Abstract.—Fisheries discard data are often characterized by a smooth distribution of positive amounts of per-set discard but with an extremely large number of zero observations. This discontinuity is difficult to fit with a standard distribution. One approach is to model per-set discard with a mixture of two distributions, with one component representing the zero observations and the other representing the observations of positive discard. In this paper, we describe such a mixture model that is suitable when the discard observations have been rounded to integer amounts. In particular, when "rounded" zeros (representing small amounts of discard) and "true" zeros (representing no discard) are indistinguishable in the data, the mixture model can be used to estimate the proportion of either. We fit this model to tuna discard data collected by observers aboard the U.S. tuna purse-seine fleet in the eastern tropical Pacific Ocean during the years 1989–92. We use the model to estimate discard per set, allowing the model parameters to depend upon fishing strategy and geographic location, and we estimate mean discard per set fisherywide.

Manuscript accepted 26 October 1995.
Fishery Bulletin 94:330–340 (1996).

A mixture model for estimating discarded bycatch from data with many zero observations: tuna discards in the eastern tropical Pacific Ocean

Peter C. Perkins
Elizabeth F. Edwards
Southwest Fisheries Science Center
National Marine Fisheries Service, NOAA
P.O. Box 271, La Jolla, California 92038

Many fisheries catch unwanted individuals of nontarget species in addition to target species. This bycatch is generally discarded and in many fisheries few, if any, individuals survive capture and discard (e.g. Joseph, 1994). Estimating the extent of such discard is increasingly important as fisheries managers contend with situations where unwanted catch in a fishery is desirable in other contexts. For example, bycatch in one fishery may include juvenile members of the target species in the same or another fishery, or individuals from threatened, endangered, or protected species (e.g. Collins and Wenner, 1988; Caillouet et al., 1991).

Despite their increasing importance, bycatch and discard remain relatively unstudied. Few fisheries routinely measure discards so that the amount of discard usually must be estimated rather than reported directly (e.g. Berger et al., 1989). The U.S. tuna purse-seine fishery in the eastern tropical Pacific Ocean (ETP) provides an opportunity to examine this problem because quantitative information on discard of tuna (including both nontarget tuna species and juveniles of target species) has been collected from the fishery since 1988.

A flexible approach is required to model tuna discard from this fishery because the purse-seine vessels capture fish using three distinct fishing strategies. These strategies are defined by the different types of sets involved: "log fishing," "school fishing," and "dolphin fishing." Log fishing catches tuna by setting purse seines around fish associated with floating objects. Log sets usually capture schools of small (30–50 cm) yellowfin tuna, *Thunnus albacares*, or mixed schools of small yellowfin and like-size skipjack tuna, *Katsuwonus pelamis*. School fishing catches tuna by setting purse seines around schools composed purely of tuna (again, usually small fish and either pure schools of yellowfin or mixed schools of yellowfin and skipjack tuna), located by surface disturbances created by the schools. Dolphin fishing catches tuna by first locating surface disturbances created by closely associated dolphins (NRC, 1992) and by setting purse seines around both tuna and dolphins. Tuna associated with dolphins almost always consist of pure schools of large (80–120 cm) yellowfin tuna (IATTC, 1989).

Log fishing generates large amounts of tuna discard in almost every set, whereas school fishing generates

moderate amounts of tuna discard and in a smaller proportion of sets than does log fishing (Joseph, 1994). Dolphin fishing generates small amounts of tuna discard and only infrequently. Thus, tuna discard data from dolphin sets are almost all zero observations, whereas data from log sets are mostly nonzero observations, and school sets are an intermediate case.

In this paper we develop a method for using a single probability distribution to model discard per set for these three disparate types of data and show how to use the model to estimate mean discard per set for each set type. The focus of the present study is development and description of the model as a solution to a common problem in discard estimation. In general, the method presented is applicable to any situation involving analysis of data characterized by subsets with varying proportions of zero observations. Detailed results of applying the model and its implications for the U.S. tuna purse-seine fishery in the ETP are the subject of a future paper.

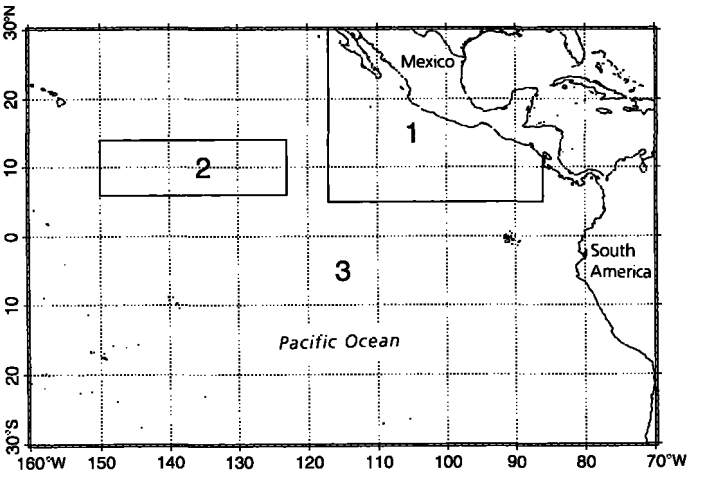

Figure 1

Geographic strata used in developing models to estimate mean discard per set for the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92 (Federal Register, 1989). Area 3 includes all ETP not explicitly included in areas 1 and 2.

Methods

Data

The data consisted of per-set estimates of total tons of tuna discarded by the U.S. fleet only. We did not have access to data on species or size composition of tuna discards nor to data on nontuna discards.

Data were collected by National Marine Fisheries Service (NMFS) or Inter-American Tropical Tuna Commission (IATTC) observers placed aboard U.S. tuna purse-seiners during routine fishing trips to the eastern tropical Pacific Ocean (Fig. 1) as part of a bycatch study initiated in 1989 by the IATTC. Observers recorded time and position of all sets made by U.S. vessels fishing in the ETP during the 31-month study period (from 1 September 1989 to 30 March 1992). Observer coverage was 100% during this period. However, during the first eleven months (from 1 September 1989 to 30 July 1990), discard information was recorded only for approximately half of the sets. Observers recorded discards for all sets during the remaining twenty months.

Because it was not feasible to weigh tuna discard directly, observers estimated the discard weight by counting the number of brailers (large fish baskets) used to empty the net after each set, multiplying by an estimated tonnage per brailer, and then multiplying by the estimated fraction of nontarget tuna in the catch. Observers estimated this fraction by observing the composition of brailers or by observing catch sorting on deck. Occasionally the majority of the catch was discarded before being brailed aboard. In these cases, observers estimated discard weight by first estimating the weight of the total catch and then subtracting an estimate of the tonnage loaded by brailer.

Observer estimates of discard tonnage were rounded to integer values, with rounding interval increasing with amount of discard (Fig. 2). There is a systematic tendency toward rounding to the nearest 5 or 10 metric tons (t) for small and medium estimates of discard and to the nearest 25 or 50 t for the largest estimates. For sets with moderately small amounts of discard, observer estimates tended to be more precise because the bycatch, as well as the target fish, were brailed aboard the vessel, then sorted on deck. This allowed the discard to be easily compared with the total catch. For sets with large amounts of discard, the fish may not have been brought on board, making precise estimates more difficult and rounding tendencies greater. For sets with very small amounts of discard, weights were rounded to the nearest ton so that it was not possible to distinguish observations with no discard from those with very small amounts of discard (less than one-half ton). Thus, "zero observations" may correspond to either case.

We did not attempt to account for the uncertainty introduced by these sources of measurement error and rounding. In the absence of data or studies for determining the ground truth of observer estimates

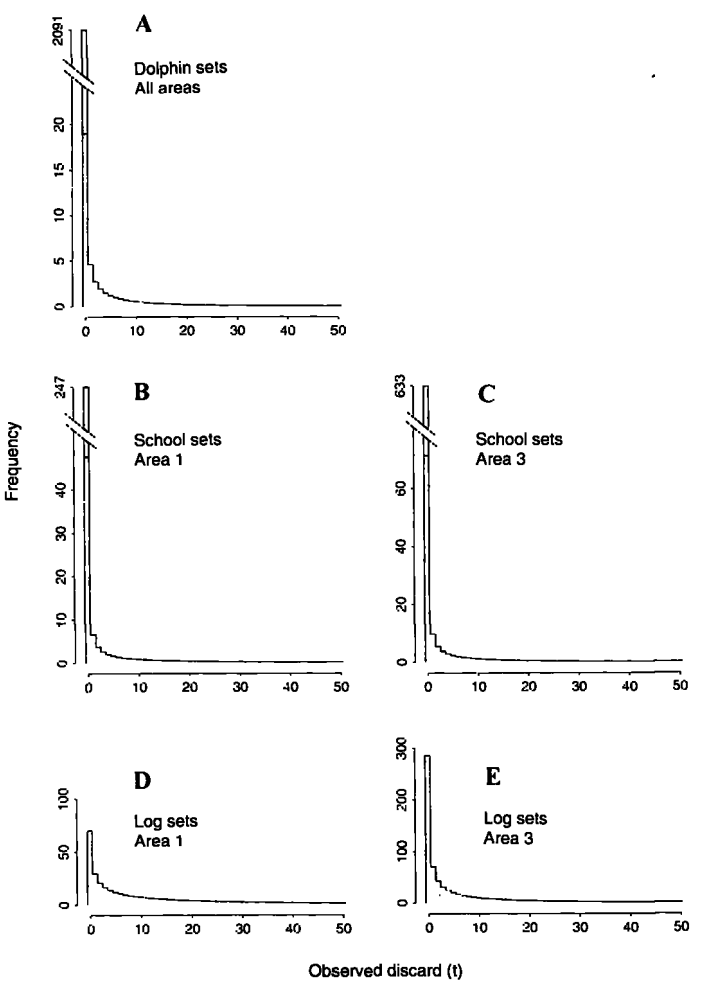

Figure 2

(**A–E**) Observed and fitted per-set tuna discards for school, dolphin, and log sets for the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92. Geographic areas are defined in Federal Register (1989) (Fig. 1). Bars indicate observed frequencies, lines indicate fitted frequencies from the negative binomial with added zeros model. Top portion (above horizontal division) of the fitted frequency for the zero bin of dolphin and school set data represents "true zeros," bottom portion represents "rounded" zeros from negative binomial (NB) component. Vertical axes have been cut and scaled to make NB components directly comparable. Not shown: (C) 3 observations >50 tons, (D) 16 observations >50 tons, (E) 12 observations >50 tons.

of discard or in the absence of a plausible model for the measurement errors, we treated the discard weight estimates as exact measurements.

Discard weight was recorded for 59% (2,110 of 3,590, Table 1) of observed dolphin sets, 76% (960 of 1,266) of observed school sets, and 75% (998 of 1,328) of observed log sets. These sets generated 134, 1,098, and 9,819 tons of reported discard, respectively. The relatively small discard totals for school and dolphin sets were due to the large numbers of those sets with zero discard reported. Positive amounts of tuna discard were reported in 65% (650 of 998, Table 1) of log sets for which discard was recorded, but in only 8% (80 of 960) of school sets and only 0.9% (19 of 2,110) of dolphin sets for which discard was recorded. We ignored log and school fishing in area 2 (see next sec-

Table 1

Fishing effort in numbers of sets for the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92. Geographic areas are defined according to Federal Register (1989) (Fig. 1). N is the total number of sets in a given area, n is the number of sets for which discard weight was recorded, and n^+ is the number of sets for which strictly positive discard was reported.

Set type	Area	N	n	n^+
Dolphin	1	2,496	1445	10
	2	498	272	5
	3	596	393	4
	Total	3,590	2,110	19
School	1	399	279	32
	2	0	0	0
	3	867	681	48
	Total	1,266	960	80
Log	1	537	326	257
	2	10	4	4
	3	791	672	393
	Total[1]	1,328	998	650

[1] Totals for log sets do not include sets in geographic area 2 because these sets were not included in our analysis. See text for explanation.

tion) for this analysis, because 0 school sets and only 10 log sets (4 with estimated discard) occurred in this area (Table 1). We also omitted 7 sets in which the entire catch (target catch plus discard) was lost owing to equipment failure.

Modelling discard per set

We chose a modified negative binomial (NB) distribution known as the negative binomial with added zeros (NBAZ) (Johnson and Kotz, 1969) to model discard per set. This distribution can accommodate the wide range in the proportion of zero observations, as well as the relatively heavy tails in the observed distributions of discard for all three set types (Fig. 2). (See Discussion section for two other models considered but rejected.)

The NBAZ is a mixture of a NB distribution and a discrete probability mass at zero. Under this model, discard per set is either exactly zero with probability p or has a NB distribution with probability $1-p$. The NB portion of this distribution can be viewed as representing strictly positive amounts of discard rounded to integer values. Thus, zero values that are part of the NB can be interpreted as observations of small amounts of discard rounded down to zero. Zero values from the probability mass can be interpreted as exact zeros. The probability function for this modified NB distribution is

$$\Pr\{Y=y\} = \begin{cases} p+(1-p)\left(\frac{1}{1+a\mu}\right)^{1/a}, & y=0 \\ (1-p)\frac{\Gamma(y+1/a)}{y!\Gamma(1/a)}\left(\frac{1}{1+a\mu}\right)^{1/a}\left(\frac{a\mu}{1+a\mu}\right)^{y}, & y=1,2,\ldots \end{cases} \quad (1)$$

where Y is an individual observation (tons of discard per set), p is the probability of an observation coming from the "true zero" state, $1-p$ is the probability of an observation coming from the NB state, and μ and a are the mean and variance parameters, respectively, of the conditional NB.[1]

The parameter a determines the shape of the distribution. As a approaches zero, the conditional NB distribution in the mixture approaches a Poisson distribution. As a increases, the conditional NB becomes more skewed, with a heavier tail and higher probability of a zero observation. The parameter p is a mixing parameter which controls the relative importance of the NB and the probability mass at zero. When p is one, the distribution is a probability mass at zero. When p is zero, the probability distribution becomes strictly NB and expected discard per set is μ (the NB mean).

The expected value and variance for individual observations from this probability distribution are

$$E[Y] = (1-p)\mu \quad (2)$$

$$\text{var}[Y] = (1-p)\left(\mu + (a+p)\mu^2\right). \quad (3)$$

We fit the NBAZ model using maximum likelihood and allowing the three model parameters to depend upon set type and geographic area. We also considered using tons of tuna loaded (i.e. commercial catch), time of day, and month as covariates, but rejected them as either unfeasible (due to sampling unbalance) or statistically unimportant. We did not attempt to account for any long-term (i.e. year to year) trend in discard rates because the data included too few years for such an analysis.

A priori, we used the same three geographic areas (Fig. 1) as those currently used to compare U.S. and non-U.S. dolphin mortality rates (Federal Register, 1989). These roughly define the major fishing areas in

[1] The NBAZ can be equivalently reparametized in terms of a "zero-truncated" NB and a probability mass at zero. The parameter corresponding to p would, in that case, denote the probability of a zero observation regardless of source. Thus, that form does not distinguish between "true" and "rounded" zeros.

the system. The total number of sets observed in each area, including sets for which discard was not recorded, represents the actual areal distribution of fishing effort during the study period. However, observation of discard was not proportional to this distribution of total effort (Table 1). In the analysis that follows, it is important to distinguish between the total number of sets, denoted by $N_{i,j}$, and the number of sets for which discard was recorded, denoted by $n_{i,j}$. The former define the actual distribution of fishing effort, whereas the latter simply reflect the sample taken. Because our sample of sets with discard recorded was not proportional to the total effort, ignoring area in the analysis could lead to biased estimates if the mean discard per set differs from area to area for a given set type.

Because there were clear differences between the the three set types in per-set discard, we included set type as a covariate for all three model parameters. Thus, with set type and geographic area as the only covariates, our analysis reduced to fitting the model (Eq. 1) independently for each set type, with p, μ, and a having possibly different values in each area. To determine an appropriate dependence upon area, we used stepwise likelihood-ratio tests to select the simplest model that could not be significantly improved by adding additional terms. We first made initial fits for each set type using no areal dependence, then progressively added dependence for more of the model parameters. At each step, we used a quasi-Newton numerical optimization algorithm to maximize the likelihood and estimate parameters. It should be noted that because this is not a linear model, significance levels (i.e. p-values) from these likelihood-ratio tests are approximate. We used the large-sample normal approximation for MLE's to compute standard errors for p, μ, and a. For comparison, we also computed bootstrap standard errors.

It can be shown from the likelihood equations for the NBAZ that estimates for the parameters a and μ depend solely on the positive observations in the data. The estimate for the parameter p depends on all the data, but most strongly upon the proportion of zero observations. Thus, the precision of the estimates for a and μ can be very poor if the data contain few positive observations, even though the precision of the estimate for p may still be very good.

Estimating mean discard per set

We used Equation 2 and the maximum likelihood estimates for p and μ from the best-fit models to estimate mean discard per set for each set type in each area. We also calculated a "pooled" estimate for each set type as the weighted average of the area-specific estimates, where weightings were proportional to total effort in each area. For example, mean discard per set of type i in area j is estimated as

$$\hat{\mathrm{E}}[Y_{i,j}] = (1 - \hat{p}_{i,j})\hat{\mu}_{i,j}, \tag{4}$$

whereas the "pooled" estimate for all areas combined is estimated as

$$\hat{\mathrm{E}}[Y_i]_{pooled} = \sum_j N_{i,j}\hat{E}[Y_{i,j}] \Big/ \sum_j N_{i,j}, \tag{5}$$

where $N_{i,j}$ is the total effort (in number of sets) of type i occurring in area j. Note that this "pooled" calculation is based on the proportion of total sets (including those for which discard was not recorded) observed in each area. This is an estimate of the mean discard per set over the entire fishery during the study period. However, it is also valid as a prediction of future discard if the proportion of effort (sets) in each area remains constant as the actual number of sets varies, assuming that other factors in the fishery, such as size and species composition of discard and style of fishing, remain the same.

While Equation 4 provides a straightforward way to compute the MLE for the product $(1-p)\mu$, the variance of that product can be difficult to estimate accurately. However, we were able to use the likelihood equations for the NBAZ to derive explicit forms for the MLE of mean discard per set. Specifically, only the product $(1-p)\mu$ need be estimated, and we derived, through algebraic manipulation of the likelihood equations, simple closed-form expressions that do not involve the individual parameter estimates. By the invariance properties of maximum likelihood estimates, these simpler forms give results that are identical to those from using Equation 4.

With no areal dependence, the MLE for the product $(1-p)\mu$ is simply the sample mean:

$$\hat{\mathrm{E}}[Y] = \bar{y} = (1/n)\sum_k y_k, \tag{6}$$

where the set type subscript i is suppressed for clarity. Similarly, with complete areal dependence, the MLE for each area reduces to the sample mean in that area, and the "pooled" estimate is computed by using Equation 5. In both of these cases, the variance for the MLE of $(1-p)\mu$ can be estimated by using the sample variance of the data.

When only the mixing probability p depends on area, the MLE for mean discard per set in area j is slightly more complicated, and reduces to

$$\hat{E}[Y_j] = (n_j^+ / n_j) \sum_k y_k^+ / n^+, \quad (7)$$

where n_j^+ and n_j are the number of positive observations and the total number of observations in area j, the y_k^+ are the positive observations in all areas, and n^+ is the total number of positive observations in all areas. Similarly, when only the NB mean μ depends on area, the MLE for mean discard per set in area j reduces to

$$\hat{E}[Y_j] = (n^+ / n) \sum_k y_{j,k}^+ / n_j^+, \quad (8)$$

where n^+ and n are the number of positive observations and the total number of observations in all areas, the $y_{j,k}^+$ are the positive observations in area j, and n_j^+ is the total number of positive observations in area j. Again, Equation 5 is used to compute "pooled" estimates in these latter two cases. Note that the estimates for different areas are not independent, because both Equations 7 and 8 involve observations from all areas. In particular, the first term in Equation 7 is an area-specific estimate of the probability of a positive observation, whereas the second term is a "pooled" estimate of the mean for positive observations. This is consistent with the areal dependence on which Equation 7 is based, and provides more precise estimates of E[Y] than simply taking the sample mean in each area. A similar interpretation holds for Equation 8.

As a consequence of Equations 6, 7, and 8, the estimate of mean discard per set $(1-p)\mu$ can be much more precise than the estimates of the individual parameters involved in it, because it does not depend solely on either the positive observations or the proportion of zeros.

While variance estimators for Equation 6 are straightforward, there is no simple analytic result for estimating the variance of Equations 7 or 8 (see Discussion). Thus, for consistency, we used bootstrap methods in all cases. Our bootstrap resampling procedure varied slightly for each set type, depending on the particular areal dependence chosen for the model parameters. When no dependence was appropriate, data were resampled across all areas. When dependence was important, data were resampled by area in the same proportions as the original observations.

Results

Modelling discard per set

Based on the results of likelihood-ratio tests, geographic area was a statistically significant predictor of discard per set for only two of the three set types (log and school sets).

Nonzero observations of discard from the third set type (dolphin sets) were reported very infrequently (19 out of 2,110 sets, Table 1). The data provided little statistical information from which to distinguish patterns in discard between geographic areas, and area failed to produce a significant improvement in the fit when included as a covariate for dolphin sets. Therefore, we selected the model with no areal dependence for any of the parameters so that the estimates for p, a, and μ for dolphin sets are fisherywide values (Table 2). The standard error of the mixing parameter p for the dolphin model is small (CV=1.53%), reflecting the high estimate for p dictated by the extremely large number of zero observations of discard. The standard errors of the parameters for the NB portion of the probability distribution (a and μ) are quite large (CV's > 90%, Fig. 3), reflecting the few positive data available for their determination.

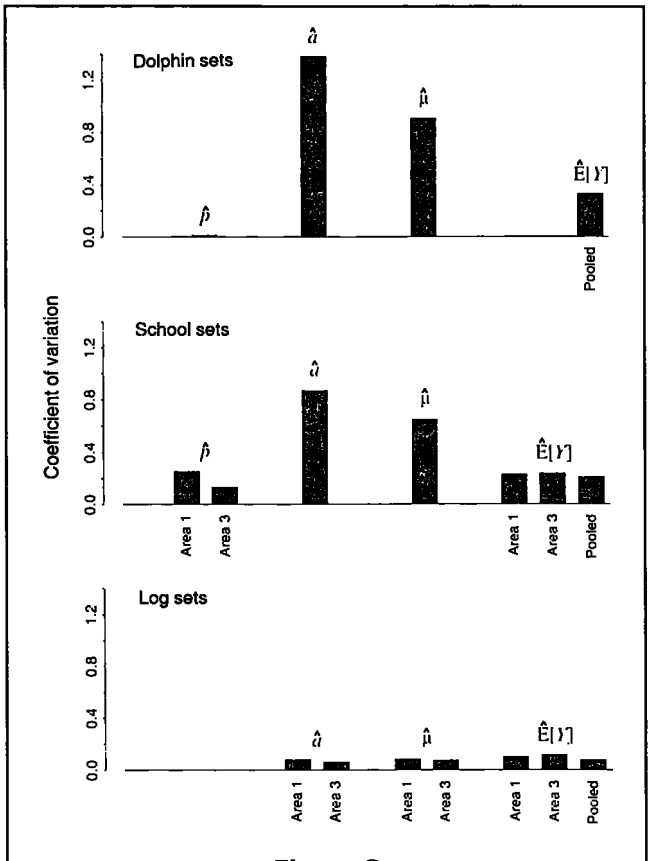

Figure 3

Coefficients of variation for estimates of the model parameters p, a, and μ, and for estimates of mean discard per set for the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92. Geographic areas are defined in Federal Register (1989) (Fig. 1). Pooled estimates are fisherywide, across all areas.

Table 2

Parameter estimates from a fit of the negative binomial with added zeros to tuna discard data from the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92. p is the mixing parameter, and μ and a are the mean and shape parameters of the conditional negative binomial. Standard errors are in parentheses. Geographic areas are defined according to Federal Register (1989) (Fig. 1).

		School sets		Log sets	
Parameters	Dolphin sets	Area 1	Area 3	Area 1	Area 3
p	0.982 (0.015)	0.715 (0.182)	0.825 (0.111)	0 (0)	0 (0)
μ	3.53 (3.21)	5.53 (3.60)		15.4 (1.3)	7.09 (0.55)
a	3.87 (5.36)	7.20 (6.28)		2.34 (0.19)	3.93 (0.25)

At the other extreme, nonzero discard observations were very frequent for log sets (650 out of 998 sets, Table 1). Using fishing area as a covariate for both the mean and shape parameters μ and a, we improved the fit significantly (p-value <0.001) over simpler models. However, the numerical optimization failed to converge to a positive value for p in either area, producing estimates of zero for p in both areas (Table 2). Thus, the estimated probability distributions effectively collapsed to unmodified NB's. Because positive observations were so abundant, estimated standard errors for the mean and shape parameters (Table 2) were small (CV's <8.5%, Fig. 3).

Discard from school sets presented an intermediate case in which we selected a model which included marginally different estimates for the mixing probability p in areas 1 and 3, but no geographic dependence for a or μ (Table 2). Because there were considerably fewer nonzero observations (80 out of 960 sets, Table 1) for school sets than for log sets, parameter estimates were much less precise. Likelihood-ratio tests indicated that fishing area should be included as a covariate for either the shape parameter a or the mixing probability p, but that including areal dependence for p and a simultaneously, or for μ, did not further improve the fit. Because the approximate p-values for adding areal dependence to the two parameters were fairly similar (0.04 for a, 0.12 for p) and the two parameters have similar effects in the model,[1] there was no clear basis for selecting one parameter over the other. We subsequently decided to include areal dependence only for p for two reasons. First, the small number of positive observations for school sets limits the precision of the shape estimate. Second, the difference in the estimated shape between areas was due mainly to two unusually large observations in area 3. Without these two observations, the difference in estimated shapes was reduced, and the significance levels of the two different models were nearly equal (approximate p-values of 0.09). As was the case for dolphin sets, the predominance of zeros in the school set discard data led to small estimated standard errors for the mixing probability p (CV's <13.5%, Fig. 3) but to large estimated standard errors for the mean and shape parameters (CV's > 65%, Fig. 3).

In our model, p may be interpreted as the probability of exactly zero discard, as opposed to small amounts of discard that have been rounded down to zero in the data. The estimates of p for the three set types imply that essentially all dolphin sets (98%) involve no discard, whereas log sets always involve at least some discard. Observer experience[2] indicates that this result is consistent with generally observed patterns for dolphin and log sets.

The estimated shape parameters varied widely between the three set types (Table 2), but the large standard error estimates for the school and dolphin shape parameters prevent us from making any strong statements about shape as a function of set type. As mentioned above, the estimated shape parameter for school sets was strongly affected by the presence of two unusually large observations (100 and 125 tons of discard) in area 3. Repeating the analysis without these two observations led to a shape estimate of 3.75 (SE=2.10), which is more similar to the shape estimates for log sets (2.94 for area 1, 3.93 for area 3).

We could not use bootstrap methods to compute standard errors for dolphin sets because there were so few sets observed with positive discard recorded (Table 1). In resampling for the bootstrap, approximately one-third of the samples contained too few positive observations for the maximum likelihood algorithm to converge. Therefore, Table 2 includes only the standard errors computed from the analytic approximation formulae.

[1] p and a are similar in the effect they have on the estimated distribution. Increasing either one increases the probability of a zero observation, although increasing a also increases the probability of a large observation.

[2] Jackson, A. 1994. Southwest Fisheries Science Center, Natl. Mar. Fish. Serv., P.O. Box 271, La Jolla, CA 92038. Personal commun.

Estimating mean discard per set

Because the model we fitted for discard from log fishing reduced to a simple NB distribution (with $\hat{p}=0$), the estimates of mean discard per log set in each fishing area are just the corresponding mean parameters μ_j. Mean discard per school or dolphin set was estimated with Equation 4.

Estimates of mean discard per log set were an order of magnitude larger than those for school sets and two orders of magnitude higher than those for dolphin sets (Fig. 4). Most of this difference is due to the wide range in the estimated proportion of sets with zero discard. By comparison (Table 2), estimated mean parameters for the NB component of the model differ by less than a factor of five. Thus, the model that we fitted indicates that, on average, there is a considerable difference among set types in per-set discard, although for sets in which discard actually occurs, there is comparatively less difference in the amount.

Mean discard for log sets was estimated at 10.5 t per set pooled over areas, ranging from 7.1 t per set in area 1 to more than double that value (15.4 t per set) in area 3 (Fig. 4). Mean discard for school sets was estimated at 1.16 t per set pooled over areas, ranging from 1.57 t in area 1 to 0.97 t in area 3. Mean discard per set for dolphin sets was estimated at 0.06 t per set fisherywide. Implications of these results for the fishery are discussed in another study (Edwards and Perkins, in prep.).

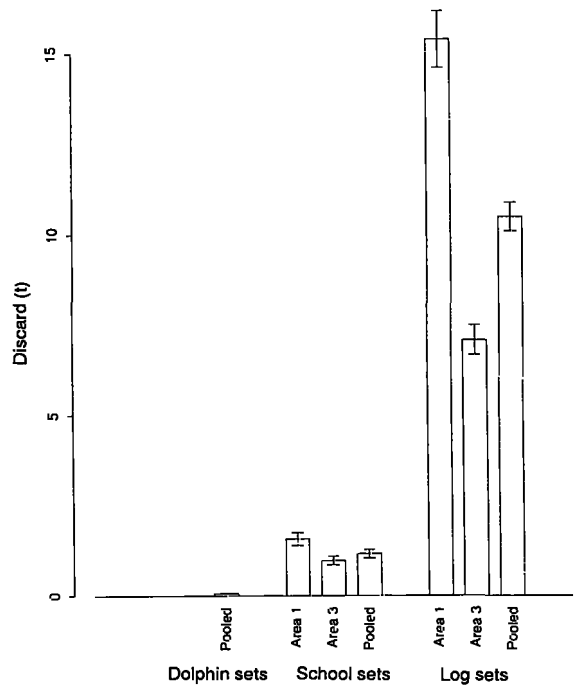

Figure 4

Estimated mean tuna discard per set for the U.S. tuna purse-seine fleet fishing in the eastern tropical Pacific Ocean, 1989–92. Geographic areas are defined in Federal Register (1989) (Fig. 1). Pooled estimates are fisherywide, across all areas. Standard errors are indicated by error bars.

The coefficients of variation (CV's) for the estimates of mean discard per school and dolphin sets (21% and 33%, respectively) are much smaller than those for the individual parameter estimates of α and μ (Fig. 3). As noted in Methods, this is because estimating mean discard per set (i.e. $(1-p)\mu$) is a more robust procedure than estimating the individual parameters. In the case of log sets, the CV's for the estimates of $E[Y]$ and μ differ (Fig. 3), even though in this case the model reduced to a NB distribution where $E[Y] = \mu$. The CV's differ because in estimating variances for the individual parameter estimates we used analytic approximations, while in estimating variances for mean discard, we used bootstrap methods (see Methods).

Where possible (i.e. log and dolphin sets), we estimated variances using the analytic expression in Equation 10 (see Discussion) and found that the results agreed with bootstrap estimates to within about 5%.

Note that the fisherywide estimates for log and school sets are not simply the average of the estimates in each fishing area. This is because the number of sets in each area for which discard was recorded was not proportional to the actual number of sets made in that area. This imbalance was an important reason for including geographic area in the analysis. Nonproportional sampling was not a factor for dolphin sets, because the estimated discard in that case was the same for all fishing areas.

Discussion

Estimating model parameters

Our approach differs somewhat from that of Mangel and Smith (1990), who used the NBAZ to estimate the total biomass of a fish stock. In their analysis, observations of catch from within a stock's geographic range were modelled with the NB component, while the probability mass at zero accounted for observations from outside the range. The sole parameter of interest was the mean μ of the NB component, and fixed values were assumed for the mixing and shape parameters p and α. They reduced the count data to "presence-absence" and derived a likelihood for μ in terms of that reduction. In contrast, we were inter-

ested in estimating the mean of all observations (including "true zeros") and in modelling per-set discard, which requires estimates of p and a. Therefore, we did not follow their approach because we did not have any a priori values for p and a. Reducing counts to simple presence-absence would have decreased the information in the sample such that estimation of the full set of parameters would not have been possible.

Estimating variances for model parameter estimates

The analytic approximation formulae that we used to estimate the variance of the individual parameter estimates are based on the asymptotic normality of MLE's. Since this method uses the estimates for p, a, and μ (rather than their unknown "true" values) in the information matrix, it suffers from the tendency for ML estimates of variance to be biased downwards (e.g. Efron, 1992). We did not attempt to "bias correct" these variance estimates.

When a variance estimate is based on a normal approximation to the sampling distribution of the parameter, the accuracy of the approximation should always be investigated. One way to help validate the normality assumption is to use results from bootstrapping to approximate the sampling distribution. Figure 5 illustrates some examples for the current data. Histograms of the bootstrap replicate parameter estimates for dolphin set data were very skewed. By implication, the normal-approximation variance estimates for the dolphin data, while convenient, are probably not satisfactory. For school set data, histograms of the replicates were slightly skewed because of a small number of unusually large observations. Bootstrap standard errors were consistently higher than the analytic approximations, indicating that the latter may be optimistic. For log set data, histograms were close to normality, and bootstrap standard errors were very similar to those from the analytic approximations. The analytic estimates in this case are probably appropriate.

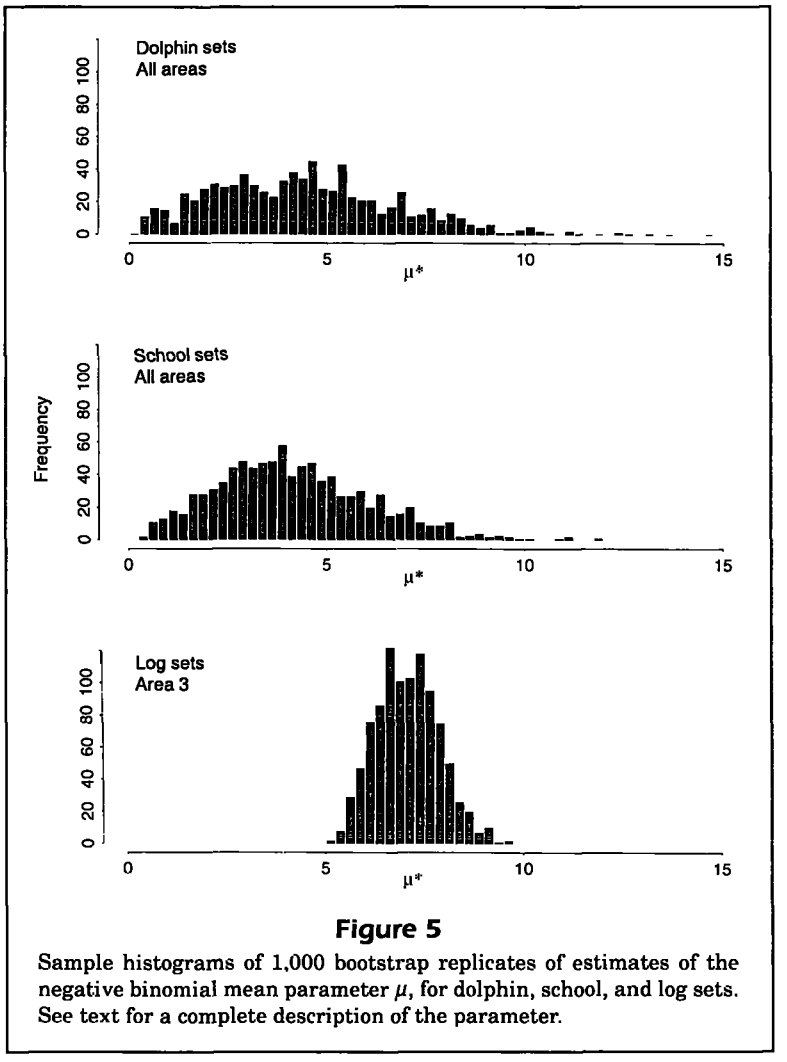

Figure 5

Sample histograms of 1,000 bootstrap replicates of estimates of the negative binomial mean parameter μ, for dolphin, school, and log sets. See text for a complete description of the parameter.

Estimating variances for mean discard per set estimates

In an attempt to derive analytic formulae for the variance of our estimates of E[Y], we manipulated the likelihood equations for the NBAZ and found simplified forms for the MLE of E[Y]. In some cases, the simplified form reduces to the sample mean, Equation 6, and the variance of that estimator is simply (suppressing area and set type subscripts for simplicity)

$$\begin{aligned}\operatorname{var}\left(\hat{\mathrm{E}}[Y]\right) &= (1/n)\operatorname{var}[Y] \\ &= (1/n)(1-p)\left(\mu+(a+p)\mu^2\right),\end{aligned} \tag{9}$$

which can be estimated by substituting MLE's for a, μ, and p. More simply, by using the fact that the estimator is just the sample mean, the minimum variance unbiased estimate of Equation 9 is the sample variance,

$$\hat{\operatorname{var}}\left(\hat{\mathrm{E}}[Y]\right) = \left[\sum_i (y_i-\bar{y})^2\right]/(n-1), \tag{10}$$

where $\bar{y}$ is the sample mean. In other cases, the simplified forms for the MLE of E[Y] are slightly more complex (Eqs. 7 and 8), and Equations 9 and 10 no longer apply. We did derive expressions, analogous to Equation 9, for the variance of Equations 7 and 8 in terms of the three model parameters p, α, and μ. However, these formulae are so complex as to be of no practical use in estimation, and no expression analogous to Equation 10 seems possible.

Rounding errors in the observations

The NBAZ model used in this study comprises two components. As noted in the description of the model, zero values derived from the NB component can be interpreted as observations of small amounts of discard, rounded down to zero, whereas zero values from the probability mass component can be interpreted as exact zeros. This interpretation is based on the assumption of an underlying continuous distribution for positive discard amounts (e.g. a gamma distribution) upon which rounding errors have been superimposed.

One consequence of this interpretation is that the mean amount of discard that should be associated with "true zeros" is zero, and the mean amount that should be associated with "NB zeros" is nonzero. Thus, strict adherence to this interpretation of zeros leads to the conclusion that Equation 4 may be an underestimate of E[Y]. However, if we assume a strictly decreasing underlying distribution for positive discard, symmetric rounding of amounts larger than one-half ton would tend to increase the estimate. In the absence of a specific model for the rounding errors, we did not attempt to correct for any bias due to rounding.

The EM algorithm for maximizing likelihood

We used a quasi-Newton algorithm to maximize likelihood for the parameters p, α, and μ. A useful alternative for mixture models, including "added zero" distributions, uses the EM algorithm to maximize likelihood (e.g, McLachlan and Basford, 1988; Lambert, 1992). In situations with many covariates, it provides a well-behaved alternative to the high-dimensional gradient search required by general optimization algorithms. The algorithm can be implemented by using standard regression techniques for generalized linear models. We applied the EM algorithm to the NBAZ using a combination of logistic regression to maximize likelihood for p and quasi-likelihood NB regression for μ and α (Lawless, 1987). However, the logistic regression failed to converge for the current data because the ML estimate of p for log sets was zero.

Alternative models considered

We considered but rejected two alternatives to the NBAZ model: 1) the Δ-distribution (a mixture of a probability mass at zero with a lognormal [Aitchison, 1955; Pennington, 1983]); and 2) a Γ-distribution mixed with a probability mass at zero (Coe and Stern, 1982). Both have been used in similar cases where the data to be analyzed have contained large numbers of zeros. The Δ-distribution assumes that the natural logs of the positive observations are distributed normally, or can be so transformed, and this assumption was not plausible. The data in this analysis were rounded to the nearest ton and the mode of the positive observations was one ton. Thus, no transformation could bring these data to even approximate normality. The gamma mixture model was not appropriate for the current data because maximum likelihood estimation for a highly skewed gamma distribution depends heavily upon small (near zero) observations. In this study, all observations in that region were rounded to either zero or one, implying a large relative measurement error and therefore potentially poor accuracy. Another more fundamental reason why we rejected these two models was that both models mix a continuous distribution on the positive numbers with a probability mass at zero and assume that observations from each component remain distinguishable. In the current data set, small positive observations are grouped together with zero observations, and using an NB in the mixture allows the model to distinguish between "true zeros" (actual absence of discard) and "rounded zeros" (discard so small that it was ignored or missed).

Conclusions

The methods developed here were used to model fisheries discard data which were rounded to integer values and which included widely varying numbers of zero observations, depending on one or more covariates. The usual models for integer-valued data (e.g. the Poisson distribution) did not fit the data at all well because of the extreme skewness of some of the observed distributions. The NBAZ is more flexible than the standard models and provided a much better fit. In general, the model is applicable to any set of integer-valued data which exhibit a large proportion of zero observations combined with long positive tails. Both categorical and continuous covariates may be used.

Modelling these data with a parametric probability distribution allowed us to describe patterns in the discard in some detail, for example, in estimating the percentage of "true zeros" in the data. Addi-

tionally, we were able to examine whether differences in mean discard were due to different proportions of zero observations or to different distributions of positive observations. In contrast, computing sampling-based estimates of population mean and variance would not give any indication of the patterns in the individual observations. While average or total discard is of significant interest, it is also important to quantify the amount of discard possible for an individual set. Assuming that the parametric model is accepted as appropriate, one can estimate, for example, the probability that, owing to random chance alone, discard from a particular boat will exceed a certain limit in a fixed number of sets. One can also estimate the percentage of zero observations which actually represent small amounts of discard. Finally, a parametric model provides a natural framework for predicting future discard.

Acknowledgments

We thank Al Jackson for his generous help and advice about data collection methods and data form interpretation, Pierre Kleiber for his comments on an earlier version of this manuscript, and the Inter-American Tropical Tuna Commission for making available their data on U.S. tuna vessel tuna discards. We especially thank Cleridy Lennert of the IATTC for her helpful suggestions. We also thank Ronald Hardy and two anonymous reviewers whose helpful suggestions added to the clarity of our manuscript.

Literature cited

Aitchison, J.
1955. On the distribution of a positive random variable having a discrete probability mass at the origin. J. Am. Stat. Assoc. 50:901–908.

Berger, J. D., R. F. Kappenman, L. L. Low, and R. J. Marasco.
1989. Procedures for bycatch estimation of prohibited species in the 1989 Bering Sea domestic trawl fisheries. U.S. Dep. Commer., NOAA Tech. Memo. NMFS-F/NWC-173, 23 p.

Caillouet, C. W., Jr., M. J. Duronslet, A. M. Landry Jr., D. B. Revera, D. J. Shaver, K. M. Stanley, R. W. Heinly, and E. K. Stabenau.
1991. Sea turtle strandings and shrimp fishing effort in the northwestern Gulf of Mexico, 1986–89. Fish. Bull. 89:712–718.

Coe, R., and R. D. Stern.
1982. Fitting models to daily rainfall. J. Appl. Meterol. 21:1024–1031.

Collins, M. R., and C. A. Wenner.
1988. Occurrence of young-of-the-year king, *Scomberomorus cavalla*, and Spanish, *S. maculatus*, mackerels in commercial-type shrimp trawls along the Atlantic coast of the southeast United States. Fish. Bull. 86:394–397.

Efron, B.
1992. Six questions raised by the bootstrap. *In* R. LePage and L. Billard (eds.), Exploring the limits of the bootstrap, p. 99–126. John Wiley and Sons, New York, NY.

Federal Register.
1989. Interim final rule with request for comments;Tuesday, 7 March 1989. Federal Register, Vol. 54(43), Rules and Regulations:9438–9451.

IATTC (Inter-American Tropical Tuna Commission).
1989. Annual Report: 1988. Inter-American Tropical Tuna Commission, La Jolla, CA.

Johnson, N. L., and S. Kotz.
1969. Distributions in statistics: discrete distributions. Houghton Mifflin, Boston, MA, 187 p.

Joseph, J.
1994. The tuna-dolphin controversy in the eastern Pacific Ocean: biological, economic, and political impacts. Ocean Developments and International Law 25:1–30.

Lambert, D.
1992. Zero inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34(1):1–14.

Lawless, J. F.
1987. Negative binomial and mixed Poisson regression. Can. J. Stats. 15(3):209–225.

Mangel, M., and P. E. Smith.
1990. Presence-absence sampling for fisheries management. Can. J. Fish. Aquat. Sci. 47(10):1875–1887.

McLachlan, G. J., and K. E. Basford.
1988. Mixture models: inference and applications to clustering. Marcel Dekker, New York, NY, 253 p.

NRC (National Resource Council).
1992. Dolphins and the tuna industry. National Academy Press, Washington, D.C., 176 p.

Pennington, M.
1983. Efficient estimators of abundance for fish and plankton surveys. Biometrics 39(1):281–286.