**Model-based estimates of effective sample size in Stock Synthesis using the Dirichlet-multinomial distribution**

James T. Thorson1, Kelli F. Johnson2, Richard D. Methot1, Ian G. Taylor1

1Fisheries Resource Assessment and Monitoring Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Blvd. East, Seattle, WA 98112, USA

2School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195-5020, USA

**Abstract**

Theoretical considerations and applied examples suggest that stock assessments are highly sensitive to the weighting of different data sources whenever data sources conflict regarding parameter estimates. Previous iterative reweighting approaches to weighting compositional data are generally ad hoc, do not propagate uncertainty about data-weighting when calculating uncertainty intervals, and often are not re-adjusted when conducting sensitivity or retrospective analyses. We therefore incorporate the Dirichlet-multinomial (DM) distribution into Stock Synthesis, and propose it as a model-based method for estimating effective sample size. This distribution incorporates one additional parameter per fleet (with the option of mirroring its value among fleets), and we show that this parameter represents the ratio of nominal (“input”) and effective (“output”) sample size. We demonstrate this approach using data for Pacific hake, where DM and iterative reweighting approaches give similar and plausible results. We also use simulation testing to demonstrate the estimation properties of this new estimator, and show that it provides approximately unbiased estimates of variance inflation when compositional samples capture clusters of individuals with similar ages/lengths. We conclude by recommending further research to develop computationally efficient estimators of effective sample size that are based on alternative, *a priori* consideration of sampling theory.

**Keywords**: Dirichlet-multinomial; integrated assessment; multinomial; statistical catch-at-age

**1. Introduction**

Stock assessment models are quantitative tools that are used to provide a scientific basis for the management of marine fishes (Walters and Martell, 2004). Assessment models increasingly incorporate biological assumptions regarding the population dynamics of fished species, and population dynamics parameters are estimated by fitting the assessment model to available data (Maunder and Punt, 2013). Fitting population models to available data is typically done using likelihood-based statistics, and the proper estimation of confidence and forecast intervals therefore generally requires accounting for heteroskedastic and correlated residuals as caused by unmodeled biological or measurement process (Thorson and Minto, 2015). Theoretical considerations and applied examples suggest that integrated statistical stock assessments are highly sensitive to the weighting of different data sources whenever sources conflict regarding parameter estimates. Consequently, estimates of stock status and productivity are often highly dependent upon the weighting of different data sources (Francis, 2011).

Stock assessment models frequently are fitted to sampling data that are informative about the proportion of the vulnerable population belonging to different observable categories. Common categories include the proportion of survey or fishery catch that is associated with different ages, lengths, and/or sexes. Compositional sampling is frequently assumed to follow a multinomial distribution, e.g., drawing 10 marbles from an urn that contains 15 red, 45 blue, and 40 green marbles. The multinomial distribution is derived from the assumption that a given compositional sample represents independent sampling of a fixed and known number of individuals (i.e., 10 marbles) from a set of possible categories, where there is a true “fixed” probabili­­ty *pc* associated with each category *c* (i.e., *pc*=0.15, 0.45, and 0.4 for red, blue, and green marbles). Each sample will not perfectly represent the true distribution, e.g., a single sample of 10 marbles might yield 1 red, 4 blue, and 5 green (i.e., where *pobs*=0.1, 0.4, and 0.5). The multinomial distribution implies that the sampling variance (i.e., variation if the sampling process was replicated) is a function of both the true probability and sample size, , where *n* is the number sampled and *p* is the true probability for each category. Thus, as *n* increases, the coefficient of variation for sampling variance for each category decreases by .

In practice, compositional data for fish populations arises from a process of sampling fish (e.g., non-extractive visual samples or by capturing and measuring fishes), and this sampling process is more complicated than assumed by a multinomial distribution. In particular, compositional data are likely to have greater variance than predicted by a multinomial distribution based on the number of individual fish that are sampled, and we term this “overdispersion” in the following. This overdispersion arises for two primary reasons:

1. Lack of statistical independence among individuals in a sample: The multinomial distribution assumes that each individual (e.g., marble) is sampled independently. This assumption is often violated, e.g., when fish schooling behavior leads to a single age being over-represented in each individual sample (McAllister and Ianelli, 1997).
2. Heterogeneity among samples: The multinomial distribution assumes that the probability *pc* associated with each category *c* is constant among many sampling units. This assumption is often violated, e.g., when juvenile or adult fish have an affinity for a particular depth range, leading to spatially varying proportions at age (Thorson, 2014) or between sampling tows (Crone and Sampson, 1997), such that data from different states or ports yield compositional data with different proportions for each category.

In practice, compositional data are standardized (sometimes termed “expanded”) to transform raw compositional sampling data into an aggregated estimate of the proportion in each category in a given year (Shelton et al., 2012; Thorson, 2014). Compositional standardization ideally results in an estimate of “input” sample size for the compositional data in a given year, where estimates are frequently a function of both (i) the number of boats and (ii) the total number of sampled fish (Crone and Sampson, 1997; Stewart and Hamel, 2014).

The multinomial distribution is often used in the likelihood function that is maximized to estimate parameters in an integrated assessment model. In this usage, the multinomial distribution is used to approximate the probability that a given composition sample arose from the fish population given proposed values for estimated parameters. Input sample size (calculated during compositional standardization) is often used as the sample size for this likelihood function, and therefore, input sample size controls the weighting of compositional data relative to other data sources. However, model misspecification may cause this input sample size to be an inappropriate measure of data weighting. As a thought experiment, imagine that all participants in a fishery made-up numbers for their reporting of fish sizes in their catch. These data would have no information about the size-composition of the population, and a stock assessment model would have optimal performance if it assigned zero weight to these data. As a less extreme example, age-composition data are often obtained by laboratory examination of fish samples (otoliths or spines), and these laboratory methods sometimes mis-identify the age of a given fish. Ageing error will cause age-composition data to be a noisy measure of the true age-composition, such that increasing the sample size of age-composition data is only useful if ageing error is absent (Coggins and Quinn, 1998). However, if the stock assessment model uses double-reading and ageing-error methods to correct for the ageing error (Methot and Wetzel, 2013; Punt et al., 2008), these data might again be a useful measure of population age structure. This example highlights that the proper weight of composition data depends upon the specification of the model, where model misspecification (e.g., neglecting the impact of ageing error) results in a lower optimal weight for compositional data.

These considerations therefore suggest that the optimal weighting of compositional data depends upon the match between the specified model and the true data-generating process. This conclusion implies that compositional weighting can be accomplished by inspecting the goodness-of-fit between compositional data and estimated proportions from the assessment model, and decreasing the sample size for data that generally do not match. This process suggested by McAllister and Ianelli (1997), who proposed estimating the “effective sample size” for compositional data via the match between predicted and observed compositional data. However, iterative reweighting approaches such as those suggested by McAllister-Ianelli (1997) require the following steps: (1) fit the assessment model to available data; (2) extract estimates of compositional proportions; (3) calculate the effective sample size; (4) input the new effective sample size; (5) iterate steps 1-4 a fixed number of times, or until subsequent iterations do not change the estimate of effective sample size. Decreasing the effective sample size has identical impact to decreasing the weight of the multinomial likelihood function by a fixed amount (Francis, 2011), such that this process is essentially reweighting the compositional data during each iteration of the algorithm. This iterative-reweighting algorithm has several draw-backs, including being infeasible to repeat for every sensitivity run, inability to incorporate into simulation designs, and difficulties when estimating likelihood profiles for stock assessment parameters.

In the following, we seek to develop a method to estimate effective sample size during parameter estimation. Ideally, this would involve estimating a new parameter which governs the ratio of input and effective sample size. Uncertainty in this parameter could then be estimated using conventional methods (Magnusson et al., 2013), and its uncertainty could be propagated and evaluated during stock projections. For this purpose, we propose using a Dirichlet-multinomial (DM) distribution as the likelihood function of compositional data in an integrated assessment model. We show that this distribution involves estimating a new parameter, and can be parameterized such that it estimates an approximately linear relationship between input and effective sample size. We incorporate this new distribution into the Stock Synthesis (SS) stock assessment software, which is widely used in the United States and internationally (Methot and Wetzel, 2013), when fitting to marginal and conditional length or age composition data. We then use a case study and simulation experiment to show that the Dirichlet-multinomial likelihood function provides estimates of effective sample size that are similar to iterative reweighting methods, but without requiring multiple iterations of running the assessment model.

**2. Methods**

*2.1 Introducing the Dirichlet-multinomial distribution*

We here use a Dirichlet-,ultinomial distribution:

where is the proportion at age in the available data such that , *N* is the total number of samples in the available data (which is restricted to any non-negative real number), is the estimated proportion at age (such that ), and is the estimated variance inflation coefficient. Here, we use the gamma function, rather than the conventional factorial function, so that the Dirichlet-multinomial is defined for all non-negative sample sizes *N*, such that it reduces to the conventional Dirichlet-multinomial distribution whenever input sample size is a whole number. We here present theory using notation for age-composition data, but note that it is fully applicable (and implemented in Stock Synthesis) for length-composition data as well. The first term does not depend upon the parameters, but ensures that the value of the Dirichlet-multinomial function converges on the value of the conventional multinomial function as ,:

such that the multinomial distribution is a special case of the Dirichlet-multinomial distribution.

*2.2 Computing the effective sample size*:

We define the effective sample size of a distribution for compositional data *c* as the sample size of a multinomial distribution that has the same variance. The variance of a single element from a multinomial distribution is

where *N* is the sample size. Defining observed proportion , we see that:

i.e., variance decreases as the reciprocal of sample size.

We next introduce a Dirichlet distribution:

where and is the true proportion at age. The Dirichlet distribution has variance:

such that is the effective sample size of the Dirichlet distribution:

Finally, the variance of the observed proportion at age for a Dirichlet-multinomial distribution is approximately:

We therefore calculate the estimated effective sample size *Neff* of a Dirichlet-multinomial distribution as:

i.e., that the effective sample size is the harmonic sum of *N* and .

*2.3 Two potential parameterizations*

Given the Dirichlet-mulinomial distribution and the closed-form computation of its effective sample size, we propose two alternative parameterizations that may be useful in practice for length- and age-composition samples in stock assessment models. These parameterizations differ in terms of the function relating input and effective sample size (Fig. 1).

*2.3.1 Parameterization #1 – Linear effective sample size*

As a default, we recommend a re-parameterization of the Dirichlet-multinomial distribution, wherein the variance-inflation parameter is replaced by a linear function of input sample size *N,* i.e., . This results in the following probability distribution function:

which has effective sample size:

When *N>>1* and <<*1*, this further reduces to:

i.e., the parameter can be interpreted as the ratio of effective and input sample size under these conditions. We recommend using the “linear effective sample size” parameterization given that previous methods for weighting compositional data have generally multiplied the likelihood of compositional data by a fixed quantity *λ<1*, and this parameterization has similar behavior for large input sample sizes (in practice *N>10*).

*2.3.2 Parameterization #2 – Asymptotic effective sample size*

As a potential alternative, analysts may instead use the original parameterization of the Dirichlet-multinomial distribution:

with effective sample size:

This parameterization can revert to the multinomial distribution with sufficiently large , i.e., when .However, it provides an upper bound on effective sample size with lower values of , i.e., when . Therefore, this parameterization could be useful whenever analysts seek to estimate an upper bound on the effective sample size for a given year.

We have implemented both parameterizations of the DM distribution in SS (version 3.30; release date Jan, 2016), an integrated age-structured stock assessment framework frequently used to conduct assessments in many parts of the world (Methot and Wetzel, 2013). In the following, however, we focus exclusively on the linear parameterization (version #1), given that it more closely mimics the action of the McAllister-Ianelli iterative reweighting method. We recommend future research developing and exploring more-complicated two-parameter forms for the Dirichlet-multinomial distribution, which could combine the characteristics of both versions.

*2.4 Case study: Pacfic hake*

To demonstrate this new data-weighting method, we compare its performance with that of other data-weighting methods when applied to a recent stock assessment for Pacific hake, *Merluccius productus* (Taylor et al., 2015). Pacific hake is a semi-pelagic schooling species of commercial importance to fisheries off of the US West Coast and Western Canada. Recent management is conducted following procedures determined by an international treaty, and are informed by annual stock assessments implemented using Stock Synthesis. Data used in the 2015 stock assessment includes (1) catches from 1966 to 2014, (2) fishery age–composition samples, (3) an index of abundance from ten acoustic surveys conducted between 1995 and 2013, (4) survey age-composition samples associated with the accounstic survey, (5) cohort-specific definitions of ageing error that specify improved ageing accuracy with larger cohorts, and (6) ‘empirical’ weight-at-age data calculated from all fisheries and the acoustic survey for years 1975 to 2014, which are assumed to be known without error (Taylor et al., 2015).

Four assessment models were fit to data for Pacific hake, where each model used a different approach to data-weighting for the fishery age-composition data: (i) unweighted, (ii) tuned using an iterative approach, (iii) estimated using the Dirichlet-multinomial distribution, and (iv) weight of zero. Option (ii) involved fitting the model to simulated data, computing the ratio of the harmonic mean of yearly “effective” sample size to the arithmetic mean of yearly input sample size of fishery age-composition data, multiplying this value by the “weighting factor” for the fishery age-composition data used during parameter estimation, and then inputing this value as the new weighting factor. This process was repeated two times and the third fit to data was used as the final estimate of parameters. The initial weighting factor was set to one and all additional weighting factors had an upper bound bound of one to ensure that effective sample size was never greater than the original input sample size. Option (iv) specifies that the stock assessment was fitted only to abundance indices and survey age-composition data, and represents the extreme case of “zero” weight assigned to fishery compositional data. Preliminary exploration showed that the input sample size is approximately equal to effective sample size for survey age-composition data (i.e., the iterative approach results in a ratio of 0.94, and the Dirichlet-multinomial results in a ratio of 1.00). We therefore chose to not down-weight the survey age-composition data (i.e., we did not estimate the Dirichlet-multinomial parameter for the survey age-composition data, nor did we tune them).

*2.5 Simulation testing*

The performance of the Dirichlet-multinomial distribution implemented in Stock Synthesis was explored using simulated data. To do so, we simplified the Pacific hake estimation model in five ways: (1) changed fishery selectivity to be stationary over time (i.e., removed time-varying selectivity parameters), (2) changed all fishery age-composition sample sizes to a single fixed value per year, (3) changed all survey age-composition sample sizes to 100 samples per year, (4) changed age-specific ageing error to be stationary over time and equal to the baseline ageing-error matrix, and (5) changed to using an “explicit-F” parameterization, wherein instantaneous, fully-selected fishing mortality in each year is estimated as a fixed effect. We made change (5) so that the simulated fishing intensity is plausible given the simulated vector of recruitment deviations for each simulation replicate. We then ran the modified Pacific hake assessment model on available data, extracted estimated parameters, and used these estimates as the “true” values during the simulation experiment (while confirming that estimated stock status and productivity is generally similar to that in the case study).

We then generate new, simulated data sets using the Stock Synthesis parametric bootstrap simulator. For each simulation replicate, we simulate a new vector of recruitment deviations, and also simulate a new fishing mortality pattern that increases linearly from *F=0.01* in the first year (1966) to *F=0.30* in the final year (2013). The bootstrap simulator then calculates the population abundance-at-age resulting from the input vector of recruitment deviations and fishing mortality, and simulates an abundance index and age-composition samples from their specified distributions (i.e., using a lognormal distribution with the input log-standard deviation for the abundance index and a multinomial distribution with the input sample size for the age-composition samples).

The simulation experiment involves a factorial design with three simulation scenarios, four levels of an inflation factor, and three estimation models. For each combination, we ran 100 simulation replicates, for a total of total estimation model runs. We define three simulation scenarios, where the “true” sample size used to simulate age-composition data for the fishery is either 25, 100, or 400 per year. Given this “true” age-composition sample, the “observed” age-composition sample is then inflated by a fixed inflation factor *θsim*, with value . We then use estimation methods (i), (ii), and (iii) defined in the section titled *Case study: Pacfic hake* (see above).

*2.6 Simulation model evaluation*

Estimation procedures were evaluated by comparing estimated parameters and derived quantities of interest to management to their true values as defined in the operating model. Estimation error was quantified using relative error (, where and are estimated and true parameter values respectively). Results were recorded for converged models, where convergence was defined as obtaining a gradient less than 0.1.

**3. Results**

*3.1 Case study application: Pacific hake*

Comparing four alternative methods for weighting compositional data in the Pacific hake assessment (Fig. 2) shows that estimates of spawning output and fishing intensity are generally bracketed by the two naïve approaches, i.e., either treating input sample size as effective sample size (“unweighted”) or removing fishery age-composition data entirely (“no fishery ages”). In particular, removing fishery age data results in a higher estimate of average unfished spawning output and lower spawning output estimates from the mid-1980s onward, while treating input sample size as effective sample size results in strong year-class strength estimates in the early 1980s and early 2000s. By contrast, the default iterative and new Dirichlet-multinomial weighting methods result in similar estimates of spawning output, with the exception of recent years (2010 onwards) when the Dirichlet-multinomial estimator results in somewhat elevated estimates of spawning output relative to the iterative method. Similarly, the iterative and Dirichlet-multinomial estimates of fishing intensity are more similar than the other weighting methods, particularly for early years (prior to 1970).

*3.2 Simulation experiment*

Estimates of the Dirichlet-multinomial parameter are hugely different among the three scenarios and three levels of the inflation factor (Fig. 3, panel a). However, estimates of effective sample size are generally similar for all levels of the inflation factor for a given scenario (Fig. 3, panel b). In general, the estimated effective sample size closely matches the true sample size for all scenarios and levels of the inflation factor. However, we detect a small positive bias in the estimates of effective sample size when the true sample size is 400 (i.e., median effective sample size estimate is close to 450). This bias is not detected in the other scenarios (true sample size of 25 or 100).

Comparison of parameter estimates from the unweighted multinomial, iterative reweighting algorithm, and the linear parameterization of the Dirichlet-multinomial distribution shows that the iterative reweighting and Dirichlet-multinomial approaches have similar precision and accuracy when estimating natural mortality and average unfished recruitment for all levels of the inflation factor. By contrast, the unweighted model has substantially degraded estimates of natural mortality and unfished recruitment for any inflation factor other than 1. We note that the Dirichlet-multinomial algorithm has a notable fraction (20 of 100) of replicates that do not converge for the highest level of the variance inflation (*θsim*=1000), but convergence is otherwise comparable between methods. We therefore conclude that the Dirichlet-mulinomial method has similar estimation performance to the previous iterative reweighting approach.

**4. Discussion**

In this study, we implemented two parameterizations of the Dirichlet-multinomial distribution in the Stock Synthesis software that is widely used to conduct stock assessments in the U.S. and internationally. We then compared the Dirichlet-multinomial distribution with an iterative-reweighting aproach that commonly used for most US West Coast groundfish stock assessments (e.g., Thorson and Wetzel (2015)) and which has previously been supported by simulation (Stewart and Hamel, 2014). We believe that the Dirichlet-multinomial approach is superior to this iterative-reweighting approach for several reasons.

1. *Slow or inconsistent exploration of alternative models*: Iterative reweighting methods (e.g., the McAllister-Ianelli method) require fitting a stock assessment model to data, extracting residuals, estimating effective sample size estimates from this fit, and then re-estimating the model. This iterative tuning procedure either slows exploration of alternative models (due to the need for re-tuning after each model change) or causes inconsistent exploration of alternative models (where analysts neglect to re-tune for every sensitivity run, and therefore compare between runs that are not tuned in a consistent manner).
2. *Failure to account for uncertainty in data weighting*: Iterative reweighting methods provide no obvious method for propagating uncertainty about data-weighting. By contrast, the Dirichlet-multinomial approach represents data-weighting via an estimated parameter, and the uncertainty in this parameter can be captured via standard statistical methods (e.g., likelihood profiles, asymptotic confidence intervals, or Bayesian posteriors, (Magnusson et al., 2013)).
3. *Clear standards for convergence*: Iterative reweighting methods require subjective decisions regarding when to stop tuning the sample size, what order to tune multiple fleets, and how to combine data-weighting information from multiple fleets. These subjective decisions are rarely documented and different decisions by different analysts may cause substantial differences in ultimate estimates of stock status and productivity in assessments where data weighting is an important axis of uncertainty (e.g., US West Coast sablefish). By contrast, the Dirichlet-multinomial method allows for a single, unambiguous definition of convergence (i.e., via maximizing the model likelihood), which can be independently replicated by different authors and does not require further documentation. Given our observation of a 20% non-convergence of the Dirichlet-multinomial method when the sample size adjustment ratio needed to be extreme (i.e., 1000), a possible approach in the event of non-convergence could be to do one model run using the iterative reweighting approach to get an initial value for the DM adjustment parameter, then proceeding to fully estimate that parameter in a final model run.
4. *Interpretable estimates of effective sample size*: Analysts have previously suggested alternative model-based methods for estimating effective sample size. For example, an analyst might use a simple Dirichlet distribution, rather than the Dirichlet-multinomial distribution used here. However, the Dirichlet distribution can have effective sample size that ranges from 0 to infinity, i.e., it can exceed the input sample size. By contrast, the Dirichlet-multinomial distribution ensures that the effective sample size can never be greater than the input sample size.

These benefits should all facilitate the development, exploring, testing, and review of stock assessment models in real-world applications.

We propose that benefit #4 (“interpretable estimates of effective sample size”) could be used as a diagnostic for model goodness-of-fit. Specifically, if the estimate of effective sample size is substantially lower than the input sample size, we envision that the analyst could subsequently explore potential hypotheses for overdispersed compositional data. Potential causes for overdispersed compositional data include time-varying or non-parametric fishery selectivity, time-varying growth, and other common types of model misspecification. When compositional data are overdispersed, an analyst could then sequentially add flexibility in these processes by treating them as random effects (Thorson et al., 2015), and could determine which change causes the magnitude of overdispersion to decrease.

Unmodeled processes (e.g., spatial variation in fishing intensity) will also generally result in residuals for compositional data that are correlated among categories (e.g., between age-1 and age-2 samples in a given year), years (e.g., between adjacent years for age-2 individuals), sexes (between males, females, and unsexed individuals for a given age and year), and fleets (between survey and fishery compositional data for a given age and year). For example, positive correlations between ages are likely to arise whenever unmodeled processes have a similar affect on juvenile individuals. However, the Dirichlet-mulinomial distribution assumes a fixed, negative correlation between residuals, and hence is unable to generically account for these correlations (Francis, 2014; Hrafnkelsson and Stefánsson, 2004). Alternative distributions have been developed to simultaneously model correlations and overdispersion in compositional data. One example is the logistic-normal-multinomial (sometimes called the Gaussian-multinomial), which Francis (2014) proposed as a general replacement for the multinomial distribution. However, Francis (2014) only explored correlations among categories (inter-class correlation), and did not attempt to account for correlations in a given category among years or fleets. We hypothesize that residuals in compositional data are correlated among years, ages, and fleets, and suspect that the best generic strategy for accounting for complicated correlations is via mixed-effects methods (Thorson and Minto, 2015). Mixed-effect methods for model building have already been developed for time-varying selectivity and individual growth, and are increasingly feasible for age-structured population models using maximum likelihood or Bayesian estimation methods (Kristensen et al., 2014; Mäntyniemi et al., 2013; Thorson et al., 2015). We therefore recommend future research to explore whether accounting for these processes can adequately account for correlated residuals in compositional data.

As with any new method, we also encourage extensive testing using a variety of operating models and harvest control rules. Different forms of spatial structure or cohort-specific selectivity will generally result in different forms of correlation among years, categories, fleets, and sexes, and therefore will likely result in better or worse performance of the Dirichlet-multinomial distribution (given its inability to account for correlated residuals).

**5. Conclusions**

In this paper, we have shown that the Dirichlet-multinomial distribution can be used to generate model-based estimates of effective sample size for age and length-compositional data in stock assessment models. Using a real-world stock assessment for Pacific hake, we showed that the Dirichlet-multinomial distribution provides similar estimates of effective sample size to a previously used iterative reweighting approach that used a harmonic mean ratio estimator. We also provide a simulation experiment to verify that it provides approximately unbiased estimates of effective sample size given that the model is otherwise specified correctly. We conclude that the Dirichlet-multinomial distribution is a reasonable way to estimate the magnitude of overdispersion in compositional data, and recommend future research combining it with mixed-effects estimates of time-varying selectivity and individual growth to account for correlated residuals among categories, years, and fleets.

**Acknowledgements**

This publication was partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement No. ?, Contribution ?.

**References**

Coggins, L.G., Quinn, T.J., 1998. A simulation study of the effects of aging error and sample size on sustained yield estimates. Fish. Stock Assess. Models 955–975.

Crone, P.R., Sampson, D.B., 1997. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition., in: Int. Symp. on Fishery Stock Assessment Models for the 21st Century, Anchorage, Alaska, EEUU. 8Á11 Oct 1997.

Francis, R.C., 2014. Replacing the multinomial in stock assessment models: A first step. Fish. Res. 151, 70–84.

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Hrafnkelsson, B., Stefánsson, G., 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. 61, 1135–1142.

Kristensen, K., Thygesen, U.H., Andersen, K.H., Beyer, J.E., 2014. Estimating spatio-temporal dynamics of size-structured populations. Can. J. Fish. Aquat. Sci. 71, 326–336. doi:10.1139/cjfas-2013-0151

Magnusson, A., Punt, A.E., Hilborn, R., 2013. Measuring uncertainty in fisheries stock assessment: the delta method, bootstrap, and MCMC. Fish Fish. 14, 325–342.

Mäntyniemi, S., Uusitalo, L., Peltonen, H., Haapasaari, P., Kuikka, S., 2013. Integrated, age-structured, length-based stock assessment model with uncertain process variances, structural uncertainty, and environmental covariates: case of Central Baltic herring. Can. J. Fish. Aquat. Sci. 70, 1317–1326. doi:10.1139/cjfas-2012-0315

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.

McAllister, M.K., Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling: importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54, 284–300.

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

Punt, A.E., Smith, D.C., KrusicGolub, K., 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. Can. J. Fish. Aquat. Sci. 65, 1991–2005.

Shelton, A.O., Dick, E.J., Pearson, D.E., Ralston, S., Mangel, M., Walters, C., 2012. Estimating species composition and quantifying uncertainty in multispecies fisheries: hierarchical Bayesian models for stratified sampling protocols with missing data. Can. J. Fish. Aquat. Sci. 69, 231–246.

Stewart, I.J., Hamel, O.S., 2014. Bootstrapping of sample sizes for length-or age-composition data used in stock assessments. Can. J. Fish. Aquat. Sci. 71, 581–588.

Taylor, I., Grandin, C., Hicks, A.C., Taylor, N., Cox, S., 2015. Status of the Pacific Hake (whiting) stock in US and Canadian waters in 2015. Prepared by the Joint Technical Committee of the U.S. and Canada Pacific Hake/ Whiting Agreement.

Thorson, J.T., 2014. Standardizing compositional data for stock assessment. ICES J. Mar. Sci. J. Cons. 71, 1117–1128. doi:10.1093/icesjms/fst224

Thorson, J.T., Hicks, A.C., Methot, R.D., 2015. Random effect estimation of time-varying factors in Stock Synthesis. ICES J. Mar. Sci. J. Cons. 72, 178–185. doi:10.1093/icesjms/fst211

Thorson, J.T., Minto, C., 2015. Mixed effects: a unifying framework for statistical modelling in fisheries biology. ICES J. Mar. Sci. J. Cons. 72, 1245–1256. doi:10.1093/icesjms/fsu213

Walters, C.J., Martell, S.J.D., 2004. Fisheries Ecology and Management. Princeton University Press, Princeton, New Jersey.

Table 1. Parameters used to generate simulated data sets (the “operating model”) and during model fitting (the “estimation model”). A modified version of the 2015 Pacific hake assessment model with 134 estimated parameters is used as both the operating and estimation model (the model uses empirical weight-at-age techniques, and therefore does not estimate individual growth parameters). Survey and fishery selectivity values are not listed but follow the non-parametric form used in Taylor et al. (2015), but without variation over time.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Operating model* |  | *Estimation model* | |
| Name | True value |  | Estimated or fixed? | Number of estimated parameters |
| Natural mortality rate | 0.217 |  | Estimated | 1 |
| Expected recruits at unfished level (natural logarithm) | 14.470 |  | Estimated | 1 |
| Beverton-Holt steepness | 0.850 |  | Estimated | 1 |
| log-standard deviation of recruitment deviations | 0.900 |  | Fixed | - |
| Additional variance for accoustic survey index | 0.313 |  | Estimated | 1 |
| Accoustic survey selectivity at age | - |  | Estimated | 4 |
| Fishery selectivity at age | - |  | Estimated | 5 |
| Recruitment deviations | - |  | Estimated | 72 |
| Instantaneous fishing mortality rates | - |  | Estimated | 49 |

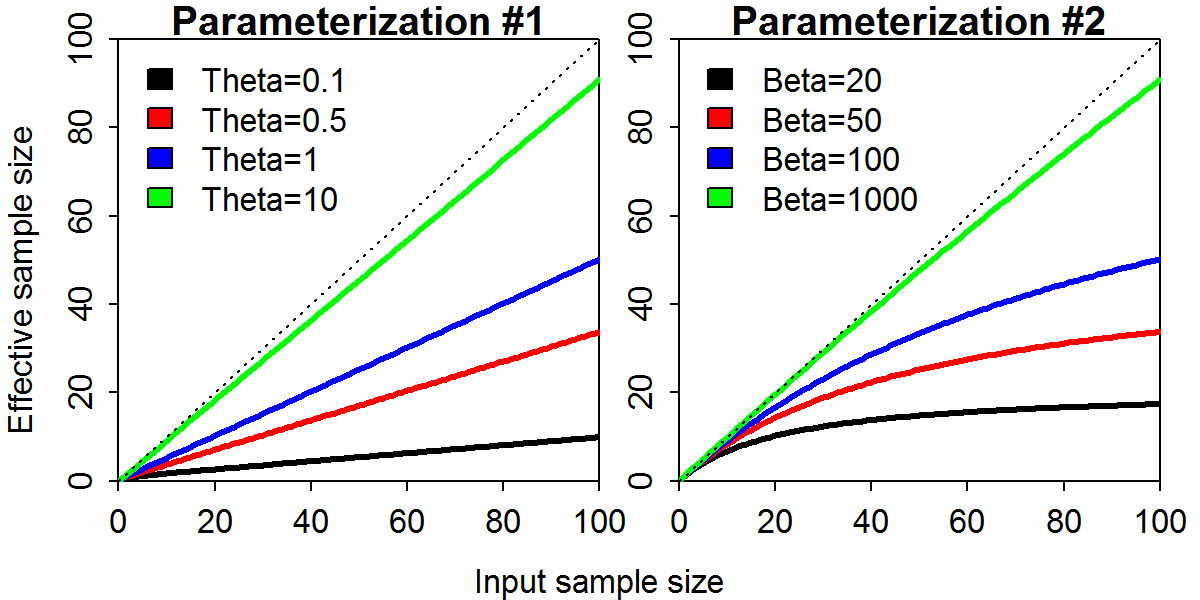
****

Fig. 1. Input sample size (x-axis) and effective sample size (; y-axis) for two paramaterizations of the Dirichlet-multinomial (DM) distribution across varying values for the DM parameter specific to each parameterization. The dashed line represents the 1:1 line where input sample size is the same as .

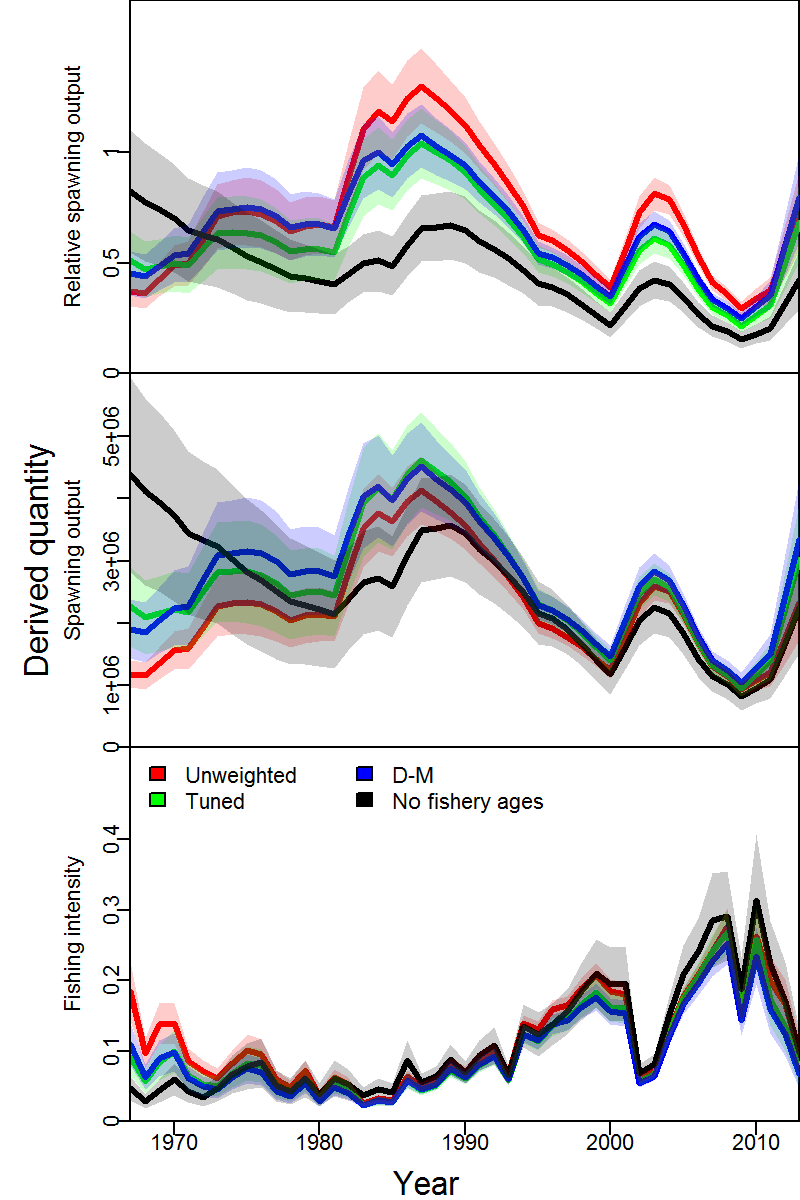


Fig. 2. Comparison of spawning output relative to average unfished levels (top), spawning output (SPB; middle), and exploitation fraction (catch divided by estimated biomass for individuals aged 3 and older; bottom) for the Pacific hake assessment given four alternative methods of weighting the age-composition data: (i) unweighted (red), (ii) tuned (green); (iii) Dirichlet-multinomial distribution (blue); and (iv) weight of zero for the age-composition data (black), where for each model we show the maximum likelihood estimates (solid line) and +/- 1 standard error (shaded region).

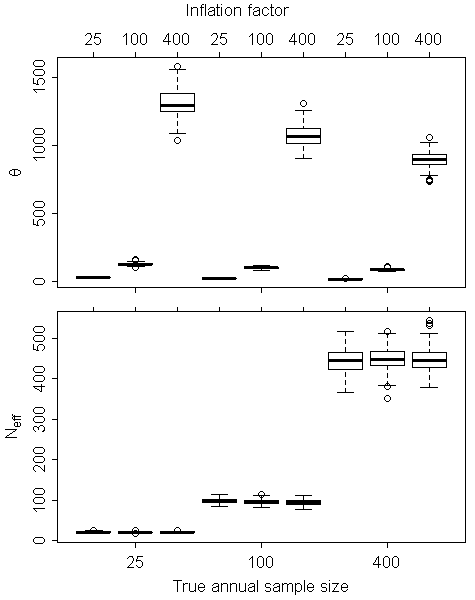


Fig. 3. Estimated Dirichlet-multinomial variance inflation parameter (top panel) and effective sample size (, bottom panel) from the “linear” parameterization (parameterization #1) of the Dirichlet-Multinomial distribution implemented in Stock Synthesis shown for three “true sample sizes” (representing an increase in available information from 25, 100, to 400 true samples per year) and three levels of variance inflation (wherein the input sample size provided to Stock Synthesis is 25, 100, or 1000 the true sample size).

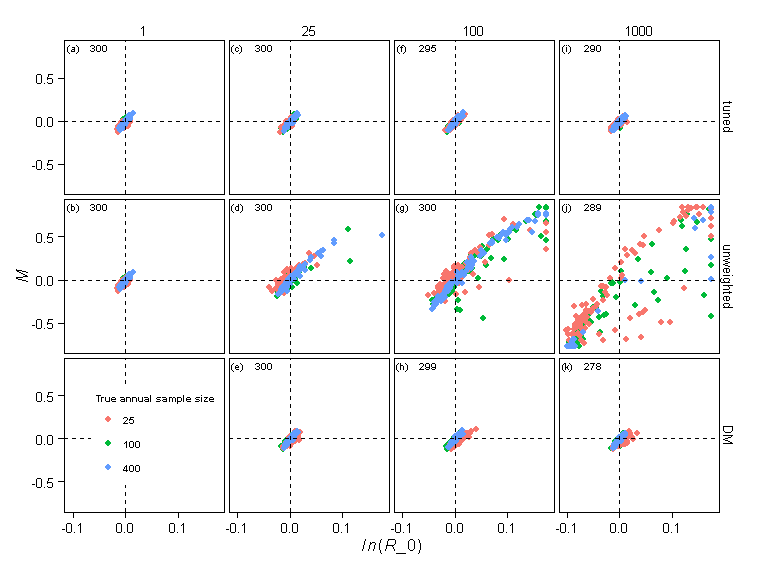


Fig. 4. Relative error in parameter estimates across estimation methods (rows; ” tuned”: using the ratio estimator of the harmonic mean to input sample size; “unweighted”: conventional multinomial treating input as effective sample size; “DM”: linear-parameterization of the Dirichlet-multinomial distribution) and levels of the inflation factor for the fishery age-composition data in the operating model (columns). Each panel depicts the maximum likelihood estimates of natural mortality rate (*M,* y-axis) and average unfished recruitment (*ln(R0)*, x-axis), where colors are used to distinguish estimates. We only show results for estimation models where the maximum final gradient was <0.1 (the number of replicates across models is indicated in each panel, where 300 implies that all 100 replicates converged for each of three estimation models), and confirm that results are qualitatively similar if using a different convergence threshold. The lower left panel is not plotted because the DM estimation method was not used when the inflation factor was one.