Simulation testing a new multi-stage process to measure the effect of increased sampling effort on effective sample size for age and length data

James T. Thorson[1], Meaghan Bryan[2], Pete Hulson[3], Haikun Xu[4], André E. Punt[5]

[1] Habitat and Ecological Processes Research Program, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA

[2] Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA

[3] Auke Bay Laboratories, Alaska Fisheries Science Center, NMFS, NOAA, Seattle, WA, USA

[4] Inter-American Tropical Tuna Commission, La Jolla, CA, USA

[5] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA, USA

Keywords: age composition data; stock assessment; data-weighting; Dirichlet-multinomial

Abstract

Ocean management involves monitoring data that are used in biological models, where estimates inform policy choices. However, few science organizations publish results from a recurring, quantitative process to optimize effort spent measuring fish age. We propose that science organizations could predict the likely consequences of changing age-reading effort using four independent and species-specific analyses. Specifically it predicts the impact of changing age collections on the variance of expanded age-composition data ("input sample size," Analysis #1), likely changes in the variance of residuals relative to stock-assessment age-composition estimates ("effective sample size," Analysis #2), subsequent changes in the variance of stock-status estimates (Analysis #3), and likely impacts on management performance (Analysis #4). We propose a bootstrap estimator to conduct Analysis #1, and derive a novel analytic estimator for Analysis #2 when age-composition data are weighted using a Dirichlet-multinomial likelihood. We then provide two simulation studies to evaluate these proposed estimators, and show that the bootstrap estimator for Analysis #1 underestimates the likely benefit of increased age reads while the analytic estimator for Analysis #2 is unbiased given a plausible mechanism for model mis-specification. We conclude by proposing a formal process to evaluate changes in survey efforts for stock assessment.

Introduction

Fisheries managers in the United States, Europe, and worldwide use stock assessment models to inform fishery regulations (Methot, 2009). Age and length data are provided to stock assessments through either fishery-independent (surveys) or fishery-dependent sources (fisheries). In modern integrated assessment models (Maunder, 2003), these age and length data inform estimates of abundance-at-age in conjunction with abundance-index, fishery catch, and other sources of information. However, collecting and interpreting information about population length- and age-structure for use in fisheries management is a complex enterprise and requires collaboration from teams that (at a minimum) conduct field samples, subsample these to record age and length, expand resulting records to account for subsampling designs, weight these data in conjunction with other data sources, and use resulting estimates to inform management changes. Data on age and length are obtained with the expectation that larger sample sizes will lead to lower variances for estimates of stock status from stock assessment models. This scientific enterprise is expensive to develop and maintain, and there are ongoing demands to quantify the amount of effort expended at each stage and for different species.

Fisheries scientists and managers must balance many competing demands for limited scientific resources which consequently determine the magnitude of age and length sampling. Surveys of fish populations often target and record information for multiple species and follow a pre-established sampling design from year to year. This design typically involves a process to select locations and times to collect samples over a fixed spatial domain, and also some process to subsample fishes in each sample for which biological data (age, length, weight) are measured. The design for sampling locations is often fixed over time; although there is more room to change the design for subsampling biological data, these subsampling designs are also often relatively consistent over time. By contrast, the importance of sampling data for each individual fish species often changes dramatically over time, either as species approach biological reference points (e.g., trigger rebuilding plans) or when the relative abundance of species changes in a given region. However, there is relatively little research regarding how to quantify the impact of changing age/length subsampling on the precision of stock assessments, and this limits the ability to objectively optimize age/length subsampling strategies.

We identify two broad approaches to fitting age and length information in stock-assessment models: a multinomial likelihood (or similar methods) for data representing proportion at age or length, e.g., in Stock Synthesis (Methot and Wetzel, 2013); or a multivariate normal likelihood for abundance-indices for each age, e.g., in SAM (Nielsen and Berg, 2014). Several methods have been suggested to account for non-independence of age and length subsamples within a given sampling unit (or haul) when using a multinomial likelihood. These methods can generically be categorized as follows:

1. *ad hoc* methods for setting input sample size, for example, one fish per haul (Pennington *et al.*, 2002);
2. using sampling or model-based estimators to measure the degree of independence in available data (Stewart and Hamel, 2014; Thorson, 2014; Thorson and Haltuch, 2018);
3. iteratively estimating effective sample size of age/length composition based residual fits to an assessment model (McAllister and Ianelli, 1997; Francis, 2011); and
4. estimating effective sample size within an extension of the multinomial distribution that permit such estimation, for example, using the Dirichlet-multinomial likelihood (Thorson *et al.*, 2017).

Alternatively, stock assessments using multivariate normal likelihood for abundance-at-age can quantify the covariance resulting from correlations among observations, similar to #2 above (Berg *et al.*, 2014), and/or the covariance in residual fits to the assessment by using a flexible specification of the likelihood, similar to #4 above (Berg and Nielsen, 2016).

In the following, we distinguish three potential "sample sizes" that define the information available in age and length composition data:

1. *Nominal sample size* (n_{reads}): The number of raw ages or lengths that are measured, either in the field (for lengths) or the laboratory (for otoliths, vertebrae, spines, or other hard parts that are informative about age).
2. *Input sample size* (n_{input}): The specified sample size for expanded age or length compositions, when fitting these as if they follow a multinomial distribution.
3. *Effective sample size* ($n_{effective}$): The sample size that results from reweighting the age or length composition data based on their fit within a stock assessment model.

These sample sizes are often measured explicitly when using a multinomial likelihood for composition data (Methot and Wetzel, 2013), and they could be approximated when using a multivariate normal likelihood for abundance-indices-at-age in other assessment models (Berg *et al.*, 2014).

Nominal, input, and effective sample size will often differ for a given stock. For example, the U.S. West Coast bottom trawl survey for canary rockfish measured lengths for 562 individuals in 43 survey tows each year on average from 2003-2010 (i.e., $n_{reads} \approx 562$), but a sample-based calculation yields a much lower input sample size ($n_{input} = 79$) as reported by Stewart and Hamel (2014). The most recent full assessment for this stock (Thorson and Wetzel, 2015) then used the Francis (2011) method for calculating effective sample size, which resulted in a yet lower sample size ($n_{effective} = 9.1$), although the alternative Ianelli-McAllister reweighting would have resulted in a small reduction from input to effective sample size ($n_{effective} = 75.4$). It is widely accepted that weighting age/length composition data by nominal sample size within a multinomial likelihood is problematic (Hulson *et al.*, 2011; Francis, 2017), although research is ongoing regarding optimal methods to calculate effective sample size (e.g., Xu *et al.*, 2019b).

Despite the importance of age and length data within statistical catch-at-age models, there has been little research regarding how stock-assessment results change with changes in age and length sampling effort. One exception is Zimmermann and Enberg (2017), which explored the impact of reduced survey frequency (including age and length subsamples) on assessment results for two Northeast Atlantic stocks. This study showed that reduced sampling frequency increased confidence interval widths, but did not specifically explore changes in age-reading effort or provide any theory for predicting changes in assessment results given changes in age subsampling. This dearth of research is surprising given that age-structured assessments are the gold-standard for stock assessment, are widely used worldwide, and remain resource-intensive due to the need to collect and read age data.

The lack of published research regarding changes in assessment model results given changes in age subsampling effort likely arises due to the complexity of the topic. We hypothesize that an increase in the number of subsampled ageing structures will result in a less-than-proportional increase in the input-sample sizes that represent the variances of the resulting measurement of age-composition, e.g., see argument in Lai (1987). The function relating age reads and input sample size is likely nonlinear because it depends not just on the number of age reads (the input sample size), but also upon the correlation between ages read in a single tow, the evenness of catch-rates between tows, and other factors (Stewart and Hamel, 2014). Similarly, the impact of increasing input-sample size on stock assessment results is likely to be nonlinear because age-composition data are typically "down-weighted" to account for model mis-specification and the non-independence among samples (Francis, 2011), and previous research has not explored how data weighting might affect the benefits of increased age subsampling effort. The net effect of these two nonlinear functions is difficult to predict; it presumably could be predicted by simulating or randomly re-sampling field-sampling data, and then exploring likely consequences on subsequent stock-assessment modelling. However, such simulation experiments are difficult to condition upon the specifics of field-sampling data for a given stock (i.e., the degree of variance in age-samplings among or within tows) or its subsequent stock assessment (i.e., the degree of down-weighting).

In this study, we therefore propose and evaluate the performance of two new estimators, predicting the likely increase in effective sample size resulting from an increase in age-reading effort. To do so, we first decompose the problem into four separate analyses, and propose new estimators for the first two that, respectively, predict resulting changes in input and effective sampling size. We then use two simulation experiments to test the accuracy of these two proposed estimators; these experiments show that we can predict a lower bound for the impact of changing age sampling size on input sample size, and can accurately predict the impact of changing input sample size on effective sample size. We conclude by discussing how these results can be used in future studies to optimize the allocation of limited age-reading effort across multiple species within a multispecies sampling design.

Methods

We develop an approach to predict how management performance would change when re-allocating resources involved with collecting age and length subsampling data. Specifically, we show that this problem can be decomposed into four separate analyses to approximate the pipeline from field-sampling to data preparation to management performance. When using a multinomial likelihood for age and length proportions, these analyses are as follows:

Analysis #1. Predict impact of changing nominal age/length sample size on the input sample size of expanded age/length composition data;

Analysis #2. Predict impact of changing input sample size for age/length composition data on the effective sample size when these data are weighted in an assessment model to account for model mis-specification; and

Analysis #3. Predict impact of changing effective sample size on the variance of stock-assessment outputs.

Analysis #4. Predict impact of changing variance for stock-assessment outputs on management performance.

Notation and further description of these four analyses are presented in Table 1. We address Analysis 1-2 in this study and recommend further research regarding Analysis 3; we note that Analysis 4 has been thoroughly studied elsewhere by a series of papers on this topic (Shertzer and Prager, 2007; Wiedenmann *et al.*, 2015; Punt *et al.*, 2016). For Analysis 1-2, we explain the theory allowing us to identify the sensitivity to changing nominal and input sample size, and also validate this theory using a simple simulation experiment. These estimators are designed for a stock assessment using a multinomial likelihood. We suspect that Analysis #1 could be adapted to measure changes in the estimation covariance in abundance-indices at age (e.g., Berg *et al.*, 2014) arising from changing sampling effort, while Analysis #2 could similarly be adapted to predict likely changes in covariance in residuals when fitting indices-at-age within an assessment model (Berg and Nielsen, 2016). However, we do not discuss further in this paper how to adapt analyses for use in a multivariate normal approach.

Analysis #1: Predicting changes in input sample size

We first conduct a replicated simulation experiment and apply a new bootstrap estimator for predicting how input sample size will be affected by a change in the nominal sample size for age/length subsamples (see Fig. 1 for visualization of simulation design). To do so, we measure input sample sizes using the Stewart and Hamel (2014) method; see Appendix A for a detailed explanation. Future analyses could use alternative method for calculating input sample size (Thorson, 2014; Thorson and Haltuch, 2018), although we do not explore the topic further here. In the following we apply the Stewart-Hamel method to data collected using a simple random sampling design for ages in each tow. The algorithm could be applied to length-stratified age samples, the use of an age-length key, or other complicated circumstances via proper bootstrapping and expansion of samples, although we do not explore these topics further here.

We apply the Stewart-Hamel method for calculating input sample size to data generated by a simple operating model. We define parameters for this operating model by fitting a multivariate spatio-temporal model to abundance-at-age for walleye pollock in the eastern Bering Sea, for each of ages 1-8 (where “age 8” is as a plus-group for all ages 8+) from ten years of data, 2009-2018. This multivariate model does not explicitly estimate any covariance in spatial or spatio-temporal components between ages, although predicted densities will be correlated among ages for some data sets; this is the default “index standardization” settings for package *VAST* (Thorson, 2019a), as demonstrated by Thorson and Haltuch (2018).

For each replicate of this simulation experiment we specifically apply the following steps:

Step 1. We simulate new data conditional on the estimated fixed and random effects from the operating model. This involves sampling the total abundance for each of approximately 360 bottom trawl samples conducted in each year. We also simulate samples of age for a random subsample of individual fish from each tow, where the size of the subsample is 20 individual fish per tow. We apply the Stewart-Hamel method to determine the $n_{input}(t)$ for each year and record these values.

Step 2. Given the simulated data set from Step 1, we then apply a bootstrap estimator to predict the sample size that would have occurred if we had instead subsampled 40 individual fish per tow. This estimator is described below, and result in prediction $n_{input}^{*}(t)$ for the input sample size that would result from increasing subsampling effort in each year and simulation replicate.

Step 3. Using the same operating model as in Step 1, we then take a subsample of 40 individual fish per tow, apply the Stewart-Hamel estimator, and record this as the true sample size $\tilde{n}_{input}(t)$ resulting from increased sampling effort.

Step 4. We replicate steps 1-3 multiple times, and compare $\tilde{n}_{input}(t)$ with $n_{input}^{*}(t)$ for each replicate and year.

We apply a "bootstrap" estimator in Step 2. This estimator applies the Stewart-Hamel estimator to a list of age/length records and associated data for each tow. However, when bootstrapping the set of age/length measurements, it samples 40 individual fish (instead of the original sample size of 20) with replacement from the 20 measured fish. On first inspection, an analyst might assume that doubling the number of age/length measurements will result in a doubling of input sample size, so we also compare the bootstrap estimator $n_{input}^{*}(t)$ with this naïve approach.

We hypothesize that doubling the number of ages will result in a less-than-proportional increase (i.e., $\tilde{n}_{input}(t)/n_{input}(t) < 2$). This is because the variance in compositions among samples (which determines the value of $\tilde{n}_{input}(t)$) arises both from imprecision when subsampling ages within each tow (which decreases with increasing age samples), but also from variance in age-composition among tows (which is less sensitive to increasing age samples). The bootstrap estimator is inspired by statistical theory and other analyses using resampling to predict the likely change in variance resulting from changing sample sizes (Anderson and Santana-Garcon, 2015). However, this particular use for bootstrap to predict a likely change in variance has not been simulation-tested before, and hence we have no prediction of its performance *a priori*. Future studies could explore alternative changes in sampling design (i.e., increasing the total number of trawl tows), or changes in the estimation model used to expand age and length-composition data (e.g., Thorson and Haltuch, 2018); we leave both as topics for future research.

Analysis #2: Predicting changes in effective sample size

We start Analysis #2 with the input-sample size $n_{input}(t)$ for each year t of a given data set, and the new input-sample size $n^*_{input}(t)$ under a proposed change to age/length subsampling effort, where the latter would be obtained using an estimator in Analysis #1. We seek to understand how changes in input-sample size affects the expected effective sample size $n^*_{effective}(t)$ when fitting these data in a stock-assessment model. The following derivation applies generically to a variety of approaches for data-weighting, but we use terminology for the Dirichlet-multinomial approach (Thorson *et al.*, 2017).

The Dirichlet-multinomial approach replaces a multinomial likelihood with an alternative Dirichlet-multinomial likelihood involving an additional parameter θ:

$$n_{effective}(t) = \frac{1}{1+\theta} + n_{input}(t)\frac{\theta}{1+\theta} \quad (1)$$

where θ governs the ratio of input and effective sample size. Estimating θ using the Dirichlet-multinomial likelihood is therefore a model-based approach to data-weighting, and data-weighting is used to ensure that the weight assigned to age-composition data in a joint likelihood is appropriate for the degree of match between data and model. It is typically justified as an important step in a process to account for model mis-specification, arising from neglecting processes (e.g., time-varying growth, survival, and fishery selectivity) that affect predicted proportion-at-age but are not frequently modeled in age-structured assessments (Francis, 2011; Thorson, 2019b).

In this paper, we propose that the estimated value for θ (and the resulting ratio of input and effective sample sizes) can be used to predict effective sample size $n^*_{effective}$ resulting from a new input sample size n^*_{input}. When $n_{input} \gg 1$ using the Dirichlet-multinomial likelihood, we derive the formula:

$$n^*_{effective}(t) = \frac{n^*_{input}(t)(1+\theta n_{input}(t))}{n^*_{input}(t) + \left(1+\theta n_{input}(t)\right)} \quad (2)$$

which is a reduced form of the Michaelis-Menten (a.k.a. Beverton-Holt) function $Y = aX/(b+X)$ where $a = b = (1+\theta n_{input}(t))$ and $X = n^*_{input}(t)$. See Appendix B for a detailed derivation; a similar formula could be derived for the Ianelli-McAllister approach for reweighting composition data, although we do not explore the topic here. This formula has two noteworthy properties (see Fig. 2): (1) $n^*_{effective}(t)$ approaches its maximum value $1+\theta n_{input}(t)$ (known as the "saturation value") as $n^*_{input}(t)$ increases asymptotically,, and (2) $n^*_{effective}(t)$ achieves half of its maximum value when $n^*_{input}(t) = 1+\theta n_{input}(t)$ (known as the "half-saturation value"). Future research could explore the variance in model mis-specification ($\hat{\sigma}^2_{model}$ in

Appendix B) resulting from imprecise estimates of $n_{effective}$, analogous to efforts for abundance indices (Kotwicki and Ono, 2019), although we do not do so here.

We conduct a simulation experiment to explore the performance of this predicted relationship (see Fig. 3 for visualization). This experiment involves an operating model (OM) that simulates age-structured population dynamics given a single fishery, as well as an estimation model that is used to predict changes in effective sample size resulting from an increase in input sample size. The operating model simulates population dynamics over twenty years for a fish population loosely based on red snapper (*Lutjanus campechanus*), tracking abundance-at-age for age-0 through age-20 as it is exploited by a developing fishery. The OM generates simulated data for a fishery-independent survey, fishery catch, as well as fishery age-composition samples with input sample size n_{input}. In particular, the OM simulates a fishery that has age-and-year specific selectivity ($S_{a,t}$):

$$S_{a,t} = S_a \exp(\varepsilon_{a,t}) \tag{3}$$

where S_a represents logistic selectivity-at-age and $\varepsilon_{a,t}$ additional variation in selectivity-at-age which follows an autoregressive process across both years and ages (Xu *et al.*, 2019a). By contrast, the estimation model (EM) assumes that fishery selectivity follows a logistic selectivity-at-age function; see Appendix C for more details. The EM therefore has mis-specified selectivity-at-age, and it also specifies that fishery age-composition data follow the Dirichlet-multinomial likelihood (Thorson *et al.*, 2017; Thorson, 2019b). It therefore responds to mis-specified selectivity by typically estimating that $n_{effective} < n_{input}$, where the degree of downweighting is a measure of model mis-specification. The simulation-experiment is conducted using R package *CCSRA* release number 1.2.0 (Thorson and Cope, 2015). We do not include the effect of ageing errors although we hypothesize that ageing error will have little effect (besides decreasing the precision of resulting stock-assessment estimates) as long as the magnitude of this error is estimated (e.g., Punt *et al.*, 2008) and appropriately included in the assessment model.

We specifically run one hundred replicates for each of three simulation scenarios that differ in the amount of fishery age-composition data that are available in each year, where $n_{input} = \{20,50,100\}$. These three scenarios use the same simulation seed for each replicate and therefore each scenario has the same population and fishery dynamics in a given replicate; scenarios differ only in how many age-composition samples are available for a given replicate. Each replicate therefore specifies an input sample size n_{input} and estimates Dirichlet-multinomial parameter θ; we apply Eq. 2 for each replicate to predict the effective sample size $n^*_{effective}$ that is expected under alternative input sample sizes $n^*_{input} = \{20,50,100\}$. We then compare this prediction with the effective sample size $n_{effective}$ occurring under that alternative n_{input} for a given replicate with identical population and fishery dynamics (which we then label $\tilde{n}_{effective}$). For the scenario when $n_{input} = 20$, we specifically compare predicted $n^*_{effective}$ under a hypothetical increase to $n^*_{input} = 50$ or 100 with $\tilde{n}_{effective}$ occurring when $n_{input} = 50$ or 100. For the scenario when $n_{input} = 50$, we similarly compare predicted $n^*_{effective}$ a hypothetical increase to $n^*_{input} = 100$ with $\tilde{n}_{effective}$ occurring when $n_{input} = 100$. In this way, we can compare predicted $n^*_{effective}$ with $\tilde{n}_{effective}$ given three levels of increasing sample size.

Results

We demonstrate results from a simulation experiment evaluating the performance of each of these two estimators.

Step 1: Expanding subsampled records

As expected, a doubling of age reads results in a less-than-doubling of input-sample size (average increase is 54%, see black histogram in Fig. 4). This less-than-linear increase occurs because some proportion of variance arises due to variability in age-composition among tows, and this portion of variance cannot be reduced by sampling more ages for a fixed set of tows. However, the bootstrap approach underestimates the increase in input sample size (average estimate is 21% increase). Therefore, the true change in input sample size (54% increase) falls between the bootstrap estimator (21% increase) and linear estimator (100% increase). Both the bootstrap approach and the true increase in input sample size have substantial variance among simulation replicates (see width of histograms in Fig. 4).

Step 2: Weighting age/length composition data given model mis-specification

We next use an age-structured population model that includes time-and-age varying selectivity as operating model for a simulation experiment exploring our analytic estimator for the sensitivity of effective sample size on changes in input sample size. Changes in selectivity in the operating model are not modeled in the estimation model, and this model mis-specification results in greater error between predicted and observed age-composition data than is expected given the input sample size n_{input}; this in turn results in downweighting of age-composition data (such that $n_{effective} < n_{input}$). Our simulation experiment shows that we can accurately predict changes in effective sample size resulting from increases in input-sample size (Fig. 5). Predicting the impact of a 400% increase in input sample size (i.e., from $n_{input} = 20$ to $n^*_{input} = 100$) generates the most variable predictions of $n_{effective}$ (Fig. 5, left column 2nd row), while predicting a 150% or

100% increase results in more precise predictions. In all three scenarios, the predicted $n^*_{effective}$ is an unbiased predictor for $n_{effective}$ (average log-error between -0.01 and 0.02).

Discussion

We have provided two new estimators that collectively predict the impact of changing age-reading effort on the amount of information used by stock assessments, as measured by the effective sample size. Although past studies have developed analytic (i.e., closed-form) formulae to predict changes in information resulting changes in the design for collecting age-composition data (e.g., Lai, 1987), these did not account for the impact of assessment model mis-specification and resulting effects of data weighting. Importantly, our approach splits this analysis into two separate algorithms; these algorithms could be updated periodically by separate research teams, which helps to spread the work (and resulting buy-in) within a given research organization. Both steps are also closely conditioned upon the specific circumstances for a given stock, either in terms of the degree of intra-haul and inter-haul correlation and variance in catch rates (Step 1) or the degree of assessment model mis-specification and resulting data weights (Step 2).

As one worked example for these analyses, we return to the most recent assessment for Canary rockfish off the US West Coast (Thorson and Wetzel, 2015). We do not have a current implementation of the software to expand length-samples to expanded length-composition data, but treating average values from Stewart and Hamel (2014) as a marginal predicted change, we predict that a 10% increase in n_{input} would require at least 56 additional length samples per year. Then, using the Ianelli-McAllister estimate of $n_{effective} = 75.4$ given $n_{input} = 79$ for length samples in the bottom trawl survey for Canary rockfish (Thorson and Wetzel, 2015) and assuming that the Dirichlet-multinomial likelihood would give a similar value (see Thorson, 2019b) such that $\log(\theta) \approx \text{logit}\left(\frac{75.4}{79}\right) = 3.04$, we predict that a 10% increase in input sample size (from $n_{input} = 79$ to $n^*_{input} = 86.9$) would result in a 9.5% increase in effective sample size (from $n_{effective} = 75.4$ to $n^*_{effective} = 82.6$). We recommend future research to validation this prediction that 56 additional length samples would result in most in a 9.5% increase in effective sample size.

By conditioning upon specifics for a given species, this approach allows analysts to compare results among species with different circumstances in their survey data and stock assessment; comparison among species is necessary for any subsequent optimization of survey effort across those same species. In particular, comparison across species could be combined with information regarding the cost of field sampling and otolith reads as well as economic benefits of improved information to provide a “net value” for proposed changes in age-reading effort. Before embarking on this larger process, however, we recommend that analysts apply both steps to a few carefully selected species (e.g., representative rockfishes vs. flatfishes) to explore whether any initial patterns emerge. For example, Stewart and Hamel (2014) showed that thornyheads (*Sebastolobus* spp.) have substantial variation in age within each trawl sample, such that reading the age of multiple individuals for each trawl tow results in a larger input sample size than for rockfishes or flatfishes. Generalities such as this could then be used to justify future “rules of thumb” to qualitatively guide resource decisions, without requiring resources to regularly update results (e.g., Gerritsen and McGrath, 2007).

The formula predicting effective sample size $n^*_{effective}$ resulting from a change in input sample size (Eq. 2) depends upon an accurate estimate of input sample size when estimating the Dirichlet-multinomial weighting parameter θ within the assessment model. To see this, note that the ratio of effective and input sample size is a logistic function of $\log(\theta)$ in Eq. 1, while the predicted effective sample size is a Michaelis-Menten function of θn_{input}. This implies that mis-specifying n_{input} during assessment-model fitting will result in a compensatory change in θ, and this will in turn result in a biased estimate of $n^*_{effective}$. Therefore, applying the theory developed here will require accurate estimates of n_{input}. There is a growing literature regarding approaches to estimating n_{input} using the bootstrap (Stewart and Hamel, 2014), simple design-based estimators (Thorson, 2014), or spatio-temporal methods (Thorson and Haltuch, 2018). Increased efforts to use these methods to estimate n_{input} in real-world assessment models is needed, such that resulting estimates of data weighting (i.e., θ) can then be used to predict $n^*_{effective}$ given changes in ageing effort. Similarly, comparing the estimation covariance for abundance indices-at-age and the covariance in residual fits to an assessment in a multivariate-normal likelihood (e.g., Fig. 1 and 3 for autumn herring in the Q1 ICES survey from Berg and Nielsen (2016)) could presumably provide analogous information regarding the magnitude of additional covariance occurring given model specification; this information could then be used to predict how assessment model fit would likely change with changing precision for abundance indices.

This study investigates only the case where the assessment model estimates selectivity that is constant over time. Xu et al. (2019b) showed that the Dirichlet-multinomial approach performs well in estimating effective sample size regardless of whether the variation in selectivity is estimated in the assessment model. Accounting for the variation in selectivity can result in more precise estimations of both spawning biomass and fishing mortality if composition data have large sample sizes (Xu *et al.*, 2019a). More importantly, estimating time-varying selectivity will reduce the magnitude of model mis-specification, which means that the ratio of effective to input sample size (i.e., $\theta/1+\theta$) will also increase. This in turn increases the expected benefit of increasing input sample size. Therefore, increasing age reads will typically allow for improved model specification (i.e., estimating additional time-varying processes), which in turn will increase the value of further age-reads. The virtuous cycle between improved model specification and increasing value for age-reading effort is one reason why it is important to periodically re-evaluate age-reading expenditures for each stock relative to other survey-design and management priorities. Alternatively, other forms of model mis-specification may not be adequately represented via down-weighting of compositional data (Punt, 2017), and we recommend further research regarding best-practices for weighting data (and resulting impacts on age-reading effort) from alternative forms of model mis-specification.

We have only addressed Steps 1-2 (age reads → input sample size → effective sample size), and not Steps 3-4 (effective sample size → assessment model variance → management performance). Predicting changes in assessment-model variance from changes in effective sample size will depend upon the assessment-model output evaluated. However, we hypothesize that the estimate of the variance of recruitment will be closely linked to changing effective sample size, and that an analytic formula could be developed. We furthermore hypothesize that such a formula should include several factors including: (1) the evenness of proportion-at-age across different age-classes, as affected by maturity and mortality-at-age, as well as recruitment variance; and (2) the degree to which information is provided by age-composition data versus other data sets (as measured; e.g., using likelihood profiles). However, we leave the elaboration of this technique as a topic for future research. In the meantime, we envision that assessment scientists will generally have some familiarity with the magnitude of changes that occur when changing the effective sample size for age-composition data. The impact of changing assessment-model variance of management performance has been widely studied, including elements involving autocorrelated errors (Wiedenmann *et al.*, 2015), delays in management actions (Shertzer and Prager, 2007), alternative approaches to rebuilding plans (Wetzel and Punt, 2016), and many other topics. We hope that our study sparks a similar interest in exploring steps 1-3 that we outline here.

This study is a key step in demonstrating how changes in the sampling intensity for compositional data influence the effective sample size for these data in stock assessment models. Understanding this relationship is essential in a highly variable funding environment that often drives reductions in total survey costs. The impact of reduced age-reading effort will not be the same across stocks, and the impact on resulting management advice from stock assessments should be used to guide this process. Open-loop simulation experiments estimating bias and precision in key assessment model parameters and management quantities would be a beneficial extension to this work. A logical step forward would be to develop a spatial operating model with flexibility to model ontogenetic movement or spatial distribution shifts linked to environmental covariates, which would of great interest given expected changes in the North Pacific and elsewhere (Morley *et al.*, 2018). The combination of a complex operating model and a simple estimation model would allow for a thorough examination of how the interaction of model misspecification and the estimation of input and effective sample size influence assessment outcomes. This open-loop simulation approach would help to determine a threshold sample size required to provide adequate management advice for multiple species and would also identify the trade-off in required sampling effort among species.

Alternatively, future research could apply a closed-loop simulation experiment (i.e., management strategy evaluation, MSE) to identify the trade-offs over a range of management objectives. The MSE approach evaluates a management strategy (a.k.a. procedure) that involves the entire pipeline from data collection, stock assessment model fitting, management advice and decision making, and implementation of the management advice (Sainsbury *et al.*, 2000; Punt *et al.*, 2016). The evaluation of this integrated pipeline is then conducted by simulating each step; this simulation is used to measure the expected management performance for a across a range of alternative circumstances. This process is a generic simulation-based approach to decision theory, and therefore could be used to evaluate the same questions explored here; that is, the impact of inter- and intra-haul correlations and variance, combined with model mis-specification, on the value of additional age-reads. Importantly, this closed-loop simulation would be useful to determine consequences when iteratively applying a decision-rule based on these analyses to continually update sample sizes for a given species (e.g., as shown in Fig. 6 when moving several times through the process in each column). Importantly, some decision-rules could result in poorly-sampled species having low value for additional samples (due to poor assessment model fits and high θ estimates), thus resulting in a lower priority for age-sampling effort and a downward spiral in resource allocation for that species. Presumably this issue could be corrected by developing optimal "decision-rules" for using results of an analysis to update funding decisions. We therefore recommend further research for how these two estimators are used to optimize funding decisions.

We conclude by noting that a management strategy evaluation (or any other integrated process for simulating population, sampling, and assessment-model components) is difficult to use when optimizing resource allocations across stocks. In particular, it is difficult to develop a simulation model that closely matches the sampling (e.g., within- and among-haul variability), population dynamics (e.g., counterfactual recruitment dynamics) and assessment-model (e.g., degree of model mis-specification) characteristics for each individual stock. Our workflow, by contrast, uses the raw sampling data (Step 1) and fitted stock-assessment model (Step 2) to automatically generate an operating model. We therefore advocate that science agencies should develop a suite of tools for evaluating the impact of changes in field-sampling effort, ranging from simple (cost efficient) to complex (expensive) following the schematic shown in Fig. 6. In particular, we hope that this suite involves analytical approaches (e.g., Step 2), as well as bootstrap methods (e.g., Step 1) that can be uniformly applied across species to explore trade-offs across species.

Acknowledgments

We thank C. Lunsford and A. Hollowed, who discussed prospects for follow-up research using these results to compare age-reading effort across species. M. Bryan, P. Hulson, and J. Thorson all identified the problem and discussed potential approaches. J. Thorson developed code for Step 1. J. Thorson derived the theoretical relationship for Step 2 and H. Xu and J. Thorson developed code to test this relationship. All authors contributed to writing and editing. We thank I. Stewart, T. Helser, and two anonymous reviewers for comments on an earlier draft.

Works cited

Anderson, M. J., and Santana-Garcon, J. 2015. Measures of precision for dissimilarity-based multivariate analysis of ecological communities. Ecology Letters, 18: 66–73.

Berg, C. W., Nielsen, A., and Kristensen, K. 2014. Evaluation of alternative age-based methods for estimating relative abundance from survey data in relation to assessment models. Fisheries Research, 151: 91–99.

Berg, C. W., and Nielsen, A. 2016. Accounting for correlated observations in an age-based state-space stock assessment model. ICES Journal of Marine Science, 73: 1788–1797.

Francis, R. I. C. C. 2011. Data weighting in statistical fisheries stock assessment models. Canadian Journal of Fisheries and Aquatic Sciences, 68: 1124–1138.

Francis, R. I. C. C. 2017. Revisiting data weighting in fisheries stock assessment models. Fisheries Research, 192: 5–15.

Gerritsen, H. D., and McGrath, D. 2007. Precision estimates and suggested sample sizes for length-frequency data. Fishery Bulletin, 105: 116–120.

Hulson, P. J. F., Hanselman, D. H., and Quinn, T. J. 2011. Effects of process and observation errors on effective sample size of fishery and survey age and length composition using variance ratio and likelihood methods. ICES Journal of Marine Science: Journal du Conseil, 68: 1548–1557.

Hutniczak, B., Lipton, D., Wiedenmann, J., and Wilberg, M. 2019. Valuing changes in frequency of fish stock assessments. Canadian Journal of Fisheries and Aquatic Sciences, 76: 1640–1652.

Kotwicki, S., and Ono, K. 2019. The effect of random and density-dependent variation in sampling efficiency on variance of abundance estimates from fishery surveys. Fish and Fisheries, 20: 760–774.

Lai, H.-L. 1987. Optimum allocation for estimating age composition using age-length key. Fishery Bulletin, 85: 179–186.

Maunder, M. N. 2003. Paradigm shifts in fisheries stock assessment: from integrated analysis to Bayesian analysis and back again. Natural Resource Modeling, 16: 465–476.

McAllister, M. K., and Ianelli, J. N. 1997. Bayesian stock assessment using catch-age data and the sampling: importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences, 54: 284–300.

Methot, R. D. 2009. Stock Assessment: Operational Models in Support of Fisheries Management. *In* The Future of Fisheries Science in North America, pp. 137–165. Ed. by R. J. Beamish and B. J. Rothschild. Springer Netherlands, Dordrecht.

Methot, R. D., and Wetzel, C. R. 2013. Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. Fisheries Research, 142: 86–99.

Morley, J. W., Selden, R. L., Latour, R. J., Frolicher, T. L., Seagraves, R. J., and Pinsky, M. L. 2018. Projecting shifts in thermal habitat for 686 species on the North American continental shelf. PLoS ONE. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0196127 (Accessed 8 September 2019).

Nielsen, A., and Berg, C. W. 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fisheries Research, 158: 96–101.

Pennington, M., Burmeister, L.-M., and Hjellvik, V. 2002. Assessing the precision of frequency distributions estimated from trawl-survey samples. Fishery Bulletin, 100: 74–80.

Punt, A. E., Smith, D. C., KrusicGolub, K., and Robertson, S. 2008. Quantifying age-reading error for use in fisheries stock assessments, with application to species in Australia's southern and eastern scalefish and shark fishery. Canadian Journal of Fisheries and Aquatic Sciences, 65: 1991–2005.

Punt, A. E., Butterworth, D. S., de Moor, C. L., De Oliveira, J. A. A., and Haddon, M. 2016. Management strategy evaluation: best practices. Fish and Fisheries, 17: 303–334.

Punt, A. E. 2017. Some insights into data weighting in integrated stock assessments. Fisheries Research, 192: 52–65.

Sainsbury, K. J., Punt, A. E., and Smith, A. D. M. 2000. Design of operational management strategies for achieving fishery ecosystem objectives. ICES Journal of Marine Science: Journal du Conseil, 57: 731–741.

Shertzer, K. W., and Prager, M. H. 2007. Delay in fishery management: diminished yield, longer rebuilding, and increased probability of stock collapse. ICES Journal of Marine Science: Journal du Conseil, 64: 149–159.

Stewart, I. J., and Hamel, O. S. 2014. Bootstrapping of sample sizes for length-or age-composition data used in stock assessments. Canadian Journal of Fisheries and Aquatic Sciences, 71: 581–588.

Thorson, J. T. 2014. Standardizing compositional data for stock assessment. ICES Journal of Marine Science: Journal du Conseil, 71: 1117–1128.

Thorson, J. T., and Wetzel, C. R. 2015. The status of canary rockfish (*Sebastes pinniger*) in the California Current in 2015. Pacific Fishery Management Council, 7700: 97200–1384.

Thorson, J. T., and Cope, J. M. 2015. Catch curve stock-reduction analysis: An alternative solution to the catch equations. Fisheries Research, 171: 33–41.

Thorson, J. T., Johnson, K. F., Methot, R. D., and Taylor, I. G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research, 192: 84–93.

Thorson, J. T., and Haltuch, M. A. 2018. Spatiotemporal analysis of compositional data: increased precision and improved workflow using model-based inputs to stock assessment. Canadian Journal of Fisheries and Aquatic Sciences: 1–14.

Thorson, J. T. 2019a. Guidance for decisions using the Vector Autoregressive Spatio-Temporal (VAST) package in stock, ecosystem, habitat and climate assessments. Fisheries Research, 210: 143–161.

Thorson, J. T. 2019b. Perspective: Let's simplify stock assessment by replacing tuning algorithms with statistics. Fisheries Research, 217: 133–139.

Wetzel, C. R., and Punt, A. E. 2016. The impact of alternative rebuilding strategies to rebuild overfished stocks. ICES Journal of Marine Science, 73: 2190–2207.

Wiedenmann, J., Wilberg, M. J., Sylvia, A., and Miller, T. J. 2015. Autocorrelated error in stock assessment estimates: Implications for management strategy evaluation. Fisheries Research, 172: 325–334.

Xu, H., Thorson, J. T., Methot, R. D., and Taylor, I. G. 2019a. A new semi-parametric method for autocorrelated age- and time-varying selectivity in age-structured assessment models. Canadian Journal of Fisheries and Aquatic Sciences, 76: 268–285.

Xu, H., Thorson, J. T., and Methot, R. D. 2019b. Comparing the performance of three data weighting methods when allowing for time-varying selectivity. Canadian Journal of Fisheries and Aquatic Sciences. https://www.nrcresearchpress.com/doi/abs/10.1139/cjfas-2019-0107 (Accessed 23 September 2019).

Zimmermann, F., and Enberg, K. 2017. Can less be more? Effects of reduced frequency of surveys and stock assessments. ICES Journal of Marine Science, 74: 56–68.

Table 1: Description of analyses that collectively assess the likely impact of changes in the number of otoliths read n_{reads} on the variance of stock-assessment model output $Var(X)$. This involves a four-stage process: (Step 1) the impact of changing n_{reads} on the input sample size n_{input} using a bootstrap experiment; (Step 2) the impact of changing n_{input} on the effective sample size $n_{effective}$ of compositional data when tuned in an assessment model; (Step 3) the impact of changing $n_{effective}$ on $Var(X)$; and (Step 4) the impact of changing assessment-model variance on management performance.

Analysis number	Input	Output	Approach to measure effect	Data needed	Responsibility for updating analysis	Processes affecting output
1	n^*_{reads}	n^*_{input}	Non-parametric bootstrap of Hamel-Stewart (2014) sample-size estimator	• Records for sampling data • Age/length subsampling measurements	Survey team	• Variance in age/length composition within and among samples

2	n^*_{input}	$n^*_{effective}$	Analytic formula: $$n^*_{effective} = \frac{\beta n^*_{input}}{\beta + n^*_{input}}$$ where $$\beta = 1 + \theta n_{input}$$	• Stock assessment model fitted to expanded data, and applying Dirichlet-multinomial likelihood	Survey or assessment team	• Variance in biomass/abundance among samples • Magnitude of error in in fit to expanded composition data (i.e., due to unmodeled variation in growth, selectivity, mortality, etc.)

3	$n^{*}_{effective}$	$Var^{*}(X)$	Parametric bootstrap of assessment model	• Stock assessment model with bootstrap simulator	Assessment team	• Which assessment model output is under consideration • Other sources of data available within the model
4	$Var^{*}(X)$	Management performance	Management strategy evaluation	• Variance and autocorrelation in assessment errors	Research community	• Delays in management (Shertzer and Prager, 2007) • Autocorrelation in errors (Wiedenmann *et al.*, 2015)

- Structure of management plan (Wetzel and Punt, 2016)
- Economic values (Hutniczak *et al.*, 2019)

Fig. 1: Graphical depiction of simulation experiment testing performance of bootstrap estimator in Analysis #1 (i.e., predicting the likely consequence of changing ageing effort on resulting input sample size). Black circles refer to data sets, whether used to parameterize an operating model ("Real-world data") or arising from the operating model; blue squares refer to models (both the operating model, and the estimator predicting changes in input sample size); red squares refer to estimators that are used to evaluate performance, i.e., by comparing predicted input sample size (n^*_{input}) with the input sample size arising from a changed number of age reads ($\tilde{n}_{input}$).

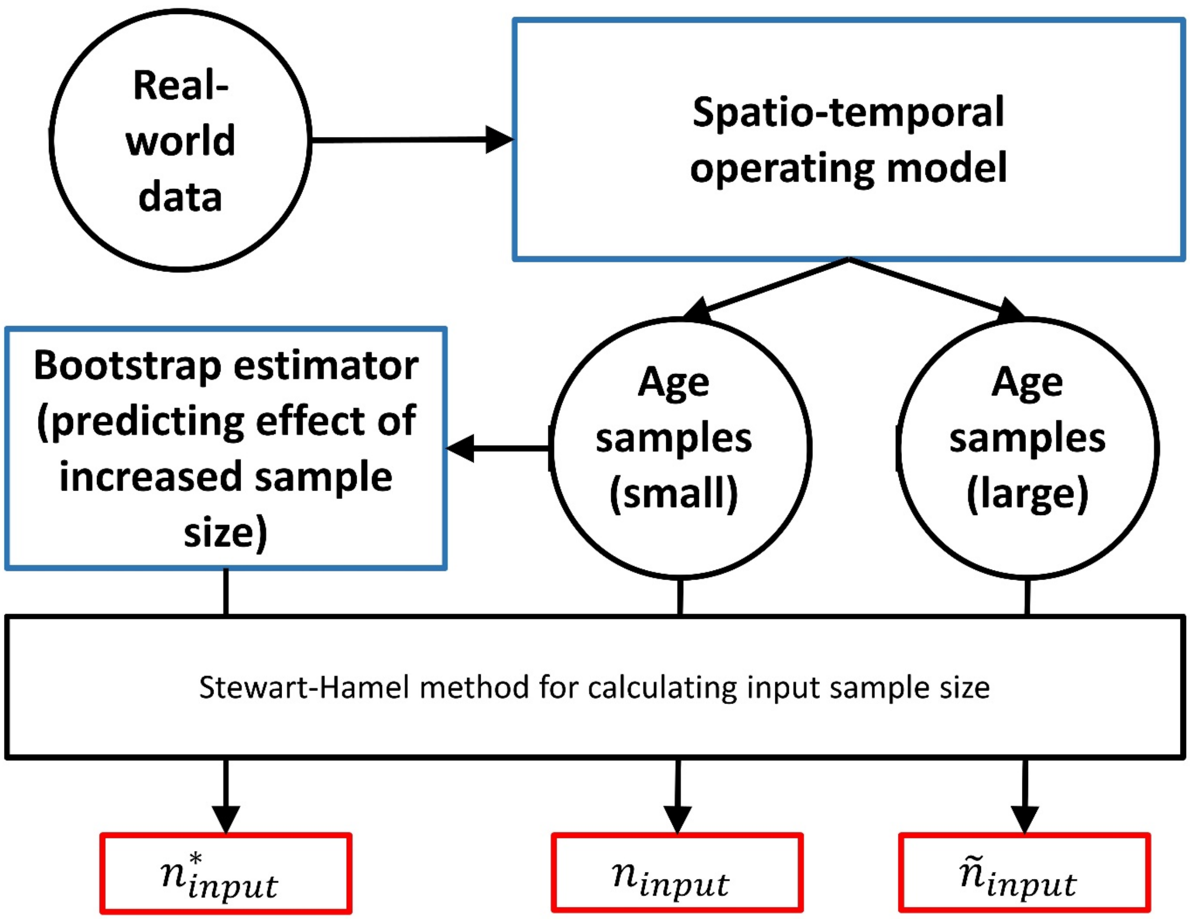


Fig. 2: Illustration of the theoretical relationship between a new input sample size n^*_{input} and the resulting effective sample size, $n^*_{effective}$ (thick solid line), for a hypothetical scenario in which the Dirichlet-multinomial parameter $\theta = 0.8$ and the original input sample size $n_{input} = 125$ (resulting in $n_{effective} = 56.1$, as indicated by the black circle). In this circumstance, the asymptotic maximum value for $\lim_{n^*_{input} \to \infty} \left(n^*_{effective} \right) = 1 + \theta n_{input} = 101$ (dashed line) and the Michaelis-Menten half-saturation value is $1 + \theta n_{input} = 101$, such that $n^*_{effective}$ achieves half of its maximum possible value when $n^*_{input} = 1 + \theta n_{input}$ (shown by the arrows connecting $n^*_{effective} = 50.5$ to $n^*_{input} = 101$).

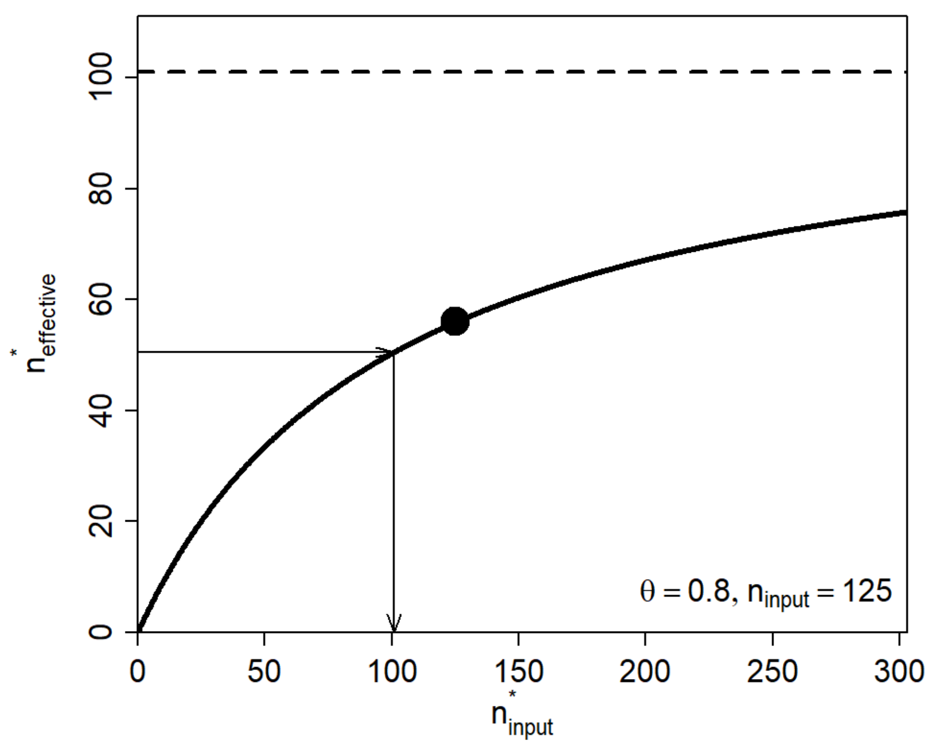


Fig. 3: Graphical depiction of simulation experiment testing performance of bootstrap estimator in Analysis #2 (i.e., predicting the likely consequence of changing input sample size on effective sample size). See Fig. 1 caption for color-code conventions. Performance is evaluated by comparing predicted effective sample size ($n^*_{effective}$) with the effective sample size occurring with a large input sample size ($\tilde{n}_{effective}$)

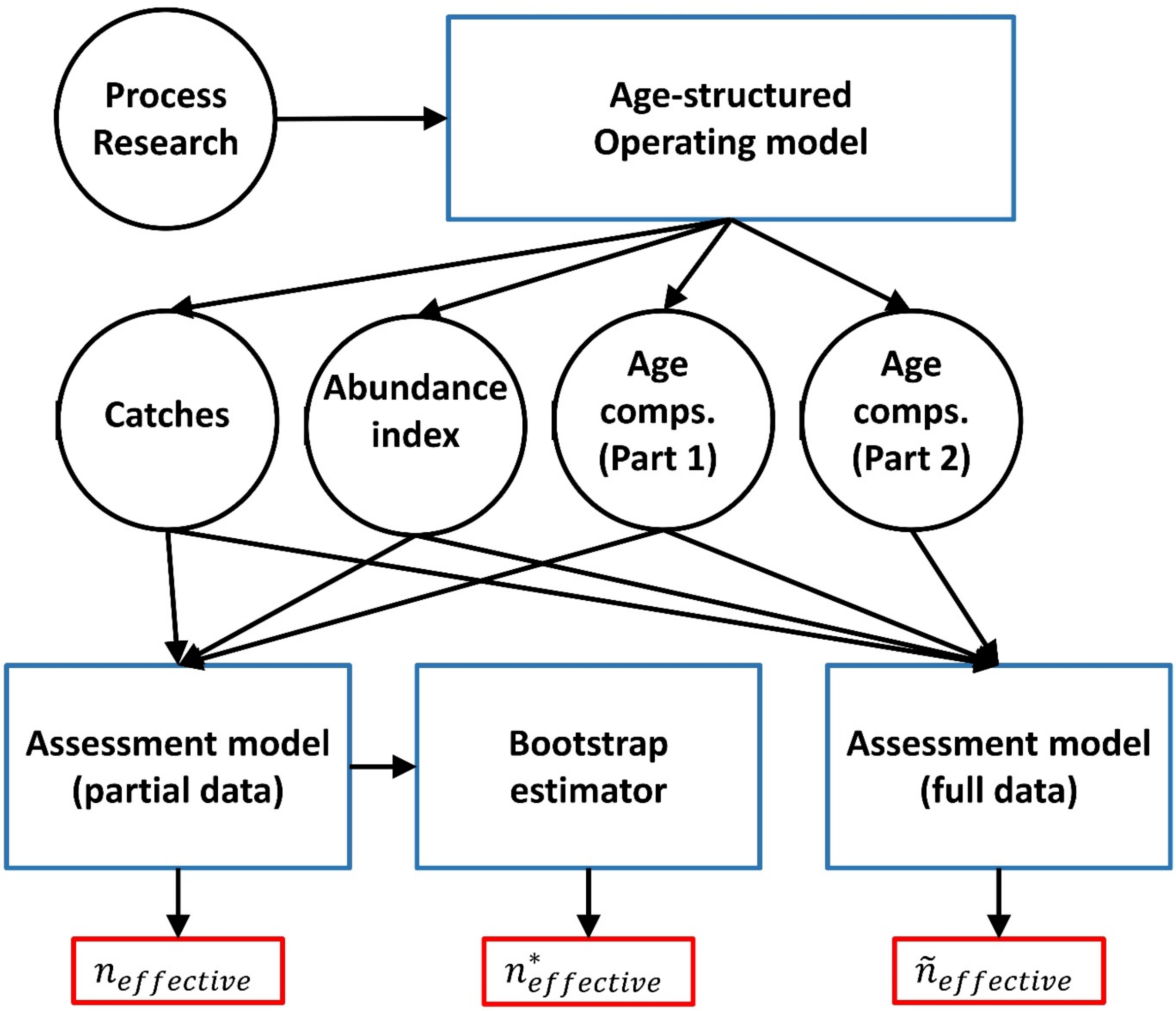


Fig. 4: Comparison of the ratio of input sample size given 40 otoliths read per tow and 20 otoliths read per tow ($\tilde{n}_{input}(t)/n_{input}(t)$, grey histogram), and the ratio of predicted sample size using a bootstrap estimator based on an input sample size of 20 otoliths per tow ($n^{*}_{input}(t)/n_{input}(t)$, red histogram) when aggregating results across 10 years and 30 simulation replicates; a linear estimator would estimate a ratio of 2.0 for all samples.

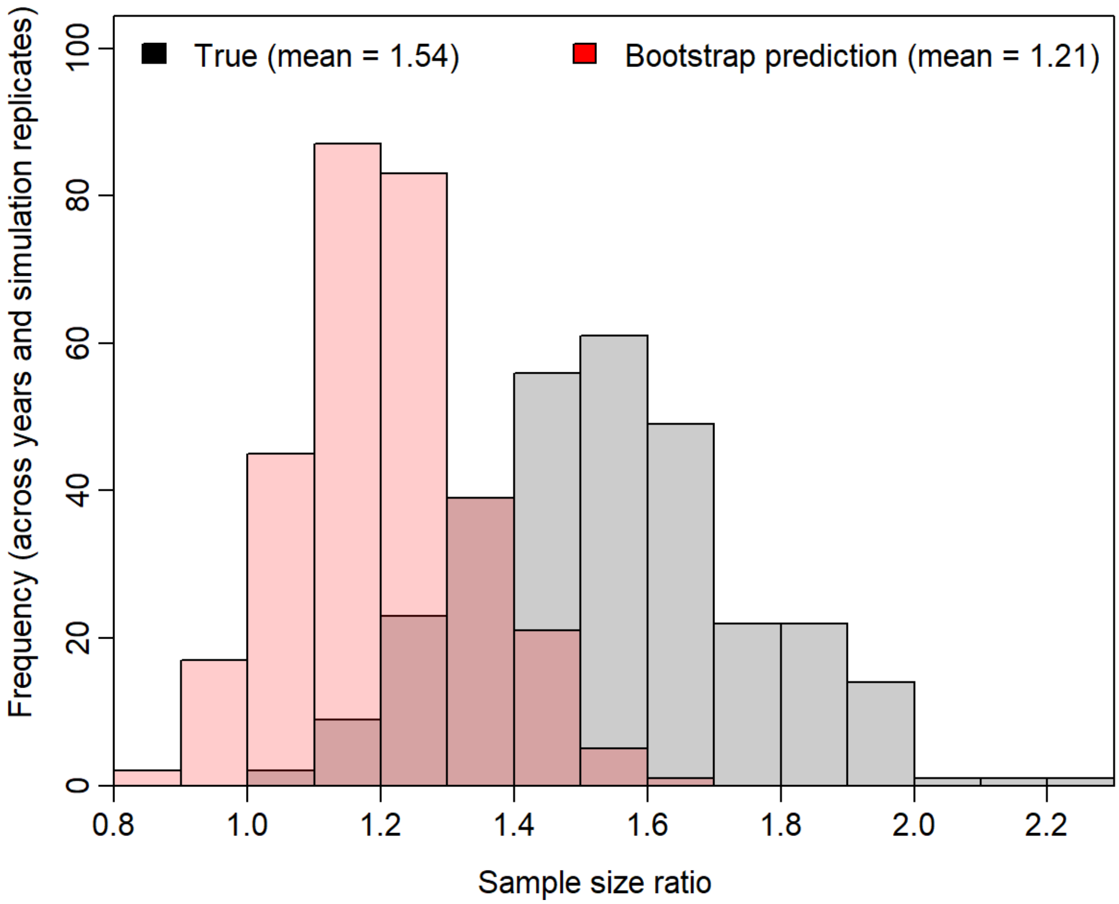
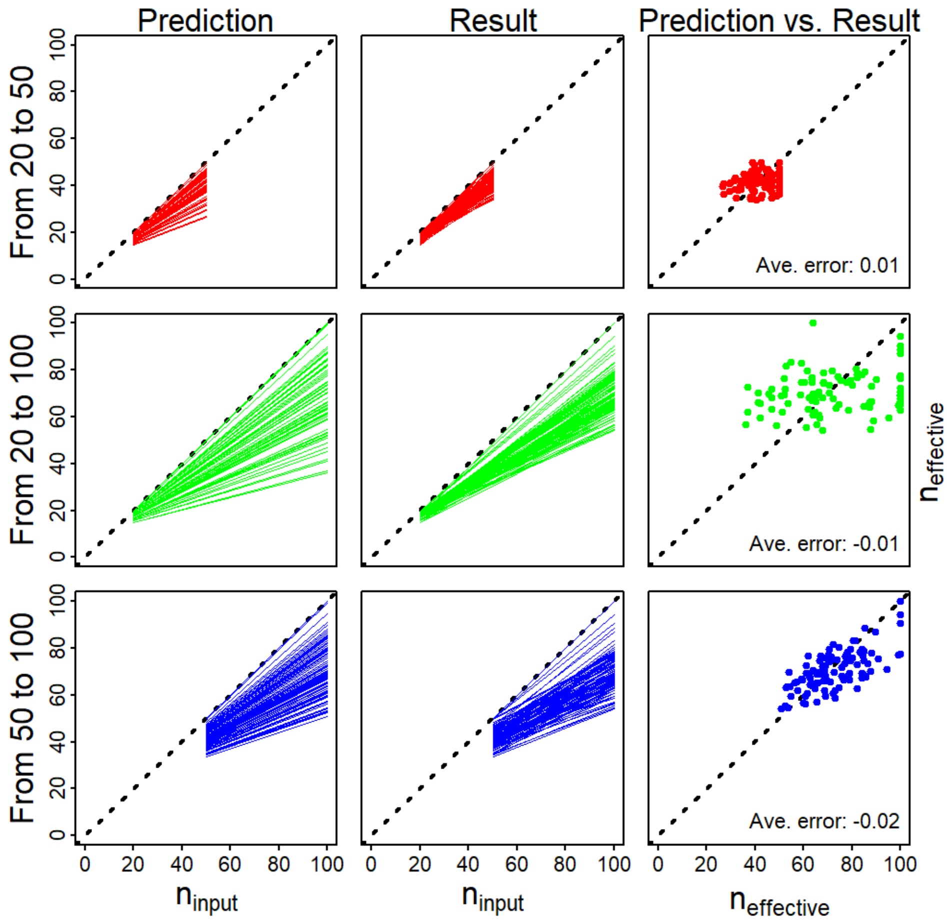


Fig. 5: Results from a simulation experiment evaluating the performance of the analytic estimator for changes in effective sample size arising from a change in input-sample size (Eq. 2), implementing 100 simulation replicates arising from three treatments. We show predictions given original sample size $n_{input} = 20$ and predicting the effect of new sample size $n^*_{input} = 50$ (top row), given $n_{input} = 20$ and predicting $n^*_{input} = 100$ (middle row), or given $n_{input} = 50$ and predicting $n^*_{input} = 100$ (bottom row). The first column shows the predictions, where a line connects the effective sample size $n_{effective}$ (y-axis) estimated using the original input sample n_{input} (x-axis) with the predicted effect sample size $n^*_{effective}$ (y-axis) for the new input sample size n_{input} (x-axis) in each simulation replicate; these lines are generally below the one-to-one line (dashed line) due to the predicted nonlinear relationship. The middle row shows the true values, where a line connects original $n_{effective}$ and n_{input} with new n^*_{input} and resulting $n_{effective}$. The right column compares the predicted $n^*_{effective}$ (x-axis) and resulting $n_{effective}$ (y-axis). A well-performing estimator will have predictions $n^*_{effective}$ and resulting $n_{effective}$ centered around the one-to-one line; we also list the average error (in log-space) at the bottom-right corner of panels in the right column.



Prediction
Result
Prediction vs. Result
From 20 to 50
From 20 to 100
From 50 to 100
Ave. error: 0.01
Ave. error: -0.01
Ave. error: -0.02
n_{input}
n_{input}
$n_{effective}$
$n_{effective}$
0 20 40 60 80 100

Fig. 6 – Schematic showing proposed workflow for periodic update of sampling effort allocated to multiple species. Step 1 (sensitivity of input on nominal sample size) and Step 2 (sensitivity of effective on input sample size) could each be updated every ΔT years to periodically update the design for age reading effort.