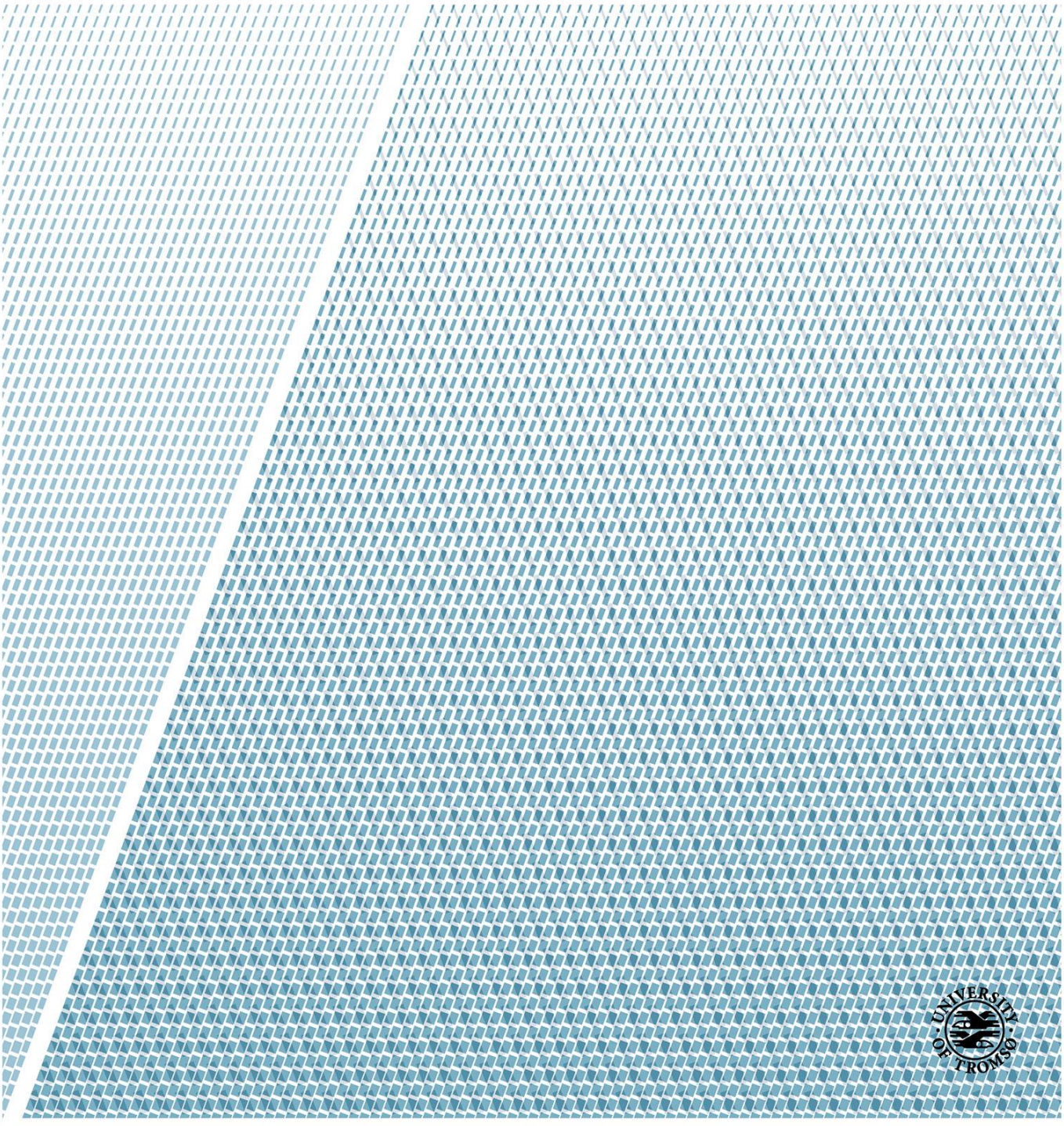
UiT
THE ARCTIC UNIVERSITY OF NORWAY

Department of Mathematics and Statistics

Separation of fish stocks by otoliths: Image representation, Fourier approximation and discrimination

Reidar Strand Hagen

STA-3900 Master's Thesis in Statistics...May 2015



Preface

This thesis was written as part of a Master in Statitics at UIT, set to be finished in spring 2015. Georg Elvebakk at UIT was the main advisor, with Alf Harbitz from the Marine Research insitute in Tromsø as co-advisor.

In writing this thesis, I have assumed the reader has strong general knowledge of statistics. A general interest in biology and computer science would be an advantage when reading this thesis, although the necessary knowledge is presented in the second chapter.

Tromsø, 2015-05-15

Reidar Strand Hagen

Acknowledgment

I would first and foremost thank Georg Elvebakk and Alf Harbitz for their excellent advice and support.

I would also like to thank Melanie Simpson for her feedback and help when writing up the results. I certainly enjoy solving problems much more than I do writing about them. Furthermore, I would like thank Helge Johansen for excellent clerical support and helping me navigate the strange world of academic bureaucracy.

R.S.H.

Summary and Conclusions

Otoliths are calcium structures found in the balance organs of all vertebrates. Their shape is dependant both on genetics and environment. For this reason analysis of their contours through Fourier Contour Analysis has become a well-known technique for separating between different stocks of fish.

In this thesis the entire process of Fourier Contour Analysis have been investigated for potential sources of bias. It has been found that specifics of how one aquires the contours from an image, such as colour and image representation, edge-traversal and image formats are largely irrelevant for the final contour created.

A probability-based approach has been proposed to substitute the stratification-based approach to resampling previously used when constructing comparable sets. This may increase the power of the comparison. Various transformations and discriminant analysis approaches have been tested, but no obvious improvements were found.

It has however been shown that numerically solving the problem of aligning otoliths , instead of aligning by the first harmonic contour, lead to better cross-validated discrimination rates. Furthermore, using the absolute values of the fourier coefficients also increased cross-validated discrimination rate. Both of these findings require further testing and work to determine whether they hold for fourier contour analysis in general or just these sets of data.

Contents

Chapter 1

Introduction

This thesis was chosen because the author finds the complexities and challenges that arise when different fields of study interact interesting and challenging. Finding the shape approximation with lowest error rate for fewest variables is interesting enough, but finding the shape approximation which best picks up on systematic biological differences in a way that makes it possible to discern using different cross-validated discrimination techniques is just plain fun.

1.1 Objectives

The overall goal of this thesis is to improve on the method of separating stocks of fish using otolith contours. In order to achieve the following objectives has been set:

1. Investigate potential sources of bias when aquiring the contour
2. Investigate to which degree true separation between the two stocks of Halibuth is possible and estimate accuracy of predictions
3. Explore different techniques for improving the basic method of Fourier Contour Analysis and assess their overall effect

1.2 Data

1.2.1 Halibuth

Our primary dataset consists of two sets of samples from the same species of halibut from two different breeding grounds. This dataset includes broken and otherwise unusable otoliths, and includes both left and right otoliths (figure 1.1).

A list of manually verified otoliths, aswell as the sex, weight and length of the Halibuths they were harvested from was provided by Alf Harbitz.

Dataset	Location	Broken	Valid	Total Samples
A_e	Eggakanten	343	828	1171
A_g	Greenland	78	83	161
A		411	921	1332

	A_e ($n = 828$)		A_g ($n = 83$)		A ($n = 921$)	
Attribute	mean	sd	mean	sd	mean	sd
Weight	1365	1161	1257	495	1355	1118
Length	49.9	11.7	50.8	5.26	49.98	11.3
Sex	0.41		0.73		0.44	

1.2.2 Cod

The second dataset consists of 1177 cod otoliths (example otolith in figure 1.2). This set contained no broken or otherwise unusable otoliths. Meta-information on catch-location, weight, length, sex, age aswell as other characteristics were made available by Alf Harbitz. This is the same dataset used in Stransky et al. (2008) and Henriksen (2013).

Cod dataset summary

Area	Code	Number of fish	Mean length	Mean age
Svalbard	SVA	32	33.7 (±3.9)	3.1 (±0.3)
Barents Sea	BAR	408	52.6 (±11.6)	4.9(±1.1)
Varanger	VAR	108	51.0 (±11.2)	5.4 (±2.1)
Nordkapp	NOK	152	46.6 (±7.1)	4.1 (±0.9)
Porsanger	POR	49	58.0 (±5.1)	6.8 (±2.2)
Balsfjorden	BAL	309	45.5 (±7.7)	5.3 (±1.5)
Vestfjorden, West	VEW	48	60.7 (±7.1)	5.0 (±0.9)
Vestfjorden, East	VEE	71	57.8 (±3.8)	4.3 (±0.9)
Total		1177	50.5 (±10.3)	4.9 (±1.5)

Density plots of length and weight have been included in figures 1.3 and 1.4. It should be clear that there are systematic differences between the groups of fish caught from different locations. This will be discussed further in chapter 6.

1.3 Limitations

The majority of this study has been limited to the dataset of halibut otoliths. However, late in the thesis the set of cod otoliths was made available. This was primeraly used for objective 3, as sufficient separation was not found in the Halibuth set to be able to compare different methodologies properly.

Furthermore, the statistic and computer science part of this process has received more attention than the biological basis for this analysis. The datasets and meta-information has been accepted as they are. Both of these limitations have been necessary to reduce the scope of this thesis.

1.4 Structure of the Report

Theory and previous work has been split into two chapters. Chapter 2 contains introductions to needed subjects in biology and computer science, while chapter 3 contains needed theory in

mathematics and statitics.

Chapter 4 focuses potential biases and problems when extracting a contour from an image. Chapter 5 focuses on whether the sets are comparable, while chapter 6 focuses on the final discriminant analysis.

Results and suggestions for further work is discussed in the final chapter.

Illustrations

Figure 1.1: Pair of halibut otoliths

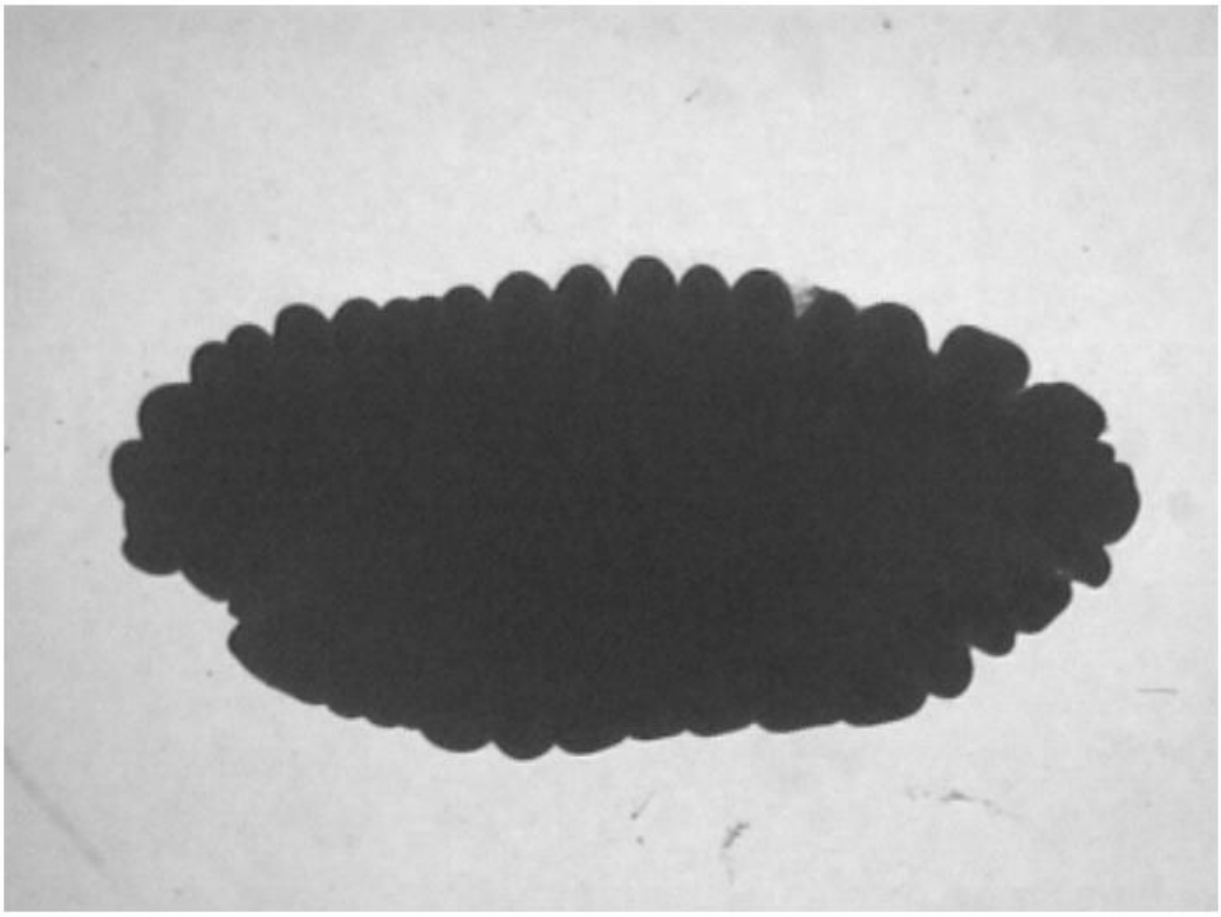


Figure 1.2: Left cod otolith

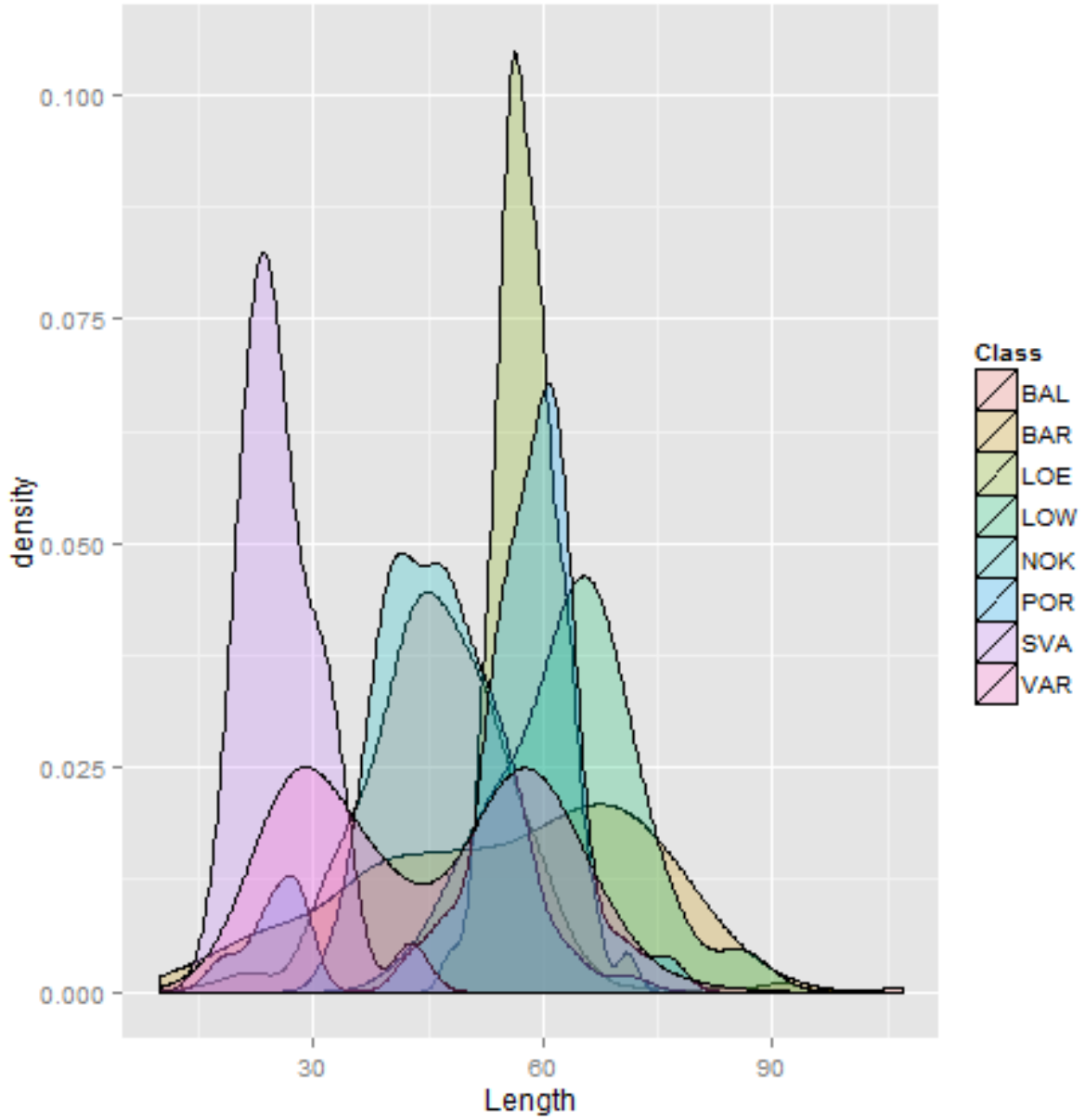


Figure 1.3: Density plot of length (in cm) of cod in dataset, grouped by location of the catch. It is clear that these groups have systematically different properties.

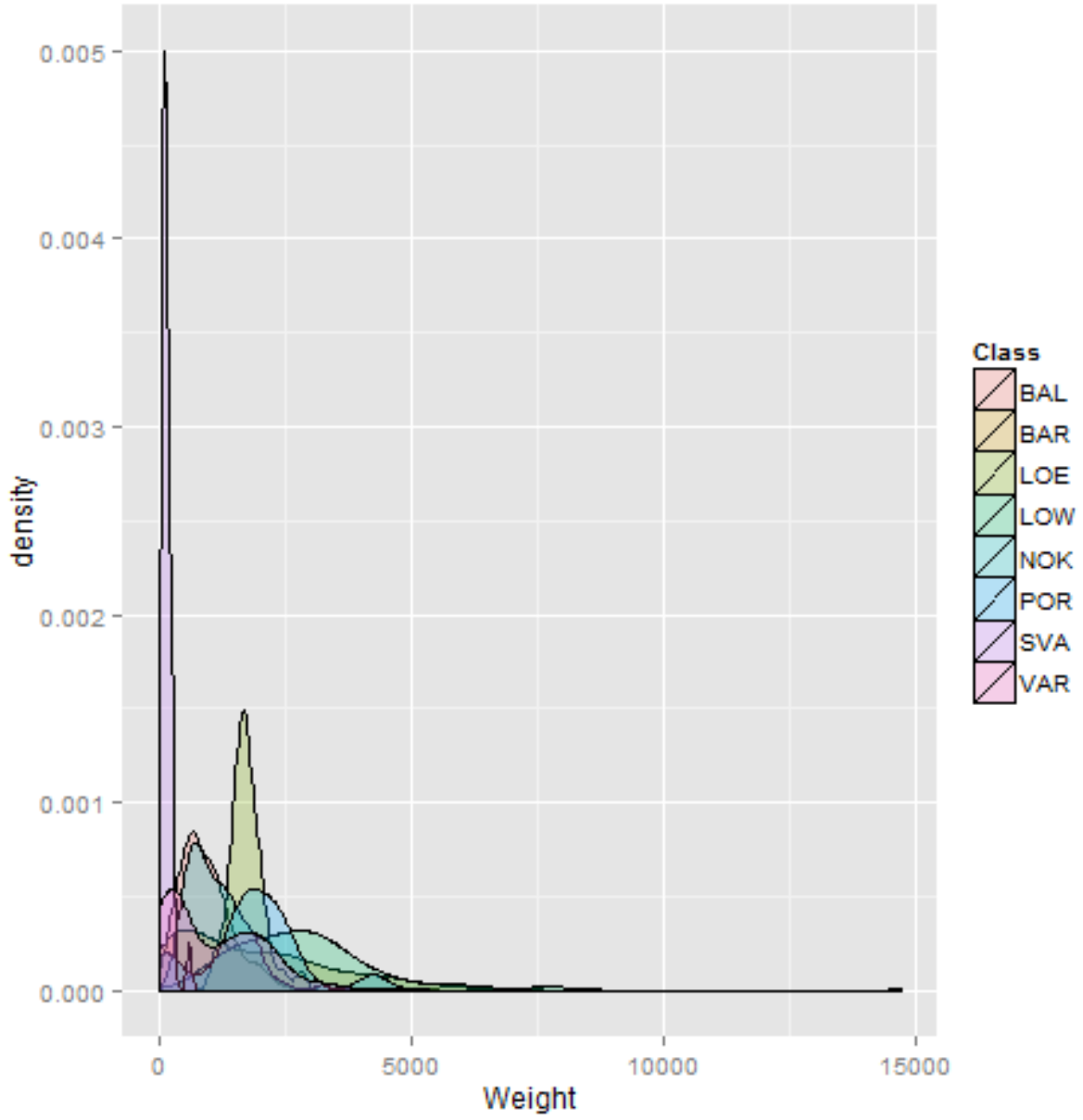


Figure 1.4: Density plot of weight (in kg) of cod in dataset, grouped by location of catch

Chapter 2

Background

This chapter contains introductions to subjects in biology and computer science relevant for this thesis. The section *Fishery Management* gives background on why analysis of fish otoliths is important. *Otolith Organs* gives a brief explanation on basic fish biology and otoliths.

The section *Otolith Imaging* explains the process of aquiring images of the otoliths, while *Image Representation* gives background on how colors in images are measured, stored and transformed. The difficulties and ambiguities in this process is why it is important to verify that systematic biases are not introduced

Lastly, the final section *Countor Tracing* outlines the algorithm used for tracing the countor and finding a valid starting point.

2.1 Fishery Management

Globally fish and fishery products intake accounts for an estimated 6.5% of human protein intake and 16.7% of human animalistic protein intake (FAO, 2012). Furthermore, for 2012 an estimated 36.7% of fishery products were traded internationally for an estimated value of 129 billions US$. This puts the overall value of total fish catch above 350 billion US$. The fishing industry is thus globally important both nutrionally and economically.

However, according to The World Bank and FAO[1] (Sun, 2009) in 2004 more than 75% of fish stocks were underperforming, as in producing lower long term yields than an optimal long term

[1] Food and Agriculture Organization of the United Nations

strategy would. The economic loss for this underperformance for 2004 was estimated to 50 billion US$. Given that the value of international fishing trade has increased by roughly 7% annually a very rough estimate for 2012 puts this value at 85$ billion US$, which entails a lost 25% increase in value of production.

Fishery management is however a very complicated subject. Fish generally, at least before they are caught, swim freely around the oceans showing little regard for internationally agreed exclusive economic zone boundaries. Fishery management is thus not only about finding an overall sustainable fishing policy, but also about how responsibility should be shared among nations. Implementing a sustainable fishing policy is a tradeoff between individual short term gain with a larger shared long term gain. Dependable and neutral information on stocks and migration patterns are thus an important building block when crafting co-operation between nations.

However, one does not only need knowledge of general population dynamics and migration patterns of fish stocks. There are also very specific challenges. A well known Norwegian example is the difference between the North East Arctic and Norwegian Coastal Cod in northern Norway (Kålås et al., 2006). The arctic cod has its nursery and feeding area in the Barents sea, and migrates each year to the coast of Norway for spawning. The coastal cod however stays along the coast of Norway all year. The latter fish stock is on IUCN (International Union for Conservation of Nature) red list as *near-threatened*, while the former is bountifull. Since they overlap during spawning season one would effectively need to stop all catch of arctic cod in order to protect the coastal cod. However this would have a huge impact on the economics of Norwegian fisheries, so fishery management have come up with other local regulations to reduce the pressure on the coastal cod stock. In order to asses the effectiveness of these measures dependable and not prohibitly expensive means of identification are necessary.

A vast array of methods and tools exist for collecting information on the health of fish stocks, however this thesis will consider only variations of otolith contour analysis.

2.2 Otolith organs

The saccule and utricle make up the otolith organ, which is present in all vertebrates (fish, amphibians, reptiles, mammals and birds). The otolith organ provides gravity, balance, movement and directional indicators in all vertebrates, and have a secondary function in sound detection in higher aquatic and terrestrial vertebrates (Popper et al., 2003). The statoconia is a combination of gelatinous matrix and calcium carbonate structures, located within either the saccule or the utricle (seen fig 2.2). The term *otolith* is used interchangingly with stataconia, but is often used to reference the calcified structure itself.

The number of statoconium vary between species, however osteichthyes species (bony fish) have three pairs of statoconium, of which the largest and the one which is used in this thesis is the *sagitta*. The calcium carbonate, which the otoliths is composed of, is derived primarily from water (Oth, 2011) and is thus dependant on both the rate of growth and water conditions. The study of otoliths can thus give information on which bodies of water a specimen has previously occupied. The most studied trace signal used is the amount of strontium (Farrell and Campana, 1996), however many other techniques can be used.

Furthermore, the growth and shape of these otoliths are dependant on both genetic and environmental influences (A.H. Weatherley, 1987). Analyzing the countor of the otolith to discriminate between different stocks and interspecies has been used successfully on a multitude of different bony fish (Parisi-Baradad et al., 2010). While using the contour for identification most likely will never be as accurate as more expensive methods, it is cheap, easy and does not require very advanced laboratory equipment making it a much more applicable method of identification.

2.3 Otolith Imaging

The sagitta is removed manually with a very sharp knife, first cutting of the skull top to reveal the dorsal part of the brain. After removing the medulla oblongata the sagitta will be exposed in the depressions in the ventrolateral sides of the braincase Oth (2011).

Furthermore, the sagitta must be sliced cross-section using a low-speed precision saw (Morales-Nin). An image is aquired using an appriopriate photographic microscope. Otoliths are however

small and somewhat brittle, so they may break if one is not carefull. A skilled operator may thus produce sets of otolith systematically different from an unskilled one.

This procedure is usually completed in controlled lab-environments, however there are no set procedures on exactly which imaging-equipment and how backlights should be set up. There may thus be significant differences between images from different laboratories.

2.4 Image Representation

The biggest challenge to accurate image representation is that colour is not an actual property of a surface. A surface has a reflection profile, which determines the proportion of different wavelengths of light which is reflected. The wavelength profile of reflected light is dependant on the original light source, which is never completely the same, except under exceptionally controlled environments.

The colour representation systems in computer science are however designed to maximize performance relative to our eyes Omer and Werman (2004). Human perception of light is 3-dimensional, with intensities measured using the 3 different wavelength profiles as shown in fig 2.4. Using one degree of freedom for colour intensity, this leaves two degrees of freedom for the colour space we can perceive out of a wide multispectral profile. Camera setups usually measure 3 distinct wavelengths and approximates these into different colorformats. The consequence of this is that different cameras will register subtly different colors as they will rely on different transformations from multispectral wavelengths to 3-dimensional color.

In practice color information in images is fairly consistant and generally trustworthy, however it is important to note that even following best practices this a field where one will still only get a best effort result. It is therefore worthwhile to investigate to which degree any analytic method is dependant on the colours reported.

2.5 Aquiring the contour

Finding the shape of the otolith was done by first converting the image to grayscale. Secondly a threshold was used on the intensity values to separate between otolith and background as in 2.1.

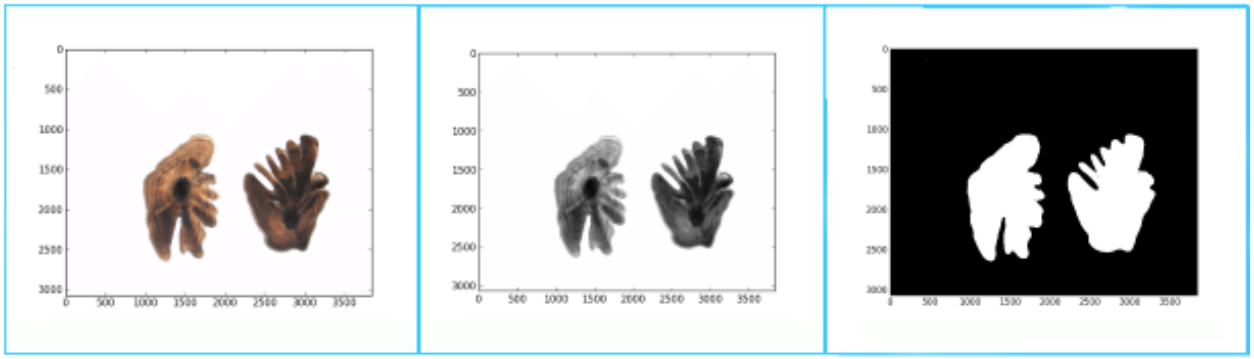


Figure 2.1:]
Thresholding illustration

The *two-pass algorithm* (Stockman and Shapiro, 2001, pp. 69–73.) was used to find the number of pieces of significant size. Their centers of mass were also found by averaging their pixel-coordinates. Images with anything other than two significant pieces were marked as invalid.

Tracing the contour of these shapes was done using the Moore Neighborhood algorithm with a modified exit criteria (Toussaint). The essence of this algorithm is to trace the contour through a series of moves. After each movement, current direction is changed -90°, and then again +45°until a valid pixel to move to is found. Given a valid starting point and direction this algorithm is guaranteed to converge. However, since a pixel may be visited several times using this technique, it is not a unique identifier for a position along the path. An edge can however only be visited once, and can be used to identify whether the contour is complete.

Potential starting positions from which to start tracing the contour were found by starting at the center of mass for each piece, and marking any edges found between center of mass and nearest edge of image. This is needed because holes in the otolith or image artifacts may create starting points that do not lead to a path around the otolith (illustrated in fig 2.5)). Creating contours from all possible starting points and keeping the longest solves this problem.

In the rest of this thesis *contour* will refer to the outline of the shape in (x_n, y_n)-space. *Path* will refer to x_n and y_n respectively.

Illustrations for chapter2

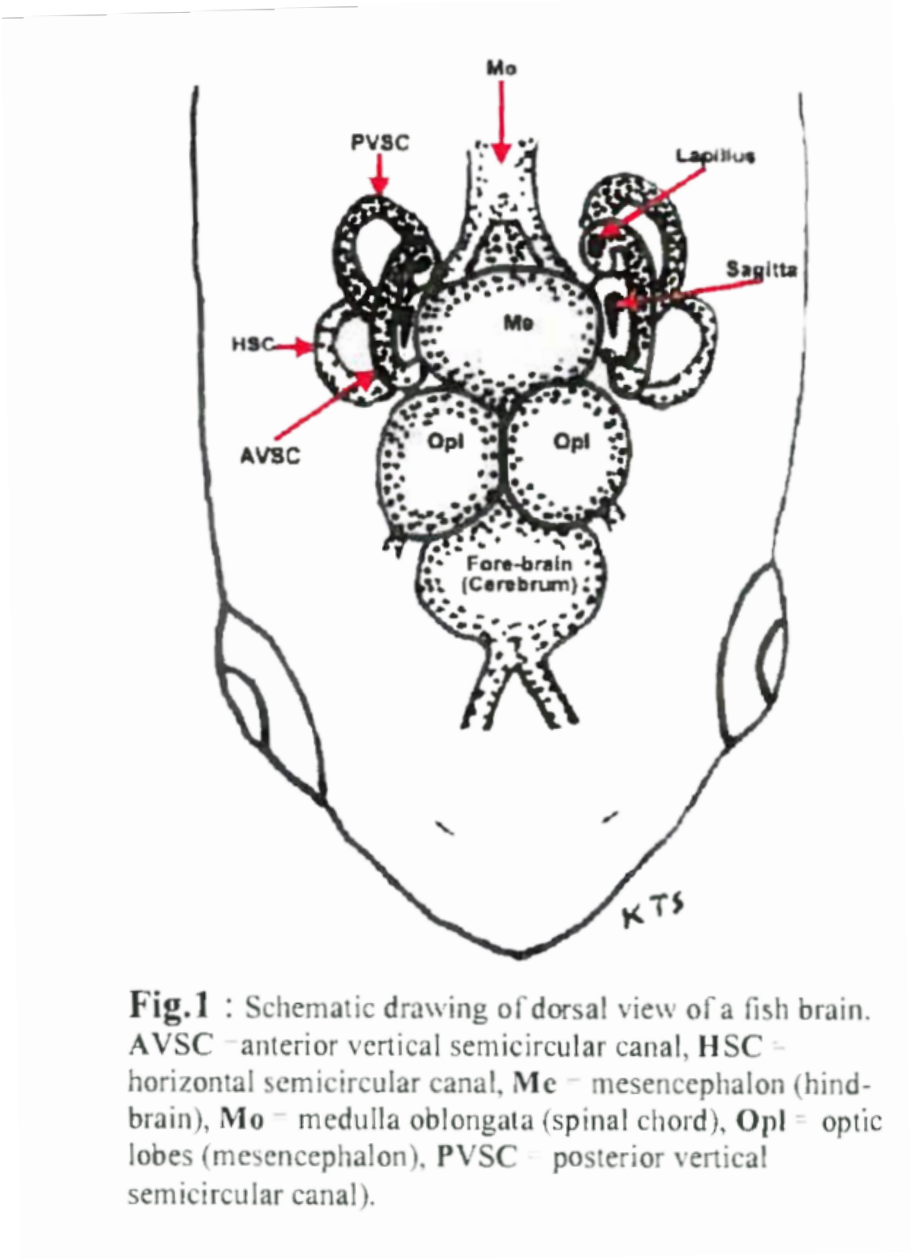


Fig.1 : Schematic drawing of dorsal view of a fish brain. **AVSC** = anterior vertical semicircular canal, **HSC** = horizontal semicircular canal, **Me** = mesencephalon (hind-brain), **Mo** = medulla oblongata (spinal chord), **Opl** = optic lobes (mesencephalon), **PVSC** = posterior vertical semicircular canal).

Figure 2.2: Overview of fish brain and otolith organs from Oth (2011)

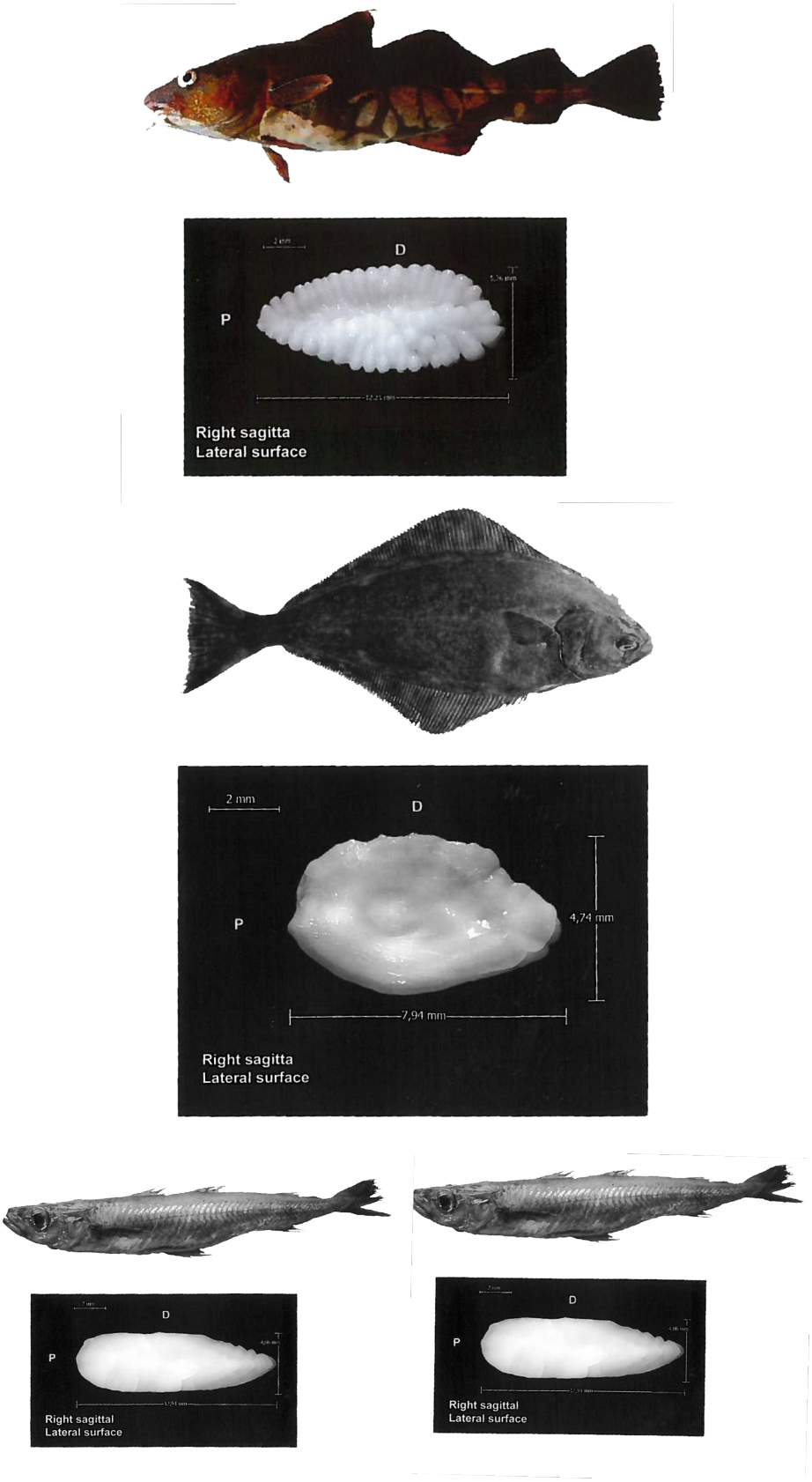


Figure 2.3: Example images of fish fish and their respective otoliths. From top to bottom, right to left (A) Cod, (B) Halibuth, (C) Herring (D) Whiting

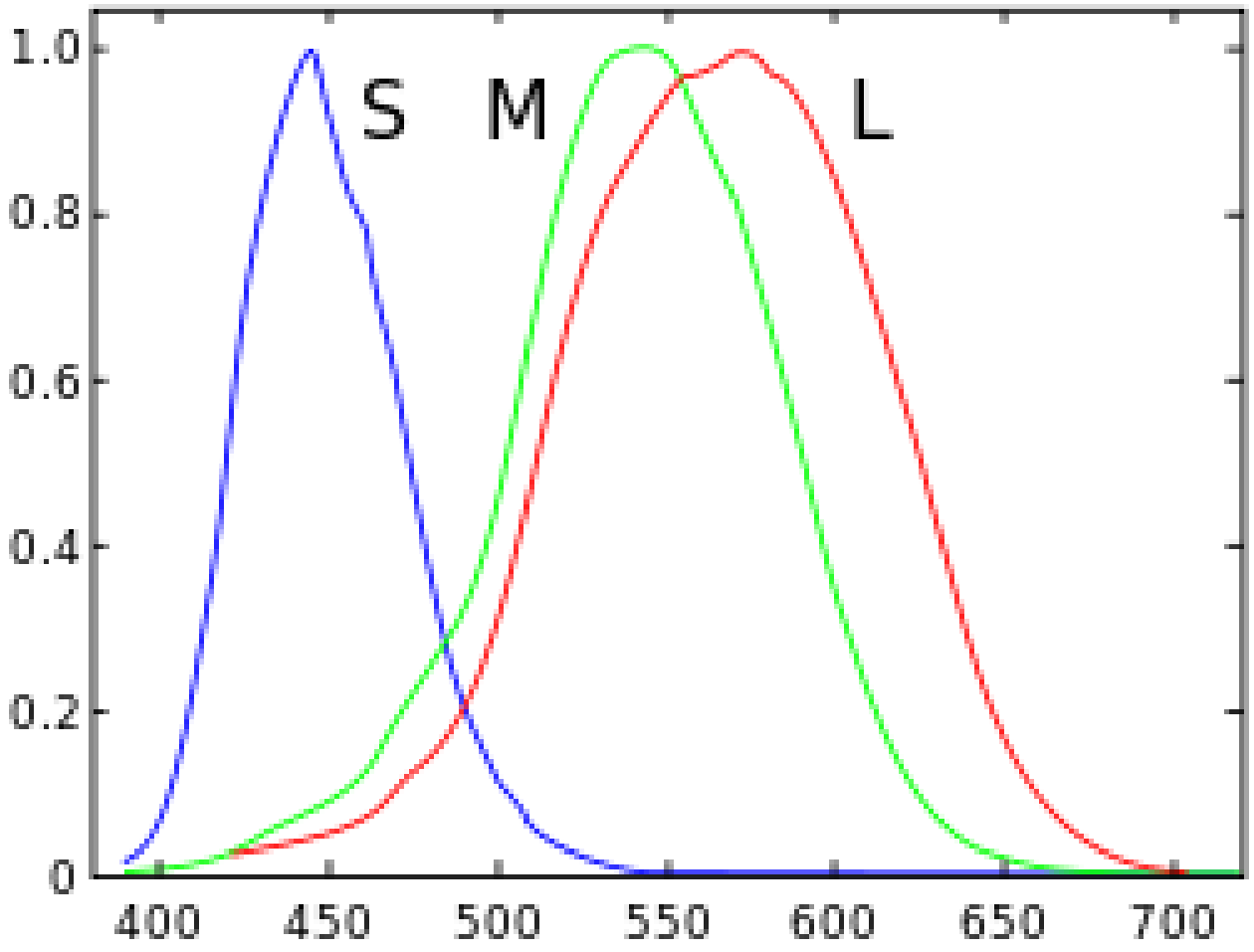


Figure 2.4: Normalized responsivity spectra of human cone cells; S, M, and L types on different wavelengths of light. Photographic equipment generally measure 3 distinct wavelengths of light to approximate the results these curves would give.

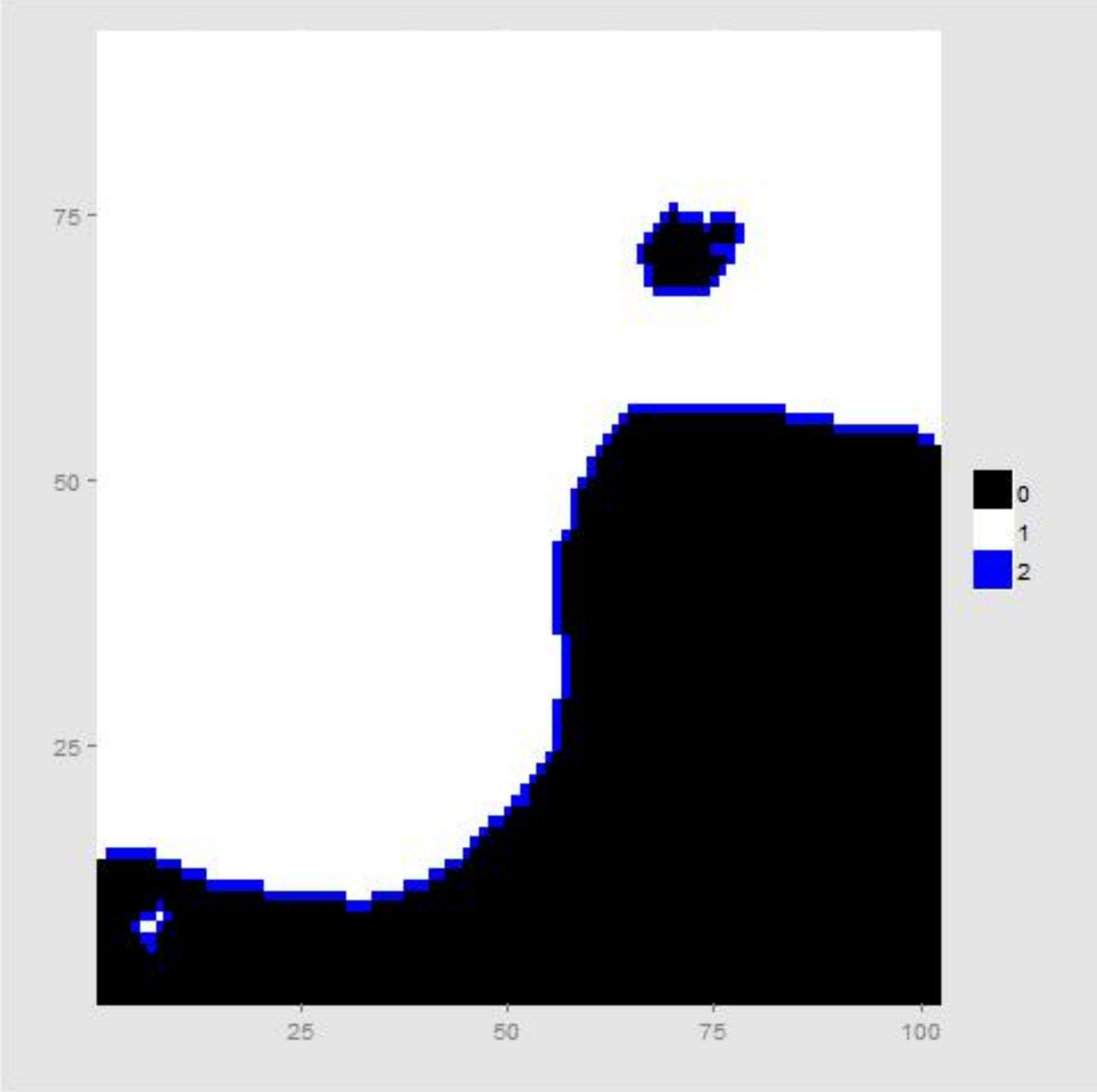


Figure 2.5: Blow-up of edge of a Halibut Otolith, for illustrating edge-traversal and starting points. The *island* of black and *lake* of white both provide false starting points for tracing the contour.

Chapter 3

Theory

This chapter contains the needed theory in mathematics and statistics.

3.1 Elliptic Fourier Descriptors

Elliptic Fourier Descriptors (EFD) is a well known method for approximating the contour of otoliths (Kuhl and Giardina, 1982). The approach itself is pretty straightforward. Given a set of co-ordinates (x_n, y_n) following the contour of a otolith, separate these into separate x_n and y_n-paths. Compute t_n using euclidic distance and approximate (x_n, t_n) and (y_n, t_n) pairs using the discrete fourier transform (DFT).

In addition to this, Kuhl and Giardina (1982) outlines a method to normalize the coefficients. This is needed because of the ambiguities introduced from using two 1D fourier-series to represent a closed 2D-contour, and because rotation and size must be considered arbitrary.

3.2 Discrete Fourier Transform

The discrete fourier transform approximates any function using a set of linearly weighted sines and cosines with increasing frequencies. In Fourier Contour Analysis this is used to approximate

two sets of paths (x_n, t_n) and (y_n, t_n). The model with $(2+4N)$ parameters will then be:

$$F_x(t) = a_0 + \sum_{n=0}^{N} a_n \cos(nt) + b_n \sin(nt) \qquad 0 < t < 2\pi \tag{3.1}$$

$$F_y(t) = c_0 + \sum_{n=0}^{N} c_n \cos(nt) + d_n \sin(nt) \tag{3.2}$$

Before transformation we have a contour described by (x_n, t_n), (y_n, t_n) with $n = \{1,2,3...K\}$, where K, the number of pixels in the countor, typically ranges between 1000 and 4000. After transformation we will have a model defined by:

$$\{a_0, c_0, a_n, b_n, c_n, d_n\} \qquad n = \{1,2,3....N\} \tag{3.3}$$

with N being the number of harmonics. Typically with N at least 10 the error is visually very small, however higher N may still contain usefull information.

3.3 Fast Fourier Transform

Fast Fourier Transform (FFT) is a specific algorithm for computing some cases of the discrete fourier transform (DFT) easily. Normal DFT scales at the order of $O(n^2)$, while FFT is both generally fast and scales better at $O(n \log n)$. However, normal DFT with 1000-2000 datapoints runs in much less than a second on a normal computer, so the total time saved is not significant when using DFT on otolith contours.

FFT requires the length of the path to be on 2^n with $n \in \mathbb{N}$, and assumes equal length between datapoints. In order to use FFT on the contours, interpolation or a similar technique would be needed to transform the paths into accepted lengths. This introduces a very small, but uneccessary error. FFT has therefore not been used in this thesis.

3.4 Creating a path and finding the coefficients

According to the methodology presented by Kuhl and Giardina (1982), if given a path x_n and y_n with $n = \{1,2,,,K\}$ then:

$$\Delta x_n = x_n - x_{n-1} \qquad \Delta x_1 = x_1 - x_N \tag{3.4}$$

$$\Delta y_n = y_n - x_{y-1} \qquad \Delta y_1 = y_1 - y_N \tag{3.5}$$

$$\Delta t_n = \sqrt{(\Delta x_n)^2 + (\Delta y_n)^2} \tag{3.6}$$

$$T = \sum_{n=1}^{N} \Delta t_n \tag{3.7}$$

$$\tag{3.8}$$

The paths to be approximated are then points x_n and y_n placed at time t_n.

$$a_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{N} \frac{\Delta x_p}{\delta t_p} (\cos \frac{2n\pi t_p}{T} - \cos \frac{2n\pi t_{p-1}}{T}) \tag{3.9}$$

$$b_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{N} \frac{\Delta x_p}{\delta t_p} (\sin \frac{2n\pi t_p}{T} - \sin \frac{2n\pi t_{p-1}}{T}) \tag{3.10}$$

$$c_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{N} \frac{\Delta y_p}{\delta t_p} (\cos \frac{2n\pi t_p}{T} - \cos \frac{2n\pi t_{p-1}}{T}) \tag{3.11}$$

$$d_n = \frac{T}{2n^2\pi^2} \sum_{p=1}^{N} \frac{\Delta y_p}{\delta t_p} (\sin \frac{2n\pi t_p}{T} - \sin \frac{2n\pi t_{p-1}}{T}) \tag{3.12}$$

with the approximated contour given by:

$$\dot{x}(t) = a_0 + \sum_{n=1}^{N} a_n \cos \frac{2n\pi t}{T} + b_n \sin \frac{2n\pi t}{T} \tag{3.13}$$

$$\dot{y}(t) = c_0 + \sum_{n=1}^{N} c_n \cos \frac{2n\pi t}{T} + d_n \sin \frac{2n\pi t}{T} \tag{3.14}$$

3.5 Standardizing size, position, rotation and traversal

In this section the same approach as presented in Kuhl and Giardina (1982) is used, however the terms σ and t_0 are included explicitly in the mathematical model, rather than as a separate

adjustment performed later.

3.5.1 Size and position

Position is centered on origo by setting a_0 and c_0 from 3.3:

$$a_0 = 0 \qquad c_0 = 0 \tag{3.15}$$

since for any real n, a_n and b_n

$$\int_{t=0}^{2\pi} a_n \cos(nt) + b_n \sin(nt)\, dt = 0 \tag{3.16}$$

This is necessary because position within the original image is arbitrary. a_0 and c_0 will thus be disregarded from now.

Kuhl and Giardina (1982) recommends standardizing size by transforming the length the of the major axis to 1. Standardizing volume of first harmonic to 1 instead may provide more aesthetically pleasing results.

3.5.2 A model with t_0 and ϕ

The final contour produced by combining the two paths is the only thing of interest when modelling otoliths. There are however many sets of x_n and y_n paths that result in the exact same contour. Since these will be treated differently in a discriminant analysis, these ambiguites must be corrected for.

The ambiguties introduced from splitting up a contour into x and y-paths are twofold. First, the direction of traversal is not of interest. Secondly, any point on the contour can be used as a starting point. Because of this, in the fourier model t may range between from any t_0 to any $\pm(t_0 + 2\pi)$ and the resultant contour will be exactly the same.

Incorporating arbitrary starting point to the fourier contour model leads to the following

$$F_x(t) = \sum_{n=1}^{N} (a_n \cos nt + b_n \sin nt) \qquad 0 < \pm(t + t_0) < 2\pi \tag{3.17}$$

$$F_y(t) = \sum_{n=1}^{N} (c_n \cos nt + d_n \sin nt) \tag{3.18}$$

Furthermore, the orientation of the otoliths is merely how they were placed during imaging. It is therefor necessary to standardize the orientation by some measure. For that reason, a rotational factor ϕ will be included in the model. To achieve this we use a rotational matrix on the form:

$$\begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X^* \\ Y^* \end{bmatrix} \tag{3.19}$$

Applying this rotational matrix yields:

$$F_x^*(t) = \sum_{n=1}^{N} (a_n \cos\phi + c_n \sin\phi) \cos nt + (b_n \cos\phi + d_n \sin\phi) \sin nt) \qquad 0 < \pm(t + t_0) < 2\pi \tag{3.20}$$

$$F_y^*(t) = \sum_{n=1}^{N} (c_n \cos\phi - a_n \sin\phi) \cos nt + (d_n \cos\phi - b_n \sin\phi) \sin nt) \tag{3.21}$$

which yields the following formulaes

$$F_x^*(t) = \sum_{n=1}^{N} a_n^* \cos(nt) + b_n^* \sin(nt) \qquad 0 < \pm t < 2\pi \tag{3.22}$$

$$F_y^*(t) = \sum_{n=1}^{N} c_n^* \cos(nt) + d_n^* \sin(nt) \tag{3.23}$$

$$a_n^* = (a_n \cos\phi + c_n \sin\phi) \cos nt_0 + (b_n \cos\phi + d_n \sin\phi) \sin nt_0 \tag{3.24}$$

$$b_n^* = -(a_n \cos\phi + c_n \sin\phi) \sin nt_0 + (b_n \cos\phi + d_n \sin\phi) \cos nt_0 \tag{3.25}$$

$$c_n^* = (c_n \cos\phi - a_n \sin\phi) \cos nt_0 + (d_n \cos\phi - b_n \sin\phi) \sin nt_0 \tag{3.26}$$

$$d_n^* = -(c_n \cos\phi - a_n \sin\phi) \sin nt_0 + (d_n \cos\phi - b_n \sin\phi) \cos nt_0 \tag{3.27}$$

The formulaes 3.22 through 3.27 can thus be used to rotate any contour by ϕ and shift starting point by t_0 at will.

3.5.3 Standardizing rotation

Going back to Kuhl and Giardina (1982), the proposed way to standardizr rotation is by standardizing the outline of the first harmonic where $N = 1$. In 5.2 a proposed method to numerically find best fit is outlined, and in 6.4.1 it is investigated if this improves discrimination. Since it is considered standard, Kuhl and Giardina (1982) will however be used for now.

A standard ellipse, with starting point set at a major axis, has the following represenation:

$$G_x(t) = c_1 \cos t \qquad 0 < t < 2\pi \tag{3.28}$$

$$G_y(t) = c_2 \sin t \qquad c_1 > c_2 \tag{3.29}$$

In order to transform eq. 3.2 to 3.29, one must thus find the t_0 and ϕ in eq. 3.22 through 3.27 that yield $b_1 = 0$, $c_1 = 0$ and $a_1 > d_1$. This yields:

$$0 = -(a_1 \cos\phi + c_1 \sin\phi) \sin t_0 + (b_1 cos\phi + d_1 sin\phi) cost_0 \tag{3.30}$$

$$0 = (d_1 \cos\phi - b_1 \sin\phi) \sin t_0 + (c_1 cos\phi - a_1 sin\phi) cost_0 \tag{3.31}$$

which results in

$$t_0 = \tan^{-1} \frac{b_1 \cos\phi + d_1 \sin\phi}{a_1 \cos\phi + c_1 \sin\phi} \tag{3.32}$$

and ϕ can be found by solving this quadratic equation

$$(b_1 d_1 + a_1 c_1) + (d_1^2 - b_1^2 + c_1^2 - a_1^2) \tan\phi + -(b_1 d_1 + a_1 c_1) \tan^2\phi = 0 \tag{3.33}$$

Furthermore, since the direction of traversal is also irrelevant it should also be standardized. After rotation and size has been normalized, direction of traversal can easily be switched by substituting $t = -t*$. Thus if $sign(a_1) \neq sign(d_1)$, we substitute $t = -t^*$. This is equivalent to setting $b_n^* = -b_n$ and $d_n^* = -d_n$.

3.6 Rotational Ambiguity

Using the normalization procedure outlined by Kuhl and Giardina (1982) there are two separate orientations that fulfill the conditions set out for normalization: rotation ϕ and $\phi + \pi$. This is referred to as the rotational ambiguity.

For sets which are generally aligned well from imaging choosing the alternative that is closest to the original rotation will work well. Using the vector from *center of mass* to the (a_0, c_0)-coefficients as a basis for a metric to choose between ϕ and $\phi + \pi$ was found to work well. This ensures that smoothest part of the contour always points the same way. If used on both left and right otoliths correct orientation was found on all cases in this set. However, the proposed solution for numerically finding best fit rotation, found in 5.2, eliminates this ambiguity.

3.7 Sinusiod Fourier Form

It should be noted that the standarized coefficients can be rewritten as a sinusiods:

$$F_x(t) = \sum_{n=1}^{N} (a_n \cos(nt) + b_n \sin(nt) = \sum_{n=1}^{N} (\alpha_n \sin(nt + \beta_n)) \qquad 0 < t < 2\pi \tag{3.34}$$

with

$$a_n = \alpha_n \cdot \sin(\beta) \qquad \alpha^2 = a_n^2 + b_n^2 \tag{3.35}$$

$$b_n = \alpha_n \cdot \cos(\beta) \qquad \beta = \sin(\frac{b_n}{\alpha_n})^{-1} \tag{3.36}$$

In 6.3.1 it is investigated whether this reparameterization leads to better results in the discriminant analysis.

3.8 Discrimant Analysis

Discrimination and classification are techniques concerned with separating distinct set of objects, aswell as creates rules for allocating new objects to previously defined groups. Unlike other more exploratory fields of analysis this field has a very specific purpose; finding good classification rules. This also means that discriminant analysis is a field where one can measure performance of an algorithm quite easily.

Typically, if one has N sets of X samples with p measured variables each, with each X_n belonging to class k of sets π_k, then the purpose of discriminant analysis is to generate a classification rule C which assign new samples of X to one of π_k. Often the number of sets to allocate to is restricted to two sets, π_1 and π_2.

3.8.1 Error rates and Cross-Validation

A model is judged on how well it classifies samples. *Apparent error rate* (APER) is the proportion of samples from the training set which are misclassified by a model. The *training set* is the set of samples which is used to build the model. If one has enough samples, some samples can be set aside as a *validation set* which is purely used to test the performance of the model. APER is a typically a very optimistic error estimate, as any random effects it picks up on will increase APER, but generally decrease actual model performance. In fact, if one adds a iid gaussian distributed noise variable APER will always increase with provably no added predictive effect.

At the heart of discriminant analysis is the concept of cross-validition [1]. The most common forms of cross-validition is Leave-one-out (LoO) and Leave-p-out (LpO). This involves classifing p-samples at the time, by building a classification model with everything but those samples. Since the model then is independant of the samples it classifies, it is gives a true estimate of the reduced models performance on that set of data. An unbiased estimate is referred to as AER (actual error rate). Actual model performance will be slightly higher, as more data generally means better models.

Whether that model performs as well on real data is however still a matter of to which degree the sets of data one has accessible is represenative of the populations they were drawn from.

[1] The terminology 'cross validition' seem to be more common in applied statistics than pure statistics. Lachenbruchs holdout procedure is exactly the same as LoO-cross validation

It is also important to introduce the concept of *model selection bias*. Cross validitions produces a probability that one sample is correctly classified. Even though cross validition produces unbiased results, there is still a random component. This means that if one select which model to use based on cross-validated results, that cross-validated result should no longer be considered unbiased in that context.

3.8.2 Confusion Matrix

After a classification rule has been found, a *confusion matrix* is commonly used to summarize the classification results. A shortened version of the confusion matrix at 6.2 showing predicted vs actual location of catch of cod otoliths is shown below. The results were found using LpO(p=10) cross-validation.

		Predicted				
		BAL	BAR	LOE	LOW	$\sum_a$
Actual	BAL	209	17	16	7	249
	BAR	54	248	30	35	367
	LOE	16	6	23	18	63
	LOW	5	8	16	10	39
	$\sum_p$	284	279	85	70	718

From this confusion matrix it is easy to compute several key statistics.

$$AER = 1 - \frac{\sum(diag)}{N} \tag{3.37}$$

Probabilities of samples being classified as x belonging to class x:

$$\frac{diag(M)}{\sum_p} \tag{3.38}$$

Probabilities of samples beloing to class x being classified as class x:

$$\frac{diag(M)}{\sum_a} \tag{3.39}$$

3.9 LDA

Linear Discriminant Analysis (LDA) is a discriminant analysis method which uses continuous independant variables to explain a categorical class-variable. Assume that $f_1(x)$ and $f_2(x)$ are multivariate normal densities. Let $\boldsymbol{\mu_1}$ be the first mean vector, $\boldsymbol{\mu_2}$ the second and $\boldsymbol{\Sigma}$ the common covariance matrix.

Now suppose that join densities of $\mathbf{X}' = [X_1, X_2 ... X_p]$ for populations π_1 and π_2 are given by:

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)} \tag{3.40}$$

Then after cancellation of common terms, the minimum ECM (expected cost of misclassification), or in other words; which density function the sample was most likely to be drawn from, becomes:

$$R_1 = -\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_3)'\Sigma^{-1}(x-\mu_2) \geq (\frac{c(1|2)}{c(2|1)}\frac{P_2}{P_1}) \tag{3.41}$$

where $c(2|1)$ and $c(1|2)$ are cost of misclassification and P_1 and P_2 are the prior probabilities of X belonging to π_1 and π_2. If there are no specific reasons to not, it is common to select equal costs of misclassification. If a sample is not classified as π_1, it is obviously classified as pi_2.

Since μ_1, μ_2 and Σ is commonly not known, it is common to use the sample mean vectors $\bar{x_1}$, $\bar{x_2}$ and S_{pooled} as estimates instead.

In general, LDA works well for fuzzy data were trends and probabilities are among the more defining characteristics. It is less good at picking up attributes that are not easily picked up in a covariance matrix, such as strong specific dependencies within a subpopulation.

Chapter 4

Validity of the contour

This chapter investigates the procedures outlined in 2.5, and attempts to investigate to which degree the contour created is dependant on these procedues. The issues raised mostly concern the realms of computer science, and it is investigated whether this affects the validity of the final discriminant analysis.

4.1 Comparing paths

In order to compare paths easily a simple metric has been used. In essence, instead of comparing contours, the x and y-paths have been compared separatly. The steps have been as following:

1. For each contour (x_n, y_n):
   (a) Change lenght of (x_n, y_n) to 1000 using linear interpolation
   (b) Combine x_n and y_n to one list of 2000 elements
2. Use correlation to measure equalness between different contours

Please refer to **??** in the appendix for further details.

4.2 Conversion to grayscale

There are many ways to convert a colour image to grayscale (Cadík, 2008). Neither are considered de-facto standard, and none more correct than the others. As outlined in 2.4, color information is not necessarily trustworthy, so to which degree the resultant path is dependant on color-choices should be investigated.

For 10 randomly selected images several methods were used for colour conversion and contours were created for each. For each contour paths were created, and the procedure outlined in 4.1 was used on each. Matlabs *rgb2gray*, which uses the luminance related to the NTSC-signal based on the BT.601 standard, was used as reference for computing correlation against.

Colour convertisions and correlation

Method	Details	Correlation
HSL - 601	$0.2989R + 0.5870G + 0.1140B$	Reference
CIE 1931 y	$0.2126R + 0.7152G + 0.0722B$	1.000
Band with highest range	R or G or B	1.000
Average of RGB	$\frac{R \cdot G \cdot B}{3}$	1.000

Different colour conversions are thus ruled out as a potential source for bias in this case. Furthermore, the effects measured seem small enough for this assertion be likely to hold in all but the most extreme cases.

4.3 Thresholding

Selecting a intensity threshold for separating background and otolith is something one would assume is easily solved. However due to images like fig 4.1, with backdrops in similar intensity to outer edge in other images, detecting the otolith is not trivial. No easily implented solutions for finding a usable thresholding automatically has been found, neither by trial nor literature search.

For the halibut dataset using the average of *mean pixel intensity* and *mode pixel intensity* worked well. Otsu's method, which is commonly used relies on the assumption of common

variance in the two groups and does not work well with our data.

However, for both the cod and the halibut sets the range of thresholds which performed well was very large. With noisy images this may however not be the case, so thresholding is probably best considered a manual step where manual verification of the contour created is necessary. This also means that computationally estimating errors in this step becomes impossible.

4.4 Image Resolution

To investigate whether images of different resolutions can be safely compared an image was rescaled using to several different resolutions using GIMP's implementation of cubic interpolation. Countors were created and compared for each as in 4.1.

Resolutions investigated

name	x	y
Orignal	3840	3072
r1920	1920	1536
r960	960	768
r480	480	384
c240	240	192

Matrix of correlations

	Original	r1920	r960	r480	r240
Original	1.0000000	0.9998851	0.9998612	0.9996346	0.9964146
r1920	0.9998851	1.0000000	0.9999901	0.9993930	0.9955669
r960	0.9998612	0.9999901	1.0000000	0.9993863	0.9955146
r480	0.9996346	0.9993930	0.9993863	1.0000000	0.9979696
r240	0.9964146	0.9955669	0.9955146	0.9979696	1.0000000

Changing the resolution does create small differences. This may be a minor cause for concern when comparing contours created from different images.

4.5 Smoothing and correlation

Smoothing can be very effective to combat distortions created by noise (Haines and Crampton, 2000). Noise is especially problematic in Fourier Contour Analysis since it not only affects the values of the contour, but the extra pixels introduced will affect the length of the contour. However, in clear and well-taken images smoothing will be less relevant.

To get a measure of the distortion introduced by smoothing, several levels of smoothing was applied to 10 known good randomly selected contours. A correlation was again computed as in 4.1. The smoothing was done as in Haines and Crampton (2000), by applying a $[0.25, 5, 0.25]$ filter on each pixel $[5, 10, 20, 80]$-times. The results were as following:

Matrix of correlations

	Original	sm5	sm10	sm20	sm80
Original	1.0000000	0.9994376	0.9991326	0.9986733	0.9975575
sm5	0.9994376	1.0000000	0.9999626	0.9998198	0.9992706
sm10	0.9991326	0.9999626	1.0000000	0.9999447	0.9995471
sm20	0.9986733	0.9998198	0.9999447	1.0000000	0.9997957
sm80	0.9975575	0.9992706	0.9995471	0.9997957	1.0000000

Very high levels of smoothing may introduce a very slight bias, however the benefits seems to outweigh the problems. Smoothing should probably be considered a standard feature when applying the fourier contour transform.

Illustrations for chapter4

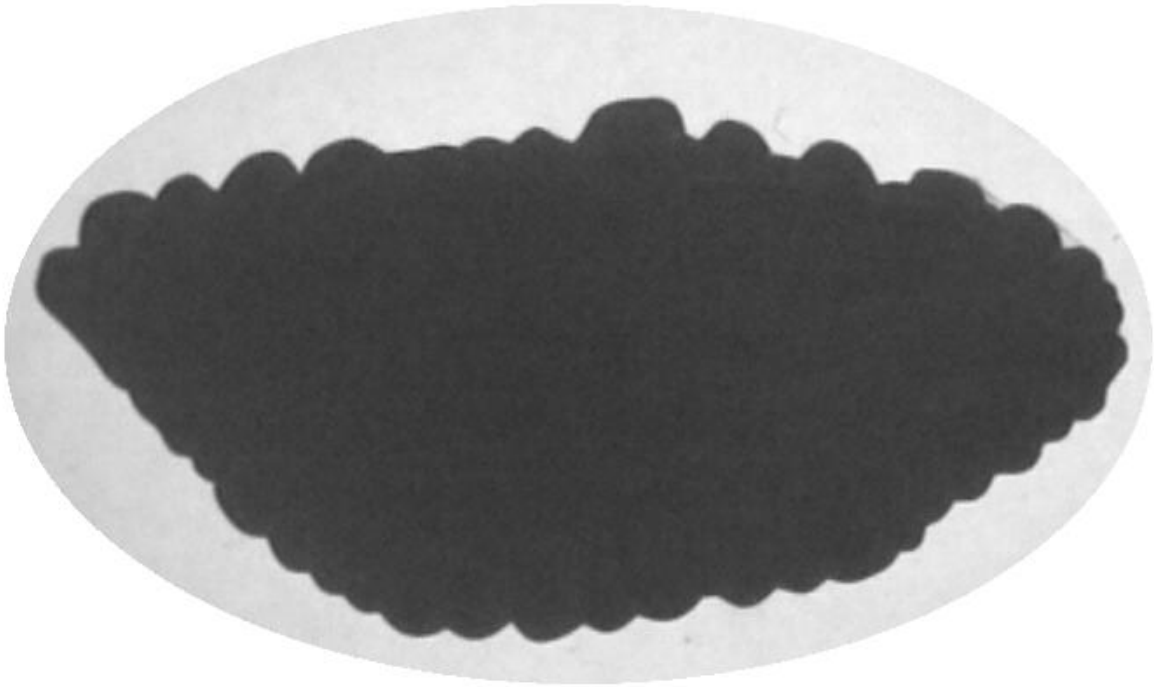


Figure 4.1: Cod otolith placed on circular backdrop.

Chapter 5

Validity of class comparison

This chapter investigates to which degree comparison of two classes of otoliths produced valid results.

5.1 Broken Otoliths

Manual inspection of the Otoliths was performed by Alf Harbiz. Out of a total of 1332 otoliths 421 were classified as not fit for inclusion in analysis. If this breakage is random this should not affect the resulting sampling distribution. As long as the sample size is still adequate random loss of samples is not a case for concern. However the process of slicing otoliths is dependant on the skill of the person cutting (Oth, 2011), and this is thus a potiental source for systematic bias.

Without knowing which otoliths were processed by which operator a detailed investigation into the potential batch-effects introduced by different operaters is not feasible. It is however possible to easily investigate whether there are systematic differences between the two halibut sets. If breakage is random, then sets of constisting of {All whole left otoliths which have whole right twin} and {All whole left otoliths which have broken right twin} should have similar characterics.

Characteristics of left otoliths

State of right otolith	n	mean otolith area	mean otolith edge
	1250	$1.035 \cdot 10^6$	5166
Whole	1184	$1.034 \cdot 10^6$	5182
Broken	66	$1.047 \cdot 10^6$	4884

The probability of a sample A^* of size 66 having length shorter than 4844 is 0.076. Likewise, the probability of a sample A^* of size 66 having area less than $1.047 \cdot 10^6$ is 0.606. Neither of these results should be very concerning. However, length of edge and size of area are positivly correlated. Accept/Reject sampling of A^* with size 66 again gives

$$P(A^*_{edge} < 4884 | A^*_{area} > 1.047 \cdot 10^6) = 0.00003 \tag{5.1}$$

This is significant, and there are thus systematic bias in which otoliths get broken. Bigger otoliths with less convuleted edges apparently tend to break more often. This is worthy of note, since intuitively it is easy to assume that smaller otoliths with strange shapes would be more likely to break.

However, while this effect is statistically significant it may not be strong enough to be biologically or analytically significant. Comparing area and length of edges of otoliths with and without broken twins graphically (5.1 and 5.2) show distributions with very comparable behaviour.

5.1.1 Distance from mean

As an addional measure, the paths of otoliths with and without broken twins were compared against an averaged path. Given K paths of $p_{ki} = [x_{k1}, x_{k2}..x_{kN}, y_{k1}, y_{k2}...y_{kN}]$ using the methodology in 4.1, and using

$$\bar{p}_i = \sum_{k=1}^{K} \frac{p_{ki}}{K} \tag{5.2}$$

$$M_k = \sum_{i=1}^{N} |p_{ki} - \bar{p}_i| \tag{5.3}$$

with M_k as a metric for equalness, the plot in 5.3 was created. No interesting patterns were found using this metric.

5.2 Rotation

Landmarks, that is distinct reoccuring features or patterns, are a common way of fixing orientation when comparing shapes. Otoliths however lack easily identifiable landmarks. This is problematic because the fourier coefficients are dependant on both starting point and orientation. Standardazing orientation and starting point in a way that best facilitates discriminant analysis is thus imperative.

Standardazing size and aligning the major axis of the first harmonic together, as done by Kuhl and Giardina (1982), is an efficient and mathematically sound method for ensuring a fairly good fit. With modern computeres it is however feasible to numerically search for orientations with better fits.

The approach outlined here first finds a good-of-fit function to compare two countors against eachother. Secondly it finds the average otolith shape to align against. Lastly it outlines how to numerically search through different rotations and starting point shifts to best align each otolith.

First a statistic for goodness of fit between two contours $P_1 = (x_n, y_n)$ and $P_2 = (x_n^*, y_n^*)$, with $n = [1,2,3..N]$ is proposed

$$T(P_1, P_2) = \sum_{n=1}^{N} \sqrt{(x_n - x_n^*)^2 + (y_n - y_n^*)^2} \tag{5.4}$$

T is thus the sum of euclidic distance between points (x_n, y_n) and (x_n^*, y_n^*), in other words a distance function. This is a good measure for difference between otolith contours as long as the two path have well-aligned starting points, however this statistic increases sharply with badly aligned starting points.

A good contour to align against is the mean contour from the previously normalized contours. Let C_k, $k = [1,2,3..K]$ be K sets of fourier coefficients with N harmonics with $C = [a_0, c_0, a_1, b_1..d_N]$. Using the inverse fourier transform, F_{inv}, the paths P_k can be recreated from the normalized coefficients. A mean contour $\bar{P}$ can thus be found as in 5.1.1. By choice we recreate the paths with 1000 points.

$$F_{inv}(C_k) = P_k = (x_n, y_n), \qquad n = [1,2,3,...1000] \tag{5.5}$$

$$\bar{P} = (\sum_{k=1}^{K} \frac{x_{kn}}{I}, \sum_{k=1}^{K} \frac{y_{kn}}{K}) \tag{5.6}$$

Remembering that equations 3.22 through 3.27 can be used to shift rotation and starting point by ρ and t_0. If we define $R(C, t_0, \rho)$ as the function that rotates contour C by ρ and shifts starting point by t_0 then finding the best orientation is just a matter of minimizing on t_0 and ρ:

$$T(\bar{P}, F_{inv}(R(C_k, t_0, \rho))) \qquad k = [1,2,3,4..K] \tag{5.7}$$

Since there are many local maxima, this was done numerically by testing every pair of 180 different possible rotation and 100 different shifts for all 2771 otoliths. For each otolith the pair of (t_0, ρ) which gave the smallest value for T was then used. This was run overnight on a standard desktop computer.

5.3 Correcting for covariates

In the halibut set there is a strong dependence between the fourier coefficients and the sex/weight/length (from now on SWL) profiles of the halibuts. Since the two sets of halibut have fairly different SWL profiles, there is big risk that any discrimination results, even cross-validated, will pick up on these differences rather than those stemming from growth conditions, water temperatures and other systematic differences between the two populations of halibut. Since the SWL profiles of these two catches most likely are not representative of the overall populations we are trying to discriminate between, we will up with a model with very little real value if that dependency is allowed to exist.

The approach used to solve this used by Are Murberg Henriksen and Alf Harbitz in Henriksen (2013) was a stratification-based resampling. The cods were placed into groups with similar characteristics, and two comparable sets were made by drawing equal numbers from each category into each set. This is a straightforward and reasonable approach.

It is proposed here that a better approach is to keep the smallest set of otoliths as is and only resample the largest set. When doing this resampling, a set of probabilities which gives the resulting distribution the same weight/length/sex characteristics as the smallest one should be used. A method , based on the accept/reject algorithm with an addional monte-carlo adjust-

ment, is proposed for finding these probabilities.

The results were as following:

Population characteristics

	Sex		Weight		Length	
Population	A_e	A_g	A_e	A_g	A_e	A_g
Initial	0.41	0.73	1365	1257	49.9	50.8
Stratified	0.5	0.5	1350	1316	51.19	51.37
Prob-based	0.735	0.733	1257	1258	50.81	50.81

Linear Regression

Using a regression [*Class* ~ *Weight+Sex+Length+Fourier Coefficents*] the following proportions of explained variance was found:

	Residuals	Sex+Weight+length	Fourier Coefficients
Initial	0.293	0.159	0.548
Stratified	0.264	0.007	0.728
P-adjusted	0.258	0.000	0.742

As seen, the stratification approach performed very well, but not perfectly. Whether the trade-off between added complexity and minor improvement is worth it is an interesting question.

5.3.1 Stratification approach

The stratification approach on this dataset was based on dividing the sets based on over/under average on weight, length and sex. This resulted in the following classes

Egga

		Weight	
		Low	High
Length	L	(224, 72)	(38, 57))
	H	(0, 0)	(226, 211)

Grenland

		Length	
		Low	High
Height	L	(14, 12)	(5, 16))
	H	(0, 0)	(3, 33)

These tables show number of males/females on characterics categorized on above or under respective means. Equal numbers were resampled from each class to form two new datasets. Category above average length and below average weight was ignored.

5.3.2 Probability adjustment

The overall goal of this algorithm is to find a set of probabilities $p(a)$, which when used when resampling set A will cause A to have the same expected value for the covariates weight, sex and length (hereby referred to as WSL) as set B.

First the WSL attributes of A and B is modeled using a kernel density function. The proportion between these two estimates is used as a measure on how the samples in A needs to be weighted up or down when sampling. Additionally, a Monto Carlo adjustment was used to get exact results. This is because when using the inverse kernel density function the expected WSL-values changes slightly.

1. Let D be any kernel density function which accepts a bandwidht bw and set of samples. Any kernel density function can be used, as long as it's additive

$$D(bw, A+B) = D(bw, A) + D(bw, B) \tag{5.8}$$

and symmetric so

$$E_{wls}(B) = E_{wls}(D(bw, B)) \tag{5.9}$$

.

In this thesis a multidimensioal gaussian density (Hayfield and Racine, 2008)[1] with bandwidth found by using the normal reference rule-of-thumb implented in the r-package np. The

[1] np: Nonparametric kernel smoothing methods for mixed data types

same fixed bandwidth should be used for both densities. This is used to estimate probability density functions for each set:

$$P_A(w,l,s) = D(bw, A) \tag{5.10}$$

$$P_B(w,l,s) = D(bw, B) \tag{5.11}$$

2. Now find the set of ratios between the probability density functions, and divide by the total to construct a new probability density function. This is equivalent to using accept/reject sampling.

$$W_{wls} = \frac{P_B(wls)}{P_A(wls)} \tag{5.12}$$

$$P_{wls} = \frac{W_{wls}}{\sum W_{wls}} \tag{5.13}$$

At this point, it should be noted that:

$$E_{wls}(B) = E_{wls}(D(bw, B)) \tag{5.14}$$

$$E_{wls}(A) = E_{wls}(D(bw, A)) \tag{5.15}$$

It should also be noted that sampling from P_a and accepting samples with probabilty P_{wls} results in exactly the same distribion of WSL as sampling from P_b.

3. Now use the inverse of the density function to find a proposed weight for each sample in A. For a sample a belonging to set A this would be

$$P(a) = \int \frac{D(bw, a)}{D(bw, A)} \cdot P_{wls} \tag{5.16}$$

Sampling using this set of probabilities will result in WSL-covariates very similar to B,however the WSL-attributes will be slightly off. This is due to P no longer being symmetrical, so this transition from continuous to discrete changes the estimates for WLS.

5.3.3 Monte Carlo adjustment

Since the inverted density function introduced a very slight error, a MC-algorithm was used to adjust the weight $P_{a,b,c}$ to minimize $E_{a,b,c}(A \cdot P_{a,b,c}) - E_{a,b,c}(B)$. The pseudo-code for the algorithm used was as following:

```
misses = 0
while(misses < 1000){
    a,b = 'draw two random samples'
    adj = random()/0.05 * p(a)
    if(letting p(b)=p(b)+adj and p(a)=p(a)-adj gives a better solution then:
        p(b)=p(b)+adj
        p(a)=p(a)-adj
        misses = 0
    if not:
        misses += 1
}
```

The score metric used assess the current model was sum of difference between estimate of standardized {a,b,c} of dataset B and weight-adjusted A.

This MC-algorithm was stable and the adjustment needed was very small. It does however complicate the algorithm and the gain in accuracy was very small. More work on testing in different scenarios of the probability based approach is necessary, both for quantification of the effects and to find best practice for bandwidth and density functions.

Illustrations

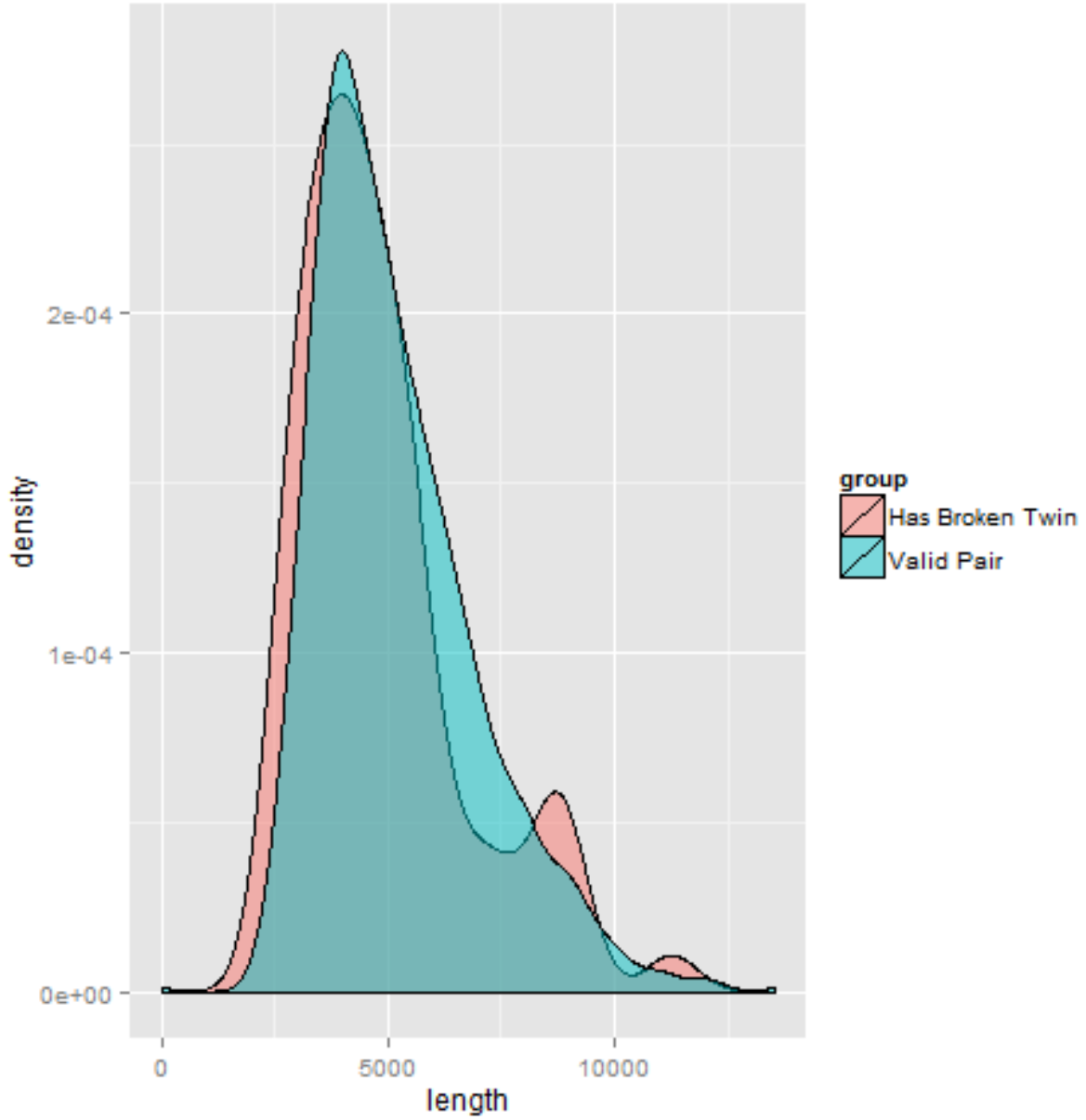


Figure 5.1: Comparison of length of contour of left halibut otoliths with whole and broken right twin

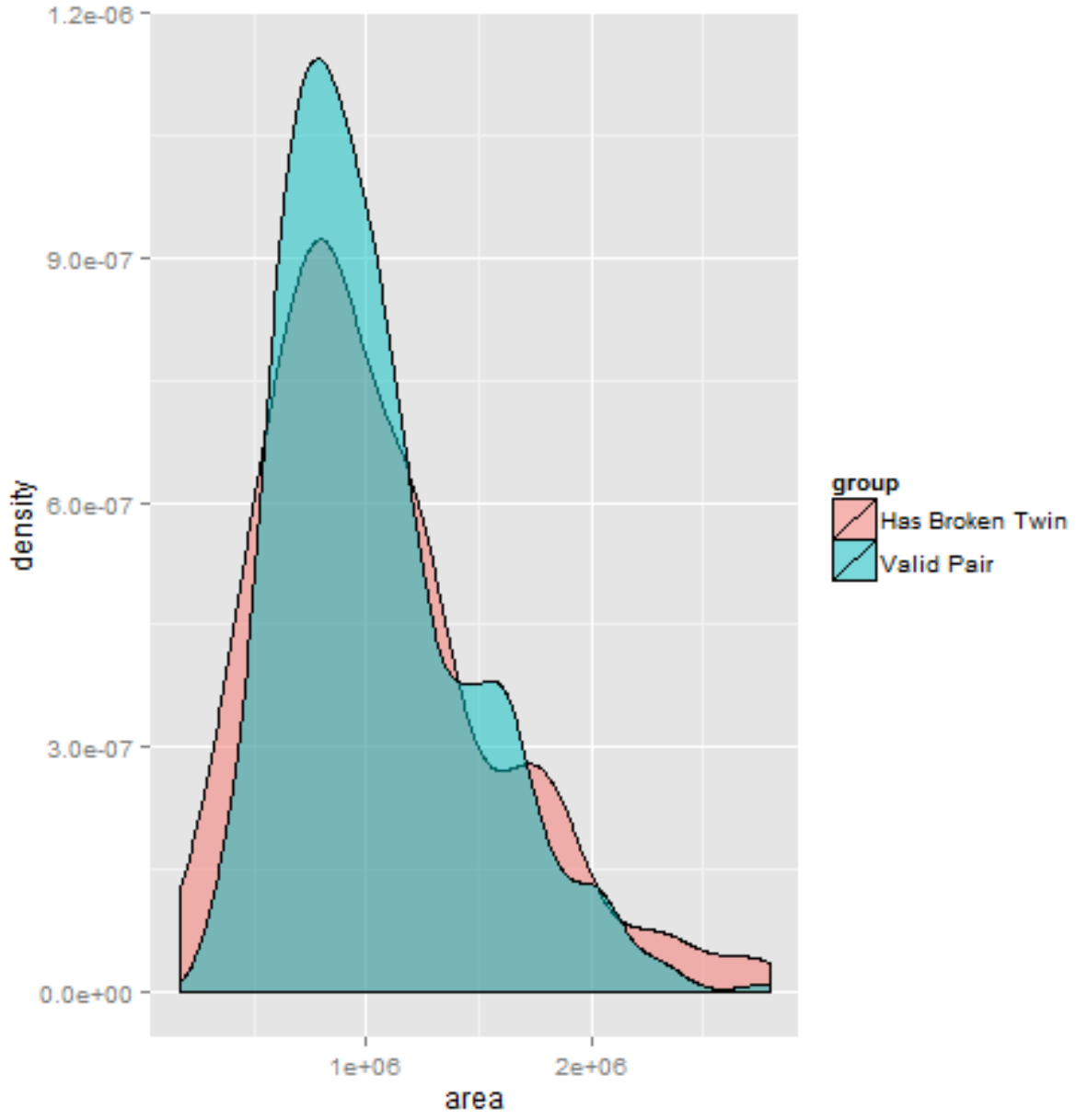


Figure 5.2: Comparison of area in pixels of left halibut otoliths with whole and broken right twin

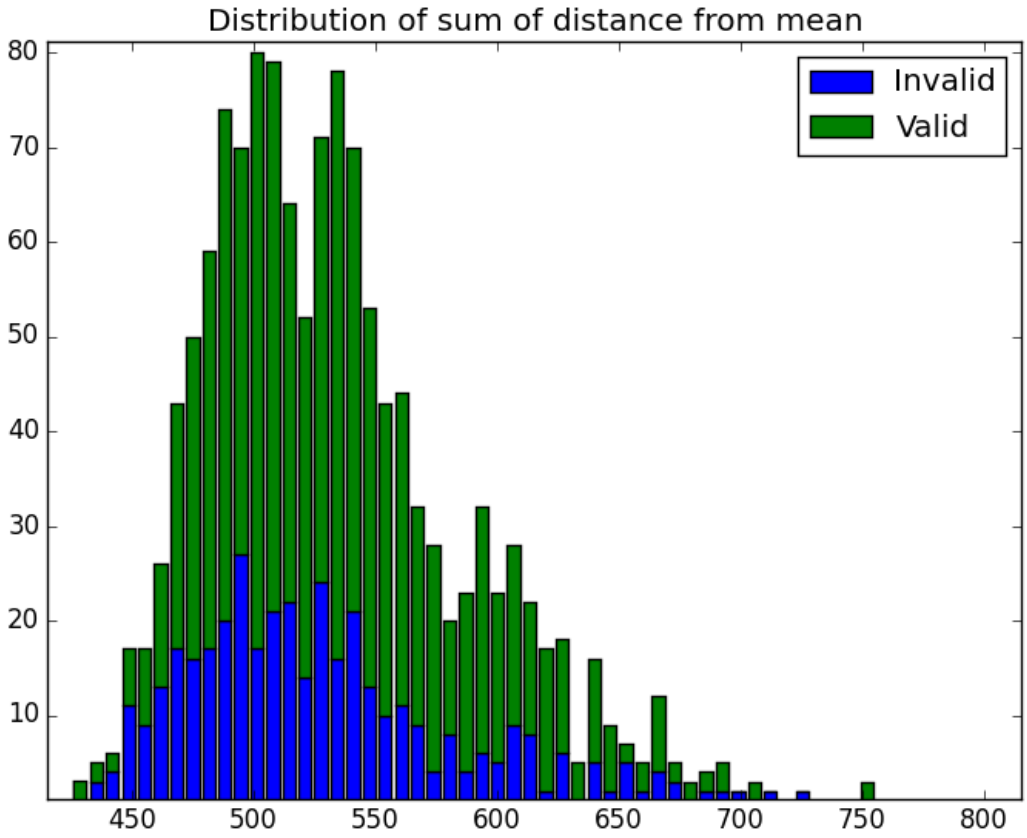


Figure 5.3: Comparison of valid and invalid otoliths using sum of distance from an averaged otolith

Chapter 6

Discriminant Analysis

6.1 Halibut Set

It was not possible to discriminate between the two sets of Halibut. The prediction rates using LDA on normalized coeffiecents with resampling as in 5.3.2 is shown in plot 6.1. It should be noted that resampling was done for each set of fourier components included. The mean overall prediction accuracy for this model was 0.502. The results were aquired using LpO ($p = 10$)-cross-validation.

It was also attempted to use the normalized coefficients as is, without any adjustments to make the sets comparable. The plot for this is shown in 6.2. It should be noted that a prior probabilty of $(0.5, 0.5)$ was used for this model. The mean prediction accuray for this model was 0.510.

The methodology used was the exact same as when achieving the same level of prediction accuracy when attempting to discriminate between 8 classes in the cod dataset.

6.2 Benchmark for Cod discrimination

Since it was not possible to discriminate between the two sets of Halibut, a cod set was made available for method testing. This is the same set used in Henriksen (2013) and Stransky et al. (2008).

In order to assess the effect of different approaches to discriminant analysis a straightforward LDA has been used as benchmark. No resampling, preprocessing or any deviations from the standard approach proposed by Kuhl and Giardina (1982) were used. All otoliths from all 8 classes were included. The class proportions were used prior probability in the LDA.

6.2.1 Optimal number of Fourier coefficients

First and foremost the number of coefficients from the Fourier Mode to use must be decided. Somewhat similar to the principal components from a PCA, latter coefficients explain less variance and adds less accuracy than the first ones (unlike PCA this is however not guaranteed with Fourier Components). As with PCA, cross-validation must be used to assess whether this added accuracy reflects noise or information, and to which degree including it leads to overfitting.

Correctly classified cod vs number of fourier components (n=1177)

1-10	-	391	508	502	527	545	564	555	569	570
11-20	577	576	591	607	616	617	614	618	623	616
21-30	608	592	591	615	608	600	589	586	600	604
31-40	603	591	592	598	599	594	609	603	604	597
41-50	592	590	599	597	593	593	593	603	596	598

Leave-p-Out with $p = 10$ cross-validation was used here, with each batch of 10 samples being drawn randomly from a pool without replacement. We observe that the results generally improve with an increased number of coefficients. We also observe that after a maximum around 15-22 the performance of the model is very stable.

Since cross-validation has a random component, the general rule is to pick the first result that has no later results which are significantly better. In this case, 19 is the number of components with the highest cross-validated discriminant efficiency. 15 is however still chosen as the optimal model, since the difference is not significant. Simpler models generally perform better and are generally less prone to random effects.

Unless otherwise noted, all tests in this chapter uses the same cross-validition, same LDA-call and same normalized fourier coefficients. Also, unless otherwise noted, any transforms are

applied to the normalized fourier coefficients.

6.2.2 Key scores for LDA Benchmark

LDA Confusion matrix for 15 fourier components

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	209	17	16	7	0	60	0	0
	BAR	54	248	30	35	1	39	0	1
	VEE	16	6	23	18	0	7	0	1
	VEW	5	8	16	10	0	9	0	0
	NOK	6	16	7	6	59	10	47	1
	POR	20	4	2	4	0	18	0	1
	SVA	0	0	0	0	13	1	18	0
	VAR	14	7	6	4	16	16	7	38

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.676	0.608	0.324	0.208	0.388	0.367	0.563	0.352

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.645	0.81	0.23	0.119	0.663	0.113	0.25	0.905

Error Rate

Overall error rate for this model is then $E = 0.471$.

The model performs very well on the two major groups of cod, however it is competely unable to distinguish between eastern and western vestfjorden (VEE and VEW). Also of note is that a rather large amount of cod is misclassified as POR.

6.3 On transforming the fourier coefficients

6.3.1 Sinusiods

As shown in 3.7:

$$F_x(t) = \sum_{n=1}^{N} (a_n \cos(nt) + b_n \sin(nt) = \sum_{n=1}^{N} (\alpha_n \sin(nt + \beta_n)) \qquad 0 < t < 2\pi \tag{6.1}$$

Described with words, the sum of a cosine and sine can be rewritten to a timeshifted sine. For harmonics with higher frequencies the timeshift should almost only be dependant on randomness, and it is therefore hoped that transforming A_n and B_n to amplitude and timeshift will separate noise from signal, and thus lead to better discrimination results.

Cross-validation however shows markedly worse results when using the transformed coefficients. The notion that separating these variables separates noise from signal thus seem incorrect.

6.3.2 Key scores for Sinusiod coefficients

Confusion matrix for Sinusiod coefficients

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	174	13	20	15	0	87	0	0
	BAR	70	155	66	71	0	41	1	4
	VEE	19	14	16	13	0	9	0	0
	VEW	7	10	10	11	0	9	0	1
	NOK	12	10	6	2	55	15	49	3
	POR	25	3	4	0	0	17	0	0
	SVA	1	0	0	0	17	1	13	0
	VAR	21	3	1	5	18	19	9	32

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.5631	0.380	0.225	0.229	0.362	0.347	0.406	0.296

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.529	0.745	0.130	0.0940	0.611	0.0859	0.181	0.800

Error Rate

Overall error rate for this model is then $E = 0.598$.

Performance has decreased markedly for the entire model. It does not seem like this transformation has any benefits.

6.3.3 Absolute Value

On a hunch it was tested whether taking the absolute value of the coefficients would improve discrimination. This was done due to a very simple concept. Numerically the positive x and the negative $-x$ are farther away from each other than from 0 (for $x \neq 0$). However looking at the contour from a biological perspective $x\sin(t)$ is more similar to $-x\sin(t)$ than either is to $0\sin(x)$.

Interestingly enough, it turns out that this actually improves discrimination rather markedly.

6.3.4 Key scores for Absolute Value

Confusion matrix for using absolute value on fourier coefficients

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	226	27	4	15	0	35	0	2
	BAR	41	280	22	25	2	36	0	2
	VEE	7	5	24	21	0	12	0	2
	VEW	8	2	18	14	0	5	0	1
	NOK	1	21	10	4	71	9	34	2
	POR	16	10	8	4	0	11	0	0
	SVA	0	0	0	0	15	1	16	0
	VAR	9	7	9	12	15	9	7	40

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.731	0.686	0.338	0.292	0.467	0.224	0.5	0.37

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.734	0.795	0.253	0.147	0.689	0.093	0.281	0.816

Error Rate

Overall error rate for this model is then $E = 0.420$.

Model performance was increased markedly in both the two major sets BAL and BAR, and slightly on the minor NOK,POR,SVA and VAR. There is still no classification between VEE and VEW. It thus seems highly likely that this should have been considered one population.

6.4 On using best fit rotation

6.4.1 No standardisation fix

Primarily to investigate what would happen LDA was run on raw pre-normalization fourier coefficients. This set has starting point selected somewhat arbitrarily by position in original the image, no orientation nor any size-standardization. These results should be regarded as highly suspicous.

6.4.2 Key scores for raw fourier coefficients

Confusion matrix for using absolute value on fourier coefficients

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	224	25	7	6	0	46	0	1
	BAR	63	266	32	32	0	13	1	1
	VEE	4	7	29	24	0	7	0	0
	VEW	2	6	17	17	0	6	0	0
	NOK	4	24	4	4	93	7	14	2
	POR	20	5	2	2	0	19	0	1
	SVA	0	0	1	0	19	0	10	2
	VAR	16	7	0	6	20	19	6	34

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.673	0.782	0.315	0.187	0.705	0.162	0.323	0.829

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.725	0.652	0.408	0.354	0.612	0.388	0.313	0.315

Error Rate

Overall error rate for this model is then $E = 0.412$, the best so far.

In this set the algorithm is allowed to pick up on systematic differences in otolith size and orientation. It was expected that this would improve results, however it is still somewhat surprising that the discrimination results are this good. It is probably not a cause for concern, as the relationship between otolith size, fish size/weight and catch location are quite strong.

6.4.3 Best Fit Rotation

Applying the best fit rotation algorithm outlined in 5.2 to the otoliths before applying LDA yields the following results:

6.4.4 Key scores for raw fourier coefficients

Confusion matrix for using absolute value on fourier coefficients

Actual Class \ Predicted Class	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
BAL	219	31	8	4	0	46	0	1
BAR	62	271	27	36	1	11	0	0
VEE	3	7	34	19	0	8	0	0
VEW	3	6	20	13	0	6	0	0
NOK	4	24	2	7	93	6	14	2
POR	18	4	2	3	0	22	0	0
SVA	0	0	0	0	19	1	10	2
VAR	22	3	2	5	20	16	5	35

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.662	0.783	0.358	0.150	0.699	0.190	0.344	0.87

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.709	0.664	0.479	0.271	0.612	0.449	0.313	0.324

Error Rate

Overall error rate for this model is then $E = 0.408$, which is slighty better than previous best.

Improvements are pretty even across the board, and no new phenoma were discovered. The model is still unable to discriminate between VEE and VEW.

6.5 PLS-DA

To investigate whether other discriminant functions would perform better, several methods were tested. From the PLS-family of functions, Partial Least Squares Regression - Discriminant Analysis, which performs PLS against categorical data was chosen. The implentation used was R's PLSDA (from Jed Wing et al., 2014).

6.5.1 Key scores for PLS-DA

Confusion matrix

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	201	19	11	9	0	69	0	0
	BAR	52	239	32	44	1	39	0	1
	VEE	14	7	23	17	0	9	0	1
	VEW	4	5	16	12	0	11	0	0
	NOK	7	18	3	6	59	11	47	1
	POR	21	4	1	3	0	20	0	0
	SVA	0	0	1	0	13	0	18	0
	VAR	18	5	7	2	16	16	8	36

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.650	0.586	0.324	0.250	0.388	0.408	0.562	0.333

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.634	0.805	0.246	0.129	0.663	0.114	0.247	0.923

Error Rate

Overall error rate for this model is then $E = 0.483$, which is comparable to initial result aquired when using LDA. PLS-DA shows very comparable behaviour to LDA on this dataset. It was hoped that PLS-DA would perform better, or atleast differently than LDA, but this seems not to be the case.

6.6 Decision trees

Decision trees are based on a making a series of discrete decision in order to classify a sample. These are represented as a tree, where one starts at the top and proceeds to an endnode. Decision trees are often visualized using dendrograms. The implentation used here was R's rpart (Therneau et al., 2012).

Since they are based on hard rules, they often perform better on data with strong clustering or stratification than more fuzzy distribution-based approaches do. It is expected that decision trees will perform poorly on these data, however it is hoped they will be able to pick up on some minor interesting dependencies.

6.6.1 Key scores for decision trees

Confusion matrix

		Predicted Class							
		BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
Actual Class	BAL	184	15	12	13	0	85	0	0
	BAR	40	215	32	48	1	71	0	1
	VEE	20	5	15	22	0	9	0	0
	VEW	8	5	13	15	0	7	0	0
	NOK	6	19	5	6	60	9	47	0
	POR	18	4	2	2	0	23	0	0
	SVA	0	0	0	1	12	0	19	0
	VAR	11	4	2	5	19	27	7	33

Probability that otolith of class X is classified as class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.595	0.527	0.211	0.313	0.395	0.469	0.594	0.306

Probability that otolith classified as class X belongs to class X

	BAL	BAR	VEE	VEW	NOK	POR	SVA	VAR
P	0.641	0.805	0.185	0.134	0.652	0.100	0.260	0.970

Error Rate

Overall error rate for this model is then $E = 0.520$, which was not surprising.

The model was not able to pick up on any differences in VEE/VEW. Performance was similar to LDA and PLS-DA, just worse.

Illustrations

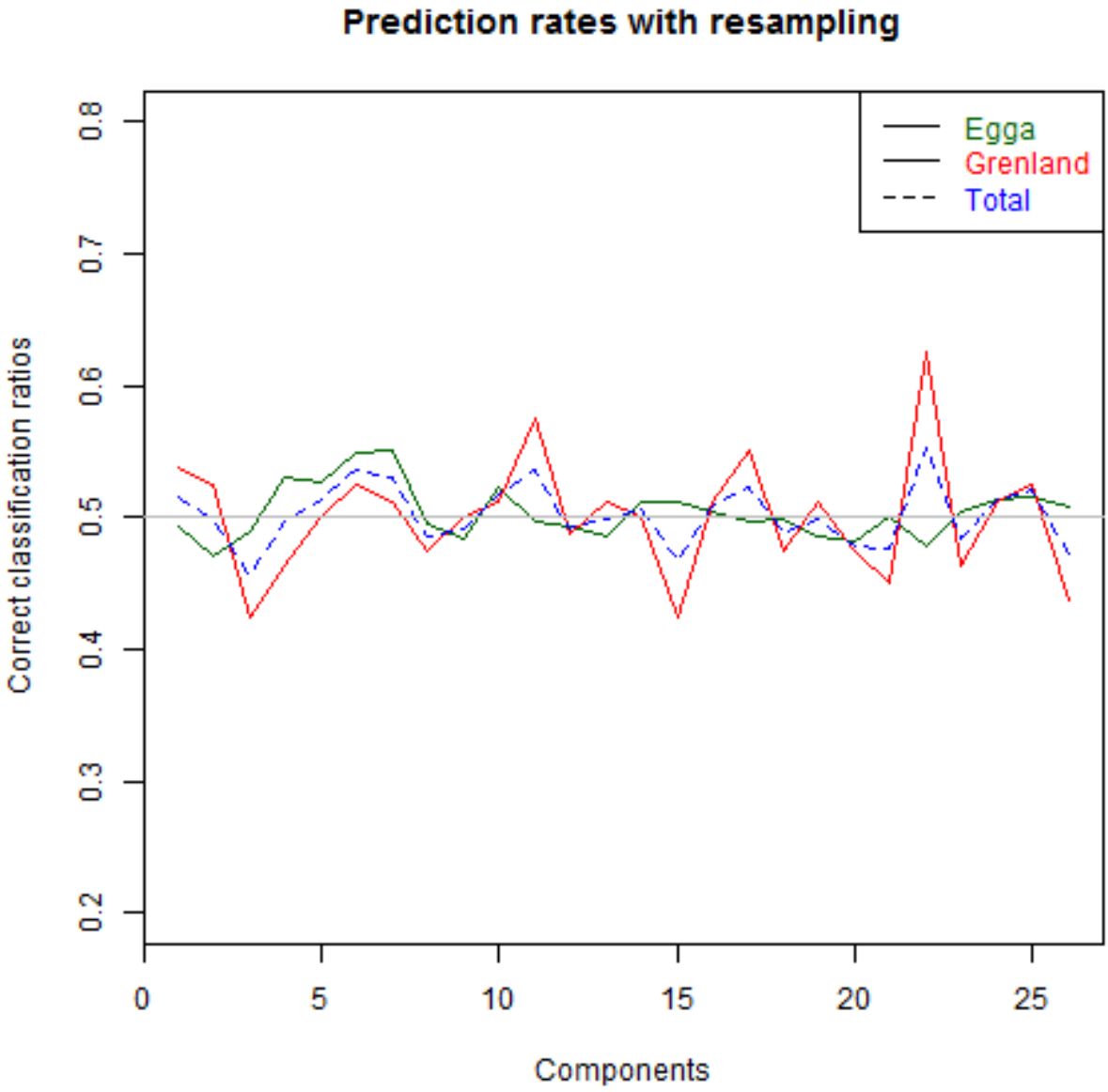


Figure 6.1: Ratio of correctly predicted Halibut from Eggakanten and Grenland, run on normalized sets of fourier coefficients using resampling to ensure comparable sets. Results pr component are independant as resampling has been run each time.

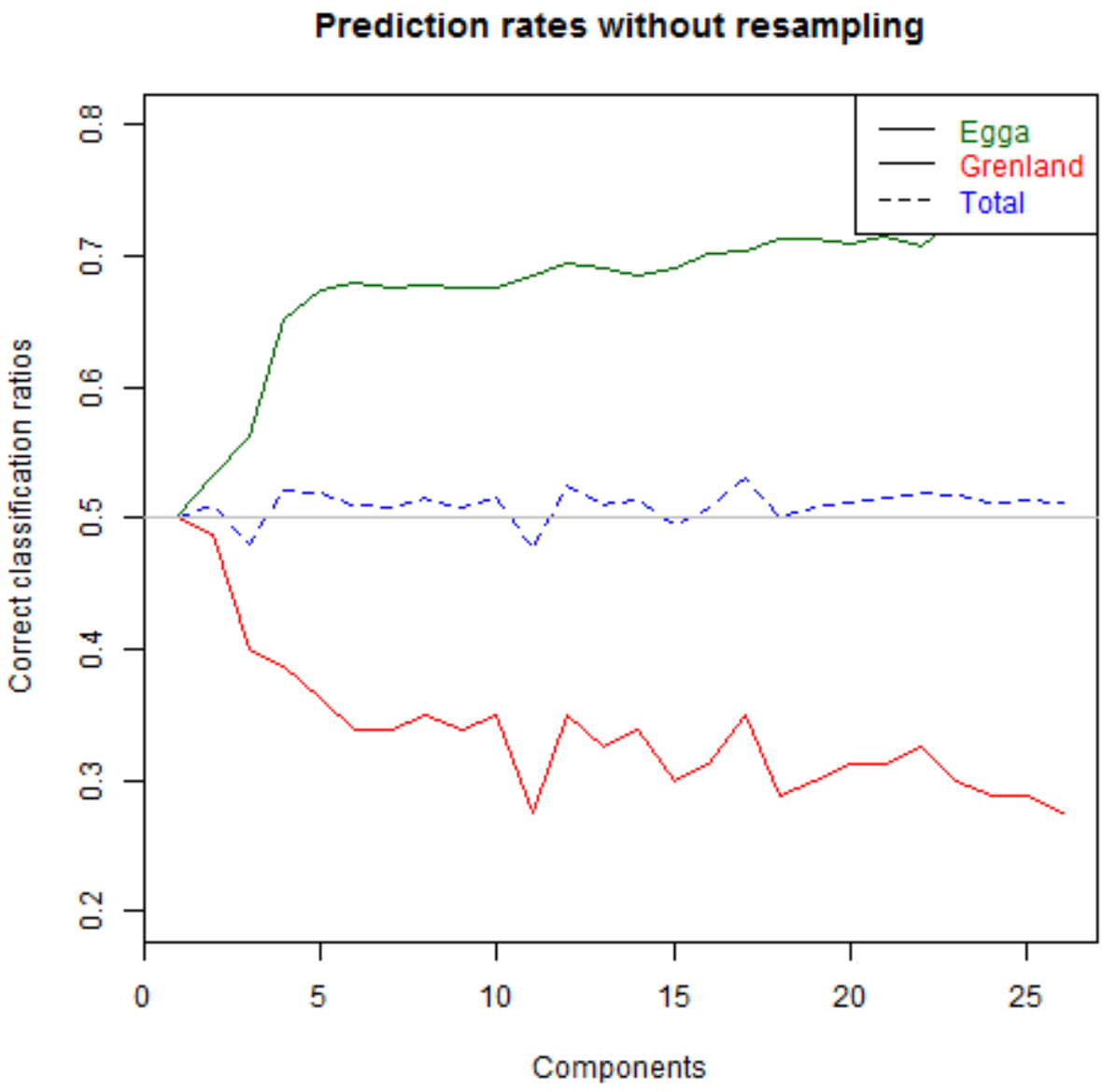


Figure 6.2: Ratio of correctly predicted Halibut from Eggakanten and Grenland, run on normalized sets of fourier coefficients. No resampling or adjustment done. Prior probability set to $(0.5, 0.5)$

Chapter 7

Summary and Recommendations for Further Work

7.1 Summary and Conclusions

The entire chain of aquiring contours with a scanner to final discriminant analysis has been investigated for potential sources of bias. The following has been found:

- The process of finding the contour from an image is robust and stable. Details of image formats and size, color represenation are largely irrelevant. Furthermore, during testing it was found that using wrong thresholds or other mistakes generally caused the process to outright fail, rather than introduce subtle biases.
- It has been found that there might be bias in which samples get invalidated through breakage. This may be a cause for subtle bias and may cause difficulty in reproducing results between sets and studies.
- A probability based resampling technique has been proposed instead using stratification. The algorithm works quite well, however it is not clear whether the added complexity is worth the fairly minor improvement.
- Different discrimination methods have been tested. Neither showed improvements over straight Linear Discriminant Analysis nor did any show any interesting characteristics. It

should also be noted that it was attempted to use Wavelets and Fourier resampling on the otoliths. These attemps gave no interesting results, but were not thourough enough to disprove their potential use. These findings were thus not included in this thesis.

- Different transformations on the fourier coefficients were tested. It was found that using the absolute value of the coefficients gave improved discrimination results. It was also found that numerically finding the best fit orientation of the otoliths after normalization further improved discrimination.

7.2 Recommendations for Further Work

It is recommended that other researchers test the findings on absolute value and numerically best orientation fit. It will be interesting to see if they hold for fourier contour analysis in general or just this set of data.

Appendix A

Code

A few selected algorithms and code-snippets have been included here. The entire codebase is available to anyone who wants, just email me at <reidar.hagen@gmail.com>. I honestly believe code is best read on a screen with proper a editor and proper formatting.

On that note, the images of the otoliths are not mine to give away. Permission to use them was given to me by Alf Harbitz.

A.1 Fourier Transform

```
fourier_transform = function(x,y,comps){
  results = matrix(0,1,4+4*comps)
  colnames(results) = c("SAMPLE","OTOLITH","A1","C1",1:(comps*4))

  N = length(x)
  dx = x - c(x[N],x[-N])
  dy = y - c(y[N],y[-N])
  dt = rep(1,N)

  # (t.to , t.from) is programming speak for t_n and t_(n-1)
  # so we need to adjust to is defined by t.from to t.to has value x[n]
  t.to = cumsum(dt)
```

```
  t.from = c(0,head(t.to,-1))

  # Tk equals T from Kuhl, since T is reserved for transpose in R
  Tk = t.to[N]
  x.previous = c(0,x[1:(N-1)])
  y.previous = c(0,y[1:(N-1)])

  # a0 and c0
  results[1,3] =  1/Tk * sum(dx/(2*dt)*(t.to^2-t.from^2)+(t.to-t.from)*(x.previous-dx/dt
  results[1,4] =  1/Tk * sum(dy/(2*dt)*(t.to^2-t.from^2)+(t.to-t.from)*(y.previous-dy/dt

  fx.comp = results[1,3]
  fy.comp = results[1,4]

  for(n in 1:comps){
    a  = Tk/(2*n^2*pi^2)*sum(dx/dt*(cos(2*pi*n*t.to/Tk)-cos(2*pi*n*t.from/Tk)))
    b  = Tk/(2*n^2*pi^2)*sum(dx/dt*(sin(2*pi*n*t.to/Tk)-sin(2*pi*n*t.from/Tk)))
    c  = Tk/(2*n^2*pi^2)*sum(dy/dt*(cos(2*pi*n*t.to/Tk)-cos(2*pi*n*t.from/Tk)))
    d  = Tk/(2*n^2*pi^2)*sum(dy/dt*(sin(2*pi*n*t.to/Tk)-sin(2*pi*n*t.from/Tk)))

    results[1,(n*4+1):(n*4+4)] = c(a,b,c,d)
  }
  return(results)
}
```

A.2 Normalize Fourier Coefficients

```
normfourier = function(coefficients){
  f = coefficients
```

```
comps = (length(f)-2)/4

x.a0 = f[1]
y.a0 = f[2]

df1 = data.frame(a = f[4*(1:comps)-1], b = f[4*(1:comps)],c = f[4*(1:comps)+1],d  = f[

# find new orientation
a = df1$a[1]; b=df1$b[1]; c=df1$c[1]; d=df1$d[1];
aa = -(b*d+a*c); bb=(d^2-b^2+c^2-a^2); cc=(b*d+a*c);
phi = atan((-bb+sqrt(bb^2-4*aa*cc))/(2*aa))

delta = atan( (b * cos(phi) + d * sin(phi)) / ((a * cos(phi) + c * sin(phi))))

rotate = function(df1,phi,delta){
  ns = 1:comps
  a = ( (df1$a * cos(phi) + df1$c * sin(phi)) * cos(ns * delta)  + (df1$b * cos(phi) +
  b = ( (df1$a * cos(phi) + df1$c * sin(phi)) * -sin(ns * delta) + (df1$b * cos(phi) +
  c = ( (df1$c * cos(phi) - df1$a * sin(phi)) * cos(ns * delta)  + (df1$d * cos(phi) -
  d = ( (df1$c * cos(phi) - df1$a * sin(phi)) * -sin(ns * delta) + (df1$d * cos(phi) -
  df1$a = a; df1$b = b; df1$c = c; df1$d = d;
  return(df1);
}
totalrotation = phi
df1 = rotate(df1,phi,delta)

# find correct direction of traversal
if(sign(df1$a[1]) != sign(df1$d[1])){
  df1$b = -df1$b;
  df1$d = -df1$d;
```

```
  }

  # find *a* major axis
  if(abs(df1$a[1]) < abs(df1$d[1])){
    df1 = rotate(df1,pi/2,pi/2);
    totalrotation = totalrotation + pi/2
  }

  # if a less than zero, change starting point to opposite side
  if(df1$a[1] < 0){
    df1 = rotate(df1,0,pi);
  }

  # normalize x-axis
  df1 = df1 / df1$a[1]

  # save result
  n = 1:comps
  f[1:2] = 0
  f[n*4-1] = df1$a[n]
  f[n*4] = df1$b[n]
  f[n*4+1] = df1$c[n]
  f[n*4+2] = df1$d[n]
  return(list(f,totalrotation))
}
```

A.3 Inverse Fourier transform

```
itransform = function(coefficients){
```

```
  a0 = coefficients[1]
  c0 = coefficients[2]
  cc = coefficients[3:length(coefficients)]
  comps = length(cc)/4
  x  = rep(0,2000) #a0
  y  = rep(0,2000) #c0

  tt = 0:1999
  TT = 2000
  for(n in 0:(comps-1)){
    x = x + cc[[1+n*4]] * cos(2*pi*(n+1)*tt/TT) + cc[[2+n*4]] * sin(2*pi*(n+1)*tt/TT)
    y = y + cc[[3+n*4]] * cos(2*pi*(n+1)*tt/TT) + cc[[4+n*4]] * sin(2*pi*(n+1)*tt/TT)
  }
  return(cbind(x,y))
}
```

A.4 Probabilty adjusted sampling

```
setwd("C:/Users/reidar/Desktop/FINAL");
library(np)
paths = read.table("DATA/paths.csv",header=TRUE,stringsAsFactors=FALSE)
centers = read.table("DATA/centers.csv",header=TRUE,stringsAsFactors=FALSE)
master =  read.table("DATA/master.csv",header=TRUE,stringsAsFactors=FALSE)
wl = read.table("DATA/wl.csv",header=TRUE,stringsAsFactors=FALSE)

group = master$GROUP[wl$SAMPLE]
g1 = (group==1)
g2 = (group==2)
```

```
t1 = data.frame(w=wl$WEIGHT[g1],l=wl$LENGTH[g1],s=wl$SEX[g1])
t2 = data.frame(w=wl$WEIGHT[g2],l=wl$LENGTH[g2],s=wl$SEX[g2])

bw1 = npudensbw(t1,bwmethod="normal-reference")
bw2 = npudensbw(t2,bwmethod="normal-reference")

# it makes more sense to use same bandwidth
p21 = fitted(npudens(bw2,t1))
p11 = fitted(npudens(bw1,t1))

acceptratio = (p21/p11)/max(p21/p11)
w = acceptratio / sum(acceptratio)

# MC-adjustment of weights to remove bias

p = w
d = as.matrix(scale(t1))
target = c(mean(wl$WEIGHT[g2]),mean(wl$LENGTH[g2]),mean(wl$SEX[g2]))
t = (target - attr(d,"scaled:center"))/attr(d,"scaled:scale")

error = colSums(diag(p) %*% d)-t
ntries = 0
N = length(p)
nums = 1:N

while(ntries < 1000){
  ntries = ntries + 1
  s = sample(nums,2)
  dp = runif(1,0,0.05) * p[s[1]]
  if(sum(abs(dp*d[s[2],]+-dp*d[s[1],]+error))<sum(abs(error))){
```

```
    p[s[1]] = p[s[1]]-dp
    p[s[2]] = p[s[2]]+dp
    error = colSums(diag(p) %*% d)-t
    ntries = 1
  }
}

write.table(p,"DATA/samplingdistr.csv",row.names=FALSE)
```

A.5 Find best rotation

```
setwd("C:/Users/reidar/Desktop/FINAL")

centers = read.table("DATA/centers.csv",header=TRUE,stringsAsFactors=FALSE)
master  = read.table("DATA/master.csv",header=TRUE,stringsAsFactors=FALSE)
paths   = as.matrix(read.table("DATA/approx_paths.csv",header=TRUE,stringsAsFactors=FALS

fourier_nrm = as.matrix(read.table("DATA/coefficients/fourier_nrm.csv",header=TRUE))
source("LIBRARY/fourierfunctions.R")

nrm_paths = matrix(0,2707,2000)
for(i in 1:210){
  nrm_paths[i,] = itransform_1000(fourier_nrm[i,3:244])
}

rot_paths = matrix(0,2707,2000)

shift = function(p,n){
  if(n==0){return(p)}
  return(cbind(c(tail(p[,1],-n),head(p[,1],n)),c(tail(p[,2],-n),head(p[,2],n))))
```

```
}

 # sum((shift(t,s)-p)^2)
score_most_shifts = function(p,t){
  score = function(s){sum(sqrt(rowSums((shift(t,s)-p)^2)))}
  return(min(sapply((0:99)*10,score)))  # check every 10 shifts
}

get_best_shift = function(p,t){
  score = function(s){sum(sqrt(rowSums((shift(t,s)-p)^2)))}
  return(which.min(sapply(0:999,score))) # check every possible
}

rotate = function(t,rot){
  return(cbind(cos(rot)*t[,1]-sin(rot)*t[,2],sin(rot)*t[,1]+cos(rot)*t[,2]))
}

find_best_rotation = function(p,t){
  score = function(i){score_most_shifts(p,rotate(t,i))}
  return((which.min(sapply((0:179)/180*2*pi,score))-1)/180*2*pi)
}

for(i in 2:2707){
  t = cbind(paths[i,1:1000],paths[i,1001:2000])
  paths[i,]= c(shift(t,get_best_shift(p,t)))
}

p = colMeans(nrm_paths)

for(i in 1:10){
```

```
  print(i)
  t = cbind(nrm_paths[i,1:1000],nrm_paths[i,1001:2000])
  rot = find_best_rotation(p,t)
  t = rotate(t,rot)
  t = shift(t,get_best_shift(p,t))
  rot_paths[i,] = c(t)
}

plot(rot_paths[1,1:1000],rot_paths[1,1001:2000],type="l")
for(i in 2:10){lines(rot_paths[i,1:1000],rot_paths[i,1001:2000])}

write.table(rot_paths,"DATA/approx_paths_rotated.csv",row.names=FALSE)
```

A.6 LDA with cross-validation

```
setwd("C:/Users/reidar/Desktop/FINAL - Torsk/")

library(MASS)
library(R.matlab)

group = as.factor(readMat("IMAGES/data/dat1177.mat")$dat1177[,7])
fourier.nrm = read.table("DATA/fourier_nrm.csv",header=TRUE)

N = dim(fourier.nrm)[1]

random_draw = list()

available = 1:N
ngroups = (N/10)
for(i in 1:ngroups){
```

```
random_draw[[i]] = sample(available,N/ngroups)
available = setdiff(available,random_draw[[i]])
}
if(available){
random_draw[[ngroups+1]] = available
ngroups = ngroups + 1
}

comp_score = rep(0,50)
for(comps in 2:50){
  scores.nrm = matrix(0,8,8)
  for(i in 1:ngroups){
    s = random_draw[[i]]
    lda_set = cbind(group[-s],fourier.nrm[,1:(2+comps*4)][-s,-c(1:5)])
    colnames(lda_set)[1] = "GROUP"
    res = lda(GROUP ~ . , lda_set ,prior=(rep(1,8)/8))
    g1 = as.numeric(group[s])
    p1 = as.numeric(predict(res,fourier.nrm[s,-c(1:5)])$class)
    for(j in 1:length(p1)){
      scores.nrm[g1[j],p1[j]] = scores.nrm[g1[j],p1[j]] + 1
    }
  }
  comp_score[comps] = sum(diag(scores.nrm))
}
print(comp_score)
comps = which.max(comp_score)

scores.nrm = matrix(0,8,8)
for(i in 1:ngroups){
  s = random_draw[[i]]
```

```
  lda_set = cbind(group[-s],fourier.nrm[,1:(2+comps*4)][-s,-c(1:5)])
  colnames(lda_set)[1] = "GROUP"
  res = lda(GROUP ~ . , lda_set ,prior=(rep(1,8)/8))
  g1 = as.numeric(group[s])
  p1 = as.numeric(predict(res,fourier.nrm[s,-c(1:5)])$class)
  for(j in 1:length(p1)){
    scores.nrm[g1[j],p1[j]] = scores.nrm[g1[j],p1[j]] + 1
  }
}
print(scores.nrm)
```

Bibliography

(2009). *Sunken Billions*. The World Bank / FAO, ISBN 978-0-8213-7790-1.

(2011). *Photographic Atlas of Otoliths (Sagittae)*. Rubin Sanson, Kristi Tangevold Sanson.

(2012). *Fao Yearbook. Fishery and Aquaculture Statistics*. Food and Agriculture Organization of the UN.

A.H. Weatherley, A. G. (1987). *The Biology of Fish Growth*. Academic Press, London (1987), pp. 209–242.

Cadík, M. (2008). Perceptual evaluation of color-to-grayscale image conversions. *Pacific Graphics 2008*.

Farrell, J. and Campana, S. E. (1996). Regulation of calcium and strontium deposition on the otoliths of juvenile tilapia, oreochromis niloticus. *Comparative Biochemistry and Physiology Part A: Physiology*, 115(2):103 – 109.

from Jed Wing, M. K. C., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., and the R Core Team (2014). *caret: Classification and Regression Training*. R package version 6.0-24.

Haines, A. J. and Crampton, J. S. (2000). Improvements to the method of fourier shape analysis as applied in morphometric studies. *Palaeontology*, 43(4):765–783.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Henriksen, A. M. (2013). Discrimination of cod otolith shapes by two different fourier methods.

Kålås, J.A., and Viken, Å. og Bakken, T. (2006). *Norsk Rødliste – Norwegian Red List.* Artsdatabanken, Norway.

Kuhl, F. P. and Giardina, C. R. (1982). Elliptic fourier features of a closed contour. *Computer graphics and image processing*, 18(3):236–258.

Morales-Nin, B.

Omer, I. and Werman, M. (2004). Color lines: image specific color representation. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–946–II–953 Vol.2.

Parisi-Baradad, V., Manjabacas, A., Lombarte, A., Olivella, R., Chic, ., Piera, J., and García-Ladona, E. (2010). Automated taxon identification of teleost fishes using an otolith online database—aforo. *Fisheries Research*, 105(1):13 – 20.

Popper, A., Fay, R., Platt, C., and Sand, O. (2003). Sound detection mechanisms and capabilities of teleost fishes. In Collin, S. and Marshall, N., editors, *Sensory Processing in Aquatic Environments*, pages 3–38. Springer New York.

Stockman, G. and Shapiro, L. G. (2001). *Computer Vision.* Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition.

Stransky, C., Baumann, H., Fevolden, S.-E., Harbitz, A., Høie, H., Nedreaas, K. H., Salberg, A.-B., and Skarstein, T. H. (2008). Separation of norwegian coastal cod and northeast arctic cod by outer otolith shape analysis. *Fisheries Research*, 90(1–3):26 – 35.

Therneau, T., Atkinson, B., and Ripley, B. (2012). *rpart: Recursive Partitioning.* R package version 3.1-55.

Toussaint, G. Grids, connectivity, and contour-tracing. *URL:< http://www-cgrl. cs. mcgill. ca/˜ godfried/teaching/pr-notes/contour. ps.*