NOAA Technical Memorandum NMFS-PIFSC-57

December 2016

doi:10.7289/V5/TM-PIFSC-57

Applications of Hawaii Longline Fishery Observer and Logbook Data for Stock Assessment and Fishery Research

William A. Walsh
Jon Brodziak

Pacific Islands Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

About this document

The mission of the National Oceanic and Atmospheric Administration (NOAA) is to understand and predict changes in the Earth's environment and to conserve and manage coastal and oceanic marine resources and habitats to help meet our Nation's economic, social, and environmental needs. As a branch of NOAA, the National Marine Fisheries Service (NMFS) conducts or sponsors research and monitoring programs to improve the scientific basis for conservation and management decisions. NMFS strives to make information about the purpose, methods, and results of its scientific studies widely available.

NMFS' Pacific Islands Fisheries Science Center (PIFSC) uses the **NOAA Technical Memorandum NMFS** series to achieve timely dissemination of scientific and technical information that is of high quality but inappropriate for publication in the formal peer-reviewed literature. The contents are of broad scope, including technical workshop proceedings, large data compilations, status reports and reviews, lengthy scientific or statistical monographs, and more. NOAA Technical Memoranda published by the PIFSC, although informal, are subjected to extensive review and editing and reflect sound professional work. Accordingly, they may be referenced in the formal scientific and technical literature.

A **NOAA Technical Memorandum NMFS** issued by the PIFSC may be cited using the following format:

Walsh, W.A. and J. Brodziak.
2016. Applications of Hawaii Longline Fishery Observer and Logbook Data for Stock Assessment and Fishery Research. U.S. Dep. Commer., NOAA Tech. Memo., NOAA-TM-NMFS-PIFSC-57, 62 p. + Appendices. doi:10.7289/V5/TM-PIFSC-57.

For further information direct inquiries to

Chief, Scientific Operations
Pacific Islands Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
U.S. Department of Commerce
1845 Wasp Boulevard
Building #176
Honolulu, Hawai`i 96818

Phone: 808-725-5331
Fax: 808-725-5532

Cover: Photograph courtesy of Stuart Arceneaux, Pacific Islands Regional Observer Program, National Marine Fisheries Service.

Pacific Islands Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
U.S. Department of Commerce

Applications of Hawaii Longline Fishery Observer and Logbook Data for Stock Assessment and Fishery Research

[1]William A. Walsh
[2]Jon Brodziak

[1]Pacific Islands Fisheries Science Center
National Marine Fisheries Service
1601 Kapiolani Boulevard, Suite 1000
Honolulu, Hawaii 96814

[2]Pacific Islands Fisheries Science Center
National Marine Fisheries Service
1601 Kapiolani Boulevard, Suite 1000
Honolulu, Hawaii 96814

NOAA Technical Memorandum NMFS-PIFSC-57

December 2016

doi:10.7289/V5/TM-PIFSC-57

CONTENTS

INTRODUCTION

This technical memorandum provides a detailed record of the primary analytical methods used for stock assessment and fishery research that have been applied to the operational and catch data collected from the Hawaii-based pelagic longline fishery by the Pacific Islands Region Observer Program (PIROP) and Fisheries Science Center (PIFSC) of the National Marine Fisheries Service (NMFS) of NOAA Fisheries. The primary purpose of this memorandum is to document assessment-related research using the Hawaii longline data and also to provide a basis for the continued application and future improvement of analytical methods by serving as a user's manual.

The contents of this technical memorandum summarize data preparation, evaluation, and analysis for the PIFSC Stock Assessment Program (SAP) as conducted throughout two decades (1995-2015). The senior SAP scientists who have directed much of the use of the data outputs and analytical results described herein have been Drs. Pierre Kleiber (retired) and Jon Brodziak. The technical memorandum is organized into three sections with four appendices. The first two sections are primarily related to use of data obtained from the federally mandated commercial logbook program initiated in November 1990, and from the PIROP, which commenced activities in February 1994. The PIROP was instituted to measure interaction rates between longline gear and protected or endangered species, especially sea turtles, following a statistical design to estimate turtle interactions (DiNardo, 1993). The last section provides conclusions and guidance on the application of Hawaii longline fishery data for stock assessment and fishery research. The information in the first two sections reflects the hierarchical use of fishery observer and logbook data for stock assessment and fishery research by the SAP of the PIFSC. The PIROP records are considered a high quality data set because the observer reports include detailed information about fishing operations as well as quantities, compositions, and dispositions of catches and are subject to thorough auditing procedures. The PIROP data include many operational fishing parameters that provide important information for conducting analyses such as standardizing the observed catch-per-unit-effort (CPUE) of pelagic fishes. The estimates of standardized CPUE using the operational parameters as covariates provide indices of relative stock abundance through time which, in turn, can be used as input information for conducting stock assessments. The PIROP data also provide information needed to measure the consistency of the self-reported logbook data, which comprise a much larger body of less detailed records, using graphical and statistical methods of comparison between observer- and self-reported commercial longline data. Thus, the quality and interpretation of the information in the logbook database depends, to some extent, on the observer information.

Based on the ongoing assessment work of the Billfish Working Group of the International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean, we have attempted to identify potential improvements in the analytical methods and documentation of applications used for billfish and other stock assessments. This focus on assessment-oriented applications is one theme of this technical memorandum. A second theme is to provide information to develop and maintain a scientific database for fishery research. In this context, the Hawaii longline fishery is characterized by multiple independent sources of data that are highly localized and hierarchically organized and that can be subjected to rigorous data checks and comparisons. Thus, one of our intended goals is to ensure that estimates of the catches, including their species- and size compositions are as accurate as possible for use as inputs to stock assessments and for general fishery research. In general, ongoing development of a scientific longline fishery database and associated applications is likely to enhance the information content of both the PIROP and PIFSC logbook databases for stock assessment and fishery research.

Section I. Preparation and Application of PIROP Fishery Observer Data

In this section, we review the preparation, evaluation, and use of PIROP data for stock assessment and fishery research. Because the PIFSC stock assessment program needs to use catch, effort, and operational data collected by the PIROP, this section focuses on accessing these data and applying them to produce technical work that is clear and reproducible. Furthermore, we review some methods used to interpret PIROP data, in particular CPUE standardization for stock assessment, and provide guidance on possible technical improvements and other uses of these data for fishery research.

The catch and operational data collected by PIROP are used for several assessment purposes, including estimation of pelagic fish catches, standardization of catch rates, and estimating other quantities of interest, such as bycatch interaction rates (McCracken, 2005). The Hawaii longline observer information is considered the best available data for stock assessment and fishery management applications. These include international assessment collaborations conducted as part of the International Scientific Committee for Tunas and Tuna-like Species in the North Pacific Ocean (ISC). Two recent ISC collaborations were the stock assessments for Pacific blue marlin *Makaira nigricans* (ISC 2016) and Western and Central North Pacific striped marlin *Kajikia audax* (ISC 2015), conducted by the ISC Billfish Working Group (BILLWG). The PIROP data have also been used by the ISC Shark Working Group, and Carvalho et al. (2014) recently used PIROP data for shortfin mako *Isurus oxyrinchus* to summarize nominal catch and effort statistics and to describe the distribution of this data-limited shark species. The PIROP data have also been used to produce fishery research which was published in the peer-reviewed literature. For example, Walsh et al. (2009) used PIROP data to quantify shark catch in the Hawaii longline fishery. Brodziak and Walsh (2013) used PIROP data and multimodel inference to standardize CPUE for oceanic whitetip shark *Carcharhinus longimanus*, a longline bycatch species, and Walsh and Brodziak (2015) conducted similar multimodel inference and CPUE standardization analyses for five species of incidentally caught billfishes.

Methodological Overview

This methodological overview section is divided into three parts. These are: (i) preparation of the PIROP data, (ii) evaluation of the PIROP data, and (iii) archival of the analytical data and results. Here we note that data preparation is a large but routine task, whereas data evaluation is not routine and requires detailed knowledge of the fishery.Both data sets and results should be archived to ensure reproducibility. In what follows, we use **boldface type** to denote ORACLE databases, files, and field names or R function or packages (R Core Development Team, 2013).

Preparation of PIROP Data

Data extraction from the ORACLE databases has typically been conducted on UNIX platforms supported at the PIFSC. Both the PIROP and the longline logbook data are located in the highly migratory species enterprise database at the PIFSC. Detailed information on the structure of the enterprise ORACLE database is available at: http://ias.pifsc.noaa.gov/lop_rpts/lodsprinttablelist and also at the InPort PIFSC Hawaii Longline Logbook Metadata Portfolio, which is available at https://inport.nmfs.noaa.gov/inport/item/2721.

The ORACLE database that stores the PIROP and logbook data is organized into schemas and tables. Of these, four schemas and nine tables are typically required to prepare stock assessment data for the Hawaii longline fishery. Preparation of the PIROP data primarily uses the schema **NEWOBS**. The associated table **CATCH_MV** contains the detailed catch and operational data from the entire history of the PIROP. This schema also includes three **LEGACY** tables, which contain fish size measurements collected during 1994-2002. The schema **ORADATA** holds the table **WALSH_MARLIN**, which contains corrected catch data for billfishes from 1994-2004 (Walsh et al. 2007). The schema **LLDS** includes the table **MOON_FRACTION_ILLUMINATED**, which contains the nightly values of the illuminated fraction of the moon from the US Naval Observatory (http://aa.usno.navy.mil/data/docs/MoonFraction.php). The latter two schemas and their associated tables can also be used to prepare the logbook data. Logbook data preparation requires use of the tables **LOGHDR** and **LOGDETAIL**, which are in the schema **OPDT**. The **LOGHDR** table contains the set-level fishery operational data that is used to uniquely identify and prepare the individual longline sets as the observations in the analytical data frame. The **LOGDETAIL** table contains species-specific catch tallies, information about the conditions of caught fishes, and information about protected species interactions.

The preparation of the PIROP data consists of nine steps (Fig. 1) in a work flow that begins with fishery data collection and evaluation by the PIROP and proceeds through protocols followed within the assessment program to initiate projects, including pelagic fish stock assessments. While the full work flow includes nine steps, the PIROP data may be first used for assessment purposes at the fourth step. The ongoing archival of data is depicted by the seventh and eighth boxes, representing data and analytical results, respectively.

An example of logbook data preparation is also provided (Appendix A), and PIROP observer data would be treated similarly in most respects. Data preparation entails the extraction of data files from an ORACLE database and importing these data files into R as data frames (**Example A 1)**. This process requires several additional steps (**Example A 2**), which include linking data frames, selecting an appropriate suite of operational and catch variables, calculating derived variables, deleting longline sets with missing predictor values (or imputing missing values if applicable), deleting outlier longline sets, and truncating the selected data to use an appropriate time period or type of fishing effort. We also note that Appendix C shows the actual data collection forms used, which include three important observer data forms as well as the current logbook form.

The preparation of an analytical file with PIROP observer data typically begins with the creation of a vector of unique longline set identifiers to specify the fishing operations to be used in the

analyses. The unique identifier for each longline set is typically assigned as the serial number of the logbook page recorded for that set. Catches of various species can be tabulated by sets and can be matched to the appropriate longline set identifier to build the analytical file. Operational variables are added similarly. It is important at this stage to ensure that the longline set identifier variable is a character variable. In particular, note that using the R command **read.table** to read a dataframe from file into R may turn the longline set identifier field (typically named **LOGBK_PG_NUM**) into a factor variable which will cause subsequent attempts to match operational and catch data to the correct longline sets to fail. It is also important to check that the longline set identifiers do not include records with the values **NA** (not available) or **<NaN>** (not a number) for the set identifier because estimates of fishing effort would be artificially inflated by these empty rows in the data set. Also, the imputation of missing operational variables may be useful to retain longline information that would otherwise be excluded due to missing values.
A typical set of procedures for manipulating an analytical data file prepared with PIROP observer data is listed and summarized in Table 1. Note that the total catch of a species is obtained by summing the numbers caught on all observed sets. The fields for individual fish condition (i.e., alive or dead) and disposition (i.e., kept or discarded) allow one to estimate the on-board handling mortality as the difference between the numbers of survivors of the haul (**CAUGHT_COND_CODE==”A”**) and the dead discards (**KEPT_RETURN_CODE==”D”**). In this context, the on-board handling mortality can be added to the kept catch to estimate the total number of dead fish caught per set. It should be noted that the descriptions of fish condition after capture and at release are subjective but may be useful for management.

One important data manipulation is the addition of the definition of the fishery sector (Table 1) or set type for each longline set. There are two types of sets in the Hawaii longline fishery and these comprise the deep-set and shallow-set fishery sectors. The sector type (**Set_type**) can be assigned to any longline set using the R **ifelse** statement to set sector type as shallow (**S**) or deep (**D**) based on the number of hooks per float (**HKS_PER_FLT**) for the longline gear as

Set_type<- ifelse (HKS_PER_FLT<15,”S”,”D”)

The importance of this manipulation stems from the fact that the Hawaii longline fishery has been managed based on deep- and shallow-set fishery sectors since 2004 (US Department of Commerce 2004). In this case, the deep-set sector is typically targeting or fishing for bigeye tuna *Thunnus obesus*, while the shallow-set sector is fishing for swordfish *Xiphias gladius*. Here the **ifelse** statement means that if the number of hooks per float on a longline set is less than 15, the effort is in the shallow-set sector and the set type is **S** (character variable); else if more than 15 hooks per float are used, then the effort is in the deep-set sector and the set type is **D**.

The size information for the Hawaii longline catch collected by the PIROP is listed as two entries (Table 1) because the morphometric measurements are stored in two different file formats. The early format is the **LEGACY** table which changed in August 2003 to the current format which is the **CATCH_MV** table. To prepare a continuous time series of size data, one can concatenate the vectors of comparable measurements obtained from the **LEGACY** (prior to August 2003) and the **CATCH_MV** tables (August 2003 to the present). In general, the morphometric measurements can be used to calculate length conversion regressions. For example, by selecting fish of any species with both fork and precaudal length measurements, a bivariate regression of the precaudal length on the fork length can be computed. This regression can then be used to predict the precaudal lengths of animals that were only measured for fork length.

The PIROP observers also record comments on their observations at sea. Observer commentary is typically available for a small fraction of the sets, but it can be useful for interpreting the PIROP data. For example, an observer might comment that there was uncertainty about the size or species composition of the catch on a longline set. In this case, the appropriate action could be to delete the observation resulting in some loss of information for an analysis, or retain the observation resulting in potential bias. In either case, the observer comment can help to make an informed decision.

The procedures involved in incorporating sea surface temperature data (SST in °C) into the PIROP data records (**Example A 3**) are presented here in full because this step requires the use of an additional record matching program. Information to use the matching program is stored in the folder **/home/las/matcher** on the Linux machine mar.pifsc.gov, which is part of the PIFSC OceanWatch Program. To begin, the script **file_checker.pl ,** which is in the folder **/home/las/matcher/input_checker/,** is run to check whether the format of the input file is correct. Given the input file format is correct, the matcher program **make_matches_V3** can be run to add the SST values to the PIROP data set. The matcher program is stored in the folder **/home/las/matcher/bin** and uses Ferret software to match various data sets as specified on the command line. A typical program run may require approximately 13 hours to process an input file of about 60,000 lines and generate the final output file. It is important to check that the input file format for the matcher is correct and includes a **negative** sign for the floating point values of longitude. It is also important to include two decimal places for the latitude and longitude values within the linking string variable. Additional variables, whether obtained from the PIROP or external sources including remote-sensing data, can also be easily matched to the longline set and catch information. Overall, these standard procedures for preparing the PIROP data enhance reproducibility and efficiency.

Evaluation of PIROP data

PIROP data are typically checked to assess whether data collected during fishing operations typical of the longline fleet were accurate. It is important to note that the longline fishing vessels can exhibit operational problems as well as changes in fishing gear. As a result, some data may be unsuitable for analytical purposes because the associated fishing operations were atypical or outliers. For example, a commercial shallow set with an unusually large catch of blue sharks (i.e., ≈ 100) or a deep set with numerous mahimahi *Coryphaena hippurus* may be unsuitable because it was not possible for the observer to get an accurate count of the total number of fish given their other duties.

Changes of fishing vessel names are not uncommon in this fishery, which can cause confusion when checking data. The permit number (**PERMIT_NUM**) is useful for data evaluation because it remains with the hull; as a result, data can be checked over time against a single number rather than one or more vessel names. These unique hull numbers are accessible at http://www.st.nmfs.noaa.gov/coast-guard-vessel-search/.

Outliers in the PIROP data that were caused by poor performance of longline fishing gear can generally be identified by checking one or more indicator variables (Table 2; Appendix A). The first indicator variable to check is the longline soak duration, which typically averages 20 hours 6 minutes and 19 hours 36 minutes in the shallow- and deep-set sectors, respectively. A relatively long soak can result from gear damage or other operational problem, such as a protected species interaction. A long soak can be investigated further using additional fields obtained directly from the **CATCH_MV** table (e.g., **SET_INTERACT_YN** or **LINE_PARTED_YN**). Long soak durations can cause distorted catch patterns. For example, the two highest blue shark catches ever recorded in the PIROP observer data (359 sharks per set) were taken on two sets with 27 and 45 hour soaks, respectively. One of these sets included a protected species interaction and yielded a catch of 5 other fish, while the other set included a breakdown that necessitated a return to port with a catch of only 9 other fish. The second indicator variable to check is the time when the longline set begins, or the begin-set time. In this case, delayed gear deployment can be expected to produce lower catches. The third indicator variable to check is the length of the mainline. In this case, a very short mainline length may indicate that the longline set was being used to test a new gear configuration.

Another important aspect of PIROP data evaluation is to check for species misidentifications. For example, Walsh et al. (2005) reported that at least 2% of the PIROP trips prior to June 2002, exhibited systematic misidentifications of billfishes by observers. These misidentifications primarily involved striped marlin *Kajikia audax* being incorrectly reported as blue marlin *Makaira nigricans*. In this context, public fish auction records may be available to provide a definitive check on the reported catches of economically valuable, incidentally caught or target species (Walsh et al. 2005). Identifications of bycatch species, including many sharks, can also be evaluated on the basis of observer comments, observer experience, photographs (**PHOTO_YN**), or specimens (**SPECIMEN_YN**), and the known distributions of pelagic fishes in the Pacific (e.g., Compagno, 1988; Mundy, 2005; Nelson, 2006).

The accrued at-sea experience of an individual observer (**O_OBSERVER_NUM**) may be an important factor for species misidentifications. In particular, misidentifications appear to occur most frequently among new observers. For example, one analysis of PIROP-reported shark catches during 1994–2006 revealed that a disproportionately high fraction (25%) of the total catches of the 10 least common species were reported in two years, from 2000 to 2001. During those years, the PIROP expanded its coverage of the longline fishery roughly fourfold, and the annual mean levels of accrued observer experience were relatively low. This example suggested that new observers, added in 2000-2001, had higher rates of species misidentifications than experienced observers and also exhibited a proclivity for errors involving uncommon species. Another important consideration regarding use of PIROP-reported data is that detailed evaluations have only been conducted for tunas, sharks, and some billfishes (Walsh et al., 2005; 2007; 2009). Hence, if the scope of stock assessments or other research is to be expanded to include more species, it would necessitate long-term, species-specific evaluations of the PIROP data. For example, a shortbill spearfish assessment may be difficult to conduct using the PIROP data because this species tends to follow a similar seasonality in the longline catch as striped marlin, resembles striped marlin, and had very high nominal catch rates during 1998–2000, which included two years of low PIROP observer coverage rates (about 5% during 1998–1999). To evaluate landings and discarded catch data for a particular species, species-specific effects can be expected and may be important. For example, the actual fishery-related effects for a target species such as bigeye tuna, particularly if subject to a quota, could involve high grading (**HIGH_GRADING_YN** or **HIGH_GRADING_COMMENTS**). This discarding practice involves selecting the most valuable fish for landing and discarding others at sea, possibly including conspecifics, to maximize economic returns while minimizing apparent fishery-related effects. Other reasons for discarding on longline trips could include spoilage (e.g., if a large catch of dolphinfish was taken early in a trip) or improved catch rates during the trip, with greater target catches later in the trip.

In summary, it is useful to evaluate the PIROP data to potentially improve the accuracy and precision of analytical data sets by identifying outliers in both catch-related and operational variables.

Archival of PIROP Data

Archival of analytical PIROP data sets commences by establishing a directory with a self-explanatory name and then storing a file with the data preparation notes therein. The data file is named and saved in text format using the **write.table** command in R. Thus, if necessary, the analytical data file could be recreated using the preparation notes, even as a copy exists in text format. Other appropriately named files can also be prepared (using **write.table**) to facilitate examination of various features of the data; e.g., a file **Disposition.txt** might contain the fields needed to calculate handling mortality with the associated set information. Statistical analyses and results can be summarized and archived in one location.

The archival of PIROP data and results is an efficient standard operating protocol, particularly in the context of periodically conducted work, such as stock assessment updates, and ensures that data files from associated previous studies are readily available. As a result, any need to re-evaluate catch or effort data before the initiation of modeling or other analytical work can be minimized.

Use of PIROP Data in Stock Assessments: Hawaii Longline CPUE Standardization

Analyses to standardize Hawaii longline CPUE have been conducted for several species of pelagic fishes taken as targets, incidental catch, or bycatch and reported as counts using PIROP data. These analyses have been conducted for both research and stock assessment purposes and have used common statistical approaches to standardize CPUE (Maunder and Punt, 2004). Calculations to standardize CPUE have typically been done using the R language (R Development Core Team, 2008) and require the use of several libraries and packages (Crawley, 2013; and Zuur et al., 2009; 2012). For example, the negative binomial distribution, a natural error structure choice for fishes reported as counts, requires the library **mass**. The negative binomial generalized linear model (GLM) can be fitted with the function **glm.nb**. Zero-inflated models, which can account for observations of excess zeros, can be fitted by calling the R package **pscl** and using the **zeroinfl** function. Other types of models can be fitted by using the standard R function **glm** and specifying a Poisson distribution with the **family** option (**family=poisson**).

The CPUE standardization analyses using the Hawaii longline data have typically treated longline sets as independent observations, as in Brodziak and Walsh (2013). This approach was based upon the assumption that within-trip relationships among individual sets can exert positive, negative, or indirect effects of varying magnitudes. Examples of negative or positive effects include vessel movements away from areas with low catches of target species but towards areas of high catches of bycatch species, or vice versa. Indirect effects also include private at-sea communications among cooperating vessels leading to ensuing vessel movements to higher catch rate areas. We note that this approach may underestimate the uncertainty about the standardized CPUE of vessels with consistently correlated catches by sets within trips and may also lead to the selection of overly complex models (Walsh and Brodziak, 2015).

Several types of GLMs have been fitted to standardize CPUE using PIROP data. Some standard GLM approaches for CPUE standardization have included counts models for the number of fish caught (Poisson, overdispersed Poisson, negative binomial), delta distribution models (delta-lognormal, delta-gamma), as well as zero-inflated models (zero-inflated Poisson, zero-inflated negative binomial). These alternative GLMs represent different hypotheses about the nature of the stochastic process of fish capture by longline gear, as reflected by the variance to the mean ratio (e.g., Brodziak and Walsh, 2013). Model selection techniques based on information criteria and the use of Akaike weights (Burnham and Anderson, 2002) have been used to assess the relative likelihoods of alternative GLMs computed for oceanic whitetip sharks (Brodziak and Walsh, 2013) and billfishes (Walsh and Brodziak, 2015).

The CPUE standardization process typically includes fitting a set of alternative fixed effects GLMs to the PIROP data using step-wise variable selection. The goal of the step-wise model fitting for the GLMs is to incorporate as many significant and important variables as possible (e.g., hook and bait types, soak duration, moon phase, begin-set time), while recognizing that patterns in fishing operations within fishery sectors can be expected to cause multicollinearity among predictive variables. For this reason, and because sample sizes of longline sets are typically large, on the order of thousands of records, step-wise variable selection has typically been conducted using reductions in the AIC per degree of freedom as an acceptance criteria (e.g., the variable is accepted if it produces a decrease of at least 5 AIC units/df), rather than AIC reductions alone, to avoid overfitting the CPUE standardization model.

Fishing year and quarter are the first variables to include in the step-wise variable selection procedure to fit the GLMs because temporal effects are paramount for CPUE standardization (e.g., Maunder and Punt, 2004). After the temporal factors have been included in the GLM, the step-wise variable selection proceeds to consider fixed effects including factors and continuous covariates along with possible interactions. Reductions in the residual deviance and Akaike Information Criterion values are used to determine whether a predictive variable is retained in the CPUE standardization model. It is important to note that the significance of a deviance reduction between two GLM models, say GLM1 and GLM2, can be tested with the R function **anova** by using the command **anova(GLM1,GLM2,test="Chisq")**. For more information on sequential testing of variables in the context of fitting GLMs, see Crawley (2013) for example. Fishing region is another factor variable that is important for virtually all CPUE standardization analyses. These have been defined as large areas of the Pacific Ocean (Brodziak and Walsh, 2013; Walsh and Brodziak, 2015), following the suggestion from Maunder and Punt (2004), stating that continuous variables with complex, potentially non-linear effects (e.g., latitude; longitude) be discretized. In this case, the region boundaries are set based on 10° latitudinal increments that are convenient, and separation of the fishery longitudinally at 160°W reflects proximity to Honolulu, Hawaiʻi rather than any underlying oceanographic feature.

The longline fishery sector is another important CPUE standardization variable. In this case, the longline set types are tested as a two-level fixed effect factor variable following significance tests for the temporal and spatial factor variables. For both blue and striped marlins, however, greater deviance and AIC reductions were attained by treating set type effects as interactions with the number of hooks per float. In the shallow-set sector, the range of hooks per float is small and the effect is minor. In the deep-set sector, however, the hooks per float range (15–40 or more) is sufficient to reduce marlin catches because high values tend to sink the longline gear below depths preferred by billfishes.

Sea surface temperature (SST) has also often been found to be an important predictive variable in CPUE standardization models, and some analyses have revealed non-linear effects. For example, SST effects were expressed as parabolic and linear terms for blue and striped marlins in the ZINB models, respectively (Walsh and Brodziak, 2015), which suggests that comparing SST effects on these billfish species could be informative.

SST was also an important predictor in the zero-inflated negative binomial (ZINB) GLM analysis of oceanic whitetip shark catches per set, with complex effects (Brodziak and Walsh, 2013). The negative coefficient for SST in the binomial model indicated that the probability of extra zeros would vary inversely with SST, and that zero catch observations in tropical waters were more likely to have been true zeros. The difference between the two histograms reflected a thermal barrier, with few positive catches below 24°C. The positive coefficient for SST in the counts process model represented a direct relationship between catch probability and SST, which would also be consistent with expectations for a tropical species.

Comparable results were obtained from other ZINB analyses conducted for several billfish species (Walsh and Brodziak, 2015). For example, the coefficients for parabolic effects of SST in the counts process and binomial models for blue marlin were positive and negative, respectively, as would be expected for the most tropical of the istiophorids (Nakamura, 2001).

The **predict** function in R can be used to calculate estimates of standardized CPUE for a fitted GLM object. For example, a fitted CPUE standardization model with the year, region, set type, and SST as predictors can be applied to a different data frame (**newdata=std_data**) comprised of a vector of fishing years (e.g., 1995–2014) and constants for the other predictors (e.g., Region 4; Set_type=D; 27°C). The output would be the estimated annual mean CPUE in the deep-set sector within Region 4 at a mean SST of 27°C. The CPUE trend can be estimated using the R command **tapply(predict(GLM$year,newdata=”std_data”),mean)**. We note that Crawley (2013) also presents code to obtain the back-transformed fitted values for several GLMs with different distributional assumptions.

In summary, recent CPUE standardizations conducted using the PIROP data from the Hawaii longline fishery have provided important information about the relative abundance of several species of large oceanic pelagic fishes. These studies have investigated and tested distributional hypotheses about the process of longline capture and this represents a general improvement in the methods to standardize longline fishery CPUE for large oceanic pelagic fishes.

Use of PIROP Data in Stock Assessments: Hawaii Longline Catches

The Hawaii longline fishery catches by species are important inputs for stock assessments of highly migratory species in the North Pacific Ocean. This catch information derived from PIROP and logbook data has been used for tuna, billfish, and shark stock assessments conducted by the ISC and the WCPFC. For example, the catch for the 2011 striped marlin stock assessment (Walsh and Ito, 2011) was compiled using corrected striped marlin catch data (Walsh et al., 2005; 2007), and PIROP data were used to estimate discarding and other fishery-related effects. This approach to estimating the Hawaii longline catch of striped marlin was consistent with the criteria for best available scientific data for stock assessments listed in Brodziak and Dreyfus (2011), who identify the need for accurate species identifications. The 2015 striped marlin stock assessment update also used the corrected catch data for 1975 to 2009 that was used in the 2011 assessment (Walsh and Ito, 2011) because the nominal and uncorrected catch data in this fishery are known to be biased by misidentifications of striped marlin (Ito, 2015). However, corrected catch data were not available for 2010–2013 in the 2015 assessment update because there was insufficient time to conduct the detailed correction process based on checking dealer reports of striped marlin catches.

Use of PIROP Data in Stock Assessments: Hawaii Longline Length Data

Fish length data collected by the PIROP have been routinely used to characterize the size composition of Hawaii longline catches for stock assessments. These length data typically consist of fork lengths (tunas), eye-fork lengths (billfishes), and precaudal lengths (sharks). Lengths and other size data (e.g., half-girths) collected by the PIROP exhibit sampling variation but this has not been evaluated in a synoptic manner. Some graphical examinations revealed variability among observers taking morphometric measurements, with much of the variation attributable to differences in length measurement techniques (S. Arceneaux, PIROP, pers. comm.). This suggests that analyses for morphometric studies, such as evaluating the ratio of half-girth to fork length as an indicator of reproductive activity, should consider within-observer data checks to ensure that the measurements were taken consistently.

Use of PIROP Data in Stock Assessments: Hawaii Longline Species Composition

The PIROP catch data have been used to characterize the species composition of Hawaii longline catches. In this context, the PIROP data can be applied to conduct multivariate CPUE analyses and other community ecology-related analyses because the commercial logbook data, the other important monitoring tool for the Hawaii longline fishery, does not include entries for many bycatch species, and also because captains often treat the logbooks as landings reports, rather than full catch reports. As a result, species composition analyses can be more reliably conducted using the more comprehensive PIROP data.

Use of PIROP Data in Stock Assessments: International Stock Assessments

The Hawaii longline fishery data collected by PIROP have been used in a number of recent stock assessments of highly migratory pelagics in the North Pacific Ocean. A considerable percentage of the stock assessment work conducted by the PIFSC, including the collaborative effort to complete the 2015 striped marlin stock assessment update, is performed in response to ongoing responsibilities to international scientific organizations (e.g., ISC Billfish Working Group). For example, in 2015, PIFSC contributed several working papers with information on the sizes of striped marlin catches, the size compositions of catches, and CPUE standardizations with time series for the Hawaii longline fishery to the ISC Billfish Working Group (Chang et al., 2015; Langseth, 2015; Walsh and Chang, 2015). These working papers were written to conform to guidelines from Brodziak and Dreyfus (2011) regarding the use of the best available scientific information for ISC stock assessments. In this context, the common purpose was to meet documentation responsibilities with transparency and scientific rigor.

The international stock assessments also provide an opportunity for new assessment research using the PIROP data. In particular, the CPUE standardization analyses for the 2015 striped marlin assessment included mixed model (Langseth, 2015), fixed effect Poisson, and delta-lognormal GLM approaches (Walsh and Chang, 2015). The mixed model analysis was a novel approach for standardizing CPUE while the Poisson and delta-lognormal models were updates from 2011 (Walsh and Lee, 2011). After considering the consistency of the CPUE standardization results across models, the Poisson GLM (Walsh and Chang, 2015) was chosen for inclusion in the stock assessment model for consistency with the previous assessment (Langseth, 2015). Nevertheless, these alternative analytical approaches highlight the ongoing work to improve CPUE standardization using the Hawaii longline data.

Use of PIROP Data for Fishery Research: Peer Review Literature

The PIFSC has published a number of papers in the peer-reviewed literature based on PIROP data. For example, Walsh and Kleiber (2001) used fishery observer data in regression tree and generalized additive model (GAM) analyses of blue shark *Prionace glauca* catch rates. Walsh et al. (2009) provided a quantitative description of observed shark catches in this longline fishery. Brodziak and Walsh (2013) used observer data to standardize CPUE for a bycatch species, the oceanic whitetip shark *Carcharhinus longimanus*. A zero-inflated negative binomial GLM (ZINB) was selected as the best-fitting standardization model, which indicated that the longline capture process was characterized by extra zeros (i.e., a greater frequency than expected under the Poisson or negative binomial counts distributions) and overdispersion in the positive catches. Walsh and Brodziak (2015) conducted similar analyses with incidentally caught billfishes; the ZINB was again found to be the best-fitting CPUE standardization model.

The latter two papers (Brodziak and Walsh, 2013; Walsh and Brodziak, 2015) were characterized by inclusion of results in the text or appendices akin to those in the working papers (e.g., descriptive effort and catch statistics, analyses of deviance, residuals plots), albeit in abridged form. This demonstrated adherence to the practice of providing as much data as possible in readily comprehensible forms to support analyses, even in research papers published in prestigious fishery journals where publishing constraints require concise presentations.

The observer data have also been used to develop methods to correct commercial fishing logbooks, which serve as the principal monitoring tool in this fishery (Walsh et al., 2005; 2007). Walsh et al. (2002) used observer data to develop a method to correct logbook data for blue shark based on use of a GAM as a surrogate observer (Walsh and Kleiber, 2001), while Walsh et al. (2005; 2007) used observer data and public fish auction sales records to do so for blue marlin and other billfishes; these analyses are described in detail in the second section of this technical memorandum.

This brief review documents multiple uses of the PIROP-reported fishery data by the SAP that have met the standards of peer-reviewed fishery journals. Such demonstrations of analytical rigor sufficient to the highest standards also reinforce confidence in working papers prepared in response to ongoing responsibilities because rigor is documented analogously in both contexts.

Review and Summary

This section has reviewed the use of PIROP data for stock assessments, including full data preparation procedures and CPUE standardizations with several statistical models. This has met our first objective, which was to ensure that PIROP data can be reproducibly prepared and that previous analytical work can be fully comprehended by using a reference document.
We have also discussed analytical caveats, reporting practices, and adherence to principles of use of best available science. Much of this experience-based commentary has been presented to meet our second objective, which is to promote scientific rigor and transparency and to facilitate improved use of fishery observer data.

Several aspects of PIROP data preparation could be improved by better R programming, and this should be seriously considered. Some procedures described herein are reproducible, but also highly inefficient from excessive caution. For example, data frame preparation can be performed by matching from data frames manually, but it would be more efficient to have a standard set of R scripts to perform the data preparation steps. If an automated routine could calculate catches from the appropriate fields (i.e., **SPECIES_NAME**, **KEPT_RETURN_CODE**) for required species, it would be very useful. This would greatly reduce data preparation work, and once verified, eliminate mistakes typical of lengthy repetitive tasks. If such a routine were then expanded to obtain the required operational variables, it would represent a major improvement compared to these procedures and might warrant recognition as an official SAP data preparation protocol.

An analogous point can be made about improvements in analytical methodology. Walsh and Brodziak (2014) standardized swordfish CPUE from the shallow-set sector of this fishery during 1995–2012, with a negative binomial GLM. The R library **MASS** and function **glm.nb** allow automation of the forward selection GLM fitting process for this distribution, which warrants evaluation for use with catch data reported as counts.

In addition to striving for technical improvements, the SAP has sought to broaden the scope of CPUE analyses, as evidenced by investigations of several models and distributions in order to gain insight into the stochastic process of longline capture. CPUE standardizations using PIROP data have been published in scientific journals (Brodziak and Walsh, 2013; Walsh and Brodziak, 2015). This attests to the conceptual merit and rigor of the analyses. International reporting responsibilities have been met comparably, by submitting thorough but concise documents specifically intended to foster the principles of use of the best available science among stock assessment collaborators.

Fishery management presently emphasizes predator-prey relationships, energy flows, habitat characteristics, and other aspects of ecosystems. In addition to information about catch rates, our ZINB analyses were also informative about the autecology of large oceanic pelagic fishes because results showed that SST acted as both a controlling factor governing metabolism and a directive factor influencing behavior of these species (Fry, 1971). Thus, the analyses yielded ecologically meaningful quantitative information about the effects of both extrinsic and intrinsic factors.

Catch compilation for stock assessments is subject to careful review. If use of corrected catch histories and PIROP catch data is to continue, adequate personnel and support and appropriate species selection will be required. Potential ramifications of discrepancies in catch totals between uncorrected logbook data and corrected data should also be considered.

Section II. Evaluation and Application of Hawaii Longline Fishery Logbook Data

This section reviews procedures used at the NOAA Fisheries Pacific Islands Fisheries Science Center (PIFSC) to evaluate, correct, and use catch data reported in mandatory fishing logbooks from the Hawaii-based pelagic longline fishery. Although methods are species-specific, all logbook data evaluation and correction procedures are based on data collected by the NOAA Fisheries Pacific Islands Region Observer Program (PIROP).

The major studies of logbook reporting accuracy for this longline fishery (Walsh et al., 2002; 2005; 2007) concentrated on blue shark *Prionace glauca* and istiophorid billfishes (blue marlin *Makaira nigricans*, striped marlin *Kajikia audax*, shortbill spearfish *Tetrapturus angustirostris*, sailfish *Istiophorus platypterus*, black marlin *Istiompax indica*). These studies were conducted by fitting generalized additive models (GAMs) to catch and operational data from the PIROP and then applying the GAM coefficients to an identical suite of predictor variables in the logbook data from unobserved fishing trips (i.e., operational variables) to predict catches. Linear regression techniques were used to identify outliers in the logbook data. Outliers in the catch data for blue shark and billfishes usually reflected non-reporting and species misidentifications, respectively. The major difference between these projects was the availability of public auction sales records to provide independent verification of the corrections applied to incidentally caught billfishes, whereas no such independent verification was available for bycatch species in this fishery, forcing reliance upon statistical inference. Although conducted for very different purposes, both studies demonstrated the feasibility of logbook correction based on use of statistical models fitted to fishery observer data for large oceanic pelagic fishes.

This section reviews the methodology employed in these studies. Several brief discussions of various aspects of logbook data characteristics and use are also included. Appendix B presents detailed examples of several aspects of the logbook data methodology.

Commercial vessels operating in the Hawaii-based pelagic longline fishery have been required to submit fishing logbooks summarizing daily effort and catch to the National Marine Fisheries Service (NMFS) in Honolulu since November 1990. The longline logbook program was initiated because the longline fishery was expanding rapidly during 1987-1991, and because other problems (e.g., interactions between Hawaiian monk seals *Monachus schauinslandi* in the Main Hawaiian Islands EEZ) had been reported (Dollar and Yamamoto, 1991. The logbook program has served as the primary monitoring tool for this longline fishery throughout the ensuing quarter-century. The logbook form is a record for a single longline set that lists species-specific tallies of kept and released fishes, as well as interactions with protected seabirds, marine mammals, and sea turtles. The logbook database at the NOAA Fisheries Pacific Islands Fisheries Science Center now contains a 25-year record, representing more than 350,000 fishing-days.

The logbooks have always been used to monitor catches of four groups of fishes: tunas,; billfishes; sharks, and other pelagic species (or miscellaneous species). This overall monitoring objective has been maintained since 1990, but the fishery has changed considerably during this period, necessitating changes in the logbook program, and the only detailed, published studies of logbook accuracy are not recent (Walsh et al., 2002; 2005; 2007). Thus, despite relative constancy of purpose, little is known about the accuracy of the longline catch data in this fishery, and it cannot be assumed that levels of accuracy have remained constant over time.

The first study of logbook accuracy (Walsh and Kleiber, 2001; Walsh et al., 2002) evaluated logbook reports of blue shark *Prionace glauca* catches from March 1994 through December 1997. Large catches of this species (e.g., > 100 per set) were not uncommon during those years, particularly by vessels targeting swordfish *Xiphias gladius* in the North Pacific Transition Zone. Hence, in simplest terms, the project investigated the accuracy of logbook reports of non-target catches that were sometimes so high as to suggest that enumeration could be inherently difficult, with the objective of identifying sources and estimating the magnitude of reporting bias.

The second project (Walsh et al., 2005; 2007) investigated logbook reporting patterns with billfishes (Family Istiophoridae). The impetus was that species misidentifications caused by superficial similarities, especially striped marlin *Kajikia audax* reported as blue marlin *Makaira nigricans* and blue reported as black marlin *Istiompax indica*, were known to be present in the logbook data, but the sources and magnitude of this reporting bias were not known.

The PIROP observer data are invaluable for direct comparison and modeling purposes, but an additional important source of information, public fish auction sales records in electronic format, has been provided by the Hawaii Division of Aquatic Resources (HDAR) to the PIFSC since January 2000. These three independent sources of information (logbook and observer reports; auction sales records), coupled with the centralized location of this longline fleet, combine to form virtually optimal monitoring circumstances (Walsh et al., 2005; 2007).

In this section, we review methodology used to prepare logbook data for analysis, identify biases and inaccuracies, and perform logbook corrections (Walsh et al., 2002; 2005; 2007). We also discuss the importance of fish auction sales records for independent verification of the corrections applied to logbook data. This section extends the first, which dealt strictly with uses of fishery observer data by describing combined uses of observer and logbook data and uses of observer, logbook, and fish auction sales data.

Background Information

Overview of catch reporting patterns

The catch data from the Hawaii-based longline fishery are characterized by several patterns. Table 3 presents catch data in a comparative format, with mean numbers of caught, kept, and released fish as reported by PIROP observers, in logbooks on the observed trips (i.e., paired observations), and in logbooks from unobserved fishing trips, to summarize these patterns. An important point to consider when comparing the data from the three sources is that most unobserved shallow-set activity occurred before 2000.

The usual pattern for logbook reporting is that the mean numbers of caught and released fish reported by fishery observers exceed those from the logbooks on observed trips, which in turn exceed those from unobserved fishing trips (Walsh, 2000). This usually reflects higher percentages of zero catches and lower numbers of released fish reported in the logbooks, even with an observer present, as exemplified by swordfish and bigeye tuna, primary targets in the shallow- and deep-set sectors, respectively. Blue shark, a bycatch species in the deep-set sector, exhibits substantial differences in catch data conforming to this pattern.

The percentages of the target and incidentally caught teleost species retained for sale in this fishery are very high in most cases. The higher percentages of zeros in the logbooks cause inflation of the apparent retention rates, but most differences are negligible (i.e., a few percentage points). The two obvious exceptions to this pattern are swordfish and albacore, the primary target in the shallow- and a secondary target in the deep-set sectors, respectively. The reason for these deviations is that each is incidentally caught in the other sector, and substantial discarding may occur as a result. This demonstrates the importance of the status of the species within the sector relative to the disposition of the catch. In contrast to the teleosts, the percentages of retained sharks were much lower, at 0.03% and 0.2% for blue shark in the shallow- and deep-set sectors, respectively, calculated from the observer data. Makos have the highest retention rates for sharks, at 28.8% and 11.7% in the deep- and shallow-set sectors, respectively. Threshers are also retained, at 9.1% and 5.5% in the deep- and shallow-set sectors, respectively.

Reporting of finned sharks is another complexity in the data comparison. When shark finning was legal (i.e., before 2000), large fractions of the sharks caught were finned (24% to 42% per year for blue shark during 1996-1999 in the shallow-set sector), but this is no longer the case. Although the effect of finning will decrease in importance as data are accumulated, it will remain important for any specific calculations made during the years when the practice was widespread. The most readily comprehensible comparison of the billfishes data is that between the observer reports and logbooks from observed trips. The blue marlin data reflect over-reporting, which represents bias caused by species misidentifications. This is a fundamentally different matter from under- or non-reporting.

In general, the logbook reports list fewer species and are often negatively biased because fewer released fish and greater percentages of zero catches are listed than in the observer data. As such, the logbook reports are more nearly akin to landings reports than catch reports.

<u>Logbook forms</u>

Several versions of longline logbook forms have been used since November 1990. Major revisions included additions and deletions of entry positions for fishes, addition and subsequent deletion of an entry position for shark fins, rearrangements of the monitored fish groups, and additions of many operational details.

The original version, used through 1994, was arranged with billfishes and tunas as the top and bottom sections, respectively. The billfishes section included five entry positions, including one for black marlin *Istiompax indica*. The kept and released fishes were recorded as tallies; total numbers of kept and released fishes were recorded as numerals.

A form introduced in 1995, was characterized by additional operational information relative to the original version, including increased information about positions and an entry for the number of hooks per float. Also, separate entry positions were added for finned, retained, and released sharks; previous versions had only included entry positions for kept and released sharks.

The form was again revised during 1997, by initiating use of the six-digit logbook page serial number as a unique set identifier. Alternating white and stippled entry positions were added, creating contrast to make the form easier to read. This six-digit logbook page serial number was fully in use by 1998.

The logbook form used from 2001–2004 was extensively modified from the previous version because litigation necessitated changes in the management of this fishery. Swordfish-targeted activity was curtailed so the tunas section was shifted to the top of the form. The black marlin entry position was removed from the billfishes section, which had the intended effect of requiring an entry for "Other Marlin" to report a black marlin. Entry positions for several bird mitigation measures and oceanic whitetip shark *Carcharhinus longimanus* were also added. The logbook form introduced during 2004 remains in use. It differs from the previous version primarily by having had the entry positions for bird mitigation techniques removed.

The remainder of this section presents background information pertaining to the groups of monitored fishes listed on the logbook form. The order of presentation (sharks, billfishes, tunas, and other pelagic species) primarily reflects accrued experience working with data from these species rather than economic value or ecological importance in this fishery.

<u>Shark catches and reporting</u>
The blue shark study (Walsh et al., 2002) was described as being, in simplest terms, an assessment of logbook accuracy when non-target catches are large. In reality, the 1994–1997 study entailed evaluation of the accuracy of logbook data from a fishery considerably different from that now in existence. Because shark finning was legal, blue (and most other) sharks represented incidental catches, with economic value attributable to fins sales. Consequently, logbook reports often included entries for finned sharks, providing some information about removals.

The expectation regarding the logbooks was that blue shark catches and removals in this longline fishery would be greater than the reported total. Results supported this expectation; the estimated under-reporting rate for the study period was 23.9% (Walsh et al., 2002).

The largest blue shark catches were consistently taken on longline sets targeting swordfish. This type of fishing effort, however, was associated with unacceptably high rates of interactions with protected sea turtles, which led to a closure of swordfish-targeted fishing from early in 2001 until April 2004 (Walsh et al., 2009). The fishery then re-opened under a two-sector management regime based upon the target fishing depth of longline gear (Department of Commerce, 2004).

The operational parameter used to define fishery sectors is the number of hooks per longline float. "Shallow" and "deep" sets use <15 and ≥ 15 hooks per float, respectively. Because a hooks-per-float field was first added to the logbook form in 1995, any attempt at retrospective two-sector analysis of data from 1990–1994 must be predicated upon imputation or inference, presumably based upon PIROP data or within-vessel fishing patterns verified during later years. The two-sector management regime has required fishery observer coverage on all shallow-set trips (i.e., swordfish-targeted fishing) since April 2004. The usual target in the deep-set sector is bigeye tuna *Thunnus obesus*, and the annual observer coverage rate has remained near 20% since

2004. Fleet-wide coverage rates during the first decade of PIROP operations were presented by Walsh et al. (2005).

A second major management change affecting sharks in this fishery was the prohibition of shark finning in 2000, by the US Shark Finning Prohibition Act and Hawaiʿi Revised Statute 188-40.5, followed by passage in 2010 of a Hawaiʿi state law banning possession of shark fins. These laws caused most sharks taken by this fleet to become bycatch without economic value. Lacking such value, post-capture release of sharks would be expected, but the change in economic status introduces uncertainty into estimation of released sharks. Figure 2 presents a comparison of logbook- and observer-reported blue shark catches during 1995–2014; the annual median difference was 8.5%.

Independent of the direct effects of the finning prohibition, the two-sector management regime led to reductions in observed shark catches and mortality and altered species composition of the shark catches (Walsh et al., 2009). The expansion of deep-set effort was associated with increased catches of some relatively deep-dwelling sharks, such as bigeye thresher *Alopias superciliosus*. In the shallow-set sector, shortfin mako *Isurus oxyrinchus* was the only species with a large increase in catch rates after the reopening (Walsh et al., 2009), which reflected increased effort northeast of Hawaiʿi (*ca*. 30°–35°N, 140°–145°W).

Southward expansion of deep-set effort, extending to near-equatorial waters, under two-sector management revealed an additional source of logbook reporting bias. Direct comparisons of logbook and observer data from tropical locales where oceanic whitetip sharks *Carcharhinus longimanus* and silky sharks *C. falciformis* would be expected to predominate among requiem sharks (Compagno, 1988; Bonfil et al., 2008) demonstrated that some captains invariably logged all sharks as blue sharks, regardless of species (Walsh, unpublished data). The total shark catches were roughly correct, but the species differences either went unnoticed or unrecorded. Thus, misidentifications of carcharhinid sharks as blue sharks apparently increased the accuracy of the estimates of blue shark catches in near-equatorial waters because the positive bias associated with misidentified catches of silky and oceanic whitetip sharks countervailed the negative bias associated with some under-reporting. This also means, however, that logbook data for tropical carcharhinids would be negatively biased by under-reporting and misidentifications.

Another known source of inaccuracy in shark data in logbook reports is due to revisions in the logbook form. In 2001, one such change in the logbook form was addition of an entry position for oceanic whitetip shark *Carcharhinus longimanus*. Figure 3 compares mean oceanic whitetip shark catches from logbooks and corresponding observer records during 2000–2014; the important feature is that five years elapsed after the logbook revision before the trends converged. Hence, an entry position may be present on the logbook form, but this does not necessarily imply its prompt adoption or correct use.

We also note that some logbook reporting errors have been definitively identified and estimated. Double-counting (i.e., reporting a shark catch twice) sometimes resulted from sharks being listed as both “Finned” and “Released” (for example, see Walsh et al., 2002; Figure 1).

The comparison of logbook reports and PIROP observer data for blue shark catches on observed trips during 1995–2014 (Fig. 2) showed that the observer-reported catches were consistently greater than those reported in logbooks, with an annual median difference of 8.5% per trip. The apparent difference reached its highest levels in 2000–2001, during the expansion of the PIROP, but decreased in 2002, after fleet-wide observer coverage rates stabilized at about 20% of trips being observed. The ongoing pattern of under-reporting of blue shark catches in the observed logbook records suggests that unobserved trips may exhibit similar biases.

The increases in observer coverage can be expected to affect logbook data correction positively because analysis of catch data for sharks in the shallow-set sector can now rely upon direct comparisons with observer reports, and preparation of corrected catch histories could entail simple substitutions. In the deep-set sector, the annual sample sizes available to model catches and evaluate data accuracy are much larger than in 1995–1999.

Although estimation of shark catches and other fishery-related effects in this longline fishery is complex, an important consideration regarding logbook accuracy evaluation for sharks is that some makos (*Isurus* spp.) and threshers (*Alopias* spp.) are sold at auction in Hawaiʻi, with sales records available for verification purposes. These sharks are also highly distinctive in appearance. Hence, it is reasonable to assume that catches of these species could be reported accurately and represent positive controls, with the connotation that these species should be reported as least as accurately as all other sharks.

There are obvious potentially serious complications in this scenario. Bigeye threshers are sometimes taken in relatively high numbers in this fishery; i.e., 1.3% of the observed sets during 1995–2014, with positive catches of bigeye threshers yielded ≥15 (Walsh, unpublished data). Thus, it could be useful to determine whether bigeye threshers are ever reported as blue sharks, as with oceanic whitetip and silky sharks, and whether bigeye thresher catch sizes are ever so large as to cause enumeration to be inherently difficult, as with blue sharks. Analogously, shortfin makos and blue sharks are very dissimilar in appearance, and the former can be extremely dangerous, but they are the two shark species taken regularly by the shallow-set sector in temperate waters. Logbooks could be compared to observer reports and auction sales records to determine whether shortfin makos are reported as blue sharks.

<u>Billfish catches and reporting</u>

This fishery takes species from both families of billfishes. Swordfish *Xiphias gladius*, monotypic in the Family Xiphiidae, is the target in the shallow-set sector and is taken incidentally in the deep-set sector. Istiophorid billfishes are primarily taken incidentally in the deep-set sector. One of the major management challenges with istiophorid billfishes in this fishery is logbook data accuracy in the context of fishery monitoring. Morphological similarities among these species (Figure 4) cause self-reporting biases attributable to species misidentifications (Figure 5[1]) that affect the means, standard deviations, or both parameters of the catch data.

A management action with very positive effects regarding the species misidentifications problem was instituted on January 1, 2000, when the Hawaiʻi Division of Aquatic Resources (HDAR)

[1] Figure 5 is taken from Walsh & Bigelow: “Where the Billfishes Were (and Were Not).” Presentation to the 56th International Conference, Lake Arrowhead, CA, May 2005.

began providing complete public fish auction records in electronic form to the PIFSC. During 1990−2000, NMFS Honolulu Laboratory monitoring staff maintained roughly a one-third coverage rate (i.e., two days per week) at the main fish auction in Honolulu. Additional records were available in HDAR trip reports, but such records were incomplete and if available were often confusing. The electronic sales data now permit efficient, convenient checks on logbook catch data, particularly species misidentifications.

This species misidentifications problem warrants careful consideration because both blue and striped marlins are managed as Highly Migratory Species. In addition, stock assessments for both are conducted under the aegis of the Billfish Working Group of the International Scientific Committee for Tuna and Tuna-like Species in the North Pacific Ocean (ISC BILLWG). The concern is that catch data inputs to stock assessments are biased by misidentifications.

Logbook data accuracy for billfishes was evaluated for a 10-year period, starting at the founding of the PIROP during 1994 (Walsh et al., 2005; 2007). Catch data correction was not attempted for the period between the outset of the logbook program (November 1990) and the founding of the PIROP (March 1994) because comparison standards were not available.

Questionable species identifications with billfishes or other species in the shallow-set sector need not be problematic for future stock assessment purposes. Direct comparisons of observer and logbook reports have been possible since 2004, and auction records have been available for verification purposes since 2000. Verification of the accuracy of catch data could be straightforward. Evaluation of logbook data accuracy for billfishes taken by the deep-set sector could also be relatively straightforward for stock assessment purposes, at least from a procedural standpoint, with reliance upon established methods.

The possibility of a future solution to the problem of billfish misidentifications has not solved the problem of bias from species misidentifications in the past. As a result of partial auction coverage and incomplete HDAR information before 2000, substantial numbers of apparent outliers that probably represented species misidentifications could not be investigated. This means that the logbook database retains considerable bias for billfishes through 1999.

This situation reflects the decision to require strong statistical evidence and verification to correct logbook data for billfishes. Given the clarity of the results, with substantial numbers of corrections despite rigorous standards (Walsh et al., 2005; 2007), it may now be appropriate to utilize circumstantial evidence in estimation of catches for stock assessments in order to re-evaluate the accuracy of the 1990–1999 logbook data for billfishes. A particularly notable example is that of blue marlin during 1997, when very large numbers were caught. It was the only year when most of the seemingly high billfish catches, identified as possible large positive outliers, were verified as blue marlin and not striped marlin misidentified as blue marlin.

Because the sales records were incomplete, many of the large logbook-reported catches could not be verified. The actual catch may have been considerably greater than the reported total, with the discrepancy caused by discarding and undercounting.

Tuna catches and reporting

Logbook data accuracy has not been investigated in detail for the five tunas taken by this fishery. Four are true tunas (bigeye tuna *Thunnus obesus*, albacore *T. alalunga*, yellowfin tuna *T. albacares*, Pacific bluefin tuna *T. orientalis*); the more primitive skipjack tuna *Katsuwonus pelamis* is also taken. These fishes can be loosely described as common to abundant (bigeye tuna, yellowfin tuna, albacore), uncommon (skipjack), and rare (Pacific bluefin tuna) in the catch of this fishery.

Walsh (2000) compared fishery observer, logbook, and auction sales data for these species from 230 observed fishing trips. The linear regressions of the logbook catches (Y) on the observer-reported catches (X) were:

$Y = 0.070+0.941X$ (bigeye tuna) with a 95% CI for slope of (0.932, 0.950)
$Y = 0.009+0.917X$ (yellowfin tuna) with a 95% CI for slope of (0.908, 0.927)
$Y = 0.015+0.767X$ (skipjack tuna) with a 95% CI for slope of (0.755, 0.779)

The regression coefficients for bigeye and yellowfin tunas approached 1, reflecting generally close agreement between the logbook- and observer-reported catches, but the 95% confidence intervals for the regression slope coefficients did not include 1. The much lower regression coefficient for skipjack reflects its smaller size and lower value compared to the other tunas.

A detailed evaluation of the accuracy of logbook reporting of bigeye tuna catches could be useful for at least two important reasons. Because bigeye tuna is the main target species in this fishery, the total of data entry errors, misidentifications as congeners, and other errors could be perceived as minimum logbook reporting inaccuracy. In addition, it would be interesting to evaluate the incidental catches of bigeye tuna in the shallow-set sector (Walsh and Brodziak, 2015) to determine whether the target species and fishery sector affect logbook data accuracy for a different high-value species.

The accuracy of yellowfin tuna logbook data would be of interest because it was formerly a target species in this fishery, but this has not generally been the case in recent years. If yellowfin tuna data accuracy has decreased, it would support the generalization that data accuracy might be affected by a change in status from a targeted- to an incidentally caught species.

Mixed catches of small bigeye and yellowfin tunas are sometimes taken on trips to tropical waters near the US territories in the Line Islands. It would be useful to check these logbook reports against available sales records because both species are valuable, but juvenile bigeye and yellowfin tuna often look very similar and can be difficult to separate by species (Itano, 1992).

Other pelagic species catches and reporting

Logbook data accuracy has not been investigated in detail for other incidentally caught pelagic species monitored in this fishery (e.g., dolphinfish (mahimahi) *Coryphaena hippurus*, opah, *Lampris guttatus*, and wahoo, *Acanthocybium solandri*). Walsh (2000) compared fishery observer, logbook, and auction sales data for the species listed in the “Other Pelagics” section of the logbook form (originally termed “Miscellaneous”) during 230 observed fishing trips. The rates of exact agreement between logbook- and observer-reported catches of opah and wahoo were 96% and 91%, respectively. These results were not surprising because both species are economically valuable, caught in low numbers, and distinctive in appearance, whereas the exact agreement rate for dolphinfish was 66%, reflecting lower value and possible counting difficulty. Ongoing evaluation of logbook data accuracy for opah and wahoo could be straightforward, particularly since 2000, by using the electronic sales information. Although these species are primarily taken in the deep-set sector, the observer sample sizes are now sufficiently large to identify sources of reporting bias (e.g., non-reporting of discards).

Evaluation of logbook data accuracy for dolphinfish, in contrast, would probably be difficult. Retention rates are likely to vary inversely with catch sizes, and any possibilities of spoilage or use of hold space intended for bigeye tuna during the remainder of the trip would also be expected to increase discarding.

It is possible that additional information could be inferred about logbook accuracy for other incidentally caught pelagic species. The logbook form revision in 1995 included additions of entry positions for oilfish *Ruvettus pretiosus* and pomfrets (Family Bramidae). Since 1995, comparison of logbook data for these species to observer data, auction records, or both should prove informative about captains who do (or do not) identify catches accurately. Moreover, it seems reasonable to expect that captains who report oilfish and pomfrets accurately also report other species taken incidentally and in low numbers (e.g., escolar *Lepidocybium flavobrunneum*, great barracuda *Sphyraena barracuda*) with similar accuracy.

Methodological Overview

Premise underlying commercial longline logbook correction

Logbook data correction at the PIFSC is based upon use of catch and operational data collected by PIROP fishery observers according to standard protocols (Pacific Islands Region Office, 2014). The observer information is considered a research quality database suitable for use as a comparison standard for the logbooks. Inferences about accuracy can be drawn by comparing logbook reports to observer data directly or by comparing logbook reports to predictions generated by statistical models fitted to observer data.

Evaluation of logbook data accuracy

Many aspects of logbook data preparation and analysis are conducted in a hierarchical manner, reflecting the availability of mutually complementary sources of information (i.e., logbooks, observer reports, sales records). In the shallow-set sector, direct comparisons of logbook data and fishery observer reports have been possible since 2004, as a result of 100% observer coverage. Evaluation of the accuracy of logbook data from unobserved fishing trips is much more difficult than direct comparisons of logbooks and fishery observer reports. It is, however, indispensable to development of corrected catch histories for large oceanic pelagic fishes. The premise underlying correction of logbook data from unobserved fishing trips is that catch and operational data collected by the PIROP can be used to fit statistical models to predict catches, thereby serving as "surrogate observers" on unobserved fishing trips (Walsh et al., 2002).

Walsh and Kleiber (2001) fitted regression tree and generalized additive models of blue shark catch rates using several operational variables (e.g., latitude, longitude, SST, date) as the predictors. Walsh et al. (2002) then applied the coefficients from a GAM to an identical suite of predictors in logbook reports and compared the reported to the predicted catches by linear regression, using an objective statistical criterion, the studentized residuals (SR), to identify outliers. However, these checks concentrated on reports of zero catches (i.e., non-reporting) and did not investigate under-reporting (i.e., positive, inaccurate reporting). A similar approach was used in the billfishes study (Walsh et al., 2005; 2007), except that fish auction sales records were also used to verify the corrections applied to logbook data.

In several logbook data studies with blue shark and billfishes (Walsh et al., 2002; 2005; 2007), strong evidence of inaccuracy was required (e.g., two sets per trip with SR>|2| or one set with SR>|3|) to consider data correction. As such, this methodology has clearly been shown to be applicable in evaluation of logbook data accuracy.

The purpose of this section is to ensure that completed and ongoing work with logbook data is fully comprehensible. To that end, preparation of logbook data files in R format is summarized, an integrated analytical structure utilizing the logbooks, fishery observer data, and auction records is described, and uses of statistical models and linear regression techniques in data correction are reviewed. Caveats about various procedures are emphasized.

The information in this section is expected to prove useful for at least three major reasons. First, the review of methodology provides a reference for logbook data use. Second, the logbook data archive primarily includes results from commercial longline sets, but there are also data from experimental fishing trips and others that began or ended outside Hawaiʻi, which should facilitate recognition that catch inputs to stock assessments should include all removals in both the PIROP and logbook archives, including experimental catches and those of far-ranging vessels. In contrast, statistical model fitting procedures are only conducted with commercial longline data reported by the PIROP, and logbook correction is analogously restricted to data from unobserved commercial fishing. Third, the procedures used to calculate catch inputs to stock assessments are not intuitively obvious because the catch totals are sums of catches reported by fishery observers,in logbooks, and possibly estimates. These calculations are explained in full.

This overview covers five topics: preparation of logbook data for analysis in an R data frame format; direct comparisons of observer and logbook data; prediction from a statistical model and regression analyses in logbook data correction; use of public fish auction records in verification of logbook data correction; and archival of past results. As in Section 1, ORACLE schemas and tables and R functions are in boldface to facilitate recognition of technique-related specifics.

<u>Preparation of longline logbook data as an R data frame</u>

The longline logbook data are obtained from three files in the PIFSC ORACLE data base, in the schemas "**OPDT**": **LOG_HEADER**; **LOG_DETAIL**; **LOG_VIEW**. Several important fields used in preparation are described in Table 4; Figure 6 presents a general illustration of logbook data preparation and use. Data preparation is presented in detail in Appendix B, with several aspects described in complete examples. The following comments pertain to those examples. The command sequence to import logbook data from ORACLE (Example A 1) is also used with PIROP data (Example A 1). To work properly, *every command in this import sequence must be exactly correct* (e.g., spaces must precede and follow the asterisk in the **select** command).

Preparation of the logbook data as a flat file (**Logsdata**) to serve as an analytical data frame in which the columns are fields and the rows are observations, is presented in Example A 2 (**DF** denotes “data frame”). The initial step entails using the **LOG_HEADER** file to create the proper dimensions. About 35 fields can provide necessary operational information usable without revision (e.g., haul dates, numbers of hooks deployed, target species, and permit number).

Additional steps require manipulations of these fields to create character variables, such as set identifiers, and factor variables to define levels of predictors in generalized linear models. Catch information is obtained from the **LOG_DETAIL table**, tabulated by sets, and matched to operational information in the data frame. The tabulation is performed by counting the listings of any species per set; i.e., **catch<-tapply(DF$SPECIES_NAME,DF$unique_set_ID,length)**, where **catch** refers to the number of records for kept and released fish of any species per set, **SPECIES_NAME** refers to a species on the logbook form, **unique_set_ID** refers to the unique set identifier used for the aggregation, and **length** is the function that counts records.

The unique set identifier (**unique_set_ID)** is prepared from three fields with the **paste** function: (**paste(LOG_DETAIL$_LAND_YR, LOG_DETAIL$TRIP_NUM, LOG_DETAIL$SERIALNUM**)). Thus, the first set of Trip 1646 in 1991 would be uniquely identified by "**1991 1646 1**". Beginning in 1997, the logbook page serial number (**Logpage**) has become the unique identifier. Therefore, the **SERIALNUM** field has been the logbook page serial number. A second unique identifier (**unique_set_ID1**) is also in the data frame, differing only in use of the original value of **SERIALNUM** rather than the logbook page serial number. Separate linking fields are prepared in both the logbook and observer data frames to permit direct paired comparisons for the observed longline sets. In other words, two fields are required in each data frame corresponding to the years through 1997, and from 1998, when the logbook serial number was adopted as the unique set identifier.

Preparation of the unique set identifiers (i.e., **Observer$Logpage**, **Logsdata$Logpage**) allows matching the corresponding fields from the two data frames. Thus, logbook accuracy for any species would usually be assessed by matching the logbook-reported value for catch on observed sets to the corresponding sets in the observer data frame and then seeking systematic differences to be investigated as sources of bias (e.g., underreporting, misidentifications). It is necessary that the linking fields in both data frames be of the same **mode** (i.e., **character**; **factor**; **numeric**). *If they are not, this procedure will fail*.

The large majority of the fields in the three data frames imported from ORACLE are not needed for logbook data analyses. As a result, learning the characteristics of the fields needed for analyses and associated calculations or revisions (e.g., reducing a character string with the **substr** function) are major challenges in logbook data preparation.

During preparation of the analytical data frame, **summary** (i.e., **summary(DF$variable)**) becomes a very useful function. It returns several descriptive statistics, including the maximum and the number of missing values for numerical variables (i.e. **NA**). Checks on maxima are useful because likely errors (e.g., two tally marks logged as 11 or three as 111) and necessary truncations (i.e., wider ranges of variables in logbooks than in observer data sets) may be revealed. This function can also be usefully applied (**summary (DF)**) in order to prepare a data frame with no missing values or to check that its dimensions are correct.

The **table** function is useful for characterizing the frequencies of character and factor variables; it returns the number of times that some categorical or character variable is present in a data frame. For stock assessment purposes, this is useful when considering set-type effects. It is helpful after invoking the **table** command to use the **unique** command (i.e., **unique(DF$Set_type)**) because the return will state whether **NA** values are present; if so, the number of missing values would be the difference between the total number of sets and the sum of the values returned by **table**.
The **unique** function is useful after selecting a species for inclusion in the data frame. The **SPECIES_NAME** field can be **NA** so a check should be performed by using a subset of the data frame:**DF1<-DF[DF$SPECIES_NAME==*species*&(!(DF$SPECIES_NAME=="NA")),]**. In words, this means that a data frame (**DF**1) selected from the original (**DF**) will include all records for a particular ***species***, but it will exclude records missing the **SPECIES_NAME** value because the exclamation point represents negation.

After preparing the analytical data frame, it is often convenient to prepare new data frames for analyses known or expected to be required. For example, data for a deep-dwelling species could be selected as **Logs_DS<-Logsdata[Logsdata$Set_type=="D",]**, where **Logs_DS** would refer to a data frame containing deep-set sector logbook data, **Logsdata** would refer to the entire analytical logbook data frame, and **Set_type=="D"** specifies the deep-set sector as the selection criterion.

Information pertaining to the SST matcher program (see Section 1) at the PIFSC and the necessary preparations are presented in Example A 3. Interested persons are referred to R. Price of the PIFSC IT and M. Abecassis of the PIFSC CoastWatch groups, respectively.
The remaining aspect of logbook data preparation entails several truncations (Example A 4). These steps limit the application data to particular types of effort in particular locales (i.e., Hawai'i-based commercial longline fishing) and result in the elimination of large numbers of missing values from various fields.

<u>Comparisons of observer and logbook data</u>
Direct comparisons of logbook and PIROP observer data are conceptually straightforward. Since 1998, the simplest method is to match the logbook data to the observer data frame using the identical unique set identifier present in both data frames. The corresponding observer and logbook values can then be plotted and any calculation(s) performed easily. For example, the sequence **Observer1<-Observer[Observer$*species* = Observer$*species*.log + 10,]** could be used to check for discrepancies between the two data sources. The sequence selects longline sets in the observer data frame (**Observer**) with an observed catch (**Observer$*species***) equal to the logbook catch plus 10 (**Observer$*species*.log***)* and saves these in a new data frame **Observer1**.
The data matching process for earlier years (i.e., 1994-1997) is more complicated. The fields available for use require careful revisions. Specifically, the logbook data (**LOGHDR**) include fields named **OBSTRIPNUM** and **OBSSETNUM.** These fields would be combined by using the **paste** function with no separator (i.e., **Logsdata$Link_95_97**<-
paste(**Logsdata**$**OBSTRIPNUM**, **Logsdata**$**OBSSETNUM, sep=""**). This new field could then be redefined as **NA** for the other years with the **ifelse** function (i.e.,

Logsdata$Link_95_97**<- **ifelse**(**Logsdata$Haulyr<**1995 | Logsdata$Haulyr**>**1997**, ”**NA**”,**Logsdata$Link_95_97)** to restrict matching to this period.

Data on the number of caught and released fishes in the shallow-set sector since 2004 can be obtained directly from the observer-reported values. Bias in logbook data associated with non-reporting of released fishes in the logbooks can be estimated directly by subtracting the logbook values from the corresponding observer values. In the deep-set sector, data accuracy evaluation for the unobserved sets (*ca*. 80%) is conducted by comparing logbook-reported catches to predictions from a statistical model fitted to PIROP observer data (Walsh et al., 2002; 2005; 2007).

Use of statistical models to predict catches

The coefficients from a statistical model can be applied to logbook data using the R **predict** function to serve as a comparison standard for unobserved fishing trips (Walsh and Kleiber, 2001; Walsh et al., 2002). *It is critically important at this step that the ranges of the predictors in the data set used to fit the model should equal or preferably exceed those in the application data set*. For this reason, some truncations of the logbook data are usually required. Sets with missing covariate values are also deleted. In practice, the truncations and deletions usually comprise a small fraction of the logbook data (*ca*. 2%). Catch data from these sets are accepted as accurate. In practice, the truncations should not approach the maxima; e.g., if the latitudinal range in the observer data (i.e., model-fitting set) was 0°-42°N, the logbook data might be truncated to 2°N-40°N. For billfishes, particularly the tropical blue marlin, SST exerts strong effects on catch rates (Walsh and Brodziak 2015; Walsh et al. 2005; 2007); therefore, if the SST range in the observer data was 17°-32°C, the logbook data might be truncated to 18°-30°C. Because SST and latitude are inversely related, both low latitude and high SST values would be associated with relatively high blue marlin catch rates. *The reason for careful truncations is that applying model coefficients to logbook data that include covariate values beyond the ranges of the model fitting data can cause very large prediction errors when predictions are back-transformed*.

An example of the use of a statistical model for logbook data correction (Appendix B: Example B 1) is presented, based upon fitting a zero-inflated negative binomial model (ZINB), which requires the R libraries **pscl** and **MASS**, to the PIROP observer data from 1995-2014. The ZINB is appropriate for data reported as counts, with more zeros than expected under the Poisson or negative binomial distributions and with overdispersion in the positive counts (Zuur et al., 2012; Brodziak and Walsh, 2013).

An important point regarding the models used to correct logbook data (Walsh et al., 2002; 2005; 2007) is that they are less complex than other models recently used to standardize CPUE of highly migratory fishes (Brodziak and Walsh, 2013; Walsh et al., 2015; Carvalho et al., 2016). This is because the logbook data do not include several operational parameters that are in the observer data. For example, the best fitting model for blue marlin (Walsh and Brodziak, 2015) computed with PIROP observer data included six factors, three continuous variables, and four interactions, but two factors and an interaction cannot be estimated from the logbook data. For this reason, a ZINB with four factors, two continuous variables, and two interactions in its negative binomial model and two continuous variables in the zero-inflation model was fitted for data evaluation. In addition, the historical changes in the operational details on the logbook form necessitate more imputation of operational variables than is required with observer data.

The sample sizes in the logbook data set are much greater than the observer data set so the ranges of operational variables are also generally greater. This disparity represents a second constraint on use of a statistical model for data correction because the procedures require prediction, and doing so outside the ranges in the model-fitting data set can cause very large errors.

The fitting methods followed Section 1; the analysis of deviance and other summary details are presented in Appendix II, Example B 1. The example also shows the effect of the presence of bias (i.e., systematic misidentifications) in the model fitting data.

The basic syntax used is **predict**(**Obsr_model**, **type=”response**”, **newdata=”Logsdata”**), where **Obsr_model** is the GLM (or GAM) object fitted to observer data, **type=”response”** is the call for the back-transform to the original units, and **newdata=”Logsdata”** specifies use of the data frame for prediction.

Applying the coefficients from a ZINB or other statistical model to the logbook data generates a vector of predicted catches. For convenience, this vector can be assigned to the data frame of interest (e.g., **Logsdata$predicted**). This facilitates regression analysis of the reported on predicted values, which entails computing the linear regression of the reported on the predicted values using the **lm** function: **lm**(**log**(**Logsdata$reported**)<-**log**(**Logsdata$predicted**)). The studentized residuals (Draper and Smith 1998) are then obtained by dividing each residual by its standard error.

The subset of observations with large SR can then be selected with **Large_SR<-Logsdata[Logsdata$SR<=-2 | Logsdata$SR>=2,**] where **Large_SR** would be a new data frame containing all longline sets with |SR|≥2. Because the large SR data are selected in approximation to the *t*-distribution, |SR| values of 2.0, 2.6, and 3.3 would correspond roughly to the 0.05, 0.01, and 0.001 probability values.

After obtaining the large SR data, it is useful to obtain a tabulation because logbook data bias tends to be concentrated among a few vessels or captains, although the sources of bias vary among species. The command **table(Large_SR$Permit**) would return a table from the **Large_SR** data frame showing the number of records per vessel. A vessel with 100 or more sets with large SR might be a source of substantial inaccuracy in the catch data. A more detailed and meaningful tabulation would take into account movements within the fleet during captains' careers by using **table(Large_SR$Permit, Large_SR$CML**). This command would return the number of large SR for each vessel by the individual captains.

Use of public fish auction sales records in logbook data correction

The premise underlying use of public fish auction data to evaluate logbook data accuracy is that the auction species identifications are definitive. In practice, this means that if the catch data from particular sets of some fishing trip have been identified as potential outliers by an objective statistical criterion, these data from the trip could be evaluated and corrected, if necessary, using the sales records. If, for example, logbook reports listed 10 blue marlin caught and kept, but sales records from the trip included only five striped marlin, then that sales total indicates that at least 5 striped marlin were misidentified as blue marlin in the logbook report. If the sales total was equal (or close) to the trip catch total, it would support changing all of the logbook-reported blue marlin to striped marlin.

The principal complexity of the logbook data correction process involves the distinction between the set-level catch data and the trip-level sales records. Specifically, the large SR from the regression analyses are used to identify longline sets with questionable catch data, but it may still be difficult to assign corrections even with support from sales records. This is particularly true during periods of high relative abundance of one or more species. For example, if a winter longline trip yielded 100 or more striped marlin, all of which are reported in the logbook as blue marlin, the sales records verified that only striped marlin had been sold for the trip, and all of the logbook sets with blue marlin catches had large SR, the correction could still be difficult because many fish might go unsold as a result of high seasonal abundance. Identification of the specific sets that require correction and allocation of the misidentified fish among them from trip-level sales records would represent the principal difficulty in this process.

This problem is compounded if multiple species are logged. Specifically, striped marlin and shortbill spearfish often go unsold when abundant. If a longline trip in winter yielded 50 striped marlin and 50 shortbill spearfish, the logbooks might list 33 blue marlin, 33 striped marlin and 33 shortbill spearfish. In this scenario, sales records might not be helpful because it would not be unusual for about half the striped marlin and shortbill spearfish to be unsold. This means that the number of blue marlin reported in the logbooks would not be contradicted by the sales records, which would preclude correction in the logbook database.

Corrected catches were allocated among sets by comparing the large SR to the values of the *t*-distribution. Thus, if the sales records indicate that 25 fish should be corrected on two sets with SR of 2.0 and 2.6, respectively, the corrections might allocate 5 fish to the first and 20 to the second set because these SR absolute values approximate the 0.05 and 0.01 probability levels of the *t*-distribution.

Figure 5 illustrates the positive effects of the availability of complete sales information. The much smaller standard deviations for blue and striped marlins in 2003 (Fig. 5c), compared to 1995 (Fig. 5a) or 1998 (Fig. 5b), reflect the feasibility of checking all apparent outliers. In contrast, much of the patchiness remaining in Figures 5a and 5b reflects lack of sales records for verification purposes.

Simple graphical presentations of logbook and auction data can also be used to identify species misidentifications. Figure 7 presents auction sales totals for blue and striped marlins from 25 longline trips plotted against the corresponding trip totals for retained fish. The diagonal trend (open black circles) represents corrected numbers of sold striped marlin plotted against numbers of retained blue marlin as reported in the logbooks; striped marlin sales totals were significantly correlated (r=0.919, df = 23, $P<<<10^4$) with the numbers of blue marlin reported in the logbooks. In contrast, the corrected blue marlin sales values were mostly zeros or other small numbers. The solid blue circles along the *x*-axis represent large numbers of blue marlin logged, but few if any sold, and the solid black circles along the *y*-axis represent few striped marlin logged, but large numbers sold. Thus, the auction records can be used graphically for exploratory purposes with the logbook data.

Estimation of a corrected catch history from the PIFSC archives

The corrected catches of blue shark (Walsh et al., 2002) and billfishes (Walsh et al., 2005; 2007) were estimated differently. In the former case, a statistical model fitted to observer data was used to predict catches; these values replaced data identified as biased by a multi-stage data evaluation process (Walsh et al. 2002). In the case of billfishes, sales records were extensively used to make the corrections as exact as possible.

The blue shark annual catch totals (or other period) were estimated as the sum of blue sharks reported by fishery observers (O), those reported in logbooks that were accepted as accurate and did not undergo correction (LB), and the model-predicted catches (M) that replaced data judged to have been inaccurate (Walsh and Kleiber, 2001) (i.e., Catch=O+LB+M). Prediction intervals were obtained by bootstrapping (Walsh et al., 2002).

Billfish catch totals (per year or other period) were estimated similarly, except that catch corrections relied upon verification by sales records. Important considerations include the fact that catches of blue marlin and other billfishes are much smaller than blue shark catches, the data checks were not restricted to zeros, and sales records were used to check the accuracy and correct the logbook catch data to the greatest extent possible. Thus, the corrected catch amount (Corrected) is a fraction of the catch total that is associated with sets identified as potential outliers by their SR and corrected on the basis of sales records rather than on model predictions as in blue shark (i.e., Catch = O + LB + Corrected). The corrected sets comprised about 5% of the catch data for both blue shark and the billfishes.

It must be emphasized that billfish catch data were not corrected in the logbook database unless direct evidence from an observer or sales records was available for verification. This means that some sets before 2000 that were very large outliers remain in the database (e.g., logbook reports of multiple blue marlin north of the Main Hawaiian Islands in winter), although the objective statistical criterion (SR) used to identify possible outliers may have indicated strongly that certain of the unobserved trips or those without sales records were inaccurate.

An important consideration regarding use of the billfishes corrected catch data stems from archival as a separate table (**WALSH_MARLIN**S) in the ORACLE database. If stock assessments or other analyses are conducted with the corrected data, the results will differ from the uncorrected logbook data. If the catch inputs to stock assessments are not computed consistently, it would be expected to affect results.

Archival of results

Analytical work related to logbook data correction and accuracy has been archived in the ONAGA system at the PIFSC (R. Price: system administrator). Appropriate backups with full results have been prepared for all major projects that led to publications. Identical output has been submitted to the ORACLE database.

The detailed set-level billfish corrections are archived in the ORACLE data base. The sequence is **Marlins_Corr<-dbGetQuery(con,"select from * from ORADATA.WALSH_MARLIN")** where **Marlins_Corr** would be an R data frame available for use.

Analytical work was conducted in S+ until *ca*. 2010. The directories (in the Walsh home directory) have self-explanatory names; e.g. **Walsh_Sharks_09** would refer to a shark project conducted during 2009. Text files in most of these directories are available to write out the working data; the exceptions would be directories for projects conducted concurrently as another project, using the same data set. To examine work from recent years, the text files can be imported directly into R using **read.table**("***text.file***",**sep=" "**) , where **read.table** imports the file as a data frame, and **sep** indicates that the separator is white space.

Section III. Conclusions

The preceding chapters and associated appendices describe the preparation and use of data from fishery observer reports and commercial logbooks in the Hawaii-based pelagic longline fishery. This detailed information and practical knowledge is intended to support the ongoing scientific work of the SAP to conduct pelagic stock assessments and research on the Hawaii longline fishery. In this final section, we present some conclusions and suggested improvements for technical work with the longline fishery data. One conclusion is that most practical research opportunities would entail retrospective investigations of the logbook data. Such efforts, in turn, would benefit from a thorough characterization of the multispecies catches and environmental conditions observed in each unique longline set, including catch history corrections as needed. It is important to develop a systematic approach to maximize the information content of the logbook database and correct data errors to the extent practicable. When considered from the perspective of ecosystem-based management, the capacity to readily analyze the multispecies longline catch data will be important for monitoring community dynamics, species abundances and potential interactions, and pelagic ecosystem processes.

Another conclusion based on previous work on bycatch and incidentally caught species (Walsh, et al. 2002; 2005; 2007; 2009) is that the necessary information is also available to assess the relationships between catches of bycatch and target species in the Hawaii longline fishery. This would be of practical interest for fishery management and also would be of interest from an ecological perspective. Increased use of the PIROP multispecies longline catch and fishing operations data can directly improve understanding and capacity to predict fishery-related interactions such as handling mortality and discarding.

In terms of suggested improvements, the primary one is to create a set of unique identifiers for each longline gear set during the entire history of the Hawaii longline fishery. Having a set of unique identifiers will remove the need for the data processing steps described in Chapter 2 and Appendix A . These steps involve identifying and matching longline sets by combining several character variables. This approach is complicated and may lead to database matching errors if done manually, without a predefined script or other software. In general, the set of unique identifiers for each longline set would be implemented in the components of both the observer and logbook databases.

The secondary suggestions for improvement are also straightforward. One suggestion is to create R language scripts for all data preparation and analytical procedures that can be efficiently encoded, tested, and verified. Finalize data processing and analytical scripts could be archived in libraries on appropriate PIFSC server. Another suggestion is to standardize the naming of fields for data preparation and analyses. This would improve documentation, reproducibility, and maintainability of data processing and analytical approaches. For example, the names assigned to variables in the ORACLE schema, **schema=ORADATA:WALSH_MARLIN**, (K.L. Sender, PIFSC) provide one possible naming convention. Another suggestion is to standardize the statistical procedures used to evaluate logbook data accuracy. The choice of the outlier prediction model and the objective criteria used to identify possible outliers can be evaluated and likely improved through simulation testing.

ACKNOWLEDGMENTS

We extend our greatest thanks to Pierre Kleiber for his long-term interest in PIROP data use at the PIFSC and for developing many of the logbook and observer data use protocols still followed with only minor changes. We also gratefully acknowledge encouragement and cooperation over many years from PIROP colleagues, especially S. Joseph Arceneaux, Kevin Busscher, and John Kelly, and from Charles Karnella of the International Fisheries Division of the Pacific Islands Region Office. We also extend our collective thanks to the FRMD and the IT staff of the PIFSC for assistance on many occasions. Special thanks are due to Brent Miyamoto for many searches of HDAR sales records, Diosdado Gonzales for discussions of and assistance with sales records, and Lucas Moxey, Russell Price, Richard Uyeda, and Christopher Tokita for computing assistance while preparing this document.

REFERENCES

Brodziak, J., and M. Dreyfus.
2011. Report of the seminar on the use of the best available scientific information. International Scientific Committee for Tuna and Tuna-Like Species in the North Pacific, ISC/11/Annex 11, p. Available at: http://isc.fra.go.jp/pdf/ISC11/Annex_11_ISC11_Seminar_FINAL.pdf

Brodziak, J. K. T., and W. A. Walsh.
2013. Standardizing catch rates of bycatch species using multi-model inference: A case study of oceanic whitetip shark in the Hawaii longline fishery. Canadian Journal of Fisheries and Aquatic Sciences 70:1723−1740.

Bonfil, R., S. Clarke, and H. Nakano.
2008. The biology and ecology of the oceanic whitetip shark, Carcharhinus longimanus. Pages 128–139 in M. D. Camhi, E. K. Pikitch, and E. A. Babcock. editors. Sharks of the Open Ocean: Biology, Fisheries and Conservation. Blackwell Scientific Publications, Oxford, UK.

Carvalho, F., G. DiNardo, and M. McCracken.
2014. Overview of catch, size, and catch rate data for shortfin mako shark Isurus oxyrinchus from the Hawaii-based pelagic longline fishery: 1995-2012. ISC/15/SHARKWG-2/06.

Carvalho, F., W. A. Walsh, and Y.-J. Chang.
2016. Standardized catch rates of blue marlin (*Makaira nigricans*) in the Hawaii-based longline fishery (1995-2014). ISC/15/BILLWG-1/05.

Compagno, L. J. V.
1988. Sharks of the Order Carcharhiniformes. The Blackburn Press. Caldwell, N.J.

Crawley, M. J.
2013. The R Book. Second edition. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, United Kingdom.

Department of Commerce.
2004. Fisheries off West Coast states and in the western Pacific; highly migratory species fisheries. National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Federal Register 69(67): 50 CFR Parts 223, 224, and 660.

Dollar, R. A., and S. S. Yoshimoto.
1991. The Federally mandated longline fishing log collection system in the western Pacific. Southwest Fisheries Science Center Administrative Report H-91-12, 35 p.

Draper, N. R. and H. Smith.
1998. Applied Regression Analysis. Third edition. John Wiley and Sons, Inc. New York.

Fry, F. E. J.
1971. The effect of environmental factors on the physiology of fish. In:Hoar, W.S., Randall, D.J. (Eds.), Fish physiology. Vol. VI. Environmental Relations and Behavior. Academic Press, New York, NY, pp. 1–98.

Itano, D.
1992. Are you sure that fish was a yellowfin? SPC Fisheries Newsletter #62 – Jul/Sep '92. South Pacific Commission, Noumea, New Caledonia. p. 29-32.

Langseth, B.
2015. Standardization of striped marlin (*Kajikia audax*) CPUE in the Hawaiian longline fishery II: Additional covariates, distribution, and use of data from the deep-set fishery sector only. ISC/15/BILLWG/1/06.

Maunder, M. N., and A. E. Punt.
2004. Standardizing catch and effort data: a review of recent approaches. Fisheries Research 70: 141–159.

McCracken, M. L.
2005. Modeling a very rare event to estimate sea turtle bycatch: lessons learned. U.S. Dept. of Commerce, NOAA Technical Memorandum NOAA-TM-NMFS-PIFSC-3, 25 pp.

Nakamura, I.
2001. Istiophoridae: billfishes. In: Carpenter, K.E., Niem, V.H. (Eds.), The Living Marine Resources of the Western Central Pacific. Volume 6. Bony Fishes Part 4 (Labridae to Latimeriidae), Estuarine Crocodiles, Sea Turtles, Sea Snakes and Marine Mammals. FAO, Rome.

Pacific Islands Region Office.
2014. Hawaii Longline Observer Program Observer Field Manual. Version LM.14.04. National Oceanic and Atmospheric Administration, Pacific Islands Region, Honolulu, Hawaiʻi.

Walsh, W. A.
2000. Comparisons of fish catches reported by fishery observers and in logbooks of Hawaii-based commercial longline vessels. Southwest Fisheries Science Center Administrative Report H-00-07. 45 pp.

Walsh, W. A. and J. Brodziak.
2014. Catch rate standardization for swordfish *Xiphias gladius* in the shallow-set sector of the Hawaii longline fishery, 1995−2012. ISC/14/BILLWG-1/05.

Walsh, W. A., and J. Brodziak.
2015. Billfish CPUE standardization in the Hawaii longline fishery: Model selection and multimodel inference. Fisheries Research 166:151–162

Walsh, W. A., and Y.-J. Chang.
2015. Standardization of striped marlin *Kajikia audax* CPUE for the Hawaii-based longline fishery during 1995–2013 using generalized linear models: An update from 2011. ISC/15/BILLWG-1/03.

Walsh, W. A., and H.-H. Lee.
2011. Standardization of striped marlin, *Kajikia audax*, CPUE with generalized linear models fitted to pelagic longline observer data from the Hawaii-based fishery: 1995-2009. ISC/11/BILLWG/1/03.

Walsh, W. A., and R. Y. Ito.
2011. A long-term corrected catch history for striped marlin, *Kajikia audax*, in Hawai'ian waters. ISC/11/BILLWG/1/08.

Walsh, W. A., K. A. Bigelow, and R. Y. Ito.
2007. Corrected catch histories and logbook accuracy for billfishes (Istiophoridae) in the Hawaii-based longline fishery. U.S. Department of Commerce. NOAA Technical Memorandum NMFS-PIFSC-13. 39 pp.

Walsh, W. A., K. A. Bigelow, and K. L. Sender.
2009. Decreases in shark catches and mortality in the Hawaii-based longline fishery as documented by fishery observers. Marine and Coastal Fisheries: Dynamics, Management, and Ecosystem Science 1:270–282.

Walsh, W.A., R.Y. Ito, K.E. Kawamoto and M. McCracken.
2005. Analysis of logbook accuracy for blue marlin (*Makaira nigricans*) in the Hawaii-based longline fishery with a generalized additive model and commercial sales data. Fisheries Research 75:175–192.

Walsh, W.A., P. Kleiber and M. McCracken.
2002. Comparison of logbook reports of incidental blue shark catch rates by Hawaii-based longline vessels to fishery observer data by application of a generalized additive model. Fisheries Research 58:79–94.

Zuur, A., Ieno, E.N., Walker, N., Saveliev, A.A., Smith, G.M.
2009. Mixed Effects Models an Extensions in Ecology with R. Springer Verlag, New York.

Zuur, A.F., Saveliev, A.A., Ieno, E.N.
2012. Zero Inflated Models and Generalized Linear Mixed Models with R. Highland Statistics Ltd. Newburgh, United Kingdom.

FIGURES

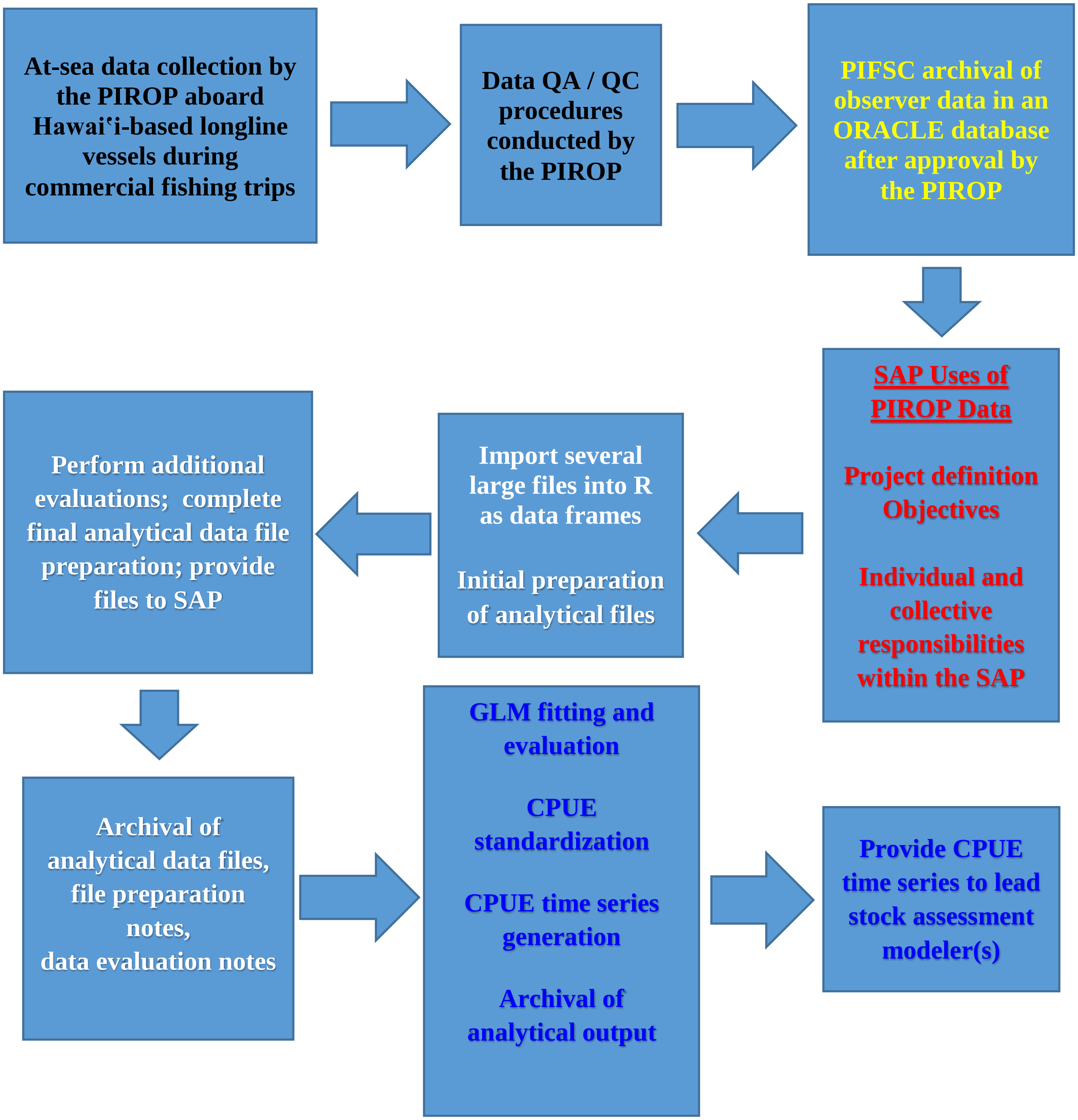


Figure 1.-- Diagrammatic representation of PIROP data use by the SAP. At-sea data collection and onshore data evaluation steps performed by the PIROP (**black**) are followed by PIFSC archival (yellow). The SAP uses PIROP data from the outset of projects, when objectives are defined and responsibilities assigned (**red**). Preparation, additional evaluation, and archival of

information pertaining to the data and the analytical files used by all SAP members (white), and completion of a CPUE analysis (**deep blue**) as a stock assessment input.

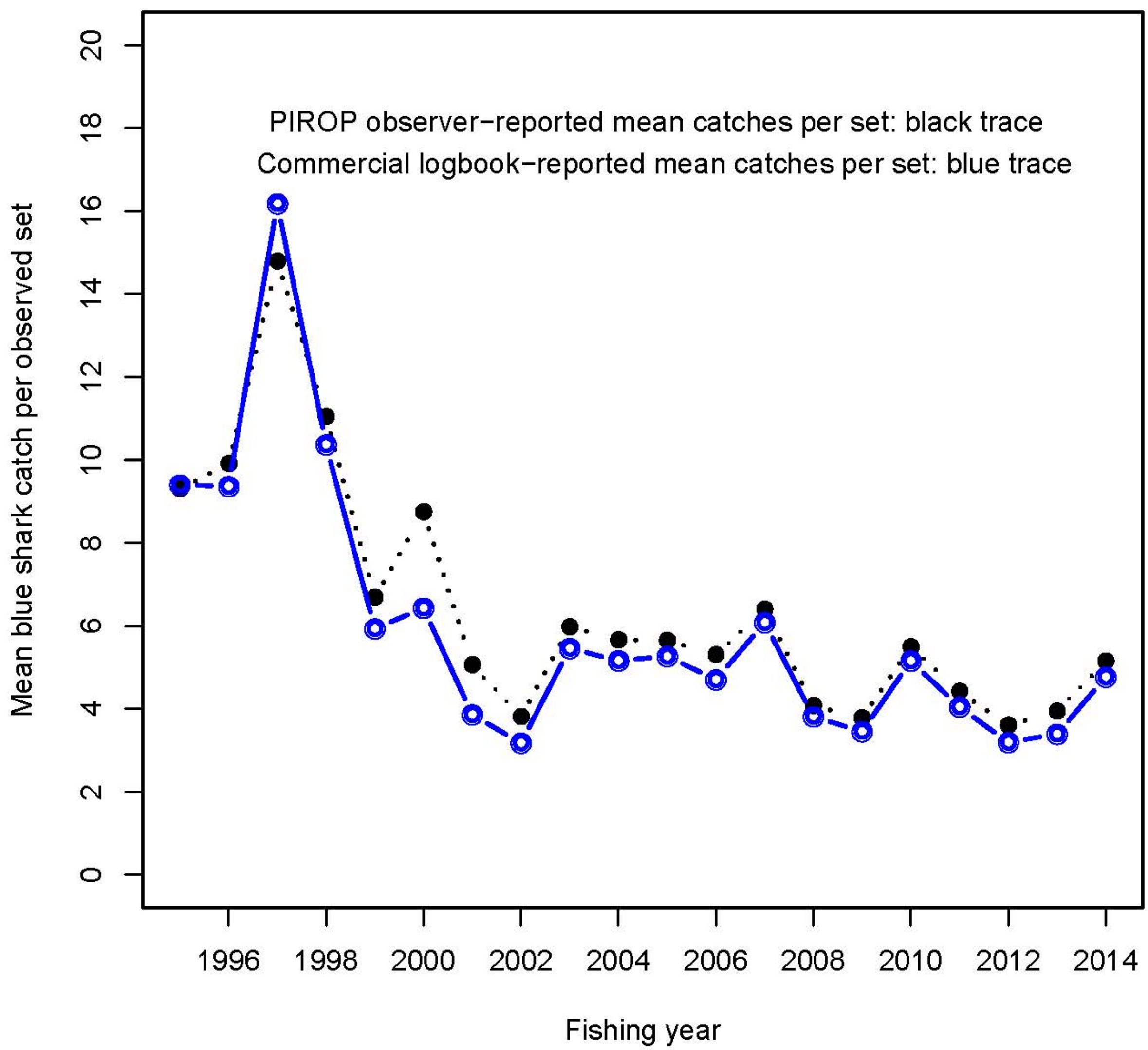


Figure 2.--Comparison of annual mean catches of blue shark *Prionace glauca* per observed longline set during 1995-2014. The two traces are data reported by PIROP observers (dotted black line) and in commercial logbooks on observed trips (solid blue line).

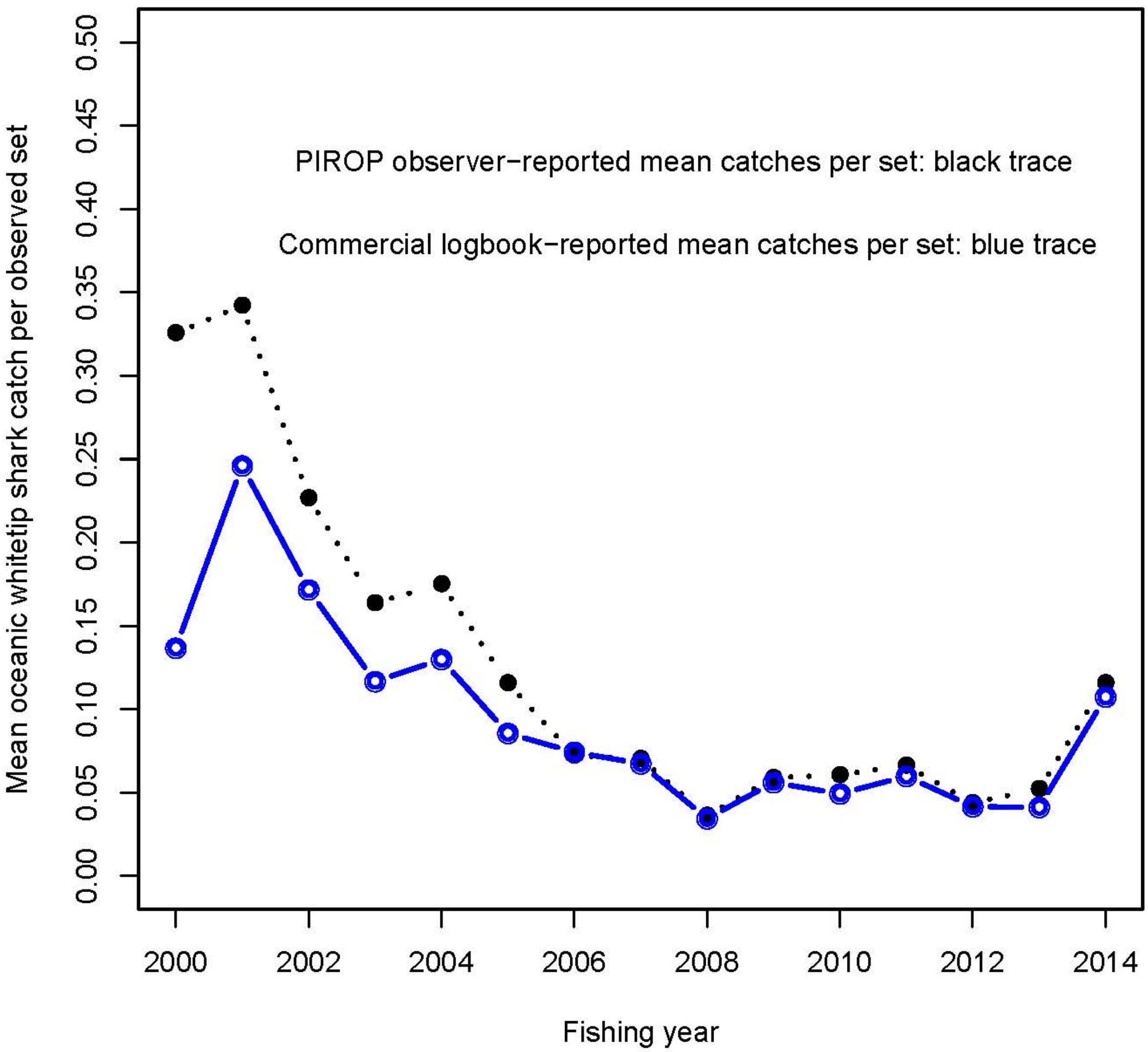
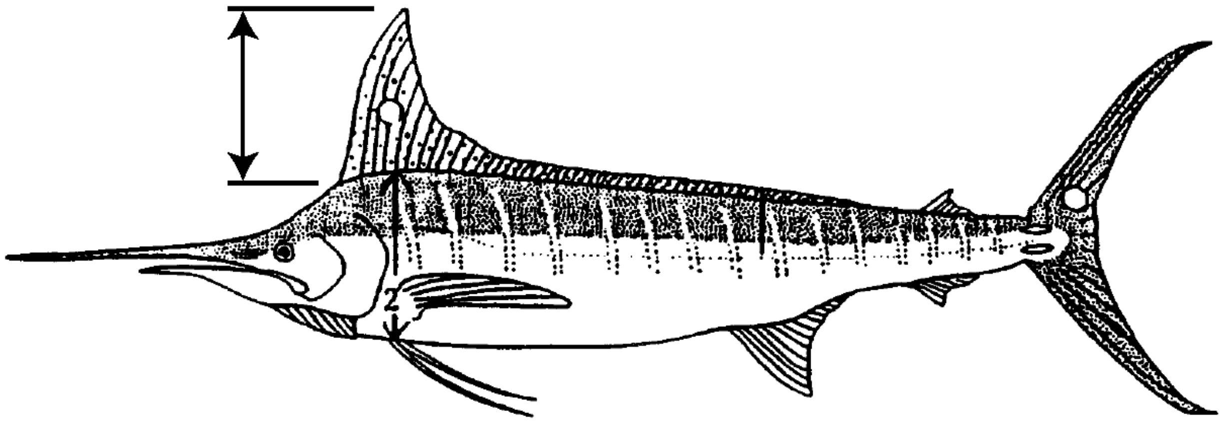


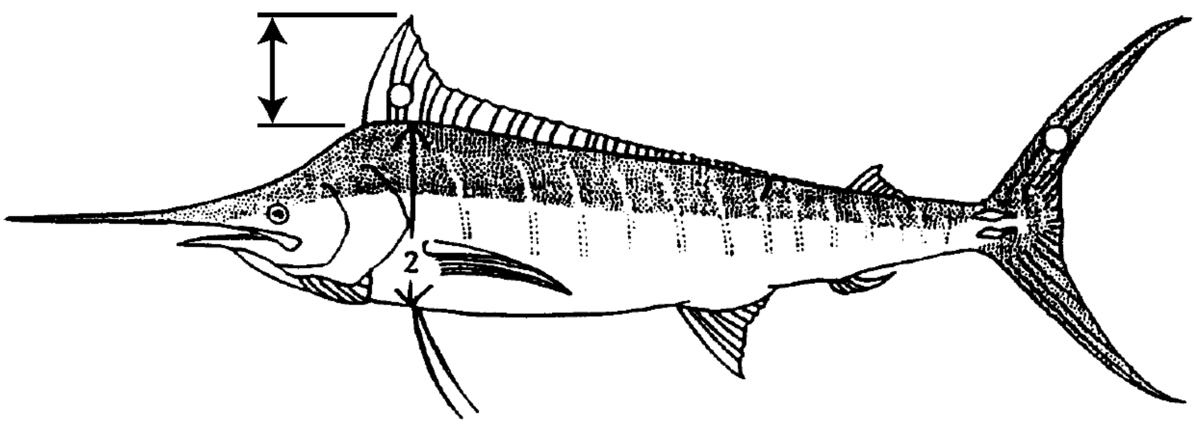
Figure 3.--Comparison of annual mean catches of oceanic whitetip shark *Carcharhinus longimanus* per observed longline set during 2000-2014. The two traces are data reported by PIROP observers (dotted black line) and in commercial logbooks on observed trips (solid blue line).

(a) Common in this fishery

Striped marlin *Kajikia audax*



Blue marlin *Makaira nigricans*



Shortbill spearfish *Tetrapturus angustirostris*

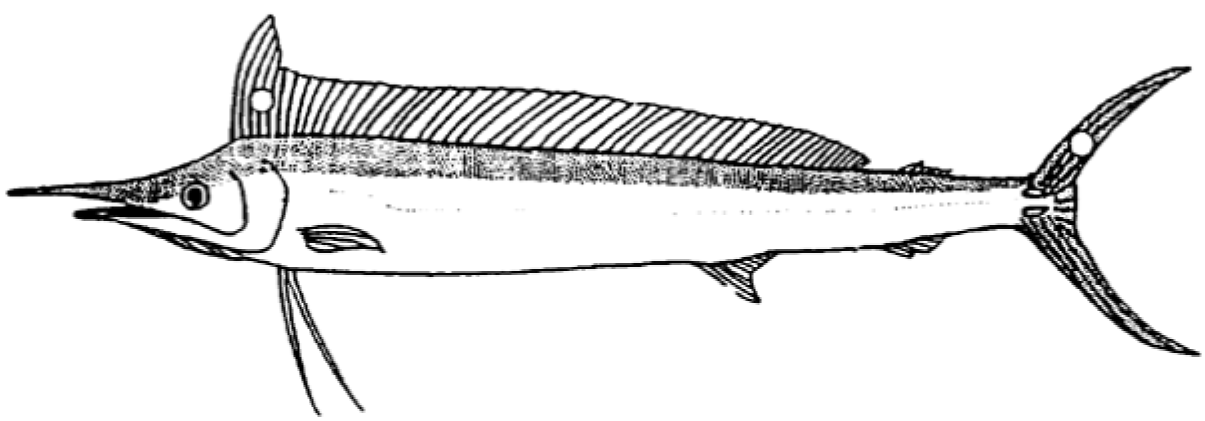
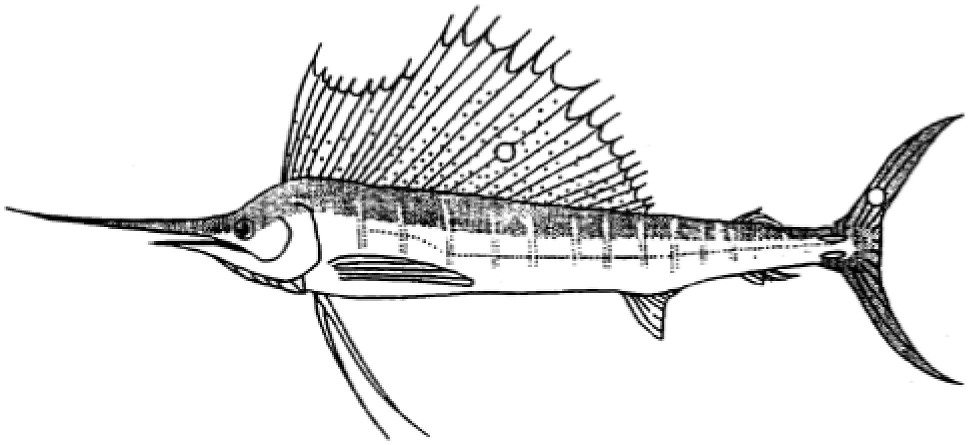


Figure 4.--Illustrations of the istiophorid billfishes taken as incidental catches by the Hawai'ian longline fishery. Useful distinguishing characteristics are shown.

(b) Uncommon or rare in this fishery

Sailfish *Istiophorus platypterus* (uncommon)



Black marlin *Istiompax indica* (rare)

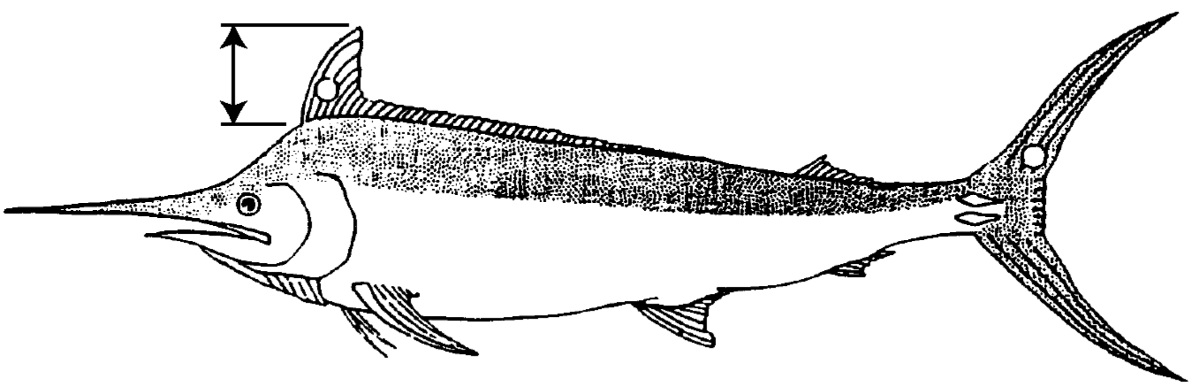


Figure 4 (cont'd).--Illustrations of the istiophorid billfishes taken as incidental catches by the Hawai'ian longline fishery. Useful distinguishing characteristics are shown.

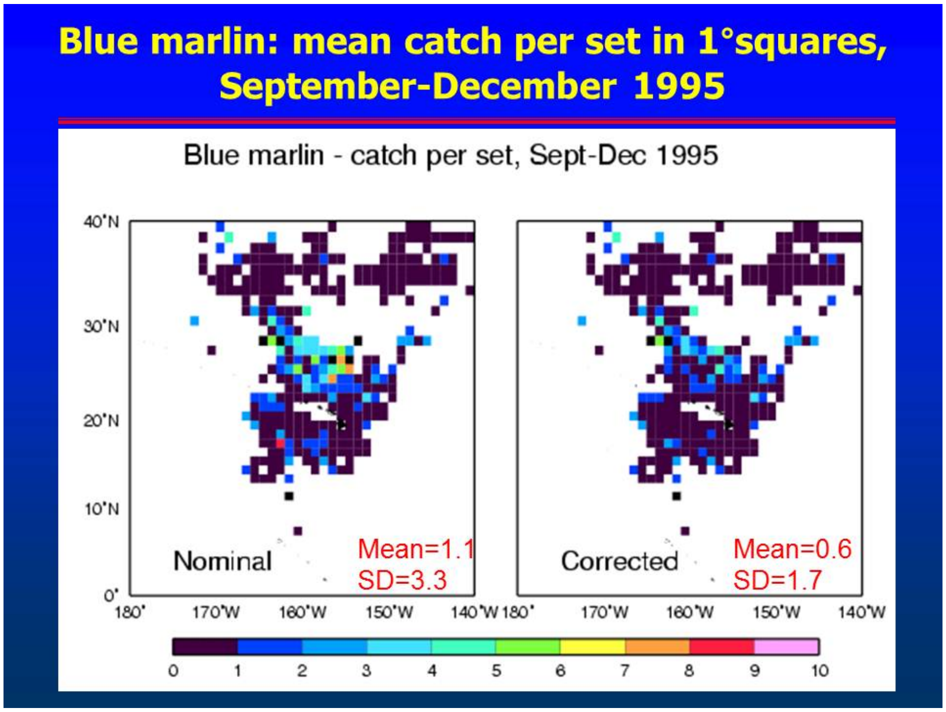


Figure 5a.--Illustration of the effects of logbook data correction with billfishes. Misidentifications during September–December 1995 affected the means and standard deviations for both species.

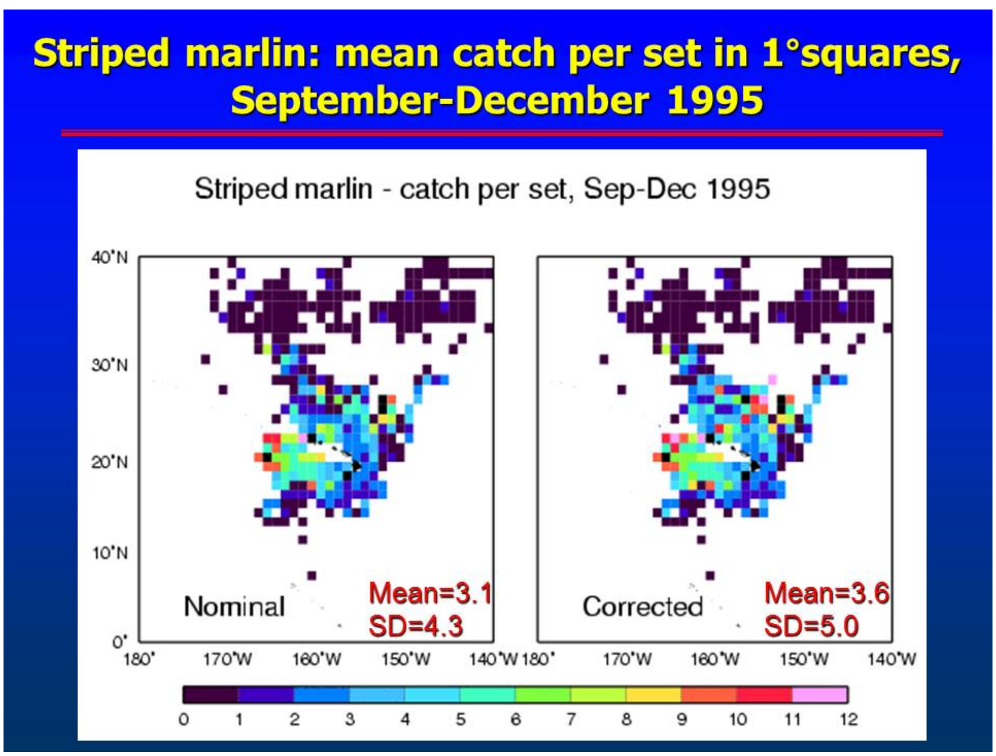
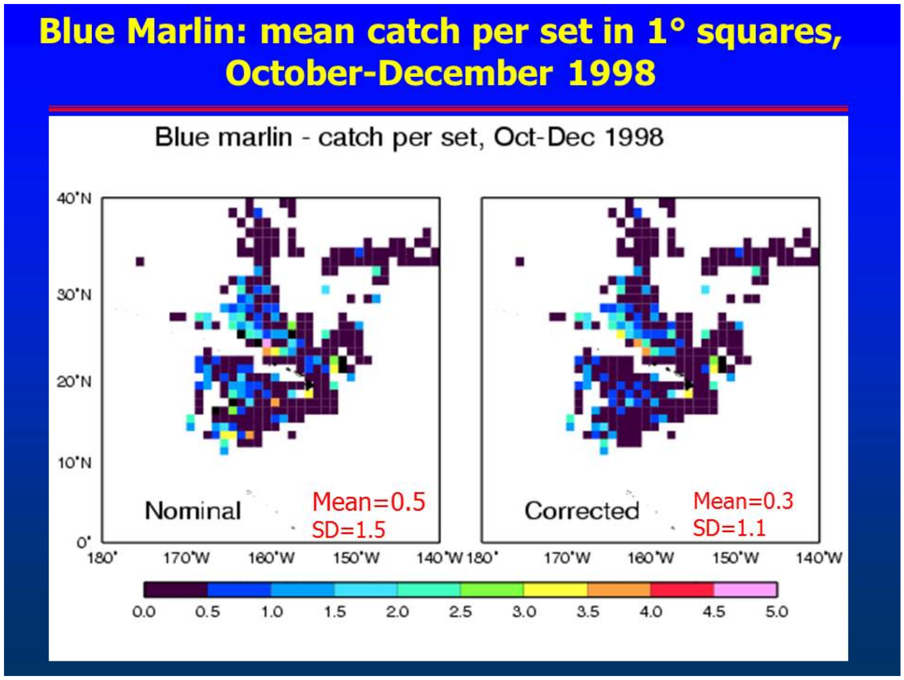


Figure 5b.--Apparent misidentifications of blue marlin during October–December 1998 affected the means and standard deviations for three species.



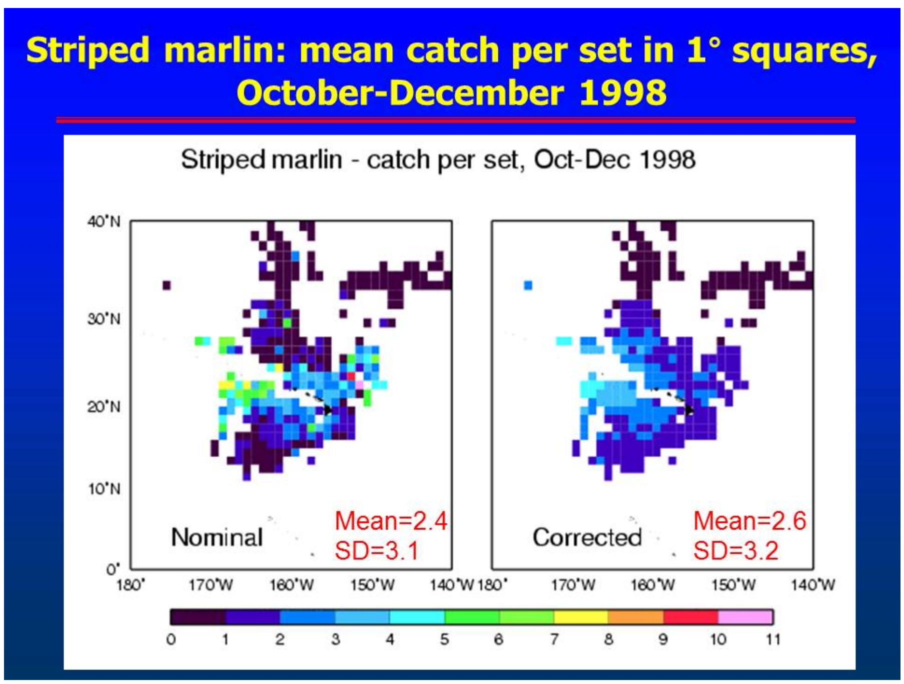


Figure 5b (cont'd).--Apparent misidentifications of blue marlin during October–December 1998 affected the means and standard deviations for three species.

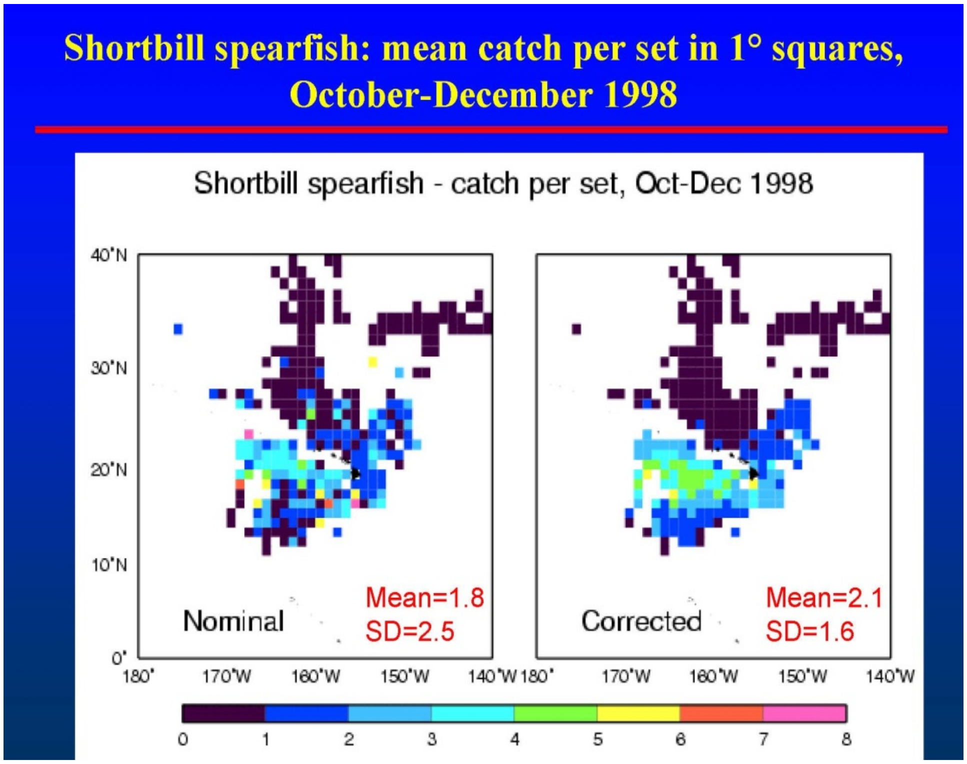
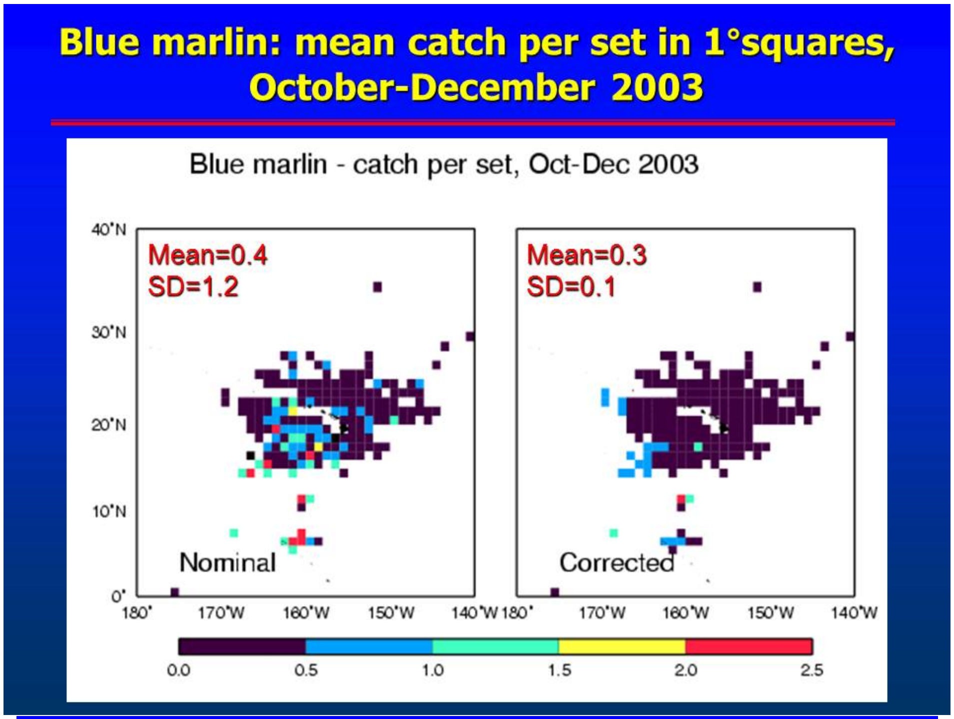


Figure 5b (cont'd).--Apparent misidentifications of blue marlin during October–December 1998 affected the means and standard deviations for three species.



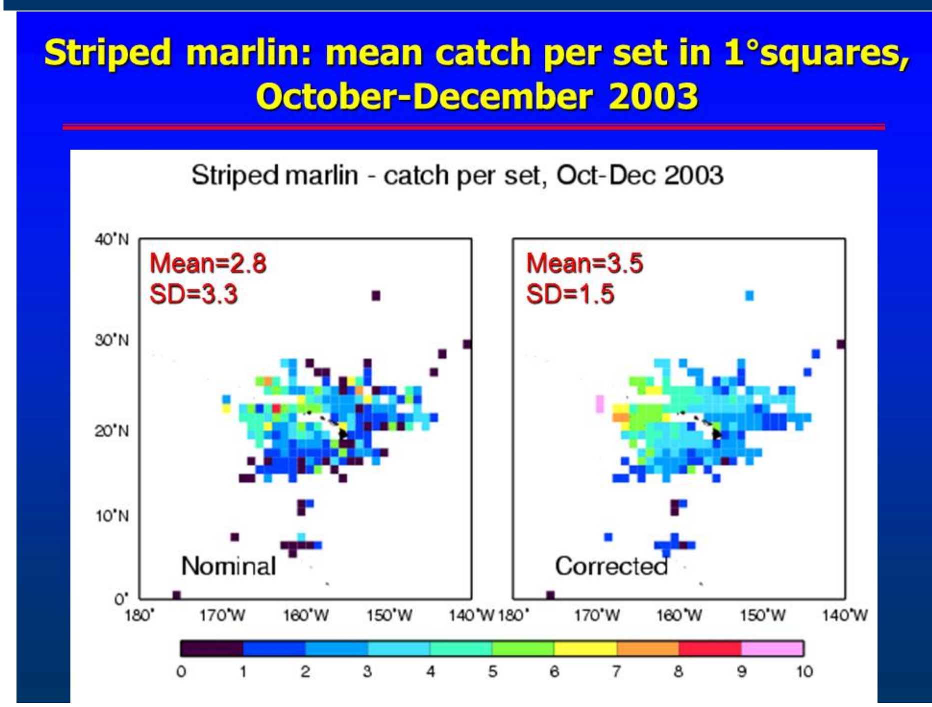


Figure 5c.--Apparent misidentifications of blue marlin during October–December 2003 inflated the blue marlin standard deviation and affected the striped marlin mean and standard deviation.

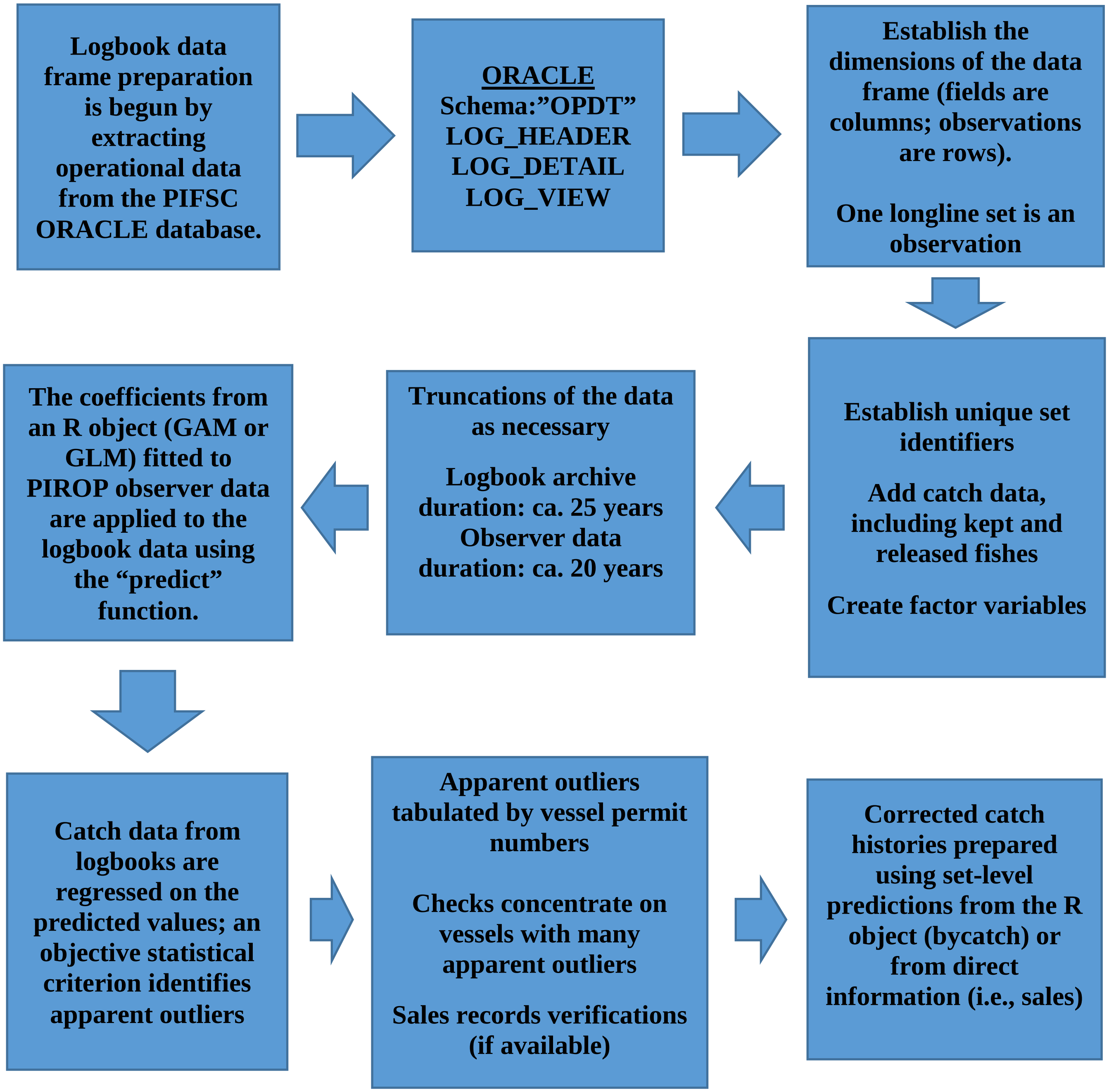


Figure 6.--Diagrammatic representation of commercial longline logbook data use. These steps generate an R data frame in a flat file format and can be used to prepare corrected catch histories.

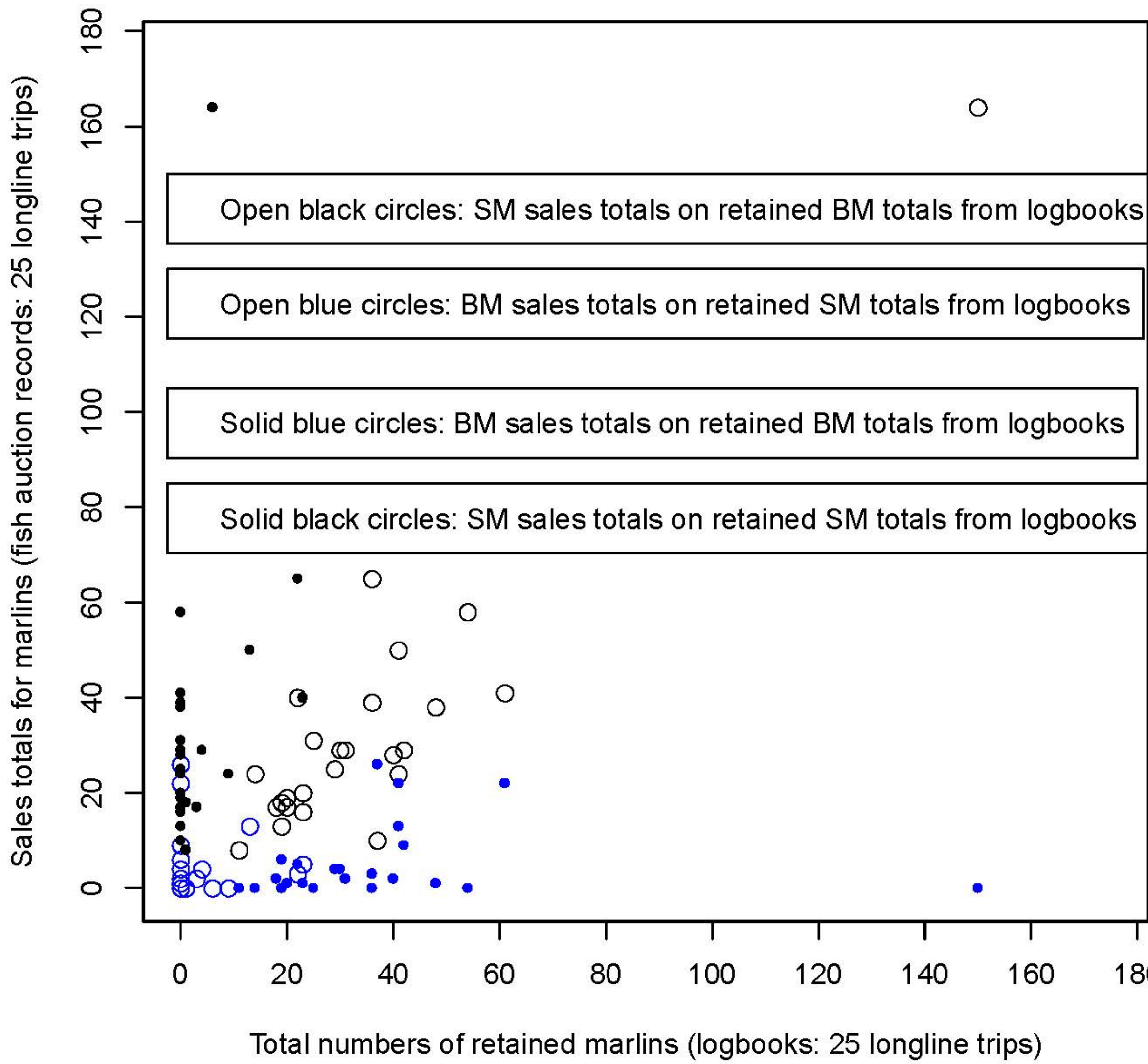


Figure 7.--Comparison plot illustrating blue and striped marlin sales totals in relation to reported totals of retained fish. Correctly identified fish fall on a 1:1 line. The solid black circles reflect low or zero sales of striped marlin caused by misidentifications as blue marlin; the solid blue circles reflect low or zero sales of blue marlin caused by misidentifications as striped marlin.

TABLES

Table 1. -- Summary of procedures used to calculate or obtain catches, catch rates, catch-related effects, sex ratios, observer commentary about operations or catches, and fishery sector definitions with PIROP data. Field names conform to those in the **CATCH_MV** table in the **NEWOBS** schema in the ORACLE database. Manipulations are performed in R after import from ORACLE as a data frame.

Variable of interest	**Characteristics**	**Required fields**	**Procedure to obtain or calculate**	**Remarks**
Total catch	Expressed in numbers of fish	ENGLISH_NAME KEPT_RETURN_CODE	Obtained directly from observer catch file Sum of kept and discarded fish (and finned, if sharks)	CHECK: ENGLISH_NAME not “NA” CHECK: Sum equal to total from ENGLISH_NAME
Nominal catch per unit effort (CPUE)	1000 × (numbers of fish caught / numbers of hooks set)	ENGLISH_NAME NUM_HKS_SET	Obtained directly from observer catch file Obtained directly from observer catch file	CHECK: ENGLISH_NAME not “NA” CHECK: NUM_HKS_SET not “NA”
Discarding rates	Expressed in percentages of total catch or numbers of fish	KEPT_RETURN_CODE	Obtained directly from observer catch file Sum of discards (live, dead and in unknown condition)	CHECK: KEPT_RETURN_CODE not “NA”
Disposition of caught fish	Expressed as the retained and discarded percentages of total catch	ENGLISH_NAME KEPT_RETURN_CODE	Calculated from two fields in the observer catch file: Retained, discarded, and finned fractions are divided by total catches.	CHECK: ENGLISH_NAME not “NA” CHECK: KEPT_RETURN_CODE not “NA”

Table 1 (cont'd)

Variable of interest	Characteristics	Required fields	Procedure to obtain or calculate	Remarks
Handling mortality rates	Expressed in percentages of total catch or of those surviving longline capture	CAUGHT_COND_CODE KEPT_RETURN_CODE	Calculated from two fields in the observer catch file: Select survivors of capture, subtract dead discards, and divide by appropriate total.	CHECK: KEPT_RETURN_CODE not "NA" CHECK: CAUGHT_COND_CODE not "NA"
Minimum mortality of caught fish	Expressed in percentages of total catch	KEPT_RETURN_CODE	Calculated from one field in the observer catch file: Retained and discarded fractions plus the handling mortality fraction are divided by total catches.	CHECK: KEPT_RETURN_CODE not "NA" CHECK: Handling mortality rates not "NA"
Sex ratios	Expressed as the ratio of males and females in sexually dimorphic or dichromatic species	GENDER_CODE	Calculated from one field in the observer catch file: Select sexed fish and divide the male and female fractions by the total.	CHECK: ENGLISH_NAME not "NA" CHECK: GENDER_CODE not "NA"
Fishery sector (deep "D" and shallow "S" sets)	The fishery sectors are defined on the basis of the number of hooks per float on a longline set.	HKS_PER_FLT	Obtained by defining set deep and shallow set types: Set_type <- ifelse(HKS_PER_FLT<15, "S", "D")	CHECK: HKS_PER_FLT not "NA" CHECK: NUM_HKS_SET (typical values for sector)

Table 1 (cont'd)

Variable of interest	**Characteristics**	**Required fields**	**Procedure to obtain or calculate**	**Remarks**
Observer commentary	Notes recorded at-sea about various aspects of longline sets or catches	CATCH_COMMENTS CATCH_COMMENTS_YN	Obtained directly from observer catch file: Select CATCH_COMMENTS_YN = "Y" Use CATCH_COMMENTS	CHECK: LOGBK_PG_NUM not "NA" CHECK: CATCH_COMMENTS_YN = "Y" and not "NA"
Sizes March 1994 – August 2003 (Trips LL001 – LL1064)	Morphometric measurements (total and fork lengths)	TOTAL_LEN FORK_LEN (PRECAUDAL_LEN =="NA")	Obtained directly from legacy files for tunas, sharks, and billfishes after their import into R as data frames.	CHECK: Measurement not "NA" CHECK: Use data from Trip LL1064 and lower trip numbers
Sizes August 2003 – present (Trips LL1068 – LL5175)	Morphometric measurements (total, fork, and precaudal lengths)	TOTAL_LEN FORK_LEN PRECAUDAL_LEN	Obtained directly from observer catch file: Complete time series prepared by concatenating legacy and recent data.	CHECK: Measurement not "NA" CHECK: Avoid duplicate legacy data Delete LL1065 – LL1067

Table 2. --Summary of criteria used to evaluate the accuracy of PIFSC logbook data and its suitability for analytical use by the SAP. Field names conform to those in the **CATCH_MV** table in the **NEWOBS** schema in the ORACLE database except species names, which are in lower case.

Longline data evaluation criterion	**Indicator of the need for longline data evaluation**	**Reasons to evaluate longline data using this indicator**	**Fields pertaining to indicator**	**Remarks and appropriate actions**
Commercial fishing using typical methods for this fleet	SOAK_TIME Longline gear soak duration (longer soak than usual)	Gear loss or other problem(s) Protected species interaction	LINE_PARTED_YN NUM_SECTIONS_ RETRVD SETGEAR_COMMENTS SET_INTERACT_YN HAUL_INTERACT_YN	Positive answers (“Y”) represent grounds to delete a set before analyses.
Commercial fishing using typical methods for this fleet	Longline gear begin-set time (later deployment than usual)	Reduced soak duration caused by gear deployment delay Reduced catch or apparently distorted catch composition	SET_BEGIN_DATETIME HAUL_BEGIN_DATETIME Catches of target species and other species of interest CATCH_COMMENTS	Reduced soak durations, unusual catches, or both may represent grounds to delete a set before analyses.
Commercial fishing using typical methods for this fleet	Longline gear mainline length (very short mainline)	A mainline less than one mile in length, especially if two or more short mainlines are deployed	MAINLN_LEN_RPTD	A short mainline may reflect attempted avoidance of regulations
Species-specific catch tallies compared to typical catches by this fleet	Unusually large catches or unexpected zero catches	Gear loss or other problem can lead to high bycatches Unexpected zeros or large catches may reflect misidentifications or some technical problem(s).	LINE_PARTED_YN NUM_SECTIONS_ RETRVD CATCH_COMMENTS	Positive answers (“Y”) represent grounds to delete a set before analyses. Check species identifications and correct mistakes if possible.
Species-specific catch tallies compared to typical catches by this fleet	Unexpected catches of rare or unusual species, or catches reported from outside a known distribution	Unexpected positive catches may indicate misidentifications.	O_OBSERVER_NUM PHOTO_YN SPECIMEN_YN	Check observer experience (i.e., number and types of previous trips) Check species identifications and correct mistakes if possible.

Table 2 (cont'd)

Longline data evaluation criterion	**Indicator of the need for longline data evaluation**	**Reasons to evaluate longline data using this indicator**	**Fields pertaining to indicator**	**Remarks and appropriate actions**
Species-specific catch tallies compared to typical catches by this fleet	Discarding unusually high numbers of commercial species	Discarding of commercial species in the context of high-grading may reflect possible spoilage or quota avoidance.	HIGH_GRADING_YN HIGH_GRADING_ COMMENTS	Positive answers ("Y") may represent grounds to delete a set before analyses if catch tally is doubtful.

Table 3.-- Summary of catch reporting patterns in the Hawaii-based longline fishery during 1995-2014 for 10 species, including the two target species, two other tunas, three billfishes, and three sharks. Table entries include mean numbers caught, kept (and percent kept), and released, and the percent zero catches per set. For sharks, the mean numbers kept (K) and finned (F) are included. Data sources are PIROP fishery observer reports (Observer), logbooks from the observed trips (Logbook (Obs)), and logbooks from unobserved trips (Logbook (No obs)). Data are presented by fishery sectors, with descriptions of the status of the various species.

Species	Fishery sector and Species status	Data source	Mean catch per set	Mean number kept per set	Mean number released per set	Percent zero catches
Swordfish *Xiphias gladius*	Deep Incidental catch	Observer	0.387	0.220 (56.8)	0.167	73.2
		Logbook (Obs)	0.338	0.260 (76.9)	0.078	80.4
		Logbook (No obs)	0.179	0.160 (89.4)	0.019	86.5
	Shallow Target species	Observer	11.881	10.580 (89.0)	1.301	2.3
		Logbook (Obs)	11.192	10.248 (91.6)	0.944	5.2
		Logbook (No obs)	8.038	7.654 (95.2)	0.384	15.8
Bigeye tuna *Thunnus obesus*	Deep Target species	Observer	9.301	8.802 (94.6)	0.500	7.0
		Logbook (Obs)	8.674	8.421 (97.1)	0.265	7.6
		Logbook (No obs)	8.460	8.299 (98.1)	0.161	8.5
	Shallow Incidental catch	Observer	1.163	1.075 (92.4)	0.089	54.3
		Logbook (Obs)	1.151	1.105 (96.0)	0.063	55.6
		Logbook (No obs)	2.916	2.857 (98.0)	0.059	38.0

Table 3 (cont'd)

Species	Fishery sector and Species status	Data source	Mean catch per set	Mean number kept per set	Mean number released per set	Percent zero catches
Albacore *Thunnus alalunga*	Deep Secondary target, Incidental catch	Observer	1.321	1.285 (97.3)	0.035	71.6
		Logbook (Obs)	1.335	1.250 (93.6)	0.016	71.3
		Logbook (No obs)	2.090	2.073 (99.2)	0.017	66.0
	Shallow Incidental catch, Bycatch	Observer	1.460	0.814 (55.8)	0.645	64.0
		Logbook (Obs)	1.185	0.674 (56.9)	0.419	69.1
		Logbook (No obs)	2.082	1.269 (61.0)	0.813	60.0
Yellowfin tuna *Thunnus albacares*	Deep Secondary target, Incidental catch	Observer	1.895	1.729 (91.2)	0.166	56.4
		Logbook (Obs)	1.786	1.703 (95.4)	0.082	57.6
		Logbook (No obs)	1.687	1.638 (97.1)	0.050	56.9
	Shallow Incidental catch	Observer	0.368	0.348 (94.6)	0.020	81.1
		Logbook (Obs)	0.344	0.332 (96.5)	0.013	82.0
		Logbook (No obs)	1.303	1.282 (98.4)	0.021	57.7

Table 3 (cont'd)

Species	**Fishery sector and Species status**	**Data source**	**Mean catch per set**	**Mean number kept per set**	**Mean number released per set**	**Percent zero catches**
Blue marlin *Makaira nigricans*	Deep Incidental catch	Observer	0.253	0.244 (96.4)	0.009	82.3
		Logbook (Obs)	0.271	0.259 (95.6)	0.005	82.5
		Logbook (No obs)	0.297	0.293 (98.7)	0.003	81.4
	Shallow Incidental catch	Observer	0.155	0.139 (89.7)	0.016	90.9
		Logbook (Obs)	0.189	0.149 (78.8)	0.008	90.2
		Logbook (No obs)	0.611	0.597 (97.7)	0.014	76.5
Striped marlin *Kajikia audax*	Deep Incidental catch	Observer	0.934	0.880 (94.2)	0.053	59.4
		Logbook (Obs)	0.857	0.817 (95.3)	0.020	61.7
		Logbook (No obs)	0.893	0.882 (98.8)	0.011	60.8
	Shallow Incidental catch	Observer	0.545	0.485 (89.0)	0.060	73.9
		Logbook (Obs)	0.482	0.425 (88.2)	0.035	76.6
		Logbook (No obs)	0.686	0.644 (93.9)	0.042	71.2

Table 3 (cont'd)

Species	Fishery sector and Species status	Data source	Mean catch per set	Mean number kept per set	Mean number released per set	Percent zero catches
Shortbill spearfish *Tetrapturus angustirostris*	Deep Incidental catch	Observer	0.844	0.780 (92.4)	0.064	58.9
		Logbook (Obs)	0.791	0.762 (96.3)	0.022	61.8
		Logbook (No obs)	0.853	0.842 (98.7)	0.011	60.5
Blue shark *Prionace glauca*	Deep Bycatch	Observer	4.259	0.009 (K) 0.180(F)	4.070	13.3
		Logbook (Obs)	3.659	0.045 (K) 0.164 (F)	3.450	24.3
		Logbook (No obs)	2.888	0.015 (K) 0.595 (F)	2.278	30.0
	Shallow Bycatch	Observer	8.680	0.003 (K) 0.722 (F)	7.941	3.9
		Logbook (Obs)	8.370	0.022 (K) 0.698 (F)	7.651	10.3
		Logbook (No obs)	10.826	0.011 (K) 3.592 (F)	7.223	13.9
Thresher sharks *Alopias* spp.	Deep Incidental catch, Bycatch	Observer	0.477	0.026 (K) 0.005 (F)	0.403	77.1
		Logbook (Obs)	0.409	0.027 (K) 0.006 (F)	0.375	76.9
		Logbook (No obs)	0.343	0.022 (K) 0.022 (F)	0.298	86.5
	Shallow Incidental catch, Bycatch	Observer	0.044	0.004 (K) 0.001 (F)	0.033	96.2
		Logbook (Obs)	0.051	0.004 (K) 0.001 (F)	0.046	96.3
		Logbook (No obs)	0.078	0.011 (K) 0.015 (F)	0.051	96.7

Table 3 (cont'd)

Mako sharks *Isurus* spp. (Observer data for Kept, Finned, and Released sharks are for shortfin makos only)	Deep Incidental catch, Bycatch	Observer	0.191	0.055 (K) 0.001 (F)	0.117	84.3
		Logbook (Obs)	0.162	0.058 (K) 0.002 (F)	0.102	87.2
		Logbook (No obs)	0.122	0.058 (K) 0.008 (F)	0.057	90.4
	Shallow Incidental catch, Bycatch	Observer	0.625	0.073 (K) 0.008 (F)	0.529	61.5
		Logbook (Obs)	0.550	0.070 (K) 0.006 (F)	0.474	67.5
		Logbook (No obs)	0.123	0.022 (K) 0.024 (F)	0.077	93.0

Table 4.-- Summary of procedures used to identify unique longline sets and calculate several variables related to effort with PIFSC logbook data. Field names conform to those in the **LOGHDR** table in the **OPDT** schema in the ORACLE database. Manipulations are performed in R after import from ORACLE as a data frame. Relevant caveats are presented. "NA" denotes "not available".

Operational parameter(s) (or other variable)	**Variable name** (**ORACLE**: **Log_Header**)	**Variable Characteristics**	**Required calculations or manipulations with variable(s)**	**Pertinent caveat(s) or remarks**
Unique set identifier (NA in ORACLE)	Unique_set_ID (NA in ORACLE)	Character	R "**paste"** function used to manipulate three fields.	Need ORACLE fields: LAND_YR, TRIPNUM, SERIALNUM
Unique set identifier (second: 1998–present)	Logpage	Character	None	This is the logbook page serial number, useful for all sets since 1998.
Dates of fishing (Haul dates)	HAULYR, HAULMON, HAULDAY	Reported since November 1990. Numeric	No required calculations or manipulations. Date calculations using combinations simple.	None These are begin-haul times, uninformative about long hauls.
Dates of fishing (Set dates)	SETYR, SETMON, SETDAY	Reported since January 1995. Numeric	No required calculations or manipulations. Date calculations using combinations simple.	Useful since 1995. These are begin-set times, uninformative about long sets.
Hooks (Number of hooks set as a measure of effort)	HOOKSSET HOOKS_LOST	Hooks set and lost reported since November 1990 and 2000, respectively. Numeric	Hooks set Hooks set - Hooks lost	Hooks set useful prior to 2000. Hooks set and lost useful since 2001. Average loss 2 hooks.
Characteristics of fishing effort	TRIP_TYPE TARGET	Character Character	None	Reported since 1990 and 1995, respectively. Not in definition of two-sector management.

Table 4 (cont'd)

Operational parameter(s) (or other variable)	**Variable name** (**ORACLE**: **Log_Header**)	**Variable Characteristics**	**Required calculations or manipulations with variable(s)**	**Pertinent caveat(s) or remarks**
Hooks per float	Hooks HOOKSSET HOOKS_LOST	Numeric	Imputed for 1990-1994 based on histories of vessels or captains after 1995.	Reported since 1995. Basis of definition of two-sector management.
Set types	Hooks per float	Numeric	Imputed for 1990-1994 based on histories of vessels or captains after 1995.	Shallow sets use < 15 hooks per float; deep sets use 15 or more hooks per float.
Spatial variables	Latitude and longitude in degrees and minutes BHLAT_DEG BHLAT_MIN	Numeric	Arithmetic averages of end-set (ES --- not shown) and begin-haul (BH) positions.	Begin-set, end-set, begin-haul, and end-haul positions available since 2001.

APPENDICES

APPENDIX A

This appendix is arranged as a series of examples to show preparation and use of longline logbook data from the NOAA Fisheries Pacific Islands Fisheries Science Center.

Example A_1. Commands used to import data from the PIFSC data base into R as data frames.
Example A_2. Summary of data frame preparation in R format for the PIFSC longline logbook database.
Example A_3. SST Data Matcher Instructions and Format Requirements
Example A_4. Truncations and final preparations applied to the logbook data set before detailed accuracy evaluations based on comparisons to predictions from a statistical model.

Example A 1. Commands used to import data from the PIFSC data base into R as data frames.

```
> library(ROracle)
> ora<-dbDriver("ROracle")
> con<-dbConnect(ora,user="user_name",password="your_pswd",dbname="PIC")
> dbListTables(con,schema="OPDT")
[1] "LOGDETAIL"  "LOGHDR"  "LOG_VW"
> DF_1<-dbGetQuery(con,"select * from OPDT.LOGDETAIL")
> class(DF1)
[1] "data.frame"
> dim(DF_1)
[1] 2260443        21
> DF_2-dbGetQuery(con,"select * from OPDT.LOGHDR")
> class(DF_2)
[1] "data.frame"
> dim(DF_2)
[1] 356457         105
> DF_3<-dbGetQuery(con,"select * from OPDT.LOG_VW")
> class(DF3)
[1] "data.frame"
> dim(DF_3)
[1] 2261454        102
> Log_Detail<-DF_1
> Log_Header<-DF_2
> Log_View<-DF_3
> objects()
[1] "Log_Detail"     "Log_Header"       "Log_View"
```

These are the data frames needed to construct the analytical data frame for the logbooks.

Example A 2. Summary of data frame preparation for the PIFSC longline logbook database. The tables used are LOGHDR and LOGDETAIL. Explanatory notes are italicized. PIROP observer data would be treated similarly.

File dimensions (*both are data frames*)

```
> dim(Log_Detail)
[1] 2260443    21
> dim(Log_Header)
[1] 356457   106
```

First field in the logbook data frame is the ***Logpage*** (*this is the logbook page serial number, which has served as a unique set identifier since 1997*).

```
> Logsdata<-data.frame(Log_Header$Logpage)
> dim(Logsdata)
[1] 356457    1
```

The linking variables and unique set identifiers must be character variables.

```
> Logsdata$Logpage<-as.character(Logsdata$Log_Header.Logpage)
> dim(Logsdata)
[1] 356457    2
```

```
> summary(Logsdata)
 Log_Header.Logpage   Logpage
 1      :  8567     Length:356457
 2      :  8334     Class :character
 3      :  8081     Mode  :character
 4      :  7740
 5      :  7432
 6      :  7030
 (Other):309273
> mode(Logsdata$Logpage)
[1] "character"
```

The extra field shows where it was obtained (i.e., copied from Log_Header) *Delete*.

```
> Logsdata$Log_Header.Logpage<-NULL
> dim(Logsdata)
[1] 356457      1
```

> *Correct*

Addition of numerical temporal variables

```
> dim(Logsdata)
[1] 356457      1
> Logsdata$Haulyr<-Log_Header$HAULYR
> Logsdata$Haulmo<-Log_Header$HAULMON
> Logsdata$Haulday<-Log_Header$HAULDAY

> Logsdata$Setyr<-Log_Header$SETYR
> Logsdata$Setmo<-Log_Header$SETMON
> Logsdata$Setday<-Log_Header$SETDAY

> Logsdata$Depart_yr<-Log_Header$DEPART_YR
> Logsdata$Depart_mo<-Log_Header$DEPART_MON
> Logsdata$Depart_day<-Log_Header$DEPART_DAY

> Logsdata$Return_yr<-Log_Header$RETURN_YR
> Logsdata$Return_mo<-Log_Header$RETURN_MON
> Logsdata$Return_day<-Log_Header$RETURN_DAY
 > Logsdata$Land_yr<-Log_Header$LANDYR
> Logsdata$Land_mo<-Log_Header$LANDMO
```

The variable additions appear correct. Zero values for dates are incorrect.

Addition of variables related to effort.

```
> Logsdata$Tripnum<-Log_Header$TRIPNUM
> Logsdata$Trip_type<-Log_Header$TRIPTYPE
> Logsdata$Target<-Log_Header$TARGET

> mode(Logsdata$Tripnum)
```

```
[1] "numeric"
> sum(Logsdata$Tripnum)
[1] 231272503
```

Correct --- advantageous to use as numeric

```
> mode(Logsdata$Trip_type)
[1] "numeric"
> table(Logsdata$Trip_type)
    B       C       M        T
 28900   7447   35085   285025
```

The letters denote: B=Broadbill (swordfish); C=Certificate (shallow-set ca. 2004; also swordfish); M=mixed species; T=tunas.

```
> Logsdata$Trip_type<-as.character(as.factor(Logsdata$Trip_type))
> table(Logsdata$Trip_type)
    B       C       M       T
 28900   7447   35085  285025
> mode(Logsdata$Trip_type)
[1] "character"
```

Correct mode change; no change in tabulation.

```
> table(Logsdata$Target)
    M      B       BM    M      B      M      T        T M    TB  TBM
 19976  19533     20     14    157    224 264314     45     51     20
> mode(Logsdata$Target)
[1] "numeric"
> Logsdata$Target<-as.character(as.factor(Logsdata$Target))
> mode(Logsdata$Target)
[1] "character"
> table(Logsdata$Target)
    M      B       BM    M      B      M      T        T M   TB  TBM
 19976  19533     20     14    157    224 264314     45     51     20
```

Correct mode change; no change in tabulation. Multiple categories introduce ambiguity.

```
> Logsdata$Setnum<-Log_Header$SETNUM
> Logsdata$ResExptCode<-Log_Header$RSCH_EXPMTL_CODE
> Logsdata$Permit<-Log_Header$PERMITNUM
> Logsdata$Vessel<-Log_Header$VESSELNAME
> Logsdata$CML<-Log_Header$CML
> Logsdata$Hooks<-Log_Header$HOOKSSET
> Logsdata$Permit<-as.character(as.factor(Logsdata$Permit))
```

Correct

```
> Logsdata$Vessel<-as.character(as.factor(Logsdata$Vessel))
```

```
> Logsdata$CML<-as.character(as.factor(Logsdata$CML))
> Logsdata$Hooks<-Log_Header$HOOKSSET
```

Identification of experimental sets and research activities.

```
> mode(Logsdata$ResExptCode)
[1] "numeric"
> unique((Logsdata$ResExptCode))
[1] <NA> R    X
Levels: R X
> table((Logsdata$ResExptCode))
  R    X
2017  223
> Logsdata$ResExptCode<-as.character(as.factor(Logsdata$ResExptCode))
> table((Logsdata$ResExptCode))
  1      2
2017  223
> mode((Logsdata$ResExptCode))
[1] "character"
> Logsdata$ResExptCode<-ifelse(Logsdata$ResExptCode=="1","R",Logsdata$ResExptCode)
> Logsdata$ResExptCode<-ifelse(Logsdata$ResExptCode=="2","X",Logsdata$ResExptCode)
> Logsdata$ResExptCode<-ifelse(is.na(Logsdata$ResExptCode),"LL",Logsdata$ResExptCode)
> table((Logsdata$ResExptCode))
   LL        R         X
354217   2017    223
```

If neither research nor experimental, commercial longline activity.

```
> 354217+2017+223
[1] 356457
> dim(Logsdata)
[1] 356457     21
```

Correct

The next two variables refer to the PIROP trip and set codes. These differed from the PIFSC assigned trip and set numbers.

```
> Logsdata$Tripnum.obs<-Log_Header$OBSTRIPNUM
> mode(Logsdata$Tripnum.obs)
[1] "numeric"
> Logsdata$Tripnum.obs<-as.character(as.factor(Logsdata$Tripnum.obs))
> Logsdata$Setnum.obs<-Log_Header$OBSSETNUM
> Logsdata$Setnum.obs<-as.character(as.numeric(Logsdata$Setnum.obs))
```

This facilitates checking to determine whether there was an observer on that fishing trip.

Unique set identifiers

This is a simple matter after 1997 because the logbook page serial number was instituted as a unique set identifier during 1997. Thus, for 1998 and beyond, **Logsdata$Logpage** *was copied within the data frame and renamed* **Logsdata$unique_set_ID**.

The unique set identifier for the earlier years was created by turning the **Logsdata$Landyr**, **Logsdata$Tripnum**, *and* **Logsdata$Serialnum** *variables into characters, and then making a unique identifier variable with the "***paste***" command*.

```
> Logsdata$Land_yr1<-as.character(as.numeric(Logsdata$Land_yr))
> Logsdata$unique_set_ID<-ifelse(Logsdata$Haulyr>1997,Logsdata$Logpage,"NA")
```

This means that if the haul year was 1998 or later, the logpage serial number is the identifier.

```
> Logsdata$Land_yr1<-as.character(as.numeric(Logsdata$Land_yr))
> Logsdata$Tripnum1<-as.character(as.numeric(Logsdata$Tripnum))
> Logsdata$Serialnum1<-as.character(as.numeric(Logsdata$Serialnum))
> Logsdata$unique_set_ID1<paste(Logsdata$Land_yr1,Logsdata$Tripnum1,
Logsdata$Serialnum1)
> length(unique(Logsdata$unique_set_ID))
[1] 356447
> dim(Logsdata)
[1] 356457     33
```

Ten sets remain unidentified.

Additional operational variables

Mainline length (nautical miles)

```
> Logsdata$ML_length<-Log_Header$MAINLINE
> summary(Logsdata$ML_length)
   Min. 1st Qu.  Median    Mean  3rd Qu.     Max.     NA's
   0.00   29.00     33.00      34.55    40.00     620.00   52992
```

The maximum may be a units problem or similar error.

Hooks per float

This is particularly important because hooks per float became the basis for two sector management after 2001.

```
> summary(Log_Header$MAXHKSFLT)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00    24.00    26.00      23.98   30.00     88.00   54772
> summary(Log_Header$MINHKSFLT)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.     NA's
   0.00    24.00    26.00      23.88    30.00     88.00   50521

> Log_Header$Hkpfl<-(Log_Header$MINHKSFLT+Log_Header$MAXHKSFLT)/2
> summary(Log_Header$Hkpfl)
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.    NA's
   0.00   24.00      26.00      23.98   30.00     88.00   54772
```

The preceding steps made the average number of hooks per float in **Log_Header***; it is now assigned to* **Logsdata**.

```
> Logsdata$Hkpfl<-Log_Header$Hkpfl
> summary(Logsdata$Hkpfl)
```

```
   Min. 1st Qu.  Median    Mean  3rd Qu.     Max.    NA's
   0.00    24.00    26.00     23.98   30.00     88.00   54772
> Logsdata$Max_Hkpfl<-Log_Header$MAXHKSFLT
> Logsdata$Hkpfl<-ifelse(is.na(Logsdata$Hkpfl),Logsdata$Max_Hkpfl,Logsdata$Hkpfl)
> summary(Logsdata$Hkpfl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    24.00    26.00     23.98   30.00     88.00   54772

> Logsdata$Min_Hkpfl<-Log_Header$MINHKSFLT
> Logsdata$Hkpfl<-ifelse(is.na(Logsdata$Hkpfl),Logsdata$Min_Hkpfl,Logsdata$Hkpfl)
> summary(Logsdata$Hkpfl)
   Min. 1st Qu.  Median    Mean 3rd Qu.     Max.    NA's
   0.00   24.00     26.00     23.88   30.00     88.00   50521

> junk<-Logsdata[is.na(Logsdata$Hkpfl), ]
> dim(junk)
[1] 50521    38
> table(junk$Haulyr)
 1990   1991   1992    1993    1994   1995 1996  1997  1998  1999  2000  2001  2003  2004  2005
 1015  12635  11546  12318   10799   1896   95     67     34     31     38     30     1      15     1
```

Most of the sets lacking hooks per float information were from 1990 through 1994. Imputation of the hooks per float may be feasible because many captains apparently habitually use a certain number of hooks per float (or possibly just always report the same value).

Set types as factors

> A two-level factor variable and corresponding character variable are created using the **hooks per float (hkpfl).** *The factor variable is intended for use in GLM analyses*.

```
> Logsdata$Set_type<-ifelse((is.na(Logsdata$Hkpfl)),"0","NA")
> Logsdata$Set_type<-ifelse(Logsdata$Hkpfl<15,"1",Logsdata$Set_type)
> Logsdata$Set_type<-ifelse(Logsdata$Hkpfl>=15,"2",Logsdata$Set_type)

> table(Logsdata$Set_type)
    1          2
 43234   262702
> 43234+262702
[1] 305936
> unique(Logsdata$Set_type)
[1] NA  "1" "2"

> Logsdata$Set_type<-ifelse(Logsdata$Set_type=="1","S","D")
> unique(Logsdata$Set_type)
[1] NA  "S" "D"
> table(Logsdata$Set_type)
    D        S
262702  43234
```

```
> Logsdata$Set_type1<-as.factor(Logsdata$Set_type)
> Logsdata$Set_type1<-ifelse(Logsdata$Set_type1=="S",1,2)
> table(Logsdata$Set_type1)
    1      2
 43234 262702
```
Correct

Temporal variables as factors

```
> Logsdata$Haulyr1<-as.factor(Logsdata$Haulyr)
> table(Logsdata$Haulyr1)
 1990  1991   1992   1993   1994   1995   1996   1997   1998   1999   2000   2001   2002
 1015 12635 11546 12318 10799 11732 11638 11846 12505 12805 12931 12186 14110
 2003   2004  2005    2006   2007   2008   2009   2010   2011   2012   2013   2014
14883 16029 18195 17302 19385 19482 18572 17948 18641 19466 19734  8754

> Logsdata$Quarter<-ifelse(Logsdata$Haulmo<4,1,"NA")
> Logsdata$Quarter<-ifelse(Logsdata$Haulmo>3 & Logsdata$Haulmo<7,2,Logsdata$Quarter)
> Logsdata$Quarter<-ifelse(Logsdata$Haulmo>6 & Logsdata$Haulmo<10,3,Logsdata$Quarter)
> Logsdata$Quarter<-ifelse(Logsdata$Haulmo>9,4,Logsdata$Quarter)

> table(Logsdata$Quarter)
   1       2         3       4
95392 92912 74558 93595

> Logsdata$Quarter1<-as.factor(as.numeric(Logsdata$Quarter))
> table(Logsdata$Quarter1)
   1        2         3       4
95392 92912 74558 93595
> mode(Logsdata$Quarter1)
[1] "numeric"
> sum(Logsdata$Quarter1)
Error in Summary.factor(c(4L, 4L, 4L, 4L, 4L, 4L, 4L, 4L, 4L, 4L, 4L,  :
  'sum' not meaningful for factors
```
Correct

Positions

```
> Log_Header$BS_latitude<-Log_Header$BSLATDEG+(Log_Header$BSLATMIN/60)
> summary(Log_Header$BS_latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   18.03     22.05     22.29   26.95    46.00      26

> Log_Header$EH_latitude<-Log_Header$EHLATDEG+(Log_Header$EHLATMIN/60)
> Log_Header$BS_latitude<-Log_Header$BSLATDEG+(Log_Header$BSLATMIN/60)
> summary(Log_Header$BS_latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   18.03     22.05     22.29   26.95    46.00      26
```

```
> summary(Log_Header$EH_latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   17.53   21.73   21.81   26.33   45.60   50206

> Log_Header$Latitude<-(Log_Header$BS_latitude+Log_Header$EH_latitude)/2
> summary(Log_Header$Latitude)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00   17.52    21.73   21.81   26.33   45.63   50207

> summary(Log_Header$BHLATDEG)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    18.0    22.0    21.8    26.0    46.0      83
> summary(Log_Header$BHLATMIN)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00   14.00   29.00   29.26   45.00   59.00      83
> summary(Log_Header$EHLATDEG)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00   17.00   21.00   21.32   26.00   45.00   50206
> summary(Log_Header$EHLATMIN)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00   14.00   29.00   29.34   45.00   59.00   50206

> Log_Header$Latitude<-
ifelse((is.na(Log_Header$Latitude)),(Log_Header$BHLATDEG+(Log_Header$BHLATMIN/60
)),Log_Header$Latitude)
> summary(Log_Header$Latitude)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00   18.03   22.06   22.29   26.95   46.63      82


> Log_Header$Longitude<-
ifelse((is.na(Log_Header$Longitude)),(Log_Header$BHLONGDEG+(Log_Header$BHLONG
MIN/60)),Log_Header$Longitude)

> summary(Log_Header$Longitude)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
    0.0   154.7   159.1   158.7   162.9   180.0      82

> Logsdata$Latitude<-Log_Header$Latitude
> Logsdata$Longitude<-Log_Header$Longitude
> Logsdata$Longitude_SST<-Log_Header$Longitude*(-1)
>
> summary(Logsdata$Latitude)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.    NA's
   0.00   18.03   22.06   22.29   26.95   46.63      82
```

```
> summary(Log_Header$Longitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0   154.7   159.1     158.7  162.9     180.0     82
 > summary(Logsdata$Longitude_SST)
   Min.   1st Qu.  Median    Mean    3rd Qu.    Max.    NA's
 -180.0   -162.9   -159.1    -158.7    -154.7       0.0      82
```

Begin-set time

```
> Log_Header$BS_Time<-Log_Header$BSHR+(Log_Header$BSMIN/60)
> summary(Log_Header$BS_Time)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.     NA's
  1.000   7.083      8.000     10.020   9.750    24.000     780

> Logsdata$BS_Time<-Log_Header$BS_Time
> summary(Logsdata$BS_Time)
   Min. 1st Qu.  Median    Mean    3rd Qu.    Max.    NA's
  1.000   7.083    8.000     10.020    9.750    24.000     780
```

Fishing Regions --- discretized and made into a factor (Brodziak & Walsh 2013)

```
> Logsdata$Region<-ifelse(Logsdata$Latitude<10 & Logsdata$Longitude<160,1,"NA")
> Logsdata$Region<-ifelse(Logsdata$Latitude<10 &
Logsdata$Longitude>=160,2,Logsdata$Region)
> Logsdata$Region<-ifelse((Logsdata$Latitude>=10 & Logsdata$Latitude<20) &
Logsdata$Longitude<160,3,Logsdata$Region)
> Logsdata$Region<-ifelse((Logsdata$Latitude>=10 & Logsdata$Latitude<20) &
Logsdata$Longitude>=160,4,Logsdata$Region)
> Logsdata$Region<-ifelse((Logsdata$Latitude>=20 & Logsdata$Latitude<30) &
Logsdata$Longitude<160,5,Logsdata$Region)
> Logsdata$Region<-ifelse((Logsdata$Latitude>=20 & Logsdata$Latitude<30) &
Logsdata$Longitude>=160,6,Logsdata$Region)
> Logsdata$Region<-ifelse(Logsdata$Latitude>=30 &
Logsdata$Longitude<160,7,Logsdata$Region)
> Logsdata$Region<-ifelse(Logsdata$Latitude>=30 &
Logsdata$Longitude>=160,8,Logsdata$Region)
```

```
> table(Logsdata$Region)
     1        2        3         4         5          6          7        8
  1034   8683  52449  73155 121465  59652  23445  16492
> mode(Logsdata$Region)
[1] "character"
> Logsdata$Region1<-as.factor(as.character(Logsdata$Region))
> table(Logsdata$Region1)
      1        2        3         4         5          6          7        8
  1034   8683  52449  73155 121465  59652  23445  16492
```

```
> mode(Logsdata$Region1)
[1] "numeric"
```
Correct

The remainder of this example presents the catch compilations and links to the operational information.

Blue shark

```
> junk<-Log_Detail[Log_Detail$ENGLISH_NAME=="BLUE SHARK",]
> dim(junk)
[1] 255380   25
> junk$unique_set_ID1<-junk$unique_set_ID
> mode(junk$unique_set_ID1)
[1] "character"
```
This allows linking this data frame with all blue shark records to the operational information.

```
> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)
> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)

> junk$Blue_shark<- junk$NUMFINNED+ junk$NUMKEPT+ junk$NUMRELEASED

> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)

> Logsdata$Blue_shark<-junk$Blue_shark[match1]
> Logsdata$BS_Finned<-junk$NUMFINNED[match1]
> Logsdata$BS_Kept<-junk$NUMKEPT[match1]
> Logsdata$BS_Released<-junk$NUMRELEASED[match1]

> summary(Logsdata$Blue_shark)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.     NA's
   0.00    2.00     4.00       6.55     7.00     800.00   101077


> summary(Logsdata$BS_Finned)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.     NA's
   0.00     0.00     0.00       0.98    0.00     700.00   101077
> summary(Logsdata$BS_Kept)
   Min.  1st Qu.  Median    Mean 3rd Qu.    Max.     NA's
   0.00     0.00      0.00       0.15    0.00    400.00   101077
> summary(Logsdata$BS_Released)
   Min. 1st Qu.  Median    Mean  3rd Qu.     Max.      NA's
   0.00    1.00      3.00        5.43    6.00       800.00   101077

> Logsdata$Blue_shark<-ifelse((is.na(Logsdata$Blue_shark)),0,Logsdata$Blue_shark)
> summary(Logsdata$Blue_shark)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
```

```
  0.000   0.000     2.000     4.693    5.000   800.000

> Logsdata$BS_Finned<-ifelse((is.na(Logsdata$BS_Finned)),0,Logsdata$BS_Finned)
> Logsdata$BS_Kept<-ifelse((is.na(Logsdata$BS_Kept)),0,Logsdata$BS_Kept)
> Logsdata$BS_Released<-ifelse((is.na(Logsdata$BS_Released)),0,Logsdata$BS_Released)

> summary(Logsdata$Blue_shark)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  0.000   0.000    2.000      4.693   5.000    800.000
> sum(Logsdata$Blue_shark)
[1] 1672690

> summary(Logsdata$BS_Finned)
    Min.   1st Qu.   Median    Mean   3rd Qu.     Max.
  0.0000   0.0000   0.0000   0.6988   0.0000 700.0000
> sum(Logsdata$BS_Finned)
[1] 249079

> summary(Logsdata$BS_Kept)
    Min.   1st Qu.   Median   Mean    3rd Qu.    Max.
  0.0000   0.0000   0.0000   0.1058   0.0000 400.0000
> sum(Logsdata$BS_Kept)
[1] 37729

> summary(Logsdata$BS_Released)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  0.000   0.000    2.000     3.888     4.000    800.000
> sum(Logsdata$BS_Released)
[1] 1385882
> 249079+37729+1385882
[1] 1672690
```
Correct

Mako shark

```
> junk<-Log_Detail[Log_Detail$ENGLISH_NAME=="MAKO SHARK",]
> dim(junk)
[1] 37883    25
> junk$unique_set_ID1<-junk$unique_set_ID
```
This allows linking this data frame with all mako records to the operational information.

```
> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)
> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)

> junk$Mako_shark<-junk$NUMKEPT+junk$NUMFINNED+junk$NUMRELEASED
> summary(junk$Mako_shark)
```

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  0.000   1.000    1.000      1.385   1.000     48.000

> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)



> Logsdata$Mako_shark<-junk$Mako_shark[match1]
> summary(Logsdata$Mako_shark)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0      1.0        1.0         1.4     1.0        48.0  318574

> Logsdata$Mako_shark<-ifelse((is.na(Logsdata$Mako_shark)),0,Logsdata$Mako_shark)
> summary(Logsdata$Mako_shark)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1472  0.0000 48.0000
> sum(Logsdata$Mako_shark)
[1] 52458

> Logsdata$Mako_Finned<-junk$NUMFINNED[match1]
> Logsdata$Mako_Finned<-ifelse((is.na(Logsdata$Mako_Finned)),0,Logsdata$Mako_Finned)

> Logsdata$Mako_Kept<-junk$NUMKEPT[match1]
> Logsdata$Mako_Kept<-ifelse((is.na(Logsdata$Mako_Kept)),0,Logsdata$Mako_Kept)

> Logsdata$Mako_Released<-junk$NUMRELEASED[match1]
> Logsdata$Mako_Released<-
ifelse((is.na(Logsdata$Mako_Released)),0,Logsdata$Mako_Released)
> sum(Logsdata$Mako_shark)
[1] 52458
> sum(Logsdata$Mako_Finned)
[1] 2657
> sum(Logsdata$Mako_Kept)
[1] 20598
> sum(Logsdata$Mako_Released)
[1] 29203
> 2657+20598+29023
[1] 52278
> 2657+20598+29203
[1] 52458
```
Correct

Thresher shark

```
> junk<-Log_Detail[Log_Detail$ENGLISH_NAME=="THRESHER SHARKS",]
> dim(junk)
```

```
[1] 41887    25

> junk$unique_set_ID1<-junk$unique_set_ID
> mode(junk$unique_set_ID1)
[1] "character"
```

This allows linking this data frame with thresher shark records to the operational information.

```
> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)
> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)
> junk$Thresher_shark<-junk$NUMKEPT+junk$NUMFINNED+junk$NUMRELEASED

> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)

> Logsdata$Thresher_shark<-junk$Thresher_shark[match1]
> summary(Logsdata$Thresher_shark)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0     1.0     1.0     2.4     2.0   200.0  314570
> Logsdata$Thresher_shark<-ifelse((is.na(Logsdata$Thresher_shark)),0,
Logsdata$Thresher_shark)
> summary(Logsdata$Thresher_shark)
    Min.  1st Qu.   Median     Mean   3rd Qu.   Max.
  0.0000   0.0000   0.0000   0.2815   0.0000 200.0000
> sum(Logsdata$Thresher_shark)
[1] 100347
> Logsdata$Thresher_Finned<-junk$NUMFINNED[match1]
> Logsdata$Thresher_Finned<-
ifelse((is.na(Logsdata$Thresher_Finned)),0,Logsdata$Thresher_Finned)
> Logsdata$Thresher_Kept<-junk$NUMKEPT[match1]
> Logsdata$Thresher_Kept<-
ifelse((is.na(Logsdata$Thresher_Kept)),0,Logsdata$Thresher_Kept)
> Logsdata$Thresher_Released<-junk$NUMRELEASED[match1]
> Logsdata$Thresher_Released<-
ifelse((is.na(Logsdata$Thresher_Released)),0,Logsdata$Thresher_Released)

> summary(Logsdata$Thresher_Finned)
    Min.    1st Qu.    Median     Mean    3rd Qu.     Max.
 0.00000  0.00000  0.00000   0.01508  0.00000 50.00000
> sum(Logsdata$Thresher_Finned)
[1] 5375
> summary(Logsdata$Thresher_Kept)
    Min.    1st Qu.    Median     Mean    3rd Qu.     Max.
 0.00000  0.00000  0.00000  0.02248  0.00000 42.00000
> sum(Logsdata$Thresher_Kept)
[1] 8013
> summary(Logsdata$Thresher_Released)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.244   0.000  200.000
> sum(Logsdata$Thresher_Released)
[1] 86959
 > 5375+8013+86959
[1] 100347
```

Correct

Oceanic whitetip shark

The logbook form first had an entry position for this species in 1995.

```
> junk<-Log_Detail[Log_Detail$ENGLISH_NAME=="OCEANIC WHITETIP SHARK",]
> dim(junk)
[1] 9557   25
> junk$unique_set_ID1<-junk$unique_set_ID
> mode(junk$unique_set_ID1)
[1] "character"

> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)
> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)

> junk$OWT_shark<-junk$NUMFINNED+junk$NUMKEPT+junk$NUMRELEASED

> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)

> Logsdata$OWT_shark<-junk$OWT_shark[match1]
> Logsdata$OWT_Finned<-ifelse(((is.na(Logsdata$OWT_Finned))&
Logsdata$Haulyr>1994),0,Logsdata$OWT_Finned)
> Logsdata$OWT_Kept<-ifelse(((is.na(Logsdata$OWT_Kept))&
Logsdata$Haulyr>1994),0,Logsdata$OWT_Kept)
> Logsdata$OWT_Released<-ifelse(((is.na(Logsdata$OWT_Released))&
Logsdata$Haulyr>1994),0,Logsdata$OWT_Released)
```

These commands differed from those used for the other sharks because this species was added to the form in 1995 so the entries from 1990 through 1994 should remain as "NA".

```
> Logsdata$OWT_shark<-
ifelse(Logsdata$Haulyr>1994,Logsdata$OWT_Finned+Logsdata$OWT_Kept+
Logsdata$OWT_Released,NA)
> summary(Logsdata$OWT_shark)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    0.00    0.00    0.06    0.00   32.00  48313
junk<-Logsdata[Logsdata$Haulyr>1994,]
> summary(junk$OWT_shark)
```

```
   Min.    1st Qu.    Median     Mean    3rd Qu.     Max.
0.00000  0.00000   0.00000  0.05524  0.00000 32.00000
> summary(junk$OWT_Finned)
   Min.      1st Qu.    Median     Mean      3rd Qu.     Max.
0.000000 0.000000 0.000000 0.001645 0.000000 9.000000
> summary(junk$OWT_Kept)
   Min.      1st Qu.    Median     Mean      3rd Qu.    Max.
0.000000 0.000000 0.000000 0.002612 0.000000 8.000000
> summary(junk$OWT_Released)
   Min.     1st Qu.     Median     Mean   3rd Qu.      Max.
0.00000   0.00000   0.00000   0.05098  0.00000 32.00000
> sum(junk$OWT_Finned)
[1] 507
> sum(junk$OWT_Kept)
[1] 805
> sum(junk$OWT_Released)
[1] 15709
> 15709+805+507
[1] 17021
> sum(junk$OWT_shark)
[1] 17021
```
 > *Correct*

Swordfish

The remaining species are teleosts so finned specimens would not be expected.

```
> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)
> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)
> sum(junk$NUMKEPT)
[1] 669960
> sum(junk$NUMRELEASED)
[1] 47252
> 669960+47252
[1] 717212
> sum(junk$NUMFINNED)
[1] 0
```
As expected

```
> junk$Swordfish<-junk$NUMKEPT+junk$NUMRELEASED
> summary(junk$Swordfish)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.00     1.00      4.00        6.67   10.00    103.00
> sum(junk$Swordfish)
[1] 717212
```
Correct

```
> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)
```

```
> Logsdata$Swordfish<-junk$Swordfish[match1]
> Logsdata$Swordfish<-ifelse((is.na(Logsdata$Swordfish)),0,Logsdata$Swordfish)

> Logsdata$Sword_Kept<-junk$NUMKEPT[match1]
> Logsdata$Sword_Kept<-ifelse((is.na(Logsdata$Sword_Kept)),0,Logsdata$Sword_Kept)

> Logsdata$Sword_Released<-junk$NUMRELEASED[match1]
> Logsdata$Sword_Released<-
ifelse((is.na(Logsdata$Sword_Released)),0,Logsdata$Sword_Released)

> summary(Logsdata$Swordfish)
   Min.  1st Qu.  Median    Mean    3rd Qu.    Max.
  0.000    0.000    0.000     2.012    1.000    103.000
> sum(Logsdata$Swordfish)
[1] 717212

> summary(Logsdata$Sword_Kept)
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.
  0.000   0.000    0.000     1.879   1.000    103.000
> sum(Logsdata$Sword_Kept)
[1] 669960

> summary(Logsdata$Sword_Released)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1326  0.0000 75.0000
> sum(Logsdata$Sword_Released)
[1] 47252
 > 669960+47252
[1] 717212
```

Correct

```
> 669960/717212
```

[1] **0.9341171** ***High retention rate as expected for a primary target species.***

Blue marlin

```
> junk<-Log_Detail[Log_Detail$ENGLISH_NAME=="BLUE MARLIN",]
> dim(junk)
[1] 67264    25

> junk$unique_set_ID1<-junk$unique_set_ID
> mode(junk$unique_set_ID1)
[1] "character"

> junk$NUMKEPT<-ifelse((is.na(junk$NUMKEPT)),0,junk$NUMKEPT)
> junk$NUMRELEASED<-ifelse((is.na(junk$NUMRELEASED)),0,junk$NUMRELEASED)
```

```
> junk$NUMFINNED<-ifelse((is.na(junk$NUMFINNED)),0,junk$NUMFINNED)

> sum(junk$NUMKEPT)
[1] 120895
> sum(junk$NUMRELEASED)
[1] 2987
> sum(junk$NUMFINNED)
[1] 0


> junk$Blue_marlin<-junk$NUMKEPT+junk$NUMRELEASED
> summary(junk$Blue_marlin)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.842   2.000  86.000

> match1<-match(Logsdata$unique_set_ID1,junk$unique_set_ID1)
> Logsdata$Blue_marlin<-junk$Blue_marlin[match1]
> Logsdata$Blue_marlin<-ifelse((is.na(Logsdata$Blue_marlin)),0,Logsdata$Blue_marlin)
> Logsdata$Blumar_Kept<-junk$NUMKEPT[match1]
> Logsdata$Blumar_Kept<-ifelse((is.na(Logsdata$Blumar_Kept)),0,Logsdata$Blumar_Kept)

> Logsdata$Blumar_Released<-junk$NUMRELEASED[match1]
> Logsdata$Blumar_Released<-
ifelse((is.na(Logsdata$Blumar_Released)),0,Logsdata$Blumar_Released)

> summary(Logsdata$Blue_marlin)
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3475  0.0000  86.0000
> sum(Logsdata$Blue_marlin)
[1] 123882

> summary(Logsdata$Blumar_Kept)
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.3392  0.0000 77.0000
> sum(Logsdata$Blumar_Kept)
[1] 120895



> summary(Logsdata$Blumar_Released)
    Min.  1st Qu.   Median    Mean  3rd Qu.     Max.
 0.00000  0.00000  0.00000  0.00838  0.00000 45.00000
> sum(Logsdata$Blumar_Released)
[1] 2987

> 120895+2987
```

[1] 123882
Correct preparation for blue marlin. Maximum probably misidentifications.
IDENTICAL PROCEDURES (not shown) ***FOLLOWED FOR THE OTHER TELEOSTS.***
Example A 3. Summary of the SST Data Matcher Instructions
(We thank Lucas Moxey, formerly of the PIFSC, author of these instructions).
In an effort to assist other divisions for matching satellite oceanographic data (SST, ocean color, etc) with particular dates and geographic locations (lat/lon), we can use a simple routine that will make use of Ferret for retrieving specific data points from the datasets that exist in the OceanWatch THREDDS server.

The matcher routines are available in MAR, under /home/las/matcher Here, the user will find 3 directories, these being "bin", "input_checker", and "work". In order to run the data matches, the user must follow these steps:

1. Check that the list of data points for which we wish to correlate the satellite data has the correct format. In order to do this, a text file with the list of dates and lat/long values must be created. In this list, each line must follow the following exact format:

    col1 has values between 1 - 31 (no text)

    col2 has values of Jan - Dec (no numbers)

    col3 has values between 1980 - 2013 (no text)

    col4 has values between -180 through 180 (no text)

    col5 has values between -90 though 90 (no text)

    This would result in a text file with a list of entries such as:

    14 Nov 2001 150.833325 20.377075

    21 Oct 2001 160.942925 14.015

    11 Sep 2001 164.974175 13.062925

    30 Nov 2001 159.262525 23.322475

    5 Oct 2001 150.41125 4.97125

    In order to ensure that the file contains the information using the correct format, the user can simply run the command:

    /home/las/matcher/input_checker/file_checker.pl <text file with list of entries>

    If there is an error or extraneous character somewhere in each line, this program will detect it and warn the user.

2. Once the list passes all the checks, the user must copy this file to the /home/las/matcher/work directory and rename the file as “sample1.xyt”

3. Then, the user must change to the /home/las/matcher/bin directory and run the matcher command as:

   ./make_matches_V3 <dataset> <interval>

   An example of matching the list with weekly AVHRR GAC SST data would be:

   ./make_matches_V3 PF weekly

Once the processing is completed, an output file with the matched data will be generated in the bin directory, and will be named “FINAL_data_out.asc”. In this output file, each of these lines will have the original “to match” information (line 1 below), as well as the mid-point date and value of the matched SST data (line 2 below):

```
14 Nov 2001 150.833325 20.377075
14 Nov 2001 150.833325 20.377075     08-NOV-2001        28.80
```

Necessary format and contents of the data to be used

```
> head(SST_data)
  Day Month Year      Long      Lat
1  12   Nov 1991 -158.2500 17.06667
2  14   Nov 1991 -157.4500 19.08333
3  15   Nov 1991 -157.7000 19.68333
4  11   Nov 1991 -158.6667 19.66667
5  12   Nov 1991 -158.8333 19.33333
6  13   Nov 1991 -158.8333 19.50000
```

```
> tail(SST_data)
       Day Month Year      Long      Lat
356452  20   Jun 2014 -164.5167 15.01667
356453  21   Jun 2014 -164.5833 15.03333
356454  22   Jun 2014 -164.5167 15.06667
356455  24   Jun 2014 -164.4417 18.04167
356456  26   Jun 2014 -161.6583 19.44167
356457  27   Jun 2014 -161.2833 19.65833
```

*Write out the text file for the SST matcher using “**write.table**”, with white spaces as separators, a three-letter character variable for the month, all other variables numeric, and longitude with a negative sign*.

```
> write.table(SST_data,"SST_matcher_data.txt",sep=" “)
```

Example A 4. Summary of truncations and final preparations conducted with the PIFSC 1990-2014 longline logbook data before initiating accuracy evaluations.

Several truncations were applied to the logbook data before initiating data accuracy evaluations. In general, these ensure that the ranges of operational variables in the logbook data are within those of the PIROP observer data and that missing values are removed.

The initial truncation of the logbook data involved five steps: the evaluation period was defined as 1995-2014; only commercial longline fishing was considered (i.e., no experimental or research fishing); all trips departed from and returned to Hawai‘ian ports; all sets were deployed >200 hooks; all sets were deployed above 1°N and between 130°-179°W. In total, these steps lead to deletion of 20.1% of all logbook data from 1990-2014. Catch data from deleted sets are generally accepted as accurate.

.*Study period*: *1995-2014*

This interval was characterized by PIROP coverage of this longline fleet. The observer data are the model-fitting data (Observer_95_14). Logbook (application) data are named accordingly.

```
> dim(Logsdata)
[1] 356457    100
> Logs_95_14_1<-Logsdata[Logsdata$Haulyr>1994,]
> dim(Logs_95_14_1)
[1] 308144    100

> 356457-308144
[1] 48313
> 100*(48313/356457)
[1] 13.55367
```

The initial truncation to the 20-year study period removes 13.6% of the sets.

```
> table(Logs_95_14_1$Haulyr)
 1995  1996  1997  1998  1999  2000  2001  2002  2003  2004  2005  2006  2007
11732 11638 11846 12505 12805 12931 12186 14110 14883 16029 18195 17302 19385
 2008  2009  2010  2011  2012  2013  2014
19482 18572 17948 18641 19466 19734  8754
```

The table is the annual number of sets since 1995.

*The first four years have been deleted --- **Correct.***

The reasons for this truncation were that 1) *1995 was the first full year of operations for the PIROP*; and 2) *1995 through 2014 was also convenient as a 20-year time series.*

Commercial longline fishing

The data to be evaluated are restricted to commercial longline fishing logbooks.

```
> table(Logs_95_14_1$ResExptCode)
    LL      R      X
305904   2017    223
```

```
> dim(Log_95_14_1)
[1] 308144    103

> Logs_95_14_2<-Log_95_14_1[(!(is.na(Log_95_14_1$ResExptCode))) &
Log_95_14_1$ResExptCode=="LL",]
> dim(Logs_95_14_2)
[1] 305904    103
> 308144-305904
[1] 2240

> dim(Logsdata)
[1] 356457     97
> 2240/308144
[1] 0.0072693
```

Deletion of research and experimental sets removes 0.8% of the longline sets since 1995.
Correct --- checked against PIROP data and numbers of research and experimental sets close.

Hooks>200

```
> Logs_95_14_3<-Logs_95_14_2[Logs_95_14_2$Hooks>200,]
> dim(Logs_95_14_2)
[1] 305904    102
> dim(Logs_95_14_3)
[1] 305652    102
> 904-652
[1] 252

> dim(Logs_95_14_1)
[1] 308144    102
> 100*(252/308144)
[1] 0.08177995
```

Deletion of very low hook numbers removes 0.08% of the longline sets since 1995.

Positions

```
> Logs_95_14_4<-Logs_95_14_3[Logs_95_14_3$Latitude>2 &
(Logs_95_14_3$Longitude>130 & Logs_95_14_3$Longitude<179),]
 > dim(Logs_95_14_3)
[1] 305652    102
> dim(Logs_95_14_4)
[1] 304798    102
 > 305652-304798
[1] 854
> 100*(854/308144)
[1] 0.2771432
```

> Sets near the extremes of the fishery comprised 0.3% of the logbook data since 1995.

SST
summary(Logs_95_14_4$SST)

```
   Min.  1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  14.25   24.05     25.28      24.83   26.18    30.90   24047 
```

The large number of NA rreturns must be evaluated, but there is nothing obviously wrong about the range, quartiles, or measures of central tendency.
No apparent need for an SST truncation.

Two important points pertain to the following: it is necessary to include the negation of NA values for arrival ports; the LOG_HEADER file in ORACLE also includes departure ports, but there are more NA values than for the Arrival ports.

Arrival ports

```
> Logs_95_14_5<-Logs_95_14_4[(Logs_95_14_4$Return_Port=="HI" |
Logs_95_14_4$Return_Port=="HNL" | Logs_95_14_4$Return_Port=="LIH") &
(!(is.na(Logs_95_14_4$Return_Port))),]
> dim(Logs_95_14_4)
[1] 304798    102
> dim(Logs_95_14_5)
[1] 284708    102
 > 304798-284708
[1] 20090
> 20090/308148
[1] 0.06519594
```

Removal of fishing that did not return to Hawaii removes 6.5% of the logbook data since 1995.
The result of these procedures is a logbook data set for 1990-2014 l in the form of a flat file as an R data frame. The Return-date fields retain large numbers of missing values but can be removed at any time. Small numbers of missing values for other fields may be imputed. ***This file is ready for <u>initial use</u>.*** *It is renamed accordingly.*
***N*=284,708 sets (79.9% of the data since 1990; 92.4% since 1995)**
> Logs_95_14<-Logs_95_14_5   (RENAMED DATA FRAME)

APPENDIX B

Analytical procedures used for logbook data accuracy evaluation

Example B 1. Summary of a zero-inflated negative binomial GLM (ZINB) fitted to PIROP observer data for use in evaluation of logbook data accuracy. The initial step entails fitting the ZINB model and checking its residuals to ensure that the model used to evaluate logbook data accuracy has been fitted to accurate observer data.

The large residuals from the initial fit were checked by trips. A total of 42 trips with multiple (2-11) large residuals (Pearson residual>5) were deleted from the ZINB-fitting data on the basis of misidentifications by PIROP observers. These errors have been reported multiple times but remain in the PIROP-reported catch data.

Deleted trips (PIROP numbering system):

“LL0094”, ”LL0105”, “LL0236”, ”LL0294”, ”LL0361”, “LL0396”, “LL0502”, “LL0507”, “LL0516”, “LL0557”, “LL0645”, “LL0657”, “LL1859”, “LL1877”, “LL2065”, “LL2279”, “LL2340”, “LL2344”, “LL2391”, “LL2402”, “LL2403”, “LL2505”, “LL2564”, “LL2718”, “LL2731”, “LL2787”, “LL2801”, “LL2865”, “LL2871”, “LL3020”, “LL3062”, “LL3290”, “LL3517”, “LL4298”, “LL4235”, “LL4264”,”LL3101”,”LL3156”, “LL3169”, “LL2694”

A second data frame without these trips with systematic misidentifications is obtained by **negating** (with an **exclamation point**) trips for selection as follows (**Tripnum** is a character variable):

```
Obsr_Corr_Data<-Observer_95_14[(!(Observer_95_14$Tripnum=="LL0094"|
Observer_95_14$Tripnum=="LL0105" | Observer_95_14$Tripnum=="LL0236"|
Observer_95_14$Tripnum=="LL0292" | Observer_95_14$Tripnum=="LL0361" |
Observer_95_14$Tripnum=="LL0396" | Observer_95_14$Tripnum=="LL0502" |
Observer_95_14$Tripnum=="LL0507"| Observer_95_14$Tripnum=="LL0516" |
Observer_95_14$Tripnum=="LL0557"| Observer_95_14$Tripnum=="LL0645" |
Observer_95_14$Tripnum=="LL0657" | Observer_95_14$Tripnum=="LL1859" |
Observer_95_14$Tripnum=="LL1877" | Observer_95_14$Tripnum=="LL2065" |
Observer_95_14$Tripnum=="LL2279" | Observer_95_14$Tripnum=="LL2340" |
Observer_95_14$Tripnum=="LL2344"| Observer_95_14$Tripnum=="LL2391"|
Observer_95_14$Tripnum=="LL2402"| Observer_95_14$Tripnum=="LL2403"|
Observer_95_14$Tripnum=="LL2505"| Observer_95_14$Tripnum=="LL2564"|
Observer_95_14$Tripnum=="LL2718"| Observer_95_14$Tripnum=="LL2731"|
Observer_95_14$Tripnum=="LL2787"| Observer_95_14$Tripnum=="LL2801"|
```

```
Observer_95_14$Tripnum=="LL2865"| Observer_95_14$Tripnum=="LL2871"|
Observer_95_14$Tripnum=="LL3020"| Observer_95_14$Tripnum=="LL3062"|
Observer_95_14$Tripnum=="LL3290"| Observer_95_14$Tripnum=="LL3517"|
Observer_95_14$Tripnum=="LL4298"| Observer_95_14$Tripnum=="LL4235"|
Observer_95_14$Tripnum=="LL4264"|Observer_95_14$Tripnum=="LL3101"|
Observer_95_14$Tripnum=="LL3156"|Observer_95_14$Tripnum=="LL3169"|
Observer_95_14$Tripnum=="LL3503"| Observer_95_14$Tripnum=="LL3507"|
Observer_95_14$Tripnum=="LL2694")),]

> dim(Obsr_Corr_Data)
[1] 63535   121
> dim(Observer_95_14)
[1] 64230   121
> 64230-63535
[1] 695
> 695/64230
[1] 0.01082049
```

1.1% of the observed sets (0.9% of the observed trips) deleted for systematic misidentifications

R commands to fit a ZINB

```
Loading required package: lattice
> library(pscl)
```

Null model (ZINB)

```
> null_ZINB<-zeroinfl(Blue_marlin~1+offset(log(Hooks)) | 1, data=Observer_95_14,
dist="negbin",link="logit")
> summary(null_ZINB)
Call:
zeroinfl(formula = Blue_marlin ~ 1 + offset(log(Hooks)) | 1, data = Observer_95_14,
dist = "negbin", link = "logit")
```

Pearson residuals:

Min	1Q	Median	3Q	Max
-0.4350	-0.3817	-0.3635	-0.2859	33.3092

Count model coefficients (negbin with log link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.98575	0.01112	-807.94	<2e-16 ***
Log(theta)	-1.21003	0.02254	-53.69	<2e-16 ***

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.55	18.93	-0.557	0.578

Theta = 0.2982

Log-likelihood: -3.746e+04 on 3 Df

Zero-inflated negative binomial GLM (ZINB).

Factors are the haul year and quarter, the fishing region, and the set type. Continuous variables are the SST, number of hooks per float, and begin-set time (*All data, including trips with systematic misidentifications*)

```
> BluMar_ZINB<-
zeroinfl(Blue_marlin~Haulyr1+Quarter1+Region1+Set_type1+SST+offset(log(Hooks)) |
Quarter1+SST+BS_time+Hkpfl,data=Observer_95_14,dist="negbin",link="logit")
> summary(BluMar_ZINB)
```

Call:

zeroinfl(formula = Blue_marlin ~ Haulyr1 + Quarter1 + Region1 + Set_type1 + SST + offset(log(Hooks)) | Quarter1 + SST + BS_time + Hkpfl, data = Observer_95_14, dist = "negbin", link = "logit")

Pearson residuals:

Min	1Q	Median	3Q	Max
-1.0279	-0.4065	-0.2794	-0.1149	41.4775

Count model coefficients (negbin with log

	Estimate	Std. Error	z value	Pr(>|z|)

(Intercept)	-15.27202	0.59296	-25.756	< 2e-16	***
Haulyr11996	-0.41770	0.10548	-3.960	7.50e-05	***
Haulyr11997	-0.35505	0.11777	-3.015	0.00257	**
Haulyr11998	-0.92677	0.11947	-7.757	8.67e-15	***
Haulyr11999	-0.97811	0.14362	-6.810	9.73e-12	***
Haulyr12000	-0.83358	0.09445	-8.825	< 2e-16	***
Haulyr12001	-0.84284	0.08666	-9.726	< 2e-16	***
Haulyr12002	-1.43783	0.08831	-16.281	< 2e-16	***
Haulyr12003	-1.20170	0.08653	-13.887	< 2e-16	***
Haulyr12004	-1.66794	0.08634	-19.318	< 2e-16	***
Haulyr12005	-1.24869	0.08453	-14.773	< 2e-16	***
Haulyr12006	-1.18272	0.08505	-13.906	< 2e-16	***
Haulyr12007	-1.86702	0.08809	-21.193	< 2e-16	***
Haulyr12008	-1.05893	0.08424	-12.571	< 2e-16	***
Haulyr12009	-1.52135	0.08595	-17.700	< 2e-16	***
Haulyr12010	-1.60631	0.08749	-18.359	< 2e-16	***
Haulyr12011	-1.56064	0.08664	-18.013	< 2e-16	***
Haulyr12012	-1.87288	0.08943	-20.942	< 2e-16	***
Haulyr12013	-1.85733	0.08965	-20.718	< 2e-16	***
Haulyr12014	-1.47693	0.08650	-17.074	< 2e-16	***
Quarter12	0.44926	0.06255	7.182	6.87e-13	***
Quarter13	0.36061	0.07182	5.021	5.13e-07	***
Quarter14	0.28693	0.07209	3.980	6.88e-05	***
Region12	-0.25868	0.12466	-2.075	0.03798	*
Region13	-0.36469	0.12484	-2.921	0.00348	**
Region14	0.23010	0.12088	1.904	0.05697	.
Region15	-0.87568	0.12756	-6.865	6.65e-12	***
Region16	-0.29218	0.12621	-2.315	0.02061	*
Region17	-1.02905	0.14564	-7.065	1.60e-12	***
Region18	-0.85285	0.14471	-5.893	3.78e-09	***
Set_type12	1.49350	0.05140	29.055	< 2e-16	***

SST	0.30532	0.02172	14.058	< 2e-16 ***
Log(theta)	0.48109	0.07214	6.668	2.59e-11 ***

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.03471	1.19872	1.697	0.0896 .
Quarter12	-0.12254	0.24281	-0.505	0.6138
Quarter13	0.54881	0.25087	2.188	0.0287 *
Quarter14	0.93950	0.22906	4.102	4.1e-05 ***
SST	-0.34985	0.04110	-8.511	< 2e-16 ***
BS_time	0.21362	0.02318	9.215	< 2e-16 ***
Hkpfl	0.14384	0.01420	10.133	< 2e-16 ***

Theta = 1.6178

Log-likelihood: -3.223e+04 on 40 Df

Zero-inflated negative binomial GLM.

Factors are the haul year and quarter, the fishing region, and the set type. Continuous variables are the SST, number of hooks per float, and begin-set time (*The next set of results are computed after deleting trips with systematic misidentifications).*

```
> null_ZINB_corr<-zeroinfl(Blue_marlin~1+offset(log(Hooks)) | 1, data=Obsr_Corr_Data, dist="negbin",link="logit")
> AIC(null_ZINB_corr)
[1] 72827.13

> BluMar_ZINB_corr<-zeroinfl(Blue_marlin~Haulyr1+Quarter1+Region1+Set_type1+SST+offset(log(Hooks)) | Quarter1+SST+BS_time+Hkpfl,data=Obsr_Corr_Data,dist="negbin",link="logit")
> summary(BluMar_ZINB_corr)
```

Call:

zeroinfl(formula = Blue_marlin ~ Haulyr1 + Quarter1 + Region1 + Set_type1 + SST + offset(log(Hooks)) | Quarter1 + SST + BS_time + Hkpfl, data = Obsr_Corr_Data, dist = "negbin", link = "logit")

Pearson residuals:

Min	1Q	Median	3Q	Max
-1.0876	-0.4043	-0.2751	-0.1050	51.3924

Count model coefficients (negbin with log link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-15.32545	0.53391	-28.704	< 2e-16	***
Haulyr11996	-0.28514	0.10885	-2.620	0.008804	**
Haulyr11997	-0.11443	0.11825	-0.968	0.333195	
Haulyr11998	-0.72964	0.12103	-6.029	1.65e-09	***
Haulyr11999	-0.75554	0.14541	-5.196	2.04e-07	***
Haulyr12000	-0.60009	0.09737	-6.163	7.14e-10	***
Haulyr12001	-0.79876	0.09189	-8.692	< 2e-16	***
Haulyr12002	-1.21696	0.09234	-13.180	< 2e-16	***
Haulyr12003	-0.93397	0.09079	-10.287	< 2e-16	***
Haulyr12004	-1.41817	0.09022	-15.718	< 2e-16	***
Haulyr12005	-1.00892	0.08870	-11.375	< 2e-16	***
Haulyr12006	-1.00131	0.08958	-11.178	< 2e-16	***
Haulyr12007	-1.64112	0.09272	-17.699	< 2e-16	***
Haulyr12008	-0.86944	0.08934	-9.731	< 2e-16	***
Haulyr12009	-1.28502	0.09017	-14.251	< 2e-16	***
Haulyr12010	-1.34100	0.09170	-14.623	< 2e-16	***
Haulyr12011	-1.28799	0.09084	-14.178	< 2e-16	***
Haulyr12012	-1.63095	0.09400	-17.351	< 2e-16	***
Haulyr12013	-1.58719	0.09357	-16.963	< 2e-16	***
Haulyr12014	-1.20032	0.09034	-13.287	< 2e-16	***
Quarter12	0.47155	0.05908	7.981	1.45e-15	***
Quarter13	0.37872	0.06758	5.604	2.09e-08	***

Quarter14	0.24664	0.06914	3.567	0.000361	***
Region12	-0.25406	0.12076	-2.104	0.035384	*
Region13	-0.45386	0.12040	-3.770	0.000163	***
Region14	0.16535	0.11658	1.418	0.156088	
Region15	-0.98345	0.12274	-8.013	1.12e-15	***
Region16	-0.40830	0.12181	-3.352	0.000803	***
Region17	-1.21818	0.14438	-8.437	< 2e-16	***
Region18	-1.03796	0.14370	-7.223	5.07e-13	***
Set_type12	1.60401	0.05046	31.790	< 2e-16	***
SST	0.30069	0.01929	15.592	< 2e-16	***
Log(theta)	0.76105	0.08625	8.824	< 2e-16	***

Zero-inflation model coefficients (binomial with logit link):

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	2.87358	1.07837	2.665	0.007705	**
Quarter12	-0.15754	0.20454	-0.770	0.441157	
Quarter13	0.39453	0.22839	1.727	0.084089	.
Quarter14	0.74329	0.21264	3.495	0.000473	***
SST	-0.35775	0.03869	-9.247	< 2e-16	***
BS_time	0.21143	0.02281	9.268	< 2e-16	***
Hkpfl	0.12627	0.01286	9.817	< 2e-16	***

Theta = 2.1405

Log-likelihood: -3.106e+04 on 40 Df

> AIC(BluMar_ZINB_corr)

[1] 62199.16

Example B 2. Summary of the use of a generalized linear model (GLM), in this case a zero-inflated negative binomial GLM (ZINB), to predict catches and of comparisons of reported to predicted catches in the context of the evaluation of logbook data accuracy.

Application of a GLM (or GAM) object to logbook data necessarily begins with a second series of truncations. The **R predict** function cannot accept "**NA**" values for model covariates, meaning that there cannot be any "**NA**" values in the logbook data frame for the variables in the GLM object. If there are "**NA"** values, these are imputed if possible or deleted if necessary.

In the present example, there were no missing values in the logbook data during 1995-2014 for the dates of fishing (haul year, haul month, haul day), numbers of hooks set, and geographic positions, and moderate numbers of "NA" values for the hooks per float (309) and begin-set time fields. These can be checked and often corrected because values may be missing for some but not all sets on a trip, allowing substitution of the trip average (or median). Also, in the case of hooks per float, this can be investigated by tabulating this field for individual captains using the Commercial Marine License number, as follows: **table(Logsdata$Hkpfl**,**Logsdata**$**CML**). If the CML returns a single value for some captain for the hooks per float, it would mean that the captain in question always sets a certain number of hooks per float (e.g., 30 to target bigeye tuna). Missing values would be replaced in light of these personal tendencies.

Extract a data frame with the NA values for the hooks per float field:

```
> junkHPF<-Logs_95_14[is.na(Logs_95_14$Hkpfl),]
> dim(junkHPF)
[1] 309 102
> summary(Logs_95_14$Hkpfl)
Min.  1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
2.00    24.00     26.00     24.17   30.00     88.00     309
```

Correct

Tabulate these missing values by captains using the CML.

This is conveniently done using the table function to create objects that can be used to trace within-captain patterns of hooks per float use. The results showed that 113 captains had at least one missing hook per float value, but four captains accounted for 29% (**red boldface**) of the total. These were traced to specific trips, as shown below.

```
> table(junkHPF$CML)
 1148  1326  1541  1586  1638  1669  1717  1749  1837  1910  2057  2066  2122
    3     1     1     4     1     1     1     1     1     2     3     2     2
 2141  2151  2189  2204  2205  2281  2285  2299  2413  2502  2503  2504  2561
    1     1     2     1     2     1     5     1     2     4    17     2     1
 2633  2676  2792  2799  2812  2927  3042  3073  3105  3163  3193  3229  3269
    2     1     1     1     6     1     4     1     1     4     4     4     1
 3319  3326  3329  3416  3443  3444  3462  3498  3524  3619  3767  3873  3888
    2     3     1     3     3     1     1     1     1     4    15     3     1
 3931  3979  3987  4022  4032  4035  4085  4189  4211  4228  4303  4361  4379
    1     6     1     1     1     1     3     1     2     3     4     1     1
 4413  4473  4475  4490  4558  4570  4574  4712  4759  4786  4809  4811  4824
    7     1     4     7     1     8     1     1     2     5     1     1    30
 5025  5082  5335  5407  5453  5639  5702  5741  5876  5941  6103  6388  6545
    3     2     4     2     1     5     1     3     1     1     1     1     1
 6741  6781  6819  6898  7340  7356  7394  7516  7713  7729  8202  8274  8481
    1     1     1     1     1     1     2     1     1     1     3     1     1
 8730  8815 10550 10933 11075 12177 12522 90005 90010
    1     1     3     1     1     1     1     1    29
```

The four licenses with 15 or more were checked.

CML 4824

```
> junkHPF1<-Logs_95_14[Logs_95_14$CML==4824 & (!(is.na(Logs_95_14$Hkpfl))),]
> dim(junkHPF1)
[1] 1401  102
> table(junkHPF1$CML,junkHPF1$Hkpfl)
        4    5    6   15   20   22   24    25    26   29
 4824 167   23    9   27    2    8   52   962   122    1
>  junkHPF1<-Logs_95_14[Logs_95_14$CML==4824 & (!(is.na(Logs_95_14$Hkpfl))),]
> table(junkHPF1$Haulyr)
2001
30
> junkHPF2<-junkHPF1[junkHPF1$Haulyr==2001,]
```

> junkHPF2$Hkpfl

[1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA 6 6 6 6 6 6 6 6 6 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 15 NA NA NA

These trips will be deleted because certain trips during 2001 were questionable regarding the actual target species.

> dim(Logs_95_14)

[1] 284708   102

> *Logs_95_14<-Logs_95_14[!((Logs_95_14$Tripnum==426 | Logs_95_14$Tripnum==508 | Logs_95_14$Tripnum==612 | Logs_95_14$Tripnum==693) & Logs_95_14$Haulyr==2001),]*

> dim(Logs_95_14)

[1] 284672   102

> 708-672[1]

36   *Correct*

CML 2503

> junkHPF1<-Logs_95_14[Logs_95_14$CML==2503 & (is.na(Logs_95_14$Hkpfl)),]

> junkHPF1$Tripnum

[1] 141 141 141 141 140 140 140 140 140 140 140 140 140 140 140 140 175

> junkHPF1$Haulyr

[1] 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996

[16] 1996 1997

> junkHPF1$Haulmo

[1] 1 2 2 2 1 1 1 1 1 1 1 1 1 1 1 1 2

> junkHPF1$Haulday

[1] 30 1 3 4 6 7 8 9 10 11 12 13 14 15 16 17 19

> junkHPF1<-Logs_95_14[Logs_95_14$CML==2503 & (!(is.na(Logs_95_14$Hkpfl))) & Logs_95_14$Haulyr==1996,]

> dim(junkHPF1)

[1] 70 102

> summary(junkHPF1$Hkpfl)

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    30      30       30        30      30       30
```

These values were corrected as follows on the basis of within-year consistency.

> Logs_95_14$Hkpfl<-ifelse((is.na(Logs_95_14$Hkpfl)) & Logs_95_14$CML == 2503, 30, Logs_95_14$Hkpfl)

CML 3767

```
> junkHPF1<-Logs_95_14[Logs_95_14$CML==3767 & (!(is.na(Logs_95_14$Hkpfl))),]
> dim(junkHPF1)
[1] 533 102
> table(junkHPF1$CML,junkHPF1$Hkpfl)
          3    4   5   6  24  25  26  27   28   30
  3767    5   69  34   1  10  44  78  23  102  139
```

> *Obtain NAs*

```
> junkHPF1<-Logs_95_14[Logs_95_14$CML==3767 & (is.na(Logs_95_14$Hkpfl)),]
> dim(junkHPF1)
[1]  15 102
> junkHPF1$Tripnum
 [1]  79  79  79  79  79  79  79  79  79  79  79  79  79  79 154
> junkHPF1$Haulyr
 [1] 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996 1996
> junkHPF1$Haulmo
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
> junkHPF1$Haulday
 [1]  6  8  9 12 13 15 16 17 18 19 20 21 22 23  3
summary(junkHPF1$Hkpfl)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    30      30      30        30      30       30
```

CML 90010

```
> junkHPF1<-Logs_95_14[Logs_95_14$CML==90010 & (!(is.na(Logs_95_14$Hkpfl))),]
```

> dim(junknew)

[1] 29 102

> *Correct*

> junknew$Tripnum

 [1] 361 361 361 361 361 361 361 361 361 361 361 361 361 476 476 476 476 476 476 476 476 476 476 476 476 476 476 476 476

> junknew$Haulyr

 [1] 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997 1997

These were two swordfish trips (i.e., shallow-set) trips during 1997. Hkpfl=4 for these trips.

> junknew$Trip_type

 [1] B B B B B B B B B B B B B B B B B B B B B B B B B B B B B

> *Logs_95_14$Hkpfl<-ifelse(Logs_95_14$CML==90010 & Logs_95_14$Trip_type== "B",4, Logs_95_14$Hkpfl*)

> junknew<-Logs_95_14[Logs_95_14$CML==90010,]

> junknew$Hkpfl

[1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 NA NA NA NA NA NA NA NA NA NA NA NA NA NA

*Correct number of changes. The main file (***Logsdata***) would be changed identically.*

The example has shown that three of the four captains with substantial numbers of missing values for hooks per float could be corrected on the basis of patterns. For simplicity, the remaining missing values will be imputed using the sector medians.

> tapply(Logs_95_14$Hkpfl,Logs_95_14$Set_type1,median,na.rm=T)

 1   2

4   27

> Logs_95_14$Hkpfl<-ifelse(((is.na(Logs_95_14$Hkpfl)) & Logs_95_14$Trip_type==T), 30,Logs_95_14$Hkpfl)

> summary(Logs_95_14$Hkpfl)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2.00	24.00	26.00	24.17	30.00	88.00	263

> *Logs_95_14$Hkpfl<-ifelse(((is.na(Logs_95_14$Hkpfl)) & Logs_95_14$Trip_type== "B") ,4, Logs_95_14$Hkpfl*)

> summary(Logs_95_14$Hkpfl)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2.00	24.00	26.00	24.16	30.00	88.00	236

> *Logs_95_14$Hkpfl<-ifelse(((is.na(Logs_95_14$Hkpfl)) & Logs_95_14$Trip_type=="T"), 27, Logs_95_14$Hkpfl*)

> summary(Logs_95_14$Hkpfl)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
2.00	24.00	26.00	24.17	30.00	88.00	67

This shows that 196 of the 263 values were replaceable.

dim(Logs_95_14)

[1] 284672 102

> Logs_95_14<-Logs_95_14[!(is.na(Logs_95_14$Hkpfl)),]

> summary(Logs_95_14$Hkpfl)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	24.00	26.00	24.17	30.00	88.00

> dim(Logs_95_14)

[1] 284605 102

Correct number of deletions.

Begin-set time

> junkBS<-Logs_95_14[is.na(Logs_95_14$BS_Time),]

> table(junkBS$CML)

1036	1534	1541	1590	1725	1726	1760	1813	1837	1839	1910	1967	2018	2030	2038	2055	2065	2066	2122
5	16	1	5	2	2	2	57	18	3	7	14	1	26	19	1	5	3	1

```
2189 2192 2205 2285 2299 2319 2321 2413 2424 2561 2709 2799 2812 2839 2864 3042 3074 3079 3083
4 6 17 9 1 2 1 36 3 1 24 21 5 1 5 21 5 7 3
3105 3218 3269 3286 3443 3496 3498 3619 3677 3767 3774 3839 3873 3929 3979 3987 4034 4085 4189
3 1 4 1 2 4 2 3 4 2 2 1 3 2 4 1 1 1 3
4211 4304 4332 4349 4379 4413 4526 4527 4574 4824 4954 4992 5117 5453 5554 5672 5953 6103 6388
1 1 12 8 8 3 3 2 4 8 3 1 5 4 1 1 5 1 5
6520 6545 6696 6741 6746 6749 6781 6884 6935 7429 8019 8202 8346 8348 8499 9638 9975 9991 10763
3 1 1 3 1 1 5 5 2 1 1 1 11 16 1 1 1 3 1
10899 11875 12522 23150 23151 27309 71921 90017
1 1 1 1 1 1 1 4
> dim(junkBS)
[1] 553 102
> 16+57+18+14+26+19+17+36+24+21+21+12+11+16
[1] 308
> 308/553
[1] 0.556962
```

56% of the missing values are traceable to 14 CMLs. If so desired, these could be traced like the hooks per float. Small numbers of missing set times can be replaced by mean or median within-trip values. For simplicity, the missing values will be deleted in this example.

```
> dim(Logs_95_14)
[1] 284605    102
> Logs_95_14<-Logs_95_14[!(is.na(Logs_95_14$BS_Time)),]
> dim(Logs_95_14)
[1] 284052    102
> 4605-4052
[1] 553
```

Correct number of deletions.

Sea surface temperature (SST°C)

This is the final and most difficult step because the amount of missing data is far greater.

```
> summary(Logs_95_14$SST)
Min.  1st Qu.  Median    Mean   3rd Qu.   Max.    NA's
14.25   24.10     25.30     24.88    26.20     30.90   20527
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> dim(junkSST)
[1] 20527   102
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2010 2011
2936 2353 2579 2830 3072 1983 3345    2      2     378     4      35     12    984    12
```

These are the missing values arranged by year. For simplicity, the period with minimal missing data is considered.

Trips with missing SST values were identified.

```
> junknew<-Logs_95_14[Logs_95_14$Haulyr==2003 & (Logs_95_14$Tripnum==12 |
Logs_95_14$Tripnum==71),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
12                71
25.10647 27.75643
```

These are the mean SST values for the other sets on Trips 12 and 71, respectively. These will be used for imputation. The following scripts do the same for five more years.

```
> junknew<-Logs_95_14[Logs_95_14$Haulyr==2004 & (Logs_95_14$Tripnum==17 |
Logs_95_14$Tripnum==47),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
      17       47
19.35235 21.18000
> junknew<-Logs_95_14[Logs_95_14$Haulyr==2006 & (Logs_95_14$Tripnum==287 |
Logs_95_14$Tripnum==301 | Logs_95_14$Tripnum==315 | Logs_95_14$Tripnum==317),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
     287        301          315          317
18.12375 18.51455 18.23833 18.19077
```

```
> junknew<-Logs_95_14[Logs_95_14$Haulyr==2007 & (Logs_95_14$Tripnum==6 |
Logs_95_14$Tripnum==15 | Logs_95_14$Tripnum==20 | Logs_95_14$Tripnum==24 |
Logs_95_14$Tripnum==36 | Logs_95_14$Tripnum==44 | Logs_95_14$Tripnum==49),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
       6       15       20       24       36       44       49
24.94143 24.75167 24.59571 24.13000 25.28100 24.22167 24.95400

> junknew<-Logs_95_14[Logs_95_14$Haulyr==2008 & (Logs_95_14$Tripnum==15 |
Logs_95_14$Tripnum==20 | Logs_95_14$Tripnum==24 | Logs_95_14$Tripnum==30 |
Logs_95_14$Tripnum==36 | Logs_95_14$Tripnum==49 | Logs_95_14$Tripnum==50),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
      15       20       24       30       36       49       50
25.55867 23.92583 25.42667 24.94125 24.23167 24.91200 24.08583

> junknew<-Logs_95_14[Logs_95_14$Haulyr==2011 & (Logs_95_14$Tripnum==17 |
Logs_95_14$Tripnum==20 | Logs_95_14$Tripnum==195 | Logs_95_14$Tripnum==203 |
Logs_95_14$Tripnum==221 | Logs_95_14$Tripnum==224 | Logs_95_14$Tripnum==229),]
> tapply(junknew$SST,junknew$Tripnum,mean,na.rm=T)
      17       20      195      203      221      224      229
23.72955 24.38273 23.39077 22.70583 22.28300 24.27692 25.33250
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.25   24.10   25.30   24.88   26.20   30.90   20527
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2003) &
(Logs_95_14$Tripnum==12)),25.1,Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25   24.10   25.30   24.88  26.20   30.90   20526
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2003) &
(Logs_95_14$Tripnum==71)), 27.8, Logs_95_14$SST)
> summary(Logs_95_14$SST)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25    24.10     25.30     24.88   26.20     30.90   20525
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2004) &
(Logs_95_14$Tripnum==17)),19.4, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2004) &
(Logs_95_14$Tripnum==47)),21.2, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2006) &
(Logs_95_14$Tripnum==287)),18.1 Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2006) &
(Logs_95_14$Tripnum==301)),18.5, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2006) &
(Logs_95_14$Tripnum==315)),18.2, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2006) &
(Logs_95_14$Tripnum==317)),18.2, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==6)),24.9, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==15)),24.8, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==20)),24.6, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==24)),24.1, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==36)),25.3, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==44)),24.2, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==49)),25, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==1434)),25.6, Logs_95_14$SST)
```

```
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2007) &
(Logs_95_14$Tripnum==1454)),25.4, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==15)),25.6, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==20)),23.9, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==24)),25.4, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==30)),24.9, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==36)),24.2, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==49)), 24.9,Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2008) &
(Logs_95_14$Tripnum==50)),24.1, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==17)),23.7, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==20)),24.4, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==195)),23.4, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==203)),22.7, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==221)),22.3, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==224)),24.3, Logs_95_14$SST)
> Logs_95_14$SST<-ifelse(((is.na(Logs_95_14$SST)) & (Logs_95_14$Haulyr==2011) &
(Logs_95_14$Tripnum==229)),25.3, Logs_95_14$SST)
> summary(Logs_95_14$SST)
```

```
   Min. 1st Qu.  Median    Mean   3rd Qu.    Max.    NA's
  14.25   24.10    25.30      24.88   26.20      30.90   20460
> 527-460
[1] 67
```

Correct number of data substitutions with within-trip mean SST values for five years.

The much larger numbers of missing values in 2005 (378) and 2010 (984) are dealt with more efficiently. Data frames are created for each year and the within-trip mean SST values are computed. A match is created between the main logbook data frame and the within-year data frames that include the within-trip means as SST values. These means are matched to the trips with missing values to substitute for the latter.

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002 2005 2010
2936 2353 2579 2830 3072 1983  3345  378  984
```

Reversed --- obtain sets with SST

```
> junkSST<-Logs_95_14[Logs_95_14$Haulyr==2005,]
> SSTmean<-tapply(junkSST$SST,junkSST$Tripnum,mean,na.rm=T)
> SSTtripnum<-tapply(junkSST$Tripnum,junkSST$Tripnum,unique)
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> mode(SSTdata$SSTtripnum)
[1] "numeric"
> sum(SSTdata$SSTtripnum)
[1] 901824
> mode(Logs_95_14$Tripnum)
[1] "numeric"
> sum(Logs_95_14$Tripnum)
[1] 182081218
 > SSTdata$Tripnum<-SSTdata$SSTtripnum
> sum(SSTdata$Tripnum)
[1] 901824
```

```
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2005
,SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25    24.10      25.30     24.88   26.20     30.90   20088
> 460-88
[1] 372
```

Check for six missing values.

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002 2005 2010
2936 2353 2579 2830 3072 1983 3345    6     984
```

Values were traced to one trip; the return was “NaN“ (i.e., not computed, not a number.)

```
> dim(Logs_95_14)
[1] 284052    102
> Logs_95_14<-Logs_95_14[!(Logs_95_14$Haulyr==2005 & Logs_95_14$Tripnum==1395),]
> dim(Logs_95_14)
[1] 284046    102
```

Correct number of deletions.

```
> junkSST<-Logs_95_14[Logs_95_14$Haulyr==2010,]
> SSTmean<-tapply(junkSST$SST,junkSST$Tripnum,mean,na.rm=T)
> SSTtripnum<-tapply(junkSST$Tripnum,junkSST$Tripnum,unique)
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2010,
SSTdata$SSTmean[match1],Logs_95_14$SST)
```

```
> summary(Logs_95_14$SST)
Min.   1st Qu.   Median    Mean  3rd Qu.   Max.    NA's
14.25    24.10     25.30      24.88   26.20    30.90   19106
> 20088-19106
[1] 982
```

Probably like 2005.

```
> dim(Logs_95_14)
[1] 284046    102
```

Years with many missing values treated comparably.

2010

```
> Logs_95_14<-Logs_95_14[(!(Logs_95_14$Haulyr==2010 & (is.na(Logs_95_14$SST)))),]
> dim(Logs_95_14)
[1] 284038    102
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.25   24.10    25.30     24.88   26.20     30.90   19098
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002
2936 2353 2579 2830 3072 1983 3345
```

1996

```
> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1996,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1996),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 1076
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1076
> SSTdata<-data.frame(SSTtripnum,SSTmean)
```

```
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1996,
SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25    24.08     25.30     24.86   26.20     30.90   16268
```

1997

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997  1998 1999 2000 2001 2002
 106  2353  2579 2830 3072 1983 3345

> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1997,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1997),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 1111
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1111
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) &
Logs_95_14$Haulyr==1997,SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.25   24.06    25.28     24.84   26.18     30.90   13991
```

1998

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002
 106   76 2579 2830 3072 1983 3345

> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1998,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1998),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 1134
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1134
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1998,
SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25    24.05     25.28      24.83   26.18    30.90   11439
```

1999

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002
 106    76     27   2830 3072 1983 3345
> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1999,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1999),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
```

```
> length(SSTmean)
[1] 1127
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1127
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==1999,
SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
14.25    24.03    25.27    24.81   26.17    30.90    8670

> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002
 106    76    27    61   3072 1983 3345
> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2000,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2000),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 1095
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1095
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2000,
SSTdata$SSTmean[match1],Logs_95_14$SST)
```

```
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
14.25    24.02     25.27     24.80   26.17     30.90    5635
```

2001

```
> junkSST<-Logs_95_14[(is.na(Logs_95_14$SST)),]
> table(junkSST$Haulyr)
1996 1997 1998 1999 2000 2001 2002 
 106    76     27     61     37   1983 3345 

> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2001,]
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2001),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 869
>  SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 869
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2001,
SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
  14.25   24.03    25.27     24.81   26.18     30.90    3672
```

2002

```
> junkSST1<-Logs_95_14[(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2002,]
```

```
> junkSST2<-Logs_95_14[(!(is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2002),]
> SSTmean<-tapply(junkSST2$SST,junkSST2$Tripnum,mean,na.rm=T)
> length(SSTmean)
[1] 1117
> SSTtripnum<-tapply(junkSST2$Tripnum,junkSST2$Tripnum,unique)
> length(SSTtripnum)
[1] 1117
> SSTdata<-data.frame(SSTtripnum,SSTmean)
> SSTdata$Tripnum<-SSTdata$SSTtripnum
> match1<-match(Logs_95_14$Tripnum,SSTdata$Tripnum)
> Logs_95_14$SST<-ifelse((is.na(Logs_95_14$SST)) & Logs_95_14$Haulyr==2002,
SSTdata$SSTmean[match1],Logs_95_14$SST)
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  14.25   24.05   25.28   24.82   26.18   30.90     397
```

Truncations performed to make prediction with R feasible:

```
> dim(Logs_95_14_5)
[1] 284708    102
> dim(Logs_95_14)
[1] 284038    102
```

Removal of remaining missing SST values.

```
> Logs_95_14<-Logs_95_14[!(is.na(Logs_95_14$SST)),]
> dim(Logs_95_14)
[1] 283641    102
> 4708-3641
[1] 1067

> 1067/284708
[1] 0.003747699
```

Missing SST values that could not be replaced by within-trip means comprised 0.3% of the sets.

Creating factors in log data

```
> summary(Logs_95_14$Haulyr)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1996    2002    2006     2006    2010    2014
> summary(Logs_95_14$Haulyr1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1996    2002    2006     2006    2010     2014
> Logs_95_14$Haulyr1<-as.factor(as.numeric(Logs_95_14$Haulyr1))
> summary(Logs_95_14$Quarter)
   Min. 1st Qu.  Median    Mean  3rd Qu.    Max.
  1.000   1.000    2.000     2.471   4.000     4.000
> summary(Logs_95_14$Quarter1)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000     2.471   4.000     4.000
> Logs_95_14$Quarter1<-as.factor(as.numeric(Logs_95_14$Quarter1))
> summary(Logs_95_14$Region)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   5.000     4.764   6.000   8.000
> summary(Logs_95_14$Region1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   5.000     4.764   6.000   8.000
> Logs_95_14$Region1<-as.factor(as.numeric(Logs_95_14$Region1))
> summary(Logs_95_14$Haulyr1)
 1996  1997   1998    1999  2000   2001   2002    2003   2004   2005  2006    2007   2008
11339 11607 12278 12508 12774 10023 13397 14109 15832 15630 15910 18720 19005
 2009   2010   2011   2012   2013   2014
18257 17443 18136 19050 19211  8412
```

Checks on continuous variables. No missing values indicates that prediction should work.

```
> summary(Logs_95_14$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.25   24.05   25.28   24.82   26.18   30.90
> summary(Logs_95_14$Hkpfl)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   24.00   26.00   24.21   30.00   88.00
> summary(Logs_95_14$BS_Time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   7.067   8.000     9.186   9.000    24.000
```

Removal of observed sets

```
> table(Logs_95_14$Observed_set)
     0         1
223984  59657
```

```
> Logs_95_14_Apply1<-Logs_95_14[Logs_95_14$Observed_set==0,]
> table(Logs_95_14_Apply1$Observed_set)
     0
223984
```

A new data frame (**Logs_95_14_Apply1**) *contains only unobserved logbook sets with no missing values for any of the covariates in the prediction model (BluMar_ZINB)?*

Example B_3. Prediction of catches using a zero-inflated negative binomial GLM for blue marlin to serve as a comparison standard for unobserved fishing trips.

```
> summary(Logs_95_14_Apply1$Latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.067  17.230  21.410  21.380  25.720  45.360
> summary(Obsr_Corr_Data$Latitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.275  18.230  23.240  23.010  28.300  44.560

> summary(Logs_95_14_Apply1$Longitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  130.0   154.8   159.4     158.7    163.0   179.0
> summary(Obsr_Corr_Data$Longitude)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  134.6   154.0    158.7     158.1   162.8    179.9

> summary(Logs_95_14_Apply1$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.25   24.17   25.33   24.97   26.20   30.70
> summary(Obsr_Corr_Data$SST)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.74   23.55    25.17     24.37   26.18    30.60
```

These summaries suggest some further truncations may be needed. Check predictions.

The **R predict** *function is applied to the GLM object; "***type=response***" obtains the response variable back-transformed to the original units; and* **newdata**="**Logs_95_14_Apply1"** *identifies the data frame to which the model coefficients are applied.*

```
> test<-predict(BluMar_ZINB_corr, type="response", newdata=Logs_95_14_Apply1,
link="logit", family="negbin")
> length(test)
[1] 223984
> summary(test)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 0.0001  0.0737  0.1628  0.2802  0.3658  7.9420     216
> summary(Logs_95_14_Apply1$Blue_marlin)
   Min.   1st Qu.  Median    Mean   3rd Qu.    Max.
 0.0000  0.0000   0.0000    0.3254  0.0000  44.0000
```

Check on additional truncation

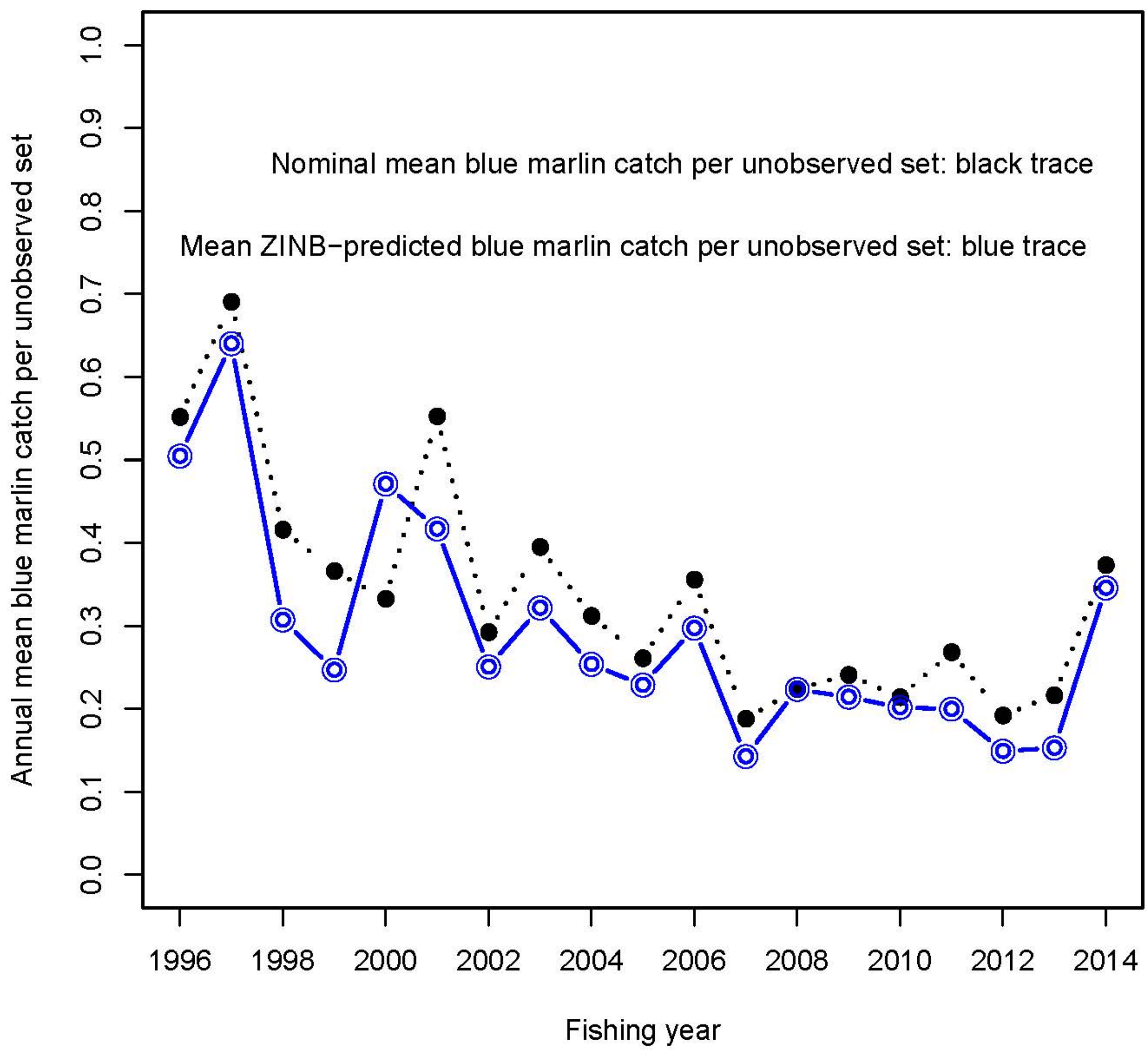
```
Logs_95_14_Apply2<-Logs_95_14_Apply1[(Logs_95_14_Apply1$Latitude>3 &
Logs_95_14_Apply1$ Latitude<43)  & Logs_95_14_Apply1$Longitude>135,]
> Logs_95_14_Apply2<-Logs_95_14_Apply1[(Logs_95_14_Apply1$Latitude>3 &
Logs_95_14_Apply1$Latitude<43) & Logs_95_14_Apply1$Longitude>135,]
> test_Apply2<-predict(BluMar_ZINB_corr, type="response", newdata=Logs_95_14_Apply2,
link="logit", family="negbin")
> summary(test_Apply2)
   Min.    1st Qu.   Median    Mean   3rd Qu.   Max.     NA's
0.00010 0.07383 0.16290 0.28020 0.36580 7.94200     215
> dim(Logs_95_14_Apply1)
[1] 223984    103
> dim(Logs_95_14_Apply2)
[1] 223701    103
> 984-701
[1] 283
```

The very small changes in the predictions indicate that the additional truncation is not required.

Use of the ZINB model to predict catches to serve as comparison standards for logbooks from unobserved fishing trips is presented in **Example_B_3_Figure 1**. The trends illustrate ongoing pattern of upward bias in the logbook data caused by species misidentifications. The apparent deviation from the trend in 2000 probably resulted from an extrinsic factor. Shortbill spearfish were extremely numerous in the catch throughout the autumn and winter months of 2000, and captains may have misidentified blue marlin as shortbill spearfish from habit because the latter were at such a high level of seasonal abundance (Walsh & Bigelow: "Where the Billfishes Were (and Were Not)." Presentation to the 56th International Conference, Lake Arrowhead, CA, May 2005).

The remainder of the procedure is straightforward. For convenience, the vector of predictions can be assigned to the data frame subject to prediction. The linear regression of the reported values on the predicted values is computed (possibly using appropriate transformations), and the studentized residuals or other objective criterion obtained from the **R** regression object. These can be tabulated by CML, with individuals with large numbers of potential outliers investigated first. See Crawley (2013) for details of prediction from a GLM in R.

Example_B 3_Figure 1. Comparison plot depicting use of a zero-inflated negative binomial GLM to predict longline catches to serve as a comparison standard for unobserved fishing.



APPENDIX C

Forms used for research in the Hawaii longline fishery

Current PIFSC logbook form

PIROP Observer Catch Event Log

PIROP Observer Gear Configuration Form

PIROP Observer Set and Haul Information Form

PIFSC Logbook Form

OMB Control No. 0648-0214, Expiration Date: 08/31/2018

PACIFIC ISLANDS DAILY LONGLINE FISHING LOG No.

VESSEL ____________ **PERMIT NUMBER** ____________

Date of Departure from Port: ___/___/_____ **Port:** ____________

Date of Return to Port: ___/___/_____ **Port:** ____________

NMFS USE ONLY
HL Trip type: ________
HL Trip no.: ________
OT: ________ OS: ____

SET INFORMATION Side set [ ] Observer on Board: [ ]

DATE OF SET: ___/___/_____ **Target species:** [ ]Tuna [ ]Swordfish

Number of Hooks Set ________ hooks **Length of Mainline Set** ______ miles **Bait Type** __________

Hooks per Float *(minimum)* ______ *(maximum)* ______ **Number of Lightsticks** ________

BEGIN SET Time: ____:____ **Position:** ______° ____' N/S Latitude; ______° ____' E/W Longitude

END SET Time: ____:____ **Position:** ______° ____' N/S Latitude; ______° ____' E/W Longitude

HAUL INFORMATION

DATE OF HAUL: ___/___/_____

BEGIN HAUL Time: ____:____ **Position:** ______° ____' N/S Latitude; ______° ____' E/W Longitude

END HAUL Time: ____:____ **Position:** ______° ____' N/S Latitude; ______° ____' E/W Longitude

Number of Hooks Lost: ______ hooks

PELAGIC SPECIES				PROTECTED SPECIES				
		NUMBER OF FISH				NUMBER RELEASED		
TUNAS:		Kept	Released			Uninjured	Injured	Dead
Albacore (*tonbo*)	15			SEALS:				
Bigeye tuna	16			Monk Seal	51			
Yellowfin tuna	17			Sea Lions	72			
Skipjack tuna (*aku*)	22			Other Seals	62			
Bluefin tuna	19			DOLPHINS:				
BILLFISH:				Bottlenose	70			
Blue marlin	1			Spinner	71			
Striped marlin (*nairagi*)	2			Other dolphins	50			
Sailfish	4			WHALES:				
Spearfish (*hebi*)	5			Humpback	66			
Swordfish (*broadbill*)	6			False Killer	52			
Other marlin (specify):	32			Other whales	58			
OTHER PELAGICS:				TURTLES:				
Mahimahi	11			Green	53			
Moonfish (*opah*)	12			Leatherback	54			
Wahoo (*ono*)	13			Loggerhead	60			
Oilfish (*walu*)	20			Olive ridley	59			
Pomfret (*monchong*)	21			Hawksbill	64			
Other pelagics (specify):	14			Unidentified hardshell	65			
SHARKS:				BIRDS:				
Blue shark	7			Black-foot Albatross	73			
Mako shark	8			Laysan Albatross	74			
Thresher shark	9			Short-tailed Albatross	75			
Oceanic white-tip shark	24			Other birds (specify):	63			
Silky shark	25			SHARK:				
Other Shark (specify):	10			Scalloped Hammerhead	77			

I certify that the above information is complete and true to the best of my knowledge:

VESSEL CAPTAIN/OPERATOR: Print name: ____________________

CML: __________ Signature: ____________________ Date: __________

Observer Catch Event Log

Observer ID

Haul Date: Day | Month | Year 2 0

DOC/NOAA Fisheries
Pacific Islands Region
Longline Observer Program

Catch Event Log

Trip No.

Set No.

This Catch Page No.

Log comments for specific Catch Log records on the back of this form.

Page No	Line No.	Species Common Name	Species Code	Float No.	Hook No.	Caught Condition Code (A, D, I, U)	Kept/Return Code (K/A, D, F, I, U)	Damaged Code	Gender Code (M, F, U)	Length 1 Code	Length 1 Measurement	Length 2 Code	Length 2 Measurement	Length 3 Code	Length 3 Measurement	✓ Tag(s)?	✓ Specimen(s)?	✓ Photo(s)?	✓ Sketch(es)?	✓ Comment(s)?
	1															☐	☐	☐	☐	☐
	2															☐	☐	☐	☐	☐
	3															☐	☐	☐	☐	☐
	4															☐	☐	☐	☐	☐
	5															☐	☐	☐	☐	☐
	6															☐	☐	☐	☐	☐
	7															☐	☐	☐	☐	☐
	8															☐	☐	☐	☐	☐
	9															☐	☐	☐	☐	☐
	10															☐	☐	☐	☐	☐
	11															☐	☐	☐	☐	☐
	12															☐	☐	☐	☐	☐
	13															☐	☐	☐	☐	☐
	14															☐	☐	☐	☐	☐
	15															☐	☐	☐	☐	☐

Most Common Fish

Code	Common Name
216	Yellowfin Tuna
212	Skipjack Tuna
215	Albacore Tuna
211	Bigeye Tuna
301	Swordfish
302	Striped Marlin
303	Shortbill Spearfish
424	Bigeye Thresher Shark
418	Blue Shark
191	Escolar
221	Wahoo
193	Snake Mackerel
144	Opah
185	Sickle Pomfret
121	Lancetfish
218	Dolphinfish

Most Common Protected Species

Code	Common Name
681	Black-Footed Albatross
682	Laysan Albatross
601	Other Identified Bird
504	Loggerhead Sea Turtle
505	Olive Ridley Sea Turtle
506	Leatherback Sea Turtle
502	Green Sea Turtle
500	Unid. Hard Shell Sea Turtle
742	False Killer Whale
746	Risso's Dolphin
743	Short-Finned Pilot Whale
731	Bottlenose Dolphin
755	Humpback Whale

Code	Damage
BD	Bird Damage
CC	Cookie Cutter damage
CO	Other Damage, see comments
MM	Marine mammal damage
SB	Shark damage to body
SH	Shark damage - Head on hook
ST	Shark damage to tail
SQ	Squid damage
UN	Undetermined source of damage
ND	Observation shows No Damage
SW	Swordfish Damage

Note: Code CO must have comments but others may need comments also.

Measure fish logged on gray lines (1,4,7,10,and 13). For all fish recorded Out of protocol (white lines), prefix the measurement code with 'O'; e.g. 'OFL'.

Measurement Codes
- AL Approximate Fork Length (ft)
- FL Fork Length (cm)
- EF Eye to Fork (cm)
- CI Clasper Inner Length (cm)
- PC Pre-Caudal

Measurement Protocols
- Sharks: FL, PC, CI
- Billfish: EF
- All Other Fish: FL

Caught Condition Codes
- A Caught Alive
- D Caught Dead
- I Caught Injured
- U Caught Condition Unknown

Kept/Return Codes
- K Kept
- A Returned alive
- D Returned dead
- F Returned Finned
- I Returned Injured
- U Returned, Unknown condition

Gender Codes
- M Male
- F Female
- U Unknown

A blank Gender field indicates Unknown.

form v. CL.10.01

Observer Gear Configuration Form

Observer ID

OMB Control No. 0648-0593 exp. 10/31/2018

DOC/NOAA Fisheries
Pacific Islands Region
Longline Observer Program

Trip No.

Set No.

Gear Configuration

Hooks/Floats

06 Other
08 Offset Round Circle
09 Offset Flat Circle

Hook Characteristics

Hook Type Code	Hook Sizes	Hook Diameter (mm)	Hook %

No. Floats
Hooks Per Float
No. Hooks Set

Fishing Techniques

Target Species Code

Name:

Bait Code 0

01 Large Squid
02 Small Squid
03 Saury (Sanma)
04 Mackerel (Saba)
05 Mixed
06 Other
07 Sardine

Light Devices

Type Code 0

00 None
01 Light Stick
02 Glow Bead
03 Other

No. Devices

Color Code

01 Blue
02 Green
03 Black
04 Pink
05 White
06 Yellow
07 Magenta
08 Mixed
09 Other
10 Clear
11 Red
12 Orange
13 Silver/Metal

Main Line

Material Code 0

01 Mono
02 Multi
03 Other

Diameter . mm

Reported Length nm

Reported Test lbs

No. Strands

Color Code

01 Blue
02 Green
03 Black
04 Pink
05 White
06 Yellow
07 Magenta
09 Other
10 Clear
11 Red
12 Orange
13 Silver/Metal

Float Line

Material Code 0

01 Mono
02 Multi
03 Other

Diameter . mm

Measured Length . m

Branch Line

Material Code 0

01 Mono
02 Multi
03 Other

Diameter . mm

Measured Length . m

Reported Test lbs

No. Strands

Color Code

01 Blue
02 Green
03 Black
04 Pink
05 White
06 Yellow
07 Magenta
09 Other
10 Clear
11 Red
12 Orange
13 Silver/Metal

Leader

Material Code 0

01 Mono
02 Wire
03 Other

Diameter . mm

Measured Length . m

Reported Test lbs

Weight Size g

form v. GC.13.02

Observer Set and Haul Information Form

Observer ID

OMB Control No. 0648-0593 exp. 10/31/2018

DOC/NOAA Fisheries
Pacific Islands Region
Longline Observer Program

Trip No.

Set No.

Logbook Page No.

Set and Haul Information

Begin Set

Date/Time: Day | Month | Year 2 0 | Hour | Minute

End No. 1

Latitude: Deg. | Decimal Min. . | N/S

Longitude: Deg. | Decimal Min. . | E/W

Weather Code

Beaufort Scale

Sea Surface Temperature . Degrees F.

End Set

Date/Time: Day | Month | Year 2 0 | Hour | Minute

End No. 2

Latitude: Deg. | Decimal Min. . | N/S

Longitude: Deg. | Decimal Min. . | E/W

Weather Code

Beaufort Scale

Sea Surface Temperature . Degrees F.

Begin Haul

Date/Time: Day | Month | Year 2 0 | Hour | Minute

Latitude: Deg | Decimal Min. . | N/S

Longitude: Deg. | Decimal Min. . | E/W

Weather Code

Beaufort Scale

Sea Surface Temperature . Degrees F.

End Haul

Date/Time: Day | Month | Year 2 0 | Hour | Minute

Latitude: Deg | Decimal Min. . | N/S

Longitude: Deg. | Decimal Min. . | E/W

Weather Code

Beaufort Scale

Sea Surface Temperature . Degrees F.

Weather Codes

01 Clear
02 Partly cloudy
03 Layers of clouds
04 Drizzle
05 Showers
06 Rain
07 Thunderstorms
08 Rain and fog
09 Fog/thick haze
10 Snow, rain/snow mix
99 Other

Beaufort Scale

00	Surface like a mirror	0 ft
01	Ripples like scales, no foam	.25 ft
02	Sm. wavelets, glassy crests	.50 ft
03	Lg wavelets, some whitecaps	2 ft
04	Sm, waves, numerous whitecaps	4 ft
05	Mod. waves, some spray	6 ft
06	Lg. waves, more spray	10 ft
07	Sea heaps up, spray & foam	14 ft
08	Mod. waves, foam in streaks	18 ft
09	High waves, rolling, reduced vis.	23 ft
10	Very high waves, hanging crests, heavy rolling	29 ft

Set/Haul Events

Haul Back Dir. Code 0

01 Begin Set
02 End Set
03 Other

Line Parted? ☐ ✓

No. Sections Retrieved

Protected Species Interactions

During Set? ☐ ✓
During Haul? ☐ ✓

Comments

form v. SH.12.12

Availability of NOAA Technical Memorandum NMFS

Copies of this and other documents in the NOAA Technical Memorandum NMFS series issued by the Pacific Islands Fisheries Science Center are available online at the PIFSC Web site http://www.pifsc.noaa.gov in PDF format. In addition, this series and a wide range of other NOAA documents are available in various formats from the National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, U.S.A. [Tel: (703)-605-6000]; URL: http://www.ntis.gov. A fee may be charged.

Recent issues of NOAA Technical Memorandum NMFS–PIFSC are listed below:

NOAA-TM-NMFS-PIFSC-56 2012 economic cost earnings of pelagic longline fishing in Hawaii.
K.O. KALBERG and M. PAN
(October 2016)

55 Hawaii Marine Recreational Fishing Survey: a summary of current sampling, estimation and data analyses
H. MA and T.K. OGAWA
(September 2016)

54 Proceedings of the 2015 international summit on fibropapillomatosis: global status, trends, and population impacts.
S. HARGROVE, T. WORK, S. BRUNSON, A.M. FOLEY and G. BALZAS
(August 2016)