5-29-2019

A Study to Assess the Effect of Tow Duration and Estimate Dredge Efficiency for the VIMS Sea Scallop Dredge Survey : Final Report

David Rudders
Virginia Institute of Marine Science

Sally Roman
Virginia Institute of Marine Science

Arthur Trembanis

Danielle Ferraro

Recommended Citation

Rudders, D., Roman, S., Trembanis, A., & Ferraro, D. (2019) A Study to Assess the Effect of Tow Duration and Estimate Dredge Efficiency for the VIMS Sea Scallop Dredge Survey : Final Report. Marine Resource Report No. 2019-04. Virginia Institute of Marine Science, William & Mary. doi: 10.25773/g9sh-qt28

Final Report

A Study to Assess the Effect of Tow Duration and Estimate Dredge Efficiency for the VIMS Sea Scallop Dredge Survey

Award Number: NA17NMF4540044
VIMS Marine Resource Report No. 2019-04

Submitted to:

National Marine Fisheries Service
Northeast Fisheries Science Center
Cooperative Research Program
166 Water Street
Woods Hole, Massachusetts 02543-1026

Submitted by:

David B. Rudders[1]
Sally Roman[1]
Arthur Trembanis[2]
Danielle Ferraro[2]

[1]Virginia Institute of Marine Science
William & Mary
Gloucester Point, Virginia 23062

[2] School of Marine Science and Policy
University of Delaware
Newark, Delaware 19716

May 29, 2019

Contents

Project Summary

For the sea scallop, *Placopecten magellanicus*, the concepts of space and time have emerged as the basis of an effective management tool. The strategy of closing or limiting activities in certain areas for specific lengths of time has gained support as a method to conserve and enhance the sea scallop resource. In the last decade, rotational area management has provided a mechanism to protect juvenile scallops from fishing mortality by closing areas based upon scallop abundance and age distribution. Approximately half of the sea scallop industry's current annual landings come from areas under this rotational harvest strategy. While this represents a management success, it also highlights the extent to which landings are dependent on the success of this strategy. The continued prosperity of scallop spatial management is dependent on both periodic and large incoming year classes, as well as, a mechanism to delineate the scale of a recruitment event and subsequently monitor the growth and abundance of these scallops over time. Current and accurate information related to the abundance and distribution of adult and juvenile scallops is essential for managers to respond to changes in resource subunits.

The sea scallop fishery is typically supported by several primary survey methods (i.e., dredge and optical surveys), which provide multiple, spatially explicit biomass estimates on an annual basis. From 2015 - 2018 significant divergence in area-specific biomass estimates between the different survey methods was noted. The divergent estimates were associated with areas of high scallop density within the Nantucket Lightship Access Area (NLCA) and the Elephant Truck Access Area (ETCA).

The working hypothesis behind the divergent estimates focused upon a potential gear saturation effect for the survey dredge that impacted dredge performance resulting in a reduction in efficiency in high density areas. If dredge efficiency is reduced as a function of increasing density, then applying a constant efficiency estimate to scale relative biomass to absolute biomass may result in an underestimate of biomass in high density areas. While several independent sources of biomass estimates are beneficial for successful management of the resource, divergent area-specific estimates can contribute to uncertainty for setting of annual specifications and assessment of the resource.

The current study consisted of two objectives. The first objective was to conduct an experiment that would result in an understanding of the underlying processes that contributed to the observed variability in dredge efficiency. The second objective was to provide an empirical basis to mitigate this deviation of performance via a the examination of the effect of tow duration on scallop catch and document the effect on dredge performance by reducing the standard 15-minute tow time to 10 minutes.

Results for the dredge efficiency portion of the study suggested that within the standardized experimental protocol, gear efficiency was observed to be reduced at scallop densities greater than two scallops per m^2. The average efficiency at high densities (>2 scallop/ $m^{2)}$ was estimated at 0.135, which is significantly lower than the assumed value of 0.40 for soft substrate. This result is similar to a previous dredge efficiency study completed by the Northeast Fisheries Science Center (NEFSC) with 2016 and 2017 data from VIMS and NEFSC. Observed dredge efficiency at the station-level, however, was rarely equal to or greater than the assumed value of 0.40. The decline in efficiency was non-linear and once efficiency attained this lower level it remained relatively constant. Results from the tow duration component of the study did not suggest that the shorter duration negatively impacted scallop catch rates. Catch rates between the two durations were similar in the Mid-Atlantic and NLCA. Given this result, we conclude that a ten-minute tow may not be short enough to address the potential gear saturation issue.

Project Background

The sea scallop, *Placopecten magellanicus*, supports a fishery that, in 2017, landed 41 million pounds of meats with an ex-vessel value of over US $511 million (NMFS, 2018). These landings resulted in the sea scallop fishery being one of the most valuable wild caught single species fishery on the East Coast of the United States. While historically subject to extreme cycles of productivity, the fishery has benefited from recent management measures intended to bring stability and sustainability, as well as a data rich situation resulting from dedicated research funded through the industry supported Sea Scallop Research Set Aside (RSA) Program.

These funding sources typically allow for several dredge and optical surveys to be conducted on an annual basis at various spatial scales. Biomass estimates from these surveys are made available to managers and stock assessment scientists for use in setting specifications for the upcoming fishing year and to manage rotational access areas on an annual basis. Beginning in 2015, divergence in SAMS (Scallop Area Management Simulator) area-specific biomass estimates between the different survey methods was observed (Figure 1). The divergence in biomass estimates seemed to exist for the high density areas in some portions of the Elephant Truck Access Area (ETCA) and Nantucket Lightship Access Area (NLCA). One suggestion put forth to explain this discrepancy is a potential gear saturation effect for dredge gear. A preliminary examination of the 2016 - 2017 area-specific biomass estimates for the VIMS dredge survey, the Northeast Fisheries Science Center (NEFSC) dredge survey and NEFSC HabCam optical survey supported the hypothesis that that dredge efficiency was reduced at higher scallop densities (NEFSC, 2018) (Figure 2).

Gear saturation may be occurring in the dredge for several reasons and affecting gear performance. One potential explanation is that there is a scallop density effect on gear performance. As the dredge becomes filled over the course of a tow in high density scallop areas, scallops may not be retained in the dredge during the latter part of the tow (Shumway and Parsons, 2006). Another possible reason for gear saturation is similar in that for areas of high sand dollar abundance the dredge may become filled with sand dollars and scallops may not be retained once the dredge is full (Shumway and Parsons, 2006). If fewer scallops are captured by the gear under these conditions, applying the assumed dredge efficiency value of 0.40 will lead to an underestimate of scallop biomass (NEFSC, 2018). Ultimately, the process of dredge filling is a candidate for the observed reduction in efficiency and biomass estimates that appear to be lower than the optical surveys.

The two main objectives of this project were to conduct a tow duration study in the Mid-Atlantic Bight (MAB) resource area and conduct a dredge efficiency study in the NLCA resource area. The project provides an analysis of the effect of a reduced tow time on the catch rate of scallops. The project also attempts to understand the

underlying processes that are possibly contributing to reduced efficiency by directly examining dredge gear performance and gaining both insight and empirical evidence to evaluate whether dredge efficiency is compromised and under what conditions this may occur. The project contributes additional information to other research conducted by the Virginia Institute of Marine Science (VIMS) focusing on survey dredge performance. This information is necessary to understand dredge gear performance, validate efficiency assumptions and understand why there has been divergent biomass estimates from various survey techniques. The project can also aid in improving dredge survey biomass results by providing an experimental basis to improve survey protocols and reduce potential bias that may be occurring at high scallop densities. This information will aid in reducing uncertainty associated with the annual specification setting process and how to treat recent survey dredge data in the assessment process.

Methods

Tow Duration

Field Study

A tow duration experiment using a paired tow design was conducted onboard the F/V *Nancy Elizabeth* in the MAB region to examine the effect of reduced tow duration on scallop catch and scallop length distribution. Tow pairs were completed within the VIMS MAB fishery independent dredge survey domain with data from the 2017 survey used to inform site selection to ensure tows would be representative of the gradient of scallop and sand dollar densities characteristic of the area. The paired tow design allows for advanced analyses like GLMMs to be utilized and minimizes between haul variability.

At each selected location, a 15-minute and 10-minute tow were conducted. The 15-minute tow represented the standard survey tow duration and the 10-minute tow duration representing a reduced tow duration time based on recommendations from the Scallop Survey Peer Review Panel (SSSMPRT, 2015). An alternating paired towing approach was used with an ABBA BAAB method, where A was the 15-minute tow and B was the 10-minute. Tows were made in the same direction and area as close in time as possible. All other procedures for fishing the sampling gear followed standard survey protocols (i.e., gear configuration, towing protocols, catch sampling). A standardized National Marine Fisheries Service (NMFS) sea scallop survey dredge, 2.4 m (8 feet) in width equipped with 2-inch rings, 3.5-inch diamond mesh twine top and a 1.5-inch diamond mesh liner was used for the project.

Sampling of the catch was performed using the protocols established by DuPaul and Kirkley (1995). For each tow pair, the entire scallop catch was placed in baskets. Depending on the total volume of the catch, a fraction of these baskets were measured for sea scallop length frequency. The shell height of each scallop in the sampled fraction was measured to the nearest millimeter (mm) using an electronic Ichthystick

measuring board. This protocol allows for the estimation of the size frequency for the entire catch by multiplying the catch at each shell height by the fraction of total number of baskets sampled. Finfish and invertebrate bycatch were quantified, with commercially important finfish and barndoor skates being sorted by species and measured to the nearest 1 mm (total length (TL).

Catch data (scallops, finfish, invertebrates, and trash) were entered into the data acquisition program Fisheries Environment for Electronic Data (FEED). Length measurements were recorded using an electronic Ichthystick measuring board integrated with the FEED program that allows for automatic recording of length measurements. The bridge log was also entered into FEED with an integrated GPS data stream. Recorded data included location, time, tow-time (break-set/haul-back), tow speed, water depth, weather and comments relative to the quality of the tow.

VIMS used the same experimental approach to conduct similar tow duration studies in the NLCA and Closed Area II (CAII) in 2016 and 2017. These studies were included as part of individual projects whose main objective was to conduct resource assessment surveys. Funding was provided by the Sea Scallop RSA program for all tow duration studies (NA16NMF4540044, NA16NMF4540042 and NA17NMF4540044). Data from all areas and years was synthesized for this analysis to allow for a larger sample size and encompass a broad range of spatial areas and resource conditions.

Analysis

Data analyses consisted of an initial visual examination of scallop and debris catch, as well as relative scallop length frequency distributions. A generalized linear model (GLMM) and a generalized additive model (GAM) were used to test for differences in scallop catch and catch at length. Scallop catch was analyzed by examining the expanded number of scallops captured, as well as the number of baskets caught. Debris was defined as all material (e.g., sand dollars, mud, rocks) left on deck after all scallops, finfish and skate bycatch were removed. Debris was put into bushel baskets to quantify catch. All analyses were conducted by area (i.e., CAII, NLCA and MAB).

A one-tailed analysis of variance (ANOVA) or a Wilcoxon rank sum test were used to test for differences in the mean scallop catch (number of animals) and debris catch (bushel baskets) between tow durations by area (Sokal and Rohlf, 1995). Assumptions required for an ANOVA (i.e., normality and homogeneity of variance) were tested for prior to implementing the appropriate test (Sokal and Rohlf, 1995). A one-tailed test was used, because there was no expectation that a 15-minute tow would catch less than a 10-minture tow. A Kolmogorov-Smimov (K-S) test as used to test for differences in the relative length frequency distributions of scallops between tow durations by area.

GLMMs and GAMs were developed following the approach of Holst and Revill (2009) and Miller (2013). GLMMs and GAMs fit the proportion of scallops caught at length in the 10-mintue tow conditioned on the total catch at length for a tow pair in both the 10 and 15-mintue tows. The Holst and Revill (2009) method uses a binomial polynomial GLMM where low order polynomial terms can be included as fixed effects to accommodate a non-linear response. The Miller (2013) approach fits several GAM variants with a cubic spline smoother across all pairs and within pairs using different error structures (i.e., binomial and beta-binomial). Fixed effects considered for GLMMs were area, length (mm), length^2, scallop catch (number of baskets), debris catch (number of baskets) and an interaction term of area and length^2. For GAMs, length was the fixed effect and area-specific models were developed. The random effect specified for both models was the tow pair. An offset term that accounted for both subsampling and differences in area swept was included in both models. Forward selection was used for GLMM model development and for both approaches the Akaike information criterion (AIC) was the basis for model selection (GLMM and GAM). The model with the lowest AIC was selected as the optimal model for both approaches. All analyses were completed in R v 3.3.2 (R Core Team, 2016).

Dredge Efficiency

Field Study

During August of 2017, a dredge efficiency experiment using a paired design was conducted onboard the F/V *Christian and Alexa* in the NLCA region with the objective of examining the effect of scallop density on dredge efficiency. Spatially, the pairs were completed within the VIMS NLCA fishery independent dredge survey domain and data from the 2017 survey were used to inform site selection to ensure tows would be representative of a gradient of scallop densities present. This paired design was similar to the design used by the NEFSC to conduct a survey dredge efficiency study in 2008 and 2009 (Miller et al., *in press*).

At each location, a paired survey dredge/autonomous underwater vehicle (AUV) pair was completed. The survey dredge was first towed following standard survey protocols, discussed above. After the survey dredge tow was completed and the catch sampled as described above, the AUV was deployed for a mission. The AUV mission covered the tow path, as well as the adjacent area around the tow path. Each mission consisted of four straight line transects 1,852 m in length, spaced approximately 5 m apart (Figure 3). The transect length of 1,852 m is similar to the nominal distance of 1,850 m covered during a standard survey dredge tow.

The survey dredge used in the study was the NMFS standard survey dredge, discussed above. The AUV utilized was the University of Delaware's Gavia, equipped with an integrated digital camera, flash strobe lighting system and side scan sonar (Figure 4) (Trembanis et al., 2017; Ferraro et al, 2017). Sensors onboard the AUV also

collected environmental data including depth, temperature, salinity and dissolved oxygen. Vehicle location, altitude, depth, pitch and roll were continuously recorded. The camera on the AUV was a Point Grey Grasshopper 14S5C/M-C model that took georeferenced photos with a Sony ICZ285AL CCD at a resolution of 1280 x 960 pixels (1.2 mega pixels). The camera was mounted within the nose module of the AUV and paired with a flash strobe on the control module for illumination. The camera was focused to take images at a distance of ~2.5 m above the seabed at an effective rate of 1.9 fps (Figure 5). At a constant vehicle altitude of 2.5 m and a viewing angle of 41.19 degrees, each image covered 1.88 m x 1.41 m (2.65 m^2) of seafloor with a resolution of 2 mm per pixel. Each image was collected in JPEG format with metadata (including latitude, longitude, depth, altitude, pitch, heading, roll) embedded in the header file. The 1,800 kHz high-frequency Marine Sonic side scan sonar acoustically imaged the seabed simultaneously with a 10 m range to image dredge scars from the survey dredge (Figure 6).

Catch sampling and data collection for the survey dredge tows were identical to the catch sampling protocols described above for the tow duration component of the project and followed protocols which have been utilized during all of VIMS scallop surveys since 2005.

Side-scan Sonar Processing

All side-scan sonar data collected were made into mosaics with SonarWiz 7 (Chesapeake Technology Inc.) and exported as georeferenced raster images (geotiff with world files). Dredge scars were visually detected in all missions by looking for a line that was roughly 2.4 m wide and smoother than the surrounding seabed, then manually digitized using the polyline tool in Sonarwiz. The outer bounds of the scar features were used to filter the image set to only those outside the scar bounds, so only the scallops outside the scar were included in density calculations (Figure 7).

Image Processing

Images were first enhanced using the multiscale retinex algorithm from Fred's ImageMagick Scripts with a color model and brightness gain to clarify the image contents (Weinhaus, 2007). University of Delaware server-side code parsed the embedded metadata from each JPEG, and both the images and associated metadata were subsequently loaded into a MySQL database. The database was integrated with a web-based image annotation system, accessible at robots.udel.edu/Scallop (Trembanis et al., 2017) (Figure 8).

Manual annotation of a subset of the images collected in this study was carried out by a set of trained human annotators for three reasons: (a) to provide a comparison of AUV scallop density to survey dredge density estimates, (b) to generate a training set for the YOLOv3 scallop detector, and (c) to generate density estimates and shell height

length frequencies for comparison with those output by the YOLOv3 detector. Every fourth image from a selected AUV mission was displayed sequentially for annotation. Every scallop in the image was counted, assigned a “healthy” or “compromised” rating, and, when clearly visible, sized by drawing a line from the hinge to the edge of the shell margin. If more than 50 percent of a scallop was not in the image, then the scallop was not counted. Compromised scallops were distinguished from healthy scallops by a shell in a nonlife position or the presence of a disarticulated or severely damaged shell. The dominant substrate, a rating of image clarity, and the presence or absence of scallops in the image was also noted. All image annotators were trained and given a sample set of 60 photos containing 150 scallops. In order to access the annotation system, annotators had to count within 5 percent of the total number of scallops in the images provided, as well as, the proportion of healthy and unhealthy scallops.

To measure shell height, annotators drew a line bisecting the scallop from the hinge to the shell margin. The scallops that were not measured were overlapped by other scallops, partially buried, a portion was out of the image frame, or otherwise obscured. This occurred primarily in areas of extremely high scallop density, where it was common to observe >100 scallops in a single image. Shell height was defined as:

$$SH = W_{pixel} \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

where SH is shell height in mm, W_{pixel} is the width of a pixel in mm, and X_2, X_1, Y_2, and Y_1 are the coordinates of the line segment annotated.

Convolutional Neural Network Development

In partnership with Dr. Christopher Rasmussen and graduate student Jiayi Zhao (University of Delaware, Department of Computer and Information Sciences), a convolutional neural network (CNN) was developed to detect and measure scallops in the AUV-derived images. Dr. Rasmussen’s lab built a CNN called “Scallopscan”, first based on the YOLOv2 architecture for object detection and classification, (Redmon and Farhadi, 2016), and recently upgraded to the improved YOLOv3 architecture (Redmon and Farhadi, 2018). The neural network offers state-of-the-art object detection at faster speeds than the live frame rate from the AUV. Unlike the image quality requirements for manual annotation, the network ran successfully on both enhanced and raw, un-enhanced images. Scallopscan was iteratively trained on a number of image sets during its development. The number of object categories it was configured to detect was one: “healthy scallop.” Scallops annotated as “unhealthy” were excluded from training, but those annotations comprised approximately 2.5 percent of the scallops annotated in non-treated (non-dredged) images. YOLOv3 uses a bounding box defined by the line annotations defined by human annotators that captured scallop size. The bounding box was defining using the line as the diameter of a circle inscribed in a bounding box. Scallopscan was implemented within the new YOLOv3 framework and

then trained with the line annotations. The training set used for YOLOv3 was comprised of 373,806 scallops from 97,344 images annotated from a previous incidental mortality RSA study (Trembanis et al., 2017; Ferraro et. al., 2017) and this study.

Convolutional Neural Network Implementation

Following training, Scallopscan was used to detect scallops in all of the images from the AUV missions in this study. Only a small fraction of the collected images were annotated, leaving the majority available for the Scallopscan run. A minimum confidence threshold of 0.3 was chosen to select detections for output, which had associated precision and recall values of 0.897 and 0.863, respectively.

Shell height for each detected scallop was estimated using the mean side length of the bounding box, defined as:

$$SH = W_{pixel} \frac{(X_2 - X_1) + (Y_2 - Y_1)}{2}$$

where SH is shell height in mm, W_{pixel} is width of a pixel in mm, and X_2, X_1, Y_2, and Y_1 are the coordinates of the upper left and lower right corners of the bounding box. This method was chosen to account for the potential of the scallop to be oriented in any direction, while the bounding box was always oriented in line with the sides of the image frame.

Density Estimates

The density of scallops for each dredge tow was calculated by dividing the total estimated number of scallops caught by the area swept (m^2) of the survey dredge. Area swept was calculated as the tow distance (m) estimated from the start and end coordinates of a tow multiplied by the survey dredge width of 2.4 m.

AUV mission scallop densities were calculated once using the manually annotated images and a second time using the images with detections from Scallopscan. First, the area photographed was calculated for all images collected. Image width was defined as:

$$W_{image} = 2\ H\ tan\left(\frac{a_h}{2}\right)$$

where W_{image} is image width in m, H is the height (altitude) of the AUV from the seafloor in m, and a_h is the underwater horizontal viewing angle of the camera, in degrees. Given the 4:3 aspect ratio of the camera, image length was calculated as:

$$L_{image} = 0.75\ W_{image}$$

Using the digitized bounds of the dredge scars, the centroid of each image was defined as being inside or outside of the scar plus a 1 m buffer using the sp package in

R (Bivand *et al.*, 2013; Pebesma and Bivand, 2005; R Core Team, 2016). Images that were completely outside of the dredge scar were used to calculate density, and the buffer was added to exclude the images immediately abutting the dredge scar perimeter. Images collected at vehicle altitudes of >10 m were removed from the data set in order to filter out any manual annotations or detections of poor quality. The Scallopscan image set was downsampled to every fourth image to remove double- or triple-counted scallops present in more than one image due to image overlap. This also allowed for a comparison of human annotated and YOLOv3 annotated data that was on the same image sampling scale. In addition, scallops in the Scallopscan data set with a shell height of <30 mm or >180 mm were removed to minimize the amount of known false positives. Scallop abundance per AUV mission-dredge pair was calculated by summing the number of scallops annotated or detected in each of the remaining images, and area imaged per pair was calculated by summing the areas. Density was defined as the number of scallops divided by the area imaged.

Analysis

Annotator Evaluation

To estimate variation of scallop count and shell height measurement between individual annotators, all but three annotators (n = 21) measured scallops in a subset of images containing 140 photos. Mean scallop count was compared across annotators, and a two factor ANOVA was used to test for differences between annotators (Sokal and Rohlf, 1995). The two factors included in the ANOVA were image and annotator. Assumptions required for an ANOVA (i.e. normality and homogeneity of variance) were evaluated (Sokal and Rohlf, 1995). Mean shell height was compared across annotators, and a heteroscedasticity-corrected Type II ANOVA was used to test for significant differences between annotators using the car package (R Core Team, 2016; Fox and Weisberg, 2011). In addition, the group measured the same five scallops in a single image ten times in order to estimate variation of replicate measurements within and between annotators. Either a one factor ANOVA or Kruskal-Wallis rank sum test was used to test for an effect of annotator on shell height measurements (Sokal and Rohlf, 1995; Kruskall and Wallis, 1952). Prior to implementing the tests, the required assumptions of an ANOVA (i.e., normality and homogeneity of variance) were evaluated (Sokal and Rohlf, 1995).

Annotator YOLOv3 Comparison

Since the number of missions annotated by individual annotators (n = 20) and YOLOv3 (n = 28) differed, comparisons were completed to determine if human annotated data and YOLOv3 data were similar. These comparisons allowed us to determine the validity of YOLOv3 annotated data, so that the entire set of paired tows could be considered for efficiency analysis to increase the sample size. Relative length frequency distributions pooled across all 20 missions that were annotated by both

humans and YOLOv3 were compared and tested for significant differences. The clus.lf function in the fishmethods package was used to perform a two-sample K-S test that accounts for a lack of independence in length measurements taken from the same station (R Core Team, 2016; Nelson, 2018). Length data were binned into 5 mm length bins. A two factor randomized block ANOVA was used to test for significant differences in the mean density estimates between the two annotation methods (Sokal and Rohlf, 1995). The ANOVA included annotator type (human or YOLOv3) and station as the two factors. Assumptions required for an ANOVA (i.e., normality and homogeneity of variance) were evaluated (Sokal and Rohlf, 1995).

Survey Dredge Annotation Comparison

Another analysis was completed to examine for differences in the relative length frequency distributions of the survey dredge data and the annotated data. This was completed for the human annotated data set and the YOLOv3 data set compared to the survey dredge data. A similar cluster K-S test using the clus.lf function was performed to test for significant differences in the length distributions (Nelson, 2018). Length data were also binned into 5 mm length bins.

A selectivity analysis using the SELECT method was completed for both the YOLOv3 and human annotated data sets to assess the assumption of 100 percent selectivity for optical survey methods (Millar, 1992; Millar and Fryer, 1999; NEFSC, 2018). We fit a logistic selection curve to the data, as this functional form has provided the best fit to scallop dredge and optical data and is the most common functional form observed for towed fishing gear selectivity studies (Millar, 1992; Yochum and DuPaul, 2008; Park et al., 2011, NEFSC, 2018). This analytical approach conditions the catch of the optical data at length l to the total catch from both gears (i.e., AUV and non-selective survey gear):

$$\Phi_c(l) = \frac{p_c \exp(a + bl)}{(1 - p_c) + exp(a + bl)}$$

where $\Phi_c(l)$ is the proportion of scallops-at-length observed, *a* and *b* are the logistic selection curve parameters, the intercept and slope respectively, l is the length of a scallop, and p_c is the split parameter and is the measure of relative efficiency for the annotated data compared to the survey dredge (Millar, 1992). The split parameter was estimated within the model because we had no *a priori* information to inform using an assumed value for the AUV data. The YOLOv3 data set model was fit using a maximum likelihood approach (Millar, 1992). The analysis was completed with the R statistical software and the trawlfunctions package (R Core Team, 2016). The trawlfunctions package documentation and code can be found at http://www.stat.auckland.ac.nz/~millar/selectware/code.html. For the human annotated data set, the model in R did not converge. These data were analyzed using an Excel

version of the SELECT method. The Excel template can be found at https://www.stat.auckland.ac.nz/~millar/selectware/EXCEL/. The model was fit using the Solver function in Excel. Parameter estimates for each data set were used to model the predicted selectivity curves for each annotation data set.

Survey Dredge Efficiency Analysis

Dredge efficiency was analyzed following a similar approach taken by the NEFSC (NEFSC, 2018). The ratio of dredge density to AUV density, also referred to as the capture efficiency, was plotted against the mean density (dredge density + AUV density/2) for each pair. A generalized additive model (GAM) was fit to the same data on the log scale using the gam function in the mgcv package to model the relationship between efficiency and density (Wood, 2011; R Core Team, 2016; Wood, 2017). The response variable was the density ratio and the explanatory variable was mean density. A thin plate regression spline was used as the smoother and smoothing functions were selected with generalized cross validation criterion.

Survey Dredge Biomass Estimation

Absolute biomass in metric tons (mt) for the survey dredge was calculated using several efficiency values to compare the relative performance of these values by SAMS area. Data from VIMS' 2018 NLCA and MAB surveys were used to estimate biomass. Biomass estimates were also compared to the NEFSC Habcam optical biomass estimates for 2018. NEFSC Habcam assumes 100 percent efficiency for scallops greater than 40 mm (NEFSC, 2018). Biomass estimates were calculated with an area swept method used by VIMS since 2015 (Rudders and Roman, 2018), following methods from Cochran (1977) for calculating a stratified random size of a population. Area-specific shell height meat weight relationships were used, based on the 2018 assessment (NEFSC, 2018). The following five approaches were used to scale relative biomass to absolute biomass for the survey dredge:

- 0.40 method - Apply the assumed 0.40 efficiency value across an entire survey domain (all stations).
- 0.13 method - Apply the lowest value of 0.135 predicted from the GAM model across an entire survey domain (all stations).
- 0.10 method - Apply a value of 0.10 across the entire survey domain. This is the value used by the sea scallop Plan Development Team (PDT) and in the 2018 benchmark assessment to adjust dredge efficiency (NEFSC, 2018).
- SAMS method - Apply either the assumed value of 0.40 or the lowest predicted GAM value of 0.135 at the SAMS-level, depending on past divergence with optical survey estimates. This approach was used by the sea scallop PDT in several of the past years to address survey dredge performance issues and in the benchmark 2018 assessment (NEFSC, 2018). A value of 0.135 was used in

the following SAMS areas in the NLCA: NL South Deep and NL NA (also referred to as NL West in 2018). In the remaining 4 SAMS areas the assumed 0.40 value was used. In the MAB, the reduced value of 0.135 was applied to the ET Flex SAMS areas, while the other 8 SAMS areas used the assumed 0.40 value.

- Station method - Apply a reduced value of 0.135 based on station-level density estimates. If the density at a station was greater than 2 scallops per m^2, the lower value of 0.135 was used. If station-level density was less than 2 scallop per m^2, the traditional 0.40 efficiency value was used.

Results

Tow Duration

Figure 9 shows the location of all tow duration pairs by area. Table 1 provides summary information by area. A total of 276 pairs were completed across the three study sites. The total expanded number of scallops caught, average scallop catch (expanded number) and results of parametric tests by area are provided in Table 2. There was no significant difference in the mean catch between the two tow durations for the MAB or NLCA, indicating the 10-minute tow caught a similar quantity of scallops compared to the 15-minute tow duration (Figure 10). There was a significant difference for CAII, with the 15-minute tow catching more scallops than the 10-minute tow (Table 2). Bland-Altman plots by area for the expanded number of scallops, debris catch and total catch (number of baskets of scallops + number of baskets of debris) are shown in Figures 11 – 13. CAII was the only area where the expectation of greater catch rates with increased tow duration held for scallop, debris and total catch. Table 3 shows debris catch, average debris catch and results of parametric tests by area. There were no significant differences in debris catch between the 10 and 15-minute tows. Relative length frequency distributions are provided in Figure 14. The K-S tests indicated there were no significant differences in length distributions between the two tow durations.

GLMM results indicated the optimal model had an interaction term of area and $length^2$, as well as, a length effect term (Table 4). The predicted proportion caught at length by area is shown in Figure 20. There was an increase in the relative efficiency for the 10-minute tow as length increased for CAII and NLCA. For the MAB, the relative efficiency was higher for the 10-minute across all length classes (Figure 15). Results from the Miller approach indicated a binomial model with an intercept and smoother of size for across pair effects and for the random effects fit the data the best for all areas (Figure 16). The predicted proportion caught at length graphs showed a similar trend for the relative efficiency of the 10-minute tow compared to the GLMM results for each area.

Dredge Efficiency

A total of 30 dredge AUV pairs were completed in the study site (Figure 17). Of those 30 pair, one pair was excluded due to image quality issues and another pair was excluded due to survey dredge catch data issues. This resulted in 28 pairs available for analysis. Twenty AUV stations were annotated by trained human annotators and all 28 stations were annotated using YOLOv3. Over 383,000 AUV images were collected during the study. Approximately 480,000 m^2 was covered by digital images and 2,555,000 m^2 was covered by side-scan sonar. The team of 24 trained annotators counted and measured 330,419 scallops in 31,089 images across the 20 missions (Table 5). One of these missions was the excluded survey dredge station. Of the 330,419 scallops counted, 298,201 were measured for shell height (90 percent). YOLOv3 annotated 294,768 images and detected a total of 6,333,478 scallops, with a mean detection confidence of 0.63. Its image processing speed was 12.4 images per second, a processing rate approximately one thousand times faster than what the annotation team could accomplish and approximately four times faster than the image acquisition rate on the AUV (Table 5).

Annotator Evaluation

Mean scallop count across annotators was 387 +/- 53 scallops (SD), or approximately 14 percent (Figure 18). Mean shell height across annotators was 95.6 +/- 4.9 mm (SD). The two factor ANOVA showed mean scallop density varied significantly between annotators (p-value <0.001). The Type II ANOVA indicated mean shell height varied significantly between annotators (p-value <0.001). Repeated measurements by a single annotator varied on average by 4.5 mm (SD) or 2.1 pixels on the screen (Figure 19). Pooled across all annotators, shell height varied on average by 7.4 mm (SD) or 3.4 pixels. The one factor ANOVAs or Kruskal-Wallis rank sum tests showed that annotator had a significant effect on shell height (p-values <0.001). Variances of shell height measurements on a scallop were not homogeneous when the scallop contrasted less sharply with the seafloor (i.e. scallop 5), or when the line of symmetry was difficult to discern (i.e. scallop 2) making shell height more difficult to precisely measure (Figure 19).

Convolutional Neural Network

YOLOv3 was tested on a reserved test set of 19,469 images that contained 72,879 scallops, and results showed an average precision (AP) value of 0.924 (Figure 20). Output images from the test set demonstrated that when scallops were unobscured and the image was annotated accurately, the neural network and annotators agreed on the number of scallops in the image. Scallopscan was challenged by images with extremely high scallop density (e.g., 100 scallops per image, or 37 scallops per m^2) where scallops were crowded, located on the perimeter of the image, or covered in a thin layer of sediment, reducing the contrast between the scallop and

the seafloor. Conversely, in some instances the neural network detected scallops that were missed due to annotator error (Figure 21).

Annotator YOLOv3 Comparison

There was no significant difference between relative length distributions for the human and YOLOv3 annotated data across the 20 stations that were annotated by each group (p-value = 0.39) (Figure 22). The YOLOv3 length distribution has a slight bimodal distribution that is observed at 67.5 mm (Figure 22). This distribution is evident at several of the stations (Figure 23). There also appears to be a knife edge increase in the number of scallops measured at 52.5 mm for the YOLOv3 annotated data (Figure 27). The human annotated length data indicates a greater number of both smaller (< 52.5 mm) and larger (> 100 mm) scallops were measured. The mean shell heights between the two groups was also similar, although at the larger shell heights YOLOv3 annotated lengths were greater than human annotated measurements (Figure 24). The ANOVA indicated no significant difference in mean density estimates between the two annotation methods (p-value = 0.67). At the station level, density estimates between the two methods was also comparable (Figure 25). At the highest densities (> 30 scallops per m^2), YOLOv3 density estimates were lower than the human annotated estimates

Survey Dredge Annotation Comparison

There was no significant difference between the different data sets relative length frequency distributions pooled across all pairs (Table 6) (Figure 26). The same bimodal pattern is present for the YOLOv3 data. There is also a similar trend of the YOLOv3 data not measuring as many small or large scallops as were measured in the survey dredge data. The human annotated length frequency distribution is similar to the survey dredge distribution, especially at the smaller and larger size classes. Even though there was no significant difference in length distributions, the difference between the YOLOv3 distribution and the other two distributions may indicate a selectivity issue with the YOLOv3 data that may need to be corrected for.

The selectivity curve for the YOLOv3 data set indicated the assumption of 100 percent selectivity was not met (Figure 27). The probability of a scallop being detected by YOLOv3 increases with length and 100 percent detection does not occur until 55 mm. This results confirms the issue raised by examining the length distributions. The human annotated data set showed the selectivity curve was equal to 100 percent across all length classes (Figure 28).

Based on the results from the data set comparisons, the final AUV data (n = 28) used to estimate dredge efficiency was a combination of human annotated and YOLOv3 data. The human annotated data included 19 stations and the remaining 9 stations

were YOLOv3 data. The human annotated data set included all high density stations. The remaining YOLOv3 data were for lower density stations.

Survey Dredge Efficiency Analysis

Dredge efficiency analysis indicated station-level efficiency was variable (Figure 29). The majority of stations (n = 20) had mean density estimates less than four scallop per m^2. The other eight stations had higher mean densities, ranging from 5.37 to 32.66 scallops per m^2. At lower densities, efficiency tended to be lower than the assumed value of 0.4, although there were 3 stations where efficiency was greater than the assumed value. Efficiency declined at approximately two scallops per m^2, as indicated by the GAM smoother and station-level efficiency values (Figure 29). Efficiency remained consistently low across the range of higher density values, and the lowest predicted GAM efficiency value was 0.135. The average efficiency value for densities greater than two scallops per m^2 was 0.12, which is consistent with the lowest predicted GAM value. The estimated values from this study are also comparable to the 0.10 efficiency value used in the 2018 assessment to scale dredge biomass estimates to account for efficiency issues (NEFSC, 2018).

Survey Dredge Biomass Estimation

Survey dredge density estimates in the NLCA survey domain ranged from 0 to 4.47 scallops per m^2 for the 2018 NLCA survey. Out of the 130 stations completed in 2018, only 9 stations had densities greater than 2 scallops per m^2. Densities greater than 2 scallops per m^2 were observed in the NL South Deep SAMS area (n = 7), NL South Shallow (n = 1) and NL NA (n =1) (Figure 30). Absolute biomass estimates were variable depending on treatment (Figure 31). Using the 0.40 method, survey dredge estimates were significantly lower than the optical method in the SAMS areas of concern (i.e., NL NA and NL South Deep). Estimates were similar to the optical estimate for the remaining three SAMS areas. With the 0.13 method, SAMS area dredge biomass estimates were comparable to the optical estimates in the NL South Deep, NL South Shallow and NL Ext SAMS areas. Biomass was severally overestimated with this approach in the NL North SAMS area and slightly lower than the optical estimate in the NL NA SAMS area. There was a similar trend when applying the 0.10 method compared to the 0.13 method. Dredge biomass was over estimated compared to the optical estimates in the NL North, NL South Deep and NL South Shallow SAMS areas. For the other two SAMS areas, dredge biomass estimates were similar to the optical estimates. Dredge and optical biomass estimates were similar for the NL North and the NL Ext SAMS areas when applying an efficiency correction with the SAMS method. This method underestimated biomass in the NL South Shallow and NL NA SAMS areas, while slightly over estimating biomass in the NL South Deep SAMS area. The Station method performed better than the 0.40 method, but worse than the other approaches for the NL NA SAMS. In the NL South Deep SAMS area,

this method slightly underestimated biomass compared to the optical estimate. The dredge estimate in the NL NA SAMS area being so low was a result of only one station this in area having a density great than 2 scallop per m^2. This method yielded similar biomass estimated compared to the optical estimates for the other three SAMS areas.

In the MAB survey domain, densities were generally lower for the VIMS 2018 survey compared to the NL survey. Density estimates ranged from 0 - 2.10 scallops per m^2 and only one station had a density estimate greater than the threshold of 2 scallops per m^2 (Figure 32). For the one SAMS area of concern in this survey, ET Flex, the 0.40 method and the Station method performed the best when comparing biomass estimates to the optical biomass estimate (Figure 33). The SAMS, 0.13 and 0.10 methods greatly overestimated biomass in this SAMS area. Biomass estimates for the dredge survey in 2018 were lower than the optical estimate, but the difference between the two estimates was not as pronounced as has been in previous years. For the other eight SAMS areas, the 0.13 and 0.10 methods overestimated biomass in seven of the SAMS areas, while the SAMS and Station methods provided biomass estimates comparable to the optical estimates.

Discussion

The tow duration experiment did not provide conclusive results regarding the impact of a reduced tow time on scallop catch rates. While catch rates of scallops in CAII were reduced in the 10-minuite tow compared to the standard 15-minute tow, the MAB and NLCA results were confounding and did not follow expectations. These are the two areas of current concern regarding survey dredge performance and catchability assumptions. It was also difficult to determine if and when dredge saturation was occurring. This is important in the context of the potential for reduced dredge efficiency at high densities. Dredge saturation may be occurring in discrete areas with extreme densities of scallops in the MAB and NLCA. A 10-minute tow duration may be not short enough in these high density areas to address dredge performance issues. The Scallop Survey Peer Review Panel had recommended testing several different tow duration lengths including 10 and 5 minute durations (SSSMPRT, 2015).

While the tow duration study was not conclusive, data from this portion of the project will be helpful to inform future tow duration discussions. There are other areas of research on dredge saturation and performance that would be helpful for future work on this topic. Placing cameras or video equipment on the survey dredge may allow for an optical assessment of dredge performance or filling. Using optical approaches in conjunction with data routinely collected for the survey may aid in narrowing in on an optimal tow time in high density areas. One potential method for determining dredge saturation would be to examine data collected from the StarOddi inclinometer placed on the survey dredge to determine if there is a threshold dredge angle that indicates dredge filling. A similar approach was taken by the NEFSC looking at warp tension

from the R/V *Sharp* and results from that analysis were presented during the 2018 benchmark assessment. Unfortunately, these results were inclusive. Pairing dredge angle information with video footage of dredge saturation would be beneficial. VIMS will continue to investigate dredge saturation during an upcoming 2019 sea scallop RSA project where cameras will be placed on the survey dredge to address this issue. There are also plans to analyze dredge angle data during the same project.

Another approach that will be taken in the near future is increasing the sample size for the paired tow duration project. The NEFSC completed paired 15 and 10-minute tows in the spring of 2018 in the MAB onboard the R/V *Sharp*. VIMS also completed 15-minute tows occupying the same areas, so that there is a three-way comparison for a 10 vs 15-minute tow duration. Increasing the sample size and including the newer data set in the analysis may provide more insight for the tow duration study.

The comparison for human annotated and YOLOv3 annotated data collected from the AUV provided evidence that these data sets are similar and that an automated annotation program can detect scallops. There were no significant differences in length frequency distributions or mean density estimates at the station-level. With that said, the YOLOv3 data tended to underestimate density as density increased and did not detect small scallops, as evidenced by both the length frequency distribution and selectivity analysis. The approach, however, is quite promising and would benefit from additional training sets focused on small scallops ranging in length range from 40 to 60 mm, as well as, continuing to train the algorithm on high density scallop areas with varying substrate types and the degree of sediment covering the scallops. While not the initial approach, the decision to use YOLOv3 annotated data for the efficiency analysis was justified based on the data set comparisons. YOLOv3 data were used for low density stations that did not have a significant impact on understanding how efficiency declines with high density. Including this data set also helped to increase the overall sample size for the study. The human annotated data set provided the majority of data for the efficiency analysis, and we feel that this data set accurately measured and quantified the number of scallops in the area of the survey dredge.

Dredge efficiency was estimated over a range of scallop densities. This analysis indicated reduced efficiency at densities greater than two scallops per m^2, and this result is consistent with preliminary analysis conducted by the NEFSC (2018). The lowest GAM efficiency value of 0.135 is similar to the 0.10 value used during the 2018 benchmark assessment to adjust dredge biomass estimates (NEFSC, 2018). Efficiency throughout the study area tended to be lower than the assumed value of 0.40, with the exception of three stations at the lower range of observed scallop densities. Also, once efficiency was reduced, it remained relatively consistent across the range of observed higher densities. This result may be beneficial in guiding efficiency adjustment

discussions in the future. Efficiency adjustments could be done based on a density threshold and there would not be a need for scaling efficiency value as a function of density. The study estimated efficiency based on 28 dredge AUV pairs, which is a modest sample size. The study is lacking samples in the mid density range (2 – 10 scallops per m^2) which limits the inference that can be made from this study alone. This data set will be added to the NEFSC data set collected from 2016 - 2018 for future analysis to provide a more robust updated efficiency estimate. The NEFSC has plans to use the Miller et al. (*in press*) approach to provide updated efficiency estimates. Adding this data set to the NEFSC data will increase the spatial coverage of paired tows for analysis.

Survey dredge biomass estimates were sensitive to efficiency values and how those values were applied to scale relative biomass to absolute biomass. This effect was more apparent in the NLCA survey area compared to the MAB survey area. This may be due to lower variability in scallop densities in the MAB survey area and with only one SAMS area in this survey domain where divergent biomass estimates have been observed in the past. In the NLCA survey domain, the best performing efficiency value varied between the two SAMS areas of current concern (i.e., NL South Deep and NL NA) with respect to the optical biomass estimate. This may indicate that applying updated efficiency values at different scales may be appropriate in this survey area to account for varying resource conditions and survey dredge performance. It may also not be suitable to use a station-level efficiency value based on survey dredge density estimates, since the density estimates from the survey dredge in certain areas are artificially low due to reduced dredge performance.

The project budget and program income is provided in Appendix A.

Engagement

Twenty-four undergraduates or recent graduates contributed to the image annotation team: Anna Abelman, Sarah Bajohr, Michelle Baptist, Emily Beale, Kristin Brubaker, Joseph Coffin, Alexander Douwes, Matthew Dunn, Samantha Dypko, Shailja Gangrade, Andrea Lock, Josette Messere, Conner McCrone, Erin Papke, Jennifer Peasnall, Alexa Perez-Krizan, Richard Rosas, Caitlin Stockwell, Molly Struble, Jack Sypher, Alexander Thomas, Sara Thomas, Jacqueline Valladares, and Cassandra Wilson. Recent graduate Peter Barron supported field operations and processed the side-scan sonar data collected during this project. Graduate student Jiayi Zhao contributed to the development of the scallop detector, as well as, the testing of other deep learning strategies. Graduate student Hunter Tipton contributed to project activities through field support, data processing and evaluation of the scallop detector.

Presentations

The following presentations were given regarding this project:

- C. Rasmussen, J. Zhao, D. Ferraro and A. Trembanis. 2017. Deep census: AUV-based scallop population monitoring. 2017 IEEE International Conference on Computer Vision Workshops, October 22 – 29, 2017, Venice, Italy: 2865-2873.
- S. Roman and D. Rudders. Effect of Tow Duration on Scallop Catch for the VIMS Scallop Dredge Survey. 2018 Benchmark Sea Scallop Assessment Data Meeting, February 5 – 9, 2018, Woods Hole, MA.
- S. Roman and D. Rudders. Updated Tow Duration Analysis. 2018 Benchmark Sea Scallop Assessment Data Meeting, March 26 – 29, 2018, Woods Hole, MA.
- D. Rudders, A. Trembanis, S. Roman, D. Ferraro and H. Tipton. Understanding density dependent effects on catchability and dredge performance for a sea scallop dredge survey. 2018 American Fisheries Society Annual Conference, August 19 – 23, 2018, Atlantic City, NJ.
- D. Ferraro, A. Trembanis, D. Rudders and D. Miller. 2018. Applications of autonomous underwater vehicle seabed imaging in fishery-independent sea scallop surveys. 2018 American Fisheries Society Annual Conference, August 19 – 23, 2018, Atlantic City, NJ.
- D. Rudders, A. Trembanis, S. Roman, D. Ferraro and H. Tipton. 2018. Understanding density dependent effects on catchability and dredge performance for a sea scallop dredge survey. 2018 ICES Annual Conference, September 24 – 27, 2018, Hamburg, Germany.
- D. Ferraro, A. Trembanis, C. Rasmussen, J. Zhao and N. Wilkinson. 2018. From deep learning to citizen science: Developing and implementing strategies for analyzing large imagery data sets. 2018 Ocean Sciences Meeting, October 11-18, 2018, Portland, OR.
- S. Roman, D. Rudders, A. Trembanis and D. Ferraro. 2019. Impact of Catchability Assumptions on Sea Scallop Survey Biomass Estimates. 2019 Pectinid Workshop, April 23-29, 2019. Santiago de Compostela, Spain.
- H. Tipton, A. Trembanis, C. Rasmussen and D. Ferraro. 2019. Assessing the performance of deep learning strategies in sea scallop (*Placopecten magellanicus*) survey imagery analysis. 2019 Pectinid Workshop, April 23-29, 2019. Santiago de Compostela, Spain.

References

Bivand, R. S., E. J. Pebesma, and V. Gomez-Rubio. 2013. Applied spatial data analysis with R, Second edition. Springer, NY. http://www.asdar-book.org

Cochran, W. G. 1977. Sampling Techniques (3rd ed.). John Wiley and Sons, New York. 428 pp.

DuPaul, W. D. and J. E. Kirkley. 1995. Evaluation of sea scallop dredge ring size. Contract report submitted to NOAA, National Marine Fisheries Service. Grant # NA36FD0131.

Ferraro, D. M., A. C. Trembanis, D. C. Miller and D. B. Rudders. 2017. Estimates of sea scallop (Placopecten magellanicus) incidental mortality from photographic multiple before-after-control-impact surveys. Journal of Shellfish Research 36: 615-626.

Fox, J., and S. Weisberg. 2011. An {R} companion to applied regression, second edition. Thousand Oaks, CA: Sage. Available at: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Holst, R. and A. Revill. 2009. A simple statistical method for catch comparison studies. Fisheries Research. 95: 254-259.

Kruskall, W. H. and W. A. Wallis. 1952 Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 47(260): 583-621.

Millar, R. B. 1992. Estimating the size-selectivity of fishing gear by conditioning on the total catch. Journal of the American Statistical Association. **87**: 962-968.

Millar, R.B. and R.J. Fryer. 1999. Estimating the size-selection curves of towed gears, traps, nets and hooks. Reviews Fish. Bio. Fish. **9**:89-116.

Miller, T. J. 2013. A comparison of hierarchical models for relative catch efficiency based on paired-gear data for US Northwest Atlantic fish stocks. Canadian Journal of Fisheries and Aquatic Sciences 70: 1306-1316.

Miller, T. J., D. R. Hart, K. Hopkins, N. H. Vine, R. Taylor. A. D. York and S. M. Gallager. Estimation of the capture efficiency and abundance of Atlantic sea scallops (Placopecten magellanicus) from paired photographic-dredge tows using hierarchical models. Canadian Journal of Fisheries and Aquatic Sciences. In press.

Nelson, G. A. 2018. fishmethods: Fishery Science Methods and Models. R package version 1.11-0. https://CRAN.R-project.org/package=fishmethods.

National Marine Fisheries Service (NMFS). 2018. Fisheries of the United States, 2017. U.S. Department of Commerce, NOAA Current Fishery Statistics No. 2017. Available at: https://www.fisheries.noaa.gov/feature-story/fisheries -united-states-2017.

Northeast Fisheries Science Center (NEFSC). 2018. 65th Northeast Regional Stock Assessment Workshop (65th SAW) Assessment Report. US Dept Commer, Northeast Fish Sci Cent Ref Doc. 18-11; 659 p.

Pebesma, E. J., and R. S. Bivand. 2005. Classes and methods for spatial data in R. R News 5 (2), https://cran.r-project.org/doc/Rnews/.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Redmon, J., and A. Farhadi. 2017. YOLOv39000: Better, faster, stronger. In: Proceedings IEEE Computer Vision and Pattern Recognition. pp. 7263-7271.

Redmon J., and A. Farhadi. 2018. YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

Rudders, David B. and Sally A. Roman. 2018. Final Report: A cooperative high precision dredge survey to assess the mid-Atlantic sea scallop resource. VIMS Marine Resource Report No. 2018-05. Submitted to the National Marine Fisheries Service. 57 pp.

Sea Scallop Survey Methodologies Peer Review Team (SSSMPRT). 2015. Summary Report of the Review of Sea Scallop Survey Methodologies and Their Integration for Stock Assessment and Fishery Management. Meeting to review Sea Scallop Survey Methodologies and Their Integration for Stock Assessment and Fishery Management 17-19 March 2015. Waypoint Event Center at the Marriott Fairfield Inn and Suites, New Bedford, MA. (http://s3.amazonaws.com/nefmc.org/Scallop-surveys-review-Summary-report_April-91.pdf

Shumway, S. E. and G. J. Parsons. 2006. (Editors). Scallops: biology, ecology and aquaculture. Elsevier, Boston. 1460 pp.

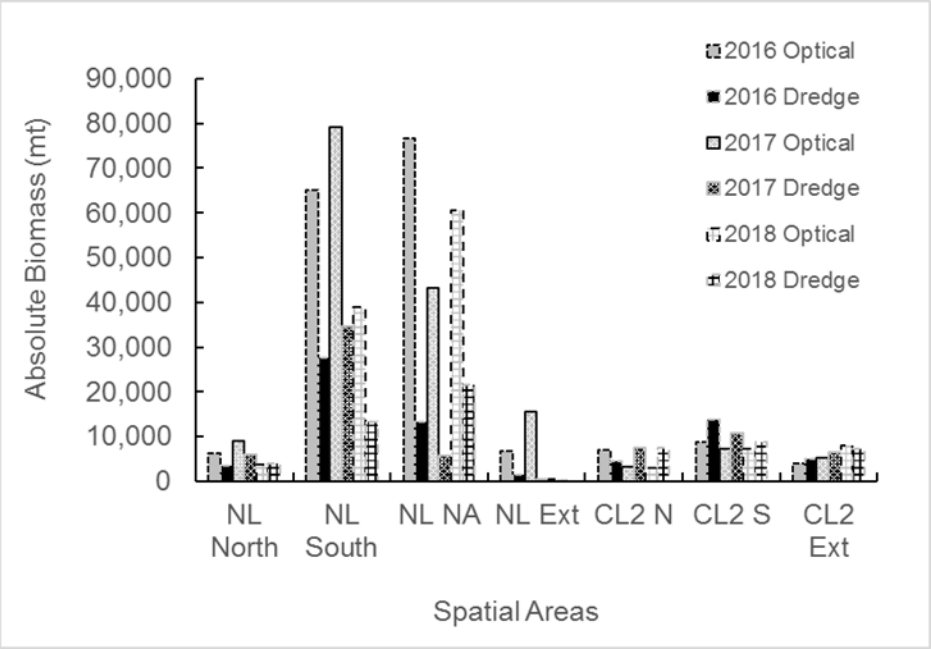
Sokal, R. R. and F. J. Rohlf. 1995. Biometry (3rd edition). W. H. Freeman and Company, New York. 887 pp.

Trembanis, A., D. Miller, D. Rudders and D. Ferraro. 2017. Final Report: Incidental mortality estimates of sea scallops from AUV based BACI surveys. URL https://www.nefsc.noaa.gov/coopresearch/pdfs/FR-14-0073.pdf.

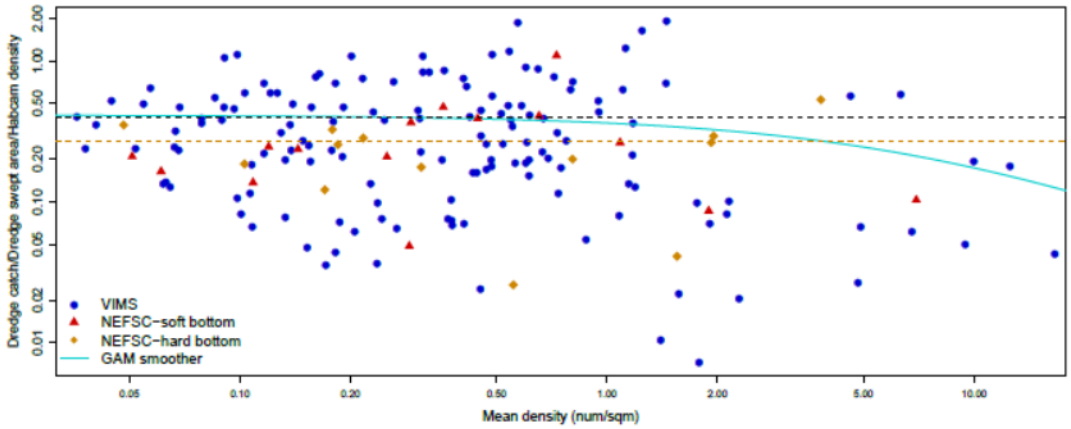
Weinhaus, F. 2007. Retinex. Available at: http://www.fmwconcepts. com/imagemagick/retinex/.

Wood, S.N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society (B) 73(1):3-36.

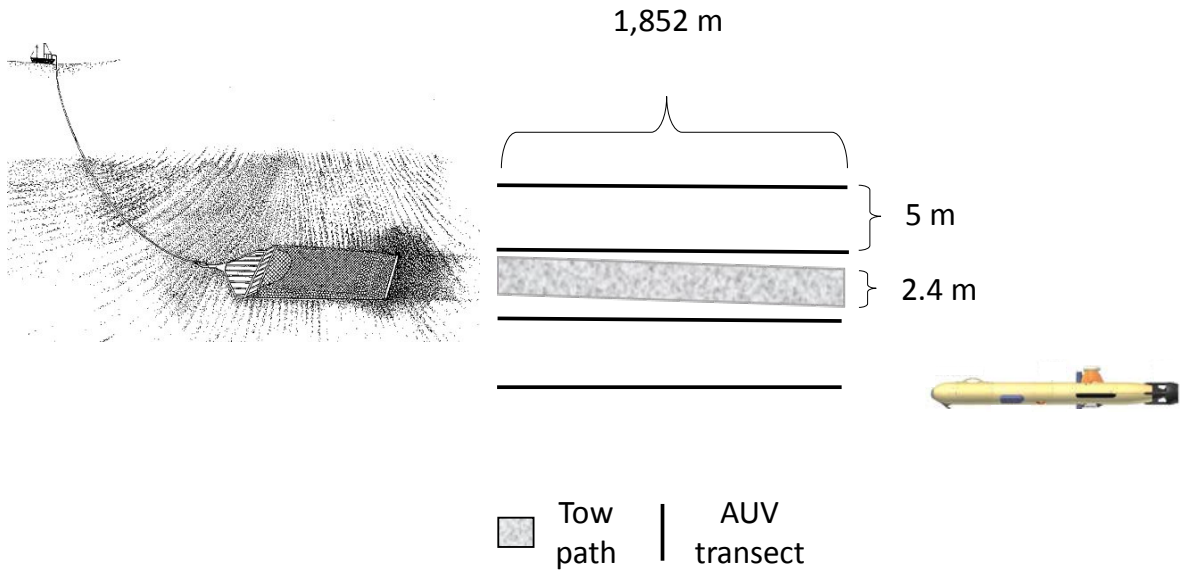
Wood, S. N. 2017. Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC.



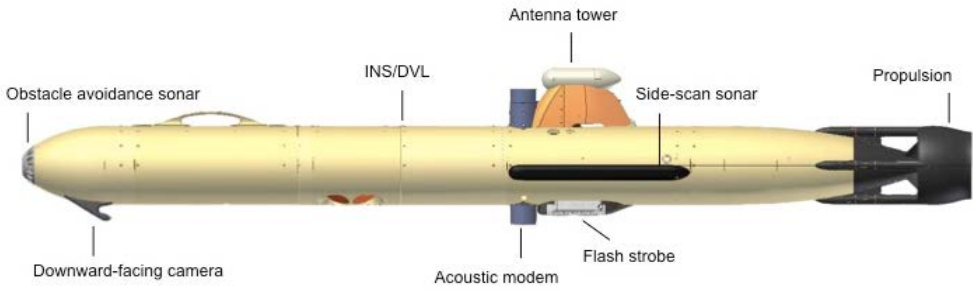
- Figure 1. Absolute biomass estimates (mt) for Georges Bank SAMS areas for the NEFSC Habcam survey and dredge survey (NEFSC and VIMS) for 2016 - 2018.



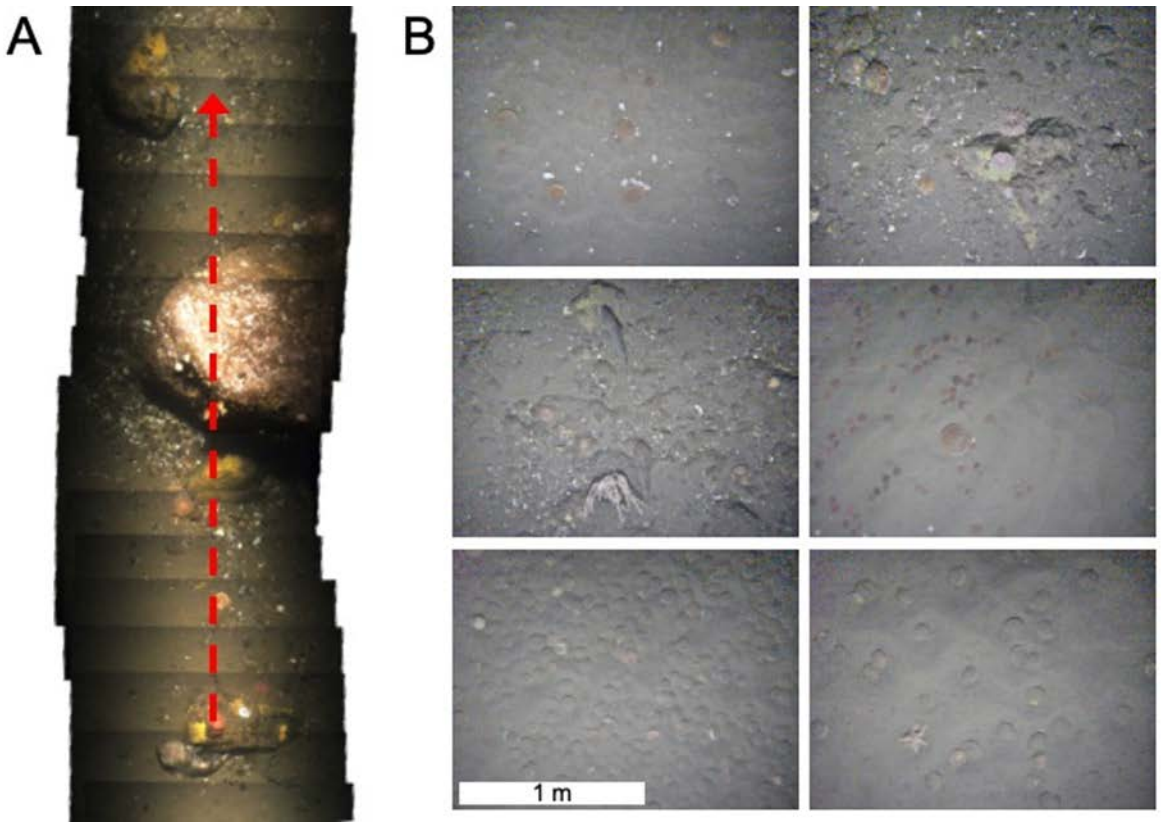
- Figure 2. Dredge to Habcam density ratio plotted against mean density (scallops/m2) for 2016 - 2017 taken from the 2018 benchmark assessment. Dredge data are from VIMS and the NEFSC. Habcam data are from the NEFSC. The solid blue line is a generalized additive model fit, the black dashed line is the assumed dredge efficiency value of 0.4 for soft substrate and the yellow dashed line is the assumed dredge efficiency value of 0.27 for hard substrate (NEFSC, 2018).



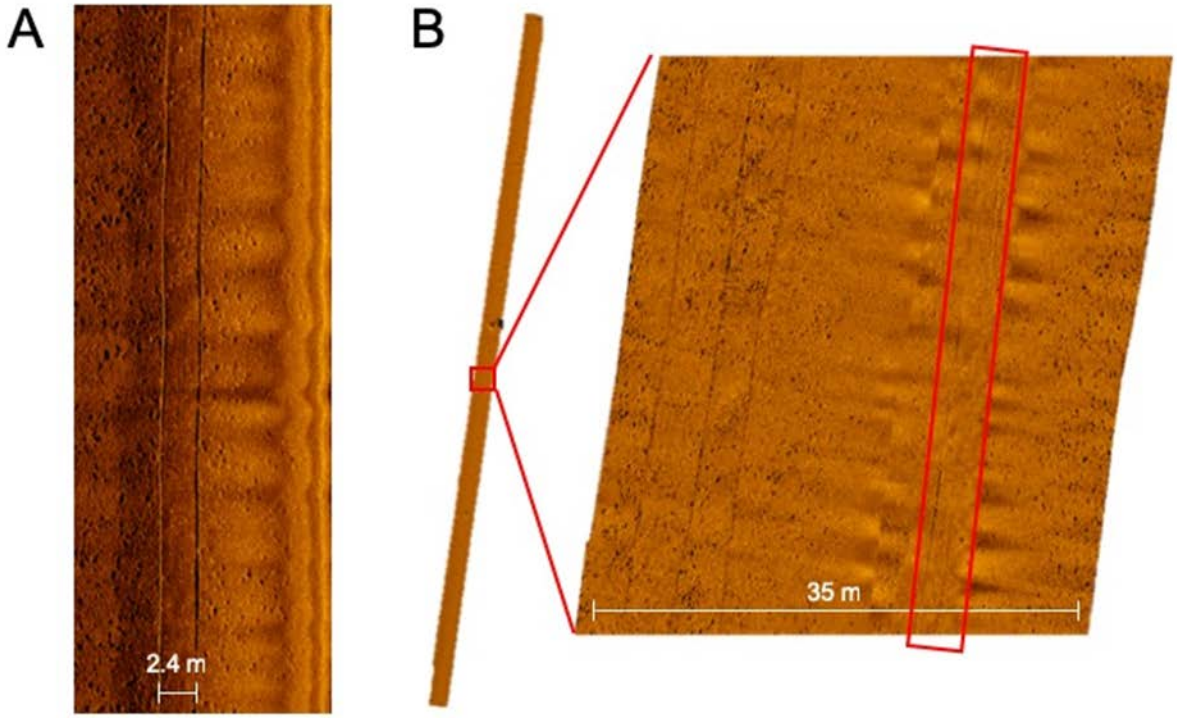
- Figure 3. Schematic of a paired survey dredge tow/AUV mission for a selected site, with a dredge tow, number and length of AUV transects and spacing between AUV transects. 2.4 m is the width of the survey dredge frame. Credit for image of vessel and dredge https://njscuba.net/artifacts/obj_dredge-trap.php



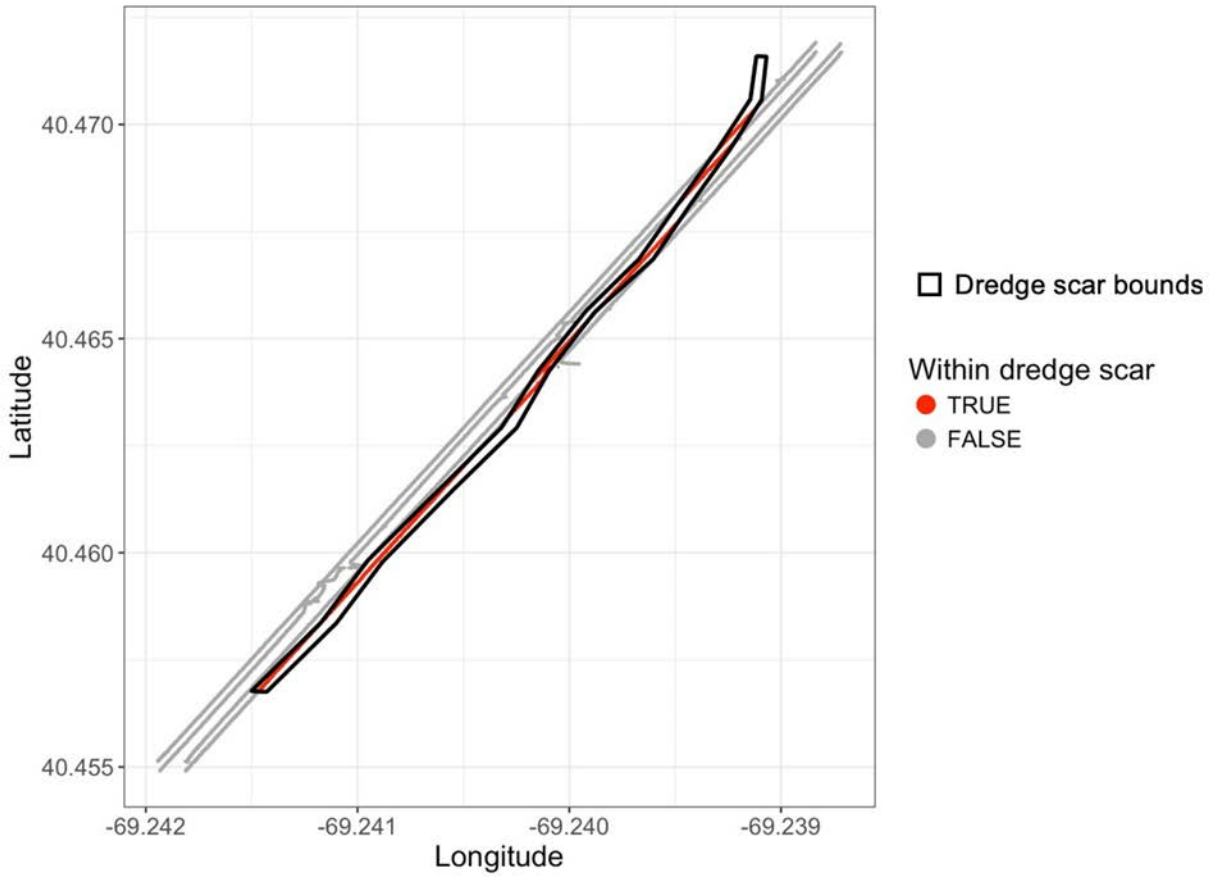
- Figure 4. University of Delaware GAVIA AUV as configured for this project.



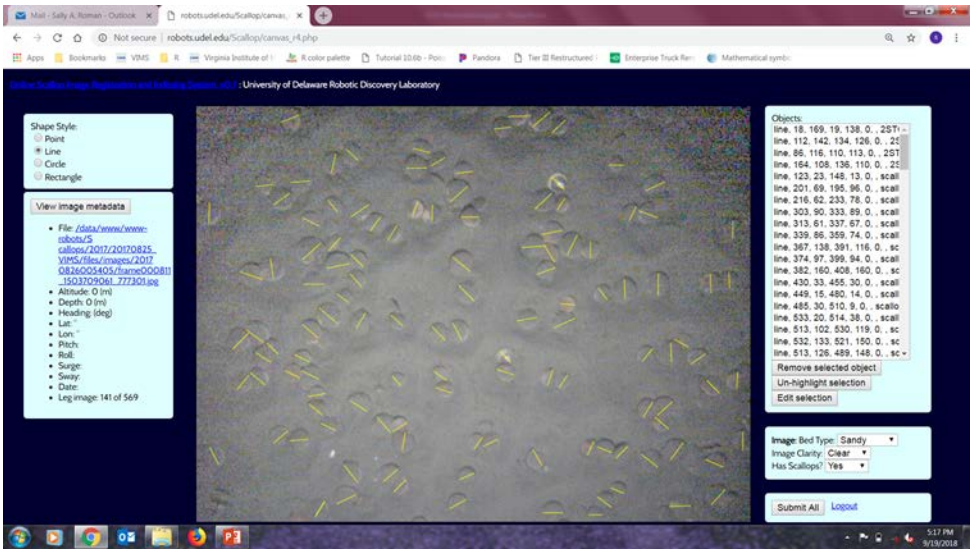
- Figure 5. (A) A sample “filmstrip” of raw images with the direction of AUV travel indicated with the red arrow. (B) Sample enhanced images collected in the Nantucket Lightship Closed Area during this study, displaying the range of scallop densities and substrate compositions encountered.



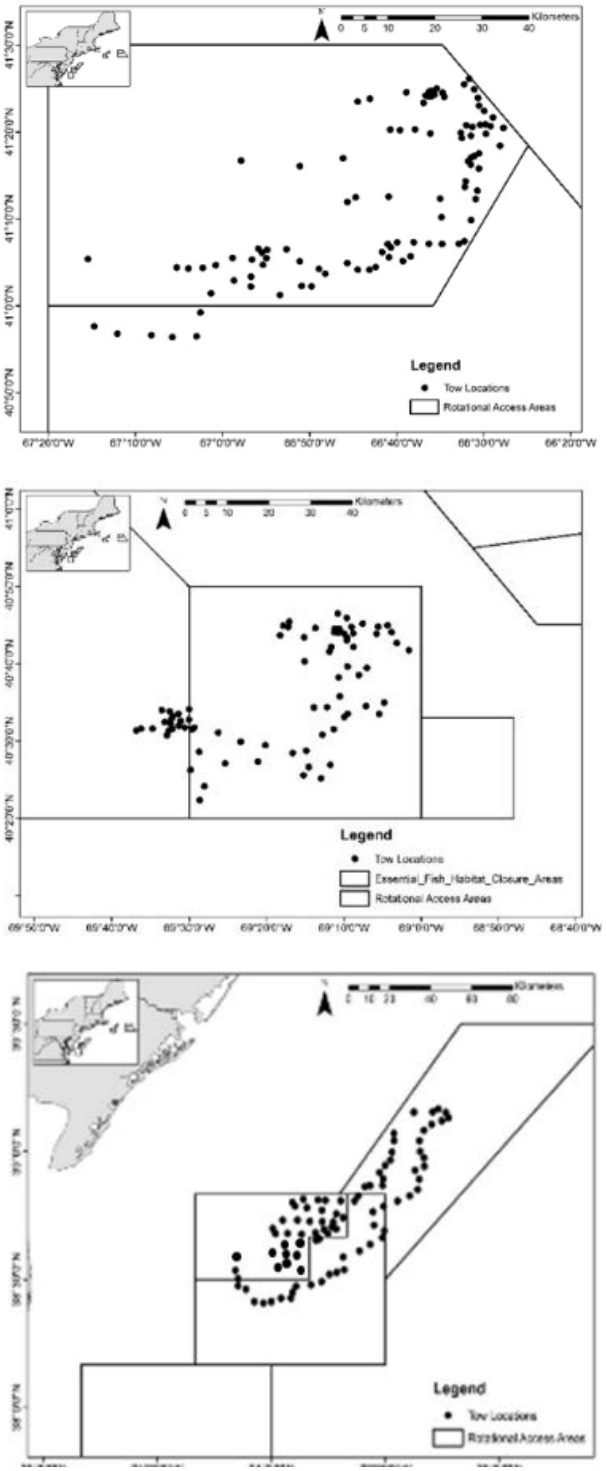
- Figure 6. (A) Example of a scar left behind by the 2.4 m (8 ft) survey dredge visible in a single raw side-scan sonar file. (B) The complete side-scan sonar mosaic from a mission with a section selected and enlarged to depict the dredge scar, highlighted with the red box.



- Figure 7. Sample map of an AUV mission depicting centroid points of the images taken during the mission. The bounds of the dredge scar with a 1 m buffer added are overlaid in black. Red points are the images that fell within the dredge scar bounds, and gray points are the images that fell outside the dredge scar bounds. Only the images outside the scar bounds were used to derive density estimates.



- Figure 8. Screen shot of the University of Delaware's custom scallop image annotation software and graphical user interface. Length measurements from scallops were recorded using the software (yellow lines)



- Figure 9. Location of all tow duration pairs by area. Top: Closed Area II, Middle: Nantucket Lightship, Bottom: Mid-Atlantic.

Figure 10. Photographs taken of paired tows in CAII (Closed Area II) (top) and MAB (Mid-Atlantic) (bottom) for a 10 and 15-minute tow.

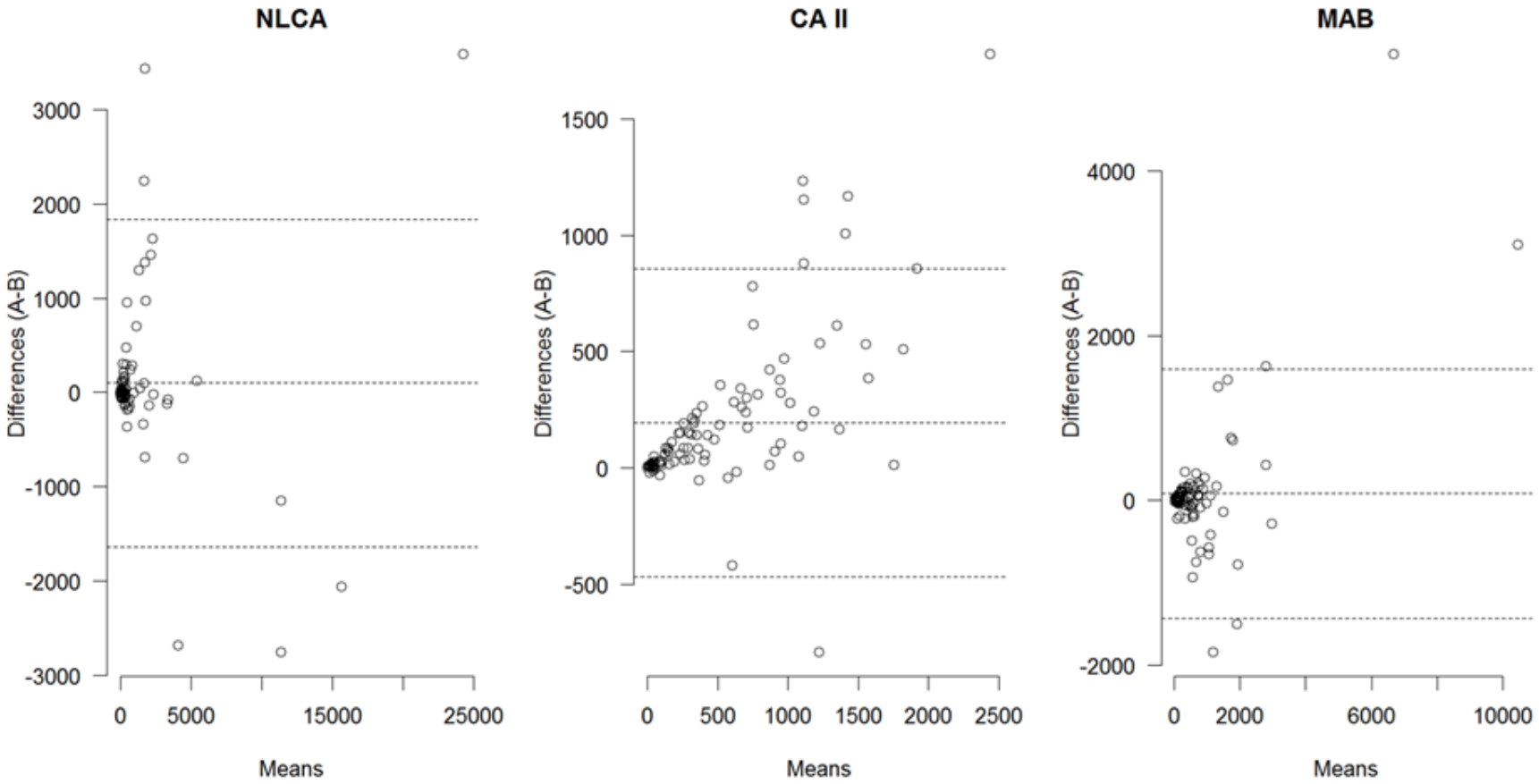


Figure 11. Bland-Altman plots by area for the expanded number of scallops. A is the 15-minute tow and B is the 10-minute tow. The x axis is the mean of the paired catch (A+B/2). The y axis is the difference between the paired catch (A-B). The middle dashed line is the mean of the difference and the upper and lower dashed lines are 95% confidence intervals. Areas: NLCA is Nantucket Lightship, CA II is Closed Area II and MAB is Mid-Atlantic.

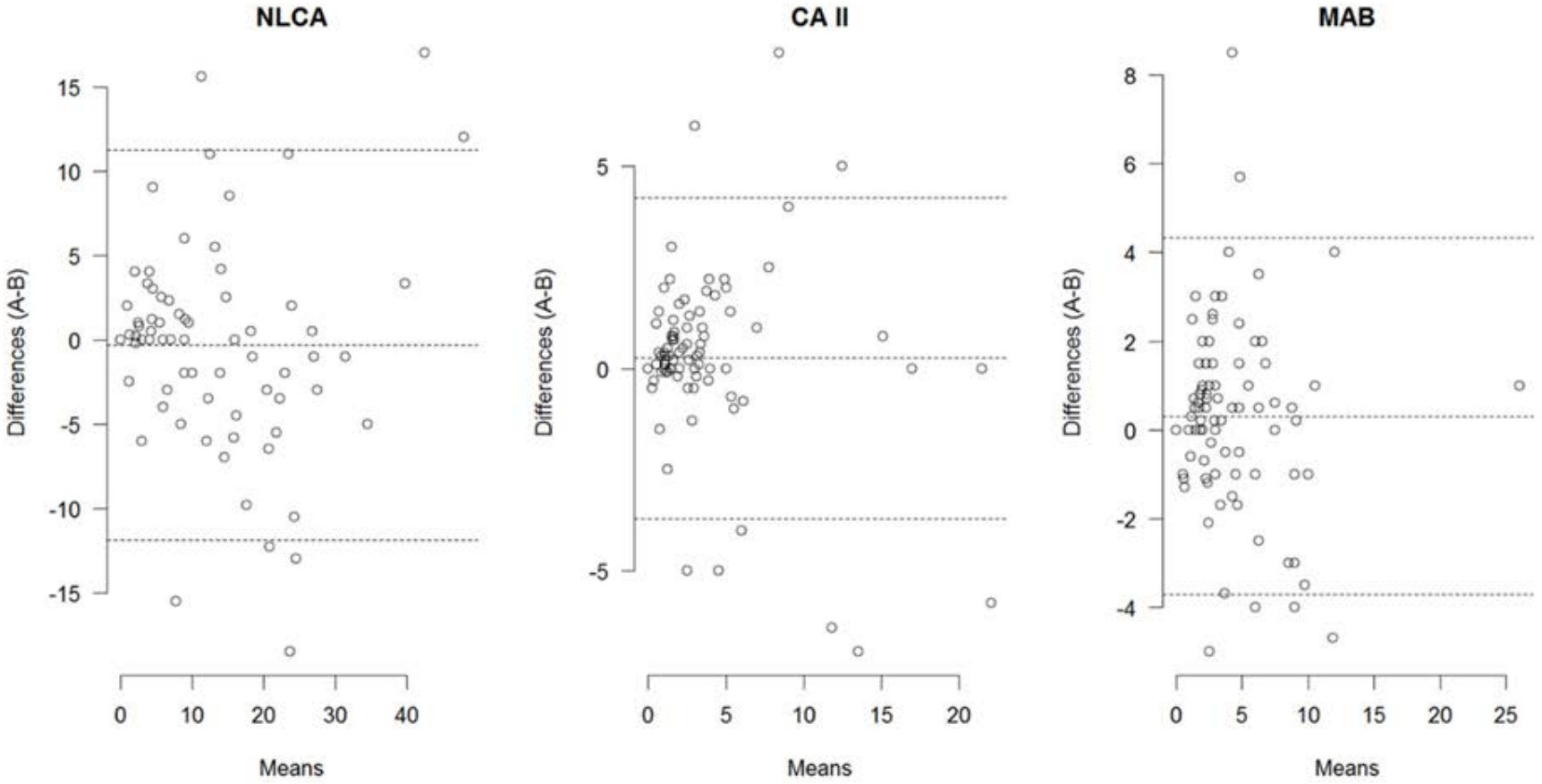


Figure 12. Bland-Altman plots by area for debris catch (baskets). A is the 15-minute tow and B is the 10-minute tow. The x axis is the mean of the paired catch (A+B/2). The y axis is the difference between the paired catch (A-B). The middle dashed line is the mean of the difference and the upper and lower dashed lines are 95% confidence intervals. Areas: NLCA is Nantucket Lightship, CA II is Closed Area II and MAB is Mid-Atlantic.

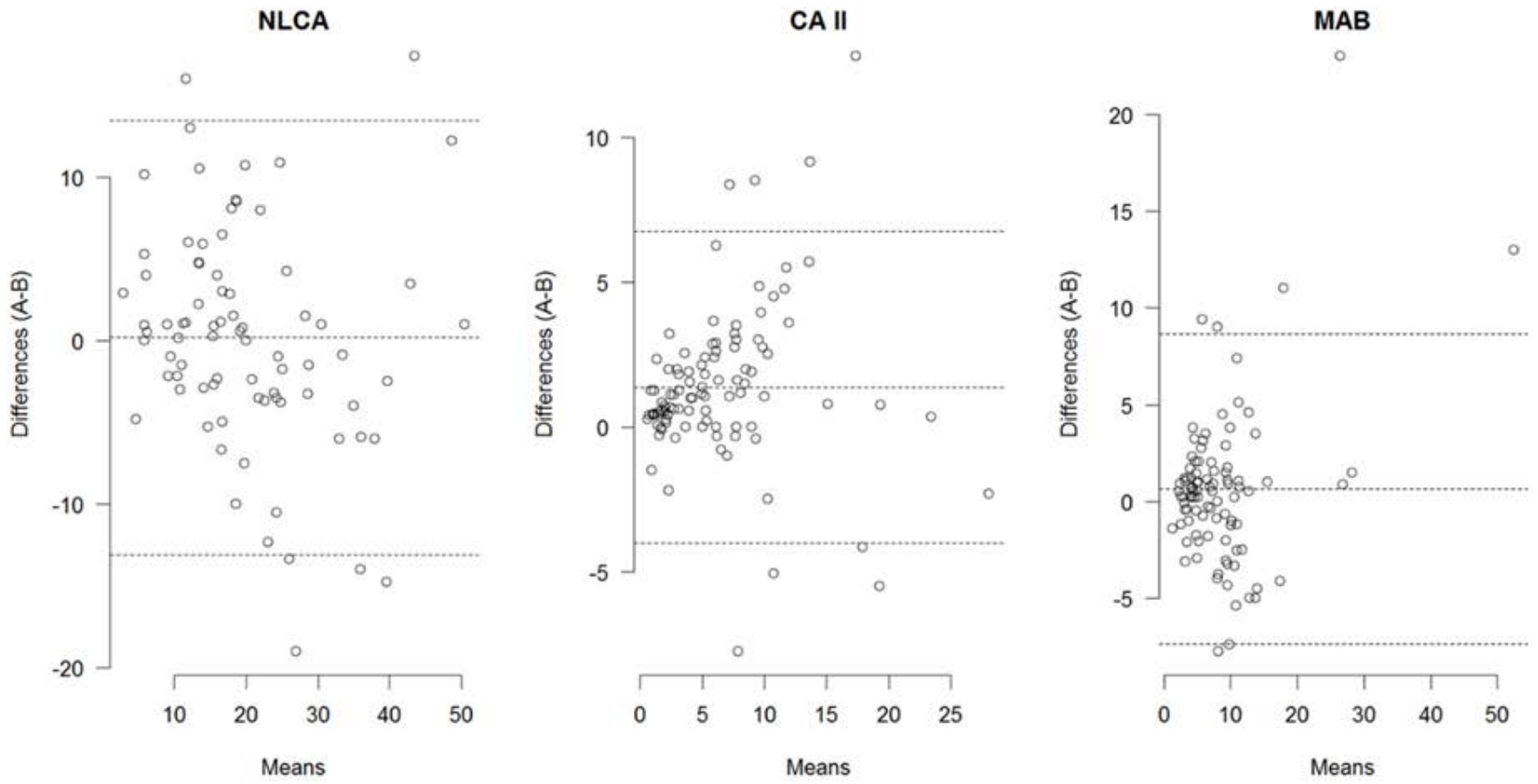


Figure 13. Bland-Altman plots by area for total catch (number of baskets of scallop catch + number of baskets of debris catch). A is the 15-minute tow and B is the 10-minute tow. The x axis is the mean of the paired catch (A+B/2). The y axis is the difference between the paired catch (A-B). The middle dashed line is the mean of the difference and the upper and lower dashed lines are 95% confidence intervals. Areas: NLCA is Nantucket Lightship, CA II is Closed Area II and MAB is Mid-Atlantic.

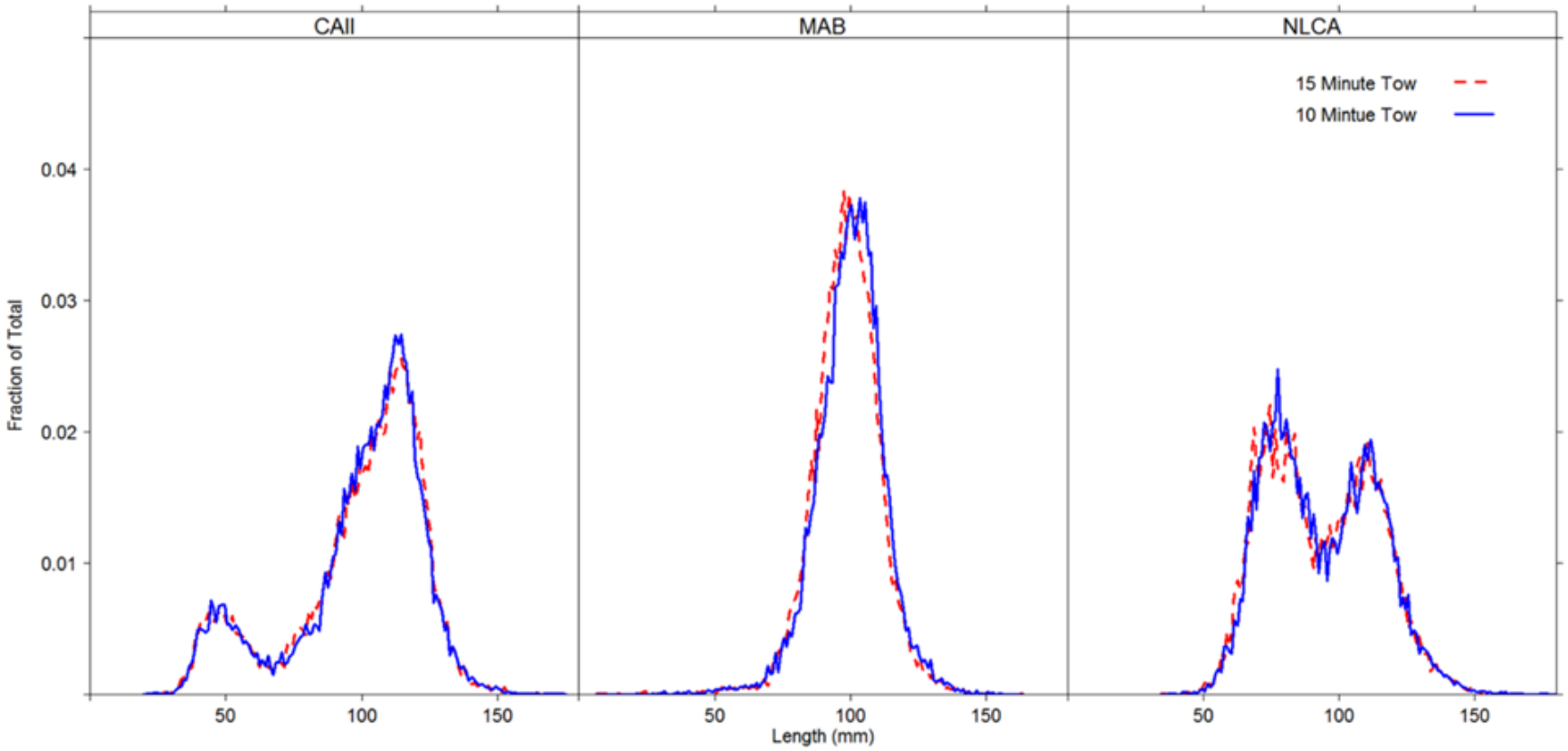


Figure 14. Relative length frequency distributions by area for the 10-minute tow (blue line) and the 15-minute tow (red dashed line). Areas: NLCA is Nantucket Lightship, CAII is Closed Area II and MAB is Mid-Atlantic.

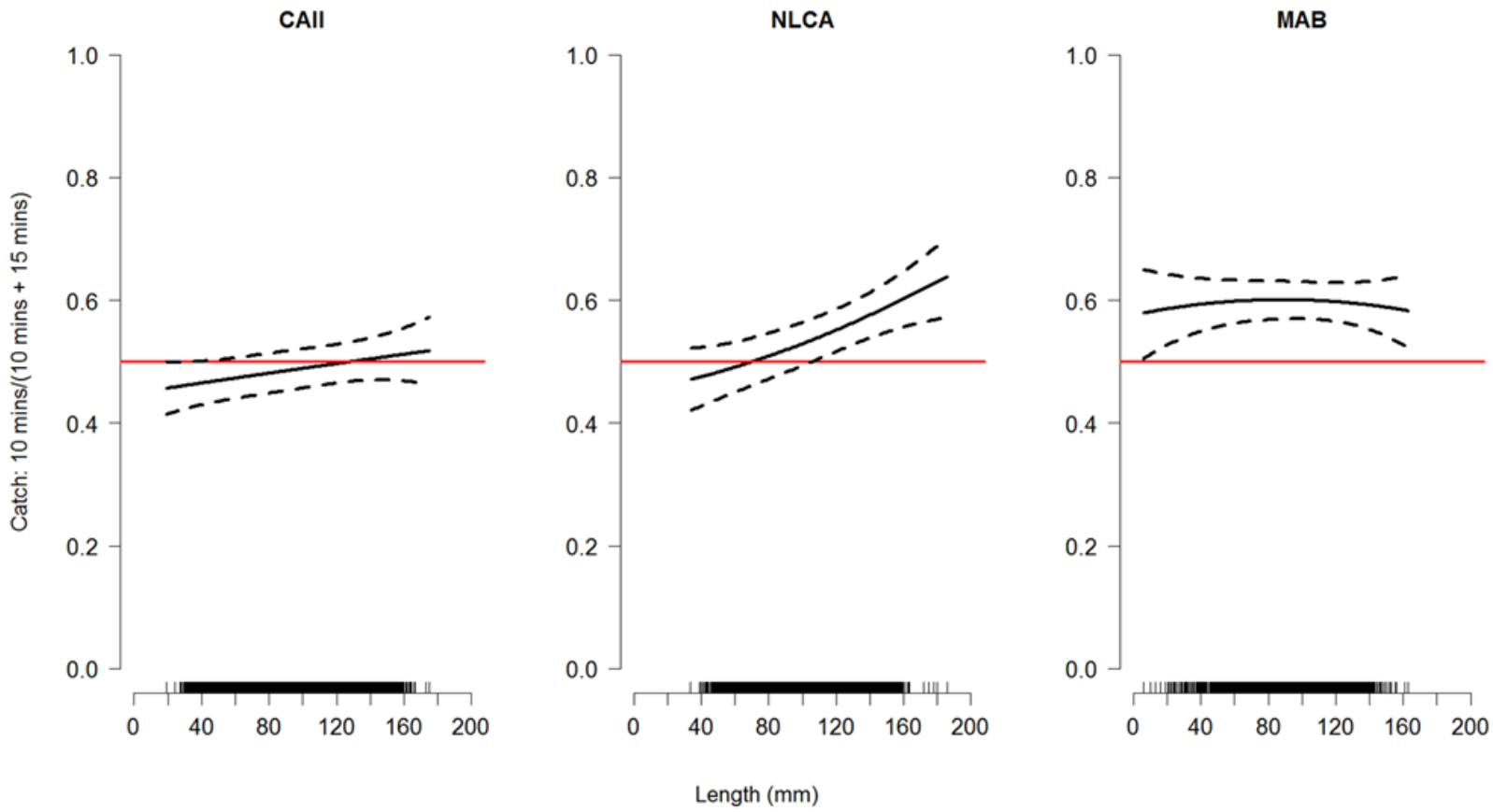


Figure 15. Predicted proportion caught at length in the 10-minute tow conditioned on total catch at length with 95% confidence intervals by area for the optimal GLMM. The red horizontal line of 0.5 indicates equal relative efficiency. A value greater than 0.5 indicates the 10-minute tow had a greater relative efficiency. The rug on the x axis are the observed lengths. Areas: NLCA is Nantucket Lightship, CAII is Closed Area II and MAB is Mid-Atlantic.

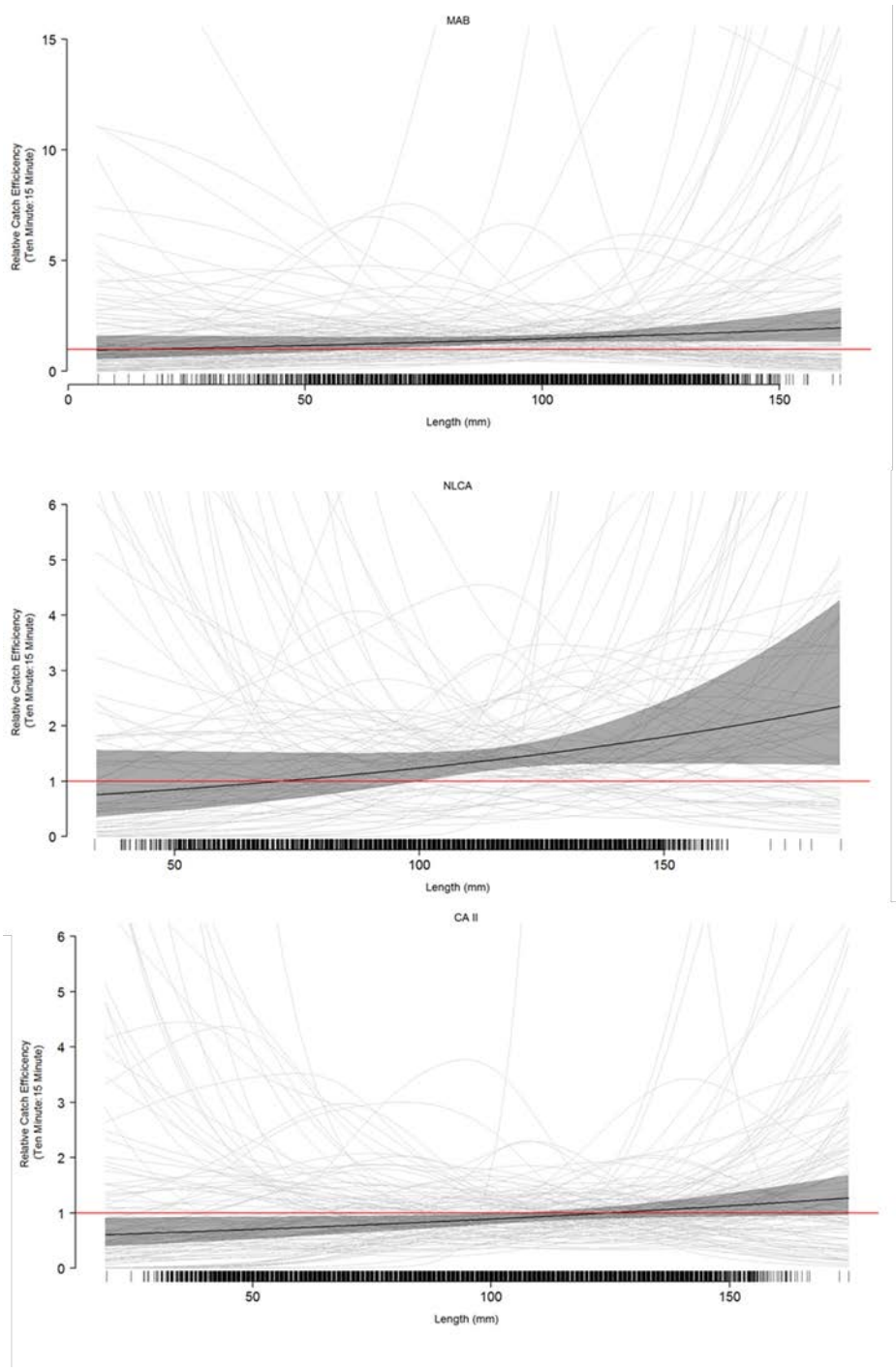


Figure 16. Predicted proportion caught at length in the 10-minute tow conditioned on total catch at length with 95% confidence intervals by area for the optimal GAM. The red horizontal line of 1 indicates equal relative efficiency. A value greater than 1 indicates the 10-minute tow had a greater relative efficiency. The rug on the x axis are the observed lengths. Top: Mid-Atlantic (MAB), Middle: Nantucket Lightship (NLCA), Bottom: Closed Area II (CA II).

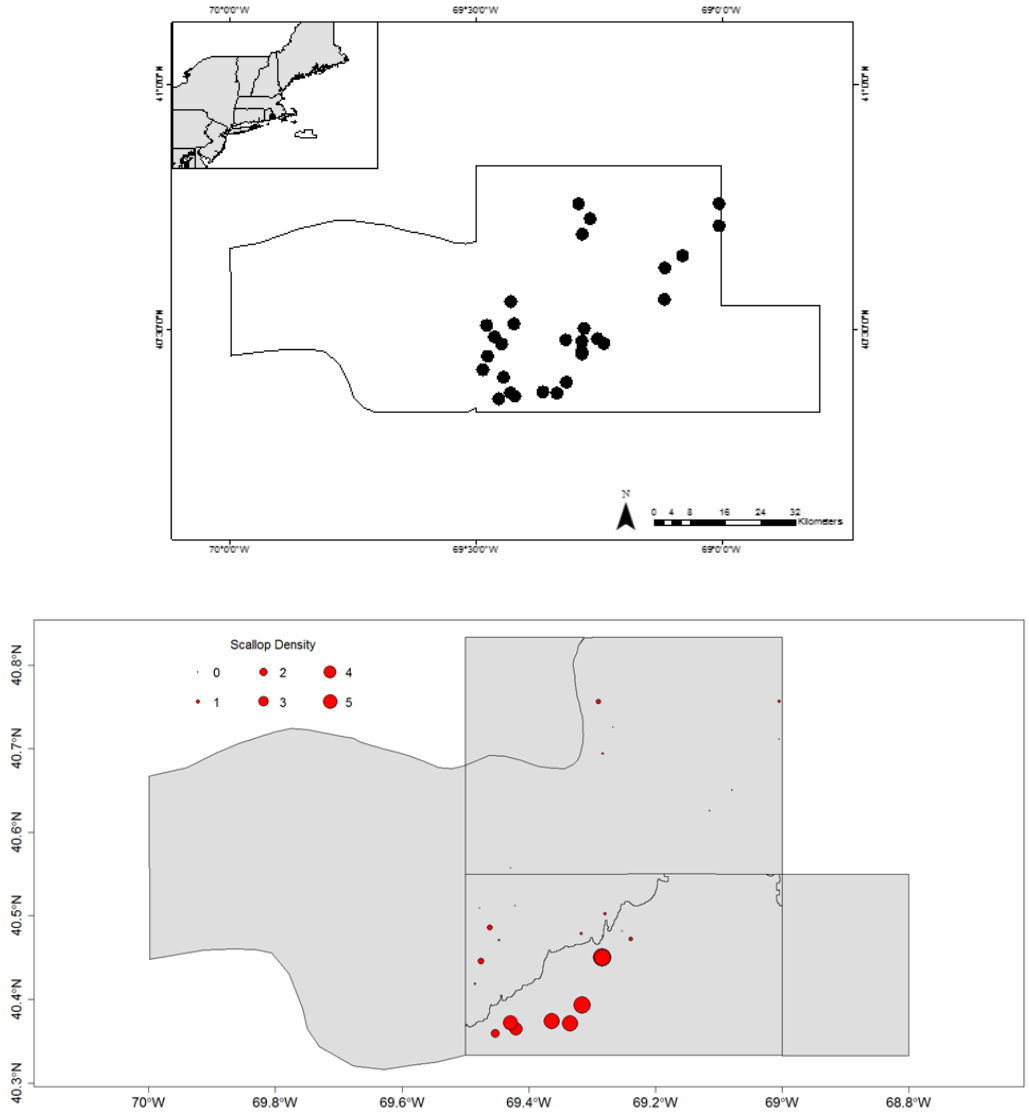


Figure 17. (Top) Location of all survey dredge AUV pairs completed in the Nantucket Lightship study area. The black outline is the VIMS 2017 survey domain. (Bottom) Bubble plot of survey dredge density estimates (scallops per m^2) for each station. The black outlines indicate the 2018 SAMS areas.

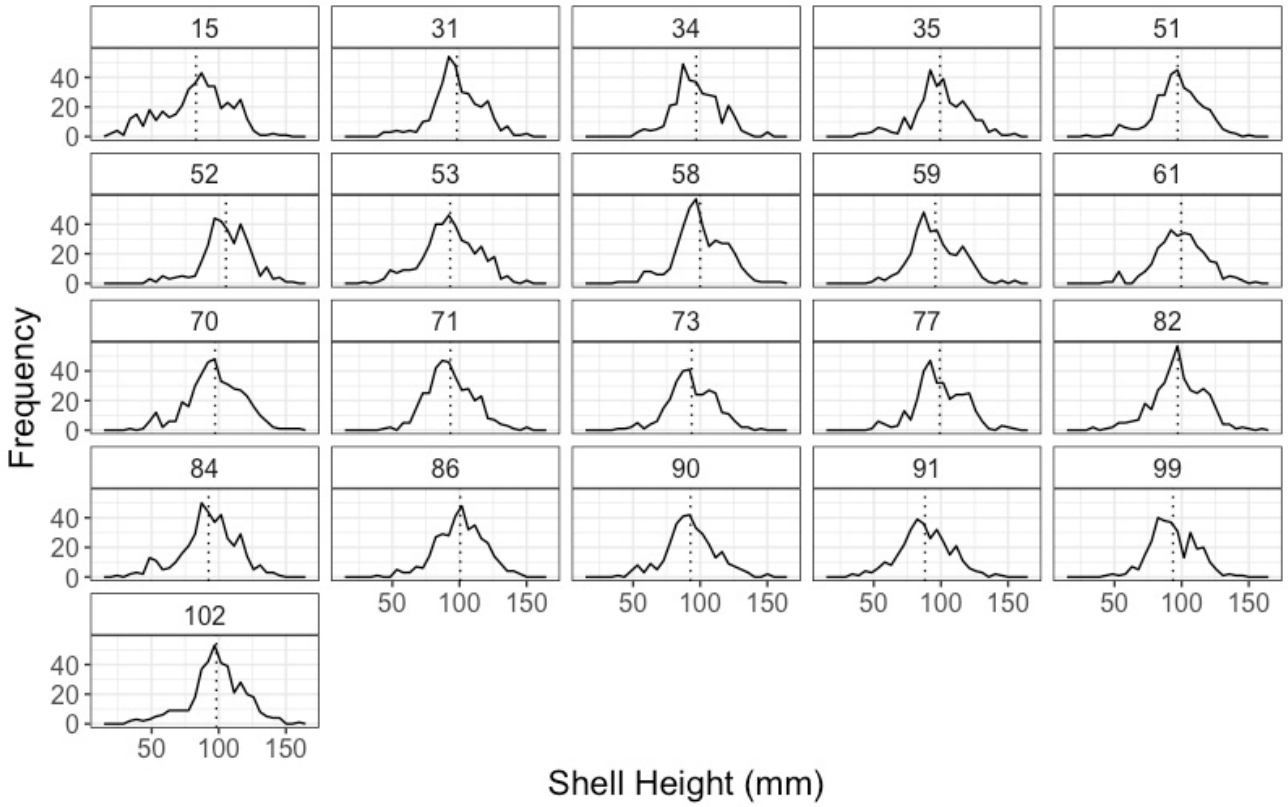


Figure 18. Distributions of scallop shell height within a standardized set of 112 images by 21 annotators. Each subplot is labeled with the annotator's unique ID, and dotted vertical lines represent the mean for each annotator.

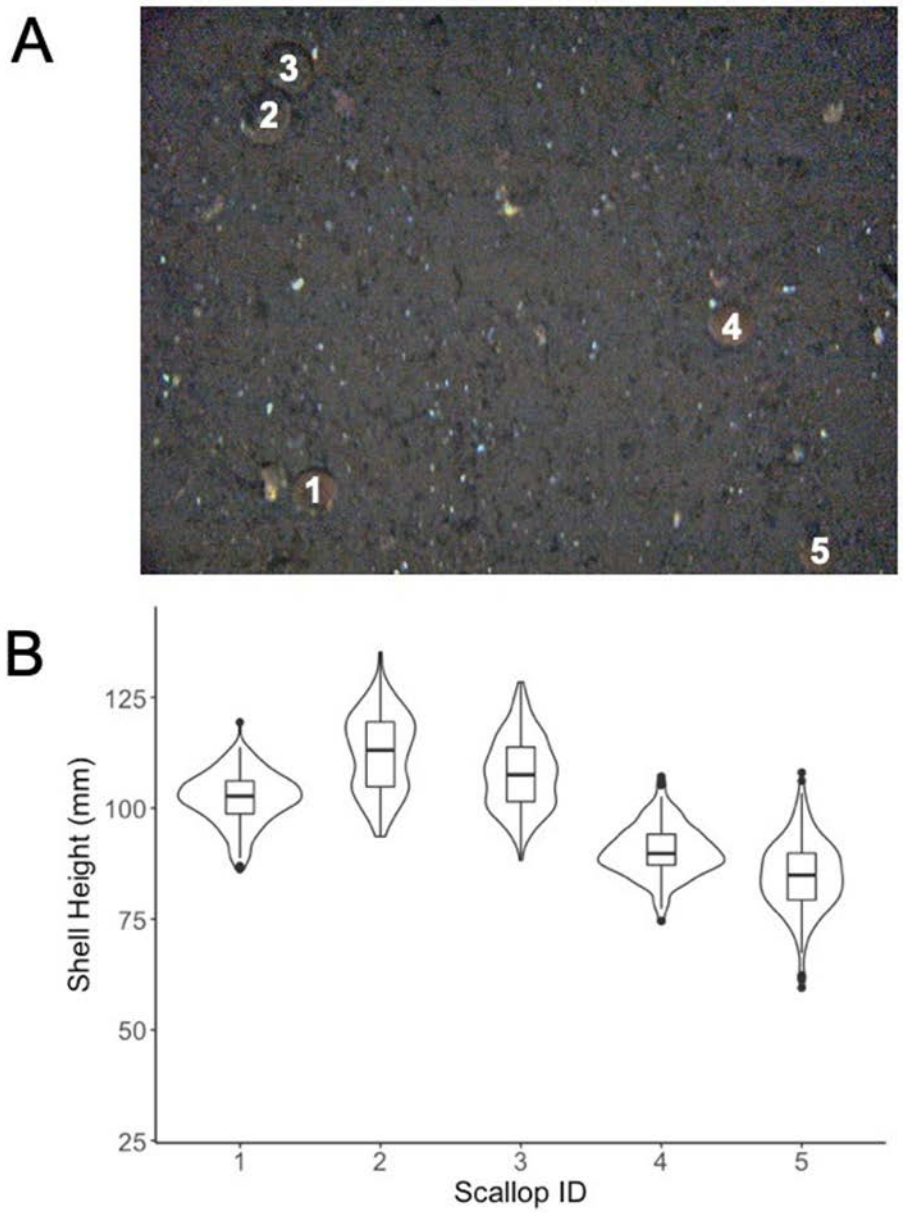


Figure 19. (A) The five scallops labeled with an identification number that were each measured 10 times by 21 annotators. (B) Distributions of shell heights for each scallop pooled across the 21 annotators.

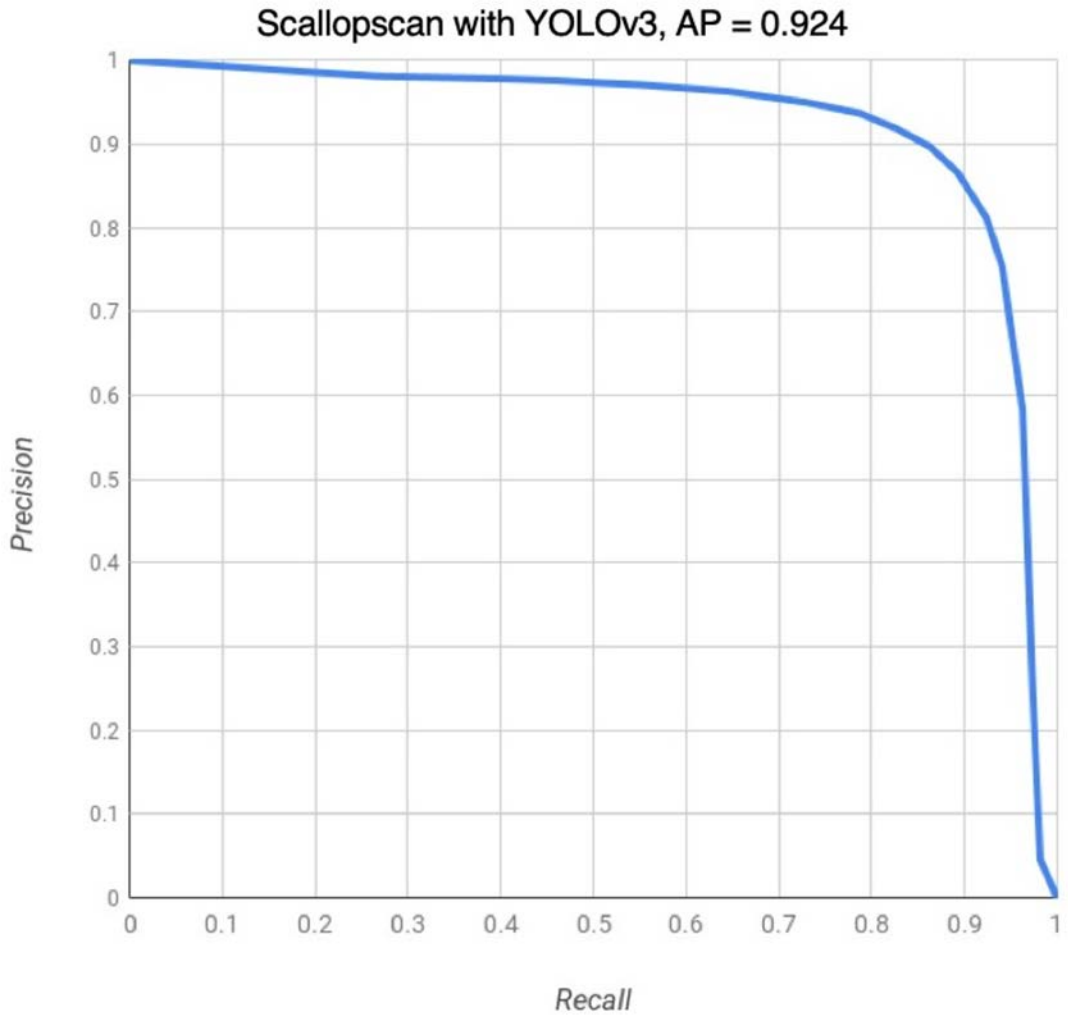


Figure 20. The precision-recall (PR) curve for Scallopscan with the YOLOv3 architecture trained for 20,000 epochs. The curve was generated from a test set consisting of 72,879 scallops from 19,469 images. Average precision (AP) was 0.924 for this image set.

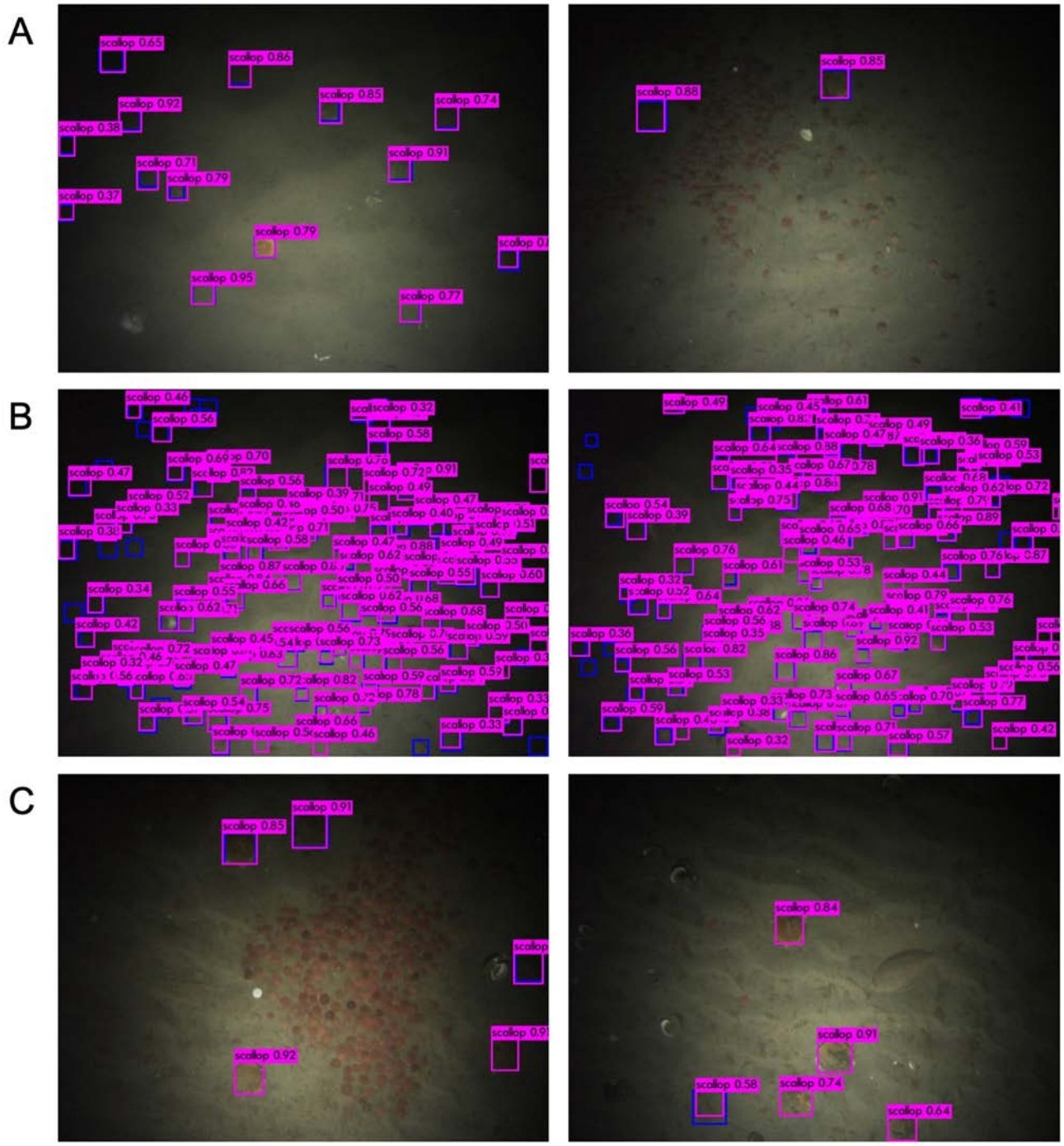


Figure 21. Sample output images from the Scallopscan YOLOv3 test set. Overlaid blue boxes are manual annotations and pink boxes are neural network detections. Each detection is labeled with a confidence value. Examples are shown of images where (A) counts between annotators and Scallopscan agreed, (B) Scallopscan missed scallops that were crowded, obscured by sediment, or located on the perimeter of the image, and (C) Scallopscan detected scallops that were inadvertently missed by annotators.

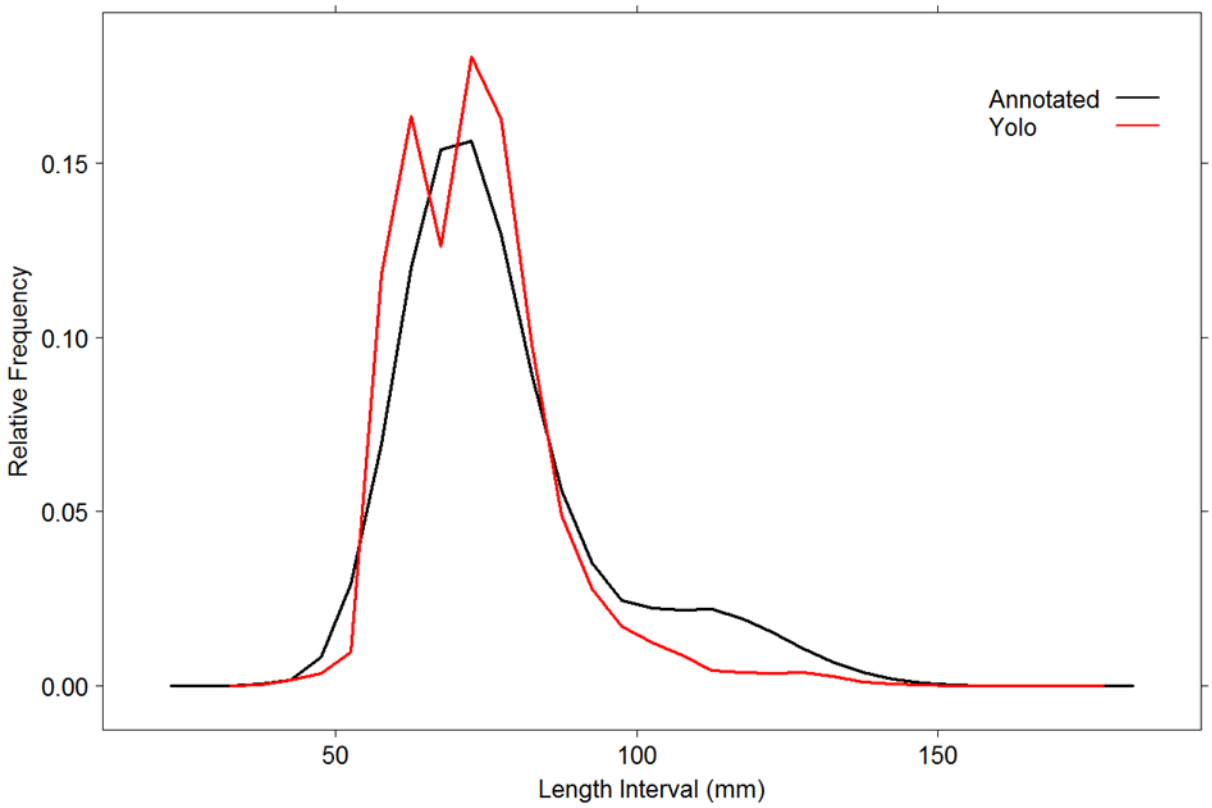


Figure 22. Relative length frequency distributions pooled across 20 stations for human annotated (black line) and YOLOv3 annotated (red line) data.

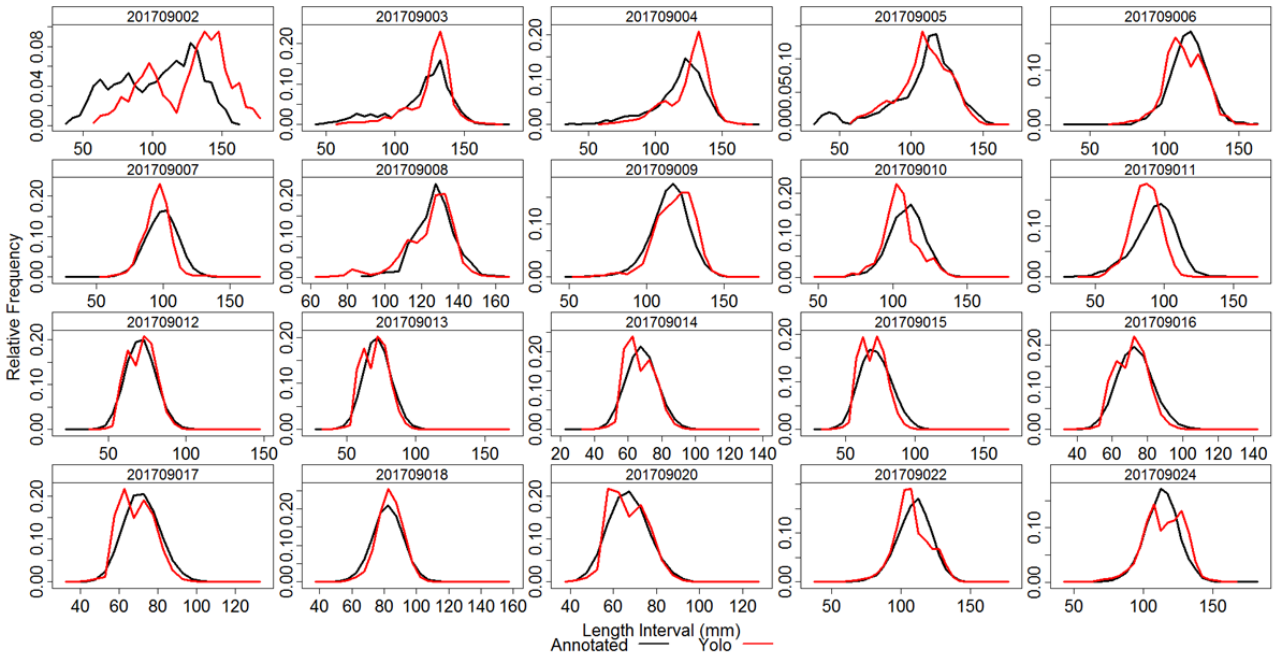


Figure 23. Relative length frequency distributions by station for human annotated (black line) and YOLOv3 annotated (red line) data.

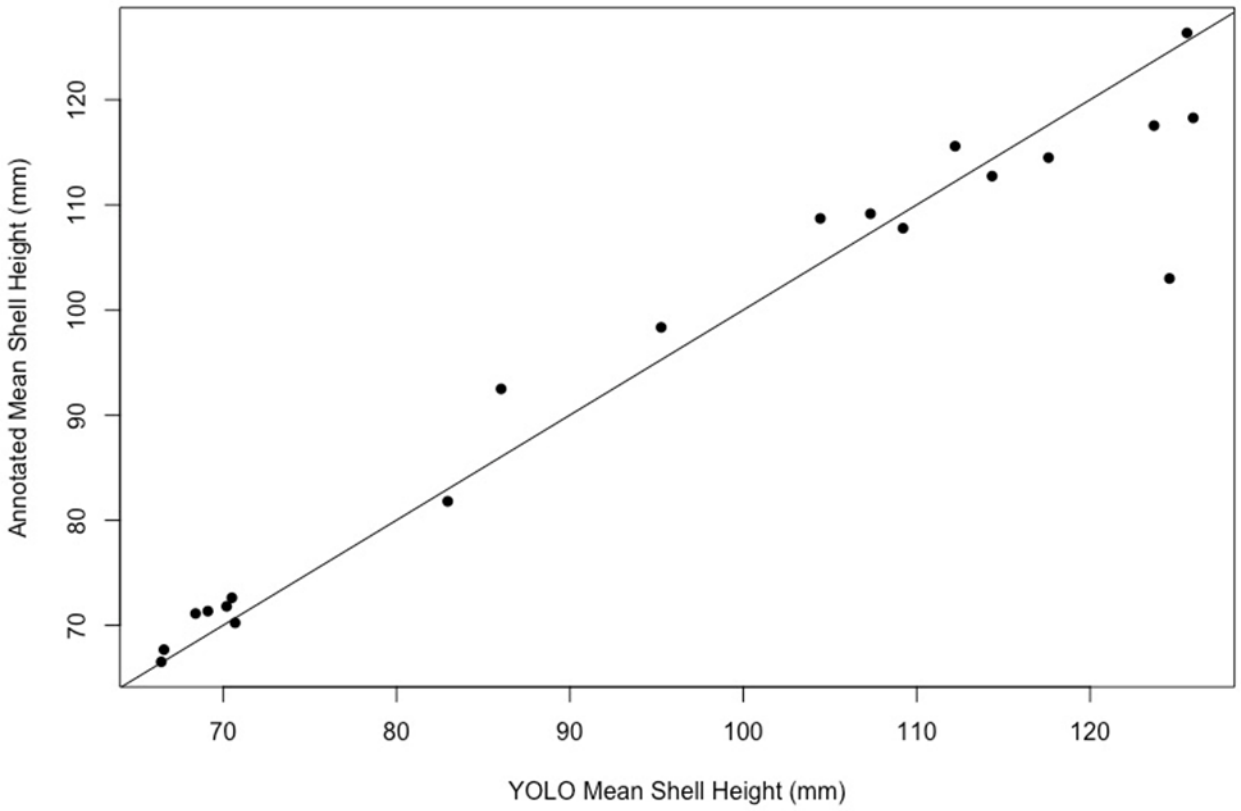


Figure 24. Mean shell heights (mm) per AUV mission from manual annotations plotted against mean shell heights (mm) per AUV mission from YOLOv3 with a 1:1 line.

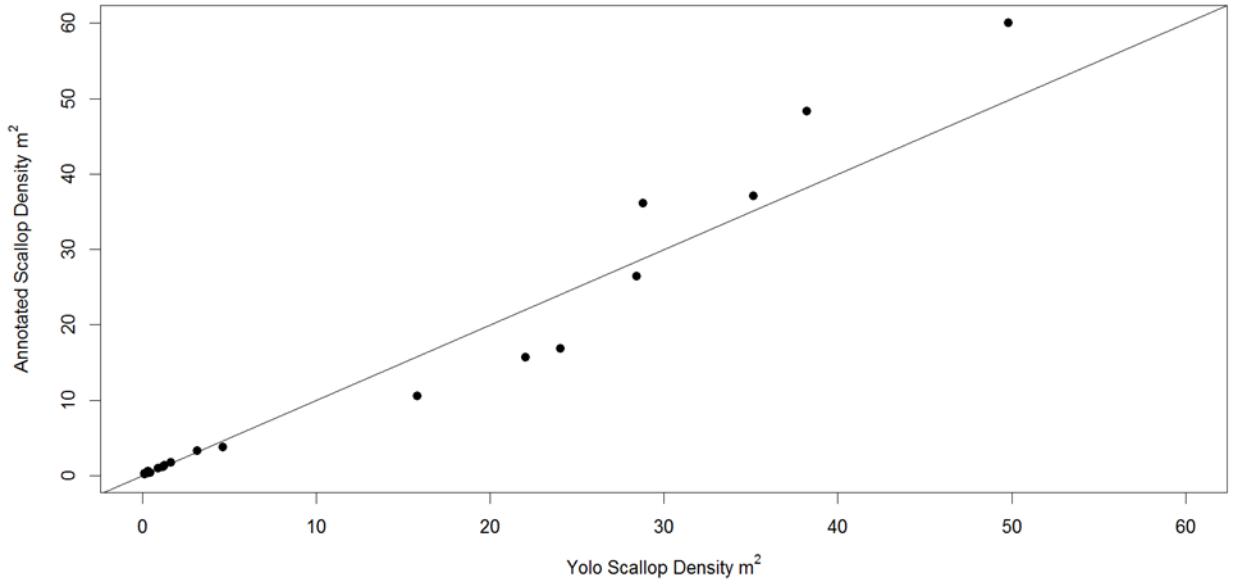


Figure 25. Human annotated scallop density estimates (scallop per m^2) plotted against YOLOv3 annotated scallop density estimates (scallop per m^2) by station with a 1:1 line.

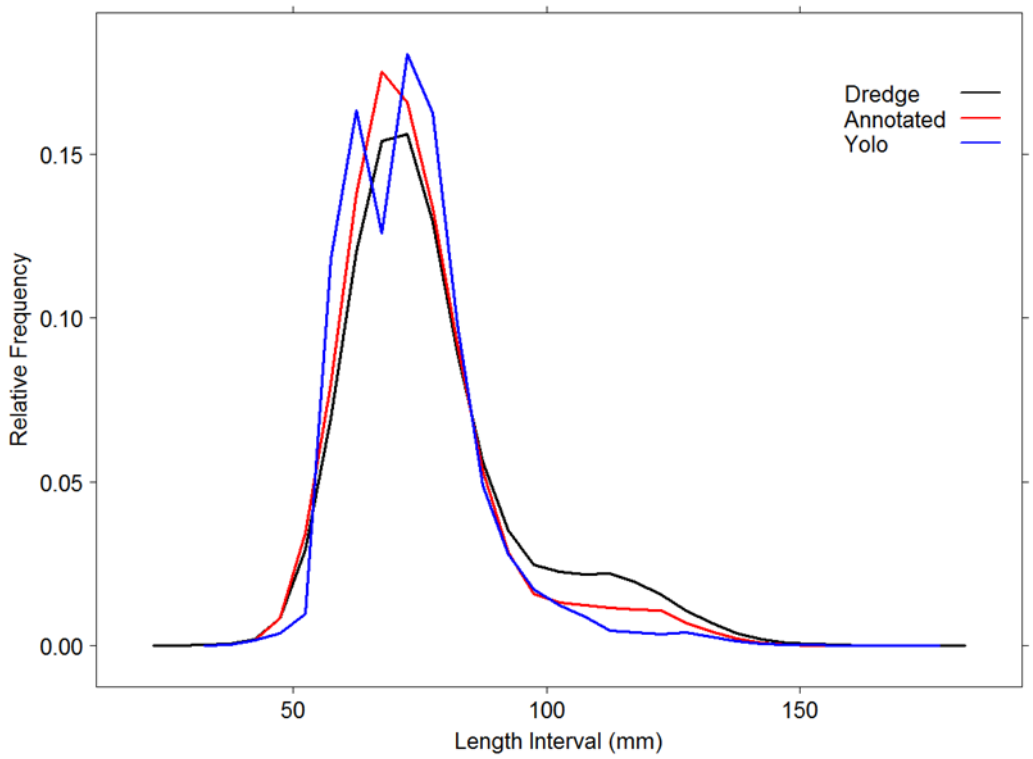


Figure 26. Relative length frequency distributions pooled across 20 stations for survey dredge (black line) and human annotated data set (red line and YOLOv3 data set (blue line).

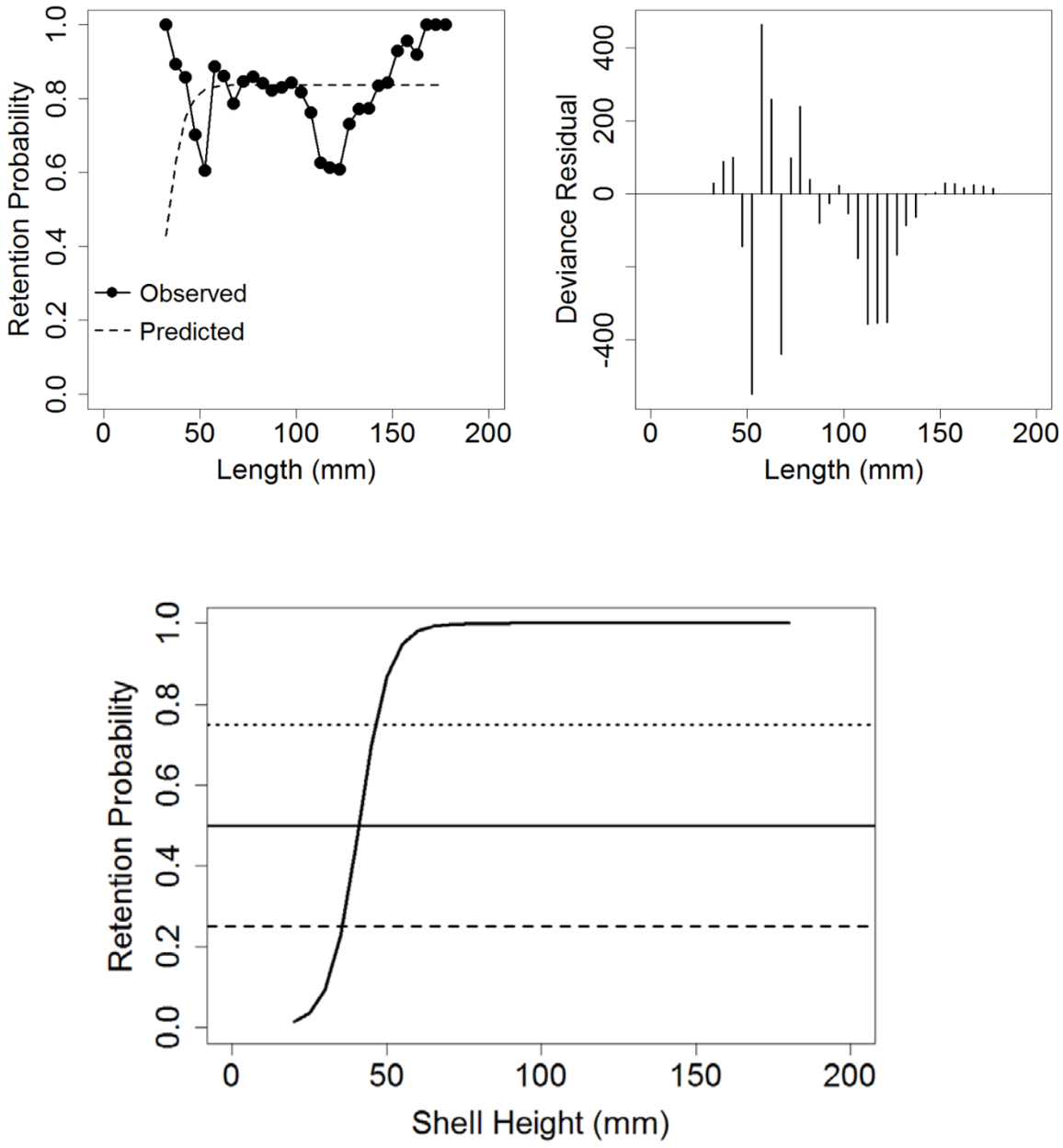
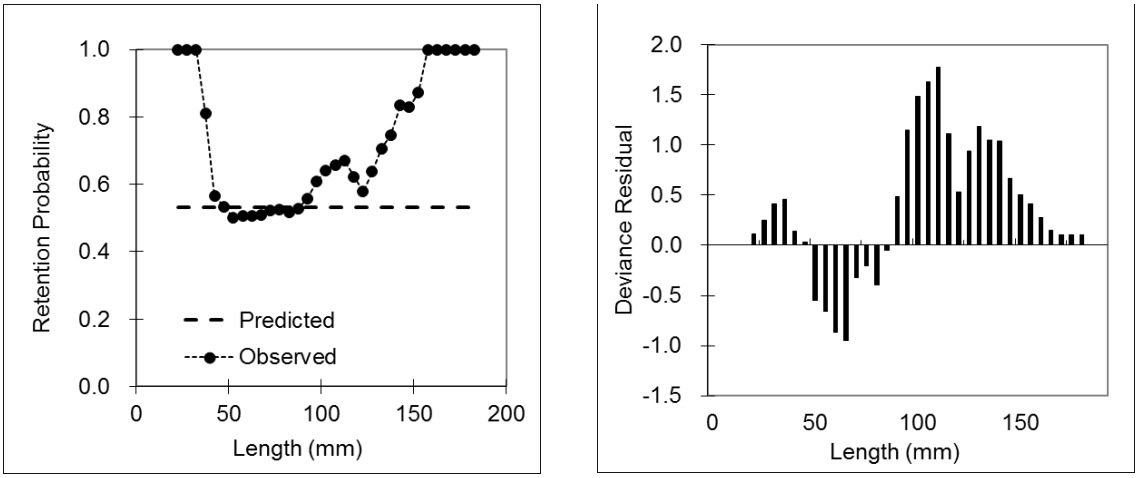


Figure 27. (Top) Predicted and observed proportion caught-at-length in the YOLOv3 annotated data set (left) and deviance residual plot (right) for the logistic SELECT model. (Bottom) Predicted selectivity curve for the YOLOv3 data set, along with the 25 percent retention probability (lower dashed horizontal line), 50 percent retention probability (middle black horizontal line) and 75 percent retention probability (upper dashed horizontal line).



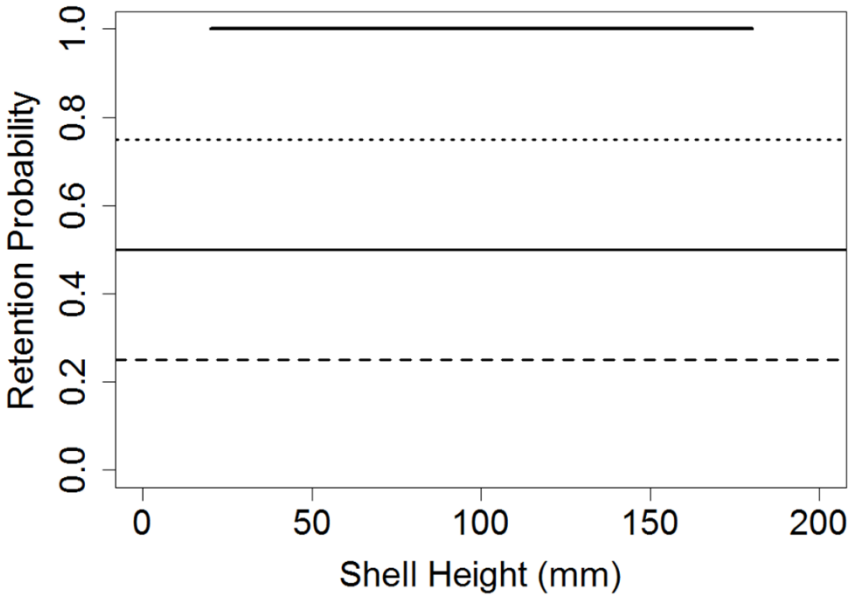


Figure 28. (Top) Predicted and observed proportion caught-at-length in the human annotated data set (left) and deviance residual plot (right) for the logistic SELECT model. (Bottom) Predicted selectivity curve for the YOLOv3 data set, along with the 25 percent retention probability (lower dashed horizontal line), 50 percent retention probability (middle black horizontal line) and 75 percent retention probability (upper dashed horizontal line).

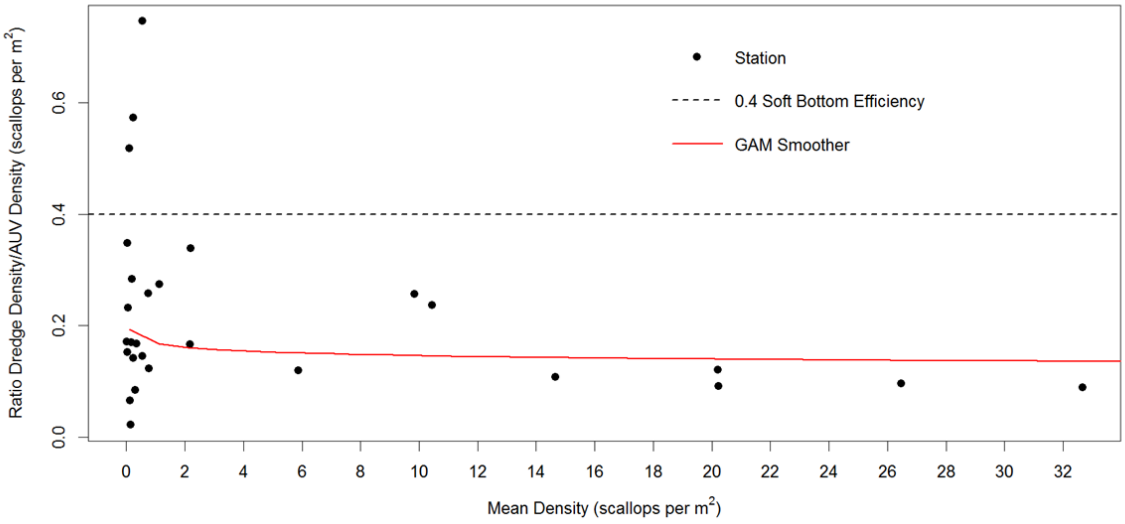


Figure 29. The ratio of dredge density to AUV density (scallops per m2) plotted against mean density (dredge and AUV) by station. The black dashed horizontal line is the assumed soft bottom dredge efficiency of 0.4. The red curve is the GAM smoother.

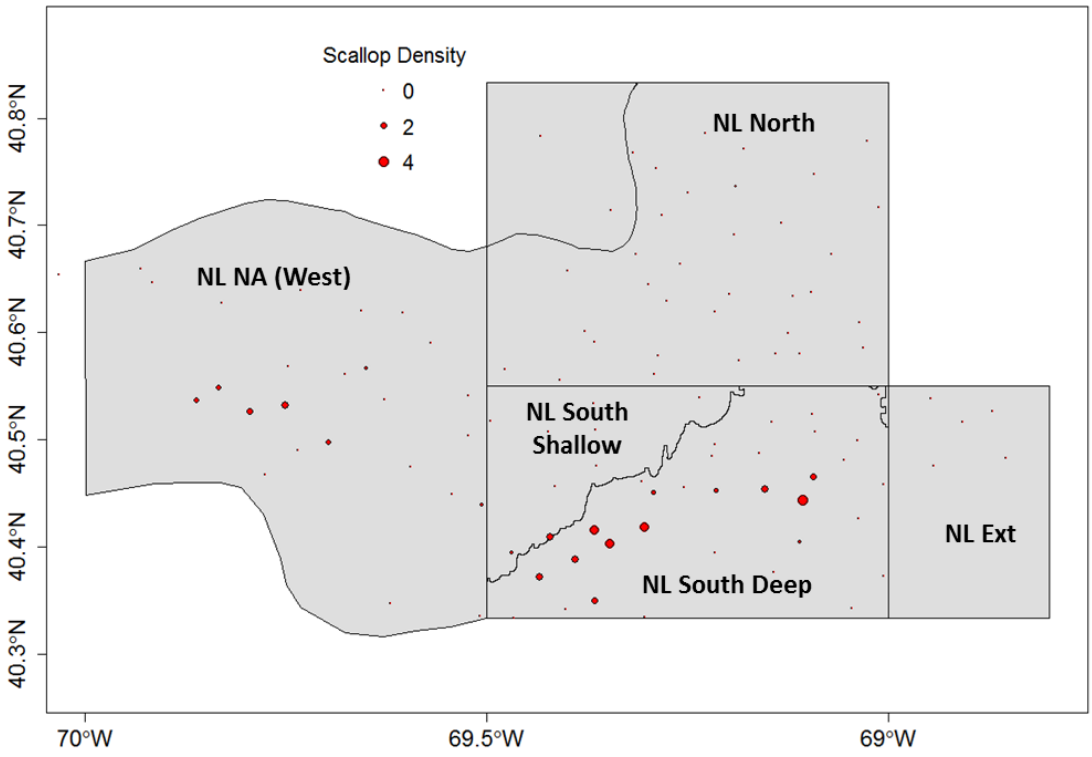


Figure 30. Density (scallops per m^2) at each station completed during the VIMS 2018 survey in the NLCA. SAMS areas are also identified.

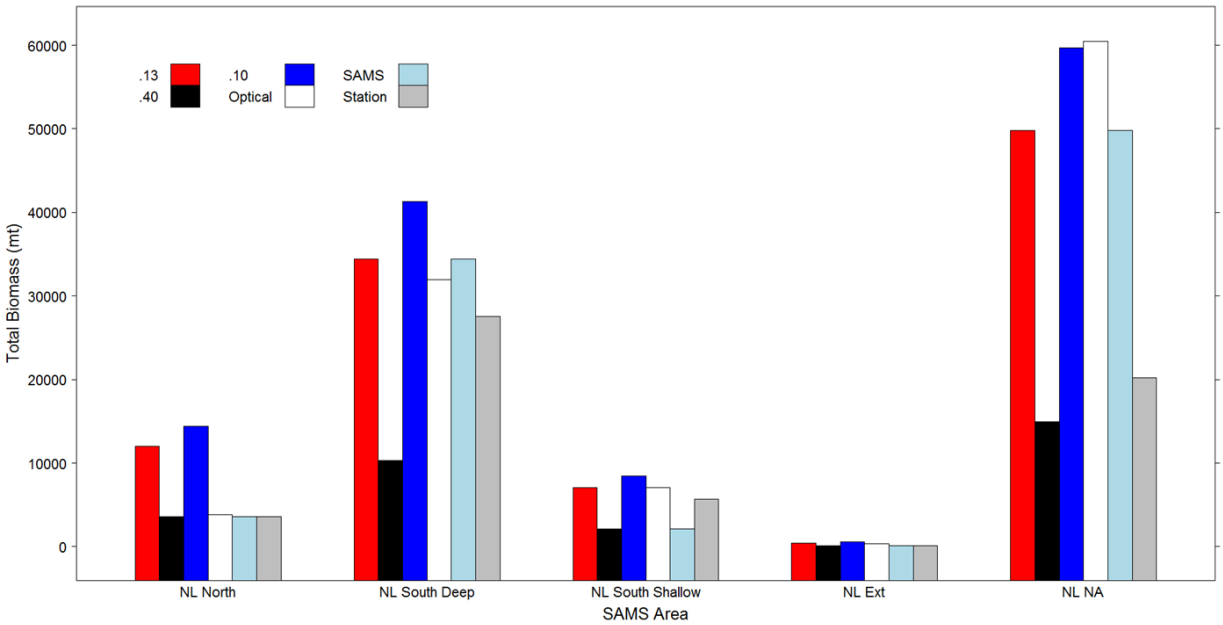


Figure 31. Absolute biomass estimates (mt) for the VIMS 2018 NLCA survey using several different efficiency assumptions plotted with the 2018 NEFSC Habcam absolute biomass estimate (white bar). 0.13 = 0.135 efficiency applied to the entire survey area. 0.40 = 0.40 efficiency applied to the entire survey area. 0.10 = 0.10 efficiency applied to the entire survey area. SAMS= either 0.135 or 0.40 efficiency value applied depending on SAMS area. Station = either 0.135 or 0.40 efficiency value applied based on a station-level density threshold of 2 scallop per m^2.

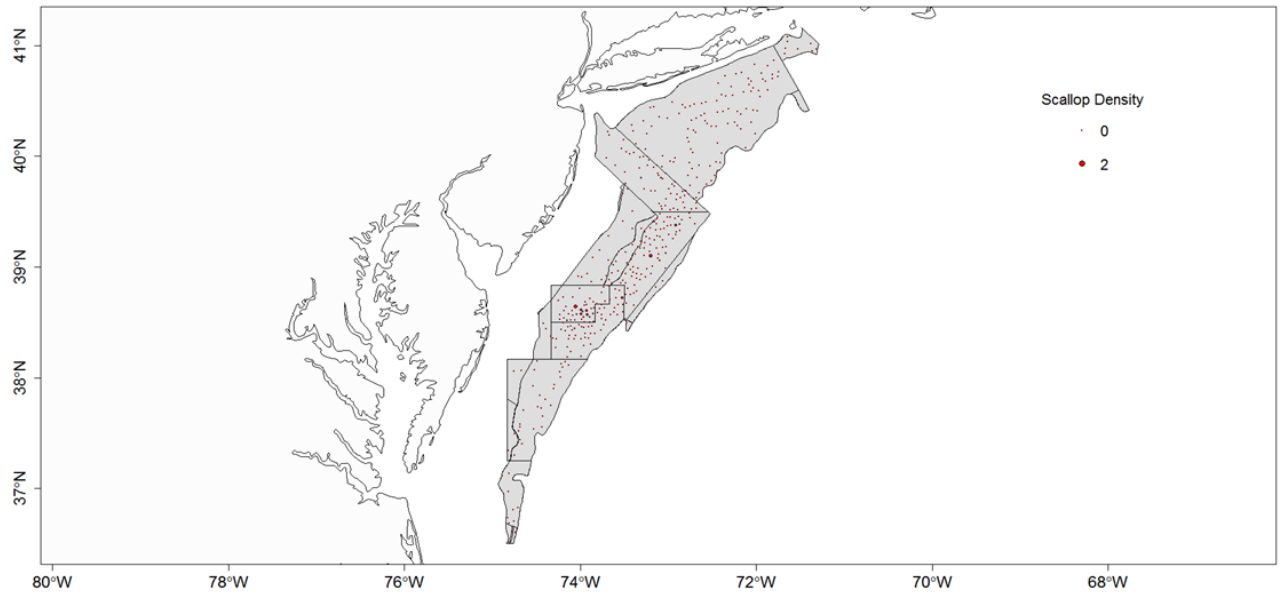


Figure 32. Density (scallops per m2) at each station completed during the VIMS 2018 survey in the MAB. SAMS areas are also identified with black outlines.

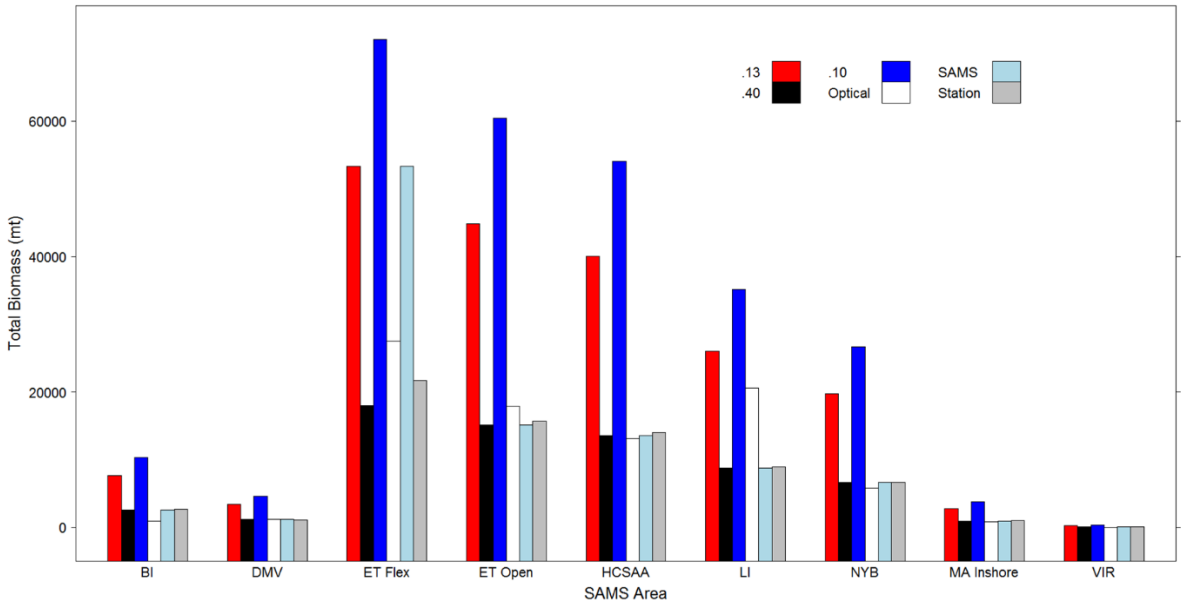


Figure 33. Absolute biomass estimates (mt) for the VIMS 2018 MAB survey using several different efficiency assumptions plotted with the 2018 NEFSC Habcam absolute biomass estimate (white bar). 0.13 = 0.135 efficiency applied to the entire survey area. 0.40 = 0.40 efficiency applied to the entire survey area. 0.10 = 0.10 efficiency applied to the entire survey area. SAMS= either 0.135 or 0.40 efficiency value applied depending on SAMS area. Station = either 0.135 or 0.40 efficiency value applied based on a station-level density threshold of 2 scallop per m2.

Table 1. Summary information for tow duration studies in Closed Area II (CAII), Nantucket Lightship (NLCA) and the Mid-Atlantic (MAB).

Area	Number of Trips	Number of Pairs	Total Number of Pairs for Area	Dates	Vessel
MAB	1	96	96	9/12/2017-9/18/2017	F/V Nancy Elizabeth
NLCA	2	40	80	6/3/2016-6/10/2016	F/V Celtic
		40		7/27/2017-8/3/2017	F/V Celtic
CAII	2	50	100	6/21/2016-6/29/2016	F/V KATE
		50		6/16/2017-6/24/2017	F/V Falvian S

Table 2. Total expanded number of scallops caught, average expanded number of scallops caught and parametric p-values by tow duration (A= 15-minute, B= 10-minute) by area: Closed Area II is CAII, Nantucket Lightship is NLCA and the Mid-Atlantic is MAB.

Area	Total Number (B)	Total Number (A)	Average Catch (B)	Average Catch (A)	P-value
CAII	42,588.55	61,900.58	425.89	619.01	0.04
MAB	67,511.95	75,609.23	703.25	787.60	0.44
NLCA	120,094.66	127,956.82	1,501.18	1,599.46	0.34

Table 3. Total baskets of debris caught, average baskets of debris caught and parametric p-values by tow duration (A= 15-minute, B= 10-minute) by area. : Closed Area II is CAII, Nantucket Lightship is NLCA and the Mid-Atlantic is MAB.

Area	Total Amount (B)	Total Amount (A)	Average Catch (B)	Average Catch (A)	P-value
CAII	313.20	339.00	3.13	3.39	0.29
MAB	371.50	400.90	3.87	4.18	0.41
NLCA	962.30	930.10	12.03	11.63	0.34

Table 4. GLMMs developed for the tow duration portion of the project. Explanatory variables included in each model, along with AIC and ΔAIC are provided. M3 in bold was the preferred model.

Model	Variables	AIC	ΔAIC
M1	~ Intercept	54,101.55	44.01
M2	~ Intercept + Area:Length^2	54,052.86	4.68
M3	**~ Intercept + Area:Length^2 + Length**	**54,057.54**	**0.00**

Table 5. Number of images annotated and number of scallops counted within those images for each AUV mission/dredge tow pair from both manual image annotation and the Scallopscan YOLOv3 detector along with density estimates.

		Manual annotations				YOLO			
Mission ID	Station ID	Num. images	Num. scallops	Imaged area	Density	Num. images	Num. scallops	Imaged area	Density
88	201709001					3215	2422	8311.5	0.29
89	201709002	1130	897	3030.23	0.30	1945	693	5105.44	0.14
90	201709003	2576	1920	6681.93	0.29	1933	1145	5012.53	0.23
91	201709004	1982	2840	5200.19	0.55	2136	1823	5580.19	0.33
92	201709005	2538	1225	6682.08	0.18	1904	748	5012.06	0.15
93	201709006	2065	1291	5491.96	0.24	1545	990	4072.76	0.24
94	201709007	2621	9508	7971.3	1.19	1975	7336	6058.11	1.21
95	201709008	143	510	368.27	1.38	119	403	311.24	1.29
96	201709009	2510	6271	6603.15	0.95	2202	5264	5733.54	0.92
97	201709010	2465	2684	6626.79	0.41	1848	2230	4961.57	0.45
98	201709011	214	5952	566.61	10.50	2672	110673	6988.94	15.84
99	201709012	629	99610	1659.84	60.01	2474	323301	6490.66	49.81
100	201709013	281	27277	755.6	36.10	3140	240920	8361.87	28.81
101	201709014	330	42163	872.34	48.33	613	61829	1616.89	38.24
102	201709015	325	13620	868.12	15.69	2998	174797	7928.55	22.05
103	201709016	367	16496	976.64	16.89	2694	170654	7096.86	24.05
104	201709017	364	26063	984.27	26.48	2482	190581	6701.91	28.44
105	201709018	2438	21236	6452.2	3.29	2325	19369	6096.44	3.18
106	201709019					2959	2268	7740.26	0.29
107	201709020	228	22103	596.3	37.07	500	45893	1305.94	35.14
108	201709021					2639	4163	6867.52	0.61
109	201709022	743	7242	1936.22	3.74	2600	31298	6733.45	4.65
111	201709024	1886	8641	4845.18	1.78	2616	11171	6790.53	1.65
112	201709025					2424	671	6239.33	0.11
113	201709026					2980	402	7712.67	0.05
114	201709027					379	20	975.77	0.02
115	201709028					3116	1347	10681.75	0.13
116	201709029					2611	417	6733.77	0.06
117	201709030					2871	5638	8964.39	0.63

Table 6. Results of the Kolmogorov-Smimov test comparing relative length frequency distributions across the human annotated (Annotated) data set, YOLOv3 data set (YOLO) and survey dredge data set (Dredge) with the associated p-values corrected for multiple comparisons.

Length Comparison	P-Value
Annotated vs Dredge	0.86
Annotated vs Yolo	0.87
Dredge vs Yolo	0.98