Full length article

Accurate aging of juvenile salmonids using fork lengths

Suresh Andrew Sethi [a,*], Jonathon Gerken [b], Joshua Ashline [b]

[a] US Geological Survey, New York Cooperative Fish and Wildlife Research Unit, 211 Fernow Hall, Cornell University, Ithaca, NY 14853, USA
[b] US Fish and Wildlife Service, Anchorage Field Office, 4700 BLM Road, Anchorage, AK 99507, USA

ARTICLE INFO

ABSTRACT

Juvenile salmon life history strategies, survival, and habitat interactions may vary by age cohort. However, aging individual juvenile fish using scale reading is time consuming and can be error prone. Fork length data are routinely measured while sampling juvenile salmonids. We explore the performance of aging juvenile fish based solely on fork length data, using finite Gaussian mixture models to describe multimodal size distributions and estimate optimal age-discriminating length thresholds. Fork length-based ages are compared against a validation set of juvenile coho salmon, *Oncorynchus kisutch*, aged by scales. Results for juvenile coho salmon indicate greater than 95% accuracy can be achieved by aging fish using length thresholds estimated from mixture models. Highest accuracy is achieved when aged fish are compared to length thresholds generated from samples from the same drainage, time of year, and habitat type (lentic versus lotic), although relatively high aging accuracy can still be achieved when thresholds are extrapolated to fish from populations in different years or drainages. Fork length-based aging thresholds are applicable for taxa for which multiple age cohorts coexist sympatrically. Where applicable, the method of aging individual fish is relatively quick to implement and can avoid ager interpretation bias common in scale-based aging.

Published by Elsevier B.V.

1. Introduction

Several Pacific salmon species (*Oncorynchus* spp.) exhibit diversity in pre-smolt freshwater rearing residency times, leading to overlap of young of year and older juvenile cohorts within freshwater habitats. Juvenile salmon life history strategies, survival, and habitat interactions may vary by age cohort (e.g. Bisson et al., 1988; Dolloff and Reeves, 1990; Keith et al., 1998; Myrvold and Kennedy, 2015), such that age-specific studies may provide more detailed insight into juvenile salmon ecology than generic stage-specific studies on mixed age cohorts. One obvious complication in such efforts is a need to age individual juvenile salmon captured as part of field studies into annual cohorts. For example, age-specific survival modeling based upon capture histories from tagged fish (Carlson et al., 2008), age-specific habitat use studies (Bradley et al., In press), and ontogenic diet studies (Richardson et al., 2016) require ages for individual fish.

Aging of juvenile salmon to an annual cohort is possible using scale reading analysis (e.g., Gilbert, 1912; Quist et al., 2012). However, the process of reading scales can be error prone and

time consuming (e.g., Campana, 2001). Furthermore, scale collection may negatively impact handled fish by increasing handling stress responses and interfering with slime and scale protection against disease (e.g., Gadomski et al., 1994), potentially influencing outcomes from studies that require repeated observations on individuals (e.g., age-specific survival estimation via mark recapture analysis). In contrast, collection of fork length measurements may be less invasive. Owing to the discrete timing of salmon spawning events, young of year and older age cohorts of juvenile salmon typically exhibit a strong size differential. An alternative method of determining ages can be implemented by characterizing the size frequency of a population of juvenile salmon of interest and then separating out age cohorts by size thresholds.

Aging by size has been implemented in a number of juvenile salmonid ecology studies. However, determination of ages is often conducted by subjective classification of animals into cohorts after visually inspecting a size frequency histogram (e.g., Keith et al., 1998; Myrvold and Kennedy, 2015). For cases with clearly separated size distributions between age cohorts, visual assessment of age-separating length thresholds can be straightforward (e.g., Fig. 1a). However, in many cases, sparse data or overlapping size distributions among age cohorts can make such visual assessments difficult (Fig. 1b; Crone and Bond, 1974). An alternative, objective approach, to aging fish based upon fork length measurements is
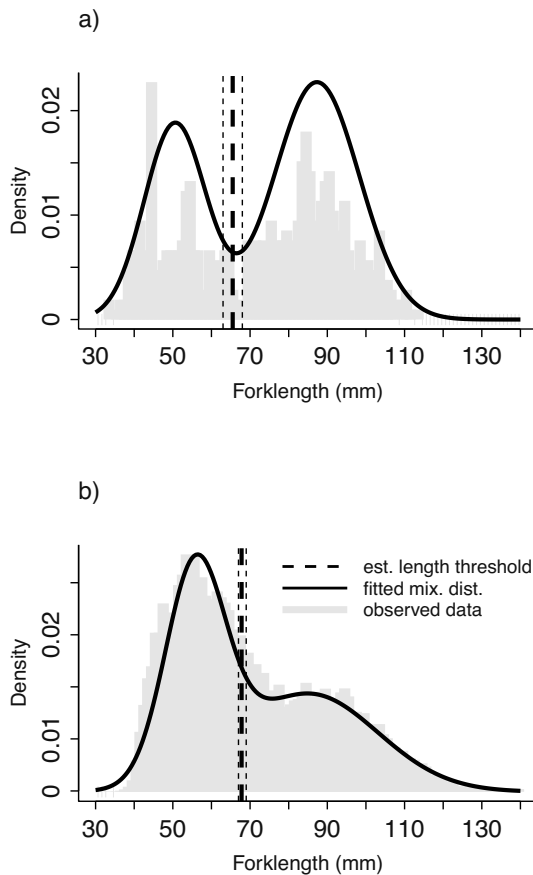
Fig. 1. Length frequency distributions and fitted two-component Gaussian finite mixture models from samples of juvenile coho salmon (*Oncorhynchus kisutch*) from the Big Lake watershed, Alaska, 2011–2015. Vertical black dashed line are the optimal age-separating length threshold based upon fitted mixture distributions and a criteria to minimize classification error with 95% confidence limits. Panel a: samples from Meadow Creek drainage lotic habitats in 2015 over the biweek period July 1 to July 15 (*n* = 523). Panel b: all samples combined over 2011–2015 in the Big Lake study region (*n* = 32,854).

to use Gaussian *K*-component finite mixture models (McLachlan and Peel, 2000) to describe the size distribution of mixed-age samples and subsequently determine an optimal age-separating size threshold based upon a misclassification error minimization criteria.

Application of finite mixture models – which describe a multimodal distribution shape as a weighted combination of component distributions – have been applied to size frequency distributions of fish species, with most applications focused on assessing the growth of marine species by tracking size modes across age cohorts, and over time if repeated samples are available (e.g., Fournier et al., 1990; Leigh and Hearn, 2000; Laslett et al., 2004; MacDonald and Pitcher, 1979; Van Beveren et al., 2014). In these applications, termed "modal analysis," finite mixture distributions are first fitted to length frequency data. Subsequently, mixture model output on mean size at age, i.e., the means of the underlying component distributions making up the mixture distribution, are then fit to a growth equation. In addition to information on growth, finite mixture models have also been used to estimate the proportions of different age cohorts in a sample of fish (e.g., Zhu et al., 2013), whereby age proportions are parameters directly estimated during finite mixture model fitting (i.e., underlying component distribution weights, or *τ* parameters from Eq. (1) below).

Here, we implement finite mixture models on length frequency information to address a separate, but related, problem, namely of aging individual juvenile fish based solely on their lengths. While finite mixture models can predict membership in a specific underlying component distribution – i.e. age cohort – for data in the sample used to fit a model, we develop a bootstrapping routine to utilize fitted mixture models to estimate optimal age-separating fork length thresholds for application to out-of-sample data. Once estimated, age-discriminating length thresholds provide a rapid means of objectively aging fish based upon length measurements for populations exhibiting sizes from two or more age cohorts. We assess the accuracy of aging based upon fork length distributions against traditional scale-based aging using data on juvenile coho salmon, *O. kisutch*, from the Big Lake watershed in southcentral, Alaska. In addition we examine whether age-separating thresholds are useful when applied across populations, and examine whether age-separating length thresholds remain accurate when applied to fish sampled in different years or times of year. Results indicate that for juvenile fish taxa exhibiting length structure across multiple sympatric cohorts such as coho salmon, aging based upon fork length data can be achieved with small to moderate sized samples of fork lengths on the order of 100 fish or greater, achieving aging accuracies of 95% or better as compared to validation samples assessed by multiple scale readers.

2. Methods

2.1. Age-discriminating fork length threshold estimation with mixture models

Fork length data samples were separated into component age cohorts using univariate *K*-component Gaussian finite mixture modeling implemented using the *mclust* package in the R statistical programming environment (Fraley and Raftery, 2012; R Core Team, 2015). *K*-component Gaussian finite mixture models represent multimodal distributions with an additive mixture of Normal component distributions. In the context of juvenile fish aging, this amounts to separating out different component length distributions specific to age cohorts from samples on a population made up of mixed ages (e.g., Fig. 1). *mclust* implements maximum likelihood methods to fit mixture models using the Expectation Maximization algorithm (McLachlan and Peel, 2000), where the likelihood is expressed as:

$$L(\boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\sigma}|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \tau_k f(y_i|\mu_k, \sigma_k) \qquad (1)$$

where $\mathbf{y} = y_1, \ldots, y_n$ is a vector of fork lengths, $\boldsymbol{\tau} = \tau_1, \ldots, \tau_k$ represents the probabilities that a randomly drawn data point comes from a particular component distribution with the constraints that $\tau_k \in [0, 1]$ and $\sum_{k=1}^{K} \tau_k = 1$ for $k = 1, \ldots, K$, and $f(y_i|\mu_k, \sigma_k)$ represents the probability of the observed data given the parameters specifying a Normal distribution for the *k*th component:

$$f(y_i|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2}\right) \qquad (2)$$

The *τ* parameters are interpreted as the proportion with which the *k*th component is present in the population (e.g., the proportion of young of year fish in the sampled population). The likelihood in Eq. (1) is maximized in choice of the proportions with which the underlying component distributions are present in the population, and their means and variances, given *K*, the total number of component distributions present, and given a specification for the covariance structure across the component distributions. With univariate data, the latter problem simplifies: either component variances are equal or they vary across components. The number of underlying component distributions and the covariance structure across component distributions are specified in advance and are