Available online at www.sciencedirect.com
SCIENCE DIRECT®
Fisheries Research 70 (2004) 299–310
FISHERIES RESEARCH
www.elsevier.com/locate/fishres

A multispecies approach to subsetting logbook data for purposes of estimating CPUE

Andi Stephens[a,*], Alec MacCall[b]

[a] *Department of Applied Mathematics and Statistics, Center for Stock Assessment Research, Jack Baskin School of Engineering, University of California, Santa Cruz, CA 95064, USA*
[b] *National Marine Fisheries Service, Southwest Fisheries Science Center, 110 Shaffer Rd., Santa Cruz, CA 95060, USA*

Abstract

An initial step in catch and effort analysis is determination of what subset of the data is relevant to the analysis. We propose an objective approach to subsetting trip records of catch and effort data when fishing locations are unknown; the species composition taken on a fishing trip is used to infer if that trip's fishing effort occurred in a habitat where the species of interest (the target species) is likely to occur. We use a logistic regression of multispecies presence–absence information to predict the probability that the target species would be present. A critical value of probability that best predicts target species presence and absence in the data set forms an objective basis for subsetting the trip records. We test this approach by applying it to a data set where individual fishing locations are known, and we show that the method is an effective substitute for information on individual fishing locations.

Keywords: CPUE; Catch-per-unit-effort; Logistic regression; Habitat; Multispecies

1. Introduction

An initial step when analyzing large data sets often involves separating the data into the subset of observations that is considered to be relevant and informative, which is retained for analysis, and the subset of observations that is considered to be uninformative, which is discarded. We refer to this process as 'subsetting' the data. In practice, subsetting is often based on ad hoc and subjective decision rules, and introduces a source of uncertainty into the analysis that is seldom evaluated. We propose an objective decision rule for subsetting catch and effort data based on the species composition of catches taken on individual fishing trips. Unlike an ad hoc decision rule, calculations based on this decision rule are reproducible by independent analysts and the results are amenable to statistical analysis, including the estimation of precision.

Fishery data in the form of landings receipts, logbooks, or catches sampled directly in the field often reflect a variety of alternative species or habitats targeted by the fishermen, even within a single fishing trip. Consequently, some of the records in a data set may not be relevant to calculating catch-per-unit-effort (CPUE)

* Corresponding author. Tel.: +1 831 459 5385; fax: +1 831 688 7087.
E-mail address: andi@soe.ucsc.edu (A. Stephens).

0165-7836/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.fishres.2004.08.009

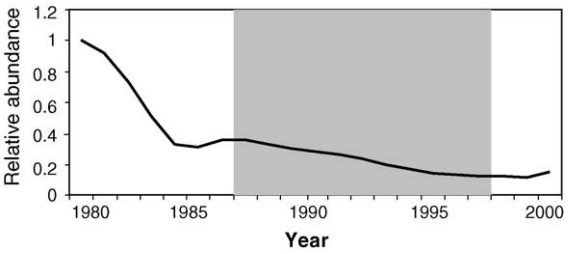

Fig. 1. Relative abundance of bocaccio over the period covered by the MRFSS and CDF&G surveys (MacCall, 2003). The shaded region denotes the period covered by the CDF&G survey.

for a particular species (referred to here as the target species). For example, the Marine Recreational Fishery Statistics Survey (MRFSS) (Osborn et al., 1996) provides records of species catch and angler effort since 1980 for recreational fishermen on the west coast of the United States. If these records are to be used as the basis of a CPUE index of abundance for a particular target species, one of the first steps in the analysis is to distinguish which of the catch and effort records are informative for that species and which are not.

Bocaccio (*Sebastes paucispinis*) forms a focus for this study. Bocaccio is a mobile species with weak site-fidelity until late maturity, although it is found in close association with similar rockfish species along rocky bottoms (Love et al., 2002). Historic abundance has been estimated by MacCall (2003) based on a number of different abundance indices (Fig. 1). The abundance of bocaccio declined severely after the early 1970s, and a current management goal is to rebuild the stock (MacCall, 2003).

A CPUE index of abundance is potentially valuable for assessing bocaccio. However, fishing trips that targeted tuna or salmon are unlikely to provide information on the abundance of a groundfish species such as bocaccio, and fishing trips that encountered these pelagic species should clearly be deleted when subsetting a data set such as MRFSS. However, even with this improvement, the data remaining may contain an unknown proportion of fishing trips that did not sample bocaccio habitat, and that proportion may vary substantially from year to year, contributing to imprecision or spurious trends in a CPUE index of bocaccio abundance. Choices of where to fish may be influenced by, for example, environmental conditions, expected catch rates, or changes in fishing regulations. The latter two influences are likely to exhibit long-term changes over time.

If fishing locations were included in the records, it would be possible to restrict the analysis to catch and effort data for only those locations known to be bocaccio habitat. However, information on fishing location may not be available. For example, the MRFSS data were usually collected dockside at the end of the fishing trip, and do not indicate where the actual fishing occurred, nor how many locations were fished. In this paper, we examine an approach to 'subsetting' that uses the species composition from fishing trips to infer whether the fishing occurred in habitat appropriate for use in CPUE calculations.

2. Materials and methods

2.1. Data

Partyboats (a.k.a. commercial passenger fishing vessels) are vessels that run regularly scheduled fishing trips for which tickets are sold to the public. Partyboats represent a major segment of the recreational fishery off the west coast of the United States. We believe that partyboat trips sample the species composition at each location visited during a fishing trip better than private boat trips because the catch from a partyboat trip usually represents the fishing effort of many more anglers.

Three data sets for partyboats off northern California are considered in the analyses of this paper: (a) catch and effort data sampled by the MRFSS program (1980–1989; 1993–1999) (MRFSS), (b) site-specific catch and effort data sampled onboard fishing vessels by the California Department of Fish and Game (CDF&G) (1987–1998) ('CDF&G site-visit'), and (c) a version of the second data set created by reorganizing the CDF&G records so that site visits are aggregated into records of (location-blind) trips ('CDF&G aggregate-trip'). After calculating CPUE for bocaccio, all CDF&G and MRFSS catch data were converted from their original values to categorical presence/absence indicators (1/0).

The data from the MRFSS program were obtained from the RecFIN database (VanBuskirk, 2003). The MRFSS data are compiled from post-fishing interviews on the dock. MRFSS aims to obtain the distribution of the catch-per-trip at the species level, the unit of

nominal fishing effort is an angler-trip, and fishing locations are not recorded. Many records, especially those from the early years, are incomplete or unclear (e.g. lacking information on date, number of anglers, or species caught). Deletion of such records prior to analysis reduced the data set by 20%. Data after 1999 were available, but were not included in the analyses because of major changes in fishing regulations, including reduced bag limits. The MRFSS/RecFIN data comprise 12 905 usable records of catch composition and fishing effort.

The CDF&G data were provided by D. Wilson-Vandenberg (CDF&G, pers. commun). The CDF&G sampling recorded catches and effort (in angler-minutes) at specific fishing sites. Data recorded by the CDF&G program include the location and duration of fishing at each site, the maximum and minimum depth at the site, and the number of each species of fish caught. We used 4544 per-site fishing observations from this dataset, comprising 458 locations and 106 species, and covering the period January 1987–December 1998. The CDF&G program did not actively sample party-boat trips targeting salmon or tuna, and thus represents a subset of the MRFSS sampling frame (although not of its data; the two programs were conducted independently).

Ideally, a set of reference locations would be chosen for estimating CPUE. These are locations known to have good catch rates for the species of interest. This precludes consideration of locations that are rarely visited and locations at which the target species is rarely caught from having undue influence on CPUE. We used only those data pertaining to locations at which bocaccio had been caught ten or more times, comprising 54 reference locations from the 458 locations fished, for comparison of CPUE estimates in the CDF&G data (Fig. 2).

The estimated abundance of bocaccio available to the central California recreational fishery declined by two thirds during the 1987–1998 period sampled by the CDF&G program and by over 80% during the 1980–1999 period sampled by MRFSS (Fig. 1).

2.2. Catch-per-unit-effort

Determining which catch and effort records pertain to a particular target species, involves discriminating between trips that fished in habitat where the target species is found (which will be referred to as target habitat) from trips that fished in non-target habitat, i.e., in which the target species was unlikely to be caught. The latter trips are not informative, and potentially contaminate the calculation of CPUE. Ideally, nominal fishing effort (E) and the fishing mortality rate (F) for a species are related by a catchability coefficient, q:

$$F = qE \tag{1}$$

and average abundance (B) is related to the CPUE by:

$$B = \left(\frac{1}{q}\right)\left(\frac{C}{E}\right) \tag{2}$$

where C is the catch.

The actual value of the catchability coefficient may not be known, but, under the assumption that it is constant, CPUE is often used as an index of relative abundance when conducting stock assessments. Ideally, the measure of nominal fishing effort is defined so as to be proportional to the fishing mortality rate that it generates (Ricker, 1975). Thus, the catchability coefficient is equal to the fishing mortality rate generated by one unit of nominal fishing effort. Fishing is unlikely to catch the target species in non-target habitat, so $C \approx 0$ and $q \approx 0$. If the catch and effort records reflect a mixture of fishing activity in both target and non-target habitats, the catchability coefficient reflects the proportions of target and non-target effort in the mixture:

$$B = \left(\frac{1}{q_{\text{mixed}}}\right)\left(\frac{C_{\text{tar}}}{E_{\text{tar}} + E_{\text{nom}}}\right) \tag{3}$$

The subscript tar in Eq. (3) indicates records from target habitat, non-indicates records from non-target habitat,

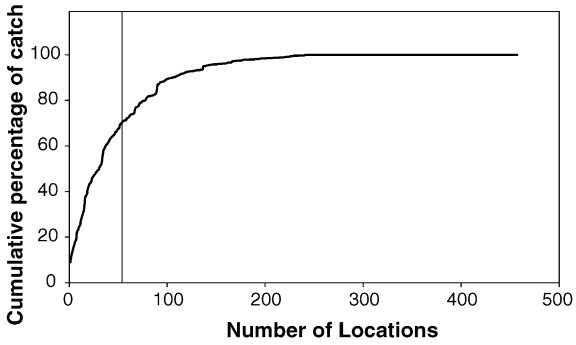

Fig. 2. The cumulative percentage of bocaccio catch vs. the number of locations in the CDF&G data. The vertical line indicates the contribution to the total catch of the 54 reference locations.

and q_{mixed} refers to the catchability coefficient that applies to the combined data. This may not pose a serious problem under some circumstances. For example, if the data contain a constant proportion of target to total effort, the value of q_{mixed} will be smaller than q_{tar} by the ratio $E_{tar}/(E_{tar} + E_{nom})$, but will still be constant. However it is unlikely that this ratio will be invariant over long periods of time because many of the factors influencing the behavior and preferences of recreational fishermen may change.

Historically, calculation of CPUE involved straightforward ratio estimators, often supplemented by complicated analyses of fishing power used to address systematic differences in the catchability coefficient among different classes of vessels in the fleet (Gulland, 1983). More recently, generalized linear models (GLMs) have been used to derive indices of abundance more directly from catch and effort data (Stefánsson, 1996). A major advantage of the GLM approach is that a wide variety of influences on the catchability coefficient can be accounted for in a relatively simple analysis. For example, the distinction of target and non-target habitats is straightforward if fishing locations are known, and this can be incorporated directly in the analysis. Using the notation in the 'R' computing language (Ihaka and Gentleman, 1996), the CPUE index can then be obtained using a GLM of the form:

$$\log(\text{CPUE}) \sim \text{year} + \text{location} + \text{other} \quad (4)$$

where the exponentiated 'year' effects estimated by the model serve as the CPUE index. The 'location' effects account for systematic differences among fishing locations, and the 'other' effects could include sources of variability such as seasonal patterns in fish abundance or availability. Although, in principle, this approach could be applied to the entire catch and effort data set, it is still advantageous to delete records for locations that rarely or never produce the target species because the GLM treats fluctuations in relative CPUE at all locations as being equally informative. For example, if CPUE declines by half at well-measured target locations, CPUE should also decline by half at locations which rarely produce any catch of the target species, even though that change would scarcely be measurable. Of course, in the case where locations are known, it is rather easy to subset the data to include records only for those locations that consistently produce catches of the target species.

In this paper, we address the problem of how to subset catch and effort data for estimation of CPUE when fishing locations are not known. The proposed method uses the observed species composition to infer whether the fishing effort occurred in a habitat in which the target species would be expected to live. This inference takes the form of a logistic regression (described below) that uses the presence or absence of other common species to estimate the probability that the target species would be encountered. Selection of a critical value allows the catch and effort data to be divided into the records in target and non-target habitat. Once the data have been 'subsetted', the CPUE index can be obtained using a GLM of the form:

$$\log(\text{CPUE}) \sim \text{year} + \text{other} \quad (5)$$

where the exponentiated 'year' effects provide the CPUE index, and 'other' refers to any additional factors. The data include numerous records for which bocaccio CPUE was zero. We used a delta-gamma GLM, where presence–absence is model and using a logistic regression (binomial family in the R computing package), and the records with non-zero values are modeled using a separate GLM assuming a gamma probability distribution (Stefánsson, 1996; Dick, 2004). Estimates of precision for the annual CPUE indices are obtained using a jackknife procedure (Belsley et al., 1980).

The model we used to calculate CPUE is a main-effects model. We investigated interaction terms and found they were rarely significant and ranged between three and five orders of magnitude smaller than the main effects, justifying their omission (Maunder and Punt, 2004).

2.3. Logistic regression

Statistical classification problems, such as the present subsetting problem, are typically addressed using either discriminant function analysis or logistic regression. Press and Wilson (1978) reviewed the properties and performance of these two approaches. Discriminant function analysis (McCullagh and Nelder, 1989) requires that the variables be normal with identical covariance matrices. Logistic regression with

maximum likelihood estimation is preferable if the explanatory variables are not multivariate normal, such as in the present case where they are categorical variables.

Although individual fishing locations may not be known, the species composition of a fishing trip provides information that can be used to infer whether the fishing trip included effort expended in target habitat. We use a logistic regression to make this inference. The species compositions from catch records are first used to estimate the parameters of the logistic regression which then used to estimate the probability that the target species would have been encountered on each trip. Those records for which the estimated probability exceeds a chosen critical value are then used in the CPUE analysis with some assurance that many of the records of catch and effort from non-target habitat have been removed.

Let Y_j be a categorical variable describing the presence/absence of the target species for trip j:

$$Y_j = \begin{cases} 1 & \text{if the target species is caught} \\ 0 & \text{if the target species is not caught} \end{cases}$$

Similarly, let x_{ij} describe the presence/absence of non-target species i in the catch during trip j.

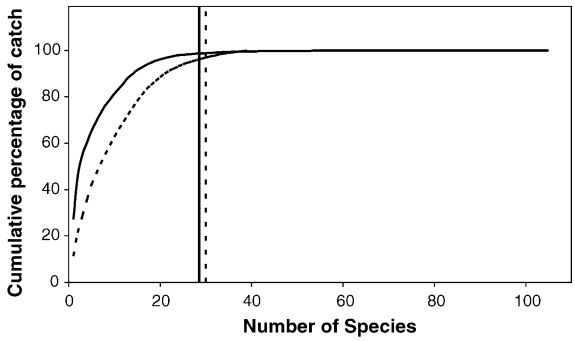

Fig. 3. The cumulative percentage of catch vs. number of species in the CDF&G (solid line) and MRFSS (dashed line) data sets. The vertical lines indicate the contributions of the species used in the analyses.

We assign a score for each trip j as a function of the species (1, 2, . . ., k) caught during that trip:

$$S_j = \exp \sum_{i=0}^{k} x_{ij}\beta_i \tag{6}$$

The coefficients β_1, β_2, . . ., β_k quantify the predictive impact of each species while β_0 is the intercept of the regression – the probability that fishing was in the habitat of the target species when none of the others species was present.

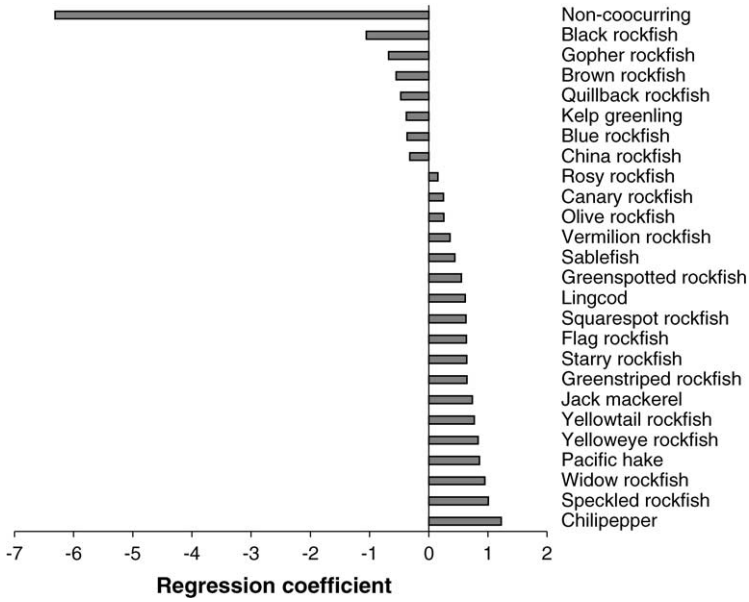

Fig. 4. Estimates of species-specific regression coefficients based on the 'CDF&G site-visits' data set.

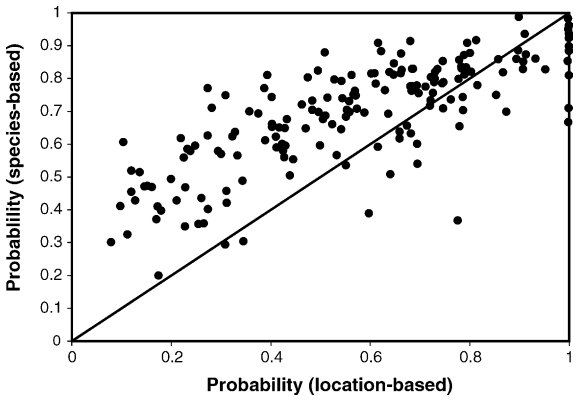

Fig. 5. Per-location probabilities of encountering bocaccio based on regressions using location (*x*-axis) and species composition (*y*-axis) as predictors.

This score is then converted into a probability of observing the target species given the vector of presences and absences of the k non-target species:

$$\pi_j = \Pr\{Y_j = 1\} = \frac{S_j}{1 + S_j} \tag{7}$$

where π_j is the predicted probability that $Y = 1$ for trip j.

Given $\beta_0, \beta_1, \ldots, \beta_k$ and the presence/absence indicators $x_{1j}, \ldots, x_{kj}$, the log-likelihood (excluding constants independent of the parameters) is the sum:

$$\boldsymbol{L}\{Y|\beta_0, \ldots, \beta_k, x_{1j}, \ldots, x_{kj}\} = \sum_{j \in j+} \log(\pi_j) + \sum_{j \in j-} \log(1 - \pi_j) \tag{8}$$

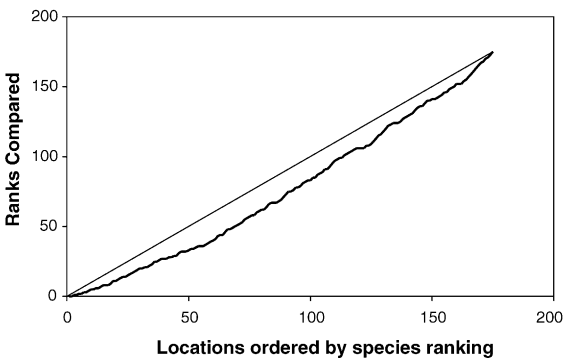

Fig. 6. Locations ranked by the species composition method (best to worst) – *x*-axis, and the number of locations ranked equally or better using Eq. (9) – *y*-axis.

where $j+$ denotes records where the target species was caught, and $j-$ denotes records where the target species was not caught.

The log-likelihood is maximized using the statistical package R (Ihaka and Gentleman, 1996). The estimated β coefficients reflect the association (positive or negative) between the non-target and the target species, and the π_j is the estimated probability that trip j occurred in the habitat of the target species.

The set of trips to be used in the CPUE analysis is defined as those for which π calculated above is less than a critical value. The critical value is selected so the number of incorrect predictions (both false positive – the target species is estimated to be found in the habitat fished during the trip when it does not, and false negatives – the target species is estimated not to be found in the habitat fished when it does) is a minimum. This

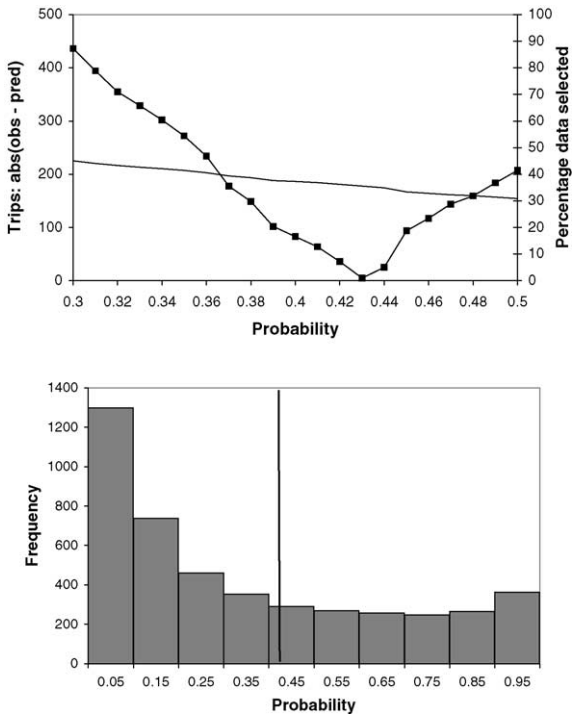

Fig. 7. Results of the application of the proposed method to the 'CDF&G site-visit' data ($n = 4544$). The upper panel plots the difference between the number of records in which bocaccio are observed and the number in which they are predicted to occur (symbols), and percentage of records retained (solid line), as a function of the critical value while the lower panel shows a histogram of the probabilities generated by the species-based regression. The vertical line indicates the critical value for which false prediction is minimized.

number is quantified by the absolute value of the difference between the number of trips observed to have caught the target species, and the number proposed to be in target habitat. We evaluate this difference as the critical value is increased from zero (all trips are in target habitat) to one (no trips are in target habitat) and identify the value that leads to the smallest absolute difference.

2.4. Validation with known locations

The 'CDF&G site visits' data set (for which location is known) was analyzed in two ways as a 'sea truth' to validate the proposed 'subsetting' approach:

(a) We fitted the following model, which includes location as a covariate, assuming a binomial error distribution, to estimate the probability of encountering bocaccio at each location:

$$Y \sim \text{location} + \text{year} + \text{season} \quad (9)$$

where Y indicates bocaccio presence/absence and there are 12 years and four (trimester) seasons. Interaction terms could be included in Eq. (9) but their inclusion was not supported statistically.

(b) We applied the proposed 'subsetting' approach to determine probability of encountering bocaccio in each location.

This validation analysis was performed for all catch records for the locations at which bocaccio occurred at least once.

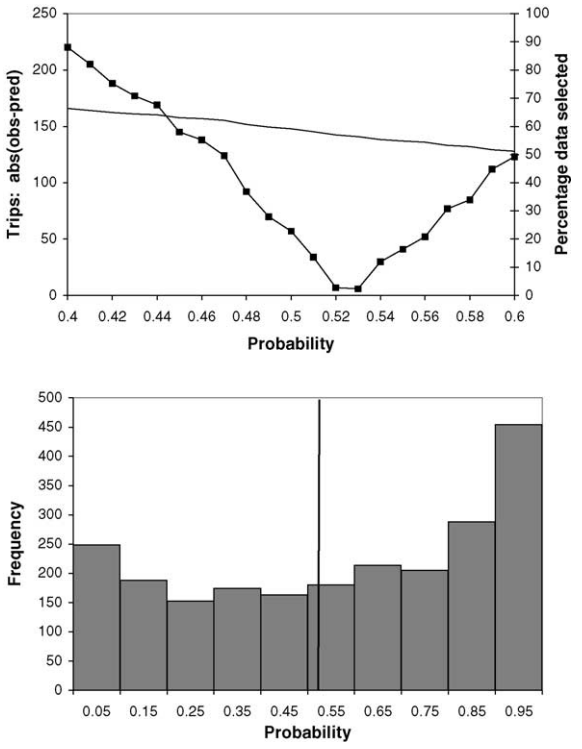

Fig. 8. Results of the application of the proposed method to the 'CDF&G aggregate-trip' data (n = 2267). The upper panel plots the difference between the number of records in which bocaccio are observed and the number in which they are predicted to occur (symbols), and percentage of records retained (solid line), as a function of the critical value while the lower panel shows a histogram of probabilities generated by the species-based regression. The vertical line indicates the critical value for which false prediction is minimized.

3. Results

3.1. Validation with known locations

We compared the performance of the proposed method for 'subsetting' catch and effort records (Section 2.3) with the location-based method (Eq. (9)) using the 'CDF&G site visits' data set. 106 species are recorded in this data set, but 30 account for 99% of the catch (Fig. 3). The two methods were therefore applied to both the full (106 species) and restricted (30 species) data sets. The results are insensitive to the number of species, so the results reported pertain to the 30 species data set only. A backwards stepwise-regression procedure was used to reduce the regressor species used by the proposed method further. Fig. 4 shows the regression coefficient for each non-target species retained for the analysis of site-visits. Species that were never caught with bocaccio are lumped into a category of 'non-coocurring species'.

Fig. 5 compares the estimated probability of encountering bocaccio for each location from: (a) Eq. (9) – x-axis, and (b) the proposed method – y-axis. The estimated probability of encountering bocaccio is higher for the proposed method than when direct account is taken of location. However, for the purposes of subsetting the data, the important issue is the relative ranking of locations and not the estimated probability of encountering bocaccio. Fig. 6 therefore plots the locations ranked by the species-based method (x-axis) against the number of locations ranked equally or better by Eq. (9) (y-axis).

Fig. 7 provides additional diagnostic statistics for the proposed method. The critical probability at which the difference between the observed and expected number of trips encountering bocaccio is minimized is clearly defined and equals 0.43 (Fig. 7, upper panel). About one-third of the records are selected for use in calculating the CPUE index, although this fraction is not particularly sensitive to the critical value in the range evaluated (Fig. 7, solid line). The distribution of the probability of encountering bocaccio among sites suggests that many site visits have very little chance of catching bocaccio (Fig. 7, lower panel). These are the least relevant records for estimating the CPUE index, and are discarded by the subsetting procedure.

3.2. Evaluation of aggregate trip data

The critical probability value increases from 0.43 to 0.53 (Fig. 8, upper panel), and the distribution of probabilities shifts to larger values (Fig. 8, lower panel) when the CDF&G data are aggregated. Actual fishing trips rarely visit only one location, and, in fact, usually visit at least two locations per trip which means that a greater percentage of the aggregate trips encounter bocaccio at some point.

Another change that occurs when the data are aggregated is that fewer explanatory species remain from the original 30 used when analyzing the site-visit data after the stepwise-regression (Fig. 9). Since the catch in an aggregate trip includes more species than an individual site-visit catch, species that were only weakly informative for site-specific data become even less informative for aggregate data.

3.3. Application to the MRFSS data

We used 30 species when applying the proposed method to the MRFSS data to be consistent with the analysis of the CDF&G data. This amounts to 75% of the species, and 97% of the catch (Fig. 3). The critical value analysis (Fig. 10, upper panel) and probability histogram (Fig. 10, lower panel) suggest that bocaccio are less prevalent in the MRFSS data set than in the CDF&G data set. This reflects a difference in the data collected. For example, the MRFSS data set includes a large number of salmon and tuna trips, which typically do not visit bocaccio habitat. Figs. 4 and 11 show that the relationships among the species are consistent (in terms of both magnitude and sign of their associated coefficients) between the MRFSS and CDF&G data.

3.4. CPUE analysis

The decline of the CPUE indices based on the full (i.e. no exclusions of non-targeted records) 'CDF&G site-visits' data (open squares in Fig. 12, upper panel) is exaggerated compared to that of the CPUE indices

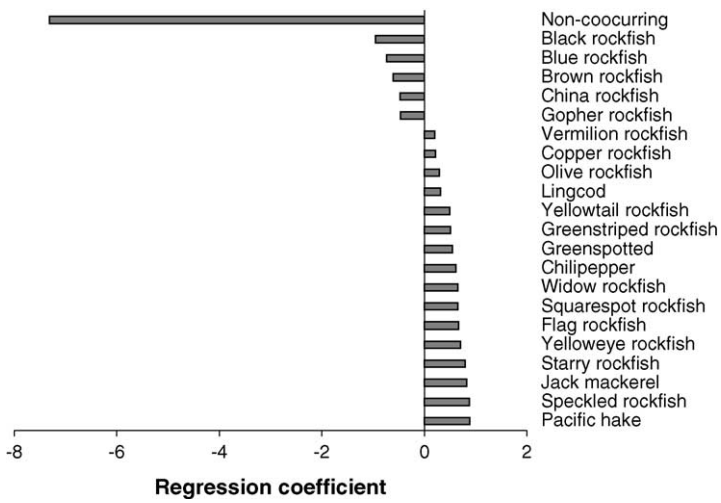

Fig. 9. Estimates of species-specific regression coefficients based on the 'CDF&G aggregate-trip' data set.

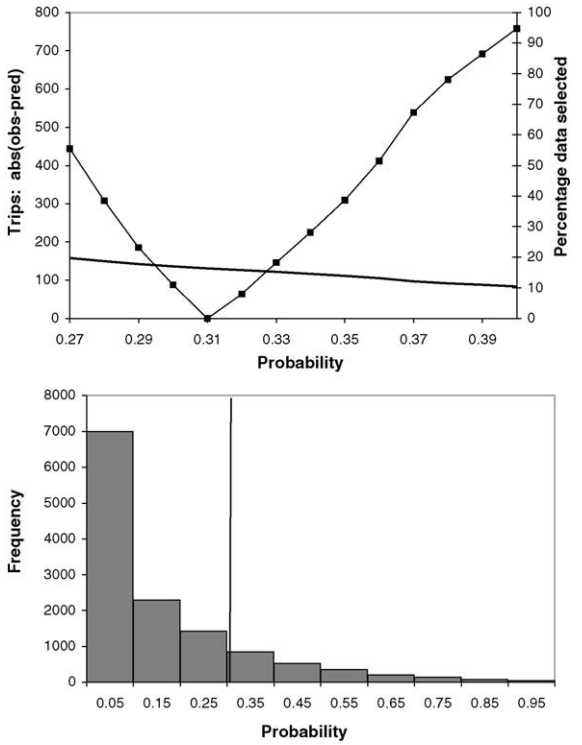

Fig. 10. Results of the application of the proposed method to the MRFSS data (n = 12 905). The upper panel plots the difference between the number of records in which bocaccio are observed and the number in which they are predicted to occur (symbols), and percentage of records retained (solid line), as a function of the critical value while the lower panel shows a histogram of probabilities generated by the species-based regression. The vertical line indicates the critical value for which false prediction is minimized.

based on data subsetted by location (open circles) or species catch composition (closed circles), particularly after 1995. A similar exaggerated decline in CPUE is apparent for the 'CDF&G aggregate-trip' data set (Fig. 12, middle panel).

Subsetting the MRFSS data changes some of CPUE indices considerably (Fig. 12, lower panel). For example, 1998 was an El Niño year, and a good year for tuna. Many partyboat trips specifically targeted tuna that year. Compared with the abundance trends in Fig. 1 (which were based on nine data sets), the CPUE index from the species regression follows the initial decline to 1984 better than the CPUE index from the full data set, and, apart from 1993 and 1994, is relatively constant during the 1990s.

Other discrepancies in these data may be explained in terms of life-history. Bocaccio recruitment is generally low with rare, large recruitments. The years 1980 and 1985 were large recruitment years (MacCall, 2003), providing large numbers of young fish for anglers in 1982 and 1986. The CPUE indices for 1982 and 1986 based on the full data set are much higher than those based on 'subsetted' data presumably because bocaccio were being caught outside the usual habitats (trips in such habitats are assigned low probabilities by the proposed method and may be discarded) as well as within them.

3.5. Site-specific changes in effort

The number of locations visited per trip in the CDF&G data, and the percentage of fishing time spent at the top 54 bocaccio locations (those at which bocaccio occurred 10 or more times) changed over time. According to the CDF&G 'site-visit' data set, the average number of locations visited during a trip rose by 45% from 1987 to 1998, while the number of visits to top bocaccio sites stayed the same, indicating an increasing diversification of fishing sites over time (Fig. 13, upper panel). The percentage of the time spent fishing the best bocaccio sites dropped by 64% over 1987–1998. In other words, during the period of bocaccio decline, vessels switched targets and progressively targeted habitats where bocaccio were less likely to be present. This target switching could not have been easily detected without the location-based data, and its effect cannot be entirely removed from species-subsetted data (Fig. 13, lower panel); the same pattern of target diversification persists, although the trend is less pronounced. There is a 40% increase in the number of sites visited per trip, and a 20% decrease in time spent fishing at the bocaccio sites.

4. Discussion

The three datasets are similar in terms of CPUE trends, critical value analyses and species selection. The species coefficients for the regressions are satisfying from a biological perspective, with regard to both magnitude and direction of influence. In particular, presence of chilipepper (*S. goodei*) is consistently a strong positive predictor of bocaccio, and the two

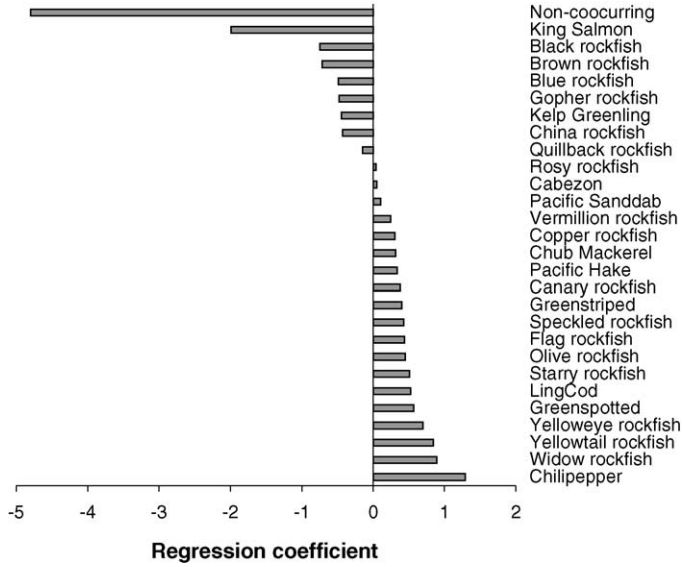

Fig. 11. Estimates of species-specific regression coefficients based on the MRFSS data set.

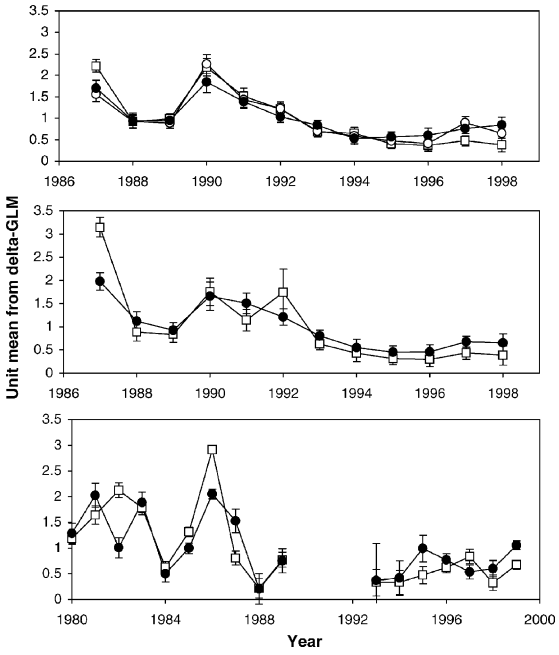

Fig. 12. Time-series of CPUE from analyses of CDF&G site-visit data (upper panel), CDF&G aggregate data (middle panel), and MFRSS data (lower panel). The CPUE indices based on all records are indicated by open squares, those from records selected using location criteria by open circles and those selected by species regression by closed circles. The errors bars indicate one standard error.

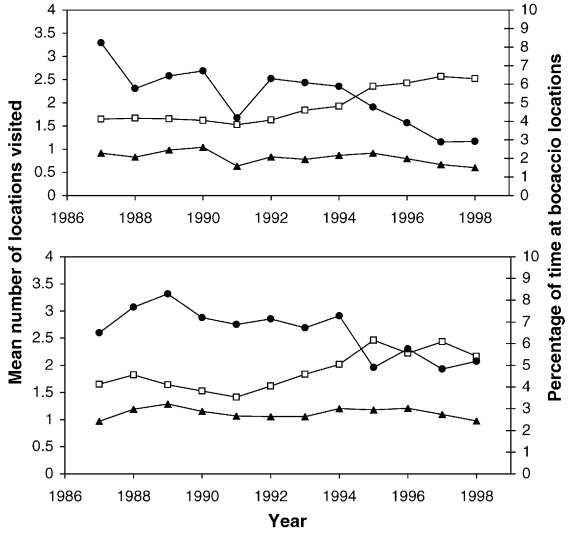

Fig. 13. Mean number of locations visited per trip (squares), mean number of visits to top bocaccio sites per trip (triangles), and percentage of time spent at top bocaccio locations (circles). Results are shown in the upper panel for the full 'CDF&G site-visit' data set and in the lower panel for the same data set after subsetting.

species are well known to co-occur in fishery landings (Williams and Ralston, 2002). In fact, they were treated as a single species in some assessments until fairly recently (Ralston et al., 1998). Presence of black rockfish (*S. melanops*), a species with a more northerly range than bocaccio (Williams and Ralston, 2002), is a negative predictor in all three datasets.

The tradeoff when selecting the critical value is between choosing more data (data quantity), which increases precision, and including less-relevant data (data quality), which decreases both precision and accuracy; these two aspects are assumed to be approximately equal in the vicinity of the proposed critical value. If issues of data quantity and quality are not of equal concern, a different critical value could be considered. The critical value and probability analyses all show that precise cutoff values can be identified to distinguish data subsets (Figs. 7, 8 and 10). Further, the probability distributions themselves identify characteristics of the data sets, such as the increased probability of bocaccio in the aggregate trip data (Fig. 8), and the predominance of low-probability trips associated with inclusion of more non-targeted fishing activity in the data collected by the MRFSS survey (Fig. 10).

We chose to restrict the subsetting analysis to categorical presence and absence data. Abundance of the explanatory species (i.e. their CPUE) could be used as explanatory variables in a similar approach. We prefer use of presence and absence data, because they should be less influenced by trends in abundance of other species. Using a rather large number of explanatory species also contributes to minimizing the effect of abundance trends.

The CPUE indices based on the 'subsetted' data sets (closed circles in Fig. 12) follow the abundance trend from the bocaccio stock assessment (Fig. 1) better than those based on the full data sets (squares in Fig. 12). The unique location-based data in the CDF&G dataset allows us to examine some of the sources of bias that can appear in aggregate trip data. Target switching (Fig. 13) would not have been directly visible without the location-based data, and its effect cannot be removed entirely from species-subsetted data. Our approach to subsetting trips has removed some, but not all, of the confounding influence of target switching.

Logistic regression of target species occurrence on presence and absence of other species provides a practical method for subsetting recreational fishing catch and effort data, and could be applied to many other types of multispecies abundance data where there is a mixture of relevant and non-relevant records (see, for example, Guisan et al. (2002), for a discussion of a similar application in terrestrial settings). This method is especially valuable in that it is reproducible by independent analysts. It also reduces the need for ad hoc decisions in stock assessments, and should contribute to improved consistency among such assessments. Subsetting the data using a species-based logistic regression also removes, or at least reduces, a common criticism about use of recreational CPUE data: that target switching can result in spurious trends in the abundance index.

Acknowledgements

The authors gratefully acknowledge Deb Wilson-Vandenberg (CDF&G) for providing data on northern California partyboat sampling. We also thank Wade VanBuskirk (Pacific States Marine Fisheries Commission) for providing the detailed RecFIN data for this project. An early version of this work was begun with Teresa Ish. E.J. Dick consulted generously on delta-GLM analyses. We thank Nick Davies and an anonymous reviewer for their helpful comments. We especially thank André Punt and Marc Mangel for numerous discussions and suggestions. Funding for this research was provided by the Centre for Stock Assessment Research.

References

Belsley, D.A., Kuh, E., Welsch, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley, New York.

Dick, E.J., 2004. Beyond 'lognormal vs. gamma': discrimination among error distributions for generalized linear models. Fish. Res. 70, 347–362.

Guisan, A., Edwards Jr., T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Gulland, J.A., 1983. Fish Stock Assessment: A Manual of Basic Methods. Wiley, New York.

Ihaka, R., Gentleman, R., 1996. R: a language for data analysis and graphics. J. Comput. Graph. Statist. 5, 299–314.

Love, M.S., Yokalovich, M., Thorsteinson, L., 2002. The Rockfishes of the Northeast Pacific. University of California Press, Berkeley.

MacCall, A.D., 2003. Status of bocaccio off California in 2003. In: Status of the Pacific Coast Groundfish Fishery through 2003 Stock Assessment and Fishery Evaluation, vol. 1. Pacific Fishery Management Council, Portland, OR.

Maunder, M.N., Punt, A.E., 2004. Standardizing catch and effort data: a review of recent approaches. Fish. Res.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, New York.

Osborn, M.F., Van Voorhees, D.A., Gray, G., Salz, R., Pritchard, E., Holliday, M.C., 1996. Marine Recreational Fishery Statistics Survey, National Marine Fisheries Service, NOAA, U.S. Dept. of Commerce. http://www.psmfc.org/recfin/data.htm.

Press, S.J., Wilson, S., 1978. Choosing between logistic regression and discriminant analysis. J. Am. Statist. Assoc. 73, 699–705.

Ralston, S., Pearson, D., Reynolds, J., 1998. Status of the Chilipepper Rockfish Stock in 1998. In: Pacific Fishery Management Council, Appendix: Status of the Pacific Coast Groundfish Fishery through 1998 and Recommended Acceptable Biological Catches for 1999: Stock Assessment and Fishery Evaluation. Pacific Fishery Management Council, Portland, OR.

Ricker, W.E., 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can. 191.

Stefánsson, G., 1996. Analysis of groundfish survey abundance data: combining the GLM and delta approaches. ICES J. Mar. Sci. 53, 577–588.

VanBuskirk, W. (Ed.), 2003. RecFIN Database. National Marine Fisheries Service, NOAA, U.S. Dept. of Commerce, http://www.psmfc.org/recfin/data.htm.

Williams, E.H., Ralston, S., 2002. Distribution and co-occurrence of rockfishes (family: Sebastidae) over trawlable shelf and slope habitats of California and southern Oregon. Fish. Bull. US 100, 836–855.