converted to centers in order to compare the histogram output to the ksdensity output, since xbins is in terms of centers for the latter function. After converting to centers, the ‘histcounts’ function retrieved the histogram PDF, normalized with the ‘probability’ function. This allows the sum of each frequency vector to add up to one, hence verifying integration over the PSD. A nested ’for’ loop constructed the matrices that contain the percentiles of interest (10 percent, 50 percent, and 90 percent) for each frequency for each time increment.

3.3.2 Confirming Smoothness of KSFE

Individual frequency bins were plotted to analyze the power spectral distribution to compare the smoothness of kernel and histogram methods. The points of interest were at 25, 50, 100, and 200 Hz. The frequency vectors associated with 25, 50, 100, and 200 Hz were isolated in the PDFs of interest. Here, that was the histogram PDF, KSFE PDFs with varying kernels, and KSFE PDFs with varying bandwidths. At each frequency, the vectors were plotted for each bandwidth at the default kernel for comparison and the vectors for each kernel at full bandwidth.

3.3.3 Segmented Data Analysis

For this analysis, there is a bug in the code that is inherent in the data. Due to an artifact from the original code used to process the raw data, the time vector data does not add up to 2880 30-second averages for each day. There are 45 seconds missing at the end of every 12 hours from the data that has been provided to me. To get around this, the data was analyzed assuming about one less day per month. So, for February of 2019, only the first 27 days were analyzed (since MATLAB cannot process half days using the following methodology).

A nested 'for' loop created a 3-D array for each PDF I wanted to analyze, where the 3rd dimension of the array is associated with the number of segments. So, the first 'for' loop indexed by time increment and the nested 'for' loop indexed by frequency. Within the nested 'for' loop, the 'ksdensity' function was called, which processes the frequency vector from the pressure dataset over an 'xbins' argument which spans the range of the pressure dataset. 'xbins' was constructed to have 761 points, which gives increments of 1/8 dB.

The resulting vector from 'ksdensity' was then normalized by the bin width. The parameters of the 'ksdensity' function were varied to output array for the default kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), box kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), triangle kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths), and Epanechnikov kernel (at full, 80 percent, 50 percent, and 20 percent bandwidths). The histogram arrays were constructed similarly to the methodology in subsection 3.3.1. This methodology was used to produce matrices segmented both by day and by 3-hour increments.

3.3.4 Pairwise Distance Analysis

From the day-by-day comparison, the MATLAB function pdist2 was used to compare multiple parameters for the pairwise distances for the different combinations of days for kernel and histogram PSDs. 'pdist2' returns the distance between each pair of observations specified in the input. This provides actual statistics to compare probability distributions. To do this, a nested 'for' loop was used to create a matrix where the columns contained the pdist2 values for each day comparison and the rows specified the frequency. An indexing function using the intersect command easily selected day comparisons for analysis.

The distance metrics used in pdist2 are the Euclidean, cosine, correlation, and hamming

metrics. Given a m-by-n matrix where x_s designates the row vectors $x_1, x_2, ...x_m$ and y_t designates the column vectors $y_1, y_2, ...y_n$, the Euclidean distance is defined as

$$d_{euc}^2 = (x_s - y_t)(x_s - y_t)' \tag{3.5}$$

The cosine distance is defined as

$$d_{cos} = 1 - \frac{x_s y_t'}{\sqrt{(x_s x_s')(y_t y_t')}} \tag{3.6}$$

The correlation distance is defined as

$$d_{cor} = 1 - \frac{(x_s - \bar{x_s})(y_t - \bar{y_t})'}{\sqrt{(x_s - \bar{x_s})(x_s - \bar{x_s})'}\sqrt{(y_t - \bar{y_t})(y_t - \bar{y_t})'}} \tag{3.7}$$

The hamming distance is defined as

$$d_{ham} = \frac{\#(x_{sj} \neq y_{tj})}{n} \tag{3.8}$$

where 'j' is an index that denotes the location in the row and column vectors.

Histogram plots of all pdist2 comparisons were constructed to compare the distributions for comparing windy vs. windy days, windy vs. non-windy days, and non-windy vs. non-windy days under the different parameters (Euclidean, Cosine, Correlation, and Hamming). The mean values for each distribution across each frequency were taken to plot the mean pdist2 values for each type of comparison (windy vs. windy, etc.). From this, a couple of frequency ranges were identified that are likely more effective than the full frequency range, and the histograms were re-plotted with these specific frequency ranges to compare the distributions. From the re-plotted histograms, it became evident that the 200-400 Hz range is optimal for analysis.

3.4 Experiment: Classification of Windy vs. Non-Windy Noise using SPD Percentiles

This is the main performance metric to compare the two methods, being the capability to extract wind events from spectral densities. This is also where the physical relevance of the statistical analysis lies. Using the METOC data from the buoy, each 3-hour segment was established as either windy or not windy by hand. To do this, the average wind speed was calculated in meters per second and if the average was above 9.5 meters per second, the segment was considered windy. The 9.5 meters per second comes from the average wind speed necessary for waves to start breaking on the surface according to the Beaufort wind scale. Assigning 1 as windy and 0 as non-windy, a vector was constructed that defines the true class labels. A custom function called 'classreport' output the precision, recall, F-value, and accuracy given a true class label vector and a classifier prediction vector.

To determine the best threshold based on the 50th percentile SPDs, a vector of the average SPDs across all frequencies for each 3-hour segment was constructed. This vector was run through a nested for-loop that stepped through threshold values from the lowest dB value to the highest dB value (in the average SPD values) for each 3-hour segment. If an SPD value was below the threshold value, it was predicted to be non-windy and assigned a value of 0 in a new vector. Otherwise, it was predicted to be windy and assigned a value of 1. Now having a true class label and various classifier prediction labels, the classreport function was called to output precision, recall, F-value, and accuracy vectors where each element is comparing the true class label to each classifier prediction label. Accuracy is defined as [20]–[22]

$$accuracy = \frac{\text{TP} + \text{TN}}{\text{total \# of positives and negatives}} \tag{3.9}$$

where TP is true positives and TN is true negatives. Similarly, precision, recall, and F-value are defined as

$$precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.10}$$

$$recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.11}$$

$$F - value = \frac{2*\text{precision}*\text{recall}}{\text{precision} + \text{recall}} \tag{3.12}$$

where FN is false negatives and FP is false positives. From this, the threshold that maximized accuracy, precision, and recall was identified. This methodology was repeated for each month by only taking the average 50th percentile SPD data for the optimal frequency range identified from the mean pdist2 value plots.

CHAPTER 4:
Results

In this section, most important results of this thesis are presented.

4.1 KSFE and Histogram PDFs

From Figure 4.1, it is evident that the kernel smoothing function estimate produces smoother outputs than the histogram output while maintaining major features in the PDF. Scattered noise at the edges of the histogram PDF were eradicated by the smoothing inherent in the KSFE. Varying the type of kernel for the KSFE presents essentially no difference by simply looking at the PDFs produced in Figure 4.2. Varying the bandwidth, as seen in Figure 4.3, shows that the KSFE PDF starts to more closely resemble the histogram PDF as the bandwidth decreases, as expected.

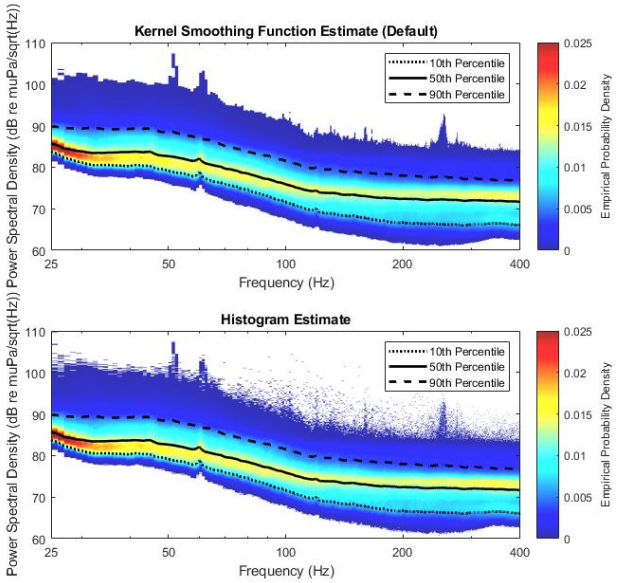


Figure 4.1. The KSFE PDF represented here is for the full bandwidth and the default (normal) kernel. Both the KSFE PDF and the histogram PDF here are for the entire month of February 2019.

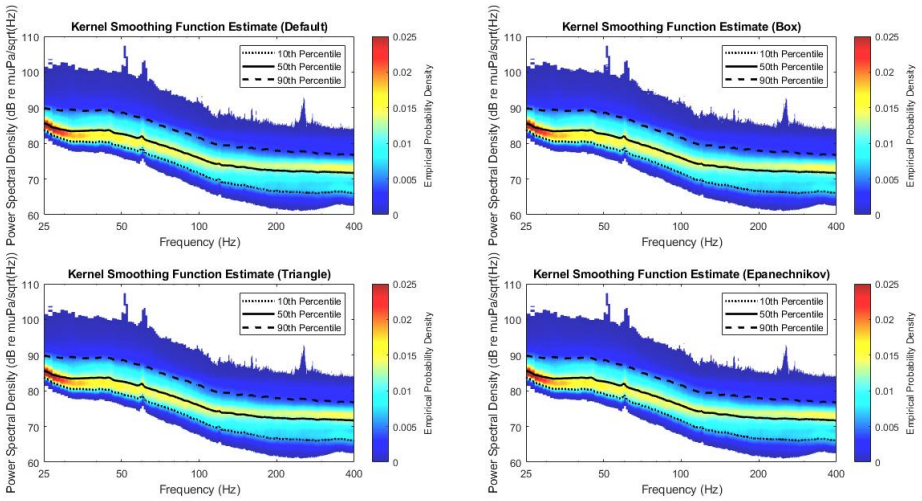


Figure 4.2. Comparison of Kernel Parameters for KSFE for February 2019

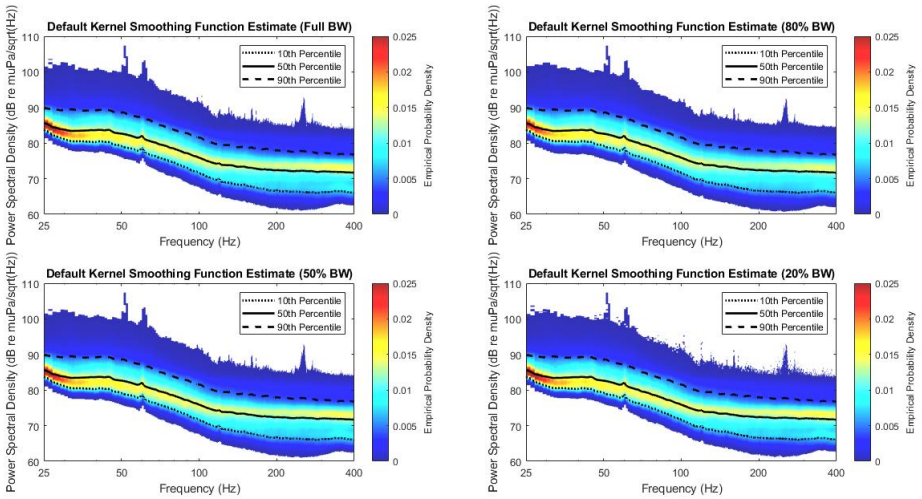


Figure 4.3. Comparison of Bandwidth Variation for KSFE for February 2019

Figure 4.4 shows the 10th, 50th, and 90th percentile SPDs from January of 2020. The 90th percentile plot picks out the very high wind events, demonstrated by how the noise levels are much higher in the 90th percentile.

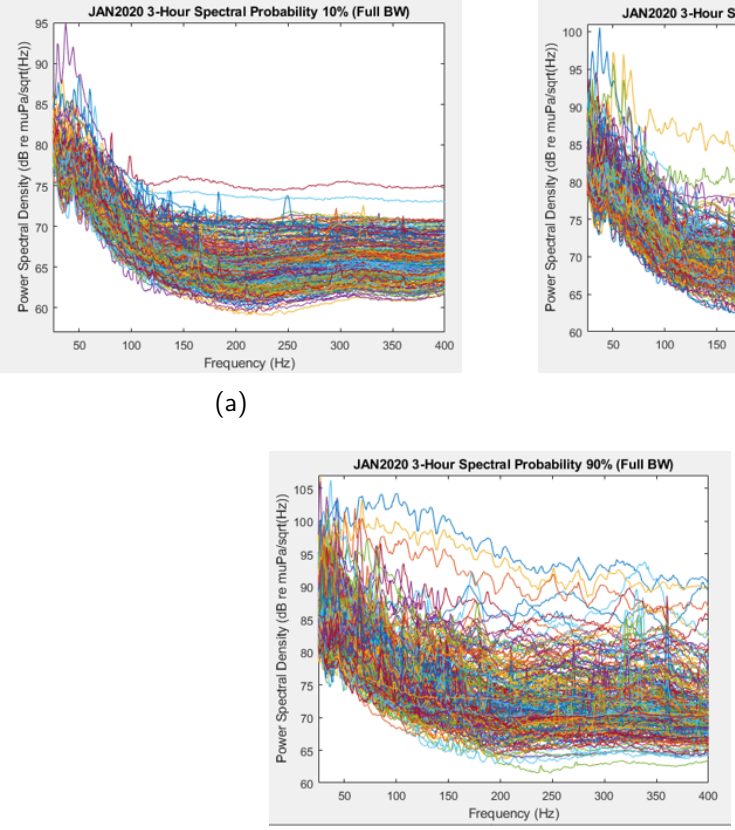


Figure 4.4. 3-Hour SPDs for JAN 2020 for the 10th (a), 50th (b), and 90th (c) Percentiles

4.2 Smoothness of KSFE

Taking 1 Hz slices of the PSD reveals how closely the various kernels approximate the histogram PSD in Figure 4.5. The box kernel is closest to the histogram method but all of the kernels are within a small margin of each other. As observed from Figure 4.5, the 20 percent bandwidth does most closely match the histogram output which was predicted from observing the PSDs.

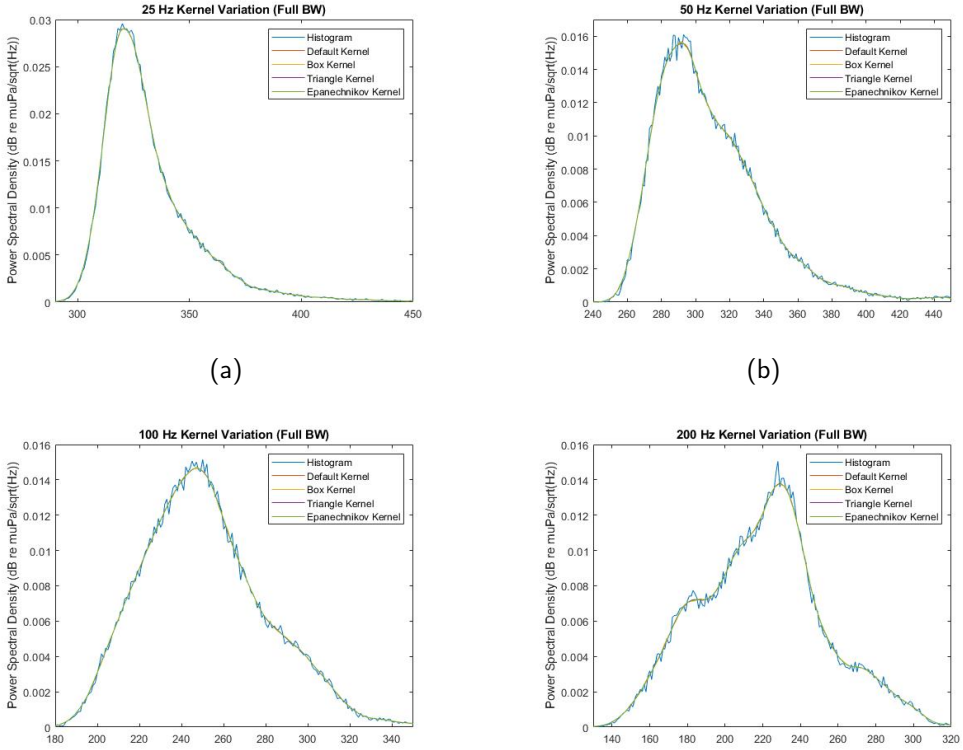


Figure 4.5. Kernel Variation Comparison for 1 Hz Slices

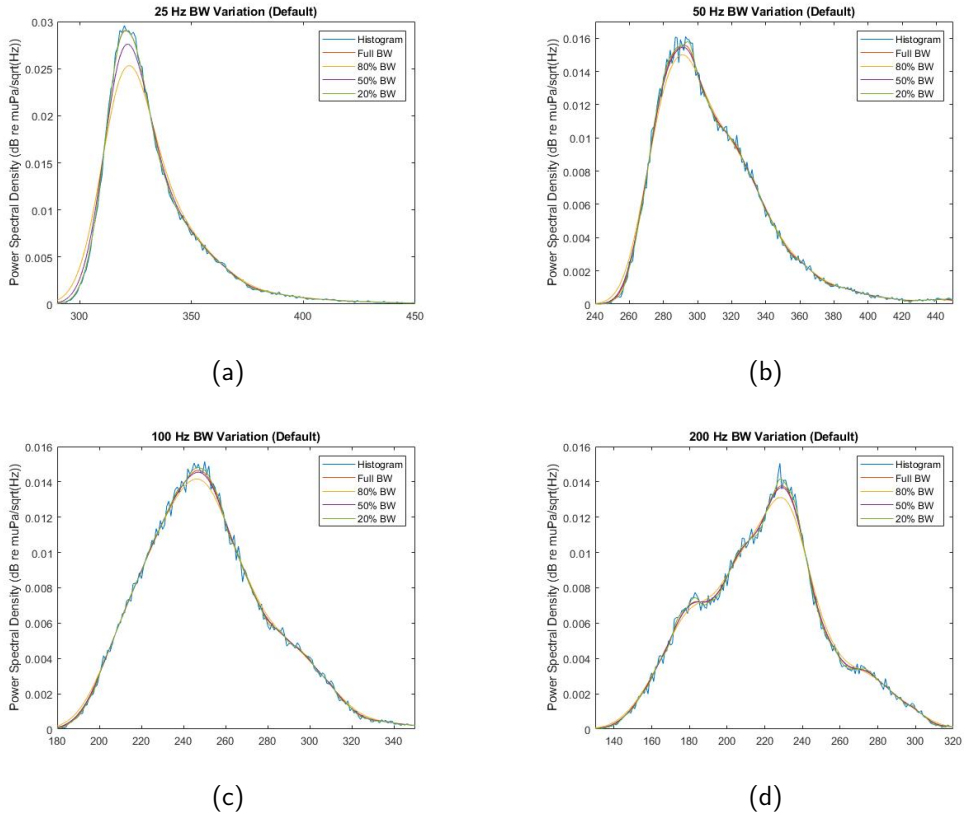


Figure 4.6. Bandwidth Variation Comparison for 1 Hz Slices

4.3 Pairwise Distance Analysis

From Figure 4.7, there are distinct differences in the windy vs non-windy distributions and the windy vs windy and non-windy vs non-windy distributions for the Euclidean, cosine, and correlation distance metrics, which indicates that there should be a threshold to predict whether a day is windy or non-windy from the SPDs. The distributions suggest that the cosine and correlation distance metrics are best for picking out rare events, but the Euclidean distance metric is useful if more sensitivity is necessary. Figure 4.8 shows that there may be an optimal frequency range in either the 200-400 Hz range or the 600-1200 Hz range. This is consistent with our understanding of how wind speed at the surface affects underwater noise levels. Noise induced by wind is prominent in the high sonic band, particularly in the 200-800 Hz range [23]. This figure also shows that there is a constant, loud tonal at about

500 Hz, which makes it difficult to analyze the data around this frequency range. This tonal is caused by a glitch in the system with an unknown source, so it is best to not process the data in the 200-400 Hz range.

As seen in Figure 4.9a, the 200-400 Hz frequency range has the most separated distributions and is therefore the optimal frequency range for analyzing noise induced by wind. While Figure 4.4 does show that there is significant noise below 200 and even 100 Hz during high wind events, those distributions are less separated from low-wind time segments compared to the 200-400 Hz range and the 600-1200 Hz range.

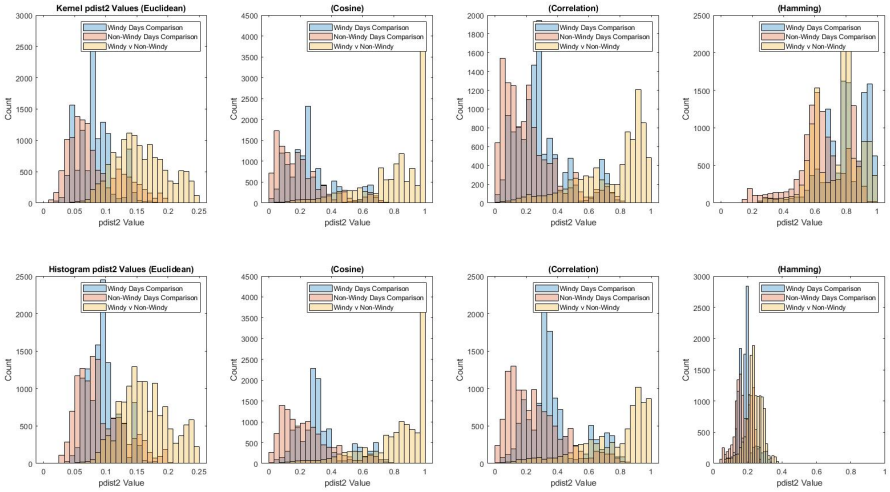


Figure 4.7. Pairwise Distance Analysis Comparison Across Entire Frequency Range

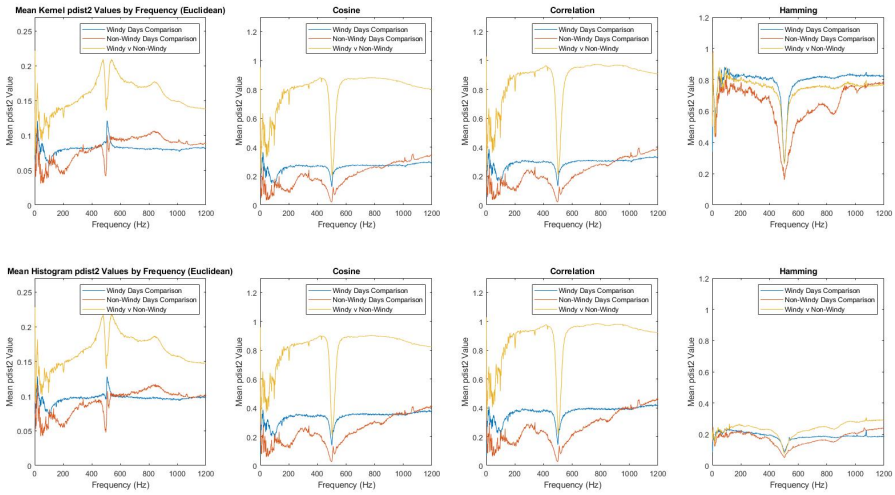
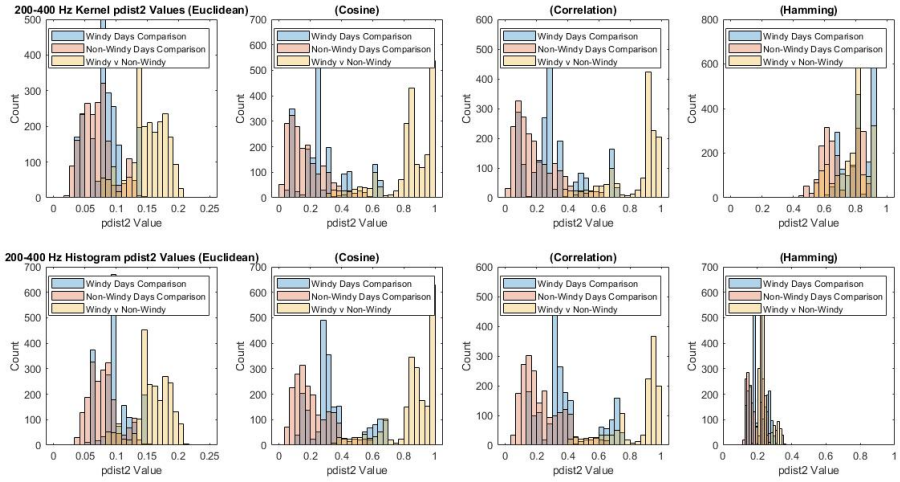
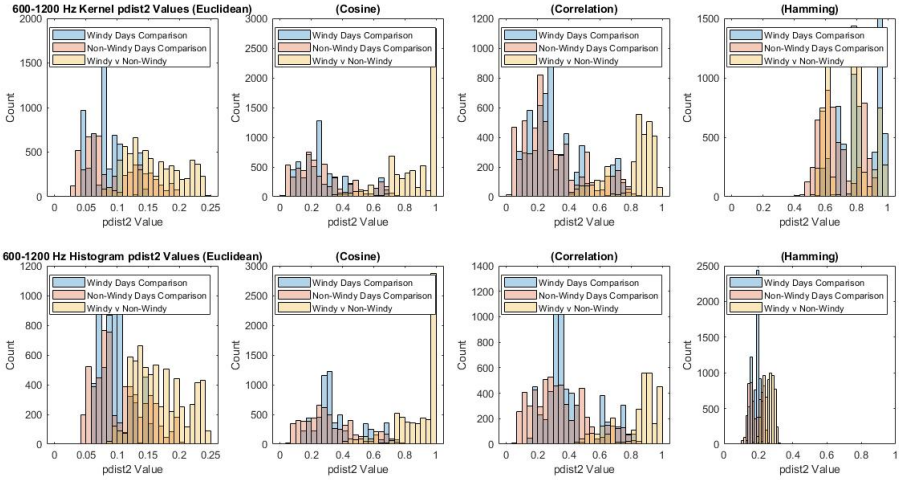


Figure 4.8. Mean Pairwise Distance Comparisons by Frequency



(a)



(b)

Figure 4.9. Pairwise Distance Analysis Comparison for 200-400 Hz Band (a) and 600-1200 Hz Band (b)

4.4 Establishing a Threshold

The accuracy in predicting windiness is maximized where precision and recall are optimized. This is consistent for both histogram and kernel methods. The average values across the 2019 data can be found in Table 4.1, the average values across the 2020 data can be found

in Table 4.2, and the average values across all the data can be found in Table 4.3. The ideal threshold to differentiate between windy and non-windy segments of time remained constant at around 64 $\mu Pa/\sqrt{Hz}$ except for the winter months for both 2019 and 2020, where it increased to 67 or 68 $\mu Pa/\sqrt{Hz}$. This could be due to changes in the sound speed profile in the cooler weather, thus affecting how sound propagates in the water.

Changing the frequency range to 200-400 Hz did improve the machine learning statistics across the board. Optimizing the frequency range also lowers the ideal threshold value. The optimal values for accuracy, precision, recall, and F-value are displayed in Tables 4.4 and 4.5 for the February 2019 data.

Machine Learning Statistic	Histogram	Kernel
Ideal Threshold	65 dB re $\mu Pa/\sqrt{Hz}$	65 $\mu Pa/\sqrt{Hz}$
Accuracy	0.6267	0.6267
Precision	0.6281	0.6281
Recall	0.9928	0.9928
F-Value	0.7677	0.7677

Table 4.1. Histogram vs. Kernel SPD Analysis Statistics for 200-400 Hz Frequency Range for 2019

Machine Learning Statistic	Histogram	Kernel
Ideal Threshold	65 dB re $\mu Pa/\sqrt{Hz}$	65 $\mu Pa/\sqrt{Hz}$
Accuracy	0.6345	0.6345
Precision	0.6362	0.6362
Recall	0.9947	0.9947
F-Value	0.7743	0.7743

Table 4.2. Histogram vs. Kernel SPD Analysis Statistics for 200-400 Hz Frequency Range for 2020

Machine Learning Statistic	Histogram	Kernel
Ideal Threshold	65 dB re $\mu Pa/\sqrt{Hz}$	65 $\mu Pa/\sqrt{Hz}$
Accuracy	0.6302	0.6302
Precision	0.6317	0.6317
Recall	0.9936	0.9936
F-Value	0.7706	0.7706

Table 4.3. Histogram vs. Kernel SPD Analysis Statistics for 200-400 Hz Frequency Range for Entire Dataset

Machine Learning Statistic	Histogram	Kernel
Ideal Threshold	72 dB re $\mu Pa/\sqrt{Hz}$	72 $\mu Pa/\sqrt{Hz}$
Accuracy	0.8194	0.8194
Precision	0.7652	0.7652
Recall	0.88	0.88
F-Value	0.8186	0.8186

Table 4.4. Histogram vs. Kernel SPD Analysis Statistics for Entire Frequency Range for FEB 2019

Machine Learning Statistic	Histogram	Kernel
Ideal Threshold	69 dB re $\mu Pa/\sqrt{Hz}$	69 dB re $\mu Pa/\sqrt{Hz}$
Accuracy	0.8472	0.8472
Precision	0.8018	0.8018
Recall	0.89	0.89
F-Value	0.8436	0.8436

Table 4.5. Histogram vs. Kernel SPD Analysis Statistics for 200-400 Hz for February 2019

It is evident from the statistics that using the kernel-based method does not improve pre-

diction accuracy compare the conventional histogram-based method. Figures 4.10 through 4.13 display the confusion matrices for the 50th percentile SPDs for February 2019. The upper left quadrant represents the number of true positives generated by the classifier, the lower left quadrant represents the false negatives, the upper right quadrant represents the false positives, and the lower right quadrant represents the true negatives. These confusion matrices demonstrate that narrowing the frequency band to the 200-400 Hz range decreased the number of false positives and false negatives.

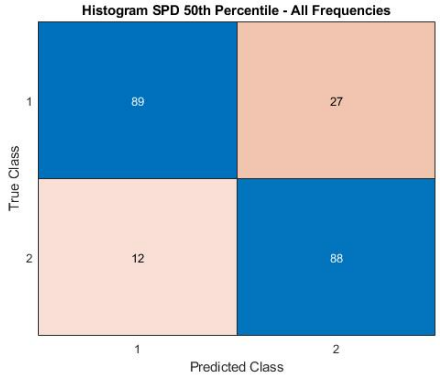


Figure 4.10. Entire Frequency Range

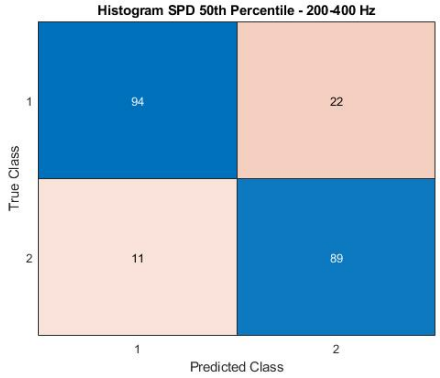


Figure 4.11. 200-400 Hz Band Only

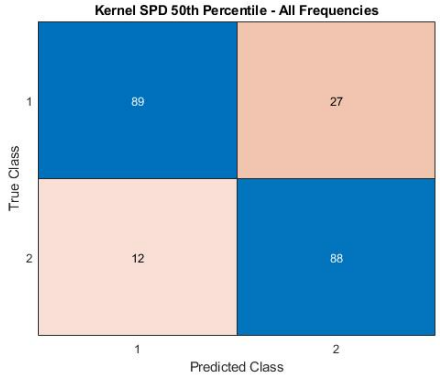


Figure 4.12. Entire Frequency Range

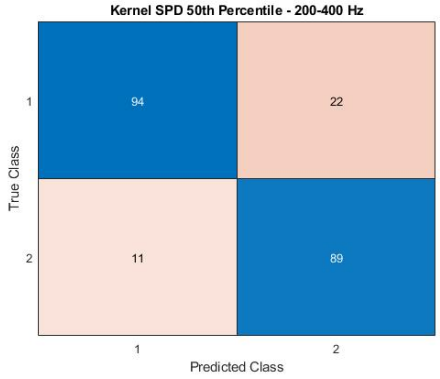


Figure 4.13. 200-400 Hz Band Only

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
Conclusion and Future Work

5.1 Conclusions

This thesis work set out to demonstrate the performance of kernel-based methods with reference to histogram methods. This performance was analyzed through the scope of a statistical analysis of physical noise sources, specifically wind speed at the surface of the water. We also sought to analyze the effect the smoothing aspect of kernel density estimators had on spectral features.

We demonstrated the ability to apply kernel smoothing estimates to produce smoother PSDs and SPDs as well as the ability to apply pairwise distance analyses to reveal categories in the ambient noise based solely on wind speed. There were no specific spectral features revealed through the kernel density estimates that were not already obvious through the histogram method. The pairwise distance analysis provided an optimal frequency range to capture ambient noise induced by wind.

Additionally, we demonstrate a simple classifier that is able to categorize ambient noise data as being associated with high wind speed or low wind speed. The proposed approach achieves a performance optimized in the 200-400 Hz range, where the statistics for the entire data-set for accuracy, precision, recall, and F-value were 0.6302, 0.6317, 0.9936, and 0.7706, respectively, with a windy vs non-windy threshold value set at 65 dB re 1 μPa/$\sqrt{Hz}$. Additionally, we found that for the chosen time duration of analysis, the window classification model had similar performance with either kernel based or histogram-based input features.

5.2 Future Work

While the efforts for this thesis examined the influence of high wind events on acoustic spectral probability density curves, follow-on efforts could lead to similar classifiers to distinguish vessel types or other signals that can impede a sonar system's performance. Since vessels not only of varying type but also of varying country of origin have potentially

distinct acoustic signatures, this route in particular could prove useful to the Navy.

The challenge with comparing kernel-based methods and histogram methods here is that the PSD resolution through the fft analysis was very high, making it so the kernel-based methods could not fully demonstrate the advantage that its smoothing capabilities provide. Future work could reduce this resolution by using large-scale automated analysis on real-time data in order to fully utilize kernel density estimation.

This thesis focused on the analysis of the omnidirectional pressure data, but future work could also implement the vector sensor channels of the MARS data. This would add a directional aspect to the analysis, which is relevant to the Navy since the ability to track contacts and detect mines necessitates a directional aspect from the data.

5.3 Final Thoughts

The pairwise distance analysis provided some of the more interesting results in this thesis. This analysis demonstrated a distinct difference between days with similar noise level characteristics and different noise level characteristics. Since days were classified by the level of wind speed at the surface, it is safe to say that noise induced by wind is the dominant ambient noise source, especially for the 200-400 Hz range. The cosine and correlation distance metrics proved to be most useful in identifying rare events. On the other hand, the Euclidean distance metric is more sensitive to smaller or lower velocity wind events.

While building classifiers proved to be moderately successful, it did prove that there was no difference between kernel and histogram methods when classifying windy and non-windy segments of time. This may be due to the volume of data available for this research such that the smoothing effect of kernel density estimation didn't have the desired impact. There is very high resolution in the PSDs with the fft analysis. These results may be different with sparse data, but underwater acoustics data is seldom sparse. Having high resolution is desirable but has the potential to be computationally expensive. There could be an advantage to utilizing large-scale automated analysis with real-time data in order to reduce fft resolution. Then kernel density estimation could be used efficiently here in a way that histogram methods could not.

List of References

[1] "Passive Acoustic Monitoring in California's National Marine Sanctuaries," *Passive Acoustic Monitoring | Office of National Marine Sanctuaries*. Available: https://sanctuaries.noaa.gov/education/teachers/passive-acoustic-monitoring.html#:~:text=Passive20acoustic20monitoring20(PAM)20is,wind2C20earthquakes)2C20and20human.

[2] N. D. Merchant, K. M. Fristrup, M. P. Johnson, P. L. Tyack, M. J. Witt, P. Blondel, and S. E. Parks, "Measuring acoustic habitats," *Methods in Ecology and Evolution*, vol. 6, no. 3, pp. 257–265, 2015.

[3] R. Behler, "Director, operational test and evaluation FY 2017 annual report," *The Office of the Director, Operational Test and Evaluation*, 2018. Available: http://www.dote.osd.mil/pub/reports/FY2017.

[4] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. CRC Press, Boca Raton, FL, USA, 2018. [Online]. Available: https://ned.ipac.caltech.edu/level5/March02/Silverman/paper.pdf.

[5] M. F. Baumgartner, K. M. Stafford, and G. Latha, "Near real-time underwater passive acoustic monitoring of natural and anthropogenic sounds," *Observing the Oceans in Real Time*, pp. 203–226, 2018.

[6] S. Siddagangaiah, Y. Li, X. Guo, and K. Yang, "On the dynamics of ocean ambient noise: Two decades later," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 10, pp. 103–117, 2015.

[7] M. A. McDonald, J. A. Hildebrand, and S. M. Wiggins, "Increases in deep ocean ambient noise in the Northeast Pacific West of San Nicolas Island, California," *The Journal of the Acoustical Society of America*, vol. 120, no. 2, pp. 711–718, 2006.

[8] B. B. Ma, J. A. Nystuen, and R.-C. Lien, "Prediction of underwater sound levels from rain and wind," *The Journal of the Acoustical Society of America*, vol. 117, no. 6, pp. 3555–3565, 2005.

[9] S. Ramji, G. Latha, V. Rajendran, and S. Ramakrishnan, "Wind dependence of ambient noise in shallow water of Bay of Bengal," *Applied acoustics*, vol. 69, no. 12, pp. 1294–1298, 2008.

[10] D. M. Farmer and D. D. Lemon, "The influence of bubbles on ambient noise in the ocean at high wind speeds," *Journal of Physical Oceanography*, vol. 14, no. 11, pp. 1762–1778, 1984.

[11] G. M. Wenz, “Acoustic ambient noise in the ocean: Spectra and sources,” *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1936–1956, 1962.

[12] W. M. Carey and R. B. Evans, *Ocean Ambient Noise: Measurement and Theory*. Springer Science & Business Media, 2011.

[13] N. D. Merchant, T. R. Barton, P. M. Thompson, E. Pirotta, D. T. Dakin, and J. Dorocicz, “Spectral probability density as a tool for ambient noise analysis,” *The Journal of the Acoustical Society of America*, vol. 133, no. 4, pp. 262–267, 2013. [Online]. doi: https://doi.org/10.1121/1.4794934.

[14] T. Hastie, J. Friedman, and R. Tisbshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer Press, New York City, NY, USA, 2017.[Online]. Available: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf

[15] P. Emanuele, R. Rudy, and S. Stefano, “Enhanced multichannel histogram equalization for speech recognition in noisy acoustic conditions,” *Frontiers in Artificial Intelligence and Applications*, vol. 234, p. 149–161, 2011. [Online]. doi: https://doi.org/10.3233/978-1-60750-972-1-149. Available: https://doi.org/10.3233/978-1-60750-972-1-149.

[16] P. K. Sahoo, A. A. Farag, and Y.-P. Yeap, “Threshold selection based on histogram modeling,” in *[Proceedings] 1992 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 1992, pp. 351–356.

[17] A. Azzalini and A. W. Bowman, “Applied smoothing techniques for data analysis: The kernel approach with s-plus illustrations,” *Oxford Statistical Science Series*, 2000. [Online]. doi: https://doi.org/10.1007/s001800000033.

[18] Monterey Bay Aquarium Research Institute. "Monterey Accelerated Research System (MARS) Cabled Observatory," Feb. 3, 2020. [Online]. Available: https://www.mbari.org/at-sea/cabled-observatory.

[19] US Department of Commerce, National Oceanic and Atmospheric Administration. "NDBC Station Page," Apr. 1, 2021. [Online]. Available: https://www.ndbc.noaa.gov/station_page.php?station=46042.

[20] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[21] A. P. Bradley, “The use of the area under the ROC curve in the evaluation of machine learning algorithms,” *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[22] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PloS One*, vol. 10, no. 3, pp. 118–432, 2015.

[23] R. Urick and W. Kuperman, "Ambient noise in the sea," 1989. Available: https://apps.dtic.mil/sti/pdfs/ADA460546.pdf

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California