University of Alberta

Modeling Zooplankton Diel Vertical Migration Patterns Based on Curve Fitting and Feature Correlation Analysis

by

Shuang Zhao

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2010
Edmonton, Alberta

Examining Committee

Jörg Sander, Computing Science

Osmar R. Zaïane, Computing Science

Sally Leys, Biological Sciences

Abstract

The goal of this thesis is to study and model the Diel Vertical Migration (DVM) pattern using machine learning methods. We choose an Almost Periodic Function as the mathematical model and fit the monthly averaged migration data into a 5-term Fourier series whose coefficients and frequency are functions of time. The resulting function captures the general characteristics of the DVM pattern whose period is similar yet undergoes gradual changes over time. Further correlation analyses show that the monthly averaged distribution of zooplankton and various environmental factors are strongly correlated. Therefore, we adjust the function so that the coefficients and frequency are functions of environmental factors. Besides, we also examine the pattern on finer time scales using classification algorithms. We build classifiers which predict zooplankton existence at different depths based on a set of environmental measurements. Experiments demonstrate that both of the above methods are valid in modeling the DVM pattern.

Acknowledgements

Many thanks to my supervisor Dr. Jörg Sander for his time, encouragement and support throughout this research study. Special thanks to Richard Dewey from University of Victoria for providing biological background information for the preprocessing of the data.

Table of Contents

1 Introduction **1**
- 1.1 Problem Definition . . . 2
- 1.2 Diel Vertical Migration . . . 3
- 1.3 Scope of the Thesis . . . 5
- 1.4 Thesis Outline . . . 7

2 Related Work and Methods **8**
- 2.1 Quantifying the Migration of Zooplankton . . . 9
- 2.2 Environmental Factors Related to the DVM Pattern . . . 12
- 2.3 Clustering . . . 15
- 2.4 Curve Fitting . . . 16
- 2.5 Almost Periodic Function . . . 18
- 2.6 Feature Selection . . . 19
- 2.7 Supervised Learning . . . 21

3 Modeling Zooplankton Diel Vertical Migration Pattern on a Large Time Granularity **23**
- 3.1 Data Pre-processing . . . 24
- 3.2 Diel Vertical Migration Path Modeling . . . 32
- 3.3 Final models for zooplankton DVM pattern . . . 39

4 Modeling Zooplankton Diel Vertical Migration Pattern on Finer Time Granularities **42**
- 4.1 Data Pre-processing . . . 43
- 4.2 Feature Selection and Classification on 60min Averaged Data . . . 46
- 4.3 Feature Selection and Classification on Finer Time Scales . . . 53
- 4.4 Feature selection on other depths . . . 57
- 4.5 Predicting zooplankton migration pictures . . . 59
- 4.6 Comparing Classification Pictures with Almost Periodic Curves . . . 60

5 Conclusion **65**
- 5.1 Contribution . . . 66
- 5.2 Future work . . . 67
  - 5.2.1 Calibration . . . 67
  - 5.2.2 Integrating migration path models . . . 67
  - 5.2.3 Interaction between fish schools and zooplankton . . . 67
  - 5.2.4 Reasoning the environmental impact . . . 67

Bibliography **68**

Appendices **72**

List of Tables

1.1 Technical specifications of ZAP [2] 5

3.1 Root mean squared errors for upper bound, lower bound and the middle depth of the migration path using different Fourier series models 37

4.1 Features and sources 44
4.2 Search methods 45
4.3 Evaluation methods 46
4.4 Top 10 features for 60 minutes averaged data 47
4.5 Top 10 features for 15 minutes averaged data 53
4.6 Feature rankings after a random feature is added 54
4.7 Top 10 features for 5 minutes averaged data 54
4.8 Top 10 features for 30 minutes averaged data 55
4.9 Feature rankings at depth 66m (depth index 550) 57
4.10 Feature rankings for depth 54m (depth index 450) 58
4.11 Feature rankings for depth 30m (depth index 250) 58

List of Figures

1.1 The ZAP transducer and pressure case [2] . . . 4
1.2 Daily 30 second averaged plot of Acoustic Backscatter Intensity. The intensity range for the zooplankton is approximately between 5 to 15 db [2] . . . 5
1.3 ADCP prior to deployment [2] . . . 6
1.4 Example of an ADCP profiler, April 2006, Saanich Inlet [2] . . . 6

2.1 Temporal derivative of acoustic backscatter at 5m depth (upper panel) and time series of acoustic backscatter strength in Lake Hallwil from 26 June to 2 July 2001 at 5 and 35m depth (lower panel) [34] . . . 11
2.2 Acoustic backscatter at the deep and shallow sites in Patricia Bay together with the sunrise and sunset times. Time is UTC. The black line indicates sunset and the red line indicates sunrise [2] . . . 13
2.3 Zooplankton abundance and changes of various environmental factors for a typical day in Lake Hallwil [34] . . . 14
2.4 Key steps in feature selection [33] . . . 20
2.5 Whole search space [9], n=4 . . . 21

3.1 Plots of ZAP intensities in Saanich Inlet, British Columbia. Each picture illustrates the migration of zooplankton for one day. The time used is UTC time, 8 hours before the local time (PST time) . . . 25
3.2 Raw data plot of April 23,2008 . . . 27
3.3 Clustering result.Epsilon=5,minObj=8,Manhattan distance . . . 28
3.4 Raw data plot of Jul01,2008 . . . 28
3.5 Clustering result.Epsilon=5,minObj=8,Manhattan distance . . . 29
3.6 Manually labeled migration path, November 16, 2008 . . . 30
3.7 Automatically generated migration path, November 10, 2008 . . . 31
3.8 The middle depth of zooplankton migration path from January 2008 to July 2009. The blue curve is used as the training set. The pink curve is used as the test set. . . . 31
3.9 The upper and lower bounds of zooplankton migration path from January 2008 to July 2009. The red and blue curves are training set for upper and lower bounds respectively. The green and black curves are test set for upper and lower bounds respectively. . . . 32
3.10 Use Fourier series to fit each period individually . . . 33
3.11 1-term Fourier series . . . 37
3.12 2-term Fourier series . . . 37
3.13 4-term Fourier series . . . 38
3.14 8-term Fourier series . . . 38
3.15 Upper bound . . . 39
3.16 Lower bound . . . 40
3.17 Middle depth . . . 40

4.1 Raw data plot for July 2nd, 2008 . . . 49
4.2 Plot of intensity values at depth 42m . . . 49
4.3 Plot of 60 minutes averaged intensity values at depth 42m . . . 49
4.4 True labels at depth 42m . . . 50
4.5 Predicted labels at depth 42m . . . 50
4.6 Raw data plot for April 9th, 2008 . . . 51
4.7 Raw data plot for January 17th, 2008 . . . 52
4.8 Plot of intensity values for depth 42m . . . 52
4.9 Plot of intensity values for depth 42m . . . 52
4.10 Part of the decision tree for 30min averaged data . . . 56

4.11 Plots of predicted and true class labels for July 26, 2009 60
4.12 The predicted picture and the predicted upper and lower bounds for April, 2009. The black curve represents the predicted upper bound and the blue curve represents the lower bound. The predicted existence of zooplankton is illustrated in green . . 61
4.13 Predicted upper and lower bounds and the true data 61
4.14 Classification results together with the true data. The pink dots illustrate the errors . 62
4.15 Plots of predicted class labels and the predicted curves. The black curve represents the predicted upper bound and the blue curve represents the lower bound. The predicted existence of zooplankton is illustrated in green 63
4.16 Plots of true class labels and the predicted curves 63
4.17 Difference between the predicted and true labels. The pink dots illustrate errors . . 64

1 Middle-depth migration path from January 2008 to July 2009 73
2 Upper bound of the migration path from January 2008 to July 2009 74
3 Lower bound of the migration path from January 2008 to July 2009 75
4 Fitted coefficient and frequency values if an 8-term Fourier series is used as the model 76
5 Monthly averaged environmental measurements 77
6 Linear correlation analysis using Pearson's correlation function 78
7 Information gain between environmental factors and coefficients 79
8 Part of the dataset for feature selection 85
9 Feature selection on 60 minutes averaged data 86
10 Feature selection on 60 minutes averaged data (continued) 87
11 Feature selection on 15 minutes averaged data 88
12 Feature selection on 15 minutes averaged data (continued) 89
13 Feature selection on 5 minutes averaged data 90
14 Feature selection on 30 minutes averaged data 91
15 Predicted and true labels for April 2, 2009 91
16 Predicted and true labels for April 22, 2009 92
17 Predicted and true labels for April 30, 2009 92
18 Predicted and true labels for July 8, 2009 92
19 Predicted and true labels for July 30, 2009 93

Chapter 1

Introduction

Plankton are minute pelagic organisms that drift with the current in a sea or lake. The animal components of the plankton are called zooplankton [35]. Many zooplankton are holoplanktonic, spending their entire life cycles within the plankton, while others are meroplanktonic, spending only part of their life cycles in the plankton, then either metamorphosing into the nekton or settling on the sea floor [36].

Zooplankton feed on phytoplankton, which are the plant components of the plankton [45]. Because of their small sizes, zooplankton can respond relatively quickly to changes of phytoplankton abundance. On the other hand, zooplankton themselves also serve as an important food source. Herrings, for example, treat zooplankton as one of their major preys [42]. Because of their important roles in the oceanic food web, studying the behaviour of zooplankton can serve as an essential method for a more complete understanding of the functioning of marine ecosystems.

Many types of zooplankton are found to undertake a diel vertical migration (DVM) pattern. They travel to the surface at night, and migrate down the water column during the daytime. Many scientists have tried to explain zooplankton's migration pattern. Among all different explanations, the most famous ones are avoidance of predators, effect of availability of nutrients and response to environmental changes (for example, light and oxygen) [41].

In this thesis, we use machine learning methods to study and model the DVM pattern of zooplankton. The models we build reflect the regularities of the migration and can be used to make predictions on zooplankton distribution in the future.

1.1 Problem Definition

The problem that we address in this thesis is the modeling and prediction of zooplankton diel vertical migration (DVM) pattern in the water column given the dynamic changes of environmental factors from various sources. We define "migration pattern modeling" as the task of predicting the positions (depths) of zooplankton in the water column in a later time based on a set of attributes. The work in this thesis takes into account the spatio-temporal aspect of the problem of modeling the DVM pattern.

Modeling the DVM pattern can help domain experts understand the behaviour of zooplankton in a systematic way. Although the advance of remote sensing techniques has made it much easier for biologists to study the spatiotemporal distribution of zooplankton, the huge size of data still makes it difficult and fallible to analyse the measured signals by direct visual observation. The work presented in this thesis use machine learning algorithms to automatically process the collected data and build migration models for different time granularities that can be viewed and understood much more easily.

In this research study, we use machine learning methods to study and model the DVM pattern of zooplankton based on data collected from 19 months. Using acoustic backscatter data measured by a Zooplankton Acoustic Profiler, the monthly averaged migration path is modeled as an Almost

Periodic Function, a function of time that shows periodicity with small variations. This mathematical representation clearly shows that the pattern is periodic on the whole, while still changing gradually with time.

Further correlation analyses reveal that various environmental and time features have important impact on the coefficients of the model. In this research study, we collect 18 features from various sources and examine their correlation with the coefficients of the almost periodic function both analytically and quantitatively. After examining their relative importance, the features are used as parameters in the functions of coefficients. The features change over time, so do the coefficients. This change allows the DVM pattern change gradually over time while still keeping the basic shape. With the help of the almost periodic model, we are able to examine and predict the general migration path on a large time granularity.

Moreover, we also examine the DVM pattern on finer time granularities. The relative importance of the features are analysed with various feature selection algorithms on 60-minute, 30-minute, 15-minute and 5-minute averaged intensity signals. The feature selection results confirm that both time and environmental factors have strong correlations with the DVM pattern.

In order to further quantify the vertical distribution of zooplankton on small time scales, we build a set of classification models for each of the depths along the water column. Given a set of environmental and time measurements as the features, the classifiers are able to predict whether or not zooplankton will exist at this particular position and time stamp. Combining the predictions of classifiers for all depths, we are able to predict the zooplankton distribution along the entire water column.

Both the almost periodic function and the classifiers are useful for modeling and predicting the DVM pattern. While the almost periodic function allows for an observation of the general shapes and gradual changes of the pattern over a long period of time, the classifiers are more suitable for finer time scale examinations. Both models provide auxiliary tools for domain experts to study the DVM behaviour in relation to environmental factors.

1.2 Diel Vertical Migration

Many species of oceanic zooplankton undertake a diel vertical migration (DVM) pattern. Migration is typically nocturnal. The zooplankton ascend to the water surface during the night, and move down in the water column during the day. The diel migration occurs during twilight hours [41].

DVM draws much attention in the biology study area because it serves as clues for predator-prey interactions and population dynamics. Also, it makes contributions to biogeochemical processes, such as transport of dissolved inorganic carbon and nitrogen to deep water [41].

There are many hypotheses regarding the reason for DVM. Predator evasion is one of the most favoured ones. According to this hypothesis, zooplankton migrate to the surface because of the abundance of food resources (phytoplankton) but have to move down the water column during the

daylight in order to avoid their predators [41].

Besides the DVM pattern, other zooplankton migration patterns are also reported. Twilight migration, where zooplankton migrate to the surface for both dusk and dawn twilight hours, moving down between dusk and dawn can be identified on occasion [41]. Some zooplankton, for example, copepods, conduct a seasonal migration pattern, drifting on the surface during winter times and descending deep in the water when weather gets warmer [30]. Moreover, not all zooplankton have a migration pattern. Some non-migration types are also found [43]. However, this thesis only focuses on the DVM pattern, because of its importance and constancy.

Because of the importance of the DVM pattern, many biologists are interested in long-time observation of zooplankton in the water column. Direct observation of the spatio-temporal distribution of zooplankton, however, is limited by the effort needed for sampling and subsequent analyses. Thus, many instruments using remote sensing techniques are designed to capture the dynamics of zooplankton distribution [34].

The Zooplankton Acoustic Profiler (ZAP) is used to monitor the presence and location of zooplankton in the water column. Figure 1.1 shows the ZAP transducer and pressure case. Deployed either on a floating buoy on the water surface or deep on the waterbed, ZAP eco-sounder transmits high frequency pulses of sound into the water column. If the signal encounters a target (zooplankton, fish schools, bubbles, etc), part of it is reflect back. By measuring the acoustic backscatter returned, ZAP can detect the presence of zooplankton and other objects [2]. The technical specification is included in Table 1.1.

Figure 1.1: The ZAP transducer and pressure case [2]

Table 1.1: Technical specifications of ZAP [2]

Frequency(kHz)	200
Range(m)	200
Beam width(degrees)	8
Sample rate(Hz)	1
Units	counts(converted to dB)

Figure 1.2 is the plot of ZAP intensity signals for one day. The ZAP is deployed on the seabed in Saanich Inlet, BC.The depth shows the distance from the instrument. Intensity values of the measured signals are illustrated in different colors. The intensity range for the zooplankton is approximately between 5 and 15 db. The DVM pattern can be clearly seen from the plot. Besides, fish signals (in red) are also visible during the daytime.

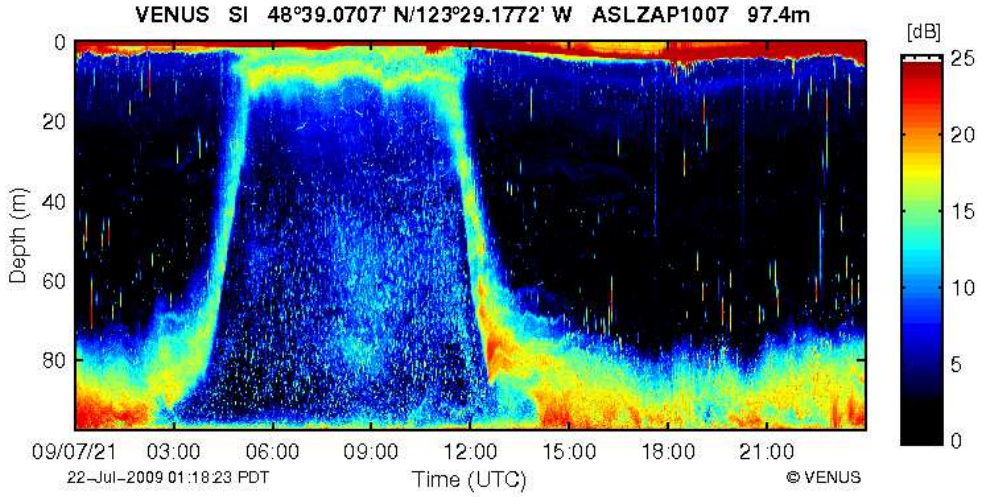


Figure 1.2: Daily 30 second averaged plot of Acoustic Backscatter Intensity. The intensity range for the zooplankton is approximately between 5 to 15 db [2]

Acoustic Doppler Current Profiler (ADCP) is an instrument that measures water column velocities. The ADCP can provide an estimate of current velocities for three directions: U (east), V (north) and W (upwards). With a high enough concentration of migrating zooplankton in the water column, the measured velocities (especially W) can be used as the estimation of the velocities of zooplankton. Figure 1.3 shows an ADCP prior to deployment.

Figure 1.4 shows the plots of velocities of three directions (U, V and W) measured by an ADCP together with the backscatter values. The DVM pattern can be seen from the vertical velocity plot.

1.3 Scope of the Thesis

This thesis presents a study of zooplankton diel vertical migration pattern modeling. The main results consist of various models that make predictions on the presence or absence of zooplankton

Figure 1.3: ADCP prior to deployment [2]

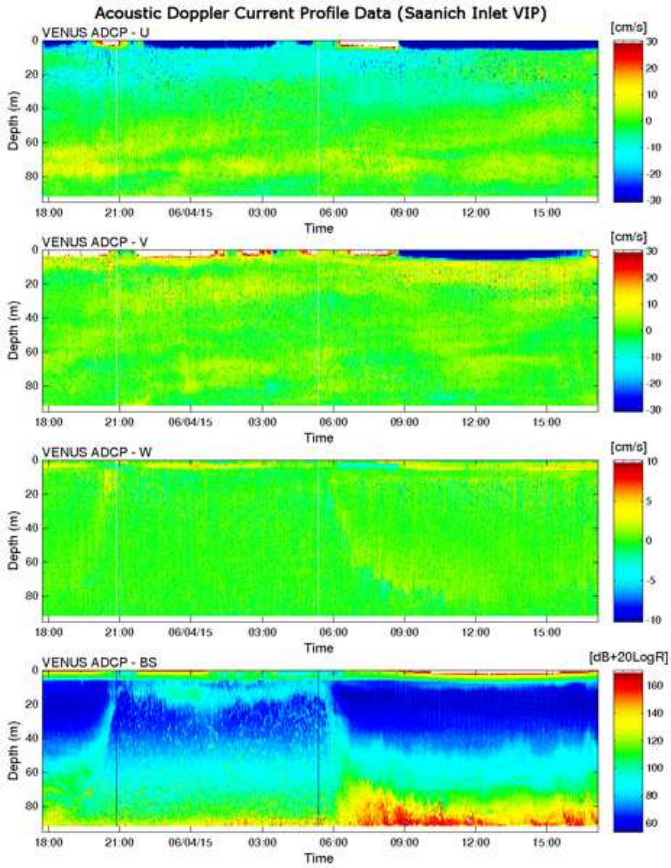


Figure 1.4: Example of an ADCP profiler, April 2006, Saanich Inlet [2]

in the water column for different time granularities. The work helps unveil the characteristics of the DVM pattern and the impact of environmental factors on it.

However, reasons for the DVM behaviour and the environmental impact are outside the scope of this research as these require further biological study.

1.4 Thesis Outline

The rest of the thesis is organized as follows: Chapter 2 outlines current work about zooplankton migration analyses in the biology research field and provides an overview of techniques we use in this research study. Chapter 3 describes the modeling of DVM pattern using an Almost Periodic Function on monthly averaged data. Chapter 4 presents correlation analyses and classification on finer time scales. We conclude our results in Chapter 5.

Chapter 2

Related Work and Methods

In this chapter, we describe the state-of-the-art research advance about zooplankton diel vertical migration (DVM) pattern in the biological research world as well as models and techniques from the research area of computing science that we apply to build our own DVM models. Section 2.1 describes current methods of DVM modeling. Section 2.2 discusses current research results on correlations between the DVM pattern and various environmental factors for different lakes and oceans around the world. Section 2.3 describes clustering methods from the area of data mining that are tried to exact migration paths in the data pre-processing stage. Section 2.4 explains techniques of curve fitting which are used to find the mathematical representation of the DVM paths for monthly averaged data. Section 2.5 describes the concept of an Almost Periodic Function, which is used as our model in the curve fitting process. Section 2.6 illustrates the techniques of feature selection, which we use to find correlations between various environmental factors and the DVM pattern. Section 2.7 gives a short review of supervised learning, which we use to model the DVM pattern on finer time granularities.

2.1 Quantifying the Migration of Zooplankton

There are several techniques found in the literature for modeling the DVM pattern, each suited to a different purpose using various measuring instruments. Rippeth and Simpson [43] use current velocities measured by an Acoustic Doppler Current Profiler (ADCP) to estimate the migration paths by zooplankton. The ADCP can measure current velocities of three dimensions (east, north and upwards), when the concentration of zooplankton in the water column is high, the measured upwards velocity can be used as an estimation of the velocity of zooplankton. In their work, a particle is assumed to move with the measured vertical velocity from the seabed to the surface and descend back to the seabed. Its moving path is used as an estimation of zooplankton migration path. In our research study, we can not directly apply this work because the data we use are collected by a different instrument (Zooplankton Acoustic Profiler, ZAP). ZAP measures only the intensity of reflected sound signals, no velocity information is recorded. There are some techniques for estimating the current velocities from intensity data [41]. However, these techniques relies on an intermediate conversion from backscatter intensity to equivalent zooplankton biomass or some measure of abundance which itself is still an active research area and may require net sampling and many subsequence effectors [18] [47] [6]. Moreover, in the presence of fish and other targets, the estimated velocity is a combination of velocities of zooplankton, fish and other moving objects at that depth. It may not always reflect the true velocity of zooplankton and the path estimated from this velocity may not always reflect the true migration path. It is also possible that the zooplankton align themselves in a certain orientation for vertical swimming that weakens the signals detected by the ADCP. For these reasons, we do not apply the particle tracking model in our research study.

Carin et al.[6] research the behaviour of zooplankton both horizontally and vertically. Horizontal research is used to confirm the impact of wind. It also shows that the DVM pattern can be

observed continuously across a 1000 km transect of the Arabian Sea. For vertical research, the biomass medians (i.e. the depth that divides the biomass in half) are generated from raw data and calibrated by net-samplings. The medians are then used to represent the DVM pattern. For each day, a cosine function is generated from the biomass medians to represent the general pattern. The root mean squared error is calculated to validate the coherence of the biomass median path to the cosine function. The paper finds out that when DVM occurs (March-April), the sinusoidal curve can closely track the biomass median. But when DVM is weak or absent (June-July), there is a big deviation between the cosine curve and the actual biomass median. The paper does not provide further analysis on the cosine curves (for example, comparing cosine curves from two consecutive days). It simply fits biomass median data into a series of cosine curves. In this research study, however, we not only want to generate functions for each day, we also want to organize these functions and generate a more compact model that shows the general behaviour of zooplankton over a long period of time. Since the DVM pattern occurs throughout the year in our dataset, the model should be able to capture the general shape of the DVM path, as well as the subtle changes at different times of the year. Moreover, besides time, we also consider a set of environmental factors by including them as parameters into the model. Because we have no calibration by net-sampling results, we can not generate the biomass median information. Instead, we use the upper bound, lower bound and middle depth of the DVM pattern generated from a semi-automatic process to build the models. Combining all of the three models, we are able to capture the temporal changes of zooplankton migration, as well as its spatial (vertical) distribution.

Besides migration path, timings of migration are also of interest. Lorke et al. [34] and Record et al. [41] estimate the ascent and descent times from backscatter intensity values. When the zooplankton move fast along the water column, the backscatter intensities at that depth also change rapidly. For a fixed depth, the time series of intensity values is collected and the temporal derivative is taken. The positive and negative peaks of the derivative are used as the estimations of timings of vertical migration. Figure 2.1 shows one example of this approach. The upper panel shows the temporal derivative of acoustic backscatter data. The lower panel is the time series of the acoustic backscatter. The peaks in the derivative correspond to the migration times. Ashjian etc [6] use peak velocity times as the migration times. These methods can calculate the estimated ascent and descent times of zooplankton for each depth, thus unveiling the whole migration path. We tried to follow this approach to generate the migration path in our work. However, in the presence of lots of noise signals and lack of net samples to calibrate with, the peaks of the derivative do not always reflect the correct migration times.

There is also some work trying to relate the acoustic backscatter intensities to meaningful biological parameters of zooplankton [6] [44]. They provide different models to characterize the size, shape, abundance, orientation and classification of zooplankton. However, the instrument we use is a single frequency ZAP. With only one frequency available, it is not possible to distinguish differ-

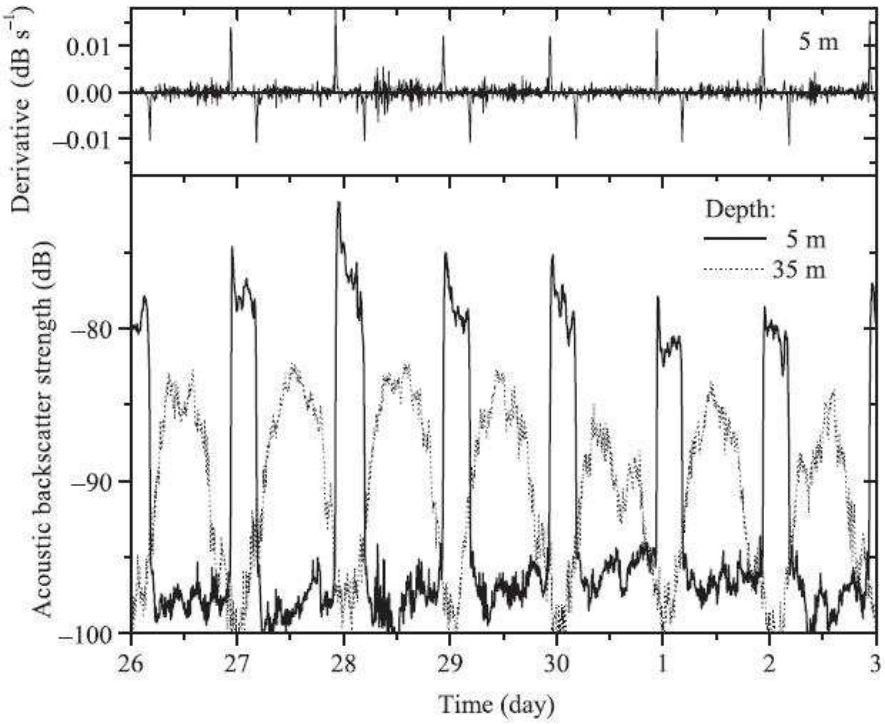


Figure 2.1: Temporal derivative of acoustic backscatter at 5m depth (upper panel) and time series of acoustic backscatter strength in Lake Hallwil from 26 June to 2 July 2001 at 5 and 35m depth (lower panel) [34]

ent species and sizes [43] [30] [34]. Moreover, the goal of our research study is to reveal the diel vertical migration paths of zooplankton, not to distinguish different types. For our purpose, the data recorded by the single-frequency ZAP is sufficient.

2.2 Environmental Factors Related to the DVM Pattern

Sunrise and sunset times have been observed to have a strong impact on zooplankton's migration times [41] [43] [30] [34] [6]. For each day, zooplankton ascend from the seabed around sunset time and move downwards to the seabed again around sunrise time. Since sunrise and sunset times vary around the year, zooplankton's migration times also change accordingly. In winter when the night is longer than the day, the zooplankton are found to spend more time on the water surface, whereas in summer, when the day length is greater, zooplankton spend much less time on the surface. Figure 2.2 shows zooplankton migration times with regard to sunrise and sunset times. The black line indicates sunset and the red line indicates sunrise. The experiment is conducted at two depths in Patricia Bay between March 22 and April 5, 2006. It clearly reveals the strong correlation between migration times and the local times of sunrise and sunset.

Besides sunrise/sunset times, many environmental factors are also believed to have impact on the DVM pattern. Record et al. [41] show that light and temperature affect the migration. In their work, cloud opacity data from a land based weather observation system is used as a measurement of light intensity. They test zooplankton's reaction to light by comparing migration times between clear days and overcast days. Each day is labelled either "clear" or "overcast" according to the cloudy opacity information. Migration times are calculated by taking the maximum and minimum temporal derivatives of intensity signals for each day. Results of t-test show that ascent/descent times on clear days are significantly different from those on overcast days. On overcast days, the zooplankton spend more time on the surface. They [41] also conducted coherence analysis between biomass median depth and temperature stratification (0 °c, -0.5 °c and -0.8 °c isotherms) for both daytime and nighttime hours. High correlations were found.

Lorke et al. [34] show in their work that water temperature, light transmissivity in water and water oxygen are correlated with zooplankton abundance. Figure 2.3 shows zooplankton distribution with respect to various environmental factors. According to the figure, the maximum abundance of zooplankton during night times occurs round depth 5m, right above the depth where light transmissivity shows a minimum and oxygen reaches the highest value. During daytimes, the highest concentration of zooplankton happens at depth 35m, above the depth where oxygen concentration drops.

Rippeth et al. [43] also mention in their work that the concentration of zooplankton is affected by the water temperature through similar analysis. Moreover, they also explain that not all species of zooplankton response to the changes of light. The Copepod Calanus helgolandicus, for example, is not sensitive to light changes caused by overcast weather.

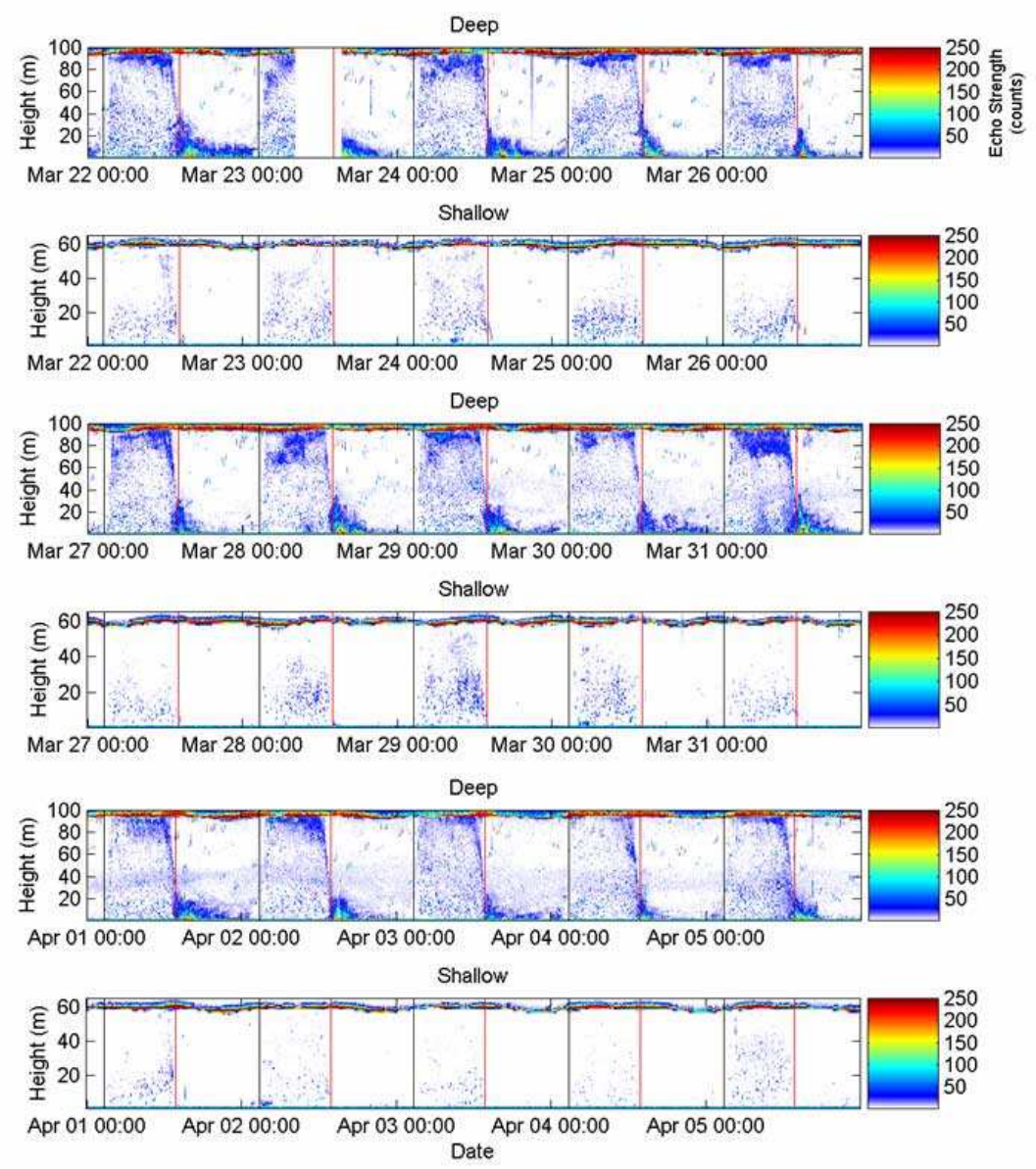


Figure 2.2: Acoustic backscatter at the deep and shallow sites in Patricia Bay together with the sunrise and sunset times. Time is UTC. The black line indicates sunset and the red line indicates sunrise [2]

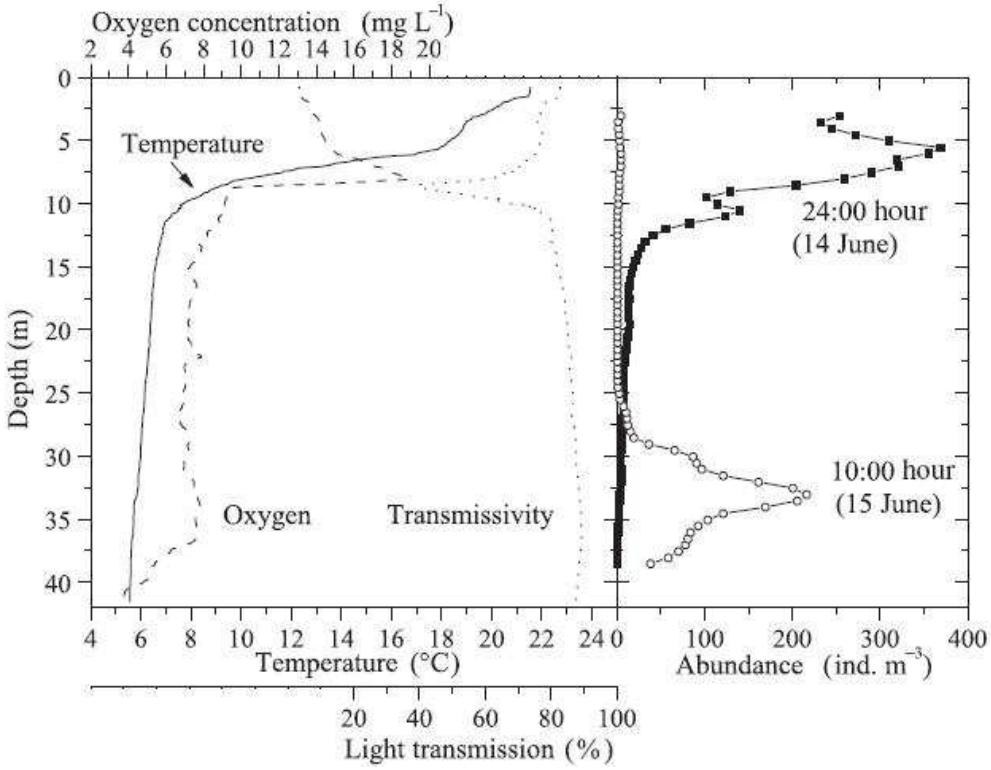


Figure 2.3: Zooplankton abundance and changes of various environmental factors for a typical day in Lake Hallwil [34]

Although the above papers have proven that environmental factors can impact zooplankton migration/distribution, none of them indicate relative importance of each of the factors. Moreover, a function describing the DVM pattern should be helpful for domain experts to do further analysis. The goal of this research study is to use machine learning methods to generate DVM models and find the quantitative correlations between environmental factors (including times) and the pattern.

2.3 Clustering

Due to the fast advance in sensor technologies and database systems, large amounts of data have been collected and stored in various media. Along with the growing availability of data is the growing demand for effective and efficient analyses of information implicitly contained in the data. As an answer to this demand, Data Mining has been an active research area for decades. The term "Data Mining" refers to the process of extracting useful information from large amount of data. It is an essential step in the process of Knowledge Discovery from Data (KDD). Besides Data Mining, the whole process of KDD also includes: data cleaning, data integration, data selection, data transformation, pattern evaluation and knowledge presentation. In practice, data mining and knowledge discovery are becoming synonyms [23].

Clustering is one important technique of data mining. It's the process of grouping objects into classes or clusters, so that objects from the same cluster have very high similarity whereas objects from different clusters have very low similarity. The major clustering methods can be classified into the following categories: partitioning methods, hierarchical methods, density-based methods, grid-based methods and model-based methods [23]. In our study, we conduct the empirical study based on several typical clustering methods: K-means, DBSCAN and STING. K-means [24] is well known as a partitioned-based clustering algorithm. Given the number of clusters k, it aims to divide a given dataset into k clusters in which each object belongs to the cluster with the nearest centroid. Ester et al. [17] proposed a density based clustering algorithm (DBSCAN) for large datasets. Two parameters Eps and $MinPts$ are used in the algorithm to control the density of clusters. DBSCAN is able to separate data noise from clusters of objects where data noise consists of objects in low density regions. DBSCAN can be used to detect clusters of any shape. However, the complexity of DBSCAN is relatively high and it requires a human participant to determine the global parameter Eps. STING [46] is a grid-based clustering method in which the data space is divided into a number of rectangular cells. Several levels of such rectangular cells are generated corresponding to different levels of resolution. These cells constitute a hierarchical structure, in which each cell at a high level can be divided into a few of cells at the next lower level. Statistical information regarding the data in each grid cell is computed and stored in advance for answering queries.

In this research study, we try to extract zooplankton migration path from raw data using clustering algorithms. An ideal result should group zooplankton and non-zooplankton signals into different clusters. After extensive experiments, however, we find that in lack of domain knowledge, it is quite

difficult to choose the best values of the parameters. The clustering results require further post-processing and the use of clustering does not make extracting migration paths easier. Therefore, in later experiments, we use other methods to extract the migration path instead.

2.4 Curve Fitting

We use Curve Fitting to build models for the DVM pattern on the monthly averaged data. The curve fitting problem involves making predictions of a dependent variable from independent ones by fitting curves to the data. The best fitting curve is determined by some statistical criterion [7]. Curve fitting includes parametric fitting and non-parametric fitting. Parametric fitting involves finding a set of parameters for one or more models that fit the data. The data is assumed to have two components: a deterministic component and a random component.

data = deterministic component + random component

The deterministic component is generated from a parametric model and the random component is caused by errors. Given a parametric model, the task of parametric fitting is to find the best set of parameters of the model that mostly fits the data.

Non-parametric curve fitting estimates values between known data points (interpolation) or creates a smooth curve through observed data (smoothing). Since the goal of this research study is to find the mathematical function of migration pattern, parametric curve fitting is used.

One famous method of parametric fitting is least squares fitting. Given a parametric model, this method tries to find the set of parameter values that minimizes the summation of squared residuals. The residual r_i for the *i*th data point is the difference between the observed response value y_i and the fitted response value $\widehat{y}_i$ from the model.

$$r_i = y_i - \hat{y}_i \tag{2.1}$$

The summed square of residuals for n data points is given by:

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.2}$$

The goodness of fit of the model can be measured by the following statistics:

Sum of Squares Due to Error: this statistic measures the total deviation from the fitted values to the corresponding observed values.

$$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \tag{2.3}$$

w_i in the above equation are the weights. The weights determine how much influence each response value has on the final parameter estimates. The weight of a high quality response value

should be higher than that of a poor quality response value. y_i and $\widehat{y}_i$ are the observed and fitted values respectively. The closer SSE is to 0, the smaller the random error the model has, and the better the fit is.

R-Square: this statistic measures the correlation between fitted response values and observed response values. It is the ratio of the sum of squares of the regression (SSR) and the total sum of squares (SST).

$$SSR = \sum_{i=1}^{n} w_i(\widehat{y}_i - \overline{y})^2 \tag{2.4}$$

$$SST = \sum_{i=1}^{n} w_i(y_i - \overline{y})^2 \tag{2.5}$$

$$R - square = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{2.6}$$

$\overline{y}$ in the above function is the mean of the observed data. An R-square value close to 1 indicates that a large proportion of variance is accounted for by the model.

Root Mean Squared Error: this statistic measures the standard deviation of the random component in the data.

$$RMSE = \sqrt{MSE} \tag{2.7}$$

$$MSE = \frac{SSE}{v} \tag{2.8}$$

v in the above function is the degree of freedom defined as the number of data points minus the number of coefficients. An MSE value closer to 0 indicate a fit this is more useful for prediction.

The above statistics measure how much variability in the data is explained by the fit, and how useful a fit will be for prediction. Besides these measures, a "good fit" should also be the one that the data could reasonably have come from and the parameters are estimated with little uncertainty. Also, the model should be easy for humans to interpret.

Besides the statistics, visual exploration of the fitted curve is also helpful for determining the goodness of fit. By examining the residuals, we can check the deviation from the fitted values to the observed values. For a good fit, the residuals should approximate the random components and thus should appear to behave randomly. If the residuals behave in a systematic pattern, it indicates that the fitted model is not a good one.

The confidence bound for the fitted parameters (coefficients) is another important indication. The confidence bound defines the lower and upper values of the coefficients and is calculated by the following function:

$$C = b \pm t\sqrt{S} \tag{2.9}$$

Where b is the estimated coefficient value, t depends on the confidence level, and is computed using the inverse of Student's t cumulative distribution function, and S is a vector of the diagonal

elements from the estimated covariance matrix of the coefficient estimates. The width of the interval of confidence bounds (difference between upper bound and lower bound) indicates the uncertainty of the fitted coefficient. The wider the confidence bound is, the more uncertain the fitted coefficient is.

2.5 Almost Periodic Function

Observation of the DVM path inspires the idea of using a periodic function as the model in curve fitting. In mathematics, a periodic function with period T satisfies:

$$x(t) = x(t+T) \tag{2.10}$$

or

$$|x(t) - x(t+T)| = 0 \tag{2.11}$$

for all t.

A periodic function can be represented by a Fourier series.The Fourier series is used to decompose a periodic function into a sum of sines and cosines, first introduced by Joseph Fourier [12]. The Fourier series representation of a periodic function is:

$$x(t) = \frac{1}{2}a_0 + \sum_{n=1}^{\infty}[a_n \cos(2\pi n f_0 t) - b_n \sin(2\pi n f_0 t)] \tag{2.12}$$

where $f_0 = \frac{1}{T}$ is the fundamental frequency. a_n and b_n are Fourier coefficients:

$$a_n = r_n \cos(\varphi_n) = \frac{2}{T}\int_{t_0}^{t_0+T} x(t)\cos(2\pi n f_0 t)dt \tag{2.13}$$

$$b_n = r_n \sin(\varphi_n) = -\frac{2}{T}\int_{t_0}^{t_0+T} x(t)\sin(2\pi n f_0 t)dt \tag{2.14}$$

where t_0 can be any time: $-\infty < t_0 < +\infty$ and the fundamental frequency f_0, the Fourier coefficients a_n, b_n, r_n, φ_n are constant.

The use of a periodic function can capture the general shape of the migration path. We can use the Fourier series representation of the function to produce a mathematical model for the DVM behaviour. If a periodic function is used, however, the resulting periods of the migration path should be perfectly identical, which fails to capture the subtle changes of the migration from day to day.

An Almost Periodic Function is a better model for our purpose. Almost Periodic Functions are functions of a real number that are periodic up to a small error, first studied by Harald Bohr. Almost Periodic Functions have many different forms in different application scenarios. We adopt the definition in signal processing, where almost periodic signals are also called quasi-periodic signals. A quasi-periodic signal is defined to satisfy:

$$x(t) \approx x(t+T(t)) \tag{2.15}$$

or

$$|x(t) - x(t + T(t))| < \varepsilon \tag{2.16}$$

where

$$0 < \varepsilon \ll \| x \| = \sqrt{\lim_{\tau \to \infty} \frac{1}{\tau} \int_{-\tau/2}^{\tau/2} x^2(t) dt} \tag{2.17}$$

T in the above function is the period and is a function of time as well.

Same as a perfectly periodic function, an Almost Periodic Function can also be represented by Fourier series:

$$x(t) = \frac{1}{2} a_0(t) + \sum_{n=1}^{\infty} [a_n(t) \cos(2\pi n \int_0^t f_0(\tau) d\tau) - b_n(t) sin(2\pi n \int_0^t f_0(\tau) d\tau)] \tag{2.18}$$

$a_n(t)$ and $b_n(t)$ in the above function are Fourier coefficients. $f_0(t)$ is the frequency [25].

Different from those of a perfectly periodic function, the frequency and Fourier coefficients in the above function are no long constant, but are also functions of time. Because of the slow variations of the coefficients and frequency, the periods of an almost periodic function look similar, while keep themselves a little different from neighbouring periods through small changes. An almost periodic function is a more proper model to use.

2.6 Feature Selection

We use feature selection to measure the importance of feature subsets. Feature selection is an important data pre-processing step for data mining and machine learning. It refers to the process of selecting a subset of features from original data in order to remove irrelevant, redundant or noise data [33]. The goal of feature selection is to select a minimum set of features so that the resulting probability distribution of the data classes is as close as possible to the original distribution when all features are used [23]. The process of feature selection is able to: (1) reduce the hypothesis search space, speeding up the follow-up algorithms (2) improve predictive accuracy (3) generate more understandable results (4) help humans have a better understanding about the data by telling them which subset of features are more important and how they are correlated (5) and in some cases, reduce storage requirement [33] [5].

The process of feature selection includes: subset generation, subset evaluation, stopping criterion verification and result validation [33]. Figure 2.4 shows the key steps.

In subset generation, a search method is used to generate a candidate feature subset. For N features, there exist 2^N candidate subsets in total. Figure 2.5 illustrates the whole search space for four features. Checking all candidate subsets is time consuming and often impractical for large numbers of N. Therefore, various strategies have been developed to reduce the search space. These search strategies can be broadly categorized into: complete search [15] [37], sequential search [31]

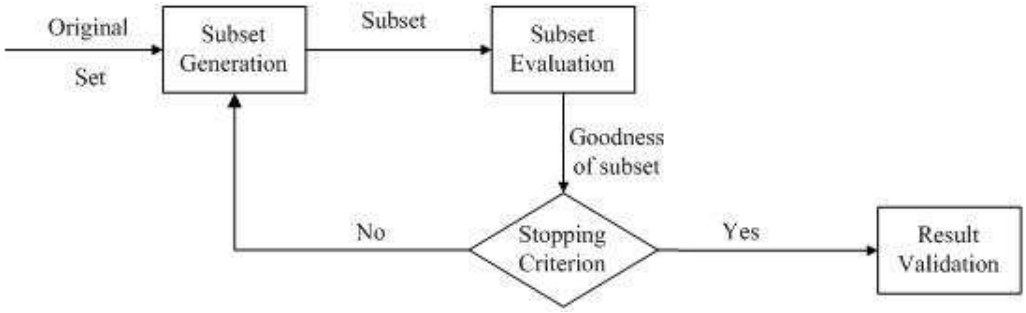


Figure 2.4: Key steps in feature selection [33]

and random search [15] [10]. Complete search guarantees to find the optimal subset. Although many methods are proposed to reduce the 2^N search space, the runtime of complete search can still be very slow. Sequential search adds (or removes) one feature at a time and evaluates the resulting feature subset until convergence. It can not guarantee to find the optimal subset, but it is quite fast (usually $O(N^2)$) [33]. Random search starts with a randomly selected feature subset and proceeds either following the sequential search or generating completely random subsets. It helps to escape from local optima but still can not guarantee to select the globally optimal subset.

Depending on the search direction, the search methods can also be divided into: forward search, backward search and random search. Forward search [29] begins with an empty feature set, and on each iteration, it adds a feature to the set on a trial basis. In detail, forward search first puts a candidate feature into the set and measures the performance of the resulting learning model by calculating its classification accuracy. The feature whose involvement most improves the accuracy is permanently added to the set. This procedure is repeatedly executed until addition of any available feature will result in reduced accuracy. Although backward search [19] is similar to forward search, it starts with a set containing all features and attempts to remove from the set the feature whose disappearance results in the highest accuracy gain. Random search [32] walks through the space of feature subsets by chance. If no start set is supplied, random search starts from a random point and reports the best subset found. If a start set is supplied, it will haphazardly search for subsets that are as good or better than the start point with the same or fewer features. In this research study, various search methods are used and their results are compared.

In subset evaluation, an evaluation criterion is used to measure the goodness of selected feature subsets. Subset evaluation methods can be broadly categorized into two groups: filters and wrappers [5]. Filters [14] [22] [32] [49] evaluate features one by one according to the intrinsic characteristics. Some of the characteristics include: separability of classes, information gain, dependency (correlation) with classes and consistency [33]. Filters evaluate feature subsets without using a mining method and are much faster than wrappers [5]. For wrappers [11] [16] [26] [28], a pre-determined mining algorithm is used to evaluate the feature subset. It gives superior performance when the

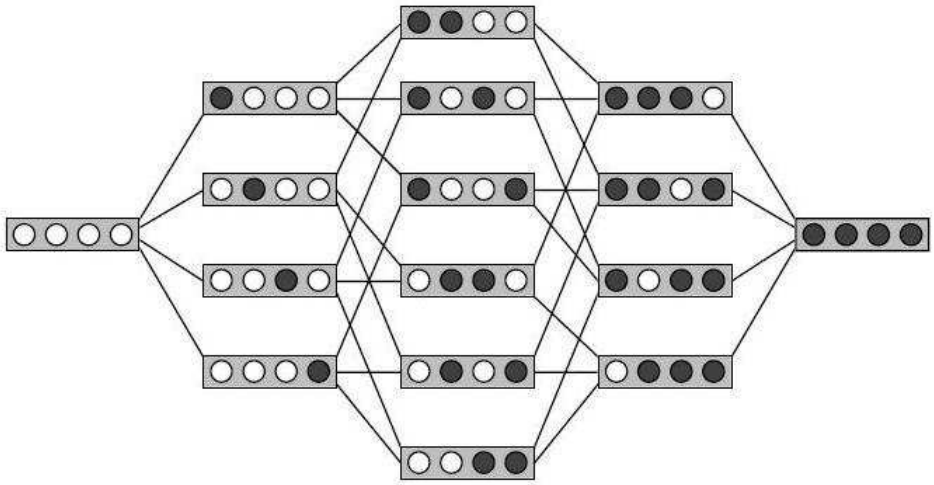


Figure 2.5: Whole search space [9], n=4

same algorithm is used for the actual mining task using the selected optimal (or near optimal) feature subset. But they are usually slower compared with filters. There are also hybrid models that try to combine filters and wrappers [13] [38] [48]. This research study includes both filters and wrapper for the evaluation process. We use filters to rank features in order to find their relative importance. This process helps us understand the features and their relationships with the intensity data better. Moreover, wrappers are used to find the feature subset that results in the highest classification accuracy.

The process of subset generation and subset evaluation repeats until a stopping criterion is met. Some stopping criteria are: (1) search is complete (2) a given bound is reached (3) subsequent subsets do not provide better evaluation results and (4) a sufficiently good subset is selected [33].

The selected best subset usually needs to be validated by either domain knowledge or tests of real datasets.

2.7 Supervised Learning

Machine learning is concerned with the design of algorithms that allow computers to learn models based on data. Once the model is learnt, it can be used to make decisions on new datasets or present the hidden pattern of the data for better understanding. Machine learning has broad applications including natural language processing, computer vision, bioinformatics, medical diagnosis, etc.

Supervised learning is one of the major categories of machine learning algorithms. In supervised learning, the whole dataset is first divided into training set and test set. Each instance of the dataset is composed of a set of features and a class label. A model is first learnt from the training set. Once the model is learnt, it is used to make predictions of class membership of the test set. The predicted

labels are compared with the true labels to measure the goodness of the model. If the model is not good enough, the learning process is repeated with some modifications (with a different learning algorithm, adding more training data, etc). If a satisfactory model is built, it is used to predict the class membership for new, unlabeled instances [8]. In our experiment, we select three machine learning algorithms to construct learning models. Decision tree [40] is well known as a typical learning model for classification accuracy. It can be used to visually represent a decision making process. In a decision tree, each interior node corresponds to one of the input features and its edges to children denote the possible values of that input feature. Each leaf represents a class given the values of the input features represented by the path from the root to the leaf. An unlabeled instance is categorized into a class if it falls into the leaf corresponding to this class. Naive Bayes Tree (NBTree) [27] uses decision tree as the general structure and deploys naive Bayes at leaves. The intuition behind it is that naive Bayes work better than decision tree when the sample dataset is small. Therefore, after several attribute splits when constructing a decision tree, it is better to use naive Bayes at the leaves than to continue splitting the attributes. A Bayesian Network [39] consists of a directed acyclic graph that encodes conditional independence among a set of attribute nodes and a class node, and a set that represents local distributions of nodes. A local distribution is typically specified by a conditional probability table. In a Bayesian Network, each attribute node is independent of its non-descendants in the graph given the state of its parents.

In this thesis, the prediction task is about finding out whether or not the zooplankton will present at a certain depth for a certain time. To perform this prediction task, we first learn models from a large number of labeled data instances. The features of data instances are obtained from different sources. Each instance represents the environmental conditions at a particular depth in the water column at a certain time. For each depth, a model is built to represent the dynamics at this particular depth. After the learning and testing process, the learnt models are used to make predictions on unlabeled data instances. Each unlabeled instance is given a class membership (label) according to its features (environmental factors). The labels indicate whether or not the zooplankton will present at that depth, for that time slice.

The selection of an appropriate set of features is crucial to the learning and prediction tasks. A good set of features can build a model that has the highest prediction accuracy. Moreover, since the ultimate goal of this research is to help biologists understand behaviour of the zooplankton better, the good set of features should also be useful for explaining the hidden dynamics of the zooplankton and can be explained and understood easily by the biologists.

Chapter 3

Modeling Zooplankton Diel Vertical Migration Pattern on a Large Time Granularity

Many types of zooplankton exercise a diel vertical migration (DVM) pattern. Around sunset times, zooplankton ascend along the water column, and spend the night on the surface. When sunrise times arrive, zooplankton descend down into the water column, and spend the daytime near the seabed. There are many hypotheses regarding the reason of the DVM pattern. One of the most famous one is that the DVM pattern of zooplankton is the result of a balance between hunting for food and avoidance of visual predators. While the sunrise and sunset times have been proven to have a strong impact on the DVM pattern, many environmental factors are also proposed to have influence on certain types of zooplankton.

The DVM pattern of zooplankton can be captured with the help of a Zooplankton Acoustic Profiler (ZAP). Mounted on the seabed, the instrument sends out high frequency sound signals into the water column, from the seabed to the surface. When the sound signal encounters an object, part of the signal is reflected back. By measuring the intensities of reflected signals, the behaviour of many marine animals, including zooplankton, can be captured. Figure 3.1 shows plots of ZAP intensities with respect to time in January 9th, April 1st, July 1st and October 10th of 2008 respectively. The measuring area is in Saanich Inlet of British Columbia, Canada.

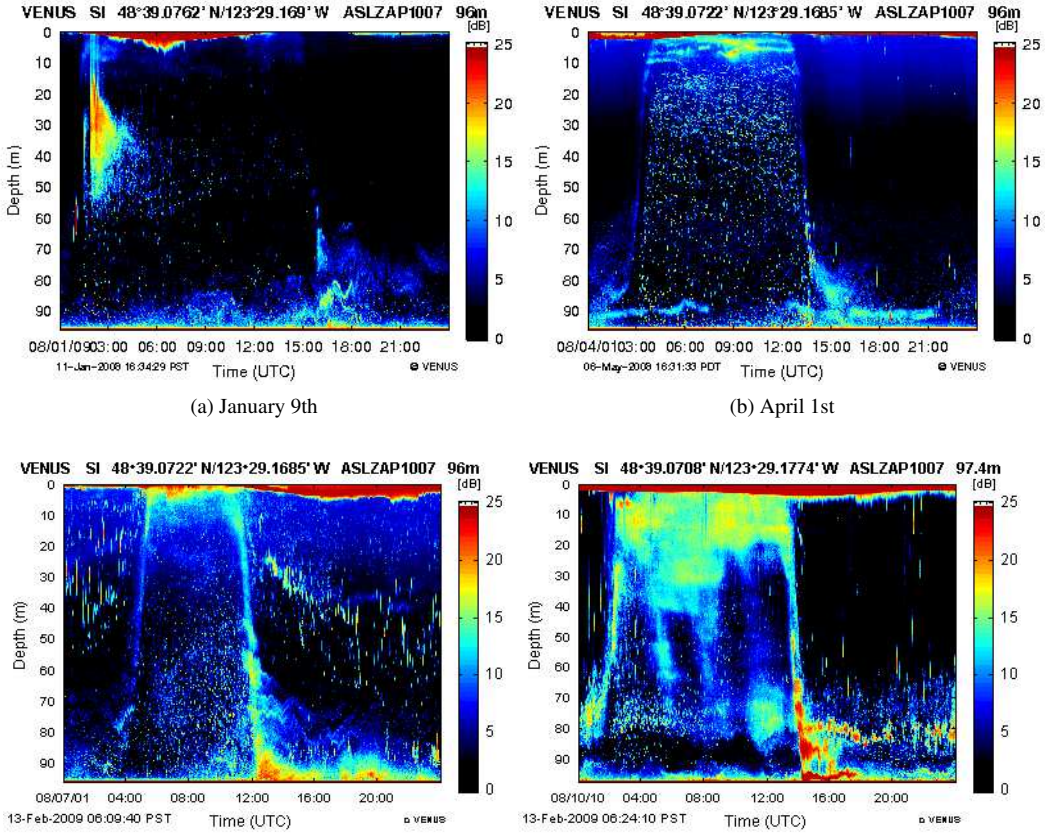
Each of the above pictures illustrates the migration of zooplankton for one day (24 hours). The time used in the pictures is UTC time, 8 hours before the local time (PST time). Therefore, 03:00 on April 1st of UTC time in the picture is 19:00 on March 31st of the local time, the time around sunset. Similarly, 15:00 on April 1st of UTC time is 07:00 on April 1st of the local time, which is also the time around sunrise. The plots show the measured intensity signals for different depth at different times of the day. They clearly illustrate the DVM pattern which goes up and down periodically.

From the pictures, we can see that zooplankton migrate around local sunrise and sunset times throughout the year. Although they migrate daily, spending night on the surface and daylight times on the seabed, their migration behaviour for each day is not exactly the same. The shape of the migration path changes gradually from day to day. Overall speaking, zooplankton spend more time on the surface in winter times and spend less time when summer comes, so the curves shown in the pictures are wider in colder times (as shown for the month of October and January) than in warmer times (as shown for the month of April and July).

Although much research has been done on the analysis of the DVM behaviour, little is done for quantitative modeling of the DVM path. The goal of this research study is to quantitatively model the DVM pattern, providing a mathematical function for the behaviour of zooplankton migration.

3.1 Data Pre-processing

In order to model the DVM pattern, we need to generate the hidden path from the ZAP (Zooplankton Acoustic Profiler) signals first. In this research study, acoustic intensity signals measured by a ZAP between 00:00:00 of January 1st, 2008 and 23:59:59 of July 31st, 2009 in Saanich Inlet are downloaded from the VENUS website [2]. The original data is sampled every second before September



(a) January 9th

(b) April 1st

(c) July 1st

(d) October 10th

Figure 3.1: Plots of ZAP intensities in Saanich Inlet, British Columbia. Each picture illustrates the migration of zooplankton for one day. The time used is UTC time, 8 hours before the local time (PST time)

2008 and every two seconds after September 2008. The whole water column is divided into 800 bins. For each time stamp, the intensity value for each bin is measured and recorded.

We compress the data by taking a 60 second averages. In order to reduce the effect of noise, medians are used instead of averages. For each day, the data is first divided along the time axis into 60 second blocks. Then for each depth in each block, a median is calculated as the representative intensity value for that particular depth during that time interval. This pre-processing step allows us to compress the data size without jeopardizing the pattern hidden inside the data.

The ZAP measures intensity backscatter as intensity values from 1 to 255. These values are converted to decibel (db) using the following formula:

$$\textit{converted value} = 10 * log_{10}(intensity)$$

The converted data have a range from 0 to 25 db. The nature of db units is that they are logarithm relative intensity scale. This conversion further helps to reduce noise included during the measurement while still preserve the DVM pattern.

With the help of great advice offered by biologists from the VENUS project [2], we know that the intensity values of zooplankton are approximately in the 5-15db range, while the intensity values of fish are often larger, in the 15-25db range. This categorization can only be used as a coarse guidance, because the actual intensity distribution is much more complex. High concentration of zooplankton, for example, can cause very high intensity values (even over 20db).

Once getting the 60-second averaged decibel values, the next step is to get the migration path for each day. This work proves to be quite challenging, because the data depicts a marine world that is much more complicated than we expected and the single-frequency instrument can not distinguish migrating zooplankton from the non-migrating ones.

Several clustering algorithms including DBSCAN, OPTICS, KMeans and STING are applied on the data set. For each algorithm, a large set of parameter values are used and the resulting clusters are carefully compared. An ideal clustering result should group migrating zooplankton signals into one cluster, separating them from the fish and water signals. After a large set of experiments, however, we can not decide which set of parameter values gives the best result. Moreover, the migration path can not easily be extracted from the clustering results.

Figure 3.2 and 3.3 show the plot of intensity values for April 23rd, 2008 and the corresponding clustering results using DBSCAN. Observing the raw data plot, we can see a non-migrating layer floating in the middle depth. The non-migrating objects have similar intensity values with migrating zooplankton and they overlap for short periods of time. A desirable clustering algorithm should be able to find the migrating path without including any non-migrating objects into it. Figure3.3 illustrates the DBSCAN results. We keep only intensities between 5 and 15 db, all other values are considered background and are ignored by the program. Each cluster is shown using a unique color.

The bounding boxes show the boundaries of the clusters. DBSCAN finds 571 clusters, most of which are small. The biggest cluster includes almost all light-blue points in the raw data plot. Both the migrating zooplankton and the non-migrating objects are considered to belong to the same cluster. The problem with DBSCAN lies in the fact that the algorithm only considers nearby neighbours without taking the "big picture" into account. When it encounters an object from the non-migrating layer, the algorithm should be able to consider the big trend (the pattern goes up first, stays on surface for some time, then moves down) and do not include this object even if its intensity is very similar to the migrating zooplankton. Unfortunately, DBSCAN can not do that. It groups all neighbouring points together if they have similar intensity values. Given the clustering result, it does not make finding the path easier. A series of post processing is required if we want to exact migrating path from the clusters. We further try a vast combination of the parameter values on data from different months. Although changing the parameters values can decrease the number of clusters, it does not change the major trend. Both migrating and non-migrating objects are grouped together.

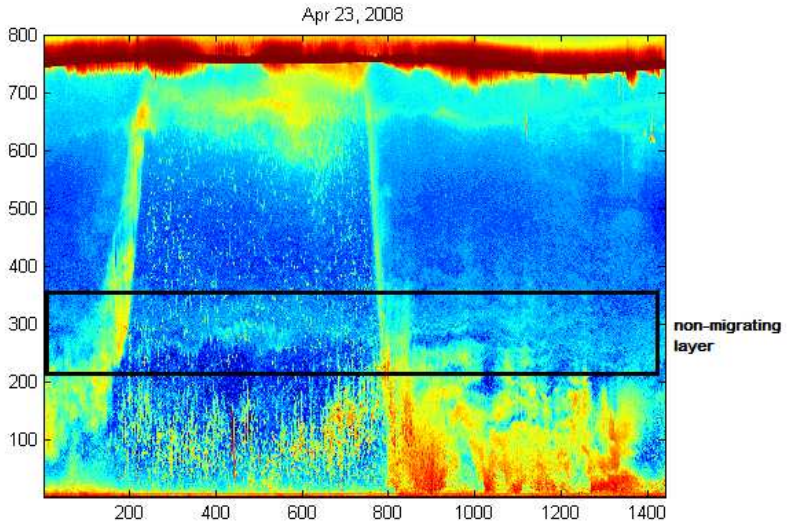


Figure 3.2: Raw data plot of April 23,2008

When fish schools are present, as are shown in Figures 3.4 and 3.5, the situation becomes even more complex. The algorithm groups fish school values into migrating cluster or breaks down the migrating objects into several clusters.

Since the presence of fish and non-migrating objects are quite common in summer months, we can't simply ignore this problem.

We also tried OPTICS, KMeans and STING with different parameter settings. All of them have the same problem as DBSCAN. The clustering algorithms treat the data as a real picture. In an actual picture, nearby similar points often belong to the same object (cluster). However, the data is not a real picture. It shows the spatial-temporal distribution of zooplankton. Although migrating

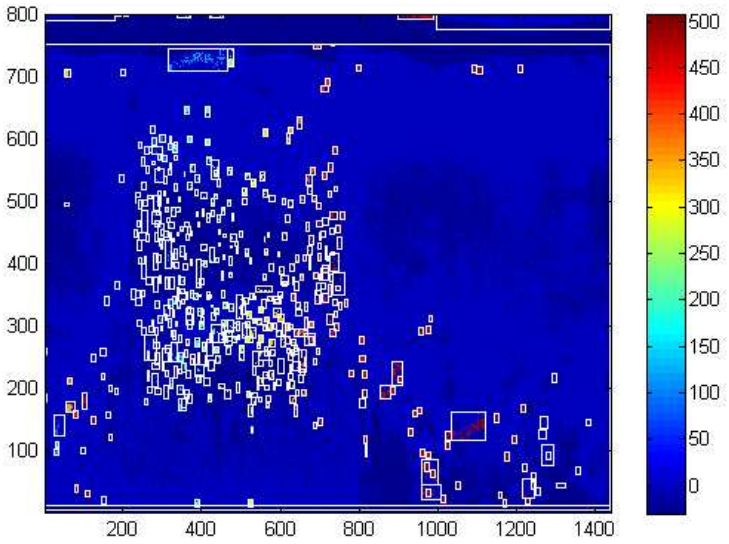


Figure 3.3: Clustering result.Epsilon=5,minObj=8,Manhattan distance

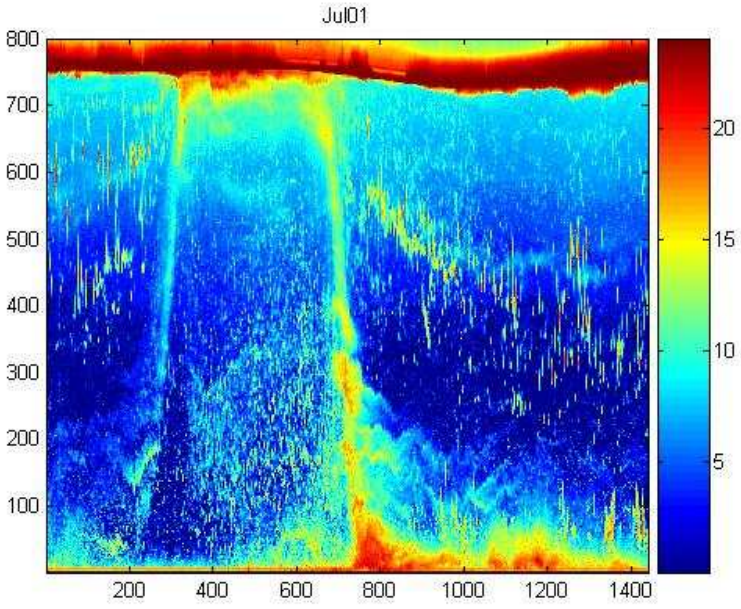


Figure 3.4: Raw data plot of Jul01,2008

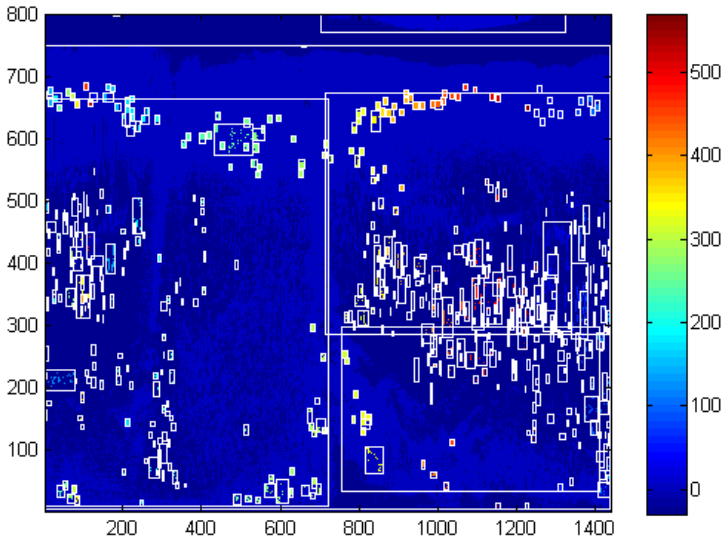


Figure 3.5: Clustering result.Epsilon=5,minObj=8,Manhattan distance

objects are often near each other and have similar intensities, the reverse is not true. Points should not be clustered only based on their distance, the trend, which goes up first, stays on surface for a while and moves down, also matters. The algorithm should only be interested in points belonging to this trend. Though it is relatively simple for visual observations, it is not easy for an algorithm to do that without learning the trend first. Clustering does not seem to be a correct solution to our problem.

In this research study, zooplankton migration path is generated using a semi-automatic way. For each month between January 2008 and July 2009, one day in that month is randomly selected. The migration depth for every 30 minutes of that day is manually generated (some of the zooplankton signals are above 20db. Using visual observations, we add them to the dataset although they are not within the 5-15db range). If the zooplankton spread along a depth interval for a particular time stamp, the middle depth of that depth interval is used as the migration depth at that time. This manually sample process is repeated for selected days.

After the manual sample process, a program is used to generate the migration path for each day using the manually sampled data as the guidance. For each time stamp in each day, depth intervals that contain intensity values between 5-15db are selected and the middle depth for each selected depth interval is calculated. The program compares each of the calculated middle depths with the one manually generated at the same time stamp in the same month. The closest middle depth is selected as the final migration depth at that time stamp. The migration path for each month is further averaged (using medians) to produce a representative path for each month. Figure 3.6 shows raw data plot of November 16, 2008 together with the manually generated migration path (the black

curve). Using this migration path, we are able to automatically generate migration path for other days in November 2008. Figure 3.7 shows the automatically generated migration path for November 10, 2008.

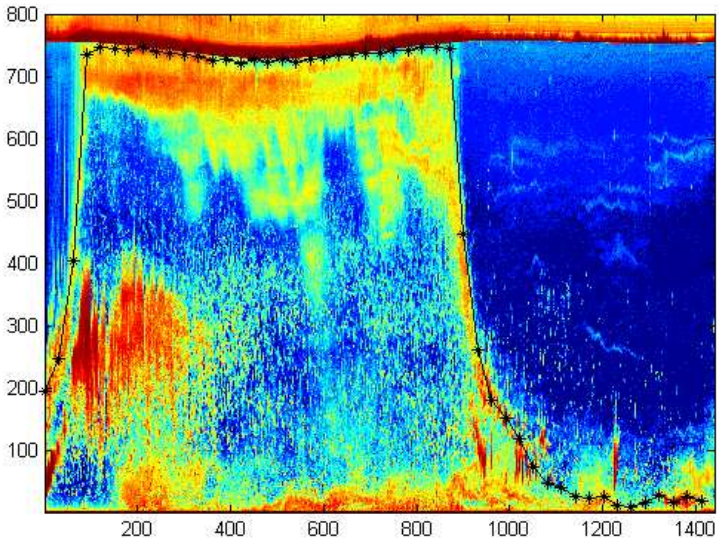


Figure 3.6: Manually labeled migration path, November 16, 2008

The whole dataset is divided into two parts. We use data from January 1st, 2008 to January 31st, 2009 as the training set and data from February 1st, 2009 to July 31st, 2009 as the test set. We want to build a migration model based on the training set and use the test set to evaluate it. Figure 3.8 is the plot of migration paths. The blue curve is used as the training set, while the pink curve is used as the test set. The time series clearly shows the DVM pattern which periodically occurs. Interestingly, the occurrences of the pattern do not have exactly the same shape. All periods of the migration paths look similar, but they also gradually change over time. A good model should be able to capture the general shape of the DVM pattern, while preserving the general changes along the time. We also include the dataset in Figure 1 in the Appendix.

There may be ways of automatically extracting the migration path from the ZAP data. However, due to the lack of ground truth in data and domain knowledge, we do not have a good method to evaluate the results and we are not sure whether the results are trustworthy. The semi-automatic process is the best way we can think of to generate reliable migration paths out of the data.

Besides the middle depth, we deploy the same semi-automatic process to generate the upper and lower bounds of the migration paths as well. Besides modeling the general shape of the DVM pattern, given the upper and lower bounds, we are also able to model the vertical distribution of zooplankton. The upper bound and the lower bound of the DVM time series are plotted in Figure 3.9. The red and green curves are the training and test data for the upper bound. The blue and black

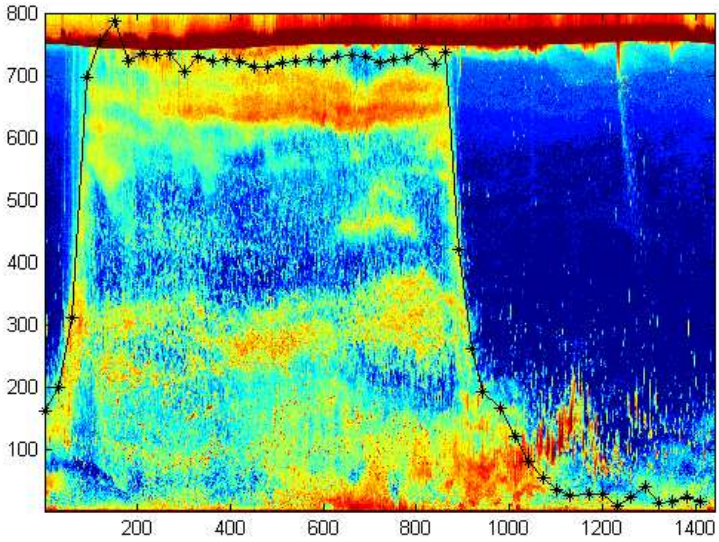


Figure 3.7: Automatically generated migration path, November 10, 2008

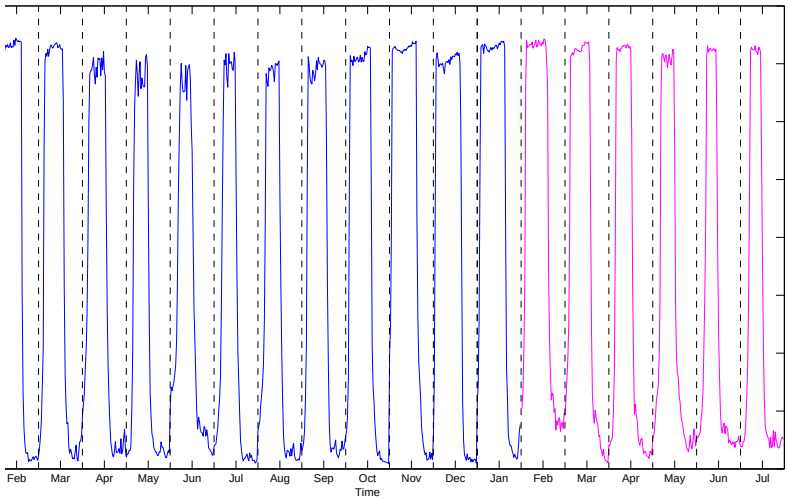


Figure 3.8: The middle depth of zooplankton migration path from January 2008 to July 2009. The blue curve is used as the training set. The pink curve is used as the test set.

curves are the training and test data for the lower bound. Both upper and lower bounds show the same trend as the middle depth. Figure 2 and Figure 3 in the Appendix show the datasets of upper and lower bounds respectively.

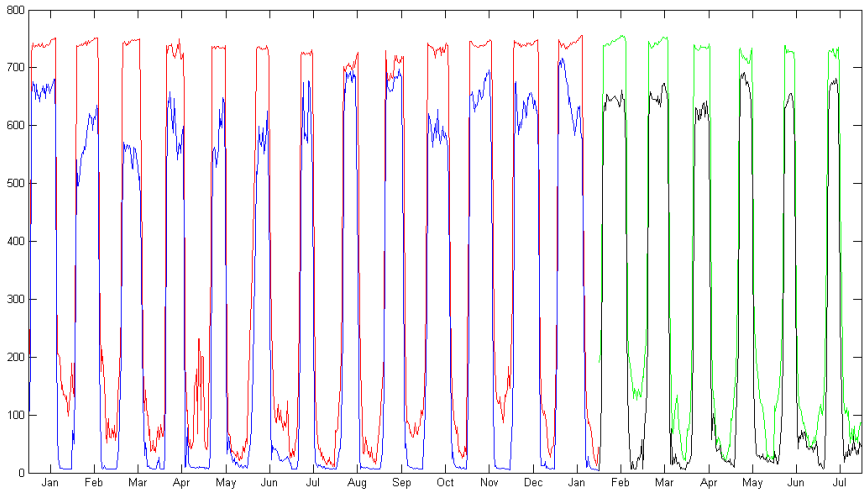


Figure 3.9: The upper and lower bounds of zooplankton migration path from January 2008 to July 2009. The red and blue curves are training set for upper and lower bounds respectively. The green and black curves are test set for upper and lower bounds respectively.

3.2 Diel Vertical Migration Path Modeling

We use an Almost Periodic Function as our model and fit the migration paths into the model. Once given a model, we need to find best values for its coefficients so that the model reflects the trend of the data as accurately as possible. The values of the coefficients of an Almost Periodic Function are learnt from zooplankton migration paths which are generated in the semi-automatic way from raw ZAP data. We apply regression in order to find the best values for the coefficients. The whole migration path is divided into two parts. We use paths from January 2008 to January 2009 as the training set to find coefficient values. The paths from February 2009 to July 2009 are used as test set to evaluate the fit of the resulting model.

Direct curve fitting for an almost periodic function is difficult, because not only does the curve change over time, all its coefficients are also functions of time. In order to finish the curve fitting more easily, we try to solve the problem using another method, with a new perspective of the periodic behaviour.

If the migration pattern is perfectly periodic, all its periods should look exactly the same. We can use the Fourier series representation of a periodic function as the target model, use the collected data to fit the model, and get estimations of all Fourier coefficients. According to the definition of a

periodic function, the resulting coefficients should all be constant.

Similarly, if we treat each period in the time series as one period from a perfectly periodic function, we can get a Fourier series representation of the periodic function by simply applying curve fitting on the data of each period. The resulting coefficients are constant. If we treat all periods in the time series this way, and try to find the Fourier series representation of each of them respectively, we will get a set of Fourier series, each of which has a set of fitted constant coefficients. Figure 3.10 illustrates this process. The red, blue and green curves are three periods from the original time series. We treat each of them as one period from a perfectly periodic function and use Fourier series to fit each of them. The table in the figure shows the fitted Fourier coefficients for the three curves.

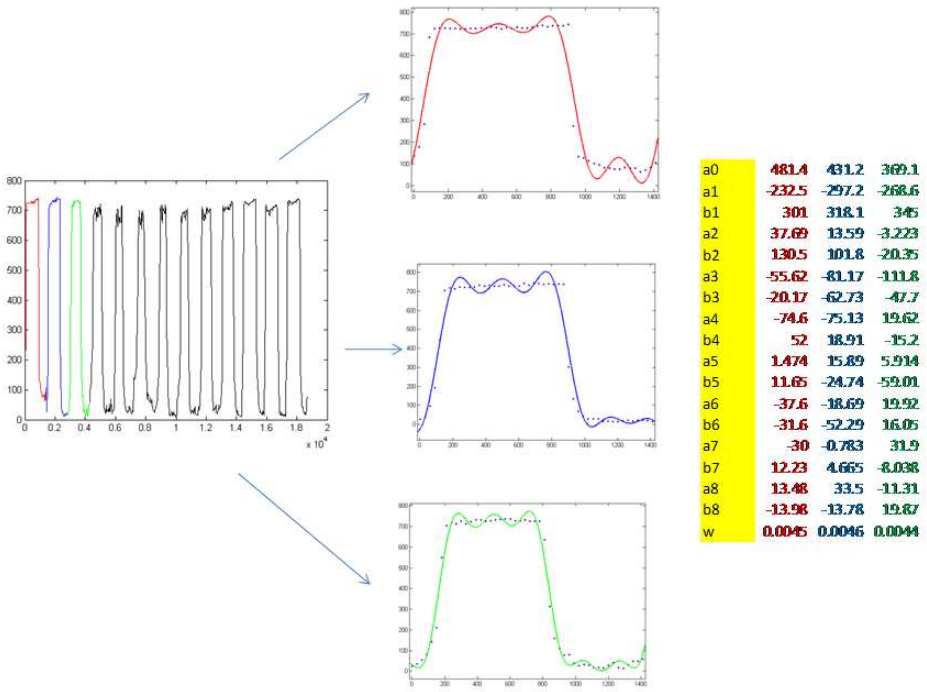


Figure 3.10: Use Fourier series to fit each period individually

Since the periods in the time series have similar behaviour (go up, stay on the surface for a while, then go down), the Fourier series representations are similar too. The slight difference between two consecutive periods lies in the slight difference of the values of the fitted Fourier coefficients. The shapes of the periods change gradually. If we treat each period in the time series as a "snapshot" of the almost periodic function, we will get the values of its coefficients at that particular time. After examining all snapshots, we will get a set of values for each of the coefficients. If we find mathematical functions to represent the changing mechanisms over time for each of the coefficients, we can use these functions as the functions of time the coefficients are generated from. Then we can

use these functions in our almost periodic model.

We apply the above idea on time series of the middle depth using Matlab's Curve Fitting tool. With the curve fitting tool, we can visually explore the dataset, fit it to different mathematical models and evaluate the goodness of fit both graphically and statistically. We use Fourier series as our model. For each period in the time series, we treat it as one period from a perfectly periodic function and try to fit the data using different terms of Fourier series (we tried Fourier series from 1 term to 8 terms). The different termed Fourier series are carefully compared and the best one is selected. The goodness of fit is measured by three statistics: (1) sum of squares due to error (SSE), (2) R-square and (3) root mean squared error (RMSE). Besides, we also examine the fit by visual observation and choose the one that is smooth with random residual and narrow uncertainty bounds.

The candidate fits are measured by SSE, R-square and RMSE. The best fit should have a SSE closest to 0, an R-square closest to 1 and a RMSE closest to 0. From the experiments, we find out that Fourier series with more terms tend to have better statistical measures and narrower uncertainty bounds on the training set. Figure 4 in the Appendix shows the fitted coefficients values if we choose the 8-term Fourier series as our model. The values are considered as snapshot values for the time varying coefficients and frequency.

A closer look at the table reveals a periodicity of the values of the coefficients (e.g. values in January 2009 and January 2008 are similar). This periodicity is consistent with what we expected, because the migration paths of January 2008 and January 2009 look almost identical. Obviously the coefficients and frequency change with time. Many research studies have also confirmed that zooplankton exercise the diel vertical migration around sunrise/sunset times. Because the times of sunrise and sunset vary throughout the year, the migration path of zooplankton also varies accordingly, shrinking in summer times and stretching in winter times.

Besides the effect of time, however, other research studies also explore the effect of environmental factors (light, temperature, etc.) to zooplankton's DVM pattern [41][34][43]. Some research studies have already proven that certain types of zooplankton can respond to certain environmental changes, although quantitative studies are not present. The almost periodicity of DVM pattern can be explained to be caused by the change of time (sunrise and sunset times), or a combination of various factors (time and environment). If the latter is true, the coefficients of the almost periodic function may be affected by both time factors and environmental factors. In order to explore the impact of various factors on DVM pattern, we collect 15 environmental measurements in the same study area of the same study period. The following data are collected from the VENUS website [2]: conductivity (s/m), temperature ($^{\circ}$c), pressure (decibar), salinity (psu), density (kg/m3), oxygen (ml/l) and transmission (%). The data are sampled every 60 seconds and recorded using UTC time. We name them water features. Besides water conditions, we also collect information of land conditions from the Weather Office of Environment Canada [3]. The weather information is measured by an on-land station near Victoria International Airport in Victoria, BC. The weather information

includes: temperature (°c), dew point temperature (°c), relative humidity (%), wind direction (10's deg), wind speed (km/h), visibility (km) and station pressure (kPa). The weather information is sampled every 60 minutes and is measured at local times. We call them land features. We also collect sunrise and sunset times from the National Research Council Canada [1]. The sunrise and sunset times are recorded in local times. We use these two times to calculate the night time for each day (the difference between sunrise and sunset times). All times are converted to UTC times. Figure 5 in the Appendix shows the monthly averaged data we collect.

In order to find the potential impact these factors have on zooplankton DVM pattern, we apply correlation analysis between each measurement and each coefficient. Correlation analysis is a common technique for finding the strength and direction of relationship between two random variables.

Figure 6 in the Appendix shows the result of linear correlation analysis. Strong correlations (p-value < 0.005) are shown in bold. The correlation measurement we use here is Pearson's correlation function [49]:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y} \tag{3.1}$$

The correlation is 1 in the case of an increasing linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

In order to measure the significance of the correlation, we also calculate the p-values for each correlation coefficient. The p-value represents the probability that the data would yield the obtained results if the null hypothesis is true. Smaller p-values indicate stronger evidence against the null hypothesis, and therefore stronger correlation.

From Figure 6 in the Appendix, we can see that night time, which is the difference between sunset and sunrise times, has strong correlations with many coefficients. This finding is consistent with our expectation that times have strong impact on zooplankton DVM pattern. Moreover, besides night time, several other factors, including both water and land measurements, also have strong correlations with the coefficients. These strong correlations suggest that, besides sunrise/sunset times, other factors may have impact on the DVM pattern as well.

We also calculate information gain to further confirm our findings. The entropy of a variable X is defined as:

$$H(X) = -\Sigma_i P(x_i) \log_2(P(x_i)) \tag{3.2}$$

The entropy of X given values of another variable Y is:

$$H(X|Y) = -\sum_j P(y_j) \sum_i P(x_i|y_j) \log_2(P(x_i|y_j)) \tag{3.3}$$

Information gain is defined as:

$$IG(X|Y) = H(X) - H(X|Y) \tag{3.4}$$

Information gain reflects additional information about X provided by Y. If information gain is high, it means that the amount of information of X provided by Y is big. In other words, X and Y are highly correlated [23].

We calculate information gain between each coefficient and each environmental measurement. Figure 7 in the Appendix illustrates the calculated information gain. Larger value of information gain indicates stronger correlation. After ranking the features according to the information gain, we find that night time, land temperature, water oxygen and water transmission have the highest rankings for almost all coefficients. On the other hand, land visibility, land wind speed and land dew point temperature almost always have the lowest ranking. According to this figure, night time, land temperature, water oxygen and water transmission have high correlations with almost all coefficients. Comparing results from linear correlation analysis, we find that night time and land temperature are highly correlated in both cases, while other land/water measurements (water oxygen, land visibility, etc) also have strong linear or non-linear correlations.

The above correlation analysis inspires us to add environmental factors into functions of coefficients. We use regression techniques to construct a linear function with environmental factors as parameters for each of the coefficients. The values of environmental factors change with time, so do the coefficients. We try the following algorithms in Weka [4]: (1) Isotonic Regression (2) Linear Regression (3) Simple Linear Regression (4) Least Median Squared Linear Regression (5) Support Vector Machine for Regression. We choose to use Support Vector Machine for Regression as the regression algorithm because it gives the smallest root mean squared error. Part D. in the Appendix shows the detailed functions for the middle depth coefficients. We also use the same technique to generate functions for upper bound and lower bound coefficients. The results are also included in Part. D.

For the almost periodic function, we try Fourier series with different terms (from 1-term to 8-term). From the experiments, we find that including more terms into the model almost always result in better statistical measurements on the training set, however, as we include more terms into the model, the resulting fitted curves have more oscillations. The almost periodicity of the curve is already given by a 2-term Fourier series, adding higher terms helps balance the amplitude to the 0-800 range. Figures 3.11 - 3.14 illustrate the fitted curve modeled by different termed Fourier series for the upper bounds (the lower bound and middle depth have similar trends). The red curve is the plot of the training data which is used to generated the model. The green curve is the plot of the test set. The blue curve is the plot of the predicted paths we generate from our model. The curve is learnt from data from January 2008 to January 2009 and is used to make predictions for February-July, 2009.

In order to choose the best model to use, we calculate the root mean squared error of the test set. Table 3.1 shows these errors.

The 5-term Fourier series is the best model among all that we have tried. Although it does not

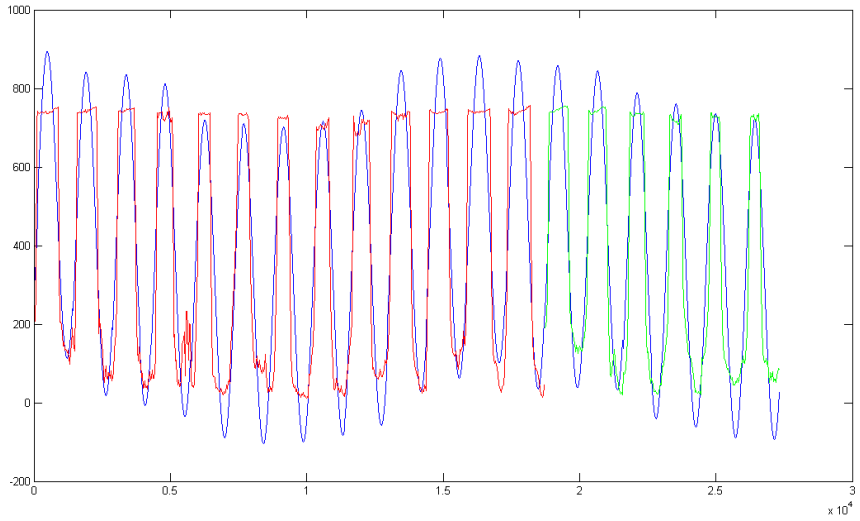


Figure 3.11: 1-term Fourier series

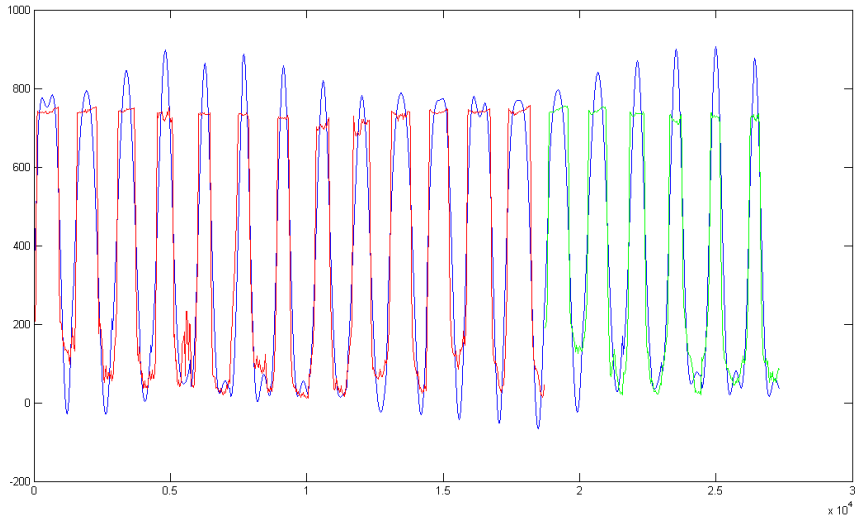


Figure 3.12: 2-term Fourier series

Table 3.1: Root mean squared errors for upper bound, lower bound and the middle depth of the migration path using different Fourier series models

Term	8	7	6	5	4	3	2
Upper Bound Error	93.48	92.13	87.52	86.73	86.72	90.89	97.87
Lower Bound Error	97.44	95.97	95.56	93	95.75	96.92	100.15
Middle Depth	86.93	86.61	83.89	82.44	84.74	85.85	95.08

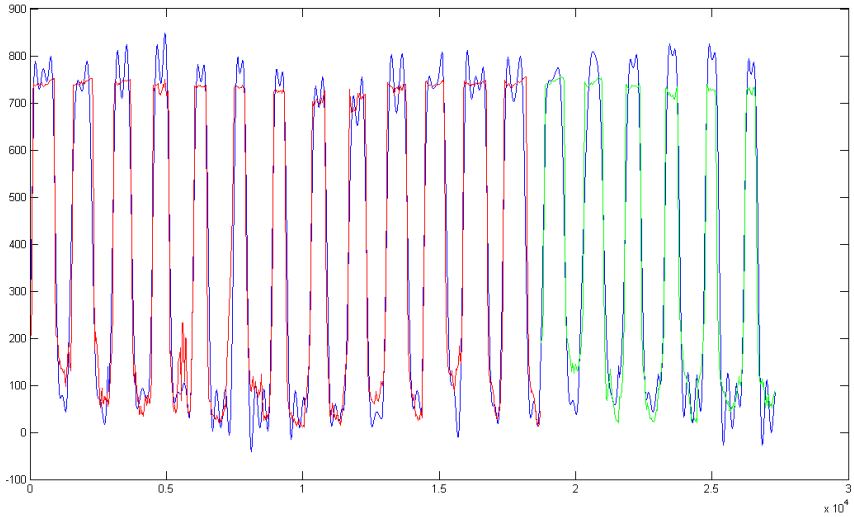


Figure 3.13: 4-term Fourier series

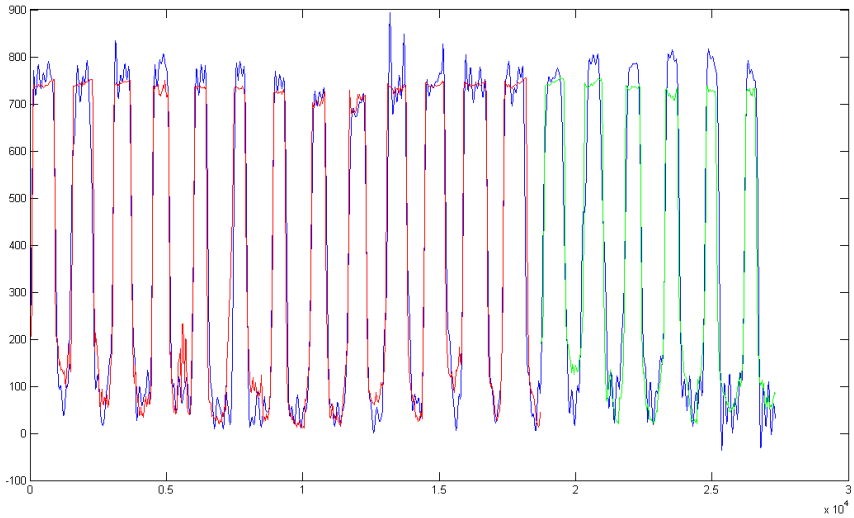


Figure 3.14: 8-term Fourier series

have the smallest statistical measurement on the training set, we still choose to use it in order to avoid data over-fitting. Figures 3.15 - 3.17 are plots of the 5-term model for upper bound, lower bound and middle depth.

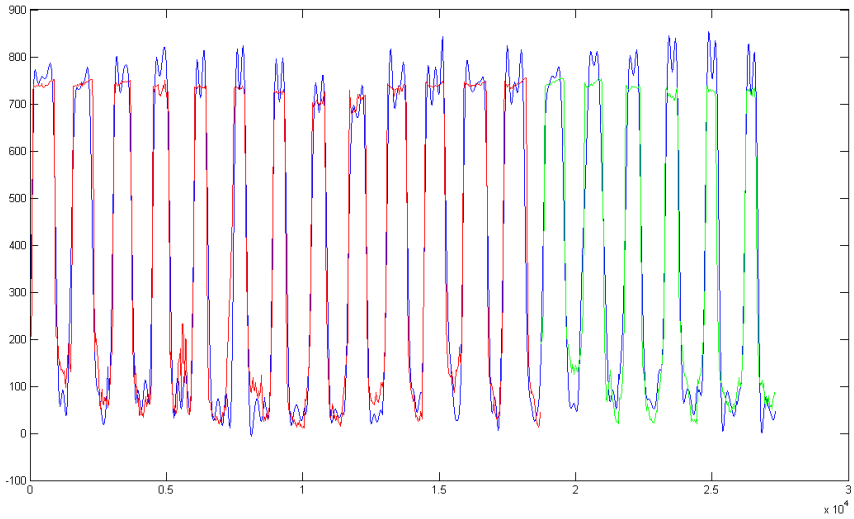


Figure 3.15: Upper bound

For the upper curve, errors occur mostly in top peaks, for the middle and lower curves, errors occur mostly in bottom perks. Using a 2-term model, we already get the basic shape of the curve, higher terms are mainly used to balance the amplitude of the curve (to make it range from 0 to 800). The errors at the perks are introduced by 3 or higher terms, as a by-product of balancing the amplitude.

Comparing prediction with the test set, we find that the fitted curve generally captures the almost periodicity of the data. All periods look similar, while each of them is a little different from neighbouring ones. The periods are wider in the winter and get narrower in the summer.

3.3 Final models for zooplankton DVM pattern

The Fourier series representation of an almost periodic function is:

$$x(t) = \frac{1}{2}a_0(t) + \sum_{n=1}^{\infty}[a_n(t)\cos(2\pi n\int_0^t f_0(\tau)d\tau) - b_n(t)\sin(2\pi n\int_0^t f_0(\tau)d\tau)] \tag{3.5}$$

Using $w(t) = 2\pi f_0(t)$, the function can also be written as:

$$x(t) = \frac{1}{2}a_0(t) + \sum_{n=1}^{\infty}[a_n(t)\cos(n\int_0^t \omega(\tau)d\tau) - b_n(t)\sin(n\int_0^t \omega(\tau)d\tau)] \tag{3.6}$$

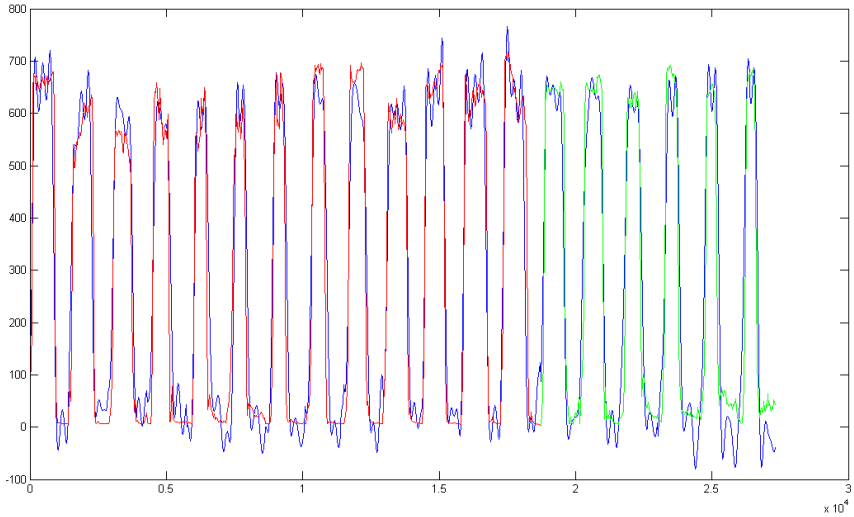


Figure 3.16: Lower bound

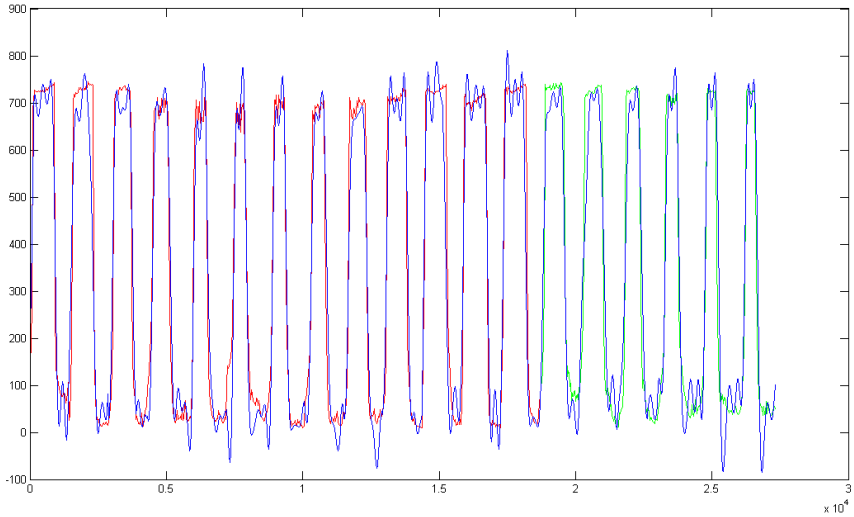


Figure 3.17: Middle depth

Because we choose to use a model with 5 terms, the Fourier function is changed to:

$$x(t) = \frac{1}{2}a_0(t) + \sum_{n=1}^{5}[a_n(t)\cos(n\int_0^t \omega(\tau)d\tau) - b_n(t)\sin(n\int_0^t \omega(\tau)d\tau)] \tag{3.7}$$

All coefficients and the frequency in the above function take the following form:

$$f(x^{(t)}) = \sum_i w_i * feature_i^{(t)} \tag{3.8}$$

Since the time scale of the above study is quite large (throughout 19 months), the environmental factors many go through very big changes during this long period of time. Their strong correlations with DVM pattern can be explained as a real correlation between the two, or a by-product of times changes. On this large time scale, it is hard to say whether or not the correlation between environmental factors and coefficients are caused by the change of time. In order to make further confirmations about the correlation between the two, we need to change to a smaller time scale, where time is not considered as a major changing factor. In the next chapter, we will explore the potential correlations using various machine learning algorithms on finer time scales.

Chapter 4

Modeling Zooplankton Diel Vertical Migration Pattern on Finer Time Granularities

The use of an Almost Periodic Function in Chapter 3 helps us model the DVM pattern in a systematic way. The function shows that the DVM pattern has similar periods on a daily base, while it gradually changes over time. The shapes of the periods in the winter are wider than those in the summer, which means zooplankton spend more time near the surface in the winter. The purpose of the almost periodic model is providing a mathematical way of predicting the DVM pattern over a long period of time.

Besides examining the DVM pattern over a long period of time, modeling the pattern on finer time scales is also desirable. Using finer time scale modeling, we are able to answer questions like "where will the zooplankton be in 15 minutes?" Though the almost periodic model is able to answer this question, since the function is more tuned for large scale modeling, the prediction it provides may be not accurate enough for finer time scale analyses.

In this chapter, we follow the typical supervised learning process to build classification models on finer time granularities. We organize the data into tuples (examples). Each of the tuples contains a set of environmental factors (features) and a class label indicating whether it represents zooplankton, fish or others. We first fix our study to a single depth, constructing models to learn zooplankton migration pattern only at this particular depth. Once the model is successfully built, we can apply the same technique on the whole water column, constructing a set of models to learn the migration pattern along the entire water column. We initially choose the depth 42m as our target depth. The study area of this depth is approximately in the middle of the water column (the whole column is 96m). The reason for choosing the middle depth is that the zooplankton migration path is the clearest to observe at this depth, with very little disturbing signals from fish, water currents, etc. Therefore, we can label the class memberships as accurately as possible.

The features of the tuples include both environmental measurements and time measurements. After extensive experiments with various feature selection algorithms on different time scales (60 minutes, 30 minutes, 15 minutes and 5 minutes), we find that a set of features, both environmental and time, are highly correlated with the distribution of zooplankton. This result further confirms our findings in the previous chapter, and provides valid reasons for using these measurements as features in the supervised learning process.

4.1 Data Pre-processing

In order to explore the correlations between environmental factors and migration path for different seasons, we use acoustic backscatter signals of seven months (January, April, July and October of 2008, January, April and July of 2009) and 18 features to conduct further correlation analyses. A new categorical land feature, weather, is added to the feature set. Using the absolute sunrise and sunset times, we calculate the relative times to sunrise and sunset for each tuple. We name the relative times and the night time as time features in our dataset. Because the absolute sunrise/sunset times are not meaningful to the zooplankton, they are not included in the dataset. Table 4.1 lists all

Table 4.1: Features and sources

Category	Features
Water features [2]	conductivity (s/m), temperature (°c), pressure (decibar),salinity (psu), density (kg/m3), oxygen (ml/l), transmission (%)
Land features [3]	temperature (°c), dew point temperature (°c), relative humidity (%), wind direction (10's deg), wind speed (km/h), visibility (km) pressure (kPa), weather
Time features [1]	night time,relative sunrise time, relative sunset time

features we include in the dataset and their sources.

We begin our study by a fixed depth (depth 42m) analysis on the intensity signals and later expand it to the entire water column. At this depth, we average the data at finer time scales (60min, 30min, 15min and 5min) then use two parameters (5 and 15db) to group the intensity signals into three classes: (1) below 5db (class 0) (2) between 5 and 15 db (class 1, zooplankton class) (3) above 15 db (class 2, fish class). Figure 8 in the Appendix illustrates part of the data we collect. Each row represents data collected for one time interval. We call each row an example. Each example has 7 water features (in yellow), 8 land features (in blue), 3 time features (in red) and one class label (in green). If a measurement is missing, it is represented by "?". The classification algorithms we use are able to build models in the presence of missing data.

We want to build a classification model that captures the relationship between environmental and time factors (features) and the intensity distribution (class labels). The model can be used to make predictions on zooplankton distribution for the depth (42m). Similarly, we can build classification models for other depths. Using these models all together, we are able to predict the zooplankton distribution throughout the entire water column. We can plot the results of prediction as a picture and this picture gives us complete information of the predicted zooplankton DVM pattern for the given time period.

In order to evaluate the validity of the classifiers, we divide the entire dataset into two parts. Data from January, April, July and October of 2008 and January of 2009 are used as the training set to build classification models and data from April and July of 2009 are used as the test set for evaluation purpose. The classifiers we build should be able to predict the zooplankton distribution for the test set according to models they learn from the training set.

Before doing classification on the training dataset, however, we first use feature selection techniques to select the best feature subset to be used for classification. The reasons for doing feature selection are:

1. We want to further confirm our results from the previous chapter. The correlation analyses on monthly averaged data show that besides time, the DVM pattern is also correlated with various environmental measurements. We want to confirm this correlation on finer time granularities.

2. We want to know which features are relevant (have correlations) to the class label. Irrelevant

Table 4.2: Search methods

	Search method name	Description
1	Greedy Stepwise	Perform a hill climbing search. The search stops when add/deleting a feature results in decreased evaluation result
2	Best First Search	Combination of hill climbing and backtracking. The subset with the highest evaluation is chosen to expand (or reduce). If the expanding does not result in improvement, the search goes back to the next best unexpanded subset and continues from there
3	Linear Forward Selection [21]	Extension of Best First Search. Restrict the number of attributes being examined.
4	Subset Size Forward Selection [21]	Extension of Linear Forward Selection
5	Exhaustive Search	Search the whole feature space starting from an empty set. The optimal subset is chosen
6	Genetic Search [20]	Use a genetic algorithm for the search
7	Random Search [32]	Hill climbing starting with a randomly selected subset
8	Rank Search	Rank the features using an evaluator first. Starting from the top of the list, subsets of increasing size are evaluated, the best set is reported.
9	Ranker	Rank features according to the individual evaluations

features may confuse the classifier (model) and reduce the classification accuracy.

3. We want to know correlations between features. If two features are strongly correlated, using only one of them should have the same effect as using two (they provide similar information). Introducing redundant features makes the classification process slower without increasing the accuracy.

4. Besides increasing accuracy, the classification process also helps us have a deeper understanding of our dataset. A classification model shows us how the features are correlated and their relationship with the class. Fewer features make it easier for us to interpret and analyze the structure of these relationships. We can use the built models together with feature selection results to find the most relevant environmental factors for zooplankton DVM pattern.

For the purpose of feature selection, we conduct both feature ranking and feature subset selection in Weka [4]. For feature ranking, we use four statistics to rank features. For feature subset selection, we use various search algorithms and evaluation criteria. In order to do feature selection, a search method is used to generate a candidate feature subset from the entire dataset. An evaluation method is then used to evaluate the goodness of the candidate set through various kinds of measurements. The search/evaluate process repeats until the stopping condition is met.

Table 4.2 lists search methods we use in the experiments. Except for ranker, all search methods in the above table are used together with a subset evaluation method. Exhaustive search guarantees to find the optimal subset, but it usually takes a long time to finish. All the other methods are generally faster than exhaustive search, but they can not guarantee that the optimal subset is found.

Table 4.3: Evaluation methods

	Evaluation method name	Description
1	CFS Subset Eval [5]	Correlation based heuristic to determine the goodness of subsets. Subsets whose features are highly correlated with the class and have very low level of inter-correlation are chosen
2	Chi Square Attribute Eval	Evaluate the goodness of individual feature by computing the chi-squared statistic with respect to the class
3	Info Gain Attribute Eval	Evaluate by computing the information gain with respect to the class.
4	Gain Ratio Attribute Eval	Evaluate by computing the gain ratio with respect to the class
5	Symmetrical Uncert Attribute Eval	Evaluate by computing the symmetrical uncertainty with respect to the class
6	One Rattribute Eval	Evaluate by using OneR classifier
7	Consistency Subset Eval [32]	Evaluate the usefulness of subsets by the level of consistency.
8	Symmetrical Uncert Attribute Set Eval [49]	Evaluate by computing the symmetrical uncertainty with respect to another set of attributes
9	Classifier Subset Eval	Evaluate the goodness of subset by a user defined classifier

Usually the locally best subset is reported. Ranker is used together with a feature ranking method to rank all features according to a measured statistic (for example, chi-squared test). The ranking method calculates statistics for features one by one, and the ranker ranks features solely according to the statistics. No inter-feature correlation is considered. No feature subset is reported.

Table 4.3 lists evaluation methods we use in the experiments. Among the various evaluators, Chi Square Attribute Eval, Info Gain Attribute Eval, Gain Ratio Attribute Eval, Symmetrical Uncert Attribute Eval and OneR Attribute Eval are used together with Ranker search to evaluate the features individually, without considering the correlations among them. In spite of their limitations, we still use these individual feature evaluators in our experiments. Because these measurements are very fast and they can provide an insight of relative importance among features. All the other evaluators are used to evaluate feature subsets. They generally take longer time to finish but inter-feature correlations can be considered. Classifier Subset Eval in particular, almost always selects a subset that results in the highest accuracy in the classification process if the same learning algorithm is used for both feature selection and training step.

4.2 Feature Selection and Classification on 60min Averaged Data

We conduct feature selection using various combinations of the above search and evaluation methods on 60 minutes averaged data. The results are shown in Figure 9 and Figure 10 of the Appendix.

Each row in the figure represents one feature selection experiment using a particular combination of search and evaluation methods. The selected features are indicated by “x”. After the best feature subset is chosen, a machine learning algorithm is used to learn a model based on the selected features. The classification accuracies of the built models are also included in the table as a measurement for the goodness of the selected subset for each search/evaluation method.

Table 4.4: Top 10 features for 60 minutes averaged data

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	nightTime	nightTime	nightTime	nightTime
2	waterConductivity	waterConductivity	waterConductivity	waterConductivity	waterConductivity
3	waterTemperature	waterTemperature	waterTemperature	waterTemperature	waterTemperature
4	waterOxygen	waterTransmission	waterOxygen	relativeSunrise	waterOxygen
5	waterTransmission	waterSalinity	waterTransmission	relativeSunset	waterTransmission
6	waterSalinity	waterOxygen	waterSalinity	waterTransmission	waterSalinity
7	relativeSunrise	relativeSunset	relativeSunrise	waterOxygen	relativeSunset
8	relativeSunset	relativeSunrise	relativeSunset	weather	relativeSunrise
9	landTemperature	landTemperature	landTemperature	waterSalinity	landTemperature
10	waterDensity	landPressure	landPressure	landTemperature	landPressure

We first examine the results of individual evaluators. From individual evaluators, top 10 important features are listed in Table 4.4.

According to Table 4.4, all individual evaluator have similar rankings. Especially, night time, water conductivity and water temperature are ranked the top 3 features by all evaluators. Water oxygen, water transmission, water salinity, land temperature, land pressure, two relative times also have high rankings. Water density has high ranking for a certain evaluator. All these evaluators do the evaluation one feature at a time. They only consider correlation between one feature and the class. No correlation between features is considered. Therefore, the above result can only indicate that the above mentioned features are important to the class. There is a possibility that certain feature also has correlation with other features. If a correlation exists between two features, one of the feature is considered redundant, and should not be selected for the classification process.

Since CFS considers both class-feature correlation and feature-feature inter-correlation, its result filters out redundant feature. We use CFS test with various search method and find out that all search methods choose relative times and night time. Water conductivity, water temperature and water transmission are also chosen by certain search methods. No land features are chosen. Comparing results with individual feature evaluator, CFS test confirms that sunrise/sunset times and night time are the most important and non-redundant features. It confirms that time is important. Water conductivity, water temperature and water transmission are also important. This finding confirms that besides time, environmental factors also have impact on the DVM pattern. Moreover, land features are less relevant to the intensity values compared with water features. The average classification accuracy (using BayesNet) is 86.74%.

Unlike individual evaluators, the subset evaluators can select the best (or locally best) feature subset. We tried three subset evaluators in our experiments: (1) Classifier Subset Evaluator (2) Consistency Subset Evaluator (3) Symmetrical Uncert Attribute Set Evaluator. For each evaluator, we also tried a set of search methods. The selected feature subsets and corresponding classification accuracies are shown in Figure 10 of the Appendix. The figure shows that both water features and land features are correlated with the class label. By comparing the classification accuracies, we find that the average accuracies for the three evaluators are 88.05%, 84.09% and 86.76% respectively. Features selected by classification subset evaluators generally result in higher classification accuracies. When decision tree is used as the learning algorithm, the accuracy reaches 89.92%, the highest of our experiments. The selected features by decision tree are: water oxygen, water conductivity, water salinity, water temperature, water transmission, land temperature, hour relative to sunrise, hour relative to sunset and night time.

We further examine the examples where the classifier makes mistakes. After closer analysis, we find that some of the mistakes are made because of the presence of fish school at certain period of time. Figure 4.1 is the plot of intensity values on July 2nd, 2008. Shortly after sunrise (at around 800), fish schools appear in the middle depth (the depth we are interested in). Some of the intensity values of the fish school are similar to those of zooplanktons (5-15db). The presence of fish school decreases the classification accuracy.

Figure 4.2 is the plot of intensity values at depth 42m. Depth 350 in the picture means depth index of 350 (the whole depth is divided into 800 bins).

When 60 minutes averages are taken, Figure 4.3 shows the plot.

Figure 4.4 shows the plot of true class labels produced from the averaged data. The black part represents class 0 and the white part represents class 1 (zooplankton class). No class 2 examples exist. The picture shows three separate white parts. The first two happen approximately between sunset and sunrise times, which correctly reflects the fact that zooplankton pass by this depth during their migration around sunrise/sunset times. The third white part is very suspicious. Comparing with the original plot of data, we find that these signals are mainly caused by fish schools. Since the intensities of the fish schools also fall into the zooplankton class range (5-15db), we label them as zooplankton signals but in fact, they should belong to the fish class.

Figure 4.5 shows the prediction of a classifier. We use ClassifierSubsetEval as the evaluator and BestFirst as the search method. BayesNet is used as the learning algorithm. The accuracy is measured 87.15% by 10 fold cross validation. The classifier labels examples between sunrise and sunset as the zooplankton class, and other examples as the non-zooplankton class. Because the classifier fails to predict the zooplankton-like fish school, its accuracy is reduced when comparing with the true class label although the prediction correctly reflects the true zooplankton distribution.

The fish schools appear frequently for data from the month of July. Their presence introduces errors to the true class labels but the classifier is still able to make correct prediction. Similar

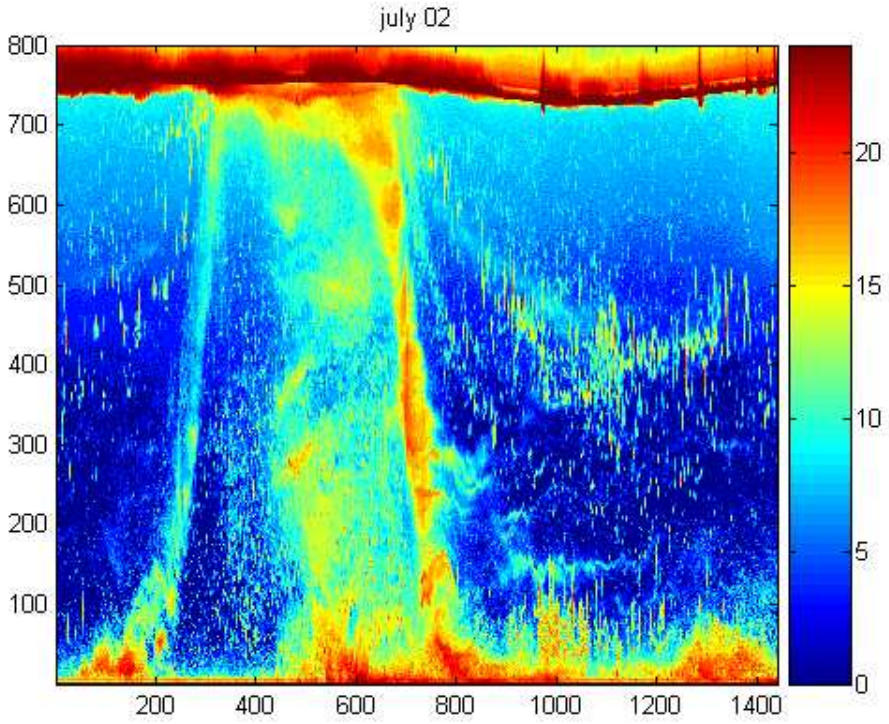


Figure 4.1: Raw data plot for July 2nd, 2008

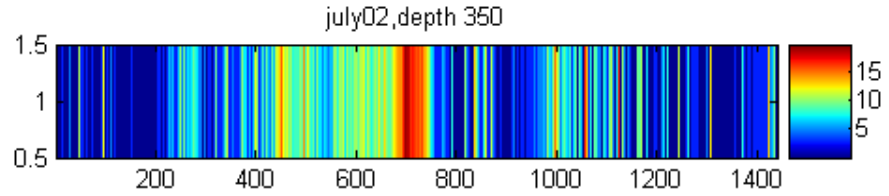


Figure 4.2: Plot of intensity values at depth 42m

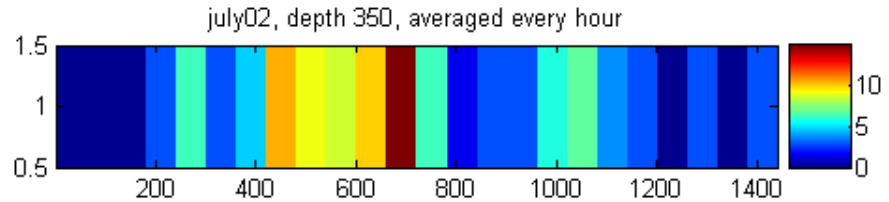


Figure 4.3: Plot of 60 minutes averaged intensity values at depth 42m

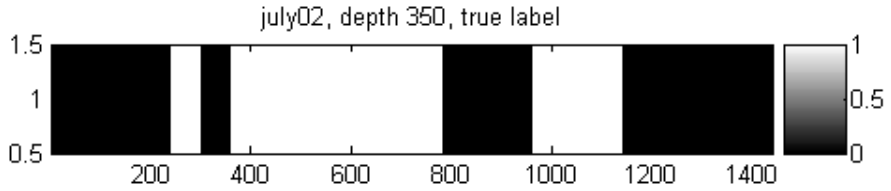


Figure 4.4: True labels at depth 42m

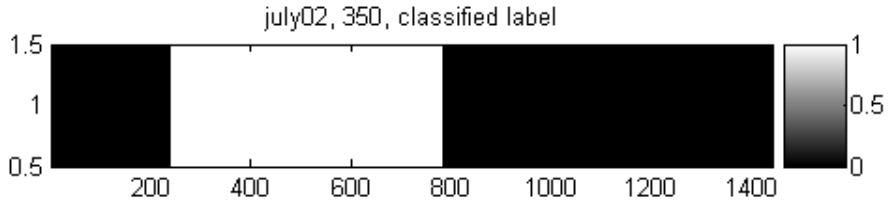


Figure 4.5: Predicted labels at depth 42m

situations happen in the month of April, when objects with the same intensity range (5-15db) stay in the middle depth of the water column for long periods of time. Figure 4.6 is one example of the case in April. Due to a lack of domain knowledge, we do not know the cause of these signals. They can be caused by other floating objects, or by different types of zooplankton. Since these objects stay in the same depth all day long without diel vertical migration, we are not interested in them and simply treat them as noise data. The noise data confuse our labeling process and reduce the classification accuracy.

Moreover, since we average the data every 60 minutes, short periods of zooplankton signals may not be correctly recorded in the labels. Figure 4.7 shows plot of intensity values on Jan 17, 2008.

The plot of intensity values for depth 42m is illustrated in Figure 4.8.

When we average the signals for every 60 minutes and do the labelling, all examples belong to class 0 (0-5db).

Figure 4.9 shows the prediction of a classifier.

According to the class labels, all examples in the white parts are considered wrong. In fact, however, most of these examples make good predictions.

Because of the presence of noise data, and the fact that we take a 60min average on the data, the class labels we create for training data may not accurately reflect the true distribution of zooplankton in the water column. If we could have an accurate dataset, we expect our accuracy would be significantly higher.

One the other hand, however, the classifier built for 60min averaged data is of limited usefulness because it can only predict the average situation in each hour. Observing the intensity plot, we find that zooplankton usually spend several minutes in this depth and then move up or down. When an average signal is taken for each hour, the short zooplankton signals can be lost. We can capture the

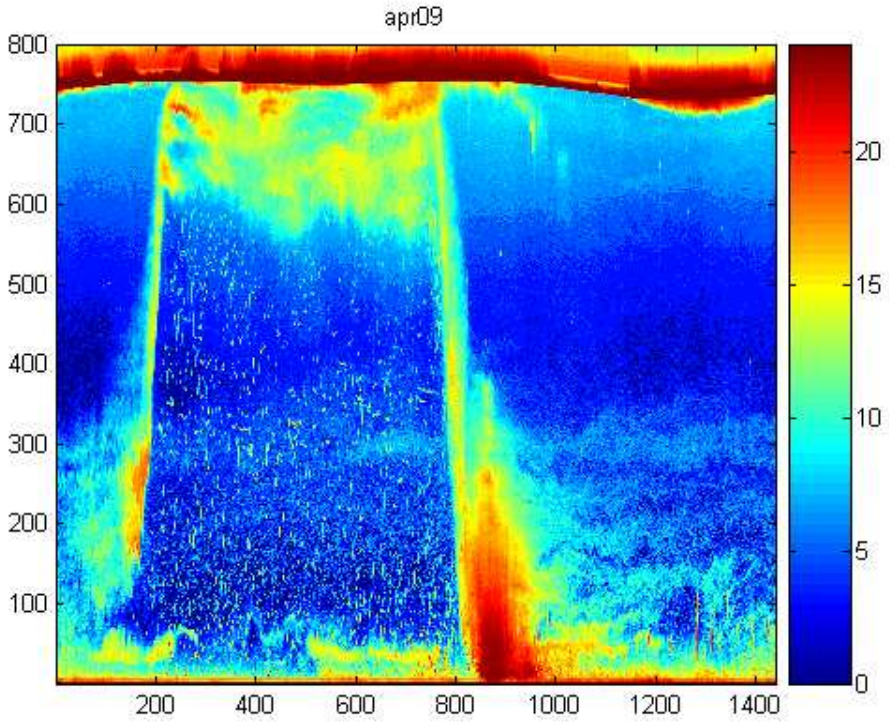


Figure 4.6: Raw data plot for April 9th, 2008

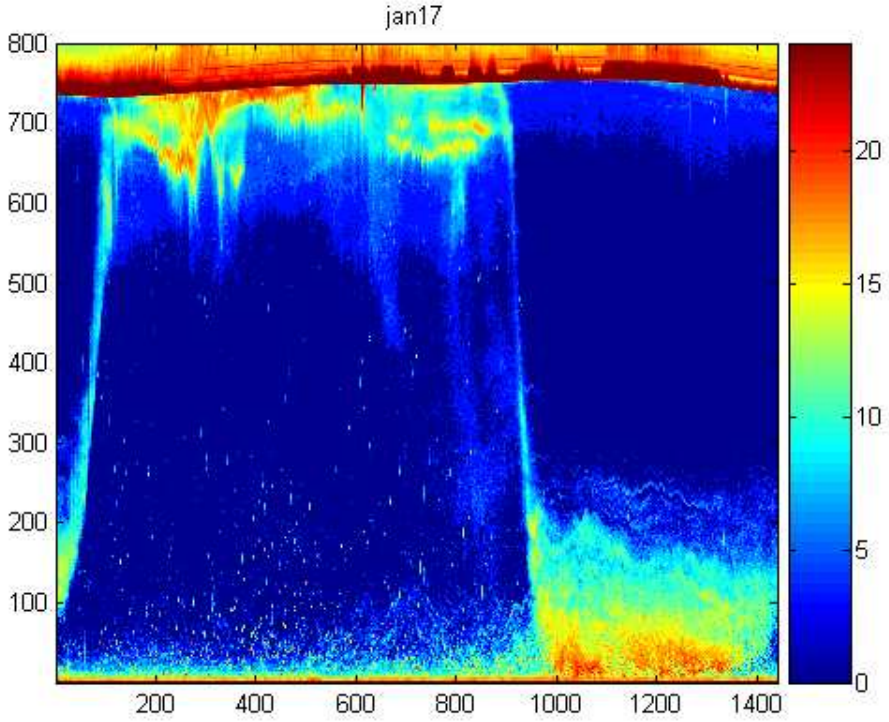


Figure 4.7: Raw data plot for January 17th, 2008

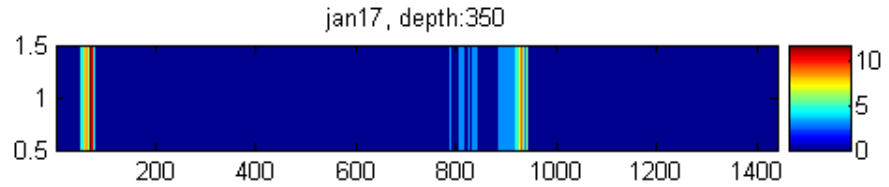


Figure 4.8: Plot of intensity values for depth 42m

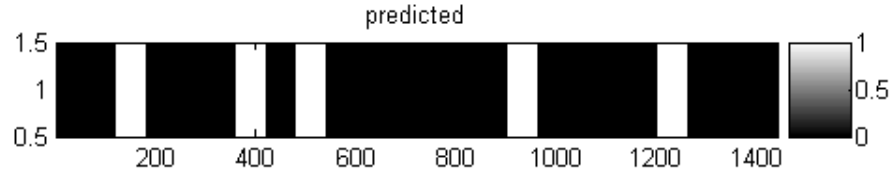


Figure 4.9: Plot of intensity values for depth 42m

Table 4.5: Top 10 features for 15 minutes averaged data

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	nightTime	nightTime	nightTime	nightTime
2	waterConductivity	waterConductivity	waterConductivity	waterConductivity	waterConductivity
3	waterTemperature	waterTemperature	waterTemperature	waterTemperature	waterTemperature
4	waterOxygen	waterSalinity	waterOxygen	relativeSunrise	waterSalinity
5	waterSalinity	relativeSunrise	relativeSunset	waterTransmission	waterOxygen
6	waterTransmission	waterTransmission	waterSalinity	waterSalinity	relativeSunrise
7	relativeSunset	waterOxygen	relativeSunrise	waterOxygen	waterTransmission
8	relativeSunrise	relativeSunset	waterTransmission	relativeSunset	relativeSunset
9	waterDensity	waterDensity	waterDensity	waterPressure	waterDensity
10	waterPressure	waterPressure	waterPressure	waterDensity	waterPressure

presence of zooplankton only if the feature selection and classification process is conducted on finer time scales (5min, 15min, etc).

4.3 Feature Selection and Classification on Finer Time Scales

Figure 11 and Figure 12 in the Appendix show the selected feature subsets and the classification accuracies for 15 minutes averaged dataset (Since the land features we can get are sampled every 60 minutes, we do not include them in our later experiments).

Table 4.5 shows feature rankings by various feature evaluators.

Comparing with feature ranking for 60 minutes averaged data, the rankings don't change too much. Night time, water conductivity and water temperature are still the most important features.

In order to verify the usefulness of various features, we also produce a random feature from uniform distribution and add it to the dataset. We rank all features using various measures. If a feature is ranked after the random feature, it is not useful and should be discarded. We use the following statistics to rank the features: Chi Squared, Gain Ratio and Symmetrical Uncert.

Table 4.6 illustrates the rankings.

According to the table, the random feature ranks as the last one for all statistics, which means all other features have stronger correlations with the class labels. The relative rankings among features do not change.

CFS tests also show similar results, marking night time, relative times and water conductivity the most important features. The average classification accuracy (using BayesNet) is 85.2%.

For feature subset selection, various classification algorithms are tried as the evaluator. The average accuracies for Classifier Subset Evaluator, Consistency Subset Evaluator and Symmetrical

Table 4.6: Feature rankings after a random feature is added

Ranking	ChiSquare	GainRatio	SymmetricalUncert
1	nightTime	nightTime	nightTime
2	waterConductivity	waterConductivity	waterConductivity
3	waterTemperature	waterTemperature	waterTemperature
4	waterOxygen	waterSalinity	waterSalinity
5	waterSalinity	relativeSunrise	waterOxygen
6	waterTransmission	waterTransmission	relativeSunrise
7	relativeSunset	waterOxygen	waterTransmission
8	relativeSunrise	relativeSunset	relativeSunset
9	waterDensity	waterDensity	waterDensity
10	waterPressure	waterPressure	waterPressure
11	random	random	random

Table 4.7: Top 10 features for 5 minutes averaged data

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	nightTime	nightTime	nightTime	nightTime
2	waterConductivity	waterConductivity	waterConductivity	waterTemperature	waterConductivity
3	waterTemperature	waterTemperature	waterTemperature	waterConductivity	waterTemperature
4	waterOxygen	waterOxygen	waterOxygen	waterTransmission	waterOxygen
5	waterSalinity	relativeSunrise	relativeSunset	relativeSunrise	relativeSunrise
6	relativeSunset	relativeSunset	waterSalinity	waterSalinity	relativeSunset
7	relativeSunrise	waterSalinity	relativeSunrise	relativeSunset	waterSalinity
8	waterTransmission	waterTransmission	waterTransmission	waterPressure	waterTransmission
9	waterDensity	waterDensity	waterDensity	waterOxygen	waterDensity
10	waterPressure	waterPressure	waterPressure	waterDensity	waterPressure

Uncert Attribute Set Evaluator are 88.01%, 83.03% and 84.54% respectively. Classifier Subset Evaluator together with J48 as the mining algorithm still gives the highest accuracy (90.68%) and is fairly fast.

We also conduct feature selection on 5 minutes (Figure 13 in the Appendix) and 30 minutes (Figure 14 in the Appendix) averaged data. Similar results are found.

Table 4.7 shows feature rankings for 5 minute averaged data.

The feature rankings for 30 minutes averaged data are shown in Table 4.8.

Comparing feature rankings for different time scales, we find that rankings from Chi Squared test, information gain ratio, information gain and symmetrical uncertainty stay almost unchanged. The most important features are: night time, water conductivity, water temperature. Water density and water pressure are less important.

The inter-feature correlations can be found by the results of the CFS evaluator. CFS tests of all the three scales (5 minutes, 15 minutes and 30 minutes) choose times relative to sunrise/sunset

Table 4.8: Top 10 features for 30 minutes averaged data

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	nightTime	nightTime	nightTime	nightTime
2	waterTemperature	waterConductivity	waterConductivity	waterConductivure	waterConductivity
3	waterConductivity	waterTemperature	waterTemperature	waterTemperature	waterTemperature
4	waterOxygen	relativeSunset	waterOxygen	relativeSunrise	relativeSunset
5	waterSalinity	waterSalinity	waterSalinity	waterSalinity	waterOxygen
6	waterTransmission	relativeSunrise	waterTransmission	waterOxygen	waterSalinity
7	relativeSunrise	waterOxygen	relativeSunset	waterTransmission	relativeSunrise
8	relativeSunset	waterTransmission	relativeSunrise	relativeSunset	waterTransmission
9	waterDensity	waterDensity	waterDensity	waterDensity	waterDensity
10	waterPressure	waterPressure	waterPressure	waterPressure	waterPressure

and night time, confirming that times are important factors for zooplankton migration pattern. For 15 minutes data, water conductivity is also chosen. By checking correlations between water conductivity and other features, we find that water conductivity is highly correlated with the following features: water oxygen, water salinity and water temperature. This correlation explains the fact that water conductivity is chosen, but water oxygen and water temperature are not, even if the latter two factors are proven to have impact on zooplankton's migration [34].

For 5 minutes averaged data, the average accuracies for Classifier Subset Evaluator, Consistency Subset Evaluator and Symmetrical Uncert Attribute Set Evaluator are 87.76%, 83.91% and 86.91% respectively. The highest accuracy is 90.32%, which is achieved by the Classifier Subset Evaluator with J48 as the mining algorithm.

For 30 minutes avaraged data, the average accuracies for Classifier Subset Evaluator, Consistency Subset Evaluator and Symmetrical Uncert Attribute Set Evaluator are 87.11%, 84.85% and 84.29%. The highest accuracy is 89.39%, which is also achieved by the Classifier Subset Evaluator with J48 as the mining algorithm. Figure 4.10 shows part of the decision tree. By visualizing the deduced tree, we find that the tree has a very complex structure. Each branch of the tree uses a different set of features to make decisions. On the other hand, relative times and night time are almost always placed in higher levels of the tree. These features are used to make the initial decisions. Water features are placed in lower levels of the tree. We choose to use J48 as the classification algorithm for further experiments.

The above results confirm that environmental factors have strong correlations with the migration of zooplankton. The strong correlations we find for monthly averaged data are caused by both time and environmental measurements. Among all factors, water oxygen, water temperature, relative sunrise/sunset times and night time are the most relevant. Water conductivity is also important

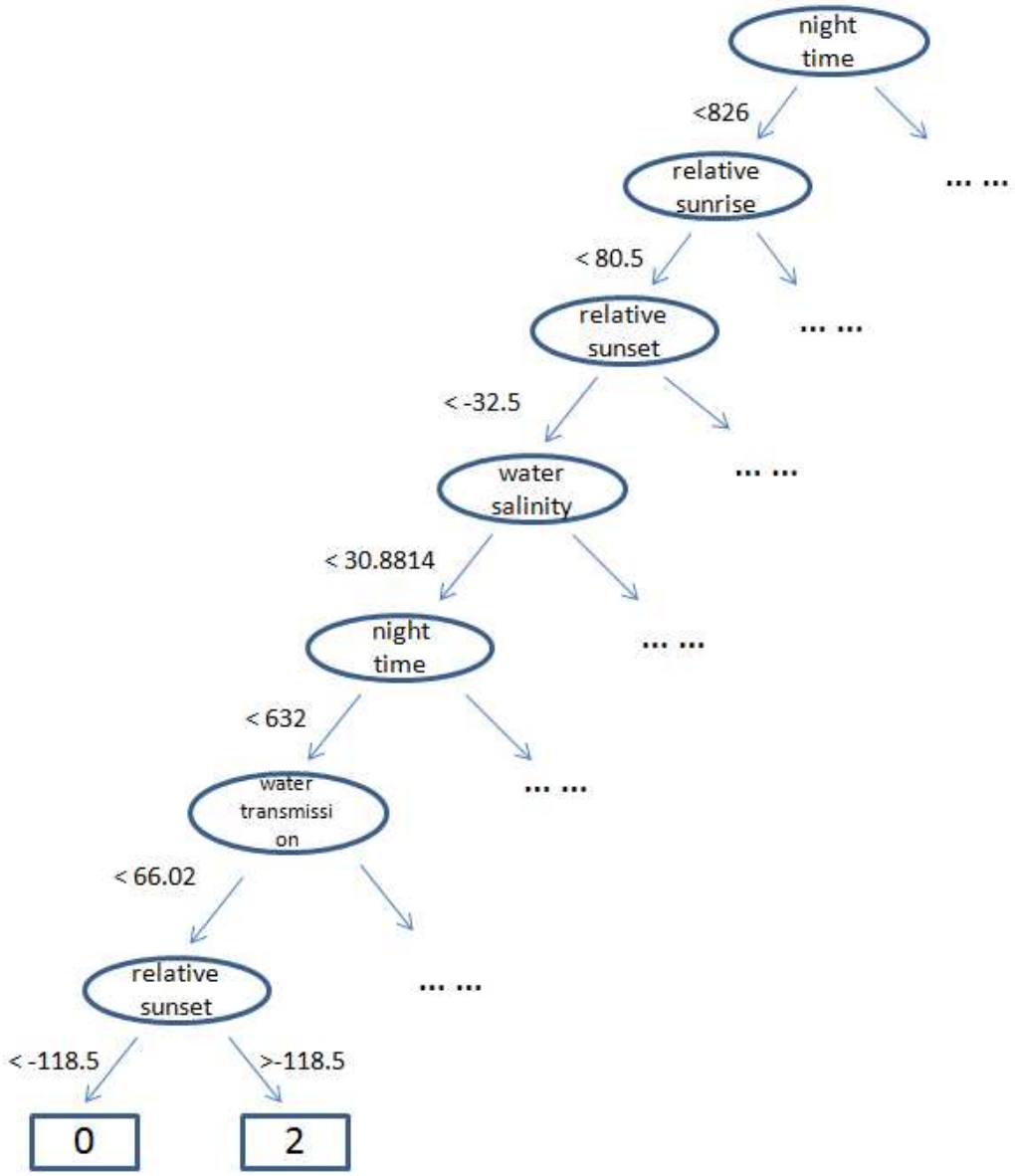


Figure 4.10: Part of the decision tree for 30min averaged data

Table 4.9: Feature rankings at depth 66m (depth index 550)

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	relativeSunset	nightTime	relativeSunrise	relativeSunset
2	relativeSunset	nightTime	relativeSunset	relativeSunset	nightTime
3	relativeSunrise	relativeSunrise	relativeSunrise	nightTime	relativeSunrise
4	waterConductivity	waterConductivity	waterTemperature	waterTemperature	waterConductivity
5	waterTemperature	waterTemperature	waterConductivity	waterConductivity	waterTemperature
6	waterOxygen	waterOxygen	waterOxygen	waterDensity	waterOxygen
7	waterDensity	waterDensity	waterDensity	waterPressure	waterDensity
8	waterPressure	waterPressure	waterPressure	waterTransmission	waterPressure
9	waterSalinity	waterSalinity	waterSalinity	waterOxygen	waterSalinity
10	waterTransmission	waterTransmission	waterTransmission	waterSalinity	waterTransmission

because it is correlated to temperature and oxygen. The relevant importance of these factors is shown in both feature rankings and the deducted decision tree.

4.4 Feature selection on other depths

In order to further confirm our findings, we also conduct feature selection on 15 minutes averaged data for the following depths: 66m (depth index is 550 out of 800, near surface), 54m (depth index: 450) and 30m (depth index: 250, near seabed). Results are shown in Figure 11 and Figure 12 of the Appendix.

Table 4.9 - 4.11 show the feature rankings for various depths.

Comparing results from different depths, we find:

1. At higher levels (depth 550 and 450, near surface), times have higher rankings. At lower level (depth 350 and 250, near seabed), water features are more important. These rankings are consistent with what we expected, because the instrument is located on the seabed. The measured water feature values should represent the situations deep in the sea better than those near the surface.

2. Time features, water conductivity, temperature and oxygen always have higher rankings. The relative rankings among the water features are almost unchanged. Conductivity, temperature and oxygen almost always have higher rankings than other water features. These results further confirm these features are the most important among all features we collect.

3. Results of CFS tests also confirm that times are more important than water conditions at higher level (near surface). Time features are selected at all depth, they are always important.

Table 4.10: Feature rankings for depth 54m (depth index 450)

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	relativeSunset	nightTime	nightTime	relativeSunset
2	relativeSunrise	relativeSunrise	relativeSunset	waterPressure	relativeSunrise
3	relativeSunset	nightTime	relativeSunrise	waterTransmission	nightTime
4	waterConductivity	waterConductivity	waterConductivity	relativeSunrise	waterConductivity
5	waterDensity	waterTemperature	waterTemperature	waterSalinity	waterTemperature
6	waterOxygen	waterOxygen	waterDensity	waterConductivity	waterOxygen
7	waterTemperature	waterDensity	waterOxygen	relativeSunset	waterDensity
8	waterSalinity	waterSalinity	waterSalinity	waterTemperature	waterSalinity
9	waterPressure	waterPressure	waterPressure	waterDensity	waterPressure
10	waterTransmission	waterTransmission	waterTransmission	waterOxygen	waterTransmission

Table 4.11: Feature rankings for depth 30m (depth index 250)

	ChiSquare	GainRatio	InfoGain	OneR	SymmetricalUncert
1	nightTime	nightTime	nightTime	nightTime	nightTime
2	waterConductivity	waterConductivity	waterConductivity	waterConductivity	waterConductivity
3	waterTemperature	waterTemperature	waterTemperature	relativeSunrise	waterTemperature
4	relativeSunrise	waterOxygen	relativeSunrise	waterTemperature	waterOxygen
5	waterOxygen	relativeSunrise	waterOxygen	waterSalinity	relativeSunrise
6	relativeSunset	waterSalinity	relativeSunset	waterOxygen	relativeSunset
7	waterSalinity	relativeSunset	waterSalinity	waterTransmission	waterSalinity
8	waterTransmission	waterTransmission	waterTransmission	relativeSunset	waterTransmission
9	waterDensity	waterDensity	waterDensity	waterPressure	waterDensity
10	waterPressure	waterPressure	waterPressure	waterDensity	waterPressure

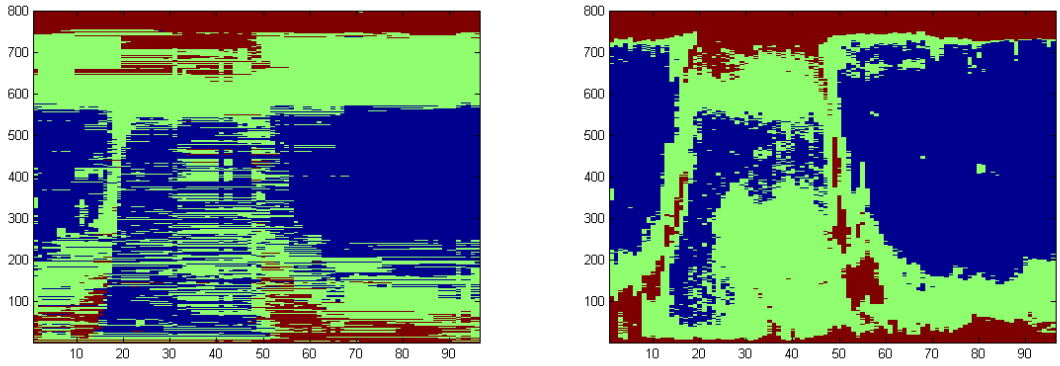
4. At lower levels, water transmission and water salinity have higher rankings compared with higher levels. At lower levels, water density has lower rankings compared with higher levels. One possible way to explain would be water transmission and water salinity change more rapidly along the water column. Therefore, the real water transmission at depth 550, for example, is very different from that at depth 250 (which is much closer to the instrument). Because of the limitation of the instrument, we don't have a way to confirm this hypothesis.

5. The deduced decision trees have similar structures. Comparing trees for different depths, we find that although the trees are not exactly the same, relative times and night time are almost always placed in higher levels. These features are used to make the initial decisions. Water features are placed in lower levels of the trees.

4.5 Predicting zooplankton migration pictures

We apply the above learning process to all depths along the water column for the training set. For each depth, water features and time features are collected, the time series is divided into 15-minute time intervals and the class labels are given using the same parameters (5 and 15db). We use the training set to build a classifier for each of the depths along the water column. J48, a typical implementation of decision tree learning algorithm [4], is used as our classification model. We evaluate this model in terms of classification accuracy via performing ten-fold cross validation ten times. In each time, the data set is divided into ten subsets, in which one of the ten subsets is used as the test set and the other nine subsets are put together to form a training set. Then the average classification accuracy across all ten trials is computed. The advantage of this method is that every instance gets to be in a test set exactly once, and gets to be in a training set nine times. The variance of the resulting estimate is reduced as the number of folds is increased. When classifiers are built, we use them to make predictions on the test set and plot the predicted class memberships as a picture. Figures 4.11 shows the predicted picture and the corresponding true picture for July 26, 2009. We include more pictures in Appendix Part G.

Comparing predictions with the true plots, we can see that the predictions are very similar to the true situations. Though sharing similar shapes, the migration path of each day is a little different from those of neighbouring days. For early April, the zooplankton signals are not very intense, so do the predictions; for middle April, the signals become more intense, and the predictions reflect this change. For late April, the signals become less intense again, and the predictions change accordingly. For July, the signals become more and more intense from early July to late July, the predictions also follow this trend. Overall speaking, the predicted pictures are very similar to the true plots. The zooplankton DVM pattern can be clearly seen from the predicted picture.



(a) Prediction. The predicted existence of zooplankton is shown in green

(b) True label

Figure 4.11: Plots of predicted and true class labels for July 26, 2009

4.6 Comparing Classification Pictures with Almost Periodic Curves

In this research study, two methods are used to model the zooplankton distribution along the water column based on the same dataset collected in the same study areas. The almost periodic curves provide a high level description of the dynamics of zooplankton distribution through a long period of time. The classification pictures, on the other hand, provide much more detailed predictions on finer time scales. Although served for different purposes, both methods are able to make predictions based on a set of related features. We want to compare these two methods in regard to accuracy and usefulness.

Each period of the almost periodic curve depicts the averaged situation for a month. On the other hand, when using classifier to make predictions, they are able to make predictions for each day. In order to compare the two methods, we first calculate a monthly averaged classification picture by taking the median values for predictions of a month. Figure 4.12 illustrates the predicted picture (averaged) together with the predicted upper and lower bounds for April 2009.

In Figure 4.12, the black curve represents the predicted upper bound and the blue curve represents the lower bound. The predicted existence of zooplankton is illustrated in green. From the picture, we can see that the path depicted by the curve deviates in some places from the predicted picture. In order to decide which prediction is closer to the true situation, we also compare the two types of predictions with the ground truth in the following pictures.

Figure 4.13 shows the upper and lower bounds together with the true data plot.

Figure 4.14 shows the difference between predicted class membership and the true situation. The pink dots illustrate the difference/errors.

From Figures 4.13 and 4.14, we can see that the classification picture is much closer to the true situation. The general shape of the migration path is correctly predicted, with most of the errors

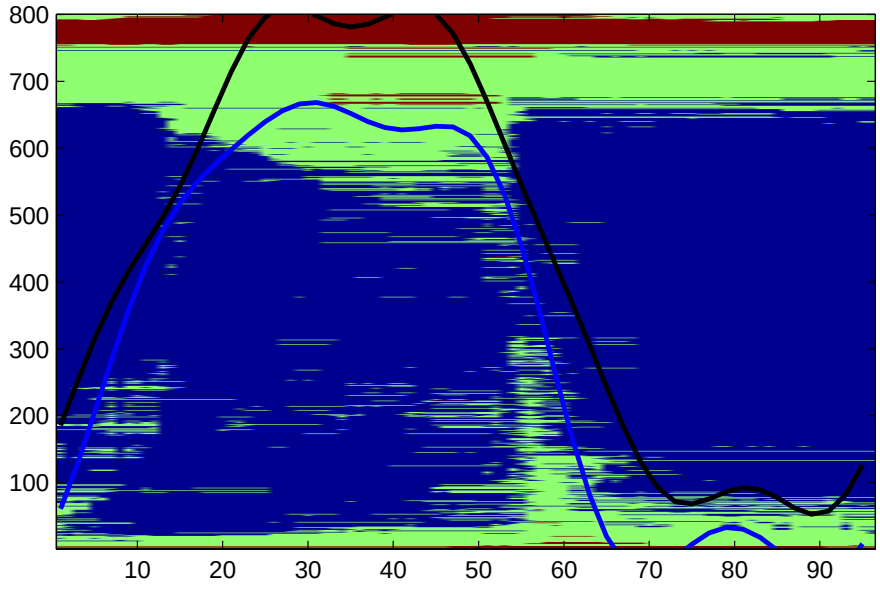


Figure 4.12: The predicted picture and the predicted upper and lower bounds for April, 2009. The black curve represents the predicted upper bound and the blue curve represents the lower bound. The predicted existence of zooplankton is illustrated in green

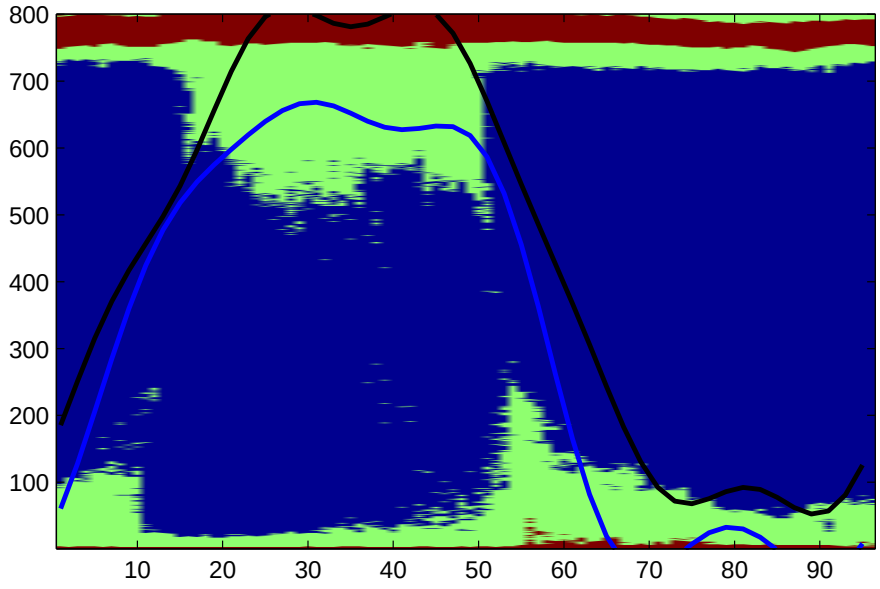


Figure 4.13: Predicted upper and lower bounds and the true data

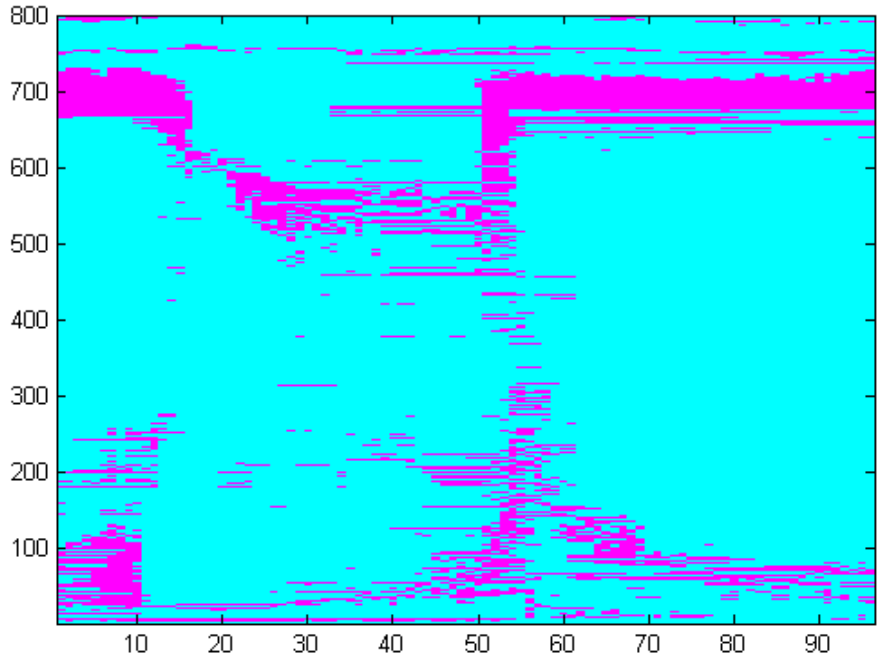


Figure 4.14: Classification results together with the true data. The pink dots illustrate the errors

occurring at the seabed and surface. The almost periodic curve, on the other hand, has relatively larger deviations.

We further conduct comparison in July. The experiments we conduct confirm that classification pictures are closer to the ground truth than the almost periodic curve. Figures 4.15 - 4.17 show the predictions and their comparison with the truth in July.

Although the almost periodic curve has more variances than the classification pictures, we still think it is a useful method for modeling the migration path. The almost periodic curve is mainly used to show the general graduations of the migration paths for a long period of time. It allows us to get a general idea of the "whole picture" and get a sense of how the curve changes over time. The classification pictures, on the other hand, are more suitable for finer time scale explorations, with detailed predictions for each day. They are useful for cases when an accurate, fine scale exploration is desired. The two methods serve for different purposes and both have values in modeling zooplankton distribution.

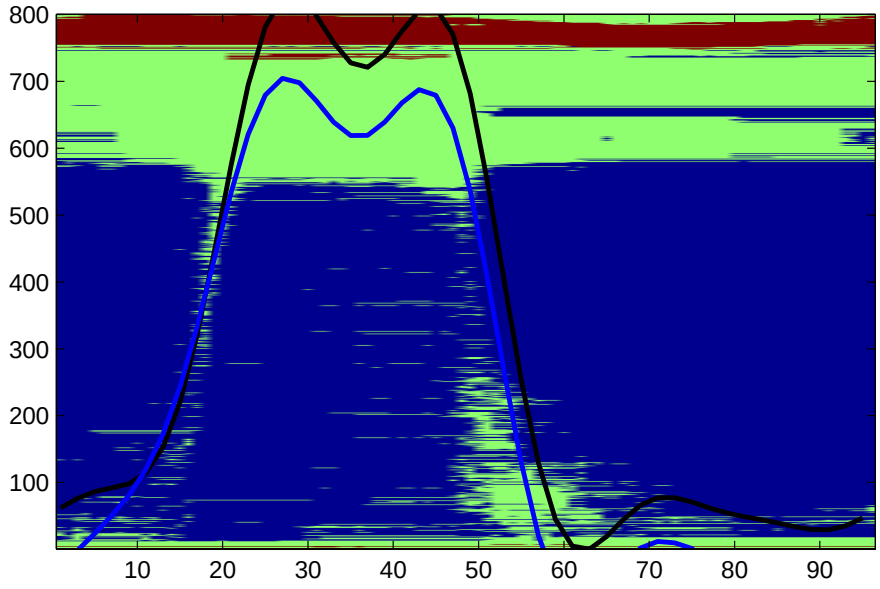


Figure 4.15: Plots of predicted class labels and the predicted curves. The black curve represents the predicted upper bound and the blue curve represents the lower bound. The predicted existence of zooplankton is illustrated in green

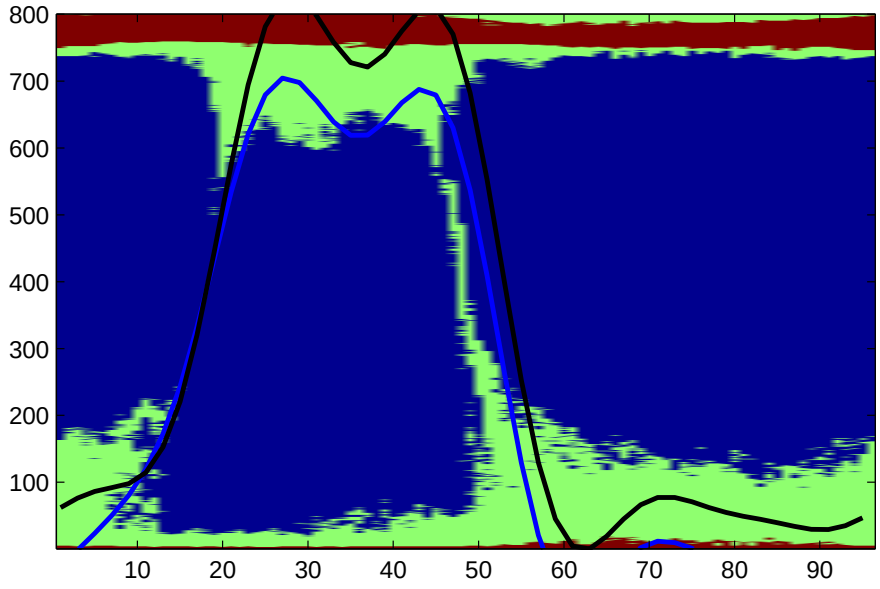


Figure 4.16: Plots of true class labels and the predicted curves

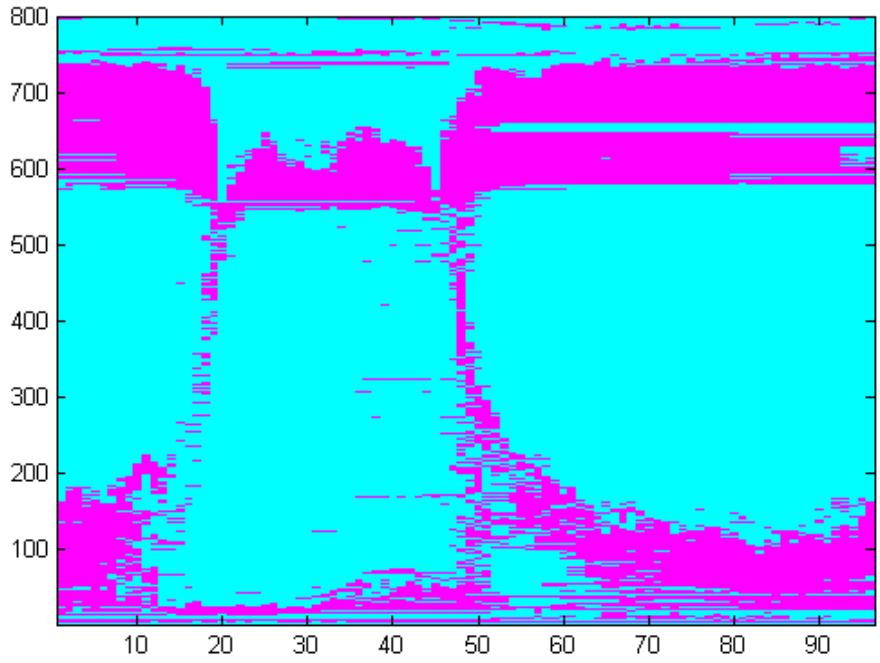


Figure 4.17: Difference between the predicted and true labels. The pink dots illustrate errors

Chapter 5

Conclusion

5.1 Contribution

Zooplankton play an important role in the oceanic food web. Studying the behaviour of zooplankton helps us understand the functioning of marine ecosystem better. In this research study, we use machine learning methods to model and predict zooplankton diel vertical migration patterns at different time granularities.

We first use an Almost Periodic Function to model the DVM pattern on a large time granularity. Acoustic backscatter data of 19 months are used to generate a monthly averaged migration path through a semi-automatic process. We use curve fitting methods to fit the migration data to the mathematical model. The almost periodic model allows us to model the DVM pattern into a generalized Fourier series whose coefficients are not constant, but functions of time. A further correlation analysis between 15 environmental measurements and the Fourier coefficients reveals that the coefficients are correlated with not only time, but environmental measurements as well. Therefore, we use environmental measurements as parameters in the functions of coefficients. The values of the environmental factors change over time, so do the coefficients. The functions of coefficients are later included into the Fourier series representation of the DVM pattern. We compare Fourier series with different terms (1 term to 8 terms) and find that the 5-term Fourier series gives the smallest error. These mathematical models clearly show the almost periodicity of the DVM pattern and can be used to predict further DVM behaviours on large time scales.

Besides modeling the general DVM pattern using monthly averaged data, we also model the pattern on finer time scales using supervised learning algorithms. Before the learning process, we first use feature selection algorithms to examine the correlation between the DVM pattern and various features on finer time granularities (60 minutes, 30 minutes, 15 minutes and 5 minutes) and various depths (near surface, middle depth and near seabed). We use acoustic backscatter data of seven months and 18 environmental measurements to conduct correlation analyses. 5 individual feature evaluators are used and the results are compared. The feature rankings produced by the evaluators are very similar. Night time, water conductivity and water temperature are ranked the top 3 features by all evaluators. For feature subset evaluation, we use 3 evaluators and 8 search methods to conduct experiments. The results show that besides time, both water and land measurements have impact on the DVM pattern. The relative importance of the measurements is shown in both feature rankings and feature subset selection results. Among all measurements, sunrise/sunset times, water temperature, water oxygen and water conductivity are the most important factors. This result further confirms our findings on monthly averaged data, and provides valid reasons for using environmental measurements as features in the learning process. For the learning process, we use decision tree as the classification algorithm and build a classifier for each depth along the water column on 15 minutes averaged data. The classifiers are able to predict the existence of zooplankton in the future. Combining predictions from all classifiers, we get a predicted migration picture. The prediction is very close to the true zooplankton distribution and clearly shows the DVM pattern on finer time

scales.

Both the almost periodic function and the classifiers are useful tools for modeling and predicting the DVM pattern. The almost periodic function is more suitable for modeling the general behaviour of the DVM pattern on a large time scale. The model allows us to explore the DVM pattern in a systematic way and make predictions on large time granularities. The classifiers, on the other hand, are more suitable for finer scale analyses. They provide a way of accurately predict the DVM pattern on small time granularities. Combining both tools, domain experts may be able to study the DVM pattern in an easier and more systematic way.

5.2 Future work

5.2.1 Calibration

The ZAP intensity data we use have not been calibrated. Calibration requires net samplings and further processes to correct what the instrument measures. The calibrated data will be closer to the real situation and we can use the calibration information to generate the biomass medians of zooplankton. The biomass medians should be a better representation of the migration path than the current paths we use from the semi-automatic process. The models learnt from the biomass medians should also represent the real situation better.

5.2.2 Integrating migration path models

In this research study, we use almost periodic functions to model the upper bound, median depth and lower bound of the DVM pattern. Using all of the models, we are able to explore the pattern both temporally and spatially. If the models are integrated into one model, it will allow examining the pattern in more compact way. Moreover, the compact model may provide additional information through the parameters.

5.2.3 Interaction between fish schools and zooplankton

In this research study, we mainly focus on the DVM behaviour of zooplankton. From the intensity plots, we can also see fish schools clearly. The zooplankton seem to avoid fish schools by descending to the seabed and during daytime, fish schools also seem to follow the zooplankton for certain months. The interaction between fish schools and zooplankton may be an interesting topic that worth further exploration.

5.2.4 Reasoning the environmental impact

This research study finds that several environmental and time measurements have strong correlations with the distribution of zooplankton. The reasons for these correlations are out of the scope of this thesis but may be interesting to domain experts.

Bibliography

[1] NRC website. http://www.nrc-cnrc.gc.ca.

[2] Venus website. http://venus.uvic.ca/.

[3] Weather office website. http://www.weatheroffice.gc.ca.

[4] Weka website. http://www.cs.waikato.ac.nz/ml/weka/.

[5] Mark A.Hall and Lloyd A. Smith. Practical feature subset selection for machine learning. *Proceedings of the 21st Australasian Computer Science Conference ACSC 98'*, pages 181–191, 1998.

[6] Carin J. Ashjian, Sharon L. Smith, Charles N. Flagg, and Nasseer Idrisi. Distribution, annual cycle, and vertical migration of acoustically derived biomass in the arabian sea during 1994-1995. *Deep-Sea Research (Part 2, Topical Studies in Oceanography)*, 49(12):2377–2402, 2002.

[7] R. Audi. *The Cambridge Dictionary of Philosophy*. Cambridge University Press, Cambridge, 1999.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[9] Avrim L. Blum and Pat Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, December 1997.

[10] G. Brassard and P. Bratley. *Fundamentals of Algorithms*. Prentice Hall, 1996.

[11] R. Caruana and D. Freitag. Greedy attribute selection. *Proceedings of the 11th International Conference on Machine Learning*, pages 28–36, 1994.

[12] Christopher Clapham and James Nicholson. *The Concise Oxford Dictionary of Mathematics*. Oxford University Press, 2005.

[13] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. *Proceedings of the 18th International Conference on Machine Learning*, pages 74–81, 2001.

[14] M. Dash, K. Choi, P. Scheuermann, , and H. Liu. Feature selection for clustering-a filter solution. *Proceedings of the IEEE International Conference on Data Mining*, pages 115–122, 2002.

[15] J. Doak. An evaluation of feature selection methods and their application to computer security. Technical report, University of California at Davis, Department of Computer Science, 1992.

[16] J.G. Dy and C.E. Brodley. Feature subset selection and order identification for unsupervised learning. *Proceedings of the 17th International Conference on Machine Learning*, pages 247–254, 2000.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[18] S. Fielding, G. Griffiths, and H. S. J. Roe. The biological validation of adcp acoustic backscatter through direct comparison with net samples and model predictions based on acoustic scattering models. *ICES Journal of Marine Science*, 61(2):184–200, April 2004.

[19] K. S. Fu and T. Y. Young. *Handbook of Pattern Recognition and Image Processing*. Academic Press, 1986.

[20] D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley publisher, 1989.

[21] M. Gutlein, E. Frank, M. Hall, and A. Karwath. Large scale attribute selection using wrappers. *Proceedings of IEEE Symposium on Computational Intelligence and Data Mining*, pages 332–339, 2009.

[22] M.A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings on the 17th International Conference on Machine Learning*, pages 359–366, 2000.

[23] Jiawei Han and Micheline Kamber. *Data Mining: concepts and techniques*. Morgan Kaufmann Publisher, 2001.

[24] J. A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

[25] M Kahrs and K. Brandenburg. *Applications of Digital Signal Processing to Audio and Acoustics*. Kluwer Academic Publisher Norwell, MA, USA, 1998.

[26] Y. Kim, W. Street, and F. Menczer. Feature selection for unsupervised learning via evolutionary search. *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 365–369, 2000.

[27] R. Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

[28] R. Kohavi and G.H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[29] P. Langley and S. Sage. Induction of selective bayesian classifiers. In *UAI*, 1994.

[30] D. D. Lemon, M. R. Clarke, J. F. Gower, and M. V. Trevorrow. The acoustic water column profiler: a tool for long term monitoring of zooplankton populations. *Proc. MTS/IEEE International Conf. Oceans*, 3:1904–1909, 2001.

[31] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining*. Kluwer Academic Publishers, MA, USA, 1998.

[32] H. Liu and R. Setiono. A probabilistic approach to feature selection-a filter solution. *Proceedings of the 13th International Conference on Machine Learning*, pages 319–327, 1996.

[33] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17:491–502, April 2005.

[34] Andreas Lork, Daniel F. McGinnis, Piet Spaak, and Alfred Wueest. Acoustic observation of zooplankton in lakes using a doppler current profiler. *Freshwater Biology*, 49(10):1280–1292, October 2004.

[35] Elizabeth Martin and Robert Hine. *A Dictionary of Biology*. Oxford University Press, 2008.

[36] M.J.Gibbons. An introduction to the zooplankton of the benguella current region. http://hdl.handle.net/1834/1252, 1997.

[37] P.M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26(9):917–922, September 1977.

[38] A.Y. Ng. On feature selection: Learning with exponentially many irrelevant features as training examples. *Proceedings of the 15th International Conference on Machine Learning*, pages 404–412, 1998.

[39] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, 1988.

[40] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 2(1), 1986.

[41] Nicholas R. Record and Brad Young. Patterns of diel vertical migration of zooplankton in acoustic doppler velocity and backscatter data on the newfoundland shelf. *Canadian Journal of Fisheries and Aquatic Sciences*, 63(12):2708–2721, 2006.

[42] A. Richardson. *Oceans*. Compass Point Books, 2001.

[43] Tom P. Rippeth and John H. Simpson. Diurnal signals in vertical motions on the hebridean shelf. *Limnology and oceanography*, 43(7):1690–1696, 1998.

[44] Timothy K. Stanton and Dezhang Chu. Review and recommendations for the modelling of acoustic scattering by fluid-like elongated zooplankton: euphausiids and copepods. *ICES Journal of Marine Science*, 57:793–807, 2000.

[45] J. Walker. *The Ozone Hole*. Stargazer Books, 2004.

[46] Wei Wang, Jiong Yang, and Richard R. Muntz. Sting: A statistical information grid approach to spatial data mining. In *VLDB*, pages 186–195, 1997.

[47] P. H. Wiebe, D. G. Mountain, T. K. Stanton, C. H. Greene, G. Lough, S. Kaartvedt, and J. Dawson. Acoustical study of the spatial distribution of plankton on georges bank and the relationship between volume backscattering strength and the taxonomic composition of the plankton. *Deep-Sea Research (Part 2, Topical Studies in Oceanography)*, 43(7-8):1971–2001, 1996.

[48] E. Xing, M. Jordan, and R. Karp. Feature selection for high-dimensional genomic microarray data. *Proceedings of the 15th International Conference on Machine Learning*, pages 601–608, 2001.

[49] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. *Proceedings of the 20th International Conference on Machine Learning*, pages 856–863, 2003.

Appendices

Part A.

Figure 1 shows the generated middle-depth migration path from January 2008 to July 2009. Data from January 2008 to January 2009 is used as the training set, the remaining data is used as the test set.

Figure 2 shows the generated upper bound of the migration path.

Figure 3 shows the generated lower bound of the migration path.

Figure 4 shows the fitted coefficient and frequency values if an 8-term Fourier series is used as the model. The table also shows the mean and standard deviation of each coefficient and the calculated statistical measurements for each fit.

Part B.

Figure 5 shows the monthly averaged environmental measurements we collect.

Part C.

Figure 6 shows results of linear correlation analysis using Pearson's correlation function. Strong correlations (p-values < 0.005) are shown in bold.

Figure 7 shows results of information gain between environmental factors and coefficients. Larger values in the table indicate stronger correlation.

Part D.

Each coefficient is modeled as a function of various measurements (features).

Here are the detailed functions for the middle depth. $a_n(t)$ and $b_n(t)$ are coefficients of a 5-term Fourier series.

a0 = 0.4982 * nightTime + 0.1409 * waterOxygen+ 0.0193 * waterConductivity-0.8838 * waterPressure-0.087 * waterDensity -0.0541 * waterSalinity+0.3196 * waterTemperature+0.9509 * waterTransmission+1.1561 * landTemperature+0.4385 * landDewPointTemperature+0.0855 * landVisibility-0.9938 * landHumidity+0.9897 * landWindDirection+2.3959 * landWindspd-0.0639 * landPressure+160.0701

a1 = -0.2953 * nightTime-1.1173 * waterOxygen+0.071 * waterConductivity+1.8055 * waterPressure+0.3435 * waterDensity+0.4814 * waterSalinity+0.247 * waterTemperature+0.1143 * waterTransmission+0.3027 * landTemperature-0.6223 * landDewPointTemperature-4.9539 * landVisibility-0.8234 * landHumidity-0.4848 * landWindDirection-1.2882 * landWindspd+0.8744 * landPressure-343.4623

time	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
1	139	26	24	94	20.5	104	39	66	32	57	194	61	117	103	93	41	53	52	39
31	180	38	36	110	25.5	141	41.5	74	37	66	246	143	144	105	119	40	57	55	38
61	284	96	55	129	26	135.5	43	82	47	90	400	452	255	134	152	48	69	55	39
91	685	193	83	171	32	143.5	58	109	65	159	722	709	526	207	189	50	93	60	47
121	725	445	142	207	31.5	149.5	70	126	120	488	727	719	729	321	244	56	117	64	48
151	727	704	210	264	39.5	163	99	147	214	714	728	707	732	665	393	101	130	86	60
181	727	718	550	393	104.5	180	120.5	180	519	704	730	695	733	740	710	168	143	103	90
211	725	714	705	600	207	223	162	257	712	706	726	699	718	728	716	394	195	138	128
241	724	722	718	682	494.5	342.5	287	419	699	697	725	700	734	730	719	706	292	211	186
271	723	720	713	688	625	482	446	683	675	713	723	698	731	733	715	728	638	300	309
301	723	721	725	689	670	575	601.5	669	668	704	723	701	729	733	720	724	715	438	503
331	725	724	713	698	706	659.5	706.5	661	665	705	723	692	723	733	726	725	718	713	723
361	724	732	723	711	681.5	700.5	698	693	689	710	720	683	723	738	726	721	711	731	729
391	726	729	732	680	644	667.5	717.5	688	683	711	721	699	720	727	724	726	703	720	725
421	724	733	732	665	684	652	676.5	690	671	703	718	704	725	731	722	727	715	725	724
451	730	728	727	694	703.5	651.5	659.5	697	681	705	724	704	728	734	723	731	714	727	724
481	728	729	728	677	657	653	717.5	695	705	715	723	702	724	741	721	730	693	723	731
511	723	736	732	710	675.5	697	700	685	690	703	725	707	728	729	721	732	706	723	721
541	728	728	734	707	664	637	703	669	711	711	723	699	726	731	723	731	716	725	720
571	724	730	736	709	660.5	688	703.5	700	702	704	726	712	729	741	732	727	704	726	716
601	732	729	736	679	660	694.5	690.5	700	700	705	728	707	732	736	732	734	698	722	729
631	727	741	728	710	710.5	698.5	689	697	699	721	731	713	732	733	732	732	712	725	725
661	727	732	729	690	716	673.5	720	701	694	715	729	713	726	731	738	727	725	616	720
691	732	744	733	721	695	595	704	705	697	717	732	712	731	737	736	730	719	390	552
721	735	741	726	686	467.5	550	517	479	706	730	730	717	736	734	736	722	551	239	372
751	739	738	725	680	269.5	407	338	323	706	728	738	715	733	743	735	573	278	161	218
781	734	740	726	451	134	279.5	203.5	202	694	729	735	720	738	742	738	313	193	100	129
811	738	738	636	278	90.5	218	148	123	566	727	735	714	739	732	704	180	165	70	76
841	738	739	314	177	60	161	96.5	57	324	487	735	718	737	699	474	94	158	62	64
871	738	737	159	123	55	101	48	35	183	208	739	702	739	658	240	112	125	74	59
901	743	303	110	79	37.5	69	32	23	79	139	434	599	730	343	127	93	114	79	54
931	275	133	78	45	33	89.5	18.5	19	43	116	261	344	416	228	96	84	89	61	53
961	134	69	80	34	29	81	14	32	29	78	183	113	173	167	79	56	83	74	36
991	127	42	39	22	29	67.5	18	31	36	53	153	48	121	119	101	45	73	72	55
1021	115	27	27	21	23	63.5	18.5	29	36	32	121	24	59	131	90	41	61	68	37
1051	102	27	34	30	23.5	62	22	22	34	23	74	20	47	128	81	51	54	45	39
1081	96	28	28	47	29.5	56.5	26	35	41	27	56	15	40	98	75	53	48	50	66
1111	83	13	18	28	29	73	18	23	50	31	42	26	40	104	56	41	36	47	39
1141	81	16	19	29	46.5	71	14	36	27	16	30	16	32	67	58	26	34	41	64
1171	76	14	17	30	37.5	56.5	27.5	44	23	18	24	18	27	83	35	30	30	46	49
1201	76	14	29	35	37.5	67.5	20	23	19	15	28	20	20	76	32	30	39	44	44
1231	83	18	41	27	31.5	57.5	23	17	25	13	15	12	25	87	23	24	58	47	37
1261	81	21	18	52	25	36.5	14.5	15	34	13	17	17	21	88	34	22	40	37	36
1291	82	16	14	26	19.5	34	15	15	32	12	17	12	17	65	21	21	35	47	38
1321	63	20	30	32	19.5	31.5	14.5	17	48	12	29	15	18	78	15	18	47	47	47
1351	73	22	47	52	26.5	29	10	15	34	11	17	17	38	80	14	31	69	48	53
1381	83	19	48.5	68	32	23.5	12	22	35	10	26.5	16	65	70	12	28	58	56	55
1411	106	15	59	33	27.5	26	15	44.5	46	9.5	16	10	76.5	96	11	25	40	47	50.5

Figure 1: Middle-depth migration path from January 2008 to July 2009

time	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
1	191	152	66	67	62.5	149	27.5	73	92	120	231	196	144	191	166	76	93	116	99
31	211	143	78	84	44	171	41.5	70	106	142	308	301	196	199	165	73.5	91	107	102
61	287	131	131	62	43	220	34	77	140	205	483	722	316	209	198	75.5	98.5	118	119
91	727	249	158	60	39.5	275	29	117	152	309	745	747	674	287	227	92	120	123	105
121	736	468	174	82	51	293	49	127	179	546	746	742	748	573	270	119	139	129	99
151	742	738	278	135	107	365	62	156	205	743	743	744	744	738	490	167	159	153	117
181	737	737	657	316	205	406	95	214	562	739	740	743	742	737	745	258	174	181	151
211	736	737	742	734	305	473	166	306	729	737	738	741	736	740	742	561	226	207	181
241	738	740	742	740	338	586	283	582	700	737	737	742	740	741	738	740	418	255	286
271	739	738	745	738	737	732	412	702	682	734	738	738	741	741	734	738	731	348	398
301	737	741	742	740	734	736	722	695	678	735	737	741	739	741	734	733	734	549	634
331	741	738	738	738	738	735	727	694	683	732	736	736	732	739	737	731	728	738	726
361	737	741	741	740	733	737	724	702	690	727	736	740	735	740	737	729	721	738	730
391	737	743	740	741	736	735	727	704	683	716	736	739	733	743	742	731	716	725	730
421	739	741	745	738	737	732	725	708	681	731	738	740	737	745	743	733	722	727	727
451	743	746	743	741	737	735	723	703	686	730	739	739	737	747	740	738	718	732	733
481	736	738	745	743	737	732	726	704	705	733	738	737	739	746	740	736	714	727	733
511	735	743	748	742	734	732	724	699	721	731	737	743	742	745	745	736	716	728	721
541	736	747	747	743	734	733	726	696	712	736	740	741	740	748	744	736	716	729	721
571	739	742	747	745	734	733	725	703	721	721	739	743	743	749	746	736	708	728	717
601	740	746	748	748	735	739	721	710	711	732	740	741	744	750	749	736	709	726	732
631	740	747	747	747	736	735	726	700	708	736	742	743	744	750	750	735	718	727	730
661	739	747	747	748	733	731	732	711	713	727	745	739	743	751	753	733	729	706	735
691	741	749	749	747	738	714	721	720	709	727	746	740	745	752	753	735	734	440	695
721	747	751	749	734	382	476	519	726	719	739	745	743	747	755	752	741	618	279	479
751	749	750	749	732	203	356	332	406	717	742	748	745	747	754	750	735	329	195	302
781	748	751	750	513	74	247	138	286	719	739	746	745	752	755	753	495	247	148	202
811	750	752	646	221	66	211	74.5	180	508	739	613	747	751	753	743	234	233	117	128
841	749	752	364	190	65.5	149	56	142	253	670	402	747	754	754	572	156	198	108	114
871	751	684	201	162	63.5	115	43.5	89	148	310	235	749	755	744	329	128	181	107	103
901	752	357	134	137	36	88.5	40	35.5	86	190	185	708	756	477	207	129	172	92	105
931	341	175	107	111	38.5	87	43	33	90	138	152	341	474	249	135	96	134	84.5	113
961	206	213	143	89	36.5	107	28.5	40.5	82	119	128	198	251	207	73	68.5	123	86.5	98
991	205	192	115	91	32	118	35.5	62	69	97	125	127	167	194	67	49	121	76.5	107
1021	199	183	99	85	30.5	92.5	22.5	53	70	85	120	112	115	186	83	61.5	118	75	75
1051	172	148	100	84	26.5	90.5	28.5	32	64	81	120	100	105	173	103	58	111	64	52
1081	145	125	65	97	36.5	109	20.5	37.5	62	68	131	107	85	159	115	55	97.5	60	76
1111	152	65	75	129	32	88	19.5	26.5	61	79	114	96	88	154	135	45.5	85	65	57
1141	134	52	37	159	24.5	83.5	13.5	29.5	73	58	98	89	81	129	108	28.5	68	46	72
1171	136	81	42	161	21	74	21.5	44	86	35	86	72	79	146	79	28.5	52	46	80
1201	134	94	56	119	28.5	90	19	27	75	38	110	48	67	138	70	30.5	25.5	44.5	54
1231	127	62	39	111	24.5	79.5	26.5	24	67	33	135	37	54	125	64	24	31	53.5	59
1261	133	60	42	161	27	92	15	18	66	25	132	34	35	138	41	23	27	52	61
1291	113	76	35	154	45.5	124	18	23.5	79	31	141	27	21	145	35	22	25	62	51
1321	98	74	56	143	35.5	78.5	15	15.5	82	28	145	42	21	138	27	28	28.5	51.5	73
1351	134	62	67	116	53	48.5	11.5	27	108	46	183	44	13	131	23	40.5	24.5	60.5	76
1381	145	82	78.5	125	58.5	25.5	15	44.5	105	41	155	47	19	132	26	49	29	70	87
1411	190	56	57	143	50	34	11	72	100	31	146	78.5	44.5	149	21	46	20	78.5	84.5

Figure 2: Upper bound of the migration path from January 2008 to July 2009

time	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul
1	89	73	7	9	9	20	6.5	7	7	48	39	41	90	4	7	17.5	14	37.5	20
31	112	65	11	7	7	36.5	7	6	13	68	145	94	129	28	38	17	15	26.5	12
61	201	50	33	7	9	40.5	6.5	23	25	115	234	199	228	61	78	17	13.5	27	16
91	658	167	42	7	8	56	6	30.5	35	185	651	500	470	146	111	22.5	12	15	13
121	662	352	103	7	8.5	78.5	6	42	83	276	653	673	700	445	149	24	10	15	9
151	676	541	199	35	7	131	6	42	154	552	659	676	709	619	253	34	39	35	7
181	670	496	445	167	10.5	176	13.5	78.5	411	620	650	634	712	639	595	97.5	92	63	8
211	670	496	499	474	38	271	28	157	598	608	635	616	701	627	647	249	135	76	35
241	645	520	571	622	149	410	122	256	692	597	623	584	717	654	658	484	218	118	92
271	671	540	564	642	403	521	224	591	669	582	636	599	712	653	655	575	556	217	187
301	653	551	556	639	554	516	474	663	663	562	634	606	703	649	647	631	677	404	385
331	653	554	563	659	563	551	571	657	659	584	635	596	687	642	645	627	685	631	604
361	641	551	567	641	557	599	624	693	660	590	645	617	678	646	646	612	689	631	654
391	657	564	566	616	528	571	675	682	662	603	661	643	673	643	645	610	681	635	668
421	659	558	560	588	540	570	577	687	660	579	662	639	668	645	644	628	691	641	675
451	661	570	565	647	550	564	588	691	668	597	645	633	637	647	641	612	692	642	671
481	669	587	558	613	561	581	579	693	683	628	650	613	648	649	651	622	686	651	678
511	657	599	527	596	629	597	570	681	684	592	652	620	634	651	637	635	670	656	672
541	664	604	525	570	600	552	678	675	681	568	671	648	628	649	645	638	675	655	675
571	647	620	562	541	608	574	676	695	684	579	682	645	613	645	633	639	676	649	682
601	671	613	559	559	597	596	659	682	682	586	677	655	601	644	646	615	665	638	678
631	665	614	560	570	649	626	592	687	697	599	690	654	583	640	669	625	651	624	668
661	671	592	546	566	639	557	581	685	686	583	688	657	596	633	668	643	641	547	644
691	656	610	536	553	488	471	496	581	687	579	689	647	602	635	673	614	587	340	492
721	663	615	530	580	229	335	285	355	685	587	696	630	620	647	673	605	447	159	270
751	664	610	506	599	33	91	66.5	157	517	580	689	644	632	661	663	408	171	82.5	125
781	671	623	512	346	45.5	23	17.5	28.5	379	572	534	643	629	649	654	233	85	41.5	78
811	673	635	464	9	62	46.5	13.5	7	135	563	331	627	634	645	639	58	68.5	62.5	40
841	680	627	233	12	50.5	42.5	9	10.5	43	481	126	627	594	646	359	87.5	34.5	59.5	18
871	680	545	129	49	35.5	40	13.5	10	50	199	36	615	579	557	122	102	29.5	73	29
901	595	193	77	77	31.5	41	13	9	34	78	11	520	578	270	15	89	28.5	67	36
931	244	31	30	17	27	38.5	14.5	8	12	24	12	232	379	145	76	80.5	30	58.5	29
961	66	8	51	15	18	36.5	11.5	9	11	36	8	40	150	90	14	46.5	24.5	70	29
991	27	7	36	12	25	31	11.5	8	10	18	8	12	35	49	16	37	28	70.5	49
1021	8	10	11	10	14.5	26.5	8.5	9.5	9	15	8	12	9	33	27	40	27	67.5	32
1051	8	12	11	9	17.5	21.5	11.5	8.5	9	12	7	12	13	7	16	40.5	32	44.5	30
1081	8	9	13	9	18	20.5	10.5	7.5	8	9	8	13	28	7	32	47	23.5	45.5	43
1111	8	7	8	9	15.5	23	7	7	7	9	7	20	14	19	15	40.5	24	42.5	39
1141	7	7	8	8	10.5	20	8.5	7	8	9	8	8	12	10	16	25.5	22.5	39.5	63
1171	7	7	7	8	10	21	7.5	7	7	8	8	11	9	6	18	23.5	27.5	45.5	42
1201	7	7	7	9	12.5	23	7.5	6	8	11	8	8	7	6	15	23.5	19	42.5	44
1231	6	7	7	9	14	23.5	7	8	8	7	7	8	7	15	7	23	17.5	45	36
1261	7	6	7	8	10	26.5	7	6.5	8	8	7	7	7	25	10	19	24	35	31
1291	7	6	7	11	8	28	7	6	7	8	7	7	5	27	7	21	20	45	38
1321	7	6	8	10	12	25	6.5	6.5	7	9	7	7	6	23	8	17.5	19	41.5	33
1351	6	6	18	10	12	21.5	6	7	7	7	7	8	4.5	30	6	21.5	18.5	45.5	53
1381	6	6	18.5	9	9	16	8	6	7	7	5	7	5	36	6	18	18	54.5	50
1411	6	6	27	9	8	17	7	6	7	7	5.5	7	4	57	5	24	16	38	43

Figure 3: Lower bound of the migration path from January 2008 to July 2009

coeffici	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan,09	averag	stdev
a0	481.4	431.2	369.1	342.4	261.3	281.20	277.7	284.8	344.6	408.8	449.8	443.7	440.2	370.48	77.068
a1	-232.5	-297.2	-268.6	-217.5	-184	-162.3	-256.9	-201.9	-236.5	-270.8	-202.8	-231	-208.8	-228.5	37.914
b1	301	318.1	345	339.8	318.5	297.9	289.1	324.6	338	326	359.7	331.7	364.5	327.22	22.773
a2	37.69	13.59	-3.223	-4.562	-81.3	-58.5	-9.108	-42.22	-7.217	12.84	62.92	57.23	60.89	3.0023	45.254
b2	130.5	101.8	-20.35	-44.41	-129.8	-103.9	-131.7	-115.5	-34.89	77	108.5	136.1	85.55	4.5308	105.12
a3	-55.62	-81.17	-111.8	-94.69	-29.33	-1.678	-31.97	-41.19	-117	-102.2	-73.84	-60.07	-94.3	-68.84	35.612
b3	-20.17	-62.73	-47.7	-13.54	-0.506	16.01	-16.47	6.922	-24.73	-43.92	15.84	-8.568	-7.891	-15.96	24.14
a4	-74.6	-75.13	19.62	35.91	38.39	34.25	65.89	55.49	22.83	-47.06	-26.88	-71.15	-34.73	-4.398	51.983
b4	52	18.91	-15.2	-23.74	-73.91	-37.47	-9.913	-42.87	-25.43	23.68	68.93	69.76	57.87	4.8167	47.021
a5	1.474	15.89	5.914	-3.038	29.32	24.75	-16.87	2.971	-7.051	2.893	-13.05	-0.385	-12.76	2.3122	14.102
b5	11.65	-24.74	-59.01	-22.42	36.83	35.03	29.01	32.46	-61.85	-49.28	-7.236	19.81	-34.74	-7.268	37.003
a6	-37.6	-18.69	19.92	25.92	11.94	-4.752	-0.774	15.82	26.82	-24.33	-40.7	-43.19	-53.41	-9.464	28.533
b6	-31.6	-52.29	16.05	19.14	-6.225	7.523	8.38	8.417	11.82	-26.87	16.06	-16.92	-4.965	-3.96	22.022
a7	-30	-0.783	31.9	-6.656	-10.75	-8.842	-23.5	-21.17	24.08	27.15	2.201	-33.68	1.551	-3.731	21.28
b7	12.23	4.665	-8.038	-0.616	36.73	8.143	-10.21	16.8	-8.577	-8.285	3.289	18.61	-4.482	4.6353	13.779
a8	13.48	33.5	-11.31	2.236	-19.85	-3.498	19.08	-7.206	2.917	15.03	-20.95	-5.65	-17.92	-0.011	16.547
b8	-13.98	-13.78	19.87	19.9	-15.63	-6.13	-5.595	-13.8	26.87	-26.94	-16.18	-9.432	-31.5	-6.641	18.021
w	0.0045	0.0046	0.0044	0.0045	0.0043	0.0052	0.0046	0.0044	0.0043	0.0046	0.0043	0.0044	0.0042	0.0045	0.0002
SSE	86298	42151	36243	17924	15006	7416	7884	22083	22526	41286	38733	36385	34979		
R	0.9805	0.9923	0.9929	0.9958	0.9964	0.9977	0.9981	0.9946	0.9952	0.9919	0.9919	0.9926	0.9929		
RMSE	53.63	37.48	34.76	24.44	22.37	15.72	16.21	27.13	27.4	37.1	35.93	34.83	34.15		

Figure 4: Fitted coefficient and frequency values if an 8-term Fourier series is used as the model

measurement	jan	feb	mar	apr	may	jun	jul	aug	sep	oct	nov	dec	jan,09	feb	mar	apr	may	jun	jul
nightTime	913	828	723	616	523	476	500	582	686	793	889	939	912	825	725	616	525	477	502
waterOxygen	1.31	2.04	2.12	2.57	2.28	2.04	1.55	1.11	0.54	0.21	0.42	0.26	1.69464	1.81	2.28	2.73	2.45	1.97	1.67
waterConductivity	3.33	3.27	3.24	3.23	3.26	3.27	3.29	3.33	3.35	3.37	3.37	3.37	3.27389	3.26	3.22	3.21	3.19	3.24	3.27
waterPressure	95.42	95.39	95.06	94.93	95.03	95.15	95.26	95.21	95.09	95.99	96.17	96.18	96.1048	95.98	96.18	96.07	96.12	96.15	96.12
waterDensity	1024.3	1024.25	1024.24	1024.24	1024.39	1024.4	1024.43	1024.46	1024.52	1024.54	1024.51	1024.45	1024.39	1024.37	1024.37	1024.39	1024.04	1024.32	1024.27
waterSalinity	30.85	30.7	30.63	30.62	30.82	30.85	30.94	31.04	31.14	31.19	31.15	31.09	30.8486	30.79	30.72	30.72	30.33	30.72	30.72
waterTemperature	9.08	8.56	8.29	8.18	8.32	8.42	8.61	8.93	9.06	9.22	9.24	9.27	8.45565	8.30	7.93	7.78	8.03	8.23	8.61
waterTransmission	65.76	57.55	64.74	70.13	64.11	64.41	59.27	52.1	46.31	68.33	66.28	62.92	49.5073	61.03	68.19	65.93	72.96	71.17	64.58
landTemperature	3.6	4.45	5.2	6.73	11.55	13.55	16.05	15.3	13.03	9.4	8.37	3.5	2.4	3.79	4.82	8.52	12.03	16.25	18.26
landDewPointTemp	1.23	2.3	1.8	1.95	7.05	7.53	10.35	11.45	9.95	6.35	6.43	0.8	0.8	0.71	0.23	3.22	6.00	8.73	11.19
landVisibility	32.2	32.2	48.3	48.3	48.3	48.3	48.3	48.3	48.3	48.3	28.15	32.2	32.2	38.17	38.21	44.78	43.62	45.77	45.58
landHumidity	87.5	91	81	75.75	76.5	69.5	70.5	76	82.5	88.5	91.25	93	90	81.61	74.17	71.77	68.26	63.28	65.56
landWinddirection	17.5	23	22.5	23.25	14	14	12	13	15	19	22.25	17	25	20.17	18.66	18.41	16.74	14.96	13.61
landWindspd	12	7	9	11	8	9	9	8	6.5	8	7	9	7	8.82	13.23	10.05	10.70	9.93	9.01
landPressure	100.93	101.69	101.82	101.52	101.54	101.68	101.4	101.18	101.64	102.08	101.64	101.28	102.6	101.39	101.45	101.67	101.46	101.21	101.42

Figure 5: Monthly averaged environmental measurements

	a0	a1	b1	a2	b2	a3	b3	a4	b4	a5	b5	a6	b6	a7	b7	a8	b8	w
nightTime/coef	**0.9746**	-0.3503	0.5067	**0.9043**	**0.9713**	-0.5094	-0.2707	**-0.8994**	**0.9071**	-0.3662	-0.3961	**-0.7611**	-0.4803	0.0418	-0.1079	0.0087	-0.3393	-0.4748
p-value	**1.69E-08**	0.2407	0.0772	**2.20E-05**	**3.33E-08**	7.54E-02	3.71E-01	**2.87E-05**	**1.87E-05**	2.18E-01	1.80E-01	**2.50E-03**	9.67E-02	8.92E-01	7.26E-01	9.78E-01	2.72E-01	1.01E-01
waterOxygen	-0.4097	0.1355	-0.1832	-0.4894	-0.4611	0.2094	-0.0783	0.3856	-0.5295	0.4403	0.1367	0.4181	0.1159	-0.0556	0.1343	0.0076	0.2332	0.1735
	0.1644	0.659	0.5491	0.0896	0.1127	0.4922	0.7994	0.1931	0.0627	0.1321	0.656	0.1551	0.7062	0.8569	0.6618	0.9803	0.4433	0.5709
waterConductivity	0.4056	-0.0583	0.0813	0.4411	0.4589	-0.0821	0.1526	-0.4103	0.5117	-0.3594	-8.26E-05	-0.4426	-0.1846	-0.074	0.0043	0.0214	-0.2976	-0.1403
	0.1692	0.8501	0.7917	0.1314	0.1147	0.7898	0.6187	0.1638	0.0738	0.2278	0.9998	0.1299	0.5461	0.8101	0.9889	0.9446	0.3234	0.6476
waterPressure	0.6975	-0.0732	0.4231	**0.7834**	**0.748**	-0.1701	0.1037	-0.6738	**0.8606**	-0.4148	-0.0968	**-0.8861**	-0.3011	-0.0564	-0.0785	-0.1655	-0.6504	-0.283
	0.008	0.8122	0.1497	**0.0015**	**0.0033**	0.5786	0.7361	0.0116	**1.59E-04**	0.1587	0.7531	**5.52E-05**	0.3175	0.8549	0.7987	0.589	0.0161	0.3488
waterDensity	-0.0874	0.231	0.08	0.0619	-0.0222	0.081	0.3923	0.0677	0.1166	-0.2848	0.0794	-0.1877	0.1598	0.0781	-0.037	-0.2112	-0.2889	-0.0491
	0.7766	0.4477	0.7949	0.8408	0.9426	0.7926	0.1849	0.826	0.7045	0.3455	0.7965	0.5392	0.602	0.7999	0.9045	0.4885	0.3384	0.8733
waterSalinity	0.1269	0.0911	0.0813	0.2401	0.1892	-4.63E-04	0.29	-0.1373	0.2992	-0.3532	0.0391	-0.3	0.0144	0.0131	-0.0387	-0.0962	-0.2947	-0.0963
	0.6795	0.7672	0.7918	0.4295	0.5359	0.9988	0.3365	0.6546	0.3207	0.2365	0.8991	0.3194	0.9628	0.9661	0.9002	0.7546	0.3283	0.7542
waterTemperature	0.4896	-0.1376	0.0552	0.4977	0.5338	-0.1114	0.0764	-0.4878	0.5673	-0.3644	-0.0187	-0.4669	-0.2553	-0.1042	-0.0037	0.0973	-0.277	-0.1427
	0.0894	0.6539	0.858	0.0835	0.0602	0.7172	0.804	0.0908	0.0432	0.2208	0.9518	0.1077	0.3998	0.7347	0.9905	0.7518	0.3596	0.642
waterTransmission	0.1094	0.0167	-0.1792	-0.0086	0.1297	0.1558	0.0057	-0.136	0.0801	0.2616	0.1645	-0.0756	-0.0435	-0.0774	0.157	0.0243	-0.0274	0.3153
	0.722	0.9567	0.558	0.9778	0.6727	0.6113	0.9853	0.6578	0.7948	0.388	0.5912	0.8061	0.8877	0.8015	0.6084	0.9373	0.9292	0.2939
landTemperature	**-0.8319**	0.3036	-0.4598	**-0.6975**	**-0.8146**	0.4701	0.3715	**0.7863**	-0.6797	0.0833	0.383	0.5693	0.4516	-0.0328	0.0183	0.0163	0.1594	0.3684
	4.20E-04	0.3132	0.1139	**0.008**	**6.97E-04**	0.105	0.2114	**0.0014**	0.0106	0.7867	0.1964	0.0423	0.1213	0.9152	0.9528	0.9578	0.603	0.2155
landDewTemp	-0.7037	0.2473	-0.3375	-0.5742	-0.6874	0.3612	0.3458	0.6766	-0.5634	-0.0118	0.2928	0.4729	0.3904	0.0318	-0.0056	-0.0026	0.0759	0.2209
	0.0073	0.4153	0.2595	0.0401	0.0094	0.2253	0.2471	0.0111	0.0449	0.9694	0.3317	0.1027	0.1872	0.9179	0.9856	0.9932	0.8054	0.4684
landVisibility	**-0.8352**	0.0952	-0.3214	**-0.7987**	**-0.8374**	0.0964	-0.0275	**0.7963**	**-0.8584**	0.2675	-0.0046	**0.8402**	0.4582	0.2965	-0.1219	0.0259	0.4666	0.3704
	3.79E-04	0.757	0.2843	**0.0011**	**3.55E-04**	0.754	0.9289	**0.0011**	**1.73E-04**	0.3769	0.9881	**3.24E-04**	0.1153	0.3253	0.6917	0.9331	0.108	0.2129
landHumidity	**0.9135**	-0.396	0.5288	**0.8166**	**0.9414**	-0.5322	-0.3307	**-0.9011**	**0.8259**	-0.2568	-0.425	-0.711	-0.5532	0.1337	-0.0421	0.0303	-0.368	-0.4975
	1.28E-05	0.1804	0.0631	**6.59E-04**	**1.59E-06**	6.12E-02	0.2698	**2.62E-05**	**5.04E-04**	3.97E-01	0.1477	0.0064	0.0499	0.6632	0.8914	0.9217	0.216	0.0836
landWindDirection	0.6585	-0.3116	**0.706**	0.6119	0.6183	-0.6675	-0.3683	-0.5026	0.5143	-0.2218	-0.6347	-0.3422	-0.1417	0.406	-0.3482	-0.0738	-0.0437	-0.3674
	1.44E-02	0.3	**0.007**	0.0262	2.43E-02	0.013	0.2156	0.0801	7.22E-02	0.4664	0.0198	0.2524	0.6443	0.1686	0.2436	0.8106	0.8873	0.2168
landWindSpd	0.0527	0.0652	-0.4306	-0.0259	-0.002	0.2332	0.0377	-0.011	0.0056	0.0576	0.2955	0.0593	-0.0112	-0.4603	0.1311	0.189	0.1922	0.2657
	8.64E-01	0.8325	0.1418	0.933	9.95E-01	0.4432	0.9026	0.9717	0.9855	0.8518	0.3269	0.8473	0.9711	0.1135	0.6694	0.5363	0.5293	0.3803
landPressure	0.1402	-0.0802	0.5597	0.2028	0.1453	-0.4492	-0.1931	-0.0973	0.1473	-0.1164	-0.5752	-0.2228	0.0336	0.6435	-0.4602	-0.2204	-0.2381	-0.145
	0.6478	0.7945	0.0467	0.5064	0.6359	0.1236	0.5273	0.7518	0.631	0.705	0.0397	0.4643	0.9132	0.0176	0.1136	0.4694	0.4335	0.6365

Figure 6: Linear correlation analysis using Pearson's correlation function

	a0	a1	b1	a2	b2	a3	b3	a4	b4	a5	b5	a6	b6	a7	b7	a8	b8	w
nightTime	2.6032	2.661	2.4116	2.1997	2.2955	2.3535	2.6612	2.5074	2.3535	2.142	2.4116	2.4493	2.2	2.3535	2.2578	2.3535	2.2578	1.93
waterOxygen	2.2955	2.354	2.2578	2.0458	1.9878	2.1997	2.3535	2.1997	2.1997	1.834	2.1039	2.1416	1.892	2.0458	2.2578	2.1997	1.7962	1.776
waterConductivity	2.0836	1.988	2.0458	1.8339	1.622	1.8339	1.9878	1.8339	1.8339	1.622	1.892	1.9297	1.68	1.8339	1.892	1.8339	1.7381	1.41
waterPressure	1.9878	1.988	1.892	1.892	1.7759	1.8339	2.2578	2.1416	1.892	1.468	1.892	1.7759	1.68	1.8339	1.7382	1.8339	1.4305	1.256
waterDensity	2.1416	2.2	2.1039	1.892	1.8339	2.0458	2.1997	2.0458	1.892	1.834	2.2578	1.9878	1.892	2.4538	2.1039	2.0458	1.7962	1.622
waterSalinity	1.9297	2.142	2.1997	1.8339	1.7759	1.9878	2.1416	1.9878	1.8339	1.622	1.892	1.9297	1.834	1.8339	2.0458	1.9878	1.7381	1.718
waterTemperature	1.9878	2.2	1.9501	1.7381	1.6801	1.892	2.0458	1.892	1.892	1.526	1.7962	1.8339	1.584	1.7381	1.7962	1.892	1.4885	1.622
waterTransmission	2.2955	2.354	2.2578	1.892	1.9878	2.3535	2.3535	2.1997	2.0458	1.988	2.2578	2.1416	1.892	2.1997	2.1039	2.1997	1.7962	1.776
landTemperature	2.4493	2.661	2.4116	2.1997	2.4493	2.3535	2.6612	2.6612	2.3535	2.142	2.4116	2.2955	2.2	2.5074	2.2578	2.3535	2.1039	1.93
landDewTemp	1.6801	1.892	1.6424	1.5843	1.5262	1.5843	1.892	1.7381	1.5843	1.372	1.6424	1.5262	1.43	1.5843	1.6424	1.7381	1.4885	1.468
landVisibility	1.027	0.873	0.7774	0.8731	1.027	0.8731	0.7193	1.2389	0.8731	0.873	1.0851	1.027	0.873	1.027	0.9312	0.5654	0.6816	0.449
landHumidity	2.0836	1.988	1.892	1.6801	1.9297	2.1416	1.9878	1.9878	1.8339	1.468	2.0458	2.0836	1.68	1.8339	1.7382	1.8339	1.892	1.41
landWindDirection	2.2955	2.354	2.2578	2.0458	2.1416	2.1997	2.3535	2.3535	2.0458	1.988	2.5654	2.1416	2.2	2.3535	2.1039	2.1997	1.9501	1.776
landWindSpd	1.5262	1.738	1.6428	1.2766	1.5843	1.4035	1.4305	1.2766	1.4305	1.219	1.4885	1.3724	1.123	1.4885	1.3347	1.4305	1.027	1.16
landPressure	1.892	2.104	2.0081	1.6424	1.892	2.1039	2.2578	1.9501	1.7962	1.584	1.8543	1.7381	1.796	1.8339	2.0081	1.9501	1.5466	1.68

Figure 7: Information gain between environmental factors and coefficients

b1 = -0.0935 * nightTime-0.4951 * waterOxygen+0.0239 * waterConductivity+0.0053 * waterPressure+0.1078 * waterDensity+0.158 * waterSalinity+0.0246 * waterTemperature-0.288 * waterTransmission-0.3807 * landTemperature+0.9448 * landDewPointTemperature+0.659 * landVisibility+3.2365 * landHumidity+4.1754 * landWindDirection-1.9758 * landWindspd-0.662 * landPressure+13.8303

a2 = 0.501 * nightTime-2.0522 * waterOxygen+0.0822 * waterConductivity+1.6243 * waterPressure+0.3771 * waterDensity+0.576 * waterSalinity+0.3864 * waterTemperature-0.093 * waterTransmission+3.1644 * landTemperature+0.1958 * landDewPointTemperature-0.6634 * landVisibility-4.5997 * landHumidity+1.4665 * landWindDirection-0.1954 * landWindspd+1.0791 * landPressure-670.0255

b2 = 0.706 * nightTime-0.1777 * waterOxygen+0.0355 * waterConductivity-0.7691 * waterPressure-0.0517 * waterDensity+0.0144 * waterSalinity+0.4321 * waterTemperature+1.8953 * waterTransmission+1.5292 * landTemperature+0.6438 * landDewPointTemperature+1.4229 * landVisibility-0.0394 * landHumidity+0.0491 * landWindDirection+2.1191 * landWindspd-0.0776 * landPressure-583.8126

a3 = -0.1776 * nightTime+0.6255 * waterOxygen+0.0134 * waterConductivity+0.6893 * waterPressure+0.1117 * waterDensity+0.1058 * waterSalinity+0.0153 * waterTemperature+0.3509 * waterTransmission-0.7259 * landTemperature+0.0795 * landDewPointTemperature-3.3063 * landVisibility+0.1871 * landHumidity-6.1455 * landWindDirection+0.962 * landWindspd+0.7699 * landPressure+9.6869

b3 = -0.1413 * nightTime-1.3491 * waterOxygen+0.0809 * waterConductivity+1.8065 * waterPressure+0.3568 * waterDensity+0.5191 * waterSalinity+0.3281 * waterTemperature+0.1088 * waterTransmission+1.0387 * landTemperature+ 0.0269 * landDewPointTemperature-3.6598 * landVisibility-1.1395 * landHumidity+0.9288 * landWindDirection-1.3527 * landWindspd+0.8765 * landPressure-326.1203

a4 = -0.1071 * nightTime-0.7382 * waterOxygen+0.0135 * waterConductivity+0.3436 * waterPressure+0.055 * waterDensity+0.0978 * waterSalinity+ 0.0338 * waterTemperature-0.242 * waterTransmission+1.6711 * landTemperature+1.9307 * landDewPointTemperature+0.3363 * landVisibility-2.7006 * landHumidity+3.0144 * landWindDirection-0.9342 * landWindspd-0.6222 * landPressure+205.1355

b4 = 0.434 * nightTime-1.8624 * waterOxygen+0.0707 * waterConductivity+2.1488 * waterPressure+0.3912 * waterDensity+0.5609 * waterSalinity+0.2759 * waterTemperature+1.1893 * waterTransmission+2.2907 * landTemperature-0.931 * landDewPointTemperature-0.8797 * landVisibility-3.2716 * landHumidity-0.1713 * landWindDirection-1.1017 * landWindspd+1.78 * landPressure-881.9389

a5 = -0.1689 * nightTime+1.2983 * waterOxygen-0.0351 * waterConductivity-0.1697 * waterPressure-0.1531 * waterDensity-0.2519 * waterSalinity-0.1547 * waterTemperature+0.3781 * waterTransmission-1.9722 * landTemperature-0.7956 * landDewPointTemperature-0.0776 * landVisibility+1.982 * landHumidity-

1.8861 * landWindDirection+0.5724 * landWindspd+0.3696 * landPressure+137.3678

b5=-0.2544 * nightTime-0.3726 * waterOxygen+0.0393 * waterConductivity+1.3349 * waterPressure+0.2281 * waterDensity+0.3024 * waterSalinity+0.0668 * waterTemperature+0.5045 * waterTransmission-1.5544 * landTemperature-1.826 * landDewPointTemperature-4.0919 * landVisibility+0.5924 * landHumidity-5.3874 * landWindDirection+0.6228 * landWindspd+0.1175 * landPressure+2.8581

w = 0.004383

The following are detailed functions for the upper bound:

a0=0.4472 * nightTime+0.2111 * waterOxygen+0.0332 * waterConductivity-1.4024 * waterPressure-0.1704 * waterDensity-0.1205 * waterSalinity+0.4887 * waterTemperature+1.9171 * waterTransmission-0.0681 * landTemperature-0.4011 * landDewPointTemperature+0.097 * landVisibility-0.7425 * landHumidity+0.5895 * landWindDirection+3.7556 * landWindspd-1.1132 * landPressure+396.1964

a1=-0.0567 * nightTime-0.8661 * waterOxygen+0.0248 * waterConductivity+0.1224 * waterPressure+0.185 * waterDensity+0.2098 * waterSalinity-0.0108 * waterTemperature-0.2329 * waterTransmission-0.9047 * landTemperature-1.0927 * landDewPointTemperature-1.3668 * landVisibility+0.5773 * landHumidity+0.5769 * landWindDirection-3.9719 * landWindspd+0.4798 * landPressure-378.3218

b1=-0.0436 * nightTime-0.2503 * waterOxygen-0.0112 * waterConductivity+1.4382 * waterPressure+0.2101 * waterDensity+0.1705 * waterSalinity-0.3598 * waterTemperature+0.0248 * waterTransmission-1.054 * landTemperature-0.5195 * landDewPointTemperature-0.3026 * landVisibility-0.4059 * landHumidity+3.2878 * landWindDirection-4.6634 * landWindspd+1.6506 * landPressure-115.7891

a2=0.249 * nightTime-1.0391 * waterOxygen+0.0368 * waterConductivity+1.7081 * waterPressure+0.2849 * waterDensity+0.3977 * waterSalinity+0.0293 * waterTemperature-0.2499 * waterTransmission+1.1798 * landTemperature-0.0609 * landDewPointTemperature-0.7131 * landVisibility-0.6643 * landHumidity+0.5431 * landWindDirection-0.1051 * landWindspd+0.6642 * landPressure-624.2422

b2=0.7439 * nightTime-0.1035 * waterOxygen+0.0413 * waterConductivity+0.1149 * waterPressure-0.0554 * waterDensity+0.0208 * waterSalinity+0.4532 * waterTemperature+0.9941 * waterTransmission+0.5993 * landTemperature+0.0156 * landDewPointTemperature+1.1545 * landVisibility+0.1736 * landHumidity-2.5601 * landWindDirection+2.4246 * landWindspd-0.2507 * landPressure-583.8455

a3=-0.1752 * nightTime-0.1387 * waterOxygen-0.0025 * waterConductivity+1.6052 * waterPressure+0.1481 * waterDensity+0.1479 * waterSalinity-0.1451 * waterTemperature+0.6922 * waterTransmission-1.0657 * landTemperature-2.9787 * landDewPointTemperature-4.0006 * landVisibility-0.2708 * landHumidity-6.773 * landWindDirection-0.3837 * landWindspd+1.4164 * landPressure-96.3242

b3=-0.0531 * nightTime-0.9217 * waterOxygen+0.0104 * waterConductivity+1.1755 * waterPressure+0.2688 * waterDensity+0.2786 * waterSalinity-0.192 * waterTemperature+0.4362 * waterTransmission+0.1862 * landTemperature-0.5931 * landDewPointTemperature-2.8729 * landVisibility-

0.4418 * landHumidity+0.413 * landWindDirection-5.0014 * landWindspd+1.5398 * landPressure-346.6738

a4=-0.3448 * nightTime-0.3158 * waterOxygen-0.0042 * waterConductivity+0.4372 * waterPressure+0.0659 * waterDensity+0.0633 * waterSalinity-0.1587 * waterTemperature-0.204 * waterTransmission+0.3043 * landTemperature+1.2292 * landDewPointTemperature-1.8616 * landVisibility+0.0726 * landHumidity+2.2855 * landWindDirection-2.2032 * landWindspd-0.3842 * landPressure+250.5848

b4=0.2518 * nightTime-1.4486 * waterOxygen+0.0563 * waterConductivity+2.2734 * waterPressure+0.2971 * waterDensity+0.4396 * waterSalinity+0.2356 * waterTemperature+0.1625 * waterTransmission+1.8328 * landTemperature-0.9759 * landDewPointTemperature-1.8936 * landVisibility-1.7241 * landHumidity-1.6949 * landWindDirection-0.877 * landWindspd+1.4936 * landPressure-633.9387

a5=-0.1796 * nightTime-0.501 * waterOxygen+0.0525 * waterConductivity+0.0128 * waterPressure-0.0598 * waterDensity+0.0428 * waterSalinity+0.5022 * waterTemperature+0.0698 * waterTransmission-1.7361 * landTemperature-3.0451 * landDewPointTemperature+1.0209 * landVisibility+2.5793 * landHumidity-2.661 * landWindDirection+1.8332 * landWindspd-0.7988 * landPressure+66.9651

b5=-0.1174 * nightTime-0.3054 * waterOxygen-0.0129 * waterConductivity+1.5915 * waterPressure+0.1946 * waterDensity+0.1731 * waterSalinity-0.3293 * waterTemperature+0.3179 * waterTransmission-1.4789 * landTemperature-2.5118 * landDewPointTemperature-3.2267 * landVisibility-1.0163 * landHumidity-7.2031 * landWindDirection-1.4513 * landWindspd+0.9261 * landPressure+1.816

w = 0.004383

The following are detailed functions for the lower bound:

a0= 0.4315 * nightTime+0.2446 * waterOxygen+0.0112 * waterConductivity-0.8882 * waterPressure-0.0955 * waterDensity-0.0774 * waterSalinity+0.2516 * waterTemperature-0.2148 * waterTransmission+0.68 * landTemperature-0.0416 * landDewPointTemperature+0.6638 * landVisibility-1.3062 * landHumidity+1.2558 * landWindDirection+2.4605 * landWindspd+0.0194 * landPressure+218.6285

a1= 0.1894 * nightTime+0.0696 * waterOxygen-0.0184 * waterConductivity+1.1165 * waterPressure+0.1543 * waterDensity+0.0977 * waterSalinity-0.3467 * waterTemperature-1.6708 * waterTransmission+0.1222 * landTemperature+1.1128 * landDewPointTemperature-1.5306 * landVisibility-3.2983 * landHumidity+4.1561 * landWindDirection-3.4908 * landWindspd+1.4583 * landPressure-363.2282

b1= 0.5104 * nightTime-1.2715 * waterOxygen+0.0637 * waterConductivity+0.2249 * waterPressure+0.115 * waterDensity+0.2288 * waterSalinity+0.4441 * waterTemperature-1.5668 * waterTransmission+2.7203 * landTemperature+2.2383 * landDewPointTemperature-1.7465 * landVisibility-6.8352 * landHumidity+1.3435 * landWindDirection-0.3598 * landWindspd-0.4839 * landPressure+510.392

a2= 0.1379 * nightTime+1.0047 * waterOxygen-0.0426 * waterConductivity-0.1157 * waterPressure-0.0074 * waterDensity-0.0815 * waterSalinity-0.3465 * waterTemperature-0.1839 * waterTransmission-0.6506 * landTemperature-0.1118 * landDewPointTemperature-0.9243 * landVisibility+1.0168 * landHumidity-1.1512 * landWindDirection+1.6889 * landWindspd+0.469 * landPressure-138.0867

b2= 0.6115 * nightTime+0.3088 * waterOxygen+0.0012 * waterConductivity-0.5501 * waterPressure-0.0861 * waterDensity-0.0839 * waterSalinity+0.1592 * waterTemperature+0.4953 * waterTransmission+0.1547 * landTemperature-0.8544 * landDewPointTemperature+2.8437 * landVisibility-1.5319 * landHumidity+0.3979 * landWindDirection+2.6427 * landWindspd+0.4066 * landPressure-413.7512

a3= -0.2935 * nightTime-0.2575 * waterOxygen+0.0139 * waterConductivity+0.5475 * waterPressure-0.0923 * waterDensity-0.0245 * waterSalinity+0.2154 * waterTemperature-0.8197 * waterTransmission-0.1061 * landTemperature-1.4205 * landDewPointTemperature-0.1367 * landVisibility+2.3777 * landHumidity-1.6842 * landWindDirection+2.1908 * landWindspd-0.8058 * landPressure+148.1451

b3= -0.0871 * nightTime-0.6463 * waterOxygen+0.0219 * waterConductivity+1.3869 * waterPressure+0.2179 * waterDensity+0.239 * waterSalinity-0.0659 * waterTemperature+1.0372 * waterTransmission+0.42 * landTemperature+0.73 * landDewPointTemperature-5.2464 * landVisibility-0.8266 * landHumidity+0.1009 * landWindDirection-4.195 * landWindspd+0.8839 * landPressure-144.077

a4= -0.3453 * nightTime-0.0725 * waterOxygen+0.017 * waterConductivity-0.3179 * waterPressure+0.058 * waterDensity+0.0865 * waterSalinity+0.0269 * waterTemperature-0.5913 * waterTransmission-0.2047 * landTemperature+1.434 * landDewPointTemperature-1.2171 * landVisibility+1.4339 * landHumidity+2.3589 * landWindDirection-0.3975 * landWindspd-0.9352 * landPressure+259.4941

b4= 0.0599 * nightTime+0.9212 * waterOxygen-0.0317 * waterConductivity-0.8872 * waterPressure-0.1573 * waterDensity-0.2084 * waterSalinity-0.098 * waterTemperature+1.0918 * waterTransmission-1.023 * landTemperature-1.0651 * landDewPointTemperature-0.907 * landVisibility+1.5208 * landHumidity-2.3067 * landWindDirection+3.3332 * landWindspd-0.6412 * landPressure+161.5931

a5= -0.0047 * nightTime-0.1912 * waterOxygen+0.0465 * waterConductivity-1.2555 * waterPressure-0.1449 * waterDensity-0.061 * waterSalinity+0.5647 * waterTemperature-0.2008 * waterTransmission-0.7892 * landTemperature-1.5181 * landDewPointTemperature+3.295 * landVisibility+0.8515 * landHumidity-1.0804 * landWindDirection+3.4273 * landWindspd-1.2535 * landPressure+195.5641

b5= -0.2546 * nightTime-0.0621 * waterOxygen+0.0016 * waterConductivity+0.9516 * waterPressure+0.1455 * waterDensity+0.1215 * waterSalinity-0.1861 * waterTemperature+1.0537 * waterTransmission-1.3518 * landTemperature-0.4805 * landDewPointTemperature-2.9956 * landVisibility+0.3997 * landHumidity-2.9756 * landWindDirection-2.6749 * landWindspd+0.489 * landPressure-1.2517

w = 0.004383

The values of environmental factors change with time and the values of coefficients change accordingly.

Part E. Feature selection dataset

Figure 8 shows part of the dataset for feature selection. Each row represents one example. Each example has 7 water features (in yellow), 8 land features (in blue), 3 time features (in red) and one class label (in green).

Figures 9 and 10 illustrate results of feature selection on 60 minutes averaged data.

Part F.

Figures 11 and 12 illustrate results of feature selection on 15 minutes averaged data. The depth in the table is denotes by the index (the whole water column is divided into 800 bins).

Figure 13 illustrates results of feature selection on 5 minutes averaged data.

Figure 14 illustrates results of feature selection on 30 minutes averaged data.

Part G.

Figures 15 - 19 show the predicted class labels and the corresponding true class labels, both plotted as pictures. Overall speaking, the predictions are coherent with the true situation and are able to capture the subtle changes from day to day.

Time:UTC	month	day	hour	oxygen median	conductivity median	pressure median	density median	salinity	temperature	transmission	temp	dew	visi	humi	windr	windspd	pres	weather	hour relative sunrise	hour relative sunset	nightTime	class
	7	4	0	2.013494	3.295355	95.481	1024.427	30.93085	8.6173	59.23213	15.7	10.9	48.3	73	12	20	100.93	MostlyClo	-12	-4	481	0
	7	4	1	2.027192	3.29517	95.86	1024.428	30.929	8.61725	59.88124	14.9	10.4	48.3	74	13	28	100.9	MostlyClo	-11	-3	481	0
	7	4	2	1.839072	3.29759	96.112	1024.452	30.95715	8.61335	61.45924	13.2	10	48.3	81	14	17	100.87	MostlyClo	-10	-2	481	0
	7	4	3	1.941052	3.29468	96.1355	1024.439	30.93855	8.60035	59.88795	12.6	10.2	48.3	85	12	7	100.84	MostlyClo	-9	-1	481	0
	7	4	4	1.980918	3.293905	96.0345	1024.431	30.92885	8.60175	56.75434	12.6	10.1	48.3	85	11	4	100.81	MostlyClo	-8	0	481	1
	7	4	5	2.070133	3.292685	95.8795	1024.418	30.9142	8.6035	58.81245	12.4	10	48.3	85	0	0	100.84	MostlyClo	-7	1	481	0
	7	4	6	2.070325	3.292605	95.701	1024.417	30.9132	8.6041	58.31331	12.1	10.1	48.3	88	4	6	100.85	MostlyClo	-6	2	481	0
	7	4	7	2.142833	3.291585	95.614	1024.41	30.9046	8.6022	60.54825	11.7	10	48.3	89	0	0	100.86	MostlyClo	-5	3	481	0
	7	4	8	1.998248	3.293875	95.6215	1024.428	30.9272	8.60405	59.30264	12	10.1	48.3	88	15	6	100.87	Cloudy	-4	4	481	0
	7	4	9	2.144846	3.291245	95.757	1024.41	30.90345	8.60045	61.12014	11.9	10.2	48.3	89	16	4	100.86	Cloudy	-3	5	481	0
	7	4	10	2.223802	3.290155	95.9195	1024.403	30.89305	8.59995	61.85094	12.4	10.7	48.3	89	14	6	100.85	Cloudy	-2	6	481	1
	7	4	11	2.130506	3.29144	95.962	1024.414	30.90715	8.59725	61.00039	12.5	11.1	48.3	91	0	0	100.85	MostlyClo	-1	7	481	1
	7	4	12	1.958995	3.293745	95.9265	1024.433	30.9311	8.60085	61.45253	12.2	10.7	48.3	91	0	0	100.89	MostlyClo	0	8	481	1
	7	4	13	1.819275	3.297695	95.687	1024.445	30.9525	8.62	62.11059	12.5	10.6	48.3	88	13	7	100.93	Cloudy	1	9	481	0
	7	4	14	1.796036	3.29817	95.1845	1024.444	30.95475	8.62225	61.58347	13.1	11	48.3	87	13	7	100.95	MostlyClo	2	10	481	0
	7	4	15	1.965742	3.29381	94.5425	1024.424	30.9284	8.60235	60.07709	14.7	11.5	48.3	81	13	13	100.96	MostlyClo	3	11	481	0
	7	4	16	1.950803	3.29424	93.844	1024.422	30.9304	8.60615	59.23773	14.7	11.4	48.3	81	12	11	100.99	MostlyClo	4	12	481	0
	7	4	17	1.853621	3.296005	93.1545	1024.428	30.94325	8.612	58.25735	14.9	11.3	48.3	79	14	20	101	MostlyClo	5	13	481	0
	7	4	18	1.972246	3.29363	92.7065	1024.41	30.92235	8.608	55.11366	14.4	11.2	48.3	81	14	22	101.01	MostlyClo	6	14	481	0
	7	4	19	2.041353	3.292175	92.573	1024.4	30.90985	8.60565	52.41651	15.9	11.7	48.3	76	13	20	101.05	MostlyClo	7	15	481	0
	7	4	20	2.238062	3.29028	92.7425	1024.379	30.8831	8.61245	56.37606	17.1	12.4	48.3	74	13	19	101.08	MostlyClo	8	16	481	0

Figure 8: Part of the dataset for feature selection

Evaluator	search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	landTemperature	landDewPointTemperature	landVisibility	landHumidity	landWindDirection	landWindspd	landPressure	weather	hourRelativeSunrise	hourRelativeSunset	nightTime	evaluator 10 CV	accuracy(%)
ChiSquaredAttributeEval	Ranker	4	2	14	10	6	3	5	9	13	16	17	11	18	12	15	7	8	1		
GainRatioAttributeEval	Ranker	6	2	15	13	5	3	4	9	12	14	16	11	18	10	17	8	7	1		
InfoGainAttributeEval	Ranker	4	2	14	12	6	3	5	9	13	16	17	11	18	10	15	7	8	1		
OneRAttributeEval	Ranker	7	2	17	18	9	3	6	10	15	11	14	13	12	16	8	4	5	1		
SymmetricalUncertAttributeEval	Ranker	4	2	14	12	6	3	5	9	13	15	17	11	18	10	16	8	7	1		
CfsSubsetEval	BestFirst																x	x		BayesNet	86.76
	Exhaustive																			TAN	
	eSearch																x	x		BayesNet	86.76
	GeneticS					x										x				TAN	
	earch																	x		BayesNet	87.08
	GreedySt																x	x		TAN	
	epwise																			BayesNet	86.76
	LinearFor																x	x		TAN	
	wardSele																			BayesNet	86.76
	ction																			TAN	
	RandomS		x				x									x		x		BayesNet	86.27
	earch																			TAN	
	RankSear																x	x		BayesNet	86.76
	ch																			TAN	
	SubsetSiz																x	x		BayesNet	86.76
	eForward																			TAN	
	Selection																				

Figure 9: Feature selection on 60 minutes averaged data

Evaluator		search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	landTemperature	landDewPointTemperature	landVisibility	landHumidity	landWindDirection	landWindspd	landPressure	weather	hourRelativeSunrise	hourRelativeSunset	nightTime	evaluator 10 CV	accuracy(%)
ClassifierSubsetEval	BayesNet K2	BestFirst															x	x	x	BayesNet K2		87.15
		GeneticSearch	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet K2		87.15
		GreedyStepwise	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet K2		87.15
		LinearForwardSelection	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet K2		87.15
		RaceSearch	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet K2		82.49
		SubsetSizeForwardSelection	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet K2		87.15
	NBTree	RankSearch x ch(GainRatio)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	NBTree		89.36
		RankSearch x ch(ChiSquare)	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	NBTree		89.36
	J48	RankSearch x ch(GainRatio)	x			x	x	x	x								x	x	x	J48		89.92
		RankSearch x ch(InfoGain)	x			x	x	x	x								x	x	x	J48		89.92
		RankSearch x ch(ChiSquare)	x			x	x	x	x								x	x	x	J48		89.92
		RankSearch x ch(OneR)	x			x	x	x	x								x	x	x	J48		89.92
ConsistencySubsetEval		BestFirst	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet TAN		84.14
		RandomSearch	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet TAN		84
		Exhaustivx Search	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	BayesNet TAN		84.14
SymmetricalUncertAttributSetEval		FCBFSearch																x	x	BayesNet TAN		86.76

Figure 10: Feature selection on 60 minutes averaged data (continued)

Evaluator	search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	RelativeSunrise	RelativeSunset	nightTime	evaluator 10 CV	accuracy (%)
ChiSquareAttributeEval	Ranker	4	2	10	9	5	3	6	8	7	1		
	depth250	5	2	10	9	7	3	8	4	6	1		
	depth450	6	4	9	5	8	7	10	2	3	1		
	depth550	6	4	8	7	9	5	10	3	2	1		
GainRatioAttributeEval	Ranker	7	2	10	9	4	3	6	5	8	1		
	depth250	4	2	10	9	6	3	8	5	7	1		
	depth450	6	4	9	7	8	5	10	2	1	3		
	depth550	6	4	8	7	9	5	10	3	1	2		
InfoGainAttributeEval	Ranker	4	2	10	9	6	3	8	7	5	1		
	depth250	5	2	10	9	7	3	8	4	6	1		
	depth450	7	4	9	6	8	5	10	3	2	1		
	depth550	6	5	8	7	9	4	10	3	2	1		
OneRAttributeEval	Ranker	7	2	9	10	6	3	5	4	8	1		
	depth250	6	2	9	10	5	4	7	3	8	1		
	depth450	10	6	2	9	5	8	3	4	7	1		
	depth550	9	5	7	6	10	4	8	1	2	3		
SymmetricalUncertAttributeEval	Ranker	5	2	10	9	4	3	7	6	8	1		
	depth250	4	2	10	9	7	3	8	5	6	1		
	depth450	6	4	9	7	8	5	10	2	1	3		
	depth550	6	4	8	7	9	5	10	3	1	2		
CfsSubsetEval	exhaustive		x						x		x	BayesNet-TAN	85.79
	depth250		x						x		x	BayesNet-TAN	82.07
	depth450								x	x	x	BayesNet-TAN	85.33
	depth550								x	x	x	BayesNet-TAN	87.59

Figure 11: Feature selection on 15 minutes averaged data

Evaluator		search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	RelativeSunrise	RelativeSunset	nightTime	evaluator 10 CV	accuracy (%)
ClassifierSubsetEval	BayesNet-K2	BestFirst								x	x	x	BayesNet-K2	84.96
		GreedyStepwise								x	x	x	BayesNet-K2	84.96
		exhaustive								x	x	x	BayesNet-K2	84.96
	NBTree	BestFirst	x	x		x	x	x	x	x	x	x	NBTree	88.03
		RankSearch(GainRatio)	x	x		x	x	x	x	x	x	x	NBTree	88.03
		RankSearch(ChiSquare)	x	x		x	x	x	x	x	x	x	NBTree	88.03
	J48	BestFirst			x	x			x	x	x	x	J48	90.68
		depth250	x		x	x	x	x	x		x	x	J48	86.07
		depth450	x	x		x	x	x	x	x		x	J48	90
		depth550	x	x	x	x			x	x	x	x	J48	90.46
		RankSearch(GainRatio)	x	x	x	x	x	x	x	x	x	x	J48	89.99
		RankSearch(ChiSquare)	x	x	x	x	x	x	x	x	x	x	J48	89.99
ConsistencySubsetEval		BestFirst	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	84.57
		depth250	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	78.24
		depth450	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	84.4
		depth550	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	84.91
SymmetricalUncertAttributeSetEval		FCBFSearch								x		x	BayesNet-TAN	85.72
		depth250								x		x	BayesNet-TAN	81.62
		depth450									x	x	BayesNet-TAN	84.19
		depth550									x	x	BayesNet-TAN	86.61

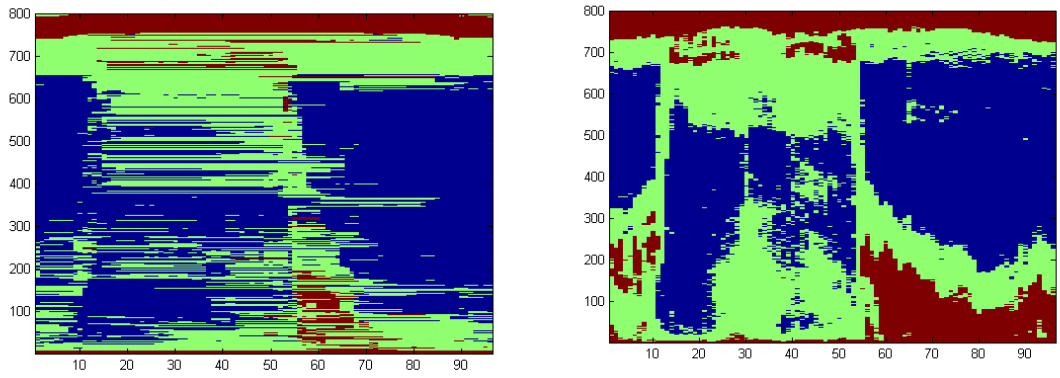
Figure 12: Feature selection on 15 minutes averaged data (continued)

Evaluator		search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	RelativeSunrise	RelativeSunset	nightTime	evaluator 10 CV	accuracy (%)
ChiSquareAttributeEval		Ranker	4	2	10	9	5	3	8	7	6	1		
GainRatioAttributeEval		Ranker	4	2	10	9	7	3	8	5	6	1		
InfoGainAttributeEval		Ranker	4	2	10	9	6	3	8	7	5	1		
OneRAttributeEval		Ranker	9	3	8	10	6	2	4	5	7	1		
SymmetricalUncertAttributeEval		Ranker	4	2	10	9	7	3	8	5	6	1		
CfsSubsetEval		exhaustive								x		x	BayesNet-TAN	86.91
ClassifierSubsetEval	BayesNet-K2	BestFirst								x	x	x	BayesNet-K2	84.62
		GreedyStepwise								x	x	x	BayesNet-K2	84.62
	NBTree	BestFirst	x	x	x	x	x	x		x	x	x	NBTree	88.66
		RankSearch(GainRatio)	x	x	x	x	x	x	x	x	x	x	NBTree	88.15
	J48	BestFirst			x		x	x	x	x	x	x	J48	90.32
		RankSearch(GainRatio)	x	x	x	x	x	x	x	x	x	x	J48	90.21
ConsistencySubsetEval		BestFirst	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	83.91
SymmetricalUncertAttributeSetEval		FCBFSearch								x		x	BayesNet-TAN	86.91

Figure 13: Feature selection on 5 minutes averaged data

Evaluator		search method	waterOxygen	waterConductivity	waterPressure	waterDensity	waterSalinity	waterTemperature	waterTransmission	RelativeSunrise	RelativeSunset	nightTime	evaluator 10 CV	accuracy (%)
ChiSquareAttributeEval		Ranker	4	3	10	9	5	2	6	7	8	1		
GainRatioAttributeEval		Ranker	7	2	10	9	5	3	8	6	4	1		
InfoGainAttributeEval		Ranker	4	2	10	9	5	3	6	8	7	1		
OneRAttributeEval		Ranker	6	2	10	9	5	3	7	4	8	1		
SymmetricalUncertAttributeEval		Ranker	5	2	10	9	6	3	8	7	4	1		
CfsSubsetEval		exhaustive									x	x	BayesNet-TAN	84.29
ClassifierSubsetEval	BayesNet-K2	BestFirst								x	x	x	BayesNet-K2	84.22
		GreedyStepwise								x	x	x	BayesNet-K2	84.22
	NBTree	BestFirst	x		x			x		x	x	x	NBTree	88.25
		RankSearch(GainRatio)	x	x			x	x	x	x	x	x	NBTree	87.34
	J48	BestFirst			x	x			x		x	x	J48	89.21
		RankSearch(GainRatio)	x	x			x	x	x	x	x	x	J48	89.39
ConsistencySubsetEval		BestFirst	x	x	x	x	x	x	x	x	x	x	BayesNet-TAN	84.85
SymmetricalUncertAttributeSetEval		FCBFSearch									x	x	BayesNet-Tan	84.29

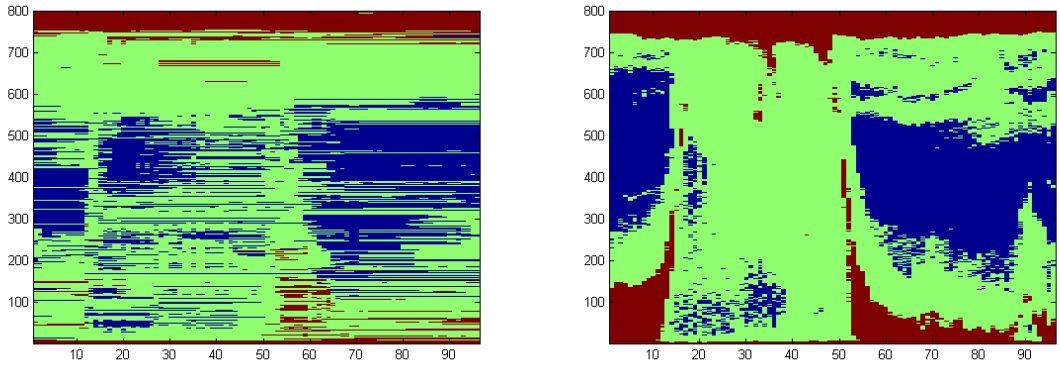
Figure 14: Feature selection on 30 minutes averaged data



(a) Prediction

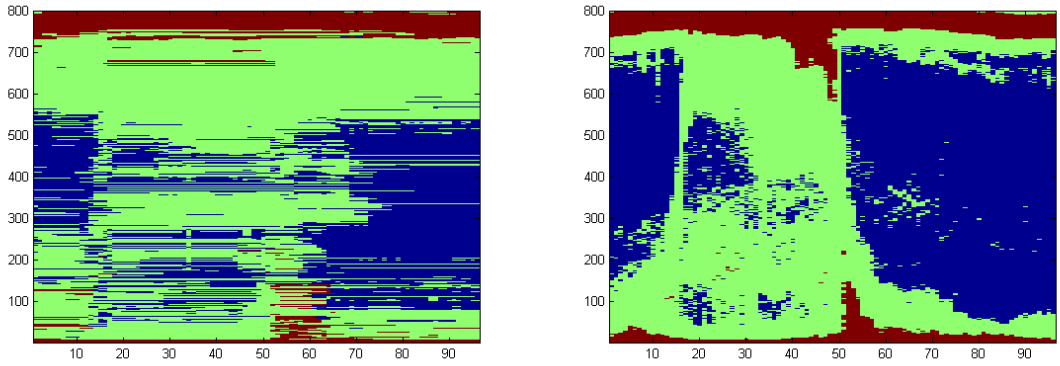
(b) True label

Figure 15: Predicted and true labels for April 2, 2009



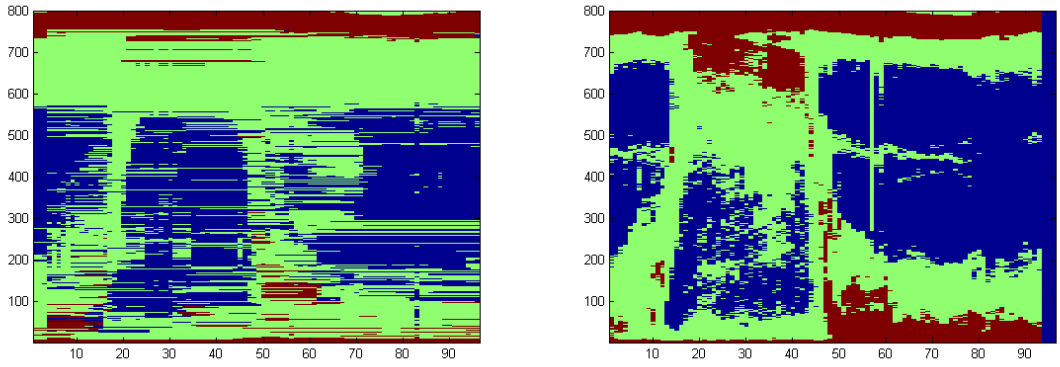
(a) Prediction (b) True label

Figure 16: Predicted and true labels for April 22, 2009



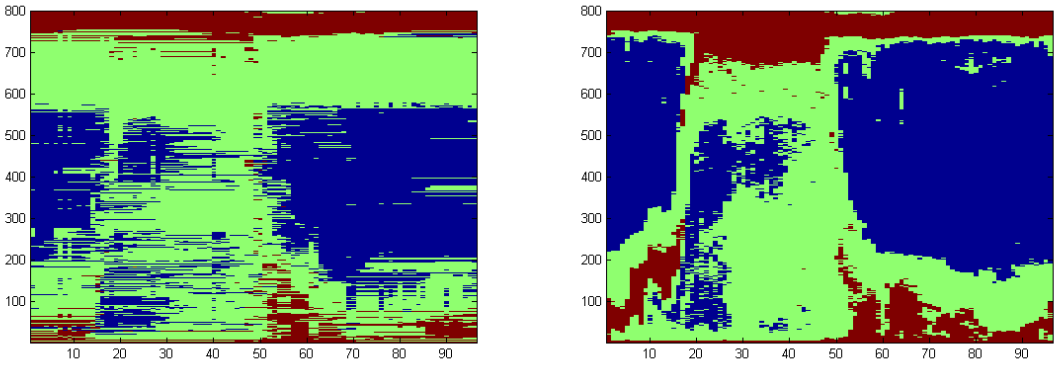
(a) Prediction (b) True label

Figure 17: Predicted and true labels for April 30, 2009



(a) Prediction (b) True label

Figure 18: Predicted and true labels for July 8, 2009



(a) Prediction

(b) True label

Figure 19: Predicted and true labels for July 30, 2009