Deliverable 4.3
New methods for post-processing biotelemetry data in aquaculture

Version 2.0

WP 4
Deliverable 4.3
Lead Beneficiary: NTNU/JU
Call identifier: Biological and Medical Sciences - Advanced Communities: Research infrastructures in aquaculture
Topic: INFRAIA-01-2018-2019
Grant Agreement No: 871108
Dissemination level: open
Date: 11.10.2023

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871108 (AQUAEXCEL3.0). This output reflects only the author's view and the European Commission cannot be held responsible for any use that may be made of the information contained therein.

Contents

1. Objective

The objective of deliverable 4.3 was to explore the potential of using unconventional post-processing methods, either from established scientific practices or based on machine learning, to derive new data types and knowledge from telemetry data. The main findings presented in this deliverable were obtained through two master theses at JU (delivery in early May 2024) and NTNU (delivery in June 2023).

2. Background

Deliverable 4.1 from WP4 in AQUAEXCEL 3.0 presented an overview of the different practices for using biosensors and biotelemetry as research tools within aquaculture. While the datasets acquired from such methods are usually interpreted by plotting or simply viewing the data as time series or statistical aggregations, it is likely that a deeper insight into the dynamics of the monitored biological systems can be obtained by further processing. Given the inherently individual based nature of the data acquired with such tools, more thorough post processing could generate new knowledge on systematic relationships and features in the datasets that are difficult to detect, identify and categorise manually. This deliverable summarises possible new methods for analysing telemetry data with the aims of providing new knowledge and insights, and that might contribute to reaching the aims of Precision Fish Farming (PFF, Føre et al., 2018a). The work is limited to looking into methods applicable to data acquired by either acoustic telemetry (i.e., that transmit their data wirelessly through acoustic signals) or data storage tags (i.e., that store measured data internally for later retrieval by the user).

3. Methodology

The work was mainly conducted through two parallel master theses conducted at JU and NTNU. An advantage of using master students for such work is that they are 100% dedicated to the task at hand during the project period, which gives a deeper insight into the methods and challenges addressed.

3.1. Master project at NTNU

The master student at NTNU started the work in February 2023 and delivered his thesis in June 2023 (Smedhaug, 2023). Data files containing 3D-positioning data from individual fish were provided to the student and formed the basis for the work. The data was collected from salmon in two different cages, with three individual fish in each cage being equipped with acoustic telemetry tags set to measure and transmit swimming depth at regular intervals (20-60 s between each transmission). The transmissions were picked up by acoustic receivers (TBR700, Thelma Biotel AS) placed along the outer perimeter of the cage. The internal clocks of the receivers were also synchronised with high precision using a surface module that had a GPS antenna, enabling the use of Time Difference of Arrival (TDoA) methods for positioning. When using TDoA on acoustic signals, it is possible to position the origin of the signal by evaluating the difference in time at which the signal is received at different receivers. Since each transmission from the acoustic tags contained the tag depth in m, this meant that a fish emitting an acoustic transmission could be positioned in 3D within the cage volume if this transmission was received by a minimum of 3 receiver units.

Traditional processing approaches where statistical methods are applied directly to the full dataset can identify large-scale responses such as distribution patterns. This work was therefore more keyed towards providing insight into individual behavioural patterns in shorter time scales. The aim of the analysis was to identify small-scale behaviours in individual fish (e.g., swimming in circles, turning abruptly, staying stationary over time) rather than looking into more overall behavioural expressions. The datasets were divided into a night and day part, to enable capturing the differences between day and night behaviour, which are known to be easily distinguishable in farmed salmon.

All results and most of the discussion related to the specific results are derived from the master thesis (Smedhaug, 2023), and note that this deliverable only presents excerpts from this work. The reader is thus referred to the full master thesis (Smedhaug, 2023) for further details on the methods, results and discussion from that study (the master thesis will be available online through NTNU Open). Moreover, any future works aspiring to use some of the findings of the master project should cite the master thesis rather than this deliverable.

3.2. Master project at JU

The master student was focused on the possibility of modelling selected changes in the behaviour of telemetrically marked fish, with the assumption that the future state depends probabilistically on the current state, and on the classification of these changes in the time series. The key object of the study is therefore a matrix of transitions between consecutive states. The overall probability is conditional and has to be evaluated.

The student used existing datasets of fish telemetry recordings collected for salmon using acoustic telemetry for almost half a year. The use of long time series enables to estimate the conditional probabilities and matrix of transitions. The objective is to identify any patterns in the dynamics, using various methods of multivariate data analysis, causality, and information theory.

The full master thesis will be available from the JU university library when it has been delivered.

4. Results and Discussion

4.1. Master project at NTNU

4.1.1. Summary of methods explored through the thesis

The processing pipeline used in the master project at NTNU consisted of three pre-processing steps and the application of three different methods for statistical analysis on the pre-processed data. Table 1 contains the definition and description of the main methods, terms and variables used in the analyses.

Table 1: Glossary for Master project at NTNU listing the most important terms and variables used in the analyses.

Name	Explanation
Trajectory	Sequence of 3D-positions with max interval of 60 s between consecutive data points
Average depth	Describes the mean depth of all samples in a trajectory
Depth difference	Summed depth variation within a trajectory divided by the number of samples

Track length	Full length of a trajectory in body lengths found by summing all segments between consecutive positions and dividing by body length and number of samples
Recreated angles	Sum of absolute changes in swimming direction through a trajectory divided by the number of samples
Mean distance from centre	Mean distance of all positions in a trajectory to the centre of the cage
Distance from centre moved	Max distance from centre minus min distance from centre in a trajectory
Scatterplots	Method for plotting pairs of variables from trajectories to identify eventual links between these.
PCA	Principal Component Analysis run using the variables derived from the 3D trajectories to identify frequently and rarely occurring behavioural expressions.
HDBScan	Method for unsupervised clustering of data applied to the variables derived from the 3D trajectories to identify clusters that could imply specific behavioural expressions

Pre-processing:

Step 1: Pre-processing started by first identifying data point sequences that could represent trajectories describing individual small-scale behaviours. These were characterised by consisting of three or more consecutive data values with a max inter-sample interval of 60 s as this was assumed to be close enough in time for the data points to be part of the same behavioural expression.

Step 2: The second step in pre-processing was to smoothen the data. Trajectories identified in the first step usually only featured a few data points, and direct plotting of these would typically yield a very sharp-edged sequence of straight lines which does not resemble the actual swimming trajectory of a fish. To Improve upon this, the second step in the pipeline was to smooth the trajectories through polynomial interpolation. Rather than seeking to fit one polynomial to all data points in a trajectory, the interpolation was made for three and three data points at a time. A second order polynomial was fitted to each triplet of data points through the entire trajectory with a stride length of 1, meaning that for a trajectory consisting of e.g., 6 data points, polynomials would be fitted to data points 1,2 and 3, then 2, 3 and 4, then 3, 4 and 5, and finally 4, 5 and 6. When all polynomials had been derived, they were used to interpolate the data points such that the time resolution of the resulting trajectory was 1 s. For regions where two polynomials overlapped (e.g., the last half of the sequence generated by the polynomial from samples 2, 3, 4 and the first half of the sequence generated by 3, 4, 5), the values were averaged. This resulted in a smooth and realistic representation of the trajectory of the fish.

Step 3: Once the trajectories were ready, the next step was to identify the variables of interest. While each trajectory consisted of a series of x, y, z values, such sequences of 3D coordinates are probably not the best measure for studying behavioural expressions. It was therefore decided that some selected collective variables describing the properties of the trajectories should be used in further analyses rather than the 3D positions directly, which resembles the processing pipeline used by behaviouralists in regular analyses. The variables derived in the analysis can be compared with

those typically used for the analyses of time series positioning data in EthoVision XT (Noldus, The Netherlands, Figure 1).

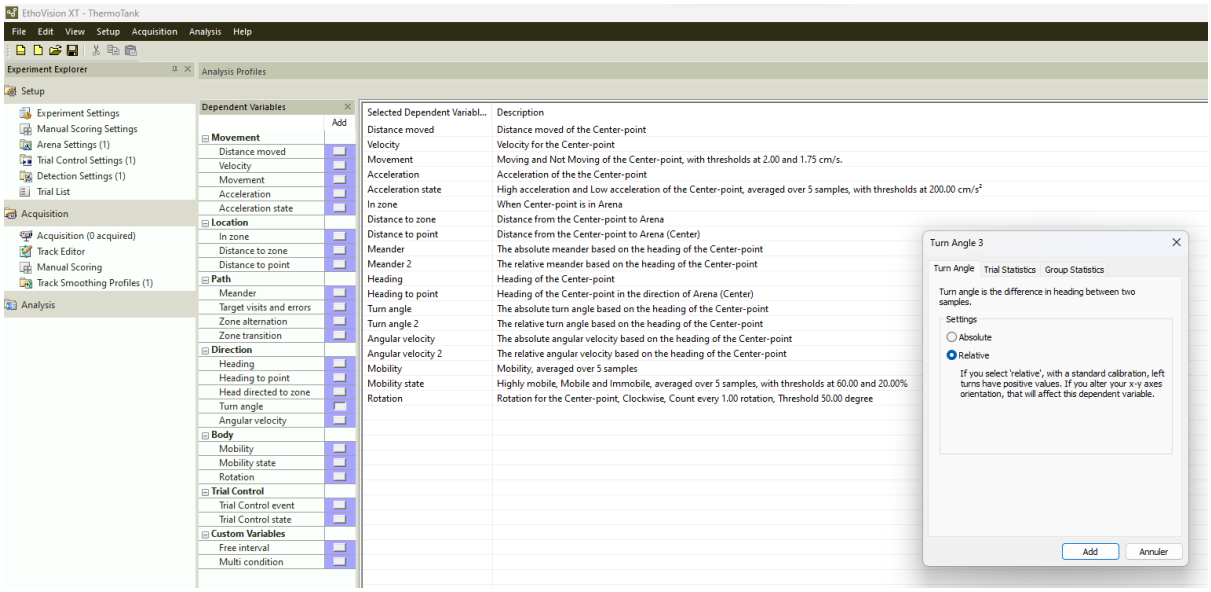

Figure 1: Screenshot from EthoVision XT (Noldus, The Netherlands) showing the various variables that are used in that tool.

The variables chosen in the present analysis and the rationale for choosing these were:

- Average depth: the average depth in m of all samples included in the smoothed trajectory (similar to location variables in EthoVision). The depth of a trajectory carries information on what depth the behavioural expression was conducted, which is often an important feature determining what the fish is doing.
- Depth difference: the depth change per sample in m, found by summing the depth difference between contiguous samples and then dividing by the number of samples (similar to movement variables in EthoVision). This variable will describe short term depth dynamics which is known to be an important feature in several aspects of salmon behaviour, particularly related to feeding behaviour.
- Track length: the full length of the track in body lengths per sample, found by summing the length of all inter-data point segments and dividing by number of samples and the individual body length (equal to “Distance moved” in EthoVision). Track length contains information about how far the fish travels in a track, which is relevant for several types of behaviour (e.g., circular/straight line cruising will have long track lengths, while more stationary behaviours, such as during feeding, will have short track lengths)
- Recreated angles: sum of the absolute changes in angle between samples divided by the number of samples (similar to meander and angular velocity variables in EthoVision). Rapid angular changes can be indicative of some types of localised behaviour and have been observed to be a feature when salmon are feeding. Few angular changes through a track could imply straight line or circular swimming patterns along the outer perimeter of the cage.
- Mean distance from centre: the mean distance from the positions in a track to the centre of the cage, found by summing the horizontal distances to the centre for each data point and dividing by the number of datapoints (equal to “Distance to point” in EthoVision). The

distance kept to the centre is a feature that can be important in many types of behaviour, for instance, some fish (e.g., salmon) will typically stay close to the centre where the feeding is conducted when engaged in feeding activities, while a fish traversing the outer perimeter of the cage will have a long distance to the centre.

- Distance from centre moved: max distance from centre minus min distance from centre in a track (similar to "Distance moved" in EthoVision). Large values in this variable could imply that the fish is moving much radially in the cage (which can be the case when feeding), while small values here could imply more steady straight lined or circular swimming patterns.

Once these variables had been derived from all identified trajectories, they represented the dataset used in the further analyses. While the interpolated trajectories in themselves were thus not used in any further analyses, they were used to visualise the trajectories.

Statistical analyses:

Method 1: The first statistical analysis was to find the mean and variance of the variables for all six fish included in the study. This represented a straightforward and basic approach that resembles the methods used in conventional processing of such data, and that could reveal some simple relationships such as overall differences between individual fish, differences between day and night and similar.

Method 2: The second statistical approach tested was to plot the selected variables and search for apparent trends in the data. This approach was also similar to conventional analyses, and consisted of providing histogram plots of the distribution of the variables for day and night per individual, and scatterplots for pairs of variables in the dataset. The latter method could reveal links between pairs of variables in the dataset that could be of interest, while the histograms would provide insight into the distribution of the variables. Moreover, similarities between individual fish or between day and night for individuals could be interpreted as implying similar behavioural expressions.

Method 3: The third approach used to analyse the data was Principal Component Analysis (PCA) which is a much used method for automatic analyses and dimensionality reduction of multivariate datasets (Dubey, 2018). PCA is a method that requires no prior knowledge about the processed dataset, and that uses linear transformations to extract a set of Principal Components (PCs) that together explain the variance in the data. Each PC is a linear weighted combination of the different variables in the dataset (in this case the abovementioned variables extracted from the trajectories), and all PCs in a PCA are orthogonal and are linearly uncorrelated, rendering them independent from each other. Specifically, a PCA will result in a list of derived PCs that together comprise the entire dataset and its variations. This list is ranked according to how much a PC explains of the total dataset. It is often such that the highest ranked PCs will explain most of the variance in the data, meaning that it is possible to recreate most of the variance using only the few most important PCs. This is the main feature of PCA, as this will effectually enable dimensionality reduction of the dataset, and simultaneously highlight which features are most important in describing the variance of the dataset. How many PCs are needed is determined by the user choosing a certain percentage (typically 80-95%) that needs to be explained. Following a PCA, further insight into the system dynamics can be obtained by analysing the variables comprising the most important PCs, and the internal scaling between these. To avoid issues due to the variables having different ranges in value

and variance, all variables were normalised. The method used to conduct the PCA was based on that published by Pedergosa et al. (2011).

Method 4: The final method of statistical analysis was to use unsupervised clustering of the variables. The method used was the HDBScan approach (Campello et al., 2013), using a public implementation provided by McInnes et al. (2017). This was primarily done with the aim of identifying eventual clustering in the collected set of trajectories as this could imply specific individual behavioural modes. Clustering was done using hyperparameters chosen from the analyses of the histograms and the scatterplots.

4.1.2. Outcomes and results

Pre-processing:

The number of tracks identified using the initial pre-processing method was found to be in the 1000s for all individual fish (see Figure 2 for example data from one of the six individuals in the study). While a relatively large proportion of the time intervals between adjacent datapoints were within the range 20-60 s (Figure 2 a), only the tracks with a length of five or more trajectories were used in the analyses, and as seen from Figure 2 b, this excluded most of the identified trajectories. However, there were still sufficient trajectories to run proper analyses of individual behaviour for all individual fish.

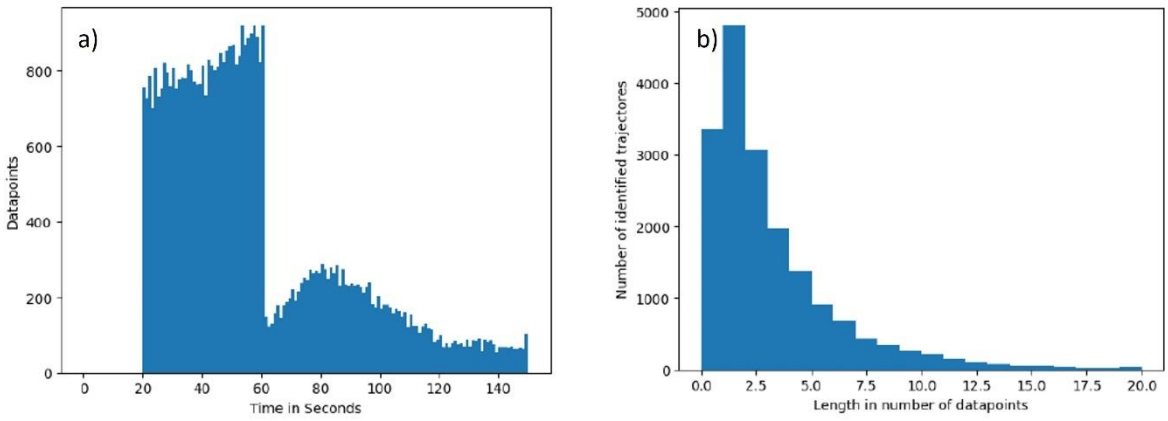

Figure 2: Histograms describing a) time intervals between adjacent data points; b) the distribution of the length of the trajectories from one individual (from Smedhaug, 2023).

The polynomial interpolation of the identified trajectories resulted in a more comprehensive dataset but also a more realistic reflection of how the fish may have moved through the behavioural expression. This is exemplified in Figure 3 where a non-interpolated trajectory (i.e., consisting only of the data points acquired from the telemetry system) is compared with the same trajectory after polynomial interpolation.

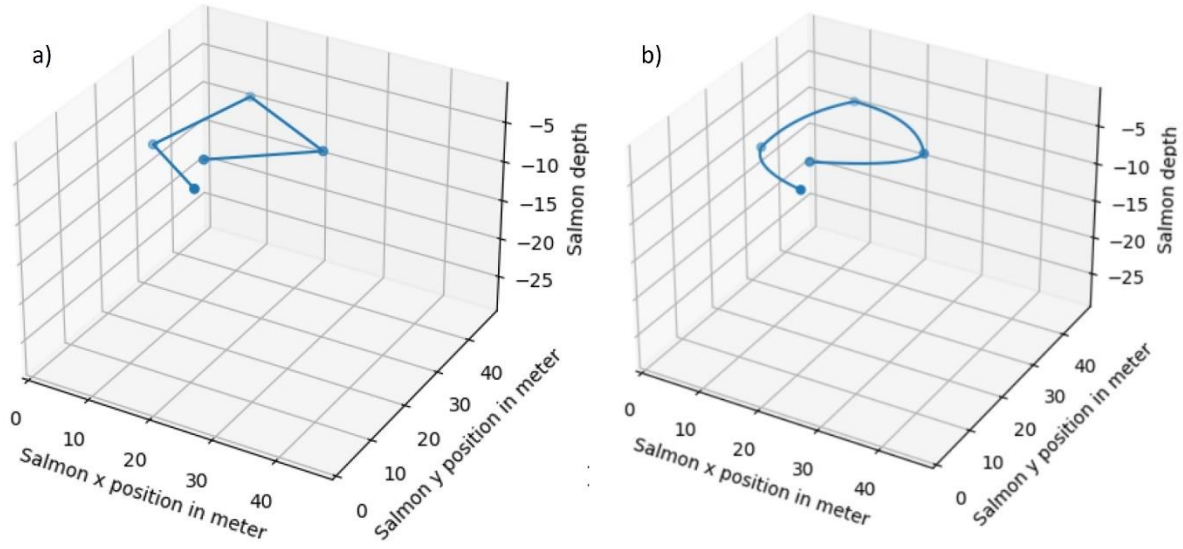

Figure 3: Illustration of the interpolation process. a) non-interpolated trajectory consisting of five datapoints; b) the interpolated version with 1 s timesteps between datapoints (from Smedhaug, 2023).

Statistical analyses:

Method 1: The mean and variance of the derived variables shed some light on the variations between the cages and individual fish, and between day and night. See Smedhaug (2023) for a full presentation of these results, as we will here only include excerpts of these.

- Mean depth varied between day and night (shallower swimming during night for most fish) and varied much between the individuals, but there seemed to be no apparent difference between cages.
- Depth differences were generally larger during day than night and varied between individuals and seemed to be larger in one of the cages than in the other.
- Track length was similar across individuals, and all fish had higher values in this variable for day than for night.
- Angle change had no consistent patterns of variation, but was higher during day than night for the fish in one cage, while the opposite was true for the fish in the other cage (i.e., higher values during night than day)
- Mean distance to centre was similar between day and night for all individuals, with distances and variances being somewhat larger for fish in one of the cages.
- Distance from centre moved was similar for day and night for all individuals, with few systematic differences except variance being higher during day than night for one cage and vice versa for the other.

Method 2: The histogram plots of the variables were made such that they distinguished between trajectories collected during daytime and night (Figure 4). The main features of the histograms matched with the mean and variance values computed in method 1, but they also provided a more in-depth view of how the variables were distributed for the individual fish. Figure 4 shows the histograms for two different individuals, illustrating some of the individual variations seen in the plots. Four of the six fish had histogram representations resembling Figure 4 a) where daytime and

night distributions were relatively similar, albeit with some propensity towards shallower swimming and less depth variations during night, while the remaining two fish had larger differences between day and night in several variables.

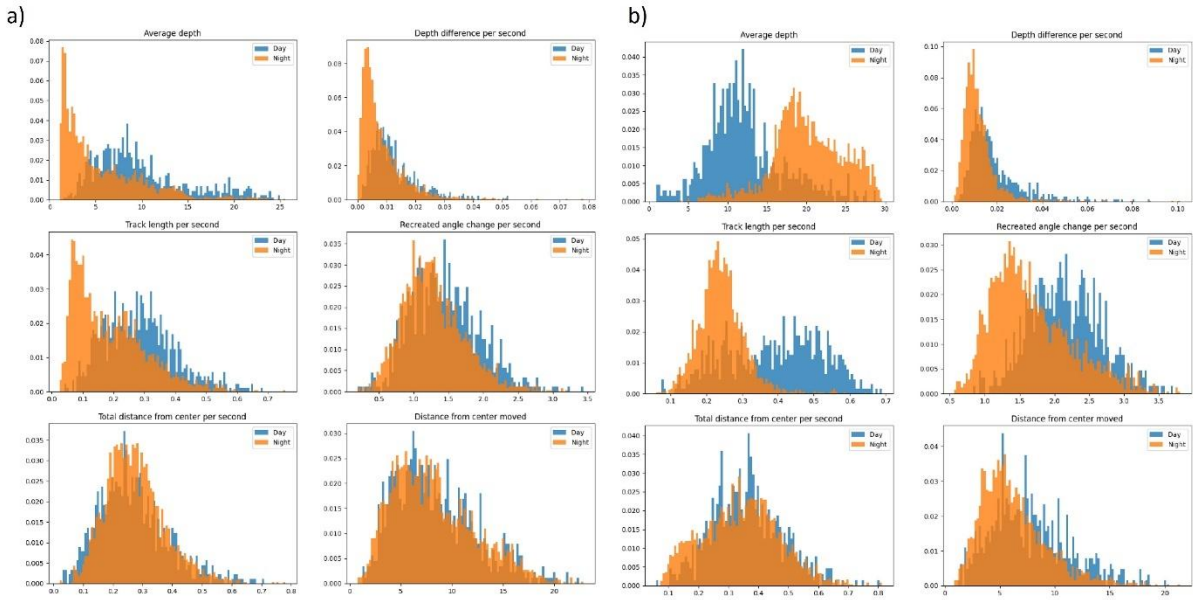

Figure 4: Histogram distributions of the six variables (av. Depth, depth difference, track length, recreated angle, total distance centre and distance from centre moved) for two individuals. Orange bars denote night and blue bars denote daytime, the horizontal axis denotes variable values and the vertical axis provides frequency of occurrence of the values across all trajectories identified for the individuals. a) Individual with relatively similar distributions of the variables across night and day; b) individual with larger variations in some variables (from Smedhaug, 2023).

The scatterplots were also made to distinguish between day and night and were used to explore several relationships between pairs of variables. The most interesting relationships were identified as being average depth vs. depth difference, track length vs. recreated angle and total distance from centre vs. distance from centre moved. For some fish, the scatterplots obtained based on trajectories acquired during daytime and night were similar (Figure 5 a), while others had difference in the distributions either with the variability being higher during day than at night (Figure 5 b) or vice versa. This method also proved able to identify interesting features in the datasets, such as the triangular shape seen in the track length vs. recreated angle plot in Figure 5, which could imply that there is a practical limit of how low the recreated angle variable can be given a certain track length.

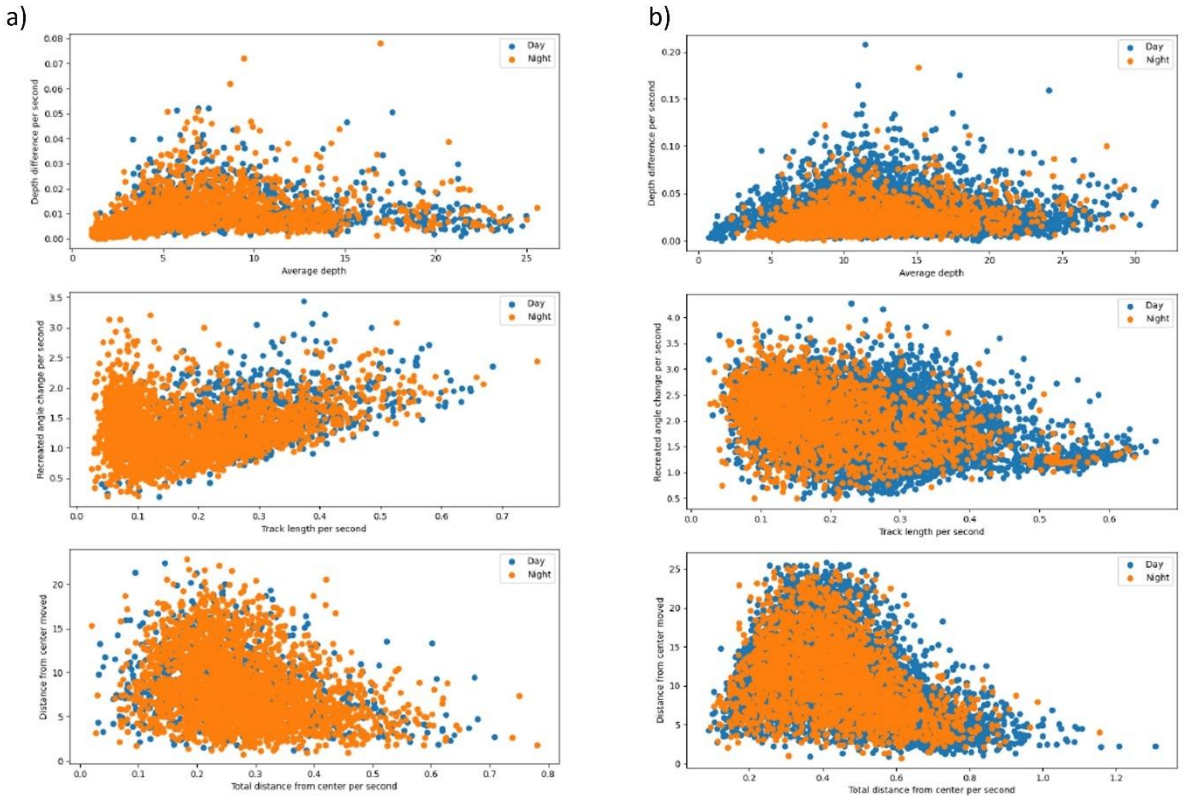

Figure 5: Examples of scatterplots relating two and two variables for two individual fish. The plots relate average depth and depth difference (top), track length and recreated angle (middle) and total distance from centre and distance from centre moved (bottom). Orange dots denote values from trajectories obtained during night, while blue dots denote daytime trajectory data. a) Individual with similar distributions for day and night; b) individual with more variations between day and night.

Method 3: PCA was run for the trajectories of all fish, and was, similarly to the other analyses, conducted separately for day and night data. For all fish from one of the cages, the first four PCs explained more than 90% of the variance in the dataset, while the first four PCs in the analyses of the fish from the other cage explained almost 90% of the variance. In addition to providing insight into the potential for dimensionality reduction, the analysis of the importance of the different variables in the most important PCs for each fish could shed light on the nature of the specific behaviours expressed by that component. This can be seen in the example in Figure 6, where about a third of the tracks observed during daytime (Figure 6 a) were explained by a PC featuring relatively shallow swimming (-0.5 on mean depth), large variations in swimming depth (1 in depth difference), and medium long trajectories (track length of 0.42), angle variations (recreated angle of 0.35) and distances kept to the centre (distance centre of 0.58), while the distance moved towards the centre were given low importance (value of 0.09).

a)

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

b)

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

Figure 6: Outcomes of PCA for one of the reference fish showing the percent of variance explained by each component (first column) and the linear weighting of each variable in each of the six components provided by the analysis, -1 implying that the PC favours low values for a variable, 0 that a variable is not important and 1 high values were important. a) results from daytime trajectories; b) results from nighttime trajectories (from Smedhaug, 2023).

Method 4: The clustering using HDBScan was conducted using the variables identified as most interesting in the histogram plots, which were average depth, depth difference, track length and angle changes. While it was hoped that this analysis would identify distinct clusters in the variables that could imply different individual behavioural expressions, the method could not easily identify clear clusters in all the cases due to a higher presence of noise points (i.e., data points that do not belong to a cluster) than points belonging to a specific cluster. While the results from this exercise were less easily interpreted than the outcomes of the PCA, it was possible to identify some interesting features by viewing the clustering plot together with example trajectories belonging to each cluster. Figure 7 shows an example of this where the daytime trajectories of an individual were analysed. This individual was the one fish out of all used in the study that had most eligible trajectories after the pre-processing and is therefore perhaps the most interesting individual to analyse since using this method as it is likely that analysing its trajectories will provide a more complete insight into the behaviours of that fish. The clustering analysis searched for clusters in the manifold spanned by the variables mentioned above, but the figure only shows average depth, depth difference and track length as visualisation is limited to three dimensions. The analysis revealed that there were four identifiable clusters in the data and that almost 50% of all trajectories were assigned to one of these (the rest being labelled noise points). Cluster 3 featured low track length, medium depth difference and high variations in angle and was the largest cluster comprising 2570 trajectories, implying that this behaviour was relatively often expressed by this individual. The second largest was cluster 1 with 134 trajectories. These were mainly characterised by long track lengths and low variations in depth. Cluster 0 ranked third with 16 trajectories, and featured low mean depth (i.e., fish being close to the surface), low track length and large changes in angle. And finally, cluster 2 featuring 15 trajectories, was categorised by medium track length and very high depth variations.

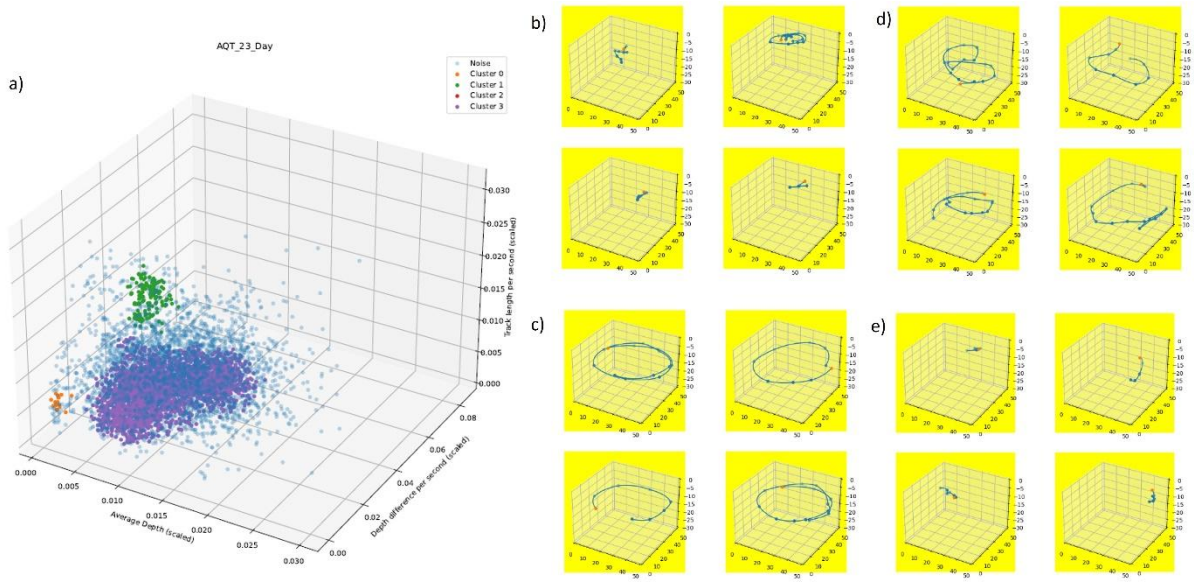

Figure 7: Data from clustering analysis for one individual during daytime. a) the clustering analysis yielded four clusters – Cluster 0 (orange, 16 trajectories), 1 (green, 134 traj.), 2 (red, 15 traj.) and 3 (purple, 2570 traj.). The rest of the points (grey) are considered noise as they do not belong to either of the clusters; b) example trajectories for cluster 0; c) example trajectories for cluster 1; d) example trajectories for cluster 2; e) example trajectories for cluster 3 (from Smedhaug, 2023).

4.1.3. Discussion

Rather than conducting an extensive discussion of the specific results from the analyses conducted by the master student, this section will focus on the properties of the methods and discuss their potential applicability in acquiring new knowledge on fish behaviour by processing telemetry data.

Pre-processing:

The pre-processing steps applied in this study (i.e., identifying contiguous trajectories, smoothing these using polynomials and deriving secondary variables from the properties of the identified trajectories) seem like reasonable steps towards achieving more detailed insights into the specific behavioural expressions of individual fish. Restricting analyses to trajectories identified by inter-data point time intervals as described earlier is useful as the causality between consecutive positions will be higher the shorter the time between these. This means that a sequence of five or more data points is likely to reflect some of the short time behavioural expressions of individual fish in aquaculture. This type of observations has proven difficult to acquire using conventional means such as computer vision, active acoustics, and visual inspection, as it is then just possible to observe an individual in very short time spans. Since telemetry provides the individual data history of the studied animals, the trajectory-based pre-processing will combine both the short-term aspects that

are often elusive in telemetry studies (where we often try to derive statistical measures rather than individual measures of behaviour) with individual histories over time. While the master project explored this approach for datasets comprised by 3D-positions, it is also feasible to consider its utility to other data types. For instance, although variables such as heart rate (e.g., Brijs et al., 2018) and stress related variables (e.g., Wu et al., 2015) are today largely collected using data storage tags (DSTs) that provide continuous datasets, their utility in commercial cages is likely to be higher if these are also possible to monitor using acoustic telemetry. In such cases, the data received from the tags on these variables would be sparser than in the DSTs due to the need for acoustic communication, and it would be likely that some data points are lost due to acoustic interference. In such cases, an approach analogue to the trajectory-based analysis could be used to identify sequences in which there are several consecutive samples close in time. These could then be used to focus an analysis of these factors towards short-term variations.

Statistical analyses:

While the thesis results demonstrated the potential of acquiring new insights into individual behaviour in salmon by applying conventional methods (mean and variance, histograms, and scatterplots), these methods are as implied conventional and well known, and will therefore not be covered in more detail here (see the master thesis Smedhaug, 2023 for discussion of the results obtained with these methods).

PCA: PCA analysis has not previously been used to analyse telemetry data in fish, and the results acquired in the master project implied that this is a method that has potential in helping identify the specifics of short-term individual behaviour in farmed fish as measured by telemetry tools. The analysis of the composition of the most dominant PCAs is particularly interesting and should be explored in future applications of this method to such data. If the variables defined can somehow be linked with recognisable behavioural expressions, such considerations could result in new insight into individual fish behaviour. In the example data provided in this deliverable, the most common PC during night time (PC0, 50% of the variance) was characterised by a deep mean depth (avg. depth 0.87), large variations in depth (depth diff 1), long track lengths (0.77) and low angle variations (0.16), distance to centre (0.15) and radial distance moved (distance to centre moved -0.03). While it is not immediately apparent what type of behaviour this reflects, a possible interpretation of this PC could be that the fish seems to prefer to stay deep in the cage on average during night, display much vertical travelling, and stay not too far from the centre of the cage. The PC also implies that the fish kept a relatively constant distance to the centre, and a long track length while angle variations were low. In summary this could imply a fish moving slowly in a pattern that is a more or less circular horizontally and that features relatively large vertical migrations. Although this type of behaviour has not been properly categorised yet, this could be a feature of salmon behaviour that is common during night-time. This pattern has some similarities with individual behaviours previously observed by e.g., Juell and Westerberg (1993) and reviewed by Oppedal et al. (2011) in that salmon maintain movement during night, albeit at a slower speed. However, few such observations have also featured the vertical migrations displayed by the fish in this case.

For daytime, the data implied a less homogeneous behavioural expression, as the most common PC during daytime (PC0) only described 33% of the variance (as opposed to 50% for the night-time trajectories). This implies that the behaviour during daytime is more variable than that during night,

which is in accordance with common knowledge on salmon and their nature as visually oriented animals (Oppedal et al., 2011). The PCs from daytime were less easily interpretable than those from night, mostly because of the greater variability in behavioural expression. For instance, the most dominant PC (PC0) was characterised by relatively shallow swimming (avg. depth -0.5), large variations in swimming depth (depth diff 1) and few radial movements (dist center 0.09), and otherwise medium track length, angles and distance to centre. Of these, only the depth difference gave a clear indication in that the fish appeared to display relatively strong vertical motions near the surface. The medium weights for the other variables render the interpretation of the exact behavioural expression more difficult.

A general observation acquired from the PCA was that the variable depth difference appeared to be the most important variable in the most significant PC (i.e., PC0) for all fish. This could imply that depth variability is an important part of the individual behaviour expression of salmon in captivity. While this may not come as a surprise to either farmers or researchers, this does illustrate yet another facet of how PCA could be used to quantify such elements of "common knowledge" that are often based on anecdotal observations. Another general observation was that there seemed to be more similarities between individuals in the PCA outcome based on trajectories collected during night. While this can probably be ascribed to the variability in behavioural expressions being higher during day than during night, it also shows the potential of using PCA to quantitatively compare the short-term behaviours of individual fish in a more practical manner than conventional approaches that may typically entail comparing time series plots of swimming trajectories.

Clustering: While the outcomes of the clustering were less clear than those from the PCA, they were nonetheless interesting, in that for some fish they gave insight into how the trajectories could be split into different data clusters that could be automatically distinguished by a computer. Moreover, the method also provided variable features for each cluster that, similarly to the weights in the PCA outcome, could be further analysed to divulge more about which real behavioural expressions the trajectories belonging to the clusters reflected. Since the latter observation was less easy to derive directly from the analysis outcome than for the PCA, this analysis was done by qualitatively reviewing trajectories that are typical members of each cluster visually. In the example data included here, it was apparent that trajectories belonging to the second most common cluster (cluster 1) reflected circular swimming patterns at relatively even depths. This strongly resembles the circular swimming patterns that has been reported by salmon farmers for decades, and that is believed to be one of the most common behaviours of caged salmon in periods where they are not fed/have low feeding motivation. However, one surprising feature in this analysis was that the supremely largest cluster in this case (cluster 3) reflects a type of behaviour that is very different from the circular patterns one might expect. This type of trajectory seems to reflect some sort of semi-stationary slow moving "idle" behaviour that the fish chose to exhibit most of the time during night. While this result is linked with one individual and hence cannot be generalised to salmon in general, it indicates the potential of using such methods to explore if fish behaviour in captivity is as we believe it to be, or if it diverges from our *a priori* assumptions. The observed pattern could indeed also be a feature that is general for farmed salmon, since most observations of circular swimming patterns are based on the observation of a specific control volume (e.g., camera or sonar) wherein such behaviour is observed. However, when observing this, we have no control of which individuals partake in this behaviour. Perhaps the truth is that circular swimming is a consistent behaviour expressed by farmed salmon,

but that each individual fish only partakes in this behaviour in short periods, and otherwise idle in the cage volume? The data obtained in this study are obviously too sparse to conclude upon this hypothesis, but nonetheless describe the potential outcome of such methods.

A final comment on unsupervised clustering that could be relevant is that although it is tempting to think about the largest clusters as most important to understand fish behaviour, the smaller clusters could be equally interesting. A cluster containing a large number of trajectories would indeed contain information on the behaviour of the fish most of the time. However, if that behaviour is mostly expressed as idling, circular cruising or other monotone patterns, these may not really provide much new knowledge on how the fish behave in captivity (beyond that such patterns are common). In such cases, reviewing the smaller clusters could be of greater interest, as this could unveil behaviours that are more rarely expressed, but nonetheless may be important aspects of the fish behaviour. For instance, feeding behaviour is typically a behaviour of particular interest for aquaculture since feeding is the main input from the farmer into the production process. During a day, it is unlikely that a salmon will eat more than 10-50 pellets, and with each pellet grasp probably taking a mere seconds, it is likely that the specific behaviour when the animal attacks the feed will be expressed very rarely compared with other behaviours. It is possible that the behaviour associated with cluster 0 in the analyses displayed in this deliverable shows exactly that. The time series plots from typical members of this cluster imply the fish staying shallow, close to the middle of the cage and changing swimming direction often, all of which could indicate that it is homing towards the feeding area.

Potential improvements and future development:

The choice of variables derived from the datasets has obvious ramifications for the success and precision of both the PCA and clustering methods. Choosing these variables should be done with care, since this is the only way in which to inject some system knowledge into the processing methods that are both inherently completely agnostic in terms of the dynamics of the process for which the data is collected. In the master study, variables were chosen such that they could target behavioural expressions assumed to be common in salmon such as circular swimming (e.g., constant distance to centre, long track length, low angle changes) and feeding (e.g., short track length, shallow swimming, large angle changes). Considering the infrequent occurrence of what could be assumed to be circular swimming and feeding behaviour in the analyses, it is possible that other variables could improve the consistency, precision, and interpretability of the outcomes of the analyses. However, identifying such variables is a non-trivial task that would require more research utilising the state-of-the-art within knowledge on salmon behaviour, analysing existing and new datasets collected with telemetry, the use of virtual experiments employing individual based mathematical models of fish behaviour, and simply trial-and-error to explore the outcomes of different variables.

More densely populated datasets with datapoints that are closer in time would also probably be beneficial for such studies, as they would render both PCA and clustering more efficient and accurate and increase the likelihood that they return features that describe the behaviours of individuals. However, when using acoustic telemetry, this could be very challenging to achieve considering the limited bandwidth of acoustic protocols underwater (ca. 1 byte per 4 s), the impacts of acoustic interference when multiple tags are transmitting at the same time, and the impacts of

ambient acoustic noise. If this was possible despite these challenges, a combination of more high frequent positioning data and more elaborate variables founded in science could contribute to rendering these methods excellent candidates for deriving new knowledge on fish behaviour in sea-cages.

4.2. Master project at JU

4.2.1. Summary of methods explored through the thesis

Table 2 contains the definition and description of the main methods, terms and variables used in the analyses.

Table 2: Glossary for Master project at JU listing the most important terms and variables used in the analyses.

Name	Explanation
Activity dataset	Activity values collected over 5 months using individual based acceleration measurements from caged salmon undergoing normal production as well as crowding and delousing.
Information source	A source that produces data values/symbols over time. A source is defined as a set of possible symbols and a set of corresponding probabilities of occurence. In this study, each subset of the total dataset used in the entropy analyses (i.e., typically 24 h subsegments of the datasets) are considered sources.
Information content	Quantification of the information content of a data value/symbol produced by a source (i.e., a subsegment of the dataset in this case). Found as the base 2 logarithm of the probability, *p*, of the symbol occurring. Since *p* is ranged from 0 to 1, this means that rarer data values (with low *p*) are considered to have higher information content than more common values (with high *p*). Information content is measured in bits.
Information entropy	Method for quantifying the average information produced by a source (i.e., a subsegment of the dataset in this case). This is found by summing the product of the probability, *p*, of each value occurring and the information content of this value across all values the source can provide. The result of this computation is such that the more equal the probabilities, *p*, are across the whole set of possible values, the higher the entropy value. Consequently, subsegments where values vary more will have a higher entropy value.
Causality	Method of evaluating transitions between consecutive samples to detect behavioural changes that could imply special events. Based on the assumption that consecutive samples will usually be close in value, meaning that a change to higher or lower values from a present value could indicate special events. In this case, it is assumed that consecutive activity values should be expected to be similar for a fish during regular/normal behaviour.
CDF	Cumulative Distribution Function, method for identifying how many samples of a dataset are needed to cover the full variability of the dataset.

KS	Kolmogorov-Smirnov goodness-of-fit-test. Method used to compare CDFs between cases (in this case 24 h periods).
Transition matrix	Tool for highlighting the transitions from one value *k* to the next value *k+1*.

M1: Information entropy

In order to analyse the fish telemetry data from the information theory point of view, it is at first necessary to address the question of reference states or state trajectory. Ideally, normal behaviour should reflect the basal information level. Unfortunately, we have no idea on how the normal behaviour should look like, since the fish are cultivated in captivity. Thus, they hardly present the true normal natural behaviour. However, they could present some kind of typical behaviour, hopefully of a similar random process.

Information theory arose from the statistics and cybernetics. The concept of information itself is based on the Shannon's measure, sometimes called Information Entropy S:

$$S(X) = -\sum_{i=1}^{n} p(X)_i \log_2 p(X)_i$$

where p(x) is the probability function of investigated phenomenon x.

In the case of fish telemetry data, the measured data already represent variables distributed in time, and there is no restriction to apply Entropy measurement to the available amount of information.

The information concept itself is additive, and the probabilities are logarithmic. It therefore shapes the tails of the distribution and the rare events become more important. The information entropy could be considered as a measure of surprise, or measure of our ignorance of the system. These are good characteristics to use it in the detection of atypical behaviour as measured using fish telemetry.

To investigate the relevance of the entropy approach, measurements from a study where fish were exposed to three delousing operations were used (Føre et al., 2018b). The results from that study implied that the measured variables (depth and acceleration) were strongly linked with the delousing operations, and it is also expected that this would manifest as a strong change in the information level.

The activity dataset consists of 5 months of values, approximately 600 measurements per day, obtained from 21 individual fish. Time series values were divided into 24 hours intervals, as an evaluation window period.

M2: Causality

Theory of causal systems states that every event is produced by its immediate cause. The transition between the states is therefore the result of conditional probabilities. The same concept is used in Markovian approaches for times series or state trajectory analysis. The systems, where events depend only on the previous state, are systems with so-called Markovian properties. Fortunately, even systems without such properties may be modelled or analysed as if they had those properties.

The transition analysis could reveal unusual patterns in dynamics, which are not accessible by classical frequency or wavelet analysis. In order to carry on the transition analysis of the time series of fish telemetry data, the probability distribution and conditional probabilities have to be investigated.

4.2.2. Outcomes and results

M1: Information entropy

To illustrate the method outcomes, only the acceleration-based activity variable is presented. Time series of activity over 24-hour periods (time intervals from 220 to 380 s between consecutive individual samples), are considered as independent distributions of acceleration variable values during such periods. To investigate their similarity, a cumulative distribution function (CDF) is estimated for each day as relative discrete histograms (Figure 8). The plot of CDFs will unveil dissimilarities between days based on the variability in their data and thus the information content in the samples. This plot suggests that each daily distribution is independent, incoherent, and uncorrelated to each other. The dissimilarity or independence could be simply proved by a Kolmogorov-Smirnov goodness-of-fit test (KS).

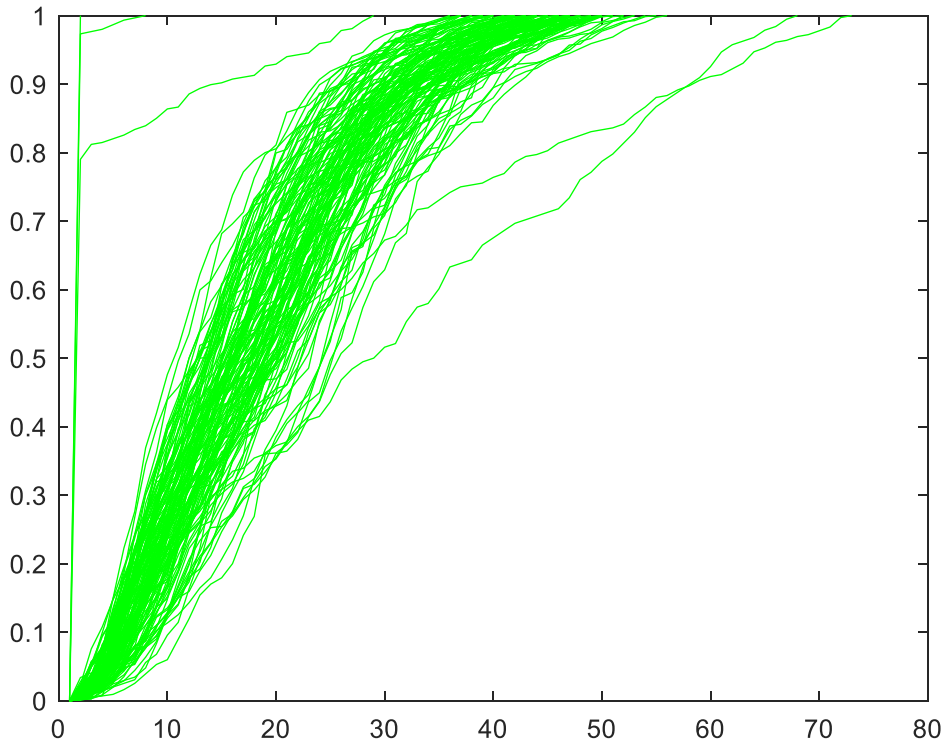

Figure 8: Cumulative distribution functions of each day acceleration values. Vertical axis denotes the total variability of the dataset (1 being max) while the horizontal axis denotes the percent of samples used. The data shows that while most days require about half the samples to cover the variability, a few days deviate from this in either having a higher variability (i.e., requiring more samples) or lower variability (i.e., requiring fewer samples). The plot also illustrates the high level of dissimilarity between the single days, and thus also high level of daily distributions mutual independency.

Therefore, there is no “typical” distribution of acceleration data from individual fish during the five months period. Since the statistical KS test is focused by principle on the distance of median values, it is actually expressing differences of common features in the data, not the differences of the rare artefacts that are on the tails of the distribution. This is why the information theory and Shannon’s entropy becomes useful due to the usage of the logarithmic function.

The information entropy was used to describe the data variability in the data points for each individual day, representing the amount of information given by the distribution of the acceleration values collected from the 21 individuals through each day (Figure 9).

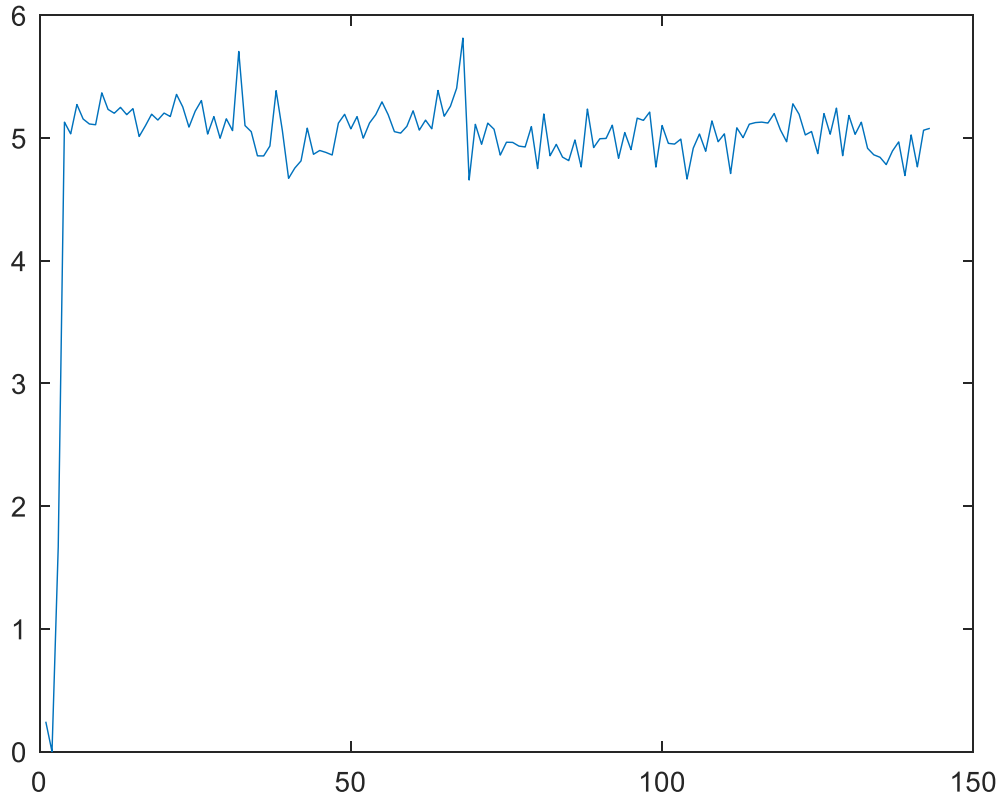

Figure 9: Information entropy values for each day of the experiment. Vertical axis: information entropy in bits; horizontal axis: experimental day number. Throughout the experimental period, the entropy value fluctuated around an information level of approximately 5 bits.

Since the information entropy is a variable itself, also the basic statistical methods, like central moments and confidence intervals, could be applied. The information entropy time evolution of Figure 9. illustrates the typical behaviour within the long measurement period as the behaviour with the informative value 4.96 bits with standard deviation of 0.67 bits.

Basically, it is difficult to define typical fish behaviour with exact mathematical attributes and their accepted range of values, conditionalities, and effects. However, it is possible to accept distributions of behavioural variables (such as acceleration) that are at a close to constant level of information as

behaviour that is not surprising or “common” behaviour. On the other hand, dramatical changes in the information level of a distribution could imply a surprise (the reaction to a sudden event). In the context of data analysis and anomaly detection, surprise refers to unexpected or atypical behaviour that deviates from the norm or expected patterns. Surprise is therefore always atypical behaviour.

Again, to illustrate the typical behaviour from an information theory point of view, two daily distributions of acceleration values were randomly chosen (day 78 and 120) with the entropy levels 4.93 and 4.97 bits respectively. The similarities in entropy imply that they should represent similar and exactly typical behaviour (Figure 10).

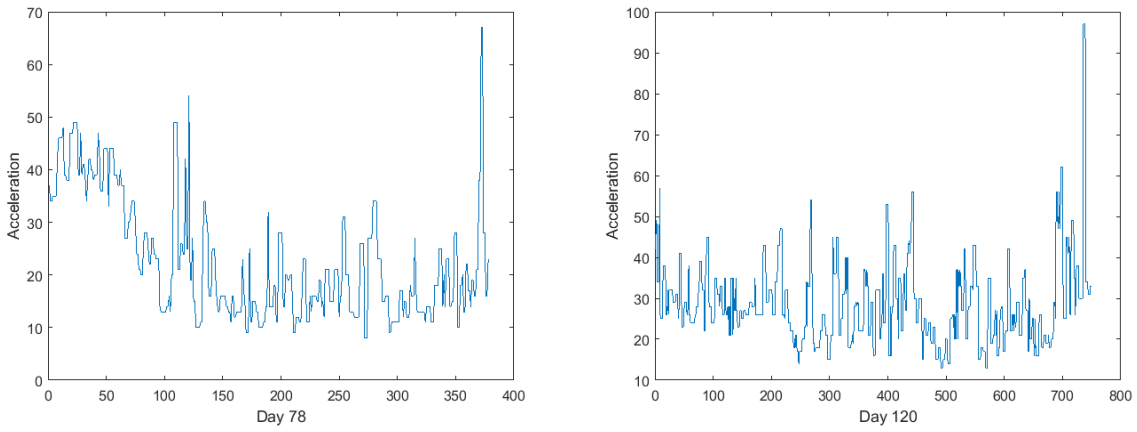

Figure 10: All acceleration values collected from the tagged fish in two separate days during the experimental period (day 78 on the left and day 120 on the right). Horizontal axis: sample number on the specific day ; vertical axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}). The corresponding entropy values (Day 78 S = 4.93 bit, day 120 S = 4.97 bits) are close but not the same, implying a larger variability on day 120.

Since the typical behaviour is partially described by the mean value of the information entropy, the identification of atypical behaviour is straightforward, using one sigma (standard deviation) interval as a criterion. To illustrate the differences between standard typical and one sigma atypical distributions, the plot of the CDFs plot was used (Figure 11).

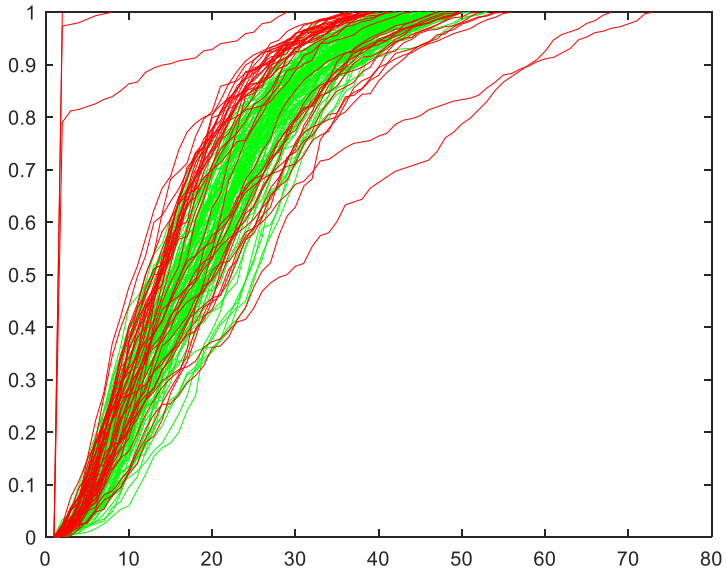

Figure 11: Cumulative distributions functions of individual days within (green) one sigma interval (4.96±0.67 bits) of information entropy and outside (red).

This implies that information entropy can be used to identify atypical distributions (red lines in the figure) based on the deviation from typical values interval (green curves in the figure). It is however important to mention, that while the typical distributions are considered typical and similar from an *information theory* point of view, statistical tests may still categorise them as different. Information entropy may thus identify elements that traditional statistics cannot detect, and thus extends the group of statistical parameters available for such analyses. This method is easy to use, and can in principle pinpoints the occurring surprises. In the case of fish telemetry, the levels of surprises are high enough to be measured by the Shannon's equation.

In our example, and selected dataset, we are searching for the extremely rare events, since they should have extreme differences in information. Changes in information simply says that something has changed and how much. Therefore, the general typical interval was extended to 3 sigma to reveal only the very rare events (Figure 12).

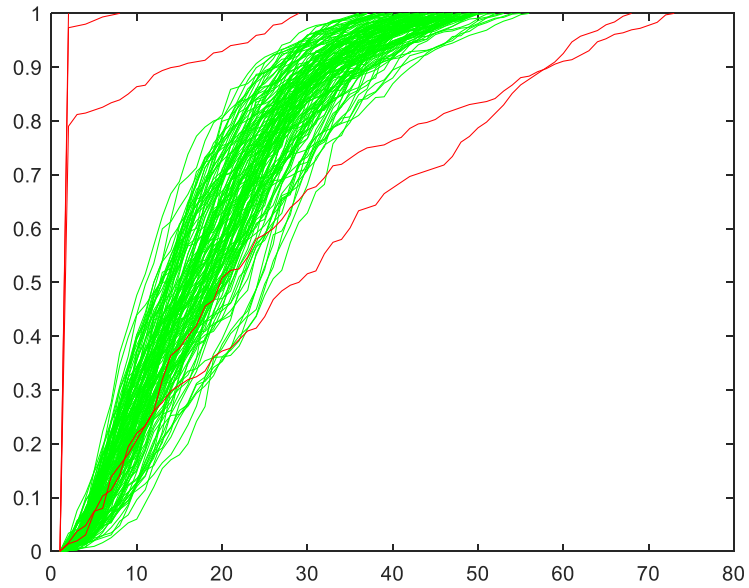

Figure 12: Cumulative distributions functions of individual days within (green) three sigma interval (4.96±2 bits) of information entropy and rare events (red).

This allows us to identify the most atypical days. It is now simple to check the distribution of acceleration values in such days, like day 32 with entropy 5.72 bits. The distribution is significantly different from the typical days (Figure 13 and Figure 14).

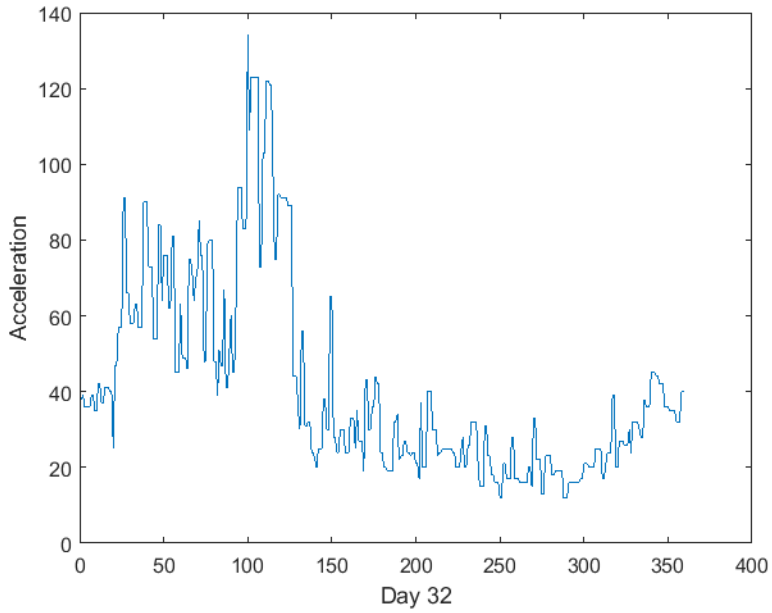

Figure 13: Example of atypical day, identified by the information entropy. The day matches with a delousing event. Horizontal axis: sample number on the specific day ; vertical axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2})

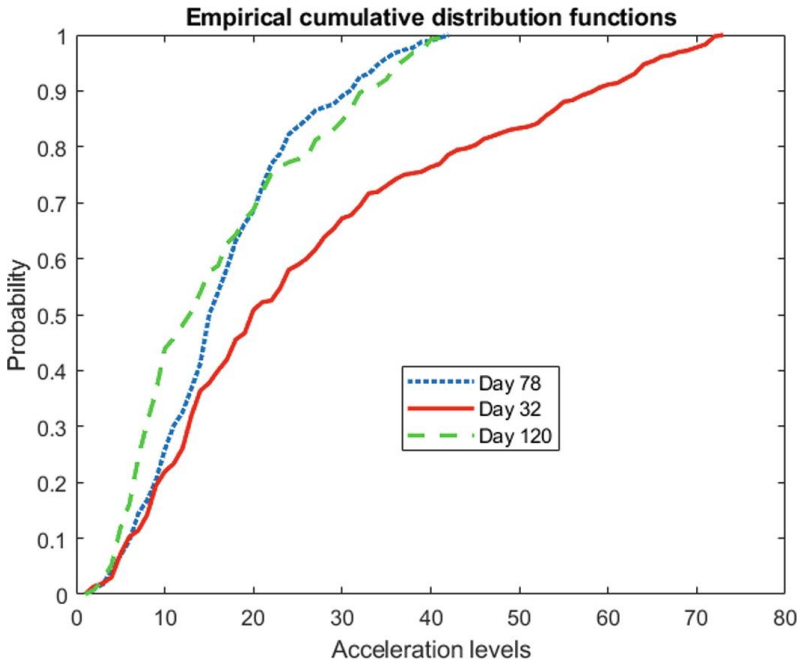

Figure 14: Differences in cumulative distribution functions of typical days (78,120) and atypical (day 32).

Using the evaluation of information entropy, four very atypical days were identified. Three of them correspond to exactly those days in which the delousing events occurred. The last atypical day was the very first day of the measurement, when fish were tagged and placed in the cage for the experiment.

This postprocessing of the telemetric data has shown the usability of the information entropy as a parameter for distinction of atypical behaviour in fish. However, in real situations such as crowding, this type of information will be needed in real time to elicit eventual warnings and therefore immediate evaluation and eventual introduction of measures to reduce stress. In order to achieve also the result in such situation, an experiment where the dataset was not divided into days, but where values were included in a computation which was processing datapoints “as they come.” With each new datapoint, the whole entropy was then recomputed (Figure 15 and Figure 16). As the computational window is longer and consecutively increasing, the absolute difference in information levels is smaller. Nevertheless, the changes in information are still significant when zooming in on the curve (Figure 16).

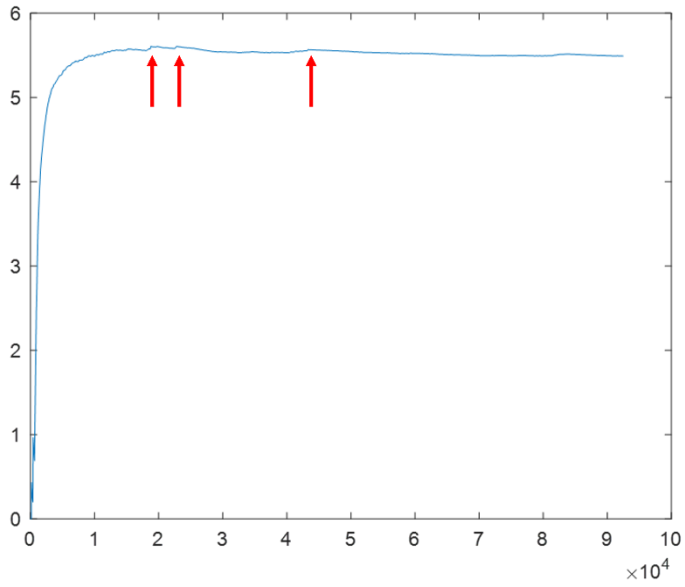

Figure 15: 'Real time' evaluation of information entropy over the entire dataset. The curve is smoother, since the changes are globally less surprising when considering all data values. Horizontal axis: data point number; vertical axis: cumulative entropy in bits. Red arrows indicate delousing events.

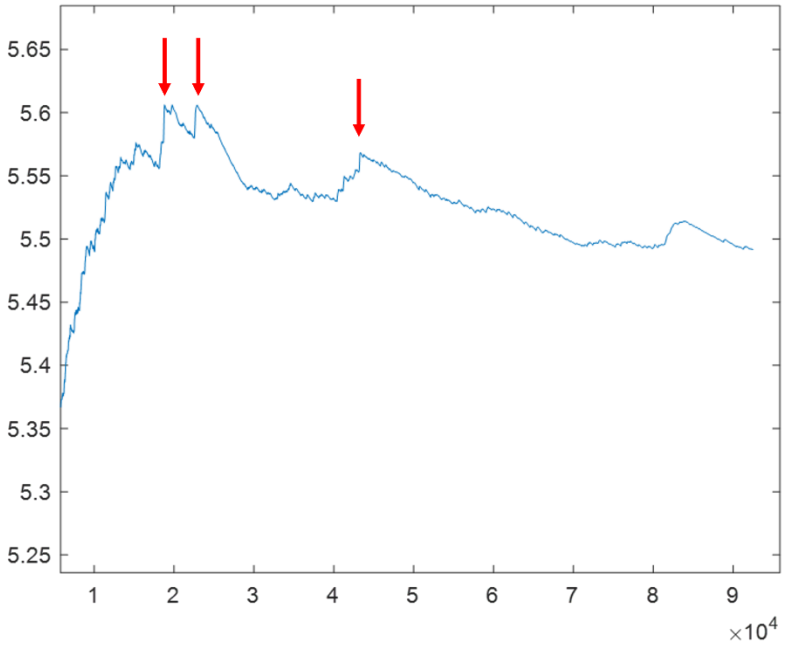

Figure 16: Details of entropy values. The three delousing events are still clearly visible. Horizontal axis: data point number; vertical axis: entropy in bits. Red arrows indicate delousing events.

M2: Causality

To illustrate approach for this analysis, where only causal relations between the states are considered, the average daily individual acceleration values were selected (Figure 17). Each level of variable values (state) has its own probability to occur (Figure 18). It is clearly seen that high acceleration levels have small (almost zero) probability of occurring (Figure 18). The shape of the

distribution is not Gaussian, yet close to it, and could probably be described as a kind of Poisson or power law process.

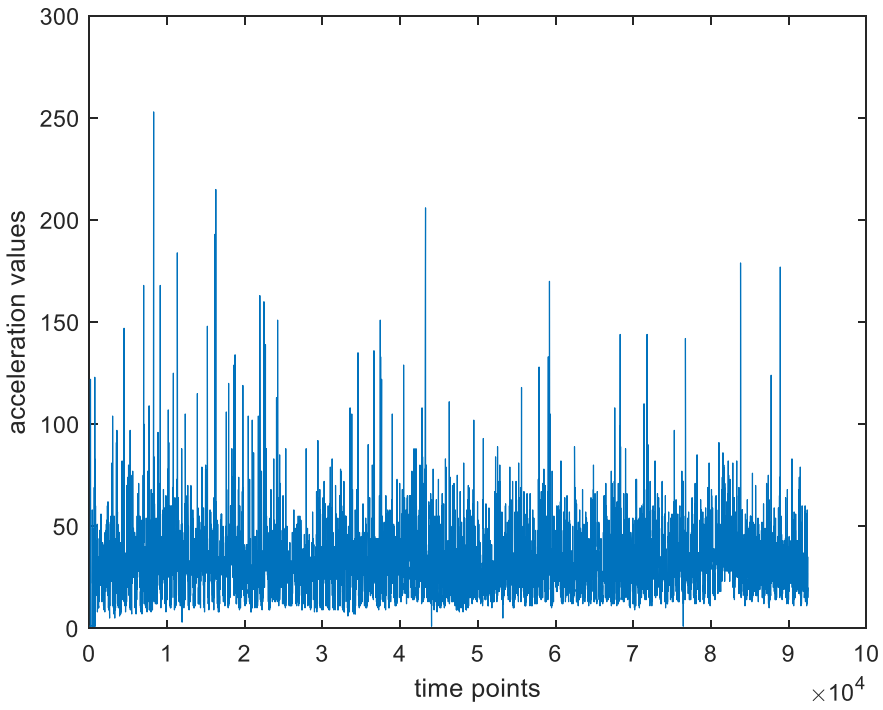

Figure 17: Whole 5 months series of acceleration values. Horizontal axis: sample number; vertical axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}).

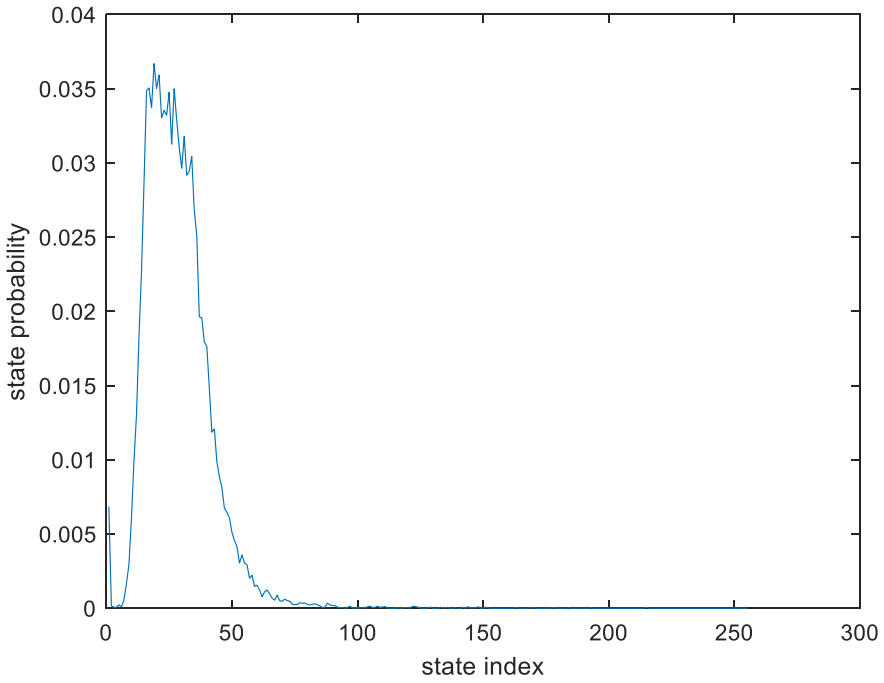

Figure 18: Probability of the fish acceleration values during the experiment. Horizontal axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}); vertical axis: probability of a specific digital value occurring.

Once the system is in a given state, the causality looks for the whole transition probability distribution, to determine which (of all) states will be the next one. For visualization, Cobweb like diagrams, where the x-axis shows the current state, and the y-axis the next state are useful (Figure 19). It is expectable, that according to the distributions, most of the points are in the lower level of the acceleration values. The obvious diagonal distribution implies strong causality between states k and k+1.

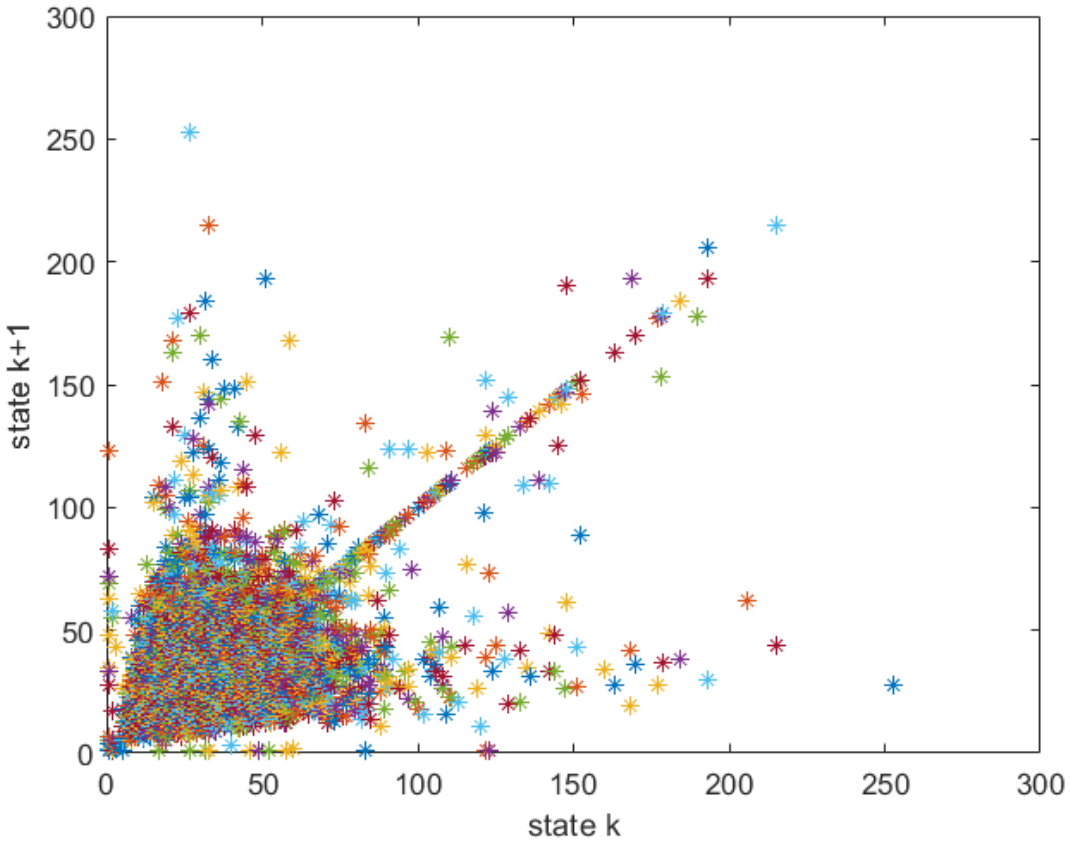

Figure 19: Diagram of state transitions. Horizontal axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; vertical axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k+1.

To clear the overlapping space in the transitions diagram, additional basic statistic evaluation could be applied, namely average k+1 state (mean), median, and modus (Figure 20). The diagonal trend in the relationship between k and k+1 is now even more apparent. However, the first central moment estimations represent only a partial information on the distributions, yet still informative. To understand more about the causal transitions, it is necessary to take into account the whole transition matrix, consisting of individual transition probabilities for each state (Figure 21). The transition matrix is relatively sparse, with not all possible transitions observed. The concentration of small probabilities in lower levels explains the swarm in the diagram. The diagonal tendency is again eminent, covering most of the highly probable transitions. The rest of the higher probabilities and most of smaller is under the diagonal.

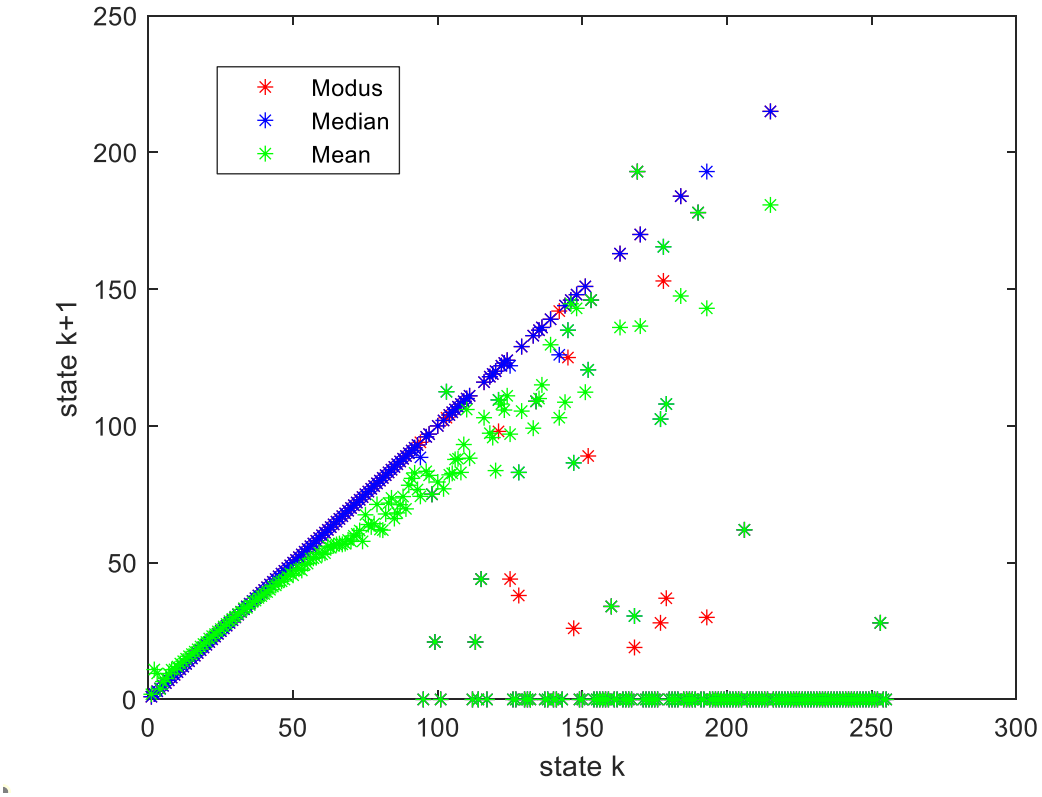

Figure 20: Mean, median, and modus diagram of state transitions. Horizontal axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; vertical axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k+1.

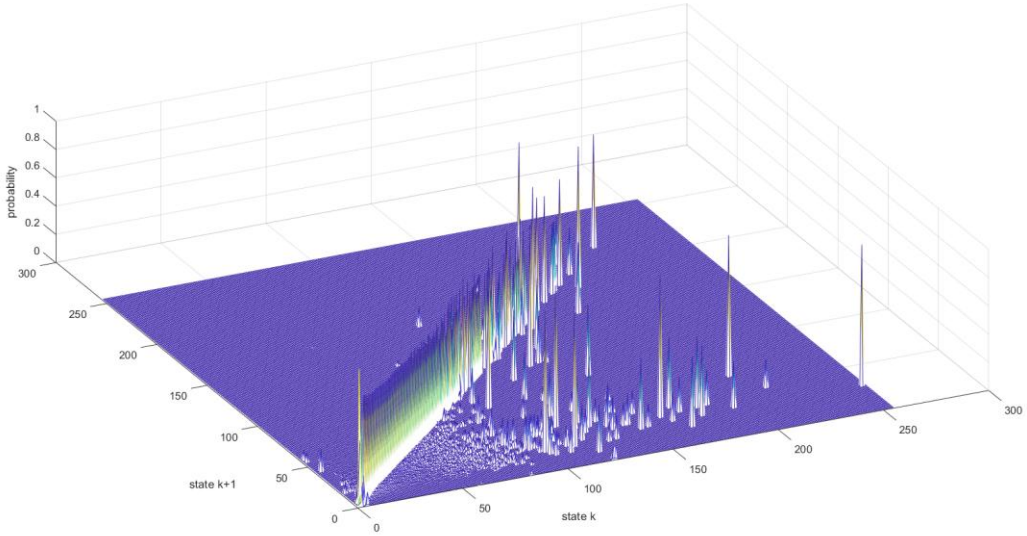

Figure 21: Transition matrix, for each state is plotted the whole probability distribution to reach the next state. Horizontal axis left to right: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; horizontal axis up-

down: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k+1; vertical axis: probability of transition from k to k+1.

The transition matrix itself has to be considered conditional, since each state k has its own probability to occur (Figure 18). To obtain the image of the whole transition distribution in the dataset, the transition matrix distributions have to be weighted by the probabilities of each state (Figure 22). The diagonal area is now correlated with the states probability. This implies that there are four rare events, with high transition probability from low probable states.

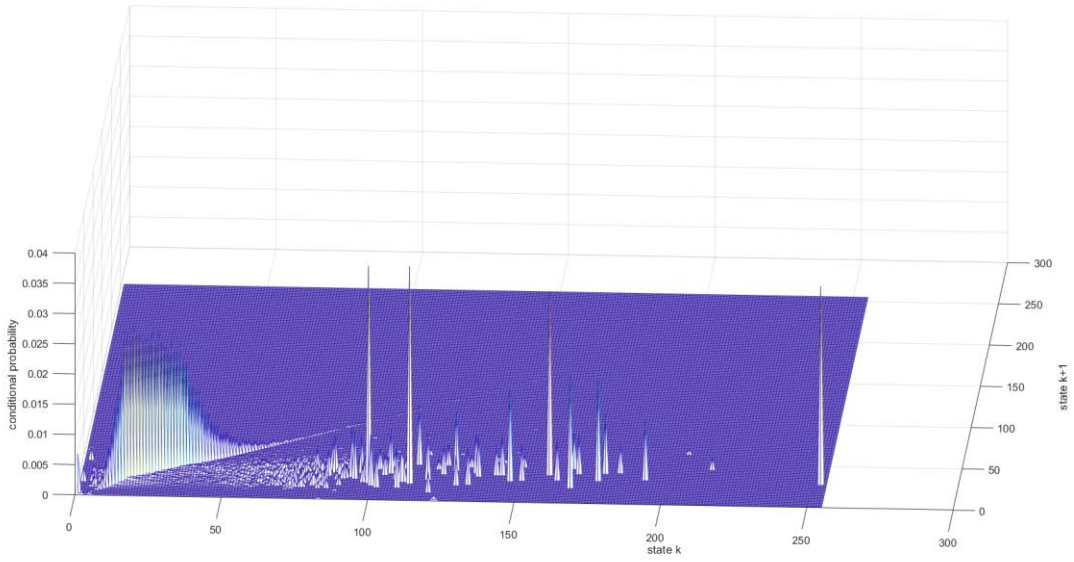

Figure 22: Weighted transition matrix showing the whole probability distribution, with four rare events. Horizontal axis left to right: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; horizontal axis up-down: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k+1; vertical axis: probability of transition from k to k+1.

To understand the diagonal trend, all transitions in the original transition matrix (still independently for each k state) were classified into three groups: i) holding transition, transition to itself, therefore implying no change of the state; ii) decreasing transition, therefore transition to the lower state level; iii) increasing transition, therefore transition to a higher state level (Figure 23). This simplification allows us to detect that there were high levels of state holding transitions, indicating that in most cases, average fish acceleration does not change. The trends of the increase and decrease are opposite, therefore some equilibristic cycles in state transitions exist. As the higher states are of lower probability, more outliers are present there. Now, as previously with the whole transition matrix, it is necessary to weigh the classification by the probability of each state occurrence to reveal the whole distribution.

The distributions of holding, decreasing, and increasing transitions all follow the trend of distribution of individual states (Figure 24). Finally, to obtain the total information about the classified groups, the weighted conditional distributions have to be summed according to the total probability law across the state levels: decrease 15.6 %, holding 68.9 %, increase 15.5 %. The probability of a state transition to result in no change is thus almost 69 percent. The rest of the transitions are divided

almost exactly half between leading to increased or decreased accelerations. The final distribution of only three classes has strongly Gaussian property (classification corresponds to one sigma borders).

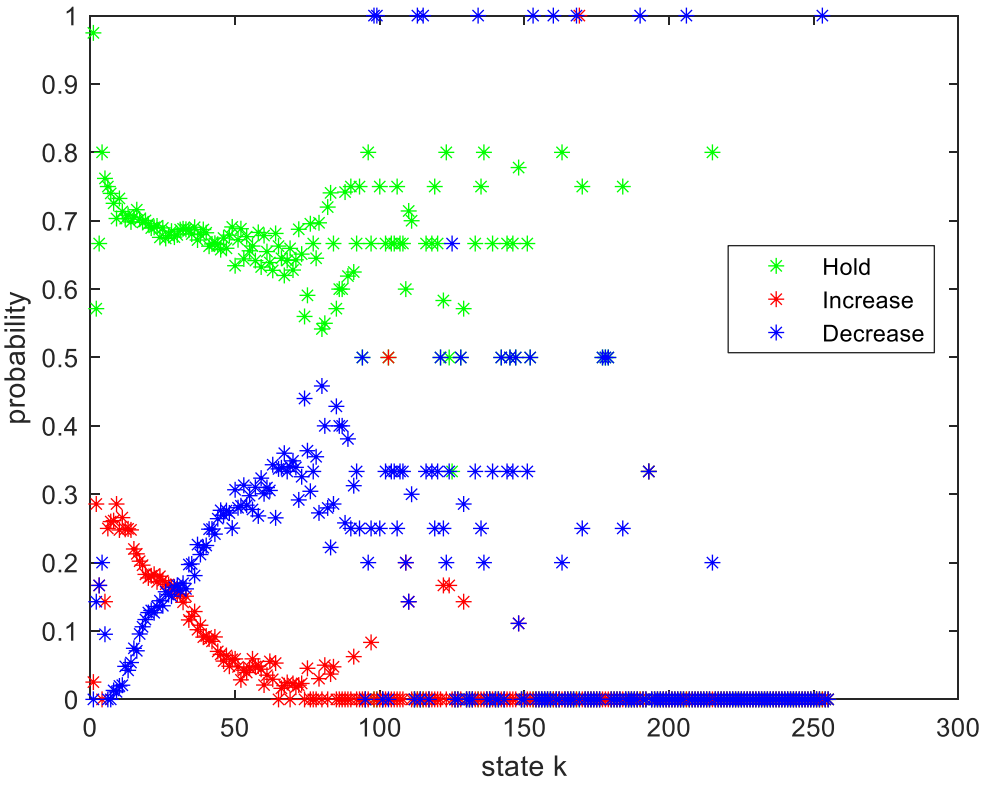

Figure 23: Classification of the state transition: holding the state level, decreasing the state level, and increasing the state level. Horizontal axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; vertical axis: probability of transition from k; colours denote whether the transition is to a lower (purple), higher (red) or the same (red) value .

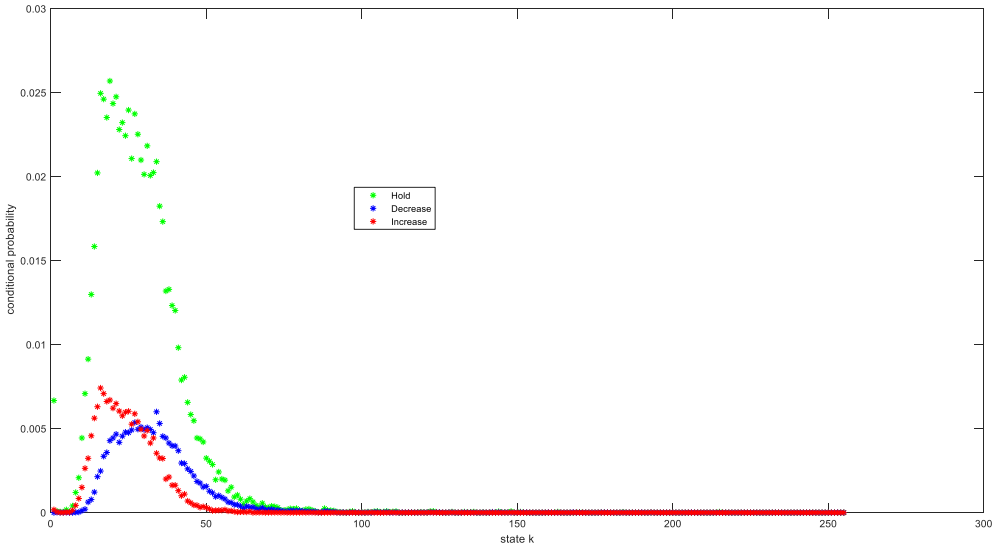

Figure 24: Total distribution of state transition classifications, all following the same tendency. Horizontal axis: digital value for acceleration (0-255 representing a range of 0-3.476 m s^{-2}) of state k; vertical axis: probability of transition from k; colours denote whether the transition is to a lower (purple), higher (red) or the same (red) value.

4.2.3. Discussion

The main problem of the definition of the typical behaviour is key for qualitative analysis, but not for quantitative analysis. First of all, the normality of data is not to be expected, and all tests have to be nonparametric. Second, due to the common fluctuations of circadian rhythms, the median of the distribution could be a moving variable, therefore even nonparametric test may reject the null hypothesis of the same population distribution. Third, the extremes with small probability will not be considered by such tests, even if they have periodical behaviour. In other words, typical behaviour does have to be one single type of distribution, but the whole set. But this is not what we are looking for *a priori*; we are rather looking for an a posteriori method which can unequivocally pinpoint the atypical behaviour.

It was already proven that entropy concept overpasses the nonparametric tests for such a task. Entropy is a clearly distinctive method with much lower computational burden (1.75 times faster in comparison of only CDF empiric estimation). Thus, methods like information entropy and causal transitions can potentially help with the application of such tools for monitoring fish welfare by analysing the movement patterns of fish, improving the efficiency of data analysis, identifying changes in behaviour or activity levels that may indicate stress or other welfare issues, helping modify experimental protocols or improve living conditions to minimize the potential for harm or distress. The entropy approach is simple in principle but offers additional avenues for evaluating positioning data and other telemetry datasets to the classical statistical analysis. This is becoming important for the fish welfare indicators estimation, since the definition of typical behaviour is yet under investigation. The entropy values analysis could be applied to a variety of simple subtasks like continuous analysis of the behaviour, division of the sets to smaller subsets, and indicating when the

fish started to behave atypically. Moreover, the analysis could help to e.g., distinguish which fish are behaving atypically, which speed, depth, or even pairs of values are atypical or typical. Other potential applications of entropy analysis on telemetry data for fish welfare studies may include analysing the variability of feeding behaviour, social interactions, changes in behaviour or welfare in response to different environmental conditions or stressors.

Both approaches have a potential to be a valuable tool in fish welfare assessment, as they allow for quantifying the complexity and variability of fish behaviour, which can be used as an indicator of their welfare. By analysing the information of fish movement patterns and behaviour, it may be possible to identify abnormal behaviour and detect potential welfare issues in aquaculture systems. Additionally, by combining additional analysis with other telemetry and sensor technologies, it may be possible to develop early warning systems for detecting and addressing welfare issues before they become serious. Overall, while it is not a comprehensive solution for assessing fish welfare, these methods can provide valuable insights into the complexity and variability of fish behaviour and may be useful tools for developing more effective management practices and promoting the welfare of captive fish. Furthermore, information, causality, and statistics-based techniques can also help to standardize telemetry data analysis and provide objective indicators of fish welfare, which can be useful for management and regulatory purposes. In conclusion, these novel approaches hold significant potential for improving fish welfare assessment and management in the context of fish telemetry.

The outcomes of this study were presented at the 10th International Work-Conference on Bioinformatics and Biomedical Engineering (Urban, 2023), Gran Canaria, Spain, and Aquaculture Europe 2023 conference, Vienna, Austria.

5. Conclusion and recommendations

5.1. New methods, their applicability, and possible new data types

Based on the findings of the master student at NTNU (Smedhaug, 2023), it appears that pre-processing in terms of identifying consistent trajectories with a high causality between the data points is a good way to refine a dataset consisting of 3D positions for more accurate analyses of short-term individual behaviours. Moreover, the results indicate that PCA has potential in shedding light on short-term salmon behaviour, and the composition of various different behavioural expressions comprising the “total” behaviour of fish. While not equally easy to interpret, unsupervised clustering also seems like a viable tool for identifying specific individual behavioural traits, and possibly also a method for better distinguishing these traits in both visual and numerical manners. While both these methods are likely to perform better and more consistently if provided a dataset where data points are denser in time, and by deriving more variables to describe the trajectories, the data used here, and variables derived from these was sufficient to demonstrate their applicability for further processing such data to achieve new insights and knowledge. We thus

recommend that PCA and clustering using HDBScan or other methods be further explored in the future, and that an extra effort is put into identifying suitable variables for such analyses.

The entropy approach, taken from information theory, is a simple, yet powerful, tool for evaluation or measure of changes in the distribution of values. Such changes always mean change in the amount of information. The classification of the information changes could distinct typical changes from atypical ones. Further investigation of the conditionality in the information promises introduction of the entropy evaluation as one of the possible welfare indicators to distinguish atypical behaviour.

The achieved results of the student at JU also suggest that the transition dynamics between the states deserves additional attention, since the conditional probabilities are essential for determining the state changes. The activity telemetry data examined showed a tendency to patterning, however the success of this method depended on how measurements are classified before analysis. Therefore, it is necessary to take into consideration classical binning problems. In the apparent randomness of the conditional probabilities, hidden cycles, with their own fluctuations, could be present. The classification methods should address such fluctuations to standardise and amplify the patterns in the transition matrix.

6. Acknowledgements

The authors of this deliverable dedicate this work mainly to the two master students, Even Åge Smedhaug at NTNU and David Laštovka at JU, as they have been instrumental in exploring the methods presented here.

7. References

Brijs, J., Sandblom, E., Axelsson, M., Sundell, K., Sundh, H., Huyben, D., Broström, R., Kiessling, A., Berg, C. and Gräns, A., 2018. The final countdown: continuous physiological welfare evaluation of farmed fish during common aquaculture practices before and during harvest. Aquaculture, 495, pp.903-911.

Dubey, A., 2018. The mathematics behind principal component analysis. available online in https://towardsdatascience. com/the-mathematics-behind-principalcomponent-analysis-f2d7f4b643.

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. and Berckmans, D., 2018a. Precision fish farming: A new framework to improve production in aquaculture. Biosystems Engineering, 173, pp.176-193.

Føre, M., Svendsen, E., Alfredsen, J.A., Uglem, I., Bloecher, N., Sveier, H., Sunde, L.M. and Frank, K., 2018b. Using acoustic telemetry to monitor the effects of crowding and delousing procedures on farmed Atlantic salmon (*Salmo salar*). Aquaculture, 495, pp.757-765.

Juell, J.E. and Westerberg, H., 1993. An ultrasonic telemetric system for automatic positioning of individual fish used to track Atlantic salmon (*Salmo salar* L.) in a sea cage. Aquacultural engineering, 12(1), pp.1-18.

McInnes, L., Healy, J. and Astels, S., 2017. hdbscan: Hierarchical density based clustering. J. Open Source Softw., 2(11), p.205.

Oppedal, F., Dempster, T. and Stien, L.H., 2011. Environmental drivers of Atlantic salmon behaviour in sea-cages: a review. Aquaculture, 311(1-4), pp.1-18.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, pp.2825-2830.

Smedhaug, E. Å., 2023. Machine learning for identification of individual salmon behaviour in aquaculture (Master's thesis, NTNU).

Urban, J., 2023. Entropy Approach of Processing for Fish Acoustic Telemetry Data to Detect Atypical Behavior During Welfare Evaluation, Lecture Notes in Bioinformatics, 2023, Springer.

Wu, H., Aoki, A., Arimoto, T., Nakano, T., Ohnuki, H., Murata, M., Ren, H. and Endo, H., 2015. Fish stress become visible: A new attempt to use biosensor for real-time monitoring fish stress. Biosensors and Bioelectronics, 67, pp.503-510.

Document Information

EU Project	No 871108	Acronym	AQUAEXCEL3.0
Full Title	AQUAculture infrastructures for EXCELlence in European fish research 3.0		
Project website	www.aquaexcel.eu		

Deliverable	N°	D4.3	Title	New methods for post-processing biotelemetry data in aquaculture
Work Package	N°	4	Title	Technological tools for improved experimental procedures
Work Package Leader	Finn Olav Bjørnson, SINTEF			
Work Participants	NTNU, JU			

Lead Beneficiary	NTNU, 10
Authors	Martin Føre, NTNU, martin.fore@ntnu.no Jan Urban, JU, urbanj@frov.jcu.cz Jo Arve Alfredsen, NTNU, jo.arve.alfredsen@ntnu.no
Reviewers	Marie Laure Begout, Marie.Laure.Begout@ifremer.fr

Due date of deliverable	31.10.2023
Submission date	30.10.2022
Dissemination level	PU
Type of deliverable	R

Version log			
Issue Date	Revision N°	Author	Change
30.10.2023	1	NTNU/JU	First version