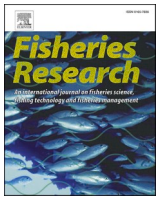
Fisheries Research 257 (2023) 106487

Contents lists available at ScienceDirect

Fisheries Research

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



Model validation for compositional data in stock assessment models: Calculating residuals with correct properties

Vanessa Trijoulet [a,*], Christoffer Moesgaard Albertsen [a], Kasper Kristensen [a], Christopher M. Legault [b], Timothy J. Miller [b], Anders Nielsen [a]

[a] National Institute of Aquatic Resources, Technical University of Denmark, Kemitorvet 201, DK-2800 Kgs. Lyngby, Denmark
[b] Northeast Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 166 Water Street, Woods Hole, MA 02543, USA

ARTICLE INFO

Keywords:
compResidual R-package
Multivariate distributions
One-step-ahead quantile residuals
Pearson
Template Model Builder

ABSTRACT

Stock assessment models are often used to inform fisheries management and need therefore to be thoroughly validated. Different diagnostics exist to validate models including the analysis of standardized residuals. Standardized residuals are commonly calculated by subtracting prediction from the observation and dividing the result with the estimated standard deviation (i.e., Pearson residuals). Many currently applied stock assessment models fit to compositional observations (e.g., age, length or stock compositions) using multivariate distributions. These distributions create correlation between observations, which are propagated in the residuals if estimated as Pearson. This study shows that using Pearson residuals to analyze goodness of the fit, when data are fitted using a multivariate distribution, is incorrect and one-step-ahead (OSA) or forecast quantile residuals should be used instead. For such distributions, OSA residuals are independent and standard normally distributed for correctly specified models. This study describes the calculation of OSA residuals specifically to de-correlate compositional observations for the multivariate distributions most commonly used in assessment models. This allows composition observations to be evaluated with the same statistical rigor as residuals from uncorrelated observations. This also prevents the possible wrong interpretation of Pearson residuals and the rejection of a correct model. We have developed an R-package that estimates OSA residuals externally to the model for models that do not include random processes. For models that use random processes, the distributions are now developed in Template Model Builder and explained in detail here for internal use.

1. Introduction

Stock assessment models use observations such as commercial catches or abundance indices from scientific surveys to estimate stock size and infer current and historic stock status and level of exploitation. Fisheries scientists use these models to predict future consequences of different fishing options, which form the basis of the scientific advice used to inform fisheries managers. It is therefore necessary to carefully evaluate the quality of these models before they are used for fisheries management. Some diagnostics such as retrospective analysis (Mohn, 1999) are specific to stock assessment models and put the focus on the consistency of the most recent estimates, which are of special interest for short-term projections. However, most used techniques are general validation techniques that could be applied to any statistical model. Notably, goodness of fit validation is routinely conducted by visually inspecting standardized residuals.

For univariate statistical models, where the observation noise is assumed independent and normally distributed, standardized residual can be calculated by subtracting prediction from the observation and dividing the result with the estimated standard deviation. These are often referred to as Pearson residuals. If the independent and univariate model is correct this will approximately lead to residuals which follow a standard normal distribution. Plotting the residuals and checking if they are normally distributed is therefore useful for validating that a model is describing the observed data well.

However, many fish stock assessment models are age- or length-structured (e.g., the State-space Assessment Model (SAM, Nielsen and Berg, 2014); Stock Synthesis (SS3, Methot and Wetzel, 2013);

* Corresponding author.
E-mail addresses: vtri@aqua.dtu.dk (V. Trijoulet), cmoe@aqua.dtu.dk (C.M. Albertsen), kaskr@aqua.dtu.dk (K. Kristensen), chris.legault@noaa.gov (C.M. Legault), timothy.j.miller@noaa.gov (T.J. Miller), an@aqua.dtu.dk (A. Nielsen).

https://doi.org/10.1016/j.fishres.2022.106487
Received 18 July 2022; Received in revised form 30 August 2022; Accepted 30 August 2022

MULTIFAN-CL (Fournier et al., 1998); the C++ Algorithmic Stock Assessment Laboratory (CASAL, Bull et al., 2005); the Beaufort Assessment model (BAM, Williams and Shertzer, 2015)); and the Age-Structure Assessment Program (ASAP, Legault and Restrepo, 1999)). These models therefore use age and/or length composition data to disaggregate observations such as total survey index, total catch, or total stomach content in the case of multispecies models (Lewy and Vinther, 2004; Begley, 2005; Trijoulet et al., 2019, 2020). In these cases, multivariate rather than univariate distributions are needed to fit to composition data and these do not assume the individual observations to be independent. Common multivariate distributions used in stock assessment models are multinomial, Dirichlet, Dirichlet-multinomial, or logistic-normal distributions (Francis, 2014; Albertsen et al., 2017). The correlation between the individual observations introduced by the multivariate distributions will be retained in the Pearson residuals and hence the Pearson residuals could appear as not independent and not normally distributed despite the fact that the data actually perfectly followed the assumed distribution and model structure (Punt et al., 2020). In this case a different way to calculate standardized residuals is needed and an option is the one-step-ahead (OSA) or forecast quantile residuals (Thygesen et al., 2017).

In this study, we describe and adapt the concept of OSA residuals specifically to de-correlate compositional observations when estimating standardized residuals and we validate the method through simulations. The method presented in this paper only concerns the calculation of the residuals for model validation, the model fitting and model results being independent from the residual calculation are therefore unaffected. The method is then applied to a fish stock, Gulf of Maine haddock. We show that using Pearson residuals to analyze goodness of the fit, when data are fitted using a multivariate distribution, is incorrect and that OSA residuals should be used instead. For fully-parametric assessment models (without random effects), we have developed an R-package to estimate OSA residuals outside of the assessment model (https://github.com/fish follower/compResidual). For state-space models using Template Model Builder (TMB, Kristensen et al., 2016), we have implemented the OSA residual estimation for the distributions described in this paper in TMB for internal use (see https://github.com/kaskr/adcomp/wiki/User-contributed-code or the direct link https://github.com/vtrijoulet/OSA_multivariate_dists). For state-space models not using TMB, this paper provides a description on how the OSA residuals can be implemented internally.

2. Material and methods

2.1. OSA residuals for compositional data: concept, examples and validation

The OSA residuals are implemented as part of TMB (see function "oneStepPredict"). To apply this concept to the multivariate distributions commonly used for fisheries composition data, it is necessary to change the calculations of these likelihood contributions to be done successively one scalar observation at a time. In this section, the concept is summarized and the necessary changes are derived.

2.1.1. OSA residuals concept

In this section, we describe how to obtain standardized residuals with correct statistical properties for use in model validation for the different types of observations that can be encountered.

OSA or Pearson residuals for continuous normal independent observations. When observations $(x_1, \cdots, x_K)$ can be assumed to be independent and Gaussian, correct standardized residuals can be calculated using Pearson or the OSA methods (described below). Pearson residuals $(r_1, \cdots, r_K)$ are defined by $r_i = \frac{(x_i - \widehat{\mu}_i)}{\widehat{\sigma}}$, for $i \in \{1, \cdots, K\}$, where $\widehat{\mu}$ is the prediction and $\widehat{\sigma}$ the standard deviation of the observations.

In this case, the OSA and the Pearson residuals will be the same and will have the correct properties, i.e., independent normally distributed.

However, when observations are not normally distributed or independent (e.g., follow a multivariate distribution), Pearson residuals will not have the correct properties and it is possible to obtain independent residuals for some multivariate distributions by calculating the OSA residuals. This concept originates from the Rosenblatt transformation, which states that any multivariate continuous distribution can be transformed to a uniform distribution on the hypercube if it can be written as successive conditional distributions (Rosenblatt, 1952). It was specifically described for residuals in Thygesen et al. (2017).

Quantile residuals for continuous non-normal independent observations. If the observations $(x_1, \cdots, x_K)$ are continuous, univariate, independent, and originating from a distribution with cumulative distribution function (cdf) F_x, which does not need to be a normal distribution, then independent and normally distributed residuals can be obtained via transformation. First, transforming the observations via the cdf will lead to quantities that follow a uniform distribution $u_i = F_x(x_i)$. This can be explained by the fact that $u_i \in (0, 1)$ and the cdf of u is $F_u(u) = P(F_x(X) < u) = P(X < F_x^{-1}(u)) = F_x(F_x^{-1}(u)) = u$, which is the distribution function for the uniform distribution. Secondly, transforming these uniformly distributed quantities $u_1, \cdots, u_K$ by the inverse cdf of the standard normal distribution, Φ^{-1}, will lead to residuals that follow a standard normal distribution if the model is correct. This can be seen by calculating the cdf for the transformed uniforms: $P(\Phi^{-1}(U) < r) = P(U < \Phi(r)) = \Phi(r)$, so the wanted cdf. Collectively these quantile residuals are defined simply as: $r_i = \Phi^{-1}(F_x(x_i))$. The model is defining the cdf of the observations, so if the model is incorrect, then the residuals will deviate systematically from a standard normal distribution (Dunn and Smyth, 1996).

Randomized quantile residuals for discrete independent observations. If the observations are discrete, but still univariate and independent, then the distribution function F_x is a step function. In this discrete case, the transformation by the distribution function needs an additional step. The probability mass at a given value x_i needs to be transformed onto the interval from $F(x_i - \epsilon)$ to $F(x_i)$. The transformation into continuous uniform(0,1) distributed quantities can be achieved by sampling from the uniform distribution for that interval, so $u_i \sim U(F(x_i - \epsilon), F(x_i))$. The final step to get standard normal residuals is the same transformation by the inverse cdf of the standard normal distribution $r_i = \Phi^{-1}(u_i)$. These randomized quantile residuals will again have the desired properties (independence and normality) if the model is correct (Dunn and Smyth, 1996).

OSA residuals to remove dependence in dependent observations. The OSA residual of the i'th scalar observation is computed using either of the quantile residual methods described above depending on the observations being continuous or discrete, but instead of using the cdf of the observation in isolation, the cdf of the predicted distribution of the i'th prediction conditioned on all previous observations is used. This allows the resulting residuals to become independent standard normal if the model is correct.

2.1.2. OSA residuals for distributions used for composition data

All of the multivariate distributions we consider here have a constraint on the sum of the vector elements. For the multinomial and Dirichlet-multinomial, the vector sums to N, the sample size. For the Dirichlet and logistic-normal the vector of proportions sums to unity. Therefore, for a K-dimensional observation vector there are only $K - 1$ OSA residuals that can be independent and standard normal. In this study, the K'th residual is therefore not presented.

Multinomial distribution. If a vector of numbers at age or at length in a

specific year $(x_1, \cdots, x_K)$ follows a discrete multinomial distribution such as $(x_1, \cdots, x_K) \sim Multi(N, (p_1, \cdots, p_K))$, with $N = \sum_i x_i$, $i \in \{1, \cdots, K\}$, K number of categories (e.g., age groups) and probability p_i, it is equivalent to express the distribution as successive binomials ($Bin(N, p)$ where N is the number of trials and p is the probability of success in each trial), as follows:

$$\begin{aligned}
x_1 &\sim Bin(N, p_1) \\
x_2|x_1 &\sim Bin\left(N - x_1, \frac{p_2}{1-p_1}\right) \\
x_3|x_{1:2} &\sim Bin\left(N - (x_1+x_2), \frac{p_3}{1-(p_1+p_2)}\right) \\
&\vdots \\
x_k|x_{1:(k-1)} &\sim Bin\left(N - \sum_{i=1}^{k-1} x_i, \frac{p_k}{1-\sum_{i=1}^{k-1} p_i}\right) \\
&\vdots \\
x_{K-1}|x_{1:(K-2)} &\sim Bin\left(N - \sum_{i=1}^{K-2} x_i, \frac{p_{K-1}}{1-\sum_{i=1}^{K-2} p_i}\right)
\end{aligned}$$

Note, for instance, that the notation $x_{1:(k-1)}$ refers to $x_1, x_2, \cdots, x_{k-1}$.

Above we see that the likelihood of the multinomial can be evaluated as a product of the successive binomials (Gelman et al., 1995). Expressing the distribution via the successive binomials allows easy access to the predictive distributions needed for estimation of the OSA residuals as explained in 2.1.1.

Dirichlet distribution. If a vector of age or length proportions in a specific year $(x_1, \cdots, x_K)$, with $\sum_i x_i = 1$ for $i \in \{1, \cdots, K\}$, follows a continuous Dirichlet distribution such as $(x_1, \cdots, x_K) \sim D((\alpha_1, \cdots, \alpha_K))$, with K number of categories (e.g., age groups) and concentration parameters α_i, it is equivalent to express the distribution as successive beta distributions ($Beta(\alpha, \beta)$ where α and β are shape parameters), as follows:

$$\begin{aligned}
x_1 &\sim Beta\left(\alpha_1, \sum_{i=2}^{K} \alpha_i\right) \\
\frac{x_2}{1-x_1}\Big|x_1 &\sim Beta\left(\alpha_2, \sum_{i=3}^{K} \alpha_i\right) \\
\frac{x_3}{1-(x_1+x_2)}\Big|x_{1:2} &\sim Beta\left(\alpha_3, \sum_{i=4}^{K} \alpha_i\right) \\
&\vdots \\
\frac{x_k}{1-\sum_{i=1}^{k-1} x_i}\Big|x_{1:(k-1)} &\sim Beta\left(\alpha_k, \sum_{i=k+1}^{K} \alpha_i\right) \\
&\vdots \\
\frac{x_{K-1}}{1-\sum_{i=1}^{K-2} x_i}\Big|x_{1:(K-2)} &\sim Beta(\alpha_{K-1}, \alpha_K)
\end{aligned}$$

The density is Jacobi transformed given that, for $K > 1$, it is not the observation x_i, but the ratio $\frac{x_i}{1-\sum_{j=1}^{i-1} x_j}$ that follows a Beta distribution.

The likelihood of the Dirichlet can therefore be evaluated as a product of the successive beta distributions (Gelman et al., 1995). Expressing the distribution via the successive beta distributions allows easy access to the predictive distributions needed for estimation of the OSA residuals as explained in 2.1.1.

The Dirichlet distribution can also be transformed into independent gamma distributions by using a random gamma distributed multiplier. While this has the advantage of resulting in an estimated residual for all composition groups, we chose not to do that to avoid using a random sample that reduces the power to detect patterns in the residuals, and to keep consistency with the other distributions for which it was not possible to do such transformation.

Dirichlet-multinomial distribution. A vector of numbers at age or at length in a specific year $(x_1, \cdots, x_K)$ follows a Dirichlet-multinomial distribution if $(x_1, \cdots, x_K) \sim Multi(N, (p_1, \cdots, p_K))$, with $N = \sum_i x_i$, $i \in \{1, \cdots, K\}$, K number of categories (e.g., age groups), and the probabilities follow a continuous Dirichlet distribution such as $(p_1, \cdots, p_K) \sim D((\alpha_1, \cdots, \alpha_K))$ with concentration parameters α_i. It is equivalent to express the distribution of observations as successive beta-binomial distributions ($BB(N, \alpha, \beta)$ where N is the number of binomial trials while α and β are beta distribution shape parameters determining the binomial success probability), as follows:

$$\begin{aligned}
x_1 &\sim BB\left(N, \alpha_1, \sum_{i=2}^{K} \alpha_i\right) \\
x_2|x_1 &\sim BB\left(N - x_1, \alpha_2, \sum_{i=3}^{K} \alpha_i\right) \\
x_3|x_{1:2} &\sim BB\left(N - (x_1+x_2), \alpha_3, \sum_{i=4}^{K} \alpha_i\right) \\
&\vdots \\
x_k|x_{1:(k-1)} &\sim BB\left(N - \sum_{i=1}^{k-1} x_i, \alpha_k, \sum_{i=k+1}^{K} \alpha_i\right) \\
&\vdots \\
x_{K-1}|x_{1:(K-2)} &\sim BB\left(N - \sum_{i=1}^{K-2} x_i, \alpha_{K-1}, \alpha_K\right)
\end{aligned}$$

The likelihood of the Dirichlet-multinomial can therefore be evaluated as a product of the successive beta-binomial distributions. Expressing the distribution via the successive beta-binomial distributions allows easy access to the predictive distributions needed for estimation of the OSA residuals as explained in 2.1.1.

Logistic-normal distribution. Contrarily to the previous distributions, it is, to our knowledge, not possible to write down the logistic-normal distribution as successive conditional distributions on closed form. However, the observations can easily be transformed to multivariate normally distributed observations of one dimension less. Maximum likelihood estimation of parameters are equivalent for the transformed variables under the multivariate normal model, but the OSA residuals are not invariant under transformation of the observations. However, the OSA residuals for the transformed observations provide the same diagnostic information as those that would be derived from the untransformed observations.

A vector of age or length proportions $(x_1, \cdots, x_K)$ in a specific year follows a logistic-normal distribution (Aitchison and Shen, 1980) if we can define $(\alpha_1, \cdots, \alpha_{K-1}) \sim N(\mu, \Sigma)$, with mean μ, covariance Σ, and $K-1$ number of categories (e.g., age groups), such as $x = \text{Logistic}^{-1}(\alpha)$, where Logistic^{-1} is either the additive

$$x = \left(\frac{e^{\alpha_1}}{1+\sum_{i=1}^{K-1} e^{\alpha_i}}, \cdots, \frac{e^{\alpha_k}}{1+\sum_{i=1}^{K-1} e^{\alpha_i}}, \cdots, \frac{e^{\alpha_{K-1}}}{1+\sum_{i=1}^{K-1} e^{\alpha_i}}, \frac{1}{1+\sum_{i=1}^{K-1} e^{\alpha_i}}\right)'$$

or multiplicative

$$x = \left(\frac{e^{\alpha_1}}{(1+e^{\alpha_1})}, \cdots, \frac{e^{\alpha_k}}{\prod_{i=1}^{k}[1+e^{\alpha_i}]}, \cdots, \frac{e^{\alpha_{K-1}}}{\prod_{i=1}^{K-1}[1+e^{\alpha_i}]}, \frac{1}{\prod_{i=1}^{K-1}[1+e^{\alpha_i}]}\right)'$$

transformation (Aitchison, 2003). To calculate the residuals for observations following a logistic-normal, the observations are transformed, via the assumed model, back to the α, which is then multivariate normally distributed if the assumed model is correct. OSA residuals for the multivariate normal distribution are then easily calculated as presented in Thygesen et al. (2017) and are already implemented in TMB, so these will

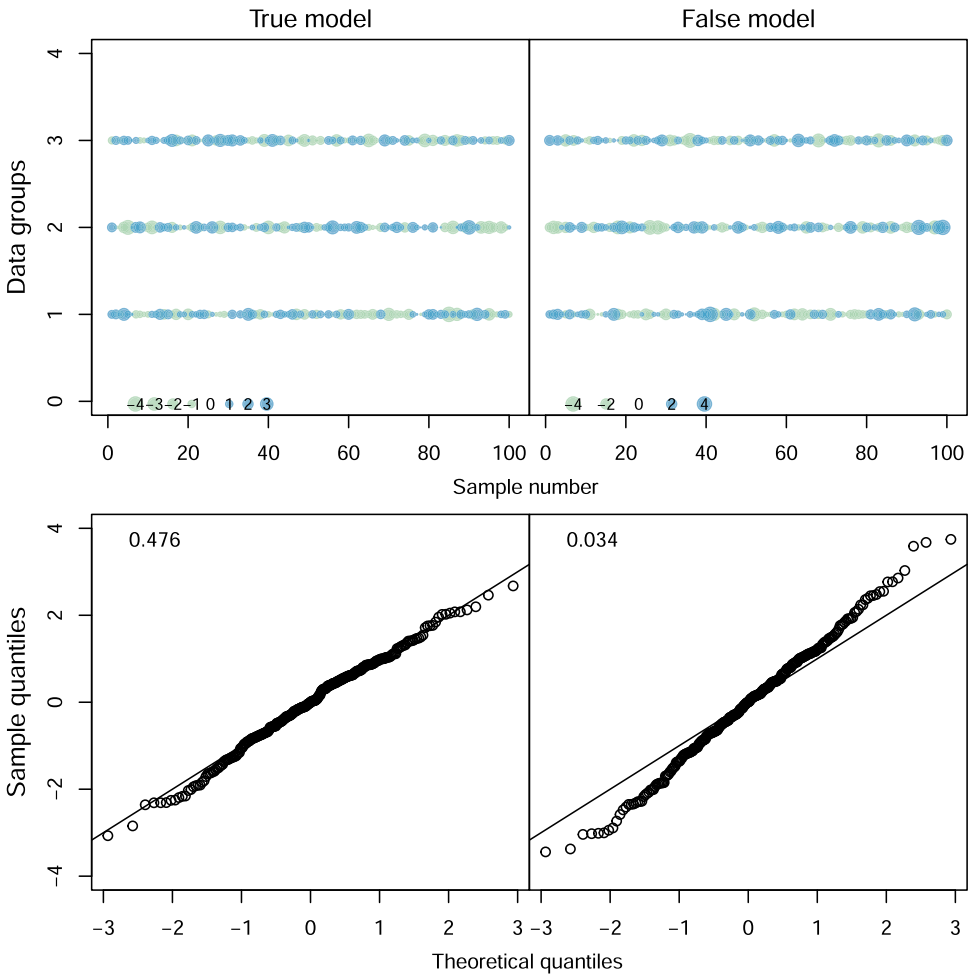


Fig. 1. Residual bubble plots (top) and normal Q-Q plots (bottom) for the simple multinomial validation example. Each bubble is scaled to the residual's size and colored given its sign (positive in blue or negative in green). The p-value of the Kolmogorov-Smirnov test for normality is given in the top left of the Q-Q plots and the line is the identity line. If the observations were for instance age compositions, the data groups would be ages and the sample number would be years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

not be discussed further in the paper. An important practical note is that in the current TMB interface the transformation from observed proportions x to $\alpha = \text{Logistic}(x)$ needs to be computed in R before the observations are passed to the compiled code. The additive logistic transformation is $\alpha = \log((x_1/x_K, \ldots, x_k/x_K, \ldots, x_{K-1}/x_K)')$ and the multiplicative logistic transformation is $\alpha = \log((x_1/(1-x_1), \ldots, x_k/(1-\sum_{i=1}^{k} x_i), \ldots, x_{K-1}/x_K)')$.

2.1.3. OSA residual validation

To validate the concept of OSA residuals for compositional observations, examples were developed where observations were simulated under both "true" (simulated from the distribution) and "false" (simulated with model violation) models for the multinomial, Dirichlet, and Dirichlet-multinomial distributions. For all distributions, 4 compositional groups were simulated for 100 simulations. In the case of age composition data, this would correspond to 4 age groups for 100 years. Observations under the true model were simulated directly from the respective true distribution. Regarding the false model, there are several ways to simulate observations which do not correspond to the prescribed distribution; we chose to simulate model violations that could realistically occur for compositions in assessment models. For the multinomial distribution, the observations under the false model were simulated by adding random noise around the true probabilities (where the model assumed constant probabilities), which would correspond to overdispersion in the observed compositions. For the remaining distributions, the model violations were constructed to mimic the situation where the selectivity was gradually changing over time (where the model assumes constant selectivity). The selectivity was arbitrarily chosen to change over a 100 year period from (0.02, 0.13, 0.25, 0.60) to (0.13, 0.02, 0.25, 0.60) for compositional group 1–4 respectively.

OSA residuals were calculated to validate if the residuals are correctly identified as independent standardized normal when the observations are from the true models and inversely when they are from the false models. To check that the OSA residuals are always correct, we replicated the process 1000 times for each model and distribution.

2.2. Application to Gulf of Maine haddock

Gulf of Maine haddock is assessed via ASAP (Anon, 2019), a statistical catch-at-age model (Legault and Restrepo, 1999), which assumes that the age composition (age proportions in the total catch multiplied by the effective sample size, ESS) follows a multinomial distribution with known ESS. The ESS is used instead of the true sample size to mimic overdispersion in the multinomial distribution. Common goodness-of-fit diagnostics used currently in ASAP are Pearson residuals. Three types of age composition observations are considered for this stock: observations for one commercial fleet (fleet 1) and two independent surveys (spring and autumn, fleets 2 and 3). The data is disaggregated into 9 ages (1–9+) and cover the period 1977–2016 for the three fleets (commercial and surveys).

2.2.1. Simulated observations

To compare OSA and Pearson residuals in a "true" case where we know the observations follow a multinomial distribution, the multinomial assumption is used in ASAP to simulate observations for the three fleets. Five replicates of observations are simulated, each covering the

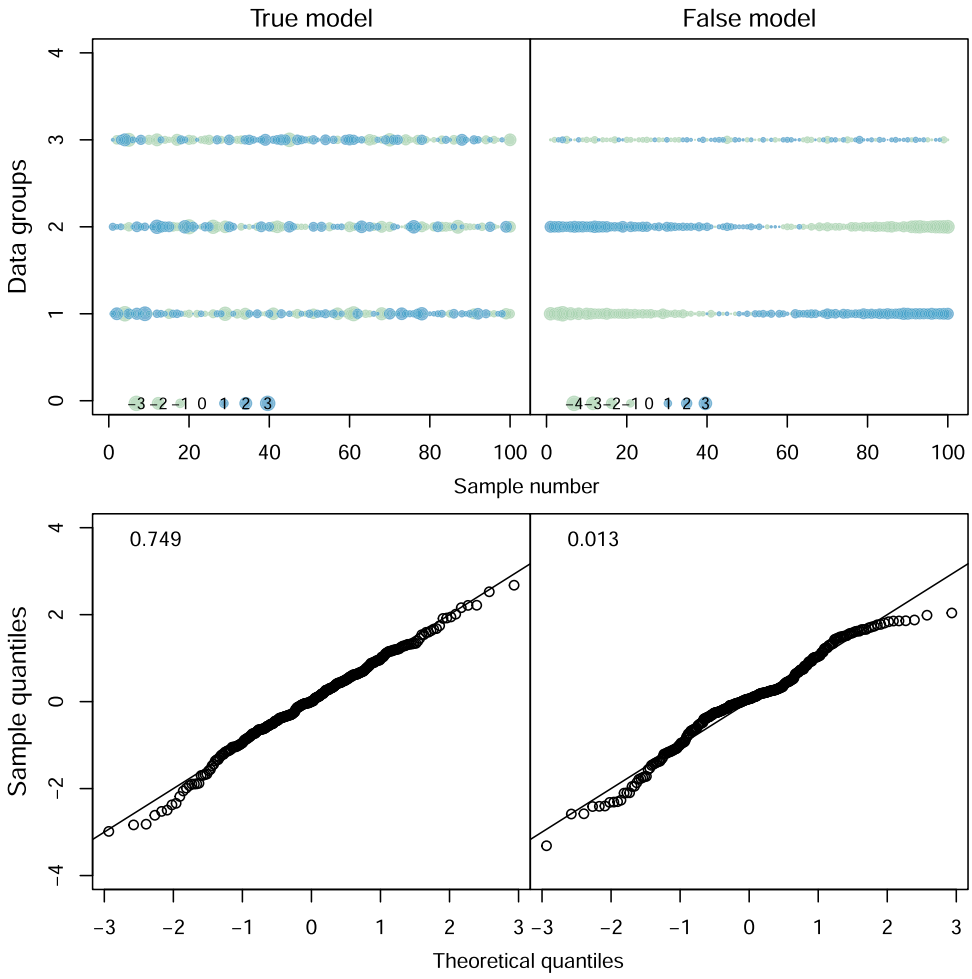


Fig. 2. Residual bubble plots (top) and normal Q-Q plots (bottom) for the simple Dirichlet validation example. See caption of Figure 1 for a detailed description of the figure.

period 1977–2016. For simplicity, the same annual ESS is assumed for a specific fleet as follows: 100 for the commercial fleet and 50 for both survey fleets. Here, we chose to use only 5 replicates per fleet because the goal is not to validate the residuals such as the validation examples above but to be sure the results are not just obtained at random. The simulated data is then used within ASAP to estimate predicted age composition assuming a multinomial distribution and both Pearson and OSA residuals are calculated outside of the model. To keep consistency with the OSA residuals, Pearson residuals for the last age group are not presented.

2.2.2. True observations

The same method is then used on true age composition data for the stock and again both Pearson and OSA residuals are estimated. The ESS for each fleet differ along the time series following the assumptions made by the assessment group in Anon (2017). The observed numbers were rounded to integers to allow multinomial residuals to be estimated. While it is not possible to know if the multinomial model is correct for true data, this might identify cases where Pearson residuals would lead to a different conclusion than the OSA residuals.

2.3. Diagnostics

For all residuals, residuals bubble plots are presented such that each bubble is scaled to the residual's size and colored given its sign (positive in blue or negative in green). To test if the residuals are normal, quantile-quantile (Q-Q) residual plots are provided as well as the p-value of the Kolmogorov-Smirnov test for standard normality. A significant p-value of ≤ 0.05 means the hypothesis of normality is rejected and the residuals are not considered normally distributed. The Kolmogorov-Smirnov test is chosen here because it can check for standard normality with mean 0 and standard deviation 1.

For the OSA residual validation examples, additional checks are performed and presented in Appendix A1. The mean and standard deviation, the correlation between the different compositional groups and the lag 1 year correlation of the OSA residuals are calculated for all iterations. In addition to the Kolmogorov-Smirnov test, two extra tests for normality are also performed: Shapiro-Wilks and Anderson-Darling. The p-values of the different statistical tests (for normality and correlation) are compared. According to the statistical theory, when the distribution of the OSA residuals is correct under the true model, the rejection probability (p-value being ≤ 0.05) should be 0.05. In contrast, the power of the statistical tests is given by the rejection probability under the false model, so that, the closer to 1 the probability of rejection is, the higher power the test has.

3. Results

3.1. OSA residual validation

For the multinomial and Dirichlet distributions, the OSA residuals are correctly identified as normally distributed (Kolmogorov-Smirnov test for normality has p-value > 0.05) under the true model and inversely (p-value ≤ 0.05) under the false model (Table A1.1 in

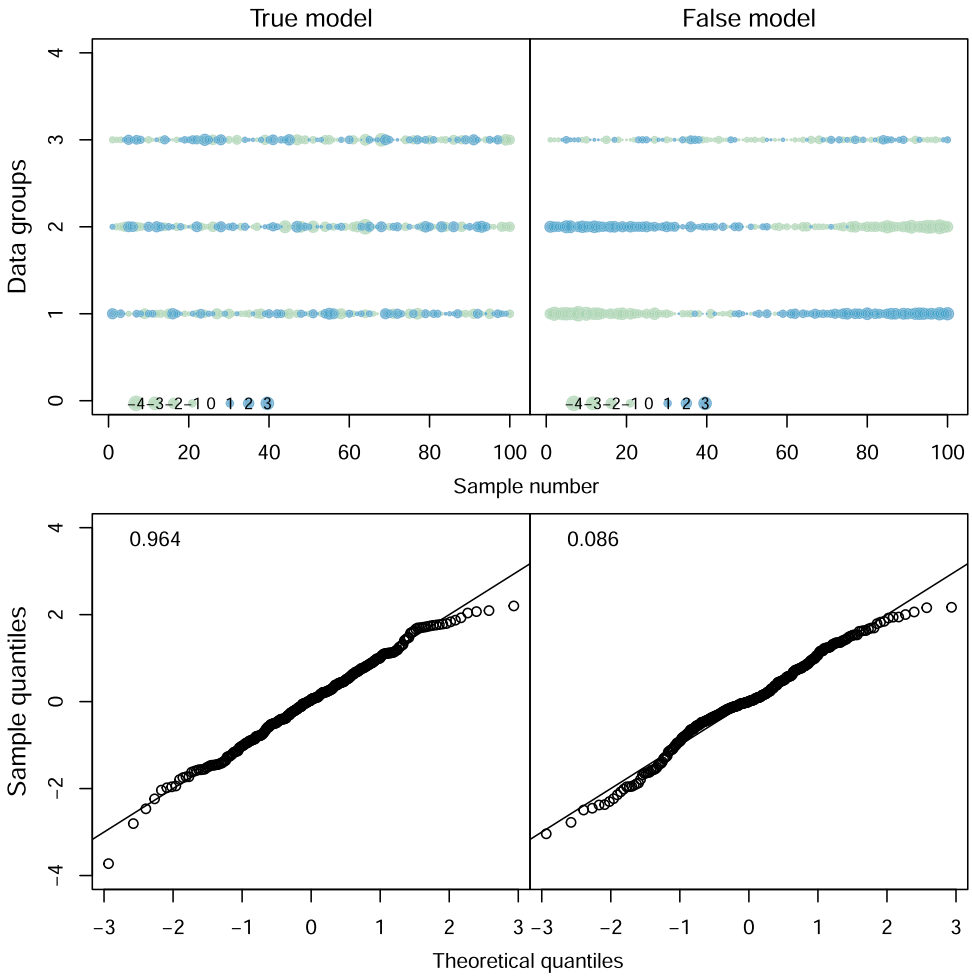


Fig. 3. Residual bubble plots (top) and normal Q-Q plots (bottom) for the simple Dirichlet-multinomial validation example. See caption of Figure 1 for a detailed description of the figure.

Appendix A1). However, the model violation is not obvious from the bubble plots of the residuals under the false model in the multinomial example (Figure 1). This is due to the fact that over-dispersion was the only model violation in this example and this mainly affects the variance of the residuals (Figure A1.1).

For the Dirichlet and Dirichlet-multinomial cases (Figures 2-3), where a change in selectivity for the first two data groups was simulated under the false model, the bubble plots clearly illustrate the bad residuals as not independent. For the first data group, large negative residuals are observed for the first half of the time series followed by large positive residuals for the other half of the time series, and inversely for the second data group. However, the statistical test for normality could not be rejected for the Dirichlet-multinomial, because the power of the Kolmogorov-Smirnov test is low due to the cumulative distribution functions between the true and false model being not distant enough (more details in Appendix A1).

The additional checks across 1000 iterations confirm the conclusions above (Appendix A1). The OSA residuals are independent (Figures A1.4-A1.9) and normally distributed with mean 0 and standard deviation 1 under the true model (Figures A1.1-A1.3). Under the false model, the OSA residuals are clearly identified as non-independent (Figures A1.4-A1.9) and non standard normally distributed (Figures A1.1-A1.3). The power of the statistical tests to identify incorrect residuals (non-normal or correlated) under the false model depends on the simulated violations and this is explained in details in Appendix A1.

3.2. Gulf of Maine haddock, simulated observations

The residuals estimated from the simulated observation example informed by the Gulf of Maine haddock assessment for the first replicate of fleet 2 appear different between Pearson and OSA residuals (Figure 4). While it is difficult to detect any clear residual patterns in the bubble plots, the OSA residuals appear independent and the magnitude of the Pearson residuals is larger. The test for normality is rejected for the Pearson residuals, but not for the OSA residuals.

The results for all the other fleets and replicates are given in Figures A2.1-A2.14. In the case of the OSA residuals, the test for normality is not rejected for any of the fleets and replicates, giving no indication that the residuals would not be normally distributed. On the contrary, Pearson residuals are rejected to be normally distributed in 7 of the 15 cases. For some of the Pearson residuals, the distribution is clearly skewed and the residuals are overall larger than the OSA residuals.

3.3. Gulf of Maine haddock, true observations

The residuals estimated from the true observations example for Gulf of Maine haddock for fleet 2 are given in Figure 5 and in Figures A3.1 and A3.2 for the other 2 fleets. The Kolmogorov-Smirnov test is not rejected for any of the fleets in the case of the OSA residuals. The Pearson residuals fail the test for normality in all cases. The residuals appear also different between the OSA and Pearson with some age effect for the Pearson residuals, while the OSA residuals appear independent.

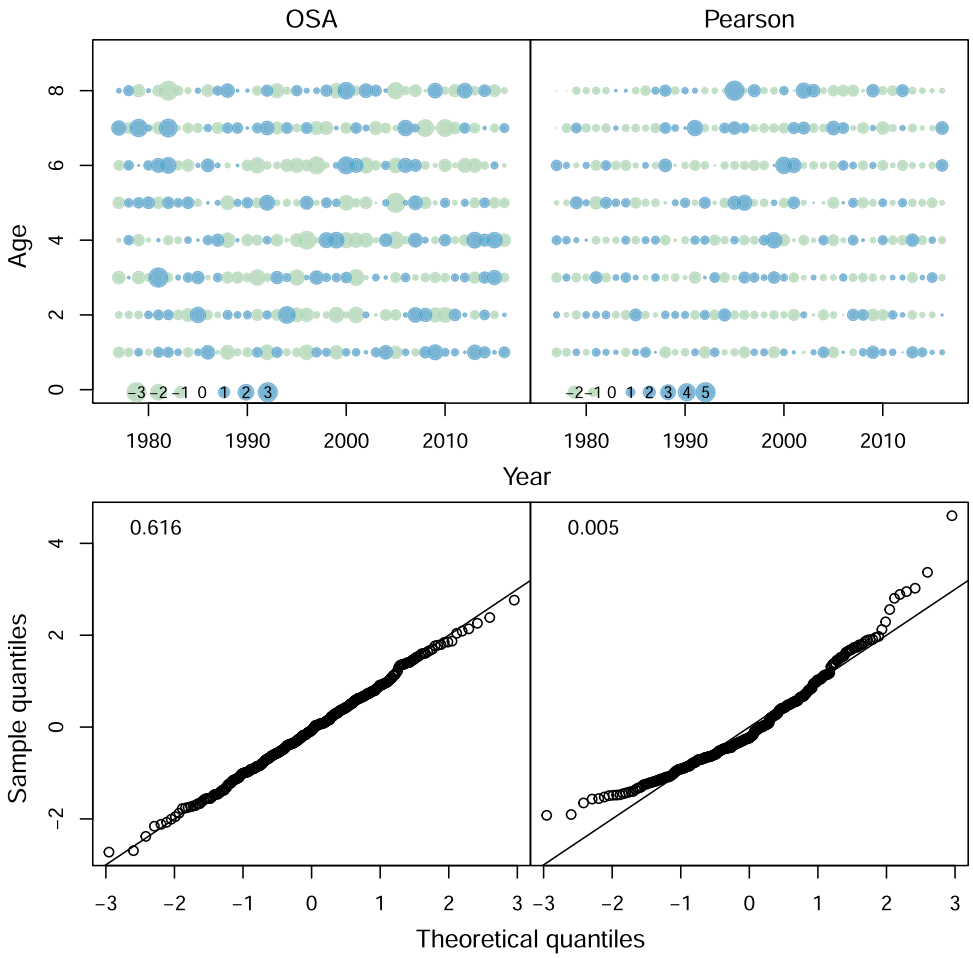


Fig. 4. Residual bubble plots (top) and normal Q-Q plots (bottom) for the 1st replicate of fleet 2 for the Gulf of Maine haddock simulated example. Each bubble is scaled to the residual's size and colored given its sign (positive in blue or negative in green). The p-value of the Kolmogorov-Smirnov test for normality is given in the top left of the Q-Q plots and the line is the identity line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4. Discussion

This paper presents the steps required to obtain OSA standardized residuals for composition observations. The OSA standardized residuals are independent (i), normally distributed (ii), with mean zero (iii), and with variance one (iv), when the model corresponds to the observations. This study demonstrates that the OSA standardized residuals do indeed have the desired properties (i-iv) for distributions commonly used for observed compositions in fish stock assessment models (multinomial, Dirichlet, Dirichlet-multinomial, and logistic normal). For correctly specified multivariate models (simulated observations), the OSA standardized residuals always have better properties than the Pearson residuals, which often do not result in independent normal residuals. When working with true observations, the correct answer is unknown but the OSA and the Pearson residuals gave different evaluations with regards to the model's ability to describe the observations. If used for validating the model, the OSA residuals would likely result in the conclusion that the model was an acceptable approximation, whereas the Pearson residuals would likely lead to the opposite conclusion. Pearson residuals could therefore be wrongly interpreted and result in rejecting a suitable model. This could have direct consequences when these residuals are used for validating stock assessment models. Poor composition residuals are typically the basis for changing estimates of selectivity, which could lead to changes in the resulting management advice. Properly validating the models proposed as basis for managing marine natural resources is important, as invalid models could lead to undesired effects on the ecosystem and management errors.

Assessment models are often integrated models (Fournier and Archibald, 1982; Maunder and Punt, 2013), which means that they are integrating many types and sources of observations (e.g., catches, age-compositions, tag-returns, survey indices). This presents the challenge of correctly weighting the different data sources when estimating quantities of interest (e.g., abundance and fishing pressure for stock assessment models). The effective weighting of the different data sources is often determined in an ad-hoc manner by assigning variances or effective sample sizes. An important part of weighting correctly is therefore validating the residuals (Francis, 2011, 2017; Maunder and Piner, 2017). This can result in an imbalance between data sources such as non-compositional observations (e.g., aggregated catch or indices) for which Pearson residuals can easily be computed and compared to an independent standard normal distribution, and compositional observations, which prior to OSA residuals could not easily be standardized and de-correlated. For example, for age-structured catches in assessment models, often the total estimated catches are validated using Pearson residuals but little attention is given to validating the age composition because Pearson residuals are known to be problematic. This gives indirectly larger weight on the total catch than the age composition data. Inversely, when Pearson residuals are considered bad, compositional observations are sometimes down-weighted to solve model misspecification (Wang and Maunder, 2017). The properties i-iv allow the OSA residuals from composition observations to be evaluated with the same statistical rigor as residuals from any other observation type (e.g., a normally distributed survey index of abundance).

The validation and simulation results illustrate the difficulty in

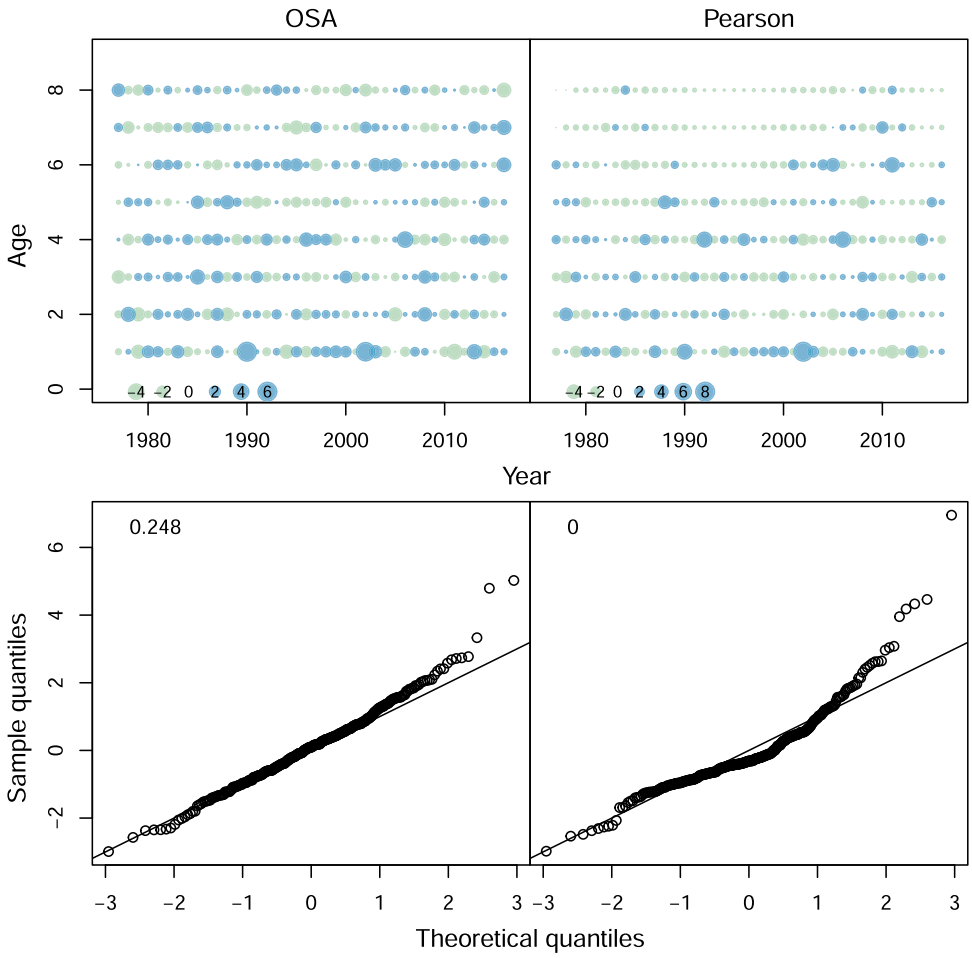


Fig. 5. Residual bubble plots (top) and normal Q-Q plots (bottom) for the 2nd fleet of the Gulf of Maine haddock true observations example. See caption of Figure 4 for a detailed description of the figure.

interpreting and validating residuals. Indeed, in most cases non-independent residuals are clearly visible in bubble plots but it might be difficult to identify residuals from a false model if only the variance is biased (e.g., multinomial example). In these cases, Q-Q plots, tests for normality and correlation could help identify nonnormal or correlated residuals. However, despite clear non-independent residuals in bubble plots, the tests for normality can sometimes conclude normality and different conclusions can sometimes be seen between different statistical tests as the power of the tests depends on the model violation. This illustrates the need to evaluate the residuals via several plots and test statistics (Carvalho et al., 2017, 2021). Discussing the choice of statistical tests is nevertheless outside the scope of this paper.

De-correlating multivariate distributions by successively predicting from one observation to the next does imply that individual scalar residuals are dependent on the order in which they are computed. The order in which composition residuals are computed is somewhat arbitrary. However, the overall properties of the residuals, and hence the conclusion about model validity should be the same.

This paper provides a tool to facilitate model validation for compositional data. Naturally, model validation based on data does not stand-alone. The model should be selected depending on the data at hand and the model properties. For instance, if accounting for essential zero in the observed compositions is needed, then distributions using proportions strictly between zero and one (e.g, Dirichlet and logistic-normal in this paper) cannot be used and an appropriate model should be chosen instead. Multivariate distributions other than the ones presented in this paper might also be relevant. Discussing the choice of multivariate distributions depending on their properties and the type of data is considered outside the scope of this paper, some literature however exist to inform modelers on the subject (Aitchison, 2003; Albertsen et al., 2017; Maunder, 2011; Francis, 2014; Thorson et al., 2017).

The OSA residuals defined here have the correct properties in the same way as well-known standardized residuals from linear normal models. Similar to standardized residuals from linear models these residuals are perfectly independent and standard normally distributed if they are calculated based on the true model parameters. In practical applications, the true model parameters are not available, so estimated model parameters are used instead, which means that the correct residual properties are asymptotic. Residuals can appear better than they should due to the parameter estimation, resulting in statistical tests for normality than can be rejected too rarely (Appendix A1). The lower the probability of rejecting the statistical test, the lower the power of the test is. This is a general problem that improves with the increase in sample size (Razali and Wah, 2011). However, in fisheries composition data, the sample size is often limited. Accounting for the use of estimated parameters instead of the true values in OSA residual calculation will be investigated in the future.

OSA residuals are notably relevant for stock assessment models, but will be equally relevant for the analysis of compositional data in other scientific disciplines. For models without random effects, these residuals can easily be calculated externally to the model and we have developed an R-package (https://github.com/fishfollower/compResidual) that can

estimate residuals given data and outputs for the distributions presented in this paper, as well as for the multivariate normal distribution. For models with random effects (e.g., state-space models), the OSA residuals need to be estimated within the model as random processes will affect the correlation of the compositional variables. All these distributions are now supplied as contributed code to TMB (see https://github.com/vtrijoulet/OSA_multivariate_dists) for internal use in state-space models based on TMB. For example, the Woods Hole Assessment Model has incorporated the distributions to allow OSA residuals to be calculated for age composition observations (Miller and Stock, 2020; Stock and Miller, 2021). For state-space models not based on TMB, this paper provides a clear description on how to implement the OSA residuals internally in section 2.1.2. The publicly available TMB code in the R-package and in the TMB contribution can also be used as examples for other coding languages.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The Gulf of Maine true observations and model predictions are provided as example in the README of the R-package "compResidual" (https://github.com/fishfollower/compResidual).

Acknowledgments

Funding: This work was supported by the European Maritime and Fisheries Foundation and the Ministry of Environment and Food of Denmark (grant ID: 33113-B-20–174; project: Genopbygning af bestandskomplekset af forårsgydende sild, GENBYGSILD).

We thank Jon Deroba for providing comments on an early version of the manuscript.

CRediT authorship contribution statement

Vanessa Trijoulet: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Christoffer Moesgaard Albertsen**: Methodology, Writing – review & editing. **Kasper Kristensen**: Methodology, Software, Validation, Writing – review & editing. **Christopher M. Legault**: Methodology, Writing – review & editing. **Timothy J. Miller**: Methodology, Writing – review & editing. **Anders Nielsen**: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fishres.2022.106487.

References

Aitchison, J., 2003. The Statistical Analysis of Compositional Data. The Blackburn Press, Caldwell, New Jersey USA.

Aitchison, J., Shen, S.M., 1980. Logistic-normal distributions: some properties and uses. Biometrika 67, 261–272. https://doi.org/10.2307/2335470. ⟨http://www.jstor.org/stable/2335470⟩.

Albertsen, C., Nielsen, A., Thygesen, U., 2017. Choosing the observational likelihood in state-space stock assessment models. Can. J. Fish. Aquat. Sci. 74, 779–789. https://doi.org/10.1139/cjfas-2015-0532.

AnonNortheast Fisheries Science Center, NEFSC, 2017.Gulf of Maine haddock 2017 Assessment Update.Unpubl. Rpt. 10 pp ⟨https://apps-nefsc.fisheries.noaa.gov/saw/sasi/sasi_report_options.php⟩.

AnonNortheast Fisheries Science Center, NEFSC, 2019.Gulf of Maine haddock 2019 Assessment Update.Unpubl. Rpt. 10 pp ⟨https://apps-nefsc.fisheries.noaa.gov/saw/sasi/sasi_report_options.php⟩.

Begley, J., 2005.Gadget user guide. Technical Report.

Bull, B., Francis, R., Dunn, A., McKenzie, A., Gilbert, D., Smith, M., Bian, R., Fu, D., 2005. CASAL (C++ algorithmic stock assessment laboratory): CASAL User Manual v2. Technical Report.

Carvalho, F., Punt, A.E., Chang, Y.J., Maunder, M.N., Piner, K.R., 2017. Can diagnostic tests help identify model misspecification in integrated stock assessments? Fish. Res. 192, 28–40. https://doi.org/10.1016/j.fishres.2016.09.018. ⟨https://www.sciencedirect.com/science/article/pii/S0165783616303113⟩.

Carvalho, F., Winker, H., Courtney, D., Kapur, M., Kell, L., Cardinale, M., Schirripa, M., Kitakado, T., Yemane, D., Piner, K.R., Maunder, M.N., Taylor, I., Wetzel, C.R., Doering, K., Johnson, K.F., Methot, R.D., 2021. A cookbook for using model diagnostics in integrated stock assessments. Fish. Res. 240, 105959 https://doi.org/10.1016/j.fishres.2021.105959. ⟨https://www.sciencedirect.com/science/article/pii/S0165783621000874⟩.

Dunn, P.K., Smyth, G.K., 1996. Randomized quantile residuals. J. Comput. Graph. Stat. 5, 236–244. https://doi.org/10.1080/10618600.1996.10474708.

Fournier, D., Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195–1207.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. 55, 2105–2116. https://doi.org/10.1139/f98-100.

Francis, R., 2017. Revisiting data weighting in fisheries stock assessment models. Fish. Res. 192, 5–15. https://doi.org/10.1016/j.fishres.2016.06.006. ⟨https://www.sciencedirect.com/science/article/pii/S0165783616301953⟩.

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

Francis, R.C., 2014. Replacing the multinomial in stock assessment models: a first step. Fish. Res. 151, 70–84. https://doi.org/10.1016/j.fishres.2013.12.015. ⟨https://www.sciencedirect.com/science/article/pii/S0165783613003093⟩.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 1995. Bayesian data analysis. Chapman and Hall.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., Bell, B., 2016. TMB: automatic differentiation and laplace approximation. J. Stat. Softw. 70, 1–21. https://doi.org/10.18637/jss.v070.i05.

Legault, C.M., Restrepo, V.R., 1999.A flexible forward age-structured assessment program 49, 246–253.

Lewy, P., Vinther, M., 2004.A stochastic age-length-structured multispecies model applied to North Sea stocks. Technical Report.

Maunder, M.N., 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: Estimating the effective sample size. Fish. Res. 109, 311–319. https://doi.org/10.1016/j.fishres.2011.02.018. ⟨https://www.sciencedirect.com/science/article/pii/S0165783611000890⟩.

Maunder, M.N., Piner, K.R., 2017. Dealing with data conflicts in statistical inference of population assessment models that integrate information from multiple diverse data sets. Fish. Res. 192, 16–27. https://doi.org/10.1016/j.fishres.2016.04.022. ⟨https://www.sciencedirect.com/science/article/pii/S0165783616301394⟩.

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74. https://doi.org/10.1016/j.fishres.2012.07.025. ⟨https://www.sciencedirect.com/science/article/pii/S0165783612002627⟩.

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. https://doi.org/10.1016/j.fishres.2012.10.012.

Miller, T.J., Stock, B.C., 2020.The Woods Hole Assessment Model (WHAM). ⟨https://timjmiller.github.io/wham/⟩.v1.0.6.

Mohn, R., 1999. The retrospective problem in sequential population analysis: an investigation using cod fishery and simulated data. ICES J. Mar. Sci. 56, 473–488. https://doi.org/10.1006/jmsc.1999.0481 (arXiv). ⟨https://academic.oup.com/icesjms/article-pdf/56/4/473/1734305/56-4-473.pdf⟩.

Nielsen, A., Berg, C.W., 2014. Estimation of time-varying selectivity in stock assessments using state-space models. Fish. Res. 158, 96–101. https://doi.org/10.1016/j.fishres.2014.01.014.

Punt, A.E., Dunn, A., Elvarsson, B.T., Hampton, J., Hoyle, S.D., Maunder, M.N., Methot, R.D., Nielsen, A., 2020. Essential features of the next-generation integrated fisheries stock assessment package: A perspective. Fish. Res. 229, 105617 https://doi.org/10.1016/j.fishres.2020.105617. ⟨https://www.sciencedirect.com/science/article/pii/S016578362030134X⟩.

Razali, N.M., Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. J. Stat. Model. Anal. 2, 21–33.

Rosenblatt, M., 1952. Remarks on a multivariate transformation. Ann. Math. Stat. 23, 470–472. ⟨http://www.jstor.org/stable/2236692⟩.

Stock, B.C., Miller, T.J., 2021. The woods hole assessment model (WHAM): a general state-space assessment framework that incorporates time- and age-varying processes via random effects and links to environmental covariates. Fish. Res. 240, 105967 https://doi.org/10.1016/j.fishres.2021.105967.

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G., 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fish. Res. 192, 84–93. https://doi.org/10.1016/j.fishres.2016.06.005. ⟨https://www.sciencedirect.com/science/article/pii/S0165783616301941⟩.

Thygesen, U.H., Albertsen, C.M., Berg, C.W., Kristensen, K., Nielsen, A., 2017. Validation of ecological state space models using the Laplace approximation. Environ. Ecol. Stat. 24, 317–339. https://doi.org/10.1007/s10651-017-0372-4.

Trijoulet, V., Fay, G., Curti, K.L., Smith, B., Miller, T.J., 2019. Performance of multispecies assessment models: insights on the influence of diet data. ICES J. Mar. Sci. 76, 1464–1476. https://doi.org/10.1093/icesjms/fsz053.

Trijoulet, V., Fay, G., Miller, T.J., 2020. Performance of a state-space multispecies model: what are the consequences of ignoring predation and process errors in stock assessments? J. Appl. Ecol. 57, 121–135. https://doi.org/10.1111/1365-2664.13515.

Wang, S.P., Maunder, M.N., 2017. Is down-weighting composition data adequate for dealing with model misspecification, or do we need to fix the model? Fish. Res. 192, 41–51. https://doi.org/10.1016/j.fishres.2016.12.005. ⟨https://www.sciencedirect.com/science/article/pii/S0165783616304143⟩.

Williams, E.H., Shertzer, K.W., 2015.Technical documentation of the Beaufort Assessment Model (BAM). Technical Report.