Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

Even Åge Smedshaug

Machine learning for identification of individual salmon behaviour in aquaculture

Master's thesis in Cybernetics and Robotics
Supervisor: Martin Føre
July 2023

NTNU
Norwegian University of
Science and Technology

Even Åge Smedshaug

Machine learning for identification of individual salmon behaviour in aquaculture

Master's thesis in Cybernetics and Robotics
Supervisor: Martin Føre
July 2023

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Engineering Cybernetics

NTNU
Norwegian University of
Science and Technology

Preface

This project is the culmination of five years of education in Cybernetics and Robotics with a specialization in biomedical cybernetics at the Norwegian University of Science and Technology (NTNU).
I would like to thank my supervisor Martin Føre for good conversations, both on and off topic, in addition to believing in the methods when results were not forthcoming. I would also like to thank my friends and family for putting up with increasingly fish-themed conversations.

Trondheim, July 2023
Even Åge Smedshaug

Abstract

The modern aquaculture industry suffers from low observability and control of conditions in the aquaculture cages. In order to better understand fish behaviour in both normal and stressful conditions, more knowledge of how fish behave is needed. The purpose of this project was to use machine learning methods to identify different modes of individual salmon behaviour with positional data from acoustic telemetry tags. Positional data from six different fish from two different cages at two different times of year was processed in order to create discrete fish swimming trajectories. Additional variables were calculated based on positional data: average depth, depth difference per second, track length per second, angle change per second, average distance from cage center per second and distance moved in relation to cage center. Every trajectory had one value for each of these variables. These trajectories were analysed based on both traditional methods and principal component analysis, in addition to being clustered with the HDBScan algorithm. In general, fish were more active at day and swam closer to the surface at night, and this was the case for every fish except one. Average depth distribution was the variable that differed most betweeen individuals. The variables that differed the least between individuals were the two variables based on distance from center, and these did not contribute to the clustering. Correlation structure at night was somewhat similar for most fish, as the variables average depth, depth difference, and track length were more correlated at night. Multiple modes of behaviour were detected, including feeding, circular swimming, and a third mode consisting of idle, non-circular swimming. For most fish, circular swimming was the most prevalent behavioural mode in the day, while short, high angle change, non-circular, idle trajectories close to the surface was the dominating swimming pattern at night. The results show that salmon behaviour has definable modes that can be detected from positional data, and that the prevalence of these modes differs from night to day. Moreover, the results show the potential of the application of machine learning methods in aquaculture.

Table of Contents

List of Tables

List of Figures

1

Introduction

1.1 Aquaculture

Aquaculture is, by definition, the 'breeding, reading, and harvesting of fish, shellfish, algae, and other organisms in all types of water environments'(US Department of Commerce and Administration (2019)). Aquaculture is a billion dollar industry in Norway (Misund (2023)), and Norway is by far the worlds largest producer of farmed Salmon. The norwegian aquaculture industry is in constant development, and the norwegian government has introduced a new class of aquaculture licences in order to facilitate the development of new technology in the industry. In turn, the norwegian aquaculture industry is developing both large offshore structures in the open ocean and closed enclosures in sheltered fjord areas (Føre et al. (2022)).

However, the aquaculture industry faces several difficulties in comparison to regular farming, several of which are associated with the fish and fish welfare. In addition the ecological dangers of escaped salmon (Hindar et al. (2006), Jensen et al. (2010)) and the destructive impact of salmon louse (Torrissen et al. (2013), Overton et al. (2019)), the poor observability of fish in the water, coupled with the sheer amount of fish in a cage, make the state of the fish extremely hard to discern for the farmer with traditional methods (e.g. visual inspection) (Føre et al. (2018a)). Whereas the traditional cattle farmer can know and measure each and every animal discretely, the aquaculture farmer cannot know or interact with their animals in the same way. Furthermore, the industrial trend of moving the farming offshore to the open sea is likely do amplify current challenges in the industry (Bjelland et al. (2015)). Therefore, there is an increasing drive to tackle these challenges with new production systems.

1.1.1 Precision Fish Farming

By utilising both pre-existing technology and modern advancements in measurement technology, computer vision, data analysis and wireless communication, the farmer can obtain a better understanding of the fish population in the cage. Variables such as feeding habits

(Føre et al. (2023)), health and welfare status (Noble et al. (2018)), lice count (Thorvaldsen et al. (2019)) and so forth gain increasingly accurate measurements as a byproduct of increasing technological intervention. By leveraging this increase in information, the farmer can make more precise decisions about feeding and delousing, and thereby may increase fish well-being, reduce antibiotics use, and liberate the fish from the dangers and stressors of excessive delousing. This is the main philosophy behind Precision Fish Farming, as described in Føre et al. (2018a).

The roots of Precision Fish Farming lie in Precision Livestock Farming. Precision livestock farming is, at its core, based on three principles (Berckmans et al. (2006)):

> 1) Animal variables (i.e. parameters related to the behavioural or physiological state of the animal) need to be measured continuously with cost-effective robust sensor technology,
>
> 2) a reliable model for predicting (expectation of) how animal variables will dynamically vary in response to external factors at any moment must be available, and
>
> 3) predictions and on-line measurements are integrated in an analysing algorithm for automatic monitoring and/or control

The difficulty of measuring the thousands of fish in an aquaculture cage, or even deciding what variables to measure, in addition to the lack of models for prediction of fish behaviour in accordance to external factors, illuminate how non-trivial the transition from terrestrial to aquatic precision farming is. However, by leveraging the increased observability gained from more precise measurements, technology enables the aquaculture industry to pivot from experience based to knowledge based production (Føre et al. (2018a)). Adapting the core principles of Precision Livestock Farming to Precision Fish Farming, the main aims of the latter are as follows (Føre et al. (2018a)):

> 1) improve accuracy, precision and repeatability in farming operations;
> 2) facilitate more autonomous and continuous biomass/animal monitoring;
> 3) provide more reliable decision support and;
> 4) reduce dependencies on manual labour and subjective assessments, and thus improve staff safety.

With more advanced monitoring and analysis techniques, the decision support system around aquaculture may be improved in order to facilitate increased fish welfare, lower food waste and less delousing which in turn increases operating efficiency.

In order to observe the fish population in the cage, the most commonly used methods are machine vision and sonar, which are less costly and time consuming than traditional sampling. Daoliang and Du (2022) details machine vision use in areas such as fish counting, behaviour recognition and biomass estimation in cages, among others. The benefits of machine vision includes non-invasivity, repeatability and relative objectivity. However, machine vision suffers from low observability in the water and very high data requirements for proper learning, in addition to the amount of expertise required for implementation (Saberioon et al. (2017)). Another tool for observing biomass and fish distribution is sonar (Ulvund et al. (2021), Johansson et al. (2006), Måløy (2020)), however, this method does not provide information on individual fish. To get the full picture of behaviour in cages, data from individual fish over time is needed.

1.2 Individual fish monitoring and Telemetry

Analysis of individual fish behaviour may be a key part in understanding the behaviour of an entire fish population. In order to identify the parameters necessary for understanding the state of the animal, in addition to creating a model for its behaviour, individual measurements can provide useful insight into how exactly individual fish operate. This knowledge can in turn be applied to the population as a whole. By aggregating measurements and reactions of different fish to the same events (e.g. delousing, feeding), the behaviour of an entire population may be modeled with greater accuracy. In turn, the fish farmer may make better decisions with regard to the operation of the fish farm, based not only on experience, but also scientific knowledge (Føre et al. (2018a)). Føre et al. (2017) shows that acoustic telemetry is a viable tool for real time monitoring of individual fish.

Telemetry is the ”automatic transfer of scientific data or other measureable quantities over larger distances by telecommunication” (Thorstad et al. (2013)). In the case of the aquaculture industry, this telecommunication is mostly done with acoustic signals, as radio waves are absorbed by saltwater. Acoustic telemetry has been used to measure heart rate and swimming activity in individual salmon by implanting them with acoustic tags with pressure sensors and accelerometers (Føre et al. (2018b)). Additionally, by implanting individual fish with acoustic tags and setting up multiple hydrophones around the cage, regular measurements of fish position can be calculated by using the time difference of arrival of the acoustic message to the different hydrophones, as described in Hassan et al. (2019).

While there are multiple studies on behaviour of entire salmon populations (Oppedal et al. (2011), Ulvund et al. (2021), Oppedal et al. (2007), Føre et al. (2011)) in response to environmental factors, and some studies on the behaviour on individual fish (Stockwell et al. (2021), Føre et al. (2017)), there are to the authors knowledge not many studies aiming to define what constitues ”standard” behaviour under regular conditions. The extraction of ”normal behaviours” is further complicated by the lack of knowledge around what variables are important for categorizing fish behaviour. This complicates the problem of quantifying fish behaviour and their reaction to external factors. However, the field of machine learning includes algorithms adept at discovering structure in data not readily apparent to the human observer, and may be an underutilised tool in the field of aquaculture.

1.3 Machine and Statistical learning

> Machine learning is a branch of Artificial Intelligence and computer science which focuses on the use of data and algorithms to imitate the way humans learn, gradually increasing its accuracy (IBM (2023a))

The field of machine learning is extensive and under constant development. It encompasses algorithms from the simplest linear regression to the most complicated AI language models. In essence, it includes all algorithms that seek to, in some way or form, to learn from data. Neural networks, clustering algorithms, decision trees and many computer vision algorithms are all part of the machine learning family (IBM (2023a), last accessed 22.06.2022). The growth of machine learning in both academia and public life is in part a result of increasing computer power, as many machine learning algorithms benefit greatly from large datasets and long runtimes. Machine learning algorithms also have the added benefit that they are adept at discovering structure in data. This structure might be difficult for a human to discern, or the dataset may be too large to be feasible analysed by a human. A data scientist can use multiple different approaches, ranging from more explainable algorithms like PCA and linear classifiers, to black box models like deep neural networks. In other words, machine learning covers multiple algorithms a person can use in order to for example understand data better, make predictions based on data, or classify datapoints.

In the field of aquaculture, machine learning is, to the authors knowledge, a relatively underutilized tool outside of machine vision (Saberioon et al. (2017)) and processing of sonar data (Måløy (2020)). The benefit is, as previously stated, that good models for fish behaviour in sea cages are lacking both in engineering and biological disciplines. To make up for this relative lack of field knowledge, the usage of machine and statistical learning techniques can cover the gap in knowledge and find patterns in data that are hard to uncover with traditional a priori modeling. While there is no guarantee that any method will yield positive or even interpretable result, as machine learning is by no means a silver bullet, the application of previously unused machine and statistical learning models models may yield increased understanding of fish behaviour and welfare in aquaculture. Machine learning techniques could be applied to acoustic telemetry data of individual fish in order to further our understanding of fish behaviour and conditions in the cage, and may ultimately be used for further behavioural modeling and decision support for the aquaculture industry.

1.4 Scope of the Project

This project aims to use machine learning techniques to qualitatively classify fish behaviour. The goal is not to create a deterministic model for fish behaviour, but rather use data exploration and unsupervised learning approaches to identify structure and distribution of data, and classify some behavioural modes, in addition to comparing these across different fish in order to identify similarities in behaviour between individuals. While there have been multiple studies on salmon behaviour as a result of environmental factors (Johansson et al. (2006), Ulvund et al. (2021), Stockwell et al. (2021), Oppedal et al. (2011), and many more), these studies use measurements of some end variable, usually depth, to quantify behavioural response to other measured external factors. This project differs from the literature on the subject as the goal is to use only positional data, divided into night and day, in order to classify different behavioural modes. The approach is qualitative, and the goal is to discover existing trends and structures in position data from farmed salmon, if any.

2

Theory

In order for the reader to better understand the machine learning methods used in this project and their respective challenges, this section will aim at giving an overview of classification in general, both supervised and unsupervised. A more in-depth explanation of cluster analysis, which is a type of unsupervised learning, will also be given, in addition to an in-depth explanation of some clustering algorithms. An introduction to PCA as an analysis tool for correlation structure in the dataset will also be given.

2.1 Classification

According to Gordon (1999), the subject of classification is concerned with the investigation of the relationships within a set of 'objects' in order to establish wether or not the data can be validly summarized by a small number of classes (or clusters) of similar objects. In data science, it encompassed the act of taking some data sample and mapping it to a *class*, or category, by some criteria. Mathematically, a classifier is a function that maps an input, x_i to a discrete class c in a set of classes C. This is the property that separates classification problems from regression problems. In regression, the target variable of the classifier function is some continous value, while in classification it is discrete. The binary classifier is the simplest example of a classifier, and has only two classes. Classification can be further divided into supervised or unsupervised learning according to the properties of the input data.

2.1.1 Supervised Classification

Supervised Classification is the branch of classification that concerns itself with training a model based on some ground truth. For every input x_i there is some value y_i, or class c_i that corresponds to it. The classifier then attempts to teach itself how to divide between classes, and often uses some sort of objective function in order to evaluate its result and improve. In binary classification, the log loss function is often used.

According to IBM (2023b) (last accessed 22.06.2023), some of the most commonly used supervised learning algorithms are:

- **Neural Networks**: This includes shallow networks, deep learning networks, convolutional networks, recurrent neural networks, and every combination thereof.
- **Naive Bayes**: Based on the principle of conditional independence between classes. Concerns itself with the equation:

$$P(\text{class}|\text{data}) = \frac{P(\text{data}|\text{class})P(\text{class})}{P(\text{data})}$$

- **Logistic Regression**: Estimates the probability of a sample being in a class. Similar to linear regression, but used for classification.
- **Support Vector Machines**: Constructs a hyperplane in n-dimensional space where the distance between the two classes is at its maximum. This plane is called the *decision boundary*, and separates the classes.
- **K nearest neighbour**: A non-parametric algorithm that classifies new datapoints according to their proximity to known datapoints.
- **Random forest**: Classification based on aggregation of multiple uncorrelated decision trees.

There are many ways to classify, and it is often beneficial to try multiple different algorithms on the same dataset in order to identify which ones yield the best results in any particular case.

2.1.2 Unsupervised Classification

Unsupervised classification differs from its supervised counterpart by the lack of ground truth. That is, whereas in supervised classification each datapoint x_i corresponds to a label or class y_i, in unsupervised classification there are no labels. In a sense, unsupervised classification is concerned with conditional densities and usually minimises some distance parameter according to chosen cluster size or amount (Haste et al. (2008), p. 185-186). As Haste et al. (2008) put it:

> With supervised learning there is a clear measure of success [...]. In the context of unsupervised learning, there is no such direct measure of success. It is difficult to ascertain the validity of inferences drawn from the output of most unsupervised learning algorithms. One must resort to heuristic arguments not only for motivating the algorithms, as is often the case in supervised learning as well, but also for judgments as to the quality of the results. This uncomfortable situation has led to heavy proliferation of proposed methods, since effectiveness is a matter of opinion and cannot be verified directly.

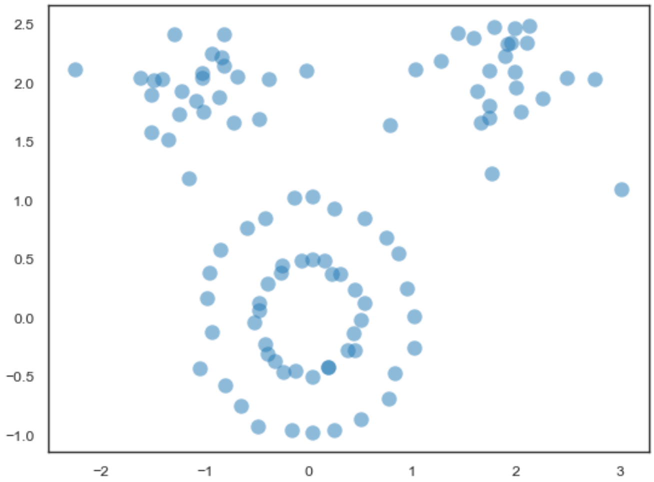


Figure 2.1: Example of two dimensional clustering problem with 4 clusters. Two cluster are present in the top half of the plot, while the circles represent a cluster each.

The lack of quantifiable results place greater emphasis on the interpretation of results and discussion thereof. In many ways, unsupervised learning is more exploratory than its supervised counterpart. However, even if specific performance measures are lacking, unsupervised classification is still an important part of machine learning and includes some powerful algorithms.

2.2 Unsupervised learning - cluster analysis

Cluster analysis, or data segmentation, aims to group or segment a collection of objects or datapoints into subsets or clusters, such that those within the cluster are more closely related than the objects assigned to different clusters (Haste et al. (2008), p.501). Cluster analysis is a type of unsupervised classification. Some types of cluster analysis, like k-means, assign every point to a cluster, while others, like DBScan, also classifies points as noise points. An example of a two dimensional clustering problem with two 'blobs' and two circles can be seen in figure 2.1. This clustering is done based on some measure of similarity or dissimilarity, not unlike the cost function in supervised classification. This measure of dissimilarity is often called the *distance measure*, and is defined between points. In short, the distance measure is a function of two points $X = [x_1, x_2, ..., x_n]^T$ and $Y = [y_1, y_2, ..., y_n]^T$ that maps to a scalar: Distance$(X, Y) : \mathbb{R}^n \longrightarrow \mathbb{R}$. This measure of similarity defines the clustering, and can be chosen based on problem structure. Popular distance measures include:

- **Euclidian**: Traditional distance, also called the L2 norm. Involves taking the square root of the sum of squared distances across every direction. N is the length of the vectors.

$$||X - Y|| = \sqrt{\sum_{i=0}^{N}(x_i - y_i)^2}$$

- **Manhattan**: Also called the L1 norm. The sum of absolute value across the distance vector:

$$|X - Y| = \sum_{i=0}^{N} |x_i - y_i|$$

- **Hamming**: A norm used for binary vectors. Counts the number of positions in which the two vectors are different:

$$\text{Hamming}(X, Y) = \sum_{i=0}^{N} x_i \oplus y_i$$

While norms like the euclidian norm work well in lower dimensions, with increasing sample size the distance between points grows larger and larger, while the distance from a point to its nearest and farthest neighbour grows similar. In addition, the amount of datapoints that need to be accessed and calculated grows exponentially as input size increases. This is sometimes referred to as the curse of dimensionality (Chen (2009)). Additionaly, cluster analysis is divided into parametric and non-parametric algorithms. Parametric algorithms typically optimize some cost function, and assumes more detailed knowledge of clusters, e.g. their distributions, while non-parametric methods are concerned with the density of data objects. As this project does not assume any prior knowledge related to how clusters are distributed, the focus will be on non-parametric cluster analysis.

2.2.1 Non-parametric cluster analysis - K-means

The perhaps most straight-forward, and one of the most popular (Haste et al. (2008), p.509), iterative descent clustering methods is the K-means algorithm. It is non-parametric, because it is concerned with the density of data, and not fitting clusters to any prior distribution. It is intended for use when all variables are quantitative (not categorical). The algorithm uses the Euclidian norm, and is concerned with minimizing the equation:

$$\sum_{k=1}^{K} N_k \sum_{i \in k} ||x_i - \overline{x}_k||^2$$

where N_k is the number of points in cluster k, K is the number of clusters, and $\overline{x_i}$ is the mean vector of the k'th cluster. In other terms, the k-means algorithm is concerned with minimizing the average dissimililarity within a cluster from the observed cluster mean (Haste et al. (2008), p. 509). The K-means algorithm requires only one input, which is the amount of clusters suspected to be present in the data.

The k-means algorithm is usually implemented by first choosing n different cluster centers at random, them assigning each data point to the closest cluster center. Then the

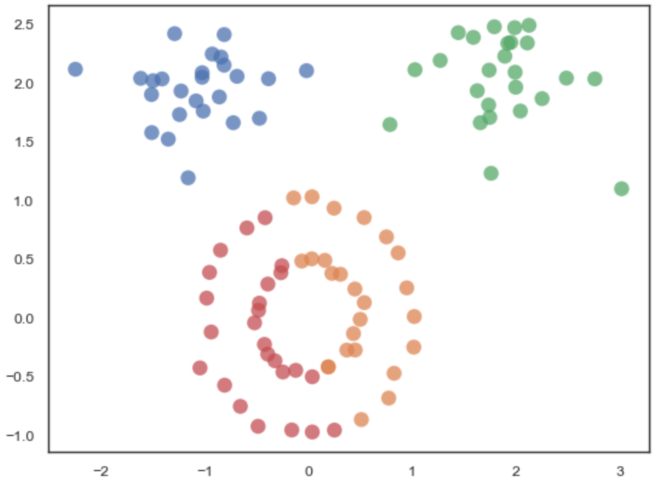


Figure 2.2: K-means clustering with n=4 clusters. Notice non-satisfactory clustering of bottom circular clusters.

means of each cluster is computed, and each point is assigned to its closest cluster mean. This process is repeated until the point assignments stop changing.

While k-means clustering is intuitive, it is not always the correct choice for the problem. The simplicity of the algorithm is also its drawback, as it cannot capture non-spherical clusters. This is illustrated in figure 2.2, where the K-means algorithm correctly classifies the two top clusters, but divides the bottom two circular clusters, as it cannot capture the circular nature of these clusters.

2.2.2 Density-based Clustering - DBScan

DBScan is an algorithm that view clusters as areas of high density separated by ares of low density. The algorithm separates clusters globally based on a density treshold, and divides datapoints into core and non-core samples. As a result, DBScan can discover clusters of varying sizes and shapes, while k-means assumes all clusters to be convex shaped (Pedregosa et al. (2011)). The two parameters of the algorithm are `min_samples` and `eps`, which signify the minimum amount of datapoints in a cluster, and the maximum distance between points in the same cluster, respectively. A point that has `min_samples` other points within `eps` range of itself is a *core point*. Any point that is withing `eps` range of a core point but does not have `min_samples` other points in its vicinity is a non-core point, and any point that lies further thatn `eps` range from a core point is classified as noise. The `eps` and `min_samples` parameters must be specified according to the input data, as illustrated in figures 2.3 and 2.4, in order to achieve a satisfactory result. The light blue color represent noise points, while the other colors represent clusters. Figure 2.3 has too high `eps` value, meaning that clusters that are ”far apart” can be classified as the same cluster. This results in three clusters being classified as the same cluster. in figure 2.4, however, a lower `eps` value and somewhat lower `min_samples` value results in all four

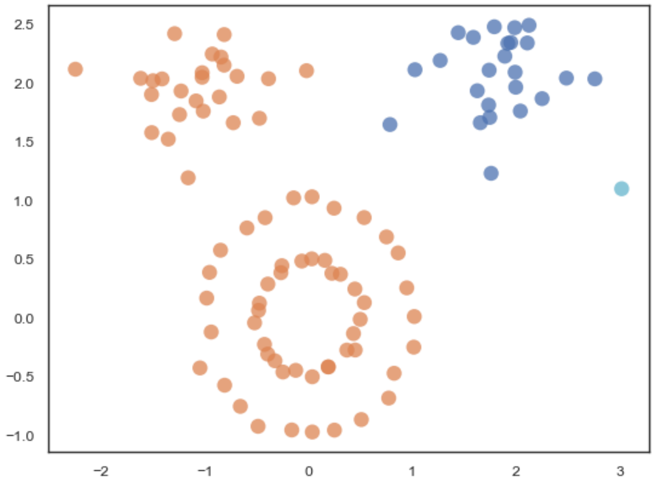


Figure 2.3: DBScan clustering with eps = .7 and min_samples = 5, Only 2 out of 4 clusters found.

clusters being correctly classified, minus some noise points in light blue. Notice how this algorithm can find circular clusters, unlike K-means, which can only find convex clusters.

2.2.3 Hierarchical Density-Based clustering - HDBScan

Campello et al. (2013) propose a density based clustering algorithm called HDBScan in their 2013 paper. This algorithm constitues a hierarchical algorithm based on an algorithm very similar to DBSCAN. The idea is that instead of specifying a global parameter for cluster density, cluster density may vary across the dataset. Many of the core concepts from DBScan are also present in HDBScan, including core objects, `min_samples` and `eps`-reachability, however, HDBScan has no concept of non-core objects, only core objects and noise.

In order to formulate the HDBScan algorithm, these concepts are defined (Campello et al. (2013)):

Let $\mathbf{X} = x_1, ..., x_n$ be a data set of n objects, and let $d(x_i, x_j)$ denote the distance measure of choice between two objects in $\mathbf{X}$. Let $m_{pts} \in \mathbb{Z} \cup [1, \infty]$ denote the single input parameter to the algorithm which signifies the minimum amount of points needed to define a cluster.

- **Core Distance**: An object $x_i \in \mathbf{X}$ w.r.t m_{pts} , $d_{core}(x_i)$ is the distance from x_i to its m_{pts} - nearest neighbour (including x_i).
- ε **-Core Object**: An object $x_i \in \mathbf{X}$ is a ε-core object for every value of ε that is greater than or equal to the core distance of x_p w.r.t m_{pts}, i.e., if $d_{core}(x_i) \leq \varepsilon$.
- **Mutual Reachability Distance**: The mutual reachability distance between jo objects $x_i \in \mathbf{X}$ and $x_j \in \mathbf{X}$ w.r.t. m_{pts} is defined as $d_{mreach}(x_i, x_j) = \max\{d_{core}(x_i), d_{core}(x_j), d(x_i, x_j)\}$.

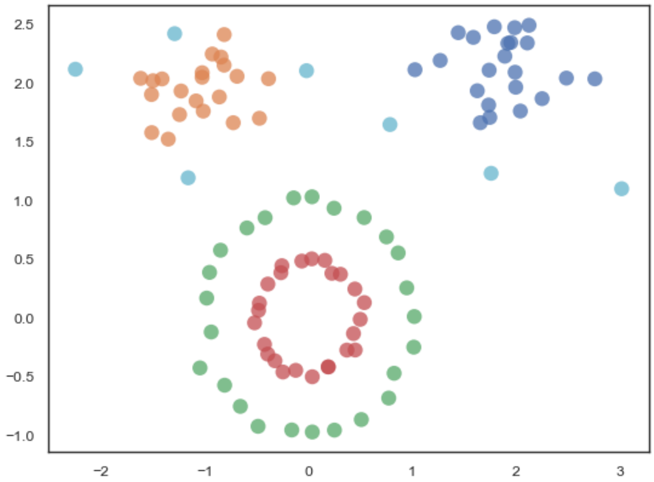
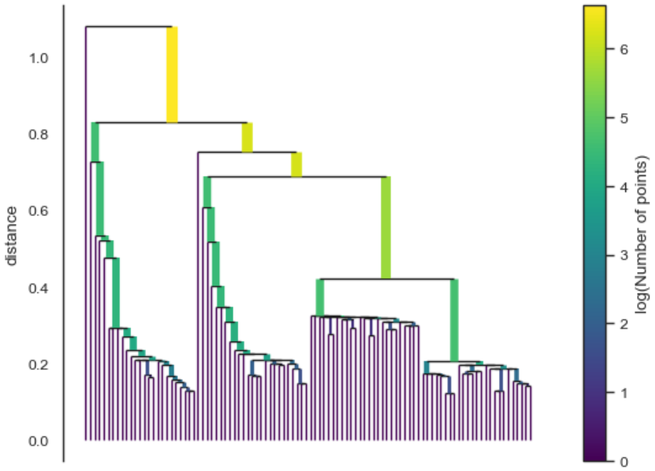


Figure 2.4: DBScan clustering with eps = .35 and min_samples = 3. All clusters found, with some noise points.

- **Mutual Reachability Graph**: The Mutual Reachability Graph is a complete graph G, in which the objects of $\mathbf{X}$ are vertices and the weight of each edge is the mutual reachability distance w.r.t. m_{pts} between respective pairs of objects

If one creates the mutual reachability graph, then removes all edges with weights greater than ε, the connected components of the resulting graph are the DBScan (not HDBScan) clusters w.r.t. m_{pts} and ε if all non-core samples are classified as noise. The unconnected objects are also classified as noise. The mutual reachability distance will keep the distance between the more densely connected points, while pushing sparser points away from their neighbours McInnes et al. (2017). The mutual reachability distance between points is *at least* their ”regular” distance, if not more.

Now that the mutual reachability graph has been aquired, a single linkage tree can be created for our dataset. For the example in figure 2.1 , the single linkage tree is shown as a dendrogram in figure 2.5. The dendrogram visualises how we can divide the dataset into clusters based on a minimum distance requirement between points. The distance requirement is denoted as ε, and is shown on the y-axis to the left of the dendrogram. At the bottom of the graph, this distance requirement is zero, which means every point is a cluster. On the top, however, the minimum distance requirements is one, and therefore every point belongs to the same cluster. In this way, one can get an overview over how the clustering of the dataset changes with different distance requirements. The color of the colored bars represent the amount of points in that cluster for that particular distance requirement. In order to make clusters from this figure, we need a minimum cluster size, and in HDBScan the minimum cluster size is the same as m_{pts}. Now, clusters can be created by traversing the dendrogram from the top downwards. We choose our distance measure to be $\lambda = \frac{1}{\varepsilon}$, in order to simplify the cluster stability calculation later, and traverse the dendrogram in figure 2.5 from the top down. When ε decreases, any cluster will either



.

Figure 2.5: Single linkage tree of mutual reachability graph. Every line is a cluster. At the bottom of the graph, every point is one cluster, while at the top, every point is a part of the same cluster. Distance on the y-axis equals ε

remain unchanged or split into different parts. If any of the split clusters have less members than the minimum cluster size (m_{pts}), all points in that cluster are labeled as noise and discarded. When the whole tree has been traversed, we end up with a tree with far fewer nodes to be used for clustering, as pictured in 2.6.

Now, we need to choose our clusters. In order to do so, Campello et al. (2013) defines the stability S of a cluster C_i: $S(C_i) = \sum_{x_j \in C_i} \frac{1}{\varepsilon_{min}(x_j, C_i)} - \frac{1}{\varepsilon_{max}(C_i)} = \sum_{x_j \in C_i} \lambda_{max}(x_j, C_i) - \lambda_{min}(C_i)$, where $\varepsilon_{min}(x_j, C_i) = \lambda_{max}(x_j, C_i)$ is the density level beyond which object x_i no longer belongs to cluster C_i and $\varepsilon_{max}(C_i) = \lambda_{min}(C_i)$ is the maximum density level at which cluster C_i exists. In other words, the stability of a cluster is dependent on how long the points that belong to it stay in the cluster as ε decreases. Starting at the leaf nodes, we can work our way up the tree. For every split, we can determine if the stability of the parent cluster is greater than the sum of stability of its children clusters. If the parent cluster is more stable, we choose the parent cluster as our cluster and unselect its children. If the sum of children cluster stabilities is greater, we set the parent cluster stability to be the sum of the children cluster stability, and continue moving up the tree, having selected the children clusters as our temporary clusters. The result of the HDBScan algorithm on the example clustering problem can be seen in figure 2.7. It captures all 4 clusters, and classifies two data points as noise (light blue).

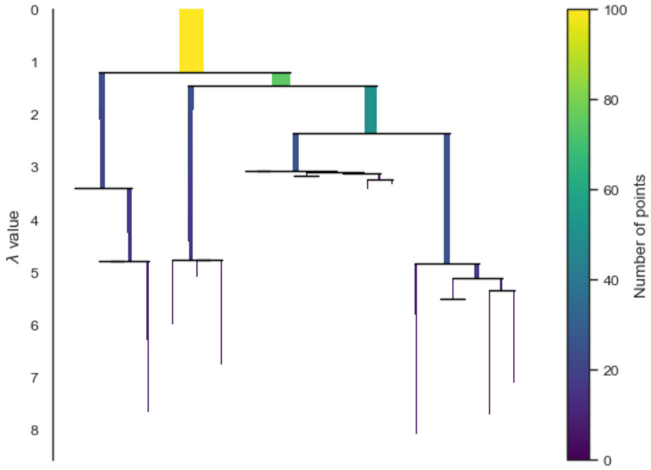


Figure 2.6: Condensed tree of mutual reachability graph

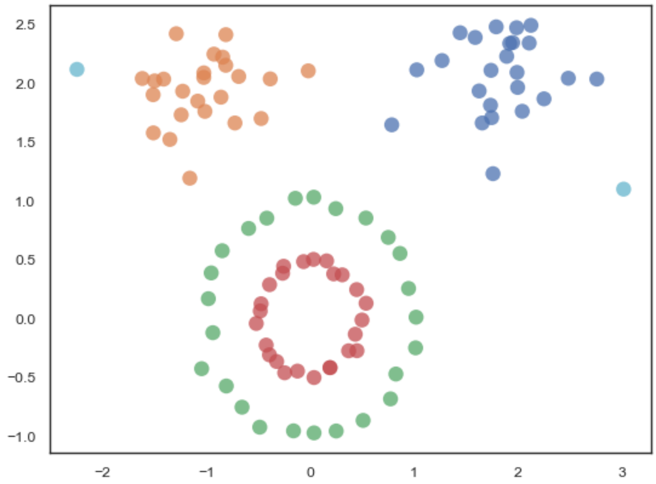


Figure 2.7: HDBScan results, min_cluster_size = 2

2.3 PCA - Principal component analysis

Principal Component analysis is a dimensionality reduction method that reduces the dimension of data while preserving as much of the variance in the dataset as possible (Dubey (2018)). This is done by transforming the variables into linear combinations of said variables, based on the correlation between them. These new variables are called **Principal Components** (Dubey (2018), last accessed 30.03.2023), and are orthogonal. The specific linear combinations of the principal components can sometimes be interpreted in order to identify which variables are correlated, and this information can in turn be used to describe the dataset. In laymans terms, PCA analysis consists of finding the directions or lines where most of the points in the dataset lie, then finding the next direction or line that the datapoints follow that is perpendicular to the first direction, and so on. This is illustrated in figure 2.8.

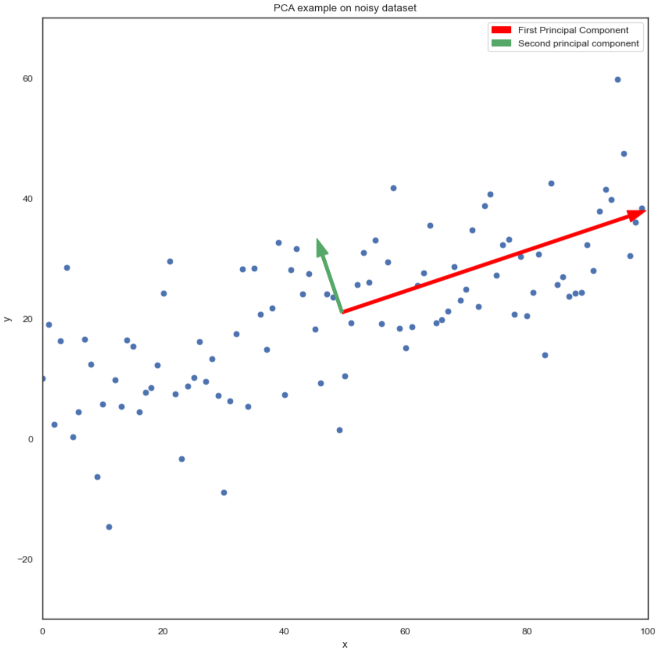


Figure 2.8: Example of using PCA on a noisy dataset. The red arrow is the first principal component, and is the linear combination of x and y that explains most of the variance in the dataset. The second principal component, the green arrow, is perpendicular to the first principal component.

Mathematically, the method can be described as such as follows (Dubey (2018), last accessed 30.03.2023). Assume the dataset X is an array with the dimensions (n_samples, n_features), and x_{ij} is the i'th sample's j'th feature, and X_i is the i'th feature.

- **Compute the mean of every feature in the dataset**:

$$\overline{X} = \{\overline{X_1}, \overline{X_2}, \ldots, \overline{X_j}\}, \qquad \text{where } \overline{X_j} = \frac{1}{n\_samples} \sum_{i=0}^{n\_samples} x_{ij}$$

- **Compute the covariance matrix of the dataset**:

$$\begin{bmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & ... & Cov(X_1, X_j) \\ Cov(X_2, X_1) & Cov(X_2, X_2) & ... & Cov(X_2, X_j) \\ \vdots & & \ddots & \vdots \\ Cov(X_i, X_1) & Cov(X_i, X_2) & \ldots & Cov(X_i, X_j) \end{bmatrix}$$

$$\text{where } Cov(X_i, X_j) = \frac{1}{n\_samples} \sum_{i=0}^{n\_samples} (x_i - \overline{X_i})(x_j - \overline{X_j})$$

$$\text{and } Cov(X_i, X_i) = Var(X_i)$$

- **Compute Eigenvectors and Eigenvalues**. The eigenvalues of a matrix A are the roots of the characteristic equation

$$\det(A - \lambda I) = 0$$

where I is the identity matrix. The eigenvectors are calculated according to this equation:

$$(A - \lambda I)\vec{v} = 0$$

which is solved for $\vec{v}$

- **Sort the eigenvectors by eigenvalues**: In order to find the principal components (the principal components **are** the eigenvectors of the covariance matrix) the eigenvectors are sorted based on their corresponding eigenvalues. These eigenvectors form the orthonormal basis of a new coordinate system based off of the PCA directions or principal components.

The first principal component, that is, the one with the highest eigenvalue, ”explains” most of the variance in the dataset. When the original data are transformed into the PCA coordinate system, the data is projected onto these principal components, which are calculated based on variable correlation. In this way, variables that vary together are combined, while variables that are uncorrelated will remain so.

Because the main purpose of PCA is to reduce the amount of variables in the dataset, it is often used in data preprocessing in order to shrink the feature space and get faster, and hopefully as good or better, results. However, due to the fact that PCA analysis is based on correlation between features in the original data, the results from PCA analysis can also be analysed and interpreted in order to gain new insight on the original data. By looking at the different PCA directions and which variables are prominent in them, it is sometimes possible to make out somewhat what these directions represent.

As evidenced above, PCA analysis is very dependent on standardisation of variables. If one variable has a much higher mean than another, it will strongly influence the final result of the analysis. If a feature has higher mean and variance, for example one variable is measured in meters and another in millimeters, the millimeter variable will have a disproportionately high impact on the PCA process. In order to remove numerical dependencies, it is very important to standardise any input data before running PCA analysis.

3

Method

3.1 The Dataset

3.1.1 Datasets

There are 6 datasets available for the analysis. 3 of the datasets are from a conventional cage, and 3 are from an experimental cage setup called Aquatraz. Aquatraz is more closed, with lice skirts, deep water intake, and possibility for the cage to be lifted completely out of the water (more info about Aquatraz at their website: aquatraz.com). Several papers on the topic show that swimming depth is sensitive to light conditions (Ulvund et al. (2021), Oppedal et al. (2007)), in addition to a study showing higher swimming activity during the night as well (Føre et al. (2018b)). In order to remove this diurnal dependency from the datasets, they were divided into night and day based on local light conditions at measurement time, bringing the grand total to 12 datasets.

The data was collected by surgically implanting fish with acoustic tags, and by using the Time Difference of Arrival algorithm on depth readings sent by the sensors to three different acoustic receivers, x and y position can be calculated from the depth measurements. These positions are given in relation to and in the same coordinate system as the hydrophones present in the cage, as explained in Hassan et al. (2019). The datasets are summarised in the following table:

Fish ID	Start date	End date	Avg. time between pos.	Cage #	# of datapoints
14	2020-07-02	2020-09-10	1.4 minutes	Aquatraz	72066
15	2020-07-02	2020-09-10	1.3 minutes	Aquatraz	79815
23	2020-07-02	2020-09-10	0.8 minutes	Aquatraz	120955
20	2021-01-08	2021-02-17	1.1 minutes	Conventional	50569
21	2021-01-07	2021-02-17	1.1 minutes	Conventional	52426
41	2021-01-07	2021-02-14	0.9 minutes	Conventional	60856

Fishes 20,21, and 41 are presumably approx. 64 cm long with a weight of around 3600g (estimated from sampling), while the aquatraz fishes are all about 38 cm long, with a weight of just under half a kilo. Both cages are places at approx (65.00, 11.74) which constitues a somewhat sheltered location north of Namsos. The aquatraz fishes are measured for two months and eight days in summer / early autumn, while the reference fishes are measured in winter, for one month and nine days, one month and ten days, and one month and 7 days, respectively.

The data was presented in the following format:

	DATE/TIME (UTC)	TIME(Unix Epoch)	XPOS(m)	YPOS(m)	ZPOS(m)
0	2021-01-07 23:05:47	1610060747	20.620000	34.230000	-6.200000
1	2021-01-07 23:06:47	1610060807	17.600000	34.150000	-7.400000
...	...	...	...	...	...

All datapoints correspond to a fish position, and include unix time and UTC time, in addition to the coordinates of the fish in x, y and z directions. Fish x and y position was normalised to have a minimum of 0 [1]. However, these datasets have varying density due to the variable transmission intervals of the tags (20s-60s), in addition to acoustic interference from other noises or tags transmitting at the same time. This increased the difficulty of working with the datasets. This is illustrated in figure 3.2, in which the time between subsequent datapoints for reference fish 41 is plotted as a histogram.

This feature of the dataset makes time series analysis of the data less feasible, as the differing measurement frequency lessens the efficacy of most time series algorithms. In order to still analyze the dataset, the first step to prepare the data for analysis was to search for sequences of data points where the time differential between subsequent points was less than sixty seconds. These sequences were called trajectories. In order to remove dependencies on trajectory length (e.g. a trajectory with more points is likely to have higher depth difference and change direction more), all variables that scale with trajectory length are divided by the length of the trajectory in seconds. The length distribution of the trajectories from the reference cage fish 41 can be seen in figure 3.2. As can be discerned from the figure, the dataset has very varying time between datapoints, which again complicates the analysis. Any trajectories of length less than five were not included in further analysis, as they were deemed too short.

3.2 Creating Trajectories

3.2.1 Recreating fish path

Now that the trajectories have been created, they can be plotted. However, the trajectories can be used to plot salmon position at different times but they do not necessarily constitute a plausible salmon trajectory. As evident in figure 3.3, the trajectory can be somewhat described, but is definitely not a true recreation of fish movement. In reality, fish move with smoother trajectories, and not in straight lines with sharp turns. Given the low and varying sampling rate of the data, speed and path are non-trivial to estimate. This is a complicating

[1] e.g. from $x \in (-5, 42.74)$ to $x \in (0, 47.74)$, as was the case with reference fish 41

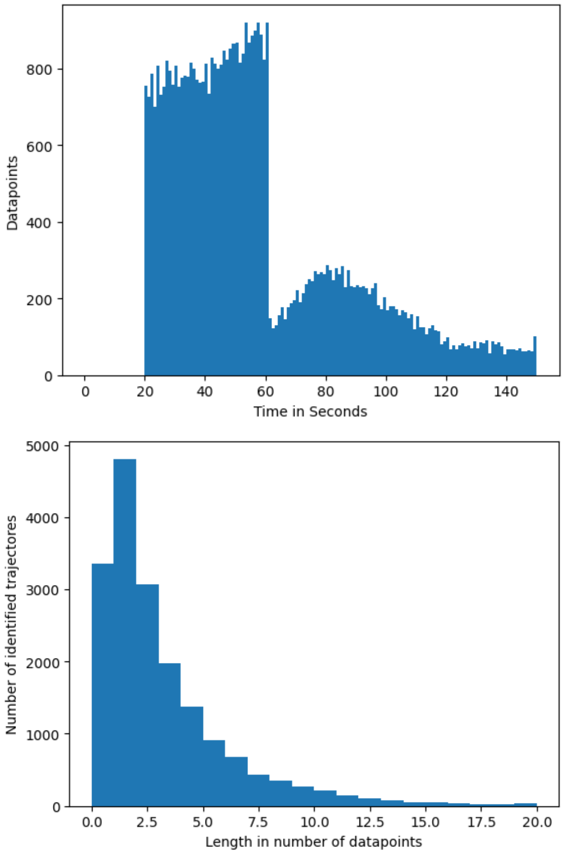


Figure 3.2: TOP: Time between subsequent datapoints for fish 41 . BOTTOM: Length of trajectories with less than 60 seconds between datapoints, fish 41 in reference cage

factor of the analysis. However, the most likely path that assumes the least movement can still be calculated based on the datapoints and time between them. In order to interpolate the trajectories and aquire more plausible values for variables like track length and angle change, a polynomial interpolation method was implemented. It interpolated three and three points, then took the average of overlapping interpolation, in order to approximate salmoon trajectories smoothly. That is, for every three points, it solved the equations :

$$
\begin{aligned}
a_0 + b_0 * t[i] + c_0 t[i]^2 &= x[i], \qquad i \in [0,1,2] \\
a_1 + b_1 * t[i] + c_1 t[i]^2 &= y[i], \qquad i \in [0,1,2] \\
a_2 + b_2 * t[i] + c_2 t[i]^2 &= z[i], \qquad i \in [0,1,2]
\end{aligned}
$$

For a, b and c, with t being the timestamps of the x, y and z positions at those times. Then, by using the resulting functions, the fish trajectory can be recreated smoothly. The overlapping points were simply averaged. This method avoids the excessive deviation from the lines between points, while still being more realistic. The recreation of a salmon trajectory can be seen in figure 3.4. By recreating the path in this way, the trajectories go from having 5+ datapoints to having a datapoint per second.

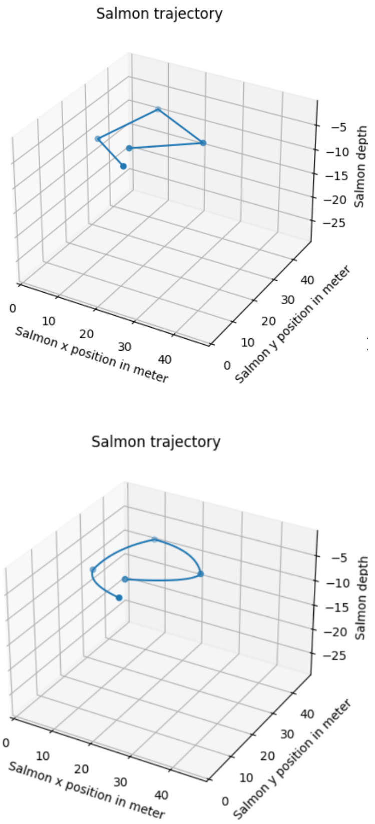


Figure 3.4: TOP: Non interpolated trajectory. BOTTOM: The same trajectory, interpolated.

3.2.2 Creating variables

In order to create data that can be used as input for models and algorithms, the data needs to accurately represent fish behaviour. However, more data is not always better, and too many variables may have a detrimental impact on performance. In order to mathematically define variables with impact on behaviour, both literature and fish movement was studied. Of note is that fact that fish in cages seem to swim in circles (Oppedal et al. (2011)), so the variables chosen as input to algorithms later needed to at least capture that kind of movement. This leads to the question: is horisontal position sufficient to understand fish behaviour? If a fish swims in circles around the cage, a human would define it as one behaviour, but mathematically speaking, the x and y variables would reach both their highest and lowest values in this single behaviour, which can confuse all linear classifiers and many non-linear ones as well. Therefore, the decision was made to not use x and y as inputs to the machine learning methods. A fish swimming in circles probably has a long track length and a constant distance from the center of the cage, in addition to a large total distance from center. Therefore, trajectory length, total distance from center, and difference in distance from center were all included as variables in the analysis.

In order to capture feeding as well, the variables angle change and average depth were included. When a fish is feeding, it may have a high total angle difference, low average depth, and low track length, in order to find and catch the pellets coming from the surface, as indicated in Føre et al. (2018b). Total depth change was also included, as changing depth may signify stress or an otherwise important event.

The following variables were calculated from the x, y and z values of the trajectories:

NOTE: x[i], y[i] and z[i] denotes the i'th values of x, y and z position, respectively. These values are the smoothed values from the polynomial trajectory interpolation mentioned above, and have a value for every second from the start of the trajectory to the end. N is the length of the trajectory in seconds.

- **Average depth**: The average depth of a trajectory, defined as

$$\frac{1}{N}\sum_{i=0}^{N}\text{z[i]}$$

- **Depth difference**: The total depth change per second in absolute value :

$$\frac{1}{N}\sum_{i=1}^{N}\text{abs}(z[i]-z[i-1])$$

- **Track Length**: Total length of trajectory in body lengths per second. Calculated as:

$$\frac{1}{N*\text{body\_length}}\sum_{i=1}^{N}\sqrt{(x[i]-x[i-1])^2+(y[i]-y[i-1])^2+(z[i]-z[i-1])^2}$$

- **Recreated angles**: Total angle sum of recreated trajectory in degrees per second. Calculated as:

$$\vec{p_i}=(x[i]-x[i-1],y[i]-y[i-1],z[i]-z[i-1])^T$$

$$\frac{1}{N}\sum_{i=1}^{N}\text{arccos}\,(\frac{\vec{p_{i+1}}}{|\vec{p_{i+1}}|}\cdot\frac{\vec{p_i}}{|\vec{p_i}|})$$

  This captures the sum of angle the fish changes from its previous direction in xy-plane, for every data point. The idea is that it is ”cheaper” for the fish to swim straight, and therefore a high angle change indicates a departure from standard behaviour. Only includes angle in xy-plane, as there is already a variable for depth change.

- **Total distance from center**: Total distance per second from center of cage, inspired by Stockwell et al. (2021), as they measured distance from center as a response variable for fish behaviour during environmental events. Calculated by first normalizing the x and y data to have a min value of 0, then setting cage centre to be half of the max value for x and y, giving the cage centre the position of (25,25,0) in x, y and z coordinates for both the reference cage and Aquatraz. The distance from center was then calculated as:

  $$c = [\text{cage\_centre\_x\_pos}, \text{cage\_centre\_y\_pos}]$$

  $$\frac{1}{N}\sum_{i=1}^{N}\sqrt{(x[i]-c[0])^2+(y[i]-c[1])^2}$$

  Z position of cage centre was not included, as the purpose was to capture distance from the center, not from the top of the cage.

- **Distance from center moved**: Max distance from center minus min distance from center. The idea is that this variable is small when circle swimming and perhaps larger when not.

3.3 Statistical analysis and visualisation

The aforementioned 12 datasets were analysed using three approaches: conventional analysis, including mean and variance analysis in addition to plotting, principal component analysis for covariance between variables, and HDBScan clustering.

3.3.1 Mean and Variance

Sample mean and sample variance can be useful measures to explore similarities and dissimilarities between datasets. Sample mean of a dataset $X = x_1, x_2, ..., x_n$can be calculated relatively straightforward:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where N is the amount of samples in the dataset. Sample variance is also scaled by number of samples. This has the effect of increasing variance with low samples, as we are less sure of the distribution. This means that two distributions, with the same apparent spread when plotted, can have very different variances, if amount of samples is different. Sample variance is calculated with the sample mean as such:

$$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$$

Sample mean and sample variance was calculated for every variable in every dataset.

This analysis had a purpose of giving insight into to what degree these variables have similar distributions for different fish, and at night versus day. The purpose of this project is not only to attempt to discover different modes of behaviour for fish, but also to see if these behaviours are consistent for different fish. After all, the usefulness of eventual results are lessened if they are not generalizable to a greater dataset.

3.3.2 Basic conventional analysis

The conventional analysis is the first stop in data exploration. This step was done by plotting all selected variables and looking for trends in the data. Different plotting methods were used, but in this report histograms and scatterplots will be presented. These were the plots made for each dataset:

- Histogram of every variable for night and day.
- Scatterplot of pairs of variables.

The histograms were made in order to determine the distribution of the variables in the identified and smoothed trajectories.

The scatterplots were made in order to identify relationships between pairs of variables. Variable pairs chosen were average depth and depth difference, track length and angle change, in addition to average distance from center and distance from center moved. Track length and angle change were chosen in order to perhaps identify circular swimming, depth and depth difference were chosen to perhaps identify feeding, and distance from center moved and average distance from center were chosen in order to uncover other dependencies.

With these plots produced, we can not only use them to discuss the behaviour of the fish, but also see if the plots are similar for the different datasets. Similar histograms and scatterplots will imply that the variables are taken from similar distributions. If the variables are from similar distributions, that means it is possible to use the behavioural data of one fish to generalize to others.

3.3.3 Principal Component analysis

After the conventional analysis, a PC analysis was made. The PCA algorithm was run on the dataset, and all principal components were extracted. Each principal component was analysed in order to ascertain if the PCA components extracts a meaningful behavioural pattern or not, for example if it constitues something similar to feeding or circle swimming. Principal components are presented with their weighted variable importances, which means that one variable will have an importance of $\pm$ 1, and the rest will have an absolute value below one. However, this analysis was mostly qualitative, as PCA is not a ”true” classification algorithm. The PCA algorithm used is part of the scikit-learn python library (Pedregosa et al. (2011)).

3.3.4 HDBScan Clustering

After PCA analysis, unsupervised clustering was performed on the datasets in order to extract meaningful trajectory prototypes. As there is no objective performance measure in this step, the clustering focused on varying parameters in order to extract meaningful partitions of data, and presented these in order to identify fish behaviour archetypes. The clustering hyperparameters were chosen based on analysis of the histograms and scatterplots, in addition to visual inspection of 3D scatterplots of clustering results, which will be included in the results. The HDBScan implementation used was created by McInnes et al. (2017)

4

Results

4.1 Mean and Variance

In order to get a measure of variable distributions, mean and variance was calculated for every dataset:

Table 4.1: Mean and variance for all datasets

	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
RE_41 Day	9.85 ± 0.39	.013 ± 0.31	0.28 ± 0.16	1.45 ± 0.11	0.27 ± 0.22	8.35 ± 0.30
RE_41 Night	5.59 ± 0.06	.008 ± 0.03	0.19 ± 0.09	1.28 ± 0.06	0.28 ± 0.06	8.25 ± 0.12
RE_20 Day	12.19 ± 0.86	.019 ± 0.20	0.39 ± 0.24	2.15 ± 0.20	0.36 ± 0.11	7.76 ± 0.33
RE_20 Night	20.17 ± 0.74	.012 ± 0.35	0.24 ± 0.31	1.68 ± 0.36	0.34 ± 0.27	6.28 ± 0.25
RE_21 Day	8.71 ± 0.46	.013 ± 0.14	0.37 ± 0.42	1.65 ± 0.18	0.45 ± 0.18	7.13 ± 0.27
RE_21 Night	2.01 ± 0.10	.005 ± 0.02	0.18 ± 0.04	1.53 ± 0.03	0.43 ± 0.00	8.05 ± 0.11
AQT_23 Day	13.06 ± 0.13	.027 ± 0.07	0.27 ± 0.12	1.88 ± 0.01	0.47 ± 0.11	10.58 ± 0.15
AQT_23 Night	11.11 ± 0.21	.021 ± 0.13	0.20 ± 0.07	2.06 ± 0.11	0.43 ± 0.15	10.96 ± 0.23
AQT_14 Day	13.93 ± 0.36	.028 ± 0.10	0.28 ± 0.14	1.69 ± 0.19	0.53 ± 0.03	8.78 ± 0.18
AQT_14 Night	11.60 ± 0.68	.019 ± 0.07	0.22 ± 0.11	1.91 ± 0.08	0.49 ± 0.21	8.40 ± 0.31
AQT_15 Day	14.56 ± 0.32	.024 ± 0.15	0.31 ± 0.03	1.56 ± 0.12	0.55 ± 0.02	8.15 ± 0.09
AQT_15 Night	12.17 ± 0.57	.02 ± 0.10	0.24 ± 0.07	1.74 ± 0.17	0.53 ± 0.27	7.86 ± 0.22

The mean and variance table 4.1 provides an overview of means and variances for the 6 included variables. Average depth for the reference fish is varies greatly between night and day, and also from fish to fish. However, for the aquatraz fish they are more similar, albeit with some differences in the variance. Average depth is lower (lower number, higher in the water) at night for all fish except fish 20. Variance in average depth is higher at night for aquatraz, and lower for the reference fish. The depth differences are lower in the reference cage, and higher in aquatraz. These somewhat lessen during the night for both cages. Track length is similar across all fish, especially during the night, where the only great discrepancy is the track length variance of fish 20. During the day, both fish 20 and 21 have a high mean, but overall this variable seems consistent across fish and cages. Track length is also shorter at night for all cases. Angle change per second is interesting in that it is lower at night for the reference fish, while it is higher at night for the aquatraz fish. There is no readily apparent consistent pattern to angle change variances. On average, angle change is pretty similar across all cases during the day. Average distance from center is very similar in night and day for all cases, but the variance is greater at night for aquatraz and fish 20, while lower for the two other reference fish. Average distance from center is somewhat greater in aquatraz. Finally, distance from center moved is somewhat similar for day and night in most cases (not fish 20 and 21). In the reference cage, variance decreases at night, while in aquatraz, variance increases at night.

4.2 Histograms of Variables

4.2.1 Reference fish 41

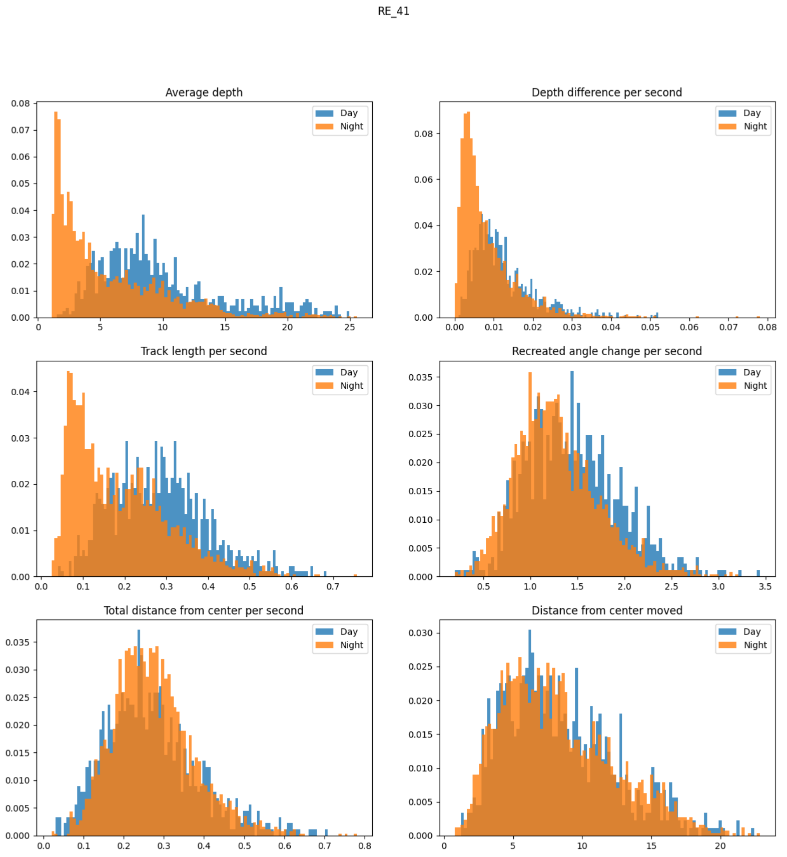


Figure 4.1: Variable distribution for reference fish 41

For fish 41, the histograms (4.1) are pretty similar between night and day for total distance from center and distance from center moved. Angle change at night is similar, but a little bit lower than in the day. The big discrepancies between night and day are found in average depth, depth difference per second, and track length per second. All these variables are a generally a lot lower in the night, but have more prominent peaks. This is consistent with the observations in the means and variances.

4.2.2 Reference fish 20

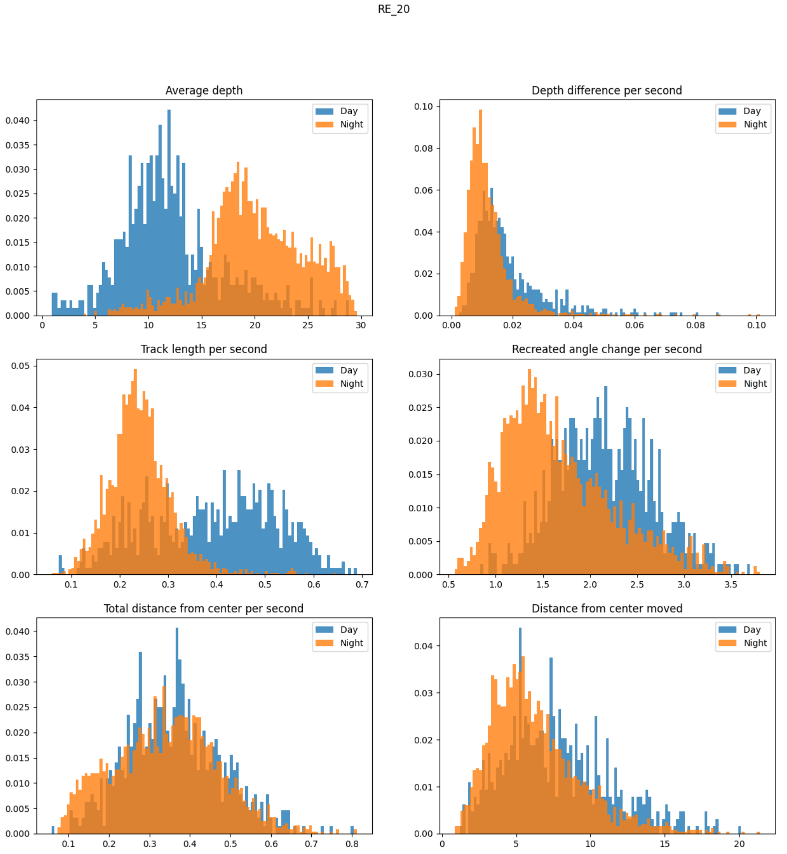


Figure 4.2: Variable distribution for reference fish 20. Notice discrepancies in track length, angle change, and average depth for day and night.

The average depth histogram (4.2) of this fish explains why the mean was so much greater in the mean and variance table. The depth distributions of average depth are very dissimilar for night and day here, which is in and of itself a very convincing argument for dividing night and day. Depth difference is somewhat lower at night, while track length is far lower at night. Total distance from center and total distance from center moved are pretty similar for day and night, with distance from center moved being somewhat skewed to be greater at day. Depth difference and track length also have more prominent peaks at night.

4.2.3 Reference fish 21

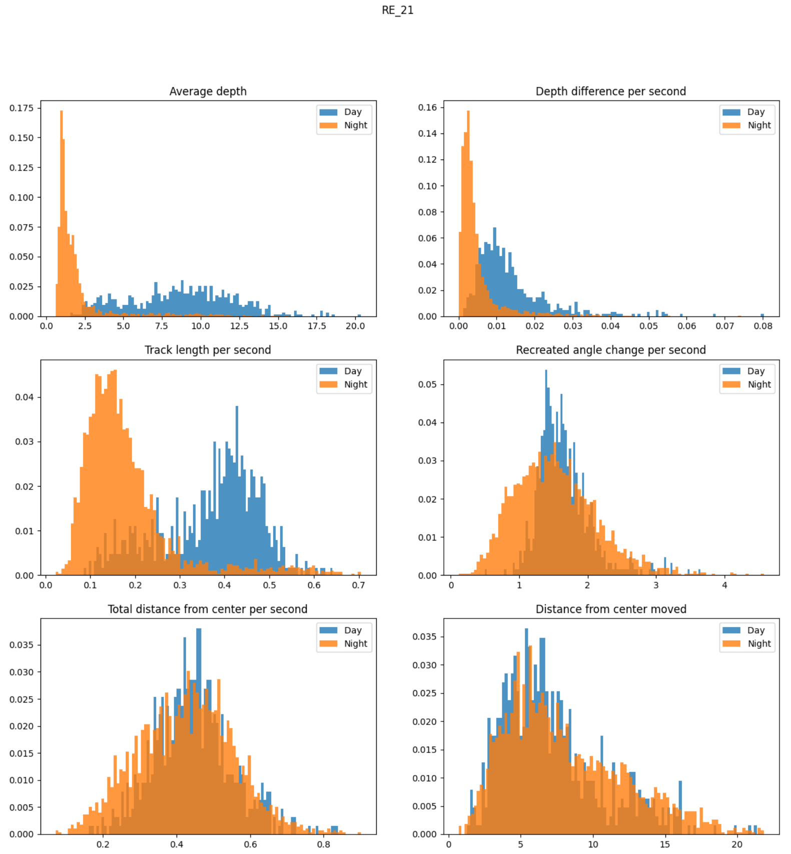


Figure 4.3: Variable distribution for reference fish 21. Notice discrepancies in average depth, depth difference, and track length for day and night

With this fish, we observe great differences between night and day for the first four variables, while the last two are similar. However, it is worth to mention that all average depth distributions (4.2) have been different, and this one is also very striking. At day fish 21 can be found from 2.5 to 15 meters, while at night it is almost always in the top 2.5 meters. Depth difference and track length are shorter at night, while angle change has a similar mean at day and night, with a very different spread. Distance from center variables are similar for day and night. Track length has two strikingly different distributions for night and day, with the mean at day being much larger. Average depth, depth difference per second, and track length per second have a lot more prominent peaks at night.

4.2.4 Aquatraz fish 23

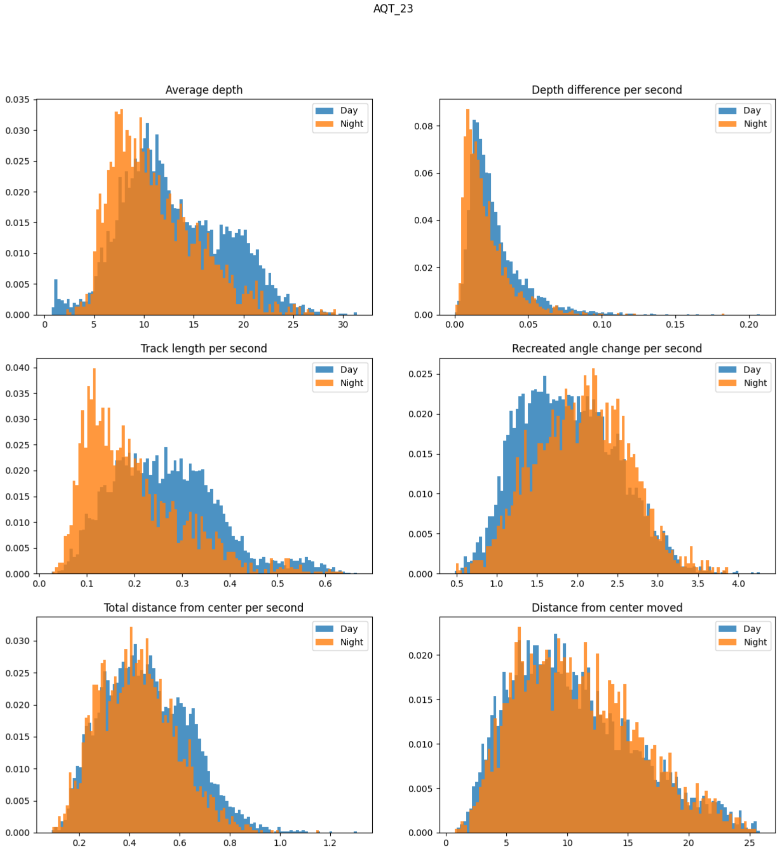


Figure 4.4: Variable distribution for aquatraz fish 23. Notice the spike at low average depth for day.

The variables pertaining to distance from center are similar (4.4), just like the previous histograms. Angle change and depth difference are also similar, but skewed to the right and left for night, respectively. Track length is lower at night. The most interesting feature of this fish is the spike at low average depth in day and not night.

4.2.5 Aquatraz fish 14

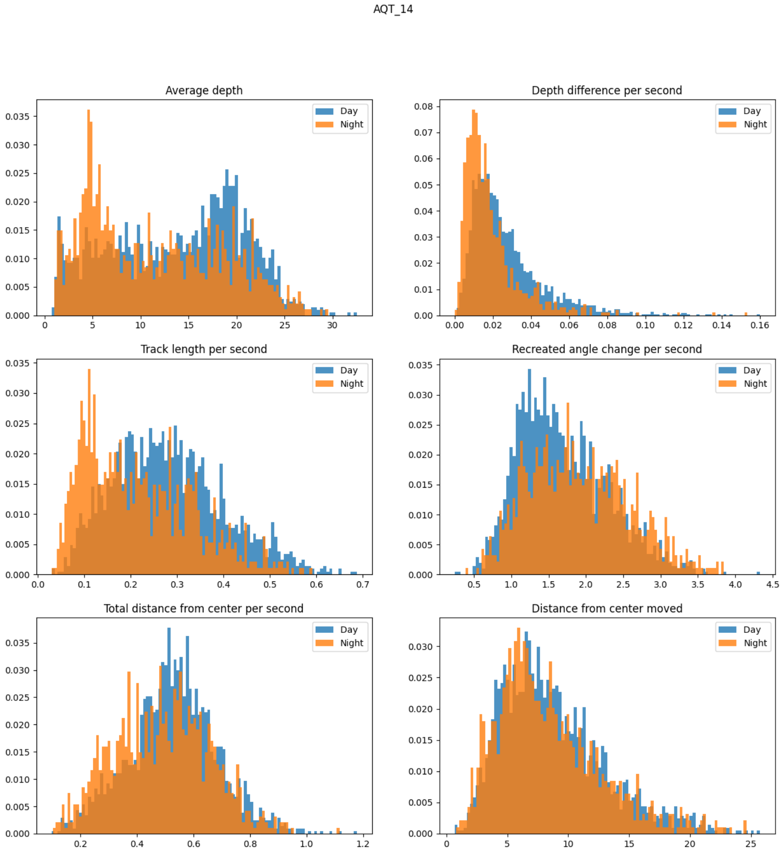


Figure 4.5: Variable distribution for aquatraz fish 14.

Total distance from center and distance from center moved are similar in these histograms (4.5). Track length is lower at night, but the difference is not as striking. Depth difference has a smaller mean and smaller tail at night, while angle change has a somewhat lower mean during the day. Average depth is flatter here than for most other fish for both day and night, but the mean is lower for night and higher for day. The prominent peaks seen in the reference fish are also present in the first three variables here.

4.2.6 Aquatraz fish 15

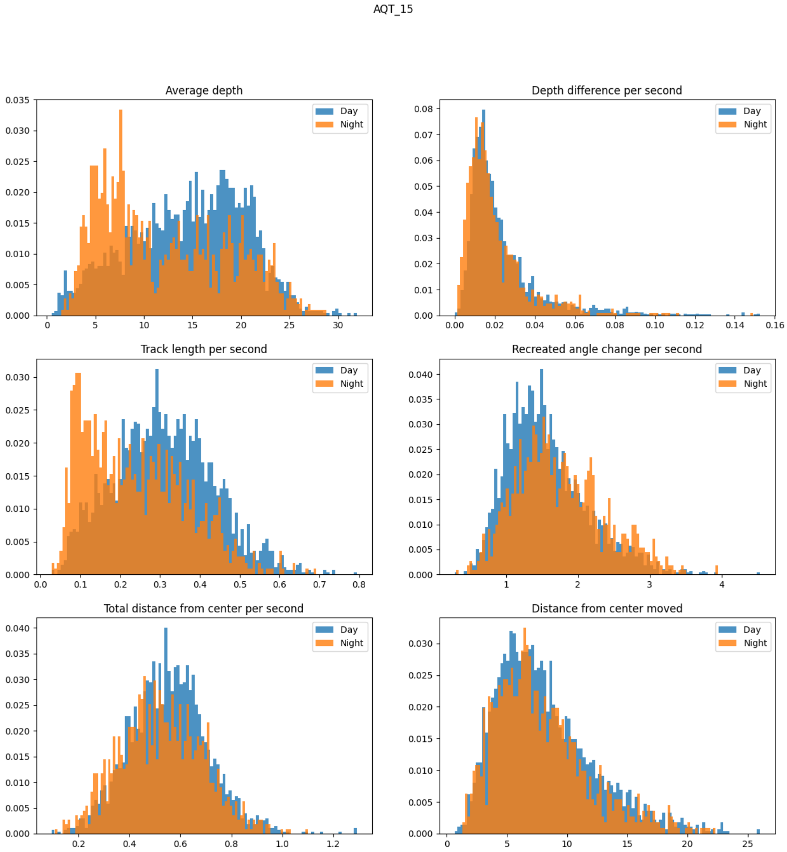


Figure 4.6: Variable distribution for aquatraz fish 15. Biggest discrepancy here is in track lengths and average depth.

Total distance from center, distance from center moved, depth difference per second, and recreated angle change are all very similar from day and night (4.6) . Track length has some overlap but is clearly lower at night. Average depth histogram is also somewhat flat, however the mean is clearly lower for night. These histograms have somewhat clearer peaks at night for average depth and track length per second.

While there are obvious differences between the distributions between night and day, and there seems to be structure beyond a standard normal distribution in the variables, this structure varies between the fish.

4.3 Scatterplots of variables

Any missing scatterplots are located in the appendix.

4.3.1 Reference fish 41

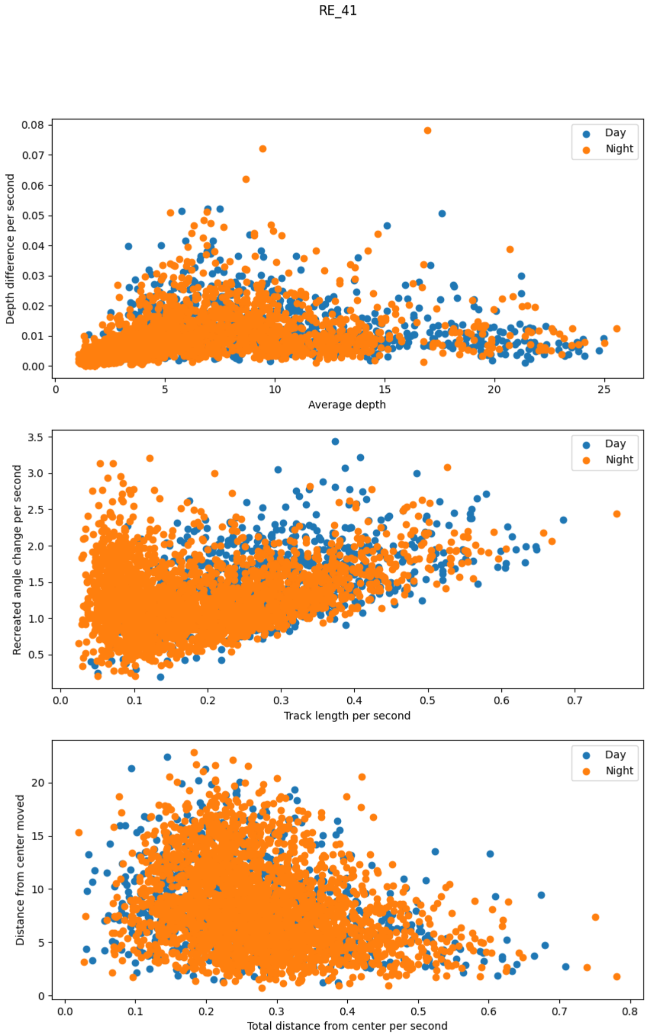


Figure 4.7: Scatterplots of variables for reference fish 41

The scatterplots 4.7 show that average depth is somewhat greater at night for fish 41. Angle vs track length is somewhat triangular, where more angle change per second is correlated with more track length per second. At night, there is some clustering with regard to low track length and variable angle change. Distance from center plots have some correlation with low distance moved but greater total distance from center.

4.3.2 Reference fish 20

In this scatterplot (6.1), we observe some clustering of low average depth at day and greater average depth at night, consistent with means and histograms. Angles vs track lengths shows a more consistent clustering in low track length at night, with longer track lengths in the day. Distance from center plots are very similar for day and night.

4.3.3 Reference fish 21

The scatterplot (6.2) shows a clear clustering of constant high depth at night. Angle and track length also shows a clear correlation in track length and angle change, in addition to a clear night cluster at low track lengths and variable angle change. There is less structure in the total distance from center variables, but both variables seem to have greater spread at night.

4.3.4 Aquatraz fish 23

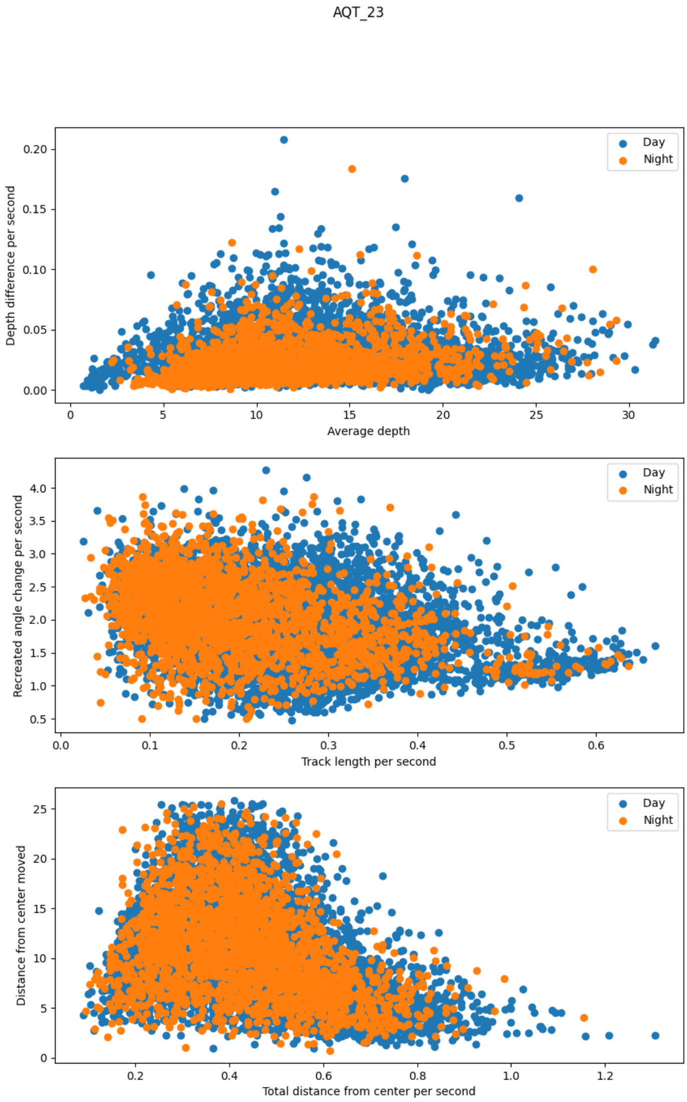


Figure 4.8: Scatterplots of variables for aquatraz fish 23. Notice cluster of low depth in the day, linear correlation between track length and angle change at long track lengths, and a clearer trend of low distance from center change with long distance from center.

From this scatterplot (4.8), a clear tendency is present in that all variables have higher mean at day. In addition, there is a cluster of small average depth in the day, as present in histograms as well. The line at small angle change but somewhat increasing track length is notable here. In distance from center, a line of constant distance from center moved with differing distance from center can be observed.

4.3.5 Aquatraz fish 14

The scatterplot (6.3) shows that in depth, night and day are pretty similar, however day trajectories have greater mean depth difference per second. The line of track length variation but low angle variance is present, while the distance from center plot is roughly triangle shaped, like the previous distributions.

4.3.6 Aquatraz fish 15

The scatterplot (6.4) shows that this fish has variables very similar to fish 14. Depth is similiar with somewhat higher depth difference mean at day, and the line of differing track length with a slight increase in recreated angles is even more clear in this plot.

While there are some clear trends in these scatter plots, there are few humanly identifiable clusters.

4.4 PCA Analysis

Any missing tables are located in the appendix.

4.4.1 Reference fish 41

Day:

Table 4.2: Principal Components for fish 41 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 33.37%	-0.50	1.00	0.42	0.35	0.58	0.09
1: 24.56%	0.24	0.74	-0.53	-0.17	-0.74	1.00
2: 19.16%	1.00	0.34	0.37	0.25	-0.10	-0.32
3: 12.46%	0.06	-0.65	0.79	0.74	-0.00	1.00
4: 6.21%	-0.39	0.05	-0.27	1.00	-0.74	-0.47
5: 4.24%	0.41	-0.12	-1.00	0.64	0.87	0.22

Table 4.2 shows the result of PCA on the data from fish 41 day. Almost 90 % of the variance in the dataset is explained by the first four principal components.
The first principal component, PC0, explains 33.37 % of the variance and describes a correlation between high depth difference, lower average depth, higher distance from center, somewhat higher track length and angle.
PC1 explains 24.56 % of the variance in the dataset and describes a correlation between high distance from center moved, low total distance from center, high depth difference, somewhat lower track length.
PC2 explains 19.16 % of the variance in the dataset and describes a correlation between high average depth, somewhat higher depth difference, track length and angle, and somewhat lower distance from center moved.
PC3 explains 12.46 % of the variance and describes a correlation between high distance from center moved, high angle and track length, lower depth difference.

Night:

Table 4.3: Principal Components for fish 41 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 49.67%	0.87	1.00	0.77	0.16	0.15	-0.03
1: 17.65%	-0.96	1.00	-0.24	0.16	0.06	0.69
2: 13.99%	0.63	-0.09	-0.34	-0.21	-0.80	1.00
3: 8.96%	0.20	0.75	-1.00	-0.64	-0.53	-0.95
4: 6.05%	0.23	0.01	-0.49	1.00	0.01	-0.10
5: 3.68%	0.37	0.00	-0.54	-0.33	1.00	0.31

Table 4.3 shows the result of PCA on the data from fish 41 night. About 90 % of the variance in the dataset is explained by the first four principal components. PC0 explains almost 50% of the variance in the dataset and describes a correlation between high depth

difference, high average depth, and high track length.
PC1 explains 17.65% of the variance in the dataset and describes a correlation between high depth diff, low average depth, and high distance from center moved.
PC2 explains 13.99% of the variance in the dataset and describes a correlation between high distance from center moved, low distance from center, and somewhat hight track length.
PC3 explains 8.96% of the variance in the dataset and describes a correlation between low track length, low distance from center moved, high depth diff, and somewhat low angle and distance from center.

4.4.2 Reference fish 20

Day:

Table 6.1 shows the result of PCA on the data from fish 20 day. The first four principal components explain more than 90 % of the variance in the dataset.
PC0 explains about 40% of the variance in the dataset and describes a correlation between high depth difference and somewhat high distance from center moved.
PC1 explains 21.12% of the variance in the dataset and describes a correlation between high distance from center, low distance from center moved, high depth diff and high track length.
PC2 explains 16.3% of the variance in the dataset and consists mostly of high average depth, with some negative contribution from distance from center moved and some positive contribution from depth difference.
PC3 explains 13.13% of the variance in the dataset and describes a correlation between low distance from center moved, low track length, low average depth and somewhat low distance fro center.

Night:

Table 6.2 shows the result of PCA on the data from fish 20 night. More than 90 % of the variance in the dataset is explained by the first four components.
PC0 explains almost 40% of the variance in the dataset and consists mostly of depth difference, with small contributions from distance from center moved and angle change..
PC1 explains 23.87% of the variance in the dataset and describes a correlation between low distance from center, high angle change, somewhat higher average depth, and somewhat lower track length and distance from center moved.
PC2 explains 20.93% of the variance in the dataset and consists mostly of distance moved from center, with a lesser negative contribution from depth difference.
PC3 explains 8.25% of the variance in the dataset and describes a correlation between high track length, high angle change, and somewhat higher distance from center.

4.4.3 Reference fish 21

Day:

Table 6.3 shows the result of PCA on the data from fish 21 day. More than 90% of the variance is explained by the first four principal components. PC0 explains 44.96% of the variance in the dataset and describes a correlation between high depth diff and somewhat higher distance from center moved.
PC1 explains 26.08% of the variance in the dataset and describes a correlation between very low distance from center moved, high average depth, high depth difference, and high track length.
PC2 explains 13.7% of the variance in the dataset and describes a correlation between high distance from center mobed, high average depth, and somewhat high track length.
PC3 explains 6.72% of the variance in the dataset and describes a correlation between very low distance from center, low track length, somewhat higher average depth, somewhat higher angle, and somewhat lower distance from center moved. .

Night:

Table 4.4: Principal Components for fish 21 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 60.59%	0.94	1.00	0.61	0.06	0.10	-0.06
1: 14.52%	-0.51	0.79	-0.38	0.39	-0.40	1.00
2: 11.11%	-0.74	0.88	-0.45	-0.25	0.36	-1.00
3: 6.92%	0.20	-0.08	-0.29	1.00	-0.07	-0.36
4: 4.64%	0.62	0.06	-0.87	-0.53	-1.00	-0.25
5: 2.21%	-0.48	0.02	0.84	0.13	-1.00	-0.39

Table 4.4 shows the results of PCA on the data from fish 21 night. More than 90 % of the variance is explained by the first four principal components.
PC0 explains 60.59 % of the variance in the dataset and describes a correlation between high depth difference, high average depth, snd somewhat high track length.
PC1 explains 14.52% of the variance in the dataset and describes a correlation between high distance from center moved, high depth difference, somewhat lower average depth, track length, and distance from center, in addition to somewhat higher angle.
PC2 explains 11.11% of the variance in the dataset and describes a correlation between low distance from center moved, high depth diff, low average depth, and somewhat lower track length.
PC3 explains 6.92% of the variance in the dataset and consists mostly of high angle change, with some negative contribution from track length and distance from center moved, with a modest positive contribution from average detph.

4.4.4 Aquatraz fish 23

Day:

Table 6.4 shows the results of PCA on the data from fish 23 day. Almost 90 % of the variance in the dataset is explained by the first four principal components.
PC0 explains 35.29% of the variance in the dataset and describes a correlation between high depth difference and somewhat high distance from center moved.
PC1 explains 22.37% of the variance in the dataset and describes a correlation between high track length, somewhat high average depth, distance from center, and depth diff, and somewhat low ange change and distance from center moved.
PC2 explains 19.66% of the variance in the dataset and describes a correlation between low average depth, low track length, and low distance from center moved, with somewhat higher distance from center and depth difference.
PC3 explains 10.96 % of the variance in the dataset and describes a correlation between high average depth and somewhat high angle change, and low distancef from center moved, low distance from center, and low track length.

Night:

Table 6.5 shows the results of PCA on the data from fish 23 night. Almost 90 % of the variance in the dataset is expained by the first four principal directions.
PC0 explains 41.3% of the variance in the dataset and describes a correlation between high depth difference and somewhat high track length and average depth.
PC1 explains 21.91% of the variance in the dataset and describes a correlation between high track length, somewhat high average depth and distance from center, and somewhat low depth difference, angle change, and distance from center moved.
PC2 explains 16.48% of the variance in the dataset and describes a correlation between low distance from center moved, somewhat high depth difference and distance from center, in addtiion to somewhat low track length.
PC3 explains 9.2% of the variance in the dataset and describes a correlation between low distance from center, high average depth, and somewhat low distance from center moved.

4.4.5 Aquatraz fish 14

Day:

Table 4.5: Principal Components for fish 14 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 36.48%	0.10	1.00	0.12	0.14	-0.15	0.42
1: 27.39%	-1.00	0.06	-0.78	0.43	-0.14	0.15
2: 15.95%	-0.04	0.51	-0.21	-0.11	0.26	-1.00
3: 8.14%	0.72	-0.14	-0.34	1.00	-0.52	-0.27
4: 7.39%	-0.34	-0.04	0.80	1.00	0.96	-0.03
5: 4.64%	0.58	0.02	-0.87	-0.06	1.00	0.44

Table 4.5 shows the results of PCA on the data from fish 14 day. Almost 90 % of the variance is explained by the first four principal directions.
PC explains 36.48% of the variance in the dataset and describes a correlation between high depth difference and somewhat high distance from center moved.
PC explains 27.39% of the variance in the dataset and describes a correlation between low average depth, low track length, and somewhat high angle.
PC explains 15.95% of the variance in the dataset and describes a correlation between low distance from center moved and somewhat high depth difference.
PC explains 8.14% of the variance in the dataset and describes a correlation between high angle, high average depth, and somewhat low distance from center.

Night:

Table 4.6: Principal Components for fish 14 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 43.65%	0.73	1.00	0.71	-0.15	0.01	0.39
1: 22.30%	-0.86	1.00	-0.61	0.49	-0.27	0.34
2: 15.15%	0.02	-0.45	0.11	0.16	-0.44	1.00
3: 7.96%	0.84	0.06	-0.65	0.27	-1.00	-0.41
4: 6.28%	0.22	-0.09	0.14	1.00	0.38	-0.05
5: 4.66%	-0.56	-0.05	1.00	0.27	-0.84	-0.53

Table 4.6 shows the results of PCA on the data from fish 14 night. The first four principal components explain almost 90 % of the variance in the dataset.
PC0 explains 43.65% of the variance in the dataset and describes a correlation between high depth difference, high average depth, and high track length, in addition to somewhat hight distance from center moved.
PC1 explains 22.3 % of the variance in the dataset and describes a correlation between high depth difference, high average depth, somewhat low track length, and somewhat high angle change.
PC2 explains 15.15% of the variance in the dataset and describes a correlation between high distance from center moved, somewhat low depth diff, and somewhat low distance from center..
PC3 explains 7.96% of the variance in the dataset and describes a correlation between low distance from center, high average depth, and somewhat low track length.

4.4.6 Aquatraz fish 15

Day:

Table 4.7: Principal Components for fish 15 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 42.05%	0.00	1.00	0.08	0.19	-0.12	0.42
1: 21.40%	-1.00	0.11	-0.96	0.39	0.05	-0.24
2: 17.22%	0.06	0.53	0.11	-0.38	0.35	-1.00
3: 7.88%	0.31	-0.05	0.17	1.00	0.05	-0.36
4: 6.64%	-0.93	-0.10	1.00	0.15	0.76	0.21
5: 4.81%	0.41	0.02	-0.45	0.02	1.00	0.33

Table 4.7 shows the results of PCA on the data from fish 15 day. The first four principal components explain almost 90 % of the variance in the dataset.
PC0 explains 42.05% of the variance in the dataset and describes a correlation between high depth difference and somewhat high distance from center moved.
PC1 explains 21.4% of the variance in the dataset and describes a correlation between low average depth, low track length, and somewhat high angle change.
PC2 explains 17.22% of the variance in the dataset and describes a correlation between low distance from center moved, somewhat higher depth difference, somewhat lower angle and somewhat higher distance from center.
PC3 explains 7.88% of the variance in the dataset and describes a correlation between high angle change, low average depth, and high distance from center.

Night:

Table 4.8: Principal Components for fish 15 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 44.81%	0.60	1.00	0.65	-0.06	-0.11	0.42
1: 21.44%	-0.92	1.00	-0.83	0.57	-0.12	0.25
2: 14.33%	-0.07	0.45	0.08	-0.27	0.42	-1.00
3: 8.25%	1.00	0.02	-0.60	0.84	-0.65	-0.61
4: 6.40%	-0.19	-0.15	0.62	1.00	0.49	-0.07
5: 4.78%	0.52	0.03	-0.58	0.00	1.00	0.35

Table 4.8 shows the results of PCA on the data from fish 15 night. The first four principal components explain almost 90 % of the variance in the dataset. PC0 explains 44.81% of the variance in the dataset and describes a correlation between high depth difference, somewhat high track length, somewhat high average depth, and somewhat high distance from center moved. .
PC1 explains 21.44% of the variance in the dataset and describes a correlation between high depth difference, low average depth, low track length, and somewhat high angle

change.

PC2 explains 14.33% of the variance in the dataset and describes a correlation between low distance from center moved and somewhat high depth difference and distance from center.

PC3 explains 8.25% of the variance in the dataset and describes a correlation between high average depth, high angle change, and somewhat low track length, distance from center, and distance from center moved.

4.5 HDBScan Clustering

Any missing figures are located in the appendix.

The histograms show that the most important variables might be average depth, depth difference, and track length, and perhaps angle change. Indeed, these were the ones who gave the most satisfactory clustering. However, ”satisfactory clustering” is, in unsupervised learning, a qualitative term. What is satisfactory or not is ultimately up to the observer, as mentioned in the theory chapter. As there are no clear metrics for clustering success, parameters were varied in order to correctly cluster humanly identifiable clusters, while simultaneously trying to identify the behaviour types we expect to find, namely feeding and circular swimming. Different values for `min_cluster_size` were tried, in addition to different variables, for example all PCA directions, a subset thereof, all variables, and subsets of all variables. Clustering on the first four variables (average depth, depth difference, track length, angle change) with a cluster size of 10 identified the humanly identifiable clusters on the aquatraz fish 23 day dataset, as this is the dataset with the most humanly identifiable clusters, and gave reasonable results on the rest. Clusters are presented with examples of each cluster. Clusters are plotted with regards to average depth, depth difference, and track length.
In general, there are few easily identifiable clusters, but when example trajectories from each cluster are plotted, the clustering starts to make some sense. Included with the clustering is a measure of how many points are clustered, and the means of variables for each cluster. While in some of the cases there are more noise points than clustered points, the important part is that the clusters are identifiable and consistent.

4.5.1 Reference fish 41

Day

Clustering for reference fish 41 during the day can be seen in figure 4.9. Of the 888 trajectories in this dataset, 42.06 % of points belong to two clusters.
Clusters look somewhat believeable, especially cluster 1. Cluster 1 (4.11) consists of 355 trajectories, and shows circular behaviour. Cluster 0 is marked with orange in the figure (4.9), and consists of 24 trajectories at a mean depth of 19 meters(4.10). These trajectories does not resemble neither feeding nor circular swimming, but rather somewhat leisurely somewhat circular behaviour. Angle change, track length, and depth diff are similar across the two clusters, but the cluster 1 is a lot higher in the water, which seems to be the primary clustering factor in this case.

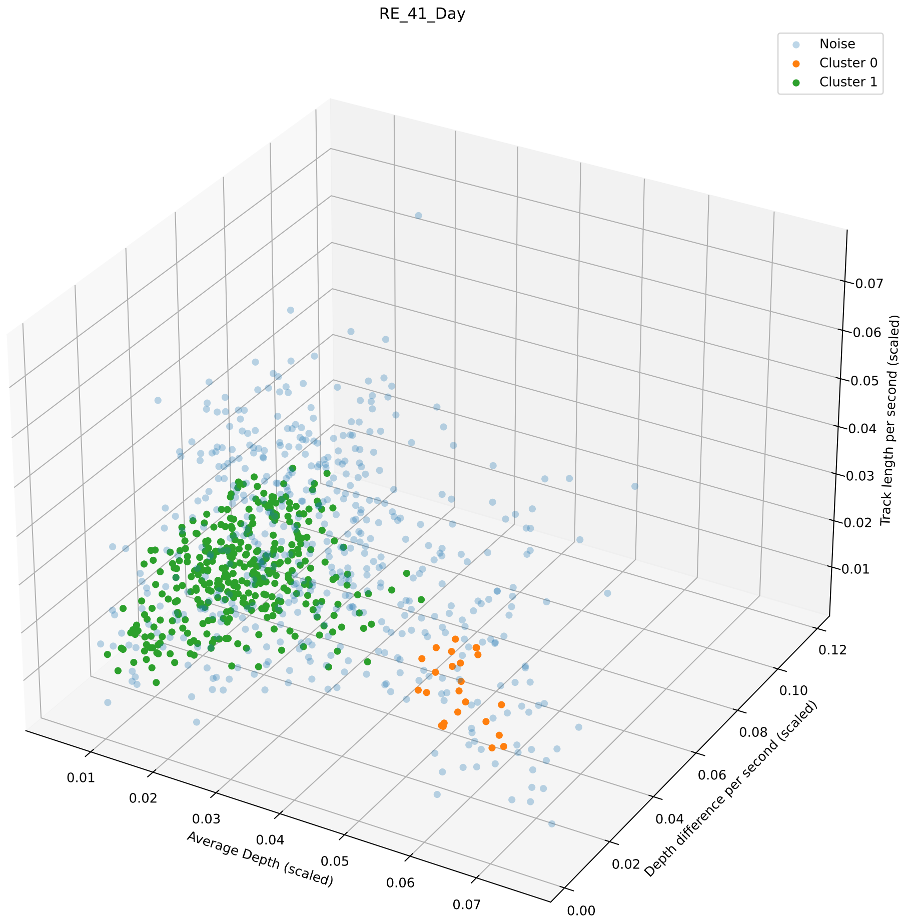


Figure 4.9: Clustering for reference fish 41 during the day. 888 trajectories in total. 42.6% of points belong to a cluster. Cluster 0 contains 24 trajectories, and cluster 1 contains 355 trajectories.

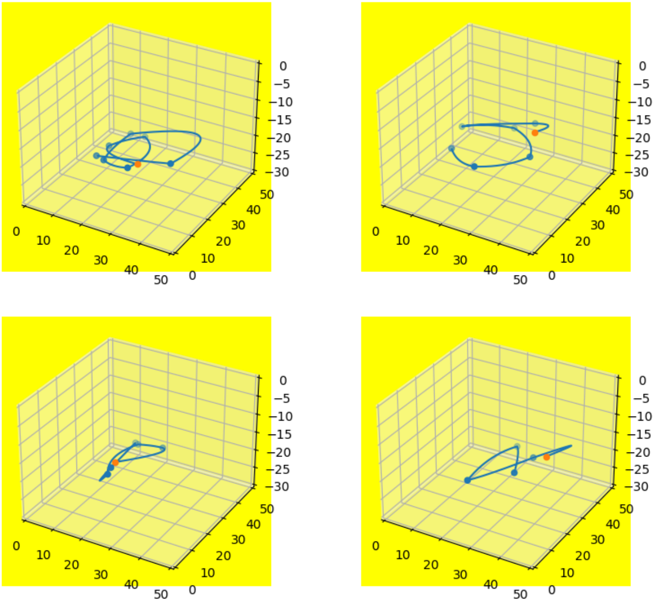


Figure 4.10: Cluster 0: 24 trajectories. Mean depth: 19.28 Mean depth diff: 0.0086 Mean track length: 0.20 Mean angle change 0.0178

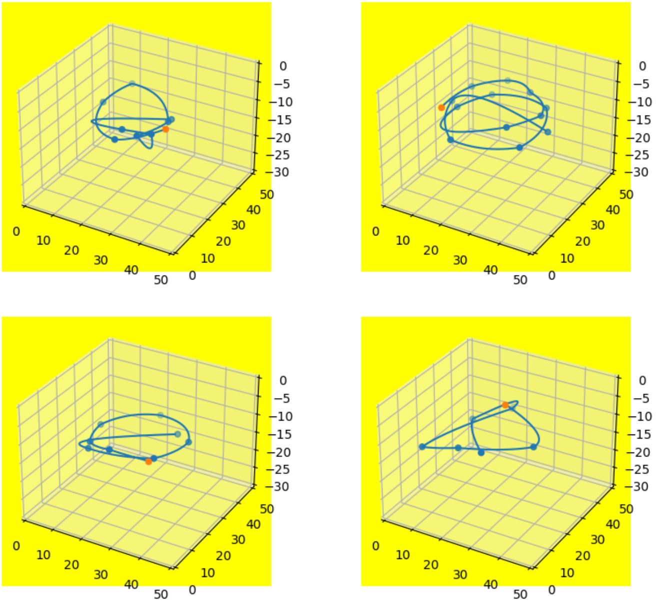


Figure 4.11: Cluster 1: 355 trajectories. Mean depth: 7.85 Mean depth diff: 0.0095 Mean track length: 0.27 Mean angle change 0.0223

Night

Clustering for reference fish 41 during the night can be seen in figure 4.12. Of 2540 trajectories in the dataset, 54% of points were clustered into two clusters.
Cluster 1 is larger than cluster 0, encompassing 1312 trajectories, which is about half the dataset. This cluster seems to capture slow, somewhat circular swimming, and close to the surface. From figure 4.12, it is also evident that this cluster is the dominating one in the dataset. Cluster 0 has 61 trajectories, and with about 50 % higher mean depth diff and more than twice the average track length this behavioural mode describes a somewhat more frantic, deeper movement than the category 1 (4.13). Category 1 is also a lot higher in the water, as can be seen in figure 4.14.

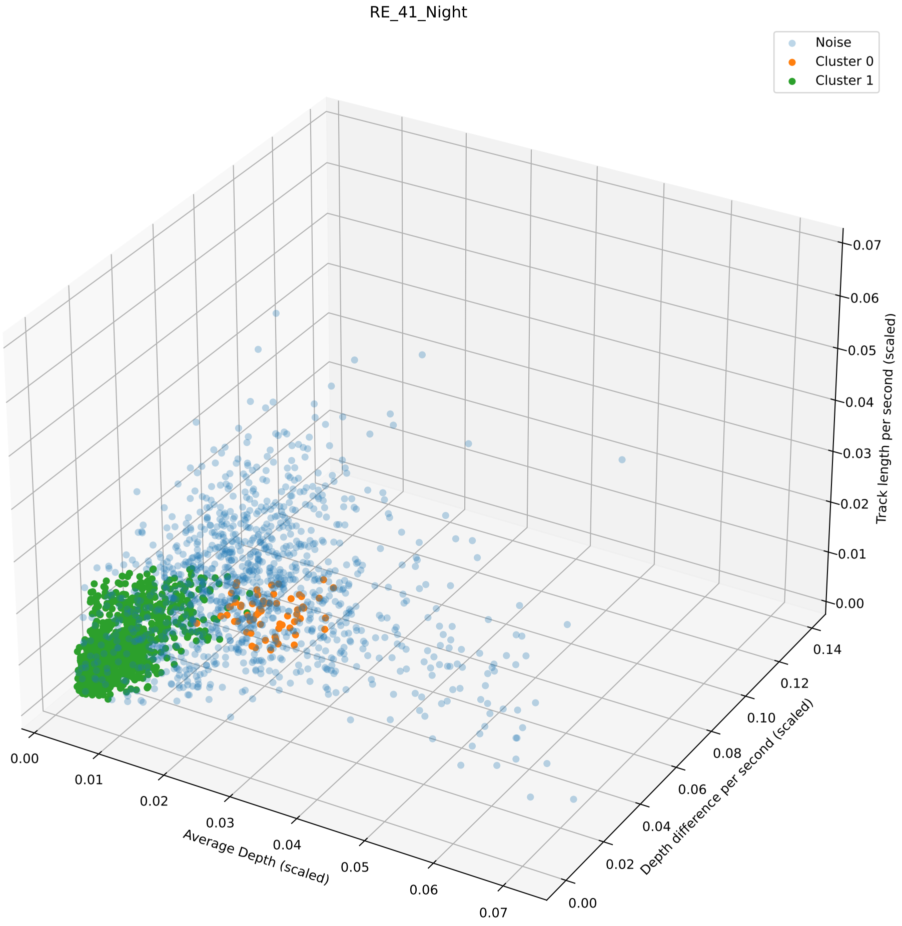


Figure 4.12: Clustering for reference fish 41 during the night. 2540 trajectories in total. 54% of points belong to two clusters. Cluster 0 contains 61 trajectories and cluster 1 contains 1312 trajectories.

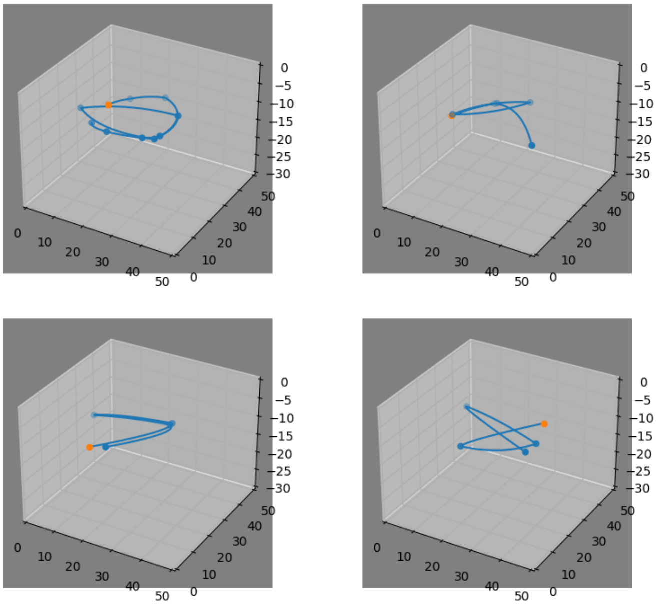


Figure 4.13: Cluster 0: 61 trajectories. Mean depth: 10.68 Mean depth diff: 0.0064 Mean track length: 0.28 Mean angle change 0.0202

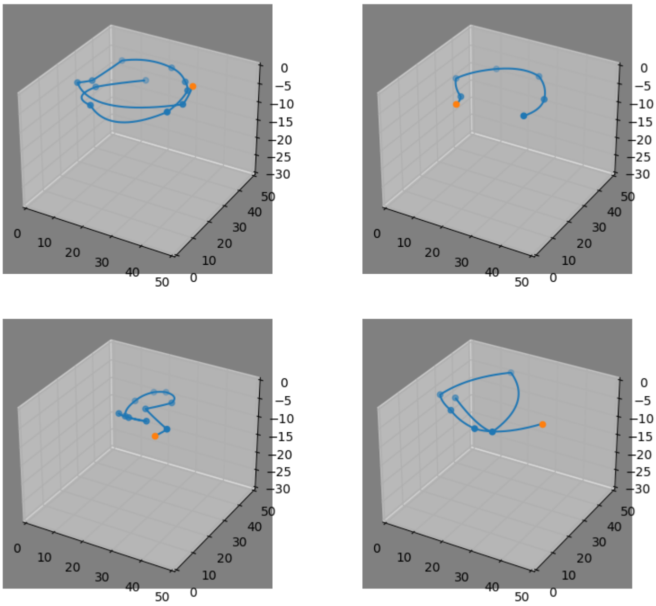


Figure 4.14: Cluster 1: 1312 trajectories. Mean depth: 2.61 Mean depth diff: 0.0042 Mean track length: 0.12 Mean angle change 0.0202

4.5.2 Reference fish 20

Day

Clustering for fish 20 during the day can be seen in figure 6.5. Of the 640 trajectories, 45.63 % of points were clustered into two clusters.

The reference fish have fewer datapoints in daylight, as the measurements were taken in winter. Nevertheless, two clusters emerge. The cluster 0 (4.15) has 277 members, and with a mean depth of 11.56 meters and a relatively long track length per second this seems to involve circular swimming, albeit with a somewhat low distance from centre. The cluster 1 (4.16) has less than half of the first category's track length, but somewhat greater mean angle change. This behaviour seems to translate to some kind of rickety back-and-forth movement. However, this cluster is relatively minor, with only 15 members. From the figure 6.5, it also seems somewhat random or noisy.

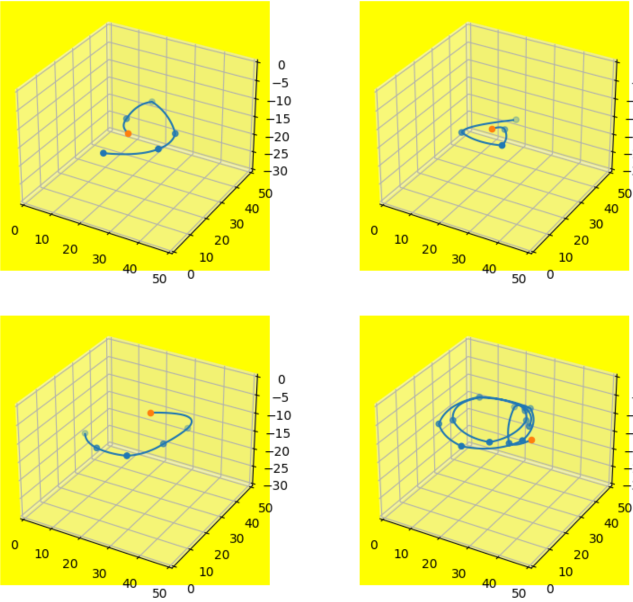


Figure 4.15: Category: 0 : 277 trajectories Mean depth: 11.56 Mean depth diff: 0.0132 Mean track length: 0.49 Mean angle change 0.0366

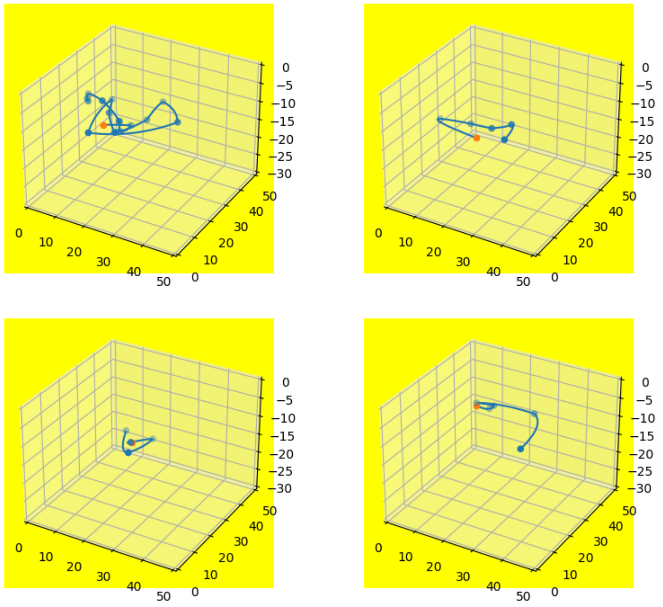


Figure 4.16: Category: 1 : 15 trajectories Mean depth: 8.08 Mean depth diff: 0.0171 Mean track length: 0.22 Mean angle change 0.0405

Night

Clustering for fish 20 during Night can be seen in figure 6.6. Of the 2439 total trajectories, 5.49 % of points were clustered into two clusters,

At night, as illustrated in 6.6, there is a clear bean-shaped distribution, similar to the cluster in 4.12. Inside this bean, HDBScan has found two clusters. Both are very deep, but this is not surprising, given this fish histogram in figure 4.2. This fish likes to swim deep and short at night, while shallow and long in the day. Both categories are very similar, as illustrated in figure 6.7 and 6.8, and it does not really make sense to split them apart. Here, it would probably be beneficial to change the `min_cluster_size` parameter og HDBScan in order to classify the whole 'bean' as one category. Nevertheless, the 5.49 % of clustered trajectories show deep, somehat slow, semi-circles. These might of course be part of full circles, but the trajectories cut off when time between measurements is greater than one minute.

4.5.3 Reference fish 21

Day

Clustering for fish 21 during day can be seen in figure 6.9. Of the 632 total trajectories, 51.11 % were clustered into two clusters
Cluster 1 (4.18) is the biggest of the two trajectories found, encompassing 306 trajectories. These have greater depth, twice the mean track length, and about 2/3's of the mean angle change of cluster 0. Visually, they resemble half circles. Cluster 0 (4.17) is comprised of only 17 trajectories, but they all have very low mean depth, somewhat short track length and somewhat high angle change.

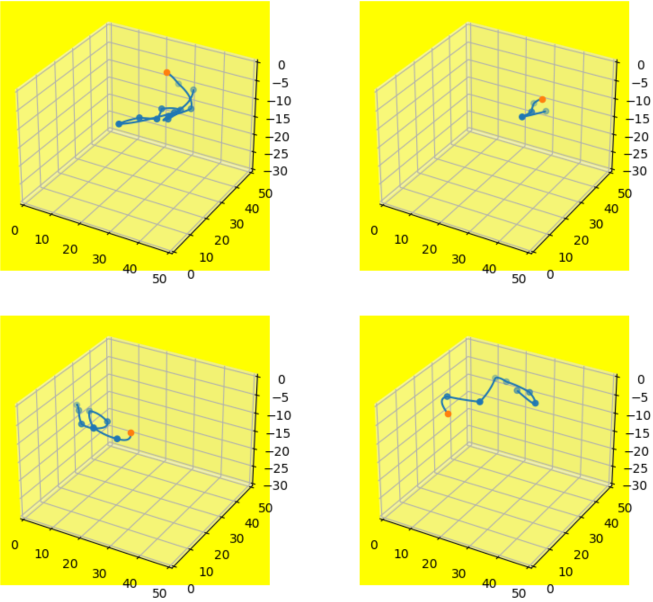


Figure 4.17: Category: 0 : 17 trajectories Mean depth: 3.82 Mean depth diff: 0.0085 Mean track length: 0.19 Mean angle change 0.0329

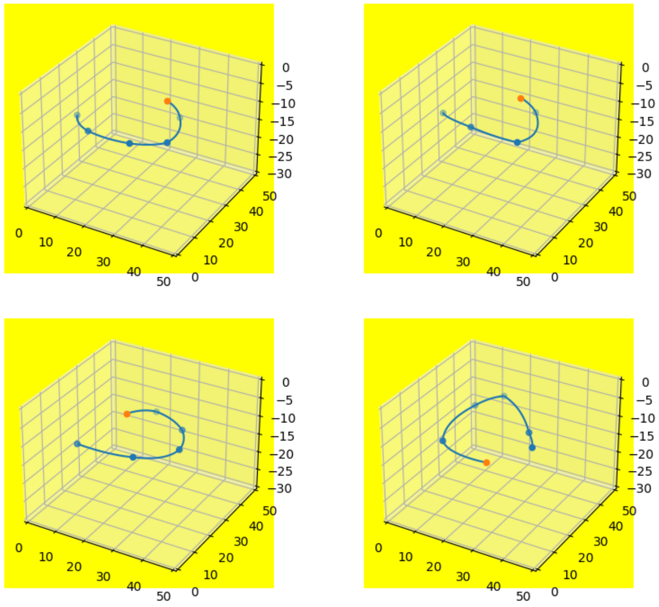


Figure 4.18: Category: 1 : 306 trajectories Mean depth: 10.04 Mean depth diff: 0.0098 Mean track length: 0.42 Mean angle change 0.0265

Night

Clustering for fish 21 during night can be seen in figure 4.19. Of the 2753 total trajectories, 92.59 % were clustered into two clusters.
At night for fish 21, we see in figure 4.19 a similar plot to 4.12. This fish likes to swim short and shallow at night, something that is evident both in cluster 1 (4.21) and in the histogram 4.3. 2505 out of 2753 trajectories belong to cluster 1, meaning 90.09 % of the dataset belongs to that category. Looking at figure 4.19, this is believeable. Category 0 however, with 44 members, captures a behavioral mode with deeper and faster swimming with higher angle change and depth difference.

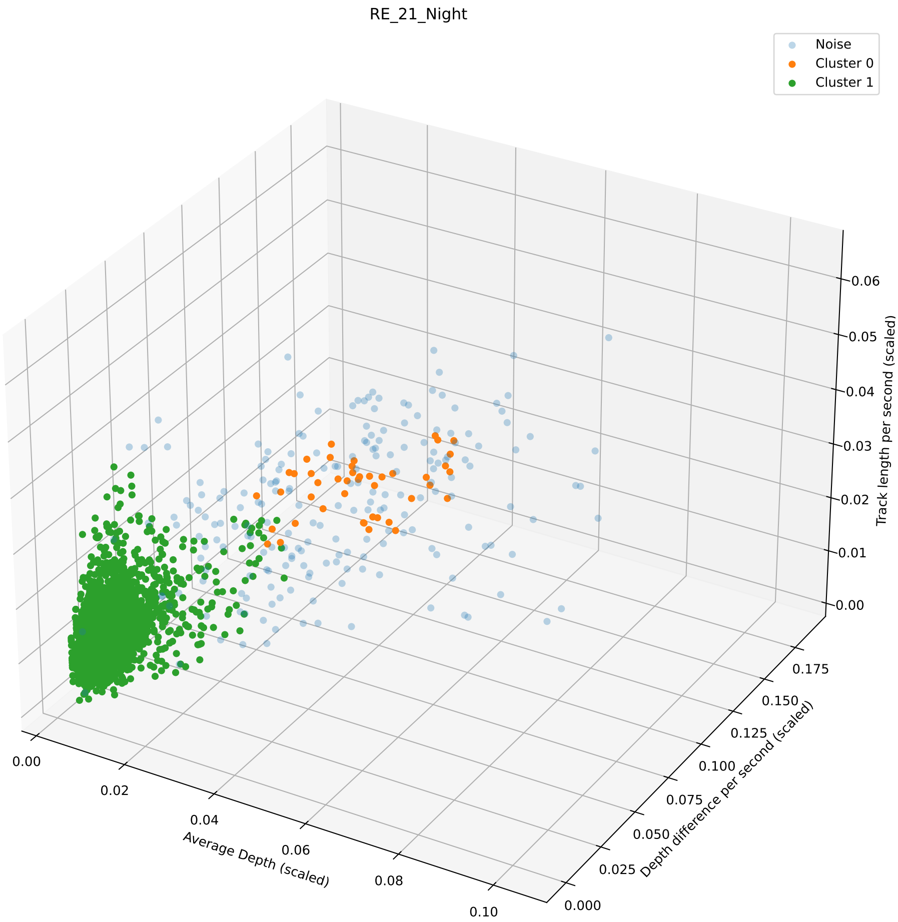


Figure 4.19: Clustering for fish 21 during Night. 2753 trajectories in total. 92.59 % of points belong to two clusters. Cluster 0 contains 44 trajectories and cluster 1 contains 2505 trajectories.

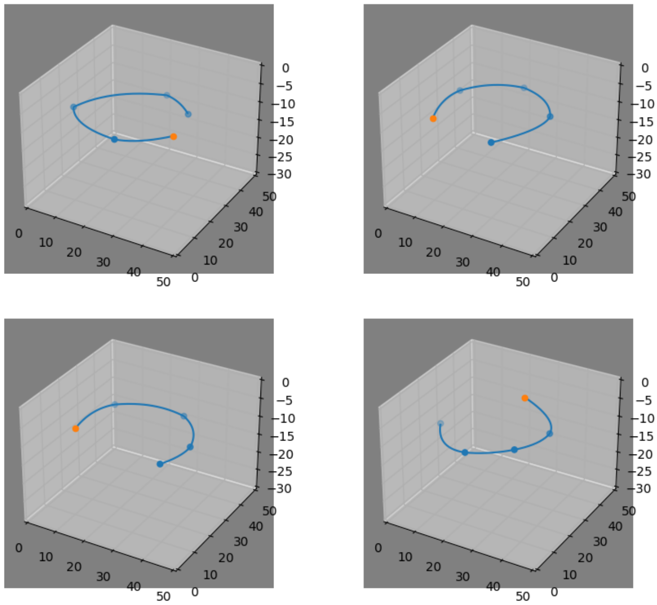


Figure 4.20: Category: 0 : 44 trajectories Mean depth: 8.92 Mean depth diff: 0.0112 Mean track length: 0.52 Mean angle change 0.0308

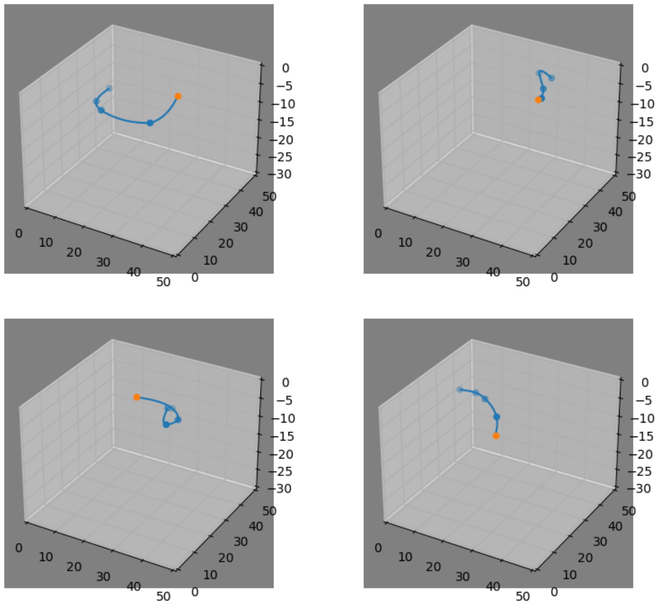


Figure 4.21: Category: 1 : 2505 trajectories Mean depth: 1.47 Mean depth diff: 0.0034 Mean track length: 0.16 Mean angle change 0.0262

4.5.4 Aquatraz fish 23

Day

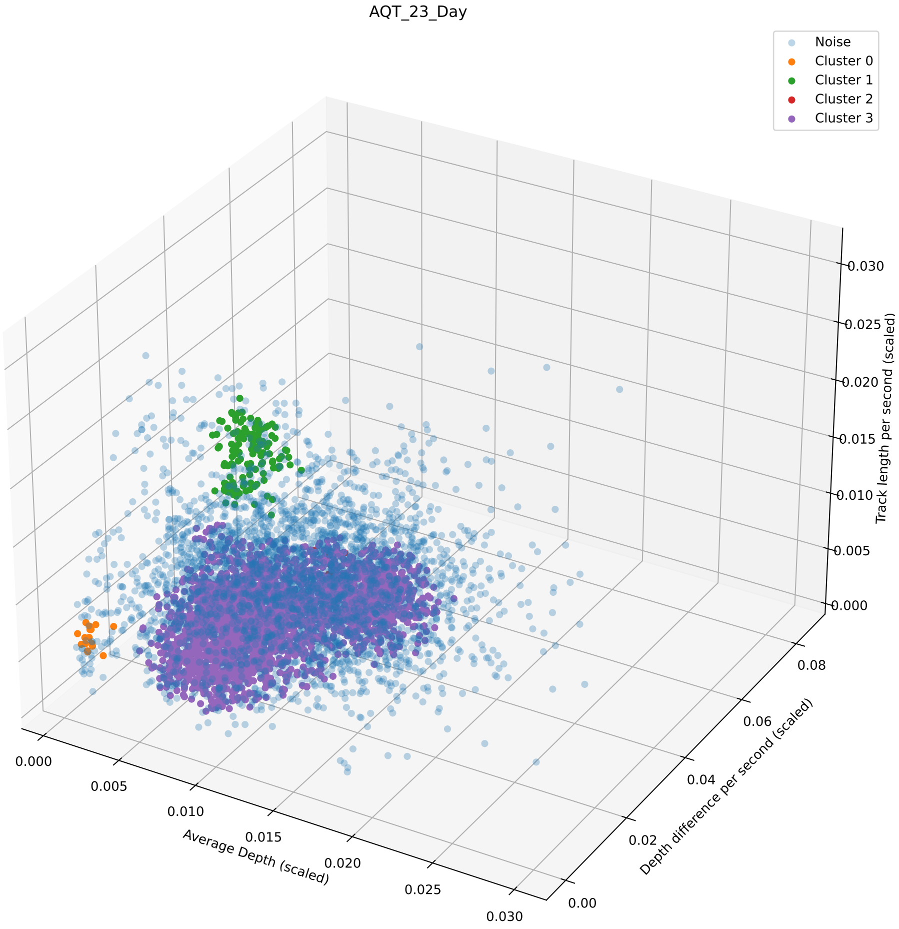


Figure 4.22: Clustering for fish 23 during Day. 5587 trajectories in total. 48.95 % of points belong to four clusters. Cluster 0 contains 16 trajectories, cluster 1 contains 134 trajectories, cluster 2 contains 15 trajectories, and cluster 3 contains 2570 trajectories,.

As is evident in figure 4.22, this fish has 3 humanly identifiable clusters. The fourth cluster is small and noisy, and behind the purple cluster. Almost 50 % of points are clustered of the 5587 total trajectories.
Cluster 3 is the biggest category here, and contains 2570 trajectories, which is almost half the dataset. These trajectories seem to describe high angle change, lower track length, and medium depth difference (4.26). The next largest cluster is cluster 1, with 134 trajectories,

which very clearly shows circular swimming (4.24), and has a high track length in comparison to other clusters. Category 0 has 16 members. With a very low mean depth, low track length, and high angle change, this trajectory might constitute feeding (4.23). This is further evidenced by the visible spike at low depths in the histogram in figure 4.4. The smallest category is somewhat mystical, as it describes a completely new mode of swimming, characterised by somewhat long track length and very high depth difference (4.25). This category has only 15 members.

An important factor in this specific case is that fish 23 day has by far the highest amount of trajectories, and this might have an impact on performance.

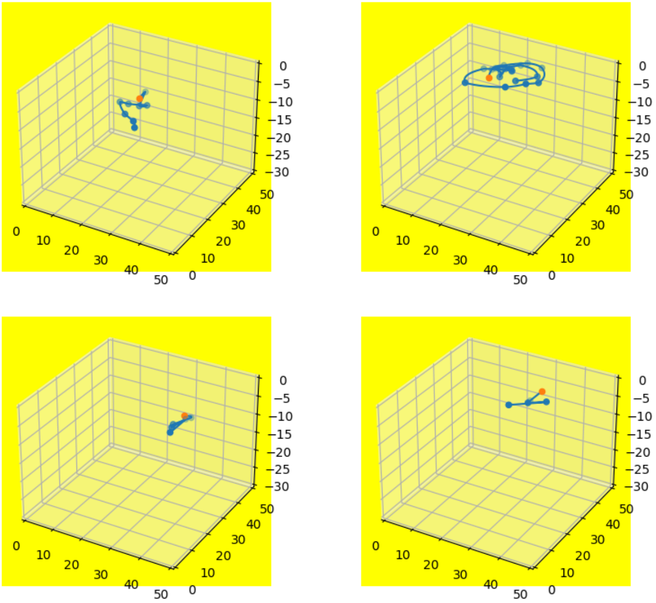


Figure 4.23: Category: 0 : 16 trajectories Mean depth: 1.38 Mean depth diff: 0.0068 Mean track length: 0.13 Mean angle change 0.0421

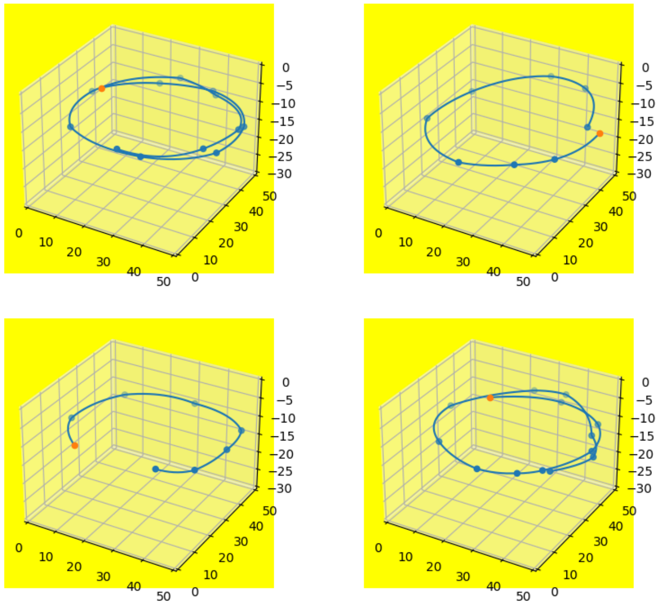


Figure 4.24: Category: 1 : 134 trajectories Mean depth: 11.73 Mean depth diff: 0.0158 Mean track length: 0.54 Mean angle change 0.0216

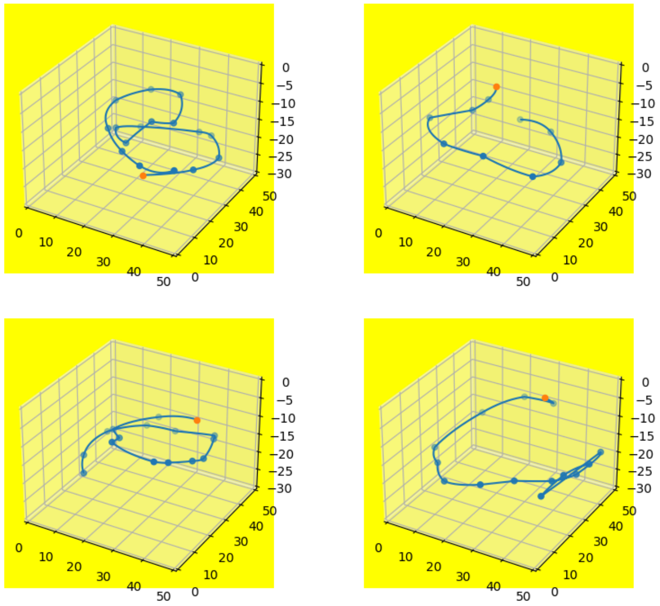


Figure 4.25: Category: 2 : 15 trajectories Mean depth: 15.52 Mean depth diff: 0.0372 Mean track length: 0.31 Mean angle change 0.0280

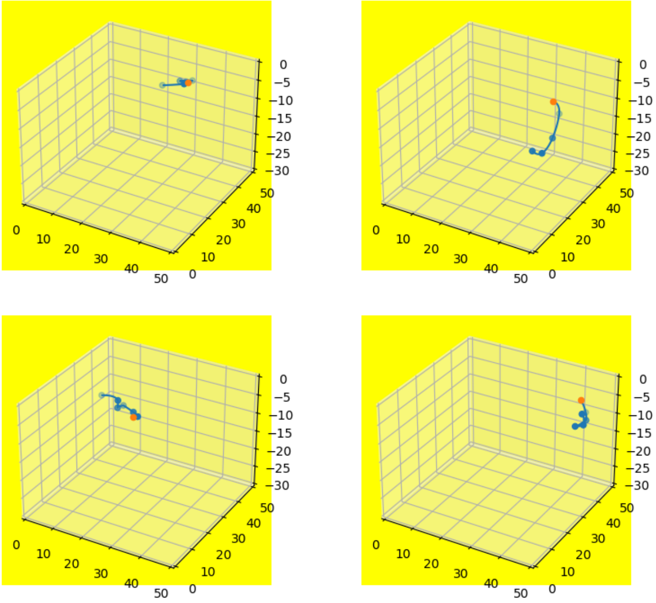


Figure 4.26: Category: 3 : 2570 trajectories Mean depth: 13.03 Mean depth diff: 0.0210 Mean track length: 0.24 Mean angle change 0.0317

Night

In figure 4.27, three clusters are present, clustering approximately 53.08 % of the data. Cluster 0 and 1 seems small and somewhat noisy, while cluster 2 contains a large part of the data. Category 2 (4.30), consists short, slow, high angle change trajectories with medium depth difference. This category contains about half of the dataset. Category 0 (4.28) seems to describe circular swimming, while category 1 (4.29) seems to contain slower, more noisy circular swimming.

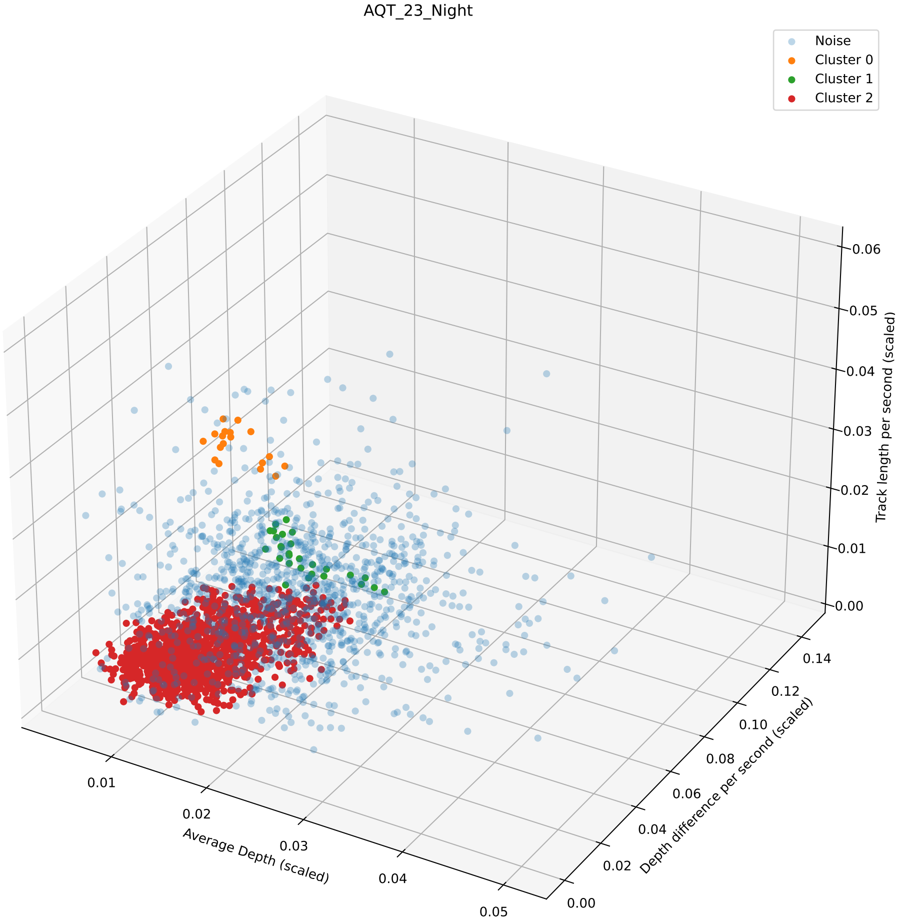


Figure 4.27: Clustering for fish 23 during Night. 2334 trajectories in total. 53.08 % of points belong to three clusters. Cluster 0 contains 18 trajectories, cluster 0 contains 28 trajectories, and cluster 1 contains 28 trajectories, and cluster 2 contains 1193 trajectories,

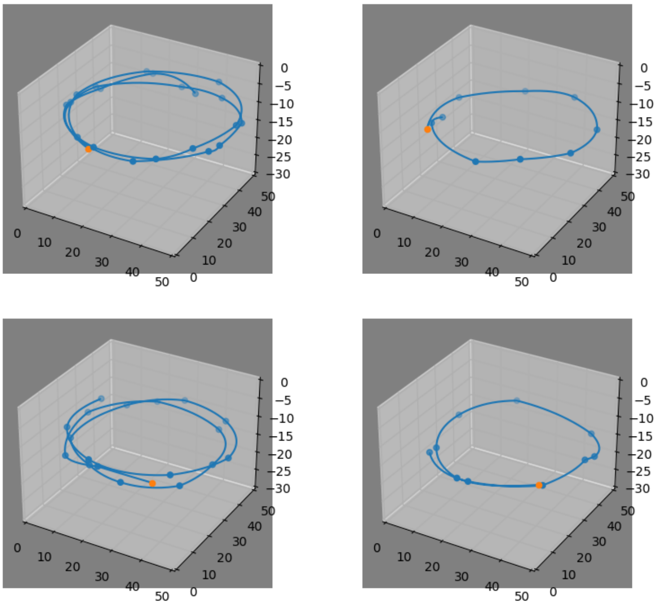


Figure 4.28: Category: 0 : 18 trajectories Mean depth: 11.25 Mean depth diff: 0.0148 Mean track length: 0.52 Mean angle change 0.0217

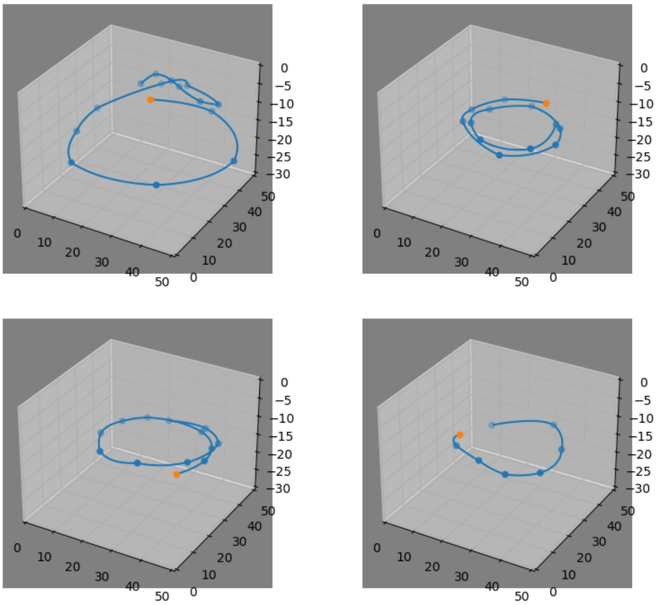


Figure 4.29: Category: 1 : 28 trajectories Mean depth: 14.68 Mean depth diff: 0.0192 Mean track length: 0.35 Mean angle change 0.0270

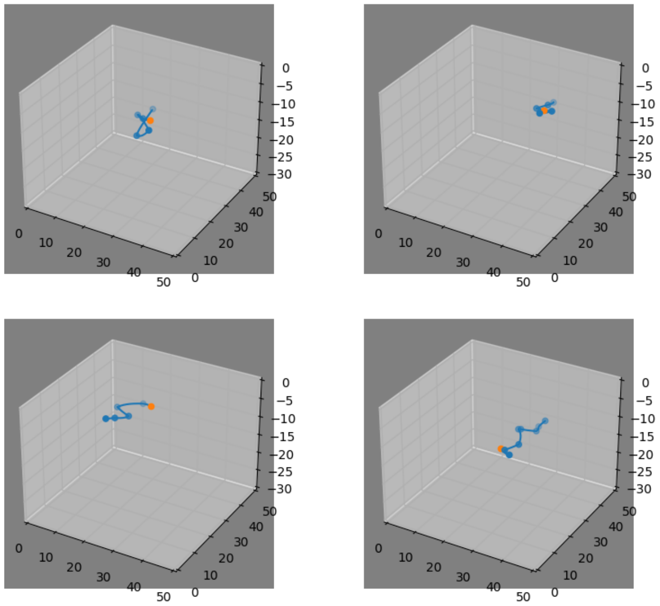


Figure 4.30: Category: 2 : 1193 trajectories Mean depth: 8.90 Mean depth diff: 0.0145 Mean track length: 0.14 Mean angle change 0.0373

4.5.5 Aquatraz fish 14

Day

32.33 % of 2069 trajectories were clustered into two clusters, as shown in figure 6.10. Cluster 1 is the biggest of the two clusters found, containing 455 trajectories. It is far deeper than cluster 0 (which is also reflected in the histogram 4.5), faster, and with lower angle change(4.32). Cluster 0, includes 214 trajectories with a low mean depth, low track length, and high angle change (4.31). Visually, they could be characterised as ”squiggly”. Depth difference is about the same for both. The second cluster describes decidedly more circular swimming.

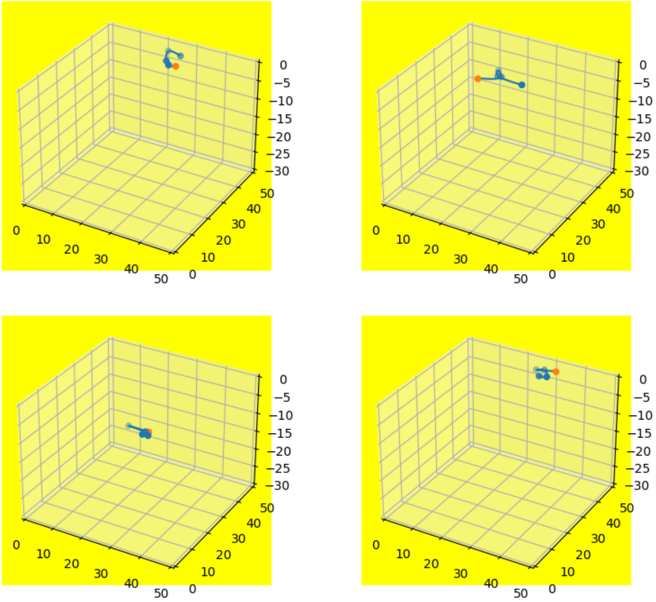


Figure 4.31: Category: 0 : 214 trajectories Mean depth: 4.96 Mean depth diff: 0.0156 Mean track length: 0.15 Mean angle change 0.0355

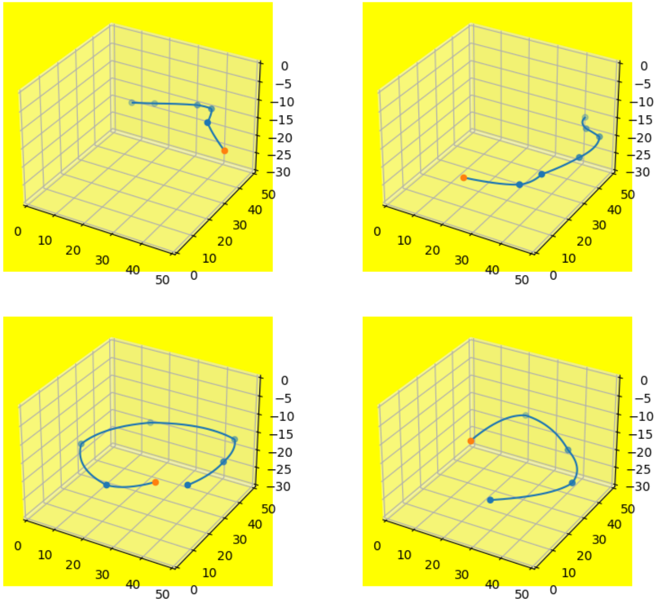


Figure 4.32: Category: 1 : 455 trajectories Mean depth: 19.04 Mean depth diff: 0.0180 Mean track length: 0.36 Mean angle change 0.0217

Night

Clustering for fish 14 during night can be seen in figure 6.11. Of the 941 total trajectories, 39.64 % were clustered into two clusters.
The cluster 0 contains 285 trajectories, with a low mean depth, low depth diff, slow swimming with high angle change (4.33). This is similar to fishes 41, 21, and 23. However, it is higher in the water than the reference fish. Cluster 1 consists of 88 trajectories, and has a much deeper mean depth. These trajectories have higher track length, lower mean angle change, higher depth diff, and are more circular (4.34).

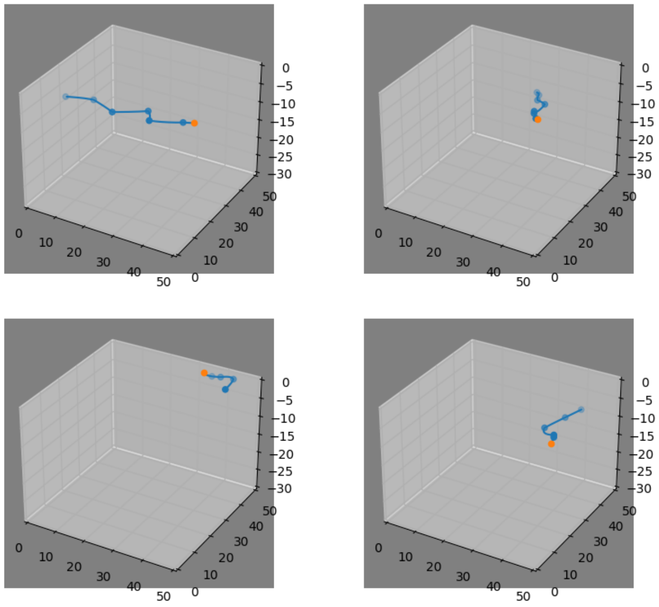


Figure 4.33: Category: 0 : 285 trajectories Mean depth: 5.66 Mean depth diff: 0.0096 Mean track length: 0.11 Mean angle change 0.0381

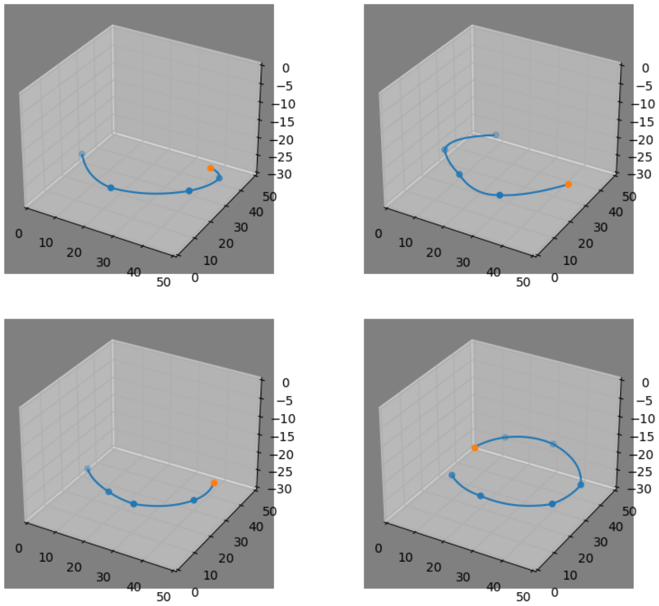


Figure 4.34: Category: 1 : 88 trajectories Mean depth: 17.84 Mean depth diff: 0.0146 Mean track length: 0.31 Mean angle change 0.0222

4.5.6 Reference fish 15

Day

Clustering for fish 15 during day can be seen in figure 6.12. Of the 2753 total trajectories, 43.73 % were clustered into two clusters.
Category 1 is by far the largest, consisting of 1123 trajectories. They have a deeper mean, higher speed, same depth diff and lower angle change (4.36) than cluster 0. The trajectories are semi-circular. Cluster 0 consists of 81 squiggly trajectories with low track length, high angle change and high mean depth (4.35).

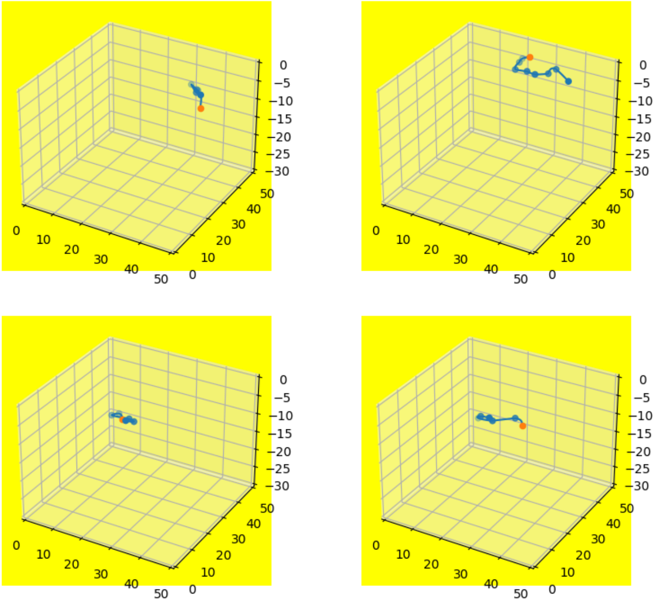


Figure 4.35: Category: 0 : 81 trajectories Mean depth: 5.78 Mean depth diff: 0.0144 Mean track length: 0.12 Mean angle change 0.0368

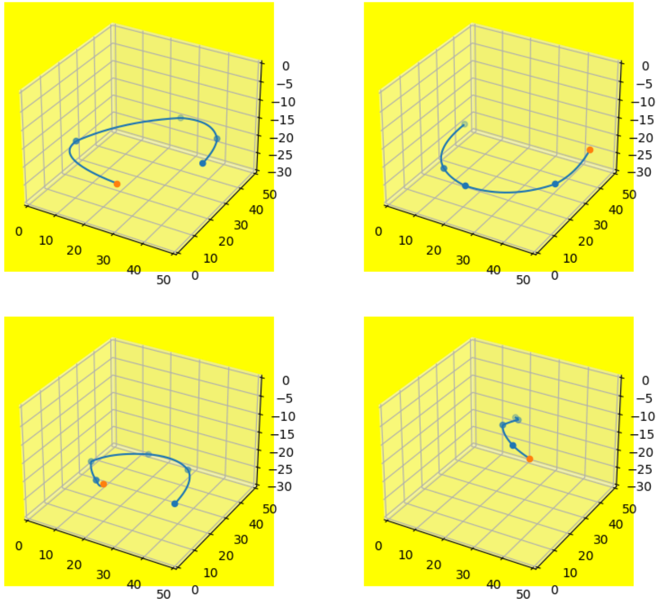


Figure 4.36: Category: 1 : 1123 trajectories Mean depth: 16.59 Mean depth diff: 0.0149 Mean track length: 0.35 Mean angle change 0.0211

Night

Clustering for fish 15 during night can be seen in figure 4.37. Of the 1110 total trajectories, 37.48 % were clustered into three clusters.

Cluster 2 shows short, slow, high in the water squiggly swimming patterns and is the largest, consisting of 344 trajectories (4.40). Cluster 1, contains 46 trajectories, and show deep, medium speed, circular paths (4.39). Cluster 0 shows longer trajectories with higher mean depth and medium angle change, and consists of 26 trajectories (4.38).

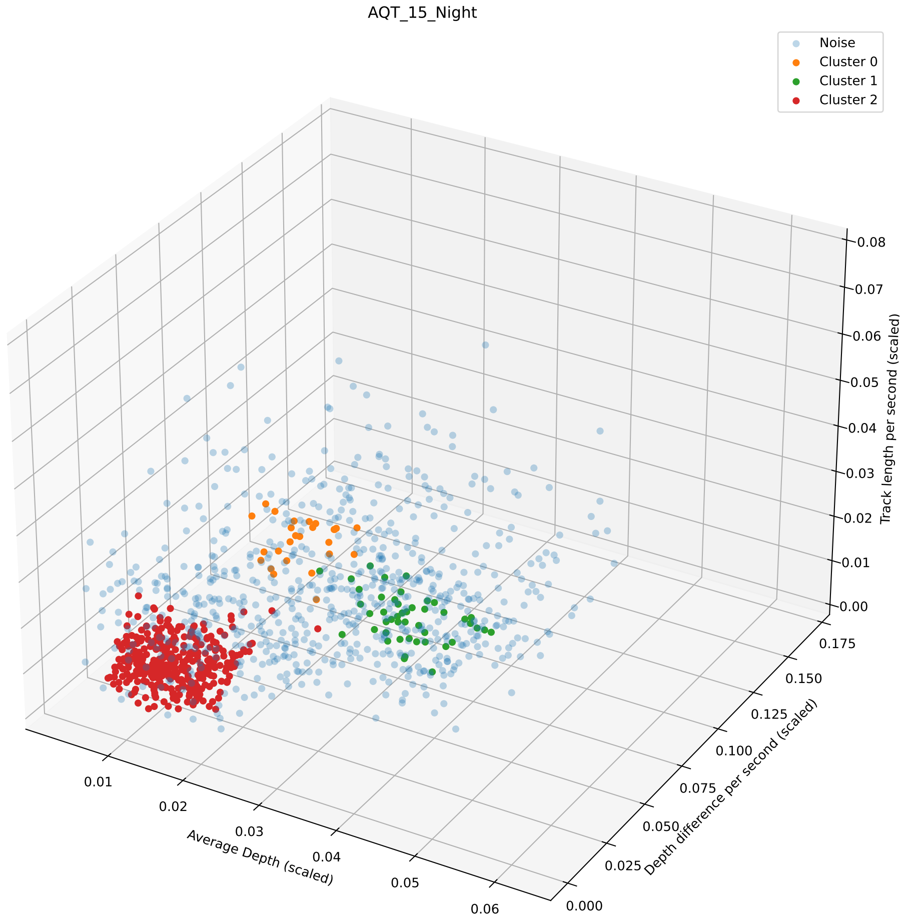


Figure 4.37: Clustering for fish 15 during Night. 1110 trajectories in total. 37.48 % of points belong to three clusters. Cluster 0 contains 26 trajectories, cluster 1 contains 46 trajectories, and cluster 2 contains 344 trajectories.

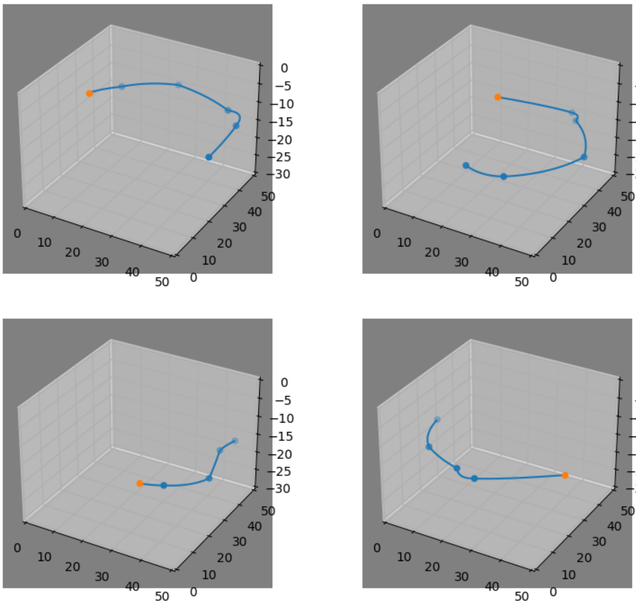


Figure 4.38: Category: 0 : 26 trajectories Mean depth: 13.54 Mean depth diff: 0.0170 Mean track length: 0.40 Mean angle change 0.0233

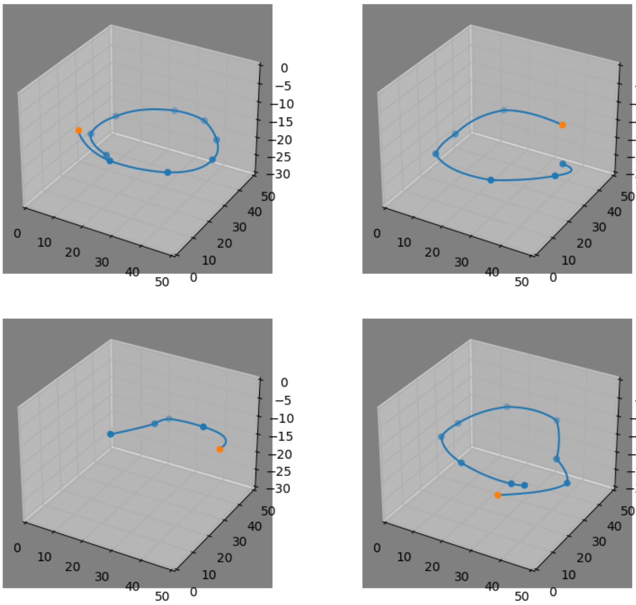


Figure 4.39: Category: 1 : 46 trajectories Mean depth: 19.85 Mean depth diff: 0.0164 Mean track length: 0.32 Mean angle change 0.0235

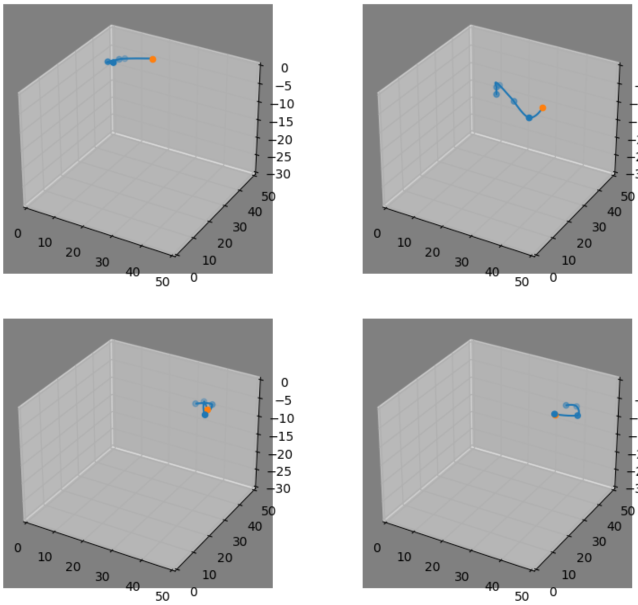


Figure 4.40: Category: 2 : 344 trajectories Mean depth: 6.40 Mean depth diff: 0.0106 Mean track length: 0.12 Mean angle change 0.0334

5

Discussion

5.1 Mean and Variance

Average distance from center is greater in aquatraz. This might be because of different fish density in aquatraz compared to the reference cage (fish densities for the different cages are not specified) or perhaps the aquatraz cage generates a rotational current that influences distribution. Lower average depth at night for almost all fish makes sense, as the salmon is a visual predator that seeks the light. This corroborates findings such as Oppedal et al. (2007), Oppedal et al. (2011), Ulvund et al. (2021). As the aquatraz measurements are summertime measurements and the reference fish are measured in wintertime, it makes sense that the similar light conditions between day and night in the Norwegian summer lessen the depth difference between night and day for the aquatraz fish. Track length and depth difference being greater at night also makes sense, as the salmon is generally more active in the day, as is also shown in Føre et al. (2018b). The fact that angle change does not show the same night and day dependency might imply that angle change is not related directly to activity, e.g. that changing angle at low speeds is relatively ”cheap” for the fish and therefore not a measure of energy expenditure, or that it is related to other factors. Distance moved relative to center is similar for night and day, but somewhat higher in aquatraz. Given that the fish generally stay closer to the edges of the cage in aquatraz, this might just be an artifact of this specific behaviour.

In general, the means and variances of the aquatraz fish are more homogenous than their reference fish counterparts. This might be due to a number of factors. For one, we know approximately how big the aquatraz fish were when they were tagged, while the reference fish sizes were estimated from sampling. Because many of these variables are scaled by length to provide more meaningful metrics, a precise fish length measurement is important for consistent results. Another factor is that the aquatraz cage might provide a more homogenous environment, as it pumps up and circulates water from below the cage, which might provide more direct control of environmental factors such as lice, currents, and algae.

In conclusion, there are definitive consistencies in the data between fish. Average depth, track length, and depth difference are greater at day and lower in the night. Angle change seems to have great individual differences, but distance from center is similar in night and day for all fish. Distance from center moved is less similar, but has variance is at least consistent within cages. The fact that average depth, depth difference, and track length are more consistent between cages and between night and day might imply that these are more statistically significant, and more important in the further analysis.

5.2 Histograms of Variables

5.2.1 Reference Fish

The histograms clearly show what was already suspected with regards to average depth, depth difference, and track length for fishes 41 and 21. This also agrees with earlier findings, which show that fish swim deeper at night (Fernö et al. (1995)). The distributions have lower mean and variance, with very prominent peaks. Fish 20, however, prefers to swim deeper in the water, even in the night. This contradicts earlier observations on ”normal” fish behaviour (Ulvund et al. (2021), Oppedal et al. (2007)), and is likely to imply that fish 20 is somewhat of an outlier. Another interesting observation can be made on the recreated angle histogram of fish 20. The mean and variance table (4.1) shows that the variance at night is smaller, but the night distribution is significantly flatter at night. This is an artifact of the way sample variance is calculated. There are significantly more datapoints at night (2753 vs 632), which makes the variance smaller, as it is divided by number of datapoints -1, as described in the theory.

Variables pertaining to distance from center moved are similar for night and day for all cases. One could argue that a similar distribution is a good sign, since it signifies that fish behave similarly between cages and individuals. However, given that the distributions usually look like skewed normal distributions, it is more likely that they just represent unnecessary noise for the algorithms. Indeed, satisfactory clustering was only achieved when these variables were not used. This may signify that distance from center average and moved are inefficient in defining and dividing fish trajectories and behaviour.

Angle change is somewhat lower at night for all fish, but distributions are different, with the most similiarities between night and day found for fish 41, and fish 20 and 21 having higher mean and lower spread, respectively.

In general, average depth, depth difference, and track length per second has the highest differences between night and day. Another interesting point is that the average depth distribution seems to be the variable that changes the most from individual to individual, even if this was not readily apparent from the mean and variance table.

5.2.2 Aquatraz Fish

For all aquatraz fish, a low depth spike (spike at > 3 m) can be seen at day in the average depth histogram. This spike is more distinct in the histogram for fish 23, as this fish does not spend any time above about two meters at night. This spike might very well indicate feeding, as this only takes place at day.

Depth difference per second is more similar both between night and day and between fish for aquatraz. This might be a result of the aforementioned more homogenous environmental factors, or the fact that light conditions are more similar between night and day for the aquatraz data, as this data is from the summer.

Variables pertaining to distance from center are also similar across fish and night/day, and do not seem to provide too much information.

Track length per second does have a clear spike at lower depths at night for all fishes however, and the spike at night at lower average depth is also present for fishes 14 and 15, similar to fishes 41 and 21.

Angle change, however, seem to be lower at day and higher at night. This is different from the reference fish, but makes sense given the mean and variance table (4.1).

5.2.3 Histograms in general

For all the histograms, the two variables pertaining distance from center are the most similar in night and day. While this does not necessarily imply that they are ineffective in defining behaviour, the assumption that behaviour is different in day and night, and the fact that these are very similar in day and night, might imply that they represent unnecessary noise that will confuse clustering algorithms. Interestingly, the average depth histograms are very different for all the fish (expect 14 and 15, where the histograms are only somewhat different). Moreover, even if the means of average depths show the same tendencies (lower at night, and they are very similar for aquatraz), the histograms show that the distributions are quite different. Studies have shown that salmon seek the light, either from the sun, moon or underwater lights, when swimming (Oppedal et al. (2007), Ulvund et al. (2021)). However, studies have shown that social hierarchy is important for weight gain (Cubitt et al. (2008), Cañon Jones et al. (2010)), and it might not be unlikely that the social hierarchy position of the salmon has an impact on swimming depth. Another reason might be that some fish seek the light at the surface, and some seek underwater lights in the same cage as Fernö et al. (1995) states that fish are more evenly distributed across the cage volume in winter. This might explain the depth preference of fish 20, if there are underwater lights in the reference cage [1].

Depth difference per second histograms are consistent within cages, where as the shape of the day distribution is different between cages (higher mean and longer tail in reference cages). Angle change is also somewhat similar in most cases, albeit not as similar as the distance from center measurements. Track length per second is clearly dependent on day or night, however, the magnitude of discrepancy varies from fish to fish.

The histograms corroborate the findings of the mean and variance analysis in the way that average depth, depth difference per second, and track length per second are the variables with the most structure, and are therefore the variables most suitable for further analysis. Indeed, the distance from center variables and angle change consistently look like a somewhat skewed normal distribution. While this does not necessarily preclude these variables from contributing to a clustering analysis, it is not very promising.

[1] This was not a part of the input data and is to the authors knowledge unknown.

5.3 Scatterplots of Variables

5.3.1 Scatterplots in general

The greatest consistencies across these scatterplots are the lines/correlations of increasing track length with low but slightly increasing angle change, and the somewhat less obvious trend of total distance from center variance with low change in distance from center moved. These trend might imply circular motions, and the fish will move very little in distance from the center of the cage, but the total distance might be very different depending on how far away from the center of the cage the fish is swimming. Similarly, swimming in circles might imply a low rate of angle change, but the speed at wich the fish swims might vary greatly.

Depth difference per second and average depth plots are somwhat interesting in the fact that these seem to not have a clear structure, despite both being calculated from the same variable (z-position of fish). Depth difference per second seems to be evenly distributed across all depths, with some of the plots having peaks in the upper half of the cage, and others have no clear peak. In addition, fish 15, 23 and 20 all have clusters of low depth and low angle change not seen at night, which might imply feeding.

5.4 PCA Analysis

5.4.1 Reference Fish

Interestingly, the most important variable in PC0 is depth difference in all cases. However, average depth does not seem to be as important. This somewhat mirrors the findings in the scatterplots, as the depth profile of the fish seems to contain a lot of structure, especially when compared to the relatively homogenous distance from center variables. However, many of the principal components are hard to define in a behavioural context. For example, PC0 for fish 41 day (4.2) describes a correlation between high depth difference, lower average depth, higher distance from center, somewhat higher track length and angle change. One could argue that this might constitute somewhat circular movement, but why is the depth difference so large? Feeding arguments are hard to make for the same reason. Other PC's however, like PC1 for reference fish 20 day, seem to make more sense as circular behaviour. In this principal component, high total distance from center, low distance from center moved, and high track length is apparent. Angle is somewhat low, but this makes sense as angle change per second might very well be lower in circular swimming than in for example feeding. However, depth difference also has a relatively high impact on this principal direction, whereas we would have expected traditional circular swimming to be relatively stable depth-wise. Looking at feeding, no PC's seem to be obviously describing feeding. A feeding PC would most certaingly constitute low average depth, so PC1 and PC2 in reference fish 21's table (6.3) come to mind of one reverses the signs of the variable importances. However, in both these principal components, distance from center moved is the most important variable, albeit with different sign between directions. As feeding may consist of mainly short movements close to the center of the cage, assuming the feed is distributed from the middle, it is unlikely that these directions describe the feeding process. In addition, they explain about 26 and 14 % of the variance in the dataset, respectively. Without knowing how much time individual salmon spend feeding, it is hard to make any assessments on wether this is a realistic or unrealistic variance explanation. For example, if the fish spends 1 % if its time feeding, feeding might be buried in noise.

At night however, an interesting structure emerges. The histograms (4.1, 4.3) for fish 41 and 21 showed more prominent peaks in the first three variables. This structure is also apparent in the principal components for the night dataset. In the tables 4.4 and 4.12, the first principal component, PC0, has high contribution from average depth, depth difference, and track length, with very low contributions from other variables. This correlation even explains more than 60 % of the variance in the data for fish 21 at night. While this observation might not constitute a behavioural mode in and of itself, the observation at least shed some light into behavioural difference at night and day.

5.4.2 Aquatraz Fish

In this analysis as well, depth difference is the most significant variable in all PC0's. The same nightly trend as in fish 41 and 21 can be observed, albeit to a somewhat lesser degree, in fishes 14 and 15 (4.6, 4.8), with average depth, depth difference, and track length taking the top spots as most significant in PC0. This is also apparent in their depth histograms (4.5, 4.6). Interestingly, these fishes (14 and 15) are also pretty similar in their PC1, with low average depth and track length, in addition to high depth difference medium angle change. These fish are also pretty similar in the first three principal components in the day. This makes sense given that these two fish have the most similar histograms. Beyond this however, it is hard to ascertain what the other principal directions describe, if anything. It seems that even if fish 23 had the ”best” result on the clustering, this does not necessarily imply that the principal directions make a lot of sense.

5.4.3 PCA in general

Beyond this, however, most of the Principal directions are hard to interpret. Many of them include contributions from many of the variables in different magnitudes, and are therefore difficult to describe with words. However, the similarities of nightly PC’s are promising, as they indicate there is some similarity between the fish in behaviour. The day principal components however, are somewhat harder to describe.

Given the result of the clustering, it might have been beneficial to remove the distance from center variables, as they might not be as indicative of behaviour, at least not when clustering. With only four variables, it is significantly easier for a human to interpret the results as well. However, as stated in the method chapter, the goal of PCA analysis was to ascertain if this method did indeed show a clear partition of the data or not, not necessarily to extract the most explainable principal directions.

5.5 HDBScan Clustering

5.5.1 Reference Fish

Fish 41 and 21 have very clear behavioral modes in the night: spend most of the time in (presumably) energy-efficient, low speed and angle change trajectories high in the water column, with some time spent swimming faster and deeper. Fish 20, however, while it has similar track lengths, rather spends its time very deep in the cage at night, and had somewhat more circular swimming patterns. Fish 20 might be an outlier for its apparent preference of swimming very deep at night, or fish 41 and 21 might randomly behave similarly. Also, if the reference cage used underwater lights, some fish swimming at the depth of these lights is not surprising, as highlighted in Ulvund et al. (2021). Interestingly, the aforementioned category of nighttime behaviour for fish 41 ($>$ 50 % of nighttime trajectories) and fish 21 ($>$ 90 % of nighttime trajectories) with low track length, low track length, and low depth depth difference exactly mirrors the nighttime peaks evident in the histograms, and principal components extracted in the principal component analysis. Additionally, fish 41' and 21's behavioural categories swim a lot higher in the water during the night. This is consistent with histograms 4.3, 4.1 and the mean and variance table 4.1 . This is likely a result of the salmon seeking the light at night, something that is also apparent in studies such as Ulvund et al. (2021) and Oppedal et al. (2007).

All day datasets in the reference fish capture something that resembles circular swimming; category 1 fish 41, category 0 fish 20, and category 1 fish 21, which constitutes roughly half of all data points for their respective datasets. These are more circular for fish 41, and more half-circular for fishes 20 and 21. This might be a behavioural difference, but it is also likely a result of fish 41 having less average time between measurements, and therefore is more likely to have longer trajectories. Fishes 20 and 21 swim faster in their circular categories, and all fish are roughly in the top half of the cage ($>$14 m).

Only fish 21 has a behaviour readily resembling feeding in category 0, while fish 20 has something that might constitute feeding in category 1. These have 17 and 15 members, respectively, and contain high angle change low track length swimming patterns, with fish 21 being at an average depth of 3.82 meters, while fish 20 stays at an average depth of 8.08 meters in its suspected feeding category. However, this might not exclude this category from describing feeding behaviour. Some fish, that are perhaps not as big or aggressive as others, might look for feeding pellets that drop deeper down in the water, to avoid the frenzy up above. However, as mentioned in the PCA part, uncertainty with regards to how much time a fish spends feeding makes it hard to determine if a category size of about 2.5 % of the total dataset is realistic for amount of time spent feeding. Furthermore, as evident in the figures 6.9 and 6.5, there are more noise points than clustered points in the areas of the suspected feeding categories, and this might indicate that only a subset of these feeding-adjacent trajectories are classified. This in turn implies that doing an analysis based on exact amount of classified points might not be a significant indicator of clustering success. Qualitatively, one can asses that the fish probably spends more time in default behaviours that may constitute circular swimming patterns than it does feeding, in which case the clustering makes sense.

Additionally, all reference fish have clusters of circular or semi-circular swimming in the night as well (cluster 1 fish 41 night, cluster 1 for fish 20 night, and cluster 0 for fish 21 night). These clusters are generally smaller at night, but present.

5.5.2 Aquatraz Fish

In general, the three fish from aquatraz are more homogenous than the three reference fish. However, it is interesting to see the same nighttime behaviour as observed for fish 41 and 21: short, squiggly trajectories high in the water dominate nighttime classification, with somewhat higher angle change in aquatraz. The fact that these trajectories are very similar to the nighttime trajectories of fish 21 and 41, albeit deeper in the water, is probably a result of the light conditions in the norwegian summer. This, as mentioned, corroborates other findings indicating that salmon in cages seek light (Ulvund et al. (2021), Oppedal et al. (2007), Føre et al. (2018b)).

In addition, circular and semicircular swimming patterns are present in both night and day for all fish in aquatraz, in addition to ”squiggly” swimming closer to the surface in the day as well. This might be feeding, but the only fish that has a very clear spike in low depths is fish 23. However, the other fish might just wait for feeding pellets to drop down, insted of competing on the surface. If we assume feeding can happen at any depth, the categories that resemble feeding but at lower depth (category 0 fish 14 (4.5), category 0 fish 15 (4.7)) are larger in aquatraz . However, these trajectories are similar to the nighttime ”idling” already discovered in in the reference fish, with the same low track length ”squiggly” appearance. Additionally, inspection of the histograms in figures 4.5 and 4.6 reveals that both fish 14 and 15 have a spike at day at low average depth, further implying that these trajectories do not imply feeding.

Fish 23 has the perhaps best results in the clustering department. The humanly identifiable clusters present in figure 4.22 are almost certain to describe feeding (category 0), circular swimming (category 1) and ”idling” (category 4). Category 2 is somewhat more mystical, and seems to also consist of circular swimming, with more depth difference and somewhat shorter. Looking at the variables, average depth is lower and depth difference is higher for category 2 than category 1. Perhaps this category is born from the fact that some trajectories are very long, and the fish might change behaviour during the course of a trajectory. For example, the bottom right trajectory in figure 4.26 might constitute a trajectory that starts with circular behaviour, but ends up with a different behaviour. Another example if this, that the behavioural mode changes within the trajectory, is evident for fish 23 night in figure 4.29. The top left trajectory starts swimming in circles (orange point is the first datapoint), then changes swimming pattern to swim back and forth, with shorter track length between points. This is similar to the bottom left trajectory of category two in fish 23 day, as evident in figure 4.25. The fish starts in a somewhat circular trajectory, but turns around at the end of the trajectory. Another possible explanation is that these trajectories constitute a ”disturbed” circular swimming, that the fish are forced to change paths based on some local conditions, for example crowding. A third explanation is that these trajectories show some kind of stress response in the fish. At night, fish 23 exhibits two categories of circular swimming, and one category of ”idling”, similar to what has been seen in all other fish except fish 20.

As to the reason why this fish had the best results, ine reason is probably that this fish had by far the most data. With 120955 datapoints and the lowest average time between datapoints at 0.8 minutes fish 23 was definitely the most promising fish going into the analysis. However, that was not the only reason. The fact that fish 23 has a very clear spike at lower depths in figure 4.4, made the isolation of suspected feeding trajectories a lot easier. It was also the only fish that had humanly identifiable clusters. In addition to this, cluster 2 for fish 23 describes a behaviour not found in any other fish. This begs the question: does fish 23 behave ”simpler” than the other fish? Is it easier to determine the different behaviours? Perhaps it is somewhat of a ”winner” fish, which competes at the top for food pellets and swims wherever it wants, reducing the impact of other fish on its behaviour, which means there is less noise. The question can also be turned the other way: does fish 23 have advanced behaviour, as there are more significant differences between behaviour archetypes? More research may be needed in order to determine what makes a salmons behaviour recognizable.

5.5.3 Clustering in general

Interestingly, the pattern of slow-moving, somewhat low depth difference, and high angle change non-circular swimming in both the night and day was an unexpected find. Perhaps this is nighttime behaviour that bleeds into the day as a result of the trajectories being classified into day and night based on light levels, or perhaps this is just some type of default behaviour for the fish. It might even be a crowding response, as the fish has no clear way to swim in absense of the clear patterns of circular swimming, or it might be something entirely different. The main takeaway is that it is a discovered feature of the data, and constitutes and unplanned find.

In addition, circular/semi-circular swimming is present in all fishes for both night and day, and is in all cases except for fish 23 the dominating daytime behaviour. This was expected, but was not as prevalent or obvious as expected. However, the fact that circular trajectories are best described by long trajectories (long sequences of points with less than 60 seconds between), and the high average time between points results in a relatively small share of the total trajectories are likely to be identified as circular. Given that most trajectories are relatively short, the fact that circular swimming appeared so distinctly for fish 23 might simply be a result of its high data quality, and not necessarily a result of its individual behaviour. Another point that should be made is the fish density in the cages. This was not a part of the input data and are therefore not taken into account. It is not unlikely that the crowding in the cages would change behaviour. A study on cod in cages found that behaviour changed drastically when crowding increased (Rillahan et al. (2011)) . Increased crowding changed cod behaviour from independent swimming into schooling behaviour. Perhaps the more ”idle” category of low depth, depth difference and track length dissapears when crowding reaches a certain level.

Another avenue of research could be investigating multiple fish and looking into fish ”social status”, and their respective behaviour. One study has shown that in a feed-restricted environment, salmon initiating aggressive interactions had less fin damage, gained more weight and attained more central positions within the school, while fish receiving aggression had more fin damage and gained less weight (Cañon Jones et al. (2010)). Another study measured serotonin and its principal catabolite on Atlantic Salmon in commercial rearing conditions with and without reduced feed condition. This study found that even when food is in excess, some subordinate fish failed to grow, showing the impact of the social hierarchy on some individuals (Cubitt et al. (2008)).

5.6 Reflections on data and applied methods

5.6.1 Data

The data leaves something to be desired. While the recreated trajectories for circular and 'long' motion are likely to be somewhat accurate, in the cases where there are only a few meters between measurements, there is no guarantee that the trajectory recreation even closely represents actual fish movement. These point might be unclassifiable, or exist as noise between clusters, essentially drowning out the otherwise detectable clusters.

If time between measurements was lower and more consistent, other algorithms and approaches might have had results. Time series analysis is an example of a technique that might have had results on such a dataset. Additionaly, information about stressful events for the fish, such as thunderstorms and delousing, might have made for an interesting approach to stressor identification.

Other variables, like acceleration and heartbeat rate, might also have provided more insight into behaviour modes. However, these variables are dependent on significantly higher sample rate than the data provided.

Lastly, more precise measurement of fish length over time would have been beneficial for the analysis, as all movement is scaled based on fish length at the time of tagging, but it is expected that the fish grow in size and weight over the tagging period, especially the aquatraz fish, that weigh less than one kilogram on the start of the data period. Additionally, less mature fish might behave differently from more mature fish. As the aquatraz fish are signifiantly smaller than the reference fish, this may have had an impact on the analysis. Also, the reference fish and aquatraz fish have data from different parts of the year. This is likely to have had an impact on the analysis, as is indicated by e.g. Fernö et al. (1995), which observes that fish distribution is different between winter and summer.

5.6.2 Conventional methods

The positive aspect of the conventional methods, namely mean and variance analysis in addition to the histograms and scatterplots, are that they are easy to understand. These methods are linked to human perception, and the results usually make sense and need little explanation.

The drawback of conventional methods in general is that they do not detect unknown structure in data. Data is understood and analysed in accordance to what we already know, which might be a limiting factor. In addition to this, clustering in four or more dimensions is impossible for a human to visualise.

Specifically in this project, there are a few potential factors that may have resulted in worse results. One factor is that mean and variance are assumed constant over the data period. Another is that other information may have come to light if other variable partitions had been used for the scatterplots.

5.6.3 PCA

PCA provides a simple approach to explore covariance in the dataset. The positive aspect of PCA is that it may discover latent correlation structure in the data, which is often hard to

discover as a human, especially when the number of dimensions exceeds 3. The extracted principal components can be analysed for composition and explained variance level.
However, these directions do not always make sense to the human observer. As seen in the PCA section, the results of the analysis was not very clear beyond a few discoveries. PCA cannot discover non-linear structures, and it is sometimes hard to ascertain which directions are more important and which are just noise. Initially, PCA was used for dimensionality reduction as well, but the results of using just the four variables mentioned earlier for clustering were simply better.

5.6.4 HDBScan clustering

HDBScan is a potent algorithm, and it is adept at extracting clusters with varying density. The ability to filter out noise points is beneficial in this type of project, as we are not only dealing with many sources of uncertainty, but also are trying to cluster a biological system, which may be determined by significantly more factors than what is practical to measure or include in the analysis.
However, the drawbacks of the method is its lack of explanation. HDBScan is a clustering tool to cluster factors based on relative density with variable minimum cluster size. There is no justification beyond distance measures for the clustering, no explanation of why different datapoints are different, and no measure of what partition of the data is most accurate. In the end, the algorithm is just a tool, and it is up to the user to determine the efficacy of the agorithm, and verify the validity of results.

5.6.5 Method

In order to limit the presence of bugs in implementation, libraries like scikit-learn Pedregosa et al. (2011), numpy and HDBScan Campello et al. (2013) have been used. However, there are still multiple decisions made that might have had an impact on results. Variables chosen have an obvious impact. Variables like total angle change (in z-direction as well), track length squared (measure of energy expenditure) and other variables that the author simply has not thought of might have had a positive (or negative) impact on performance. More advanced clustering algorithms, more in depth analysis of temperature, wind conditions, and presence of delousing and other stressful event might have resulted in new avenues of investigation and discovery.

Another potential limiting factor is the fact that one trajectory is assumed to contain only one behavioural mode. As mentioned in the discussion for aquatraz fish 23, fish may change behavioural mode mid-trajectory for some of the longer trajectories, as seems to be the case for the top left trajectory in figure 4.29. This is a result of the way trajectories are created. Because the amount of trajectories with more than five points are so few in comparison to amount of data points, it was deemed not practical to spend time and effort in dividing trajectories with a switch in swimming pattern. If data becomes available with more consistent times between datapoints, detecting switches in behavioural mode becomes more important. Time series analysis could be a useful tool for this, as mentioned above.

Imagination is the only limiting factor in data exploration. Multiple plots have been generated but not presented. Box plots, activity per hour of day, and PCA direction magnitude per hour of day were created but not presented, as the decision was made that they did not provide enough useful information. Activity analysis based on hour of day might give rise to some discoveries, for example by adding hour of day as an integer variable to every trajectory. Additionaly, choosing new hyperparameters for each dataset (instead of using the same 4 variables and `min_cluster_size` that worked best for dataset fish 23 day on every dataset) is almost guaranteed to yield better results. Especially `min_cluster_size` could have been varied, as this is highly dependent on dataset size, which range from 632 (fish 21 day) to 5587 (fish 23 day). However, in the absence of clear performance metrics and lack of similar work on the field, the value of a simpler analysis done properly and consistently was deemed higher than compilation and presentation of every thinkable metric without the same depth and coherence.

6

Conclusion

The analysis has shown clear trends in some of the variables across night and day and between different fish. Furthermore, this project has shown that average depth, depth difference, track length and angle change may be used to differentiate between humanly classifiable salmon trajectories, such as feeding and circular swimming. However, the distance from center variables were not shown to be significant in differentiating trajectories with this approach. HDBScan showed that out of non-noise trajectories, most of the fish spends the majority of their time in circular or semi-circular swimming patterns in the day, and in shorter, non-circular, more idle trajectories at night. However, fish 23 spends most of its time in these idling trajectories at both day and night, while fish 20 spends most of its time in more circular, deeper trajectories at night, highlighting individual differences between fish. Trajectories resembling feeding were detected in two fish, though suspected in more from analysis of depth histograms. Depth histograms were found to be very individual, though showing similar trends with regards to night and day dependency across fish. In general, conventional analysis, covariance analysis and clustering found similar structure in the data, the main which corroborate existing research on the topic of salmon behaviour with regards to diurnal behaviour differences (Ulvund et al. (2021), Oppedal et al. (2011), Føre et al. (2018b), Fernö et al. (1995)). In conclusion, it is possible to use machine learning techniques to classify different modes of salmon behaviour that make sense to a human observer based on positional data, with the most prominent modes of swimming behaviour being circular swimming and idling.

Further research in the field might be focused on enhancing behaviour detection methods, linking behaviours closer to real-world events, or applying similar methods to other datasets. This project has shown that it is possible to use machine learning methods and telemetry to detect salmon behaviour, and may be a first step towards leveraging methods in data science for autonomous and reliable decision support in addition to reduce dependencies on manual labour and subjective assesments in aquaculture, as outlined in the description of Precision Fish Farming in Føre et al. (2018a).

Bibliography

Berckmans, D., et al., 2006. Automatic on-line monitoring of animals by precision livestock farming. Livestock production and society 287, 27–30.

Bjelland, H.V., Føre, M., Lader, P., Kristiansen, D., Holmen, I.M., Fredheim, A., Grøtli, E.I., Fathi, D.E., Oppedal, F., Utne, I.B., Schjølberg, I., 2015. Exposed aquaculture in norway, in: OCEANS 2015 - MTS/IEEE Washington, pp. 1–10. doi:10.23919/OCEANS.2015.7404486.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based clustering based on hierarchical density estimates, in: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 160–172.

Cañon Jones, H.A., Hansen, L.A., Noble, C., Damsgård, B., Broom, D.M., Pearce, G.P., 2010. Social network analysis of behavioural interactions influencing fin damage development in atlantic salmon (salmo salar) during feed-restriction. Applied Animal Behaviour Science 127, 139–151. URL: https://www.sciencedirect.com/science/article/pii/S0168159110002352, doi:https://doi.org/10.1016/j.applanim.2010.09.004.

Chen, L., 2009. Curse of Dimensionality. Springer US, Boston, MA. pp. 545–546. URL: https://doi.org/10.1007/978-0-387-39940-9_133, doi:10.1007/978-0-387-39940-9_133.

US Department of Commerce, N.O., Administration, A., 2019. What is aquaculture? URL: https://oceanservice.noaa.gov/facts/aquaculture.html.

Cubitt, K.F., Winberg, S., Huntingford, F.A., Kadri, S., Crampton, V.O., Øyvind Øverli, 2008. Social hierarchies, growth and brain serotonin metabolism in atlantic salmon (salmo salar) kept under commercial rearing conditions. Physiology Behavior 94, 529–535. URL: https://www.sciencedirect.com/science/article/pii/S0031938408000826, doi:https://doi.org/10.1016/j.physbeh.2008.03.009.

Daoliang, L., Du, L., 2022. Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. URL: https://link.springer.com/article/10.1007/s10462-021-10102-3#citeas, doi:https://doi.org/10.1007/s10462-021-10102-3.

Dubey, A., 2018. The mathematics behind principal component analysis. URL: https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643. last accessed 30.03.2023.

Fernö, A., Huse, I., Juell, J.E., Åsmund Bjordal, 1995. Vertical distribution of atlantic salmon (salmo solar l.) in net pens: trade-off between surface light avoidance and food attraction. Aquaculture 132, 285–296. URL: https://www.sciencedirect.com/science/article/pii/004484869400384Z, doi:https://doi.org/10.1016/0044-8486(94)00384-Z.

Føre, M., Alver, M.O., Frank, K., Alfredsen, J.A., 2023. Advanced technology in aquaculture–smart feeding in marine fish farms, in: Smart Livestock Nutrition. Springer, pp. 227–268.

Føre, H.M., Thorvaldsen, T., Osmundsen, T.C., Asche, F., Tveterås, R., Fagertun, J.T., Bjelland, H.V., 2022. Technological innovations promoting sustainable salmon (salmo salar) aquaculture in norway. Aquaculture Reports 24, 101115. URL: https://www.sciencedirect.com/science/article/pii/S2352513422001119, doi:https://doi.org/10.1016/j.aqrep.2022.101115.

Føre, M., Alfredsen, J.A., Gronningsater, A., 2011. Development of two telemetry-based systems for monitoring the feeding behaviour of atlantic salmon (salmo salar l.) in aquaculture sea-cages. Computers and Electronics in Agriculture 76, 240–251. URL: https://www.sciencedirect.com/science/article/pii/S0168169911000536, doi:https://doi.org/10.1016/j.compag.2011.02.003.

Føre, M., Frank, K., Dempster, T., Alfredsen, J., Høy, E., 2017. Biomonitoring using tagged sentinel fish and acoustic telemetry in commercial salmon aquaculture: A feasibility study. Aquacultural Engineering 78, 163–172. URL: https://www.sciencedirect.com/science/article/pii/S0144860917300432, doi:https://doi.org/10.1016/j.aquaeng.2017.07.004.

Føre, M., Frank, K., Norton, T., Svendsen, E., Alfredsen, J.A., Dempster, T., Eguiraun, H., Watson, W., Stahl, A., Sunde, L.M., Schellewald, C., Skøien, K.R., Alver, M.O., Berckmans, D., 2018a. Precision fish farming: A new framework to improve production in aquaculture. Biosystems Engineering 173, 176–193. URL: https://www.sciencedirect.com/science/article/pii/S1537511017304488, doi:https://doi.org/10.1016/j.biosystemseng.2017.10.014. advances in the Engineering of Sensor-based Monitoring and Management Systems for Precision Livestock Farming.

Føre, M., Svendsen, E., Alfredsen, J., Uglem, I., Bloecher, N., Sveier, H., Sunde, L., Frank, K., 2018b. Using acoustic telemetry to monitor the effects of crowding and delousing procedures on farmed atlantic salmon (salmo salar). Aquaculture 495, 757–765. URL: https://www.sciencedirect.com/science/article/pii/S0044848617324407, doi:https://doi.org/10.1016/j.aquaculture.2018.06.060.

Gordon, A.D., 1999. Classification. Chapman Hall/CRC.

Hassan, W., Føre, M., Urke, H.A., Kristensen, T., Ulvund, J.B., Alfredsen, J.A., 2019. System for real-time positioning and monitoring of fish in commercial marine farms based on acoustic telemetry and internet of fish (iof), in: The 29th International Ocean and Polar Engineering Conference, OnePetro.

Haste, T., Tibshirani, R., Friedman, J., 2008. The Elements of Statistical Learning. Springer.

Hindar, K., Fleming, I.A., McGinnity, P., Diserud, O., 2006. Genetic and ecological effects of salmon farming on wild salmon: modelling from experimental results. ICES Journal of Marine Science 63, 1234–1247. URL: https://doi.org/10.1016/j.icesjms.2006.04.025, doi:10.1016/j.icesjms.2006.04.025, arXiv:https://academic.oup.com/icesjms/article-pdf/63/7/1234/29125575/

IBM, 2023a. What is machine learning? URL: https://www.ibm.com/topics/machine-learning. last accessed: 20.06.2023.

IBM, 2023b. What is supervised learning? URL: https://www.ibm.com/topics/supervised-learning. last accessed at 22.06.2023.

Jensen, Ø., Dempster, T., Thorstad, E., Uglem, I., Fredheim, A., 2010. Escapes of fishes from norwegian sea-cage aquaculture: causes, consequences and prevention. Aquaculture Environment Interactions 1, 71–83.

Johansson, D., Ruohonen, K., Kiessling, A., Oppedal, F., Stiansen, J.E., Kelly, M., Juell, J.E., 2006. Effect of environmental factors on swimming depth preferences of atlantic salmon (salmo salar l.) and temporal and spatial variations in oxygen levels in sea cages at a fjord site. Aquaculture 254, 594–605. URL: https://www.sciencedirect.com/science/article/pii/S0044848605006113, doi:https://doi.org/10.1016/j.aquaculture.2005.10.029.

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2. URL: https://doi.org/10.21105%2Fjoss.00205, doi:10.21105/joss.00205.

Misund, B., 2023. Fiskeoppdrett. URL: https://snl.no/fiskeoppdrett. last accessed 22.06.2023.

Måløy, H., 2020. Echobert: A transformer-based approach for behavior detection in echograms. IEEE Access 8, 218372–218385. doi:10.1109/ACCESS.2020.3042337.

Noble, C., Gismervik, K., Iversen, M.H., Kolarevic, J., Nilsson, J., Stien, L.H., Turnbull, J.F., AS, N., et al., 2018. Welfare indicators for farmed atlantic salmon: tools for assessing fish welfare.

Oppedal, F., Dempster, T., Stien, L.H., 2011. Environmental drivers of atlantic salmon behaviour in sea-cages: A review. Aquaculture 311, 1–18. URL: https://www.sciencedirect.com/science/article/pii/S0044848610007933, doi:https://doi.org/10.1016/j.aquaculture.2010.11.020.

Oppedal, F., Juell, J.E., Johansson, D., 2007. Thermo- and photoregulatory swimming behaviour of caged atlantic salmon: Implications for photoperiod management and fish welfare. Aquaculture 265, 70–81. URL: https://www.sciencedirect.com/science/article/pii/S0044848607001214, doi:https://doi.org/10.1016/j.aquaculture.2007.01.050.

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T.S., Gismervik, K., Stien, L.H., 2019. Salmon lice treatments and salmon mortality in norwegian aquaculture: a review. Reviews in Aquaculture 11, 1398–1417.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Rillahan, C., Chambers, M.D., Howell, W.H., Watson, W.H., 2011. The behavior of cod (gadus morhua) in an offshore aquaculture net pen. Aquaculture 310, 361–368. URL: https://www.sciencedirect.com/science/article/pii/S0044848610007453, doi:https://doi.org/10.1016/j.aquaculture.2010.10.038.

Saberioon, M., Gholizadeh, A., Cisar, P., Pautsina, A., Urban, J., 2017. Application of machine vision systems in aquaculture with emphasis on fish: State-of-the-art and key issues. Reviews in Aquaculture 9, 369–387. doi:10.1111/raq.12143.

Stockwell, C.L., R., F., J., G., 2021. Determining the effects of environmental events on cultured atlantic salmon behaviour using 3-dimensional acoustic telemetry. Frontiers in animal science 2.

Thorstad, E.B., Rikardsen, A.H., Alp, A., Økland, F., 2013. The use of electronic tags in fish research–an overview of fish telemetry methods. Turkish Journal of Fisheries and Aquatic Sciences 13, 881–896.

Thorvaldsen, T., Frank, K., Sunde, L.M., 2019. Practices to obtain lice counts at norwegian salmon farms: status and possible implications for representativity. Aquaculture Environment Interactions 11, 393–404.

Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O.T., Nilsen, F., Horsberg, T.E., Jackson, D., 2013. Salmon lice–impact on wild salmonids and salmon aquaculture. Journal of fish diseases 36, 171–194.

Ulvund, J., Engebretsen, S., Alfredsen, J., Kristensen, T., Urke, H., Jansen, P., 2021. Behavioural response of farmed atlantic salmon (salmo salar l.) to artificial underwater lights: Wavelet analysis of acoustic telemetry data. Aquacultural Engineering 95, 102196. URL: https://www.sciencedirect.com/science/article/pii/S0144860921000522, doi:https://doi.org/10.1016/j.aquaeng.2021.102196.

Appendix

Appendix

A Scatterplots

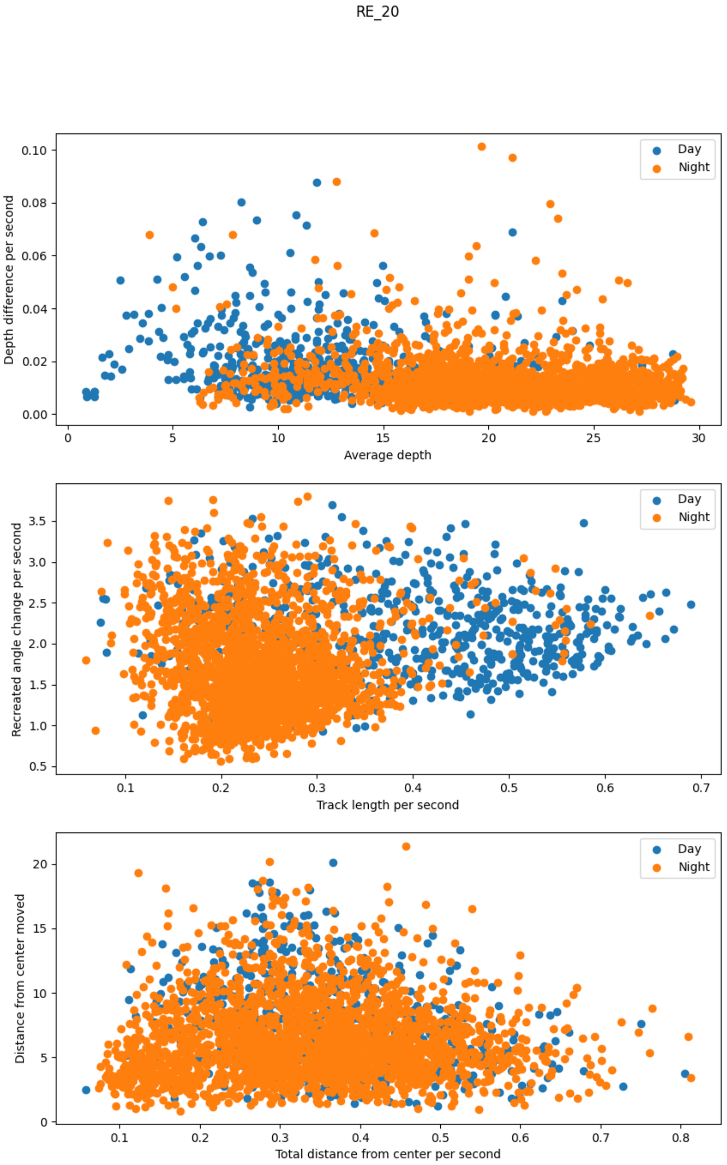


Figure 6.1: Scatterplots of variables for reference fish 20. Notice clusters of lower depth by day and lower track lengths (but not angle change) at night.

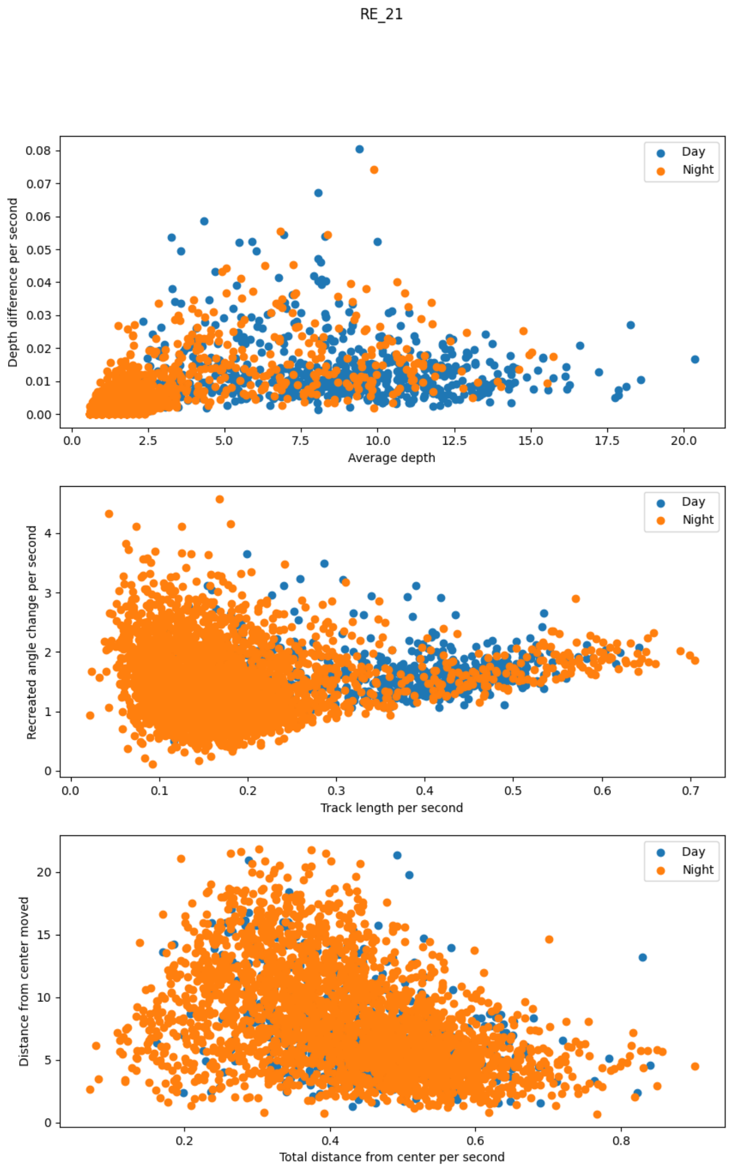


Figure 6.2: Scatterplots of variables for reference fish 21.Notice that it is deeper in the day and has a long linear correlation in track length per second and angle change per second.

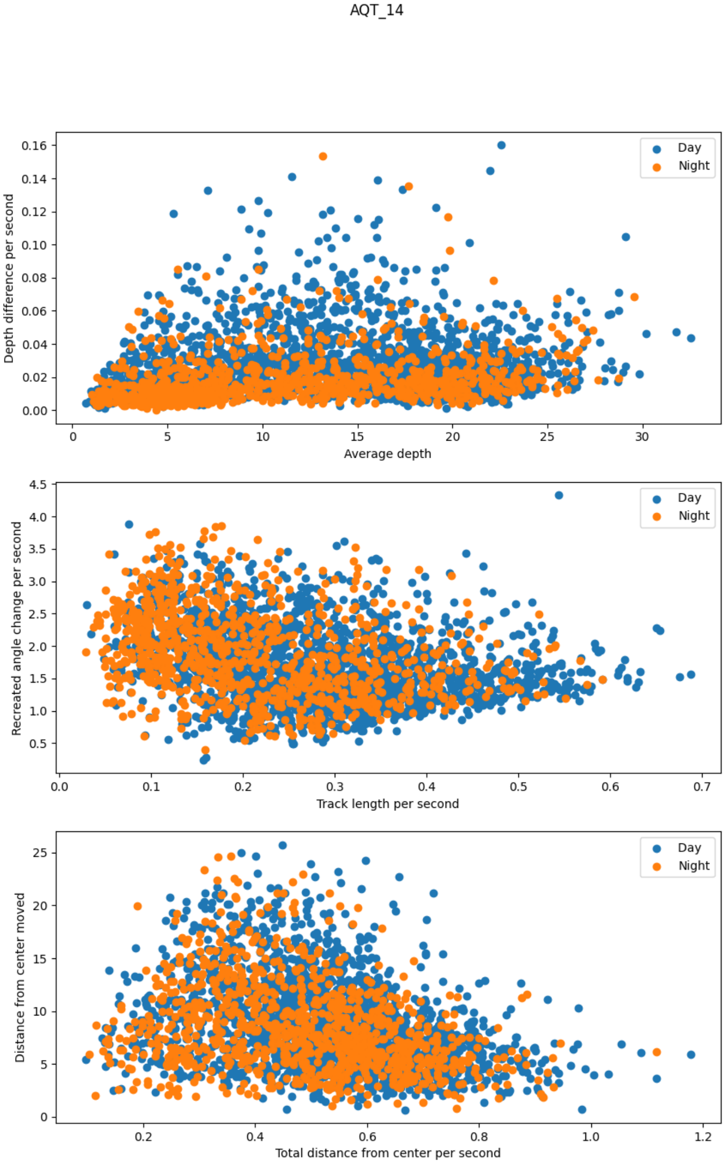


Figure 6.3: Scatterplots of variables for aquatraz fish 14. Some correlation in track length and angle change.

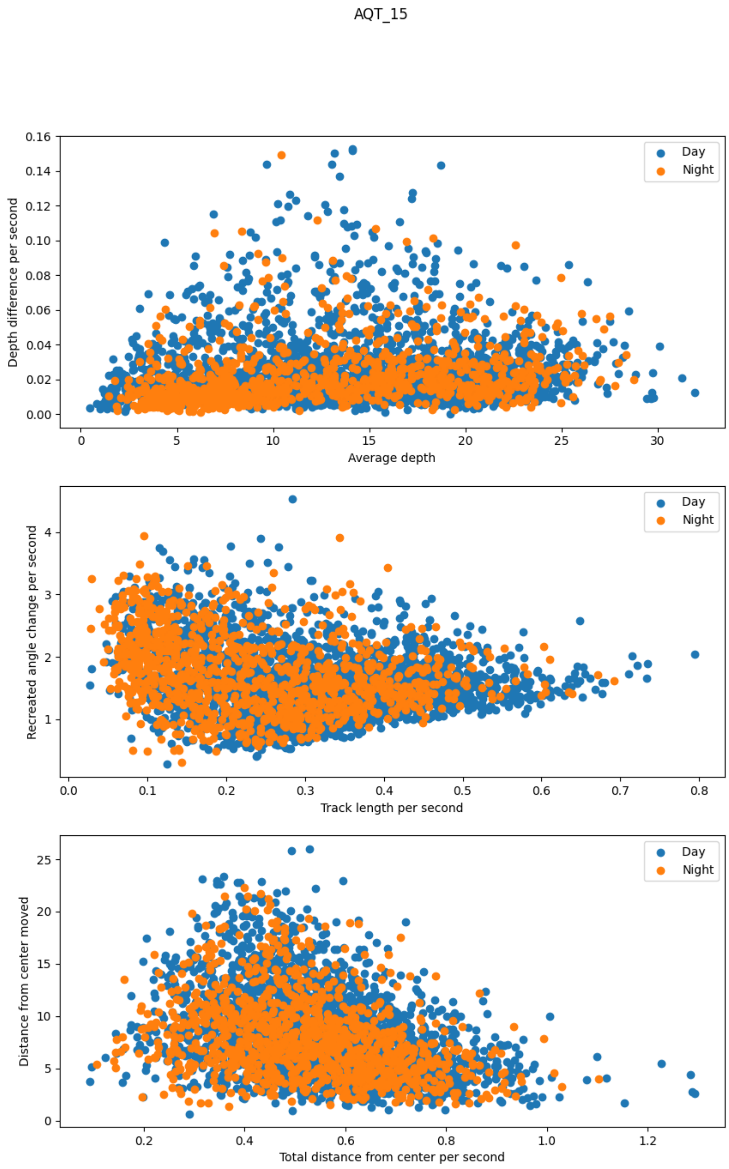


Figure 6.4: Scatterplots of variables for aquatraz fish 15. Some correlation at high track length low angle change, same at high distance from center and low distance from center moved.

B PCA Tables

Reference fish 20 Day

Table 6.1: Principal Components for fish 20 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 40.21%	-0.19	1.00	-0.24	0.06	-0.11	0.50
1: 21.12%	-0.25	0.67	0.73	-0.38	1.00	-0.81
2: 16.30%	1.00	0.32	-0.17	-0.02	-0.15	-0.37
3: 13.13%	-0.62	0.14	-0.81	0.01	-0.46	-1.00
4: 6.84%	0.07	-0.15	-0.61	-1.00	0.39	0.24
5: 2.40%	0.10	-0.12	-0.72	0.86	1.00	0.05

Reference fish 20 Night

Table 6.2: Principal Components for fish 20 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 39.62%	-0.06	1.00	0.06	0.18	-0.05	0.34
1: 23.87%	0.51	-0.02	-0.34	0.70	-1.00	-0.30
2: 20.93%	0.03	-0.38	0.07	0.21	-0.15	1.00
3: 8.25%	-0.02	-0.13	1.00	0.93	0.39	-0.26
4: 4.97%	-0.55	-0.08	-1.00	0.79	0.62	-0.01
5: 2.35%	-1.00	-0.07	0.19	-0.02	-0.56	-0.09

Reference fish 21 Day

Table 6.3: Principal Components for fish 21 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 44.96%	-0.26	1.00	-0.17	0.11	-0.07	0.50
1: 26.08%	0.91	0.89	0.68	-0.09	0.36	-1.00
2: 13.70%	0.97	-0.18	0.55	0.14	-0.24	1.00
3: 6.72%	0.55	0.08	-0.80	0.39	-1.00	-0.37
4: 5.44%	-0.12	-0.07	0.15	1.00	0.23	-0.07
5: 3.10%	-0.57	0.03	0.96	0.01	-1.00	-0.20

Aquatraz fish 23 Day

Table 6.4: Principal Components for fish 23 Day

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 35.29%	-0.03	1.00	-0.07	0.17	-0.21	0.48
1: 22.37%	0.61	0.68	1.00	-0.60	0.64	-0.75
2: 19.66%	-1.00	0.47	-0.80	-0.00	0.67	-0.87
3: 10.96%	1.00	0.20	-0.68	0.56	-0.73	-0.98
4: 6.25%	-0.55	-0.01	0.85	1.00	-0.28	-0.37
5: 5.47%	0.49	-0.08	-0.17	0.81	1.00	0.32

Aquatraz fish 23 Night

Table 6.5: Principal Components for fish 23 Night

PC no./ exp. var	Avg. depth	Depth diff	Track length	Angle	Dist. center	Dist. center moved
0: 41.30%	0.40	1.00	0.56	-0.04	-0.04	0.18
1: 21.91%	0.50	-0.63	1.00	-0.49	0.52	-0.69
2: 16.48%	-0.25	0.55	-0.45	0.06	0.50	-1.00
3: 9.20%	0.97	-0.10	-0.37	0.19	-1.00	-0.62
4: 6.39%	0.95	-0.17	-0.36	0.95	1.00	0.40
5: 4.72%	-0.42	-0.08	0.59	1.00	-0.25	-0.27

C HDBScan Clustering

Reference fish 20 Day

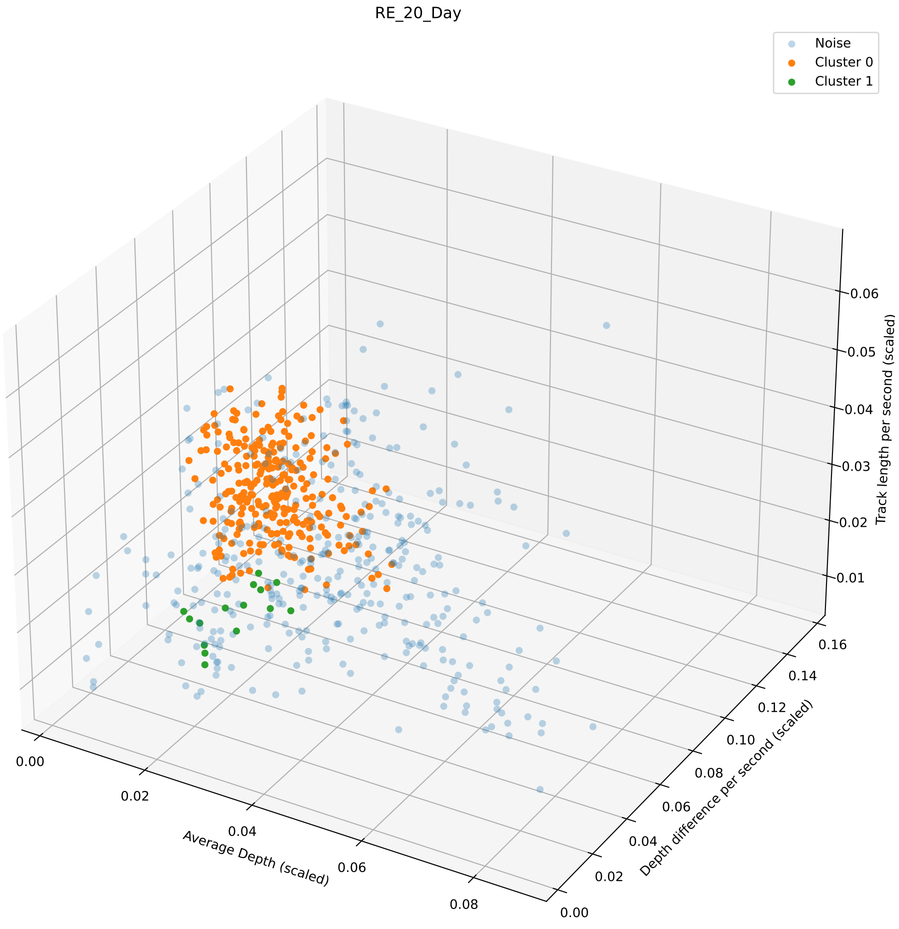


Figure 6.5: Clustering for fish 20 during Day. 640 trajectories in total. 45.63 % of points belong to two clusters. Cluster 0 contains 277 trajectories and cluster 1 contains 15 trajectories. .

Reference fish 20 Night

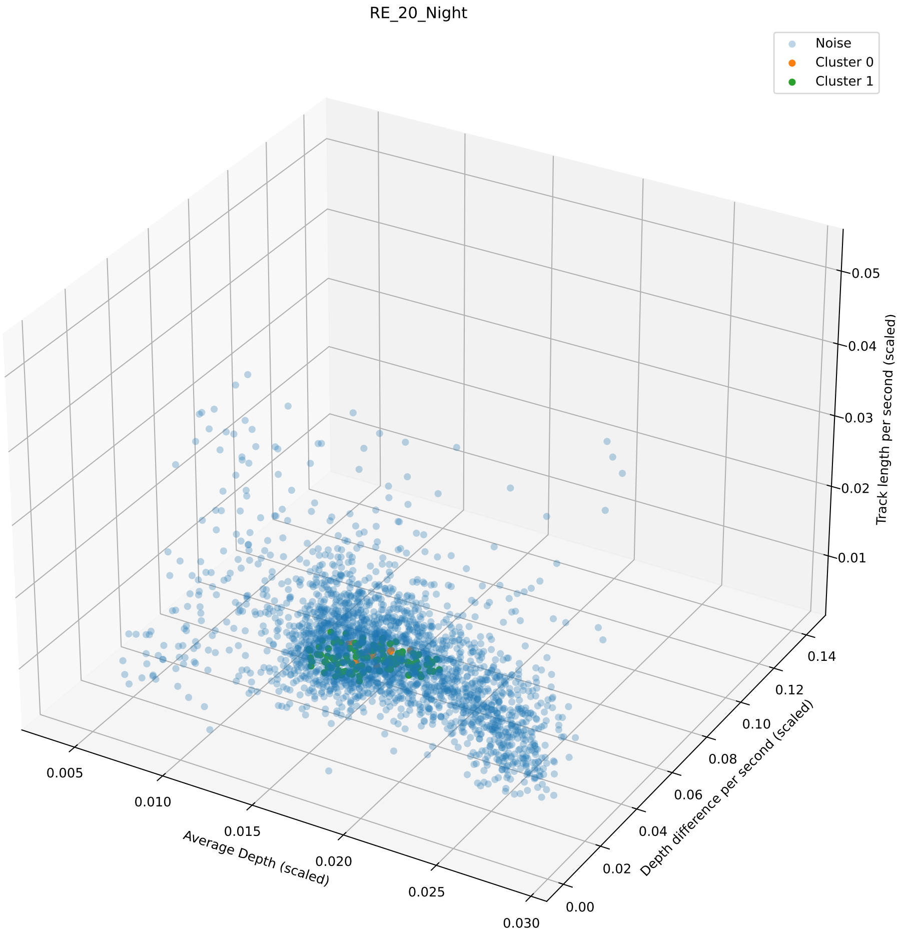


Figure 6.6: Clustering for fish 20 during Night. 2439 trajectories in total. 5.49 % of points belong to two clusters, where cluster 0 contains 17 trajectories and cluster 1 contains 117 trajectories.

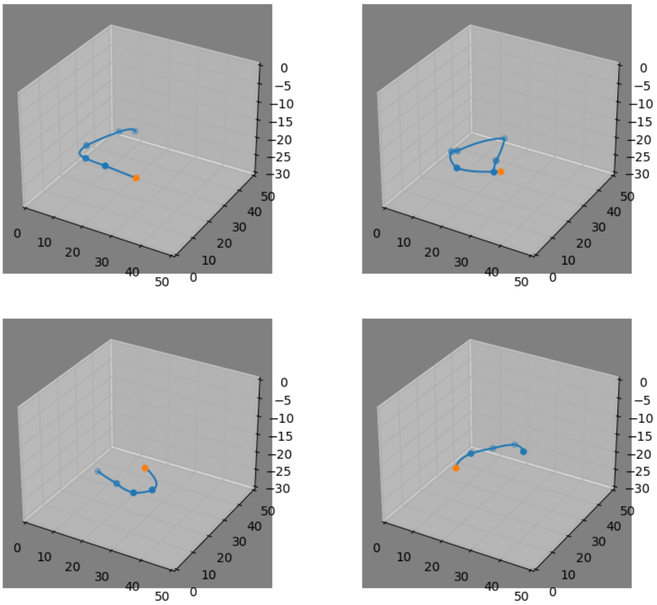


Figure 6.7: Category: 0 : 17 trajectories Mean depth: 19.67 Mean depth diff: 0.0112 Mean track length: 0.25 Mean angle change 0.0227

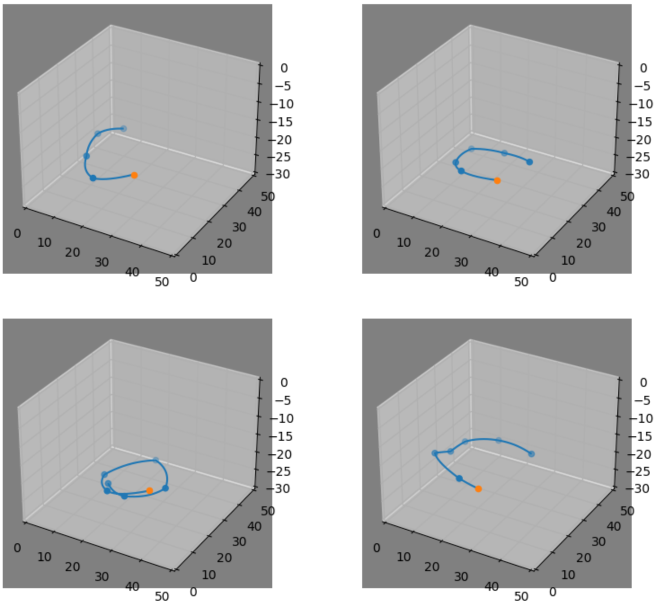


Figure 6.8: Category: 1 : 117 trajectories Mean depth: 19.81 Mean depth diff: 0.0080 Mean track length: 0.25 Mean angle change 0.0220

Reference fish 21 Day

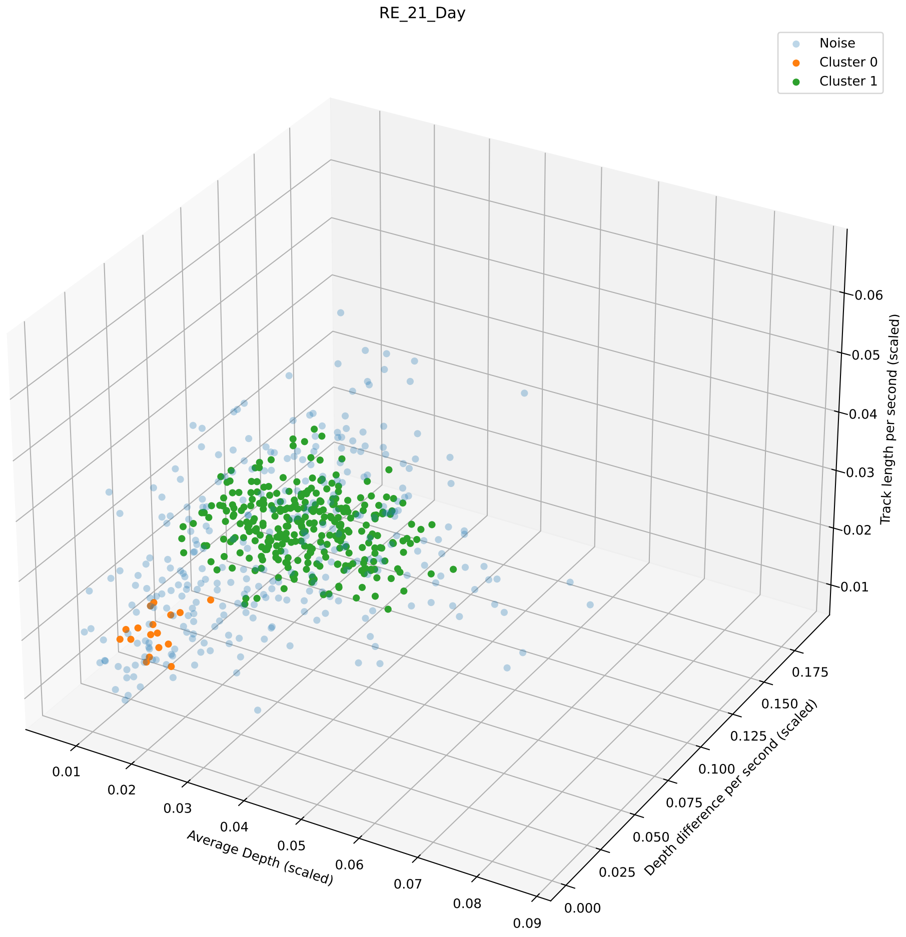


Figure 6.9: Clustering for fish 21 during Day. 632 trajectories in total. 51.11 % of points belong to two clusters. Cluster 0 contains 17 trajectories and cluster 1 contains 306 trajectories.

Aquatraz fish 14 Day

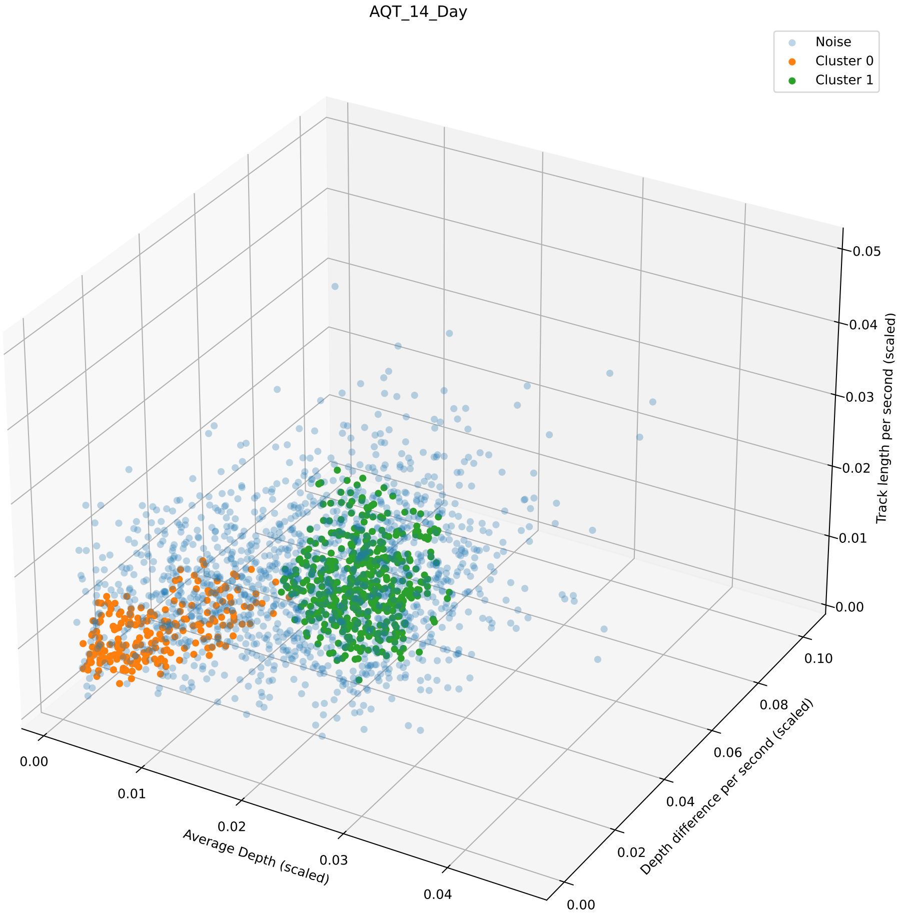


Figure 6.10: Clustering for fish 14 during Day. 2069 trajectories in total. 32.33 % of points belong to a cluster .

For the reference fish 41, in figure 4.9, 42.6 % of the data points were clustered in two clusters. Cluster 0 contains 214 trajectories, and cluster 1 contains 455 trajectories.

Aquatraz fish 14 Night

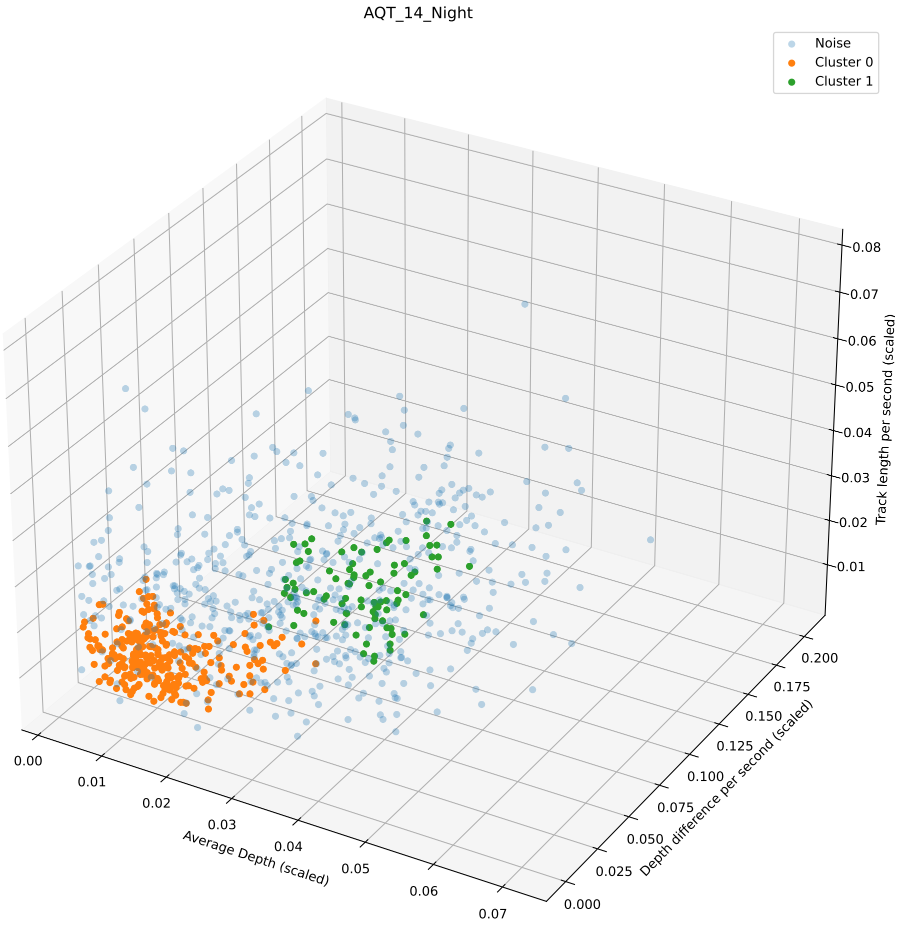


Figure 6.11: Clustering for fish 14 during Night. 941 trajectories in total. 39.64 % of points belong to two clusters. Cluster 0 contains 285 trajectories, and cluster 1 contains 88 trajectories.

Aquatraz fish 15 Day

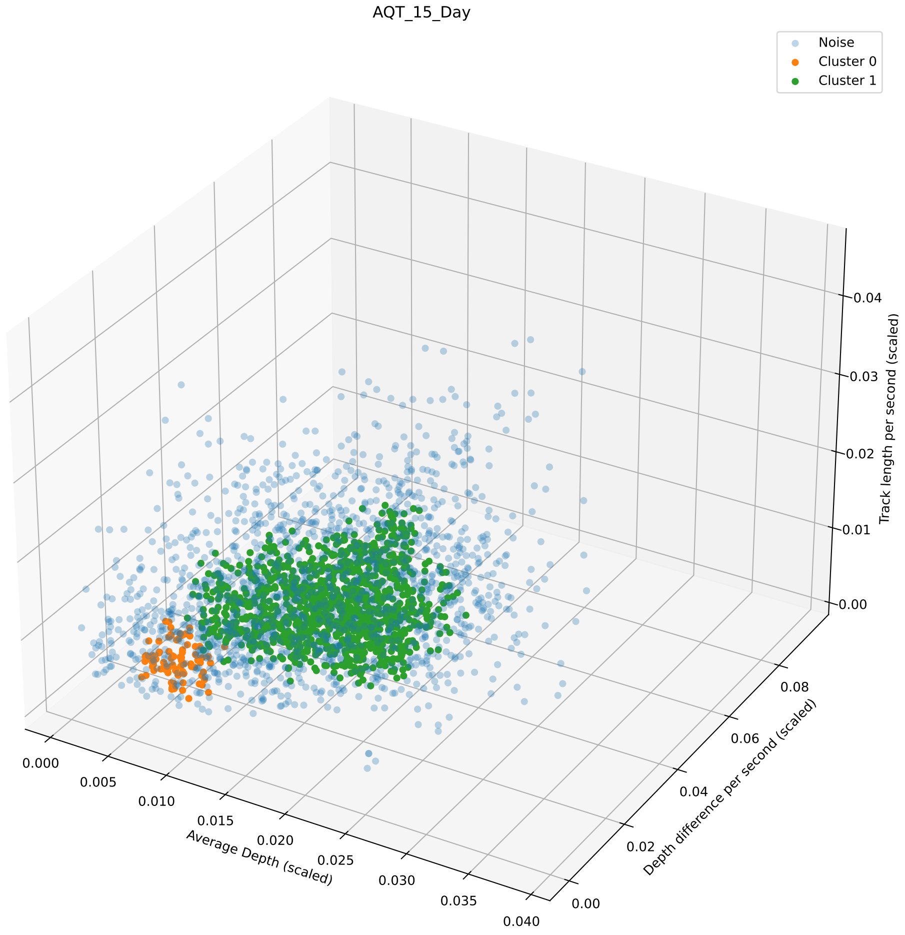


Figure 6.12: Clustering for fish 15 during Day. 2753 trajectories in total. 43.73 % of points belong to two clusters. Cluster 0 contains 81 trajectories, while cluster 1 contains 1123 trajectories.

NTNU

Norwegian University of
Science and Technology