VISUAL ANALYSIS OF OCEANIC MULTIVARIATE SPECIES TIME-SERIES DATA

by

Vegu Shree Rama Kamal Kumar

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

at

Dalhousie University
Halifax, Nova Scotia
June 2023

I dedicate this thesis to the amazing oceanic species found in Bay of Fundy from which the data was collected by Marine research organizations, making this research possible. Their existence is a testament to the incredible diversity of life in our oceans, seas, and coasts.

I am deeply grateful to Professor Stephen Brooks for his mentorship and guidance. His expertise and dedication have been instrumental in shaping this thesis.

To my family and friends, I extend my heartfelt gratitude for their unwavering support throughout this academic journey. Their encouragement has been invaluable in helping me overcome the challenges and celebrate the successes.

Last but not least, I want to express my appreciation to the experts for their collaboration, insights, and encouragement. Without their contributions, this research would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The goal of this project is to provide a data visualization dashboard which allows users to interactively analyze sequential movement patterns and rules of sharks and fishes. This will enhance the ability of oceanic experts to make educated decisions in order to minimize risks for marine ecosystems. The dashboard will present oceanic data from detection stations in Nova Scotia’s Bay of Fundy, which allows researchers to discover patterns in the movements of oceanic species. In order to determine the efficacy of the interactive dashboard, domain experts performed an evaluation using exploratory methods. This evaluation compared the dashboard’s ability to fulfill research questions and goals against the existing tools, and determined if this application would be more beneficial for existing practitioners, based on the feedback received and insights gained.

LIST OF ABBREVIATIONS USED

VMSP	Vertical Mining of Maximal Sequential Patterns
GUI	Graphical User Interface
API	Application Programming Interface
ID	Identity
ARM	Association Rule Mining
DM	Data Mining
FIM	Frequent Itemset Mining
MTS	Multivariate Time Series
SPM	Sequential Pattern Mining
SPMF	Sequential Pattern Mining Framework
SRM	Sequential Rule Mining
VA	Visual Analytics

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Dr. Stephen Brooks, for his invaluable advice and continuous support throughout this research and for the insightful comments and suggestions provided at every stage of this research.

I also express my sincere gratitude to the esteemed professors, Dr. Thomas Trappenberg, Dr. Fernando Paulovich, Dr. Vlado Kessel and Dr.Rita Orji who have tirelessly taught and guided me throughout my master’s degree program. I am grateful to all my course instructors, who have been instrumental in equipping me with the knowledge, skills, and confidence necessary to conduct my research and complete my thesis successfully.

In particular, I extend my heartfelt appreciation to my thesis defense committee, Professors Dr. Stephen Brooks, Dr. Derek Reilly and Dr. Joseph Malloch, for their insightful feedback, constructive criticism, and guidance, which have been invaluable in shaping my research work and refining my ideas.

I want to thank the Fundy Ocean Research Centre for Energy (FORCE) responsible for this project’s motivation and data. I wish to show my gratitude to Dr. Joanna Mills Flemming, Dr. Charles Bangley, Dr. Daniel Hasselman and all the other domain experts for all their feedback and reviews during the research.

I also would like to extend my gratitude to my colleagues at the lab, especially Mr. Leonardo Christino, for his support inside and outside this project. I would also like to thank my parents for their support and belief in me. I am deeply grateful for the support of my friends from home and the new friends I made at Dalhousie University.

CHAPTER 1

Introduction

Ocean data collection and access are undergoing a revolution with new applications and technologies applied to manage and understand our ocean. In particular the advance in development of underwater sensors for data collection has spawned a plethora of new ways to monitor and understand marine ecosystems. With an unprecedented ability to gather the data of the movements of sharks and fishes in local regions and analyze the environment, these new technologies open possibilities for advancements in research and decision-making.

There are several challenges faced while visualizing a large amount of data, including the need to render high density layers and to perform data preprocessing of large datasets and time consuming statistical calculations. In the Oceanic Dashboard Visualization project, we set out to track and display aquatic movements of sharks and fishes in the Bay of Fundy region, after extrapolating sequential patterns and rules using data mining techniques. With the dashboard created, companies and governing bodies can make more informed decisions on periodic functioning of turbines, which could reduce the possibility of entanglement of aquatic animals, and help create a sustainable marine ecology.

The dashboard comprises four modules, with the main interface providing the geographical visualization with a view to visualize selected patterns or rules based on different stations or regions. The ordering options help to refine the visualization based on the station, region, sharks or fishes, and patterns or rules. Subsets of the patterns and rules are filtered for exploration in the heat-map and timeline diagram, where a timeline plot is also used to visualize and filter the data through time. The proposed dashboard helps users who have some knowledge of the sequential patterns or rules to navigate and find the relevant information. The usability of the proposed dashboard to visualize and analyze the sequential rules is illustrated and then confirmed from the feedback received by conducting a user test with domain experts. By developing the dashboard, we hope to visualize aquatic animals' movement data to expedite the thought process of oceanic experts to make informed decisions and maximize insights.

1.1 Research Questions

In the development of the Visual Analytics dashboard, we explored the following research questions:

R1. How can we make time-series data representations in an accessible manner for easy interpretation while preserving and retrieving the authenticity which is crucial to maintain the integrity and accuracy of the underlying data?

R2. Does the visualization of pattern mining results on fish movements provide more insights to users than their current means of analysis?

R3. Which Pattern Mining algorithm is best for our ocean application of sharks and fish movement detections dataset?

R4. How can we visualize time-series tides data and patterns in fish movements in an expert-friendly way that meets the preferences of experts and maximizes the insights from the collected oceanic data?

R5. How to use the visualization of pattern mining to allow users to better analyze the movement of fish or sharks than the current means?

This project dashboard is an exploratory design research project. It seeks to improve visualizing plots and gain better insights in a reduced amount of time.

1.2 Definitions

Here we present some background and definitions of the data mining concepts used in the thesis.

- <u>Data Mining</u>: Data mining is the process of identifying patterns and extracting implicit information from large data sets using techniques of statistics, machine learning, and database systems.

  - <u>Predictive Data Mining</u>: Emphasizes drawing conclusions from the data in order to make predictions. For instance, regression and classification fall within this spectrum.

  - <u>Descriptive Data Mining</u>: Outlines knowledge gleaned from data. The emphasis is more on relationships, correlations, and patterns in the data than on a specific target variable. For instance, association rules and clustering fall within this spectrum.

- <u>Sequential Pattern Mining</u>: Identifies statistically significant patterns within data instances when the values are presented sequentially. The temporal order of the items is important in many situations, including the analysis of the genome, web clickstreams, and consumer purchasing patterns. In such cases, this method is suggested for finding sub-sequences among a group of sequences.

- Closed Sequential Patterns: Patterns that are not included in another pattern having the same support. They are lossless and this set is still quite large for some applications.

  For example, let's consider a sequence of transactions in a grocery store:

  Transaction 1: {Bread, Milk, Eggs}
  Transaction 2: {Bread, Butter}
  Transaction 3: {Bread, Milk, Butter}
  Transaction 4: {Bread, Butter, Jam}

  Let's assume we want to mine closed sequential patterns with a minimum support of 2. In this case, the closed sequential patterns would be:

  {Bread} (support:4),
  {Butter} (support:3),
  {Bread, Butter} (support:3)

  Here, {Bread} is a closed sequential pattern because it cannot be extended to include additional items without changing its support. Similarly, {Butter} and {Bread, Butter} are also closed sequential patterns.

- Maximal Sequential Patterns: patterns that are not included in another pattern. They are lossless with an extra database scan and generally much smaller than closed patterns. Fig. 1.3 displays the holistic subset depiction view of closed and maximal sequential patterns [24].

  Continuing with the grocery store example, let's consider the same transactions as before. The maximal sequential patterns would be:

  {Bread} (support:4),
  {Milk} (support:2),
  {Butter} (support:3),
  {Jam} (support:1),
  {Bread, Butter} (support:3),
  {Bread, Milk, Butter} (support:1)

In this case, all the closed sequential patterns we discussed earlier are also maximal sequential patterns. Additionally, {Milk}, {Jam}, and {Bread, Milk, Butter} are also maximal sequential patterns because they cannot be extended without changing their support.

Overall, closed sequential patterns capture unique patterns that are not subsumed by other patterns with the same support, while maximal sequential patterns capture patterns that are not subsumed by any other pattern, regardless of their support. Maximal patterns are generally smaller than closed patterns and provide a more concise representation of frequent sequential patterns.

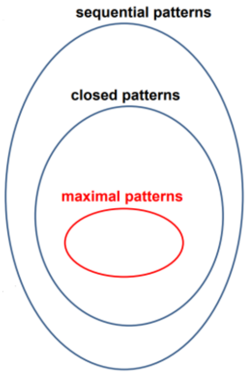


Figure 1.2: Broad View of Sequential Patterns [24]

- Sequential Rule Mining: Searches large sequences for relationships between variables. It aims to uncover robust rules found in databases by employing metrics. The purpose of rule mining in sequential databases is to determine the likelihood or probability of an object appearing in a subsequent location based on its occurrence in a preceding location, and vice versa.

- SPMF - VMSP and TRuleGrowth: SPMF is a Java-based open-source data mining toolkit that focuses on pattern mining, or finding patterns in data such as association rule mining, sequential pattern mining, sequential rule mining, clustering and classification and many more. It is made available under the GPL version 3 license.

After reading in-depth of the algorithms and investigating which are suitable for our oceanic multivariate dataset we opted for Efficient Vertical Mining of Maximal Sequential Patterns (VMSP) for mining maximal sequential patterns and T-Rule Growth for mining Sequential rules.

VMSP discovers all frequent maximal sequential patterns that occur in a sequence database [17]. Because of redundancy of multiple patterns they are very time-consuming to analyze and require much more storage space. So we are using VMSP as it counts only maximal sequential patterns and is better and faster than the rest of the similar SPMF algorithms [16].

1.3 Motivation and Objective:

Motivated by an exploratory research data set including the movement of marine life such as fishes and sharks in the Bay of Fundy, this research paper depicts a dashboard which is developed from the dataset and evaluated with oceanic experts to gain insights as to whether it is environmentally safe for a turbine technology company to run a turbine underwater in the Bay of Fundy region. The company's initial approach to setting up the turbines were resisted by environmental conservationists on the claim that mass destruction of marine mammals would occur if the turbines were successfully set up in the regions. Therefore, stakeholders were interested in finding out if there is any marine movement in the region where they plan to run the turbine. In the case that there is very minimal to no movement in the region, the turbine could be arguably set up for the benefit of electricity generation.

Using the dashboard as the medium for visualization, the domain experts will be able to deduce the patterns and rules of fish and shark movement within the three relevant regions and reason about the turbine's future in the Fundy Bay.

CHAPTER 2

Related Work

The dashboard that is currently developed takes inspiration from various research areas such as visual analytics, data visualization and domain specific research. Only the most pertinent works will be discussed and summarized in this section.

2.1 Time Series Data and Study Analysis

A time-series database is made up of sequences of values or events obtained through repeated time measurements. The values are typically measured at equal time intervals. Time-series databases are widely used in a variety of applications, including stock market analysis, economic and sales forecasting, budgetary analysis, utility studies, inventory studies, yield projections, workload projections, process and quality control, observation of natural phenomena, scientific and engineering experiments, and medical treatments [15]. Our work is related to observation of natural movements and the phenomenon of sharks and fishes in the Bay of Fundy region. Christopher et al. also studied a curated open access database of shark and ray life history traits and trends (Fig. 2.1.1) and termed it Sharkipedia [14].

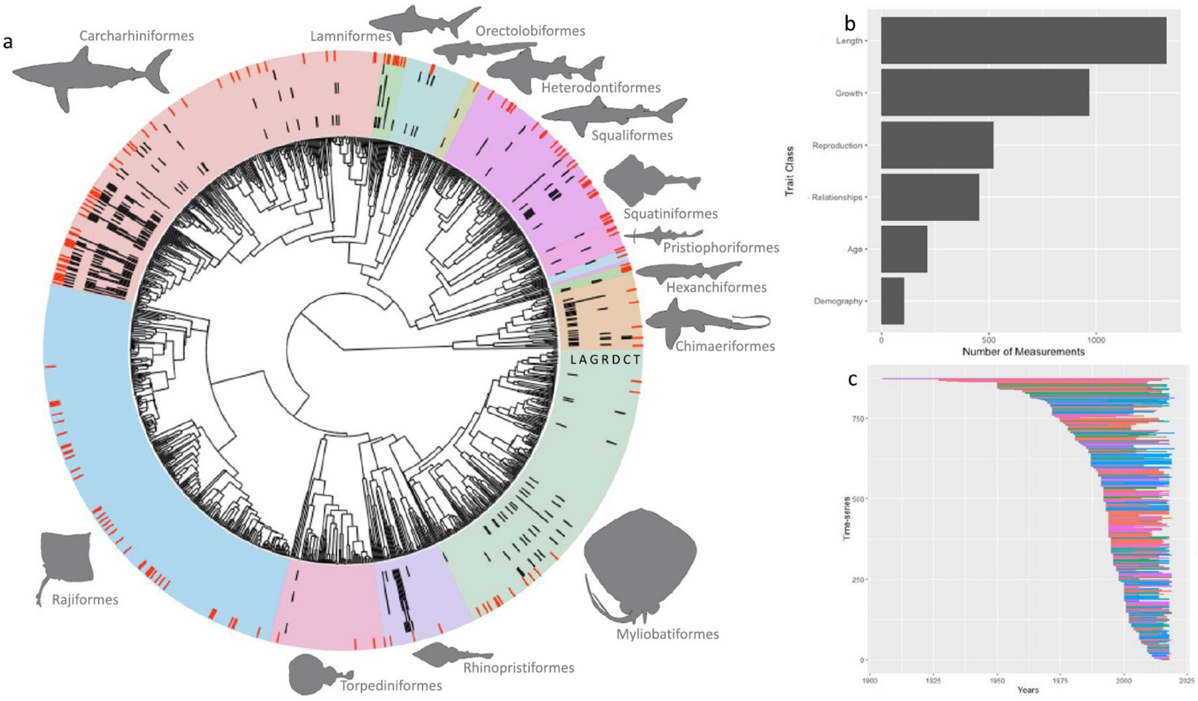


Figure 2.1.1: Summary of traits and trends of available length of time-series data [14]

Even though Sharkipedia gives a description of quantities, locations, and other data on marine life, it does not analyze patterns using SPMF algorithms. In addition, Sharkipedia does not provide information on the statistical likelihood of the patterns.

Oceanic time series analysis has two major goals which are modelling time series and forecasting time series. Trend analysis [5] has four major movements or components for characterizing time series data:

- Long term or Trend Movements: indicates the general direction of fishes or sharks in which they are moving over a long or short interval of time through a trend curve or line.
- Cyclic Variations or Cyclic Movement: long term oscillations about a trend curve or line which may or may not be periodic.
- Seasonal Movements or Variations: these are calendar related or systematic and vary with respect to seasons. For example related to cold weather conditions and abundance of food in certain regions.
- Irregular or Random Movements: series formed due to change of events, randomly due to storms or other external factors in the ocean. For example movements changed from installing a bridge, turbine or fishing patterns.

The aforementioned characteristics were instrumental in creating the patterns and rules that are showcased in our visualization dashboard. Outliers such as irregular and random movements were eliminated within the patterns. In contrast to using a treemap [5] for visualizing the movements, we opted to analyze them using a geo map since the dataset contains geographical coordinates.

2.2 Time Series Data to Sequential Pattern and Rule Mining

In the process of finding the long term or trend movements, we used sequential patterns and sequential rule mining for deducing the trends, so that we could analyze and visualize them clearly.

2.2.1 Sequential Pattern Mining

Analyzing and displaying sequence data for various analytical purposes has been the subject of a significant body of study. We examined the literature to identify the key components for mining sequential patterns, and based on our research, we have chosen to utilise SPMF's VMSP algorithm due to its high efficiency. For incorporating the display of sequential patterns into our visual design of the patterns we studied the findings mentioned in this section. Patterns can be easily displayed in the timeline plot [23] when there are few patterns (Fig. 2.2.1.1 & Fig. 2.2.1.2).

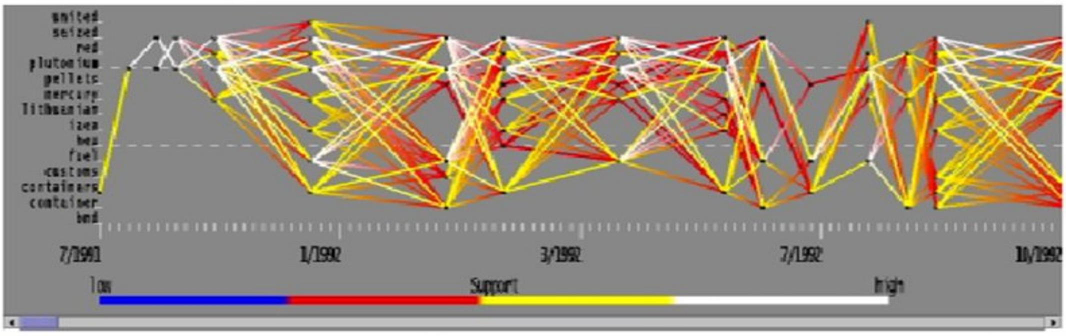


Figure 2.2.1.1 Snapshot of Sequential Patterns using monthly binning topic sets [23]

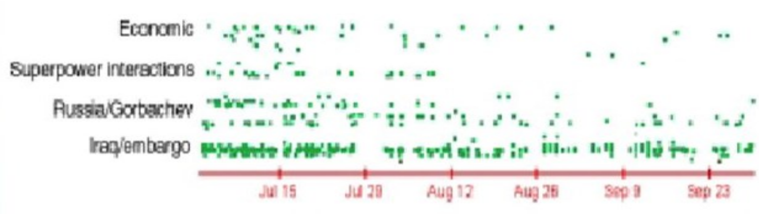


Figure 2.2.1.2 Topic combinations of Interest Vs Time [23]

However, when there are many patterns, as there are regularly in our dataset, these methods can run into scaling issues. The length of the sequences must also be taken into account in addition to the number of patterns. Numerous studies [3] examined various clustering methods to address problems with pattern visualization scalability and give a condensed picture for locating genomic sequence alignments using Gantt charts (Fig. 2.2.1.4). They all acknowledge that short, drawn-out episodes are preferable for their strategy but they have either computational or visual problems (Fig 2.2.1.3) while dealing with long sequences [11].

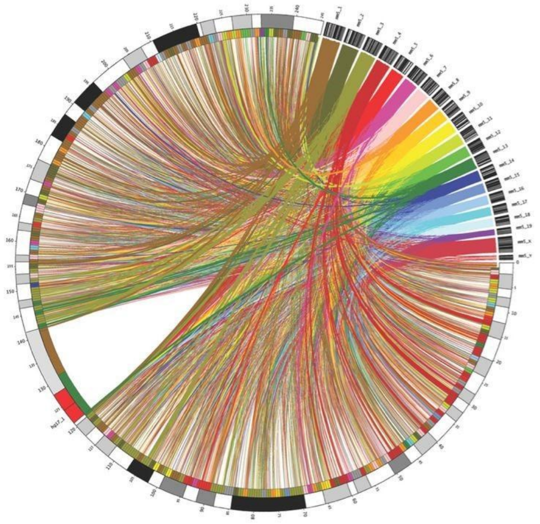


Figure 2.2.1.3: A visualization of chromosomal relationships within one genome. [11]

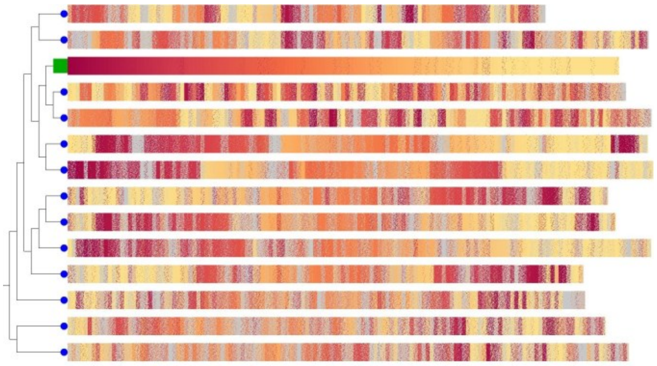
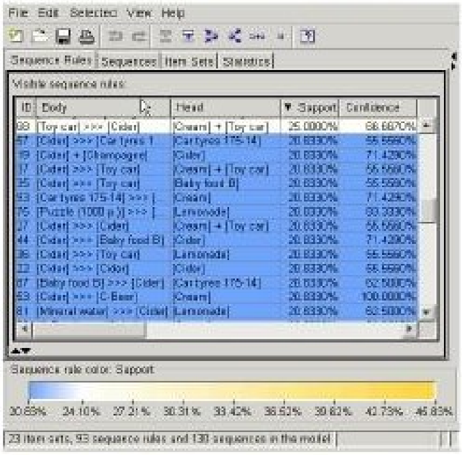


Figure 2.2.1.4: Genome sequence alignment visualization [3]

2.2.2 Sequential Rule Mining

Although mining is done using SPMF TRule Growth algorithm, with a large number of rules it is difficult to spot trends. However, there are some package visualizations that were created, which enabled us to visualize and provide high level plots to help in comparing directions (such as from one location to another), support (the relative frequency that the rules show up) and lift (the ratio of the observed support to that expected if the two rules were independent) among different factors in the data. DB2 Intelligent Miner tool [1], IBM offers a sequence rule visualizer in a tabular format [2] (Fig. 2.2.2.1). Even SPMF provides a SPMF viewer which uses Graphical User Interface (GUI) for all the rule and pattern mining algorithms [8] (Fig. 2.2.2.2). These have issues such as readability and understanding the nuances with large datasets as they are not graphically represented.



File Edit Selected View Help
Sequence Rules | Sequences | Item Sets | Statistics

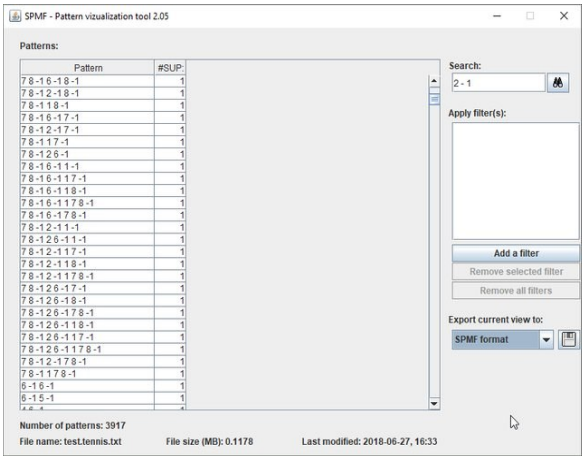
Visible sequence rules:

ID	Body	Head	▼ Support	Confidence
88	[Toy car] >>> [Cider]	[Cream] + [Toy car]	25.0000%	66.6670%
57	[Cider] >>> [Car tyres 1	[Car tyres 175-14]	20.8330%	55.5560%
19	[Cider] + [Champagne]	[Cider]	20.8330%	71.4290%
17	[Cider] >>> [Toy car]	[Cream] + [Toy car]	20.8330%	55.5560%
35	[Cider] >>> [Toy car]	[Baby food B]	20.8330%	55.5560%
53	[Car tyres 175-14] >>> [	[Cream]	20.8330%	71.4290%
75	[Puzzle (1000 p.)] >>> [	[Lemonade]	20.8330%	83.3330%
27	[Cider] >>> [Cider]	[Cream] + [Toy car]	20.8330%	55.5560%
44	[Cider] >>> [Baby food B]	[Cider]	20.8330%	71.4290%
36	[Cider] >>> [Toy car]	[Lemonade]	20.8330%	55.5560%
22	[Cider] >>> [Cider]	[Cider]	20.8330%	55.5560%
67	[Baby food B] >>> [Cider]	[Car tyres 175-14]	20.8330%	62.5000%
53	[Cider] >>> [C-Beer]	[Cream]	20.8330%	100.0000%
81	[Mineral water] >>> [Cider]	[Lemonade]	20.8330%	62.5000%

Sequence rule color: Support

23 item sets, 93 sequence rules and 130 sequences in the model

Figure 2.2.2.1: IBM Sequence Rules View [2]



Patterns:

Pattern	#SUP:
7 8 -1 6 -1 8 -1	1
7 8 -1 2 -1 8 -1	1
7 8 -1 1 8 -1	1
7 8 -1 6 -1 7 -1	1
7 8 -1 2 -1 7 -1	1
7 8 -1 1 7 -1	1
7 8 -1 2 6 -1	1
7 8 -1 6 -1 1 -1	1
7 8 -1 6 -1 1 7 -1	1
7 8 -1 6 -1 1 8 -1	1
7 8 -1 6 -1 1 7 8 -1	1
7 8 -1 6 -1 7 8 -1	1
7 8 -1 2 -1 1 -1	1
7 8 -1 2 6 -1 1 -1	1
7 8 -1 2 -1 1 7 -1	1
7 8 -1 2 -1 1 8 -1	1
7 8 -1 2 -1 1 7 8 -1	1
7 8 -1 2 6 -1 7 -1	1
7 8 -1 2 6 -1 8 -1	1
7 8 -1 2 6 -1 7 8 -1	1
7 8 -1 2 6 -1 1 8 -1	1
7 8 -1 2 6 -1 1 7 -1	1
7 8 -1 2 6 -1 1 7 8 -1	1
7 8 -1 2 -1 7 8 -1	1
7 8 -1 1 7 8 -1	1
6 -1 6 -1	1
6 -1 5 -1	1

Search: 2 - 1

Number of patterns: 3917

File name: test.tennis.txt File size (MB): 0.1178 Last modified: 2018-06-27, 16:33

Figure 2.2.2.2: SPMF Viewer [8]

By calculating the similarity between web pages using multidimensional scaling and showing the order of website visits by connecting section nodes in the scatter plot (Fig. 2.2.2.3), D'Ambrosio et al. [7] studied web usage and structure mining. Based on my own observation and analysis, I noticed that the visualization, which only displayed the antecedents and consequents for the rules, faced challenges related to clutter and readability.

In order to demonstrate sequence principles and forecast user browsing behaviour on websites, Siciliano et al. [21] employed a regression tree structure (Fig. 2.2.2.4). The user must dive down and study all branching sub patterns if they are interested in the events that take place close to the end of the sequences, which could take some time. An approach for creating sequential rules according to a user-specified temporal order in clinical circumstances was put forth by Shrestha et al. [22]. The consecutiveness of the elements in the rules is one of the restrictions given by the algorithm. This restriction prevents the generation of meaningful rules. To arrange the rules into hierarchies, the authors used dendrograms (Fig. 2.2.2.5). For situations when we do not require that every item in the rule be consecutive, this technique is constrictive.

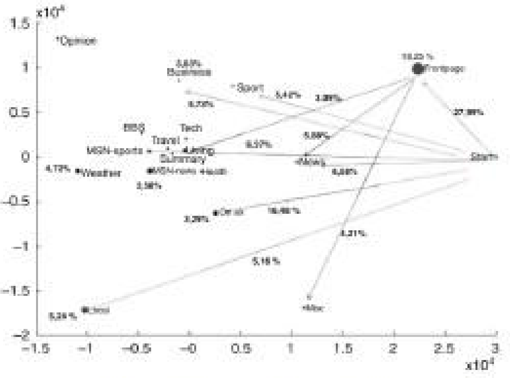


Figure 2.2.2.3: Scatter Plot by D'Ambrosio et al [7]

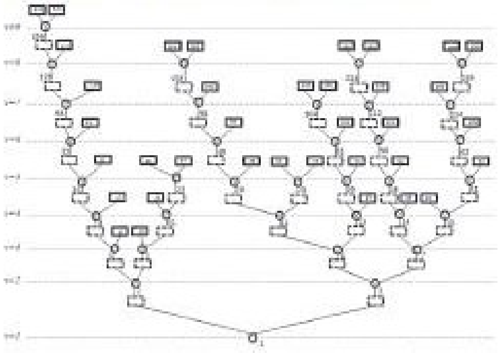


Figure 2.2.2.4: Tree by Siciliano et al. [21]

Figure 2.2.2.5: Dendrogram by Shrestha et al [22]

2.3 Importance of Support, Confidence and Minimum Gap in Sequential Pattern and Rule Mining

There are several objective metrics of patterns and rules discovered. These are based on the structure of found patterns and the statistics underpinning them. Rule support is an objective measure for sequential rules [15] of the type X and Y, expressing the percentage of movement from a sequence database that the provided rule fulfils. This is taken to be the probability P(X∪Y), where X∪Y signifies a movement from X to Y, that is, the union of regions X and Y. Another objective measure for sequence rules is confidence, which quantifies the degree of certainty of the detected link movement. This is taken to represent the conditional probability P(Y|X), or the likelihood of a movement from X to Y. Similarly for a sequential pattern only Support is being calculated by the algorithm which signifies the percentage number of times a sequence must appear to that of the total sequences in the dataset.

- support (X ⇒ Y) = P(X∪Y)
- confidence (X ⇒ Y) = P(Y|X)

The concept of a minimum gap was incorporated as a parameter into the pattern analysis algorithm. The minimum gap was set to zero, which means that only direct patterns would be included, as opposed to those in which the marine animal was detected at another location in between the locations of the pattern. A minimum gap of one (1) would allow for one other location, and so on.

Each Interestingness Measure (which is selecting and ranking patterns according to potential interest to the user), is paired with a threshold that the user can control from the frontend, which makes the dashboard interactive and helps users to visually analyze. For example Patterns and Rules that do not meet a support criterion of, say, 20% are regarded uninteresting and termed as outlier rules or patterns. Patterns and Rules below the threshold are more likely to represent noise, exceptions, or minority cases and are therefore less valuable to visualize. Even keeping higher thresholds [4] makes the algorithm run longer for noisy instances which results in slower application responses, and therefore, we naturally tend to avoid them.

2.4 Visualization using Various Plots with Interactivity

Even though from the previous section by tuning the parameters such as support, confidence and minimum gap, we could filter and evaluate the sequential patterns and rules, there are a lot of these complex patterns and rules which need to be visualized for better understanding.

One way to visualize data is through multiple views. Using a multidimensional explorer, one can view data from various perspectives. Aline Menin developed [9] a multidimensional graph (Fig. 2.4.1) explorer to provide a multi-dimensional perspective of a single data collection. They have portrayed a single data set in this work by employing many visualization approaches simultaneously, allowing the learner to view the varied perspectives. The user benefits from reduced cognitive time to process data, improved recognition of patterns and helps to evaluate data efficiently and communicate it better.

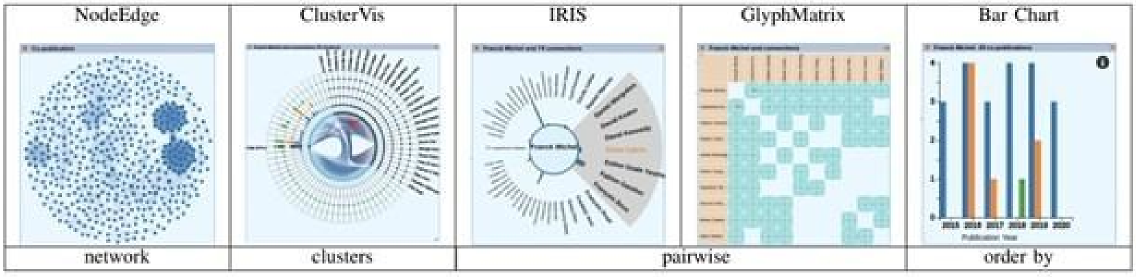


Figure 2.4.1 : Visualization Techniques available in MG Explorer [9]

The objective of this project is to offer a comprehensive view of the oceanic dataset from various perspectives. To achieve this, interactivity between multiple charts has been integrated to facilitate a multi-faceted exploration of the data. Using this visualization strategy as a foundation, we are building a project dashboard that integrates a geoplot, a heatmap, radial charts and timeline plots simultaneously. Using several complementary data visualization techniques together may increase the chances of the reader absorbing the data more efficiently with greater understanding and analysis. Furthermore, using various forms of representation ensures that the user faces minimal loss when analyzing data.

2.5 Role of Timeline plots, Heatmaps and Radial charts

A common approach for visualization of time series data is using timeline plots. It displays our data of stations and the regional movement of fishes and sharks through time. To differentiate and stand out from the competitors' user interfaces, data visualization was performed using Tableau which is a proprietary software. Data of uber rides in New York [18] was analyzed with the help of Tableau. In this, they used map chart visualization with several features such as colour filtering, effects, and borders and developed a timeline plots with the data available from the Uber rides. They appended a colour changing and text description feature. These features make the dashboard user-interface helpful in analyzing the data at any point in time. As for our project dashboard we used plotly timeplot and radial charts which provides the perspective of data through time. We do not believe that a text description feature [18] would be appropriate for proper comprehension of the data in our scenario because it would overload the visualization dashboard with information.

There are models where particular focus is given to studying a feature of marine mammals. The author of the track plot [20] utilizes digital recording tags attached to the marine mammals being observed in terms of their movement (Fig. 2.5.1) and vocalization over 24 hours. The tags consist of accelerometers, magnetometers and pressure sensors that use spiral ring visualization to map the trajectory of a fish in one region at a time and aid the user in demonstrating and analyzing the kinematic patterns from the marine mammals' movements.

This design highlights that the movements are rapidly georeferenced to a set of locations. Overall, this model aids in detecting marine mammals in a specific location at a particular time which is georeferenced if the surface fixed locations exist. Viewing it from an aggregate model perspective where multiple marine mammals within the same region are recorded, one can expect the 3D spiral model to overlap each other and create a data visualization conglomerate which would be hard to deduce the information from the overlapping structures. Since this paper focuses on the movement of multiple marine mammals within the same water region, it utilizes the basic idea of a bird's eye view [19]. It uses it to facilitate a 2D view of the origin and destination of the fish. This way, the user can view multiple trajectories of fishes across the water body and can select to zoom in or out, view the timeline plot, and observe the movement using radial maps, heatmaps and depth analysis in a particular region.

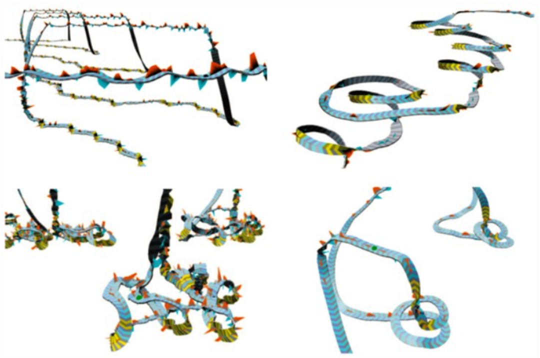


Figure 2.5.1: Track plot [20]

Mattias Persson utilized [12] different visualization techniques for Spatio-temporal data such as event maps, density trajectory maps, raster maps, heat maps (Fig. 2.5.2), choropleth maps, and line graphs to infer to the user about the climate data of different regions. They have used heat maps for visualizing the temperatures through time, similarly we have implemented heatmaps to visualize fish and shark detections through different phases of the day such as morning, evening and night. We used a heat map also to analyze and visualize the number of patterns or rules followed through various stations and regions. It aids the user in assessing the fish pattern distribution in different regions. Our approach was to position the heat map on the left-hand side of the dashboard.

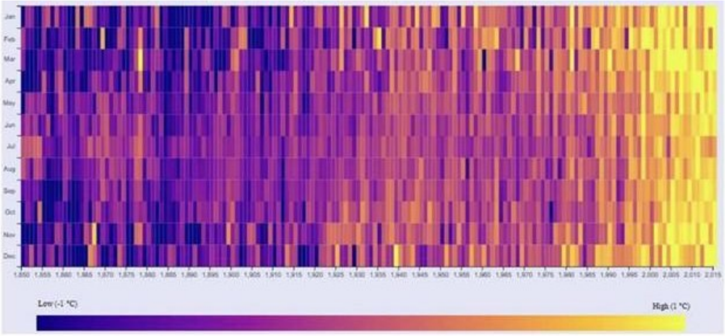


Figure 2.5.2: Heat map of the monthly temperatures 1850-2015 in the world [12]

Based on the above research heatmaps, timeline plots and radial charts have been chosen for better visualization of our Oceanic Dashboard for gaining accurate insights.

2.6 Geo Referencing with Data Clusters

Our data is centered around fish and shark movements in the Bay of Fundy. Consequently, when combined with latitude and longitude data, these patterns are visualized on a geo plot. However, mapping these patterns presents certain challenges, such as the high density of patterns in certain areas. To address this issue, clusters are utilized to effectively handle such situations. Razaei and Franti depicted [10] a real-time system using a server-side cluster. One of the most effective approaches for computing a vast amount of data relatively quickly is visualizing the data using clustering as shown in Fig 2.6.1. It aids in the effective processing of data and the reduction of latency during internet data transfer. The cluster visualization calculates the available cluster data and creates an interactive map and aids in assessing the relationship between the data [9]. We can find and divide the data set with the same features using this type of visualization. Using the cluster view and pattern count displayed, a user can perceive where the region of frequent movement is happening according to our oceanic time series dataset.

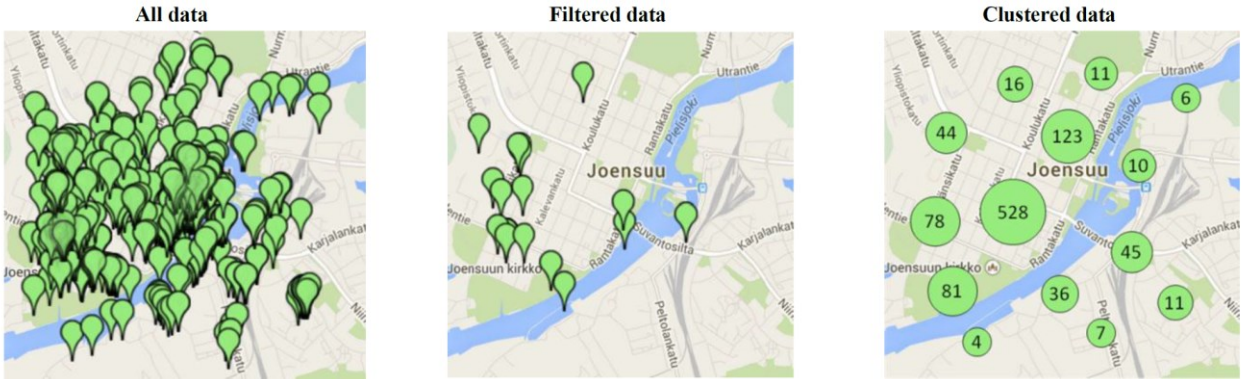


Figure 2.6.1: Clutter of 1000 geo-referenced data on the map (left), filtered by lasso select of 20 objects (middle) and clustering (right) [10]

2.7 Visualizations through Time Filter and Zone Reduction

In Visual Exploration of Temporal Origin-Destination data [13], Boyandin et al. has origins and the destinations displayed in two separate geo-maps, and the changes over time are represented in a heatmap. Implementation of a Lasso tool (Fig. 2.7.1) to select the origins is included such that it updates the heatmap with those origins. Similarly we have used the technique to select the regions and stations on our geo plot and it updates our heatmap and timeline plots. They also implemented a time filter (Fig. 2.7.2) in the heatmap so that countries are highlighted according to the year selected. We have incorporated the same technique but kept the time filter in the timeline plot, after which our oceanic time series data is filtered according to the specified time and relevant patterns and rules are formed and displayed. With this type of feature, it is helpful for the user to see the data at a particular frame of time. The inclusion of two features, namely lasso select and filtering heatmap, in our dashboard enhances user-friendliness and facilitates data analysis at specific time frames respectively.

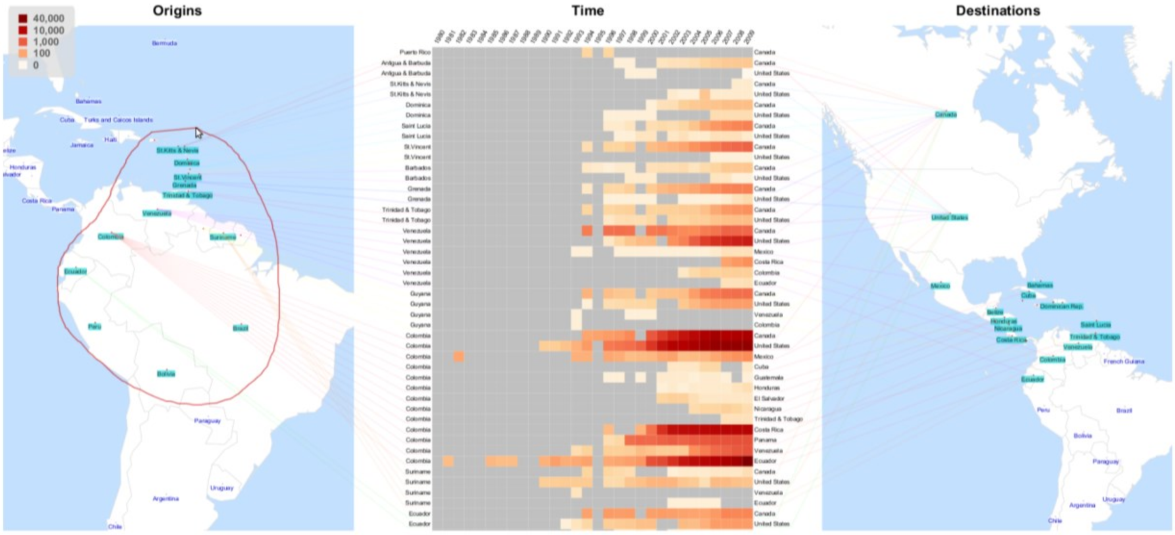


Figure 2.7.1: Selecting origins using lasso: When a selection is made, the heatmap is updated, so that only the flows between the selected origins and destinations are displayed. [13]

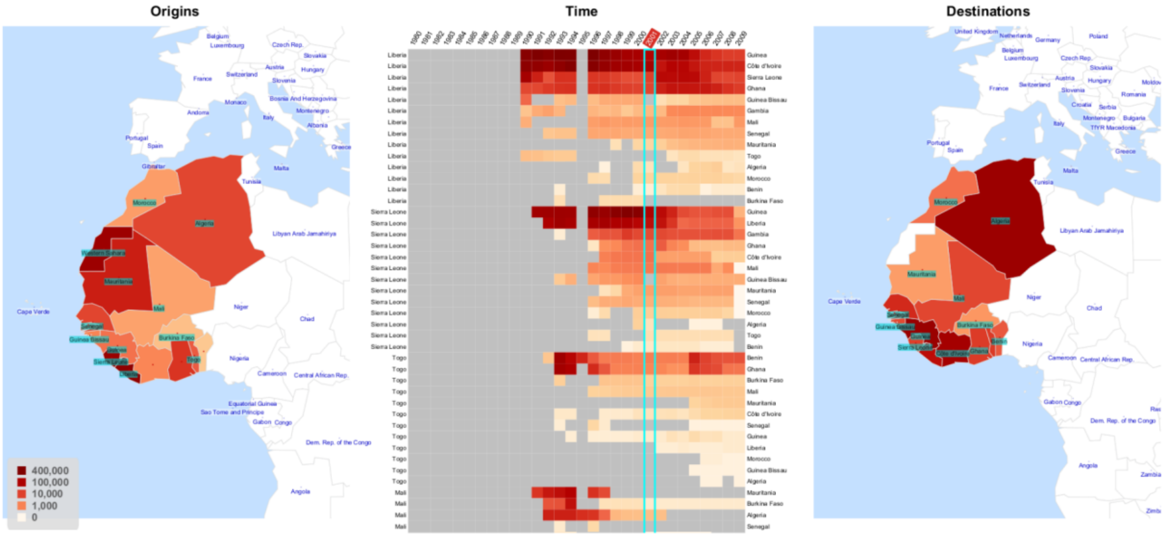


Figure 2.7.2: Selecting a year: Here the year 2001 is selected in the heatmap header, so the countries in the geographic maps are colored according to the total magnitudes of the outgoing and incoming flows in 2001. [13]

CHAPTER 3

Methodology

3.1 High Level Architecture

The high level architecture illustrates an overview of our visual analytics solution for exploring detections, sequential patterns and rules for our Sharks and Fishes dataset. The first step is to pre-process the data (Fig. 3.1 (1)). For our target domain in this study, we removed unnecessary columns, Unicode characters, special characters and kept identifiable column names based on guidance from domain experts. Then we dropped all the duplicate detections of the fishes and sharks in the data and created lists of sequences of regions and stations of the particular sharks and fishes based on their animal ID. Next, we employed data mining techniques, specifically VMSP to find sequential patterns and TRuleGrowth for rules in list sequences using SPMF frameworks(Fig. 3.1 (2)). Subsequently, the patterns and rules were post-processed to analyze the patterns from the visualization interactively (Fig. 3.1 (3)). Finally, we visualize sharks and fishes based on detections, data mining sequential patterns and rules (Fig. 3.1 (4)). All the phases are explained in the following sections.

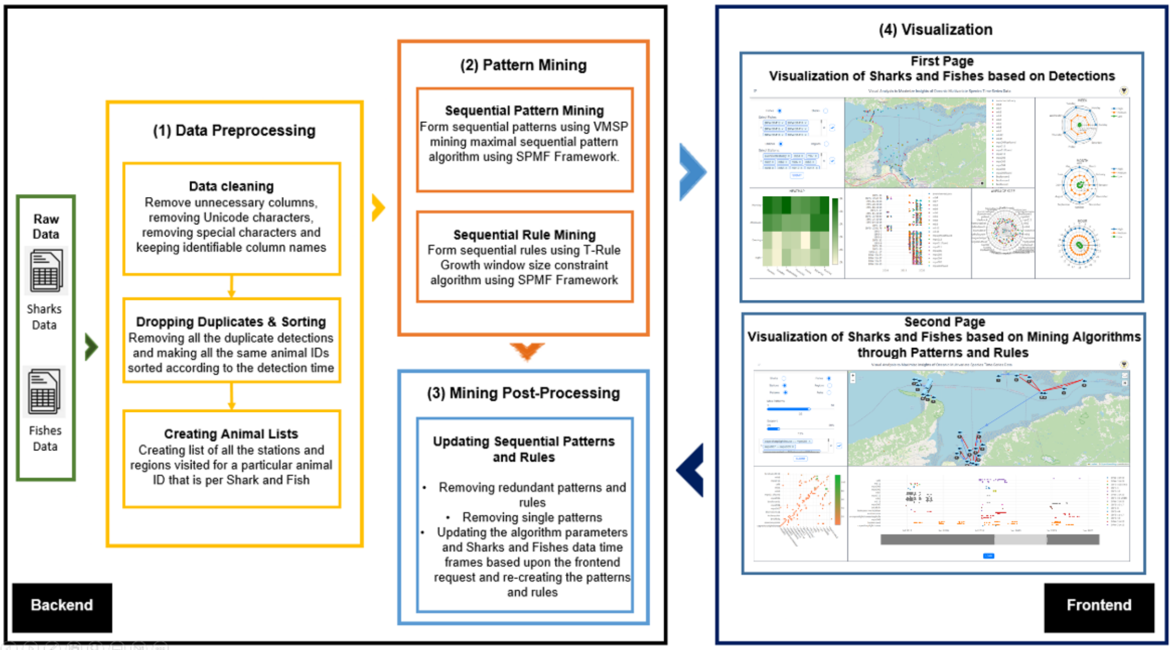

Raw Data
Sharks Data
Fishes Data
(1) Data Preprocessing
Data cleaning
Remove unnecessary columns, removing Unicode characters, removing special characters and keeping identifiable column names
Dropping Duplicates & Sorting
Removing all the duplicate detections and making all the same animal IDs sorted according to the detection time
Creating Animal Lists
Creating list of all the stations and regions visited for a particular animal ID that is per Shark and Fish
(2) Pattern Mining
Sequential Pattern Mining
Form sequential patterns using VMSP mining maximal sequential pattern algorithm using SPMF Framework.
Sequential Rule Mining
Form sequential rules using T-Rule Growth window size constraint algorithm using SPMF Framework
(3) Mining Post-Processing
Updating Sequential Patterns and Rules
Removing redundant patterns and rules
Removing single patterns
Updating the algorithm parameters and Sharks and Fishes data time frames based upon the frontend request and re-creating the patterns and rules
Backend
(4) Visualization
First Page
Visualization of Sharks and Fishes based on Detections
Second Page
Visualization of Sharks and Fishes based on Mining Algorithms through Patterns and Rules
Frontend

Figure 3.1: An Overview from Data Mining to Oceanic Dashboard Architecture

3.2 Preprocessing

For analysis and visualization, the data was collected from the shark and fish detections at the stations of the three regions in Bay of Fundy as sensors were fitted on their bodies, and when they come within a certain proximity to the station they are detected and noted. The detections are recorded and presented as rows in the Excel Workbook. This data is preprocessed and stored in the database which in-turn is reformatted and preprocessed to extract meaningful patterns that can be used for further analysis.

We designed a data pipeline for preprocessing and each stage produces a set of files that are used at various stages of analysis. For documentation we have further divided them into stages as given in Table 3.2.1 .We are doing all these pre-processing stages in a Python Web API on the server so as to have a minimal load on the front-end system. After data cleaning and removing unnecessary fields were left with seven fields which are animal id, detection time, latitude, longitude, region, station, and tide level (Table 3.2.2).

Preprocessing Stages	Steps Performed
1	Data cleaning such as removing irrelevant fields, null value rows, special and encoded characters etc.,
2	Dropping Duplicates and Sorting according to detection time
3	Creating Aquatic Animal Lists, for each animal_id for stations and regions.
4	Mining Sequential Patterns and Rules using VMSP and TRuleGrowth respectively.
5	Updating the Patterns and Rules Mining based on the support and max patterns and rules, timeframe updated from the frontend.

Table 3.2.1: Preprocessing steps for analysis and creating patterns and rules used for visualization

Animal ID	Detection Time	Latitude	Longitude	Station	Region	Tide Level
MMFSRP-20911	8/13/2017 9:56:00 AM	45.32678	-64.4341	blomidon no6	passage	0.543329
MMFSRP-19-15	2/25/2018 5:42:00 PM	45.33515	-64.4955	capesplit	passage	-0.62584
MMFSRP-20910	6/19/2019 3:13:00 AM	45.3293	-64.4064	mb11	avon	2.689851
MMFSRP-20910	8/28/2019 6:23:00 PM	45.33242	-64.4049	brickkiln	mainbasin	2.685773

Table 3.2.2: Samples of the Shark Dataset

In the oceanic time series data the tides are normalized between the range of -5 to +5 according to the preference of experts. We have divided this into low, medium and high and labelled them accordingly to have a better understanding and effective visualization of the data in the dashboard making distinct views in the radial charts of hour, day and month as represented in Table 3.2.3.

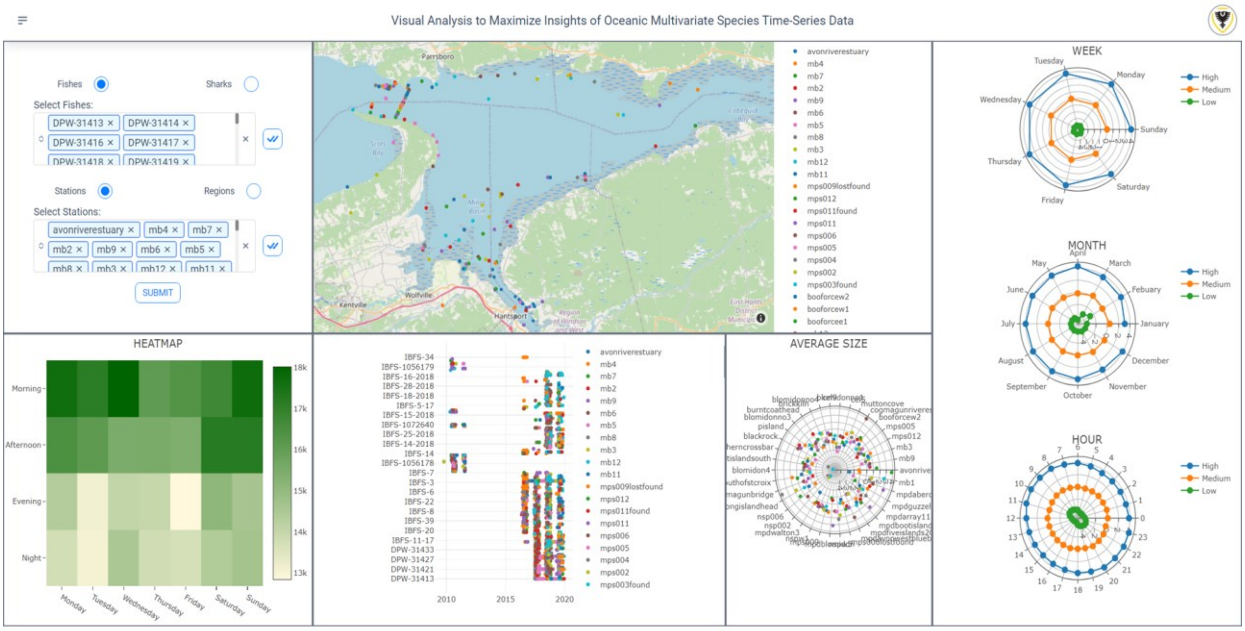
Height of Tides	Range Names
-5<=x<=-3	Low
-3<x<3	Medium
3<=x<=5	High

Table 3.2.3: Ranges of Labels for Tides

The characteristics of the data that we need to consider for the dashboard design are
as follows:

- **Vast gaps in timeline**: The time intervals between detections of a specific animal ID are significantly longer, resulting in large gaps of time in sequential detections.
- **Multivariate:** Each animal ID has a number of detections at stations and regions, and sequential patterns and rules are more prominent at these particular places so the data needs to be filtered and this option is given to the experts to change and analyze.
- **High Cardinality:** The number of unique animal IDs for fishes in data is high.

3.3 Visualization of Sharks and Fishes based on Detections



Visual Analysis to Maximize Insights of Oceanic Multivariate Species Time-Series Data

Fishes ● Sharks ○

Select Fishes: DPW-31413 × DPW-31414 × DPW-31416 × DPW-31417 × DPW-31418 × DPW-31419 ×

Stations ● Regions ○

Select Stations: avonriverestuary × mb4 × mb7 × mb2 × mb9 × mb6 × mb5 × mb8 × mb3 × mb12 × mb11 ×

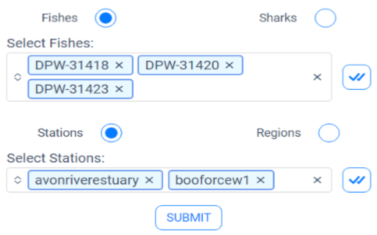
SUBMIT

Figure 3.3: An Overview of Visualization of Sharks and Fishes based on Detections

The complete overview of the Visualization of Sharks and Fishes based on Detections is presented in Fig 3.3. The selection panel is arranged in the top left of the page for customization by the user. After which all the plots such as geomap, heatmap, timeline plot, radial plots are generated and displayed. The radial plots are kept towards the right of the page and geo map and timeline plot are placed in between the selection panel and radial plots at the top and bottom respectively.

All the frontend dashboard is coded using the Svelte framework and JavaScript. Svelte framework is used as it is used in building fast web applications and building interactive user interfaces. Backend Web APIs are coded using python. Data is stored and retrieved from the PostgreSQL database. This database is used to handle multiple user requests for Web API and for smooth functioning of the overall dashboard.

3.3.1 Selection of Views through Stations and Regions



Fishes ● Sharks ○

Select Fishes: DPW-31418 × DPW-31420 × DPW-31423 ×

Stations ● Regions ○

Select Stations: avonriverestuary × booforcew1 ×

Figure 3.3.1: Selection panel for detections customization

After the selection of Fishes or Sharks, we implemented the feature of multiselect for selecting particular sharks or fishes from the dataset. We have also appended a “Select All” button for selecting all at once. In the similar fashion we have to select regions and stations. Following the customization we click on the “Submit” button to render all the plots.

3.3.2 Creation of Geoplot, Heatmap, Timeline plot and Radial Charts

The various plots and their functions are represented below.

- In the Geo Map plot we are displaying the location of the selected regions or stations (Figure 3.3.2.2).
- In the Heatmap we are calculating the number of detections found with an Animal ID in the time frame of Weekdays. When a hover is made the exact number of detections are displayed (Figure 3.3.2.1).
- In the timeline plot the X-axis is the Timeframe and Y-Axis has the Animal ID and colors are presented according to the stations or regions selected. These colors are matched with the colors represented in the Geo Map plot.

- In the radial chart “Average Tides”, the radius would be the average of the tide values calculated according to the stations and regions. The circular axes of the various stations selected in the options menu are displayed in a manner where most of the average tide size in radius are not collided. When we hover on the particular point the average tide data of that particular station is displayed.
- To the right the three radial charts the average tides values are plotted according to weekdays, month, and hour (Figure 3.3.2.3).
- These charts are again updated according to the selection of the timeline diagram based on the values presented in the selection panel.

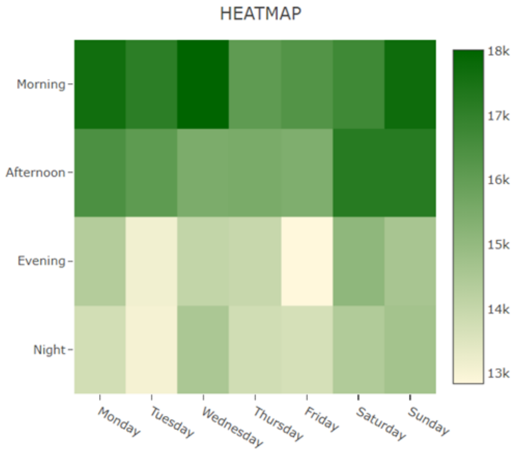


Figure 3.3.2.1: Displaying Detections count for various phases of day

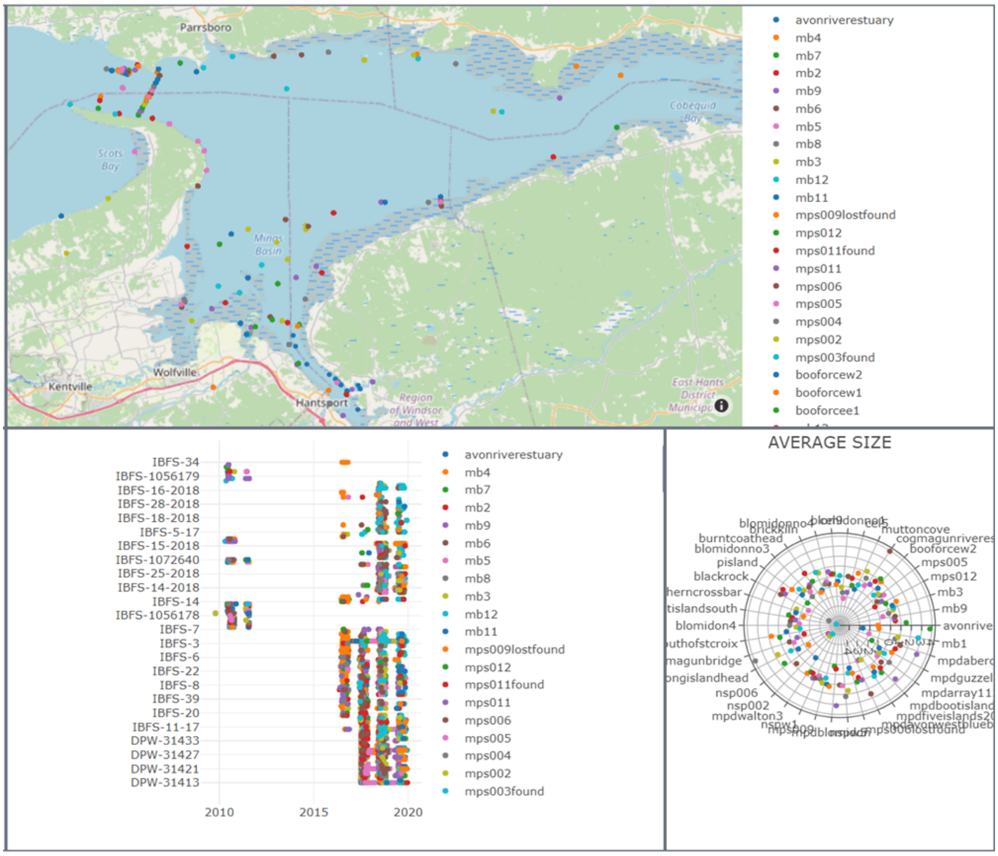


Figure 3.3.2.2: Displaying geo-map, timeline plot and average tides in a radial plot

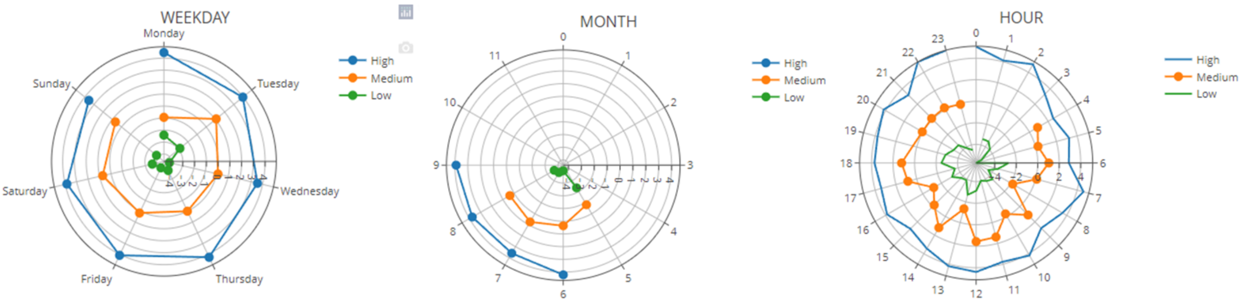
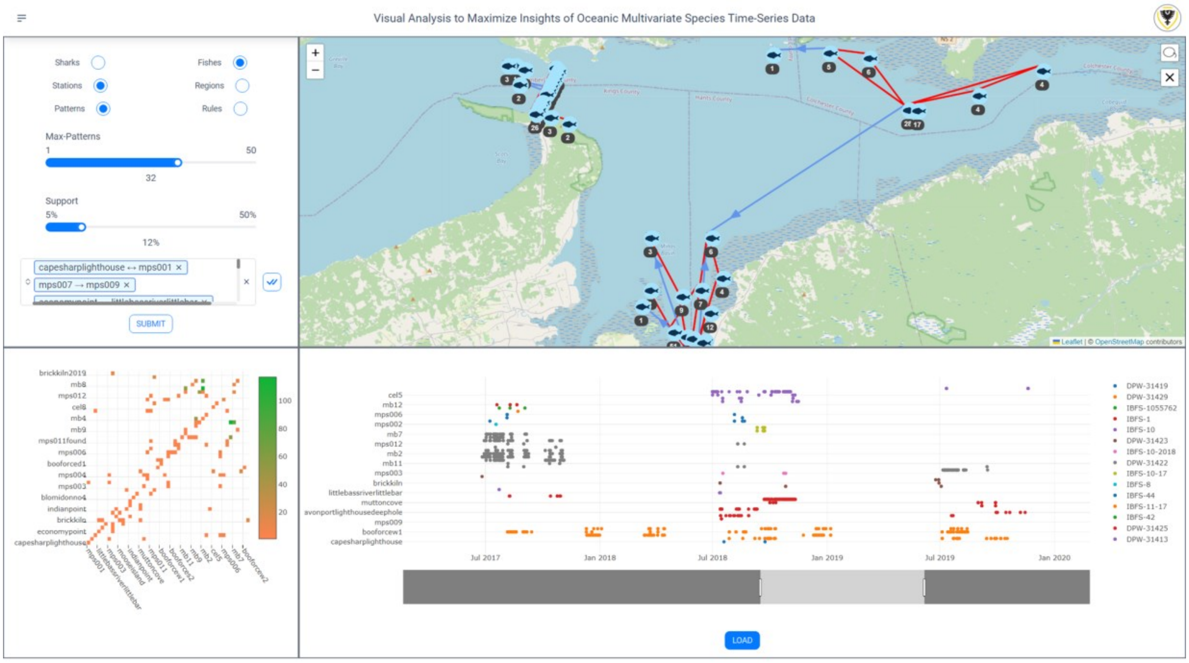


Figure 3.3.2.3: Displaying average tides in radial plots through weekday, month and hour

3.3.3 Challenges Faced

In the visualization of detections we faced a challenge in plotting the timeline diagram for the fish dataset as it was very large and the above graphs are also crowded with data. So for the fish and sharks data we decided to implement data mining algorithms to reduce the burden on the frontend to visualize efficiently and for faster processing.

3.4 Visualization of Sharks and Fishes based on Mining Algorithms through Patterns and Rules



Sharks ○ Fishes ●
Stations ● Regions ○
Patterns ● Rules ○

Max-Patterns 1 50: 32

Support 5% 50%: 12%

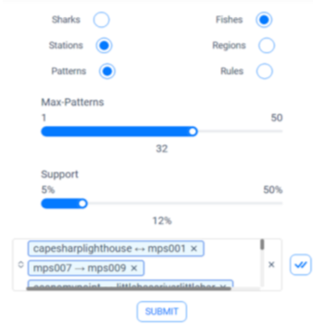
capesharplighthouse ↔ mps001 ×
mps007 → mps009 ×

Figure 3.4: An Overview of Visualization of Sharks and Fishes based on Mining Algorithms

In the visualization (Fig 3.4) we modified the solution and shifted the focus from fishes to stations. In other words now we wanted to study what are the patterns and rules of the fishes and sharks for each region and in multiple stations. This was achieved by using mining algorithms, specifically the VMSP SPMF algorithm to find the patterns and T-Rule Growth algorithm. By mining and analyzing in such a manner we could reduce the complexity from counting and viewing each detection to counting patterns and rules formed in each station and region. We further enhanced the ability to analyze and visualize these patterns discussed in further sections.

3.4.1 Selection of Views through Patterns and Rules

After the selection of Fishes or Sharks, we implemented the feature of multiselect for selecting stations or regions, and patterns or rules to generate the mining results from the dataset. We also added two additional parameters (max-patterns or rules, and support percentage) to provide additional customizability. Thereafter clicking on load the front end executes the Web API which generates patterns and rules accordingly and displayed in the multi select. We also added a “Select All” button for selecting all patterns or rules. Following the customizations we click on the “Submit” button to render all the plots such as Geo Map, timeline plot and heatmap.



Sharks ○ Fishes ●

Stations ● Regions ○

Patterns ● Rules ○

Max-Patterns: 1 – 50, 32

Support: 5% – 50%, 12%

capesharplighthouse ↔ mps001 ×

mps007 → mps009 ×

Figure 3.4.1.1: Selection panel for patterns and rules customization

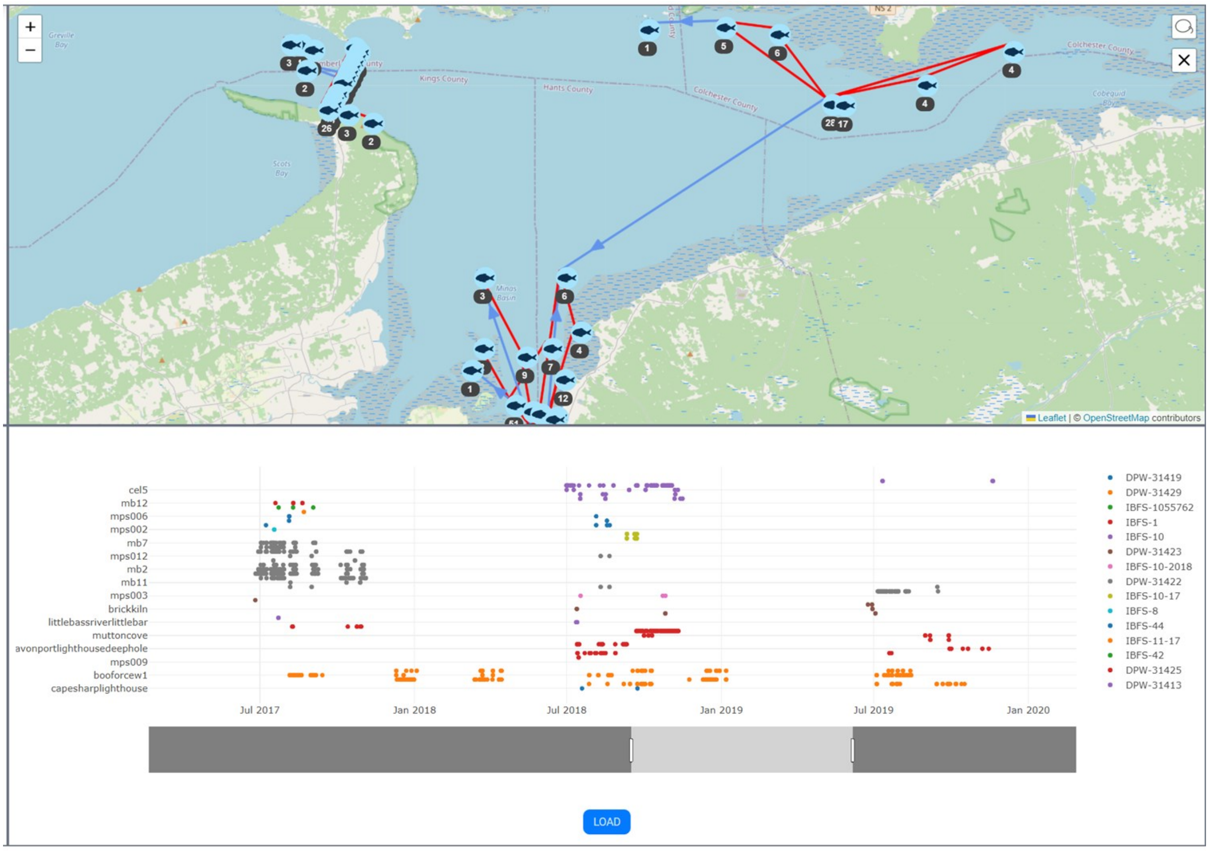


Figure 3.4.2.1: Displaying patterns in geo-map and accordingly timeline plot

3.4.2 Choosing Customizability and Exploration Strategies

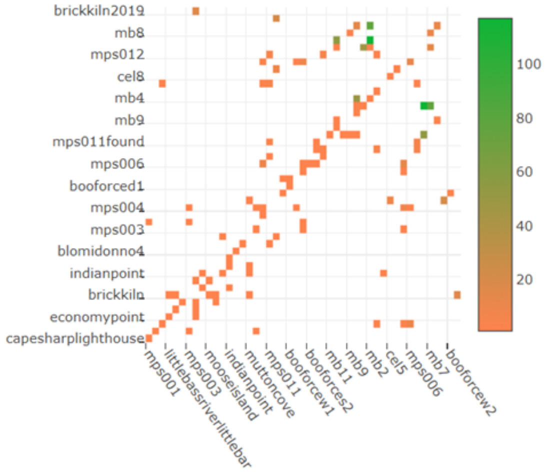


Figure 3.4.2.2: Displaying Patterns count for various stations

The main functionality and features present in the visualization based on mining patterns and rules are as follows:

- The Geo map which displays the patterns and rules selected with directions between regions and stations. The red line indicates the bi-directional pattern or rule between that particular station or region. The green line with the arrow indicates the one way direction of the pattern or rule. There is a lasso select to select particular stations or regions to visualize (Fig 3.4.2.1).
- The heat map shows the total number of patterns or rules counted between stations or regions selected by the user to analyze.
- The timeline diagram which has X-axis as time and Y-Axis as Stations, or Regions and colors indicate Animal IDs. Overall, it displays which stations were active in which period of time.

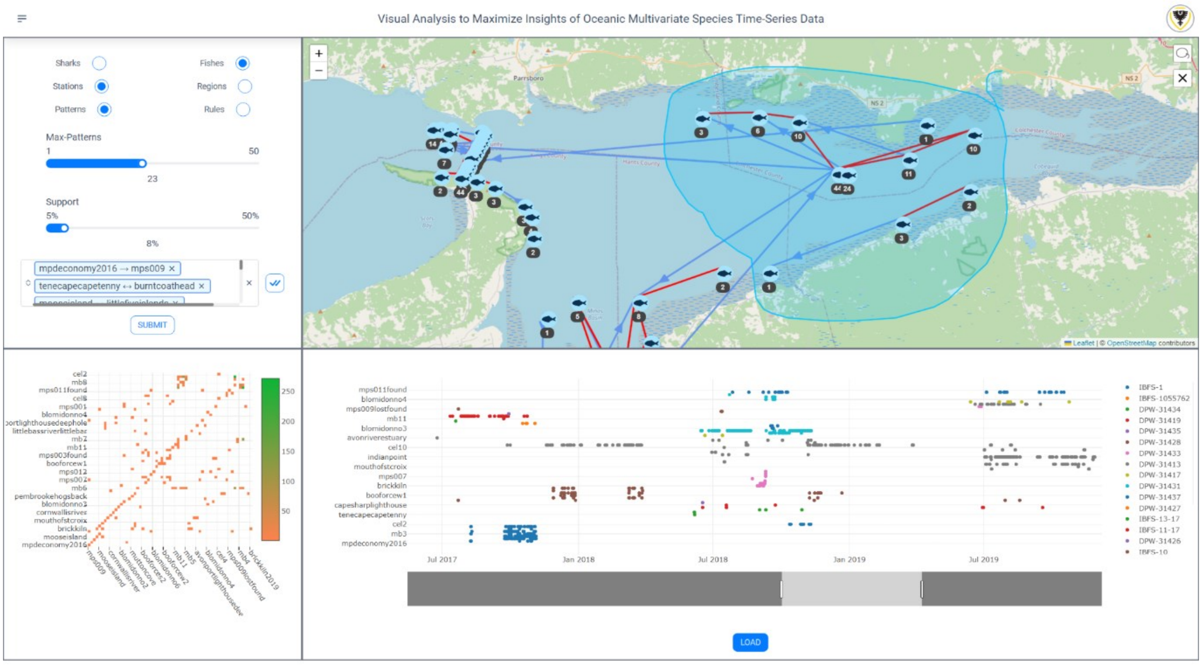
3.4.3 Interactivity among Geoplot, Heatmap and Timeline plot

There are some linkages implemented between the Geoplot, heatmap and timeline plot for effective analyzing and visualization:

- **Lasso Select:** Using the lasso select in the geographical plot to filter respective stations or regions, only filtered stations or regions are updated in heatmap and timeline plot for visualization (Fig 3.4.2.2).
- **Time Filter:** Filtering time in the timeline plot and clicking on load beside the timeline plot updates the patterns or rules according to the options selected with the new timeframe and geographical plot and heatmap gets updated. These patterns/rules get

updated and are re-plotted in the map and heatmap dynamically after running the specific mining algorithms again with the new parameters in the Web API.

This is done to analyze the patterns and rules based on changing the values in the dataset and algorithms just to provide more interactivity and deep analysis. Here deep analysis refers to the thorough examination and investigation of the patterns and rules, uncovering valuable insights and gaining a deeper understanding of the data through the lasso select and time filter exploratory techniques.



Visual Analysis to Maximize Insights of Oceanic Multivariate Species Time-Series Data

Sharks ○ Fishes ●
Stations ● Regions ○
Patterns ● Rules ○

Max-Patterns: 23 (1–50)

Support: 8% (5%–50%)

mpdeconomy2016 → mps009 ×
tenecapecapetenny ↔ burntcoathead ×

Figure 3.4.3.1: Using Lasso select to filter the stations for specific visualization

After the completion of the dashboard, it will be presented to experts for review and discussed in detail about the pros and cons, how different this is from their current way of analyzing and if we are getting particular details right.

CHAPTER 4

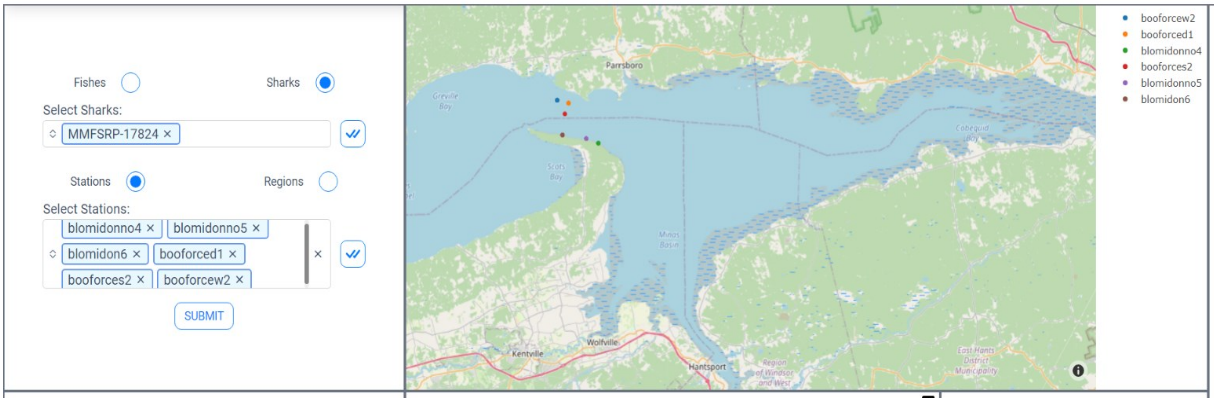
Use Cases

Two use cases discussed below demonstrate how the oceanic dashboard can be utilized to gather insights on the oceanic data of sharks and fishes. In the first use case, we look for interesting features regarding the number of detections gathered at various times of the day, average of tides at various zones, and average of tides at various days, months and hours. The second use case is to gain insights from sequential patterns and rules plotted for various stations and regions in relation to time.

4.1 Gaining Insights on Sharks and Fishes based on Detections

In one scenario Nora, a fictional user, wants to analyze shark activity and the average tides at “blomidonno” and “booforced” stations across days, hours and months. Nora does not have time to go through each of the detections by filtering from the Excel sheets, so she decides to use our oceanic dashboard to find in which specific time most activity is observed. Nora will also be able to observe what is the effect of tides on marine activity.

She first selects the sharks dataset to analyze. Next she selects a particular shark by its animal ID. After clicking on “Submit” the Geo map, heatmap and radial charts are updated. For example, the shark “MMFSRP-17824” is selected, we can see that this particular shark was only visiting the passage regions and never entered the Avon and Main Basin regions in the complete timeframe of the dataset, since there are no stations highlighted for the other regions as shown in Fig. 4.1.1.



Fishes ○ Sharks ●

Select Sharks: MMFSRP-17824 ×

Stations ● Regions ○

Select Stations: blomidonno4 × blomidonno5 × blomidon6 × booforced1 × booforces2 × booforcew2 ×

Figure 4.1.1: Selecting one shark to check the trend

In the second scenario Nora selects all the sharks for all stations and tries to study the visualizations. Here she notices that there were more detections happening on Monday nights, compared to other days. She also notices from the weekday radial chart that on Mondays the average high tides were higher than on other days. From the hour radial chart she notices average high tides are happening during the night from 22:00 hrs to 00:00 hrs as shown in Fig. 4.1.2. It may be deduced that due to high tides on Monday nights the shark activity was higher, thus resulting in more detections between stations. She also deduces that the shark activity is higher on Saturday, Sunday, and Monday than the rest of the week.

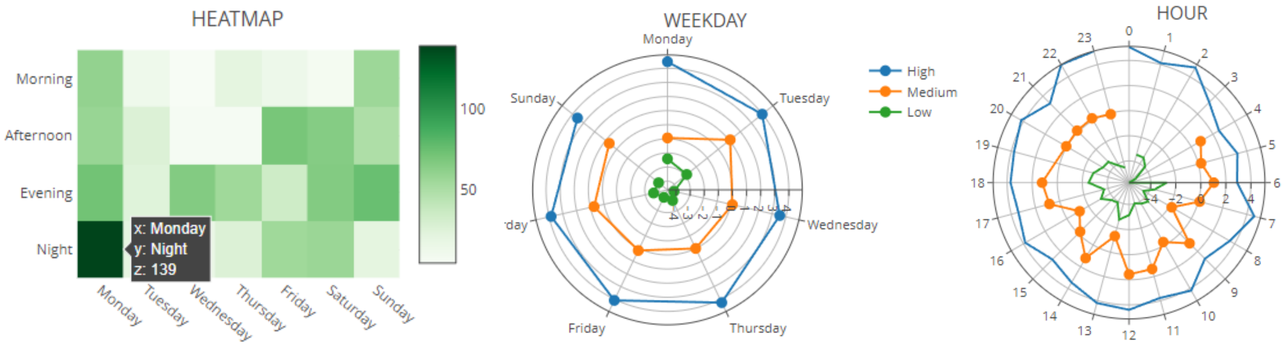


Figure 4.1.2: Analyzing Heatmap, Weekday Radial and Hour Radial charts

4.2 Gaining Insights on Sharks and Fishes based on Mining Algorithms through Patterns and Rules

In this scenario Alice wants to analyze the fishes and sharks mined with sequential patterns and rules focusing on stations. Alice finds it time consuming to view the raw patterns and rules formed by the SPMF mining algorithms, so she decides to use our oceanic dashboard to find which locations are most active by studying the number of patterns or rules formed in specific locations. She is also interested to know where the movements are happening such as the directions of the fishes and sharks.

Initially, she selects the sharks dataset and their locations according to stations. She then selects patterns to be mined with support at 5% and maximum patterns set to seven and clicks on the “load” button to start the mining. After the mining is done the patterns are updated in the multi-select tool for her to choose. Here she selects “Select All” and submits. The Geo map, heatmap and timeline plots are then updated. She then does the same steps with the fishes dataset, and by comparing both fishes and sharks as shown in Fig. 4.2.1 she deduces following insights:

- Fishes exhibit high movements for short distances and are nearer to the coast lines. Sharks are moving for long distances and travelling across the center regions of Bay of Fundy.

- Fishes are entering into the smaller channels of the water bodies towards south whereas Sharks movements are confined to larger areas of water bodies.
- There is less bidirectionality movement between stations for sharks when compared to that of fishes. Fishes are mostly revolving in circles between the stations of closer proximity.

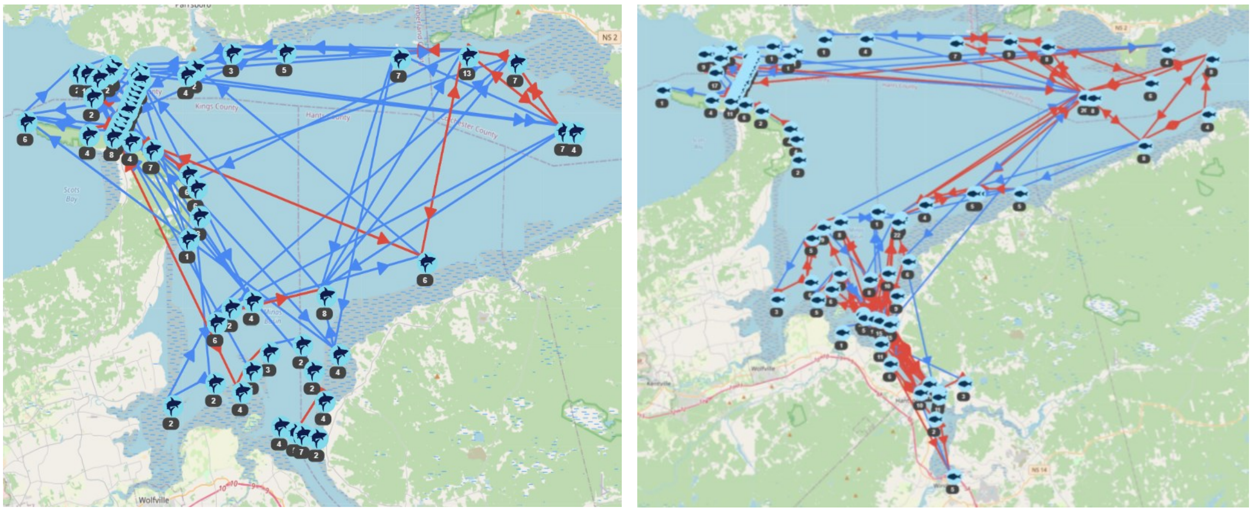


Figure 4.2.1: Comparing sequential patterns of Sharks (Left) and Fishes (Right)

She also notices that most of the fishes and shark detections in the timeline plot are happening around the summer season which is from June to October for the years analyzed as shown in Figure 4.2.2.

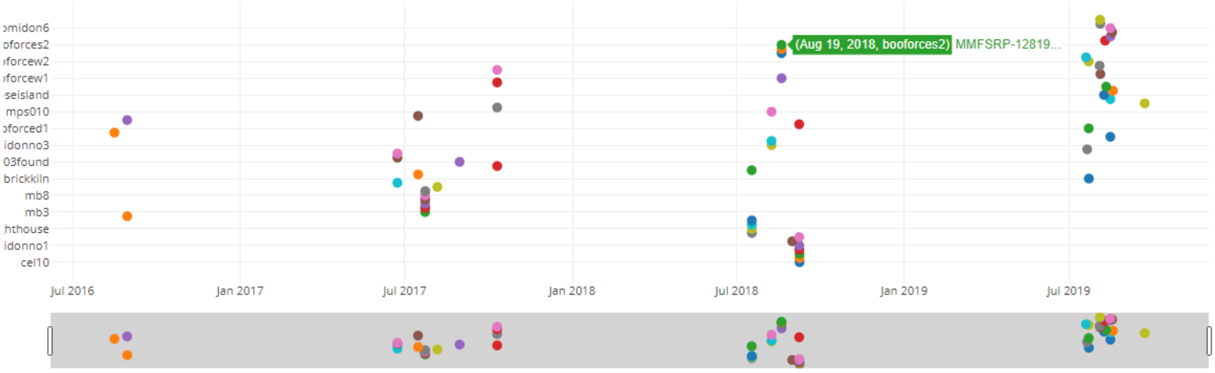
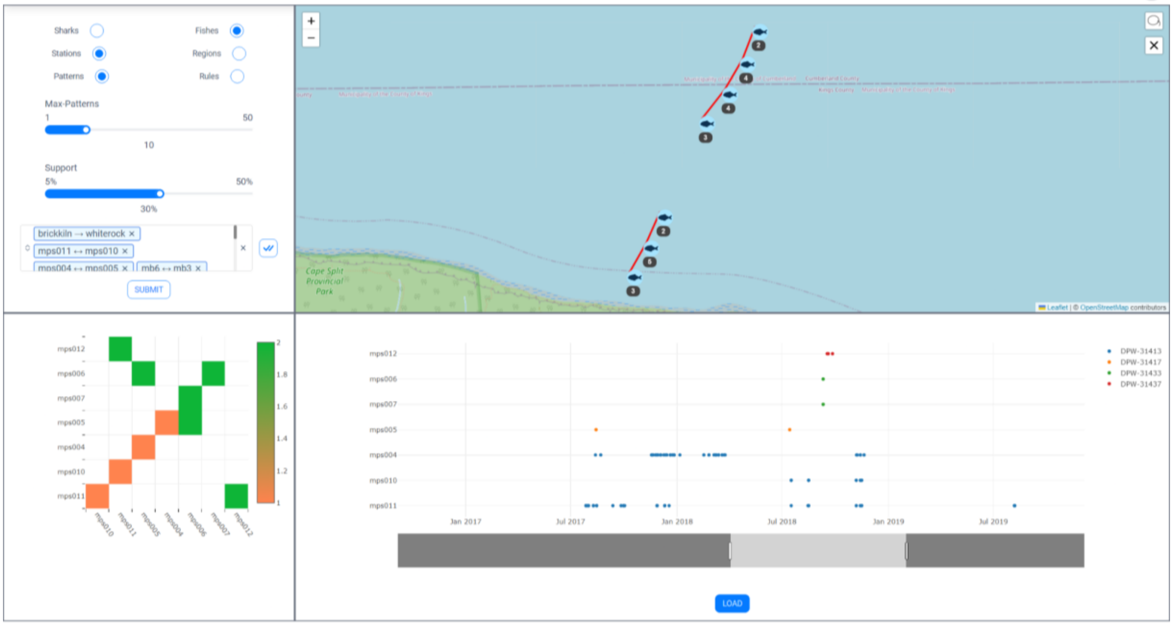


Figure 4.2.2: Stations visited by patterns are plotted across the detection time.

In the second scenario, Alice selects fishes, stations and patterns, with maximum patterns set to 10 and support at 30%, and the plots are updated. Furthermore, she uses a lasso tool to select a few stations in the Geo map to update heatmap to check only the detections happening in that particular location and consults the timeline plot to know the respective times as shown in Fig. 4.2.3. She notices that if the support was increased only the most prominent stations were displayed where the maximum fish activity was happening.



Sharks ○ Fishes ●

Stations ● Regions ○

Patterns ● Rules ○

Max-Patterns: 10 (1–50)

Support: 30% (5%–50%)

brickkiln → whiterock ×

mps011 ↔ mps010 ×

mps004 ↔ mps005 × mb6 ↔ mb3 ×

Figure 4.2.3: Using Lasso to select certain stations and check updated heatmap and timeline plots

CHAPTER 5

Evaluation

In addition to the data analysis methods discussed in section 4, we conducted domain expert interviews after receiving a letter of ethics approval from Dalhousie University ethics board (Appendix G) to obtain feedback on the usability, functionality, performance, customizability, and other related aspects of our oceanic dashboard. We also recorded the instance activity, user videos, in-session comments, and session conditions to gain a more complete understanding of user experiences with the dashboard. The evaluation procedure is described in the rest of the sections.

5.1 Domain Expert Interviews

We interviewed five professional experts referred to as P1 to P5 from the ocean domain, including professors who provided the datasets at Canadian universities, research scientists and directors who are currently working in the ocean sector and have expertise in analyzing marine related datasets. The number of participants was constrained by the availability of experts with the requisite knowledge of statistics, marine biology, and oceanography.

After signing the consent form (Appendix B) and filling the screening questionnaire (Appendix D), the interview sessions began with an approximate 10-minute presentation on the study's overview and a synopsis of Sequential Pattern and Rule Mining and SPMF. We then demonstrated the overview of the dashboard and use cases in section 4 for about 10 minutes. We asked the experts for their feedback on the dashboard, specifically: whether they found the patterns or rules in the use cases insightful, whether they would be interested in using the oceanic dashboard, and whether they had any suggestions to improve it.

Each of the interview sessions took about 60-90 minutes in total. To ensure that the subject matter experts were fully comfortable with the dashboard, we provided them with training. After this training session, we asked them to perform exploratory analysis on the dashboard and complete the Dashboard Interface Rating questionnaire (Appendix E) and Software Usability Questionnaire (Appendix F).

The instance activity and in-session comments are particularly valuable for gaining insight into the users' interactions with the dashboard. By recording their activities and the decisions they make during the analysis process, we can identify areas where the system can be improved, such as providing more contextual information or making certain features more discoverable.

The feedback obtained from the domain expert interviews after recording the activities performed was invaluable in improving the oceanic dashboard. We incorporated their suggestions and recommendations to improve the dashboard's usability, functionality, and performance. The interview results are summarised in the following sections:

5.2 User Demographic

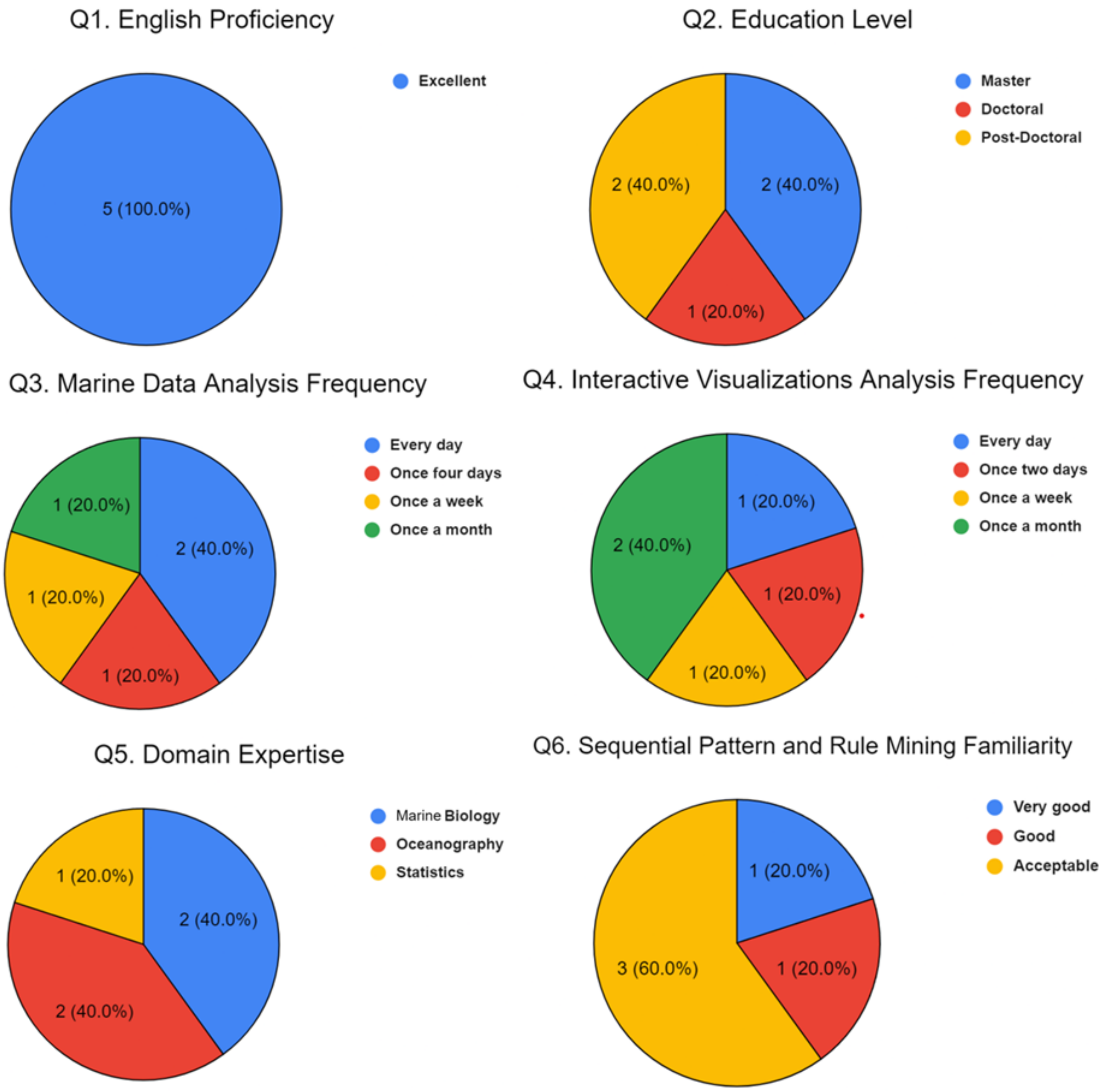


Figure 5.2.1: Pie charts providing an overview of user demographic

Figure 5.2.1 shows the demographics of the experts which analysed the oceanic dashboard. The majority of experts have an excellent understanding of English. They possess a range of educational backgrounds, with a significant portion having completed masters or post-doctoral studies. Regarding the analysis of marine data with Interactive visualizations, the frequency varies among experts. While some analyze marine data every day, others do so on a weekly or monthly basis. Additionally, it is important to note that oceanography primarily focuses on the study of the physical, chemical, and geological aspects of the ocean, while marine biology specifically concentrates on the study of organisms and their interactions within marine ecosystems. The experts who were interviewed had expertise in marine biology, statistics, and oceanography. In terms of familiarity with sequential pattern and rule mining, experts with acceptable knowledge possessed a basic understanding, experts with good knowledge had a more comprehensive understanding and application ability, while experts with very good knowledge exhibited an exceptional level of expertise and handled complex analyses and advanced sequential mining techniques.

5.3 Time to Completion for User Activities

Due to the limited number of available experts, we cannot compute statistically significant analyses of interaction times, we nevertheless noted the times taken during the various tasks. The record of tasks are provided in Figure 5.3.1 and Figure 5.3.2 while exploring the Detections Visualization and Patterns and Rules Visualization respectively.

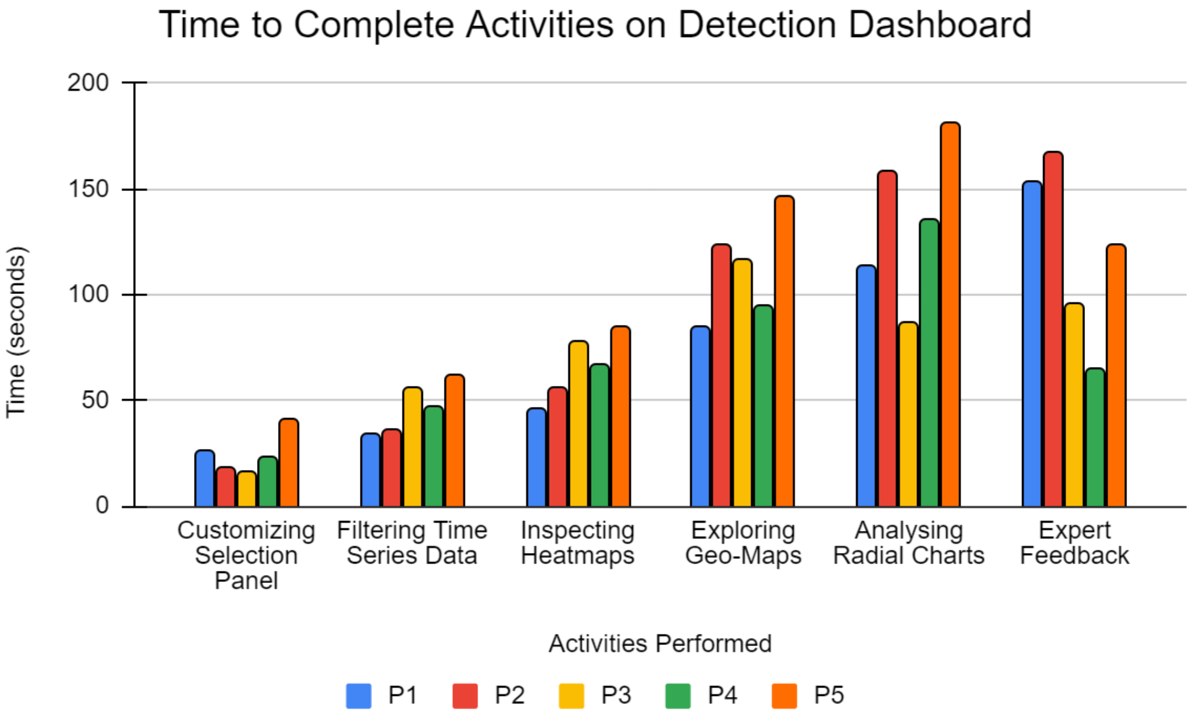


Figure 5.3.1: Time to complete activities on detections visualization

Figure 5.3.1 and Figure 5.3.2 represents the time to completion in seconds for activities [27] by page on the oceanic dashboard in the form of a barchart with each of the five participants referred from P1 to P5. For both the categories, one insight is that the action of exploring geo maps, analyzing radial charts, patterns and rules tend to have higher recorded times, suggesting that users tend to engage in these activities frequently. Customizing the selection panel and filtering the time series data indicates a moderate level of involvement. Lastly, expert feedback for patterns and rules visualization is higher than expert feedback for detections visualization which demonstrates that experts were mostly interested to know about the patterns of sharks and fishes rather than detections.

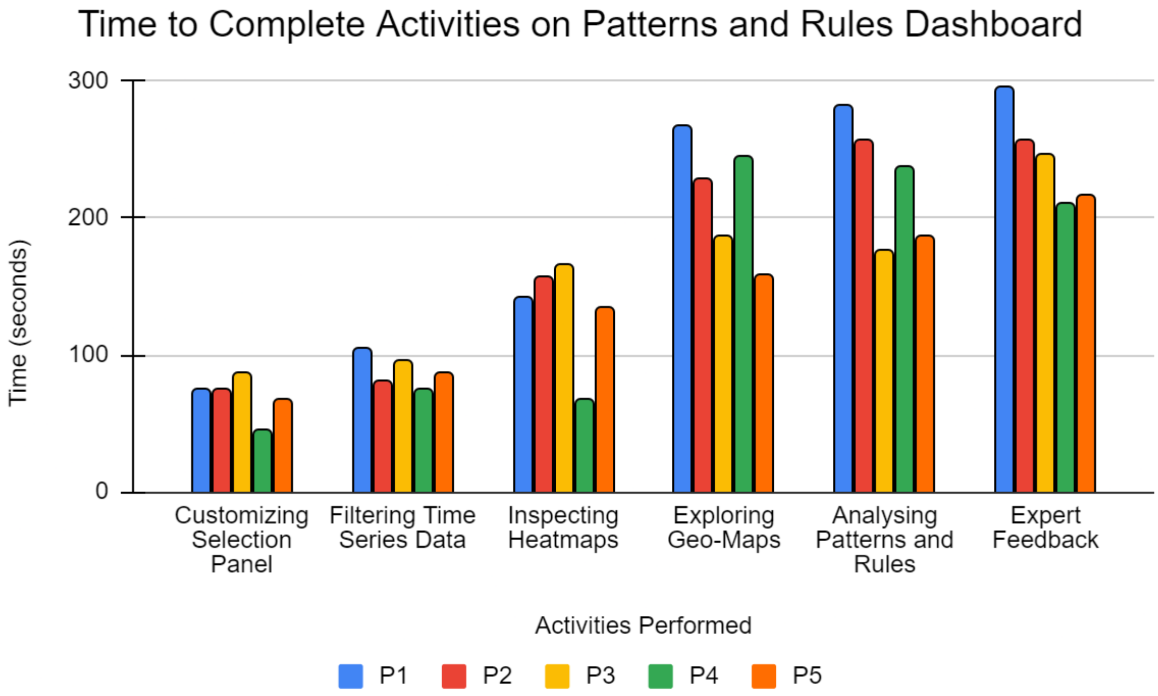


Figure 5.3.2: Time to complete activities on patterns and rules visualization

5.4 Questionnaire Responses

Analysis of the responses to the questionnaires assess user perceptions of the dashboard interface and software usability. The Dashboard Interface Rating Questionnaire in Figure 5.4.1 and Figure 5.4.2 focuses on aspects such as multi-select interface overwhelmingness, dashboard ease of use, understanding of colour and arrow meanings in geo maps, differentiation of stations, understanding of data distributions, and control over data filtering. The Software Usability Questionnaire [28] in Figure 5.4.3 evaluates factors like system usage frequency, complexity, need for technical support and learning curves. Users claimed that while they originally found difficulty in selecting options and analysing the radial charts, they eventually became more adept at doing so. The users responded positively for the lasso select tool to filter out the various patterns. The oceanic dashboard earned compliments for its aesthetic appeal, with users characterising it as a tidy and pleasing interface.

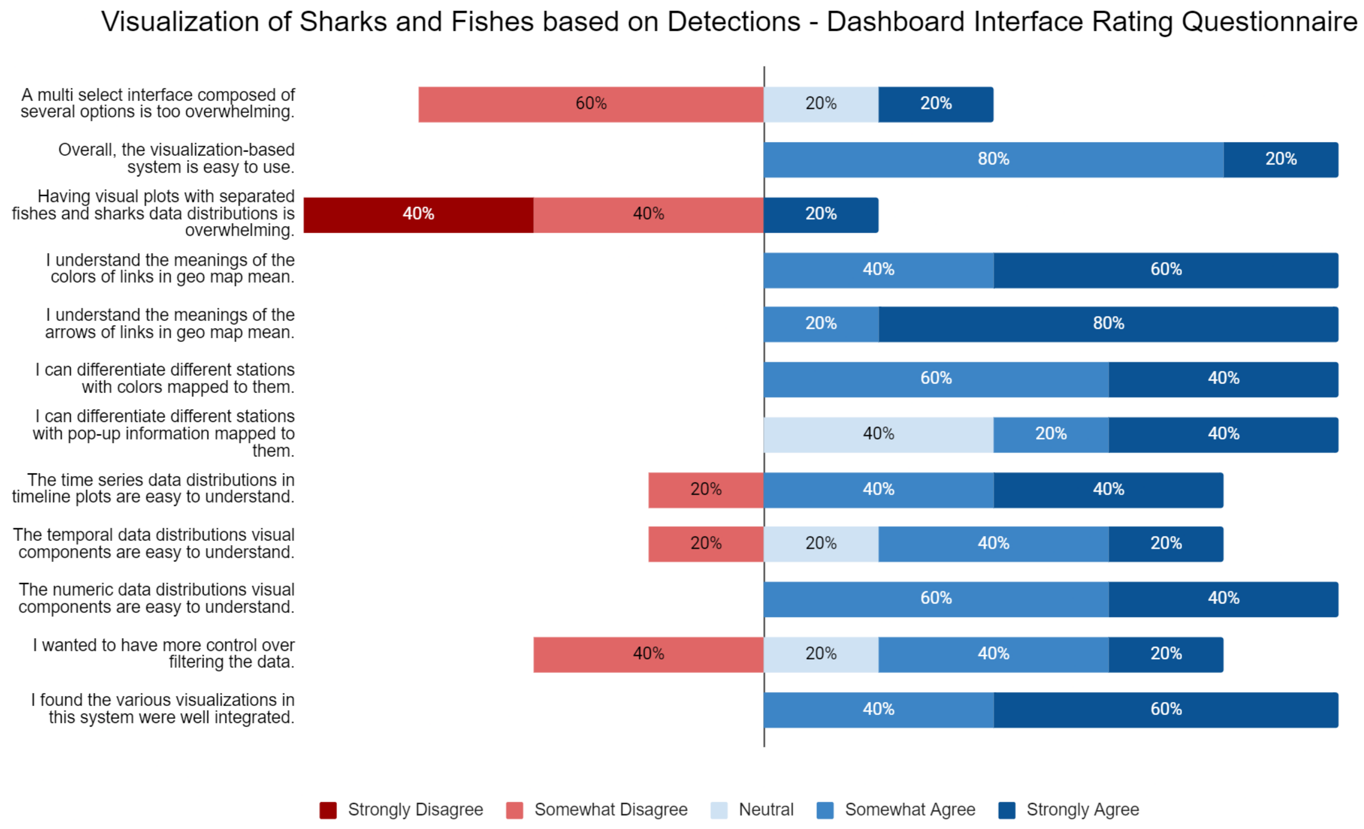


Figure 5.4.1: Detections Visualization - Interface Rating Questionnaire Responses

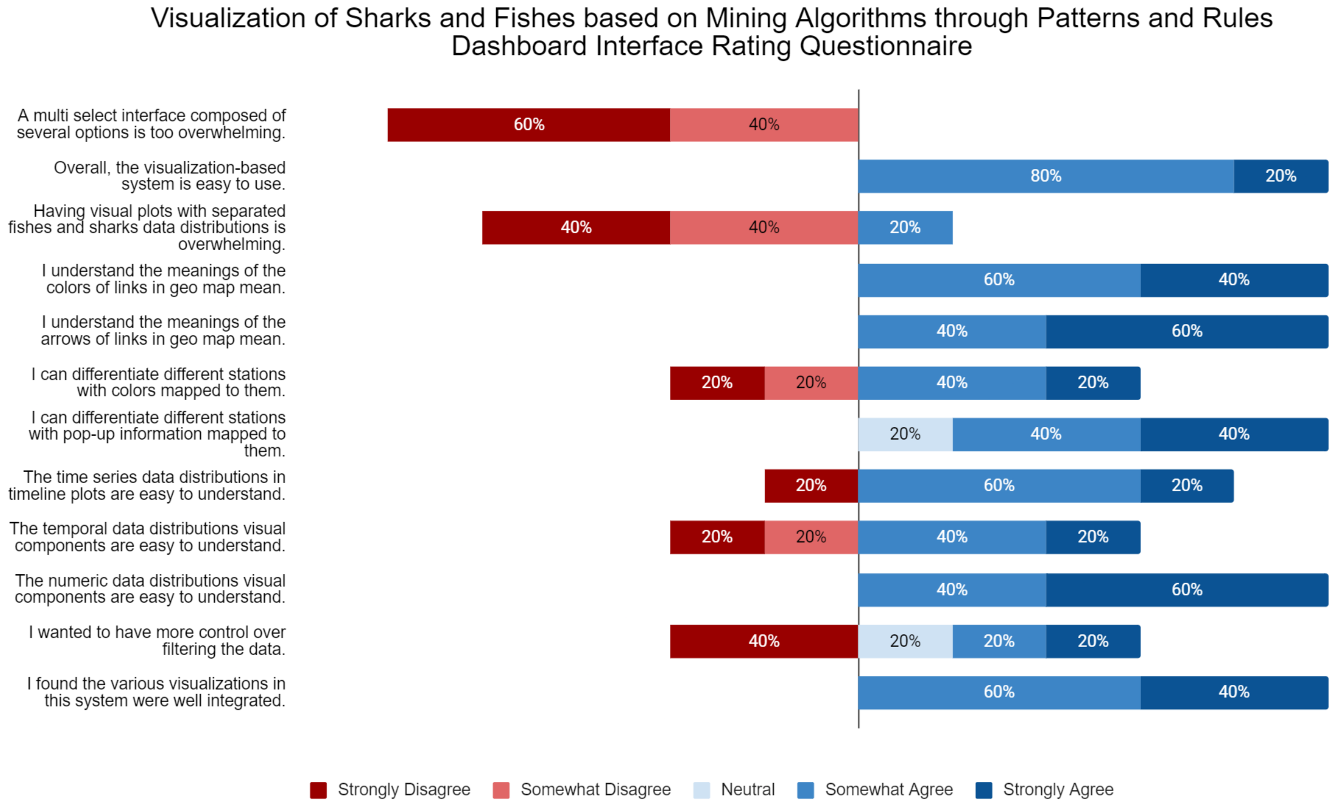


Figure 5.4.2: Patterns and Rules Visualization - Interface Rating Questionnaire Responses

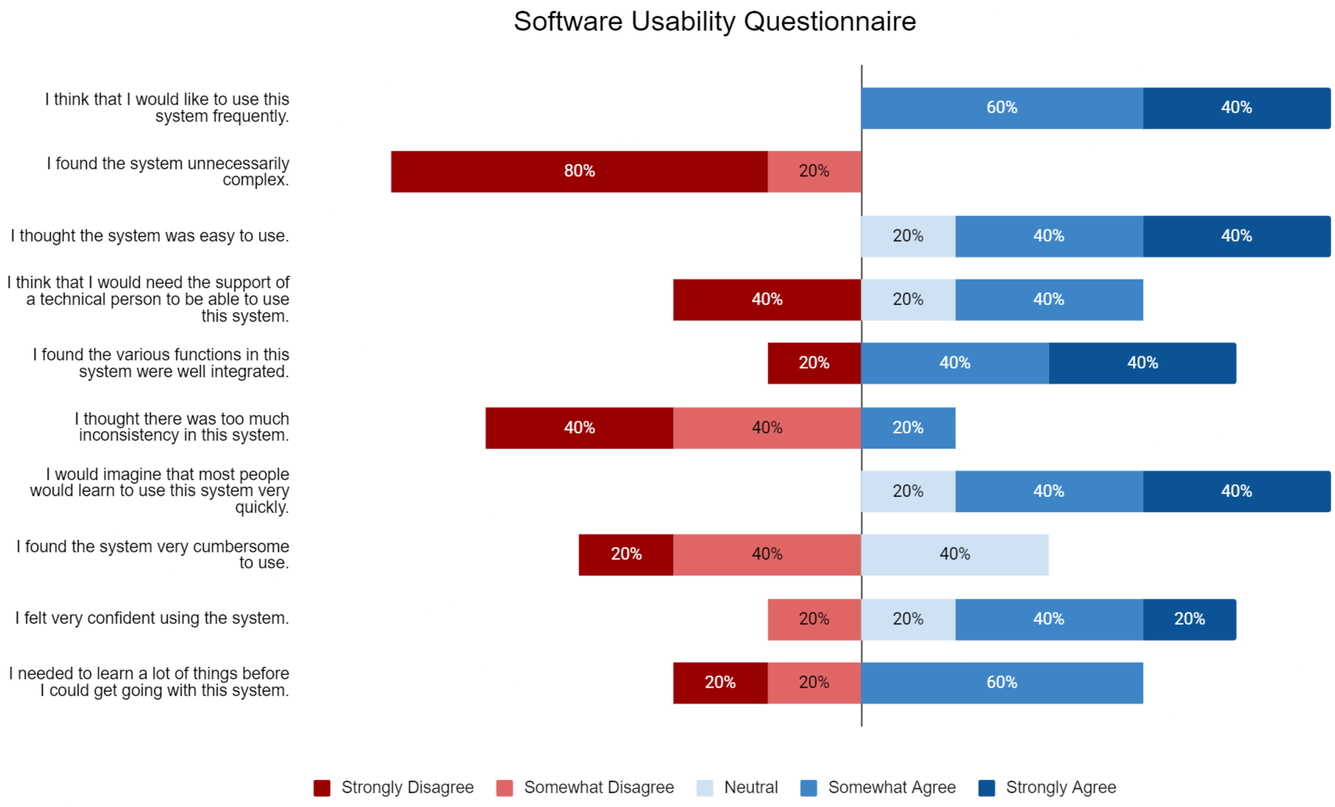


Figure 5.4.3: Software Usability Questionnaire Responses [28]

5.5 Insights gained through expert study

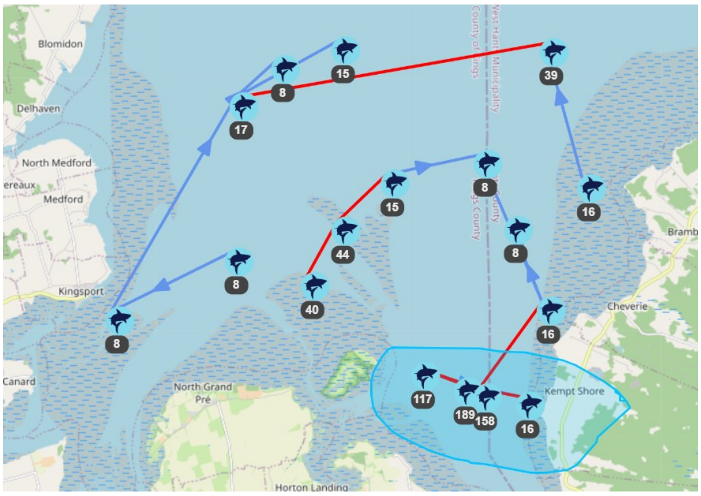


Figure 5.5.1: Sharks patterns count near Kempt Shore

Figure 5.5.1, illustrates an interesting scenario depicted by P4 where the maximum patterns are limited to eight and the support is set at five percent. Upon analyzing the data near the Kempt Shore, it was observed that there was a significantly high frequency of shark patterns formed in this location. P4 stated that this can be attributed to the abundance of diverse marine life and prey species that sharks can consume, which makes it a favourable feeding ground.

Consequently, the sharks tend to frequent this area more often than other locations, leading to the formation of more shark patterns in this region. This finding highlights the importance of considering ecological factors while analyzing patterns and trends in a given area.

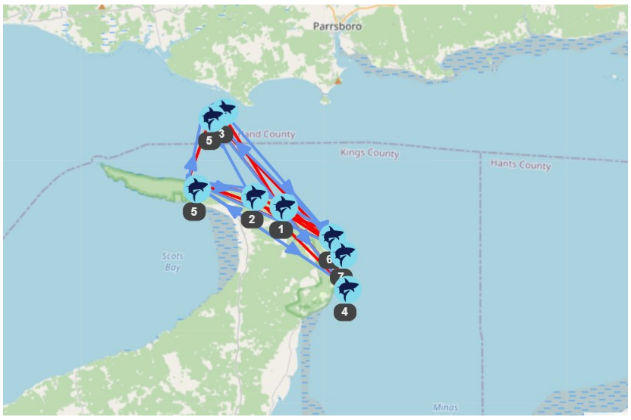


Figure 5.5.2: Sharks rules aligned near Passage

In Figure 5.5.2, an intriguing scenario is depicted by P4 where all the rules align with the passage, serving as the sole entry and exit point for sharks and fishes entering the Bay of Fundy. This alignment of rules with the passage is of significant interest as it highlights the critical role played by this entrance in regulating the movement of aquatic life within this ecosystem. The passage serves as the primary gateway for all marine creatures into the Bay of Fundy, and as a result, all rules and regulations related to the shark movement are channelled through it. This means that the passage region is vulnerable to man-made disruptions, such as installing turbines or dams. Hence, caution must be exercised while planning any structures in the passage region to protect the ecosystem.

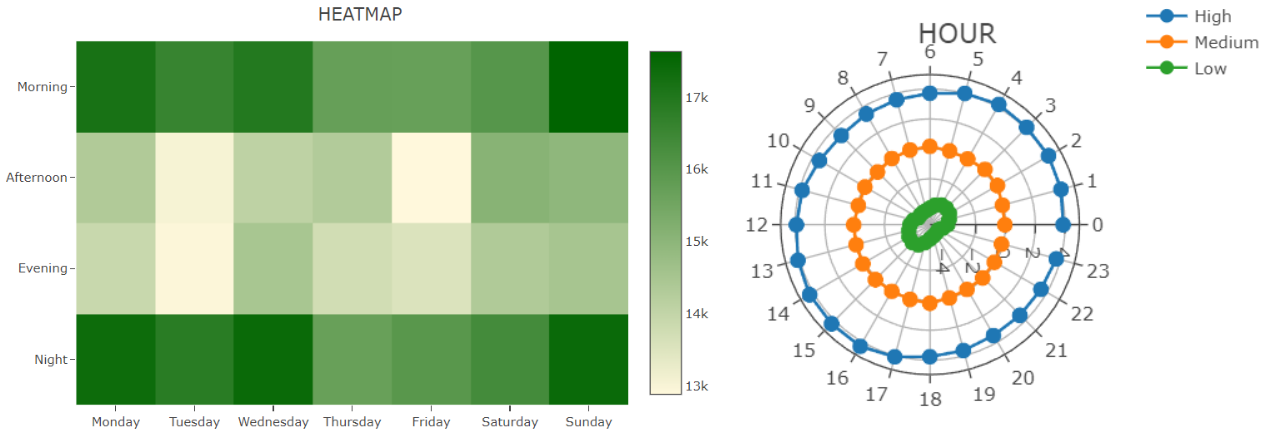


Figure 5.5.3: High levels of fish detections at night and early mornings

Figure 5.5.3, illustrates a notable scenario where the significant number of fish movements detected during nighttime and early mornings suggests that these periods may be critical for their feeding or spawning activities. This emphasizes the importance of preserving the natural cycles and habitats of these aquatic species, particularly during any construction or operation of turbines, to ensure their survival and maintain the health of the ecosystem.

5.6 Data Precision

For the accuracy and efficiency of the analysis and decision-making, the quality of the data shown in the dashboard is essential. The following comment was made in regards to the same, *"The explanation of the data and the system is very important to understanding how to use it." - P3*. As shown earlier, the data's validity and the sources from which it was gathered were already examined by the experts who provided the data. To ensure our dashboard's accuracy and relevance to the work at hand, the data would need to be thoroughly examined and cleansed before being used in the dashboard. Also the final output should be precise and crisp for the user to grasp so filtering techniques were in place also noticed by an expert, *"...the interactive ability of the dashboard and the lasso select are extremely intuitive for filtering purpose (sic)..."* and that *"it's really nice to be able to filter by both space and time." - P3*

5.7 Visualization Effectiveness

P3 while analyzing the geo-plots asserts that, *"The colours and arrows are a very simple way to show direction and patterns between sites",* which suggests that using simple arrows and colours to represent directionality would be an important aspect of the dashboard's effectiveness. Consequently the visualizations such as the geographical visualization and the timeline plot also enhance dashboards's effectiveness which is described by P3 as *"... very informative and a very cool visualization tool for a complex and long-term detection dataset."* In turn, these visualizations can help users to identify patterns and trends in the data, and make sense of large amounts of information.

While checking the average tide heights of a particular location, P1 raised the concern that, *"Some of the terminology could be changed to be less confusing."*, as the location names were in scientific names and the expert was unable to map with the real name of the location. During this case the tooltips would be useful to check for more information. Furthermore tools used to filter patterns such as lasso select, assigned various colour schemes between directionality, renaming stations, and refining the visualizations, could be altered to further improve access to the desired information. This could include *"... changing tide size to tide height or water level... [and] more information and/or a quick description on how the different patterns were classified" - P1* as per an expert suggestion.

5.8 Navigation and Interactivity

The dashboard, enhanced by its data mining capability and usage of various plots such as radial charts with heatmaps, avoids what P1 calls the *"hectic"* activity of *"...working with ... caustic tag detection datasets like these". – P1* Given the visual nature of this dashboard, a user-friendly navigation and interactivity among various plots is key for the dashboard to be easily used. The ability to refine the heatmaps and average tide heights based on different time filtering criteria also enhances the user experience. Notably, P1 identified that when *"detections and movements are related to that specific time period to the tide height," - P1* this helps to reveal patterns and gain in-depth insights.

5.9 Data Analysis and Authenticity of Insights

One of the most important components of the dashboard's efficiency is its capacity to facilitate data analysis and gain insights. While an expert was analyzing the patterns, he revealed that, *"Some... linear patterns are actually could be genuine (sic)" - P4.* Users can acquire insights and improve decision-making by utilizing the dashboard's ability to reveal patterns and rules of sharks and fishes.

The dashboard's capability to assist data analysis was enhanced through the inclusion of a variety of radial charts for analyzing and examining tide averages at various time periods. Furthermore, patterns were mapped on geographical maps to visualize the generated patterns within specific time intervals. These additions provided valuable insights and facilitated in-depth exploration of the data. This may be the motivation behind P1's positive comment when they declared that it is, *"Like I'm looking at this dashboard and my brain is just spinning thinking about what these animals are actually doing. This is a really cool visualization of this data." - P1,* P1's statement about the *"coolness"* of this dashboard demonstrates how its visualization capability improved data exploration and analysis by making it simpler to examine changes. In short, a picture is more enabling than raw data.

5.10 Efficient Data Collection and Pre-Processing

The sensor at the receiver's location records all data by default such as tide heights, wind speed, wind direction, salinity, atmospheric pressure, etc. unless instructed otherwise. This feature enables easy identification of data but may be too large or demanding for processing. Efficient exclusion and better pre-processing of data needs to be done before visualization.

Dashboard's performance and speed are crucial factors since they can influence how well a user can work with the data.

P5 brings up an important detail regarding the dashboard's performance. Since each *"...sensor on the receivers records everything by default unless you specify otherwise," -P5* these sensors may be providing a significant amount of unnecessary data to Dashboard. This feature potentially demands additional pre-processing or may result in a reduction of dashboard's performance. However, the dashboard could be updated to filter information that is unnecessary to the dashboard's users, and information which, *"...is too large or requires too much effort for processing." - P5.* It would also be beneficial for the users to specify to the sensors on the receivers, in advance, that information collection be minimized.

5.11 Customizability

Customizability of the dashboard is beneficial for users. As P3 points out, the ability to use various filters, *"Allowing the user to look at each different pattern or rule (sic) one at a time, is helpful."* This function allows users to configure the tool to their specific needs and preferences in order to better evaluate and analyse the data. Similarly, when P2 states that, *"The way that would get more insights into this kind of oceanic dashboard is being integrated by tweaking the minimum gap, is amazing,"* they identify the value of this ability to tweak the minimum gap.

5.12 Expanding Data Horizons

After the dashboard exploration is done an expert stated that, *"What we don't know is how to scale up the population. So if we were able to tag more sharks, that is, how many are actually out there?" – P5* Clearly, fish population dynamics and behaviour can be better understood by having more fish with tags. Undoubtedly, extending tagging efforts to add additional marine creatures would help to understand the marine ecology in the Bay of Fundy more thoroughly. In an effort to promote additional collection and transparency of the data, P5 suggests that allowing the public companies to append more data would be beneficial. Moreover, showcasing, *"the number of marine animals tracked, [with information] made publicly available to add more data." - P5* could expand more data horizons. Such scaling of the data source would then encourage and enable further marine biology research.

5.13 Overall Comments

Based on expert user feedback, the Oceanic Dashboard Visualization Project addresses a specific need in the field of oceanography and has the potential to be a valuable asset for oceanic experts.

The intriguing findings related to the movement patterns of the sharks and fishes within the Bay of Fundy will conceivably give rise to a generation of ecological papers and further research opportunities. Significantly, the dashboard has the potential to advance our comprehension of ecological dynamics in the ocean as an expert advocates that, *"…there's a real potential to maybe get some ecology papers and stuff out of this as well, because like, what you're finding is really interesting stuff related to how they're moving around in this system." - P4*

More information on the data scalability, sequential rule simplification, and rare patterns are mentioned in later sections.

CHAPTER 6

Discussion

The Oceanic Dashboard Visualization project has been successful in creating a useful and effective tool for analyzing the movement patterns and rules of sharks and fishes in the Bay of Fundy region. The dashboard has been well-received by domain experts, who have found it to be easy to use and reliable with useful features and tools. The geographical visualization and additional modules have proven to be particularly useful in helping experts understand and analyze the data, and the feedback received has been used to make improvements to the dashboard.

One of the main strengths of the dashboard is its ability to handle relatively large amounts of data and perform complex analyses without experiencing noticeable issues or delays. This is important, as oceanic data collection is rapidly increasing and the amount of data available is likely to continue growing in the future. The dashboard's compatibility with different devices and operating systems is also a major advantage, as it allows experts to access and use the tool from anywhere.

During the evaluation, valuable lessons were learned. One key change is the need to modify tide representation, specifically regarding normalisation. While the normalisation of tides was initially implemented, feedback from experts emphasised the importance of incorporating a toggle button. This toggle button would enable users to visualize tides both in their normalized and non-normalized forms. By incorporating these lessons into future iterations of the dashboard, it is anticipated that the tool's effectiveness and usability will continue to improve, providing even greater value to oceanic researchers and analysts.

There are also several areas for improvement and future work for the dashboard. One potential avenue for development is to continue to enhance the data visualization and make it even more effective. This could involve adding new chart types or visualization options, or improving the existing visualizations to make them more effective at communicating the data. Another potential area for improvement is to continue to refine and optimise the performance of the dashboard, particularly as the amount of data grows. Finally, it may be useful to explore ways to make the dashboard more accessible and useful to a wider range of users, including those who may not have a background in oceanic research.

CHAPTER 7

Limitations and Future Work

One potential direction for future work on the Oceanic Dashboard Visualization project is to expand the scope of the dashboard to include data from additional regions or species. This could provide a more comprehensive view of marine ecosystems and allow experts to draw more robust conclusions and make more informed decisions.

Another possibility is to incorporate additional data sources, such as satellite data or biological samples, to provide a more complete picture of the ocean environment. This could allow experts to gain a more nuanced understanding of the interactions and processes occurring within marine ecosystems.

It may be useful to explore ways to integrate the dashboard with other tools and platforms, such as GIS software or online communities, to facilitate collaboration and information sharing among experts. This could allow experts to more easily share their findings and work together to solve complex problems and challenges facing marine ecosystems.

7.1.1 Performance

One potential performance limitation of the Oceanic Dashboard Visualization project is that VMSP and T-Rule Growth algorithm's from the SPMF library used for pattern and rule mining on the pre-processed data takes longer compute times if the selected filter in the visualization is large. This could increase the time complexity and negatively affect the user experience.

During experiments, it has been noticed that the parameter support has a huge impact on execution time of producing results. As the support decreases the number of patterns and rules increases, the pattern and rule mining visualization user experience may not be seamless. To overcome this problem and improve the performance, the plan is to store the pattern or rule mining in a database and retrieve them when needed. This needs an extra space in the system. The other way is to enhance the computing power by using a larger Graphics processing unit for faster results.

This limitation can be observed in the Figure 7.1.1 provided, where the execution time increases as the support decreases and the number of patterns and rules increases.

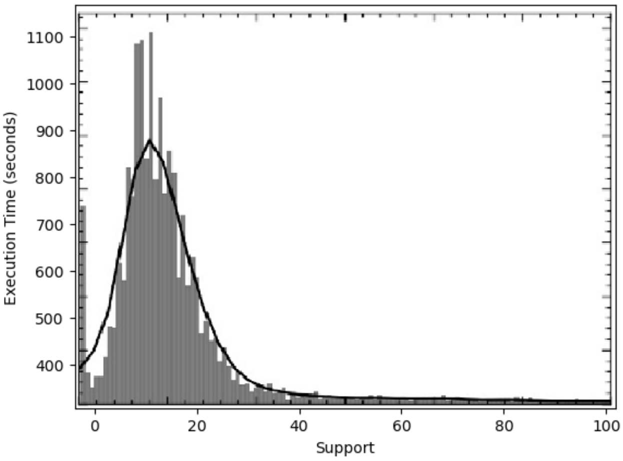


Figure 7.1.1: SPMF’s VMSP Execution time vs. support

7.1.2 Scalability

As the complexity of the data increases, such as with an increase in the number of species or detection stations, the dashboard's ability to effectively analyze and make sense of the data may be impacted. This could limit the dashboard's ability to provide meaningful insights and support data exploration and analysis.

Likewise, as the number of users accessing the dashboard increases, the dashboard's ability to handle the increased demand for data analysis and visualization may be impacted. This could limit the dashboard's ability to provide a smooth and responsive user experience as PostgreSQL can handle limited users at a given point of time.

As the volume and complexity of the data increase, the system resources required to run the dashboard, such as memory, GPU and CPU, may also increase. This could limit the dashboard's ability to handle large amounts of data and perform complex computations in a timely manner.

The scalability of the algorithms used in the dashboard may also be a limitation. As the volume and complexity of the data increase, the algorithms may become less efficient and require more computational resources. This could limit the dashboard's ability to provide accurate and timely analysis of the data.

7.1.3 Accuracy of the Analysis

The quality of the data being collected from the detection stations in the Bay of Fundy region can impact the accuracy of the analysis performed by the dashboard. For example, if the data is not properly cleaned or validated, or if there is a high level of noise or missing data, this can lead to inaccurate or unreliable results.

The tuning parameters applied to the models while using the dashboard can also impact the accuracy of the analysis. For example, if the models assume a certain distribution of the data that does not match the actual data, this can also lead to inaccurate or unreliable results.

Human error can also be a factor that can impact the accuracy of the analysis. For example, if the users of the dashboard are not properly trained or if they make mistakes when interacting with the tool, this can lead to inaccurate or unreliable results. So the dashboard is recommended for the marine experts and scientists for exploration at this time.

The limited scope of data can also be a limitation that may impact the accuracy of the analysis. For example, if the data is only collected from a few detection stations or for a limited time frame, the analysis may not be fully representative of the whole region or the whole period. For example, the data was not collected in the winter seasons from the stations in Bay of Fundy of the sharks and fishes as there were holiday periods for the staff.

7.1.4 Rare Patterns

The dashboard may have limitations in identifying rare patterns in the data, as these patterns occur because the data mining and analysis algorithms used in the dashboard may have a support or confidence threshold, which filters out patterns that do not meet a certain level of frequency or reliability. This means that rare but interesting patterns may be overlooked, and valuable insights may be missed.

One workaround for this limitation could be to use algorithms specifically designed for finding rare patterns, such as rare itemset mining algorithms or rare event prediction algorithms [26]. These algorithms can be more effective at identifying and analyzing rare patterns, and can help to uncover valuable insights that may be missed by traditional data mining and analysis techniques.

Another approach to overcome this limitation is by increasing the support threshold or by including all the patterns and then providing filtering options to the user to filter out the less interesting patterns.

7.1.5 Sequential Rule and Pattern Simplification

When mining rules with only one consequent, the representation of the data may not be as comprehensive. For example, if the rules with more than one consequent would have provided a more complete representation of the data, then the limitation of having only one consequent will make the analysis less representative.

The limitation of using the pattern mining algorithm with a fixed maximum length from one to two hundred, and no gaps in between, can have a few implications on the insights that can be generated from the data. By limiting the maximum length of the patterns to one to two hundred, the algorithm may not be able to identify long-term patterns that occur over a longer time frame. This could limit the ability of the dashboard to provide insights into the long-term trends and movements of the sharks and fishes.

Not allowing gaps in between the patterns, the algorithm may miss patterns that have breaks or interruptions in the sequence. This can limit the ability of the dashboard to identify patterns that are not continuous, which may be important for understanding the movement patterns of the sharks and fishes.

By limiting the maximum length of the patterns to two to hundred, the algorithm may not be able to identify complex patterns that involve multiple events occurring in sequence. This could limit the ability of the dashboard to provide insights into the more complex movement patterns of the sharks and fishes.

7.1.6 Conclusion

In conclusion, the Oceanic Dashboard Visualization project represents a significant step forward in the field of visual analytics of oceanic multivariate species time series data. By providing an effective tool for analyzing the movement patterns and rules of sharks and fishes in the Bay of Fundy region, this dashboard has the potential to help marine biologists, oceanic scientists to make informed decisions and minimize risks to marine ecosystems. The feedback received from domain experts has been very positive, and the dashboard has been found to be reliable, user-friendly, and packed with useful features and tools.

Moving forward, there is still room for improvement in terms of scalability, sequential rule simplification, and rare pattern detection. However, these issues can be addressed through ongoing research and development.

Overall, the Oceanic Dashboard Visualization project represents a significant step forward in the realm of data visualization and analysis for marine data. By harnessing the power of new frontend technologies and data mining techniques, this dashboard provides a comprehensive view of the movements of sharks and fishes in the Bay of Fundy region. Its ease of use, reliability, and ability to handle large amounts of data make it an invaluable tool for oceanic experts, and its potential to improve our understanding of marine ecosystems is notable. As such, this project represents a hopeful step towards a more sustainable and informed approach to oceanic research and management.

Bibliography

[1] International business machines (IBM) corporation. IBM sequence rules view. https://www.ibm.com/docs/bg/db2/9.7?topic=visualizer-sequence-rules-view. 2021.

[2] International business machines (IBM) corporation. the JAVA - based mining visualizer. https://www.ibm.com/docs/en/db2/10.5?topic=types-java-based-miningvisualizer. 2021.

[3] Danielle Albers, Colin Dewey, and Michael Gleicher. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. IEEE Transactions on Visualization and Computer Graphics, 17(12):2392 – 2401, Dec 2011.

[4] Amira Abdelwahab and Nesma Youssef. Performance evaluation of sequential rule mining algorithms. Applied Sciences, 12(10), 2022.

[5] Alain F. Zuur, Elena N. Ieno, and Graham M. Smith. Analyzing ecological data, 265-315 2007.

[6] Antoine Clarinval and Bruno Dumas. Intra-city traffic data visualization: A systematic literature review. IEEE Transactions on Intelligent Transportation Systems, pages 1–18, 2021.

[7] Bernard Fichet, Domenico Piccolo, Rosanna Verde, and Maurizio Vichi, editors. Classification and Multivariate Analysis for Complex Data Structures. Springer Berlin Heidelberg, 2011.

[8] Philippe Fournier-Viger, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Cheng-Wei Wu, and Vincent S. Tseng. Spmf: A JAVA open-source pattern mining library. Journal of Machine Learning Research, 15(104):3569–3573, 2014.

[9] Aline Menin, Ricardo Cava, Carla Maria Dal Sasso Freitas, Olivier Corby, and Marco Winckler. Towards a visual approach for representing analytical provenance in exploration processes. In 2021 25th International Conference Information Visualisation (IV), Melbourne/Virtual, Australia, 4-5, July 2021.

[10] M. Rezaei and P. Franti. Real-time clustering of large geo-referenced data for visualizing on map. Advances in Electrical and Computer Engineering, 18(4):63–74, 2018.

[11] Martin Krzywinski, Jacqueline Schein, Inan¸c Birol, Joseph Connors, Randy Gas- coyne, Doug Horsman, Steven J. Jones, and Marco A. Marra. Circos: An information aesthetic for comparative genomics. Genome Research, 19(9):1639–1645, June 2009.

[12] Mattias Persson. A survey of methods for visualizing spatio-temporal data. In Visualizing Spatio-Temporal Data, pages 22–65, 2020.

[13] Lya Boyandin, Enrico Bertini, Peter Bak, and Denis Lalanne. Flowstrates: An approach for visual exploration of temporal origin-destination data. Computer Graphics Forum, 30(3):971–980, June 2011.

[14] Christopher G. Mull, Nathan Pacoureau, Sebastian A. Pardo, Luz Saldana, Ruiz, Emiliano Garćıa-Rodŕıguez, Brittany Finucci, Max Haack, Alastair Harry, Aaron B. Judah, Wade VanderWright, Jamie S. Yin, Holly K. Kindsvater, and Nicholas K. Dulvy. Sharkipedia: a curated open access database of shark and ray life history traits and abundance time-series. Scientific Data, 9(1), September 2022.

[15] David R. Brillinger. Time Series: Data Analysis and Theory. Society for Industrial and Applied Mathematics, USA, 2001.

[16] Philippe Fournier-Viger, Cheng-Wei Wu, Vincent S. Tseng, and Roger Nkambou. Mining sequential rules common to several sequences with the window size constraint. In Advances in Artificial Intelligence, pages 299–304. Springer Berlin Heidelberg, 2012.

[17] Philippe Fournier-Viger, Cheng-Wei Wu, Antonio Gomariz, and Vincent S. Tseng. VMSP: Efficient vertical mining of maximal sequential patterns. In Advances in Artificial Intelligence, pages 83–94. Springer International Publishing, 2014.

[18] Vanessa Leung. Data visualization of uber rides with tableau. Link: https://towardsdatascience.com/data-visualization-of-uber-rides-with-tableau-67988f61f712?gi=cc7c11c00459/, 2022.

[19] Thomas Kapler and William Wright. GeoTime information visualization. Information Visualization, 4(2):136–146, June 2005.

[20] Colin Ware. Trackplot, https://ccom.unh.edu/vislab/projects/trackplot/, 2013.

[21] Roberta Siciliano, Antonio D'Ambrosio, Massimo Aria, and Sonia Amodio. Analysis of web visit histories, part II: Predicting navigation by nested STUMP regression trees. Journal of Classification, 34(3):473–493, October 2017.

[22] Aashara Shrestha, Dimitrios Zikos, and Leonidas Fegaras. An annotated association mining approach for extracting and visualizing interesting clinical events. International Journal of Medical Informatics, 148:104366, April 2021.

[23] Pak Chung Wong, W. Cowley, H. Foote, E. Jurrus, and J. Thomas. Visualizing sequential patterns for text mining. In IEEE Symposium on Information Visualization. INFOVIS 2000. Proceedings. IEEE Comput. Soc, 105-114, Utah, 2000.

[24] Rosana Veroneze. Enumerating all maximal biclusters in numerical datasets. PhD thesis, 44, June 2016.

[25] Sandra G Hart. Nasa-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting, volume 50, pages 904-908. Sage publications Sage CA: Los Angeles, CA, 2006.

[26] Sadeq Darrab, David Broneske, and Gunter Saake. RPP Algorithm: A Method for Discovering Interesting Rare Itemsets. In the International Conference on Data Mining and Big Data, pages 14-25. Springer, 2020.

[27] Lea Mets Kiritsis, Improving the onboarding experience A qualitative analysis of onboarding slides and tooltips. Kth Royal Institute of Technology, Stockholm, page 37-39, Sweden 2022.

[28] Brooke, John. SUS-A quick and dirty usability scale. Usability Evaluation in Industry, 189-194, 1996.

Appendix A - Email Recruitment Notice

Ocean data collection and access are undergoing a revolution with new applications and technologies applied to manage and understand our ocean. In particular the advance in development of underwater sensors for data collection has spawned a plethora of new ways to monitor and understand marine ecosystems. With an unprecedented ability to gather the data of the movements of sharks and fishes in local regions and analyze the environment, these new technologies open possibilities for advancements in research and decision-making.

In our Oceanic Visualization project, we set out to track and display aquatic movements of sharks and fishes in the Bay of Fundy region, after extrapolating sequential patterns and rules using data mining techniques. To validate our proposal, we are recruiting domain experts and research scientists from oceanography, marine biologists, computer science and statistics departments and companies to take part in our research study to evaluate if the proposed functionalities of our dashboard allow better interpretation of our marine datasets.

Participants should be those who understand the English language and have general computer proficiency in order to interact with the software dashboard designed.

After giving consent to participate, participants will be asked to watch a training video about how to interact with the dashboard. The study will be conducted online, and it will take about 60-90 minutes in total to complete. Prospective participants will interact with a visualization dashboard to perform several analytical tasks and perform exploratory analysis.

After using our dashboard, you will fill out and submit an evaluation questionnaire on your experience with the dashboard. This study is an opportunity to evaluate the user experience of a novel visualization system and learn more about those different areas out there such as sequential patterns and rules, which might be of interest to those in the field of marine biology.

If you are interested in participating or have any questions, please contact the lead researcher of this study: Vegu Shree Rama Kamal Kumar (kamal@dal.ca).

Appendix B - Consent Form

CONSENT FORM

Project title: Visual Analysis to Gain Insights of Oceanic Multivariate Species Time-Series Data

The lead researcher is Vegu Shree Rama Kamal Kumar who is a Computer Science master student at Dalhousie university, Halifax, Canada. His email is kamal@dal.ca, and his phone number is 902-448-5070.

This research is under the supervision of:

- Dr. Stephen Brooks (sbrooks@cs.dal.ca), professor at Faculty of Computer Science at Dalhousie University, Canada

We invite you to take part in a research study being conducted by Vegu Shree Rama Kamal Kumar, who is a Computer Science master student at Dalhousie University, Halifax, Canada. Choosing whether to take part in this research is entirely your choice, and participants are free to withdraw at any time without repercussions (you can withdraw by informing the lead researcher). The information below tells you about what is involved in the study, what you will be asked to do, and about any benefit, risk, inconvenience, or discomfort that you might experience. You should discuss any questions you have about this study with Vegu Shree Rama Kamal Kumar (kamal@dal.ca). Please ask as many questions as you like.

Purpose and Outline of the Research Study

Ocean data collection and access are undergoing a revolution with new applications and technologies applied to manage and understand our ocean. In particular the advance in development of underwater sensors for data collection has spawned a plethora of new ways to monitor and understand marine ecosystems. With an unprecedented ability to gather the data of the movements of sharks and fishes in local regions and analyze the environment, these new technologies open possibilities for advancements in research and decision-making.

In our Oceanic Dashboard Visualization project, we set out to track and display aquatic movements of sharks and fishes in the Bay of Fundy region, after extrapolating sequential patterns and rules using data mining techniques. To validate our proposal, we are recruiting domain experts and research scientists from oceanography, marine biologists, computer science and statistics to take part in our research study to evaluate if the proposed functionalities of our dashboard allow better interpretation of our marine datasets.

Who Can Take Part in the Research Study?

You may participate in this study if you are a domain expert or research scientist from oceanography, marine biology, computer science and statistics. Participants must be people who understand the English language and must have general computer proficiency in order to interact with the dashboard designed.

What You Will Be Asked to Do?

If you choose to participate in this research, you will be asked to perform pre-set operations and exploratory analysis on our oceanic visualization system and anonymously answer questions regarding its usability, which are listed below. The study should take approximately 60-90 minutes.

- You will sign the consent form.
- You will complete a screening questionnaire.
- You will be given a tutorial on how to use the software.
- You will be walked through a use-case session with the software.
- You will perform exploratory analysis with the software.
- You will submit the post-study questionnaires and comments.

If you wish to participate in the study, then please:

- Sign this consent form, and
- Give or mail it to the principal investigator.

The principal investigator will keep the original copy of this consent form.

Possible Benefits, Risks, and Discomforts

Your participation will be appreciated, and we expect that it will help us to learn how our proposed visualization helps experts interpret patterns, rules, insights from a vast volume of marine dataset. The risks associated with this study are minimal; there are no known risks for participating in this research beyond being bored, feeling tired, or frustrated in the case the tasks are new for you. No additional risks are predicted than the risk of use of computers in daily life. Your name will not be connected to the data collected from you.

Compensation / Reimbursement

Participants are not being compensated for being part of this study.

How your information will be protected:

Your participation in this research will be known only to the lead researcher. The information that you provide to us will be kept confidential. Only the research team at Dalhousie University will have access to this information. The people who work with us must keep all research information confidential. We will use a participant number (not your name) in our written and computer records so that the research information we have about you contains no names. During the study, all electronic records will be kept secure in an encrypted file on the researcher’s password-protected computer.

We will describe and share our findings in diverse ways: thesis, presentations, and journal articles. We will only report group results and not individual results. This means that you will not be identified in any way in our reports.

Anonymized data collected during the user study is going to be stored in the lead researcher's OneDrive account. This data is not going to be shared since it is specific for this project alone. With this, we expect future researchers to enhance our proposed system and replicate our results and contribute within the academic field.

Where the data will be stored:

The participant will be allowed to use the visualization system and explore patterns and rules of sharks and fishes. After exploring all the functionalities of the system, the user will be asked a set of questions about the usability and interface of the dashboard. Here the user must answer those questions and the responses will be collected. These documents were then transferred to the lead researcher's OneDrive account within a week. The data will be kept for a maximum of five years, for the purpose of revisions to a journal or conference submission, after which it will be permanently deleted. Only Vegu Shree Rama Kamal Kumar and Prof. Stephen Brooks will have access to the study results.

Here only the responses of the users are alone stored. Any information about the user like email address, name and other personal details are not stored. This is because we are not going to send the results of the user study to individual users using their email address.

The responses of the users will be kept in the researchers OneDrive account for a maximum of 5 years for the future research. The data will not be maintained indefinitely.

The researchers will use their Dalhousie University credentials for the Microsoft Teams meeting, which will ensure that the Teams meeting recordings are securely stored in Canada. During the live Teams meeting, audio and video content is routed through the United States, and therefore may be subject to monitoring without notice, under the provisions of the US Patriot Act while the meeting is in progress. After the meeting is complete, meeting recordings made by Dalhousie are stored in Canada and are inaccessible to US authorities.

Questions:

Please feel free to ask the lead researcher about anything to do with the study before (or after) you give consent to participate. You should discuss any questions you have about this study with Vegu Shree Rama Kamal Kumar or Prof. Stephen Brooks. My contact information is kamal@dal.ca.

If you have any ethical concerns about your participation in this research, you may contact Research Ethics, Dalhousie University at (902) 494-3423, or email ethics@dal.ca (and reference REB file # 2023-6482).”

If you agree to complete and participate in the study, please reply to this email with “I accept the consent agreement”, and the lead researcher will get in touch with future steps.

Note:

1. **No direct benefits are provided to the participants.**
2. **The users can withdraw from the study at any time by informing the lead researcher.**
3. **If you complete your study and you change your mind later, I will not be able to remove the information you provided as I will not know which response is yours.**

Appendix C - Overall Flowchart for Questionnaires

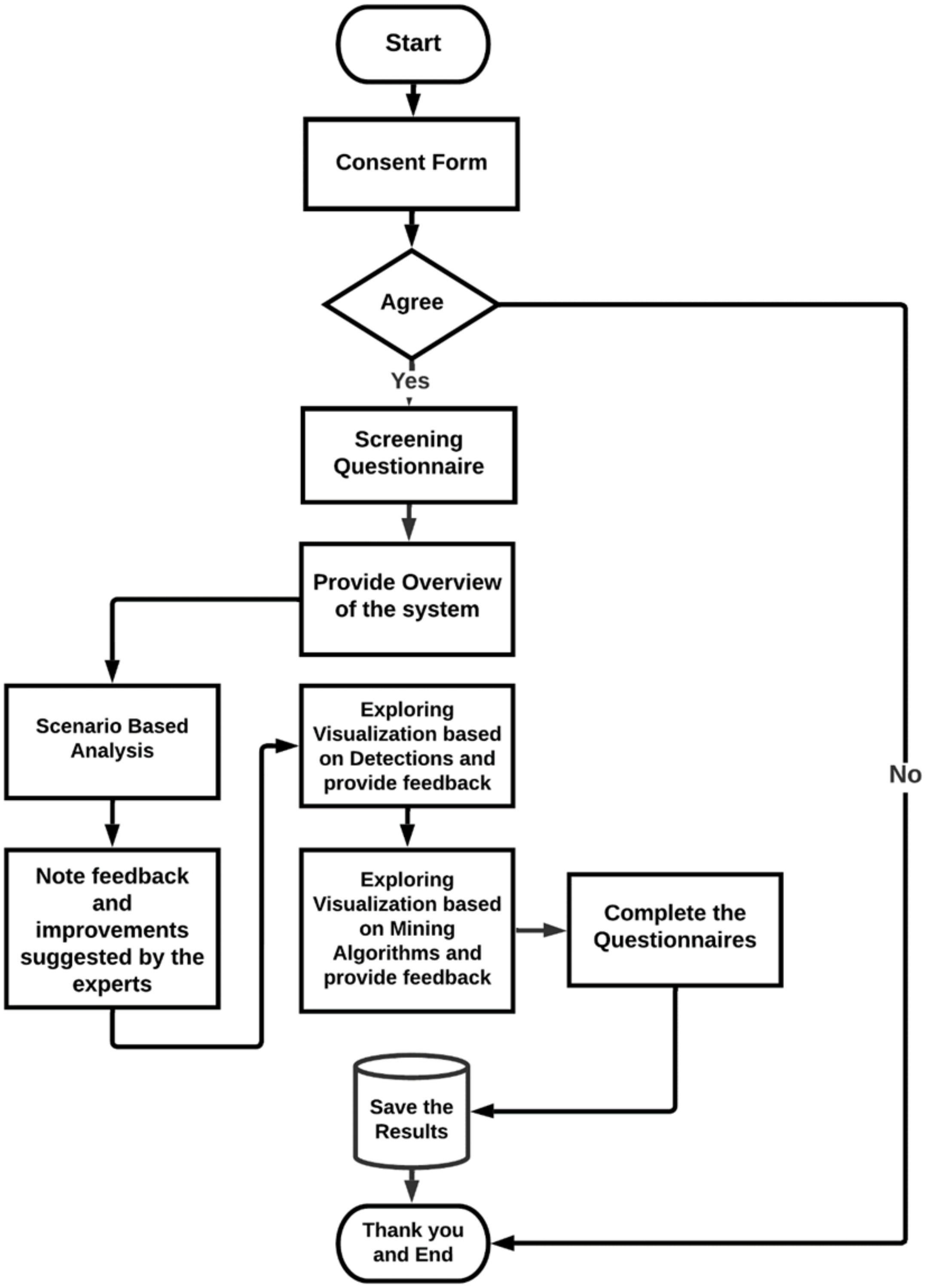

Start
Consent Form
Agree
Yes
No
Screening Questionnaire
Provide Overview of the system
Scenario Based Analysis
Note feedback and improvements suggested by the experts
Exploring Visualization based on Detections and provide feedback
Exploring Visualization based on Mining Algorithms and provide feedback
Complete the Questionnaires
Save the Results
Thank you and End

Figure 3: The study steps to be conducted for our oceanic system to be explored and evaluated by experts.

Appendix D - Screening Questionnaire

Identification number: _________________

1. At what level do you think your understanding of written English is?
    - Excellent
    - Very good
    - Good
    - Acceptable
    - Bad
    - Very bad
2. What is the highest level of education you have completed?
    - Post-Doctoral
    - Doctoral
    - Master
    - College or university
    - High school or equivalent
    - Little or no formal education
3. How often do you analyze marine data?
    - Every day
    - Once two days
    - Once four days
    - Once a week
    - Once a month
    - Once a year
    - Never
4. How often do you use interactive visualizations?
    - Every day
    - Once two days
    - Once four days
    - Once a week
    - Once a month
    - Once a year
    - Never
5. What is your domain expertise? (Survey question for internal use)
    - Computer Science
    - Statistics
    - Marine Biology
    - Oceanography
    - Other______________
6. How much are you familiar with sequential pattern and rule mining?
    - Excellent
    - Very good
    - Good
    - Acceptable
    - Bad
    - Very bad

Appendix E – Dashboard Interface Rating Questionnaire

Identification number: ________________

Please respond to the following statements using the given scale (circle response):

1.	A multi select interface composed of several options is too overwhelming.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
2.	A multi select interface composed of several options is too overwhelming.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
3.	Having visual plots with separated fishes and sharks data distributions is overwhelming.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
4.	I understand the meanings of the colors of links in geo map mean.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
5.	I understand the meanings of the arrows of links in geo map mean.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
6.	I can differentiate different stations with colors mapped to them.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
7.	I can differentiate different stations with pop-up information mapped to them.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
8.	The time series data distributions in timeline plots are easy to understand.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
9.	The temporal data distributions visual components are easy to understand.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
10.	The numeric data distributions visual components are easy to understand.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
11.	I wanted to have more control over filtering the data.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
12.	I found the various visualizations in this system were well integrated.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*

Appendix F – Software Usability Questionnaire

Identification number: ______________

Please respond to the following statements using the given scale (circle response):

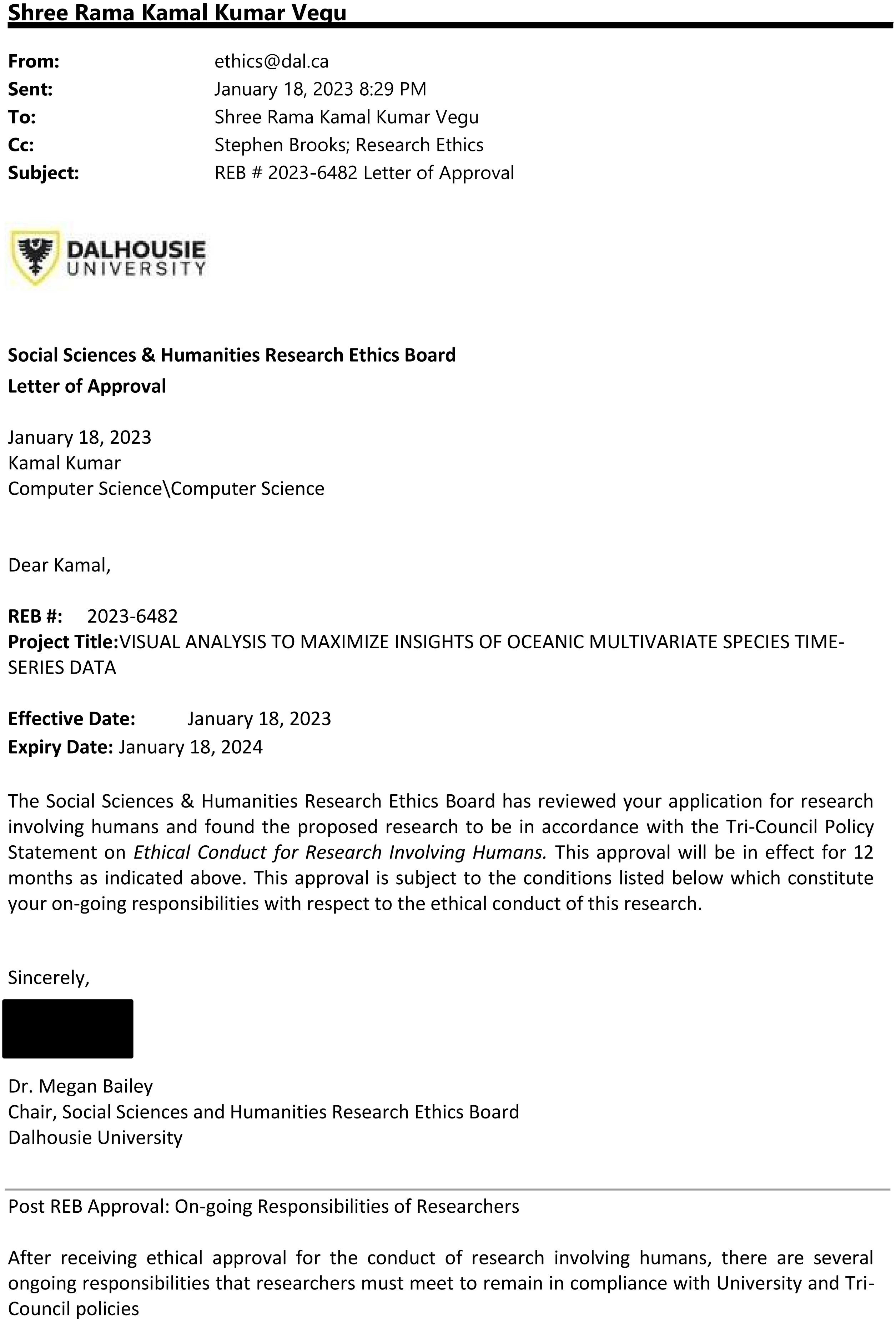
1.	I think that I would like to use this system frequently.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
2.	I found the system unnecessarily complex.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
3.	I thought the system was easy to use.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
4.	I think that I would need the support of a technical person to be able to use this system.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
5.	I found the various functions in this system were well integrated.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
6.	I thought there was too much inconsistency in this system.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
7.	I would imagine that most people would learn to use this system very quickly.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
8.	I found the system very cumbersome to use.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
9.	I felt very confident using the system.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*
10.	I needed to learn a lot of things before I could get going with this system.	***1***	***2***	***3***	***4***	***5***
		Strongly Disagree	*Somewhat Disagree*	*Neutral*	*Somewhat Agree*	*Strongly Agree*

[1] Brooke, John. "SUS-A quick and dirty usability scale." Usability evaluation in industry 189, no. 194 (1996): 4-7.

1. Please give us more comments about the system:

2. Is there any functionality that you expect to be included but was not available?

Appendix G – REB Ethics Letter of Approval



Shree Rama Kamal Kumar Vegu

From:	ethics@dal.ca
Sent:	January 18, 2023 8:29 PM
To:	Shree Rama Kamal Kumar Vegu
Cc:	Stephen Brooks; Research Ethics
Subject:	REB # 2023-6482 Letter of Approval

Social Sciences & Humanities Research Ethics Board
Letter of Approval

January 18, 2023
Kamal Kumar
Computer Science\Computer Science

Dear Kamal,

REB #: 2023-6482
Project Title:VISUAL ANALYSIS TO MAXIMIZE INSIGHTS OF OCEANIC MULTIVARIATE SPECIES TIME-SERIES DATA

Effective Date: January 18, 2023
Expiry Date: January 18, 2024

The Social Sciences & Humanities Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans.* This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Sincerely,

[redacted]

Dr. Megan Bailey
Chair, Social Sciences and Humanities Research Ethics Board
Dalhousie University

Post REB Approval: On-going Responsibilities of Researchers

After receiving ethical approval for the conduct of research involving humans, there are several ongoing responsibilities that researchers must meet to remain in compliance with University and Tri-Council policies

1. Additional Research Ethics approval

Prior to conducting any research, researchers must ensure that all required research ethics approvals are secured (in addition to Dalhousie approval). This includes, but is not limited to, securing appropriate research ethics approvals from: other institutions with whom the PI is affiliated; the institutions of research team members; the institution at which participants may be recruited or from which data may be collected; organizations or groups (e.g. school boards, Indigenous communities, correctional services, long-term care facilities, service agencies and community groups) and from any other responsible review body or bodies at the research site.

2. Reporting adverse events

Any significant adverse events experienced by research participants must be reported **in writing** to Research Ethics **within 24 hours** of their occurrence. Examples of what might be considered "significant" include: a negative physical reaction by a participant (e.g. fainting, nausea, unexpected pain, allergic reaction), an emotional breakdown of a participant during an interview, report by a participant of some sort of negative repercussion from their participation (e.g. reaction of spouse or employer) or complaint by a participant with respect to their participation, report of neglect or abuse of a child or adult in need of protection, or a privacy breach. The above list is indicative but not all-inclusive. The written report must include details of the situation and actions taken (or proposed) by the researcher in response to the incident.

3. Seeking approval for changes to research

Prior to implementing any changes to your research plan, whether to the risk assessment, methods, analysis, study instruments or recruitment/consent material, researchers must submit them to the Research Ethics Board for review and approval. This is done by completing the amendment request process (described on the website) and submitting an updated ethics submission that includes and explains the proposed changes. Please note that reviews are not conducted in August.

4. Continuing ethical review - annual reports

Research involving humans is subject to continuing REB review and oversight. REB approvals are valid for up to 12 months at a time (per the Tri-Council Policy Statement (TCPS) article 6.14). Prior to the REB approval expiry date, researchers may apply to extend REB approval by completing an Annual Report (available on the website). The report should be submitted 3 weeks in advance of the REB approval expiry date to allow time for REB review and to prevent a lapse of ethics approval for the research.

Researchers should note that no research involving humans may be conducted in the absence of a valid ethical approval and that allowing REB approval to lapse is a violation of the University Scholarly Misconduct Policy, inconsistent with the TCPS and may result in the suspension of research and research funding, as required by the funding agency.

5. Final review - final reports

When the researcher is confident that all research-related interventions or interactions with participants have been completed (for prospective research) and/or that all data acquisition is complete, there will be no further access to participant records or collection of biological materials (for secondary use of information research), a Final Report (available on the website) must be submitted to Research Ethics. After review and acknowledgement of the Final Report, the Research Ethics file will be closed.

6. Retaining records in a secure manner

Researchers must ensure that records and data associated with their research are managed consistent with their approved research plans both during and after the project. Research information must be confidentially and securely retained and/or disposed of in such a manner as to comply with confidentiality provisions specified in the protocol and consent forms. This may involve destruction of the records, or continued arrangements for secure storage.

It is the researcher's responsibility to keep a copy of the REB approval letters. This can be important to demonstrate that research was undertaken with Board approval. Please note that the University will securely store your REB project file for 5 years after the REB approval end date at which point the file records may be permanently destroyed.

7. Current contact information and university affiliation

The lead researchers must inform the Research Ethics office of any changes to contact information for the PI (and supervisor, if appropriate), especially the electronic mail address, for the duration of the REB approval. The PI must inform Research Ethics if there is a termination or interruption of their affiliation with Dalhousie University.

8. Legal Counsel

The Principal Investigator agrees to comply with all legislative and regulatory requirements that apply to the project. The Principal Investigator agrees to notify the University Legal Counsel office in the event that they receive a notice of non-compliance, complaint or other proceeding relating to such requirements.

9. Supervision of students

Faculty must ensure that students conducting research under their supervision are aware of their responsibilities as described above and have adequate support to conduct their research in a safe and ethical manner.

Appendix H – Supplementary Materials

1. Visual Analysis of Oceanic Multivariate Species Time-Series Data Presentation

A power point presentation describing the overview of the complete thesis, which was used for thesis defence has been uploaded as an electronic supplement on DalSpace.

2. Visualization of Sharks and Fishes based on Detections Demo

A video demo showcasing the various use cases of shark and fishes based on detections, as mentioned in chapter 3, has been uploaded as an electronic supplement on DalSpace. The video provides visualizations and demonstrations of how the dashboard works in identifying sharks and fishes based on detections in different scenarios. This video may also be found at https://drive.google.com/file/d/1AAR6jPMHmB5lRkVXqfFFMi7Wfni2RWlP/view?usp=sharing

3. Visualization of Sharks and Fishes based on Mining Algorithms through Patterns and Rules Demo

A video demo showcasing the various use cases of shark and fishes based on mining algorithms through patterns and rules, as mentioned in chapter 3, has been uploaded as an electronic supplement on DalSpace. The video provides visualizations and demonstrations of how the dashboard works in identifying patterns and rules of sharks and fishes based on mining algorithms in different scenarios. This video may also be found at the below shared link
https://drive.google.com/file/d/1Ow9A8u52Kucwke6JpsYw_lTwE85TBZgE/view?usp=sharing