Fisheries Research 159 (2014) 34–44

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

A nonparametric model of empirical length distributions to inform stratification of fishing effort for integrated assessments

Steven Candy*, Philippe Ziegler, Dirk Welsford

Department of Sustainability, Environment, Water, Population and Community, Australian Antarctic Division, 203 Channel Highway, Kingston 7050, Tasmania, Australia

ARTICLE INFO

Article history:
Received 12 July 2013
Received in revised form 6 May 2014
Accepted 7 May 2014
Handling Editor A.E. Punt
Available online 12 June 2014

Keywords:
Length frequency data
Generalised Additive Mixed Models
Fishery stratification

ABSTRACT

Length frequency data (LFD) are an important input to integrated stock assessments, and statistical tests for variables that significantly influence the length distribution of fish can assist in the definition of effort strata, typically denoted as fisheries or sub-fisheries, in order to account for important systematic differences due to availability and/or gear-specific selectivity of size classes. Here, a nonparametric model of the probability density function of lengths is described which, instead of fitting to LFD directly, is fitted to the set of length quantiles for a pre-determined set of corresponding probabilities *p* (in this instance 0.05, 0.1–0.9 in 0.1 increments, and 0.95). These length quantile data (LQD) can be constructed with individual hauls as sampling units or after pooling hauls to sampling units defined by combinations of covariates such as gear type, spatial block, depth strata, or the sex of sampled fish. The length quantiles are modelled as a Gaussian response variable using a Generalised Additive Mixed Model (GAMM) with smoothing splines fitted for each combination of the covariates (i.e. gear type, depth strata and sex). Graphical presentation of the fitted splines along with standard error of difference bounds were used to investigate where differences were significant in order to assist in the optimal definition of sub-fisheries. The model has the advantage of greater generality and sensitivity in detecting differences compared to modelling a single quantile such as the median. In addition, fitting splines allows flexible and parsimonious modelling of length distributions of any shape. The model is demonstrated using LQD from commercial fishing for Patagonian toothfish at Heard Island.

1. Introduction

Integrated fish stock assessments using software such as MULTIFAN-CL (Fournier et al., 1998), CASAL (Bull et al., 2005), and Stock Synthesis (Methot and Wetzel, 2013) attempt to estimate parameters of a fish population dynamics model by minimising an objective function for observations and predictions. These softwares account for the attributes of commercial fishing which both removes significant numbers of fish from the population and provides an opportunistic sampling process for generating observations of stock attributes. Random sub-samples of the "sample" of fish captured in commercial hauls measured for length and aggregated to seasonal or yearly totals by effort strata and length bins, are typically denoted length frequency data (LFD). LFD are one of a number of different data types fitted in integrated assessments.

To account for the interaction between the ability of a gear type to capture and retain fish of certain sizes and the availability of fish size classes to the fishing gear, integrated assessments typically apply selectivity functions specific to each 'fishery' or "sub-fishery" (e.g. Hillary et al., 2006; Candy and Constable, 2008). For example, to account for differing availabilities of particular length classes across the overall fishery, sub-fisheries can be defined based on spatially defined units such as particular fishing grounds or seafloor depth.

Toothfish (*Dissostichus* spp.) fisheries in the Southern Ocean have features that result in strong heterogeneity in the length frequency data (LFD). These include biological factors, such as the tendency for toothfish to move deeper as they grow, and sexual dimorphism in growth and movement rates (Collins et al., 2010; Phillips et al., 2005; Welsford et al., 2011). This availability of size classes with depth interacts with where and how the sub-fisheries target the toothfish stocks. At South Georgia (CCAMLR subarea 48.3) and Heard Island and McDonald Islands (CCAMLR Division 58.5.2) small juvenile Patagonian toothfish (*Dissostichus eleginoides*) have been targeted by trawling in waters of less than 1000 m depth, while larger fish are caught by longlines in deeper waters up to

* Corresponding author. Tel.: +61 3 62323135.
E-mail address: steve.candy@aad.gov.au (S. Candy).

http://dx.doi.org/10.1016/j.fishres.2014.05.002

2500 m. Within each sub-fishery, toothfish LFD are typically unimodal (Hillary et al., 2006; Candy and Constable, 2008).

In stock assessments, sub-fisheries are pre-defined and used to represent a stratification of fishing effort. For each sub-fishery, available LFD are aggregated across hauls for each fishing year and fitted using a multinomial likelihood. Previous work using LFD has concentrated mostly on defining and then estimating an appropriate "effective sample size" for such LFD at the sub-fishery by year level so that the uncertainty in such data due to sampling and model error is adequately accounted for by a nominal multinomial likelihood (Candy, 2008; Francis, 2011; Maunder, 2011).

In contrast, very little research has been carried out on objective methods to pre-define sub-fisheries using LFD. Typically, no formal statistical basis for defining sub-fisheries is used. For example, Hampton and Fournier (2001) simply define 16 sub-fisheries as combinations of 7 regions and 6 gear types for use in an integrated assessment of yellowfin tuna (*Thunnus albacares*) for the western and central Pacific. Notable exceptions are Phillips et al. (2005) who stratified effort for the Ross Sea Antarctic toothfish (*Dissostichus mawsoni*) stock assessment with tree-based regression methods described in Breiman et al. (1984). This was done for a single quantile (i.e. the median) of the haul-level length sample including predictors of small spatial scale research units and depth. Lennert-Cody et al. (2010) generalised this approach by modelling LFD using a multivariate regression tree approach. They applied their method to yellowfin tuna caught in the purse-seine fishery of the eastern Pacific Ocean, with sample units defined by the aggregate of hauls within 5° latitudinal by 5° longitudinal spatial units combined with temporal "quarters" where each were defined as continuous variables.

Similarly to Phillips et al. (2005) and Lennert-Cody et al. (2010), this study has as main objective the description of a statistical approach for pre-stratification of effort and thus the definition of sub-fisheries using length data, where effort is represented by individual hauls or groupings of hauls. Our method of pre-stratification is similar to that of Lennert-Cody et al. (2010) in that it does not assume an underlying parametric functional form (e.g. lognormal) for the probability density function (PDF) of lengths in the sampled population and therefore can be described as nonparametric. However, in contrast to their classification-based approach of data-driven splits on continuous predictor variables, our approach uses the classical hypothesis testing approach of pre-specifying independent factors that in combination split the samples a priori and test the statistical significance of these factors including any interactions.

The method described here involved fitting a model to a set of length quantiles, *L*, corresponding to a pre-determined set of corresponding probabilities $\boldsymbol{p}$ (in this instance 0.05, 0.1 to 0.9 in 0.1 increments, and 0.95). These length quantile data (LQD) are constructed by either treating individual hauls as the sampling units or by pooling hauls to sampling units defined by combinations of discrete-valued covariates such as gear type, spatial block, depth strata, and sex. The length quantiles are modelled as a Gaussian response variable using a Generalised Additive Mixed Model (GAMM) (Wood, 2006) with smoothing splines fitted for each combination of the covariates (i.e. gear type, depth strata, and sex). For comparison of model fit Akaike's Information Criterion (AIC) (Akaike, 1973) is used, while for a given model comparisons between fitted splines for each factor combination are carried out graphically using approximate 95% confidence bounds for differences between factor levels for given values of probability. The model and methods of model testing are demonstrated using length data obtained for trawl, longline, and trap gear types from commercial fishing for Patagonian toothfish at Heard and McDonald Islands.

2. Materials and methods

2.1. Statistical approach

The approach uses Generalised Additive Mixed Models (GAMMs) to fit smoothing splines to LQD for a fixed set of associated probabilities (Appendix A) to parsimoniously represent the underlying family of cumulative probability density functions (CDFs) for length. To do this we applied the `gamm` function in the R-software (R Core Team, 2013) library `mgcv` (see Wood, 2006).

The expected value of the response variable of length quantile *L* conditional on pre-specified probability levels, $\boldsymbol{p}$, and combinations of discrete-levelled factors such as depth strata, gear type, and sex, expressed as a composite factor *C*, is given by:

$$E(L_{kji}|C_j, p_i) = \beta_0 + \sum_{j,j'=2}^{J} \beta_{j-1} I_{jj'} + s(C_j : p_i, \kappa) \tag{1a}$$

where the subscript *i* indexes the specific quantiles chosen, for example for $\boldsymbol{p} = vec(0.05, 0.1, 0.2, \ldots, 0.9, 0.95)$; there are *J* levels of factor *C* with parameter corresponding to its first level aliased with β_0; $I_{jj'}$ is the element of the indicator matrix that takes the value 1 if $j \equiv j'$, corresponding to C_j, and zero otherwise; the subscript *k* indexes spatial blocks which are crossed (i.e. possibly only partially crossed) with level of *C*, in order to define individual LQD sample units (SUs) (note that each SU could correspond to individual hauls or be defined as all hauls pooled within each combination of spatial block and factor level of *C*); the spline terms are denoted by $s(C_j: p_i, \kappa)$; and the term $C_j: p_i$ represent the model with separate splines for each level of *C*. The parameter κ is the dimension of the basis used to represent the smooth term (Wood, 2006) (see also `choose.k` in `gam{mgcv}`help for the `mgcv` package).

The `mgcv` library allows only a single factor to be used within the spline term in this way. This requires that the composite factor *C* must be constructed to give unique levels for each combination of the factors of interest for estimation, for example for each combination of the levels of sex and depth class. In order to investigate the contribution of each factor and determine the minimal significant model in terms of their main effects and interactions, *C* must be constructed for each version of the model. Model selection should proceed by backward selection of first the interactions and then the main effects in a way that preserves marginality constraints (Nelder and Lane, 1995). In addition, to investigate the contribution of each of these factors in a model, the predictions for each factor level that make up the composite factor *C* are obtained by appropriate cross-referencing (see R-code "calculate SE" in Supplementary Material).

The random component of the model is given by:

$$L_{kji} = E(L_{kji}|C_j, p_i) + \nu_k + \tau_{kj} + \varepsilon_{kji} \tag{1b}$$

where

$$\varepsilon_{kji} \sim N(0, \sigma^2/N_{kj}), \quad \mathrm{cor}(\varepsilon_{kji}, \varepsilon_{kji'}) = \varphi^{|p_i - p_{i'}|}$$

$$\nu_k \sim \mathrm{NID}(0, \sigma_\nu^2), \quad \tau_{kj} \sim \mathrm{NID}(0, \sigma_\tau^2)$$

where ν_k is the *k*th element of the set of random effects representing spatial blocks and are considered normally and independently distributed (NID), the τ_{kj} represents a set of random effects for the SUs with each SU consisting of a sample of N_{kj} lengths, the ε_{kji} represent the combined within-SU model error and measurement error (Appendix B) and are approximated by a first-order continuous "time" autoregressive process (CAR). For example for each spatial block, a sample of females considered separately to a sample of males taken from the same block, $k = k'$, would have two levels $j = 1, 2$ within k'. The case of depth classes is more complicated because they also depend on the size of the blocks and