




## Contribution to the Themed Section: 'Applications of machine learning and artificial intelligence in marine science'

### Original Article

# Automated classification of schools of the silver cyprinid *Rastrineobola argentea* in Lake Victoria acoustic survey data using random forests

Roland Proud <sup>1\*</sup>, Richard Mangeni-Sande<sup>1,2</sup>, Robert J. Kayanda<sup>3</sup>, Martin J. Cox<sup>1,4</sup>, Chrisphine Nyamweya<sup>5</sup>, Collins Ongore<sup>1,5</sup>, Vianny Natugonza<sup>2</sup>, Inigo Everson<sup>1,6</sup>, Mboni Elison<sup>7</sup>, Laura Hobbs<sup>8,9</sup>, Benedicto Boniphace Kashindye<sup>7</sup>, Enock W. Mlaponi<sup>7</sup>, Anthony Taabu-Munyaho<sup>2,3</sup>, Venny M. Mwainge<sup>5</sup>, Esther Kagoya<sup>2</sup>, Antonio Pegado<sup>10</sup>, Evarist Nduwayesu<sup>2</sup>, and Andrew S. Brierley<sup>1</sup>

<sup>1</sup>*Pelagic Ecology Research Group, School of Biology, Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, St Andrews KY16 8LB, UK*

<sup>2</sup>*National Fisheries Resources Research Institute (NaFiRRI), PO Box 343, Jinja, Uganda*

<sup>3</sup>*Lake Victoria Fisheries Organization (LVFO), PO Box 1625, Jinja, Uganda*

<sup>4</sup>*Australian Antarctic Division, 203 Channel Highway, Kingston, Tasmania 7050, Australia*

<sup>5</sup>*Kenya Marine and Fisheries Research Institute (KMFRRI), Mombasa, Kenya*

<sup>6</sup>*School of Environmental Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK*

<sup>7</sup>*Tanzania Fisheries Research Institute (TaFiRI), PO Box 475, Mwanza, Tanzania*

<sup>8</sup>*Scottish Association for Marine Science, Oban, Argyll PA37 1QA, UK*

<sup>9</sup>*Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, UK*

<sup>10</sup>*Instituto de Investigacao Pesqueira (IIP), Maputo, Mozambique*

\*Corresponding author: tel: +44 (0)1334 46 3401; e-mail: [rp43@st-andrews.ac.uk](mailto:rp43@st-andrews.ac.uk).

Proud, R., Mangeni-Sande, R., Kayanda, R. J., Cox, M. J., Nyamweya, C., Ongore, C., Natugonza, V., Everson, I., Elison, M., Hobbs, L., Kashindye, B. B., Mlaponi, E. W., Taabu-Munyaho, A., Mwainge, V. M., Kagoya, E., Pegado, A., Nduwayesu, E., and Brierley, A. S. Automated classification of schools of the silver cyprinid *Rastrineobola argentea* in Lake Victoria acoustic survey data using random forests. – ICES Journal of Marine Science, 77: 1379–1390.

Received 7 July 2019; revised 5 March 2019; accepted 10 March 2020; advance access publication 9 May 2020.

Biomass of the schooling fish *Rastrineobola argentea* (dagaa) is presently estimated in Lake Victoria by acoustic survey following the simple “rule” that dagaa is the source of most echo energy returned from the top third of the water column. Dagaa have, however, been caught in the bottom two-thirds, and other species occur towards the surface: a more robust discrimination technique is required. We explored the utility of a school-based random forest (RF) classifier applied to 120 kHz data from a lake-wide survey. Dagaa schools were first identified manually using expert opinion informed by fishing. These schools contained a lake-wide biomass of 0.68 million tonnes (MT). Only 43.4% of identified dagaa schools occurred in the top third of the water column, and 37.3% of all schools in the bottom two-thirds were classified as dagaa.

© International Council for the Exploration of the Sea 2020.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

School metrics (e.g. length, echo energy) for 49 081 manually classified dagaa and non-dagaa schools were used to build an RF school classifier. The best RF model had a classification test accuracy of 85.4%, driven largely by school length, and yielded a biomass of 0.71 MT, only c. 4% different from the manual estimate. The RF classifier offers an efficient method to generate a consistent dagaa biomass time series.

**Keywords:** artificial intelligence, big data, dagaa, Lake Victoria, machine learning, *Rastrineobola argentea*, school analysis, species identification, stock assessment

## Introduction

In recent years, and parallel to the development of ever-cheaper computer-processor power, machine learning and artificial intelligence (AI) methods have been applied increasingly in ecology to ask “big questions” of “big data”. These methods have delivered promising results in species identification, biodiversity mapping and animal behaviour studies (Christin *et al.*, 2019). Active acoustic data collected during fish stock assessment surveys are a form of big data. A typical month-long survey can gather tens of gigabytes of data per narrowband frequency and, with the increasing inclusion of broadband echosounders and multibeam sonars in fish stock assessment surveys, this will increase by at least tenfold per vessel in the future (Demer *et al.*, 2017). Multiple autonomous platforms including wave gliders (Bingham *et al.*, 2012; Greene *et al.*, 2014) and saildrones (Mordy *et al.*, 2017; De Robertis *et al.*, 2019) are increasing the temporal and spatial coverage of fish stock monitoring, and the volume of data now being collected in some ecosystems exceeds institutional capacity for manual processing. Machine learning methods can potentially be utilized to automate data analysis pathways and, at the same time, reduce human error-induced uncertainty in stock biomass estimates.

During the analysis of acoustic survey data, visual scrutinization—classification by eye of features on echograms [the two-dimensional (2D) plots showing echo energy by depth and distance/time along track]—is often used to partition echo energy between species, but results can be operator dependent. Efforts to overcome this by the application of rigid “rules” can also be unsatisfactory. Identification of Antarctic krill (*Euphausia superba*), for example has been achieved by a simple “dB difference” approach that uses the difference in backscattering intensity between two frequencies as a diagnostic characteristic (Madureira *et al.*, 1993) but has in some areas of continental shelf been susceptible to erroneous inclusion of echoes from ice fish (Channichthyidae; Fallon *et al.*, 2016). In Lake Victoria, the silver cyprinid (*Rastrineobola argentea*; known locally as “dagaa”) is identified using a simple depth distribution rule that holds that most of the backscattered echo energy from the top third of the water column is from dagaa (LVFO, 2006). This approach—that was incorporated in to the Lake Victoria acoustic analysis standard operating procedure (SOP; LVFO, 2012) in a period when limited resources precluded anything more sophisticated—is, however, known to be flawed: not all the fish obey the rule. Here, we apply AI to the identification of echoes from schools of Lake Victoria fish in an effort to illustrate an example of the potential for AI in fisheries ecology and to improve the accuracy of stock assessments for the lake.

## The Lake Victoria fishery

Lake Victoria is the world’s largest tropical lake (68 800 km<sup>2</sup>). Fisheries are vital for local food provision and for export earnings and contribute 2–3% to the gross domestic products of

the lake’s three riparian states (Uganda, Kenya, and Tanzania). Sustainable fisheries management is a regional priority, and there is an aspiration to move towards ecosystem-based fisheries management (LVFO, 2018).

Dagaa are a small (maximum length c. 9 cm) pelagic zooplanktivorous fish (Wanink, 1999) native to Lakes Victoria, Nabugabo, and Kyoga in East Africa. It is one of the few species in Lake Victoria to remain abundant following the introduction of the Nile Perch (*Lates niloticus*) in the 1950s (Goudswaard *et al.*, 2008; Sharpe and Chapman, 2014). Dagaa make up ~60% of the total annual Lake Victoria catch, with ~0.6 million tonnes (MT) being landed in 2015 (Mangeni-Sande *et al.*, 2019). Typically 100–500 kg of dagaa can be caught per boat per night using small seine nets and light attraction (LVFO, 2016a), and ~18 700 boats target dagaa (LVFO, 2016b). Dagaa fishing employs ~70 500 fishermen, and ~16 500 women are engaged in the labour-intensive drying of the catch (Okedi, 1981; LVFO, 2016b); fish are spread out for drying—often simply on the sand—in the sun and are turned regularly by hand. Dagaa are sold into the local and regional markets and consumed almost exclusively in southern and eastern Africa: dagaa is a cheap source of animal protein for the rural poor. High-quality dried fish are sold for human consumption, and lower-quality products (~70% of the total catch) are used for animal feed (Odongkara *et al.*, 2016).

The emphasis of research on fisheries in Lake Victoria has to date been largely on Nile Perch because of its importance in generating foreign currency revenue (US\$300 million; LVFO, 2018). However, for economic and ecological reasons, it is essential to establish effective management for sustainable exploitation of other species as well, including dagaa (Kolding *et al.*, 2019), and accurate estimates of stock biomass are an essential prerequisite for that.

## Present estimates of dagaa biomass

Estimates of dagaa stock biomass are determined from acoustic data collected during bi-annual lake-wide fish stock assessment surveys. Dagaa, which are superficially similar to anchovy, are an obligate schooling pelagic species that possess swim bladders: as such dagaa is highly suitable for acoustic assessment. During daylight, dagaa aggregate into small schools (a few metres in length and height) that appear as distinct needle-like features in echograms when observed at typical survey setting, i.e. vessel speed of between 8 and 10 knots and ping intervals of between 0.2 and 0.5 s (Getabu *et al.*, 2003). Dagaa are presently evaluated by echo integration of 120 kHz data from the top one-third of the water column. It is assumed that all echo energy remaining in the top third of the water column after single-target detections (which are all attributed to Nile Perch; Kayanda *et al.*, 2012) have been removed arises from dagaa, and only dagaa. It is clear though, even from just a cursory reference to Getabu *et al.* (2003), that this is a false assumption: dagaa occupy a broader depth range than just the top third, and other species are known to inhabit

the top third. A new method for dagaa identification is needed urgently to improve the accuracy of stock assessment and, eventually, to improve the management that stems from the biomass estimate. Since the objective of the acoustic survey is to allocate all energy correctly, improving dagaa allocation will lead to improvements in the assessment of other species as well.

### Fish school analysis using acoustic data

In order to establish reliable and reproducible methods to identify and discriminate species detected acoustically during surveys, we need first to identify acoustic characteristics, or sets of characteristics, that are unique to particular target species and that are therefore diagnostic. For schooling species, these characteristics can be at the school level (rather than at the level of the individual fish), and the physical shape, echo intensity, frequency response and behaviour of schools of different fish species can be diagnostic (Coetzee, 2000; Reid *et al.*, 2000; Lawson, 2001; Bertrand *et al.*, 2008; Fernandes, 2009; Paramo *et al.*, 2010). Since the development of standard methods for extracting school characteristics (Barange, 1994; Coetzee, 2000; Reid *et al.*, 2000; Diner, 2001), analyses have been conducted to study the shapes and behaviours of schools of many species of fish (Lawson, 2001; Fernandes, 2009; Fallon *et al.*, 2016) and the swarm characteristics of krill (Tarling *et al.*, 2001; Klevjer *et al.*, 2010; Cox *et al.*, 2011). Such analyses are now being used to aid species identification, and hence to reduce uncertainty around estimates of fish stock biomass (e.g. for herring and mackerel; Fernandes, 2009).

Schools of a specified minimum size (horizontal and vertical dimensions) and echo intensity can be extracted automatically from acoustic observations [both 2D observations from conventional vertical echosounders and three-dimensional (3D) observations from multibeam sonar surveys], and school metrics pertaining to morphology, position, and acoustic scattering properties (e.g. echo energy across different frequencies) can be collated to characterize schools (Barange, 1994). Performing such an automated school extraction process for a typical month-long Lake Victoria vertical echosounder survey results in over 100 000 extracted schools. These include schools of dagaa, small (<10 cm) Nile Perch, and haplochromine cichlids (aggregations of the pelagic crustacean *Caridina nilotica* are also apparent). More than 25 acoustic surveys have been conducted on Lake Victoria over the past 20 years (Taabu-Munyaho *et al.*, 2014), and school data within them offer an incredibly valuable resource for examining potential change as a function of, for example, fishing pressure and environmental variability (Brierley and Cox, 2010, 2015). Fundamental to these types of analyses and indeed to fish stock assessment are consistent and reproducible methods to identify and discriminate species, including dagaa. It is impractical to attempt to use manual visual scrutinization to discriminate dagaa schools from the more than 2.5 million estimated schools now potentially accessible from the combined 25-survey database. Therefore, the main objective of the work reported here was to develop a robust and consistent approach that was both cost- and time-effective, and that used machine learning/AI to perform the automatic classification of dagaa schools (e.g. Fernandes, 2009; Cox *et al.*, 2011; Fallon *et al.*, 2016; Escobar-Flores *et al.*, 2019).

### Machine learning

It is now common practice to use machine learning techniques to classify data (Malde *et al.*, 2019). Features isolated in acoustic

survey data, such as schools, scattering layers, and single targets, have been classified using a wide range of machine learning techniques including mixture models (Fleischman and Burwen, 2003; Escobar-Flores *et al.*, 2018), artificial and convolutional neural networks (Haralabous and Georgakarakos, 1996; Simmonds *et al.*, 1996; Korneliusson *et al.*, 2016; Brautaset *et al.*, 2020), decision trees, random forests (RFs), and boosted regression trees (Fernandes, 2009; D'Elia *et al.*, 2014; Fallon *et al.*, 2016; Escobar-Flores *et al.*, 2018, 2019), discriminant-function analysis and principal components analysis (Nero and Magnuson, 1989; Scalabrin *et al.*, 1996; Brierley *et al.*, 1998; Lawson, 2001), and k-means clustering (Tegowski *et al.*, 2003; Proud *et al.*, 2017). Ensemble tree methods (e.g. RF and boosted regression trees) have only been adopted in the past decade but have been found to be particularly good (having high accuracy) for classifying fish schools (Fernandes, 2009; D'Elia *et al.*, 2014; Fallon *et al.*, 2016).

### Objective of the present study

The objective of this study is to develop a robust, automated method to identify echoes from dagaa schools in echosounder data collected during Lake Victoria fish stock assessment surveys. Previous work (Getabu *et al.*, 2003; LVFO, 2006), and a large accumulation of local experience, suggests that dagaa form schools that have a distinct needle-shaped (vertically tall, horizontally narrow) appearance in underway echograms. We set out first to confirm that needle-shaped acoustic features are in fact dagaa schools, and then to develop a machine learning method to identify dagaa schools amongst all extracted schools. In this study, we make no attempt to classify aggregations of the other common Lake Victoria pelagic species because there is presently not enough ground-truth data (e.g. trawl data) to underpin such an analysis.

### Methods

#### Determining the characteristics of dagaa schools

We conducted target fishing during an October 2019 field study in a coastal region (c. 40 m lakebed depth) of the Ugandan sector of the lake. We fitted a standard Lake Victoria bottom trawl with a fine-mesh cod-end cover and used this to target needle-like and non-needle-like pelagic echogram features. The net had an estimated vertical opening of <10 m, and was fished at 4 knots for 15 min at each sampled depth. Catch samples were sorted into species groups and the individuals in each group were counted, measured and weighed. The acoustic data recorded during each trawl were resampled to typical survey settings (vessel speed = 9 knots and ping interval = 0.2 s) to reconstruct echograms that would have been produced had the fished schools been encountered at typical survey speed.

#### Acoustic survey data collection

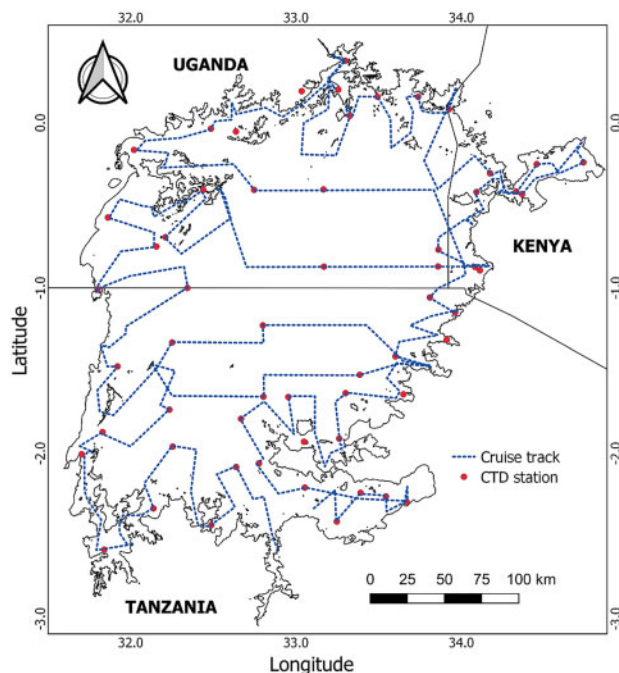
Acoustic and environmental data collected during the November 2015 fish stock assessment survey in Lake Victoria (LVFO, 2015) were used to build a dagaa school classifier. That survey was selected because, when this work began, it was the most recent survey that had been processed. The survey was conducted from research vessel (RV) *Victoria Explorer* between 1 and 29 November (including 4 days breaks for provisioning). It covered c. 4 000 km of survey track over most of the lake, across a range of lakebed depths between 2 and 70 m. Most of the survey was conducted in daylight hours, and sampling effort was highest

in the more productive inshore regions of the lake (Figure 1). The night-time vertical distribution and echogram appearance of dagaa schools may differ from the daytime distribution and form, and so night-time observations were excluded from the analysis: only acoustic data collected between sunrise and sunset (excluding astronomical twilight) were analysed. Acoustic data were collected using two hull-mounted Kongsberg (Horten, Norway) Simrad EK60 scientific echosounders operating at 70 and 120 kHz, both with a 7° nominal beam width. A pulse length of 0.256 ms was used, with a ping interval of 0.2 s. A standard split-beam echosounder calibration (Foote *et al.*, 1983; Demer *et al.*, 2015) was carried out prior to the survey. In this study, we use and report only 120 kHz data because one objective of the work is to develop a route for the reanalysis of historic surveys, and early surveys only used 120 kHz. However, at an early stage of this study, we investigated the benefit of including 70 kHz data as well but found no improvement in decision tree-based school classification using two frequencies.

Hydrographic measurements were taken at predefined stations ( $N=58$ , see Figure 1) using a Sea and Sun Conductivity, Temperature, and Depth (CTD) probe and a YSI 650 multi-parameter sonde to measure temperature (°C), dissolved oxygen concentration (DO,  $\text{mg l}^{-1}$ ), conductivity ( $\mu\text{S cm}^{-1}$ ), pH, turbidity [Formazin Turbidity Units (FTU)], and chlorophyll *a* concentration ( $\mu\text{g l}^{-1}$ ).

### School extraction and manual classification

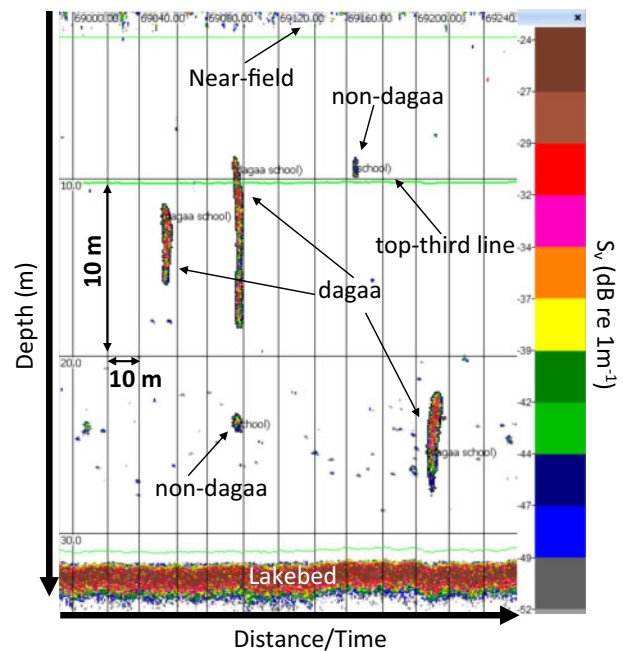
The “Schools Detection Module” in Echoview software (v9; Myriax, Hobart, Tasmania) was used to extract all schools from the echosounder data. Before running the school detection algorithm, echosounder data were thresholded at  $-54 \text{ dB re } 1 \text{ m}^{-1}$  (i.e. any samples below this value were excluded from analysis).



**Figure 1.** Map of Lake Victoria, East Africa, showing the cruise track and the CTD stations (solid points) for the November 2015 fish stock assessment survey.

Recalling that  $S_v = 10 \times \log_{10}(10^{(TS/10)} \times \text{packing density})$ , the threshold was set with the consideration of an expected mean target strength ( $TS$ ) at 120 kHz of dagaa with a mean length of 5.3 cm of  $-57.6 \text{ dB re } 1 \text{ m}^2$  and a very conservative minimum school packing density of c. 2 fish per  $\text{m}^3$  (Tumwebaze, 2003).

The school detection algorithm [Shoal Analysis and Patch Estimation System (SHAPES)] is based on the work of Barange (1994) and Coetzee (2000) and requires a number of parameters to be set. From preliminary analysis, local prior knowledge, and the work of Getabu *et al.* (2003), dagaa schools were perceived to be characteristically very narrow (a few pings) relative to vertical extent (tens of samples) in echograms, dense, and compact (i.e. without any vacuoles or holes), and so the SHAPES algorithm parameters were set conservatively to ensure that all schools of this nature, as well as schools with the more usual rounded echogram appearance, would be captured. Thus, all school detection parameters, except for the horizontal linking distance, were set to their minimum possible values, i.e. the minimum candidate height, minimum candidate length, minimum school height, minimum school length, and vertical linking distance were all set to 1 m. The maximum horizontal linking distance was set to 5 m to ensure that, given the vessel speed and ping rate, consecutive pings could be linked. The SHAPES algorithm was run across the entire acoustic dataset. It identified schools with a diversity of forms, from small compact needle-like schools typical of dagaa (see Figure 2) to large amorphous schools hundreds of metre in length that were layer like in appearance. The expert view is that



**Figure 2.** Echoview-generated 120 kHz echogram of automatically detected and manually labelled schools (grid size 10 by 10 m). Three dagaa schools are labelled and have the characteristic needle-like echogram appearance (vessel speed = 9 knots and ping interval = 0.2 s). In this example, the lakebed is at 32 m and the top third line (that under the existing standard operating procedure would demarcate the lower limit of the dagaa habitat) is at 10.7 m: the dagaa schools extend deeper than the top third line and the non-dagaa schools can be seen above the top third line, well illustrating the inability of depth alone to differentiate dagaa and non-dagaa.

these layers are comprised of haplochromine cichlids and small Nile Perch (<10 cm). Any “schools” that were longer than 100 m in length were deemed to be scattering layers (Proud *et al.*, 2015) and were excluded from further analysis. All remaining schools were examined manually and categorized by eye (visual scrutiny) as either dagaa or non-dagaa.

School metrics (Table 1) were exported from Echoview. Environmental variables (Table 1) were ascribed to each school as those at the nearest CTD station by distance.

### Using machine learning to build a school-based classifier

An RF model was built using a subset of the manually identified schools. During the survey, on-transect vessel speed varied between 6 and 13 knots, but c. 84% of effort was between 8 and 10 knots (which is the typical range of survey speeds across all historic acoustic surveys on the lake). Historic manual classification has identified needle-like echo traces as dagaa schools, but the aspect ratio (height to width) of echogram features is of course a function of vessel speed and ping rate. This raises the possibility that schools detected at slow speed would be rejected by eye as dagaa because they would appear too rounded: this is an important illustration of one of the weaknesses of visual classification methods. To avoid incorporating any potential speed-related bias in the RF model of dagaa school characteristics, only visually classified schools detected in the range of usual survey speeds (8–10 knots) were used to build the RF model.

Following the standard RF protocol (Breiman, 2001), schools remaining after speed filtering were split randomly into a training dataset (80% of data are typically used to train an RF classifier, and we adhered to that) and a test dataset (20% of data are typically used to test RF classifiers). The R packages “caret”, “party”, and “trees” (Strobl *et al.*, 2008, 2009; Kuhn, 2019; Ripley, 2019) were used to build RF models. RF algorithms have two tuning parameters: these are *mtry*, the number of variables to select randomly from the total available list of school metrics (Table 1) when splitting data at each node in a tree, and *ntrees*, the number of trees to build. In this study, *mtry* was initially set to 4 and *ntrees* to 500 (these are the default values), but a range of different

*mtry* and *ntrees* values was also used to assess their impact on RF classification accuracy.

We used repeated (three times) tenfold cross-validation to assess the accuracy of the RF (Stone, 1974; Breiman, 2001). This validation process involved splitting the training dataset into ten equally sized subsets (or folds), building the RF model using a dataset containing nine of the tenfolds, and then validating the model on the other remaining fold. This process was repeated ten times such that each fold acted as the validation dataset once. This process was repeated three times (with random, so probably different, tenfold splitting on each of the three occasions), and the accuracy of the model was calculated by taking an average over the resultant 30 accuracy values ( $3 \times 10$  folds).

### Assessment of the RF model

The RF model was assessed using the mean and standard deviation of the training accuracy, and the kappa statistic  $\kappa$  (the proportion of classification agreement beyond that expected to occur by chance, where  $\kappa = 0$  is suggestive of classification only matching what would be expected by random chance assuming a binomial distribution; Cohen, 1960). RF models are difficult to interpret, since they are typically comprised of hundreds of fully grown decision trees. In the majority of cases, RF models are assessed by accuracy metrics and the importance of each predictor (each school metric in this case) is assessed by single specific or multiple so-called “importance metrics” (Breiman, 2001). Here, we use conditional variable importance (Strobl *et al.*, 2008) to assess each predictor’s ability to discriminate between target classes (i.e. dagaa or non-dagaa): unlike other importance measures (e.g. mean decrease in accuracy), conditional variable importance is robust against correlated variables (Fallon *et al.*, 2016), e.g. water temperature and school depth are likely to be correlated.

### Dagaa stock biomass estimates

The RF model, which was built using a subset of the extracted schools, was used to classify the entire dataset of schools from the 2015 survey. School-based estimates of dagaa stock biomass were then calculated using both the manually classified schools and the

**Table 1.** School metrics used to build an RF model to classify detected schools

School metric	Description	Unit
Length	Mean length of school corrected for beam width	m
Depth	Mean depth of school	m
Height	Mean height of school corrected for pulse length	m
Image compactness	School perimeter squared/ $(4 \times \pi \times$ school area); for a perfectly circular school this would be 1	Unitless
NASC	School NASC is an historic acoustic unit that is the average amount of echo energy produced by the school per $\text{m}^2$ of lake surface, scaled up to an area of 1 nautical mile squared	$\text{m}^2 \text{nmi}^{-2}$
Lakebed depth	Depth of lakebed as detected by the 120 kHz echosounder	m
Temperature	Measured value at school depth obtained from the closest CTD station	$^{\circ}\text{C}$
DO	Measured value at school depth obtained from the closest CTD station	$\text{mg l}^{-1}$
pH	Measured value at school depth obtained from the closest CTD station	Unitless
Turbidity	Measured in Formazin Turbidity Units. Measured value at school depth obtained from the closest CTD station	FTU
Chlorophyll <i>a</i> concentration	Measured value at school depth obtained from the closest CTD station	$\mu\text{g l}^{-1}$
Longitude	Taken from vessel GPS	Degrees East
Latitude	Taken from vessel GPS	Degrees North
Time of day	Decimal time, calculated from vessel GPS	Hours

RF classified schools. Echo energy from schools classified as dagaa (either manually or via the RF model) was converted into biomass following the Lake Victoria Fisheries Organization SOP for stock assessment (LVFO, 2012). Accordingly, mean dagaa nautical area scattering coefficient (NASC) values were determined for each of the 18 SOP-defined lake areas, which are split by country (3; Uganda, Tanzania, and Kenya), lake quadrant (4; NW, NE, SE, SW), and depth (3; “inshore” <10 m; “coastal” 10–40 m, and “deep” >40 m). These 18 mean NASC values were converted to biomass density ( $T\ m^{-2}$ ) using the mean dagaa  $TS$  per kg ( $TS_{kg}$ , i.e. the amount of 120 kHz echo energy produced by 1 kg of dagaa) of  $-29.4\ dB\ kg^{-1}$  (Tumwebaze, 2003). Biomass densities were multiplied by associated areas to scale to biomass (T) in each of the 18 areas, and these were summed to give a whole-lake value. This process was repeated 1000 times, resampling with

replacement dagaa school NASC values by area on each iteration (i.e. bootstrapping), and 95% confidence intervals were calculated.

## Results

A total of 120 181 schools (larger than 1 m in length and height) were detected by the SHAPES algorithm in the echosounder data. Schools in “bad data” regions (e.g. sections of transect with no GPS) and schools detected at night were removed reducing the useable dataset to 115 778 schools.

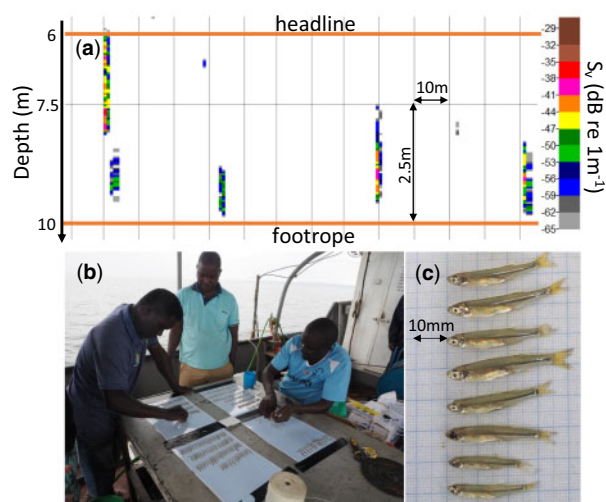
### Confirmation that needle-like echo traces are dagaa schools

It is generally believed, based on the work of Getabu *et al.* (2003) and on accumulated local expert opinion, that schools with a needle-like appearance in 120 kHz underway echograms are dagaa schools. To this end, needle-like schools were fished during an October 2019 field study (Figure 3 and Table 2).

A total of 93 schools were detected acoustically during Haul 1 (near surface), and 99.5% of the total catch by number (399 fish) was dagaa. The only non-dagaa component of the combined catch was two haplochromine cichlids, each just 4 cm long (the mean length of dagaa was c. 3.8 cm). These cichlids would have contributed c. 1% to integrated trawl echo energy (estimated using haplochromine  $TS = 20\log L - 66.65$ ; LVFO, 2015). Dagaa and needle-like schools were also present in Hauls 2–5 along with similar numbers of haplochromines, conforming with the view of Getabu *et al.* (2003) that dagaa are not restricted to the near-surface layer (Table 2). However, since catch obtained from Hauls 2–5 was likely contaminated during time spent at the surface whilst deploying and recovering the net, these observations were not quantitatively assessed.

### Manual classification of schools using the lake-wide 2015 survey data

A total of 56 079 of the 115 778 schools passed for manual visual identification were classified as dagaa. The remaining 59 699 schools, judged by experts to be non-dagaa, would have contained haplochromines, *Tilapia* spp., small Nile Perch (<10 cm) and other species, but the present state of knowledge is



**Figure 3.** Example dagaa trawl (Haul 1) during the October 2019 field study in Ugandan waters aboard the RV *Ibis*: (a) 120 kHz echogram showing needle-like schools, which are commonly believed to be dagaa schools; (b) catch from needle-like schools being sorted; and (c) dagaa, which comprised >99% of the catch by number.

**Table 2.** Net haul and catch information (numbers of individual fish)

Haul	Wire out (m)	Headline depth (m)	Dagaa (N)	Haplochromines (N)	Needle-like schools (N)
1	25	6	399	2	93
2	50	11	288	1	16
3	75	26	73	86	0
4	100	30	86	103	0
5	125	34	68	22 <sup>a</sup>	0

The water depth was 40 m.

<sup>a</sup>The decapod *Caridina nilotica* was also present, in small number, in the catch from Haul 5.

**Table 3.** Distributions of dagaa and non-dagaa schools by depth according to expert manual classification, and median lake-wide biomass estimates (bootstrapped 95% confidence intervals given in square brackets)

Depth zone	Dagaa schools (N)	Non-dagaa schools (N)	Total schools (N)	Dagaa school biomass (T)
Top third	24 357	6 403	30 760	370 701 [361 388–379 408]
Bottom two-thirds	31 722	53 296	85 018	312 707 [301 747–324 400]
Total	56 079	59 699	115 778	683 107 [668 957–697 721]

**Table 4.** Distributions of dagaa and non-dagaa schools by depth according to RF classification, and median lake-wide biomass estimates (bootstrapped 95% confidence intervals given in square brackets)

Depth zone	Dagaa schools (N)	Non-dagaa schools (N)	Dagaa school biomass (T)
Top third	26 171 (+7.44%)	4 589 (−28.33%)	394 373 [385 006–403 609] (+6.38%)
Bottom two-thirds	32 534 (+2.56%)	52 484 (−1.52%)	315 853 [305 435–327 424] (+1.01%)
Total	58 705 (+4.68%)	57 073 (−4.40%)	710 547 [695 426–725 205] (+4.02%)

Brackets indicate percentage change relative to manual classification.

insufficient to classify them by species; that will be the task for a subsequent project.

Only 43.4% of the manually classified dagaa schools occurred in the top third of the water column, but 89.3% of non-dagaa schools occurred in the bottom two-thirds: together these proportions give the “top third” method an overall school classification success rate by number of c. 72.6% (see Table 3). Dagaa school biomass was found to be almost equally distributed between the top third and bottom two-thirds of the water column. The dagaa stock biomass estimate arising from the manual classification was 0.68 MT (see Table 3).

### RF model

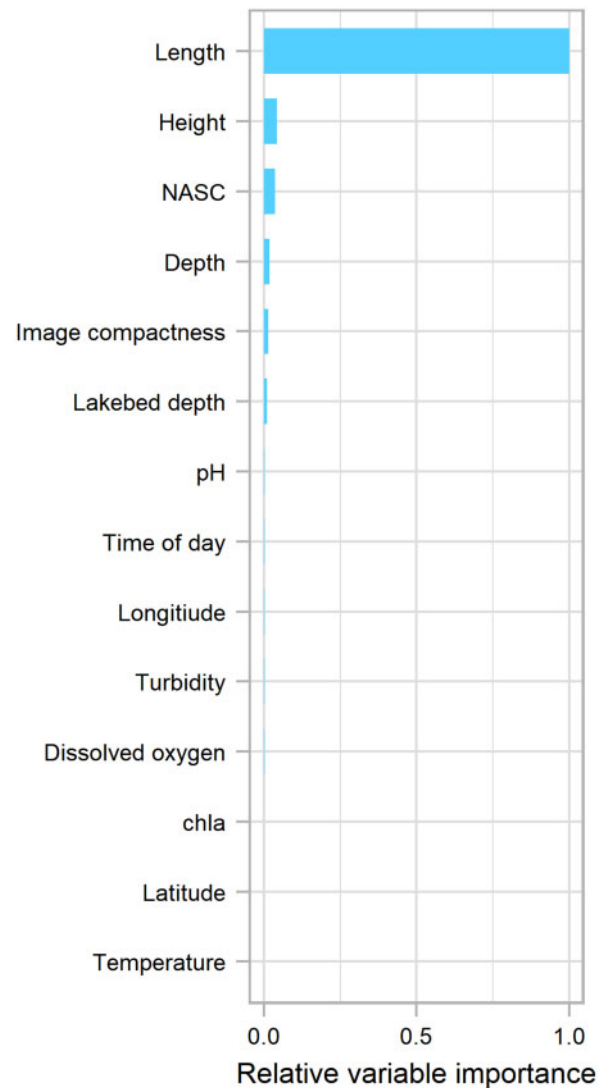
The 49 081 manually classified schools remaining after filtering for vessel speed were split into a training dataset (13 547 dagaa schools, 25 718 non-dagaa schools) and a test dataset (3 319 dagaa, 6 497 non-dagaa). The training dataset was used to train the RF classifier. An RF classifier was constructed using all 14 available school and environment metrics (Table 1). The default values of *mtry* (4) and *ntrees* (500) produced the best model as evaluated by model accuracy; other *mtry* and *ntrees* parameter values were tested (*mtry*: 2–8 and *ntrees*: 200–2000) but provided no improvement in accuracy. The RF model had a training classification accuracy of 85.0% (*SD* = 0.49%), a test classification accuracy of 85.4%, and a  $\kappa$ -value of 0.66 (*SD* = 0.011).

### RF predictions

The RF model was used to classify all schools in the full dataset of 115 778 schools (i.e. not just the schools that passed the speed filter). Since school dimensions were determined from GPS position, there were no speed-related artefacts in the automatically extracted school metric values. Schools classified by the RF model as dagaa were used to estimate lake-wide biomass, and are summarised in Table 4. The RF-derived biomass value differed by only 4.02% from the manual school classification result (see Tables 3 and 4). The largest difference between manual classification and the RF model classification was of non-dagaa schools in the top third depth zone. The manual scrutinization classified 1 814 more schools as non-dagaa. We believe that this occurred when slow vessel speed served to stretch observations of dagaa schools horizontally, giving them a non-dagaa appearance in the echogram. The RF approach takes school dimensions from GPS locations so is not “misled” by variability in vessel speed.

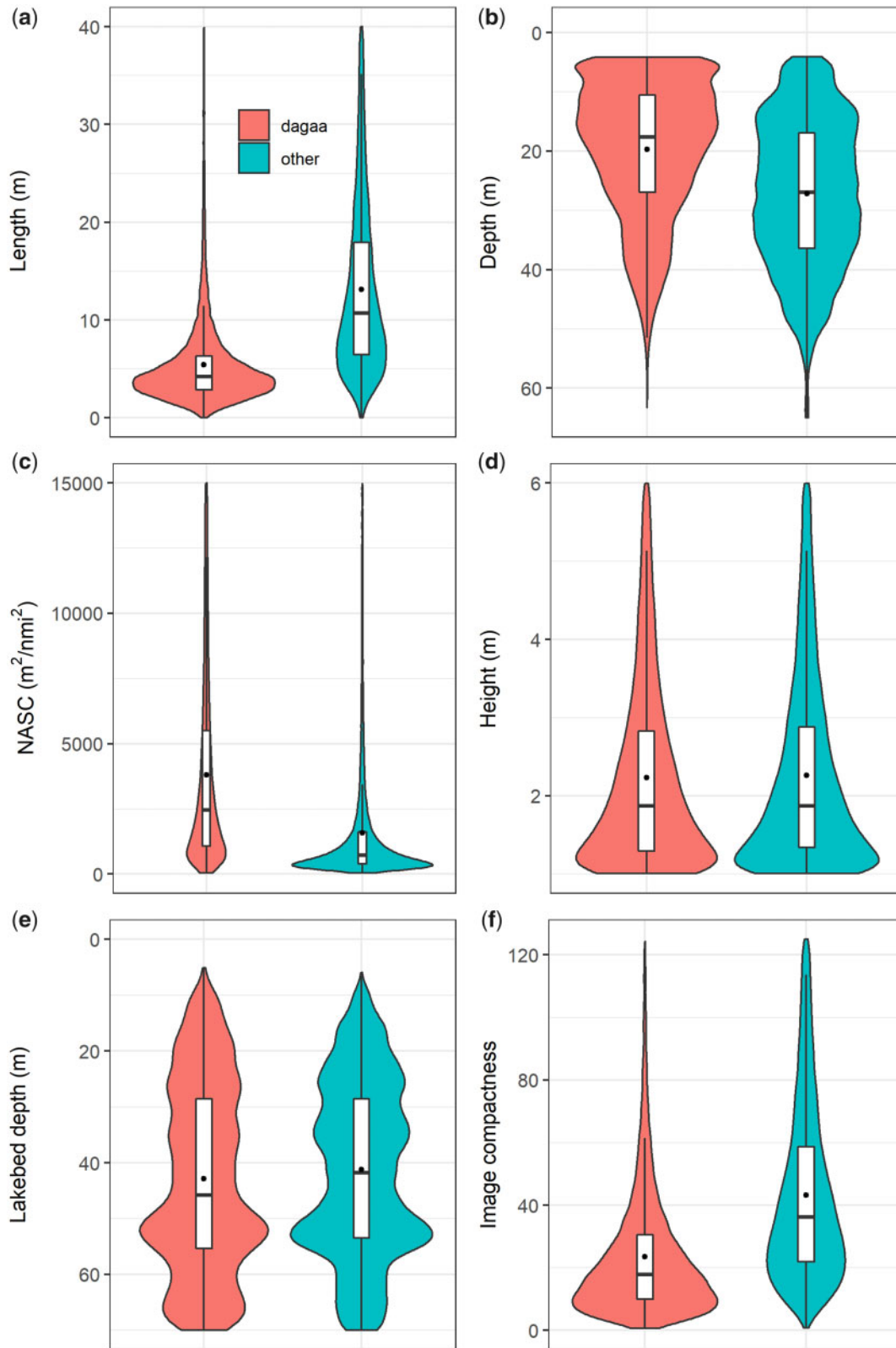
### Importance of different school metrics to overall RF model effectiveness

Evaluating the importance of each school metric (Table 1) to the RF model, regardless of any correlation between the metrics (known as “conditional variable importance”), showed that



**Figure 4.** Relative variable importance (conditional variable importance normalized between 0 and 1) for school metrics used to build the RF model.

school length was the most important metric, followed by school height, school NASC, school depth, school image compactness, and lakebed depth (Figure 4). Environmental variables other than lakebed depth contributed very little to the overall predictive power of the model, and when all environmental information was removed, the overall RF accuracy reduced by only c. 1%. This suggests that, during the 2015 survey, school structure was not influenced strongly by environmental variability across Lake Victoria.



**Figure 5.** Violin plots, which show smoothed probability density, with boxplots overlaid, for important school metrics used in the RF model to classify dagaa schools. Plotted school metrics are: (a) school length, (b) school depth, (c) school nautical area scattering coefficient (NASC) value, (d) school height, (e) lakebed depth and (f) school image compactness. Black filled circles show distribution means.



### School metrics

Distributions of the school metrics found to be important for dagaa classification were plotted as violin (Hintze and Nelson, 1998) and box plots, displaying the first quartile (Q1), median (M), third quartile (Q3), and probability density of each distribution (Figure 5). Dagaa school length (Figure 5a) (Q1 = 2.88 m; M = 4.23 m; Q3 = 6.34 m) was found to be significantly different to non-dagaa school length (Q1 = 6.82 m; M = 11.63 m; Q3 = 20.87 m; Kolmogorov-Smirnov (KS) test:  $p < 0.001$ ;  $D = 0.53$ ), a finding which provides quantitative support for the descriptive picture painted by Getabu *et al.* (2003) of dagaa schools as “needles”. Both dagaa and non-dagaa schools were found across all lake strata (inshore, coastal, and deep—see Figure 5b), but dagaa schools were typically found shallower in the water column (Q1 = 10.51 m; M = 17.63 m; Q3 = 26.94 m) than the non-dagaa schools (Q1 = 16.97 m; M = 27 m; Q3 = 36.48 m), which is some limited endorsement of the simple “top third” rule (but note that there are many dagaa schools in deeper water that the third-rule does not capture). School heights were similar between dagaa and non-dagaa schools (Figure 5d), but image compactness values of dagaa (Q1 = 10.02; M = 18.03; Q3 = 31.3) were significantly smaller (KS test:  $p < 0.001$ ;  $D = 0.38$ ) than non-dagaa (Q1 = 23.38; M = 40.29; Q3 = 72.19), i.e. in equidimensional  $x, y$  space, dagaa schools were paradoxically actually the more circle like in appearance: although appearing as needle-like features in echograms, if the aspect ratio of the image was to be set to 1:1, dagaa schools would in fact appear as squashed circles with a median length and height of 4.23 and 1.68 m, respectively (see Figure 5).

### Discussion

We have developed a new automated and standardized method to classify schools extracted from Lake Victoria echosounder data as either dagaa or non-dagaa using an RF model. The RF model had a school classification accuracy of 85.4% as judged against a test dataset of 9 816 manually classified schools. When used to classify all detected schools, the RF model picked out schools that resulted in a total lake-wide biomass of c. 0.71 MT, which was within c. 4% of the biomass derived from schools classified manually as being dagaa (0.68 MT): bootstrapped confidence limits for biomasses arising from manual and RF classification overlapped.

### Implications for fish stock management

The dagaa biomass estimate reported here of c. 0.7 MT is likely to be an underestimate for several reasons, such as: (i) since vertical echosounders are used to collect the data, and because dagaa are known to occupy shallow depths, some of the signal will be lost in the acoustic near-field (approximately the top 1.85 m for the 120 kHz transducer presently used); (ii) a component of the fish population may respond to the vessel (most likely avoiding, but possibly being attracted) (Brehmer *et al.*, 2019); and (iii) dagaa schools observed at present survey settings are relatively narrow (just a few pings in length) and it is possible that some particularly narrow schools (<1 m in length in some dimension) are not detected because the distance along survey track between consecutive pings, from beam edge to beam edge, is >school length. Although conservative, the biomass is determined by what will be a reproducible method that will be able to deliver an internally consistent relative index of variability over the years that will be valuable for management under the precautionary approach

(Francis, 1996). Suggestions for progressing towards absolute dagaa stock estimation are given below, and include the use of multibeam sonar to sample the near surface.

### Method performance

The RF classifier provides a robust and consistent means of dagaa school classification that, assuming software capable of performing school extraction is available, is both time- and cost-effective: the RF approach can achieve in minutes a classification task that, for the November 2015 survey (classifying manually 120 181 schools), took c. 100 person hours. The RF method will enable repeatable estimates of dagaa stock biomass to be calculated (estimates that would not be subject to any potential expert operator bias) and make this component of the stock assessment process resilient to the loss of expertise that might arise due to changes in personnel. Assuming stability in school morphology over time (and there is evidence from stocks of other pelagic species that this is likely; Cox *et al.*, 2011; Brierley and Cox, 2015), the RF method will enable the reanalysis of historic data (there are ~20 pre 2015 surveys), and future surveys (surveys are accumulating at ~2 per year presently) in an equivalent manner to produce robust and consistent time series.

One of the strengths of the RF classifier is that it uses actual length/widths/echo energies of schools to identify them, rather than relying on a visual interpretation of a feature the appearance of which will be influenced by vessel speed, ping rate, colour scale, feature depth, and echosounder beam angle (see Diner, 2001). In recognition of these potential impediments to successful and reliable visual classification, the RF model was built using only schools detected at usual survey speeds (8–10 knots), so avoiding the distortion in school appearance at the extremes of vessel speed that we believe is at the root of the differences in numbers of schools classified as dagaa/non-dagaa by RF and manual methods. In future studies relying on visual identification to test AI approaches, prior to visual scrutiny, echosounder data should be resampled in distance such that ping width is constant and consistent with typical survey settings. Changes in school width with depth (as the acoustic beam widens) should also be accounted for (Diner, 2001).

### Potential for future development

Vertically orientated echosounders commonly used in fish stock assessment (Fernandes *et al.*, 2002; Simmonds and MacLennan, 2005) have very narrow beam widths at short range (at 10 m range, the acoustic beam of the 120 kHz echosounder used in this study has a width of c. 1.2 m) and so offer a limited window of observation on species that inhabit the near surface. Consequently, the pelagic trawl used to fish near-surface dagaa schools in this study (Table 2 and Figure 3) would likely have encountered many schools that were not detected acoustically. Near-field effects also mean that echo returns from close to the transducer (c. 1.85 m in the case of the 120 kHz transducer used here) are not quantitatively reliable, such that typical surveys are effectively blind to the top few metres, potentially missing biomass. Use of multibeam sonar, instruments that typically sample a fan of acoustic beams spanning up to 180° beneath the vessel, or horizontally oriented echosounders, can open a window on the near surface (Gerlotto *et al.*, 1999; Paramo *et al.*, 2010). Multibeam has been used to make 3D measurements of fish schools at or close to the surface, and has also delivered valuable

data on the scale of avoidance by schools of research vessels (Gerlotto *et al.*, 2004). Incorporating multibeam instrumentation into Lake Victoria fish stock assessment surveys would effectively increase the volume of the lake sampled, provide valuable information with regard to school morphology, lead to more school detections for a given area (which could be readily integrated into the RF model) and hence reduce uncertainty in fish stock biomass estimates.

The RF model was trained and tested using data collected during a single survey (it was impractical to try to manually classify schools from more than one survey given available resources), but a future objective is to apply the RF classifier to the full range of available survey data (1997–present). We will need then to be wary of the potential for seasonal and/or annual changes in school characteristics. Lake Victoria shows strong seasonal physical change between fully mixed in the rainy season and stratified in the dry season. Deeper waters can become oxygen depleted in stratified times (Njiru *et al.*, 2012), and this may serve to vertically restrict dagaa habitat. Vertical habitat compression has been reported in the seas off Peru when the oxycline shallows (Bertrand *et al.*, 2008). Year-to-year variability in school structure may be less important: work on a variety of species over years spanning strong fluctuations in stock biomass has suggested that school shape does not vary significantly, but rather that it is the number of schools that varies with fluctuations in stock biomass (Brierley and Cox, 2015).

Between 2005 and 2014, total Lake Victoria fish stock biomass (including dagaa, Nile Perch, haplochromine cichlid species and others) has, on the basis of echosounder data analysis, appeared to be stable at c. 2.5 MT (Taabu-Munyaho *et al.*, 2016): the biomasses of Nile Perch and dagaa have both appeared to fluctuate from year to year, but in opposite directions. It is questionable how this apparent total biomass invariability can be ecologically possible given the greatly varying sizes, trophic levels, and ages-at-maturity of dagaa and Nile Perch. How much of this apparent zero-sum game is an artefact of fish not obeying the “top third” rule remains to be determined and will be the subject of an investigation that the repeatable RF classifier developed here will enable.

The next step will be to recalculate the time series of dagaa biomass from school information extracted from 20 years’ worth of acoustic survey data. This will be achieved by (i) pre-processing of the historic acoustic survey data (e.g. filtering noise spikes, which may resemble dagaa schools) and collating calibration results; (ii) building a new training dataset, composed of schools manually classified in different seasons and years, to study temporal changes in dagaa distribution, and investigate the validity of the “top third” method and drift in RF model parameters across the time series; (iii) applying geostatistical and or maximum entropy methods (Petitgas, 2001; Brierley *et al.*, 2003) to map dagaa echo intensity; and (iv) converting echo intensity to biomass using the latest measurements of dagaa *TS* and length–weight relationships derived from catch data. The new Lake Victoria dagaa biomass time series will enable any emerging interannual fluctuations in biomass to be considered in light of annual catches and environmental variability.

### Concluding remarks

The work reported here is a first step in moving Lake Victoria fisheries data analysis towards a fully automated processing chain built on machine learning and AI methods. Due to the automated

nature of these methods, time-series reanalysis will no longer be impractical and a severe drain on resources, but will be achievable rapidly with minimal manual effort. This will pave the way for a spectrum of studies on spatial and temporal variability in species distributions and progress along the road to ecosystem-based management of Lake Victoria fisheries, and to underpinning sustainable economy and food security in East Africa (Kolding *et al.*, 2019).

### Acknowledgements

We thank the participants of a workshop (Hydro-acoustic data analysis using R) funded by the University of St Andrews and the University of Strathclyde held at the Kenya Marine and Fisheries Research Institute (KMFRI) 17–21 June 2019 for insightful discussion of the RF method.

### Funding

Acoustic assessments of fish stocks in Lake Victoria have been supported by multiple funders over the years. The dagaa classification reported here was supported specifically by several Scottish Funding Council Global Challenge Research Fund (GCRF) grants from the University of St Andrews and the University of Strathclyde, by a GCRF Networking Grant to ASB and RJK from the UK Academy of Medical Sciences (GCRFENG\100371), and a Royal Society International Collaboration Award to ASB and Rhoda Tumwebaze, LVFO (ICA\R1\180123).

### References

- Barange, M. 1994. Acoustic identification, classification and structure of biological patchiness on the edge of the Agulhas Bank and its relation to frontal features. *South African Journal of Marine Science*, 14: 333–347.
- Bertrand, A., Gerlotto, F., Bertrand, S., Gutiérrez, M., Alza, L., Chipollini, A., Díaz, E., *et al.* 2008. Schooling behaviour and environmental forcing in relation to anchoveta distribution: an analysis across multiple spatial scales. *Progress in Oceanography*, 79: 264–277.
- Bingham, B., Kraus, N., Howe, B., Freitag, L., Ball, K., Koski, P., and Gallimore, E. 2012. Passive and active acoustics using an autonomous wave glider. *Journal of Field Robotics*, 29: 911–923.
- Brautaset, O., Waldeland, A. U., Johnsen, E., Malde, K., Eikvil, L., Salberg, A., and Handegard, N. O. 2020. Acoustic classification in multifrequency echosounder data using deep convolutional neural networks. *ICES Journal of Marine Science*, 77: 1391–1400.
- Brehmer, P., Sarré, A., Guennégan, Y., and Guillard, J. 2019. Vessel avoidance response: a complex tradeoff between fish multisensory integration and environmental variables. *Reviews in Fisheries Science & Aquaculture*, 27: 380–391.
- Breiman, L. 2001. Random forests. *Machine Learning*, 45: 5–32.
- Brierley, A. S., Gull, S. F., and Wafy, M. H. 2003. A Bayesian maximum entropy reconstruction of stock distribution and inference of stock density from line-transect acoustic-survey data. *ICES Journal of Marine Science*, 60: 446–452.
- Brierley, A. S., and Cox, M. J. 2010. Shapes of krill swarms and fish schools emerge as aggregation members avoid predators and access oxygen. *Current Biology*, 20: 1758–1762.
- Brierley, A. S., and Cox, M. J. 2015. Fewer but not smaller schools in declining fish and krill populations. *Current Biology*, 25: 75–79.
- Brierley, A. S., Ward, P., Watkins, J. L., and Goss, C. 1998. Acoustic discrimination of Southern Ocean zooplankton. *Deep Sea Research Part II: Topical Studies in Oceanography*, 45: 1155–1173.

- Christin, S., Hervet, É., and Lecomte, N. 2019. Applications for deep learning in ecology. *Methods in Ecology and Evolution*, 10: 1632–1644.
- Coetzee, J. 2000. Use of a shoal analysis and patch estimation system (SHAPES) to characterise sardine schools. *Aquatic Living Resources*, 13: 1–10.
- Cohen, J. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20: 37–46.
- Cox, M. J., Watkins, J. L., Reid, K., and Brierley, A. S. 2011. Spatial and temporal variability in the structure of aggregations of Antarctic krill (*Euphausia superba*) around South Georgia, 1997–1999. *ICES Journal of Marine Science*, 68: 489–498.
- D’Elia, M., Patti, B., Bonanno, A., Fontana, I., Giacalone, G., Basilone, G., and Fernandes, P. G. 2014. Analysis of backscatter properties and application of classification procedures for the identification of small pelagic fish species in the Central Mediterranean. *Fisheries Research*, 149: 33–42.
- De Robertis, A., Lawrence-Slavas, N., Jenkins, R., Wangen, I., Mordy, C. W., Meinig, C., Levine, M., *et al.* 2019. Long-term measurements of fish backscatter from saildrone unmanned surface vehicles and comparison with observations from a noise-reduced research vessel. *ICES Journal of Marine Science*, 76: 2459–2470.
- Demer, D. A., Andersen, L. N., Basset, C., Berger, L., Chu, D., Condiotty, J., Cutter, G. R., *et al.* 2017. 2016 USA–Norway EK80 Workshop Report: Evaluation of a Wideband Echosounder for Fisheries and Marine Ecosystem Science. ICES Cooperative Research Report No. 336. 69 pp.
- Demer, D. A., Berger, L., Bernasconi, M., Bethke, E., Boswell, K. M., Chu, D., Domokos, R., *et al.* 2015. Calibration of Acoustic Instruments. ICES Cooperative Research Report No. 326. 133 pp.
- Diner, N. 2001. Correction on school geometry and density: approach based on acoustic image simulation. *Aquatic Living Resources*, 14: 211–222.
- Escobar-Flores, P. C., Ladroit, Y., and O’Driscoll, R. L. 2019. Acoustic assessment of the micronekton community on the Chatham Rise, New Zealand, using a semi-automated approach. *Frontiers in Marine Science*, 6:507.
- Escobar-Flores, P. C., O’Driscoll, R. L., and Montgomery, J. C. 2018. Predicting distribution and relative abundance of mid-trophic level organisms using oceanographic parameters and acoustic backscatter. *Marine Ecology Progress Series*, 592: 37–56.
- Fallon, N. G., Fielding, S., and Fernandes, P. G. 2016. Classification of Southern Ocean krill and icefish echoes using random forests. *ICES Journal of Marine Science*, 73: 1998–2008.
- Fernandes, P. G., Gerlotto, F., D. V. Nakken, O., and Simmonds, E. J. 2002. Acoustic applications in fisheries science: the ICES contribution. *ICES Marine Science Symposia*, 215: 483–492.
- Fernandes, P. G. 2009. Classification trees for species identification of fish-school echotraces. *ICES Journal of Marine Science*, 66: 1073–1080.
- Fleischman, S., and Burwen, D. L. 2003. Mixture models for the species apportionment of hydroacoustic data, with echo-envelope length as the discriminatory variable. *ICES Journal of Marine Science*, 60: 592–598.
- Footo, K. G., Knudsen, H. P., and Vestnes, G. 1983. Standard calibration of echo sounders and integrators with optimal copper spheres. *Fiskeridirektoratet, Havforskningsinstituttet*, 17: 335–346.
- Francis, J. M. 1996. Nature Conservation and the Precautionary Principle. *Environmental Values*, 5: 257–264.
- Gerlotto, F., Castillo, J., Saavedra, A., Barbieri, M. A., Espejo, M., and Cotel, P. 2004. Three-dimensional structure and avoidance behaviour of anchovy and common sardine schools in central southern Chile. *ICES Journal of Marine Science*, 61: 1120–1126.
- Gerlotto, F., Soria, M., and Fréon, P. 1999. From two dimensions to three: the use of multibeam sonar for a new approach in fisheries acoustics. *Canadian Journal of Fisheries and Aquatic Sciences*, 56: 6–12.
- Getabu, A., Tumwebaze, R., and MacIennan, D. N. 2003. Spatial distribution and temporal changes in the fish populations of Lake Victoria. *Aquatic Living Resources*, 16: 159–165.
- Goudswaard, P. C., Witte, F., and Katunzi, E. F. B. 2008. The invasion of an introduced predator, Nile perch (*Lates niloticus*, L.) in Lake Victoria (East Africa): chronology and causes. *Environmental Biology of Fishes*, 81: 127–139.
- Greene, C., Meyer-Gutbrod, E., McGarry, L., Hufnagle, L., Chu, D., McClatchie, S., Packer, A., *et al.* 2014. A wave glider approach to fisheries acoustics: transforming how we monitor the Nation’s Commercial Fisheries in the 21st century. *Oceanography*, 27: 168–174.
- Haralabous, J., and Georgakarakos, S. 1996. Artificial neural networks as a tool for species identification of fish schools. *ICES Journal of Marine Science*, 53: 173–180.
- Hintze, J. L., and Nelson, R. D. 1998. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52: 181.
- Kayanda, R., Everson, I., Munyaho, T., and Mgaya, Y. 2012. Target strength measurements of Nile perch (*Lates niloticus*: Linnaeus, 1758) in Lake Victoria, East Africa. *Fisheries Research*, 113: 76–83.
- Klevjer, T., Tarling, G., and Fielding, S. 2010. Swarm characteristics of Antarctic krill *Euphausia superba* relative to the proximity of land during summer in the Scotia Sea. *Marine Ecology Progress Series*, 409: 157–170.
- Kolding, J., van Zwieten, P., Marttin, F., Funge-Smith, S., and Poulain, F. 2019. Freshwater Small Pelagic Fish and Fisheries in Major African Lakes and Reservoirs in Relation to Food Security and Nutrition. FAO Fisheries and Aquaculture Technical Paper No. 642. Rome. 124 pp.
- Korneliussen, R. J., Heggelund, Y., Macaulay, G. J., Patel, D., Johnsen, E., and Eliassen, I. K. 2016. Acoustic identification of marine species using a feature library. *Methods in Oceanography*, 17: 187–205.
- Kuhn, M. 2019. caret: Classification and Regression Training. R package version 6.0-84.
- Lawson, G. 2001. Species identification of pelagic fish schools on the South African continental shelf using acoustic descriptors and ancillary information. *ICES Journal of Marine Science*, 58: 275–287.
- LVFO. 2006. Lake Victoria Acoustic Survey August 2005: Report of an Analysis Workshop 3–14 October 2005 Mwanza, Tanzania and further analysis during March 2006. Jinja, Uganda. 58 pp.
- LVFO. 2012. The Standard Operating Procedures for Acoustic Surveys. Jinja, Uganda. 63 pp.
- LVFO. 2015. A Report of the Lake-Wide Hydro-Acoustic Survey 2015. Jinja, Uganda. 30 pp.
- LVFO. 2016a. Regional Catch Assessment Survey Synthesis Report June 2005 to December 2015. Jinja, Uganda. 35 pp.
- LVFO. 2016b. Regional Status Report on Lake Victoria Biennial Frame Surveys between 2000 and 2016. Jinja, Uganda. 87 pp.
- LVFO. 2018. Fisheries Management Plan III (FMP III) for Lake Victoria Fisheries 2016–2020. Jinja, Uganda. 52 pp.
- Madureira, L. S. P., Everson, I., and Murphy, E. J. 1993. Interpretation of acoustic data at two frequencies to discriminate between Antarctic krill (*Euphausia superba* Dana) and other scatterers. *Journal of Plankton Research*, 15: 787–802.
- Malde, K., Handegard, N. O., Eikvil, L., and Salberg, A. 2019. Machine intelligence and the data-driven future of marine science. *ICES Journal of Marine Science*, 77: 1274–1285.
- Mangeni-Sande, R., Taabu-Munyaho, A., Ogutu-Ohwayo, R., Nkalubo, W., Natugonza, V., Nakiyende, H., Nyamweya, C. S., *et al.* 2019. Spatial and temporal differences in life history parameters of *Rastrineobola argentea* (Pellegriin, 1904) in the Lake Victoria basin in relation to fishing intensity. *Fisheries Management and Ecology*, 26: 406–412.

- Mordy, C., Cokelet, E., De Robertis, A., Jenkins, R., Kuhn, C., Lawrence-Slavas, N., Berchok, C., *et al.* 2017. Advances in ecosystem research: saildrone surveys of oceanography, fish, and marine mammals in the Bering Sea. *Oceanography*, 30: 113–115.
- Nero, R. W., and Magnuson, J. J. 1989. Characterization of patches along transects using high-resolution 70-kHz integrated acoustic data. *Canadian Journal of Fisheries and Aquatic Sciences*, 46: 2056–2064.
- Njiru, M., Nyamweya, C., Gichuki, J., Mugidde, R., Mkumbo, O., and Witte, F. 2012. Increase in anoxia in Lake Victoria and its effects on the fishery. *In* *Anoxia*, pp. 99–128. Ed. by P. Padilla. InTech, Rijeka.
- Odongkara, K., Yongo, E., and Mhagama, F. 2016. The State of Lake Victoria Dagaa *Rastrineobola argentea*: Quantity, Quality, Value Addition, Utilization and Trade in the East African Region, for Improved Nutrition, Food Security and Income. Report of the EAC and Lake Victoria Fisheries Organisation. 87 pp.
- Okedi, J. 1981. The Engraulicrpris “dagaa” fishery of Lake Victoria with special reference to the southern waters of the lake. *In* *Proceedings of the Workshop of the Kenya Marine and Fisheries Research Institute on Aquatic Resources of Kenya*, 13–19 July. Mombasa.
- Paramo, J., Gerlotto, F., and Oyarzun, C. 2010. Three dimensional structure and morphology of pelagic fish schools. *Journal of Applied Ichthyology*, 26: 853–860.
- Petitgas, P. 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. *Fish and Fisheries*, 2: 231–249.
- Proud, R., Cox, M. J., and Brierley, A. S. 2017. Biogeography of the global ocean’s mesopelagic zone. *Current Biology*, 27: 113–119.
- Proud, R., Cox, M. J., Wotherspoon, S., and Brierley, A. S. 2015. A method for identifying sound scattering layers and extracting key characteristics. *Methods in Ecology and Evolution*, 6: 1190–1198.
- Reid, D., Scalabrin, C., Petitgas, P., Masse, J., Aukland, R., Carrera, P., and Georgakarakos, S. 2000. Standard protocols for the analysis of school based data from echo sounder surveys. *Fisheries Research*, 47: 125–136.
- Ripley, B. 2019. *tree: Classification and Regression Trees*. R package version 1.0-40.
- Scalabrin, C., Diner, N., Weill, A., Hillion, A., and Mouchot, M. C. 1996. Narrowband acoustic identification of monospecific fish shoals. *ICES Journal of Marine Science*, 53: 181–188.
- Sharpe, D. M. T., and Chapman, L. J. 2014. Niche expansion in a resilient endemic species following introduction of a novel top predator. *Freshwater Biology*, 59: 2539–2554.
- Simmonds, E. J., and MacLennan, D. N. 2005. *Fisheries Acoustics: Theory and Practice*. 2nd edn. Blackwell Science, Oxford. 437 pp.
- Simmonds, E. J., Armstrong, F., and Copland, P. J. 1996. Species identification using wideband backscatter with neural network and discriminant analysis. *ICES Journal of Marine Science*, 53: 189–195.
- Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36: 111–133.
- Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, 9: 307.
- Strobl, C., Hothorn, T., and Zeileis, A. 2009. Party on! *The R Journal*, 1: 14.
- Taabu-Munyaho, A., Marshall, B. E., Tomasson, T., and Marteinsdottir, G. 2016. Nile perch and the transformation of Lake Victoria. *African Journal of Aquatic Science*, 41: 127–142.
- Taabu-Munyaho, A., Nyamweya, C. S., Sitoki, L., Kayanda, R., Everson, I., and Marteinsdóttir, G. 2014. Spatial and temporal variation in the distribution and density of pelagic fish species in Lake Victoria, East Africa. *Aquatic Ecosystem Health & Management*, 17: 52–61.
- Tarling, G. A., David, P., Guerin, O., and Buchholz, F. 2001. The Swarm Dynamics of Northern Krill (*Meganyctiphanes norvegica*) and Pteropods (*Cavolinia inyexa*) during Vertical Migration in the Ligurian Sea Observed by an Acoustic Doppler Current Profiler, 16: 1–16.
- Tegowski, J., Gorska, N., and Klusek, Z. 2003. Statistical analysis of acoustic echoes from underwater meadows in the eutrophic Puck Bay (southern Baltic Sea). *Aquatic Living Resources*, 16: 215–221.
- Tumwebaze, R. 2003. *Hydroacoustic Abundance Estimation and Population Characteristics of Rastrineobola argentea in Lake Victoria*. University of Hull, Hull, UK.
- Wanink, J. H. 1999. Prospects for the fishery on the small pelagic *Rastrineobola argentea* in Lake Victoria. *Hydrobiologia*, 407: 183–189.

Handling editor: Olav Godo