

# Otolith shape and size: The importance of age when determining indices for fish-stock separation.

James Mapp<sup>a</sup>, Ewan Hunter<sup>bc</sup>, Jeroen Van Der Kooij<sup>c</sup>, Sally Songer<sup>c</sup>, Mark Fisher<sup>a</sup>.

Affiliations:

<sup>a</sup> School of Computing Sciences, University of East Anglia, Norwich, Norfolk, NR4 7JT, UK.

<sup>b</sup> School of Environmental Sciences, University of East Anglia, Norwich, Norfolk, NR4 7JT, UK.

<sup>c</sup> Centre for the Environment, Fisheries and Aquaculture Science, Pakefield Rd, Lowestoft, Suffolk, NR33 0HT, UK

Corresponding author

{j.mapp, mark.fisher}@uea.ac.uk, {ewan.hunter, jeroen.vanderkooij, sally.songer}@cefas.co.uk

## Abstract

Stock-separation of highly mobile Clupeids (sprat - *Sprattus sprattus* and herring - *Clupea harengus*) using otolith morphometrics was explored. Analysis focused on three stock discrimination problems with the aim of reassigning individual otoliths to source populations using experiments undertaken using a machine learning environment known as WEKA (Waikato Environment for Knowledge Analysis). Six feature sets encoding combinations of size and shape together with nine learning algorithms were explored. To assess saliency of size/shape features half of the feature sets included size indices, the remainder encoded only shape. Otolith sample sets were partitioned by age so that the impact of age on classification accuracy could be assessed for each method. In total we performed 540 experiments, representing a comprehensive evaluation of otolith morphometrics and learning algorithms. Results show that for juveniles, methods encoding only shape performed well, but those that included size indices held more classification potential. However as fish age, shape encoding methods were more robust than those including size information. This study suggests that methods of stock discrimination based on early incremental growth are likely to be effective, and that automated classification techniques will show little benefit in supplementing early growth information with shape indices derived from mature outlines.

## Keywords

Fourier Descriptors; Curvature Scale Space; Otolith shape; Otolith size; Stock discrimination

# 1 Introduction

Otoliths are the calcium carbonate structures forming the inner ear of many vertebrates. Teleost fish have three types of otoliths, the larger sagittal otolith being the most commonly used for aging and classification studies, as they are easier to prepare, observe, and measure (Campana and Casselman 1993). Otolith morphology varies markedly between species, however separate stocks of the same species, often identical physically, can sometimes be discriminated through subtle differences in otolith morphometrics. Expert otolith readers have drawn on otoliths to discriminate between: different ages or cohorts (Burke et al. 2008; Burke et al. 2009); sex (Cardinale et al. 2004); diet (Gagliano and McCormick 2004) and of course stock (Begg and Brown 2000; Begg et al. 2001; Brophy and Danilowicz 2002; Mériqot et al. 2007; Duarte-Neto et al. 2008). Some of these distinctions are more complex and more important to fisheries management (Begg et al. 2005), which requires accurate measurements of stock composition/mixing, or stock movement, to inform decision making (Stransky 2005).

Shape analysis forms a major part of otolith science, and many studies have analysed shape with a view to separating stocks. Methods include statistical analysis of general shape parameters such as circularity, eccentricity, area, perimeter length, form-factor, and annular growth increments. (Burke et al. 2008; Burke et al. 2008; Paul et al. 2013; Keating et al. 2014). In some cases these measurements are supplemented with, or normalised by, measures such as fish length or weight (DeVries et al. 2002; Stransky 2005; Mériqot et al. 2007). Otolith boundaries are also extracted and represented, or encoded in different ways (transformed) prior to analysis with methods such as Fourier transforms (Begg and Brown 2000; Galley et al. 2006; Bani et al. 2013); and Elliptical Fourier transforms (Campana and Casselman 1993; Burke et al. 2008; Duarte-Neto et al. 2008). Other methods of otolith boundary representation include Wavelets (Parisi-Baradad et al. 2005), Curvature-Scale-Space (Begg et al. 2005; Parisi-Baradad et al. 2005) and the more recent Shapelet transform method (Lines et al. 2012; Mapp et al. 2013; Hills et al. 2014).

Unlike previous studies, where otolith morphology has been used to achieve stock separation, here we approach the problem of fish stock discrimination from a machine learning standpoint, applying tools used for pattern classification in other biological fields, such as microscopy image analysis (Shamir et al. 2010), and (human) bone age classification (Bagnall and Davis 2014). Machine learning methods have been used previously for otolith age classification (Fablet and Le Josse 2005; Bermejo et al. 2007), and underlying methods have been used for stock separation with varying success.

The aim of this study is *not* focused on the absolute classification accuracy of stock identification per se. Rather we compare the viability of morphometric methods that include, and those that exclude, size information as representations of otolith boundary, in an attempt to assess their potential efficacy for application in fish stock management. To do this three methods of otolith boundary representation (or 'data transforms') were compared: Elliptical Fourier Descriptors; Curvature Scale Space; and a method that represents otolith boundary as basic shape indices (e.g. area, perimeter length). The methods for comparison were chosen based on their prior

application, and the immutability of transform methods (or parameters), regardless of the data to be transformed.

The Waikato Environment for Knowledge Analysis, (WEKA (Hall et al. 2009)) is a freely available library of machine learning tools which are widely adopted by the machine learning community, and which have previously been used for otolith classification (Hills et al. 2014). The library comprises many statistical and modelling tools together with learning algorithms. WEKA provides a framework allowing the construction of large numbers of classifiers, using different learning algorithms that can be used alone or can be integrated into other software packages. For this study we make extensive use of WEKA's library to build multiple classifiers of otolith shape, using the following learning algorithms which represent the spectrum of classifier types: Naive Bayes, Bayesian Networks, Logistic, HyperPipes, J48/C4.5, Random and Rotation Forests, Nearest Neighbour, and Support Vector Machines (Brownlee 2013).

To investigate whether otolith shape or size has more potential to discriminate between the otolith characteristics of different stocks, we conducted three stock-discrimination experiments: Separation of Irish-Sea and Celtic-Sea herring; Separation of North-Sea/Thames Herring; and separation of North-Sea/Western-Channel sprat. Each problem was subdivided into single-age sets so that each experiment was carried out within a single age-class. Otoliths from herring captured in the Irish/Celtic Seas belong to one of two spawning stocks: Autumn spawning (Irish-Sea) and autumn/winter spawning (Celtic-Sea). Irish and Celtic Seas stocks have previously shown to be separable using otolith morphology (Burke et al. 2008). Herring caught in the southern North-Sea may belong to one of two stocks: Thames herring (spring spawning) and North-Sea herring (autumn/winter spawning). As with The Celtic/Irish herring these have been previously shown to be separable using otolith morphology (Mapp et al. 2013). Two spawning components compose the third set: Western Channel sprat, spawning in spring and captured in the English Channel; and North-Sea sprat which spawn in the spring to late summer and were captured in the southern North-Sea.

## 2 Materials and methods

Machine learning aims to build generic models (*classifiers*) with which to classify further otolith samples. Each classifier is constructed using *training data* and a *learning algorithm*. Whilst the algorithms are sometimes called classification algorithms, they perform no classification themselves. Rather they are used to build the classifiers, which are then used to classify new (otolith) samples from the *test data*.

A graphical representation of dataflow through our system is shown in Figure 1, which shows the flow from each source dataset (where images are already represented as rotation-normalised outlines), through boundary transformation, model construction and testing.

Whilst this study combines several complex methods of otolith boundary representation and data classification, in-depth knowledge of each of the methods is not required, and description of all transform methods and classification algorithms in detail would not be feasible. We therefore give brief overviews of methods used in the relevant sections, and refer to previous texts for further detail of methods.

## 2.1 Otolith Datasets

We were able to source three matched datasets of clupeid otolith images (one sprat, and two herring), each representing discrete paired spawning stocks which cannot easily be discriminated visually (e.g. by visual inspection of external morphology), two of which (herring sets) have previously been reported as having distinct otolith morphology (Burke et al. 2008; Mapp et al. 2013).

Age-Set	Contents of Set	Samples	Reference & Notes
<b>BB0</b>	Age-0 Herring set. Celtic/Irish Sea	40/class	Trawls conducted by the Agri-Food and BioSciences Institute (AFBI) & the Marine Institute (MI) (Burke et al. 2008)
<b>BB1</b>	Age-1 Herring set. Celtic/Irish Sea	13/class	
<b>S0</b>	Age-0 Sprat set. North-sea/Western Channel	40/class	ICES Surveys 2013/2014. (ICES 2013; ICES 2014). Age-0 set partially created from age-1 samples using trace of the age-0 annuli
<b>S1</b>	Age-1 Sprat set. North-sea/Western Channel	25/class	
<b>S2</b>	Age-2 Sprat set. North-sea/Western Channel	11/class	
<b>S3</b>	Age-3 Sprat set. North-sea/Western Channel	13/class	
<b>H2</b>	Age-2 Herring set. North-sea/Thames	11/class	
<b>H3</b>	Age-3 Herring set. North-sea/Thames	27/class	ICES survey 2008/2009. (Mapp et al. 2013)
<b>H4</b>	Age-4 Herring set. North-sea/Thames	21/class	
<b>H5</b>	Age-5 Herring set. North-sea/Thames	15/class	

Table 1: Table showing species, provenance and number of each 'age-set' and their sources. 'Age-set' is the descriptive identifier given to each set, e.g. BB1 refers to the age-1 partition of the Celtic-Sea/Irish-Sea Herring. See text for further explanations.

Each datasets was subdivided into single-age subsets, creating ten 'age-sets' (Table 1, Figure 1). This creates ten smaller stock discrimination problems which are assessed during this study. For example, we attempt to separate age-0 sprat from the North-Sea and Western Channel using each of the described transform methods combined with each machine learning algorithm, attempt to separate age-1 sprat from the North-Sea/Channel using each method etc.

## 2.2 Segmentation/Boundary extraction

Boundaries were extracted from otolith images using simple supervised thresholding, scripted in MATLAB software. Images were converted to greyscale and binarised using an automatically chosen threshold (Otsu 1975). Further refinement was attempted using histogram back-projection (Swain and Ballard 1991) with strict confidence thresholds. The resulting mask was then displayed, and the process supervisor could make interactive adjustments e.g.: change the initial threshold; discard the histogram refinement; or manually amend the mask. In extreme cases where the outline is poorly defined, the supervisor outlined the entire otolith manually. Due to otolith image variability across datasets, it was not possible to construct a uniform process to perform adequately on all sets without pre/post supervision. The sample sets were captured and processed by different research groups, and stored/preserved using a variety of techniques. Image capture was performed under variable lighting conditions and using different equipment, which is not currently standardised between otolith laboratories. Examples of image variability are shown in Figure 2.

Masks are flipped and/or rotated so that the ventral edge is uppermost, the otolith rostrum to the left of the image. The major axis of the mask is calculated using MATLAB's 'regionprops' function, and used to align images so that all are normalised for rotation. Whilst this is the standard method of normalising outlines for Elliptical Fourier analysis, performing the normalisation prior to boundary extraction ensured that the same boundary orientation was used for each of the transform methods. This avoids any potential transform bias based on orientation of the boundary.

Once standardised, each mask was processed using a morphological 'closing' to remove noise and smooth the edges. The boundaries were extracted by performing a logical XOR function between the denoised masks and their morphological dilation with a 3x3 structuring element (Gonzales and Woods 2002) which removes interior pixels. This results in 4-connected boundaries (where all pixels are adjacent to each other orthogonally) which were superimposed over the original images and displayed (so that corrections could be made). Otolith boundaries were then stored for further processing as series of ordered coordinate pairs, ordered anti-clockwise from the upper-leftmost points of the boundaries.

## 2.3 Transform Selection and methods

Before attempting to build classifiers with which to separate classes, it is often advantageous to represent otoliths in a more manageable manner through data transformation. An obvious first transform is to simply represent the images digitally, and then extract the outline of each otolith (Section 2.2). Using further data transforms allows representation of the otolith boundaries in different ways, often simplifying or reducing the information presented. For example, a simple method

(tested in this study) measures shape indices such as the length and width, thus representing each otolith by just two numbers.

Three 'static' or 'non-data-adaptive' transformation methods previously used for otolith shape analysis were selected (Bagnall et al. 2006). Static transforms are defined as those where the transformed version of each boundary is dependent only on that instance of transformation, and not on other samples in the dataset. For example, Principle Component analysis (PCA (Pearson 1901)) relies on knowledge of other samples and their classes in order to set parameters by which to transform each individual instance, and are therefore 'data-adaptive' or 'non-static'.

The application of individual transform methods to each of the single-age otolith sets, creates multiple 'train/test' sets which were used to build classifiers using a selection of learning algorithms. In total sixty train/test sets were created using these methods. This constitutes six sets for each of the ten single-age 'age-sets', seen under 'train/test sets' in Figure 1. Half of the train/test sets held size or size and shape data (size-inclusive sets), the remaining half contained only shape data (size-exclusive sets).

### **2.3.1 Curvature Scale-Space (CSS)**

Curvature Scale Space (Mokhtarian and Mackworth 1992) forms the basis for contour based shape descriptors as part of the MPEG-7 standard (Bober 2001; Zhang and Lu 2003; Zhang and Lu 2003). As such it is an ideal starting point for boundary based shape classification of otoliths and has been used for several other studies in this field (Abbasi et al. 1999; Parisi-Baradad et al. 2005; Jalba et al. 2006). Research has shown that CSS encoding can be an effective and robust (to noise, scale and rotation) method of matching query images to database instances, when combined with global parameters such as *circularity* and *eccentricity* (Abbasi et al. 1999; Amanatiadis et al. 2011).

The CSS process iteratively smooths a boundary contour until it is entirely convex. As each concavity disappears during the smoothing process, its location is recorded as both the point along the boundary from which it disappeared, and the smoothing iteration or 'evolution' which removed it. This produces a set of variables which can be projected onto the original boundary to show approximation of location and magnitude of boundary concavities. A simplified example of this projection (landmarking) can be seen in Figure 3 (smaller concavities not landmarked).

For the purpose of this study we used a built-in-house CSS transform system (Mapp et al. 2013). Boundaries were first subsampled to five-hundred points, and transformed by the system using methods described by (Abbasi et al. 1999). This produced a set of maxima coordinate-pairs (distance along boundary, evolution

magnitude) for each otolith, which were fully invariant to scale (size-exclusive). We constructed two datasets from the resulting transformed boundaries using only the largest seventeen (by evolution magnitude) maxima. As the learning algorithms require an equal number of indices per instance we only used the minimum number of maxima (17) held by any transformed boundaries in that image-set. The methods of ordering the data in the train/test sets are given below and an example shown in Table 2:

- bou – CSS maxima co-ordinate pairs; ordered according to point along the boundary. (see Table 2)
- evo – CSS maxima co-ordinate pairs; ordered according to evolution magnitude.

---

**Set of maxima feature-pairs extracted from example in Figure 3**

---

(27,81), (57,271), (108,58), (169,139), (218,121), (265,74), (323,68), (340,125), (365,89), (401,159), (447,113)

---

ordering: by boundary (bou)

**27, 81, 57, 271, 108, 58, 169, 139 ...**

ordering: by evolution magnitude (evo)

**57, 271, 401, 159, 169, 139, 340, 125 ...**

---

Table 2: Showing the CSS feature pairs (point along boundary, evolution magnitude) for the boundary (top); the ordering of the pairs using boundary order, and evolution magnitude ordering methods (bottom).

### 2.3.2 Elliptical Fourier Descriptors (EFDs)

Elliptical Fourier descriptors are commonly used for class-separation by otolith boundary, and many studies use this method for stock discrimination (Bird et al. 1986; Doering and Ludwig 1990; Castonguay et al. 1991). Burke et al. (2008) successfully discriminate two populations of herring using selected elliptical Fourier coefficients. However in contrast to Burke et al. (2008) we did not perform any statistical selection of harmonics/coefficients, as this was performed by the learning algorithms we employed.

To transform the boundaries a modified version of MATLAB scripts for generating EFDs (available on the MATLAB file exchange website (Manurung 2011)) was used, generating harmonics for each boundary. As boundaries were pre-normalised for rotation at the extraction phase, we needed only normalise for translation, and normalise coefficients for scale.

Each otolith boundary, when reconstructed using its first fifty Fourier harmonics, showed a mean pixel disparity between reconstruction and original boundary contour <1. We therefore generated only the first fifty harmonics for each outline and recorded the transformed boundaries both before and after scale-normalisation, so

that both size-variant and size-invariant harmonics could be assessed. We use the generated harmonics to construct two train/test sets (per age-set) for classification:

- Fou50 – First fifty Fourier Harmonics, non-normalised for scale (size-inclusive) ordered  $H_1C_1, H_1C_2, H_1C_3, H_1C_4, H_2C_1, \dots, H_{50}C_4$ . Where H = Harmonic; C = Coefficient.
- Fou50n – First fifty Fourier harmonics; ordered as above but normalised for scale. (size-exclusive)

### 2.3.3 Shape Parameters (SPa)

Each age-set was transformed using simple shape/size measurements of the boundary and its enclosed region (SPa transform). The majority of this study is conducted using MATLAB, and so we use the built-in region properties function ('regionprops') to generate the following metrics from the otolith mask (the 'filled' boundary): Perimeter Length; Area (of the enclosed region); Convex Area (the area of the convex hull); Solidity (Area/Convex Area); Height (of the calculated bounding box); Width (of the calculated bounding box); Extent (Area/(Height\*Width)); Major Axis Length; Minor Axis Length; Eccentricity (ratio of the distance between the foci of the ellipse and major axis length) and Equivalent Diameter (of a circle with the same area). We discard any variables generated by the regionprops method which are not mentioned, as they are either unusable for classification or hold redundant data. Two train/test sets (per age set) were compiled using the transformed series. Each set contained size information for the boundaries (size-inclusive) and are denoted:

- Box – Sub-selection of the SPa transform; Retaining only Height and Width.
- STAT – Full Statistic block; ordered: Perimeter, Area, ConvexArea, Solidity, Height, Width, Extent, MajorAxisLength, MinorAxisLength, Eccentricity, EquivDiameter.

## 2.4 Learning Algorithms

The Train/test sets were used to build and test classifiers using a number of available algorithms to compare and assess the transformation methods. Whilst Curvature Scale Space has a unique matching algorithm, we proceeded to classify CSS transformed otoliths using the same algorithms as for the other train/test sets (created using EFDs/SPa) so that we directly compared transform methods and not the associated matching algorithm.

Learning algorithms by which to train classifiers were selected from the Waikato Environment for Knowledge Analysis (WEKA (WEKA 2002)) and implemented using

MATLAB's java interface. In this process each otolith from a set was removed sequentially, Classifiers trained using remaining otoliths in that set (the training set), before the removed otolith (the test set) was classified. This process, known as 'leave one out cross validation' (LOOCV) is widely accepted as good practice within the machine learning community.

There are a number of algorithms available. Each algorithm learns from, or processes, training data in its own way. Learning algorithms fall into two predominant categories, 'supervised' and 'unsupervised' learning. As with previous studies we concentrate on supervised learning, meaning that the class membership of otoliths in the training set is known by the learning algorithms we use. We build 540 classifiers using each of the 60 train/test sets described in conjunction with each of the following nine learning algorithms (using each algorithms default settings), they are denoted as:

- **NB** – Naive Bayes (Lewis 1998)
- **BN** – Bayesian Networks (Friedman et al. 1997)
- **Log** – Logistic Regression (Le Cessie and Van Houwelingen 1992)
- **HP** – HyperPipes (WEKA 2002)
- **J48** – J48/C4.5 (Quinlan 1993)
- **RaF** – Random Forest (Breiman 2001)
- **IBk** – k-Nearest Neighbours (Aha et al. 1991)
- **SMO** – Support Vector machine (Platt 1999)
- **RoF** – Rotation Forest (Rodriguez et al. 2006)

Classification algorithms are further grouped according to similarity of methods used to construct classifiers. The algorithms above were chosen to cover a range of the classifier groups. Both Naive Bayes and Bayesian Networks algorithms belong to the *Bayesian* group of classifiers, which construct classifiers by explicitly applying Bayes' Theorem. Logistic Regression belongs to the *regression* group, which iteratively refines the classifier measures of error in predictions made by the classifier. Hyperpipes and k-Nearest Neighbour algorithms are *instance based methods*, also called *memory-based* as they compare new instances to those within the training data to calculate similarities with 'previous' cases. J48, Random forest and Rotation Forest are *decision tree* algorithms. Each constructs classifiers that make series of 'decisions' based on new instances' variables (harmonic coefficients, CSS maxima etc.). Additionally, random and rotation forest algorithms also qualify as *Ensemble Methods*, where multiple (potentially) weaker classifiers are used independently, and where a further decision based on their predictions is used to classify the new instance. Finally Support Vector Machines are a *kernel method* where input data is used to create a higher dimensional information space which is then used to classify new instances.

## 2.5 Statistical Testing

To test whether otolith age, or retaining size information during transformation returns distinct results, we performed statistical testing on our results using two methods.

Mann-Whitney U-tests were performed for results for each age-set ( $p > 0.05$ ). Size-inclusive transform results are assigned to one group, and size-exclusive results to the other group. Tests that reject the null hypothesis show significant difference between size-inclusive and size-exclusive methods for that age-set.

N-way analysis of variance testing was performed, to determine which factors (learning algorithm, otolith age, boundary transformation) significantly impact classification accuracies. Associated post-hoc testing is used to graphically represent ANOVAN results.

## 3 Results

### 3.1 Relative Performance of Learning Algorithms

When comparing learning algorithms across ages, no recognisable patterns were found, and results do not show any significant difference between learning algorithms used for modelling the data using any of our age-sets, and can be seen in Table 3 ( $p = 0.7107$ ). Figure 4 shows results of ANOVAN post-hoc testing for the learning algorithms across all age categories, and the average result (across learning algorithms) for classification of each train/test set is shown in Table 4.

	<b>Sum Sq</b>	<b>d.f.</b>	<b>Mean Sq</b>	<b>F</b>	<b>Prob&gt;F</b>
<b>age</b>	55176.3	5	11035.3	81.93	0.0000
<b>transform</b>	2795.5	5	559.1	4.15	0.0011
<b>algorithm</b>	731.1	8	91.4	0.68	0.7107
<b>age*transform</b>	18335.9	25	733.4	5.45	0.0000
<b>age*algorithm</b>	2894.0	40	72.4	0.54	0.9911
<b>transform*algorithm</b>	3324.1	40	93.1	0.62	0.9688
<b>Error</b>	56034.5	416	134.7		
<b>Total</b>	142390.0	539			

Table 3: Results of N-way analysis of variance (ANOVAN) of all results across: otolith age, boundary transform method, and learning algorithm. Also shown are interaction tests between factors.

### 3.2 Size-inclusive Vs Size-exclusive Transforms

Table 3 shows that different boundary transform methods produce classifiers which return significantly different stock separation accuracies ( $p=0.00$ ). Further variance testing of results shows that groupings of size-inclusive and size-exclusive transform results are also statistically distinct across all age categories ( $p=0.00$ ). Checkmarks in Table 4 show age-sets that reject the null hypothesis ( $p>0.05$ ) and therefore show significant difference between size-inclusive and size-exclusive transform results. These sets are plotted in Figure 5 and Figure 6 using solid markers; whilst those that accept the null hypothesis (that results between groups are indistinct) are shown using 'open' markers (both at  $p>0.05$ ).

	<b>BB0</b>	<b>BB1</b>	<b>S0</b>	<b>S1</b>	<b>S2</b>	<b>S3</b>	<b>H2</b>	<b>H3</b>	<b>H4</b>	<b>H5</b>
<b>bou</b>	64.3	46.2	56.9	57.3	49.5	47.9	49.0	49.0	54.2	54.4
<b>evo</b>	68.8	56.4	64.4	64.2	50.0	53.0	52.5	52.3	46.6	54.1
<b>Fou50n</b>	81.5	54.7	60.1	68.0	47.5	63.2	42.9	49.0	54.2	53.3
<b>Fou50</b>	81.9	61.1	61.7	76.7	34.8	53.4	50.5	45.5	43.9	52.6
<b>Box</b>	98.5	79.5	65.0	85.8	56.6	46.2	46.0	45.7	47.6	45.9
<b>STAT</b>	98.8	77.4	65.1	83.6	57.6	43.6	43.3	47.5	53.2	47.8
<b>Size Exc</b>	71.5	52.4	60.5	63.2	49.0	54.7	48.1	50.1	51.7	54.0
<b>Size Inc</b>	93.1	72.6	63.9	82.0	49.7	47.7	46.6	45.4	48.2	48.8
<b>U-test (5%)</b>	✓	✓		✓	✓					

Table 4: Table showing average classification results (averaged over algorithms) for each train/test set. Checkmarks show age-sets where size-inclusive and size-exclusive methods were distinct from one another using U-tests ( $p>0.05$ ). Also shown are average classification accuracies over size inclusive/exclusive methods for each age-set.

Figure 5 shows the average rank of size-inclusive and size-exclusive methods by age for each of the single-age image sets. For the younger age-sets, size-inclusive transform methods produce significantly better results than size-exclusive, or shape-only methods. As the age of the samples increases, the average rank of size-inclusive transforms increases while the rank of exclusive methods decreases; this proceeds until exclusive (shape only) methods perform better on average (have lower ranks) than size-inclusive methods.

However, whilst at age-3 to age-5, size-exclusive methods perform better than size-inclusive methods, the results themselves are not significantly different (between size inclusive/exclusive methods). Figure 6 shows average classification accuracy by age for each of the image sets. Whilst at higher ages size-exclusive methods do appear to perform better (than size-inclusive methods), the results are not as significant as those for younger image sets and perform only marginally better than random assignment of class.

### 3.3 Relative Performance by Otolith Age

N-way analysis of variance including interaction tests (Where potentially interacting factors are adjoined by a \* in Table 3), shows strong interaction between transform method and sample age ( $p=0.00$ ), whilst age itself shows as contributing most to classification accuracy. Interaction between age and transform methods is comparable ( $p=0.00$ ) when transform methods are grouped into size inclusive/exclusive classes, and results of size inclusive/exclusive post-hoc testing can be seen in Figure 7 (combined and modified for clarity) where it can be seen that as age increases, the relative positions of size-inclusive and size-exclusive transforms reverse.

Figure 8 illustrates post-hoc ANOVAN testing by age. It shows that classification of younger otoliths (ages 0/1) returns results statistically distinct from each other and from results using older samples (age 2+). Results of age 2+ test are indistinct (from each other).

## 4 Discussion

Our results show significant differences between size-inclusive and size-exclusive transform methods for otolith discrimination, depending on the age-category of otoliths modelled, demonstrating that type of transform must be considered when modelling otolith boundaries. We also show that the age of the sample otoliths impacts results when single-age sets are used to construct and test classifiers regardless of the classification methods used in this study. Analysis of our learning algorithms was limited given the difference in sizes of our age-sets; however we demonstrate that no significant difference exists between algorithms for these particular problem sets. Further analysis using consistent dataset sizes may yield better results in this area.

Results for the age-0 and age-1 sets show a significant difference between size-inclusive and size-exclusive transformation methods. Specifically, size based methods achieve up to 100% accuracy for age-0 herring (using individual classification algorithms), and averaging as high as 93.0% across all learning algorithms for size-inclusive transform methods on the same age-set. These results are far better than for size-exclusive transforms of the same set which achieved up to 91.3% and averaged 71.8%. This disparity in results between size-inclusive and size-exclusive methods can be seen across all age-0 and age-1 sets; indicating that for younger otolith samples, size is perhaps a more useful metric for stock discrimination.

Differences in the shape and size of otoliths as they grow are regulated by both endogenous and exogenous factors (Lombarte et al. 2003; Vignon and Morat 2010; Mille et al. 2015), with a tendency towards increasingly complex shape as the fish

ages. Our study shows clear differences in classification potential between using shape, and classification using size indices, subject to age of the samples. This observation supports previous studies where exogenous factors (such as separate spawning grounds) were found to influence early growth, with later growth being constrained by endogenous factors (Lombarte et al. 2003). Endogenous effects are more likely to influence shape-based classification performance of more mature specimens, potentially introducing additional boundary curve complexity (Vignon 2012; Sadighzadeh et al. 2014) as a result of secondary growth (Morales-Nin 2000).

The use of size may be an obvious method when distinguishing between stocks that spawn at different times of the year, where we expect age-specific inter-class sizes. However, even when metrics recorded using the SPa transform method were normalised (by otolith length), we achieve similar result patterns and accuracies. For the Irish/Celtic Sea dataset, the larger otoliths of the two classes (Celtic-Sea herring) are from fish that spawn later in the year than the smaller class (Irish-Sea) but have a much faster initial growth rate. Further investigation is therefore required to determine the age at which size-inclusive methods become incapable of separating spawning components for this particular discrimination problem.

Whilst for age-0 and age-1 sets we return favourable results, both in terms of classification accuracy and disparity between inclusive/exclusive methods; we find that as sets increase in age, classification accuracies reduce, becoming comparable to random assignment of class. For the eldest of our age categories size-inclusive and size-exclusive transform methods do appear to perform differently, however statistical testing using the full result suite shows that these too are indistinct. Shape may therefore be the primary factor to consider for stock classification using otolith outlines for older fish, but further testing should be carried out in this area, as after age partitioning, the datasets used in our study were small and inconsistent in size across ages-categories. Testing using larger sets may not show such a large drop in classification accuracies, and may clearly credit or discredit the change from size-inclusive to size-exclusive methods, both for the stocks studied here, and for additional stocks and/or species.

In machine learning, parameter calculations should be performed using a separate training set (Ambroise and McLachlan 2002; Cawley and Talbot 2010). Failure to use a separate set to calculate parameters risks building classifiers that only represent the data used to create them (over-fitting), rather than building generic classifiers. For example, when constructing a classifier using PCA the components should be calculated using a single set of data (otoliths), and the classifier built using a second set (using the components calculated using the first set). Testing the constructed classifier is then carried out using a third set. By using only static transform methods in this study, the need to calculate parameters was avoided. This was an important consideration as our age partitioned datasets are small. Further division required during validation of non-static methods would result in smaller sets with which to build and test classifiers, reducing robustness.

Results achieved using EFDs for the age-0 partition of the Irish/Celtic Sea herring set (Burke et al. 2008) did not match their published figure of 97%. However, this is largely down to two factors. First we used only a portion of the original dataset, and we split the portion further into two single age-sets; whereas the results reported by Burke et al. (2008) were based on the larger age-0 set, boosted by age-0 ring traces from the age-1 otoliths. Second we remove test instances from the set prior to parameter selection (regression analysis) and to building classifiers; whilst the original work performs parameter selection on the full set (including test samples). The validation in this work therefore reduces accuracy, but goes some distance to reduce over-fitting classifiers to the training set, and would produce classifiers which are more reliable when classifying further otolith samples. Our results are consistent with those in the previous work however, confirming that it is certainly possible to separate stocks of juvenile herring using Fourier analysis.

The variable numbers of samples per class during these tests may introduce some bias among test results, whereby smaller training sets produce less robust classifiers. However, when compared with previous classifications of the same dataset (Mapp et al. 2013) Curvature Scale Space methods performed to a similar degree when separated into single-age experiments, showing no significant change in accuracy, despite the training sets used here being only a fraction of the number used previously. Further experiments using larger training sets may yield increased classification accuracies however, and may show the curvature transform method to be an important tool when classifying datasets containing otoliths from older fish.

Whilst most of our results appear to be correlated with tests on other sets; our age-0 sprat tests return much lower classification accuracies, and size-inclusive/exclusive methods are not statistically distinct. This is despite the set being the same size as BB0 (40 instances per class). This may be due to one class of the set being comprised of age-0 ring traces from age-1 otoliths, in the absence of available age-0 samples for that task. The drop in accuracies may be down to several reasons in this instance: accuracy of outlining; or visible rings being inconsistent or uncorrelated with age-0 growth. Further, age-0 ring extraction has yet to be validated for this species (Torstensen et al. 2004).

Results from the single age testing show a significant reduction in classification accuracies as the age of otoliths used to build and test classifiers increases, regardless of the methods used to transform the otolith data. However the class sizes of age 2, 3 and 5 sets used in this study were relatively small, and therefore, classification accuracies may be impacted by insufficient training set sizes. Further testing with larger datasets is required to establish whether either method is able to perform with adequate certainty to allow clear stock discrimination in older fish, or whether clear results are the reserve of younger otoliths.

Poor quality of images is often a significant problem for shape orientated studies (Gonzales and Woods 2002) and introduced additional complexity in this study.

Initial segmentation/outlining of samples proved difficult due to difference in otolith storage/preparation, and without a great deal of supervision from an expert reader would not have been possible. The variability of otolith image capture methods, therefore, introduces additional complexity that may be reduced or eradicated by improving the quality of images and increased standardisation.

## **5 Conclusions**

We show that separation of young clupeids is possible using size-inclusive or size-only transformations with little or no feature selection. Whilst classification of juvenile otoliths is certainly possible using complex transformations such as Fourier or Curvature Scale Space, the addition of simple measurements such as otolith length, height or aspect-ratio, or refraining for performing scale normalisation of Fourier harmonics, may increase classification accuracies. The importance of early growth for classification supports methods where early incremental growth are used for stock separation.

The age of otoliths used for classification studies is likely to impact classification accuracies. Experiments using mixed age otoliths may suffer due to the differing defining feature (shape or size) depending on otoliths age composition of the dataset. This impact may be overcome by separating samples and applying separate classification methods, or potentially by performing size-inclusive tests using inner ring traces for the older otoliths, where ring extraction has been validated.

## **6 Acknowledgements**

This study was carried out under the umbrella of the Collaborative Centre for the Sustainable Use of the Seas (CCSUS), which provided a catalyst for collaborative work and we would like to thank both the University of East Anglia (UEA), and the Centre for Environment, Fisheries and Aquaculture Science (Cefas) for their support. In addition we would like to thank Defra project MF1112 'Pelagic Ocean Science: Ecology and Interconnectivity of Diverse Ocean Networks (POSEIDON)' for provision of sprat and herring otoliths, and Deirdre Brophy of the Galway-Mayo Institute of Technology (GMIT) for herring otolith images. Finally we would like to thank the Cefas Seedcorn project 'Trawling Through Time' for additional support throughout.

This research was carried out as part of a Ph.D research study at the University of East Anglia, and as such did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

- Abbasi, S., F. Mokhtarian and J. Kittler (1999). "Curvature scale space image in shape similarity retrieval." Multimedia systems **7**(6): 467-476.
- Aha, D. W., D. Kibler and M. K. Albert (1991). "Instance-based learning algorithms." Machine learning **6**(1): 37-66.
- Amanatiadis, A., V. Kaburlasos, A. Gasteratos and S. Papadakis (2011). "Evaluation of shape descriptors for shape-based image retrieval." IET Image Processing **5**(5): 493-499.
- Ambroise, C. and G. J. McLachlan (2002). "Selection bias in gene extraction on the basis of microarray gene-expression data." Proceedings of the national academy of sciences **99**(10): 6562-6566.
- Bagnall, A. and L. Davis (2014). "Predictive Modelling of Bone Age through Classification and Regression of Bone Shapes." arXiv preprint arXiv:1406.4781.
- Bagnall, A., E. Keogh, S. Lonardi and G. Janacek (2006). "A bit level representation for time series data mining with shape based similarity." Data Mining and Knowledge Discovery **13**(1): 11-40.
- Bani, A., S. Poursaeid and V. M. Tuset (2013). "Comparative morphology of the sagittal otolith in three species of south Caspian gobies." Journal of fish biology **82**(4): 1321-1332.
- Begg, G. A. and R. W. Brown (2000). "Stock identification of haddock *Melanogrammus aeglefinus* on Georges Bank based on otolith shape analysis." Transactions of the American Fisheries Society **129**(4): 935-945.
- Begg, G. A., S. E. Campana, A. J. Fowler and I. M. Suthers (2005). "Otolith research and application: current directions in innovation and implementation." Marine and Freshwater Research **56**(5): 477-483.
- Begg, G. A., W. J. Overholtz and N. J. Munroe (2001). "The use of internal otolith morphometrics for identification of haddock (*Melanogrammus aeglefinus*) stocks on Georges Bank." Fishery Bulletin **99**(1): 1-1.
- Bermejo, S., B. Monegal and J. Cabestany (2007). "Fish age categorization from otolith images using multi-class support vector machines." Fisheries Research **84**(2): 247-253.
- Bird, J. L., D. T. Eppler and D. M. Checkley Jr (1986). "Comparisons of herring otoliths using Fourier series shape analysis." Canadian Journal of Fisheries and Aquatic Sciences **43**(6): 1228-1234.
- Bober, M. (2001). "MPEG-7 visual shape descriptors." IEEE Transactions on circuits and systems for video technology **11**(6): 716-719.
- Breiman, L. (2001). "Random forests." Machine learning **45**(1): 5-32.
- Brophy, D. and B. S. Danilowicz (2002). "Tracing populations of Atlantic herring (*Clupea harengus* L.) in the Irish and Celtic Seas using otolith microstructure." ICES Journal of Marine Science: Journal du Conseil **59**(6): 1305-1313.
- Brownlee, J. (2013). "A Tour of Machine Learning Algorithms." Machine Learning Mastery.
- Burke, N., D. Brophy and P. King (2008). "Shape analysis of otolith annuli in Atlantic herring (*Clupea harengus*); a new method for tracking fish populations." Fisheries Research **91**(2): 133-143.

Burke, N., D. Brophy and P. A. King (2008). "Otolith shape analysis: its application for discriminating between stocks of Irish Sea and Celtic Sea herring (*Clupea harengus*) in the Irish Sea." ICES Journal of Marine Science: Journal du Conseil **65**(9): 1670-1675.

Burke, N., D. Brophy, P.-J. Schön and P. A. King (2009). "Temporal trends in stock origin and abundance of juvenile herring (*Clupea harengus*) in the Irish Sea." ICES Journal of Marine Science: Journal du Conseil: fsp140.

Campana, S. E. and J. M. Casselman (1993). "Stock discrimination using otolith shape analysis." Canadian Journal of Fisheries and Aquatic Sciences **50**(5): 1062-1083.

Cardinale, M., P. Doering-Arjes, M. Kastowsky and H. Mosegaard (2004). "Effects of sex, stock, and environment on the shape of known-age Atlantic cod (*Gadus morhua*) otoliths." Canadian Journal of Fisheries and Aquatic Sciences **61**(2): 158-167.

Castonguay, M., P. Simard and P. Gagnon (1991). "Usefulness of Fourier analysis of otolith shape for Atlantic mackerel (*Scomber scombrus*) stock discrimination." Canadian Journal of Fisheries and Aquatic Sciences **48**(2): 296-302.

Cawley, G. C. and N. L. Talbot (2010). "On over-fitting in model selection and subsequent selection bias in performance evaluation." Journal of Machine Learning Research **11**(Jul): 2079-2107.

DeVries, D. A., C. B. Grimes and M. H. Prager (2002). "Using otolith shape analysis to distinguish eastern Gulf of Mexico and Atlantic Ocean stocks of king mackerel." Fisheries Research **57**(1): 51-62.

Doering, P. and J. Ludwig (1990). "Shape Analysis of Otoliths—a Tool for Indirect Ageing of Eel, *Anguilla anguilla* (L.)?" Internationale Revue der gesamten Hydrobiologie und Hydrographie **75**(6): 737-743.

Duarte-Neto, P., R. Lessa, B. Stosic and E. Morize (2008). "The use of sagittal otoliths in discriminating stocks of common dolphinfish (*Coryphaena hippurus*) off northeastern Brazil using multishape descriptors." ICES Journal of Marine Science: Journal du Conseil **65**(7): 1144-1152.

Fablet, R. and N. Le Josse (2005). "Automated fish age estimation from otolith images using statistical learning." Fisheries Research **72**(2): 279-290.

Friedman, N., D. Geiger and M. Goldszmidt (1997). "Bayesian network classifiers." Machine learning **29**(2-3): 131-163.

Gagliano, M. and M. I. McCormick (2004). "Feeding history influences otolith shape in tropical fish." Marine Ecology Progress Series **278**: 291-296.

Galley, E. A., P. J. Wright and F. M. Gibb (2006). "Combined methods of otolith shape analysis improve identification of spawning areas of Atlantic cod." ICES Journal of Marine Science: Journal du Conseil **63**(9): 1710-1717.

Gonzales, R. C. and R. E. Woods (2002). "Digital Image Processing." New Jersey: Prentice Hall **6**: 681.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009). "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter **11**(1): 10-18.

Hills, J., J. Lines, E. Baranauskas, J. Mapp and A. Bagnall (2014). "Classification of time series by shapelet transformation." Data Mining and Knowledge Discovery **28**(4): 851-881.

ICES (2013). "Report of the international bottom trawl survey working group (IBTSWG)." ICES CM 2013/SSGESST 10(272).

ICES (2014). "Report of the working group of international pelagic surveys (WGIPS)." ICES CM 2014/SSGESST 1(360).

Jalba, A. C., M. H. Wilkinson and J. B. Roerdink (2006). "Shape representation and recognition through morphological curvature scale spaces." IEEE Transactions on Image Processing 15(2): 331-341.

Keating, J. P., D. Brophy, R. A. Officer and E. Mullins (2014). "Otolith shape analysis of blue whiting suggests a complex stock structure at their spawning grounds in the Northeast Atlantic." Fisheries Research 157: 1-6.

Le Cessie, S. and J. C. Van Houwelingen (1992). "Ridge estimators in logistic regression." Applied statistics: 191-201.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. European conference on machine learning, Springer.

Lines, J., L. M. Davis, J. Hills and A. Bagnall (2012). A shapelet transform for time series classification. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Lombarte, A., G. Torres and B. Morales-Nin (2003). "Specific Merluccius otolith growth patterns related to phylogenetics and environmental factors." Journal of the Marine Biological Association of the UK 83(02): 277-281.

Manurung, A. (2011). "Elliptic fourier for shape analysis."

Mapp, J., M. Fisher, A. Bagnall, J. Lines, S. Warne and J. S. Phillips (2013). Clupea Harengus: Intraspecies Distinction using Curvature Scale Space and Shapelets-Classification of North-sea and Thames Herring using Boundary Contour of Sagittal Otoliths. ICPRAM.

Mérigot, B., Y. Letourneur and R. Lecomte-Finiger (2007). "Characterization of local populations of the common sole *Solea solea* (Pisces, Soleidae) in the NW Mediterranean through otolith morphometrics and shape analysis." Marine Biology 151(3): 997-1008.

Mille, T., K. Mahe, M. Villanueva, H. De Pontual and B. Ernande (2015). "Sagittal otolith morphogenesis asymmetry in marine fishes." Journal of fish biology 87(3): 646-663.

Mokhtarian, F. and A. K. Mackworth (1992). "A theory of multiscale, curvature-based shape representation for planar curves." IEEE Transactions on Pattern Analysis and Machine Intelligence 14(8): 789-805.

Morales-Nin, B. (2000). "Review of the growth regulation processes of otolith daily increment formation." Fisheries Research 46(1): 53-67.

Otsu, N. (1975). "A threshold selection method from gray-level histograms." Automatica 11(285-296): 23-27.

Parisi-Baradad, V., A. Lombarte, E. García-Ladona, J. Cabestany, J. Piera and O. Chic (2005). "Otolith shape contour analysis using affine transformation invariant wavelet transforms and curvature scale space representation." Marine and Freshwater Research **56**(5): 795-804.

Paul, K., R. Oeberst and C. Hammer (2013). "Evaluation of otolith shape analysis as a tool for discriminating adults of Baltic cod stocks." Journal of Applied Ichthyology **29**(4): 743-750.

Pearson, K. (1901). "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 **2**(11): 559-572.

Platt, J. C. (1999). "12 fast training of support vector machines using sequential minimal optimization." Advances in kernel methods: 185-208.

Quinlan, J. R. (1993). "C4. 5: Programming for machine learning." Morgan Kaufmann: 38.

Rodriguez, J. J., L. I. Kuncheva and C. J. Alonso (2006). "Rotation forest: A new classifier ensemble method." IEEE transactions on pattern analysis and machine intelligence **28**(10): 1619-1630.

Sadighzadeh, Z., T. Valinassab, G. Vosugi, A. Motallebi, M. R. Fatemi, A. Lombarte and V. M. Tuset (2014). "Use of otolith shape for stock identification of John's snapper, *Lutjanus johnii* (Pisces: Lutjanidae), from the Persian Gulf and the Oman Sea." Fisheries Research **155**: 59-63.

Shamir, L., J. D. Delaney, N. Orlov, D. M. Eckley and I. G. Goldberg (2010). "Pattern recognition software and techniques for biological image analysis." PLoS Comput Biol **6**(11): e1000974.

Stransky, C. (2005). "Geographic variation of golden redfish (*Sebastes marinus*) and deep-sea redfish (*S. mentella*) in the North Atlantic based on otolith shape analysis." ICES Journal of Marine Science: Journal du Conseil **62**(8): 1691-1698.

Swain, M. J. and D. H. Ballard (1991). "Color indexing." International journal of computer vision **7**(1): 11-32.

Torstensen, E., A. Eltink, M. Casini, W. McCurdy and L. Clausen (2004). "Report of the Work Shop on age estimation of sprat." Institut of Marine Research, Flø devigen, Arendal, Norway: 14-17.

Vignon, M. (2012). "Ontogenetic trajectories of otolith shape during shift in habitat use: Interaction between otolith growth and environment." Journal of Experimental Marine Biology and Ecology **420**: 26-32.

Vignon, M. and F. Morat (2010). "Environmental and genetic determinant of otolith shape revealed by a non-indigenous tropical fish." Marine Ecology Progress Series **411**: 231-241.

WEKA. (2002). "The HyperPipes Classifier."

Zhang, D. and G. Lu (2003). "A comparative study of curvature scale space and Fourier descriptors for shape-based image retrieval." Journal of Visual Communication and Image Representation **14**(1): 39-57.

Zhang, D. and G. Lu (2003). "Evaluation of MPEG-7 shape descriptors against other shape descriptors." Multimedia Systems **9**(1): 15-30.

Figure 1: Experimental data flow through, from normalised otolith outlines (datasets) to classification results. Each age-set was transformed using each transform method to create individual train/test sets for each age-set (60 in total). Each train/test was used in turn with each of the algorithms to build and test classifiers using leave one out cross validation (LOOCV).

Figure 2: Three examples with overlaid boundary. Showing variability in image qualities obtained for this study. Samples from (l-r) (Burke et al. 2008), (ICES 2013), (Mapp et al. 2013). The left image shows consistent background and a well-defined otolith edge. The central image shows an instance of severe background texturing. On the right an image showing examples of foreground noise, as bubbles trapped in the storage medium.

Figure 3: Image showing the CSS feature points extracted; as point along boundary, and annotated with evolution (curvature) magnitude. Only the eleven points with largest evolution magnitude are shown in this example.

Figure 4: Learning algorithms – Results of post-hoc ANOVAN testing showing marginal mean classification accuracies for different learning algorithms.

Figure 5: Average rank of size-inclusive and size-exclusive transform methods for each age-set. Tests where size-inclusive/exclusive results are statistically different (5%) are shown using solid markers.

Figure 6: Average classification result of size-inclusive and size-exclusive transform methods for each age-set. Tests where size-inclusive/exclusive results are statistically different (5%) are shown using solid markers.

Figure 7: Results of post-hoc ANOVAN testing showing interaction between sample age and transform grouping (size-inclusive/exclusive). Modified/coloured for clarity, size-inclusive results (red), size-exclusive (blue).

Figure 8: Results of post-hoc ANOVAN testing showing sample age impact on classification accuracies. Age 0/1 tests are statistically distinct from one another.

SYSTEM FLOW















