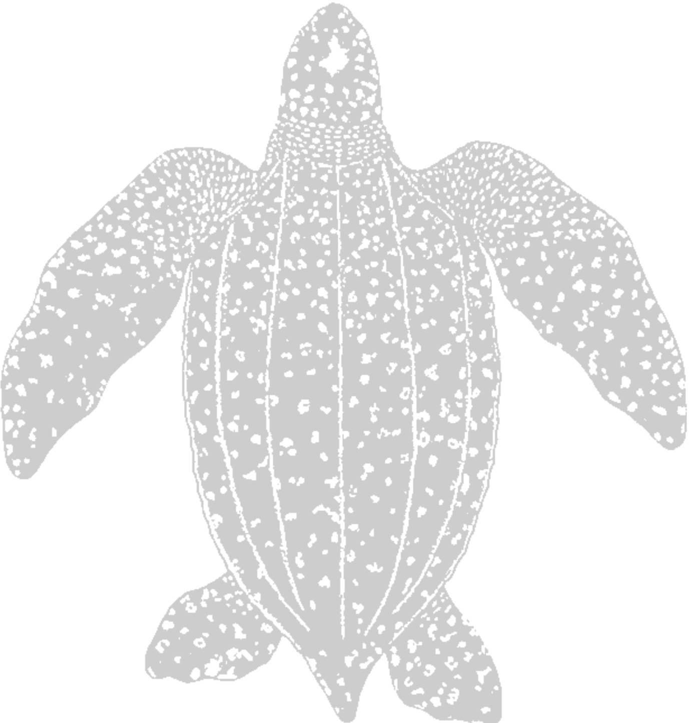
NOAA Technical Memorandum NMFS-PIFSC-3

November 2004

Modeling a Very Rare Event to Estimate Sea Turtle Bycatch: Lessons Learned



Marti L. McCracken

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:

> Author. Date. Title. U.S. Dep. Commer., NOAA Tech. Memo., NOAA-TM-NMFS-PIFSC-*XX*, *xx* p.

For further information direct inquiries to

Chief, Scientific Information Services
Pacific Islands Fisheries Science Center
National Marine Fisheries Service
National Oceanic and Atmospheric Administration
U.S. Department of Commerce
2570 Dole Street
Honolulu, Hawaii 96822-2396

Phone: 808-983-5386
Fax: 808-983-2902

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

Modeling a Very Rare Event to Estimate Sea Turtle Bycatch: Lessons Learned

Marti L. McCracken

Pacific Islands Fisheries Science Center
2570 Dole Street
Honolulu, Hawaii 96822-2396

NOAA Technical Memorandum NMFS-PIFSC-3

November 2004

MODELING A VERY RARE EVENT TO ESTIMATE SEA TURTLE BYCATCH: LESSONS LEARNED

MARTI L. MCCRACKEN

ABSTRACT

Estimation of sea turtle bycatch in the Hawaii-based pelagic longline fishery is discussed in the context of modeling a very rare event using heirarchical catch data collected by longline vessel captains and NMFS observers. Problems in bycatch model formulation, identification of efficient predictor variables, model selection, and model diagnostics are explored in detail. Models to predict bycatch of leatherback, olive ridley, and loggerhead sea turtles are developed using a variety of statistical tools including classification trees, generalized linear models, and generalized additive models. Prediction intervals for bycatch are derived using a nonparametric bootstrap algorithm. The statistical methods are applied to estimate annual bycatch and corresponding prediction intervals for all three turtle species in the years 1994-1999. Problems encountered in all aspects of the research and their resolution are discussed at length. Unresolved statistical issues are identified and suggestions for improving turtle bycatch estimation methods are offered.

Key Words: AIC; BIC; Bootstrap simulation; Bycatch; Classification trees; Generalized additive models; Generalized linear models; Hierarchical data; Longline fishery; Prediction intervals.

INTRODUCTION

Methodology for modeling the occurance of a rare event is well established, but when the event is extremely rare and the data are hierarchical, many of the commonly used modeling techniques are unsatisfactory. In the literature, there are comments concerning modeling a very rare event but there is no comprehensive paper on modeling a very rare event when the data are hierarchical. This problem was encountered when using a model-based approach to estimate total sea turtle bycatch in the Hawaii-based pelagic longline fishery. Bycatch is counted and recorded for each fishing operation within a fishing trip so the data are hierarchical; furthermore, turtle bycatch is a very rare event with 96% to 99% of the counts being zero. Herein, the dilemmas encountered when modeling hierarchical data of an extremely rare event and strategies to overcome these predicaments are discussed. Topics discussed include possible model types, creating more efficient predictor variables, model selection, and model diagnostics. To approximate prediction intervals for turtle bycatch a bootstrap algorithm is developed. Prediction models for turtle bycatch are selected, fitted, and used to estimate the total annual turtle bycatch for the years from 1994 to 1999.

The reason for modeling turtle bycatch is to derive required annual estimates of the total bycatch for all trips landing in Hawaii from 1994 to 1999 by registered Hawaii-based longline fishing vessels. Loggerhead, leatherback, olive ridley, and green turtles are occasionally hooked or entangled on the longline gear accidently during a fishery operation (set). Because these four species of turtles are listed as endangered a Biological Opinion authorizing the fishery is required under Section 7 of the Endangered Species Act. The Biological Opinion includes an incidental take statement based on the anticipated take for the fishery. Annual estimates of bycatch are compared to these numbers to determine if the fishery has exceeded the anticipated take. If so a formal review is undertaken to determine if the fishery is threatening the survival of the species of concern. Herein, a turtle bycatch (take) is defined as a turtle that is hooked or entangled on the longline upon retrieval. To obtain records of turtle bycatch for a sample of trips, an observer program was established in 1994. This program places trained observers on selected trips of the longline fishery. During these trips, observers record turtle bycatch and other pertinent information for each fishing operation.

This document is an updated and expanded version of a previous report (McCracken, 2000) concerning these estimates. It describes subsequent work undertaken to further understand the complexities of modeling hierarchical data of an extremely rare event. The bycatch of green turtles is not discussed in this document because there were too few observed takes to pursue a model with explanatory variables (McCracken, 2000).

Estimating a characteristic of the whole population using observations from part of the population is a familiar problem. One common approach is to draw a probability sample and use an estimator based on the sampling probabilities to estimate the characteristic. In practice, it is often difficult to design an affordable, fair, and practical probability sampling survey of a fishing fleet's activities. Some of the typical obstacles are that (1) a list of trips and their departure dates typically does not exist beforehand, (2) the composition and activity level of a fleet can change over time, and (3) the availability of observers is limited and can fluctuate.

Alternatively, an estimator based on an assumed prediction model can be used. An appropriate model to assume is one that accurately represents the process generating the characteristic of interest; i.e., it is model-unbiased or the bias is small enough to be inconsequential. If an appropriate model does not exist beforehand, as in our case, the idea is to use available empirical data to find a good approximating model for bycatch. This model-based approach does not require a probability sample, although some manner of randomization in the selection of a sample is typically advantageous (Hansen, Madow, and Tepping, 1983).

Because of limited funding, only about 5% of the annual trips by Hawaii-based longliners were sampled from 1994 to 1999. During this period different sampling schemes were used, but a probability sample was not obtained. Therefore, to estimate total bycatch I used a prediction model that predicted a set's bycatch of the species of concern. During model development, decisions were made concerning suitable model types, selecting and fitting models, and how to estimate uncertainty in the predicted bycatch. When making these decisions it was important to consider the data structure and properties of the response variable (the quantity to be modeled). Three characteristics of the bycatch data to be considered were that (1) the response variable, turtle bycatch, was a rare event, (2) the data were hierarchical, and (3) many of the potential predictor variables were related. Because of these complexities, model building was not straightforward and modeling techniques commonly used were unsatisfactory. In the next section, the data structure is described in more detail. The model types that I considered are presented in Section 3. For the selected model types, model fitting is reviewed in Section 4. Creating prediction variables, model selection, and estimation of prediction intervals are discussed in Sections 5, 6, and 7, respectively. In all these sections, I emphasize the effect that modeling hierarchical data of a very rare event had on the process and the dilemmas that were faced. In Section 8, the predictions of bycatch and a brief discussion concerning the interpretation of the prediction models are presented. I conclude this paper with a discussion of how the estimates of turtle bycatch might be improved in the future.

THE RESPONSE VARIABLE AND DATA STRUCTURE

To select an appropriate model type the properties of the response variable, turtle bycatch, and the data structure must be considered. The data used to develop prediction models for bycatch were created by matching the records of observed longline sets to corresponding daily logbook records kept and reported by longline vessel captains for any fishing operation. Possible predictor variables from the logbook records and the recorded bycatch from the observer records were then extracted. Turtle bycatch was a rare event: out of 3,107 sets from 266 observed trips from 1994 to 1999, only 32 olive ridley, 33 leatherback, and 142 loggerhead bycatches were recorded. Since turtle bycatch represents counts of a rare event, a natural distribution to consider is the Poisson distribution. The zero-inflated Poisson (ZIP) distribution should also be considered. A ZIP variate arises when the stochastic process generates a Poisson variate or a value of zero. If no turtles are exposed to the longline, then by default the bycatch is zero, and the variable bycatch may exhibit more zeros than expected for a Poisson variate.

Concerning the data structure, information on turtle bycatch was collected by placing observers on selected vessels for the duration of a trip. During observed trips, the bycatch for each longline set was recorded. Thus, the independent primary unit of observation was the trip, and the sets within a trip were subunits. Data with this structure are commonly referred to as hierarchical. Since sets within a trip are typically close together in time and space and similar in fishing style, we expect bycatches within a trip to be more closely related than bycatches across trips. Therefore, when creating a model I considered the possibility of stochastic dependence among sets from the same trip. For simplicity, I could have avoided this dependence and modeled bycatch at the trip level, but information concerning the predictor variables would have been lost. For example, latitude and longitude were recorded for each set, and these values would have needed to be summarized for the trip. Hence, modeling at the set level is likely to be more informative, but results can be misleading if the hierarchical structure of the data is ignored and independent observations assumed.

In addition to the data structure, we need to be concerned with how much information is present in the data. Turtle bycatch is a rare event; in fact, it appears to be a very rare event. For all three species, the data contained a very high percentage of sets with zero takes: 98.8% for leatherbacks, 99.1% for olive ridleys, and 96.2% for loggerheads. Because of the rarity of the event, we need to consider how much information is available to model bycatch since this determines how complex a model can be reliably fitted. As the effective sample size increases, the number of model parameters we can estimate accurately increases. If the specified model has more parameters than can be estimated reliably from the data, we say the model is overparameterized. If we want to gather information concerning a rare event, a large sample size is needed to increase the likelihood of observing an event. A sample of 100 may be sufficient for a common event, but a rare event may not occur even once in 100 observations. An equivalent-sized sample of a common and a rare event does not provide an equal amount of information. Given the rarity of bycatch and the scope of the area and time involved, our sample of 266 trips (3,107 sets) is not large. Therefore, we need to be meticulous not to specify an overparameterized model and careful when assuming asymptotic distributions. For example, if relying on the Student's t distribution, the asymptotic distribution of maximum likelihood estimates, the degrees of freedom (typically the number of independent observations minus the number of parameters being estimated) needs to be modified to improve the correspondence between the distribution of the test statistic and the asymptotic distribution (McCullagh and Nelder, 1989). With hierarchical data, the formulas used to adjust the degrees of freedom for independent data (see McCullagh and Nelder, 1989) are inappropriate; therefore, we do not have a good estimate of the effective degrees of freedom.

POSSIBLE SUITABLE MODEL TYPES

When modeling independent Poisson counts, the generalized linear model (GLM) known as the log-linear model is commonly used. This model assumes that an observed count y_i is generated from a Poisson distribution with mean u_i and that all counts are independently generated. In practice, counts often exhibit more variation than expected for a Poisson variate. The relationship between the mean and variance for a Poisson variate is a fixed relationship; specifically, the variance equals the mean. Therefore, a Poisson distribution may seem suitable for a response, but when we inspect the data the variance will not appear to equal the mean. If the variance is larger, we say that the data exhibit overdispersion. When modeling the subunits in a hierarchical data set, we expect overdispersion since the subunits are clumped within primary units. Failure to account for overdispersion can lead to serious underestimation of standard errors and misleading inferences about the form of the linear predictor.

The negative binomial model, the ZIP model, and the generalized linear mixed model (GLMM) are three model types that can be used to model overdispersed Poisson counts. The ZIP model is appropriate when overdispersion is caused by more zeros than expected. The GLMM model is appropriate when the data are hierarchical. The negative binomial distribution is appropriate if counts are generated from a Poisson distribution having a random mean with a gamma distribution.

Another approach to model overdispersed data is not to specify the distribution exactly but to only specify a model for the mean of the data and the relationship between the mean and the variance. This is the prerequisite for quasi-likelihood estimation. For an overdispersed Poisson variate, Y, with expected value μ, a log-link and $Var(Y) = \phi\mu$ are frequently assumed. The parameter ϕ is typically called the dispersion parameter. Advantages of quasi-likelihood estimation are that point estimates of the model's coefficients do not depend on the value of ϕ, and they have properties similar to those of MLEs. Under quite general conditions, they are consistent and asymptotically normal and retain relatively high efficiency as long as the degree of overdispersion is moderate (McCullagh and Nelder, 1989). The disadvantage is that for the estimation of ϕ, quasi-likelihood does not behave like a log-likelihood. For Poisson data, a conventional estimator for ϕ is $X^2/(n-p)$ where $X^2 = \sum_{i=1}^{n}((y_i - \mu_i)^2/\mu_i)$ is the generalized Pearson statistic, n is the number of independent observations, and p is the number of parameters estimated. This estimator is the moment estimator and is consistent, but is known to perform poorly if a sizeable proportion of the observed counts are small (McCullagh and Nelder, 1989), as in our data. In fact, when modeling turtle bycatch the estimates of the dispersion parameter based on the Pearson statistic appeared to underestimate dispersion. This was confirmed by generating Poisson variates, $\phi = 1$, assuming a particular model, and then refitting the assumed model and estimating ϕ. The estimated dispersion parameter was consistently smaller than one. Another estimator of dispersion is the residual deviance divided by the degrees of freedom, but for the turtle bycatch data this estimator behaved like the moment estimator. In conclusion, quasi-likelihood may provide suitable estimates of the parameters in the predictor but the commonly used estimates of dispersion are unreliable. If these estimators of dispersion are used for determining uncertainty or in model selection, the results might be highly misleading.

All the models mentioned above assume that the predictor is additive and linear. The linearity restriction will not always provide a good approximation model, and a model with no assumptions concerning the shape of the additive function can be advantageous. This is the motivation behind the development of GAMs. GAMs still assume additivity in the predictors, but they allow the additive smooth function to take on any shape ranging from a straight line to nonparametric curves of increasing complexity. An advantage of working in the GAM environment is that linear, polynomial, smooth, and step functions of the predictors can all be fitted and then compared, albeit crudely. Beyond the common generalized models, GAMs have recently been developed for the negative binomial model (Thurston, Wand, and Wiencke, 2000) and the generalized mixed model (Fahrmeir and Lang, 2001).

Negative binomial, generalized mixed, or ZIP models that fit nonparametric curves, or are linear if appropriate, sound appealing, but they require more information from the data and more complex algorithms to fit the model. Significantly, it is possible to generate data as rare as that observed with turtle bycatch by just using a Poisson generator. In fact, a log-linear model and its GAM and quasi-likelihood counterparts converged quickly and provided good fits to the bycatch data when a small number of parameters were specified. A quick attempt was made to fit a GLMM and a negative binomial linear model. For both of these models, there were problems with parameter estimates diverging; when they did converge, the fit was often poor. I did not spend further time trying different estimators or algorithms for fitting these two models, a ZIP model, or their GAM counterparts as it was becoming apparent that the rarity of turtle bycatch, and not overdispersion, was the overwhelming factor that had to be handled and a more complex model was likely not appropriate. In summary, a log-linear model and its GAM and quasi-likelihood counterparts seemed to be the most suitable approach for modeling turtle bycatch.

FITTING A GLM AND GAM

Because of the rarity of turtle bycatch, an estimator's large-sample properties are likely not applicable and should not be automatically assumed. To estimate the unknown parameters in a GLM, we typically use maximum likelihood estimators (MLEs). Given an appropriate model and independent samples, MLEs generally have the optimal large-sample properties of being consistent and asymptotically efficient. Loosely speaking, for a large independent sample, MLEs are as good an estimator as there is. For a small independent sample, however, MLEs are often biased, and methods commonly used for interval estimation and model selection are frequently inappropriate since they assume asymptotic distributions (i.e., $n \to \infty$).

When fitting a GAM one needs to decide on (1) the methodology used to fit the smoother, (2) the type of smoother, and (3) the degree of smoothing. There are several ways to fit a GAM but none has been shown to be optimum. The most widely used method for fitting a GAM is probably an iterative algorithm that uses backfitting in combination with local scoring (Schimek and Turlach, 2000). Although this algorithm is popular, it has limitations that are relevant to the structure of our data. This algorithm, and most others, assume independent errors and can be adversely affected if the errors are dependent. It can also have severe problems with increasing collinearity or concurvity (concurvity refers to nonlinear dependence) among predictor variables. Additionally, the bias and variance of the smoothers cannot be derived theoretically except in special cases. Thus, to estimate variances and confidence intervals, resampling techniques are commonly used (Schimek and Turlach, 2000). There are newer algorithms that may be more robust to these problems, but they are not readily available and their behavior is not yet well understood. For a review of these algorithms see Schimek and Turlach (2000).

There are several different types of smoothers, but the type of smoother used is generally less critical than the methodology used to fit the smoother (Schimek and Turlach, 2000). The most widely used smoother is probably the scatterplot smoother known as the cubic smoothing spline, fitted by satisfying a penalized log-likelihood criterion using the backfitting algorithm in combination with local scoring. The intention of the penalized log-likelihood criterion is to optimize the fit while penalizing roughness to some prespecified extent. This is the smoother and algorithm I used to fit the GAMs considered in this paper.

When fitting a GAM the degree of smoothing for each dimension (predictor) must be specified. Instead of assuming a parametric form, a smoother uses the data to determine the shape of the functional relationship. The shape is determined by the degree that the data are smoothed, and this is calibrated by a quantity known as the equivalent degrees of freedom. As the degrees of freedom is increased, the smoothing function gains flexibility and becomes 'rougher,' enabling the display of more hills and valleys and more complex shapes. When there is more than one predictor, finding a global minimum is difficult for data-driven methods and selection of the degree of smoothing remains a troubling issue (Schimek and Turlach, 2000). For further details and references concerning GAMs, Hastie and Tibshirani (1990) and Schimek and Turlach (2000) are good sources.

CREATING PREDICTOR VARIABLES

To create possible predictor variables I started with the variables recorded in the logbooks, listed in Table 1. I then used classification trees (Chambers and Hastie, 1993) to suggest possible beneficial transformations of these variables. Classification trees express the relationship between the response variable and the predictor variables as a step function. This step function is created by partitioning the sample space into distinct regions defined by the predictors. Within a region the predicted bycatch is constant. The predictor variables and how they split the sample space are selected to provide the best fit. Trees have the advantage of not assuming additivity and expressing some interactions more efficiently, including nonadditive interactions.

Classification trees assume that the response variable is a multinomial variate. The variable turtle bycatch is not a multinomial variate. But suppose we defined a new variable to equal zero if there were no bycatches in the longline set and one if there was at least one bycatch in the set. Then our new response variable would resemble a binomial variate, a special case of the multinomial distribution. Because the vast majority of observed bycatches would be zero or one, little information would be lost. However, we would no longer be modeling the magnitude of bycatch but the presence or absence of bycatch. For this reason, I did not use trees to predict bycatch but rather to suggest new categorical predictor variables, defined by the distinct regions created by a tree.

Particularly, I was interested in defining new categorical variables that (1) pooled levels of an existing categorical variable, (2) split a continuous variable into categories, or (3) captured an interaction. Since the levels of categorical variables are expected to indicate constant levels of bycatch, trees are a natural way to explore pooling levels of a categorical variable. Using trees in this manner is particularly helpful when we are restricted to a low-dimensional model and including a categorical variable with several levels results in overparameterization. A continuous variable should only be split into categories if doing so provides a better approximating model. If the relationship between bycatch and a continuous variable is a continuous smooth function, categorization can introduce unwanted bias and be less efficient. Trees can capture a nonadditive interaction or express an interaction more efficiently (require fewer parameters). For example, if turtles are very rare in an area their rarity may be the overwhelming factor affecting bycatch, versus fishing practices, and adding another predictor variable will account for little if any unexplained variability. In another area, where turtles are more common, bycatch is likely to be higher and more variable and adding further predictors may explain some of this variability. Trees provide a framework to easily look for these patterns and express them. If a tree captures an interaction, we can create a new variable that captures this interaction. This new variable can then be used in a GLM or a GAM.

I used the tree function in S-PLUS (Insightful Corp., 2001) to grow and prune trees. When growing a tree for a rare event, the minimal observations required in each categorical region should be sufficiently large to prevent regions from having estimated probabilities of zero without sufficient samples to support this result. For more information about tree-based models see Chambers and Hastie (1993) or Hastie, Tibshirani, and Friedman (2001).

Although trees may be very useful in defining new categorical variables, when comparing the performance of these new variables to other predictors the significance levels for the categorical variables will likely be inflated and favored in a stepwise selection routine. This is because we have used the data being modeled to suggest and create these categorical predictors.

The variable *day* was treated differently than the other variables because it is circular. To model day I used a piecewise polynomial known as a periodic B-spline. Piecewise polynomials are parametric linear functions but are more flexible than ordinary polynomials. A piecewise polynomial is obtained by dividing the values of the predictor into contiguous intervals and fitting a separate polynomial in each of these intervals. The endpoints of these intervals are known as knots. Additional restrictions can be placed on the piecewise polynomials so they are connected at the knots, resulting in a continuous curve throughout the range of the predictor. With further restrictions, the curve can be forced to be a smooth periodic function, meaning the function is continuous as it wraps around a circular variable. A B-spline of order M refers to a continuous piecewise polynomial with continuous derivatives up to order $M-2$. I used cubic splines as they are considered the lowest-order spline where the knot discontinuities are typically not visible and there is seldom any good reason to go beyond them (Hastie, Tibshirani, and Friedman, 2001)

MODEL SELECTION

Concurrently with selecting the model type, the structural aspects of the model must be determined. This involves selecting the predictor variables and the form of the additive function for each predictor. The true relationship between bycatch and the predictor variables is likely complex and probably includes effects of different magnitudes. Some of these effects are probably not even measured in our data, but proxies may exist. For example, the number of turtles exposed to a longline set is a likely factor in turtle bycatch, but this variable is unknown. However, the variables latitude, longitude, and day in the year are likely proxies for the unknown spatial and temporal distribution of turtle density. Therefore, we do not expect to unearth the true relationships but aim to select the best approximating model among a finite set of candidate models.

Our ability to detect associations of smaller magnitudes and fit more complex models increases as the sample size increases. With over 3,000 independent observations, fitting a complex high-dimensional model would generally not be a problem. However, our 3,107 longline sets are not independent, and the sample size is relatively small considering the rarity of the event and the scope of the area and time involved. With so few positive bycatches there is little information to fit a complex model, but a simple model may be sufficient since takes are so rare.

When several predictor variables expressed in different ways are being considered as in the turtle bycatch analysis, a stepwise procedure is a convenient way to start the selection process. Before progressing with a stepwise procedure, the criterion for model selection needs to be decided. Considerable debate exists about how model selection should be carried out, but many of the most common approaches are based on the likelihood function. For GLMs, model selection based on functions of the likelihood is supported by asymptotic distribution theory, but for GAMs this generally is not true and we rely on approximations and heuristics (Chambers and Hastie, 1993). For example, for a GLM the deviance statistic has an asymptotic χ^2-distribution but for a GAM it does not. Hastie and Tibshirani (1990) provide some empirical evidence that supports using the χ^2-distribution when evaluating the deviance of a GAM, and in practice, the deviance and other functions of the likelihood remain useful tools for evaluating and selecting models when large sample properties are applicable. Two widely used likelihood-based criteria are Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). Bayesian, cross-validation, and bootstrap methods have been the focus of recent research, but these methods generally are more computer intensive and not feasible if the set of candidate models is not small.

To develop a subset of plausible models for turtle bycatch, I used stepwise selection. The final model was selected based on model diagnostics and scientific judgment. In this section, I first describe the stepwise procedure used and the theory behind it. I then describe how potential predictors were introduced into the stepwise procedure and how models were evaluated.

Stepwise Selection

Stepwise methods are easy to use and inexpensive computationally, but they must be used and interpreted with caution. There are several pitfalls. First, the final model may not optimize a reasonable criterion for choosing a model. Second, the apparent ordering of the selected predictors is an artifact of the method and need not reflect the degree of their association with the dependent variable. Finally, the significance of predictors may be seriously overstated.

When using a stepwise procedure to select a model, one needs to choose a computational algorithm that provides a systematic technique for visiting and comparing candidate models. I used the S-PLUS stepwise GAM procedure *step.gam* (Insightful Corp., 2001). This algorithm is based on a modified form of AIC, denoted as GAMAIC herein, and allows one to step through different models along a prespecified path (Chambers and Hastie, 1993). The visited model with the smallest value of GAMAIC is the selected model.

The modified form of $\text{AIC} = -2l(\hat{\beta} \mid x) + 2p$ that *step.gam* uses is $\text{GAMAIC} = -2l(\hat{\beta}, \hat{s} \mid x) + 2p\hat{\phi}$, where $l(\cdot)$ is the natural logarithm of the likelihood function, $\hat{\beta}$ is the vector of maximum likelihood estimates for β, $\hat{s}$ is the fitted smoother, x is the set of predictors, and p is the sum of the degree of freedom of the smoother and the number of linear parameters fitted. The addition of $\hat{\phi}$ in the second term is based on the principles of quasi-likelihood.

As detailed in Burnham and Anderson (1998), Akaike used the relationship between the maximized log-likelihood and the Kullback-Leibler distance to derive an approximately unbiased estimate of the expected Kullback-Leibler distance (or information). If the true model has infinite parameters, the minimum-AIC criterion is asymptotically efficient in the sense that the prediction error is asymptotically minimized (Shibata, 1981). Although the smoothers of GAMs are not maximum likelihood estimates, based on the principles of GAMs and quasi-likelihood estimation we would expect GAMAIC to behave similarly.

Although AIC has been widely used, it can have serious shortcomings. Foremost, if the sample size is small or the candidate model has too many parameters in relation to the size of the sample, AIC is a strongly negatively biased estimator of the expected Kullback-Leibler distance (Hurvich and Tsai, 1989, 1995), and this bias can lead to overfitting. Therefore, unless the sample size is large a bias adjustment is strongly recommended. The exact or best small-sample bias correction term varies by model (Findley, 1985). Since our sample size is relatively small, a small-sample bias correction term is likely needed. Small-sample bias correction terms have been proposed for generalized and quasi-likelihood models (Hurvich and Tsai, 1995, Hurvich et. al., 1998), but these corrections require knowledge of the sample size n. This brings us back to the predicament concerning the value of n. In our case, is it the number of trips sampled or the number of observed longline sets? Since sets within a trip are not independent, the amount of information in the data is likely between these two numbers. Even if observations were independent, turtle bycatch is a very rare event, and the effective sample size is likely smaller than the actual sample size.

AIC has frequently been compared to BIC $= -2l(\hat{\beta} \mid x) + log(n)p$ (Schwarz, 1978). As when using AIC, the model with the smallest value of BIC is the selected model. Theoretically, selecting the model with minimum BIC yields the model that is most probable under the posterior distribution (the conditional density of β given x). Assuming that the true model is included in a finite set of candidate models, BIC is consistent, meaning BIC will asymptotically select the correct model. In practice, BIC tends to underfit, especially as n increases. If we use BIC, we still face the dilemma concerning the value of n.

AIC and BIC assume independent observations. Pan and Le (2001) observed that for binary clustered data AIC and BIC work well if the independence assumption is satisfied, or nearly so, but perform poorly when this assumption is seriously violated. For dependent observations, they proposed a bootstrap method for model selection. As a way to select a model for predicting bycatch, this method was not computationally feasible due to the large number of candidate models being considered. Fortunately, the dependence among sets within a trip appeared to be slight to moderate.

A controversial interpretation of AIC is that the first term is a measure of the model's lack of fit and the second term $2p$ is a penalty for increasing the size of the model. Expanding on this idea, the General Information Criterion is defined as GIC $= -2l(\hat{\beta} \mid x) + \alpha p$ (Atkinson 1980, 1981), where α is either constant or a function of n. GIC includes AIC ($\alpha = 2$) and BIC ($\alpha = log(n)$). Additionally, since most of the small-sample bias correction modifications to AIC can be expressed as AIC plus a term that is a function of n and p, these modified forms of AIC are also included. Using similar modifications to GAMAIC, we can express GIC as GAMGIC $= -2l(\hat{\beta}, \hat{s} \mid x) + \alpha p \hat{\phi}$. For linear regression models with independent observations, Atkinson (1981) suggests using a range of reasonable values for α to provide a set of plausible initial models for further analysis. If n is unknown, using a range of values for α is supported by the fact that altering n in the bias correction term is equivalent, in practical terms, to changing the value of the penalty. Because of the dilemma concerning the value of n and $\hat{\phi}$, I incorporated this idea into GAMGIC, where GAMGIC was computed as if observations were independent. Because these ideas are exploratory, I only used GAMGIC and *step.gam* to create a set of plausible models.

Within *step.gam* you cannot directly adjust α, but the value of $\hat{\phi}$ can be specified using the *scale* command. If *scale* is interpreted as $scale = \hat{\phi}\alpha$, then both $\hat{\phi}$ and α are specified and α can be adjusted by means of this command. After a few trials, it was clear that values of $scale > 8$ tended to underfit models; hence, I used values of *scale* ranging from 2 (AIC) to 8 (BIC assuming independent Poisson variates). Observation indicated that AIC tended to overfit and BIC tended to underfit the models. This observation was supported by simulations where independent and dependent correlated Poisson variates were generated using models similar to those being considered. The generated values were then subjected to similar model selection procedures. The simulation suggested that AIC could grossly overfit the model, particularly when the stepwise procedure could visit overparameterized models. Proceeding with caution under this framework, stepwise selection using *step.gam* was a useful tool.

As potential predictors, I considered modeling continuous variables using linear terms, categorizations, smoothers with different degrees of freedom, and polynomials. For categorical variables, I considered their original form and any pooling suggested by a classification tree. I also considered potential interactions between predictors. To understand the relationships between predictor variables, I considered them individually, in groups of related variables, and in subsets that showed possibilities for prediction. Because the outcome is likely to be influenced by the starting model and the order in which term formulas are specified, both of these factors were investigated.

Model Diagnostics

Once a model was fitted, informal verification of the goodness of fit was obtained through residual plots. Figure 1 is an example of a diagnostic residual plot supplied by S-PLUS and used extensively in this analysis. The solid line on the figure is the smooth curve fitted to the loggerhead bycatch data for the variable latitude; only latitudes greater than 22°N were included in the data modeled. If observations are independent and pairwise correlations among the predictor are not high, the dashed lines lie approximately two standard errors away from the central curve on either side (Chamber and Hastie, 1993). Because the bycatch data are hierarchical and the correlation among predictors may be high, the standard error bands are likely too narrow, but they still give a rough indication of how uncertainty may be spread throughout the curve. In

Figure 1, the standard error band follows the general pattern of the fitted curve but flares out near the endpoints of the observed latitudinal range. The fact that we cannot draw a horizontal line across the plot without going way outside the band provides evidence that bycatch was associated with latitude.

Parts of the plot where the standard error band is particularly wide suggest that the fit is unstable; that is, there is a high level of uncertainty in the fit. Unstable fits can be caused by sparse data or by a choice of degrees of freedom of the smoother that is too low to give an adequate representation of the relationship. Sparse data can be detected using the rug plot along the bottom of Figure 1. The rug plot gives the frequency of the corresponding predictor by placing a short vertical line at each observed value. In Figure 1, the rug plot indicates there were few observations at the higher latitudes; hence, the standard error band there is very wide. The band probably widens at the lower range of latitudes because the reliability of a GAM fit is always reduced at the endpoints of the range (Hastie and Tibshirani, 1990).

The black circles on Figure 1 are the partial deviance residuals (Chambers and Hastie, 1993). The deviance residual for observation y_i is the square root of its contribution to the overall deviance multiplied by 1 if y_i is greater than its fitted value and -1 if y_i is less than its fitted value. A satisfactory diagnostic plot typically has residuals distributed evenly and randomly above and below the fitted curve. Because of the very high frequency of zeros in our data, a satisfactory diagnostic plot will look different. First, we expect $\hat{\mu}_i$ to be positive but consistently near zero. Since an observed count of one or greater is large when compared to a fitted value of zero, observed zeros will have small negative residuals and observed positive bycatches will have large positive residuals. Therefore, we do not expect residuals to be symmetrically distributed around the fitted curve. Instead, a good fit is indicated by the positive residuals loosely following the pattern of the fitted curve and the negative residuals falling farther below the curve as the frequency of positive residuals increases.

PREDICTION

After selecting and fitting the predictive model, the next step was to use the model to predict the bycatch of each unobserved longline set. For each set i, the predicted bycatch was $\hat{Y}_i = \hat{\mu}_i = exp(\hat{\eta}_i)$, where $\hat{\eta}_i$ was the number obtained when using set i's recorded logbook predictor values in the fitted model. Total bycatch was predicted by adding the observed bycatch to the sum of the predicted bycatch for all unobserved sets.

At this juncture we have only a point estimate of total bycatch. Because this estimate is subject to sampling error and random fluctuations, it should be accompanied by a measure of uncertainty. This measure is frequently expressed in the form of a prediction interval. A prediction interval is different from what is typically called a confidence interval: an interval that expresses the uncertainty in the parameter estimates, such as $\hat{\mu}$, because of sampling fluctuations. When a stochastic model is used to predict total bycatch, we assume that a set's bycatch is a random variable. This assumption implies that total bycatch is also a random variable that fluctuates randomly around its mean. Hence, a predictor has two sources of variation: the variation in $\hat{\mu}$ and the variation of Y.

Although quasi-likelihood was used to fit the models, to estimate uncertainty I did not assume any large sample properties or use the quasi-likelihood estimate of the dispersion parameter. Instead, I used a nonparametric bootstrapping algorithm based on algorithms in Davison and Hinkley (1997) to approximate the prediction intervals for predicted total bycatch. To account for overdispersion, I did not assume an error structure but mimicked the error structure of the original data by resampling residuals. Specifically, I resampled the standardized Pearson residuals (Davison and Hinkley, 1997). The standardized Pearson residuals were calculated as

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi}\hat{\mu}_i}} \qquad (i = 1, ..., n), \tag{1}$$

where $\hat{\mu}_i$ was the fitted value for longline set i.

The basic steps of the algorithm were as follows. For $r = 1, \ldots, R$ bootstrap replications: (1) a new response y_i^* was generated for each set using the equation

$$y_i^* = \hat{\mu}_i + \epsilon_i^* \sqrt{\hat{\phi}\hat{\mu}_i} \qquad (i = 1, ..., n), \tag{2}$$

where ϵ_i^* was randomly sampled from the set of standardized Pearson residuals; (2) the model was refitted using the y_i^*s; (3) total bycatch Y_{yr}^* was predicted for $yr = 1994, \cdots, 1999$ using the refitted model; (4) for each year, standardized prediction errors were calculated as

$$d_{yr}^* = \frac{Y_{yr}^* - \hat{Y}_{yr}^*}{\sqrt{\hat{\phi}\hat{Y}_{yr}^*}}, \tag{3}$$

where Y_{yr}^* and $\hat{Y}_{yr}^*$ were the bootstrap-generated total bycatch and the predicted total bycatch, respectively, for the current bootstrap replication. Finally, the R values of d_{yr}^* were ordered so that $d_{yr,(1)}^* \leq \cdots \leq d_{yr,(R)}^*$. The 95% prediction interval was approximated as

$$(\hat{Y}_{yr} + d_{yr,((R+1)*.05)}^* \sqrt{\hat{Y}_{yr}\hat{\phi}}, \hat{Y}_{yr} + d_{yr,((R+1)*.95)}^* \sqrt{\hat{Y}_{yr}\hat{\phi}}\ ). \tag{4}$$

The probability that the prediction intervals for 1994 to 1999 all enclose the true bycatch is not 0.95, as this requires simultaneous prediction intervals. I used $R = 999$ as it is recommended that $R \geq 999$ when using bootstrapping to approximate confidence or prediction intervals (Davison and Hinkley, 1997).

Notice that an estimate of ϕ is used in four places. In Equation 2, the selected standardized Pearson's residual is multiplied by the same value it was divided by in

Equation 1; therefore, the value of ϕ is negated. Similarly, Equation 4 nullifies the standardization of the prediction errors in Equation 3. Because the value of $\hat{\phi}$ is used to standardize a random variable and then negated within the bootstrap process, a poor estimate of $\hat{\phi}$ could cause difficulty when trying to mimic the error structure. In practice, I found it best to use the estimate of ϕ based on the Pearson statistic if this estimate was greater than one; otherwise, I used $\phi = 1$. Since a value less than one was probably due to sparseness, not underdispersion, $\hat{\phi} = 1$ is likely a better estimate.

To mimic the error structure successfully, the sampling protocol used to resample the residuals in the first step is very important. Drawing a simple random sample of residuals with replacement was not appropriate in this situation for two reasons: (1) the residuals were not homogenous, and (2) this protocol assumes that sets were independent. Trips, not sets, were the independent observational unit, and the correlation between sets within a trip needed to be captured. To model this correlation I used an adaptation of block resampling with trip as the sampling block. Block resampling has proved useful when bootstrapping time-series data (Davison and Hinkley, 1997). The idea of block resampling is to capture the autocorrelation structure by sampling, with replacement, blocks of consecutive observations. These blocks are then pasted together to form a new series. Although this algorithm may mimic the correlation, it will not capture the heteroscedasticity present in the residuals.

The heteroscedasticity present in the residual plot in Figure 2 was typical for the models used to predict turtle bycatch. Notice that the residuals for the smaller fitted values have more extreme values than those for the larger fitted values. This is because a recorded count of one divided by a fitted value near zero produces a large residual. Therefore, if only small residuals are applied to small fitted values, the occasional positive bycatch in the midst of zeros would not be recreated. Similarly, if a large residual is assigned to a large fitted value, an extremely large and unrealistic value of bycatch would be generated. To prevent creating unrepresentative residuals and to produce the general pattern of heteroscedasticity present in the residual plots, I stratified the residuals based on their corresponding fitted values.

Finally, to mimic the correlation and heteroscedasticity in the residuals I merged the idea of block resampling and stratification of residuals. Within each stratum, trips were sampled with equal probability and replacement until the required number of ϵ_i^*s was generated. The generated residuals from selected trips were then pasted from end to end to form a new bootstrap series.

Three further details concerning the sampling protocol need mentioning: (1) To prevent splitting sets from the same trip into different strata, I used the minimal amount of stratification necessary to reproduce the heteroscedasticity. Using this strategy, most sets from a trip tended to fall in the same stratum, but some trips were split. I considered assigning sets within a trip to the same stratum based on a summary statistic, such as the mean or median, of a trip's fitted values. However, since the heteroscedasticity of residuals was a bigger problem than the correlation between residuals, splitting trips appeared preferable. (2) The inherent bias in GAMs and most nonparametric regression methods distorts the residuals and fitted values, thus naive substitution of $\hat{\mu}_i$ and ϵ_i^* from the fitted model into Equations 1 and 2 is apt to provide inaccurate prediction intervals. To avoid bias, Davison and Hinkley (1977) recommended taking residuals from an undersmoothed curve and fitted values from an oversmoothed curve. This is the strategy I used. In practice, I found that the oversmoothing should be slight (the degrees of freedom of the smoother should be reduced only slightly), and the average deviance of the undersmoothed and oversmoothed curves should be approximately equal to the deviance of the fitted model. (3) One drawback of this algorithm was that y_i^* could take on negative and non-integer values. To fix this, a constant was added to y_i^* and the new value was rounded to the nearest non-negative integer. The constant was selected so that the average total bycatch of the bootstrap-generated data sets was comparable to the observed total bycatch.

After the first step was completed, each longline set had a generated y_i^*. In the second step, y_i^* was used in place of the recorded observed bycatch and the prediction model refitted, meaning that the unknown parameters were reestimated for each bootstrap sample. If a variable created by a classification tree was included in the prediction model, for each bootstrap replication, a tree was fitted and pruned to the same size as the selected model and the classification of the variable redefined before the model was refitted. This procedure was consistent with the idea of refitting a smoother but not redetermining the degree of smoothing.

Finally, I checked the suitability of the algorithm by confirming that the distributions of the residual deviance and $\hat{\phi}$ for the bootstrap replications were centered near their values for the original prediction model. Furthermore, for a few bootstrap replications, I generated residual plots similar to Figure 2 and checked for the same general pattern as in the original plot.

An advantage of using the statistic d^* is that bias is implicitly adjusted for in the bootstrap distribution (Davision and Hinkley, 1997); therefore, it is usually not necessary to incorporate an empirical bias adjustment into the numerator of d^* when approximating prediction intervals. However, if the point estimates are going to be given, a bias adjustment should be considered. I estimated bias as the mean of $\hat{Y}_r^* - Y_r$ for $r = 1, \ldots, R$, where $\hat{Y}_r^*$ and Y_r were the predicted and actual total bycatch for bootstrap replication r.

RESULTS

For leatherback and olive ridley turtles there were few observed bycatches, and only a simple model could be fitted. However, the bycatch on individual longline sets appeared to be consistently small over time and space, suggesting that there were few factors with large effects. For loggerhead turtles there were more observed bycatches, and a more complex model could be fitted. In this section, for each turtle species I compare different prediction models and present their predicted bycatch.

For predicting leatherback bycatch, the variable latitude appeared to be a good predictor, but the best way to express the relationship between bycatch and latitude was less clear. Although I thought a smoother with six degrees of freedom provided the best approximating model throughout the latitudinal range, the fit in the middle (Fig. 3), specifically between 15°N and 25°N, appears as if it was oversmoothed. Between these two latitudes, there were numerous observed sets throughout the year but only one positive bycatch observed. The fitted curve between these two latitudes steadily declined, reaching a local minimum near 20°N, near the one positive bycatch, and then steadily inclined. The data seem to suggest that the curve should be discontinuous or fall more dramatically, then be relatively flat, and then rise more dramatically. One possible explanation for this pattern may be the effect of fishing practices on the bycatch. Most of the fishing between these latitudes was tuna fishing, but this was also true south of 15°N. Although leatherbacks are thought to occur throughout the fishing grounds of the Hawaii-based longline fleet, their density and how it fluctuates seasonally are unknown. However, the North Equatorial Current flows through the southern part of the fishing grounds, and we would expect a higher biomass of food in this area. Therefore, turtle density might be higher in this area, resulting in a higher rate of bycatch. If the type of fishing affected the bycatch rate and the different types of fishing were confined to distinct latitudinal regions, a realistic curve may be discontinuous or characterized by very steep slopes and relatively flat areas. A curve with these characteristics contradicts the shape of a cubic spline and is difficult to fit using a cubic spline smoother. As an alternative to a smooth curve, I split latitude into four categories as suggested by a classification tree and then fitted a GLM. Table 2 provides the intervals of latitude defining the four categories. Although expressing latitude as a categorical variable probably resulted in some model bias, it is not clear the bias was greater than that produced by using a GAM. Table 3 gives the point estimates and the 95% prediction intervals for the GAM and GLM. The point estimates and lengths of the prediction intervals are very similar. The increased magnitude and wider prediction intervals for leatherback bycatches in 1998 and 1999 can be explained numerically by the increased percentage of trips near the northern boundary of observed latitudes. Both models probably introduced some model bias, but due to the rarity of leatherback bycatch this bias was probably small, especially when compared to the uncertainty in the predictions.

To predict olive ridley bycatch, fitting a GLM using a categorization suggested by a classification tree proved useful. The categorization had four levels defined by the variables sea surface temperature, number of hooks, and latitude. Table 2 provides the definition for each category. This categorization appears to capture an interaction using few parameters. Olive ridley turtles are believed to be rarer in colder water, and the majority of observed bycatches were in warmer water. The definition of the first category therefore seems logical. The next three categories were defined by the number of hooks and latitude. The relationship between olive ridley bycatch and number of hooks was interesting. If the number of hooks increased, the predicted number of bycatches decreased. At first this seems counterintuitive, but the observer data include longline trips with different objectives; some were targeting tuna, others swordfish, and others mixed species. The targeted fish usually determines the fishing practice, and fishing practice is a suspected factor in the bycatch rate. The variable trip type was supposed to capture the species being targeted, thus the style of fishing, but it is suspected that the mixed category was sometimes incorrectly checked in the logbooks. If the variable trip type was strongly associated with bycatch but was recorded inaccurately, this association may not be captured in the data. Typically, more hooks are used when targeting tuna than swordfish; therefore, number of hooks may have been a proxy for fishing practice. If number of hooks was recorded with greater accuracy than trip type, it may appear to have a stronger association with bycatch, regardless of whether this was true. The point estimates and 95% P. I. for olive ridley bycatch are given in Table 4.

Although still a rare event, loggerhead bycatches were more common than for the other two species. Figure 1 shows that there were no positive loggerhead bycatches in the southern part of the fishing grounds; 24.4°N was the southernmost location of a positive bycatch. This is not surprising, as loggerhead turtles are thought to occur only in the northern region of the fishing ground. If loggerheads do not occur in the southern region, then the probability of a loggerhead bycatch there is zero, and these structural zeros should be excluded when modeling bycatch. Including them can result in an unstable model and produce unrealistic positive bycatch probabilities, since $\hat{\mu} > 0$ in a log-linear model. I truncated the data at 22°N as this was where the fitted curve tended to 'flatten out' to a very small probability. To predict total bycatch, I considered two models, each with three predictors. Both models included the variables day and latitude, but the third predictor was sea surface temperature in one model and number of hooks in the other model. Latitude entered both models as a smooth cubic spline with 3 d.f., and day entered both models as a B-spline with one knot (5 d.f.). Modeling day as a circular variable captured the seasonal variation in bycatch; hence, day was likely a proxy for the seasonal density of loggerhead turtles in the fishing grounds or for the seasonality of the different fishing practices. Sea surface temperature entered the first model as a smooth cubic spline with 4 d.f., and the number of hooks entered the second model as a categorical variable with two levels. Since swordfishing sets typically take place where sea surface temperatures are colder and also involves fewer hooks, the variables sea surface temperature and number of hooks were related. If one variable was entered into the model the other appeared to provide little additional information. The model with number of hooks as a predictor appeared to fit the data slightly better according to GAMGIC. Similar to the olive ridley model, the category with more hooks had a smaller predicted loggerhead bycatch. Sea surface temperature was not recorded in the logbooks but estimated by interpolating data collected by satellite-borne temperature sensors. Although sea surface temperature may have had a stronger relationship with bycatch, number of hooks may be a better predictor because it had less measurement error. The estimates of bycatch for both models are given in Table 5.

The estimated bias of predicted total bycatch for loggerheads was slightly higher than for the other two species (Tables 3, 4, and 5). As mentioned previously, smoothers typically have increasing bias at the boundaries. As the number of predictors increases in a GAM, the number of boundaries also increases and therefore we expect the bias to increase. The GAM used to model leatherback bycatch had one predictor, but the GAM in the first model for loggerhead bycatch had three predictors; therefore, we would expect it to have higher bias. Finally, compared to the other years, in 1994 there were more longline sets in northern latitudes where loggerhead bycatches were predicted to be higher.

CONCLUSIONS

Although simple models appeared sufficient to predict leatherback, olive ridley, and loggerhead bycatch, several different avenues can be taken to improve the predictions of bycatch. Foremost, the prediction intervals reported in Tables 3, 4, and 5 do not account for the uncertainty in model selection. As model selection was part of the prediction process, prediction intervals that also account for this uncertainty would be more realistic. We now have a better understanding of the process generating turtle bycatch and should reconsider the set of candidate models before modeling any future data. With a smaller set of candidate models, it is easier to quantify the uncertainty in model selection. Additionally, any advancement in the methodology for fitting and selecting a model for hierarchical data of a very rare event would be advantageous. Finally, a probability sample would provide the option of using a design-based estimator.

Currently, Hawaii-based longliners can only participate in swordfishing if a trained NMFS observer is onboard. With complete observer coverage of the swordfish fishery there is no longer a need to estimate turtle bycatch for this sector. For the rest of the fleet, observer coverage of the Hawaii-based longline fishery is now around 20%, and a quasi-probability sample protocol is being followed. The bycatch of turtles in this sector of the fleet has diminished to the point where modeling bycatch is no longer reasonable and the Horvitz-Thompson estimator is being used to estimate total bycatch.

ACKNOWLEDGMENTS

I am grateful to Jerry Wetherall, Samantha Strindberg, and an anonymous referee for their valuable contributions and to Judith Kendig and Francine Fiust for their editing and preparation of the final version.

REFERENCES

Atkinson, A. C.
1980. A note of the generalized information criterion for choice of a model. Biometrika 67:413-418.

Atkinson, A. C.
1981. Likelihood ratios, posterior odds and information criteria. J. of Econometrics 16:15-20.

Burnham, K. P. and D. R. Anderson.
1998. Model selection and inference: A practical information-theoretic approach, 353 p. Springer, New York, NY.

Chambers, J. M., and T. J. Hastie.
1993. Statistical models in S, 608 p. Chapman and Hall, New York, NY.

Davison, A. C. and D. V. Hinkley.
1997. Bootstrap methods and their application,582 p. Cambridge University Press, New York, NY.

Fahrmeir, L. and S. Lang.
2001. Bayesian inference for generalized additive mixed models based on Markov random field priors. Appl. Stat. 50:201-220.

Findley, D. F.
1985. On the unbiasedness property of AIC for exact or approximating linear stochastic time series models. J. Time Series Analysis 6:229-252.

Hansen, M. H., W. G. Madow, and B. J. Tepping.
1983. An evaluation of model-dependent and probability-sampling inferences in sample surveys. J. Am. Stat. Assoc. 78:776-793.

Hastie, T. J. and R. J. Tibshirani.
1990. Generalized additive models, 335 p. Chapman and Hall, New York, NY.

Hastie, T. J. and R. J. Tibshirani, and J. Friedman.
2001. The elements of statistical learning: Data mining, inference, and prediction, 533 p. Springer, New York, NY.

Hurvich, C. M., and C. L. Tsai.
1989. Regression and time series model selection in small samples. Biometrika 76:297-307.

Hurvich, C. M., and C. L. Tsai.
1995. Model selection for extended quasi-likelihood models in small samples. Biometrics 51:1077-1084.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai.
1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. of the Royal Stat. Soc., Series B: 60:271-293.

Insightful Corp.
2001. S-PLUS 6 for Windows, Insightful Corporation, Seattle, WA.

McCracken, M. L.
2000. Estimation of sea turtle take and mortality in the Hawaiian longline fisheries. NOAA-NMFS-SWFSC Administrative Report H-00-06.

McCullagh, P. and J. A. Nelder.
1989. Generalized linear models, 2nd edition, 511 p. Chapman and Hall, New York, NY.

Pan, W. and C. T. Le.
2001. Bootstrap model selection in generalized linear models. J. of Agric. Biol. and Environ. Stat. 6:49-61.

Schimek, M. G. and B. A. Turlach.
2000. Additive and generalized additive models. *In* Smoothing and regression approaches, computation, and application (M. G. Schimek, ed.), p. 277-327. Wiley, New York, NY.

Schwarz, G.
1978. Estimating the dimension of a model. Ann. Stat. 6:461-464.

Shibata, R.
1981. An optimal selection of regression variables. Biometrika 68:45-54.

Thurston, S. W., M. P. Wand, and J. K. Wiencke.
2000. Negative binomial additive models. Biometrics 56:139-144.

Table 1. Explanatory variables considered for predicting total take

Variable	Notes
latitude	degrees north
longitude	degrees east
year	1994-1999
month	January-December
day	from 1 to 365 and represented as a circular variable using a periodic B-spline
hooks	number of hooks on longline
sea surface temperature	calculated by interpolating satellite records for the recorded latitude, longitude, and date (by week)
vessel length	registered length
trip type	3 categories (swordfish, tuna, mixed)

Table 2. Models for predicting bycatch

Species	Model class and predictor variables (hooks=number of hooks, lat=latitude, sst=sea surface temperature)
Leatherback	GLM with a categorical variable defined by $[lat \leq 14.95°N]$, $[14.95°N < lat \leq 24.84°N]$, $[24.84°N < lat \leq 33.82°N]$, $[lat > 33.82°N]$ GAM with s(lat,6)
Olive Ridley	GLM with a categorical variable defined by $[sst \leq 24.22°C]$, $[sst > 24.22°C$ and $hooks \leq 1073.5$ and $lat \leq 19.71°N]$, $[sst > 24.22°C$ and $hooks \leq 1073.5$ and $lat > 19.71°N]$, $[sst > 24.22°C$ and $hooks > 1073.5]$
Loggerhead	GAM 1 with s(lat,3), day as a circular variable, and s(sst,4) GAM 2 with s(lat,3), day as a circular variable, and hooks in two categories: $[hooks \leq 1225]$, $[hooks > 1225]$

Table 3. Leatherback bycatch estimates (Est.) with approximate 95% prediction intervals (PI). The estimated average percent bias was 2% for the GAM and −1% for the GLM. The models are defined in Table 2.

	GLM		GAM	
Year	Est.	95%PI	Est.	95%PI
1994	109	[79-143]	106	[74-136]
1995	99	[71-133]	101	[69-131]
1996	106	[75-147]	105	[79-139]
1997	89	[65-121]	97	[70-127]
1998	139	[87-193]	129	[87-173]
1999	132	[89-183]	125	[86-162]

Table 4. Olive ridley bycatch estimates (Est.) with approximate 95% prediction intervals (PI). The estimated average percent bias was 1%. The models are defined in Table 2.

Year	Est.	95%PI
1994	137	[82-228]
1995	119	[77-183]
1996	147	[97-207]
1997	127	[89-177]
1998	118	[82-164]
1999	105	[67-140]

Table 5. Loggerhead bycatch estimates (Est.) with approximate 95% prediction intervals (PI). The estimated average percent bias was −4% for the GAM 1 and −3% for the GAM 2. The models are defined in Table 2.

	GAM 1		GAM 2	
Year	Est.	95%PI	Est.	95%PI
1994	491	[231-765]	445	[314-621]
1995	387	[180-540]	401	[280-519]
1996	412	[267-544]	401	[312-546]
1997	320	[227-424]	344	[243-450]
1998	409	[292-595]	396	[311-528]
1999	418	[281-556]	346	[258-477]

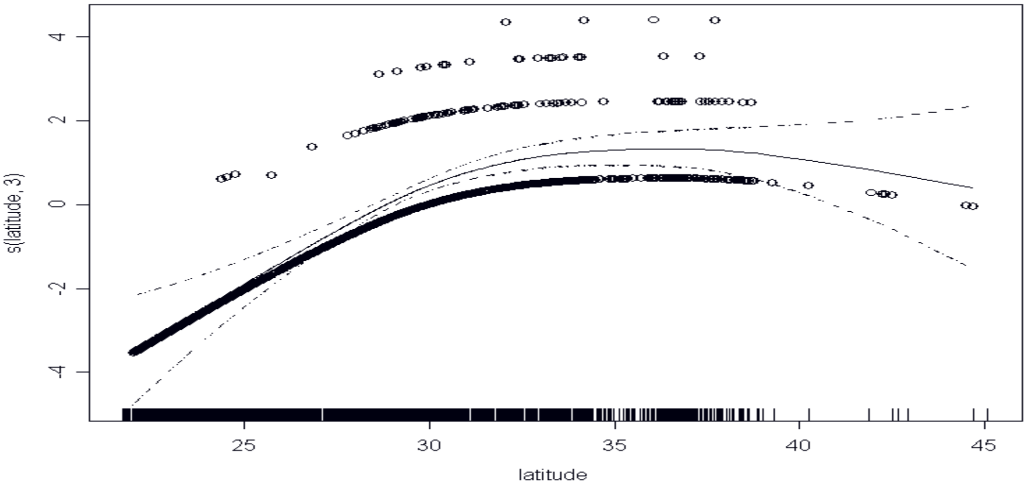


Figure 1. The fit of loggerhead bycatch to the scaled latitude (lat), for latitudes greater than or equal to 22°N. The solid line represents the fitted smooth curve with 3 degrees of freedom, the dashed lines denote the fitted smooth plus or minus 2 standard errors (approximate) and demarcate a "standard error band," the black circles represent partial deviance residuals, and the bars on the x-axis are the rug plot. The residuals are well distributed above and below the curve and follow the basic line of the curve. The standard error band shows a definitive curve and is narrow in the center of the curve but wider at the right endpoints where there are few observations.

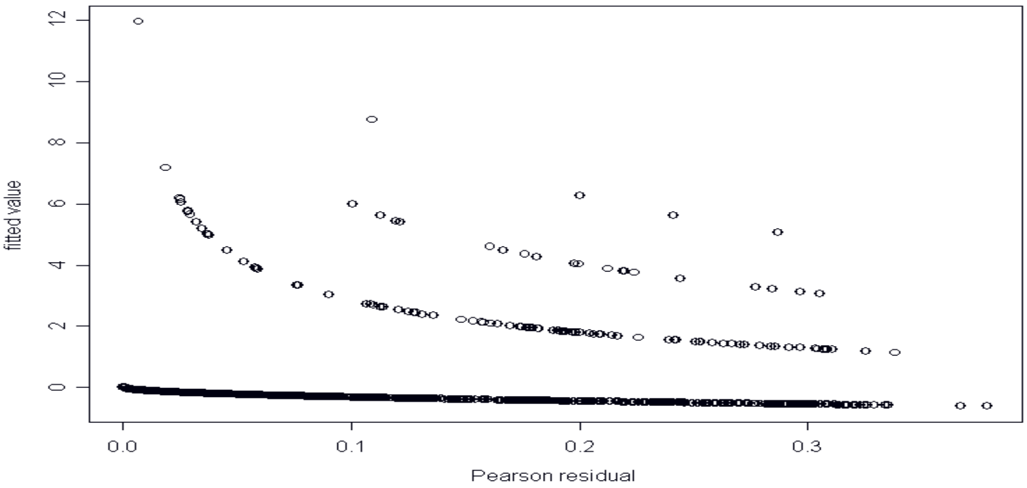


Figure 2. The residual plot of the fitted values versus the Pearson residuals for the loggerhead bycatch prediction model that used latitude, day, and hooks as predictor variables (see Table 2).

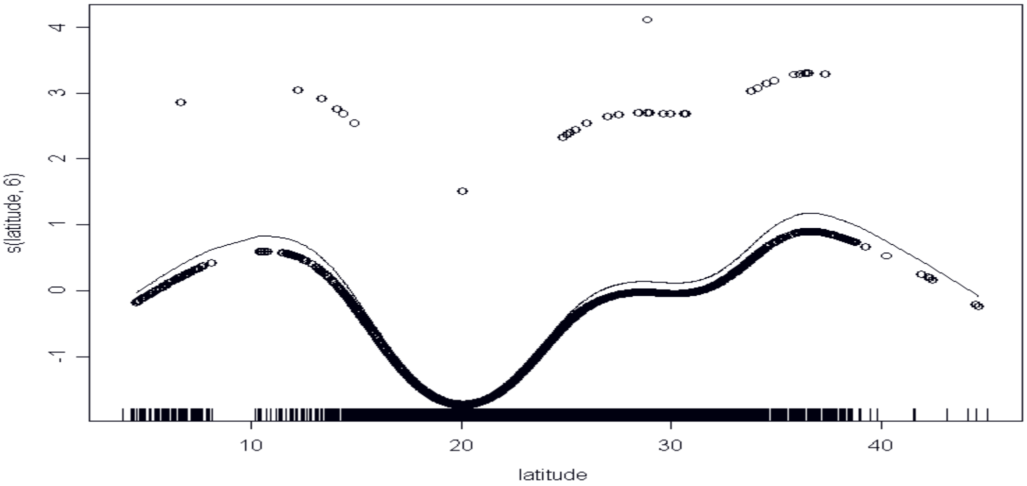


Figure 3. The fit of leatherback bycatch to latitude (lat). The solid line represents the fitted smooth curve with 6 degrees of freedom, the black circles represent partial deviance residuals, and the bars on the x-axis are the rug plot. The fit seems reasonable at the boundaries but appears to be oversmoothed in the middle.

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-1 The Hawaiian monk seal in the Northwestern
Hawaiian Islands, 2001.
T. C. JOHANOS and J. D. BAKER (comps. and eds.)
(April 2004)

2 Contingency plan for Hawaiian monk seal Unusual
Mortality Events.
P. K. YOCHEM, R. C. BRAUN, B. RYON,
J. D. BAKER, and G. A. ANTONELIS
(May 2004)