Masteroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for teknisk kybernetikk

Armon Hakimi

Optimization of measurement setup for the surveillance of loads, movements, and deformation in fish farms

Masteroppgave i Kybernetikk og Robotikk
Veileder: Martin Føre
Medveileder: Pascal Klebert
Juni 2023

NTNU
Kunnskap for en bedre verden

SINTEF

Armon Hakimi

Optimization of measurement setup for the surveillance of loads, movements, and deformation in fish farms

Masteroppgave i Kybernetikk og Robotikk
Veileder: Martin Føre
Medveileder: Pascal Klebert
Juni 2023

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi og elektroteknikk
Institutt for teknisk kybernetikk

NTNU

Kunnskap for en bedre verden

Abstract

This thesis presents an investigation into the application of dimensionality reduction techniques, specifically Principal Component Analysis (PCA), to sensor data from a fish farm sea cage. The goal of the thesis is to find a more sparse sensor setup that retains most of the information of the full setup. The primary focus lies on depth sensor data, but accelerometer and load shackle data is also analysed.

The study reveals that the PCA can be effectively applied to depth sensor data, indicating the existence of optimized sensor setups that don't lose significant amounts of information regarding depth dynamics. However, the PCA's application to accelerometer and load shackle data proved more challenging, highlighting the limitations of the PCA and the potential presence of non-linear relationships in these sensors.

Alternative approaches and techniques, such as t-SNE, UMAP, autoencoders, and wavelet transforms are discussed as they might be better suited to uncover the patterns in the accelerometer and load shackle data. The importance of future research in this area is also discussed.

While this study is not comprehensive enough to crown one single sensor setup as most optimal, it offers valuable and novel insights that can help optimize sensor setups in the aquaculture industry. Thus, the findings presented in this thesis contribute to the ongoing academic efforts, while laying the foundation for future research aimed at improving the optimization of sensor setups in the fish farming industry.

Acknowledgements

I must begin by extending my deepest gratitude to my main supervisor Martin Føre. His genuine and deeply humane interest for my well-being, beyond the confines of this thesis, instilled an immense sense of comfort, and had a profound impact on me. I cannot explain it.

A big thanks must also be given to my co-supervisor, Pascal Klebert, who was invaluable in providing crucial information and fervently suggesting possible approaches.

That being said, I wouldn't even have made it here without the support of my mom, dad, and sister. Bless them, they really do mean the world to me.

Special shout-out to my study buddies who distracted me all too often; who added a spark of life to my daily routine.

...and finally: wow. i did it:))

Table of contents

Figures

Tables

Chapter 1

Introduction

1.1 Motivation

Aquaculture is a growing global industry that has been highlighted as one of the key future providers of food to support a growing world population. In 2013, the World Bank predicted that 62% of all seafood would be farm raised by 2030 [1]. Through its growth over the last decades, the Norwegian salmon industry has evolved to become one of the most important national industrial segments, with an export of 1.25 million metric tons of salmon in 2022, at a record-high total value of 105.8 BNOK [2]. These values are predicted to reach 5 million metric tons at a value of 500 BNOK by the year 2050 [3], although the stagnation experienced in recent years might put a stopper to this goal.

The last couple of decades has seen a trend within fish farming where farm sites are increasingly being established at more remote sites, mainly because there is a shortage of coastal sites suitable for fish farming, but also because the conditions further from shore may be more beneficial for the fish. For more information regarding the challenges and considerations in moving offshore, the reader is referred to [4, 5].

Suffice to say, these "offshore-sites" are typically more exposed and therefore more likely to experience larger and more extreme environmental forces than previous sites. This, together with the reduced capability for human intervention (due to remoteness) means that structural monitoring is even more important for exposed sites than for conventional fish farms. However, due to the flexibility of the sea cages it can be difficult to identify exactly *which sensors are needed* and *where they should be placed* to best understand how the environmental forces cause structural movements and deformations, and how this affects the fish farm. This thesis aims to produce novel insights to this problem.

1.2 Scope of work

The scope of this work is restricted to the application and assessment of dimensionality reduction techniques, specifically Principal Component Analysis (PCA), to sensor data collected from

a fish farm sea cage located in Buholmen in Norway. The primary focus is on depth sensor data, but the work also includes various attempts at reducing the dimensionality of accelerometer and load shackle data.

The study starts with the application of PCA to depth sensor data in several ways, investigating the potential of the PCA in reducing the number of sensors while retaining as much information as possible. Through this exploration, the thesis delves into finding an optimal sensor setup in the fish farming industry.

Next, the effectiveness of the PCA is investigated when applied to accelerometer and load shackle data. Given the lack of dimensionality reduction when applied to these sensor types, this thesis then goes on to examine assumptions of the PCA, before discussing the potential need for alternative approaches better suited to handling non-linear data. This is limited to a brief discussion of techniques such as t-SNE, UMAP, autoencoders, and wavelet transforms.

Finally, while this thesis presents some significant and novel findings, it does not deliver one single optimal sensor setup that is necessarily applicable to all kinds of fish farms. Rather, this thesis presents the findings in light of what they indicate based on the conditions from which they were derived. While this can be useful in and of itself, it acts mainly as a foundation for further research. The findings are presented with the hope that future work will be able to build upon and further refine the results.

1.3 Main contributions

This thesis makes several key contributions that help understand how the measurement setup on fish farms can be optimized.

The first and perhaps most significant contribution comes from using the Principal Component Analysis (PCA) on depth sensor data. The analyses carried out on these sensors highlight the potential in using PCA as a dimensionality reduction tool for depth sensors. Furthermore, the results clearly point to certain sensors being more "important" than others. This novel finding could have important implications for the optimization of measurement setup on fish farms. Some of these implications include potential reductions in costs through less equipment, and consequently also less maintenance.

The second contribution of this thesis lies in the limitations of the PCA when applied to accelerometer and load shackle data. The lack of dimensionality reduction for these sensors might suggest that they contain non-linear relationships, or that they indeed all are "equally important" in understanding the dynamics at the fish farm. These insights highlight the fact that there is still a need for more work, especially in trying alternative dimensionality reduction techniques that can handle non-linear data, or other approaches altogether.

Thus, this thesis paves the way for future research, listing and discussing some promising methods that can be explored, as well as demonstrating the utility of the PCA. These are mentioned briefly above and expanded upon in later sections. The discussion of these future research directions can also be considered a contribution, providing a good starting point for further studies in the area.

1.4 Structure of the report

Chapter 2 - Background: This chapter is dedicated to providing the reader with background material that gives context to the rest of the report. This includes a short rundown of the history of fish farming in Norway and a brief discussion of related works. Then comes an overview of the various dynamics that affect sea cage structures, how each of these are measured, and some of the associated challenges. This is followed by a quick discussion of the various approaches to optimizing sensor setups in the context of this thesis. Finally, the tools that will be used in this study, namely PCA and Butterworth filters, are explained in sufficient detail.

Chapter 3 - Methodology and Experimental Setup: This chapter starts by presenting the site from which data is collected, followed by an overview of the sensor setup and how each sensor collects data. It then presents the preprocessing steps that were applied to each of the sensor types, before giving a detailed description of the various analyses that were carried out.

Chapter 4 - Results: This chapter presents and briefly comments on the results of applying PCA to sensor data in a multitude of different ways. It also presents the results of trying to use PCA to reconstruct depth sensor data.

Chapter 5 - Discussion: This chapter starts with a summary of all the findings, before providing a more in-depth interpretation of all the observed results. This chapter then goes on to discuss the implications of these findings, along with the limitations of the study. Finally, alternative approaches and future research is discussed.

Chapter 6 - Conclusion: A summary is provided of the key findings of the thesis and their implications. This chapter then briefly recounts the possible explanations for the observed results before reiterating on the future work that be done to build upon the presented results.

Chapter 2

Background

2.1 Norwegian salmon farming and the shift to more exposed farming sites

Fish farming in Norway has a long and rich history, dating back to the viking age. At some point, it is thought that in addition to traditional fishing, vikings developed simple ponds to produce fish. While fish farming has existed for several centuries, it wasn't until the late 1960s that modern fish farming really began taking shape in Norway. During this time, more and more attention shifted towards marine-based fish farming, starting what would lead Norway on the road to becoming the largest salmon farming nation.

Norway's first fish farm was deployed off the island of Hitra in 1970, by brothers Ove and Sivert Grøntvedt. They devised a system where the rearing of smolt was conducted in land-based facilities, while most of the growth would be achieved in marine fish farms that were deployed in the sea. After their first successful harvest in 1971, several other salmon farms followed suit. The first decade of fish farming in Norway was summarized by rapid growth, with annual production increasing from 500 tons in 1971 to 8,000 tons in 1980. [6]

The 1980s were marked with a similar growth, largely fuelled by the introduction of Norwegian salmon in Japanese sushi. During this decade illness among fish started becoming a problem, spawning various research projects that aimed to better understand and improve fish health. By the time the decade was over, production had increased from 8,000 tons to 170,000 tons in 1990. [6]

The 1990s saw great developments in the use of vaccines for fish farming. Prior to this decade, disease outbreaks could lead to significant losses and were starting to become a big issue. Vaccines made the use of antibiotics in salmon farming almost obsolete.

The 2000s saw the introduction of stricter regulations to ensure environmental sustainability. In 2005, the Norwegian parliament passed several laws that were designed to facilitate sustainable development [7]. With the new system, aquaculture facilities were required to carry out comprehensive environmental assessments considering factors such as water quality and

impact on wild fish populations before being granted the right to operate. Aquaculture farms were now also required to report disease outbreaks and take certain measures to prevent the spread of diseases between farms. Efforts were also made towards reducing the number of fish that escape farms. In 2006, more than 900,000 fish escaped Norwegian fish farms, although this number has been greatly reduced in the years since thanks to stronger structures and better monitoring of farms [8].[1] This started marking the need for better structural monitoring as a means to prevent escapes.

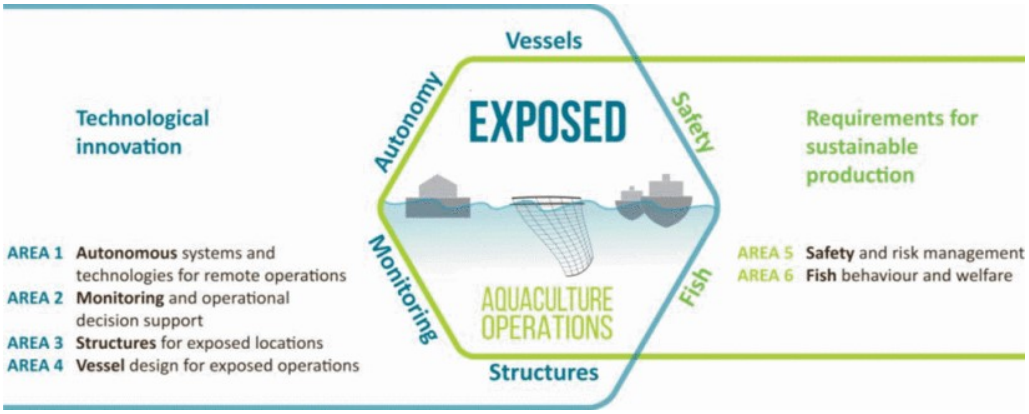


Figure 2.1: Autonomy and monitoring are two of the areas within technological innovations that will allow aquaculture operations to move to more exposed locations. Image courtesy of Bjelland *et al.* [4]

One area that received a lot of attention in the 2010s and that continues to be a problem to this day is lice infestations. As with disease, lice infestations among fish pose one of the biggest threats to the industry [11]. More recent developments in methods aimed at preventing lice infestations among farmed salmon are reviewed in [12].

As the demand for farmed salmon continued to grow and stricter regulations were put in place, Norwegian salmon farmers also had to start moving their fish farms further offshore in the 2010s. This led to new environmental challenges that still make it harder to monitor and maintain fish farms. In particular, stronger currents are a big issue as they can severely deform the cages, causing fish to die from a lack of space [13]. Several areas of focus have been identified as key components in enabling fish farms to move further offshore. Some of these are shown in Figure 2.1, including technologies such autonomous feeding systems and automated cage cleaning systems [4]. The progress of autonomous solutions is perhaps best appreciated when considering that this decade also saw the introduction of the first remotely operated fish farm [14]. Although most of these technologies remain largely under development to this day, they have proven that they can help improve efficiency and sustainability [15].

All in all, considering the continued efforts to reduce fish escapes and the myriad challenges associated with moving fish farms to more exposed locations, it is clear that more sophisticated

[1] Readers interested in learning more about escapes from fish farms and its various causes are referred to [9, 10]. These articles delve much deeper into the issue, presenting more numbers and detailed analyses of the various causes for fish escapes in the period from 2000 to 2018.

structural monitoring will be crucial.

2.2 Related works in structural monitoring

Marine fish farms consist of several sea cages where all the components (i.e., ropes, nets, floating collars) are flexible such that they comply with, rather than resist, environmental forces. This effectively reduces the strain on components due to environmental excitation during harsh weather and other demanding events, but it also has the adverse effect of making sea cages particularly difficult to monitor. Structural monitoring is an important element in fish farming as it is used to evaluate the structural integrity of existing facilities, or for planning the development of future farms and their location. In both cases, the main aim of structural monitoring is to predict/detect events such as net ruptures, which can ultimately lead to fish escapes, loss of equipment/infrastructure, and impair the welfare of the fish.

While somewhat limited in number, some studies have been conducted to monitor the forces and their effects on full-scale fish farms, using various sensor setups. Fredriksson *et al.* used a non-invasive optical measurement system to measure a sea cage's heave, surge, and pitch response to different wave elevations [16]. The tension in the anchor line was also measured, and all this was used to validate various numerical models. In another study, Lader *et al.* used an acoustic current meter and depth sensors to study net deformations relative to various incoming currents, emphasizing the importance of multiple current measurements due to complex eddies at one of their sites [17]. In yet another study, DeCew *et al.* combined acoustic sources, hydrophones, and current meters to investigate current-induced shape changes in a small-scale fish cage [18]. These studies nicely demonstrate how different sensors can be used to investigate sea cage deformations in response to different or varying current profiles, as well as its importance in ensuring fish welfare.

A great amount of effort has also been put into modelling and simulating sea cages with various designs and conditions, ranging from basic studies of net panels in flow [19] to full scale simulations of cage dynamics [13]. One such paper examines the wake effect on aquaculture nets with different angles of attack and current velocities [20]. In another research project, Moe-Føre *et al.* examine how different cage models can yield different deformation predictions [21]. Specifically, they test the triangle and spring models in FhSim[2] as well as a truss model using ABAQUS and MATLAB, concluding that each model has it's strengths and weaknesses.

In a paper by Endresen and Klebert, different flexible cage designs were tested to see which ones fit best with physical models, all while varying the loads used in each design [24]. Amongst other discoveries, they find that using lighter weights on the different cage designs yields inaccurate numerical results, likely due to the global deformation of the cages and subsequent breakdown of model validity.

[2]Many of the articles mentioned here use FhSim. Developed by SINTEF Ocean, FhSim is an extremely flexible and efficient simulation tool. Its primary purpose is to aid in the design, analysis, and operation of marine systems, including aquaculture facilities. It uses a highly modular modeling approach to enable the simulation of interconnected systems with complex dynamics. This includes mechanics, hydraulics, electric power, and control systems among others. FhSim's inner workings won't be discussed further in this thesis, but can be found in articles such as [22, 23], or at their website `https://fhsim.smd.sintef.no/`.

More recent attempts at using sensor data to predict deformation include a recent paper by Su *et al.* where they examine the use of real-time measurements from an underwater positioning system to predict the cage's deformation [25]. The positioning system used in their study consists of three acoustic sensors that are mounted on the cage at different depths and locations. Their work validates that the approach is well suited for general-purpose monitoring of cage deformations. However, they also conclude that more sensors (or a combination of different sensors) are needed for higher accuracy. This leads nicely into the research that has been undertaken and shall be presented in this thesis.

2.3 Structural monitoring methods

In this section, the reader presented with the various dynamics that are important to consider when monitoring the structural integrity of a sea cage. In addition, the parts that follow briefly discuss how each aspect is attempted to be captured through the use of a specific type of sensor. The experimental setup of each sensor type is given in Section 3.1.

2.3.1 Wave induced deformations

Wave induced deformations represent an important aspect of structural monitoring in floating fish farms. Waves cause deformations that strain the fish farm structure, leading to damage or potentially failure over time. Observing and understanding these deformations plays a key role in allowing the development of future strategies that can minimize or mitigate some of the damage.

Accelerometers can be used to measure the vertical movement of the floating collar due to waves. This in turn says something about the wave induced deformations and the overall severity of the incoming waves. The main challenge with accelerometers is that their signals can be quite noisy, often caused by various environmental factors, such as wind, or other small fluctuations. Such disturbances can affect the accuracy and reliability of the collected data. When higher accuracy is needed, various signal processing techniques can be used to filter away some of the noise, improving the quality of the measured data.

2.3.2 Structural loads and tensions

In addition to causing deformations, waves also strain the system that is responsible for holding the sea cage in place: the mooring system. This is a crucial aspect that must be monitored to ensure the safety and stability of the fish farm. In particular, operators must be able to detect when the load exerted on the system is so high that it can cause damage to the sea cage. To this end, load shackles are used to monitor the forces acting upon the mooring system. Their purpose is to provide a reliable and accurate way to measure the weight and force being exerted on the system.

Some of the challenges associated with using load shackles lie in the difficulty of installing and maintaining them. As opposed to the accelerometers that are positioned on the floating collar, above water, the load shackles are partly submerged in water, weigh much more, and

are always under some tension when operating. This makes the inspection or replacement of malfunctioning load shackles much more challenging than with accelerometers. Furthermore, while accelerometers are unlikely to experience a sudden failure, load shackles are constantly withstanding great environmental forces, making them more likely to malfunction.

2.3.3 Net cage deformations

As opposed to the two previous dynamics, net cage deformation is exclusively the result of underwater phenomena, specifically: currents. Observing and limiting net cage deformation is of utmost importance as severe deformations can lead to big losses of internal volume. This has been shown to negatively affect fish in several ways, particularly when the stocking density is high [25, 26]. In extreme cases, mass mortalities of up to 40 tons of fish have been observed [13]. Depth sensors play a vital role in detecting these deformations.

Being the only sensors that are mounted on the net, depth sensors provide crucial data to quantify deformation. As sections of a net cage bend inwards (or outwards) due to the forces induced by currents, mounted depth sensors move with it, registering a change in depth as the net curves slightly upwards. The measured change in depth serves as a way to quantify the degree of cage deformation.

One of the challenges with using depth sensors is that they don't say anything about the way in which the cage is deforming: it could be deforming inwards, outwards, or anything in-between. This makes it difficult to draw a *direct* connection between depth sensors measurements and the shape of the net cage without a good model to describe their relationship. Nonetheless, monitoring and processing the vertical movement of each depth sensor is a critical stepping stone in understanding the net cage's deformation and ensuring its correct functioning.

2.3.4 Environmental conditions

As explained in the previous subsections, waves and currents can strain the mooring lines, as well as cause the entire sea cage to deform in various ways. In addition to measuring each of the effects as mentioned above, it is crucial to measure the severity of the environmental conditions themselves, that is, the "strength" of the waves and currents.[3] Furthermore, as strong winds can cause wind-induced waves (as well as directly influencing the emergent parts the sea cage, although to a lesser degree), they should also be taken into account when examining the environmental conditions. In this study, these factors were measured using a weather buoy, with the goal of giving context to the analyses of the data gathered from the other sensors.

One of the main challenges associated with the weather buoy lies in determining exactly how to measure each of the environmental conditions of interest. Due to the impracticality of measuring individual wave profiles, one will have to rely on spectral analyses and the resulting *average* wave heights and wave periods. Likewise, similar approaches must be taken when considering wind speeds. This can make it difficult to examine the effects of a single wave or a single gust of wind on accelerometer and load shackle readings. This is however not an issue when considering current speeds, which tend to vary only on larger time-scale.

[3]The notion of "strength" is specified and further elaborated in Section 3.1.

2.4 Dimensionality reduction and other approaches to optimizing sensor setups

While the sensors mentioned above are capable of measuring various important aspects of sea cage deformations, it is not at all clear how many are needed, or what the optimal placements are. These issues can be approached in a multitude of different ways.

One might for instance approach the problem by applying a dimensionality reduction technique, such as t-SNE (t-distributed Stochastic Neighbor Embedding) or a Principal Component Analysis (PCA) to name two. Unlike PCA, t-SNE is able to reduce the dimensionality of non-linear datasets, but it is also more complex, requiring very careful tuning, as the "wrongful" selection of hyperparameters has been shown to produce misleading results [27, 28]. PCA on the other hand is only able to reduce the dimensionality of linear datasets, but is much simpler and cannot produce similarly misleading results as there are no hyperparameters to tune. This makes the use of PCA as an introductory analysis more attractive than other, more complex techniques.

One might also explore completely different approaches to optimizing sensor setups, such as examining the data through the lens of a wavelet transform. However, this technique is relatively complex and requires a solid mathematical understanding of the underlying principles in order to be used. This left it somewhat impractical given the scope of this project.

After some consideration, it was decided that PCA, a well-established statistical technique for dimensionality reduction would be used in this study. It is simple and does not require any careful hyperparameter tuning, making it ideally suited for this scenario. Furthermore, it excels at clearly revealing the internal structure of data. Thus, PCA was selected as the primary tool in this study. Seeing as PCA is unable to handle noisy data, some sort of filtering will also be included as a secondary/auxiliary tool. This is further elaborated in Section 2.5.

2.4.1 Principal Component Analysis

Principal Component Analysis (PCA) is a widely used unsupervised machine learning technique for dimensionality reduction and feature extraction. Forming the basis for multivariate data analysis, it was first introduced by Pearson in 1901 [29] and later developed independently by Hotelling in 1933 [30]. The main goal of PCA is to transform a high-dimensional dataset into a lower-dimensional dataset while retaining as much of the original information as possible. Strictly speaking, it tries to create a new dataset that preserves as much variance as possible. This is particularly useful when dealing with large datasets, as reducing the number of dimensions can help improve computational efficiency and facilitate data visualization [31]. It is therefore frequently used in areas such as pattern recognition and signal processing.

PCA achieves dimensionality reduction by identifying linear combinations of the original variables, known as principal components (PCs), which capture the maximum amount of variance in the data. The first principal component (PC1) is computed as the linear combination that accounts for the largest proportion of the dataset's total variance. The next principal components are found in a similar way, with one important constraint: they must be orthogonal to the

preceding components, ensuring that the PCs are uncorrelated. In this way, a PCA generates a new coordinate system where the new axes correspond to the principal components. Each data-point from the original dataset is then projected onto this lower-dimensional plane, as visualized in Figure 2.2.

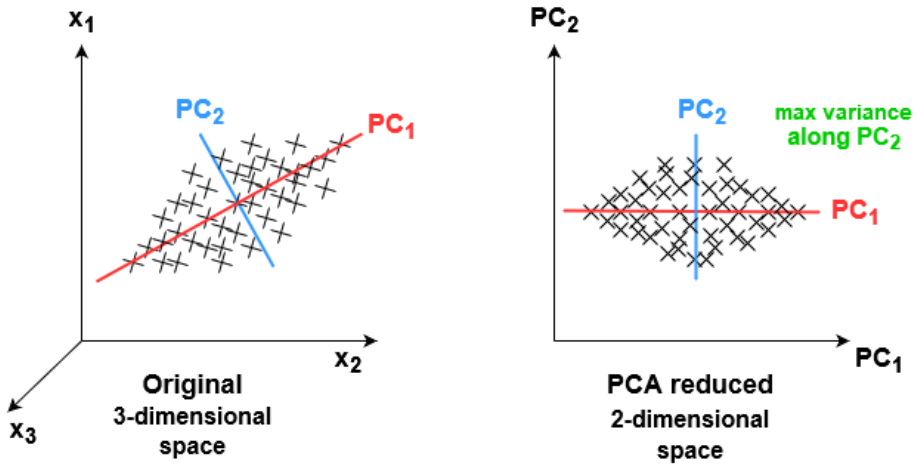


Figure 2.2: Sketch showing how PCA reduces the dimensionality of a dataset. As seen on the left side, despite existing in three dimensions, the dataset seems to lie completely on a two dimensional plane. Thus, one can represent the dataset using only two axes as seen on the right side. Figure is modified from Serafeim Loukas' Towards Data Science article.

The main advantage of PCA is its ability to reveal underlying patterns data, highlighting redundancies in the original variables. However, by focusing on the most significant components it also filters out some of the noise as a bi-product, although this is most often not what it is used for.

PCA also has some drawbacks that one should be aware of. PCA assumes linear relationships between the variables and may not be suitable for handling non-linear dependencies. Trying to use PCA on a dataset that contains highly non-linear dependencies will fail, meaning that no dimensionality reduction will be achieved. Furthermore, the method generally doesn't perform well on datasets that contain a lot of noise or outliers, as these make the dataset more non-linear.

Another drawback of the PCA is that the principal components that are returned are often far less interpretable than the original variables. This is because they represent combinations of features rather than directly measurable quantities. As such, PCA is perhaps best used not as an end in itself, but merely as a tool to guide further investigation [32].

Nonetheless, the PCA is a powerful and versatile technique for dimensionality reduction. By capturing the essential structure of high-dimensional data in a lower-dimensional space, PCA provides a means to reveal hidden patterns, reduce computational complexity, and improve the results of other machine learning algorithms. Readers interested in more intuitive explanations of the PCA are referred to the excellent turorial by Jonathan Shlens [33].

2.4.2 Mathematical workings of PCA

To perform a PCA, one first needs to calculate the covariance matrix of the dataset. The covariance matrix $\mathbf{\Sigma}$ captures the relationships between the variables, where each element σ_{ij} represents the covariance between variables i and j. For a dataset $\boldsymbol{X}$ with n samples and p features, the covariance matrix is given by:

$$\mathbf{\Sigma} = \frac{1}{n-1}\boldsymbol{X}^T\boldsymbol{X} \tag{2.1}$$

In practice, it is very common to use the centered matrix $\boldsymbol{X}^*$ where the mean of each column is subtracted from elements in that column, i.e. $x^*_{ij} = x_{ij} - \overline{x}_j$. This doesn't change the covariance matrix as it by definition is the average squared deviation from the mean. However, it does make implementations easier and provides cleaner geometric interpretations.

Furthermore, in addition to using centered matrices, it is very common to also standardize the dataset before calculating the covariance matrix. Standardization ensures that all variables are on the same scale, hindering variables with larger scales from dominating the principal components simply because they have larger scales. To standardize the dataset, one must divide each column in the centered dataset by their respective standard deviations, σ_j. Thus, one can replace $\boldsymbol{X}$ in Equation (2.1) with the standardized $\boldsymbol{Z}$ as calculated by:

$$\boldsymbol{Z}_j = \frac{\boldsymbol{X}_j}{\sigma_j} = \frac{\boldsymbol{X}^*_j - \overline{x}_j}{\sigma_j} \tag{2.2}$$

where $\boldsymbol{Z}_j$ form the columns of $\boldsymbol{Z}$. Again, this ensures that all features contribute equally to the analysis, and also makes it easier to visualize, interpret, and analyze the data. After calculating the covariance matrix of the standardized dataset, one may proceed with the rest of the PCA.

Next, the eigenvectors and corresponding eigenvalues of the covariance matrix $\mathbf{\Sigma}$ should be computed. The eigenvectors represent the principal components, that is, the directions of maximum variance in the data, while the eigenvalues act as "weights" that indicate the amount of variance explained by each eigenvector. Mathematically, this can be represented by:

$$\mathbf{\Sigma}\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \tag{2.3}$$

where $\boldsymbol{e}_i$ is the i-th eigenvector and λ_i is the corresponding eigenvalue. The larger the eigenvalue, the more variance can be explained by its corresponding eigenvector. By sorting the eigenvalues in descending order and applying the same transformation to the eigenvectors, one can nicely determine the importance of each eigenvector.

Together, the sum of all the eigenvalues represents the total explained variance in the original dataset. The proportion of variance explained by a specific principal component is given by the ratio of its eigenvalue to the total explained variance. This can be computed as:

$$\pi_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}, \tag{2.4}$$

where p is the number of variables in the dataset. Furthermore, one can now define the *cumulative* explained variance using the k most important principal components as:

$$\text{Cumulative explained variance} = \sum_{i=1}^{k} \pi_i. \tag{2.5}$$

Note that both the ratio of explained variance π_i and cumulative explained variance can be (and often are) represented as percentages by multiplying by 100%. By using only the eigenvectors associated with the largest eigenvalues, it is possible to project the data onto a lower-dimensional space, effectively reducing the dimensionality while preserving the majority of the information in the original data.

To reduce the dimensionality of the dataset $\boldsymbol{X}$, one simply needs to multiply it by the matrix $\boldsymbol{E}$ which contains the selected eigenvectors. That is,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{E} = \boldsymbol{X}[\boldsymbol{e}_1 \quad \boldsymbol{e}_2 \quad \dots \quad \boldsymbol{e}_m], \tag{2.6}$$

where the m eigenvectors explain the desired amount of variance. To decide exactly how many principal components to use, it is common to define a threshold and use however many components that are required to surpass said threshold. One might for instance require that the cumulative explained variance of the principal components be above 95%. A higher threshold means that more of the variance will be explained by the data, but it will also require more principal components. Thus there is a trade-off between cumulative explained variance and the dimensionality of the reduced sensor setup.

2.4.3 Using PCA to reconstruct data

As explained in [34], once the PCA is completed, the data $\boldsymbol{Y}$ can also be used to reconstruct the full dataset.

Given an original dataset $\mathbf{X}$, one can standardize it and perform PCA, resulting in a matrix $\mathbf{Y}$ of transformed data and a matrix $\mathbf{E}$ of eigenvectors or principal components as described in Equation (2.6).

One may then select a subset of the most important principal components (columns) from the matrix $\mathbf{E}$, denoted $\mathbf{E}_{\text{reduced}}$, along with the corresponding transformed data (columns) from the matrix $\mathbf{Y}$, denoted $\mathbf{Y}_{\text{reduced}}$.

To reconstruct the original dataset, one needs to undo the transform in Equation (2.6). This can be done by multiplying the reduced data $\mathbf{Y}_{\text{reduced}}$ by the transpose of the selected eigenvectors $\mathbf{E}_{\text{reduced}}^T$:

$$\mathbf{X}_{\text{reconstructed}} = \mathbf{Y}_{\text{reduced}}\mathbf{E}_{\text{reduced}}^{T}. \tag{2.7}$$

Keep in mind that this reconstructed dataset is still in standardized form. Thus, to obtain the original scale of the dataset, one must multiply the reconstructed dataset by the standard deviation and add the mean of the original dataset:

$$\mathbf{X}_{\text{rescaled}} = \mathbf{X}_{\text{reconstructed}} \odot \boldsymbol{\sigma} \oplus \overline{\mathbf{X}}, \tag{2.8}$$

where $\odot$ denotes element-wise multiplication, $\oplus$ denotes element-wise addition, $\boldsymbol{\sigma} = [\sigma_1 \quad \sigma_2 \quad \dots \quad \sigma_j]$ is the standard deviation of each column in the original dataset $\boldsymbol{X}$, and $\overline{\mathbf{X}} = [\overline{x}_1 \quad \overline{x}_2 \quad \dots \quad \overline{x}_j]$ is the mean of each column in the original dataset $\boldsymbol{X}$.

The resulting dataset, $\mathbf{X}_{\text{rescaled}}$, is an approximation of the original dataset, reconstructed using only the selected principal components. Note that this process involves losing information due to the reduced number of principal components used in the reconstruction.

While reconstructing the original dataset based on fewer features isn't the main purpose of PCA, it can be used to provide a more interpretable way of assessing the reduced dataset. One can reconstruct the original dataset from the reduced dimensions and then calculate the error between the reconstructed dataset and the true dataset to get a sense for how good the reduction is.

For the sake of clarity, consider this specific example. Say there are 20 depth sensors located across various points on a net cage. When one moves due to currents, others are likely to do the same. In other words, their movement is somehow correlated. You perform a PCA and find that 5 sensors can explain 95% of the variance. To check whether they really "capture" the movement of the other sensors, you reconstruct (and rescale) the data for the 20 sensors, based on the data from the 5 sensors that explain 95% of the variance. The rescaled data can then be compared to the original data to see how closely it follows the original data.

2.4.4 Importance of dataset size

There has historically been some debate as to how much data is required to gain stable[4] PCA results. However, most of these debates have lacked solid experimental grounding. There are two main schools of thought, those who think the recommended sample size N can be given as a number, and those who think that the sample size N should be given as a *ratio* to how many features (variables) are in the dataset. Both schools of thought agree that more data is better.

According to [35], few articles examine the issue comprehensively enough to be definitive. In their study, they conclude that stability is likely to be the result of an interaction between both of the schools of thought: a large sample size *and* a high samples-to-features ratio is likely

[4] In the context of this study, stability can be understood as applying a PCA to data collected from different time periods of equal length by the same set of sensors, and obtaining the same, or at least similar reduced sensor setups.

what leads to the best outcome. The view that it is an interaction between the two is backed by several articles that examine the effects of increasing either the sample size or the ratio [36, 37]. Furthermore, and crucially to coming discussions, both schools of thought also agree that too few data points can cause *unstable* results.

2.5 Filtering

Recall that wave induced deformations will be measured using accelerometers, which can often be quite noisy, and that the PCA struggles with such data. This raises the possible need for filtering to remove noise.

Several techniques exist for filtering data, each with their own advantages and disadvantages. In this study, three types of filters were mainly considered to be used, namely Butterworth, Chebyshev and Elliptic filters. Both Chebyshev and Elliptic filter have steeper roll-off characteristics than Butterworth filters, but this is not too important in the context of this thesis, as the frequency of noise is likely to be *much* higher than the frequency at which waves strike and thus deform the sea cage. Furthermore, Chebyshev filters have ripples in the pass-band and Elliptic filters have ripples in both the pass-band and the stop-band. Butterworth filters have no such unwanted ripples.

In articles such as [38] where all three filters are considered, it is concluded that "the Butterworth filter is the best compromise between attenuation and phase response." More generally, the Butterworth filter is a widely used type of signal processing filter that is known for its maximally flat frequency response in the passband [39].

By applying the Butterworth filter to the sensor data, high-frequency noise should be attenuated, while the relevant lower-frequency information should still be preserved. All things considered, Butterworth filters present a suitable approach to noise reduction prior to running PCA in certain situations in this study.

2.5.1 Description of Butterworth filters

Named after its inventor, Stephen Butterworth, this filter has found applications in various fields such as audio processing, communication systems, and control systems, where a smooth frequency response is desired.

One of the key advantages of the Butterworth filter is its ability to provide an optimal trade-off between the flatness of the passband and the rate of attenuation in the stopband. In other words, the filter has a smooth transition between the passband and stopband regions, while having a rapid roll-off rate in the stopband. This characteristic ensures that the desired frequency components of the input signal are preserved with minimal distortion, while the unwanted frequency components are effectively attenuated.

Butterworth filters can be designed as low-pass, high-pass, band-pass, or band-stop filters, depending on the desired frequency response. The order of the filter determines the steepness of the roll-off in the stopband, with higher-order filters offering a faster rate of attenuation.

The cutoff frequency defines the boundary between the passband and stopband, specifying the frequency at which the filter's gain drops to half its passband value (approximately -3 dB).

In this study, the Butterworth filter was used to preprocess the data obtained from accelerometers and load shackles. These sensors are prone to capturing noise and high-frequency fluctuations that may obscure the underlying trends in the data. The goal of using the Butterworth filter is to eliminate noise while still keeping the essential features of the sensor measurements. The Butterworth filter has been widely used in various applications, such as signal processing, data analysis, and especially medical data preprocessing, often to great success [40, 41]. The coming section will delve into the relevant mathematical details of the Butterworth filter and its implementation in the data preprocessing stage. Readers who are interesting in learning more about the filter are referred to [42].

2.5.2 Mathematical workings and Python implementation of Butterworth filters

The Butterworth filter is a type of IIR (Infinite Impulse Response) filter, meaning that its impulse response continues indefinitely, approaching, but never quite reaching 0. The filter is given by its order and cutoff frequency. The cutoff frequency, ω_c, defines where the filter transitions from pass-band to stop-band, while the order of the filter, n, determines how aggressively the filter attenuates signals that are in the stop-band. The Butterworth filter's frequency response, $H(\omega)$, can be expressed as:

$$H(\omega) = \frac{1}{\sqrt{1 + (\frac{\omega}{\omega_c})^{2n}}} \tag{2.9}$$

For a more intuitive understanding, Figure 2.3 shows visually what the filter's response looks like.

To apply the Butterworth filter to the sensor data, it must be implemented as a discrete-time system. The continuous-time filter is first designed by specifying the order n and cutoff frequency ω_c, and then transformed into a discrete-time filter. In Python, this transformation was implemented using the `butter()` function from `scipy.signal`. Once run (with the correct arguments), it returns the filter coefficients a and b that define the filter's behavior in the discrete domain. Once these coefficients are obtained, they can be used to filter the sensor data. In the Python implementation, this was done using the `filtfilt()` function, also from `scipy.signal`.

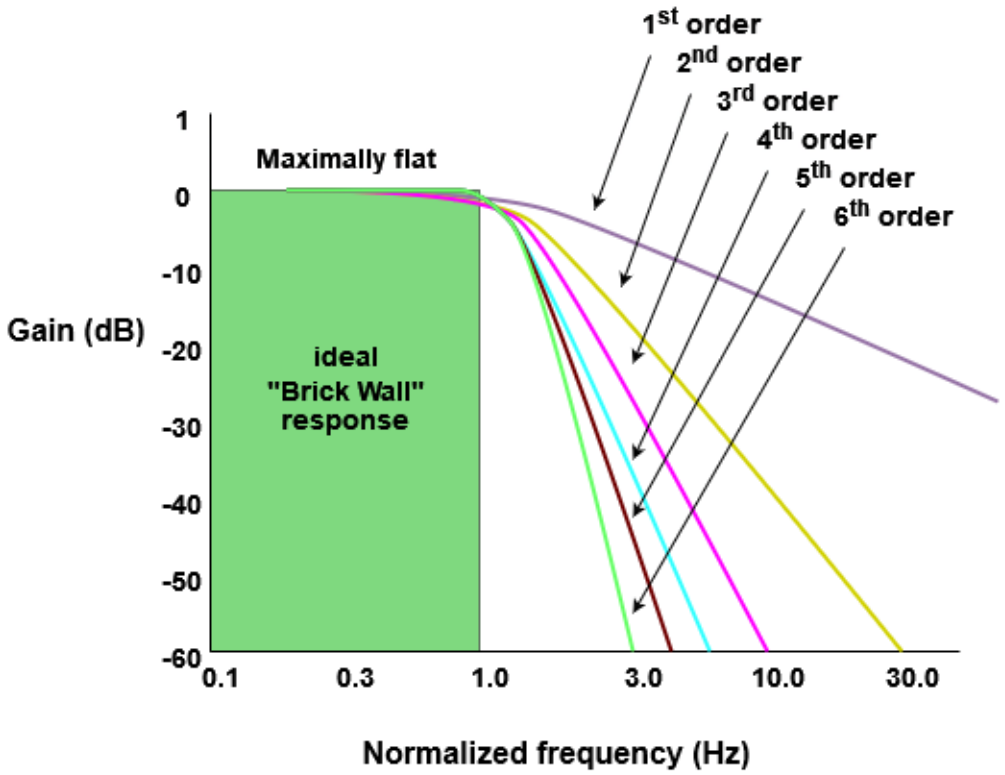


Figure 2.3: Diagram showing the frequency response of a Butterworth filter around the transition band. Normalized frequency simply indicates division by the Nyquist frequency[5] of the signal first. Figure is modified from AnalogueDialogue.

Chapter 3

Methods and Experimental Setup

3.1 Sensor setup and data collection

The sea cage used to collect data is located at Buholmen and is equipped with accelerometers, load shackles, and depth-sensors. Together, these sensors provide a good basis to analyze structural deformations and the effect of waves. This section begins by providing an overview of Buholmen fish farm, before presenting the experimental setup of each of the sensors, along with how their data is recorded. The same is done for the weather buoy.

3.1.1 Buholmen fish farm

All the data used in this project was gathered from Buholmen Fish Farm. SINTEF Ocean has quite a few sites dedicated to conducting research and collecting data. While the three main sites are Rataren, Tristeinen and Korsneset, they also have access to farm sites that are less frequently used for research. One of these is Buholmen fish farm, located off the coast of Åfjord as shown in Figure 3.1. It has been operational since mid 2013 but is in the process of being shut down at the time of writing (mid 2023) due to problems regarding repeated disease outbreaks.

Being located in a fairly exposed area of the coast means that Buholmen experiences widely varying weather conditions, making it a good candidate for data gathering. Wind speeds typically lie in the range of 3-15 m/s, although gusts can reach speeds of 25 m/s or more, while current speeds typically stay below 50 cm/s. Temperatures vary from 5°C in months of January-March to 15°C in the months of July-September.

In total, Buholmen fish farm consists of 10 sea cages, all of which are owned and operated by Salmar. However, one of these cages contains fish that are used by SINTEF Ocean for research. This sea cage has a diameter of 50m and depth of 30m, and is only used for raising Atlantic salmon.

Figure 3.1: Map showing the locations of different SINTEF fish farms. Buholmen can be seen in the middle, towards the top. Trondheim is located at the Sealab pin (for reference). Photo courtesy of SINTEF ACE.

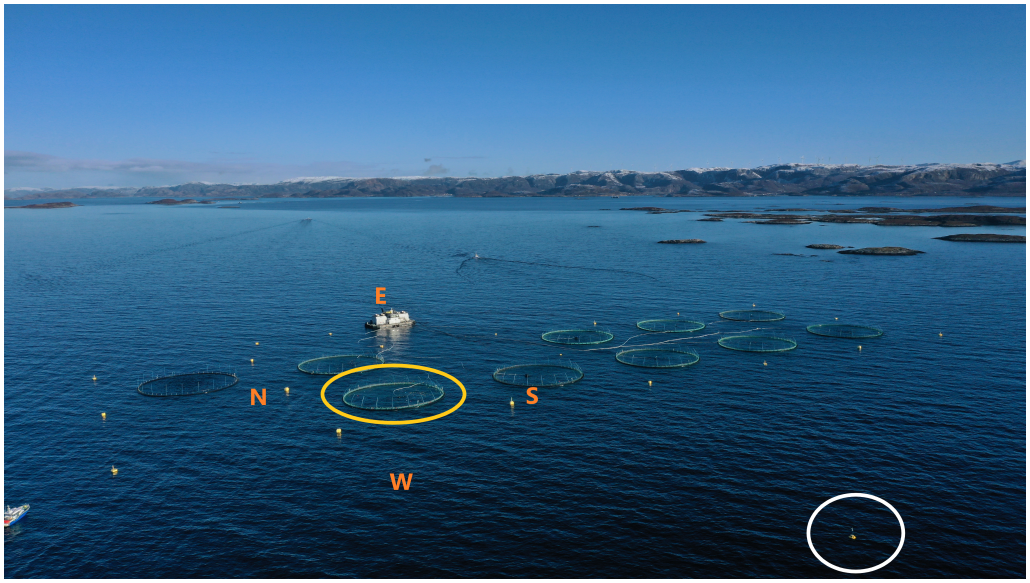


Figure 3.2: Image showing Buholmen fish farm. Orange circle shows the sea cage that is used for research by SINTEF Ocean. The white circle shows the location of the weather buoy that collects weather and environmental data. North, east, west and south included for reference. Photo courtesy of SINTEF ACE.

3.1.2 Accelerometers

In this project, a total of 8 G-Link-200-OEM accelerometers (produced by LORD, MicroStrain Sensing Systems) were placed in a circular configuration along the floating collar, above the

water surface, as shown in Figure 3.4 and Figure 3.3. Each accelerometer records and saves x, y, and z-acceleration at a frequency of 8Hz and has done so from November 2019 to April 2020. These sensors were stored individually, but synchronized to provide the same time-stamp across all devices. Although there are few errors in this time period, notable outages include sensor 4 not recording any data from early January to mid February.

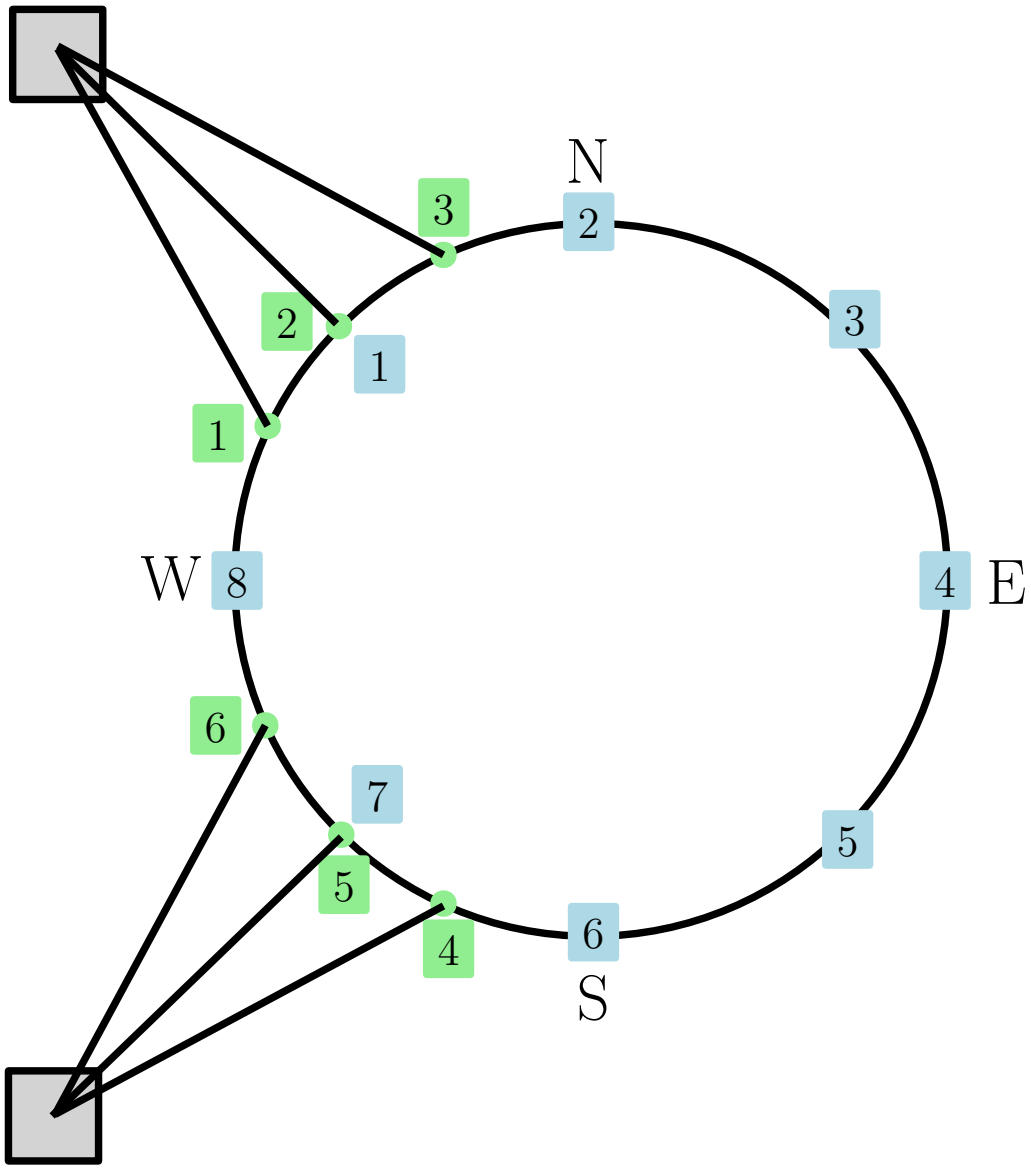


Figure 3.3: Sketch showing the sea cage from above. The placements of the load shackles are shown in green and accelerometers in blue.

3.1.3 Load shackles

In this project, 6 "type 4991" load shackles (produced by James Fisher Strainstall) were used to measure the tension in the mooring system. These were placed along the western side of the sea cage as seen in Figure 3.3. This is the side that faces the ocean and is most likely to be hit by waves. Furthermore, there are no other sea cages that obstruct the waves coming from this direction, as seen in Figure 3.2.

The load shackles record the force they experience at a frequency of 4Hz and were synchronized to do so at the same timestamps. However, due to some technical error, they only saved data in sporadically occurring two-hour intervals throughout each day. That is to say, during one day, the shackles saved data from 01:09 to 03:09, 03:09 to 05:09, 07:09 to 09:09, 11:09 to 13:09 and 21:09 to 23:09, while another day, they only saved data from 09:09 to 11:09 and 15:09 to 19:09. The load shackles collected data like this from November 2019 to March 2020. On top of this, shackle number 1 broke down in early January 2020, leaving only shackles 2-6.

Finally, the load shackles only save the voltages they measure. A linear calibration equation is given for each load shackle. These need to be applied to every measurement to yield measured force in tons.

3.1.4 Depth sensors

To measure net cage deformations, 16 milli-F Data Storage Tags (DST) (produced by Star-Oddi) were mounted on the net cage at various depths. The milli-F DST is a small cylindrical logger that is often used for monitoring the movements of fish and other marine animals. It's a high-precision depth and temperature data logger, with a depth accuracy of ±0.4% of the selected measurement range and a temperature accuracy of ±0.1°C.

The 16 depth-sensors were attached to the net cage in two circular configurations at 7m and 15m, as shown in Figure 3.4. Instead of being located along the upper circular plane, sensor number 2 is located at the bottom. Besides this exception, all the other depth sensors are spread apart by 45°.

These sensors only record and save depth (and temperature) every 4 minutes, but this is not an issue as water currents don't tend to change much during such short time periods.[1] Unfortunately, no data is available from sensor number 15 due to technical difficulties. However, the other sensors diligently recorded data every 4 minutes from mid. December 2019 to early April 2020. Like the accelerometers and load shackles, the depth sensors were also synchronized so as to save data at the same timestamps. For some unknown reason, sensors number 5 and 12 were offset by two and one minutes respectively.

3.1.5 Weather buoy

The weather buoy is located right next to the facility, as seen in Figure 3.2, and is made up of a Wavesense sensor (produced by Fugro OCEANOR) that is highly programmable. It meas-

[1] There are several complex factors affecting how water currents change, but these typically cause changes on the timescale of hours and days (or more). [43]

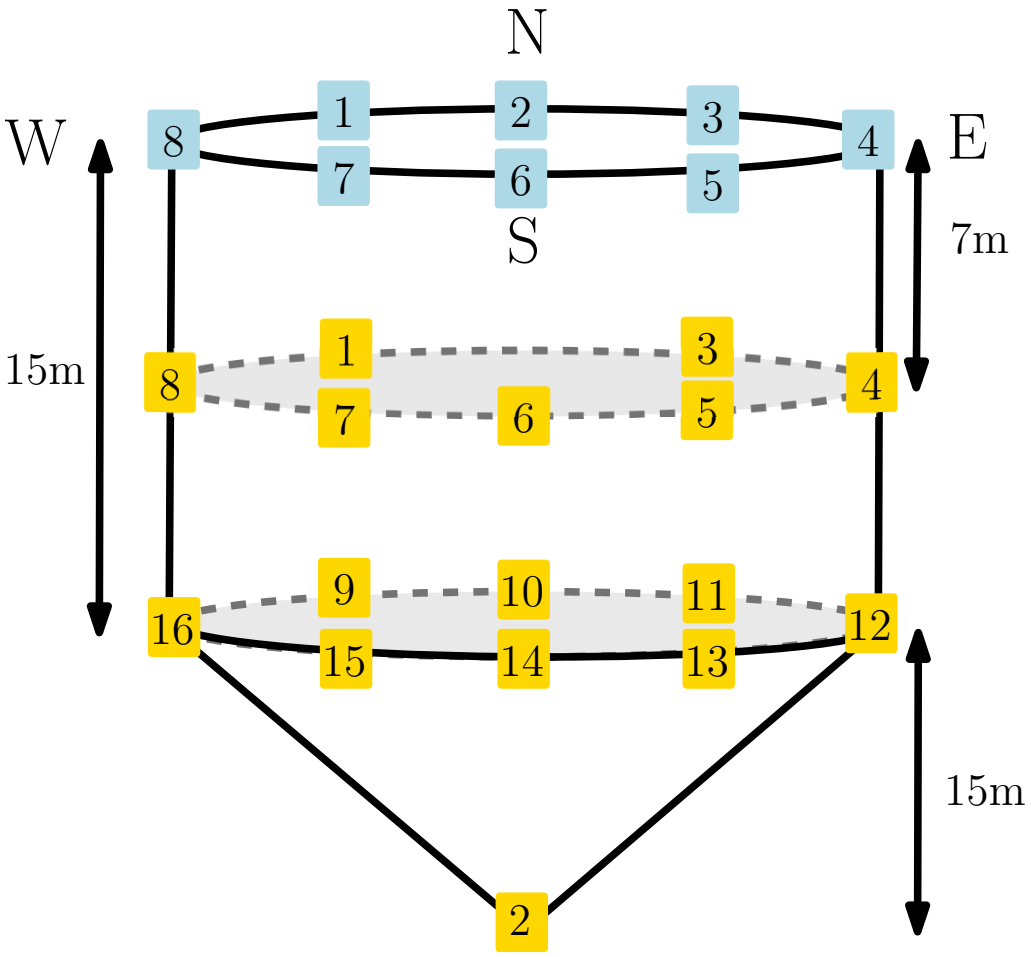


Figure 3.4: Figure showing the placements of the depth sensors in yellow and the accelerometers in blue. Geographic north, south, east and west are included for reference.

ures a range of environmental conditions, some of which are crucial to understanding how phenomena that affect sea cages arise. These are given below.

Amongst other things, the weather buoy measures current speed and direction at 3-meter intervals down to 60 meters. Without this data, the analysis of depth sensor measurements would be have to be conducted without knowing whether there were weak or strong currents present. This would in turn limit our ability to examine whether there is a connection between environmental conditions and optimal sensor setups.

Additionally, the buoy is equipped with sensors to measure wave height and period, offering crucial information about wave-induced movements and loads. Again, this provides context to the analysis of data from the accelerometers and load shackles. Lastly, the weather buoy also measures wind direction and speed, although only the wind speed is used in this study. These factors significantly influence wave formation and can impact the movement/deformation of the floating collar.

It is worth mentioning that the weather buoy also measures a multitude of other conditions, such as O_2 concentration, salinity, air temperature, and water temperature at different depths. These parameters are vital in understanding the environmental conditions within the fish farm and how they might affect fish health and growth. However, these parameters are far less important to our analysis and will therefore not be used in this thesis.

All the parameters measured by the weather buoy are saved every whole hour. In order to do a spectral analysis of the waves, the buoy at Buholmen collects data points at a frequency of 4Hz over 40 minutes - 20 minutes before each whole hour and 20 minutes after - before performing calculations and then saving the data. While the weather buoy's data is complete (no missing values at any point), it is only available from mid January 2020 to early March 2020.

In this study, the data from the weather buoy was used to select time-periods where the mentioned conditions span the entire range of low-high values. It is within these time-periods that data from accelerometers, load shackles, and depth sensors shall be analyzed. This approach enables one to examine if there is a connection between the optimal sensor setup and the harshness of the environmental conditions in a given time period.

3.1.6 Sensor overview

All the variation in when and how each sensor collects data can be hard to keep track of. The table that follows is meant to act as a summary, making it easier to compare how each sensor collects and stores data.

Sensor	Data collection period	Sampling rate	Errors	Saving format
8 Accelerometers	Dec. 2019 - Apr. 2020	8 Hz	Sensor 4 down in January.	Binary
6 Load shackles	Nov. 2019 - Mar. 2020	4 Hz	Sensor 1 down from January. Records data sporadically in 2-hour intervals.	.Txt
16 Depth sensors	Dec. 2020 - Apr. 2020	Every 4 min	Sensor 15 down.	.Dat
Weather buoy	Jan. 2020 - Mar. 2020	Every 1 hour		.Txt

Table 3.1: Table showing the data collection period, sampling rate, errors, and saving format of each sensor.

3.2 Selecting time periods (cases) for analysis

To conduct a focused analysis, ten specific time periods were chosen by examining the data from the weather buoy. Altogether, these ten time periods cover the different environmental conditions observed at the fish farm, specifically in terms of current speed, wave amplitude and period, and wind speed. This selection process helped reduce the dataset size, while also

allowing the investigation of whether certain sensors only were necessary under specific conditions. To capture the full spectrum of conditions experienced at the fish farm, the following criteria were considered when selecting the time periods:

- Current speed: time periods with low, moderate, and high current strength at 7m, 16m, and 31m depth were selected in an attempt to understand how different water flow conditions might affect the fish farm structures and sensor measurements. The current strength was generally in the range of 0-40 cm/s in the time period where the weather buoy recorded data. This data is shown in Figure 3.5.
- Wave height and period: time periods with different wave heights and periods were selected to examine the impact of different wave conditions on the sea cage movements and mooring system. The *significant wave height*[2] was generally in the range of 0-2m while the wave period was in the range of 3-6s. This data is shown in Figure 3.6.
- Wind speed: time periods with calm, moderate, and strong winds were included to evaluate the effects of wind-induced surface waves and currents on the fish farm. These might affect load shackle readings and potentially the current strength at shallow depths, which in turn affects depth sensor reading. The wind strength generally stayed in the range of 0-16m/s.

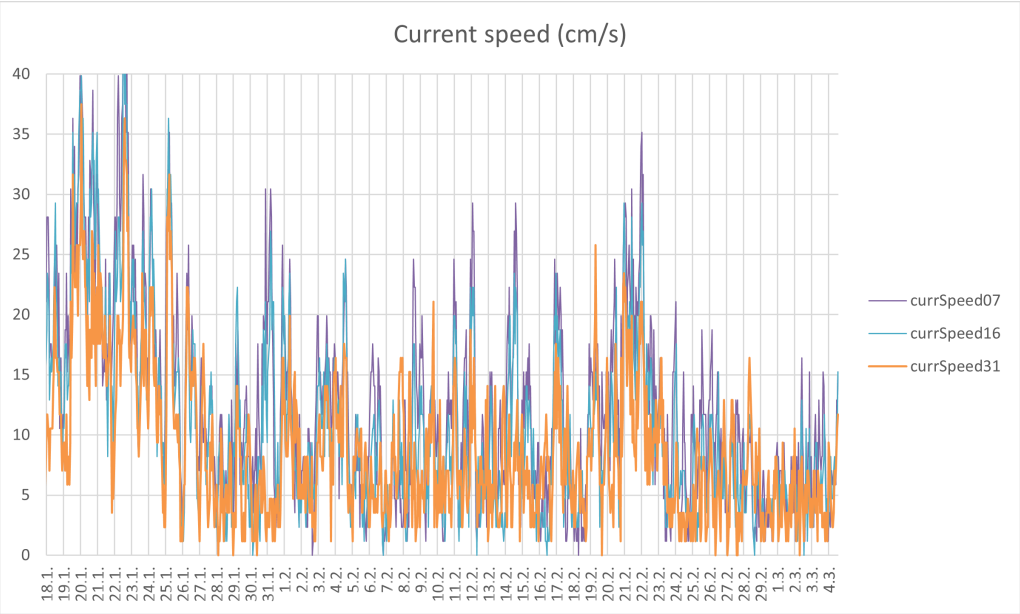


Figure 3.5: Graph showing the current speeds measured by the weather buoy at a depth of 7m (in purple), 16m (in blue) and 31m (orange).

In addition to analyzing each factor individually, time periods with various combinations of these environmental parameters were included, enabling the exploration of their combined

[2]The significant wave height is defined as the average height of the highest one-third of waves. This can also be calculated in other ways. This study uses H_{m0} as an alternate way of calculating the significant wave height. Mathematically, H_{m0} is the square root of the zeroth moment of the wave spectrum.

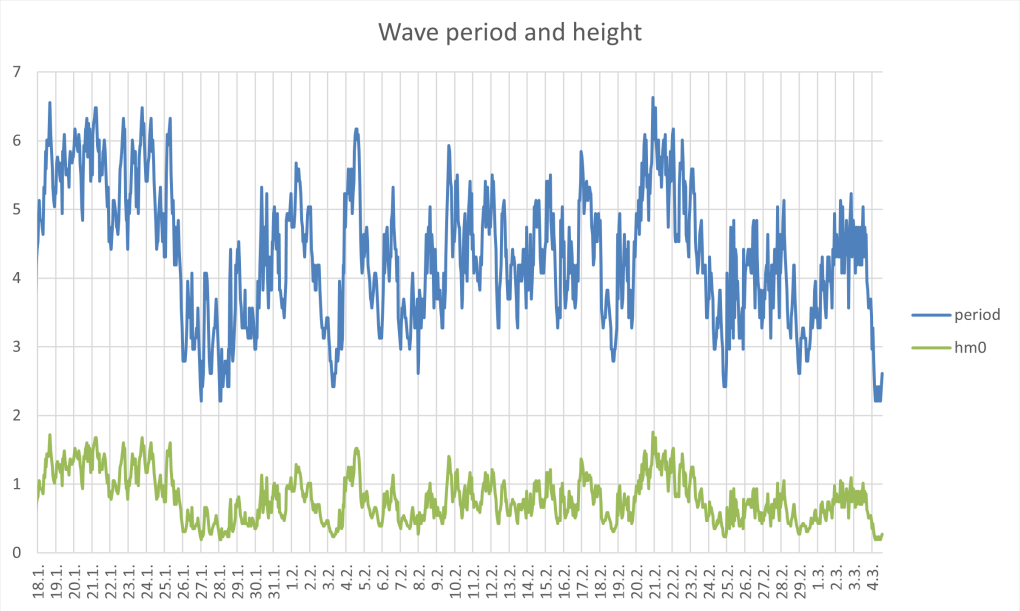


Figure 3.6: Graph showing the wave period (blue, in seconds) and significant wave height (green, in meters) measured by the weather buoy.

influence on the fish farm structures and sensor readings. This however turned out to be slightly difficult as those periods with strong winds also tended to have strong currents and big waves, and vice versa.

The time periods that were chosen for analysis are shown in Table 3.2. Each of the cases span a 1.5 hour time window. This was done for two main reasons.

Firstly, recall that the weather buoy gathers data in 40 minute windows around a given whole hour. This means that the data it records is most representative of the weather conditions within that time frame. When a "time period" entry reads "09:15 - 10:45", the rest of the values in the row indicate that the data was collected by the weather buoy from 09:40 to 10:20, but saved with the timestamp "10:00".

Secondly, the load shackles only record data in sporadic 2-hour intervals. This places an upper limit on the time span of each of the cases. Considering these two pieces of information, it seemed sensible to only analyze data from 1.5 hour windows, as listed in Table 3.2. Throughout Chapter 4, when referring to "case X" in the various graphs, the reader should understand it as the time periods given in Table 3.2.

As seen in Table 3.2, the selected cases cover a range of stormy, moderate and calm weather conditions. Some cases also have a mix of high values in one column and low values in other columns. This was done in an attempt to cover most of the wide range of weather conditions that can be experienced by exposed fish farms.

All the selected time periods fall within January 2020 - early March 2020 as this was the time within which most sensors were recording data. More importantly, it is the time period within

Case no.	Time period in 2020	Wave period (3-6s)	Wave height H_{m0} (0-2m)	Current speed (0-40 cm/s)			Wind speed (0-15m/s)
				7m	16m	31m	
1	18.01 09:15 - 10:45	3.12	0.39	9.38	4.69	3.52	4.58
2	19.01 11:15 - 12:45	5.13	1.06	9.38	8.20	8.20	10.53
3	20.01 15:15 - 16:45	6.17	1.52	39.84	39.84	37.5	16.68
4	21.01 07:15 - 08:45	6.33	1.60	38.67	35.16	25.78	13.47
5	23.01 03:15 - 04:45	5.32	1.13	42.19	41.02	19.92	14.08
6	25.01 17:15 - 18:45	5.5	1.21	29.3	30.47	28.13	13.47
7	26.01 19:15 - 20:45	2.96	0.35	16.41	15.23	12.89	3.21
8	28.01 15:15 - 16:45	2.96	0.35	2.34	3.52	4.69	3.01
9	21.02 17:15 - 18:45	5.93	1.41	21.09	16.41	11.72	16.13
10	01.03 13:15 - 14:45	3.28	0.43	3.52	1.17	1.17	8.41

Table 3.2: Table showing the weather pattern in 10 different time periods, as given by the weather buoy. Data from the sensors will be analyzed within these time periods. Numbers in parenthesis indicate the range within which *most* values tend to lie, as well as units of measurement. Orange color scheme is used to convey the severity of values, with lighter shaded squares indicating a (relatively) low value and darker squares indicating a (relatively) high value. Darker shaded rows taken as a whole point to stormy weather while lighter shaded rows point to calm weather.

which the weather buoy was recording weather conditions. Furthermore, most of the selected time periods lie in late January. This is because the weather buoy's data was initially only made available for late January - early February. One additional month of data (February - early March) was made available some time after the time periods for analysis had already been selected. However, this is not an issue as the additional month doesn't contain any novel weather patterns that don't appear in late January - early February.

After identifying the ten cases of interest, the corresponding data was extracted from the accelerometers, load shackles, and depth sensors. This targeted approach enabled more efficient data analysis and the possibility of examining the interplay between environmental conditions and sensor measurements. The results from such an approach could provide valuable insights into the optimal deployment and utilization of sensors under different conditions, ultimately leading to improved monitoring strategies for fish farms.

3.3 Preprocessing

In this thesis, data from accelerometers, load shackles, and depth sensors all had to be preprocessed to make sure they were in a format suited for subsequent analysis. This section discusses the specific steps taken in preprocessing each type of sensor.

3.3.1 Accelerometer data

For the accelerometer data, the raw binary files were first imported into a program called SensorConnect. This allowed for the visualization of the data as well as the option to export it to CSV format. Only the z-components (vertical motion) of each sensor were extracted (horizontal effects of waves are captured by the load shackles) within the relevant time periods and exported to CSV files.

Sensor number 4 hadn't recorded any data during most of the exported time periods and was therefore removed completely from the dataset, to ensure consistency. Any rows containing NaN values were eliminated, and the numbers were rounded down from 15 decimal places to 4. The dataset was then saved as a new CSV file with timestamps. Preprocessed data from case 1 is plotted in Figure 3.7.

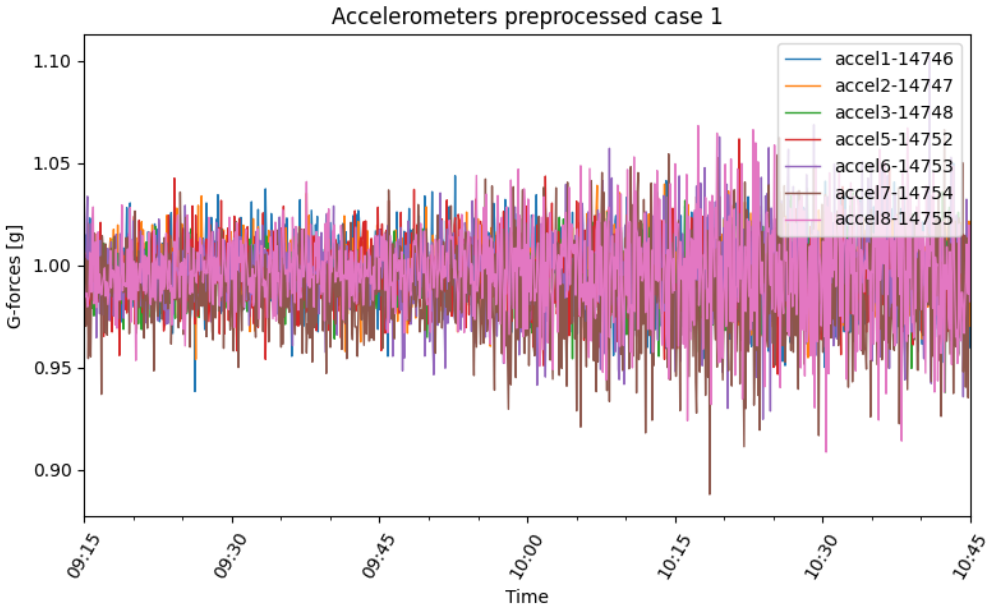


Figure 3.7: Plot showing the preprocessed accelerometer data from the first time period listed in Table 3.2. Note that only every 40th data point (every 5 seconds) is used in this plot for visualization purposes.

3.3.2 Load shackle data

The load shackles required a different preprocessing approach. Each load shackle came with its own set of linear calibration equations, that is, equations of the form "a*x + b", where "a" and "b" are decimal numbers and "x" is the voltage measured by the load shackle. Only after applying the corresponding equation to the corresponding sensor in the corresponding time period would one obtain the force experienced in tons.

As the load shackles recorded data in sporadic 2-hour intervals, the appropriate data files had to be located first. This was done manually by finding the 10 files that corresponded to the datetimes given in Table 3.2. Then, a script was written to apply the respective calibration equations to each load shackle during each time interval.

Sensor number 1 (or ch0 in subsequent plots) was found to be malfunctioning during most of the time periods and was therefore removed completely from the dataset. Furthermore, all the recorded data had timestamps that ended in 0.095s, 0.345s, 0.595s and 0.845s. As this might have troublesome for later cross-sensor analyses, all the timestamps were offset by 0.095s to make sure they ended in 0.0s, 0.25s, 0.50s and 0.75s. Finally, the calibrated data was rounded down from 10 decimal places to 4 and then saved to a CSV file with timestamps. Preprocessed data from case 1 is plotted in Figure 3.8.

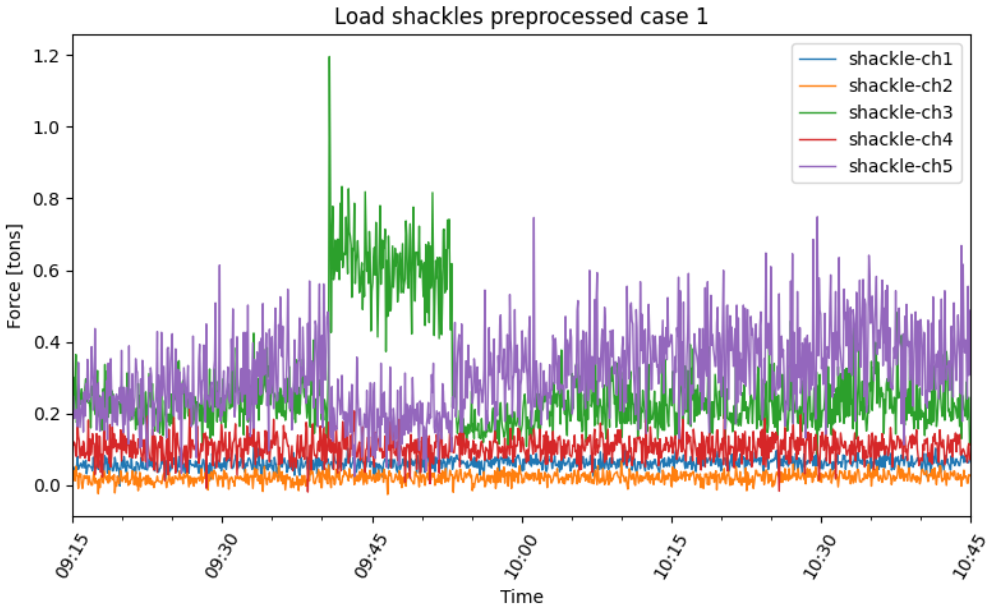


Figure 3.8: Plot showing the preprocessed load shackle data from the first time period listed inTable 3.2. Note that only every 20th data point (every 5 seconds) is used in this plot for visualization purposes.

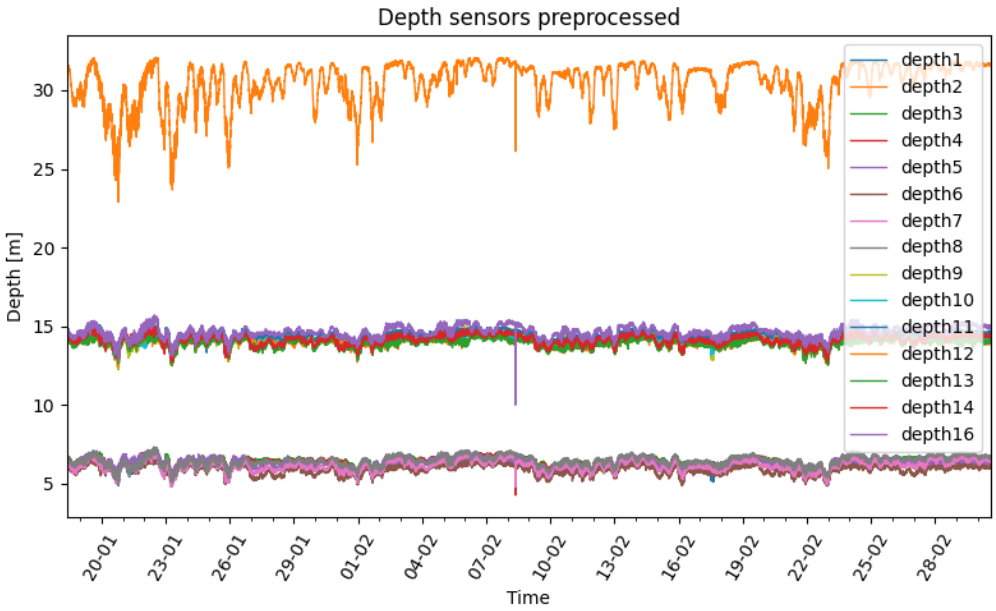


Figure 3.9: Plot showing the preprocessed depth sensor data. As opposed to the two previous plots, this plot shows all the data from the first date to the last date listed in Table 3.2. Furthermore, every data point is used in this visualization.

3.3.3 Depth sensor data

Lastly, each depth sensor had its data stored in separate .Dat files. These contained some metadata as well as recorded depths (with 2 decimal places precision) and water temperature. Recall from Section 3.1 that data from sensor number 15 was missing, meaning that it had no .Dat file and was dropped from this stage.

First, the .Dat files were stripped so as to only include depth measurements and timestamps. Then, all the data from the sensors was concatenated and saved as one text file. In this process, it became apparent that sensors no. 5 and 12 were offset by two and one minutes, respectively, from the rest of the sensors. To align these sensors with the others, the offsets were corrected *before* saving all the depth sensor measurements to a single text file. Rows containing NaN values, which were located at the very beginning of December or end of April, were also removed. Preprocessed depth sensor data is plotted in Figure 3.9.

3.3.4 Summary of preprocessing

The data preprocessing steps for each sensor type involved data extraction, cleaning, and calibration to ensure consistency and accuracy in the analysis. By following the steps outlined in this section, the data was transformed into a format suitable for further analysis in the context of this thesis. The steps taken to preprocess the data from the different sensors is neatly summarized in Table 3.3 for future reference. After preprocessing, three CSV files remained, each containing timestamped measurements from accelerometers, load shackles, or depth sensors.

Sensor	Sensors dropped	Preprocessing steps
8 Accelerometers	Sensor no. 4	Z-component exported from binary to CSV files using SensorConnect. Numbers rounded from 15 decimal places to 4. CSV files concatenated into one CSV file.
6 Load shackles	Sensor no. 1	Calibration eqs. applied to each sensor in each time period. Numbers rounded from 6 decimal places to 4. 0.095s subtracted from each timestamp to ensure nice timestamps.
16 Depth sensor	Sensor no. 15	Metadata stripped from .Dat files. Sensor no. 5 offset by 2 min to match other timestamps. Sensor no. 12 offset by 1 min to match other timestamps. Data joined on timestamps and saved into one CSV file.

Table 3.3: Table showing a summary of the preprocessing steps applied to each of the sensors.

No preprocessing was needed on the data from the weather buoy, as this was simply used to select time periods within which the data from the other sensors would be analyzed. The data from the weather buoy itself would not be used in any analysis.

3.4 Principal Component Analysis

PCA was implemented in Python using the `sklearn` library. Before running PCA, the relevant dataset (data from one of the sensor types) was always standardized first, so that each variable was centered and scaled to have a mean of 0 and a standard deviation of 1. In each of the coming analyses, the PCA would identify the fewest number of sensors that were required to explain 95% of the given dataset's variance. These sensors, which will also be loosely referred to as "most important sensors" or "optimal sensor setup", would then be returned in a list by the Python function.

The rest of this section outlines the different approaches and methods that were attempted used in combination with the PCA. It gives a description of the various ways in which the PCA was applied to the different datasets obtained from depth sensors, accelerometers, and load shackles. This section also discusses the reconstruction of the full dataset based on a reduced dataset, although this was only done for the depth sensors.

3.4.1 Principal Component Analysis on individual datasets

Initially, PCA was performed separately on each type of sensor data from each of the 10 cases. This meant running the PCA separately on measurements from depth sensors, accelerometers, and load shackles. This approach aimed to identify whether dimensionality reduction could be achieved for each sensor type independently. Seeing as the depth sensors had continuous measurements available in the entire time period between case 1 and case 10, it was decided that a PCA would also be run on all the depth sensor data collected during this month-and-a-half long time period.

3.4.2 Aggregated analysis of depth sensor data

To further verify the results of the initial analysis and assess the consistency of the reduced sensor setup, a more detailed analysis was conducted. The entire depth sensor dataset (mid. December to early April) was divided into segments of equal lengths (e.g. fourteen one-week segments). A PCA was then carried out on each segment independently, identifying and returning the most important sensors.

A tally was kept of how many times each sensor was included in the list of most important sensors. Then, a bar chart was created to visualize this data. This approach made it possible to assess the stability of the "optimal" sensor setups that the PCA would return, by looking at the heights of the bars in the bar chart.

Several segment lengths were examined. For starters, the initial analysis (that was conducted on data from the 10 cases listed in Table 3.2) is completely equivalent to conducting a PCA on ten 1.5 hour segments. This however doesn't utilize the entire dataset. Thus, after trying fourteen one-week segments (from 21st of December to 28th of March), the segment-length was gradually increased from one day to four weeks, one day at a time.

As one might imagine, there is a trade-off between segment-length and total number of segments. With shorter segments, e.g. one-day segments, there will be roughly 100 separate segments in total (3 months or $\sim$100 days), but each sensor will only contain 360 data points in each segment, as the sampling rate is 4 minutes. On the other hand, with longer segments, e.g. 14-day segments, each sensor will contain 5040 data points, but there will only be 7 separate segments to run PCA on.

An additional layer of flexibility was introduced by varying the degree to which the analysis window was shifted each time. For the sake of clarity, consider the following example. When using segments of length 14 days, the "default" way would be to run the PCA on weeks 1-2, then weeks 3-4, then 5-6, then 7-8, then 9-10, and then 11-12, and finally weeks 13-14. However, one could also run the PCA on weeks 1-2, then weeks 2-3, then 3-4, all the way up to weeks 12-13 and finally 13-14. In the first example, the window of analysis is shifted by the same amount as the segment length (14 days) each time, while in the latter example, the window of analysis is only shifted by half the segment length (7 days) each time. While the first example yields 7 separate segments to run PCA on, the latter example yields 13 *overlapping* segments to run PCA on.

For a clearer presentation and discussion of the results, *segment length* shall be referred to as L_{seg} (in days), *window shift* as L_{shift} (in days), and the resulting *no. of segments* for analysis as N_{seg}. With these variables in place, the first example given above would be given by $L_{seg} = 14$, $L_{shift} = 14$, $N_{seg} = 7$, while the second example would be given by $L_{seg} = 14$, $L_{shift} = 7$, $N_{seg} = 13$.

3.4.3 Butterworth filtering and PCA

Here, a Butterworth filter was applied to both accelerometer and load shackle data before running a PCA in an attempt to remove some of the noise and improve the PCA results. The

Butterworth filter was tried with several cutoff frequencies ω_c and an order of $n = 4$. However, different cutoff frequencies presented challenges; low frequencies effectively wiped the data, while high frequencies did not result in any significant dimensionality reduction. This is reiterated and further discussed in Chapter 4.

In the end, a cutoff frequency of 0.8Hz was chosen as it is slightly higher than the highest wave frequency, which is around 0.5Hz (recall that *most* waves hit the fish farm with a period of 3-6s). This ensures that the filter preserves the important frequency components of the signal while eliminating higher frequency noise that could negatively impact the PCA results.

3.4.4 Rolling window averages and PCA

An alternative noise-reduction approach involved applying rolling window averages to the accelerometer and load shackle data before running a PCA. This method aimed to reduce noise by smoothing out the signals over time.

The rolling average, or rolling window average technique involves replacing each data point with the average of all values within a defined "window" of size n, starting from the data point in focus. This technique has the potential to smooth out short-term fluctuations and highlight longer-term trends in the data, potentially improving the results of the PCA.

Different window lengths were tried, but as with the Butterworth filter, there was one main challenge; short windows preserved patterns but did not lead to dimensionality reduction, while longer windows "smeared" the data in time, effectively destroying it. Again, this is further discussed in Chapter 4.

3.4.5 Combining datasets

Accelerometers and load shackles

In an attempt to explore potential relationships between the datasets, load shackle and accelerometer data was combined into a single dataframe before running a PCA. This made intuitive sense as waves have an effect on both of the measurements, pointing to a potential link between the two datasets. Thus, this was and attempt investigate whether the combined data could lead to better dimensionality reduction or reveal any hidden patterns across the datasets.

To perform this joint analysis, the accelerometer and load shackle data was combined using an intersection join, which only kept every other accelerometer measurement to match the lower sampling rate of the load shackles. This ensured that the combined dataset had an equal number of measurements from each type of sensor. As before, a PCA was run on each of the cases given in Table 3.2.

All sensors combined

Additionally, all datasets, including depth sensors, accelerometers, and load shackles, were combined into one large dataframe to investigate the possibility of joint dimensionality reduc-

tion. This makes less intuitive sense, but was easy to implement and couldn't do any harm (as will be explained in Chapter 5). This approach was mainly meant to catch any hidden relationships that might have been present between the datasets, and analyze their impact on the total explained variance.

The disparate sampling rates of the different sensors posed a challenge. The depth sensors sample once every four minutes, which is significantly slower than the 8Hz and 4Hz sampling rates of the accelerometers and load shackles, respectively. Nonetheless, to create a combined dataframe, an intersection join was used. This effectively downsampled all data to the sampling rate of the depth sensors.

3.4.6 Reconstructing dataset from a subset of sensors

When successful, the PCA returns a reduced sensor setup. By starting with a reduced sensor setup and essentially running the PCA in reverse, it is possible to recreate data for the rest of the sensors not included in the reduced setup. This was only attempted with depth sensor data.

The reconstruction process was performed in two steps. First, the dataset was split into two, using the time period between all the days in Table 3.2 as a training set and a few hours or days thereafter as a test set. Then, PCA was applied to the training set to obtain the principal components corresponding to the sensors that explained 95% of the cumulative variance. Next, (data from) these sensors were extracted from the test set and the reconstruction process was applied to this part that the PCA had not seen before. This was done to validate the reconstruction method on unseen data.

The whole process as it was implemented in Python can be described as follows:

1. Split the entire depth sensor dataset into training and test sets. In this thesis, the time period between 10:00 on the 18th of January and 14:00 on the 1st of March was used as the training set. A given period of time starting immediately after the training set was used as the test set.
2. Standardize both datasets by subtracting the mean and dividing by the standard deviation.
3. Perform PCA on the standardized training set. Use the results to find out which sensors are able to explain 95% of the training set's variance, obtaining principal components in the process.
4. Filter the test set to only include measurements from the sensors found in the previous step. Perform PCA on this reduced standardized test set.
5. Reconstruct the sensors that were filtered away (in the test set, in the previous step) using the principal components that were found in step 3. This is done by multiplying the PCA-transformed test set ($\boldsymbol{Y}_{test}$) by the transpose of the principal components ($\boldsymbol{E}_{training}$) of the training set (found in step 3). This is in essence the same as computing Equation (2.7), but using $\boldsymbol{Y}_{reduced}$ from the test set and $\boldsymbol{E}_{reduced}$ from the training set.
6. Rescale the test data using the training set's mean and standard deviations, as in Equation (2.8)

7. (Optional to increase accuracy) Overwrite the estimations of those sensors that were kept in step 4. In step 4, one essentially pretends to have less than 15 sensors, wanting to predict the values of the rest. Since the measurements of the sensors that are *kept* are known, one doesn't need to estimate these through the reconstruction process. Thus, one might as well overwrite whatever is estimated, using the values from the original measurements. This step *is* executed in the context of this study.

By following this process, the full test-dataset was reconstructed using only the sensors that explained 95% of the cumulative variance in the training-dataset. This helped intuitively validate the effectiveness of PCA for dimensionality reduction.

Chapter 4

Results

In this chapter, the results of the various analyses are presented. What follows below is a quick recap of all the analyses done, in order of appearance in this chapter.

Initially, PCA was run separately on accelerometers, load shackles, and depth sensors, using data from each of the time periods given by the 10 cases in Table 3.2. This provided a great starting point, and branched the coming analyses into two: further analysis of depth sensor data and filtering of accelerometer and load shackle data.

Following the initial analyses, a PCA was run on depth sensor data from a 1.5-month time period, in order to further verify the dimensionality reduction capabilities found in the initial analysis. One drawback with this new approach was that there was no way to assess the stability of the returned optimal sensor setup. Only running the analysis once made it somewhat difficult to trust that the sensors found to be important *really were* important. Thus, another experiment - the aggregated analysis - was conducted to bring more "certainty" or "credibility" to the reduced sensor setups provided by the PCA.

Next, the reader is presented with the results found when first applying a Butterworth filter before running a PCA. This was only done with accelerometer and load shackle data. Afterwards, another attempt was made where a rolling average was applied (instead of the Butterworth filter) to accelerometer data and load shackle data before running a PCA.

Up until this point, sensor data from accelerometers, load shackles and depth sensors had been analysed separately. In the combined analysis, data from these sensors would be combined into a single dataframe for analysis, in an attempt to uncover possible cross-sensor patterns.

Finally, the reader is presented with results from the reconstruction experiment. Here, a reduced depth sensor setup was used to reconstruct data for the missing sensors, further verifying the capabilities of PCA and shedding light on potential other use-cases of the technique.

4.1 Principal Component Analysis

This section presents the results from running the principal component analysis on each of the different sensors, without any filtering or rolling window averages. The PCA plots show the cumulative explained variance when including n sensors. In this analysis, it was decided that a cumulative explained variance above 0.95 (or 95%) would be considered a successful reduction in dimensionality, preferably with as few sensors as possible.

It is imperative that this limit be approached with a bit of caution. Consider the case where one is trying to reduce the dimensionality of measurements from 20 sensors by only using the sensors that explain 95% of the variance. Even when there is absolutely no underlying pattern, each sensor will account for an equal share of the total variance, or roughly 5% in this case. In this scenario, 19 (or less) of the sensors will always be able to explain 95% of the total variance in the dataset despite there being no underlying pattern. Selecting these 19 sensor based on the fact that they explain 95% of the total variance will not yield any "meaningful" reduction. This will be more akin to discarding 5% of the information.

That being said, what follows are the results of the principal component analysis.

4.1.1 Accelerometers

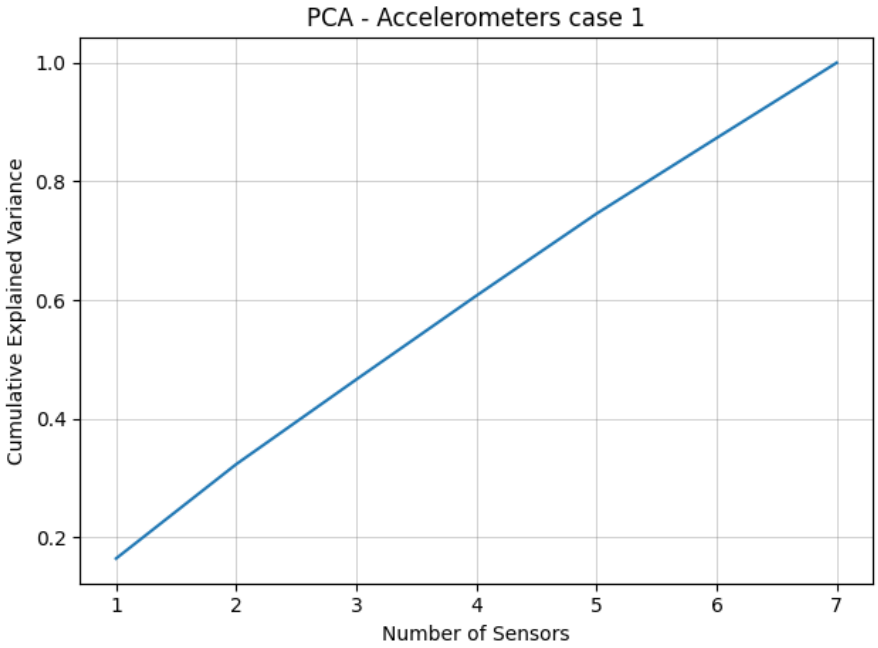


Figure 4.1: Plot showing the cumulative explained variance when using n accelerometer sensors. The analysis is performed on data from the time period given by case 1 in Table 3.2, that is, from 09:15 to 10:45 on the 18th of January.

As seen in Figure 4.1, each sensor accounts for roughly one-seventh of the total variance, thus

forming a (almost) straight diagonal line when plotted cumulatively. This is an indication that the analysis didn't find any underlying patterns and that every sensor is equally important in explaining variance. It is important to note that this doesn't necessarily mean there *is no* underlying pattern, but that the analysis wasn't able to find any.

While the included plot only shows the results of the analysis on data from one of the 10 cases, it is important to mention that the analysis was conducted on data from within all of the 10 time periods. The analyses of the other 9 cases yielded very similar results, with each sensor only accounting for roughly one-seventh of the total explained variance.

4.1.2 Load shackles

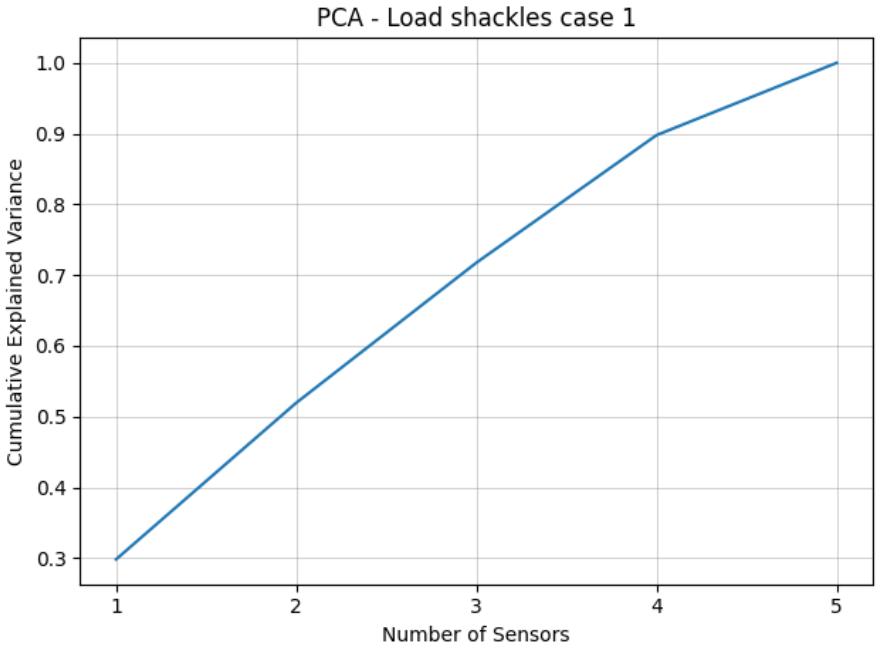
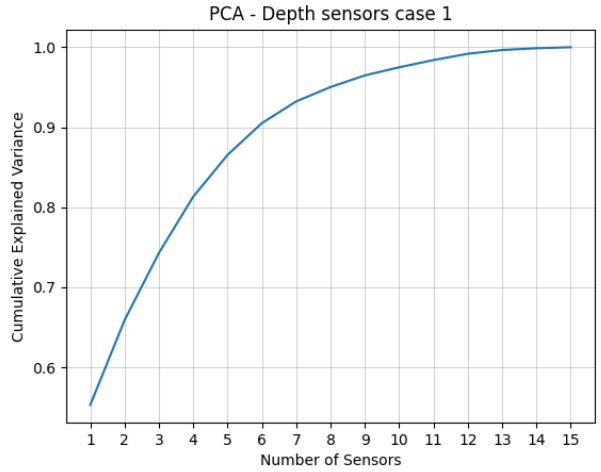


Figure 4.2: Plot showing the cumulative explained variance when using n load shackles. The analysis is performed on data from the time period given by case 1 in Table 3.2.

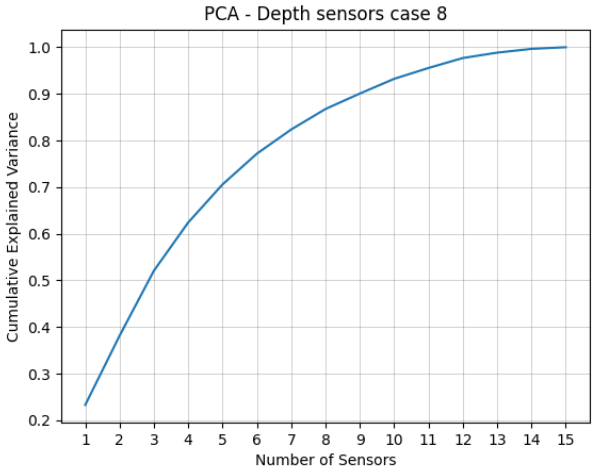
Figure 4.2 shows the results of applying a PCA to load shackle data from case 1 in Table 3.2. No one sensor is able to explain more than 30% of the total variance in the dataset. Furthermore, the least important sensor still accounts for 10% of the total variance. This again indicates that the analysis wasn't able to find any linear connection between the different sensors, or that they are all roughly equally important in explaining the dataset's variance.

The analysis was repeated on data from all the 9 other cases. However, most of them gave similar results. Data from day 6 indicated that 4 of the 5 sensors were able to explain 95.3% of the total variance (these were sensors 1, 2, 3, and 5), but this seems mostly due to random chance, given the fact that none of the other days produced similar results.

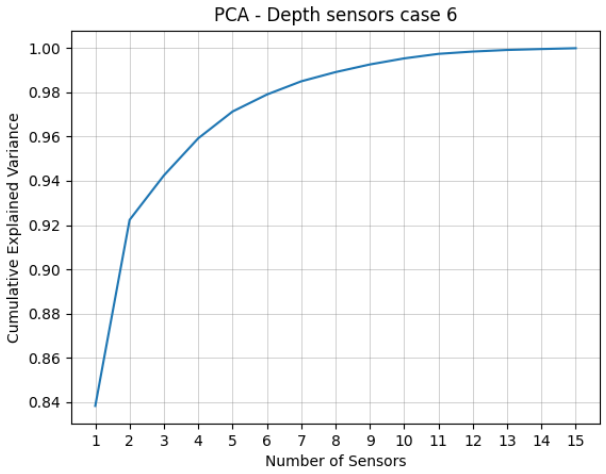
4.1.3 Depth sensors



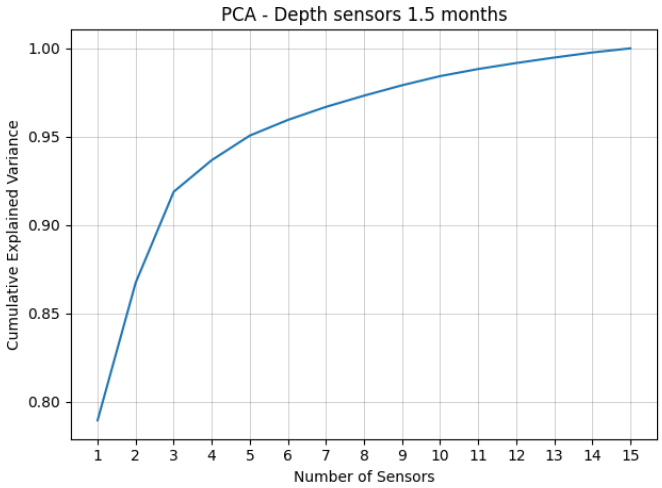
(a) Result of PCA on depth sensor data from case 1.



(b) Result of PCA on depth sensor data from case 8.



(c) Result of PCA on depth sensor data from case 6.



(d) Result of PCA on *all* the depth sensor data starting from case 1 to case 10 (including the time in-between cases).

Figure 4.3: Figures showing the results of running PCA on depth sensor data from various time periods. (a) shows the results from case 1, as has been done with accelerometer and load shackle data. (b) shows the results from case 8, where the analysis required the highest number of sensors (11) to explain 95% of the total variance. (c) shows the results from case 6, where the analysis required the lowest number of sensors (4) to explain 95% of the total variance. (d) shows the results of the analysis when using all the data from 18.01 to 01.03 (2020).

The PCA yielded much more interesting and varied results when applied to data from the depth sensors. When run on data from case 1, it indicated that 8 out of 15 sensors were required to explain 95% of the variance in the dataset (as seen in Figure 4.3a). As opposed to the results from the accelerometer and load shackles, one depth sensor alone is able to account for slightly less than 60% of the variance in the dataset, greatly exceeding the $\frac{1}{15} \approx 6.67\%$ one

would expect if the analysis had failed to find any connection between the readings.

However, the analysis yielded mixed results when comparing performance across all the different cases. The analysis showed that 11 out of 15 sensors were needed to explain 95% variance on data from case 8 (as seen in Figure 4.3b), while only 4 out of 15 sensors were needed on data from case 6 (as seen in Figure 4.3c): these were sensors no. 1, 5, 8 and 16.

More sensors were generally needed to explain 95% of the variance on days with calm weather (like case 1, case 8 or case 10). On the other hand, the analysis indicated that some stormy days required comparatively few sensors (like case 5 or 6), while other stormy days required many sensors (like on case 4) to explain 95% of the variance. Overall, there didn't appear to be any clear correlation between the environmental conditions and sensors required to explain 95% of the variance.

1.5-Month analysis

Sampling data once every 4 minutes meant that each of the depth sensors only made 23 measurements in each 1.5 hour case that was analysed. This seemed to be quite sparse, especially considering the importance of dataset size as discussed in Section 2.4.4. With this in mind, it was decided that a PCA would be run on all the data gathered in the time period starting from the beginning of case 1 and lasting until the end of case 10 (Table 3.2). It was hoped that this would further confirm that dimensionality reduction could be achieved for the depth sensors. Using data from a 1.5 month time period amounted to roughly 15500 data points for each depth sensor.

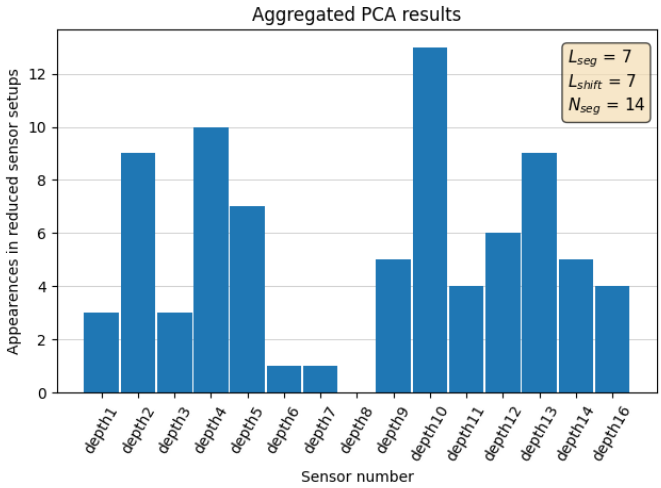
The result of applying the PCA to all the depth sensor data from the 18th of January to the 1st of March is shown in Figure 4.3d. Here, the analysis indicated that only 5 sensors were needed to explain 95.1% of the dataset's total variance. These were sensors no. 1, 2, 4, 10, and 12.

4.1.4 Aggregated analysis of depth sensors

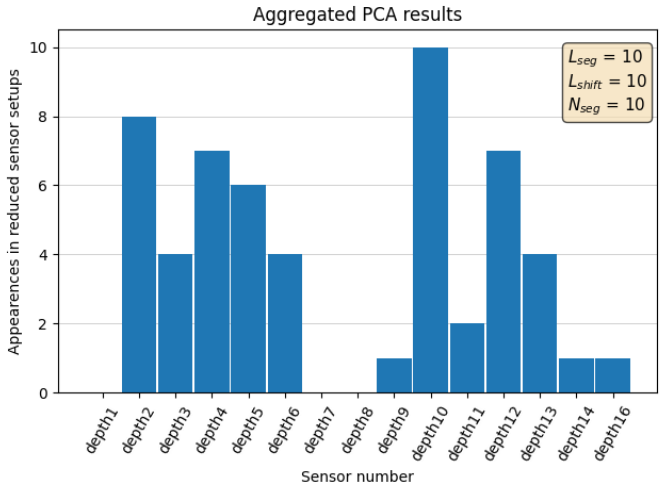
In this section, the entire available depth sensor dataset was split into segments of equal length and analysed individually. As before, the IDs of the sensors that were required to explain 95% of the variance in each of the segments were returned. By counting how many times each of the sensors was included in the list of returned sensors, a bar chart could be made. These charts, using various segment lengths L_{seg} and shift lengths L_{shift} are shown in Figure 4.4.

Sensor no. 10 seems to be the most important. Further inspection reveals that sensor no. 2 also seems to be quite important, as it consistently appears in most of the reduced sensor setups. Finally, sensors 4, 5, 12 and 13 also appear to be quite important, but with somewhat varying scores across the tests. All in all, sensors 2, 4, 5, 10, 12, and 13 seem to be most important according to the aggregated analysis, with sensors no. 2 and 10 being especially important.

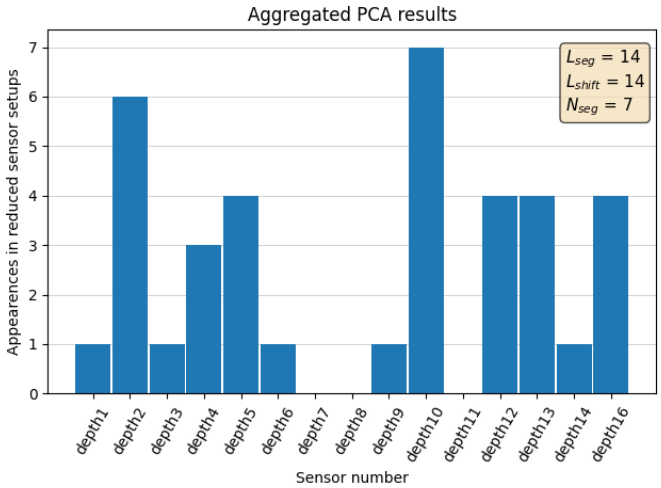
While the choice of L_{seg} and L_{shift} seems to influence the importance of some sensors, others score low across all the tests. Notably, sensors 1, 6, 7, 8, 9, 11, and 14 all have a low score on at least three of the four charts. This indicates that they aren't as important as some of the other sensors.



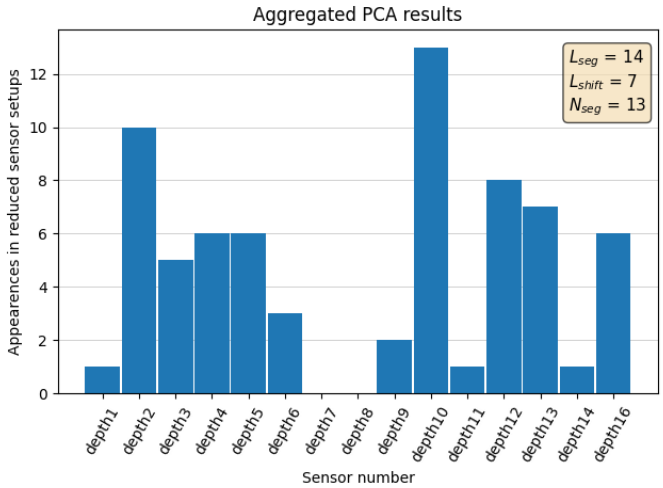
(a) Aggregated results with $L_{seg} = 7$, $L_{shift} = 7$, $N_{seg} = 14$.



(b) Aggregated results with $L_{seg} = 10$, $L_{shift} = 10$, $N_{seg} = 10$.



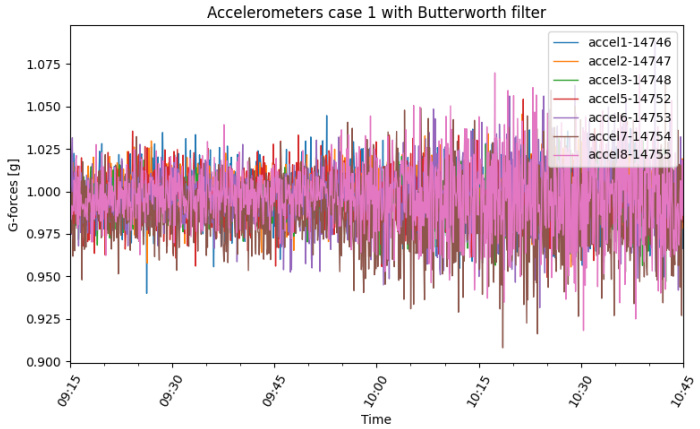
(c) Aggregated results with $L_{seg} = 14$, $L_{shift} = 14$, $N_{seg} = 7$.



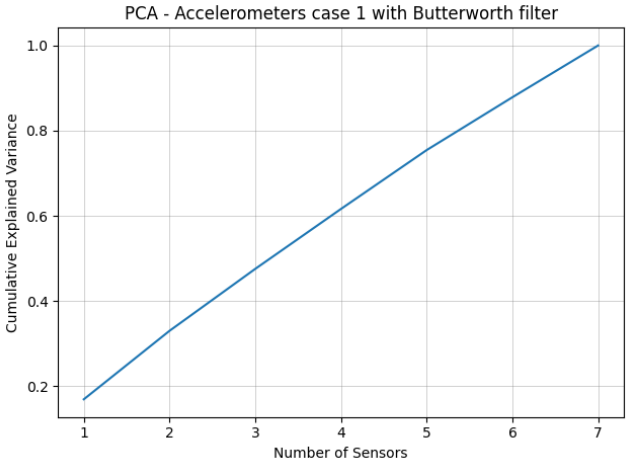
(d) Aggregated results with $L_{seg} = 14$, $L_{shift} = 7$, $N_{seg} = 13$.

Figure 4.4: Figures showing aggregated PCA results from depth sensor data using various segment lengths and shifts. All these analyses were done on data starting from the 21st of December.

Shorter and longer segment lengths than those shown were also tried. However, clear patterns didn't seem to emerge before using a segment length L_{seg} of at least 5 days. On the other hand, using segment lengths above 14 days meant that there would be quite few segments to analyse, resulting in vertically squished graphs. This would make it much more difficult to clearly distinguish between the importance of some sensors.



(a) Accelerometer data after applying a Butterworth filter, using a cutoff frequency of 0.8Hz and filter order of 4. Only every 40th data point (every 5 seconds) is shown in this plot for visualization purposes.



(b) Result of performing PCA on Butterworth filtered accelerometer data.

Figure 4.5: Figure that summarizes the effect of applying a Butterworth filter to the accelerometer data. (a) shows the filtered accelerometer data. (b) shows the results of the PCA when using the filtered data.

4.2 Butterworth filter

Figure 4.5 shows the results of applying the Butterworth filter with a 0.8 Hz cutoff frequency to the accelerometer data. Figure 4.5a shows the filtered data, while Figure 4.5b shows the cumulative explained variance. The fact that the cumulative explained variance closely resembles that of the unfiltered accelerometer data (Figure 4.1) indicates that the Butterworth filter has little effectiveness in improving PCA results compared to the unfiltered data.

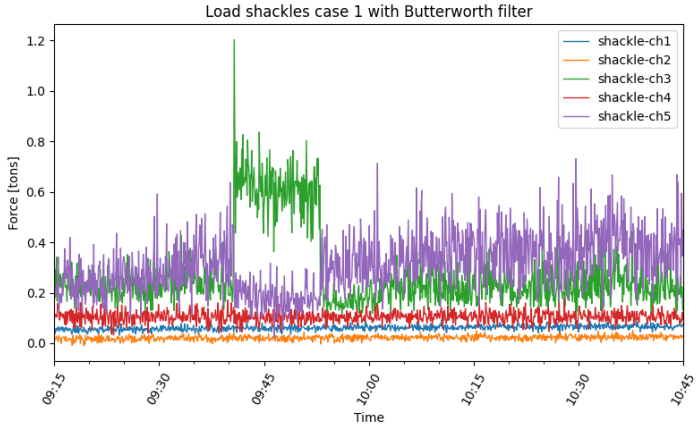
Several other cutoff frequencies were also tested, as well as using data from other time periods to investigate whether the filter could improve the PCA results. When higher cutoff frequencies were used, there were minimal changes to the data, but each sensor would only be capable of explaining roughly one-seventh of the total variance. As expected, this indicates that higher cutoff frequencies are ineffective at improving the PCA outcome, as more noise will be included.

On the other hand, when using cutoff frequencies below 0.2 Hz, the PCA was able reach 95% explained variance with only 6 out of 7 sensors. However, at such low cutoff frequencies, the data was essentially destroyed, losing crucial information from the original signal. Indications that the dimensionality of the dataset could be reduced wouldn't be properly grounded in the original data anymore. This demonstrated that low cutoff frequencies weren't suitable for preserving the integrity of the data while achieving meaningful dimensionality reduction.

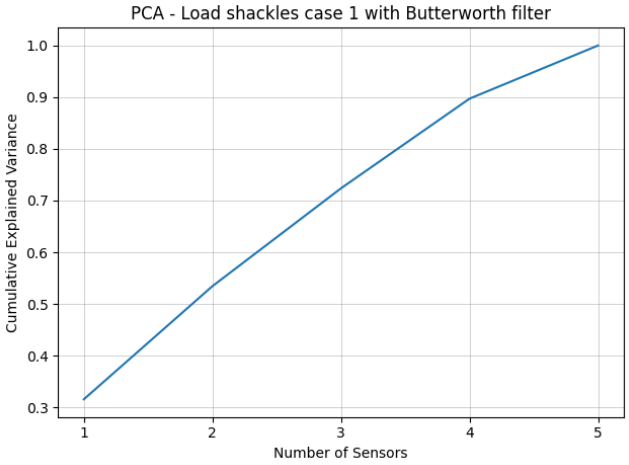
By considering the preservation of crucial signal components and the fact that the accelerometers were intended to capture the effect of waves on the fish farm, a cutoff frequency of

0.8Hz was determined to be a suitable choice for the Butterworth filter that was used in this thesis. However, it's effectiveness in improving PCA results was very limited.

The Butterworth filter was also applied to data from the load shackles before performing PCA, as shown in Figure 4.6. Like with the accelerometers, the load shackles didn't benefit from being filtered. As was done with the accelerometer data, various cutoff frequencies were tested on data from various time periods, but with little success. The points that have been briefly discussed in this section regarding choice of cutoff frequency and lack of improvement in PCA results apply directly to the load shackles as well.



(a) Load shackle data after applying a Butterworth filter, using a cutoff frequency of 0.8Hz and filter order of 4. Only every 20th data point (every 5 seconds) is shown in this plot for visualization purposes.



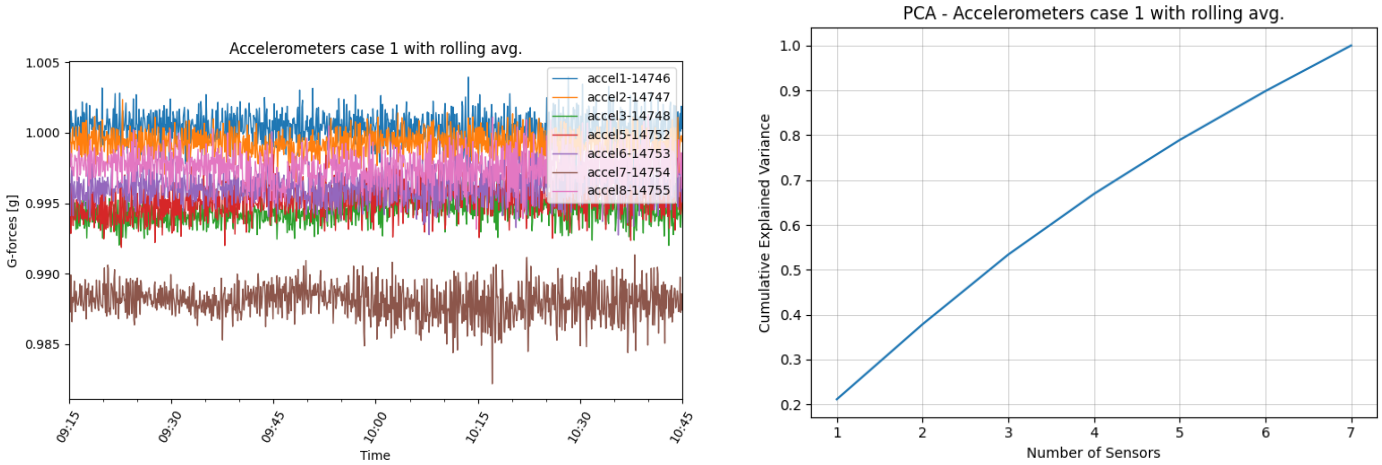
(b) Result of performing PCA on Butterworth filtered load shackle data.

Figure 4.6: Figure that summarizes the effect of applying a Butterworth filter to the load shackle data. (a) shows the filtered load shackle data. (b) shows the results of the PCA when using the filtered data.

4.3 Rolling averages

In addition to running the PCA on Butterworth filtered data, it was also run on data that had first been smoothed by a rolling average. The results of this analysis are shown in Figure 4.7 and Figure 4.8. When rolling average windows were applied to the accelerometer data, no reduction in dimensionality was achieved before extending the window size to roughly 10s. At this point, data from case 6 managed to obtain a cumulative explained variance of 95.1% using only 6 out of 7 sensors. However, this is not representative of the result obtained on other days. When using rolling averages with a window size of 10s, accelerometer data from most other cases performed similarly to that shown in Figure 4.7b.

The results were slightly better when using load shackle data, albeit not substantially. When using a window size of 1s, most cases performed similarly to that shown in Figure 4.8b. Only data from case 6 was able to surpass 95% cumulative explained variance without all the sensors.



(a) Accelerometer data after applying a rolling average with a window size of 10s (80 measurements). Every 40th measurement plotted (every 5s) for visualization purposes.

(b) Result of performing PCA on accelerometer data that had first been smoothed using a rolling average.

Figure 4.7: Figure summarizing how smoothing accelerometer data with a rolling windows average affects the analysis. (a) shows how the raw data is transformed by the filter. (b) shows the results of the PCA when using the rolling-averaged data.

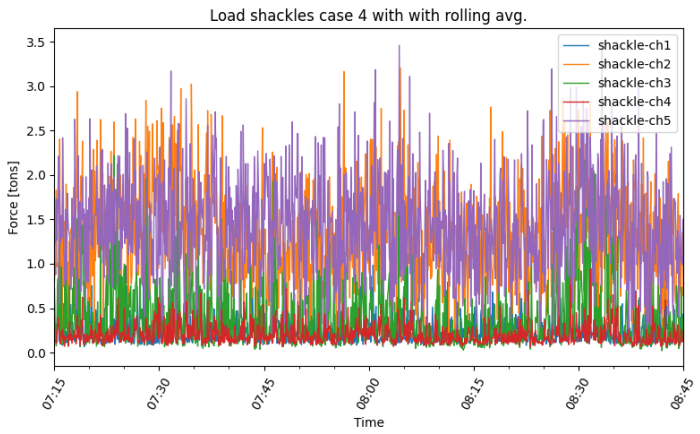
When increasing the window size to 5s, data from cases 1, 5, 6, 7, and 8 were all able to surpass 95% cumulative explained variance without needing all 5 sensors. However, the results from these cases were split when it came to which sensor was least important in explaining variance. The results from cases 1 and 7 indicated that sensor no. 4 carried the least information, results from cases 5 and 6 indicated that sensor no. 1 carried the least information, while the result from case 8 indicated that sensor no. 3 carried the least information. All this is to say that no conclusion can be drawn with certainty.

In the context of this study, applying a rolling average to the accelerometer and load shackle data did not substantially improve the PCA results. As mentioned, several different window sizes were tested, from short windows that preserved more detail in the data, to longer windows that provided more smoothing.

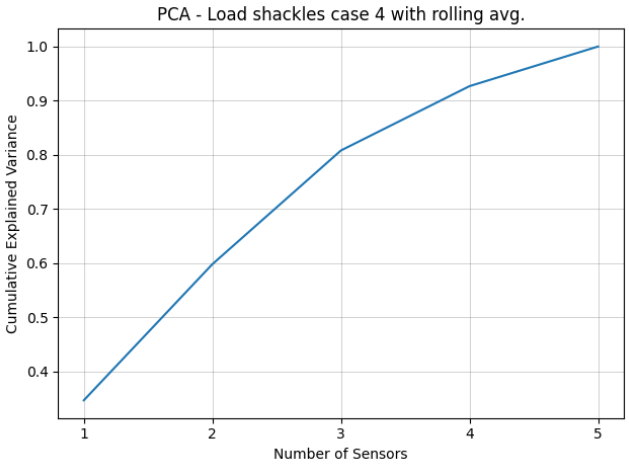
For shorter window sizes, the PCA results were not significantly improved and the analysis wasn't able to eliminate any of the sensors. As the window size was increased, the PCA began to yield results, with the explained variance increasing as the window size was expanded. However, this approach has an important drawback. The larger the window size, the more the data is "smeared" out over time. This results in a loss of detail in the data, as the original signal is increasingly averaged out. At these larger window sizes, the filtered data began to lose meaningful connection to the original data.

Despite *some* results when applied to load shackle data, the use of a rolling average filter cannot be deemed successful in substantially or meaningfully improving the PCA results for the accelerometer and load shackle data in this study. The trade-off between smoothing the data (and consequently improving the PCA results) and preserving the integrity of the original

signal could not be satisfactorily resolved using this method.



(a) Load shackle data after applying a rolling average with a window size of 1s (4 measurements). Every 20th measurement plotted (every 5s) for visualization purposes.



(b) Result of performing PCA on load shackle data that had first been smoothed using a rolling average.

Figure 4.8: Figure summarizing how smoothing load shackle data with a rolling window average affects the analysis. (a) shows how the raw data is transformed by the filter. (b) shows the results of the PCA when using the rolling-averaged data.

4.4 Combined data analysis

In this section, an investigation is launched to determine whether more sparse sensor setups can be uncovered by first combining different sensor data into a single dataframe and then analysing the combined data. The goal is to explore whether combining the datasets may reveal any shared patterns of variance that could be exploited to achieve a more effective dimensionality reduction with the PCA. By performing the analysis on this joint dataset, an attempt is made at leveraging the potential of PCA to uncover patterns of variance that span multiple sensor types, potentially leading to a better representation and consequent analysis of the sensor data.

4.4.1 Accelerometers and load shackles

One might reasonably expect that accelerometer data and load shackle data could be correlated due to the shared effect waves have on both of these. By analyzing the accelerometer and load shackle data together, it was hoped that the PCA would be able to identify a smaller number of principal components that effectively capture the joint variability of these sensors.

Unfortunately, the results of the PCA on the combined dataset did not show substantial improvements over the separate analyses. Most of the results were similar to that of case 7, shown in Figure 4.9. Here, 11 out of 12 sensors were needed to reach a 95% explained variance. There were occasional instances where one sensor could be dropped while still explaining a signific-

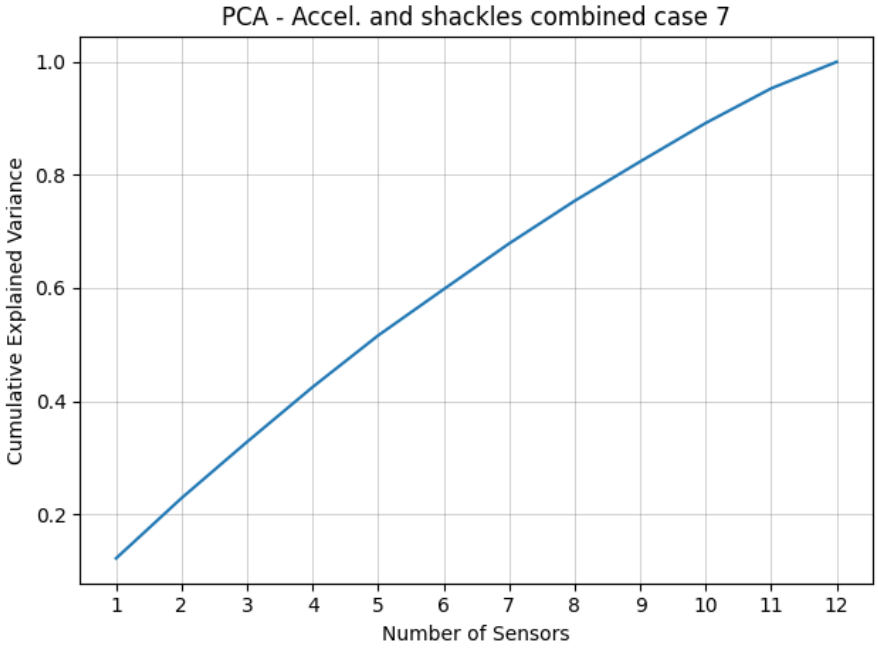


Figure 4.9: Figure showing cumulative explained variance when running PCA on combined accelerometer and load shackle data. The plot only shows the results of running the analysis on data from case 7. Most other cases yielded similar results.

ant portion of the variance, but these occurrences were inconsistent and appeared to be due to random chance rather than a meaningful pattern in the data.

Given the results, it doesn't appear as though combining the accelerometer and load shackle data has a substantial effect on the results of the PCA. This suggests that the underlying factors driving the variability in these sensors' measurements may be independent, or that both datasets simply exhibit some pattern that cannot be uncovered by the PCA. Thus, combining the datasets did not yield the desired improvements in the effectiveness of the PCA.

4.4.2 All sensors

Although it is more difficult to reasonably justify, an experimental analysis was also conducted where all the sensor data was combined into a single dataframe, before running a PCA. This included measurements from the accelerometers, load shackles, and the depth sensors.

Applying PCA to this combined dataset resulted in a reduction of approximately 6-10 sensors in each of the 10 cases chosen for analysis. Results are shown for case 3 in Figure 4.10. Upon further inspection, it was found that this reduction was essentially the same as what was achieved when PCA was applied *independently* to each of the different sensors. In other words, the act of combining data didn't contribute to a better explanation of the variance. Instead, it seemed that the structure found in the depth sensor data alone dominated the PCA results.

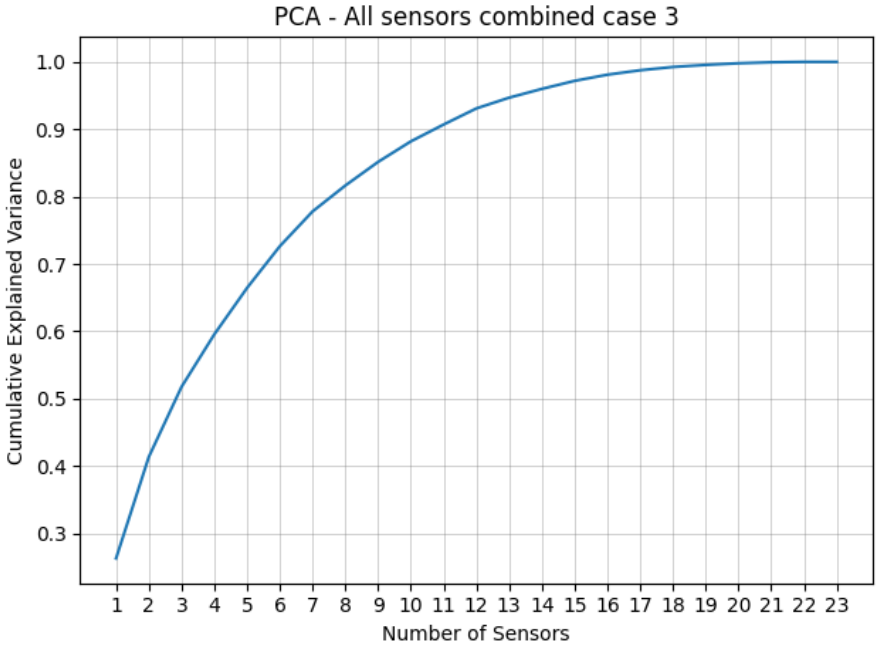


Figure 4.10: Figure showing cumulative explained variance when running PCA on data from all the sensors combined. The plot only shows the results from case 3. Most other cases yielded similar results.

Despite this attempt to find underlying patterns in the combined dataset, no further dimensionality reduction was achieved beyond what was observed with the depth sensor data alone. Combining all the sensor data into one dataset before performing a PCA did not reveal any additional patterns that could be leveraged for better results. Nonetheless, this analysis confirmed the importance of the information contained in the depth sensor data, supporting the results of the previously conducted independent analysis in Section 4.1.

4.5 Reconstruction of depth sensor data

In this analysis, PCA was run on data from a 1.5-month period (from 10:00 on the 18th of January to 14:00 on the 1st of March) to identify the most important sensors. As presented earlier, the results can be seen in Figure 4.3d. Again, the analysis indicated that sensors 1, 2, 4, 10, and 12 were the most important sensors. The principal components corresponding to these sensors were then extracted.

To ensure that the reconstruction process was applied to unseen data, a separate dataset was created. This reduced dataset was created by only extracting the measurements from sensors 1, 2, 4, 10, and 12 over a 2-hour time period immediately following the end of the 1.5-month period. The reconstruction process involved using the extracted principal components to essentially do "PCA in reverse" and transform the reduced dataset back to the original high-

dimensional space. The error between the original and reconstructed sensor values is shown in Figure 4.11.

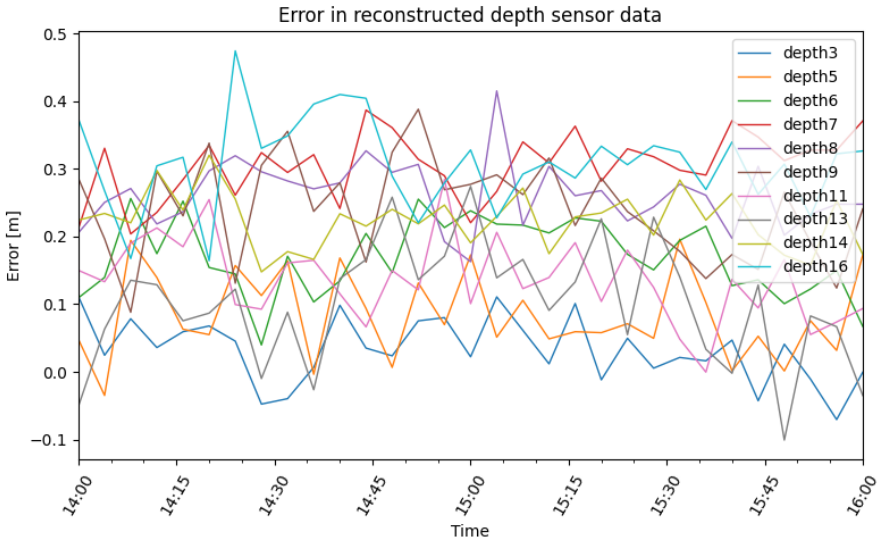


Figure 4.11: Figure showing the error in reconstructed depth sensor data starting from 14:00 on the 1st of March. Reconstruction was done using data from sensors 1, 2, 4, 10, and 12; the sensors found to be most important in the 1.5 month analysis.

While the reconstructed data does not perfectly match the original data, it nonetheless represents a reasonable approximation. However, it is essential to note that despite the mathematical possibility of this reconstruction, its practical application is very limited due to the inherent loss of information during the dimensionality reduction process. This is more evident when trying to reconstruct data over longer periods of time. To highlight this point, the experiment was repeated twice more, with the aim of reconstructing data over a 5-day and 10-day period.

This experiment was repeated using the sensors that were found to be most important in the aggregated analysis, namely sensors 2, 4, 5, 10, 12, and 13. The PCA was run on the same data, but this time, the principal components corresponding to these sensors were extracted and kept. Measurements from these sensors were then extracted over a 5-day period (instead of 2 hours), and the principal components were used to reconstruct data for the remaining sensors. The results are shown in Figure 4.12.

Although not shown, the experiment was repeated one final time (using the same sensors as those found in the aggregated analysis), with the aim of reconstructing data over a 10-day period. In this case, errors peaked at slightly above one meter.

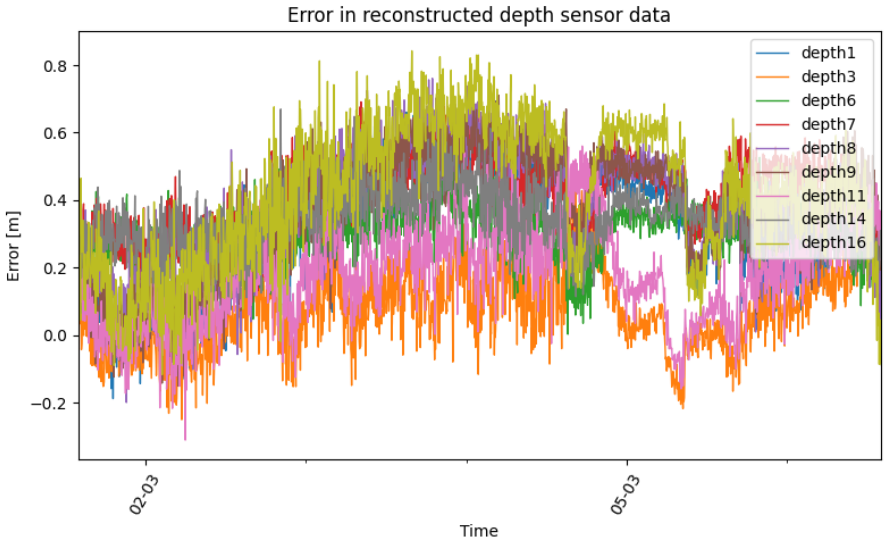


Figure 4.12: Figure showing the error in reconstructed depth sensor data over 5 days starting from 14:00 on the 1st of March. This time, the reconstruction was done using data from sensors 2, 4, 5, 10, 12, and 13; the sensors found to be most important in the aggregated analysis.

Chapter 5

Discussion

5.1 Summary of findings

This master's thesis has delved into the potential of using Principal Component Analysis (PCA) as a method for identifying more sparse sensor setups in fish farming operations. What follows here is a brief summary of the all the different analyses and the findings. A more thorough discussion of the results is given in Section 5.2.

Depth sensors:

The initial analysis consisted of performing PCA on the 10 cases listed in Table 3.2. Here, reduced sensor setups were obtained that were able to explain at least 95% of the given data-set's variance. However, these reduced sensor setups varied greatly both in terms of how many sensors were required (mostly 7-11 sensors, down from 15), and which sensors these were. Furthermore, there didn't seem to be any connection between the optimal sensor setups and the environmental conditions in the different cases. It nonetheless indicated that there was potential for finding reduced sensor setups using the PCA.

In the 1.5 month analysis, it was found that sensors 1, 2, 4, 10, and 12 were key contributors, explaining around 95% of the total variance in the dataset. This was a big reduction in the number of sensors as compared to the initial analysis, and seemed to be more reliable as it was based on a much larger number of depth measurements.

Afterwards, the aggregated analysis was performed. This analysis pointed very strongly towards sensor no. 10 being essential, followed closely by sensor no. 2. Besides these two, sensors 4, 5, 12, and 13 were also found to be quite important.

Finally, mostly as an academic exercise, an attempt was made at reconstructing data based on various reduced sensor setups. Comparing the reconstructed data with the real data revealed relatively high errors, indicating limited utility. While this procedure *is* sometimes used in other academic applications (albeit infrequently), it is not well suited in the context of this thesis, as will be further elaborated shortly, in Section 5.2

Accelerometers and load shackles:

When PCAs were applied to accelerometer and load shackle data, the results were less encouraging. Each sensor was found to be equally important in explaining the dataset's variance. In spite of also using noise-reduction techniques such as the Butterworth filter and rolling window averages, no substantial dimensionality reduction was achieved. The variance in these datasets was more evenly distributed across all sensors, indicating the absence of a small subset of dominant sensors akin to what was found in the depth sensor data. It is unclear if this is because each sensor fundamentally captures different components of the dynamics at the fish farm, or whether the sensors simply contain non-linear relationships that can't be picked up by the PCA.

Combining datasets:

In the first attempt, accelerometer and load shackle data were combined before running the PCA. This didn't yield any significant sensor reductions. Next, data from all the different sensors (accelerometers, load shackles and depth sensors) were combined before running the PCA. This required downsampling all data to the sampling rate used by the depth sensors (4 minutes). No further reduction was observed in dimensionality beyond what was observed for each of the sensor types individually, further underscoring the challenge posed by the accelerometer and load shackle datasets.

5.2 Interpretation of results

5.2.1 Depth sensors

Initial analysis

Recall that initially, PCAs were performed on ten distinct 1.5-hour time intervals, each with different environmental conditions. This approach was chosen in an attempt to investigate whether the sensor importance varied significantly based on the environmental state at the time of measurements.

The initial PCA results for the depth sensors were somewhat inconsistent, reflecting different "important" sensors across the ten selected cases. This is highlighted by Figure 4.3, which shows that 11 sensors were required to explain 95% of the dataset's variance in case 8, while only 4 sensors were required in case 6. This observed variability could be attributed to several factors.

One key element that was likely causing much of the variation was the relatively short duration of each of the cases: 1.5 hours. As mentioned before, this only amounts to 23 data points for each sensor. This might not have provided a sufficiently representative sample of the normal operating conditions. Recall from Section 2.4.4 that research shows that too little data can make PCA results unstable, although the amount of data required to avoid such instability isn't well defined. Another factor that could help explain some of the variation is the inherent stochasticity in the dynamics of the sea cage in the presence of various weather conditions.

Connection between environmental conditions and reduced sensor setup

Unfortunately, no correlation was observed between the environmental conditions in the different cases and the reduced sensor setups that were found in each case. The initial analysis mainly seemed to indicate that more data would need to be analysed in order to obtain stable sensor setups. While this would be achieved in the coming analyses, the greatly prolonged time periods that had to be used (upwards of 5 days) made it practically impossible to analyse the impact of different environmental conditions and the optimal sensor setups.

It would be interesting to see if this problem could be solved by having more measurements (i.e. more frequent sampling rate) available in each case. One might reasonably hypothesize that a more frequent sampling rate would make the reduced sensor setups found in each 1.5-hour case more stable, but this would need to be tested. This would potentially enable one to analyse the impact of different weather conditions on the optimal sensor setup.

1.5-month analysis

The findings from this analysis (as seen in Figure 4.3d) indicated that sensors 1, 2, 4, 10, and 12 collectively explained approximately 95% of the total variance. This sensor setup was much more sparse than most of those generated by running PCA on each of the 10 cases, suggesting that five depth sensors could in theory be enough to provide a near-comprehensive understanding of the depth dynamics at the fish farm. An encouraging finding in terms of potential cost savings and sensor management. This analysis also supported the suspicion that 23 data points per sensor were too few.

While this analysis was a success in terms of finding a highly sparse sensor setup, its indication that the sensors 1, 2, 4, 10 and 12 (and not some other sensors) were most important could well be questioned as it was only performed once.

Aggregated analysis

In the aggregated analysis, the entire available dataset was split into segments of equal lengths, and a PCA was run on each of these segments, counting how many times each sensor was included in the reduced setups. Various segment lengths and shift lengths were attempted as seen in Figure 4.4. Varying these parameters also had the added benefit of shedding light upon whether the reduced sensor setups were dependent on the methodology employed (segment/shift lengths) or not.

By employing the aggregated analysis, the variability observed in the initial cases was reduced. The analysis pointed to sensor no. 10 as the most critical, closely followed by sensor no. 2. Sensors 4, 5, 12, and 13 were also noted as important, although to a slightly lesser extent. This was in close agreeance with the result of the 1.5 month analysis, except for one major deviation: in this analysis, sensor no. 1 was found to be quite unimportant. The fact that sensor no. 1 was almost never included in the reduced sensor setups *despite* trying varying segment lengths strogly undermines its importance. Furthermore, it indicates that the inclusion of sensor no. 1 in the results of the 1.5 month analysis was likely "due to chance".

The results of the aggregated analysis reflect a more consistent understanding of sensor importance, and should be weighted more heavily. Nonetheless, the 1.5 month analysis also pointed to sensors no. 2, 4, 10, and 12 being important, something that the aggregated analysis also supports.

Unfortunately, the principal component analysis does not provide any intuitive reasoning behind its "chosen" dimensionality reductions. This is why it is often referred to as a "exploratory analysis". As mentioned in Section 2.4.1, the PCA is perhaps best used to guide further investigations or to reduce the dimensionality of a dataset before further analyses. That being said, one might reasonably speculate why some of these sensors are included.

Optimal sensor setup speculation

The inclusion of sensor no. 2 is perhaps the one which is easiest to understand intuitively. Looking at Figure 3.4, one can see that sensor no. 2 is the *only* sensor that is located at a depth of 30m. This might make its dynamics distinct from the others, which are located higher up and along vertical net cage segments. In terms of the PCA, recall that it in essence tries to maximize the amount of information retained while performing a dimensionality reduction. Sensor no. 2 might carry information that none of the other sensors have. This essentially makes the information sensor no. 2 carries much more "important".

As a counter-example, consider sensor no. 7. If it were to be discarded, sensors no. 6 or 8 could likely substitute it due to their close horizontal proximity and identical vertical placement. If sensor no. 2 were to be discarded, there wouldn't be any other sensors in the same horizontal plane that could substitute it in the same way.

By the same logic, it is fairly easy to understand the inclusion of at least one sensor from 7m depth and at least one sensor from 15m depth. This would at least partly explain the inclusion of sensor no. 4 or 5 (from 7m depth), and sensor no. 10, 12 or 13 (from 15m depth).

Unanswered questions raised by aggregated analysis

A few things clearly still remain as mysteries. If one examines Figure 4.4 and Figure 3.4 once more, one might wonder: why exactly is sensor no. 10 marked as critical in importance across all the tests? What makes this specific sensor so important?

One might hypothesize that parts of these questions can be answered by closely examining the general direction of currents over the time span of the analysis windows. If for instance the current tends to flow mostly from north to south (see Figure 3.2 and Figure 3.4), then one would expect sensor no. 10 (being placed on the northern side) to be displaced more than sensors on the south side. The inclusion of sensor no. 10 would therefore cover most of the variation seen in the dataset. This seems intuitively sound, especially as waves are generally reported to hit Buholmen from the north. Nonetheless, this explanation would need much more thorough investigation.

Furthermore, notice that sensors 4 and 5 are right next to each other. Likewise, sensors 12 and 13 are also right next to each other. Why are sensor pairs that are right next to each other

being marked as important? Intuitively speaking, shouldn't sensors that are more spread out better explain the variability observed in readings from across the entire net cage?

One might hypothesize that the apparent inclusion of two pairs of sensors that are right next to each other can be explained through further examination of the results of the aggregated analysis. Notice that in each of the subfigures in Figure 4.4, the number of times sensors 4-5 are included seem to roughly add up to N_{seg}. The same applies to sensors 12-13. It might be the case that sensors no. 5 and 13 are excluded from the reduced setups whenever sensors no. 4 and 12 are included, and vice versa. This would indicate that only one of each pair is needed, but that aggregating the results can make it *seem* as though both are needed.

Finally, notice that sensors 4 and 5 are located right above sensors 12 and 13. What makes sensors in this particular quadrant so special in terms of explaining the dataset's variance? Intuitively speaking, wouldn't it be beneficial to also have sensors present in another quadrant of the net cage, e.g. sensors 4-5 and 9-16?

This could potentially also be explained by examining the general direction of currents, but would contradict the inclusion of sensor no. 10 if currents tend to flow from southeast to northwest.

Unfortunately, there is no way to know the answers to these questions with certainty as long as one uses the PCA.

Note on shifting sensor timestamps

Recall that sensors no. 5 and 12 were initially offset by two and one minutes, respectively, from the rest of the sensors. This was rectified by subtracting two and one minutes from their timestamps (respectively) to match those of the rest of the sensors. Some might suspect this to introduce errors, but this is most unlikely. It is generally well understood that neither ocean currents nor coastal currents vary much on the time scale of minutes. The mechanisms that drive changes in currents tend to occur on timescales of days or longer, and are discussed in much more detail in textbooks such as [43–45].

There are a few phenomena that can cause changes on the time scale of hours in places like Buholmen. Most notably, these include winds, which mainly affect surface currents, tides, that occur roughly on a bi-daily cycle, and coastal up- and down-welling, that mainly affect currents over several hours. Finally, studies such as [46] that examine temporal current changes often use sensors that record data once an hour or so, again underscoring the fact that currents don't change much over the course of a few minutes. Thus, the temporal shift applied to depth sensors no. 5 and 12 is highly unlikely to have distorted the results in any way.

Linearity of depth sensor data

Before wrapping up the discussion of depth sensor results, it is worth briefly considering why the analysis proved successful at all. This can be helpful in understanding why the same analysis failed on accelerometer and load shackle data. One central assumption that lays the foundation for how the PCA works is that the data must be linearly dependent for the analysis to

work. Depth sensor data must evidently adhere to this condition, at least to some degree.

The relatively stable underwater-environment could well explain why the depth sensors contain linear relationships, as opposed to the accelerometers and load shackles that are being bombarded by chaotic and rapidly changing waves. Depth sensors are almost exclusively affected by currents, which crucially tend to vary far less abruptly in both direction and strength. This is indeed why they only return measurements once every 4 minutes: more frequent measurements aren't necessary in terms of capturing big deformation changes. The vastly slower dynamics of the currents could be allowing the depth sensors to reach stable equilibrium positions that are linearly dependent.

Summary

In summary, when performed on depth sensor data, especially in the aggregated analysis, the PCA demonstrated its potential in identifying a reduced sensor setup that could explain a majority of the dataset's variance. While the results of this analysis can be used as they are, there is room for other analyses that might provide more *intuitively understandable* explanations behind the optimal sensor setup. This analysis primarily paves the way for further investigations into optimizing sensor deployment, potentially leading to cost-effective and efficient sensor management at the fish farm. It also highlights the importance of selecting appropriate time intervals and having enough test cases for the PCA to return stable results that can be trusted.

5.2.2 Accelerometers and load shackles

The application of PCA to accelerometer and load shackle data resulted in a quite different outcome from the depth sensors. With data from these sensors, PCA yielded no significant reduction in sensor setup. As seen in Figure 4.1 and Figure 4.2, each sensor seems to explain roughly an equal share of the dataset's variance, leaving no opportunity for sensor elimination without significant loss of information. While somewhat disappointing, this finding presents an interesting point of discussion. Specifically, it can lead to a better understanding of the inherent nature of the data coming from these sensors.

Possible explanations for lack of dimensionality reduction

One possible explanation for this result is the absence of any underlying patterns in the data from these sensors. Accelerometers measure the dynamic motion of the floating collar, which is heavily influenced by the chaotic nature of ocean waves. While one often imagines waves as following a wave-front and moving in one direction, waves are typically much more chaotic. Similarly, load shackles measure the tension forces on the sea cage, which are also affected by the unpredictable nature of wave and wind activity[1]. Given the stochastic and unpredictable characteristics of these forces, it is completely plausible that no strong correlation exists

[1]As opposed to currents, wave and wind dynamics are *much* more chaotic and unpredictable. These dynamics are far outside the scope of this thesis, but readers interesting in delving deeper into wave and wind dynamics are referred to [47] and [48].

between the sensors. This would indicate that each sensor is equally important in giving a full picture of the conditions at the fish farm.

Another possible explanation might be attributed to the inherent limitations of the PCA. Recall that the PCA works by finding *linearly independent* components that successively maximize variance. As such, there might *exist* a correlation between the sensors, but this won't be picked up by the PCA if it is a non-linear relationship. Despite their chaotic and unpredictable nature, wind and wave-induced motions might give rise to non-linear relationships in the accelerometer and load shackle data that the PCA is unable to capture.

Non-linearity of data

Even in the presence of relatively well-behaved and uniform wave-fronts, both accelerometer data and load shackle data might simply be containing non-linear relationships. Consider what happens to the accelerometers as a single wave hits the fish farm. As it takes time for the wave to travel from one side of the floating collar to the other, sensors will rise and fall asynchronously. While one set of sensors are rising, others may be stationary or even falling. Considering the size of the farm and the sampling rate, this delay is large enough to make the data non-linear. The same logic can be applied to explain non-linear relationships in load shackle data. A wave likely doesn't hit all the points where load shackles are placed simultaneously. Even if the delay were half a second, this would be enough to offset it in time by two measurements.

If this is the case, it should in theory be possible to shift individual sensor measurements in time before running the PCA to obtain better results. The problem with such a procedure is the chaotic nature of waves: they rarely (if ever) hit the fish farm at a constant angle and constant frequency for prolonged periods. Waves in the sea tend to exist in a chaotic and superposed manner [48]. This makes it exceedingly difficult to estimate the order in which sensors are hit by waves and the time delays before other sensors are hit by the "same" wave.

Limited number of sensors

Another point worth considering is the limited number of sensors. With only 7 accelerometers and 5 load shackles, it might be challenging to achieve a substantial reduction in dimensionality. While having a high number of features isn't a requirement *per se*, most studies applying PCA often deal with significantly higher dimensions, reducing them to a more manageable number while still capturing the majority of the variance.

In [49], PCA is used to improve the performance of neural networks. In this context, 50 features were attempted to be reduced with a PCA, and it was noted that it struggled with non-linear data. In another paper, the authors used upwards of 1500 features [50]. These were then reduced to a size of roughly 1000 features. It's possible that the relatively low number of sensors used in this thesis, coupled with the significant role each sensor plays in capturing the sea cage's state, leads to an equal distribution of explained variance across the sensors.

Furthermore, using more accelerometers and load shackles would decrease the distance between each sensor. This would help counteract the delayed impact waves can have on sensors that are far apart, while also giving the PCA more features to work with.

Dataset size

It is worth clearly stating that the length of each case, 1.5 hours, should easily have been sufficient. Wave dynamics are much quicker than the dynamics of underwater currents. To be able to capture wave dynamics, the accelerometers and load shackles sampled data with frequencies of 8Hz and 4Hz (respectively), as opposed to the vastly slower depth sensors. This meant that when running a PCA on data from one of the cases in Table 3.2, each accelerometer would contain 43200 data points and each load shackle would contain 21600 data points. These numbers far exceed the amount of data points examined in any of the depth sensor analyses.

Note on dropping NaN values and rounding numbers

Furthermore, it is also worth clarifying that dropping NaN values and rounding numbers to 4 decimal points does *not* hinder the PCA from reducing the dimensionality of data. NaN values, though scarce and comprising less than 0.1% of data points in accelerometer data, were omitted. This should not impact the PCA as the it doesn't rely on continuous data. Dropping entire rows doesn't disrupt any potential linear relationships that might be present.

Additionally, the rounding of accelerometer values to 4 decimal points (from 15), and load shackle values from 10 to 4, does not significantly alter the data distribution. Considering that the measurements from these sensors are on the order of 0.1-1, the level of rounding applied in this thesis simply optimizes computation time without substantially affecting the output of the PCA.

Even in a hypothetical scenario where accelerometer or load shackle data were to be on a much smaller scale than that being rounded to, rounding would *aid* the PCA in finding a linear relationship by rounding different numbers to the same values. Thus, rounding and dropping NaN values cannot be said to account for the lack of dimensionality reduction in the accelerometer and load shackle data.

Summary

In summary, the results from the accelerometer and load shackle data indicated that a straightforward PCA might not be the optimal approach for sensor reduction in these cases. The equal importance of each sensor, the chaotic nature of the data, and the limitations of PCA as a linear method all contribute to this conclusion.

This invites the exploration of alternative dimensionality reduction techniques, potentially non-linear ones, to tackle the challenges presented by accelerometer and load shackle data. Alternatively, one might look into various methods for augmenting the data. As shall be presented next, methods for noise reduction were attempted, but with little success in improving PCA results. Despite the lack of immediate success, these results offer valuable insights into the inherent complexities of the data and guide the future direction of research.

5.2.3 Effectiveness of Butterworth filter and rolling window averages

Following the initial analysis of accelerometer and load shackle data, two techniques were employed to investigate if the lack of dimensionality reduction from the PCA could be attributed to the presence of noise. The Butterworth filter and rolling window averages were applied (separately) to the data before running the PCA in an attempt to improve the dimensionality reduction results. The PCA is known to struggle with noisy measurements and outliers, as these can cause otherwise linear relationships to appear non-linear. In this context, neither method resulted in any significant improvements in the PCA results. This indicates that noisy measurements are not to blame for the PCA's inability to generate more sparse sensor setups, a finding that warrants further discussion.

The Butterworth filter, a signal processing tool designed to offer a flat frequency response in the passband, was used in an attempt to remove high-frequency noise from the sensor data. It was expected that the removal of such noise would enhance the detection of any underlying patterns within the data during PCA. However, in this context, the Butterworth filter's impact was far less significant than hoped for. Despite trying several cut-off frequencies to ensure that noise would be removed while retaining the original signal, no real improvements were achieved. This suggests that high-frequency noise was not substantially obscuring the primary patterns within the data.

Likewise, the application of rolling window averages, a technique commonly employed to smooth short-term fluctuations, did not lead to considerable improvements in PCA outcomes. Again, one explanation for this is that the underlying patterns within the accelerometer and load shackle data do not seem to follow a consistent long-term trend. It appears as though the effect of waves on the fish farm is inherently chaotic and dynamic. This means that the data collected by the accelerometers and load shackles contains patterns that are more complex than what rolling window averages can help clarify.

As mentioned previously, in both of these methods, a trade-off had to be made between the degree of data-augmentation and preservation of the original signals. With the Butterworth filter, it can be noted that a sufficiently low cut-off frequency and sufficiently high filter order *will* lead to a substantial improvement in PCA results. However, at such low frequencies, one is essentially distorting the initial signal to such a degree that the PCA might notice linear relationships where there in reality is none. In the same way, a sufficiently long rolling window *will* lead to a substantial improvements in PCA results, but again, this is not desirable if there is no underlying linear relationship to be found.

In conclusion, the attempts at removing noise from the accelerometer and load shackle data did not yield significant improvements in PCA-based dimensionality reduction. This indicates the absence of noise, or that if there is noise present, it certainly is not obscuring an underlying linear relationship. Furthermore, this analysis highlights the complexity of the data, and instead prompts the consideration of alternative dimensionality reduction methods that might be better suited to these particular datasets.

5.2.4 Combination of sensor data

The idea of combining data from different sensors into a single dataframe was fundamentally an attempt at uncovering patterns that might not have been apparent when analysing each sensor type separately. Considering the way in which waves affect the fish farm, it seemed completely plausible that analysing accelerometer and load shackle data simultaneously might have improved PCA results. Thus, two attempts were made: one where only accelerometer data and load shackle data were merged, and another where depth sensor readings were included as well. In both cases, the sampling frequencies were downsampled to that of the slowest sensor involved. None of these combined datasets were able to significantly improve PCA results.

Combining accelerometer and load shackle data for a joined analysis seemed logical as an incoming wave will both strain the mooring lines and elevate the floating collar simultaneously, leading to higher readings among both sensor types. While this may be the case, the combination did not yield any improvements in dimensionality reduction from the PCA, as seen in Figure 4.9. This result signals the absence of a strong linear correlation between the different sensor types. The chaotic nature of waves might be such that the relationship between these sensors is non-linear.

Next, all the different sensor types were combined into a single dataframe that was then used in a PCA. The justification for this approach stemmed from the understanding that a larger dimensional space can only enhance the PCA's performance. Even if there doesn't exist an intuitively obvious correlation between the sensors, the worst possible outcome would simply be the lack of improvement.

Consider the scenario of initially having data from two variables that don't seem to form a straight line and can't be reduced. Adding a third variable (dimension) can either align all data points on a plane or reveal that they are spread out in the third dimension as well. In the latter case, all three dimensions are necessary to describe the variables, but in the former, one dimension can be eliminated. This illustrates that combining data will at worst maintain the same level of dimensionality *reduction* or improve it, but never worsen it.

Given the minimal additional effort of combining datasets, it was seen as a worthwhile endeavor, even if there was no immediately apparent correlation between the different sensor types.

All that being said, combining all the data and downsampling it to the sampling rate of the depth sensors (one measurement per 4 minutes) didn't seem to yield any improved results. The results seen in Figure 4.10 simply pointed to the fact that some of the depth sensors could be removed, a finding that was already previously discovered. All in all, the results obtained from the combined analysis reflected those obtained when running the PCA on each of the sensor types separately. Again, this points to the lack of linear relationships between the various sensor types. However, it also highlights and reinforces the findings regarding reduction of depth sensor setup.

In summary, while combining data from different sensors into a single dataset might provide a more comprehensive view of the system's state, it did not seem to facilitate a more success-

ful PCA application. The inherent differences in the nature of the data and the limitations of PCA in capturing complex relationships may both contribute to this result. These findings further emphasize the need to consider alternative dimensionality reduction techniques if further reduction is desired beyond what was observed for the depth sensors.

5.2.5 Reconstruction of depth sensor data

The reconstruction of depth sensor data was another experiment conducted in this study, not out of direct necessity, but mostly out of curiosity. Given the results of the PCA on the depth sensor data, a reduced sensor setup was used to "generate" data for the remaining sensors. This was mostly done to get an intuitive understanding of how well the reduced sensor setup was able to represent the remaining sensors.

As seen in Figure 4.11, the results of the reconstruction were intriguing, but not necessarily immediately applicable. The reconstructed data was reasonably accurate over short time periods, with errors ranging from 0cm to 40cm. This served as a reassurance, showcasing the potential of the reduced sensor setup in representing the original data. Furthermore, this also implied that despite reducing the number of sensors, significant portions of the original information were still preserved.

While PCA-based reconstruction of data can be beneficial in specific scenarios, it's not appropriate in this thesis' context. For instance, one study suggests using PCA for reconstructing measurements of a faulty sensor from a set of other linearly correlated sensors [51]. However, this method requires high certainty in sensor correlation and is optimal when only a couple of sensors are removed. Here, we're removing over half the depth sensors, leading to significant reconstruction errors for extended periods.

Furthermore, the mentioned method requires data availability from all sensors that are to be reconstructed prior to the reconstruction. Recall that one first needs to run the PCA on a dataset where all the sensors are present in order to extract principal components with correct dimensions. Contrary to this requirement, this thesis aims for a permanently reduced sensor setup. If one wanted to reconstruct data for all the 16 depth sensors, these would all have to be installed on the net cage first. This would essentially completely contradict the goal of identifying a reduced setup in the first place, since all the sensors would still need to be installed on the net cage to enable their reconstruction.

Thus it becomes clear that such a reconstruction may not have immediate practical applications in the context of this thesis. Nonetheless, the process helps visualize the extent to which a reduced set of sensors can capture the measurements of the discarded sensors. It is also worth noting that while the reconstruction was not without errors, the range of these errors reaffirms the potential of PCA for effective sensor setup optimization. Furthermore, it raises interesting possibilities for future research, where further enhancements to the reconstruction method (or the investigation of other methods) may possibly result in more accurate approximations.

5.3 Implications of findings

The results of this study have significant implications for the fish farming industry, particularly regarding the use of sensor data for monitoring and decision-making purposes. The successful application of PCA on the depth sensor data suggests that it is possible to operate with a reduced set of sensors without losing significant information about the state of the fish farm. Specifically, sensors 10 and 2 emerged as the most important ones, closely followed by sensors 4, 5, 12, and 13. Such a reduction could potentially lead to lower costs and overall more effective operations.

The concept of reconstructing data from a reduced sensor setup, although academically interesting, seems to have limited practical application in the specific context of fish farming. The method tends to perform best when only a few sensors are removed. As shown in this thesis, removing more than half of the depth sensors can lead to relatively large errors, limiting the feasibility of data reconstruction. Furthermore, the necessity for data to be available from all sensors prior to commencing reconstruction starkly contradicts the aim of permanently operating with a reduced sensor setup. Thus, while reconstruction might be useful in specific scenarios, such as temporary sensor failure, it doesn't seem suitable for achieving long-term operation with fewer sensors.

However, the PCA did not perform well on all types of sensor data. In particular, the accelerometer and load shackle analyses did not result in any significant dimensionality reduction. This suggests that these types of sensors all provide unique and necessary information that cannot be captured by a subset of sensors, or that they contain non-linear data that can't be reduced through a PCA. This facilitates the need for further investigations.

Overall, these findings highlight both the potential and the limitations of PCA in processing sensor data from fish farms, offering valuable insights for future research in this area. Given the limited success of PCA with non-depth sensor data, other dimensionality reduction techniques might be explored in future studies.

5.4 Limitations

The PCA algorithm, which was extensively used in this thesis, makes certain assumptions that may not always hold true in real-world settings. As mentioned throughout this thesis, PCA operates on the premise that there are linear relationships between the variables in the dataset. Given the results, this appears to be a reasonable assumption for the depth sensors, but it may not hold true for the accelerometers and load shackles, which are influenced by the more complex and chaotic waves.

Even when the PCA is successful in reducing the dimensionality of a dataset, it is very hard to intuitively explain why some sensors are discarded while others are kept. The PCA's inability to effectively reduce the dimensionality of non-linear data or provide an explanation when it is successful has proved to be central limitation in this project.

Furthermore, these analyses have been somewhat constrained by the limited number of sensors

available. PCA's effectiveness often becomes more apparent when dealing with even higher dimensional data. While, in theory, it is completely possible to reduce a dataset consisting of 7 accelerometers or 5 load shackles, it would likely have been easier if more sensors were available in closer proximity to each other. This highlights another limitation in the application of the method.

Finally, it's important to acknowledge that these findings are specific to data gathered from Buholmen, and may not generalize to all scenarios or other fish farms. Other sea cage sizes or environmental conditions might provide entirely different results. In other words, sensor importance might depend on a variety of unaccounted factors. As explained in Chapter 1, the aquaculture industry in Norway has in recent years seen a trend where fish farms are being moved to more exposed location. It's completely conceivable that the "importance" of different sensors (as indicated by the analyses) could shift when moving to even more exposed locations, where weather patterns are known to be more extreme. Hence, care should be taken before extrapolating the findings given in this thesis to other situations.

In light of these limitations, future work could derive great benefit from exploring the use of other dimensionality reduction techniques that do not rely on the same assumptions as PCA, such as non-linear dimensionality reduction techniques (this shall be further discussed shortly). Furthermore, increasing the number of sensors, if possible, could offer a richer dataset for PCA and other similar algorithms to work with. Lastly, it's important to validate these findings with additional case studies to understand whether the results found in this thesis are generalizable to other fish farms.

5.5 Alternative approaches

In addition to the Principal Component Analysis (PCA), other dimensionality reduction techniques could be explored in future research.

If accelerometer and load shackle data indeed contain non-linear relationships, then techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) [52] and Uniform Manifold Approximation and Projection (UMAP) [53, 54] might be better suited for achieving dimensionality reduction. These have proven to be more effective on non-linear data in part due to their ability to preserve the local and global structure of data in high-dimensional spaces. These techniques could potentially be more successful than the PCA when it comes to reducing the dimensionality of accelerometer and load shackle data. As discussed, this is because the PCA is unable to capture the more complex non-linear relationships that might be present in these sensors.

Autoencoders, a type of artificial neural network, represent a machine learning approach to the dimensionality reduction problem [55]. In short, an autoencoder learns to compress data from the input layer into a hidden layer, before using this lower dimensional representation to reconstruct the full data in the output layer. The output layer is then compared to the input layer to train the weights in the network. The amount of neurons in the hidden layer can be gradually decreased until the dimensionality is low enough, or until the model's performance drops. The compression and decompression functions are learned in an end-to-end manner,

enabling the autoencoder to capture more complex relationships in the data. However, the values in the hidden layer are only a lower dimensional representation of the input data. This approach cannot be used directly to determine exactly which sensors should be kept/discarded in a reduced sensors setup.

In order to determine exactly which sensors should be kept in an optimal sensor setup, completely different approaches may also be employed. One such approach would be to apply various transforms to the dataset in an attempt to identify other underlying components. One might for example utilize a wavelet transform: an improved version of the well-known Fourier transform that allows for the analysis of non-stationary components of a signal [56]. Readers that are interested in an in-depth description of the wavelet transform and its application in various fields are referred to [57] and [58]. When applied to time-series data from sensors, wavelet transforms could potentially reveal underlying patterns, trends, and other components that might be otherwise difficult to see. If it turns out that many of the sensors are in essence capturing the same components, then this could provide significant assistance in determining which sensors can safely be discarded in order to obtain an optimal setup.

These are some of the alternative approaches that can be considered in future research to find the most effective ways of reducing dimensionality/uncovering optimal sensor setups, especially when it comes to the accelerometers and load shackles.

5.6 Future research

The research presented in this thesis represents an initial exploration into the dimensionality reduction of sensor data from a marine fish farm. While some of the findings are promising, it is clear that more can still be done to refine and expand upon this work.

One of the main areas that can be further explored is in refining the application of the PCA itself. While PCA provided insightful results with the depth sensor data, it was clear that it struggled with accelerometer and load shackle data. Further work is needed in determining whether this is due to these sensors containing non-linear relationships, or because each sensor really is equally important in the context of monitoring conditions at a fish farm.

Future research could also explore the use alternative dimensionality reduction techniques such as t-SNE or UMAP. Autoencoders or wavelet transforms might also help gain insight into accelerometer and load shackle data.

While the results of the depth sensor analyses provided lots of insight, future work could benefit greatly from having more data available, whether through quicker sampling rates or longer periods of data collection. This could provide great help in further supporting the findings of this thesis. One could also try entirely different methods for dimensionality reduction, such as those mentioned before, and compare the results.

Future work would certainly also benefit from a more consistent flow of load shackle data. Recall that these sensors only recorded data in sporadic 2 hour intervals. In the context of this thesis, this was not too big an issue, as the high sampling rate meant that there were

enough data points in each 2 hour interval. However, other approaches might require a more consistent data flow in order to yield useful results over longer periods of time.

Part of this thesis was aimed at investigating whether there was a link between severity of weather conditions and optimal sensor setup. The fact that the PCA only yielded useful results when run on depth sensor data from several days meant that no conclusions could be drawn regarding the effect of specific environmental conditions on the optimal sensor setup (for any of the sensor types). This aspect could also be further investigated, perhaps most simply by making more frequent depth readings, so that one has enough data to run the PCA on shorter segments where the conditions are reasonably static.

While the research presented in this thesis offers promising initial insights, it also indicates that various future work can be conducted for improving the optimization of sensor setups in fish farming. It is hoped that future work will build upon these findings in order to provide a better understanding of the optimal sensor setup.

Chapter 6

Conclusion

This thesis has explored the use of Principal Component Analysis (PCA) in reducing the dimensionality of sensor data from fish farms. The aim was to identify a reduced sensor setup that could provide sufficient information to enable decision-making without compromising the effective functioning of the sea cage. The research presented in this thesis has yielded valuable insights. As shown, there is both potential and limitations in the use of PCA as a dimensionality reduction tool in this specific context.

The analyses on depth sensor data showed highly promising results. The PCA showed that one can reduce the dimensionality without losing significant amounts of information, with certain sensors emerging as the most important. Specifically, the analysis *indicated* that sensors [1, 2, 4, 10, 12] or [2, 4, 5, 10, 12, 13] were the most important, depending on the way the PCA was applied to the data and how one interprets the results.

The cost-saving implications of these findings for the fish farming industry are significant as they could potentially lead to reductions in equipment, installation, maintenance, and data processing expenses. It was also found that the method for reconstructing data from a reduced sensor setup was not practically feasible for long-term operations, but could prove useful in specific scenarios like temporary sensor failure.

On the other hand, using PCA on accelerometer and load shackle data was less successful, indicating one of two things: that these types of sensors all provide unique and necessary information, or that they contain non-linear relationships. Both conclusions point to the necessity of using alternative approaches and different methods of dimensionality reduction to further examine these sensor types.

The PCA algorithm and its assumptions formed the basis of this thesis, which presented some limitations. As mentioned, the inability of the PCA to reduce the dimensionality of non-linear data was one of the main limitations. Another big limitation is that the PCA provides no intuitive explanation for the chosen reduction even when successful. Furthermore, the limited number of sensors available for reduction might have been a limitation. Despite these, the work undertaken provides a good starting point for further investigations into optimal sensor

setups in the fish farming industry.

Looking ahead, there is a good amount of opportunity for future research to expand upon this work. Refined applications of PCA, the exploration of alternative dimensionality reduction techniques, and better data availability can all form the basis of future studies. Future studies could also aim at uncovering the relationship between environmental conditions and reduced sensor setups, if such a relationship exists. This would also help understand whether the results found in this thesis are generalizable to other fish farms at more exposed locations. These points highlight the importance of additional case studies.

In conclusion, this thesis provides an insightful foundation upon which future work can be built. While some challenges have been identified, they do not completely hinder the potential of PCA and other similar techniques in optimizing sensor setups in fish farming. Although it is a complex problem, finding an optimal sensor setup in the aquaculture industry offers many promising paths to explore. This thesis has attempted to explore one of these paths, with the hope that many more will be explored in future works.

Bibliography

[1] *Fish to 2030: Prospects for Fisheries and Aquaculture - World | ReliefWeb*, en, Feb. 2014. [Online]. Available: https://reliefweb.int/report/world/fish-2030-prospects-fisheries-and-aquaculture (visited on 14/03/2023).

[2] *Norway's seafood exports worth NOK 151.4 billion in 2022*. [Online]. Available: https://en.seafood.no/news-and-media/news-archive/norways-seafood-exports-worth-nok-151.4-billion-in-2022/ (visited on 16/06/2023).

[3] T. Olafsen, U. Winter, Y. Olsen and J. Skjermo, *Value created from productive oceans in 2050*, en, Dec. 2012. [Online]. Available: https://www.sintef.no/en/latest-news/2012/value-created-from-productive-oceans-in-2050/ (visited on 13/03/2023).

[4] H. V. Bjelland, M. Føre, P. Lader, D. Kristiansen, I. M. Holmen, A. Fredheim, E. I. Grøtli, D. E. Fathi, F. Oppedal, I. B. Utne and I. Schjølberg, 'Exposed Aquaculture in Norway,' in *OCEANS 2015 - MTS/IEEE Washington*, Oct. 2015, pp. 1–10. DOI: 10.23919/OCEANS.2015.7404486.

[5] M. Holmer, 'Environmental issues of fish farming in offshore waters: Perspectives, concerns and research needs,' en, *Aquaculture Environment Interactions*, vol. 1, no. 1, pp. 57–70, Aug. 2010, ISSN: 1869-215X, 1869-7534. DOI: 10.3354/aei00007. [Online]. Available: http://www.int-res.com/abstracts/aei/v1/n1/p57-70/ (visited on 14/03/2023).

[6] *Celebrating 50 years of modern aquaculture*, en, Mar. 2023. [Online]. Available: https://en.seafood.no/news-and-media/news-archive/celebrating-50-years-of-modern-aquaculture/ (visited on 13/03/2023).

[7] *Aquaculture Act*, en. [Online]. Available: https://www.fiskeridir.no/English/Aquaculture/Aquaculture-Act (visited on 21/06/2023).

[8] O. Lekve, *Norwegian aquaculture*, en, Mar. 2012. [Online]. Available: https://www.barentswatch.no/en/articles/norwegian-aquaculture/ (visited on 13/03/2023).

[9] Ø. Jensen, T. Dempster, E. B. Thorstad, I. Uglem and A. Fredheim, 'Escapes of fishes from Norwegian sea-cage aquaculture: Causes, consequences and prevention,' en, *Aquaculture Environment Interactions*, vol. 1, no. 1, pp. 71–83, Aug. 2010, ISSN: 1869-215X, 1869-7534. DOI: 10.3354/aei00008. [Online]. Available: https://www.int-res.com/abstracts/aei/v1/n1/p71-83/ (visited on 16/06/2023).

[10] H. M. Føre and T. Thorvaldsen, 'Causal analysis of escape of Atlantic salmon and rainbow trout from Norwegian fish farms during 2010–2018,' en, *Aquaculture*, vol. 532, p. 736 002, Feb. 2021, ISSN: 0044-8486. DOI: 10.1016/j.aquaculture.2020.736002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0044848620315684 (visited on 16/06/2023).

[11] M. J. Costello, 'How sea lice from salmon farms may cause wild salmonid declines in Europe and North America and be a threat to fishes elsewhere,' *Proceedings of the Royal Society B: Biological Sciences*, vol. 276, no. 1672, pp. 3385–3394, Jul. 2009, Publisher: Royal Society. DOI: 10.1098/rspb.2009.0771. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rspb.2009.0771 (visited on 14/03/2023).

[12] L. T. Barrett, F. Oppedal, N. Robinson and T. Dempster, 'Prevention not cure: A review of methods to avoid sea lice infestations in salmon aquaculture,' en, *Reviews in Aquaculture*, vol. 12, no. 4, pp. 2527–2543, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/raq.12456, ISSN: 1753-5131. DOI: 10.1111/raq.12456. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/raq.12456 (visited on 16/06/2023).

[13] P. Klebert, Ø. Patursson, P. C. Endresen, P. Rundtop, J. Birkevold and H. W. Rasmussen, 'Three-dimensional deformation of a large circular flexible sea cage in high currents: Field experiment and modeling,' en, *Ocean Engineering*, vol. 104, pp. 511–520, Aug. 2015, ISSN: 0029-8018. DOI: 10.1016/j.oceaneng.2015.04.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801815001262 (visited on 13/03/2023).

[14] *ABB enables first remote-controlled submersible fish farm in the Arctic Ocean*, en, May 2019. [Online]. Available: https://new.abb.com/news/detail/24385/abb-enables-first-remote-controlled-submersible-fish-farm-in-the-arctic-ocean (visited on 21/06/2023).

[15] S. J. Ohrem, E. Kelasidi and N. Bloecher, 'Analysis of a novel autonomous underwater robot for biofouling prevention and inspection in fish farms,' in *2020 28th Mediterranean Conference on Control and Automation (MED)*, ISSN: 2473-3504, Sep. 2020, pp. 1002–1008. DOI: 10.1109/MED48518.2020.9183157.

[16] D. W. Fredriksson, M. R. Swift, J. D. Irish, I. Tsukrov and B. Celikkol, 'Fish cage and mooring system dynamics using physical and numerical models with field measurements,' en, *Aquacultural Engineering*, vol. 27, no. 2, pp. 117–146, Feb. 2003, ISSN: 0144-8609. DOI: 10.1016/S0144-8609(02)00043-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144860902000432 (visited on 18/06/2023).

[17] P. Lader, T. Dempster, A. Fredheim and Ø. Jensen, 'Current induced net deformations in full-scale sea-cages for Atlantic salmon (Salmo salar),' en, *Aquacultural Engineering*, vol. 38, no. 1, pp. 52–65, Jan. 2008, ISSN: 0144-8609. DOI: 10.1016/j.aquaeng.2007.11.001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144860907000957 (visited on 18/06/2023).

[18] J. DeCew, D. W. Fredriksson, P. F. Lader, M. Chambers, W. H. Howell, M. Osienki, B. Celikkol, K. Frank and E. Høy, 'Field measurements of cage deformation using acoustic sensors,' en, *Aquacultural Engineering*, vol. 57, pp. 114–125, Nov. 2013, ISSN: 0144-8609. DOI: 10.1016/j.aquaeng.2013.09.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0144860913000903 (visited on 18/06/2023).

[19] G. Løland, *Current forces on and flow through fish farms*. Institutt for Marin Hydrodynamikk, 1991.

[20] P. C. Endresen, M. Føre, A. Fredheim, D. Kristiansen and B. Enerhaug, 'Numerical Modeling of Wake Effect on Aquaculture Nets,' en, American Society of Mechanical Engineers Digital Collection, Nov. 2013. DOI: 10.1115/OMAE2013-11446. [Online]. Available: https://asmedigitalcollection.asme.org/OMAE/proceedings-abstract/OMAE2013/55355/270685 (visited on 13/03/2023).

[21] H. Moe-Føre, P. Christian Endresen, K. Gunnar Aarsæther, J. Jensen, M. Føre, D. Kristiansen, A. Fredheim, P. Lader and K.-J. Reite, 'Structural Analysis of Aquaculture Nets: Comparison and Validation of Different Numerical Modeling Approaches,' *Journal of Offshore Mechanics and Arctic Engineering*, vol. 137, no. 4, Aug. 2015, ISSN: 0892-7219. DOI: 10.1115/1.4030255. [Online]. Available: https://doi.org/10.1115/1.4030255 (visited on 13/03/2023).

[22] K.-J. Reite, M. Føre, K. G. Aarsæther, J. Jensen, P. Rundtop, L. T. Kyllingstad, P. C. Endresen, D. Kristiansen, V. Johansen and A. Fredheim, 'FHSIM — Time Domain Simulation of Marine Systems,' en, American Society of Mechanical Engineers Digital Collection, Oct. 2014. DOI: 10.1115/OMAE2014-23165. [Online]. Available: https://asmedigitalcollection.asme.org/OMAE/proceedings-abstract/OMAE2014/45509/279039 (visited on 13/03/2023).

[23] B. Su, K.-J. Reite, M. Føre, K. G. Aarsæther, M. O. Alver, P. C. Endresen, D. Kristiansen, J. Haugen, W. Caharija and A. Tsarau, 'A Multipurpose Framework for Modelling and Simulation of Marine Aquaculture Systems,' en, American Society of Mechanical Engineers Digital Collection, Nov. 2019. DOI: 10.1115/OMAE2019-95414. [Online]. Available: https://asmedigitalcollection.asme.org/OMAE/proceedings-abstract/OMAE2019/58837/1067839 (visited on 13/03/2023).

[24] P. C. Endresen and P. Klebert, 'Loads and response on flexible conical and cylindrical fish cages: A numerical and experimental study based on full-scale values,' en, *Ocean Engineering*, vol. 216, p. 107 672, Nov. 2020, ISSN: 0029-8018. DOI: 10.1016/j.oceaneng.2020.107672. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801820306661 (visited on 13/03/2023).

[25] B. Su, E. Kelasidi, K. Frank, J. Haugen, M. Føre and M. O. Pedersen, 'An integrated approach for monitoring structural deformation of aquaculture net cages,' en, *Ocean Engineering*, vol. 219, p. 108 424, Jan. 2021, ISSN: 0029-8018. DOI: 10.1016/j.oceaneng.2020.108424. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801820313317 (visited on 13/03/2023).

[26] J. Turnbull, A. Bell, C. Adams, J. Bron and F. Huntingford, 'Stocking density and welfare of cage farmed Atlantic salmon: Application of a multivariate analysis,' en, *Aquaculture*, vol. 243, no. 1, pp. 121–132, Jan. 2005, ISSN: 0044-8486. DOI: 10.1016/j.aquaculture.2004.09.022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0044848604005538 (visited on 17/06/2023).

[27] Z. Yang, Y. Chen and J. Corander, *T-SNE Is Not Optimized to Reveal Clusters in Data*, arXiv:2110.02573 [cs, stat], Oct. 2021. DOI: 10.48550/arXiv.2110.02573. [Online]. Available: http://arxiv.org/abs/2110.02573 (visited on 18/06/2023).

[28] M. Wattenberg, F. Viégas and I. Johnson, 'How to Use t-SNE Effectively,' en, *Distill*, vol. 1, no. 10, e2, Oct. 2016, ISSN: 2476-0757. DOI: 10.23915/distill.00002. [Online]. Available: http://distill.pub/2016/misread-tsne (visited on 18/06/2023).

[29] K. Pearson, 'LIII. On lines and planes of closest fit to systems of points in space,' *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901, Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720, ISSN: 1941-5982. DOI: 10.1080/14786440109462720. [Online]. Available: https://doi.org/10.1080/14786440109462720 (visited on 28/04/2023).

[30] H. Hotelling, 'Analysis of a complex of statistical variables into principal components,' *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933, Place: US Publisher: Warwick & York, ISSN: 1939-2176. DOI: 10.1037/h0071325.

[31] I. T. Jolliffe and J. Cadima, 'Principal component analysis: A review and recent developments,' *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20 150 202, Apr. 2016, Publisher: Royal Society. DOI: 10.1098/rsta.2015.0202. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202 (visited on 28/04/2023).

[32] S. Wold, K. Esbensen and P. Geladi, 'Principal component analysis,' en, *Chemometrics and Intelligent Laboratory Systems*, Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists, vol. 2, no. 1, pp. 37–52, Aug. 1987, ISSN: 0169-7439. DOI: 10.1016/0169-7439(87)80084-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0169743987800849 (visited on 28/04/2023).

[33] J. Shlens, *A Tutorial on Principal Component Analysis*, arXiv:1404.1100 [cs, stat], Apr. 2014. DOI: 10.48550/arXiv.1404.1100. [Online]. Available: http://arxiv.org/abs/1404.1100 (visited on 09/06/2023).

[34] H. Abdi and L. J. Williams, 'Principal component analysis,' en, *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, 2010, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.101, ISSN: 1939-0068. DOI: 10.1002/wics.101. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/wics.101 (visited on 28/04/2023).

[35] J. Osborne and A. Costello, 'Sample size and subject to item ratio in principal components analysis,' *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, Nov. 2019, ISSN: 1531-7714. DOI: https://doi.org/10.7275/ktzq-jq66. [Online]. Available: https://scholarworks.umass.edu/pare/vol9/iss1/11.

[36] S. S. Shaukat, T. A. Rao and M. A. Khan, 'Impact of sample size on principal component analysis ordination of an environmental data set: Effects on eigenstructure,' en, *Ekológia (Bratislava)*, vol. 35, no. 2, pp. 173–190, Jun. 2016. DOI: 10.1515/eko-2016-0014. [Online]. Available: https://sciendo.com/article/10.1515/eko-2016-0014 (visited on 28/05/2023).

[37] M. Björklund, 'Be careful with your principal components,' *Evolution*, vol. 73, no. 10, pp. 2151–2158, Oct. 2019, ISSN: 0014-3820. DOI: 10.1111/evo.13835. [Online]. Available: https://doi.org/10.1111/evo.13835 (visited on 29/05/2023).

[38] P. Podder, M. M. Hasan, M. R. Islam and M. Sayeed, *Design and Implementation of Butterworth, Chebyshev-I and Elliptic Filter for Speech Signal Analysis*, en, Feb. 2020. DOI: 10.5120/17195-7390. [Online]. Available: https://arxiv.org/abs/2002.03130v2 (visited on 19/06/2023).

[39] S. Butterworth *et al.*, 'On the theory of filter amplifiers,' *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930. [Online]. Available: https://www.changpuak.ch/electronics/downloads/On_the_Theory_of_Filter_Amplifiers.pdf.

[40] S. K. Jagtap and M. D. Uplane, 'The impact of digital filtering to ECG analysis: Butterworth filter application,' in *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, Oct. 2012, pp. 1–6. DOI: 10.1109/ICCICT.2012.6398145.

[41] R. G. T. Mello, L. F. Oliveira and J. Nadal, 'Digital Butterworth filter for subtracting noise from low magnitude surface electromyogram,' en, *Computer Methods and Programs in Biomedicine*, vol. 87, no. 1, pp. 28–35, Jul. 2007, ISSN: 0169-2607. DOI: 10.1016/j.cmpb.2007.04.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260707000983 (visited on 19/06/2023).

[42] J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications, Fifth Edition*. Dian zi gong ye chu ban she, 2022, ISBN: 978-7-121-43981-0. [Online]. Available: https://books.google.no/books?id=9eN6zwEACAAJ.

[43] S. Pond and G. Pickard, *Introductory Dynamical Oceanography* (Pergamon international library of science, technology, engineering, and social studies). Elsevier Science, 1983, ISBN: 978-0-7506-2496-1. [Online]. Available: https://books.google.no/books?id=5pQf8dBYxIUC.

[44] M. Tomczak and J. S. Godfrey, 'Regional Oceanography: An Introduction,' 1994.

[45] R. H. Stewart, *Introduction to physical oceanography*, en. Robert H. Stewart, 2008, Accepted: 2017-04-10T21:04:33Z Artwork Medium: Electronic Interview Medium: Electronic. [Online]. Available: https://oaktrust.library.tamu.edu/handle/1969.1/160216 (visited on 20/06/2023).

[46] K. A. Orvik, 'Long-Term Moored Current and Temperature Measurements of the Atlantic Inflow Into the Nordic Seas in the Norwegian Atlantic Current; 1995–2020,' en, *Geophysical Research Letters*, vol. 49, no. 3, e2021GL096427, 2022, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL096427, ISSN: 1944-8007. DOI: 10.1029/2021GL096427. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL096427 (visited on 20/06/2023).

[47] B. Kinsman, *Wind Waves: Their Generation and Propagation on the Ocean Surface*, en. Courier Corporation, Jan. 1984, ISBN: 978-0-486-64652-7.

[48] L. H. Holthuijsen, *Waves in Oceanic and Coastal Waters*, en. Cambridge University Press, Feb. 2010, Google-Books-ID: 7tFUL2blHdoC, ISBN: 978-1-139-46252-5.

[49] H. Wang, G. Li, Z. Ma and X. Li, 'Image recognition of plant diseases based on principal component analysis and neural networks,' in *2012 8th International Conference on Natural Computation*, ISSN: 2157-9563, May 2012, pp. 246–251. DOI: 10.1109/ICNC.2012.6234701.

[50] F. Song, Z. Guo and D. Mei, 'Feature Selection Using Principal Component Analysis,' in *Engineering Design and Manufacturing Informatization 2010 International Conference on System Science*, vol. 1, Nov. 2010, pp. 27–30. DOI: 10.1109/ICSEM.2010.14.

[51] R. Dunia, S. J. Qin, T. F. Edgar and T. J. McAvoy, 'Identification of faulty sensors using principal component analysis,' en, *AIChE Journal*, vol. 42, no. 10, pp. 2797–2812, 1996, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690421011, ISSN: 1547-5905. DOI: 10.1002/aic.690421011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690421011 (visited on 01/06/2023).

[52] L. v. d. Maaten and G. Hinton, 'Visualizing Data using t-SNE,' *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008, ISSN: 1533-7928. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html (visited on 01/06/2023).

[53] L. McInnes, J. Healy and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv:1802.03426 [cs, stat], Sep. 2020. DOI: 10.48550/arXiv.1802.03426. [Online]. Available: http://arxiv.org/abs/1802.03426 (visited on 01/06/2023).

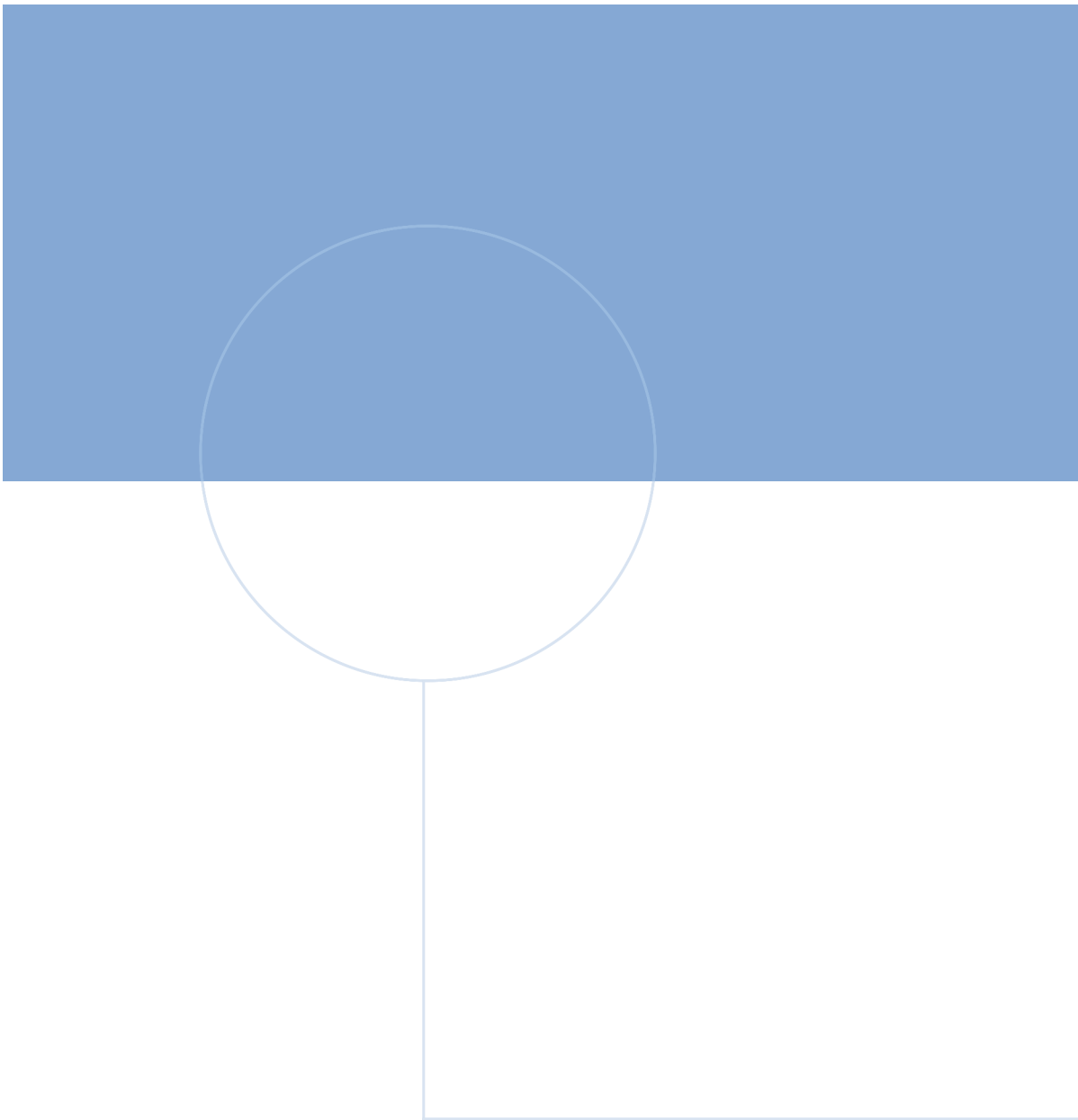
[54] T. Sainburg, L. McInnes and T. Q. Gentner, *Parametric UMAP embeddings for representation and semi-supervised learning*, arXiv:2009.12981 [cs, q-bio, stat], Aug. 2021. DOI: 10.48550/arXiv.2009.12981. [Online]. Available: http://arxiv.org/abs/2009.12981 (visited on 01/06/2023).

[55] G. E. Hinton and R. R. Salakhutdinov, 'Reducing the dimensionality of data with neural networks,' eng, *Science (New York, N.Y.)*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, ISSN: 1095-9203. DOI: 10.1126/science.1127647.

[56] M. Sifuzzaman, M. R. Islam and M. Z. Ali, 'Application of Wavelet Transform and its Advantages Compared to Fourier Transform,' en, 2009, Accepted: 2016-12-22T17:15:58Z Publisher: Vidyasagar University , Midnapore , West-Bengal , India, ISSN: 0972-8791 (Print). [Online]. Available: http://inet.vidyasagar.ac.in:8080/jspui/handle/123456789/779 (visited on 01/06/2023).

[57] P. S. Addison, *The Illustrated Wavelet Transform Handbook: Introductory Theory and Applications in Science, Engineering, Medicine and Finance, Second Edition*, en. CRC Press, Jan. 2017, Google-Books-ID: wBoNDgAAQBAJ, ISBN: 978-1-4822-5133-3.

[58] A. N. Akansu and R. A. Haddad, 'Chapter 6 - Wavelet Transform,' en, in *Multiresolution Signal Decomposition (Second Edition)*, A. N. Akansu and R. A. Haddad, Eds., San Diego: Academic Press, Jan. 2001, pp. 391–442, ISBN: 978-0-12-047141-6. DOI: 10.1016/B978-012047141-6/50006-9. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780120471416500069 (visited on 20/06/2023).



NTNU
Kunnskap for en bedre verden

SINTEF