

**Unsupervised Machine Learning as a Tool for Exploratory Analysis of
Acoustic Telemetry Data: A Case Study With Northern Pike in
Toronto Harbour.**

by

Adogbeji Agberien

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in
partial fulfillment of the requirements for the degree of

Master of Science

in

Biology with Specialization in Data Science

Carleton University
Ottawa, Ontario

© 2021, Adogbeji Agberien

Abstract

Spatial ecology aims to further knowledge of an organism's relationship with its environment and guide decision-making related to conservation. The advancement of biotelemetry has facilitated this goal, however, data management, from its acquisition to its utilization, is central to its success. Standard analysis may include separating tagged individuals into predefined groups based on biometrics or capture location, and then comparing relationships among groups, environmental measures, and their seasonal habitat choices. While effective in that it informs on the relationship among variables, this approach is computationally intensive, and the insight provided is limited to behaviour among predefined groups. This study effectively and efficiently leverages machine learning methods - hierarchical clustering and principal component analysis - to explore animal behaviour, thus providing an efficient, alternative method to analyzing acoustic telemetry data. A by-product of this project is software development that can facilitate analysis of acoustic telemetry data.

Acknowledgements

I'd like to dedicate this thesis to my mum and dad, thanks for your continued support. To my siblings, Vanessa, Fejiro, and Ephraim, you motivate me more than you could ever imagine, so this is for you too.

I have come to understand that ever so often, it takes a village to raise a child, and on that note, thanks to the extended family, aunties, and uncles who supported me through this academic and life journey. To my friends who listen to me rant and go on tangents on scientific topics, those who contribute to the discussions and those who listen, thank you. Special thanks to Lucas Brydges who also contributed to the edit of this paper.

To my supervisors, Dr Steven Cooke, Dr Jonathan Midwood, and Dr Patrick Farrell, I appreciate your guidance through this project, but most importantly, appreciated the freedom to freely explore the statistical learning initiatives in this project.

Table of Contents

Abstract	i
Acknowledgements	ii
List of Tables	iii
List of Figures	iii
1 Introduction	1
1.1 Aquatic Ecology, Acoustic telemetry, and its Data	1
1.1.1 Background	1
1.1.2 Why Acoustic Telemetry	4
1.1.3 Machine Learning as a Tool for Efficient and Effective Data Analysis	5
1.2 Unsupervised Machine Learning as a Tool for Analysing Acoustic Telemetry Data	8
1.2.1 Principal Component Analysis	9
1.2.2 Hierarchical Clustering	11
1.3 Acoustic Telemetry and the Great Lakes	12
1.3.1 The Great Lakes Areas of Concern and Toronto Harbour as the Study Site	12
1.3.2 Northern Pike as the Study Species	14
1.4 Research Statement	15
2 Methods	16
2.1 Habitat Conditions and Receiver Deployment	16
2.2 Northern Pike Information	18
2.3 Determining Receiver Clusters Based on Environmental Data	19
2.4 Analysis of Detection Data: Data Pre-Processing	20
2.5 Analysis of Detection Data: Clustering of Northern Pike and Receiver Groups	21
3 Results	23
3.1 Habitat and Receiver clusters	23
3.2 Hierarchical clustering of Northern Pike based on the time spent at the different receiver groups	29
3.2.1 Temporal Behaviour of Northern Pike in Cluster One	30
3.2.2 Temporal Behaviour of Northern Pike in Cluster Two	32
4 Discussion	35

Appendix	39
4.1 Code I	39
4.2 Code II	66
Bibliography	81

List of Tables

3.1	Receiver groups in Toronto Harbour, their environmental characteristics, and habitat clusters and receiver clusters to which they belong	25
3.2	Cross-tabulation of habitat clusters and receiver clusters	26
3.3	Correlation table informing the proportion of nodes which each hierarchical clustering method shares with the others.	30

List of Figures

1.1	A largemouth bass being implanted with a LOTTEK acoustic tag (Smith-Root electric fish handling gloves on colleague as lake-water is being pumped over the gills during surgery. Credit: Alice Abrams	3
1.2	A diagram illustrating the basic concept of acoustic telemetry as the presence of a fish is being detected by a receiver. Credit: GLATOS (https://glatos.glos.us/Acoustic)	3
2.1	Map of Toronto Harbour and the position of acoustic receivers (grey circles) within the harbour. The inset map indicates the position of Toronto Harbour within Lake Ontario and relative to the other Great Lakes.	18
3.1	Violin plots illustrating the distribution of respective environmental measured within each habitat cluster.	24
3.2	Violin plots illustrating the distribution of respective environmental measured within each receiver cluster.	26
3.3	Bar plots indicating the percentage of variance explained by each principal component	28
3.4	PCA bi-plot illustrating temporal receiver preference within Toronto Harbour.	29
3.5	Dendrogram illustrating results of several hierarchical clustering methods	30
3.6	Bar graph illustrating the percent of time spent at the different receiver groups by northern pike in cluster one.	31
3.7	Line graphs and bar plots illustrating monthly detection activity within Toronto Harbour for northern pike in cluster 1. Plot A represents the mean monthly depths of the waters at which these northern pike resided; plot B represents the mean monthly percent SAV of the waters at which these northern pike resided; plot C represents the mean monthly stratification temperature of the waters at which these northern pike resided; plot D represents the percentage of time spent relative to overall time in the harbour; plot E represents percentage of time spent at the respective receiver cluster during the indicated month.	32
3.8	Bar graph illustrating the percent of time spent at the different receiver groups by northern pike in cluster two.	33
3.9	Line graphs and bar plots illustrating monthly detection activity within Toronto Harbour for northern pike in cluster 2. Plot A represents the mean monthly depths of the waters at which these northern pike resided; plot B represents the mean monthly percent SAV of the waters at which these northern pike resided; plot C represents the mean monthly stratification temperature of the waters at which these northern pike resided; plot D represents the percentage of time spent relative to overall time in the harbour; plot E represents percentage of time spent at the respective receiver cluster during the indicated month.	34

Chapter 1

Introduction

1.1. Aquatic Ecology, Acoustic telemetry, and its Data

1.1.1 Background

Earth is a vast and dynamic system with a complex array of diverse and interacting life. The waters cover approximately seventy percent of earth, and these aquatic ecosystems - freshwater and marine - are generally teeming with life; some of which include fish, insects, mammals, and plants. Within these aquatic ecosystems, there are complex interactions among organisms, and between the organisms and their environment (Marshall 2013). The study of aquatic ecosystems is important as they provide important resources such as food, and ecosystem services such as climate regulation, for humans and other organisms (Lynch et al. 2016). This study of the relationships pertaining to an aquatic ecosystem, including the interaction among the organisms, and between these organisms and their environment is defined as aquatic ecology. Some challenges of aquatic ecology studies include the vast size of the aquatic ecosystem, water clarity, and the dynamic and complex interactions within these ecosystems (Hussey et al. 2015). The advent of telemetry has mitigated this challenge and its continued development has facilitated projects in which it is utilized (Cooke et al. 2013).

Broadly, telemetry can be defined as the collection and transmission of data, and it is a tool which

can be used across remote locations. Within aquatic ecosystems, it primarily involves the use of electronic tags that transmit signal to monitor aquatic organisms. There are various types of telemetry (*e.g.*, acoustic, radio, and satellite) and they are classified based on their method of signal transmission (Cooke et al. 2013). Each type of telemetry has its pros and cons, and considerations on the sufficient method to utilize is better done during the project design phase (*i.e.*, prior to data collection; Cooke et al. (2013)). This thesis utilizes data acquired via acoustic telemetry, however, the analytical concept can be extended to other telemetric data analysis.

Acoustic telemetry follows the logic of sonar, a World War I development. The technology has however continued in advancement which has led to its utilization within the fisheries industry (Hockersmith and Beeman 2012). Acoustic telemetry can be simply defined as the use of sound signals to attain information on the presence of an electronically tagged organism. There are two main components of acoustic telemetry - transmitters and receivers (Cooke et al. 2013). The transmitters are electronic tags that are either implanted or externally attached to organisms, and send sound pulses at a pre-determined rate through water. Each transmitter has a uniquely assigned ID which acts as a proxy for the organism to which it is affixed. Some transmitters are also equipped with sensors that measure environmental conditions of their location. The sound pulses these transmitters send are heard by the “receivers” within range; thus the receivers are data-loggers, typically within the waterbody, that detect the pings of transmitters and log information such as the identity of the organism and if capable, other environmental information such as the water temperature at the time when the organism is detected.

Figure 1.1, shows the process of fish-tagging as a transmitter is being surgically implanted in the largemouth bass (*Micropterus salmoides*), and Figure 1.2 illustrates the basic action of a receiver detecting pings emanating from the transmitter in the tagged fish.

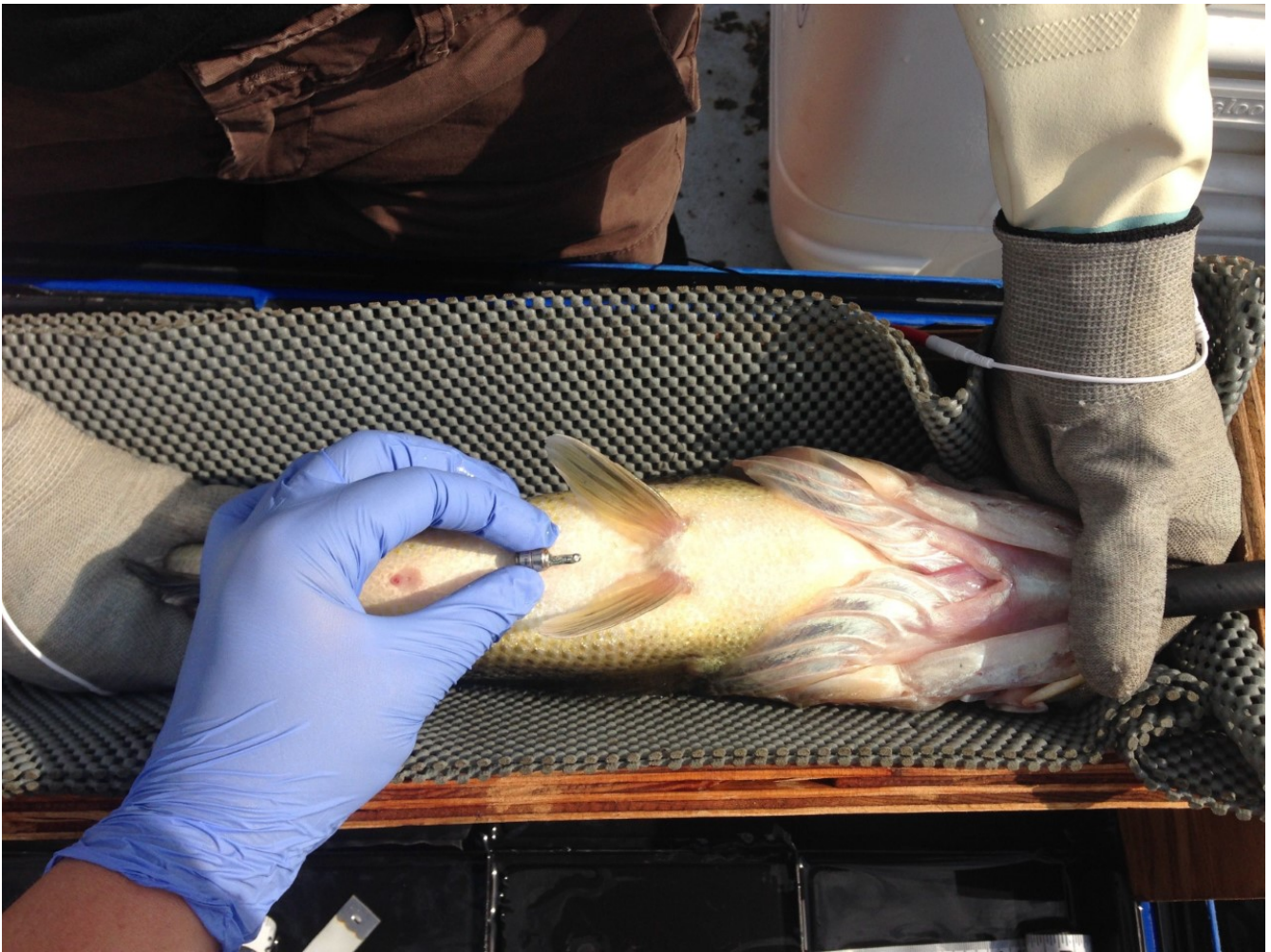


Figure 1.1: A largemouth bass being implanted with a LOTEK acoustic tag (Smith-Root electric fish handling gloves on colleague as lake-water is being pumped over the gills during surgery. Credit: Alice Abrams



Figure 1.2: A diagram illustrating the basic concept of acoustic telemetry as the presence of a fish is being detected by a receiver. Credit: GLATOS (<https://glatos.glos.us/Acoustic>)

1.1.2 Why Acoustic Telemetry

As mentioned in Cagnacci et al. (2010), “*Movement is the glue that ties ecological processes together*”. Aquatic ecosystems are dynamic, with lots of inhabitant movements and interactions across spatio-temporal scales; and the study of movements within these systems can inform on the state of the ecosystem (Cagnacci et al. 2010; Hussey et al. 2015; Miller et al. 2019). Telemetry facilitates the study of movement by constantly logging spatio-temporal details of the tagged species, as such, information such as resource requirements and the changing state of an ecosystem can be attained through the study of population-level movements. During the 1950s, researchers began to use acoustic telemetry to study fish, sufficiently performing research duties such as monitoring salmon migration (Hockersmith and Beeman 2012). Acoustic telemetry has since become a well-sought tool for monitoring animal movement because it is relatively cheap yet efficient in long-term, continuous data collection within a specific study site (Cagnacci et al. 2010; Brownscombe et al. 2019).

Acoustic telemetry is capable of providing insight into home ranges, spawning sites, and residency, however, the accuracy of these insights are dependent on the robustness of the project design. Essentially, while acoustic telemetry is an efficient tool for monitoring studies, it is not without its cons. First, the researcher must recognize that with acoustic telemetry, the transmitter as opposed to the organism which it is affixed to is what is detected, as such, false conclusions may occur if the transmitter becomes detached from the organism or if the organism is preyed upon. To reduce the likelihood of such conclusions, it is essential to pre-process the data. Some other concerns to be accounted for during project design and analysis include battery life of the tags, the detection range, precision and accuracy of the receivers, and potential false detections caused by signal collisions due to multiple transmitters that operate at the same frequency being within range of a receiver (Gjelland and Hedger 2013; Brownscombe et al. 2019). Also, while telemetry facilitates the study of movement, conclusions on activity are limited to inferences, except in cases where there was direct observation of activity while logging telemetry data. Additionally, although a short-term effect, the method of capture and tagging, and tag type may influence the behaviour of the tagged species (Brownscombe et al. 2019). Given that the primary reason for a transition to

acoustic telemetry is its efficiency in data-collection, of which said data-collection may be affected by the aforementioned factors, a robust project design is thus paramount to the success of acoustic telemetry. With proper project design, the pros of acoustic telemetry still outweigh its cons, thus making it a frequently utilized tool within the fisheries science.

Telemetry finds use in behavioural research, and is useful for evaluating the success of habitat restoration projects. As an example, hardening of shorelines caused by development of boat slips can result in a loss of fish habitat and productivity, and to examine the success of rehabilitation, tagged fish can be monitored to study their use of the environment, and ultimately, conclusions and data-driven decisions can be made based on the attained information (Veilleux et al. 2018). Acoustic telemetry also facilitates collaboration among researchers who use compatible tracking devices and share their data (Cooke et al. 2011; Krueger et al. 2018), thus facilitating studies across broad spatial scales (Block et al. 2016; Krueger et al. 2018). This collaboration permits more expansive studies, thus gaining more data to either validate knowledge or gain new insights on behaviour of the tagged species. Additionally, the capability of telemetry to concurrently collect information on environmental variables (*e.g.*, depth and water temperature) entices its use as the primary data-collection tool. This advantage of efficient data collection consequently means a proliferation in data, and also presents a challenge associated with a big data - efficient management and processing.

1.1.3 Machine Learning as a Tool for Efficient and Effective Data Analysis

Due to its data collection capability, the fisheries industry and research and management that support it have become ever-present and high-volume adopters of acoustic telemetry. This efficiency of acoustic telemetry has led to a proliferation in data as transmitters have rapid ping rates (*e.g.*, second to minutes) and can be detected simultaneously on numerous receivers. On realizing the benefit of this data collection capability, informal and formal networks that facilitate data sharing among researchers have arisen; examples include the Great Lakes Acoustic Telemetry Observation System (GLATOS; Krueger et al. (2018)) and the Ocean Tracking Network (OTN; Cooke et al.

(2011)).

Just as we have efficient methods of data collection, we need efficient methods of data analysis. Much of the data we collect are high-dimensional, with a number of unknown interactions among variables. They may also have non-linear dependencies and violate assumptions of classic statistical methods. For example, acoustic telemetry data violates assumption of independence as successive observations of the same individual are spatio-temporally correlated. So often the task may not require the intensiveness of classic statistics methods as it simply aims to explore underlying behaviour and make guided decisions based on the insight attained. On that note, a definition for classic statistics and machine learning is due.

There is a substantial overlap between classic statistics and machine learning as both aim to build mathematical representations of the underlying data. From my perspective, the underlying mathematical methods are the same in both fields with the primary difference being the syntax and terminology being used; for example, what is referred to as predictor variables in statistics are referred to as input features in machine learning, or what is referred to as mean squared error in statistics is referred to as L2 loss in machine learning. That said, there are differences between both with the primary goal of statistics being to draw population inferences from a sample, and that of machine learning being to find generalizable predictive patterns (Bzdok, Altman, and Krzywinski 2018). Simply put, machine learning models focus on the making the best possible predictions while statistical models focus on inferences about the relationship among variables. As an example, linear regression is an algorithm which is used in both statistics and machine learning. When implementing linear regression in statistics, one builds a model in which inferences on the relationship between the predictor variables and the response variable can be made, whereas while implementing linear regression in machine learning, the data is split into training and test sets so as to improve the performance on the model while making predictions on the test set. While the statistical model may be able to make predictions on previously unseen data, that is not its value, just as the machine learning model may provide insight on relationships but that is not its value. On that note, both statistical and machine learning approaches have value in ecological studies, and determining what approach suffices for a particular task depends on the purpose of

the project.

When seeking to understand the relationship between a set of predictor variables and a target variable, the classic statistical approach takes the helm as its aim is to accomplish such (Rogers and White 2007; Whoriskey et al. 2019). An analytical work-flow may include separating tagged individuals into groups based on physical features (e.g., length), or their original capture location, along with a set of environmental variables that are empirically or theoretically associated with behaviour (Midwood et al. 2019). A statistical analysis of the relationship among time spent, the environmental variables, season of detection, and the physical or physiological variables are then evaluated. To make valid conclusions on these relationships however, assumptions based on the method used must be confirmed (*e.g.*, normality and independence of each variable). On the other hand, the goal of a project may be to make predictions so as to guide decision-making. For example, in the project by Brownscombe et al. (2020), they collect information such as tracking data, environmental data, and visual surveys of marine fish species, permit (*Trachinotus falcatus*). Their data was of a small sample size and had a large amount of measured features, of which some of these features were correlated. The data violated assumptions and were plagued with the curse of dimensionality (*i.e.*, large number of variables and fewer observations), as such, provided analytical challenges for a classic statistical approach. Given the goal was to predict spawning sites as opposed to understanding the relationship between collected variables and spawning sites, the authors implemented machine learning models (classic and conditional random forests, and fuzzy k -means) and evaluated analysis based on the predictive performance of each model.

Other example of machine learning utilization include wildebeest identification documented in (Valletta et al. 2017); as counting the population of wildebeest is a necessary but tedious task for sufficient monitoring of ecosystem health in the Serengeti National Park, Tanzania (Estes 2014). To alleviate this challenge of a tedious count, Random Forests, a SML method, was used to identify wildebeest in aerial photos of the Serengeti (Valletta et al. 2017).

ML methods fall into one of two categories - supervised learning methods or unsupervised learning methods. Supervised Machine Learning (SML) is such that for each observation of the predictor

variable(s), x_i , there is an associated response variable y_i , while Unsupervised Machine Learning (UML) is such that for each observation of the predictor variable(s), x_i , there is no associated response variable, y_i . UML methods are more challenging to validate than SML methods; this is because for SML, one can validate on prediction results using metric such as precision, false negative rate, or F-score, while for UML, there is no such metric to evaluate the accuracy of groupings. Essentially, SML algorithms can be readily validated because of the presence of response data that “supervise” analysis while UML algorithms lack response variables and as such results are not readily verified by a calculated metric (James et al. 2013). That said, subject matter expertise plays a vital role in UML because one can look at the content/overarching characteristics of each group and decipher biological reliability of groups formed via UML. SML methods can be used to identify the relationship between predictor and response variables and can also be used for prediction and classification tasks while UML methods are used for identifying groups or patterns in unlabelled data, and can be used for dimensionality reduction *i.e.*, to reduce the number of predictor variables.

1.2. Unsupervised Machine Learning as a Tool for Analysing Acoustic Telemetry Data

Ever so often, analysis of ecological data requires the implementation of UML algorithms, an example of which is the case study in this thesis, *i.e.*, to discover behavioural patterns in the tagged species within the harbour. Passive acoustic telemetry generates spatio-temporal data that is unlabelled, and when the goal is to discover groups and patterns in the data, UML suffices for such tasks. UML can be used to provide insights through visualization, discover relationships among variables, and discover groups in a set of observations. Among the algorithmic components of UML, there are dimensionality reduction/visualization techniques, such as principal component analysis (PCA), and clustering techniques, such as hierarchical clustering.

1.2.1 Principal Component Analysis

The goal of the dimensionality reduction technique, PCA, is to take a set of points in high-dimensional space and find a representation of these points in low-dimensional space (preferably 2-D), so that results of analysis are easily comprehensible to the analyst (James et al. 2013). The mathematical overview of PCA is shown below.

Given an $n \times m$ dimensioned-matrix, such that:

$$X_{n \times m} = \begin{bmatrix} X_{1,1} & X_{1,2} & \dots & X_{1,m} \\ X_{2,1} & X_{2,2} & \dots & X_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,2} & \dots & X_{n,m} \end{bmatrix}$$

where n represents the number of rows (observations) and m represents the number of measurements (variables), the goal of PCA is to reduce the dimension of the matrix, yet retain most of the information in the large $n \times m$ -sized dataset. This goal is beneficial as reducing the size from $n \times m$ to say, a 2-D or 3-D matrix, we are then able to visually comprehend our data. To achieve this goal, some important steps are; feature scaling (*i.e.*, standardizing the data), computing the variance-covariance matrix, and computing the eigen-decomposition of the variance-covariance matrix.

PCA achieves its goal by determining which variables maximize the variance in the dataset, and failing to standardize the data will skew the results of analysis. For example, say we have the length of fish in metres (cm), and it varies less than their weight in kilograms (kg), PCA will determine that the direction of the maximal variance is more toward their weight, and thus represent the weight more than the height. Standardizing the data makes the weight and length comparable by scaling such that each has a mean of 0 and standard deviation of 1, thus preventing a biased result. To standardize the data, we calculate the z -score which is simply:

$$z = \frac{X - \mu}{\sigma}, \text{ where } \sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Standardization is done based on the goal of analysis (*i.e.*, comparing variables or observations). Using the standardized values, we compute the variance-covariance matrix, $Cov(Z_{m \times m}) = \frac{1}{N} Z_{m \times n}^T Z_{n \times m}$. This step is done to be able to evaluate the relationship between variables as some variables may be highly correlated, thus making them redundant. The variance-covariance matrix will be a symmetric $m \times m$ matrix where the diagonal elements are the variances of standardized variables Z_1 to Z_m , and the elements above and below the diagonal elements are the covariances informing the relationship between variables. A positive covariance informs a positive correlation, and a negative covariance informs inverse correlation.

The third step is eigen-decomposition of the variance-covariance matrix to attain the eigenvectors, W and the eigenvalues, λ ; of which the W s are the principal components and λ s inform on the magnitude or effect size of that principal component. By organizing the principal component matrix, $W_{m \times m}$ by their corresponding λ s in decreasing order, the first principal component W_1 will explain the most amount of variation in the data, and the last principal component W_m will explain the least amount of variance in the data set. By multiplying the original data matrix $X_{n \times m}$ with the principal component matrix $W_{m \times m}$, we attain a data matrix $T_{n \times m}$, which is a representation of $X_{n \times m}$ but with points that can be plotted along the principal component axis, and provide a comprehensible representation of the data set. At a basic level, the validity of the method can be observed in the fact that we start with a data matrix of X , sized at $n \times m$, and end with a data matrix T , also sized at $n \times m$. By selecting the first k components of T , PCA may sufficiently provide a representation of the data set in k -dimensional space, thus providing the analyst with readily comprehensible visuals and potentially speeding up downstream analysis as the amount of computation is reduced from m dimensions to k dimensions.

The capability of PCA to sufficiently reduce the dimension of the data set and provide easily comprehensible visuals makes it a well sought technique in subjects such as ecology that deal with large, high-dimensional data (Midwood et al. 2018). This usefulness of PCA is further demonstrated in this thesis.

1.2.2 Hierarchical Clustering

Clustering is the process of finding subgroups, or rightly termed “clusters” in a data set. The goal is to partition observations such that the observations within a cluster are more similar to themselves than the observations within another cluster. For example, given a data set of characteristics of several fish species (*e.g.* their weight, length, scale count, fin position, *e.t.c*), and say we do not have identifier labels of each species (*e.g.*, coho salmon (*Oncorhynchus kisutch*) and northern pike (*Esox lucius*)), an effective clustering algorithm would work such that it would rightly group similar species into the same cluster, all without having known the identity of the species. Some examples of clustering algorithms include k -means, Hierarchical Clustering (HC), Gaussian Mixture Models (GMM), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). For this thesis, hierarchical clustering was utilized because the primary goal of the project is to automatically identify clusters within the data set (*i.e.*, without pre-specifying the number of clusters). While a method such as DBSCAN does not require that the number of clusters be defined, it does require that we specify ϵ which is the distance that defines a neighbourhood, and we specify θ which is the minimum number of points required to form a cluster (Ester, Kriegel, and Xu 1996). Additionally, hierarchical clustering assumes that every data point is relevant; and given how we aggregate our data, we want each data point to contribute in determining which cluster an observation belongs to (more on data restructuring and aggregation in the methods section). Hierarchical clustering simply requires that we specify a method for computing distances between observations and subsequently between clusters, and is also less sensitive to the distance metric of choice (James et al. 2013). Ultimately, hierarchical clustering was settled upon because of its ease-in-use when developing an analytical software that requires the user to manipulate parameters. If the user so chooses to pre-specify the number of clusters in the data set, that is also viable with hierarchical clustering.

Hierarchical clustering techniques are divided into two types - agglomerative hierarchical clustering and divisive hierarchical clustering (Hastie, Tibshirani, and Friedman 2009). Agglomerative hierarchical clustering, the technique used in this thesis, is a bottom-up approach to clustering; that is, it begins with each observation being identified as its own cluster, thus implying n clusters in a

data set with n observations. For each data point, it finds its closest data point and merges them together, thus forming $n - 1$ clusters. Subsequently, for each cluster, it finds the closest cluster and merges them together, thus forming $n - 2$ clusters. The preceding steps are repeated until all clusters are tied into one large cluster (a cluster containing all sub-clusters, and ultimately all observations). A relatively important parameter to specify when utilizing agglomerative hierarchical clustering is the linkage parameter (often referred to as the method). The linkage parameter is important because it is the basis on which the dissimilarity between clusters is computed. Common linkage methods are - Complete linkage, Single linkage, Average linkage, and Ward linkage. Hastie, Tibshirani, and Friedman (2009) provides comprehensive descriptions of the common linkage methods, but in summary, single linkage is based on minimal inter-cluster dissimilarity and uses the minimum of the distances between all observations in the two clusters; average linkage is based on mean inter-cluster dissimilarity and uses the average of the distances of each observation of the two clusters; complete linkage is based on maximal inter-cluster dissimilarity and uses the maximum distances between all observations of the two clusters; finally, Ward linkage is a method based on the analysis of variance (ANOVA) and appoints observations to clusters based on the least total within-cluster-variance (*i.e.*, the pairs of clusters with the minimum between-cluster distance are merged).

1.3. Acoustic Telemetry and the Great Lakes

1.3.1 The Great Lakes Areas of Concern and Toronto Harbour as the Study Site

The Great Lakes are of environmental and socio-economic importance to Canada and the United States of America (USA); however, these waters have Areas of Concern that were determined based on pollution, overuse of resources, urbanization, and ultimately, their deteriorated environmental health (Environment and Climate Change Canada 2015). In 1972, the Canadian and USA governments signed an agreement “to restore and maintain the chemical, physical, and biological integrity of the Waters of the Great Lakes.” This agreement was referred to as the Great Lakes

Water Quality Agreement (GLWQA). In 1987, the GLWQA identified 43 Areas of Concern (AOC) and since then, the Government of Canada (GOC) has supported initiatives to clean up the AOC by providing significant funds to local initiatives (“Great Lakes: Areas of Concern” 2007). The most recent update of the GLWQA recognizes the need to “anticipate and prevent emerging threats to the quality of the Waters of the Great Lakes” (Canada 2012). In order to adhere to the GLWQA, there are constant monitoring programs within the Great Lakes that collect information pertaining to environmental health. Among the AOC, the focus of this project is the Toronto and Region AOC which is situated along the North Shore of Lake Ontario.

Toronto Harbour and some of the areas east and west of it are part of the Toronto and Region AOC, and they had suffered significant losses of aquatic habitat due to urbanization. Since their listing as AOC, the Toronto and Region AOC Remedial Action Plan (RAP) was developed, and it involved improvement of fish habitat and installation of acoustic arrays to monitor movement and habitat selection within the harbour (Veilleux et al. 2018; Midwood et al. 2019; Barnes et al. 2020). The waters have undergone rehabilitation actions to improve and ensure habitats capable of sustaining native populations. Some of the restoration efforts included creating coastal wetlands, restoring habitat lost through urban development, and establishing of new habitats to support fish and wildlife life-cycle (Barnes et al. 2020). Northern pike (*Esox lucius*) were one of the target species for much of the habitat restoration works within the harbour, as they are native to the area and are recreationally important species (Midwood et al. 2019).

Monitoring and documenting their use of the harbour can inform on the success of the restoration and can facilitate future restoration projects. A comprehensive summary of residency, habitat selection, and within-harbour movements of fishes over a five-year period, starting in 2010, was thus performed by Midwood et al. (2019). A significant by-product of that project is a standard operating procedure for pre-processing of the data, and readily available summaries for comparison with the results of this project.

1.3.2 Northern Pike as the Study Species

Northern pike are a keystone predatory fish species that provide top-down control in freshwater ecosystems (Paukert and Willis 2003; Craig 2008). They are generally found in northern latitudes of Asia, Europe, and North America, and their occurrence could be natural or due to fish-stocking practices (Craig 2008; Paukert and Willis 2003). Northern pike are subject to vast amounts of research because of their importance to commercial and recreational fisheries, extensive distribution, and pivotal role in fisheries. They are a hardy species that can tolerate a broad range of environmental conditions (Pierce 2012), but generally have their activity, feeding, physiology, and reproduction affected by factors such as temperature, dissolved oxygen, and pH (Casselman 1996). Accordingly, there are optimal conditions and factors needed for northern pike to thrive. Among the factors known to influence northern pike habitat choice are mean wind fetch, mean depth, mean summer stratification water temperature, and percent cover of submerged aquatic vegetation (% SAV) (Pratt and Smokorowski 2011).

Northern pike are cool-water species that spawn in the spring, when the water has warmed to between 8-12°C (Casselman 1996). As it warms in the mid-summer, primarily when the water exceeds 20°C - 25°C, northern pike usually seek out cooler, deeper waters (Casselman 1996; A. Kobler et al. 2008). Depth and mixing of water do play important roles as the water temperature profile changes. Wind fetch is defined as the distance wind blows over water without obstruction. An increased wind fetch is directly proportional to increased wave action because of the absence of obstruction as wind blows across the water. The shape of the waterbody and presence of vegetation can thus attenuate wind action. Decreased wave action implies calmer waters and influences biomass production, but extreme wave action affects the survival of eggs. Wave action can facilitate mixing as the water warms, and is thus beneficial in the early spring during spawning season. Casselman (1996) devised a ranking chart of the physical requirements of habitat used by spawning northern pike; this ranking was based on extensive field observations, however, the numerical ranks were subjective, with 1 representing the least important feature and 9 representing a highly important feature. In this rank, wave action was rated 2, thus implying a lesser important physical feature, however, in the presence of high water, moderate wave action was considered

better than little or no wind exposure, or extreme wave action.

Depth and % SAV are directly proportional because shallower waters imply more sunlight penetration to support plant life. Northern pike prefer these shallower ($< 12m$), waters with increased % SAV for several reasons including that it supports their opportunistic predatory behaviour, and importantly is a good area for spawning as their eggs readily attach to the submerged aquatic vegetation (Cook and Bergersen 1988; Craig 2008). Some works have however identified distinct intra-population behaviours with selection of distinct vegetation types in two groups and a more mobile morph in the third (Alexander. Kobler et al. 2009). Both depth and % SAV have high scores in the ranking system by Casselman (1996), with depth scoring 9 and % SAV scoring 8. This ranking is in regards to spawning habitat and informs that their optimal spawning habitat is 10-70cm deep with between 40% - 90% dense submergent and emergent aquatic vegetation. The quality of the spawning habitat declines as the water gets deeper or shallower, or as vegetation gets sparser or denser.

Northern pike are among the dominant predatory fish within Toronto Harbour that provide top-down control within the ecosystem, as such, monitoring them makes practical sense for the restoration project. They are also effective study species to validate the method used in this paper because the substantial amount of information published on the species provides a benchmark for comparison of results.

1.4. Research Statement

The purpose of this study is to explore an alternative approach to studying animal behaviour using acoustic telemetry. Specifically, the study aims to investigate the behaviour of northern pike based on the detection data as opposed to using a hypothesis-based method to explore behaviour. Subsequently, to validate the effectiveness of the method, results are compared to published information on northern pike behaviour.

Chapter 2

Methods

The acquired data were pre-processed, visualized, and analysed using R software (R Core Team 2021). Mapping was also done using QGIS (QGIS Development Team 2021). For easy reproducibility as described in Xie, Allaire, and Golemund (2018), Xie, Dervieux, and Riederer (2020), and Allaire et al. (2021), data analysis was done in an R markdown environment, concurrently with the project write-up. A significant portion of the analysis was conducted in R markdown because it speeds up the process of citing packages used for analysis. Also, given that the report is written in the same environment the analysis is conducted it circumvents a continuous switch of the working environment. The data pre-processing was not done in R markdown because it would take a longer run time to produce the analytical report given the original data set is larger than the cleaned data set. Essentially, it results in a longer run time when knitting the document into the desired output (*e.g.*, pdf). The code used for analysis after pre-processing is shown in the Appendix, Code I section; complete code including preprocessing can be seen on the github page, <https://github.com/dijiagberien/ExplorationOfAnimBehavTorHarNorthernPike>

2.1. Habitat Conditions and Receiver Deployment

Several environmental variables were measured to investigate the habitat preference of northern pike. These variables included mean wind fetch, mean depth, mean summer stratification water

temperature, and percent cover of submerged aquatic vegetation (% SAV). These variables were selected based on their known influence on fish behaviour (Casselman 1996; Pratt and Smokorowski 2011) as well as the availability of these data within the Toronto and Region AOC (Midwood et al. 2019).

Detection data were acquired using autonomous underwater acoustic receivers with integrated hydrophones (VR2W, Innovasea). The receiver arrays were initially organized to cover a variety of habitat conditions as well as select locations of interest where rehabilitation works had occurred or were planned, but the spatial organization of deployments have since expanded as the focus shifted to covering new areas of interest.

In some locations, such as Embayment C, Spadina, and Peter slip, there were more receivers than at other locations (Figure 2.1). This was because the initial project design was primarily aimed at assessing and informing the extent to which restored areas were used (Veilleux et al. 2018). To facilitate analysis, locations with multiple receivers and relative habitat homogeneity were grouped together. For example, the five receivers within Peter slip were grouped as “Peter Slip,” and any detection on any of the five receivers deployed in Peter slip was noted as a “Peter Slip” detection. This collection of multiple receivers in locations with relative habitat homogeneity was therefore referred to as a receiver group.

There were gaps in data collection at certain receiver groups due to factors such as; the receiver groups not being deployed, becoming disconnected, being removed, or becoming lost. Midwood et al. (2019) provide details on deployment history of receiver groups in the harbour; information pertaining to the brand, battery life, and mass of the different receivers are also provided.



Figure 2.1: Map of Toronto Harbour and the position of acoustic receivers (grey circles) within the harbour. The inset map indicates the position of Toronto Harbour within Lake Ontario and relative to the other Great Lakes.

2.2. Northern Pike Information

Between 2010 and 2018, a total of 158 northern pike were tagged in Toronto Harbour. The northern pike in the study were tagged at a different times, as such, it was expected that the total number of detections or total duration varied. To understand northern pike habitat preferences through time, the northern pike clusters in the harbour were firstly determined as detailed in the section on *Analysis of Detection Data: Clustering of Northern Pike and Receiver Groups*. Following the clustering of northern pike based on their temporal activity in the harbour, a weighted average of the monthly depth and % SAV at the receiver groups where these northern pike resided was calculated. The weighted average was calculated based on the percent time spent at the different

receiver groups; for example, to calculate the average depth at which the northern pike resided in December, the time spent at the different receiver groups at which the northern pike were detected is calculated, and the mean depth of each receiver group is also noted. The formula is then applied as shown in equation (2.1):

$$\bar{E} = \frac{T_1 E_1 + T_2 E_2 + \dots + T_n E_n}{\sum_{i=1}^n T_i} \quad (2.1)$$

where \bar{E} refers to the average of the environmental variable of interest (*i.e.*, depth or %SAV), E refers to the environmental variable of interest, T refers to the percent time spent, i refers to the specific receiver group when there are n receivers groups present in the receiver cluster. By applying this weighted average formula, we are able to better incorporate a temporal aspect into their behavioural analysis.

2.3. Determining Receiver Clusters Based on Environmental Data

Hierarchical clustering was used to determine the receiver clusters based on environmental measures - mean wind fetch, mean depth, mean percent submerged aquatic vegetation, and mean stratification temperature. Prior to clustering, each environmental variable was scaled, so as to reduce the overshadowing effect any variable may have while implementing the algorithm. Hierarchical clustering was performed using the base R “hclust” function with a “Ward.D” method and the “complete” method using a dissimilarity structure based on euclidean distances (R Core Team 2021). The optimal number of receiver clusters was determined based on the within cluster sum of squares.

2.4. Analysis of Detection Data: Data Pre-Processing

The first step of analysis was data-preprocessing. The original dataset contained 12,648,092 observations from 41 variables, of which some of these variables were redundant. An example of redundancy is “Location = EMC,” “Station no = 10,” “Station = EMC-010,” in which case both “Location” and “Station no” were excluded. Some other columns contained information that were not required for analysis, such as common name of the organism and its scientific name. Based on these assessments, several columns were removed leaving only six columns – animal ID, detection date-time, location name, longitude, latitude, min_lag. The min_lag is another recorded parameter by the receiver and it informs the time between subsequent detections of the same pike at the same receiver group.

After reducing variables to those deemed necessary for this analysis, the next step was to remove repeated detections and potential false detections. Removing repeated detections was done by sorting the data based on northern pike ID and detection time, and subsequently excluding repeat observations. False detections on the other hand can arise when multiple transmitters are located within detection range of a receiver and their emitted signal is detected at the same time thus producing erroneous ID. The erroneous ID may be the same as a valid ID that has not been deployed in this study site but rather somewhere else (a Type A false positive). In this scenario, this ID could be easily identified and removed given its irrelevance to our study. A more challenging circumstance is identifying a Type B false positives, *i.e.*, erroneous IDs that are the same as others within our study site (Simpfendorfer et al. 2015).

Perfectly removing Type B false positives is almost impossible given the inherent difficulties of differentiating between such an erroneous ID and a true positive ID. On that note, a commonly utilized technique for excluding such potentially false detections is omitting detections of the same code, on the same receiver that are separated by 30 times the nominal delay. The logic behind such an approach is that given subsequent detections of the same individual, on the same receiver, the detections with shorter interval are more probable than those with long intervals. While true, the measure of “30 times” the nominal delay is subjective, but has become a rule of thumb for pre-

processing of telemetry data. It is also understood that one may remove true positive detections when utilizing this approach, however, the loss of data will only be a small fraction of the complete detection data. The GLATOS R package contains a function called “false_detections,” and it implements the aforementioned method, only requiring that the arguments be specified [*i.e.*, a threshold time interval of 3600 seconds; Thomas, Hayden, and Holbrook (2018); Holbrook et al. (2019)]. In this case, 1.3% of the data were identified as potentially false detections and these observations were removed from the dataset.

Other pre-processing steps included calculation of timing between detections for each northern pike as this was needed for both determining potentially false detections and for calculating total residency duration in the harbour. Single detections at receiver groups that lasted less than six hours (singletons) and long absences (greater than 24 hours between detections) were also removed. The timing between detections was then recalculated following exclusion of singletons and long absences from the array. Computations on the data frame were done using a combination of R’s `data.table` (Dowle and Srinivasan 2020) and `tidyverse` packages (Wickham et al. 2019) as they are efficient tools when working with tabular data.

2.5. Analysis of Detection Data: Clustering of Northern Pike and Receiver Groups

To determine the groups of northern pike within the harbour, the data were summarized in a tabular format such that for each northern pike, the average time spent at each receiver group was calculated. This simply meant converting the data from a long format to a wide format and aggregating based on the time spent at each receiver group. Finally, for each pike, the data on time spent was standardized to have a mean of zero and standard deviation of one. This was done because inherently, each northern pike resided in the harbour for different durations given their different tagging time frames. Furthermore, the purpose of the subsequent clustering was to determine groups of pike based on their preferred locations within the harbour as opposed to determining pike that resided more frequently within the harbour.

Hierarchical clustering was performed using the base R “hclust” function with a “Ward.D” method and a dissimilarity structure based on euclidean distances (R Core Team 2021). The optimal number of clusters was determined based on the within cluster sum of squares, and was calculated and visualized using the “fviz_nbclust” function from the “factoextra” package (Kassambara and Mundt 2020).

To determine receiver groups that were utilized in similar manner, a temporal scale of interest was first chosen; in this case by month, because of interest in seasonal behaviour. Next, the data were summarized in a tabular format such that for each receiver group, the cumulative time spent by the northern pike for each month was computed. To mitigate the effect of discrepancies in the data collected (for example missing data due to receiver being removed and reinstalled), a two-sided circular moving average was calculated using three months of data (2.2):

$$\bar{T}_i = \frac{T_{i-1} + T_i + T_{i+1}}{3} \quad (2.2)$$

where \bar{T} is the average time spent, T is the time spent, and i is the current month. Three months were chosen as the length of moving average because of the estimated number of months per season. Explicitly, the time spent at a particular receiver group in January was computed as the average of the time spent at that receiver group in December, January, and February. Finally, for each receiver group, the time spent was standardized to have a mean of zero and standard deviation of one.

Hierarchical clustering was then performed using the base R “hclust” function with a “Ward.D” method and a dissimilarity structure based on euclidean distances (R Core Team 2021). The optimal number of receiver clusters was determined based on the within cluster sum of squares. This was also calculated and visualized using the “fviz_nbclust” function from the “factoextra” package (Kassambara and Mundt 2020). Finally, results from clustering based on environmental data were cross-tabulated against results based on detection data, and PCA was used to visualize the receiver preferences through time.

Chapter 3

Results

3.1. Habitat and Receiver clusters

For clarity, the clusters based on the measured environmental variables are referred to as habitat clusters as they are solely based on the integrated habitat conditions, and the clusters based on detection data are referred to as receiver clusters as they are solely based on detections at the receiver groups.

The mean wind fetch, depth, % SAV, and stratification temperature of each of the receiver groups are shown in Table 3.1. As determined by hierarchical clustering, there are primarily two habitat clusters within the harbour with one habitat cluster containing 19 receiver groups and the other habitat cluster containing 15 receiver groups. As illustrated in the violin plots on figure 3.1, receiver groups in habitat cluster one (*habOne*) are characterized by relatively deep waters, low % SAV, and relatively low stratification temperature. While some *habOne* receivers have high mean wind fetch, they are indeed outliers relative to other receiver groups within the cluster. Table 3.1 indicates all the receiver groups, their environmental characteristics, and the habitat cluster to which they belong. Comparatively, receiver groups in habitat cluster two (*habTwo*) are characterized by shallow waters, high % SAV, and relatively high stratification temperature.

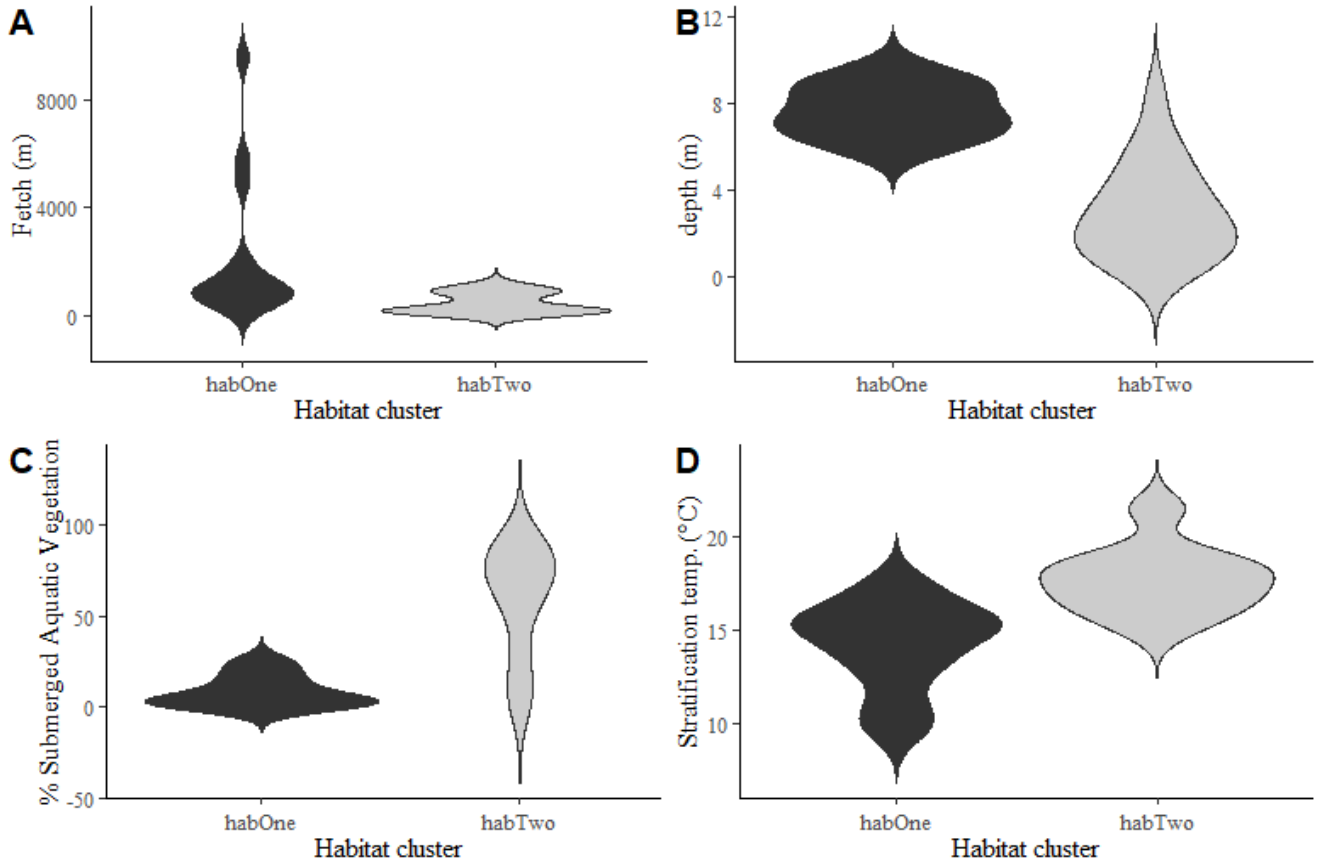


Figure 3.1: Violin plots illustrating the distribution of respective environmental measured within each habitat cluster.

Based on hierarchical clustering on the detection data (*i.e.*, mean monthly time spent at the different receiver groups), there are also two receiver clusters within the harbour, with the one receiver cluster containing 22 receiver groups and the other receiver cluster containing 12 receiver groups. As illustrated in the violin plots in Figure 3.2, much like the results of hierarchical clustering on the environmental variables, receiver groups in receiver cluster one (*recOne*) are primarily characterized by relatively deep waters, low % SAV, and relatively low stratification temperature, while receiver groups in cluster two (*recTwo*) are primarily characterized by relatively shallow waters, high %SAV, and relatively high stratification temperature. Table 3.1 indicates all the receiver groups, their environmental characteristics, and the receiver cluster to which they belong.

Table 3.1: Receiver groups in Toronto Harbour, their environmental characteristics, and habitat clusters and receiver clusters to which they belong

receiver group	Receiver cluster	Habitat cluster	Fetch	Mean depth	% SAV	Stratification temp.
Billy Bishop East	recOne	habTwo	862	6.1	83.2	16.9
Billy Bishop West	recOne	habOne	9585	5.7	0.1	15.3
Cell 1	recOne	habTwo	155	0.9	63.6	21.3
Cell 2	recTwo	habTwo	181	1.3	6.0	18.8
Cell 3	recOne	habOne	299	7.9	12.5	17.3
Cherry Beach	recOne	habOne	1547	6.5	24.5	9.7
Cherry Beach 2b	recOne	habOne	5854	6.1	5.2	10.1
Curtain	recOne	habOne	4903	6.5	0.6	12.8
Don River	recOne	habTwo	110	1.1	3.7	19.3
Don River Mouth	recTwo	habOne	702	7.2	3.7	13.8
E Western Gap	recOne	habOne	695	8.2	2.3	15.5
Embayment A	recOne	habTwo	222	2.4	28.1	14.8
Embayment B	recOne	habTwo	1161	1.2	3.6	17.6
Embayment C	recTwo	habTwo	234	2.9	23.3	16.0
Embayment D	recTwo	habTwo	276	0.1	69.4	21.8
Jarvis	recOne	habOne	922	8.9	8.2	15.6
Mid.Waterfront	recOne	habOne	1050	8.5	0.2	17.3
N Eastern Gap	recOne	habOne	898	9.2	17.8	15.4
OHM	recOne	habTwo	520	4.3	56.1	15.4
Parliament	recOne	habOne	650	7.3	2.3	13.4
Peter Slip	recOne	habTwo	867	8.4	56.6	16.3
S Eastern Gap	recOne	habOne	2026	9.7	16.3	14.5
Spadina	recOne	habOne	830	6.9	25.0	16.0
TOI-027	recOne	habTwo	181	3.0	90.1	18.0
TOI-040	recTwo	habTwo	94	2.4	74.9	18.8
TOI-041	recTwo	habTwo	140	2.2	77.7	18.5
TOI-042	recTwo	habTwo	78	1.5	81.1	17.9
TOI-043	recTwo	habTwo	105	0.6	84.8	18.4
TOI-044	recTwo	habTwo	782	3.5	85.0	15.9
TOI-045	recTwo	habTwo	862	6.1	83.2	16.9
TOI-046	recTwo	habTwo	977	4.8	62.3	17.5
TOI-047	recTwo	habTwo	1016	4.3	82.3	16.9
Turning Basin	recOne	habOne	157	8.9	7.0	11.2
W Western Gap	recOne	habOne	1276	7.4	2.4	15.2

Table 3.2: Cross-tabulation of habitat clusters and receiver clusters

	habOne	habTwo
recOne	14	8
recTwo	1	11

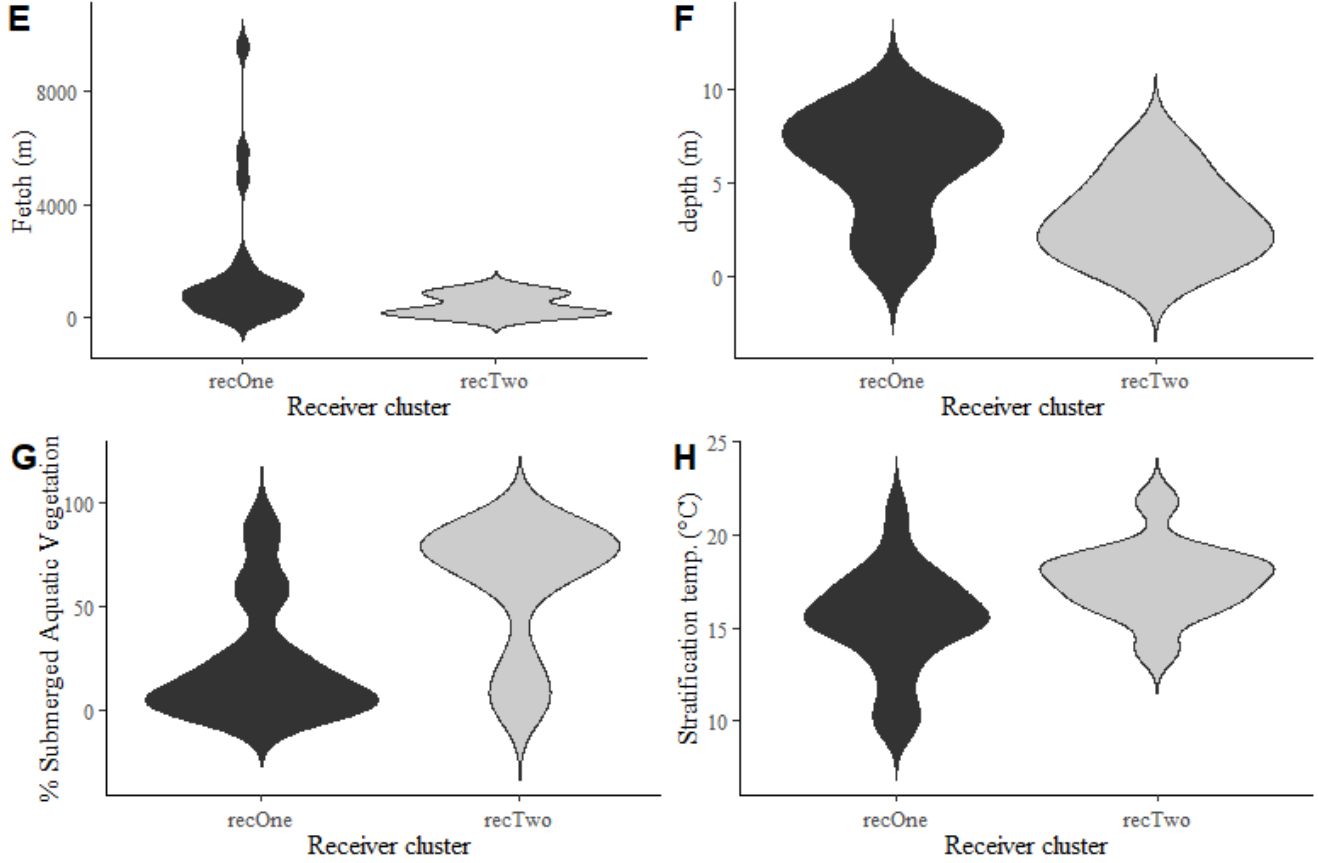


Figure 3.2: Violin plots illustrating the distribution of respective environmental measured within each receiver cluster.

Results of hierarchical clustering based on the habitat environmental characteristics and hierarchical clustering based on the mean monthly time spent at each location were cross-tabulated and are shown in Table 3.2. Cross-tabulation informs approximately 73% cluster similarity when comparing the results of clustering based on measured environmental variables with results of clustering based on the average monthly time spent at these receiver groups (Table 3.2).

The *recTwo* receiver that was characterized as a *habOne* receiver is Don River Mouth, while the eight *recOne* receivers that were characterized as *habTwo* receivers are Billy Bishop East, Cell 1,

Don River, Embayment A, Embayment B, OHM, Peter Slip, TOI-027. The overall characteristics of these receiver groups suggest that they were rightly characterized as *habTwo* receivers (*i.e.*, shallow waters, high % SAV, and relatively high stratification temperature.). For example, although Don River has a low % SAV of 3.7%, it has a stratification temperature, mean depth, and fetch that are all below the harbour average.

The receiver clusters (*i.e.*, *recOne* and *recTwo*) are characterized based on temporal detections at corresponding receiver groups, and according to PCA, the first two principal component are enough to explain approximately 80% of the variation in the data set (Figure 3.3). The PCA bi-plot (Figure 3.4) then illustrates that northern pike primarily reside in the waters of receiver cluster one in the colder months (*i.e.*, January to March, and October to December), and increase their residency in receiver cluster two in the warmer months (*i.e.*, April to September). This information is sufficiently represented along principal component one, which explains 50% of the variance in the data. When interpreting the PCA bi-plot, it should be noted that because of the dimension along which the observations were scaled, the plot is representative of when each receiver group experienced increased detections as opposed to informing which locations were utilized most frequently.

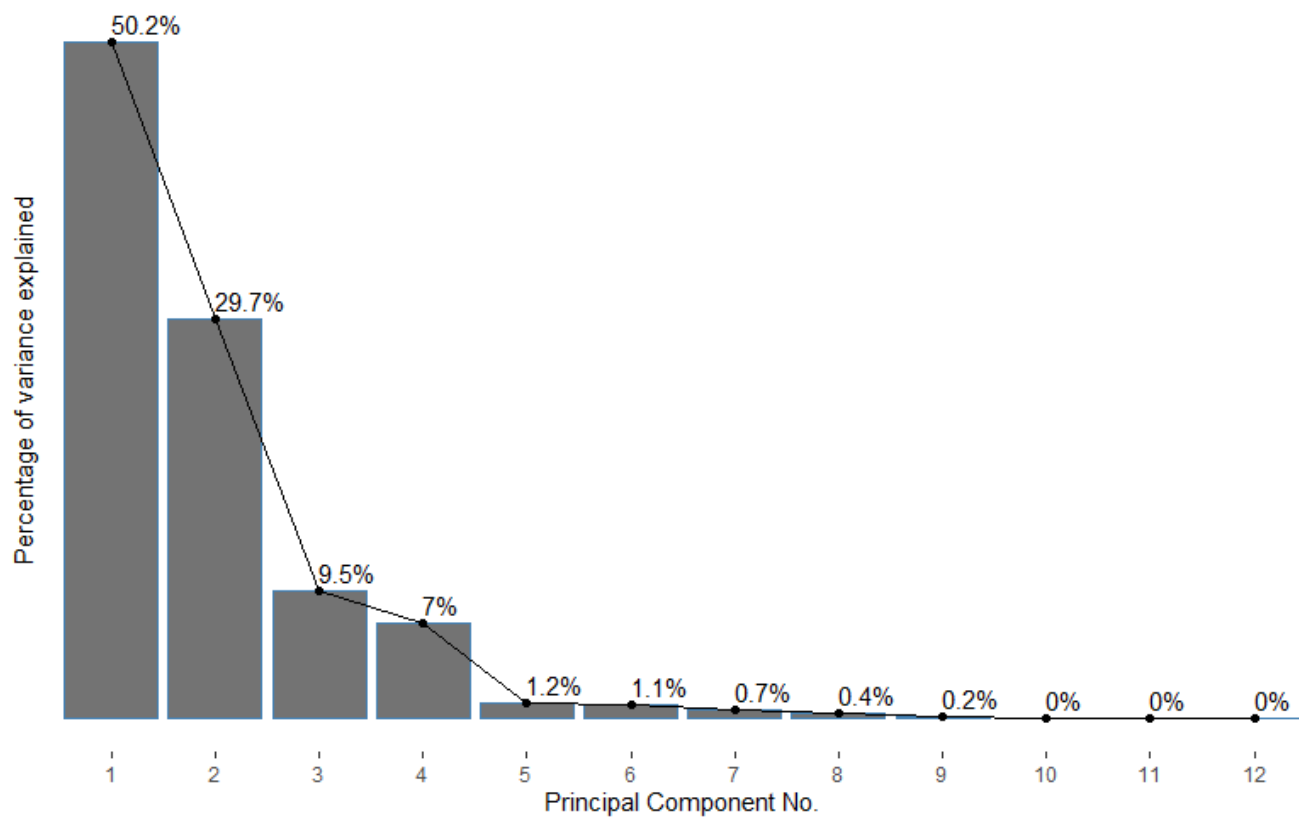


Figure 3.3: Bar plots indicating the percentage of variance explained by each principal component

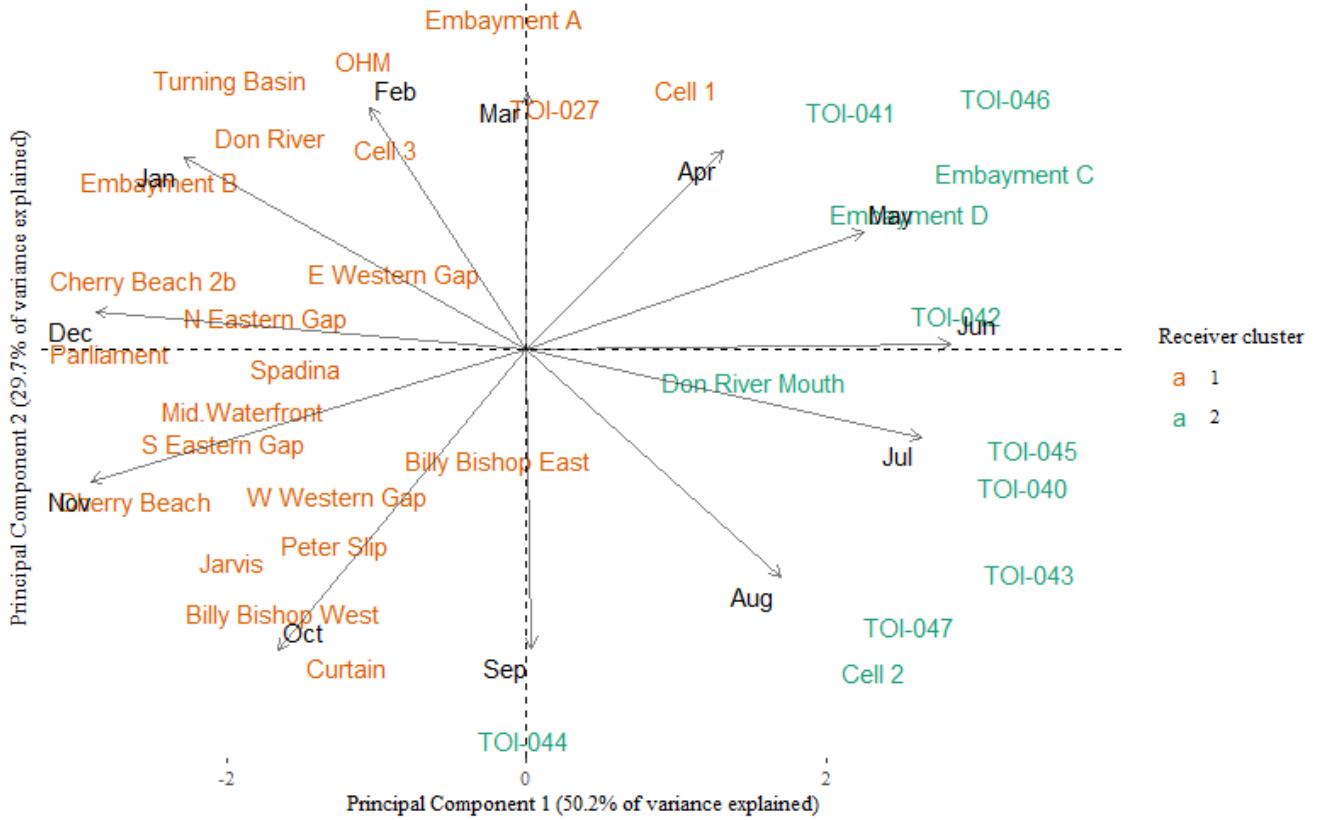


Figure 3.4: PCA bi-plot illustrating temporal receiver preference within Toronto Harbour.

3.2. Hierarchical clustering of Northern Pike based on the time spent at the different receiver groups

The optimal number of northern pike clusters was determined to be two based on the within cluster sum of squares, and from Figure 3.5, we see that each hierarchical clustering method - Ward, Single, Complete, and Average - results in identification of northern pike clusters within the harbour. However, from Table 3.3, we see that these clusters varied in terms of their members. Ward's method was ultimately chosen because of the ease with which it permits a user to visually examine cluster similarity.

Table 3.3: Correlation table informing the proportion of nodes which each hierarchical clustering method shares with the others.

	Ward	single	complete	average
Ward	1.00	-	-	-
single	0.68	1	-	-
complete	0.82	0.68	1	-
average	0.79	0.69	0.8	1

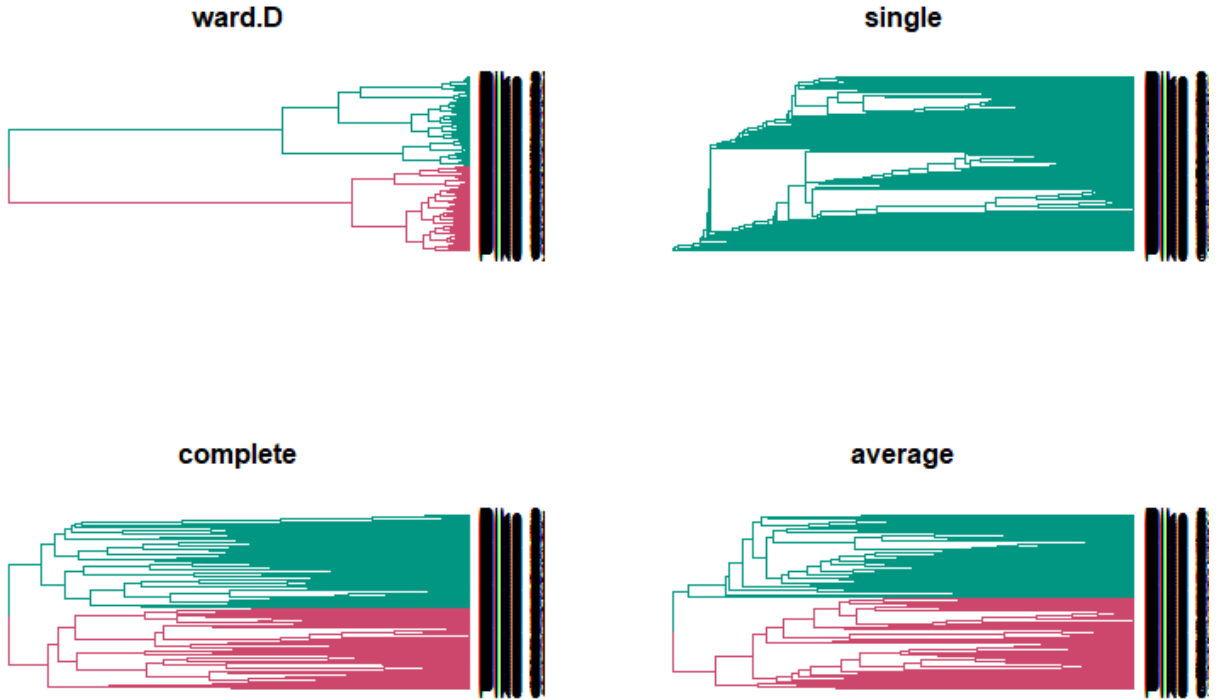


Figure 3.5: Dendrogram illustrating results of several hierarchical clustering methods

3.2.1 Temporal Behaviour of Northern Pike in Cluster One

Northern pike in cluster one primarily resided in the inner harbour (Figure 2.1; Figure 3.6), spending 67.9% of their time at the *recOne* receivers and 32.1% of their time at the *recTwo* receivers. As illustrated in Figure 3.6, the top three receiver groups at which they resided are Spadina (14.4%),

Mid. Waterfront (13.7%), and TOI-041 (12.4%). Table 3.1 shows that Spadina and Mid. Waterfront are primarily characterized by their relatively deep waters (6.9m and 8.5m respectively), low % SAV (25% and 0.2% respectively), while TOI-041 is characterized by its shallow waters (2.2m) and high mean % SAV (77.7%).

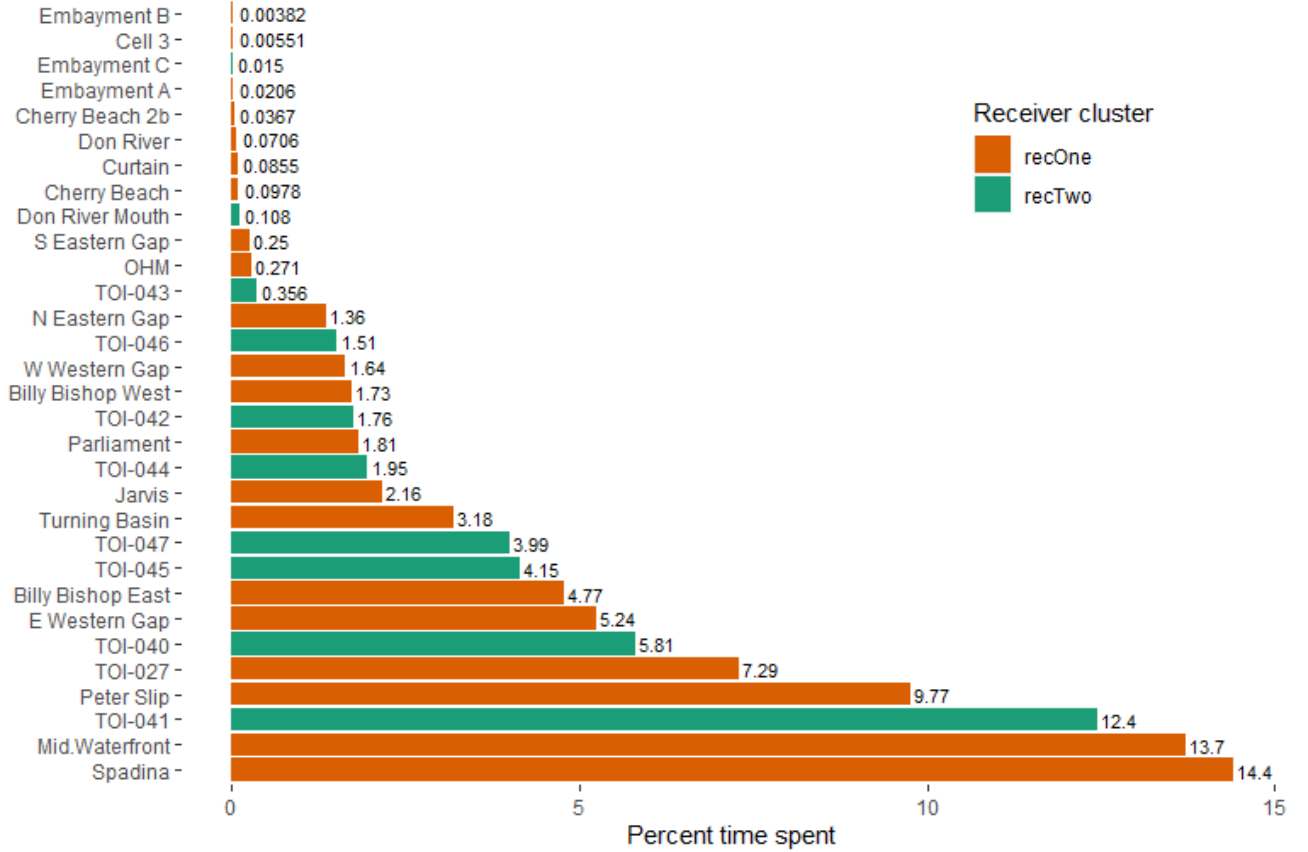


Figure 3.6: Bar graph illustrating the percent of time spent at the different receiver groups by northern pike in cluster one.

From plot D in Figure 3.7, we observe that residency in the harbour is at its peak in the later months of the year (*i.e.*, October, November, and December), and in the early months of the year (*i.e.*, March, April, and May). We also observe from plot E that although these northern pike resided primarily in the cluster 1 receivers, their residency in the cluster 2 receivers gradually increases from January until June, after which it gradually decreases until December. Additionally, from plots A, B, and C in Figure 3.7, we observe that during the spring and early summer months, the tagged northern pike spend more time in shallow waters with relatively high % SAV and increased stratification temperature, and then increase their residency in deeper waters with relatively low

stratification temperature in the late summer and fall months.

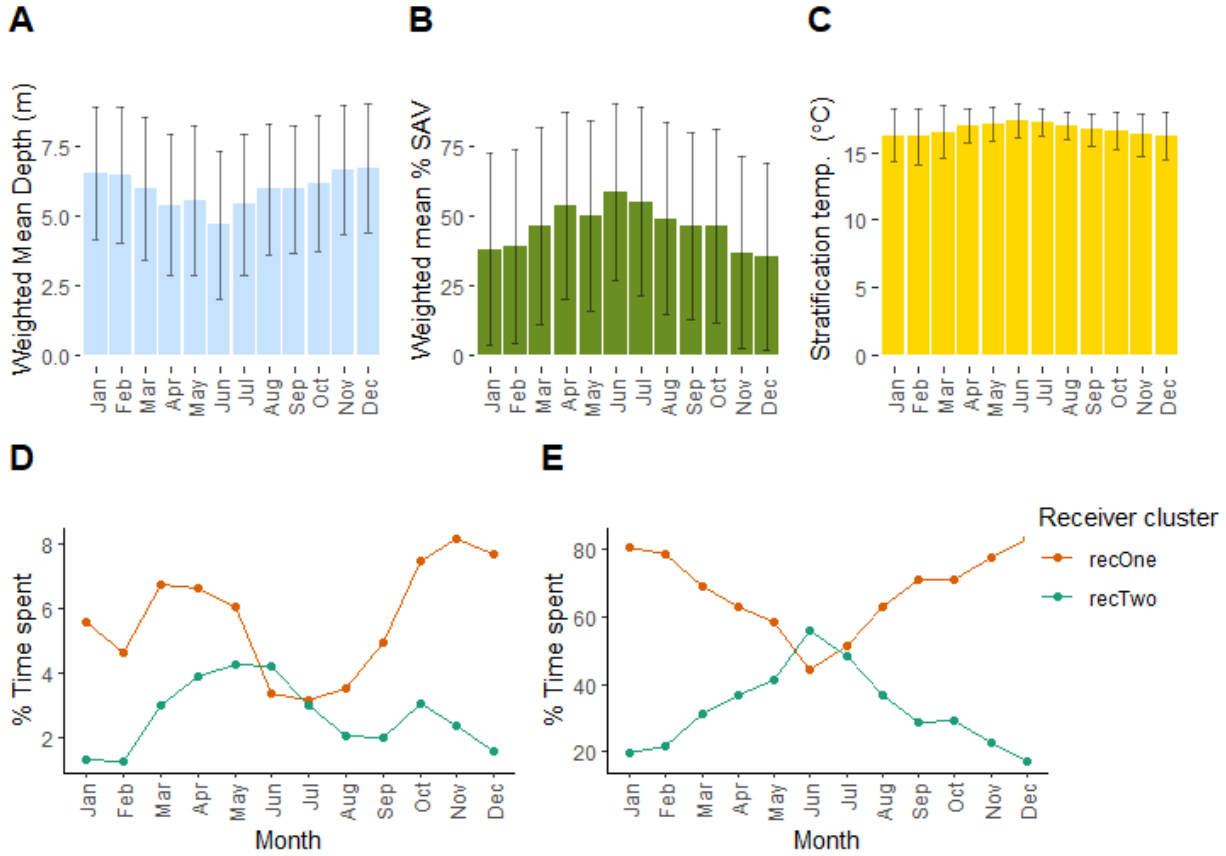


Figure 3.7: Line graphs and bar plots illustrating monthly detection activity within Toronto Harbour for northern pike in cluster 1. Plot A represents the mean monthly depths of the waters at which these northern pike resided; plot B represents the mean monthly percent SAV of the waters at which these northern pike resided; plot C represents the mean monthly stratification temperature of the waters at which these northern pike resided; plot D represents the percentage of time spent relative to overall time in the harbour; plot E represents percentage of time spent at the respective receiver cluster during the indicated month.

3.2.2 Temporal Behaviour of Northern Pike in Cluster Two

Northern pike in cluster two primarily resided in the outer harbour (Figure 2.1). These northern pike spent almost equal amount of time at both receiver clusters, spending 52.9% of their time at the *recOne* receivers and 47.1% of their time at the *recTwo* receivers. As illustrated in Figure 3.8, these northern pike primarily resided in Embayment C (36.9%) and Cell 3 (15.7%); they did however reside at Cell 2 (8.82), Cherry Beach (8.32), Jarvis (7.02), OHM (6.79), and Curtain (6.45) in almost equal proportions. Table 3.1 shows that Embayment C is characterized by its shallow

waters (2.9m), relatively low % SAV (23.3%), and moderate stratification temperature (16°C).

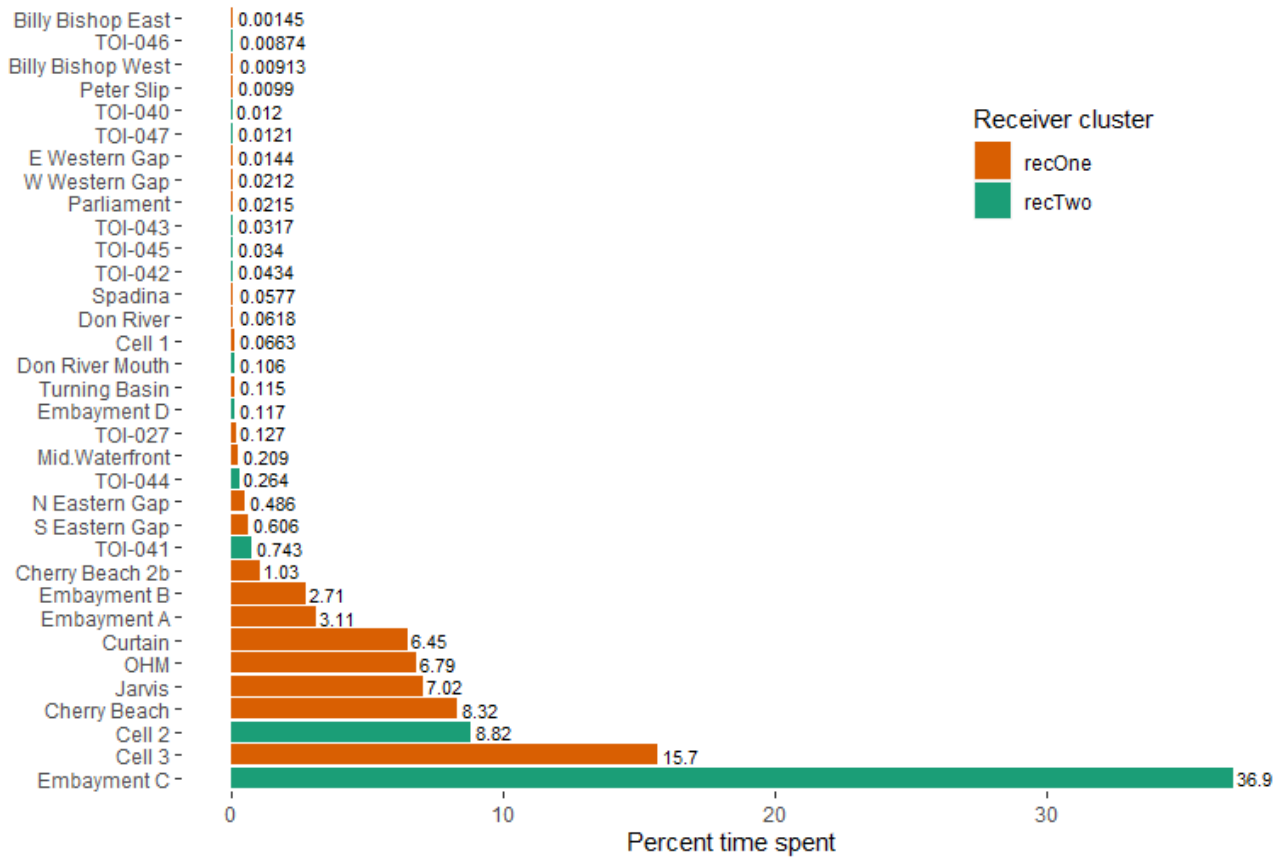


Figure 3.8: Bar graph illustrating the percent of time spent at the different receiver groups by northern pike in cluster two.

From plot D in Figure 3.9, we observe that residency in the harbour is at its peak in the spring and fall months, and from plot E we observe a preference for the waters of *recTwo* from May until August, and a preference for the waters of *recOne* in the other months. Additionally, from plots A, B, and C in Figure 3.9, we observe that during the fall months, they primarily reside in relatively deep waters; during the spring months, they primarily reside in waters with increased % SAV; and during the summer months, they reside primarily in waters with increased stratification temperature.

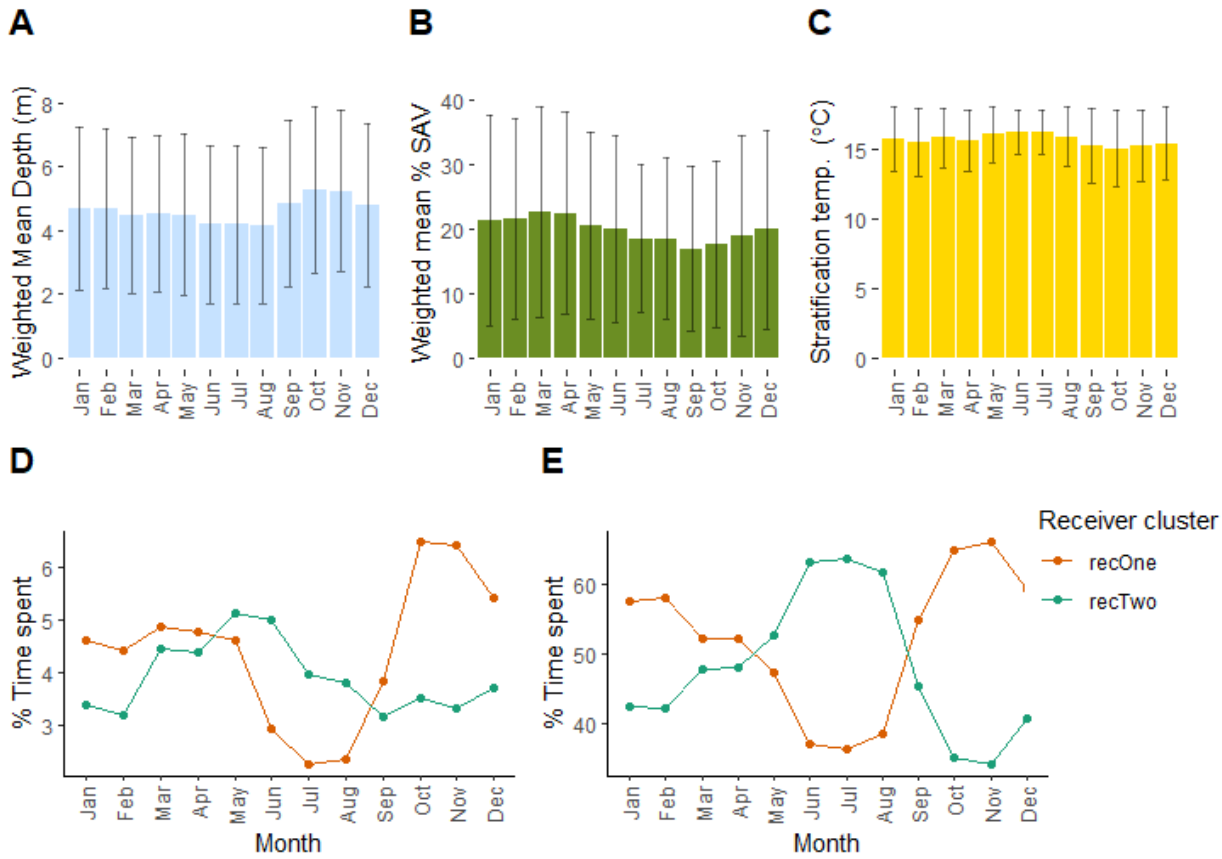


Figure 3.9: Line graphs and bar plots illustrating monthly detection activity within Toronto Harbour for northern pike in cluster 2. Plot A represents the mean monthly depths of the waters at which these northern pike resided; plot B represents the mean monthly percent SAV of the waters at which these northern pike resided; plot C represents the mean monthly stratification temperature of the waters at which these northern pike resided; plot D represents the percentage of time spent relative to overall time in the harbour; plot E represents percentage of time spent at the respective receiver cluster during the indicated month.

Chapter 4

Discussion

Our understanding of how aquatic animals are distributed in space and time has historically been challenging given the inherent difficulties of continually observing animals that live underwater. However, developments in biotelemetry that involve the attachment of small electronic tags to animals has allowed researchers to study aquatic organisms and quantify movements at unprecedented scales (Brownscombe et al. 2019). Data management, from its acquisition to analysis and utilization in decision-making, is central to the success of biotelemetry. This project focuses on the analysis of biotelemetry data, and illustrates that through unsupervised machine learning, we are able to efficiently observe the underlying behaviours within a study system. We were able to objectively define northern pike and receiver (habitat) clusters within Toronto Harbour, and subsequently evaluate the behaviour of the northern pike clusters within the harbour.

Published information on northern pike behaviour suffices as valid benchmark to validate the results of the UML methods used in this project. Accordingly, as illustrated in Figure 3.7 and Figure 3.9, we observed increased residency in the spring and fall months, but decreased residency in the harbour in the mid-late summer months. This result makes logical sense because of the increased water temperature in the mid-late summer. As informed in the introductory section, *Northern Pike as the Study Species*, and in publication by Casselman (1996), northern pike are cool water species that seek cooler water as the temperature increases above 20°C - 25°C. This decreased detection

is thus likely a shift to cooler, deeper waters and is in agreement with information published by (A. Kobler et al. 2008). For each northern pike cluster, we also observed increased residency in shallow waters with higher %SAV in the spring months, an information in agreement with that provided by Casselman (1996), and is likely due to better spawning conditions.

Results of exploring behaviour based on the detection data can inform interesting characteristics of northern pike. For example, cross-tabulation indicated a 73% overlap when comparing results of clustering receiver groups based on environmental characteristics against results when clustering based on detection data. The logic here is that if indeed 100% of their movement and residency could be explained by habitat characteristics, then the results of cross-tabulation between habitat and receiver clusters will result in 100% similarity as opposed to 73% similarity (Table 3.2). While no model may result in 100% cluster similarity (*i.e.*, due to stochastic components of movement and residency), an improved model will result in an increased percentage similarity when compared to the clusters based on detection data. An important note here is that the simplistic clustering utilized here can provide basic information on the underlying behaviour within the system, however, it does not completely capture the intricacies of behaviour. Explicitly, the benefit of this model was such that we had 2 clusters which captured a temporal behaviour of northern pike as they moved between basic habitat types, however the limitation is that there are not strictly two habitat types (*i.e.*, deep with low % SAV or shallow with high % SAV). Habitats may follow an environmental gradient and may be a mix of multiple biotic and abiotic factors that are not mutually exclusive, as such, a more comprehensive study to evaluate and understand behaviour will require that more clusters be formed.

More importantly however, the purpose of this project was to shed light and reinvigorate the utilization of unsupervised machine learning as a tool to examine biotelemetric data. The reach of the method supersedes a focused study of a sole species and can be utilized to explore interactions of a community, or more, with their environment. The method has the benefit of being rapid and easily tuned to include more parameters depending on the scope of the study. It is thus a data-centric approach to discerning groups and providing insight into an organism's (or group of organisms') ecology and behaviour. Moreover, this analytical approach can be used to support

management activities by providing streamlined and consistent data processing and visualization. The method is not without its limitations, and a primary challenge is that simply utilizing the time and location of detection without collection of other characteristics will not further insight into understanding behaviour. Without collection of other biotic and abiotic characteristics, we may know or be able to predict certain actions but will not understand the reason for such actions. As such, much like the classic statistical methods, it requires that additional information pertaining to studied species and their environment be collected. In essence, collecting data of other biotic and abiotic factors alongside detection data can improve our understanding of animal behaviour. The effectiveness of the method is also dependent on the quality of the acquired data. For improved models, the analyst has to account for the nuances in environmental conditions that may influence the quality of acquired data. This is because signals may be disrupted by structures (physical or environmental noise) on the transmission pathway (Gjelland and Hedger 2013). There could also be signal collisions due to multiple transmitters that operate at the same frequency being within range of a receiver (Brownscombe et al. 2019). Such scenarios may result in false detections being included in the data, which would impact the quality of information attained. Additionally, although short-term, the capture-method, tagging-method, and tag type are factors that may influence behaviour, and must be considered during project design (Brownscombe et al. 2019). In essence, while the method is robust, as with other methods, the validity of information attained is dependent on the project design.

Future work of this project will incorporate software development for exploratory analysis of telemetry data. There has been continued use of acoustic telemetry in ecological studies, and consequently, a proliferation in data, which if analysed sufficiently will facilitate ecological studies and related project management duties. Efficient and effective data analysis must thus accompany acoustic telemetry development in order to maximize insight into data. It will be beneficial for project managers to have software that can rapidly analyse acoustic telemetry data and provide results that can address their project goals. On that note, future work will focus on discussions with project managers to understand their project needs (*e.g.*, the specifics on habitat rehabilitation, or targeted questions such as the environmental characteristics that

need to be modified so as to improve biodiversity within a habitat). Software will be further developed to include control tools to permit users with statistical machine learning knowledge or biological knowledge of their study species to be able to tweak hyper-parameters of the model. This software project will be open source and allow other developers to contribute to development as required by the broad needs within the ecological community. The software is currently available at <https://adogbejiagberien.shinyapps.io/ExpFishBehav/>, and the current code for the software is shown in the Appendix section, Code II. Updates will continually be made and available on github at <https://github.com/dijiagberien/Web-App-For-Exploring-Telemetry-Data/blob/main/ExpFishBehav/app.R>

Appendix

4.1. Code I

```
# Show a random subset of the detection data  
detection.data[sample(nrow(detection.data), 5)]
```

```
# Show a random subset of the habitat environmental data  
habitat.data.clean[sample(nrow(habitat.data.clean), 5)]
```

```
# Perform receiver clusters based on the  
# environmental measures of habitats  
habitat.hclust <- habitat.data.clean %>%  
  as.matrix(rownames = ("receiver group")) %>%  
  scale() %>%  
  dist() %>%  
  hclust(method = "ward.D2") %>%  
  as.dendrogram()
```

```
# Cut the tree at 2 branches  
habitat.hclust.treeCut <- cutree(habitat.hclust, k = 2)
```

```
# Check the count in each cluster  
table(habitat.hclust.treeCut)
```

```

# Convert the receiver groups to factors
habitat.groups <- levels(as.factor(habitat.hclust.treeCut))

# Merge habitat clusters with habiat data
habitat.grps.hclust <- cbind(
  habitat.data.clean,
  as.factor(as.character(habitat.hclust.treeCut))) %>%
  data.table(keep.rownames = T) %>%
  setnames(old = c("V2"), new = c("Habitat cluster")) %>%
  dplyr::select(`receiver group`, `Habitat cluster`)

# Determine receiver clusters based on the detection activity
# Convert to wide format such that receiver group ~ data,
# aggregate by total time spent
receiver.month.wide <- dcast(
  detection.data,
  `receiver group` ~ lubridate::month(date, label = T),
  value.var = "time spent (secs)",
  fun = sum) %>%
  as.matrix(rownames = "receiver group") %>% t()

# Perform some arithmetic on the data
# Calculate moving average based on the monthly time spent
ma <- function(x, n = 3){
  stats::filter(x, rep(1 / n, n), sides = 2, circular = T)}

receiver.month.wide[1:12, ] <- ma(
  receiver.month.wide[1:12, ], )

# Scale the data based on monthly detections at the

```

```
# respective locations
```

```
receiver.month.wide <- receiver.month.wide %>%  
  scale() %>% t()
```

```
# Perform hierarchical clustering
```

```
receiver.hclust <- receiver.month.wide %>%  
  dist() %>%  
  hclust(method = "ward.D2") %>%  
  as.dendrogram()
```

```
# Cut the tree at 2 branches
```

```
receiver.hclust.treeCut <- cutree(receiver.hclust, k = 2)
```

```
# Get the number of receivers within each branch
```

```
table(receiver.hclust.treeCut)
```

```
# Convert the receiver groups to factors
```

```
receiver.groups <- levels(  
  as.factor(receiver.hclust.treeCut))
```

```
# Select the receivers and their cluster groups
```

```
receiver.grps.hclust <- cbind(  
  receiver.month.wide,  
  receiver.hclust.treeCut) %>%  
  data.table(keep.rownames = T) %>%  
  setnames(old = c("rn", "receiver.hclust.treeCut"),  
           new = c("receiver group", "Receiver cluster")) %>%  
  dplyr::select(`receiver group`, `Receiver cluster`)
```

```
# Convert the receiver cluster to factors
```

```
receiver.grps.hclust$`Receiver cluster` <- as.factor(  
  receiver.grps.hclust$`receiver group`)
```



```
receiver.grps.hclust$`Receiver cluster`)
```

```
# Merge the receiver clusters based on detection data  
# and the receiver clusters based on environmental vars
```

```
receiver.cluster <- merge(  
  receiver.grps.hclust,  
  habitat.grps.hclust,  
  by = "receiver group")
```

```
receiver.cluster <- merge(  
  receiver.cluster,  
  habitat.data.clean,  
  by = "receiver group")
```

```
# Show first 3 rows in the data set
```

```
receiver.cluster[1:3, ]
```

```
# Rename cluster observations to avoid confusion
```

```
receiver.cluster$`Receiver cluster` <- mapvalues(  
  receiver.cluster[, `Receiver cluster`],  
  c("1"), c("recOne"))
```

```
receiver.cluster$`Receiver cluster` <- mapvalues(  
  receiver.cluster[, `Receiver cluster`],  
  c("2"), c("recTwo"))
```

```
receiver.cluster$`Habitat cluster` <- mapvalues(  
  receiver.cluster[, `Habitat cluster`],  
  c("1"), c("habTwo"))
```

```
receiver.cluster$`Habitat cluster` <- mapvalues(  
  receiver.cluster[, `Habitat cluster`],  
  c("2"), c("habOne"))
```

```
# Set the factor orders
```

```
receiver.cluster$`Receiver cluster` <- factor(  
  receiver.cluster$`Receiver cluster`,  
  c("recOne", "recTwo"))
```

```
receiver.cluster$`Habitat cluster` <- factor(  
  receiver.cluster$`Habitat cluster`,  
  c("habOne", "habTwo"))
```

```
# Visualize environmental characteristics of  
# each habitat cluster
```

```
habitat.fetch <- ggplot(data = receiver.cluster,  
  aes(x = `Habitat cluster`,  
      y = Fetch,  
      fill = `Habitat cluster`)) +  
  geom_violin(trim = F) +  
  scale_fill_grey(start = 0.2, end = 0.8) +  
  labs(x = "Habitat cluster", y = "Fetch (m)") +  
  theme_classic(base_size = 12,  
    base_family = "Times New Roman") +  
  theme(legend.position = "none")
```

```
habitat.depth <- ggplot(data = receiver.cluster,  
  aes(x = `Habitat cluster`,  
      y = `Mean depth`,  
      fill = `Habitat cluster`)) +  
  geom_violin(trim = F) +  
  scale_fill_grey(start = 0.2, end = 0.8) +  
  labs(x = "Habitat cluster", y = "depth (m)") +
```

```

theme_classic(base_size = 12,
              base_family = "Times New Roman") +
theme(legend.position = "none")

habitat.sav <- ggplot(data = receiver.cluster,
                    aes(x = `Habitat cluster`,
                        y = `% SAV`,
                        fill = `Habitat cluster`)) +

geom_violin(trim = F) +
scale_fill_grey(start = 0.2, end = 0.8) +
labs(x = "Habitat cluster",
     y = "% Submerged Aquatic Vegetation", fill = "Group") +
theme_classic(base_size = 12,
              base_family = "Times New Roman") +
theme(legend.position = "none")

habitat.strat <- ggplot(data = receiver.cluster,
                      aes(x = `Habitat cluster`,
                          y = `Stratification temp.`,
                          fill = `Habitat cluster`)) +

geom_violin(trim = F) +
scale_fill_grey(start = 0.2, end = 0.8) +
labs(x = "Habitat cluster",
     y = expression(paste("Stratification temp. ", " (", degree, "C", ")")),
     fill = "Group") +
theme_classic(base_size = 12,
              base_family = "Times New Roman") +
theme(legend.position = "none")

```

```
# Viz environmental vars by receiver cluster
```

```
hab.clust.env.vars <- ggdraw() +  
  draw_plot(habitat.fetch, 0, 0.5, 0.5, 0.5) +  
  draw_plot(habitat.depth, 0.5, 0.5, 0.5, 0.5) +  
  draw_plot(habitat.sav, 0.02, 0, 0.5, 0.5) +  
  draw_plot(habitat.strat, 0.5, 0, 0.5, 0.5) +  
  draw_plot_label(c("A", "B", "C", "D"),  
                  c(0, 0.5, 0, 0.5),  
                  c(1, 1, 0.5, 0.5))
```

```
# Visualize environmental characteristics of each
```

```
# receiver cluster
```

```
receiver.fetch <- ggplot(data = receiver.cluster,  
                        aes(x = `Receiver cluster`,  
                            y = Fetch,  
                            fill = `Receiver cluster`)) +  
  geom_violin(trim = F) +  
  scale_fill_grey(start = 0.2, end = 0.8) +  
  labs(x = "Receiver cluster", y = "Fetch (m)") +  
  theme_classic(base_size = 12,  
                base_family = "Times New Roman") +  
  theme(legend.position = "none")
```

```
receiver.depth <- ggplot(data = receiver.cluster,  
                        aes(x = `Receiver cluster`,  
                            y = `Mean depth`,  
                            fill = `Receiver cluster`)) +  
  geom_violin(trim = F) +  
  scale_fill_grey(start = 0.2, end = 0.8) +
```

```

labs(x = "Receiver cluster", y = "depth (m)") +
theme_classic(base_size = 12,
               base_family = "Times New Roman") +
theme(legend.position = "none")

receiver.sav <- ggplot(data = receiver.cluster,
                      aes(x = `Receiver cluster`,
                          y = `% SAV`,
                          fill = `Receiver cluster`)) +
geom_violin(trim = F) +
scale_fill_grey(start = 0.2, end = 0.8) +
labs(x = "Receiver cluster",
     y = "% Submerged Aquatic Vegetation",
     fill = "Group") +
theme_classic(base_size = 12,
               base_family = "Times New Roman") +
theme(legend.position = "none")

receiver.strat <- ggplot(data = receiver.cluster,
                        aes(x = `Receiver cluster`,
                            y = `Stratification temp.`,
                            fill = `Receiver cluster`)) +
geom_violin(trim = F) +
scale_fill_grey(start = 0.2, end = 0.8) +
labs(x = "Receiver cluster",
     y = expression(paste(
       "Stratification temp. ", " (", degree, "C", ")")),
     fill = "Group") +

```

```
theme_classic(
  base_size = 12, base_family = "Times New Roman") +
theme(legend.position = "none")
```

```
# Viz environmental vars by receiver cluster
```

```
rec.clust.env.vars <- ggdraw() +
  draw_plot(receiver.fetch, 0, 0.5, 0.5, 0.5) +
  draw_plot(receiver.depth, 0.5, 0.5, 0.5, 0.5) +
  draw_plot(receiver.sav, 0.02, 0, 0.5, 0.5) +
  draw_plot(receiver.strat, 0.5, 0, 0.5, 0.5) +
  draw_plot_label(c("E", "F", "G", "H"),
    c(0, 0.5, 0, 0.5),
    c(1, 1, 0.5, 0.5))
```

```
# Cross tabulate both receiver cluster and habitat cluster
```

```
table(receiver.cluster$`Receiver cluster`,
  receiver.cluster$`Habitat cluster`)
```

```
# Get the receiver groups that vary
```

```
receiver.cluster[(`Receiver cluster` == "recOne" &
  `Habitat cluster` == "habTwo") |
  (`Receiver cluster` == "recTwo" &
  `Habitat cluster` == "habOne")]
```

```
# Visualize members of each cluster based
```

```
# on hc on the habitat data
```

```
habitat.dendrogram <- fviz_dend(
  habitat.hclust, k = 2,
  k_colors = c("#1B9E77", "#D95F02"),
  type = "phylogenetic", repel = T) +
theme_dendro()
```

```
# Visualize members of each cluster based on
```

```
# hc on the detection data
```

```
receiver.dendrogram <- fviz_dend(  
  receiver.hclust, k = 2,  
  k_colors = c("#1B9E77", "#D95F02"),  
  type = "phylogenetic", repel = T) +  
  theme_dendro()
```

```
# Perform PCA of receiver ~ month
```

```
receiver.pca <- PCA((receiver.month.wide),  
  scale.unit = F,  
  graph = F,  
  ncp = length(receiver.month.wide))
```

```
# Data visualization
```

```
varExplained <- fviz_eig(receiver.pca, addlabels = T,  
  ncp = length(receiver.month.wide),  
  barfill = "gray45") +  
  
  theme_classic() +  
  labs(title = "") +  
  xlab("Principal Component No.") +  
  ylab("Percentage of variance explained") +  
  theme(axis.line.x = element_blank(),  
    axis.line.y = element_blank(),  
    axis.text.y = element_blank(),  
    axis.ticks.y = element_blank())
```

```
# Bi-plot visualization
```

```
receiver.month.biplot <- fviz_pca_biplot(  
  receiver.pca,  
  habillage = as.factor(receiver.hclust.treeCut),  
  geom = "text",  
  repel = T,  
  col.var = "black",  
  alpha.var = 0.5,  
  palette = c("#D95F02", "#1B9E77"),  
  addEllipses = F,  
  title = " ",  
  legend.title = "Receiver cluster") +  
  theme_classic(  
    base_size = 10,  
    base_family = "Times New Roman") +  
  labs(title = "") +  
  xlab(  
    "Principal Component 1 (50.2% of variance explained)") +  
  ylab(  
    "Principal Component 2 (29.7% of variance explained)") +  
  theme(axis.line.x = element_blank(),  
        axis.line.y = element_blank(),  
        axis.text.y = element_blank(),  
        axis.ticks.y = element_blank())  
  
receiver.cluster.legend <- get_legend(receiver.month.biplot)  
  
dendro.pcaBiplot <- ggdraw() +
```



```

draw_plot(receiver.dendrogram, 0.2, 0.5, 0.5, 0.5) +
draw_plot(receiver.month.biplot +
          theme(legend.position = "none"),
          0.2, 0, 0.5, 0.5) +
draw_plot(receiver.cluster.legend, 0.70, 0.3, 0.2, 0.5) +
draw_plot_label(c("A", "B"), c(0.15, 0.15), c(1, 0.5))

```

```

# Determine the pike clusters in the harbour
# Convert to wide format such that animal ID ~ receiver group,
# aggregate by mean time spent
pike.receiver.wide <- dcast(detection.data,
                           `animal ID` ~ `receiver group`,
                           value.var = "time spent (secs)",
                           fun = mean) %>%
as.matrix(rownames = "animal ID")

```

```

# Replace missing values with 0, implying no detection
pike.receiver.wide[is.nan(pike.receiver.wide)] = 0

# Scale the data relative to each pike
pike.receiver.wide <- pike.receiver.wide %>%
  t() %>%
  scale() %>%
  t() %>%
  data.frame()

```

```

# Perform PCA on the detection data
pike.pca <- PCA(
  (pike.receiver.wide),
  scale.unit = F,

```

```
graph = F,  
ncp = length(pike.receiver.wide))
```

```
# Visualize the potential number of clusters
```

```
no.of.clusters <- fviz_nbclust(  
  pike.receiver.wide,  
  FUNcluster = hcut,  
  method = "wss",  
  linecolor = "black") +  
  labs(title = "",  
        x = "Number of clusters, k",  
        y = "Total Within Cluster Sum of Squares")
```

```
# Perform hierarchical clustering using several methods
```

```
# Put the methods in a list
```

```
hclust.method <- c("ward.D", "single", "complete", "average")
```

```
# Create an empty list to be populated
```

```
pike.dendlist <- dendlist()
```

```
# Iterate through the methods list and perform
```

```
# hierarchical clustering using the appropriate method
```

```
for(i in seq_along(hclust.method)) {  
  hclust.pike <- hclust(  
    dist(pike.receiver.wide),  
    method = hclust.method[i])  
  pike.dendlist <- dendlist(  
    pike.dendlist,  
    as.dendrogram(hclust.pike))  
}
```

```

}

names(pike.dendlist) <- hclust.method

# Return the populated list
pike.dendlist

# Plot a dendrogram of the different methods
#par(mfrow = c(2,2))

#allDendrograms <- for(i in 1:4) {
#   pike.dendlist[[i]] %>%
#     set("branches_k_color", k=2) %>%
#     plot(axes = FALSE, horiz = TRUE)
#   title(names(pike.dendlist)[i])
#}

# Get the correlation among the different methods
pike.dendlist.correlation <- cor.dendlist(
  pike.dendlist, method = "common")

# Perform hierarchical clustering and utilize ward method
pike.hclust <- pike.receiver.wide %>%
  dist() %>%
  hclust(method = "ward.D2") %>%
  as.dendrogram() %>%
  color_branches(k = 2, col = c("#7570B3", "#E7298A"))

# Cut the tree at 2 branches

```

```

pike.hclust.treeCut <- cutree(pike.hclust, k = 2)

# Get the number of pike within each branch
table(pike.hclust.treeCut)

# Covert each pike group to a factor
pike.groups <- levels(as.factor(pike.hclust.treeCut))

# Make plot of the hierarchical clustering dendrogram
# par(mfrow = c(1,1))
#hclustWardPikeClust <- plot(pike.hclust, horiz = T, leaflab = "none")

# Add legend
# hclustLegend <- legend("topleft",
#       legend = pike.groups,
#       fill = c("#7570B3", "#E7298A"),
#       bty = "n",
#       horiz = T,
#       title = "Northern pike clusters:")

# Select the pike and their clusters
pike.grps.hclust <- cbind(
  pike.receiver.wide,
  as.factor(pike.hclust.treeCut)) %>%
  data.table(keep.rownames = T) %>%
  setnames(old = c("rn", "as.factor(pike.hclust.treeCut)"),
           new = c("animal ID", "Northern pike cluster")) %>%
  dplyr::select(`animal ID`, `Northern pike cluster`)

```

```

# Merge the pike clusters to the data table
detection.data <- merge(
  detection.data, pike.grps.hclust, by = "animal ID")

# Merge receiver cluster and habitat cluster to the
# data table
detection.data <- merge(
  detection.data,
  receiver.cluster,
  by = "receiver group")

# Calculate time spent by groupings
overall.time.spent <- detection.data[
  , by = .(`Northern pike cluster`,
          `Receiver cluster`,
          `receiver group`),
  .(`time spent` = as.numeric(sum(`time spent (secs)`)))]

# Visuals for the receiver detections
receiver.activity <- overall.time.spent[
  , by = .(`Northern pike cluster`,
          `receiver group`,
          `Receiver cluster`),
  .(`Total time spent` = sum(`time spent`))] %>%
  .[, by = .(`Northern pike cluster`), ":= "
    (`Percent time spent` = (
      `Total time spent`/sum(`Total time spent`))*100)] %>%
  .[order(`Northern pike cluster`, `Total time spent`)]

```

```

# Bar plot of time spent at the receiver groups by
# cluster 1 pike
pikeOneActivity <- receiver.activity[`Northern pike cluster` == 1] %>%
  mutate(`receiver group` = fct_reorder(
    `receiver group`, -`Percent time spent`)) %>%
  ggplot(aes(
    y=`Percent time spent`, x=`receiver group`, fill = `Receiver cluster`)) +
  geom_bar(stat="identity") +
  geom_text(aes(
    label = signif(`Percent time spent`, 3)), hjust = -0.1, size = 3) +
  coord_flip() +
  xlab("") +
  scale_fill_manual(values = c("#D95F02", "#1B9E77")) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        legend.position = c(0.8, 0.8))

```

```

# Percent time spent at the receiver clusters by
# cluster 1 pike ----
time.spent.at.rec.cluster <- receiver.activity[
  `Northern pike cluster` == 1, by = .(`Receiver cluster`),
  .(`time spent` = sum(`Total time spent`))]

time.spent.at.rec.cluster[,
  `percent time spent` := `time spent`/sum(`time spent`)]

```

```

# Calculate mean monthly depths, %SAV, and
# stratification temperature
# based on pike activity
pike.harbour.use <- detection.data[
  , by = .(`Northern pike cluster`,
    lubridate::month(date, label = T)),
  .(`weighted mean depth` = wtd.mean(
    x = `Mean depth`,
    weights = as.numeric(`time spent (secs)`)),
  `weighted st. dev. depth` = sqrt(
    wtd.var(x = `Mean depth`,
      weights = as.numeric(`time spent (secs)`))),
  `weighted mean sav` = wtd.mean(x = `% SAV`,
    weights = as.numeric(
      `time spent (secs)`)),
  `weighted st. dev. sav` = sqrt(
    wtd.var(x = `% SAV`,
      weights = as.numeric(`time spent (secs)`))),
  `weighted mean strat` = wtd.mean(
    x = `Stratification temp.`,
    weights = as.numeric(`time spent (secs)`)),
  `weighted st. dev. strat` = sqrt(
    wtd.var(x = `Stratification temp.`,
      weights = as.numeric(
        `time spent (secs)`))))] %>%
setnames(old = "lubridate", new = "Month")

```

```

# Visualize monthly depth preference for cluster one pike
# Group 1 pike depth preference through time
grpOnePikeBehavDepth <- ggplot(
  data = pike.harbour.use[`Northern pike cluster` == 1],
  aes(x = Month)) +
  geom_bar(aes(y = `weighted mean depth`,
               stat = "identity", fill = "slategray1") +
  geom_errorbar(aes(
    ymin = `weighted mean depth` - `weighted st. dev. depth`,
    ymax = `weighted mean depth` + `weighted st. dev. depth`,
    width = 0.3, colour = "black", alpha = 0.5) +
  labs(y = "Weighted Mean Depth (m)") +
  theme_classic() +
  theme(axis.line.x = element_blank(),
        axis.line.y = element_blank(),
        axis.title.x = element_blank(),
        axis.text.x = element_text(angle = 90,
                                    vjust = 0.5, hjust=1))

```

```

# Group 1 pike SAV preference through time ----
grpOnePikeBehavSAV <- ggplot(data = pike.harbour.use[
  `Northern pike cluster` == 1], aes(x = Month)) +
  geom_bar(aes(y = `weighted mean sav`,
               stat = "identity", fill = "olivedrab") +
  geom_errorbar(aes(
    ymin = `weighted mean sav` - `weighted st. dev. sav`,
    ymax = `weighted mean sav` + `weighted st. dev. sav`,
    width = 0.3, colour = "black", alpha = 0.5) +
  labs(y = "Weighted mean % SAV") +

```



```

theme_classic() +
theme(axis.line.x = element_blank(),
      axis.line.y = element_blank(),
      axis.title.x = element_blank(),
      axis.text.x = element_text(angle = 90,
                                  vjust = 0.5, hjust=1))

```

```

# Group 1 pike strat preference through time ----
grpOnePikeBehavStrat <- ggplot(
  data = pike.harbour.use[`Northern pike cluster` == 1],
  aes(x = Month)) +
  geom_bar(aes(y = `weighted mean strat`),
           stat = "identity", fill = "gold") +
  geom_errorbar(
    aes(
      ymin = `weighted mean strat` -
            `weighted st. dev. strat`,
      ymax = `weighted mean strat` +
            `weighted st. dev. strat`),
    width = 0.3,
    colour = "black", alpha = 0.5) +
  labs(y = "Weighted mean Strat (celcius)") +
  theme_classic() +
  theme(axis.line.x = element_blank(),
        axis.line.y = element_blank(),
        axis.title.x = element_blank(),
        axis.text.x = element_text(
          angle = 90, vjust = 0.5, hjust=1))

```

```

# Temporal activity for cluster one pike
group.1.pike.temporal.summary <- detection.data[
  `Northern pike cluster` == 1,
  by = .(`Receiver cluster`,
    lubridate::month(date, label = T)),
  .(`Total time` = as.numeric(sum(`time spent (secs)`))) %>%
  setnames(old = "lubridate", new = "Month") %>%
  .[, by = Month, "!="
    (`Percent of monthly time` = (
      `Total time`/sum(`Total time`))*100)] %>%
  .[, `Percent of yearly time` := (
    `Total time`/sum(`Total time`))*100]

g1.percent.yearly.res <- ggplot(
  group.1.pike.temporal.summary,
  aes(x = Month,
    y = `Percent of yearly time`,
    group = `Receiver cluster`,
    color = `Receiver cluster`)) +
  geom_line() +
  geom_point() +
  scale_color_manual(values = c("#D95F02", "#1B9E77")) +
  labs(y = "% Time spent") +
  theme_classic() +
  theme(legend.position = "none",
    axis.text.x = element_text(
      angle = 90, vjust = 0.5, hjust=1))

```

```

g1.percent.monthly.res <- ggplot(
  group.1.pike.temporal.summary,
  aes(x = Month, y = `Percent of monthly time`,
      group = `Receiver cluster`,
      color = `Receiver cluster`)) +
  geom_line() +
  geom_point() +
  scale_color_manual(values = c("#D95F02", "#1B9E77")) +
  labs(y = "% Time spent") +
  theme_classic() +
  theme(legend.position = "none",
        axis.text.x = element_text(
          angle = 90, vjust = 0.5, hjust=1))

g1.temporal.rec.legend <- get_legend(
  g1.percent.monthly.res + theme(legend.position = "right"))

```

```

# Activity visualization ----
pikeOneBehavPlots <- ggdraw() +
  draw_plot(grpOnePikeBehavDepth, 0, 0.5, 0.3, 0.4) +
  draw_plot(grpOnePikeBehavSAV, 0.3, 0.5, 0.3, 0.4) +
  draw_plot(grpOnePikeBehavStrat, 0.6, 0.5, 0.3, 0.4) +
  draw_plot(g1.percent.yearly.res, 0, 0, 0.4, 0.4) +
  draw_plot(g1.percent.monthly.res, 0.4, 0, 0.4, 0.4) +
  draw_plot(g1.temporal.rec.legend, 0.8, 0.3, 0.1, 0.1) +
  draw_plot_label(c("A", "B", "C", "D", "E"),
                  c(0, 0.3, 0.6, 0, 0.4),
                  c(1, 1, 1, 0.5, 0.5))

```

```

# Bar plot of time spent at the receiver groups by
# cluster 2 pike ----
pikeTwoActivity <- receiver.activity[`Northern pike cluster` == 2] %>%
  mutate(`receiver group` = fct_reorder(
    `receiver group`, -`Percent time spent`)) %>%
  ggplot(aes(
    y=`Percent time spent`,
    x=`receiver group`,
    fill = `Receiver cluster`)) +
  geom_bar(stat="identity") +
  geom_text(aes(label = signif(
    `Percent time spent`, 3)),
    hjust = -0.1,
    size = 3) +
  coord_flip() +
  xlab("") +
  scale_fill_manual(
    values = c("#D95F02", "#1B9E77")) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        legend.position = c(0.8, 0.8))

```

```

# Percent time spent at the receiver clusters
#by cluster 2 pike ----
time.spent.at.rec.cluster <- receiver.activity[
  `Northern pike cluster` == 2, by = .(`Receiver cluster`),
  .(`time spent` = sum(`Total time spent`))]

```

```
time.spent.at.rec.cluster[
  , `percent time spent` := `time spent`/sum(`time spent`)]
```

```
# Visualize monthly depth preference for cluster two pike
# Group 2 pike depth preference through time
```

```
grpTwoPikeBehavDepth <- ggplot(data = pike.harbour.use[
  `Northern pike cluster` == 2], aes(x = Month)) +
  geom_bar(aes(y = `weighted mean depth`,
    stat = "identity", fill = "slategray1") +
  geom_errorbar(aes(ymin = `weighted mean depth` -
    `weighted st. dev. depth`,
    ymax = `weighted mean depth` +
    `weighted st. dev. depth`),
    width = 0.3,
    colour = "black",
    alpha = 0.5) +
  labs(y = "Weighted Mean Depth (m)") +
  theme_classic() +
  theme(axis.line.x = element_blank(),
    axis.line.y = element_blank(),
    axis.title.x = element_blank(),
    axis.text.x = element_text(
      angle = 90, vjust = 0.5, hjust=1))
```

```
# Group 2 pike SAV preference through time ----
```

```
grpTwoPikeBehavSAV <- ggplot(data = pike.harbour.use[
  `Northern pike cluster` == 2], aes(x = Month)) +
  geom_bar(aes(y = `weighted mean sav`,
    stat = "identity",
```

```

    fill = "olivedrab") +
geom_errorbar(aes(ymin = `weighted mean sav` -
                  `weighted st. dev. sav`,
                  ymax = `weighted mean sav` +
                  `weighted st. dev. sav`),
              width = 0.3, colour = "black", alpha = 0.5) +
labs(y = "Weighted mean % SAV") +
theme_classic() +
theme(axis.line.x = element_blank(),
      axis.line.y = element_blank(),
      axis.title.x = element_blank(),
      axis.text.x = element_text(
        angle = 90, vjust = 0.5, hjust=1))

```

Group 2 pike strat preference through time

```

grpTwoPikeBehavStrat <- ggplot(
  data = pike.harbour.use[`Northern pike cluster` == 2],
  aes(x = Month)) +
geom_bar(aes(y = `weighted mean strat`),
         stat = "identity", fill = "gold") +
geom_errorbar(aes(ymin = `weighted mean strat` -
                  `weighted st. dev. strat`,
                  ymax = `weighted mean strat` +
                  `weighted st. dev. strat`),
              width = 0.3, colour = "black", alpha = 0.5) +
labs(y = "Weighted mean Strat (celcius)") +
theme_classic() +
theme(axis.line.x = element_blank(),
      axis.line.y = element_blank(),

```

```
axis.title.x = element_blank(),
axis.text.x = element_text(
  angle = 90, vjust = 0.5, hjust=1))
```

Temporal activity for group 2 pikes

```
group.2.pike.temporal.summary <- detection.data[
  `Northern pike cluster` == 2,
  by = .(`Receiver cluster`,
    lubridate::month(date, label = T)),
  .(`Total time` = as.numeric(sum(`time spent (secs)`)))] %>%
setnames(old = "lubridate", new = "Month") %>%
.[, by = Month, ":="
  (`Percent of monthly time` = (
    `Total time`/sum(`Total time`))*100)] %>%
.[, `Percent of yearly time` := (
  `Total time`/sum(`Total time`))*100]
```

```
g2.percent.yearly.res <- ggplot(
  group.2.pike.temporal.summary,
  aes(x = Month,
    y = `Percent of yearly time`,
    group = `Receiver cluster`,
    color = `Receiver cluster`)) +
  geom_line() +
  geom_point() +
  scale_color_manual(values = c("#D95F02", "#1B9E77")) +
  labs(y = "% Time spent") +
  theme_classic() +
  theme(legend.position = "none",
```

```

axis.text.x = element_text(
  angle = 90, vjust = 0.5, hjust=1))

g2.percent.monthly.res <- ggplot(
  group.2.pike.temporal.summary,
  aes(x = Month, y = `Percent of monthly time`,
      group = `Receiver cluster`, color = `Receiver cluster`)) +
  geom_line() +
  geom_point() +
  scale_color_manual(values = c("#D95F02", "#1B9E77")) +
  labs(y = "% Time spent") +
  theme_classic() +
  theme(legend.position = "none",
        axis.text.x = element_text(
          angle = 90, vjust = 0.5, hjust=1))

g2.temporal.rec.legend <- get_legend(
  g2.percent.monthly.res + theme(legend.position = "right"))

pikeTwoBehav <- ggdraw() +
  draw_plot(grpTwoPikeBehavDepth, 0, 0.5, 0.3, 0.4) +
  draw_plot(grpTwoPikeBehavSAV, 0.3, 0.5, 0.3, 0.4) +
  draw_plot(grpTwoPikeBehavStrat, 0.6, 0.5, 0.3, 0.4) +
  draw_plot(g2.percent.yearly.res, 0, 0, 0.4, 0.4) +
  draw_plot(g2.percent.monthly.res, 0.4, 0, 0.4, 0.4) +
  draw_plot(g2.temporal.rec.legend, 0.8, 0.3, 0.1, 0.1) +
  draw_plot_label(c("A", "B", "C", "D", "E"),
    c(0, 0.3, 0.6, 0, 0.4),

```



```
c(1, 1, 1, 0.5, 0.5))
```

4.2. Code II

```
# App for analysis of acoustic telemetry data to explore fish behaviour

# Load libraries
library(factoextra)
library(FactoMineR)
library(lubridate)
library(tidyverse)
library(data.table)
library(leaflet)
library(leaflet.extras)
library(shinyRGL)
library(shinythemes)
library(bslib)

options(shiny.maxRequestSize = 5*1024^2)

ui <- fluidPage(

  # Select theme

  theme = bs_theme(bootswatch = "lux"),

  titlePanel(title = "Exploration of fish activity"),

  # Interface design...with sidebar and main panel
```

```

sidebarLayout(

  # Add sidebar
  sidebarPanel(
    fileInput(
      inputId = "detection.file", label = "Upload detection data"),
    hr(),
    helpText(
      "CSV FILE ONLY!",
      "Please summarize data by daily detections prior to upload
      ...it reduces the size
      of the data set and speeds up analysis."),
    hr(),
    helpText(
      "Currently only able to analyze data containing
      the following columns:
      \"animal ID\", \"date\", \"receiver group\",
      \"longitude\", \"latitude\", \"time spent (secs)\",
      hr(),
      h4("UPDATES COMING IN DUE TIME!"),
      hr(),
      tags$li(class = "dropdown",
        tags$a(
          href = "https://github.com/dijiagberien/ExpFishBehavApp",
          icon("github", "Source code", target = "_blank"))),
      tags$li(class = "dropdown",
        tags$a(href = "https://adogbejiagberien.netlify.app/",
          icon("globe", "Website", target = "_blank")))
  )
)

```

```

),

# Add main panel ----
mainPanel(tabsetPanel(

# Tab design ----
type = "tab",

# Map and basic summary tab ----
tabPanel("Map and basic summary",
  hr(),
  helpText(
    "Hover over icon to show pertaining info."),
  helpText(
    "Toggle icon at top right of map
    to hide receiver labels"),
  leafletOutput("data.map", height = 400),
  hr(),
  tableOutput("data.summary")),

# High level summary based on fish residency ----
tabPanel("Fish: PCA-3D plot and high level summary",
  tabsetPanel(
    type = "tab",
    tabPanel(
      "PCA", hr(),
      plotOutput("fishPCA", height = 700), hr()),
    tabPanel(

```

```

        "Summary", DT::dataTableOutput("fish.summary"))
    )
  ),

  # High level summary based on receiver detections ----
  tabPanel("Receivers: PCA-2D plot and high level summary",
    tabsetPanel(
      type = "tab",
      tabPanel(
        "PCA", hr(),
        plotOutput("receiverPCA", height = 700), hr()),
      tabPanel(
        "Summary", DT::dataTableOutput("receiver.summary"))
    ))
  ))
)
)

server <- function(input, output){

  # Import the detection data
  fish.detection.data <- reactive({
    fish.detection.file <- input$detection.file
    if (is.null(fish.detection.file)) {
      return()
    }
  })
}

```

```

fread(fish.detection.file$datapath)
}))

# Plot map of receivers and add metadata to receiver groups
output$data.map <- renderLeaflet({

  if (is.null(fish.detection.data())) {
    return()
  }

  receiver.data <- fish.detection.data()[
    , by = .(`receiver group`),
    .(`longitude` = mean(longitude),
      `latitude` = mean(latitude),
      `total time spent` = sum(`time spent (secs)`),
      `fish count` = uniqueN(`animal ID`),
      `earliest detection date` = min(date),
      `latest detection date` = max(date),
      `possible no. of days present` = difftime(max(date), min(date)),
      `no. of. days with detections` = uniqueN(date)
    )] %>%
    .[, c("percent time spent") := (
      `total time spent`/sum(`total time spent`)) * 100] %>%
    .[, !c("total time spent")]

  receiver.data$label <- paste(
    "<p>", "Receiver group: ", receiver.data$`receiver group`, "</p>",

```

```

"<p>", "Fish count: ", receiver.data$`fish count`, "</p>",
"<p>", "Percent use by tagged fish: ",
round(receiver.data$`percent time spent`, 2), "</p>",
"<p>", "Possible no. of days present: ",
receiver.data$`possible no. of days present`, "</p>",
"<p>", "No. of days with detections: ",
receiver.data$`no. of. days with detections`, "</p>",
"<p>", "Earliest detection date: ",
receiver.data$`earliest detection date`, "</p>",
"<p>", "Latest detection date: ",
receiver.data$`latest detection date`, "</p>")

leaflet() %>%
  addTiles() %>%
  setView(lng = mean(receiver.data$longitude),
          lat = mean(receiver.data$latitude),
          zoom = 14) %>%
  addMarkers(lng = receiver.data$longitude,
             lat = receiver.data$latitude,
             label = lapply(receiver.data$label, HTML),
             labelOptions = labelOptions(noHide = F),
             group = "Detailed info.") %>%
  addCircleMarkers(lng = receiver.data$longitude,
                   lat = receiver.data$latitude,
                   label = receiver.data$`receiver group`,
                   labelOptions = labelOptions(noHide = T),
                   group = "Receiver groups",
  ) %>%

```

```

addLayersControl(
  overlayGroups = c("Detailed info.", "Receiver groups"))
})

# Summary table of activity at the study location
output$data.summary <- renderTable({
  if (is.null(fish.detection.data())) {
    return()
  }

  detectionBriefSummary <- fish.detection.data()[
    , by = .(`receiver group`, year(date)),
    .(`No. of detected fish` = uniqueN(`animal ID`),
      `No. of receiver groups` = uniqueN(`receiver group`),
      `Total time spent` = sum(`time spent (secs)`))] %>%
    .[order(year, -`Total time spent`)]

  leastUtilizedLocations <- detectionBriefSummary %>%
    group_by(year) %>%
    slice(tail(row_number(), 3)) %>%
    select(`receiver group`) %>%
    data.table() %>%
    .[, by = .(`year`),
      .(`Least utilized receiver groups` =
        paste(`receiver group`, collapse = ", "))]

  mostUtilizedLocations <- detectionBriefSummary %>%
    group_by(year) %>%

```

```

slice(1:3) %>%
select(`receiver group`) %>%
data.table() %>%
.[, by = .(`year`),
  .(`Most utilized receiver groups` =
    paste(`receiver group`, collapse = ", "))]

detectionBriefSummary <- detectionBriefSummary[
  , by = .(year),
  .(`No. of detected fish` =
    sum(`No. of detected fish`),
    `No. of receiver groups` =
    sum(`No. of receiver groups`))]

detectionBriefSummary <- merge(
  detectionBriefSummary,
  mostUtilizedLocations, by = "year")

detectionBriefSummary <- merge(
  detectionBriefSummary,
  leastUtilizedLocations, by = "year")

detectionBriefSummary
})

# Clustering detected fish by the time spent at
# the different receiver groups ----

```



```

# Convert detection data to wide format
# animal ID ~ receiver group
fish.receiver.wide <- reactive({
  if (is.null(fish.detection.data())) {
    return()
  }
  fish.detection.data() %>%
    dcast(`animal ID` ~ `receiver group`,
          value.var = "time spent (secs)",
          fun = mean, fill = 0) %>%
    as.matrix(rownames = "animal ID") %>%
    t() %>% scale() %>% t() %>% data.frame()
})

```

```

# Perform hierarchical clustering
fish.groups <- reactive({
  if (is.null(fish.receiver.wide())) {
    return()
  }
  fish.groups.hclust <- HCPC(
    fish.receiver.wide(),
    nb.clust = 0,
    graph = F)$data.clust

  data.table(`fish cluster` =
    fish.groups.hclust$clust,
    `animal ID` =
    rownames(fish.groups.hclust))

```

```

}))

# Determine fish groups and return
# a 3-D PCA plot of fish ~ receiver ----
# Map not used because it may take time
# to render all points if data set is large
output$fishPCA <- renderPlot({
  if (is.null(fish.receiver.wide())) {
    return()
  }

  fish.pca <- PCA(
    (fish.receiver.wide()),
    scale.unit = F,
    graph = F,
    ncp = length(fish.receiver.wide()))

  fviz_pca_biplot(
    fish.pca, geom = "point", geom.var = c("text"),
    col.ind = fish.groups()$`fish cluster`, repel = T,
    col.var = "black")

}))

# Convert to side format receiver ~ month,
# calculate moving averages, and cluster receivers ----
receiver.month.wide <- reactive({

```

```

if (is.null(fish.detection.data())) {
  return()
}

# Convert to wide format receiver group ~ month
receiver.month.wide <- fish.detection.data() %>%
  dcast(`receiver group` ~
        lubridate::month(date, label = T),
        value.var = "time spent (secs)",
        fun = sum,
        fill = 0) %>%
  as.matrix(rownames = "receiver group") %>% t()

# Calculate moving average
ma <- function(x, n = 3){stats::filter(
  x, rep(1 / n, n), sides = 2, circular = T)}

receiver.month.wide[1:12, ] <- ma(receiver.month.wide[
  1:12, ], )

receiver.month.wide <- receiver.month.wide %>%
  scale() %>% t()
})

# Table of receiver groups and their cluster
receiver.groups <- reactive({
  if (is.null(receiver.month.wide())) {
    return()
  }

```

```

}

receiver.groups.hclust <- HCPC(
  data.frame(receiver.month.wide()),
  nb.clust = 0,
  graph = F)$data.clust

data.table(`receiver cluster` =
            receiver.groups.hclust$clust,
            `receiver group` =
            rownames(receiver.groups.hclust))
})

# PCA plot of receiver group ~ month
output$receiverPCA <- renderPlot({
  if (is.null(receiver.month.wide())) {
    return()
  }

  receiver.pca <- PCA(
    (receiver.month.wide()), scale.unit = F,
    graph = T, ncp = length(receiver.month.wide()))

  fviz_pca_biplot(receiver.pca, geom = "text",
                  col.ind = receiver.groups()$`receiver cluster`,
                  col.var = "black",
                  repel = T,
                  legend.title = "Receiver cluster",
                  ggtheme =

```

```

        theme_classic(
            base_size = 12,
            base_family = "Times New Roman"))

    })

detection.table.with.groups <- reactive({
    if (is.null(fish.detection.data())) {
        return()
    }
    fish.detection.data <- merge(
        fish.detection.data(), fish.groups(), by = "animal ID")
    fish.detection.data <- merge(
        fish.detection.data, receiver.groups(), by = "receiver group")
})

output$fish.summary <- DT::renderDataTable({
    if (is.null(detection.table.with.groups())) {
        return()
    }

    detection.table.with.groups()[
        ,
        by = .(`fish cluster`, `animal ID`),
        .(`no of days present` = uniqueN(date),
          `location count` = uniqueN(`receiver group`),
          `first detection` = as.character(min(date)),
          `last detection` = as.character(max(date)),

```

```

    `% days present` = round(
      100*(uniqueN(date)
        /as.numeric(
          difftime(max(date),
            min(date)) + 1)),
      2),
    `locations` = paste(
      unique(`receiver group`),
      collapse = ", ")] %>%
    .[order(`animal ID`)]
  })

output$receiver.summary <- DT::renderDataTable({
  if (is.null(detection.table.with.groups())) {
    return()
  }
  detection.table.with.groups()[
    ,by = .(`receiver cluster`, `receiver group`),
    .(`fish count` = uniqueN(`animal ID`),
      `detection count` = sum(`detection count`),
      `First detection` = as.character(min(date)),
      `Last detection` = as.character(max(date)),
      `date range` = difftime(max(date), min(date)),
      `Percentage use` = round(100 * (uniqueN(date)/
        as.numeric(difftime(max(date), min(date))))), 2))
  ] %>%
    .[order(-`Percentage use`)]

```

```
    })  
  }  
  
shinyApp(ui = ui, server = server)
```

Bibliography

- Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2021. *Rmarkdown: Dynamic Documents for r*. <https://github.com/rstudio/rmarkdown>.
- Barnes, Kaylin, Lyndsay Cartwright, Rick Portiss, Jon Midwood, Christine Boston, Monica Grana-dos, Thomas Sciscione, Colleen Gibson, and Olusola Obembe. 2020. “Evaluating the Toronto Waterfront Aquatic Habitat Restoration Strategy.” Technical {Report}. Toronto; Region Con-servation Authority.
- Block, Barbara A., Christopher M. Holbrook, Samantha E. Simmons, Kim N. Holland, Jerald S. Ault, Daniel P. Costa, Bruce R. Mate, et al. 2016. “Toward a National Animal Telemetry Network for Aquatic Observations in the United States.” *Animal Biotelemetry* 4 (1): 6. <https://doi.org/10.1186/s40317-015-0092-1>.
- Brownscombe, Jacob W., Lucas P. Griffin, Danielle Morley, Alejandro Acosta, John John, Susan K. Lowerre-Barbieri, Aaron J. Adams, Andy J. Danylchuk, and Steven J Cooke. 2020. “Ap-plication of Machine Learning Algorithms to Identify Cryptic Reproductive Habitats Using Diverse Information Sources.” *Oecologia* 194 (1-2): 283–98. <https://doi.org/10.1007/s00442-020-04753-2>.
- Brownscombe, Jacob W., Elodie J. I. Lédée, Graham D. Raby, Daniel P. Struthers, Lee F. G. Gutowsky, Vivian M. Nguyen, Nathan Young, et al. 2019. “Conducting and Interpreting Fish Telemetry Studies: Considerations for Researchers and Resource Managers.” *Reviews in Fish Biology and Fisheries* 29 (2): 369–400. <https://doi.org/10.1007/s11160-019-09560-4>.

- Bzdok, Danilo, Naomi Altman, and Martin Krzywinski. 2018. “Statistics Versus Machine Learning.” *Nature Methods* 15 (4): 233–34. <https://doi.org/10.1038/nmeth.4642>.
- Cagnacci, Francesca, Luigi Boitani, Roger A. Powell, and Mark S. Boyce. 2010. “Animal Ecology Meets GPS-Based Radiotelemetry: A Perfect Storm of Opportunities and Challenges.” *Philosophical Transactions of the Royal Society B: Biological Sciences* 365 (1550): 2157–62. <https://doi.org/10.1098/rstb.2010.0107>.
- Canada, Environment and Climate Change. 2012. “2012 Great Lakes Water Quality Agreement.” International treaties. *Aem*.
- Casselman, J. M. 1996. “Age, Growth and Environmental Requirements of Pike.” In *Pike: Biology and Exploitation*, edited by John F. Craig, 69–101. Fish and Fisheries Series. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-015-8775-4_4.
- Cook, Mark F., and Eric P. Bergersen. 1988. “Movements, Habitat Selection, and Activity Periods of Northern Pike in Eleven Mile Reservoir, Colorado.” *Transactions of the American Fisheries Society* 117 (5): 495–502. [https://doi.org/https://doi.org/10.1577/1548-8659\(1988\)117%3C0495:MHSAAP%3E2.3.CO;2](https://doi.org/https://doi.org/10.1577/1548-8659(1988)117%3C0495:MHSAAP%3E2.3.CO;2).
- Cooke, Steven J., Sara J. Iverson, Michael J. W. Stokesbury, Scott G. Hinch, Aaron T. Fisk, David L. VanderZwaag, Richard Apostle, and Fred Whoriskey. 2011. “Ocean Tracking Network Canada: A Network Approach to Addressing Critical Issues in Fisheries and Resource Management with Implications for Ocean Governance.” *Fisheries* 36 (12): 583–92. <https://doi.org/10.1080/03632415.2011.633464>.
- Cooke, Steven J., Jonathan D. Midwood, Jason D. Thiem, Peter Klimley, Martyn C. Lucas, Eva B. Thorstad, John Eiler, Chris Holbrook, and Brendan C. Ebner. 2013. “Tracking Animals in Freshwater with Electronic Tags: Past, Present and Future.” *Animal Biotelemetry* 1 (1): 5. <https://doi.org/10.1186/2050-3385-1-5>.
- Craig, J. F. 2008. “A Short Review of Pike Ecology.” *Hydrobiologia* 601 (1): 5–16. <https://doi.org/10.1007/s10750-007-9262-3>.

- Dowle, Matt, and Arun Srinivasan. 2020. *Data.table: Extension of 'Data.frame'*. <https://CRAN.R-project.org/package=data.table>.
- Environment and Climate Change Canada. 2015. “Great Lakes Water Quality Agreement.” Transparency - other. *Aem*.
- Ester, Martin, Hans-Peter Kriegel, and Xiaowei Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 226–31.
- Estes, Richard D. 2014. *The Gnu's World: Serengeti Wildebeest Ecology and Life History*. 1st edition. University of California Press.
- Gjelland, Karl Ø, and Richard D. Hedger. 2013. “Environmental Influence on Transmitter Detection Probability in Biotelemetry: Developing a General Model of Acoustic Transmission.” *Methods in Ecology and Evolution* 4 (7): 665–74. <https://doi.org/https://doi.org/10.1111/2041-210X.12057>.
- “Great Lakes: Areas of Concern.” 2007. Program results. *Aem*.
- Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. “Unsupervised Learning.” In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, edited by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 485–585. Springer Series in Statistics. New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7_14.
- Hockersmith, Eric, and John Beeman. 2012. “A History of Telemetry in Fishery Research.” In *Telemetry Techniques-A User Guide for Fisheries Research*, 7–19. American Fisheries Society.
- Holbrook, Christopher, Todd Hayden, Thomas Binder, and Jon Pye. 2019. *Glatos: A Package for the Great Lakes Acoustic Telemetry Observation System*. <https://gitlab.oceantrack.org/GreatLakes/glatos>.
- Hussey, Nigel E., Steven T. Kessel, Kim Aarestrup, Steven J. Cooke, Paul D. Cowley, Aaron T. Fisk, Robert G. Harcourt, et al. 2015. “Aquatic Animal Telemetry: A Panoramic Window into the Underwater World.” *Science* 348 (6240): 1–10. <https://doi.org/10.1126/science.1255642>.

- James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York. <https://doi.org/10.1007/978-1-4614-7138-7>.
- Kassambara, Alboukadel, and Fabian Mundt. 2020. *Factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. <https://CRAN.R-project.org/package=factoextra>.
- Kobler, A., T. Klefoth, C. Wolter, F. Fredrich, and R. Arlinghaus. 2008. “Contrasting Pike (Esox Lucius L.) Movement and Habitat Choice Between Summer and Winter in a Small Lake.” *Hydrobiologia* 601 (1): 17. <https://doi.org/10.1007/s10750-007-9263-2>.
- Kobler, Alexander., Thomas Klefoth, Thomas Mehner, and Robert Arlinghaus. 2009. “Coexistence of Behavioural Types in an Aquatic Top Predator: A Response to Resource Limitation?” *Oecologia* 161 (4): 837–47. <https://doi.org/10.1007/s00442-009-1415-9>.
- Krueger, Charles C., Christopher M. Holbrook, Thomas R. Binder, Christopher S. Vandergoot, Todd A. Hayden, Darryl W. Hondorp, Nancy Nate, et al. 2018. “Acoustic Telemetry Observation Systems: Challenges Encountered and Overcome in the Laurentian Great Lakes1.” *Canadian Journal of Fisheries and Aquatic Sciences* 75 (10): 1755–63. <https://doi.org/10.1139/cjfas-2017-0406>.
- Lynch, Abigail J., Steven J. Cooke, Andrew M. Deines, Shannon D. Bower, David B. Bunnell, Ian G. Cowx, Vivian M. Nguyen, et al. 2016. “The Social, Economic, and Environmental Importance of Inland Fish and Fisheries.” *Environmental Reviews* 24 (2): 115–21. <https://doi.org/10.1139/er-2015-0064>.
- Marshall, S. J. 2013. “Hydrology.” In *Reference Module in Earth Systems and Environmental Sciences*. Elsevier. <https://doi.org/10.1016/B978-0-12-409548-9.05356-2>.
- Midwood, Jonathan D., Lee F. G. Gutowsky, Bogdan Hlevca, Rick Portiss, Mathew G. Wells, Susan E. Doka, and Steven J. Cooke. 2018. “Tracking Bowfin with Acoustic Telemetry: Insight into the Ecology of a Living Fossil.” *Ecology of Freshwater Fish* 27 (1): 225–36. <https://doi.org/https://doi.org/10.1111/eff.12340>.

- Midwood, Jonathan D., Andrew M Rous, Susan Elisabeth Doka, and Stephen J Cooke. 2019. “Acoustic Telemetry in Toronto Harbour: Assessing Residency, Habitat Selection, and Within-Harbour Movements of Fishes over a Five-Year Period.” *Canadian Technical Report of Fisheries and Aquatic Sciences*, 3331: xx + 174 p.
- Miller, Harvey J., Somayeh Dodge, Jennifer Miller, and Gil Bohrer. 2019. “Towards an Integrated Science of Movement: Converging Research on Animal Movement Ecology and Human Mobility Science.” *International Journal of Geographical Information Science* 33 (5): 855–76. <https://doi.org/10.1080/13658816.2018.1564317>.
- Paukert, Craig P., and David W. Willis. 2003. “Population Characteristics and Ecological Role of Northern Pike in Shallow Natural Lakes in Nebraska.” *North American Journal of Fisheries Management* 23 (1): 313–22. [https://doi.org/https://doi.org/10.1577/1548-8675\(2003\)023%3C0313:PCAERO%3E2.0.CO;2](https://doi.org/https://doi.org/10.1577/1548-8675(2003)023%3C0313:PCAERO%3E2.0.CO;2).
- Pierce, Rodney B. 2012. *Northern Pike: Ecology, Conservation, and Management History*. Minneapolis: University of Minnesota Press.
- Pratt, Thomas, and Karen Smokorowski. 2011. “Fish Habitat Management Implications of the Summer Habitat Use by Littoral Fishes in a North Temperate, Mesotrophic Lake.” *Canadian Journal of Fisheries and Aquatic Sciences* 60 (April): 286–300. <https://doi.org/10.1139/f03-022>.
- QGIS Development Team. 2021. *QGIS Geographic Information System*. Open Source Geospatial Foundation. <http://qgis.osgeo.org>.
- R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Rogers, Kevin B, and Gary C White. 2007. “Analysis and Interpretation of Freshwater Fisheries Data.” In *Analysis of Movement and Habitat Use from Telemetry Data*, 625–76. Bethesda, Maryland: American Fisheries Society.
- Simpfendorfer, Colin A., Charlie Huveneers, Andre Steckenreuter, Katherine Tattersall, Xavier

- Hoenner, Rob Harcourt, and Michelle R. Heupel. 2015. “Ghosts in the Data: False Detections in VEMCO Pulse Position Modulation Acoustic Telemetry Monitoring Equipment.” *Animal Biotelemetry* 3 (1): 55. <https://doi.org/10.1186/s40317-015-0094-z>.
- Thomas, Binder, Todd Hayden, and Christopher M. Holbrook. 2018. “An Introduction to R for Analyzing Acoustic Telemetry Data.” Great Lakes Acoustic Telemetry Observation System.
- Valletta, John Joseph, Colin Torney, Michael Kings, Alex Thornton, and Joah Madden. 2017. “Applications of Machine Learning in Animal Behaviour Studies.” *Animal Behaviour* 124 (February): 203–20. <https://doi.org/10.1016/j.anbehav.2016.12.005>.
- Veilleux, M. A. N., J. D. Midwood, C. M. Boston, N. W. R. Lapointe, R. Portiss, M. Wells, S. E. Doka, and S. J. Cooke. 2018. “Assessing Occupancy of Freshwater Fishes in Urban Boat Slips of Toronto Harbour.” *Aquatic Ecosystem Health & Management* 21 (3): 331–41. <https://doi.org/10.1080/14634988.2018.1507530>.
- Whoriskey, Kim, Eduardo G. Martins, Marie Auger-Méthé, Lee F. G. Gutowsky, Robert J. Lennox, Steven J. Cooke, Michael Power, and Joanna Mills Flemming. 2019. “Current and Emerging Statistical Techniques for Aquatic Telemetry Data: A Guide to Analysing Spatially Discrete Animal Detections.” *Methods in Ecology and Evolution* 10 (7): 935–48. <https://doi.org/https://doi.org/10.1111/2041-210X.13188>.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemond, et al. 2019. “Welcome to the Tidyverse.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemond. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.
- Xie, Yihui, Christophe Dervieux, and Emily Riederer. 2020. *R Markdown Cookbook*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown-cookbook>.