ICES Journal of Marine Science

ICES Journal of Marine Science (2015), 72(8), 2209–2222. doi:10.1093/icesjms/fsv117

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

Original Article

General state-space population dynamics model for Bayesian stock assessment

Samu H. P. Mäntyniemi[1]*, Rebecca E. Whitlock[2], Tommi A. Perälä[1], Paul A. Blomstedt[1‡], Jarno P. Vanhatalo[1], Margarita María Rincón[3], Anna K. Kuparinen[1], Henni P. Pulkkinen[4], and O. Sakari Kuikka[1]

[1]*Fisheries and Environmental Management Group (FEM), Department of Environmental Sciences, University of Helsinki, Viikinkaari 2, PO Box 65, FIN-00014 Helsinki, Finland*
[2]*Department of Aquatic Resources, Swedish University of Agricultural Sciences, Stångholmsvägen 2, 178 93 Drottningholm, Sweden*
[3]*Instituto de Ciencias Marinas de Andalucía, Consejo Superior de Investigaciones Científicas, ICMAN-CSIC, Puerto Real, Cádiz, Spain*
[4]*Natural Resources Institute Finland, Paavo Havaksen tie 3, PO Box 413, FI-90014 Oulu, Finland*

*Corresponding author: tel: + 358 50 4151098; fax: + 358 2941 58257; e-mail: samu.mantyniemi@helsinki.fi
‡Helsinki Institute for Information Technology HIIT, Deparment of Computer Science, Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland

Mäntyniemi, S. H. P., Whitlock, R. E., Perälä, T. A., Blomstedt, P. A., Vanhatalo, J. P., Rincón, M. M., Kuparinen, A. K., Pulkkinen, H. P., and Kuikka, O. S. General state-space population dynamics model for Bayesian stock assessment. – ICES Journal of Marine Science, 72: 2209 – 2222.

Received 24 February 2015; revised 22 May 2015; accepted 11 June 2015; advance access publication 7 August 2015.

This study presents a state-space modelling framework for the purposes of stock assessment. The stochastic population dynamics build on the notion of correlated survival and capture events among individuals. The correlation is thought to arise as a combination of schooling behaviour, a spatially patchy environment, and common but unobserved environmental factors affecting all the individuals. The population dynamics model isolates the key biological processes, so that they are not condensed into one parameter but are kept separate. This approach is chosen to aid the inclusion of biological knowledge from sources other than the assessment data at hand. The model can be tailored to each case by choosing appropriate models for the biological processes. Uncertainty about the model parameters and about the appropriate model structures is then described using prior distributions. Different combinations of, for example, age, size, phenotype, life stage, species, and spatial location can be used to structure the population. To update the prior knowledge, the model can be fitted to data by defining appropriate observation models. Much like the biological parameters, the observation models must also be tailored to fit each individual case.

Keywords: Dirichlet-Multinomial distribution, effective population size, Markov chain Monte Carlo, stock assessment, uncertainty.

Introduction

Because the size of a fish stock can very seldom be observed directly without error, the credibility of a population assessment can be only assessed as a combination of two criteria: realistic assumptions and the ability to predict observations using the assessment model (Kuparinen *et al.*, 2012). From these, the latter is of secondary importance: it is possible to formulate a flexible statistical model that may fit the data very well with a few parameters that would not even have the hidden status of the stock as a parameter to be estimated. This study deals with the former by proposing a generic population dynamics model which can be customized for a specific population by representing a wide variety of different assumptions and life history choices.

While generic formulations of this kind already exist, they are limited in their ability to model and admit uncertainty about process error (Bull *et al.*, 2002; Mäntyniemi *et al.*, 2013a; Methot and Wetzel, 2013). Models that account for both process variation and observational uncertainty are called state-space models

© International Council for the Exploration of the Sea 2015.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

(Newman *et al.*, 2006; Buckland *et al.*, 2007). While recognized as superior to deterministic models, the practical application of state-space models has been hindered by computational difficulties (Newman *et al.*, 2009). One of the most recent advances in the area is the age-structured length-based model we presented in Mäntyniemi *et al.* (2013a) and applied in Mäntyniemi *et al.* (2013b). The model included multiple parameters for process error variances of different kinds, which in turn were treated as uncertain. The model formulation was based on explicitly defined random effects, which were also estimated using Markov chain Monte Carlo simulation. In this study, our approach is slightly different: we seek to reduce the number of process error variance parameters and to integrate the random effects out analytically whenever possible.

It is widely recognized that grouping behaviour, clustered environment, and environmental stochasticity lead to overdispersed probability distributions for organism counts in demography, time, and space (e.g. Richards, 2008; Linden and Mantyniemi, 2011; Dorazio *et al.*, 2013), compared with baseline situation representing independent behaviour under constant environment. As recently shown by Linden and Mantyniemi (2011), the variance structure of the count distribution depends on the assumptions made about these processes. While the processes leading to overdispersion can be modelled explicitly by formulating hierarchical model structures, both Bayesian and maximum likelihood estimation of model parameters become more difficult when the number of parameters increases with the introduction of random effects (Richards, 2008). In cases where the random effects are not of direct interest, the parameter estimation can be greatly expedited if the random effects can be integrated out analytically. Lindén and Mäntyniemi (2011) showed that when the interest lies in the modelling of rate or intensity parameters, the negative binomial distribution can be often used as the marginal distribution of univariate count data either exactly or approximately.

In this study, we consider the case where the parameters of interest are the proportion vector and/or the order of a multinomial model and overdispersion is known to be present in the process under study. This includes the simple Binomial model as a special case. We propose to use the Dircihlet-Multinomial (DM) distribution (also known as multivariate Pólya or Dirichlet compound multinomial distribution) as an overdispersed alternative for the standard multinomial model. The DM distribution is used regularly in some other areas of science, for example, in modelling the burstiness of words in a text (Xu and Akella, 2008) and to estimate the ammunition allocation in military combat modelling (Kvam and Day, 2001). However, except the univariate special case called Beta-Binomial distribution (e.g. Mäntyniemi and Romakkaniemi, 2002; Richards, 2008; Dorazio *et al.*, 2013), the use of this distribution seems uncommon in ecology and fisheries literature (Hulson *et al.*, 2012). Another notable exception is the use of the DM distribution in the context of determining the effective sample size (Hulson *et al.*, 2012). It is noteworthy that the cases above deal with the observational processes but not population dynamics.

The remainder of the study is structured as follows. Next section introduces and derives the process error distribution from assumptions about the dependent behaviour of fish. The third section shows how this process error model can serve the general population dynamics model and shows a way to reparameterize the model from a computational viewpoint. Appendices cover the relevant properties of the DM distribution and its approximation, an exemplary growth matrix, and a simulation study with known parameter values.

Process error distribution

The variance of the conditional probability distribution that describes the transition of the population structure and abundance over a period is often called "process error". It serves to measure the uncertainty that the analyst would have about the next state of the population in the case that the current state was known. While state-space models are becoming more common in fisheries stock assessment, the process error distribution has gained relatively little attention. A common choice has been to use a lognormal distribution either for abundances or for instantaneous mortality rates (e.g. Mäntyniemi *et al.*, 2013a; Maunder *et al.*, 2015). The purpose of this section is to look into the transition process in more detail and study the properties of the transition distribution under different assumptions about the process. We start from simplistic assumptions and work towards more realistic cases. The exemplary context is survival, but the same principles hold for capture, detection, and migration events. After deriving the process error distributions, we propose approximations that can be expected to ease the computational load of the models. Equation numbers used in this section refer to the equations in Table 1, which summarizes the resulting model structures.

Independent survival

We start by considering the univariate transition in a survival process, where each of N_t individuals survives (or not) to the next time-step, so that the population size in the next time-step is $N_{t+1} \leq N_t$. Each fish can be thought to have their own tendency $\varphi_1, \ldots, \varphi_{N_t}$ to survive, which could also be called survival probability through a time-step. It should be noted at this point that this is not a Bayesian probability but a parameter that describes the property of the individual: we think that fish i with $\varphi_i = 0.27$ would survive in 27% of all potential and equally likely conditions that may occur during the time-step. Analyst's knowledge of φ_i can then be quantified using the Bayesian degree of belief in the form of prior $p(\varphi_i)$. The event of survival of fish i can be described using an indicator variable z_i, which has Bernoulli distribution with parameter φ_i (Figure 1a). The degree of belief about the survival of fish i is then $P(z_i = 1) = \int \varphi_i p(\varphi_i) d\varphi_i = E(\varphi_i) = \mu$, which is just the mean of the prior distribution. It is worth noting that the prior variance of φ_i does not affect this probability (Mäntyniemi *et al.*, 2005).

If all fish are assumed to be independent and exchangeable in respect to their survival, then the distribution of the survivors is simply a binomial distribution [Equation (1), Mäntyniemi *et al.*, 2005]. This includes the case where all fish are believed to have exactly the same probability of survival, i.e. $V(\varphi_i) = 0$.

Temporal variation in survival

Let us now consider the case where the mean survival probabilities μ_t of different time-steps are seen as exchangeable, and consequently as conditionally i.i.d random draws from a distribution with mean $E(\mu_t) = \nu$ and variance $V(\mu_t) = \delta$. A practical functional form for the temporal variation is a Beta-distribution with parameters $\alpha = \nu\eta$ and $\beta = (1 - \nu)\eta$, so that $\delta = \nu(1 - \nu)(\eta + 1)^{-1}$. Then, the annual mean survival probabilities μ_t can be integrated out analytically (Figure 1c, when $s_t = s_k = 1$) and the marginal distribution of survivors given the previous population size and the parameters that describe the temporal variation is a Beta-Binomial distribution where parameter η controls the temporal variation (Equation 2). When η is large, annual mean survival probabilities are highly

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

Table 1. Parameterization of the Beta-Binomial distribution under different assumptions about the correlation in the survival process.

Case	Number of survivors	Parameters
(1) Independent survival	$N_{t+1}\|N_t, \mu_t \sim \text{Bin}\ (N_t, \mu_t)$	N_t = number of individuals at time t μ_t = mean survival probability of individuals at time t
(2) Temporal variation in independent survival	$N_{t+1}\|N_t, \nu, \eta \sim \text{Beta} - \text{Bin}(N_t, \nu\eta, (1-\nu)\eta)$	N_t ν = expected mean survival probability over time η = concentration of mean survival probability
(3) Schools of equal size and temporal variation	$N_{t+1}\|N_t, \nu, \eta, s_t \sim \text{Beta} - \text{Bin}(N_t, \nu\eta^*, (1-\nu)\eta^*)$ $\eta^* = \frac{(N_t - s_t)\eta}{(s_t - 1)\eta + N_t - 1}$	N_t ν η s_t = group size
(4) Correlated individuals	$N_{t+1}\|N_t, \mu_t, \rho_t \sim \text{Beta}\ - \text{Bin}\ (N_t, \mu_t\eta^{**}, (1-\mu_t)\eta^{**})$ $\eta^{**} = \frac{1}{\rho_t} - 1$	N_t μ_t ρ_t = correlation in survival
(5) Random schools	$N_{t+1}\|N_t, \mu_t, k_t, \psi_t \sim \text{Beta} - \text{Bin}\ (N_t, \mu_t\eta^{***}, (1-\mu_t)\eta^{***})$ $\eta^{***} = \frac{N_t - \xi_t/(N_t\mu_t(1-\mu_t))}{\xi_t/(N_t\mu_t(1-\mu_t)) - 1}$ $\xi_t = k_t\sigma' + k_t(k_t-1)\sigma''$ $\sigma' = \frac{\mu_t}{k_t}\left(\frac{N_t^2(1-\mu_t)}{k_t} + (N_t - k_t)\left(1 - \frac{1}{k_t}\right)(1 + (N_t - k_t - 1)\psi_t)\right)$ $\sigma'' = -\frac{\mu_t^2}{k_t^2}(N_t - k_t)(1 + (N_t - k_t - 1)\psi_t)$	N_t μ_t k_t = number of groups ψ_t = correlation in grouping process

concentrated around the mean, and small values indicate a higher degree of variation.

Schooling behaviour

When fish form schools either actively or passively, the assumption of independence in survival becomes violated. A school of fish may be able to feed and evade predators more effectively, so that the success of the group benefits all of its members (Magurran, 1990). On the other hand, the failure of the grouping strategy means reduced chance of success for all fish in the school. This is expected to increase the variance of the number of survivors. This kind of correlated survival process can be modelled in number of ways, of which three are presented here.

Schools of equal size

An extreme assumption is that each group contains exactly $s_t \leq N_t$ individuals, which all either survive or die together. The population can then be thought to consist of $k_t = N_t/s_t$ independent groups that have mean survival probability μ_t. If this mean survival probability is believed to vary in time (Figure 1c) in the same way as in the previous case, the mean and variance of the number of survivors are

$$E(N_{t+1}) = N_t\nu,$$
$$V(N_{t+1}) = N_t\nu(1-\nu)s_t\frac{N_t/s_t + \eta}{1+\eta}.$$

When $s_t = 1$, the distribution is exactly the same Beta-Binomial as in the previous case. Binomial distribution is the limiting model of this model, when group size is one and there is no temporal random variation in the mean survival probability, i.e. when $\eta \to \infty$. While it is hardly realistic that all schools of fish would have the same number of individuals, this can be seen as a practical way for approximating the effect of schooling on the number of survivors by using a Beta-Binomial distribution with these mean and variance (Equation 3).

Conditional probability and correlation

Probably, many different mechanisms can lead to a situation where a successful decision of an individual can increase the chances of the others to survive. For example, when a predator spots a batch of prey, the other predators nearby can use that information and follow to the same batch. A prey, on the other hand, may find a way to escape from the predators and others may follow the example. Similarly, a fish that finds a way out from a fishing trap can also lead the whole group to the freedom. Consider two individuals randomly chosen from the population, and assume that both have survival probability μ_t. If the survival of one or the other does not change the survival probability of the other one, then fish can be modelled as independent and the binomial model results for the total number of survivors. However, if the survival of one or the other changes the survival probability of the other one to $\lambda_t > \mu_t$, then the survivals are positively correlated (Figure 1b) and the assumptions of the binomial distribution no longer hold. In this case, the correlation between the survivals of the two randomly chosen fish is $\rho_t = (\lambda_t - \mu_t)/(1-\mu_t)$ and the covariance is $(\lambda_t - \mu_t)\mu_t$. The variance of the total number of survivors is then obtained as the sum of all elements of the covariance matrix between all the fish in the population. As shown by Hisakado *et al.* (2006), in this case, the distribution of N_t is exactly Beta-Binomial (Equation 4).

Independent random schools

Another possibility is to derive the mean and variance of the number of survivors by thinking about the size of the schools as a random process. Assume that the population consists of k_t groups, where by definition each group is assumed to have at least one individual, and other individuals then join these groups randomly. Hence, each group consists of $y_{j,t} + 1 \geq 1, j = 1, \ldots, k_t$ individuals, where $y_{j,t}$ is the number of additional fish in the group. Individuals may show correlated or independent behaviour when joining these groups. For the sake of generality, we consider the correlated case here: assume that the correlation between individuals in the grouping

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

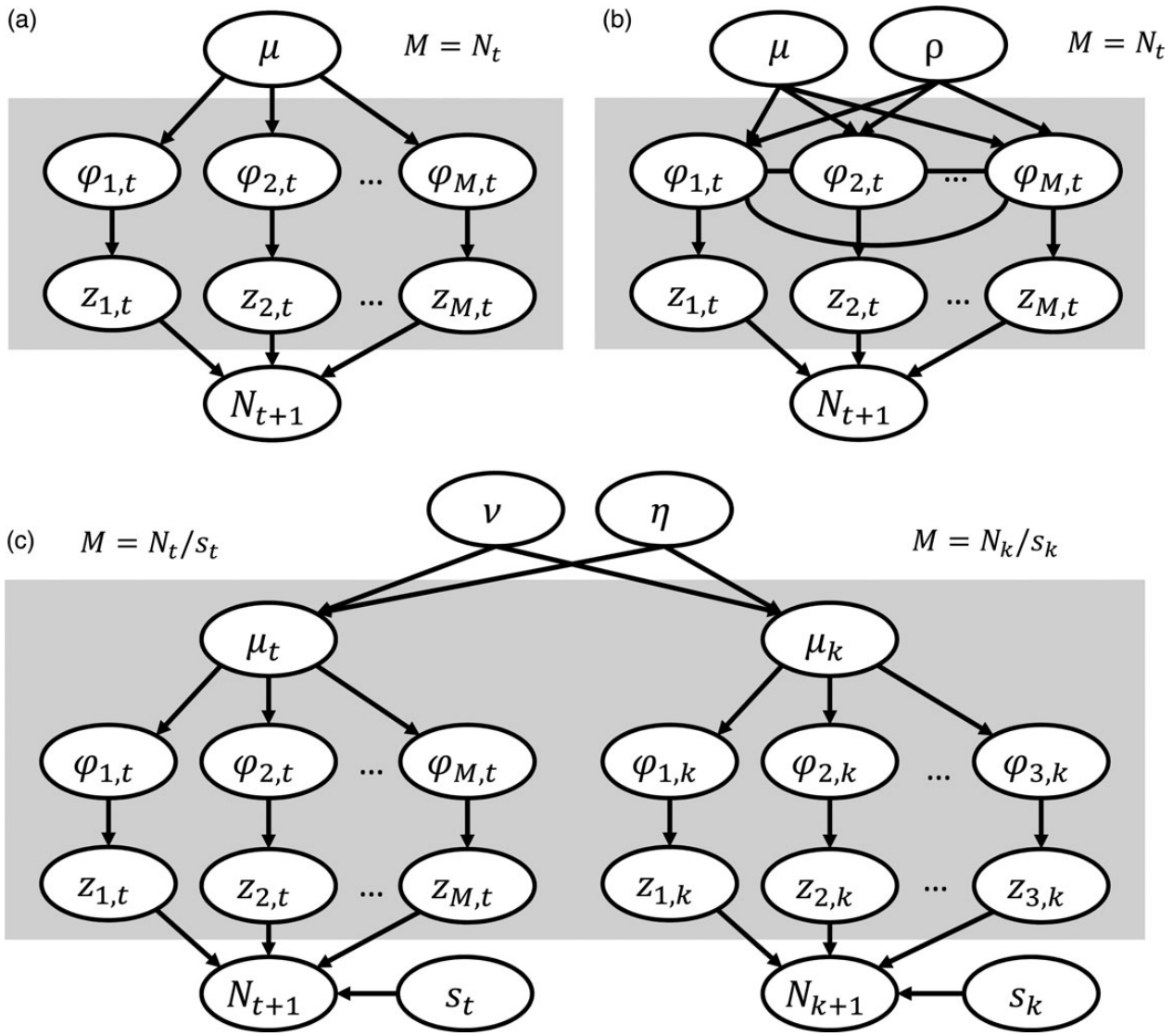

Figure 1. Graphical models describing different assumptions about the survival process. Ovals represent uncertain variables that are given a prior distribution. Arrows represent the direction of conditional specification: arrows point from conditioning factor to the dependent variable. Multivariate distributions are denoted using undirected arcs. Variables that are integrated out from the final model are shown on grey background. The resulting probability distributions and their parameters are given in Table 1. Model (a) corresponds to independent survival of individuals, model (b) specifies schooling behaviour with the concept of correlation, and model (c) represents both temporal variation of mean survival and schooling behavior.

process is ψ_t. In other words, if individual i happens to choose group j, then a randomly chosen individual chooses the same group with probability $1/k_t + (1 - 1/k_t)\psi_t$. In this case, the vector of additional fish $(y_{1,t}, \ldots, y_{k_t,t})$ in each group is a Dirichlet-Multinomial distribution (details in Appendix 1), which is a multivariate extension of the Beta-Binomial

$$(y_{1,t}, \ldots, y_{k_t,t})|k_t, \psi_t, N_t \sim \mathrm{DM}\left(N_t - k_t, \left(\frac{1}{k_t}, \ldots, \frac{1}{k_t}\right)(1/\psi_{t-1})\right)$$

and the total number of fish in each group is $x_{j,t} = y_{j,t} + 1$. Next, we assume that each of these groups would survive independently with probability φ_t and use a Bernoulli-distributed indicator variable $z_{j,t}$ to denote whether group j survived or not. The number of survivors from each group is then obtained as $z_{j,t}x_{j,t}$. Finally, the total number of survivors is the sum of survivors from each group (Figure 2). Our suggestion is to approximate this distribution using a Beta-Binomial distribution (Equation 5).

Estimation of the variance parameters

In the previous sections, we introduced a few parameters that can be used to describe the process variance. In the Bayesian context, the parameter estimation is a process of gathering existing knowledge of the parameters and formulating this knowledge using a prior probability distribution. Depending on the assumed model structures and the type and amount of data available, this prior distribution may or may not become updated. In our view, the choice of the parameterization should be primarily made based on the biological knowledge of the population under study: this should make it easier to express the prior knowledge in a biologically meaningful way.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

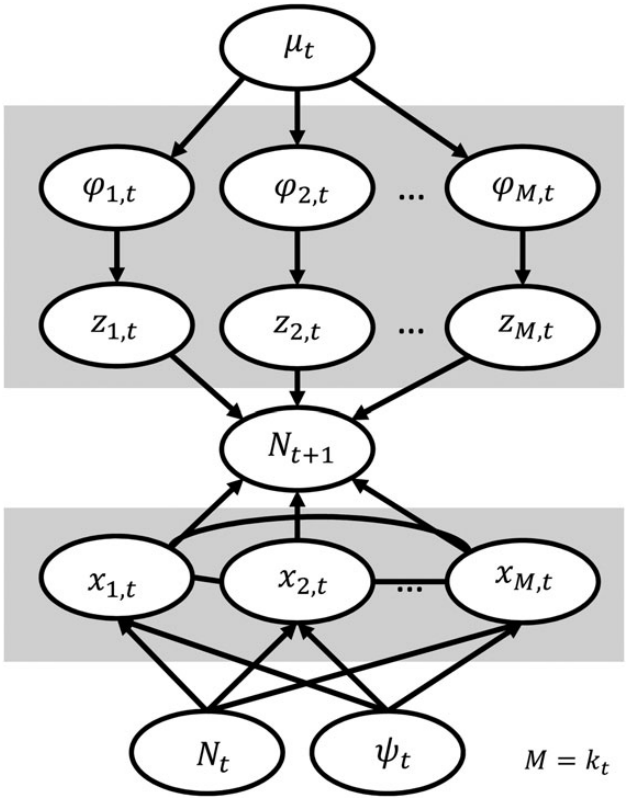

Figure 2. Graphical model describing a survival process where randomly formed schools of fish survive independently. The resulting probability distribution for the number or survivors is given in Table 1.

For example, instead of assigning a prior distribution for the size of the schools, it may be more relevant to think about the number of schools and let the size of the schools then depend on the population size and the number of schools.

For random schools, there are two parameters to think about, which both relate to the schooling process. For fixed schools and random variation in the survival rate, the size (or number) of the schools is the biologically meaningful parameter, whereas the amount of random variation in the survival rate is representing researchers lack of knowledge on upcoming environmental conditions and their effect on the survival. Thus, the former defines the minimum variance in the survival process, which cannot be reduced by improved understanding about the effects and future values of environmental covariates. This has obvious impacts on the testing of hypotheses about the effects of environmental covariates using state-space population dynamic models (Maunder *et al.*, 2015).

In other words, in addition to such a mechanistic interpretation, the overdispersion parameter η can be seen to represent unexplained variation (Warton and Hui, 2011) in multinomial regression, which means that the DM distribution can be considered as an alternative model to random effects logistic regression, where the random effects have been analytically integrated out.

In the context of the estimation of population age or size structure, the concept of effective sample size is often used to describe the amount of overdispersion in the sampling process compared with dispersion of data expected under the assumption of independence of individuals (Hulson *et al.*, 2011, 2012). The cause for this overdispersion is that instead of being independent, the fish sampled either from catch or from population tend to be dependent because fish of similar size and age are caught together at the same sampling occasion. Because of this dependence, the effective sample size is considered to be lower than the actual sample size. Perhaps, the easiest way to understand effective sample size is to think about the variance of the proportion of samples in a certain class. Under the assumption of independence, the distribution of counts is Binomial $x|N, p \sim \text{Bin}(N, p)$ and the variance of the proportion is $V(x/N) = p(1-p)/N$. However, if the variance appears to be larger, this can be matched by $V(x/N) = p(1-p)/N_{\text{eff}}, N_{\text{eff}} < N$. The full distribution can be modelled using a Beta-Binomial $x|N, p, N_{\text{eff}} \sim \text{Beta—Bin}(N, pv, (1-p)v)$, where $v = N(1-N_{\text{eff}})/(N_{\text{eff}}-N)$. In the context of population dynamics, it is analogous to think about effective population size regarding the variation in the transition process. For example, for a fixed school size and random variation in the transition rate, the effective population size can be obtained as $N_{\text{eff},t} = N_t s_t(1-\eta)/(N_t + s_t^2\eta)$.

The general population dynamics model

The scope of the modelling approach is a set of populations that are either physically and/or mentally connected. The mental connection here means that the person assessing the populations believes that information about one population is relevant to the knowledge of other populations. By physical connection, we mean movement or other interaction between populations and also the situation where all populations are affected by same external factors such as environmental conditions and fishery operations. An extreme example of minimal physical connection might be a mixed fishery of non-interacting species. At the other extreme, we could define the set of populations using a spatial grid where each cell of the grid interacts with its neighbours in terms of migration: fish of all sizes and ages might move between these "populations".

Within each population, individuals can be further structured according to at least one attribute, such as age, size, or maturity stage. Modelling the dynamics within each population boils down to defining the transition that may occur during a time-step.

The setup of the problem is the following. The total population is assumed to be closed in the sense that there is no immigration and no emigration, but there may be movement between subpopulations. Multiple fishing fleets are harvesting the population at the same time, although the harvesting pressure can be highly different in different periods of time.

For the sake of concreteness, we derive the model structure in the context of a size-structured population. However, without loss of generality, the size classes can be substituted by any other means of structuring the population. For the simplicity of notation, we first derive the model for one subpopulation and describe the exchange between subpopulations after that. This section concentrates on the core population dynamic equations that we envisage would be the same for all case studies. Parameters that are not explicitly defined here are assumed to be given a case-specific prior distribution, which can include anything from setting a fixed value to a complex hierarchical model with functions of environmental covariates and/or other model parameters.

While the size distribution of the population can be best understood in a continuous domain, we specify the model in terms of discretized size distribution. The vector of breakpoints is denoted as $(I_1, I_2, \ldots, I_{k^*+1})$, where k^* is the number of size classes and I_k denotes the lower bound of kth size bin. The upper bound of kth size bin is I_{k+1}. The state of the population at the beginning of time-step t is then summarized by vector $\mathbf{n_t} = (n_{t,1}, \ldots, n_{t,k^*})$, where $n_{t,k}$ denotes the number of individuals in size class k.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

Growth

The somatic growth of individuals is assumed to take place instantly at the beginning of each time-step. Within a discrete size distribution, the individual variation in growth is reflected as random movements of individuals between length bins. Each individual may stay in the same bin or move to higher bins. Each individual would have different probabilities of moving to higher bins depending on their age and individual growth parameters. However, in a size-structured model, there is no bookkeeping of individuals and their ages, which makes the modelling of growth variation a challenging task.

From the point of view of individuals, it would make sense to use a triangular transition matrix to move individuals between length bins. However, this leads to undesirable behaviour of the length distribution at the population level: eventually, all individuals of a cohort would belong to the highest length class which contradicts the usual situation where the variance of size at age increases with age and reaches an asymptote at the same rate as the mean size. It turns out that a square transition matrix can partially solve the problem at the population level, although such a matrix does not work as a growth model for an individual. The growth matrix is denoted as

$$\mathbf{g_t} = \begin{pmatrix} g_{t,1,1} & g_{t,1,2} & \cdots & g_{t,1,k^*} \\ g_{t,2,1} & g_{t,2,2} & \cdots & g_{t,2,k^*} \\ \vdots & \vdots & \ddots & \vdots \\ g_{t,k^*,1} & g_{t,k^*,2} & \cdots & g_{t,k^*,k^*} \end{pmatrix},$$

where $g_{t,i,j}$ denotes the probability of an individual to move to bin j from bin i. Appendix 2 contains a brief example on the derivation of this matrix for the von Bertalanffy growth model.

If individuals are assumed to jump between size classes independently, then each row vector of $\mathbf{g_t}$ acts as a probability vector for a multinomial distribution and elements of $\mathbf{n_t}$ provide the order parameter for each multinomial:

$$(z_{t,k,1}, z_{t,k,2}, \ldots, z_{t,k,k^*}) \sim \text{Multi}(n_{t,k}, (g_{t,k,1}, g_{t,k,2}, \ldots, g_{t,k,k^*})).$$

The state of the population after the growth is then found as the element-wise sum of all such vectors:

$$\mathbf{n_t^{(G)}} = \left(\sum_{i=1}^{k^*} z_{t,i,1}, \sum_{i=1}^{k^*} z_{t,i,2}, \ldots, \sum_{i=1}^{k^*} z_{t,i,k^*} \right).$$

Given the size frequencies at the beginning of the time-step and the growth matrix, the new vector of expected size frequencies after the growth is given by

$$E(\mathbf{n_t^{(G)}}) = \mathbf{n_t} \times \mathbf{g_t}$$

and the covariance matrix is

$$\text{COV}(\mathbf{n_t^{(G)}}) = \begin{pmatrix} \sum_{k=1}^{k^*} n_{t,k} g_{t,k,1}(1 - g_{t,k,1}) & -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,1} g_{t,k,2} & \cdots & -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,1} g_{t,k,k^*} \\ -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,2} g_{t,k,1} & \sum_{k=1}^{k^*} n_{t,k} g_{t,k,2}(1 - g_{t,k,2}) & \cdots & -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,2} g_{t,k,k^*} \\ \vdots & \vdots & \ddots & \vdots \\ -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,k^*} g_{t,k,1} & -\sum_{k=1}^{k^*} n_{t,k} g_{t,k,k^*} g_{t,k,2} & \cdots & \sum_{k=1}^{k^*} n_{t,k} g_{t,k,k^*}(1 - g_{t,k,k^*}) \end{pmatrix}.$$

Mortality and survival

Natural and fishing mortality are modelled as instantaneous rates that are assumed to stay constant throughout the time-step. Additional multipliers for these rates are provided in the model structure, so that lengths of the time-steps do not need to be equal and repeating seasonal patterns can be accounted for. If the patterns are not exactly known, prior distribution can be used to reflect this uncertainty. All mortality rates can have hierarchical priors with covariates as in Mäntyniemi *et al.* (2013a, b). The random variation in the mortality process is modelled using the concept of correlated survival events as described in "Process error distribution".

Individuals in the same size class are assumed to have same instantaneous annual natural mortality rate $M_{t,k}$ and fleet-specific annual fishing mortality rate $F_{t,k,j}$, where j denotes the fleet. These rates are assumed to be constant throughout the time-step. Total instantaneous mortality rate is then

$$Z_{t,k} = \delta_t^{(M)} M_{t,k} + \delta_t^{(F)} \sum_{j=1}^{j^*} F_{t,k,j},$$

where j^* denotes the number of fishing fleets. Effective length of time-step (as a fraction of a year) for natural mortality is denoted as $\delta_t^{(M)}$ and for fishing mortality as $\delta_t^{(F)}$. Based on the instantaneous rates, the probabilities of individuals dying naturally ($\rho_{t,k}$), being caught by fleet j ($\gamma_{t,k,j}$) and surviving to the next time-step ($\pi_{t,k}$) can be found using the results of Baranov (Xiao, 2005):

$$\pi_{t,k} = \exp(-Z_{t,k}),$$

$$\gamma_{t,k,j} = \frac{\delta_t^{(F)} F_{t,k,j}}{Z_{t,k}} (1 - \exp(-Z_{t,k})),$$

$$\rho_{t,k} = \frac{\delta_t^{(M)} M_{t,k}}{Z_{t,k}} (1 - \exp(-Z_{t,k})).$$

As discussed in the previous sections, positive correlation between individuals may arise as a consequence of social behaviour, such as schooling, and from varying environmental conditions that affect all the individuals in a same way. Such a joint variation can be modelled using a DM distribution (Appendix 1), which can be explicitly parameterized using the correlation κ between individuals

$$(c_{t,k,1}, \ldots, c_{t,k,j^*}, n_{t,k}^{(S)}, d_{t,k}) \sim \text{DM}\left(n_{t,k}^{(G)}, (1/\kappa - 1)(\gamma_{t,k,1}, \ldots, \gamma_{t,k,j^*}, \pi_{t,k}, \rho_{t,k})\right),$$

where $c_{t,k,j}$ denotes the number of individuals of length class k caught by fleet j, $n_{t,k}^{(S)}$ is the number of individuals of length class k that survive to next time-step, and $d_{t,k}$ represents the number of individuals that died from natural causes.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

Reproduction

The number of eggs spawned within each population is defined as the sum of eggs laid by each class of the population. For each class, the number of eggs produced is a function of the proportion of the population that spawns at that time-step, proportion of mature females, and number of eggs produced by a mature female. Each of these quantities can be given a prior distribution, which in turn may depend on other variables such as female weight and relative fecundity. The prior can have a hierarchical structure with hyperparameters and covariates. Spawning is assumed to take place at the beginning of a time-step, before growth. The number of eggs laid at spawning is calculated as

$$E_{t,1} = \delta_t^{(E)} \sum_{k=1}^{k^*} n_{t,k} m_{t,k} f_{t,k} r_{t,k} w_{t,k},$$

where $m_{t,k}$ is the proportion of mature individuals in size class k, $f_{t,k}$ is the mean number of eggs per unit of female weight, $r_{t,k}$ is the proportion of females in size class, $w_{t,k}$ is the mean weight of females and $\delta_t^{(E)}$ is the proportion of mature population that spawns at time t. The proportion of females can be made a function of other variables in the model. It is also possible to model males and females separately as subpopulations.

Depending on the chosen length of time-steps, and length intervals I_k, recruits resulting from a specific spawning time can be added to population at the beginning of the next time-step, or the recruitment to the population can take place multiple time-steps after the spawning. To implement such a delay without breaking the Markov property of the model specification, the egg cohorts are tracked until they reach the desired age for recruitment:

$$E_{t+1,a+1} = E_{t,a}, \quad a = 1, \ldots, a^*.$$

The number of new recruits is then obtained as

$$R_{t+1} = s_t E_{t,a^*},$$

where s_t is the proportion of eggs that survive to recruits. Thus, eggs laid within some earlier time-step are used to predict the recruitment using a model for the survival of eggs in the same manner as in Mäntyniemi *et al.* (2013a), Methot and Wetzel (2013), and Pulkkinen and Mäntyniemi (2013). The survival function can take any functional form defined by the analyst (e.g. Ricker, Beverton–Holt, Shepherd, and Hockey-Stick), and uncertainty about the form can be accounted for by using Bayesian model averaging (Mäntyniemi *et al.*, 2013a; Pulkkinen and Mäntyniemi, 2013).

The recruits are added to the population at the beginning of each time-step, so the total population size at the beginning of time-step $t+1$ is $R_{t+1} + \sum_{k=1}^{k^*} n_{t,k}^{(S)}$ and the population state is obtained as

$$\mathbf{n_{t+1}} = (n_{t,1}^{(S)} + p_{t+1,1}^{(R)} R_{t+1}, n_{t,2}^{(S)} + p_{t+1,2}^{(R)} R_{t+1}, \ldots, n_{t,k^*}^{(S)} + p_{t+1,k^*}^{(R)} R_{t+1}),$$

where $p_{t+1,k}$ is the proportion of recruits that enter size class k. The prior distribution for these proportions must be case-specific to match the current understanding.

Movement between populations

When $G > 1$ subpopulations or spatial areas are included, a movement matrix $\mathbf{\Omega}_{k,t}$ must be specified for each class k (e.g. age, size, or stage). This matrix describes how large proportion of individuals of the class stay in their current subpopulation/area and how large proportion moves to each of the other populations/areas. Rows of the matrix must sum up to 1. The population state after the movement can then be obtained as

$$(n_{t+1,k,1}, \ldots, n_{t+1,k,G}) \times \mathbf{\Omega}_{k,t},$$

where $n_{t+1,k,g}$ denotes the number of individuals in size class k in subpopulation g after adding the new recruits to each of the subpopulations.

The elements of the matrix can be treated as fully known or they can be given a prior distribution. The matrix can stay constant over time or it may vary according to hierarchical model with hyperparameters and explanatory variables. If the populations represent fish of, say, different growth rates, then the movement matrix can be used to model the degree of heritability using an identity matrix for all other classes than the recruits.

Reparameterization

Typical fisheries data contain more information about relative temporal changes in population size than about the absolute abundance. Therefore, all the state variables are expected to have high posterior correlation with the overall abundance, which can be formulated as mean abundance or as the abundance of a chosen reference time-step or as a function of the carrying capacity of the environment (Millar and Meyer, 2000).

Similar kind of relationship is likely to exist between the population size at the beginning of time-step and the size structure of the population. Good information about the size distribution is often available, but the total population size is not well known.

The two points above suggest that it is beneficial to reparameterize the model, so that the temporal changes in abundance are treated as uncertain parameters scaled by a single parameter that represents the overall abundance. In the same manner, the state of the population should be seen as a product of total population size and a vector of proportions of size classes.

We define the temporal change in total population size by

$$N_{t+1} = N_t q_t + R_{t+1},$$

where N_t is the population size relative to the population size at the beginning of the first time-step $\left(N^* = \sum_{k=1}^{k^*} n_{1,k}\right)$, q_t is the proportion that survives to the next time-step (derived later), and R_{t+1} is the number of recruits relative to N^*. A prior must be assigned to N^*. All subsequent population sizes are defined by N^* through the above deterministic equation. The stochastic transition is then modelled by defining the recruitment and survival processes as nonlinear Markovian stochastic processes.

We start the derivation of the model by defining the state of the population at the beginning of a time-step in terms of relative size class frequencies $\mathbf{\Phi_t} = (\phi_{t,1}, \ldots, \phi_{t,k^*})$ and the total population size $N_t N^*$. After growth, the vector of probabilities that describe the chance to find a randomly chosen individual in any of the length classes can be computed as

$$\mathbf{\Phi_t^{(G)}} = \mathbf{\Phi_t} \times \mathbf{g_t}.$$

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

The fishing and natural mortality processes further divide the potential fates of individuals to multiple categories. The probability vector for surviving to next time-step is given by

$$\mathbf{\Theta_t} = (\phi^{(G)}_{t,1}\pi_{t,1}, \ldots, \phi^{(G)}_{t,k^*}\pi_{t,k^*}),$$

the probability vector for getting killed by fishing fleet j is

$$\mathbf{\Lambda_{t,j}} = (\phi^{(G)}_{t,1}\gamma_{t,1,j}, \ldots, \phi^{(G)}_{t,k^*}\gamma_{t,k^*,j}),$$

and the probability vector for natural mortality is

$$\mathbf{P_t} = (\phi^{(G)}_{t,1}\rho_{t,1}, \ldots, \phi^{(G)}_{t,k^*}\rho_{t,k^*}).$$

The stacked vector $(\mathbf{\Theta_t}, \mathbf{\Lambda_{t,1}}, \ldots \mathbf{\Lambda_{t,j^*}}, \mathbf{P_t})$ then sums up to 1 and represents the fate of an individual that has been randomly selected from the population at the beginning of the time-step. The next step is to consider the situation in the end of the time-step in terms of the number of individuals belonging to each potential category. The number of individuals at size class that survive to the next time-step is represented by vector

$$\mathbf{n^{(S)}_t} = (n^{(S)}_{t,1}, \ldots, n^{(S)}_{t,k^*}),$$

the numbers at size class killed by fishing fleet j are

$$\mathbf{c_{t,j}} = (c_{t,1,j}, \ldots, c_{t,k^*,j}),$$

and the numbers at size of individuals that died for natural causes are

$$\mathbf{d_t} = (d_{t,1}, \ldots, d_{t,k^*}).$$

Now, assuming exchangeability between individuals, and assuming that the fates of individuals are correlated with coefficient κ, then the stacked vector of numbers at size class can be modelled using the DM distribution

$$(\mathbf{n^{(S)}_t}, \mathbf{c_{t,1}}, \ldots, \mathbf{c_{t,j^*}}, \mathbf{d_t}) \sim \mathrm{DM}\big(N^*N_t, (1/\kappa - 1)(\mathbf{\Theta_t}, \mathbf{\Lambda_{t,1}}, \ldots, \mathbf{\Lambda_{t,j^*}}, \mathbf{P_t})\big).$$

The distribution of the total number of survivors is then Beta-Binomial (Appendix 1)

$$\sum_{k=1}^{k^*} n^{(S)}_{t,k} \sim \mathrm{Beta-Bin}\Big(N^*N_t, (1/\kappa - 1)\sum_{k=1}^{k^*}\phi^{(G)}_{t,k}\pi_{t,k}, (1/\kappa - 1)\Big(1 - \sum_{k=1}^{k^*}\phi^{(G)}_{t,k}\pi_{t,k}\Big)\Big).$$

In other words, $E(q_t) = \sum_{k=1}^{k^*}\phi^{(G)}_{t,k}\pi_{t,k}$. The above Beta-Binomial distribution can be approximated very closely using a Beta-distribution (Appendix 3):

$$N_{t+1} = \sum_{k=1}^{k^*} n^{(S)}_{t,k} \approx q_t N^*N_t,$$

$$q_t \sim \mathrm{Beta}(E(q_t)\eta^*_t, (1 - E(q_t))\eta^*_t),$$

where $\eta^*_t = N^*N_t/(\kappa(N^*N_t - 1) + 1)$.

Given the number of survivors, the numbers at size of the survivors are again DM

$$\mathbf{n^{(S)}_{t,k}} \sim \mathrm{DM}\Big(\sum_{k=1}^{k^*} n^{(S)}_{t,k}, (1/\kappa - 1)\mathbf{\Theta}_t\Big),$$

which can be very closely approximated by using a Dirichlet distribution

$$\mathbf{n^{(S)}_{t,k}} \approx \boldsymbol{\theta_t}\sum_{k=1}^{k^*} n^{(S)}_{t,k},$$

$$\boldsymbol{\theta_t} \sim D(\mathbf{\Theta_t}\eta^*_t).$$

As can be seen from above, this model formulation allows one to write the population transition with a single multivariate distribution for which the probability mass function is known.

Tailoring the model to a specific case

The model above describes a generic stochastic population dynamics life cycle that may fit to a large variety of different populations. The process of tailoring the model to describe the knowledge of a particular population involves formulation of prior knowledge of the life history traits such as somatic growth rates, maximum size, size at maturation, relative fecundity, survival of eggs, and size of recruits. For example, A. Gårdmark *et al.* (unpublished material) used this modelling framework and structured subpopulations by age and defined subpopulations by different growth parameters. They tracked length at age for both groups and modelled the recruitment as a function of environmental covariates while allowing for uncertain autoregressive unexplained variation at the same time.

The second category of information includes knowledge of the exploitation process. For fishery, this includes background knowledge of the availability of fish to the fishery and about the size selectivity of the fishing gear. For Northern Baltic herring, the three fleets were assumed to have constant fishing mortality over all sizes, but the mechanical selectivities of the gears were given prior distributions based on literature and expert elicitation (A. Gårdmark *et al.*, unpublished material). This approach made it possible to account for the fact that most herring that go through the mesh die to their wounds.

The third task in each case study is to link the available data to relevant states of the population. Typically, the state of the population cannot be observed without an error, but can be seen through potentially selective sampling processes. The key is then to create appropriate statistical models to describe how the data become collected. As an example, A. Gårdmark *et al.* (unpublished material) developed observation models for total catch in numbers, age distribution from catch sampling for each fleet, and for population biomass observed in acoustic surveys. See also Appendix 4 and Whitlock *et al.* (2015) for examples about biological and observational models.

In addition to the herring case study, assessments of Bothnian Sea herring stock using this model were also reported to the ICES benchmark assessment of pelagic stocks in 2012 (ICES, 2012). The assumptions included, for example, autocorrelated recruitment deviations, density-dependent catchability, and structural uncertainty about the form of the stock–recruitment function. According to the results, catchabilities of all three fleets were estimated to be density-dependent to some extent and the Beverton–Holt stock–recruitment model gained almost 100% posterior probability.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

Discussion

The integrated state-space stock assessment framework that we have described in this study can be seen to merge existing approaches in a new way. The implementation within a general-purpose Bayesian estimation software is similar to Mäntyniemi *et al.* (2013a) and brings a huge variety of building blocks available for the stock assessment scientist: almost every population dynamic parameter can be modelled using a hierarchical model that uses environmental covariates, but still assumes existence of variation from other sources as in Mäntyniemi *et al.* (2013b). The unexplained variation can also have autocorrelation, which can be given a prior distribution. Uncertainty about functional forms and about covariates to be included or excluded can be taken into account using the Carlin and Chib (1995) approach to computation by expressing the model as a mixture of the alternative models as in Mäntyniemi *et al.* (2013b). This work combines the generic structure of the Stock Synthesis (SS) framework (Methot and Wetzel, 2013) with the flexibility in modelling the process errors and covariates described in Mäntyniemi *et al.* (2013a). The model described in Mäntyniemi *et al.* (2013a) was age-structured and tracked the mean length for each cohort using a growth model, but lacked the possibility to admit that not all fish have the same growth parameters. This is possible in SS (Methot and Wetzel, 2013; Taylor and Methot, 2013) and the same is now possible within the model presented here.

Here, we have formulated a generic state-space population dynamics model that can be efficiently reparameterized and closely approximated by using the scaled Dirichlet distribution. The derivation of process error was based on the degree of dependence between the individuals, a novel concept in population dynamics modelling.

However, the original model and its reparameterization make different assumptions about how the individuals are correlated, at least when size structure is used. The original formulation assumes that all fish within a size class are correlated, but independent of fish in the other size classes. This might make sense if the individuals show a very high degree of size aggregation. But, there is also a theoretical drawback: the discrete length classes were introduced in the first place to approximate the continuous length distribution. The approximation can be then made more precise by increasing the number of length classes. In the limit, each class contains only one or zero individuals, which means that all individuals would actually be independent of each other. In contrast, the approximation assumes that all fish are correlated irrespective of their size. The dependence is assumed to be similar in both the survival and growing processes, which may not be realistic: growth is likely less correlated than the survival.

Within the approximating model, the problem can be potentially fixed by expressing the population transition using a generalized DM mixture model, which would make it possible to have different degree of correlation in the growth process compared with the survival process. However, such a generalized covariance structure can lead to computational problems. In such a case the generalized DM mixture can be potentially approximated by a scaled generalized Dirichlet distribution (Tzu-Tsung, 1998) in the same way that the DM can be approximated using a scaled Dirichlet (Appendix 3).

The model structure was presented here as a size-structured process. However, the same equations can be applied also for other structuring variables. In principle, the only change needed is to define the growth matrix in a suitable way depending on the structuring variable. In an age-structured model, the growth matrix would be interpreted as ageing matrix, and in a stage-structured model, the growth matrix would take the role of the state transition matrix. If an age-structure is used, it is still possible to track the mean size of individuals at age using a suitable growth model, although the effects of size variation cannot be fully taken into account.

While the fact that the DM can handle the overdispersion compared with multinomial distribution with only one additional parameter is conceptually simple, this can also be seen as a restriction of the approach. The built-in assumption is that whatever causes the overdispersion, the resulting correlation of the underlying events is equal for all the categories. A counter example could be found in the context of the example of individuals moving and distributing to different areas. If, for example, one of the areas is a reproduction area and others are feeding areas, then the grouping behaviour might be stronger in the movement between feeding areas, but the movement to reproduction area might be more independent process for mature fish (or the other way around) and therefore less (or more) overdispersion could be expected in counts of individuals found in the reproduction area.

This type of situation could be modelled by breaking the process down to conditional DM distributions, where the population would be first split between reproduction and feeding areas with smaller overdispersion parameter then the further split within feeding areas could be modelled using a DM with higher overdispersion. The resulting count vector including all areas cannot be expressed as a DM distribution. However, based on the properties of the generalized Dirichlet distribution (Tzu-Tsung, 1998), it seems likely that the distribution of the count vector could have a closed form expression as a mixture of a generalized Dirichlet and a multinomial distribution. The generalized Dirichlet distribution has twice the number of parameters than a Dirichlet distribution, and can therefore represent more complex covariance structures (Tzu-Tsung, 1998).

Another direction for future research is to study if different simultaneous sources of overdispersion can be explicitly separated by using the DM distribution (or its potential generalization) in the same way as is possible with the negative binomial distribution (Linden and Mantyniemi, 2011). We took the first step on that direction by presenting the case where both the schooling behaviour and the temporal variation of the transition rate were affecting the process variance at the same time. This possibility to explicitly account for schooling behaviour in population dynamics also opens new research questions for basic research: how much do we know about the schooling behaviour of different species in different situations?

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 244706/ECOKNOWS project and from the Academy of Finland (through a grant to AK). However, the study does not necessarily reflect EC views and in no way anticipates the Commission's future policy in the area. We thank the two anonymous referees whose constructive comments greatly improved the quality of the manuscript.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

References

Bi, J. 2006. Sensory Discrimination Tests and Measurements Statistical Principles Procedures and Tables. Blackwell Publishing, Iowa, USA.

Buckland, S. T., Newman, K. B., Fernández, C., and Thomas, L. 2007. Embedding population dynamics models in inference. Statistical Science, 22: 44–58.

Bull, B., Francis, R., Dunn, A., and Gilbert, D. J. 2002. CASAL (C++ algorithmic stock assessment laboratory): CASAL user manual v1.02.2002/10/21. NIWA Technical Report 117, Wellington.

Carlin, B. P., and Chib, S. 1995. Bayesian model choice via Markov-chain Monte-Carlo methods. Journal of the Royal Statistical Society, 57: 473–484.

Dorazio, R. M., Martin, J., and Edwards, H. H. 2013. Estimating abundance while accounting for rarity, correlated behavior, and other sources of variation in counts. Ecology, 94: 1472–1478.

Hisakado, M., Kitsukawa, K., and Mori, S. 2006. Correlated binomial models and correlation structures. Journal of Physics A: Mathematical and General, 39: 15365.

Hulson, P. 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, 68: 1548–1557.

Hulson, P. F., Hanselman, D. H., and Quinn, T. J. 2012. Determining effective sample size in integrated age-structured assessment models. ICES Journal of Marine Science, 69: 281–292.

ICES. 2012. Report of the benchmark workshop on pelagic stocks (WKPELA 2012). ICES CM 2012/ACOM: 47.

Kuparinen, A., Mäntyniemi, S., Hutchings, J. A., and Kuikka, S. 2012. Increasing biological realism of fisheries stock assessment: Towards hierarchical Bayesian methods. Environmental Reviews, 20: 135–151.

Kvam, P., and Day, D. 2001. The multivariate polya distribution in combat modeling. Naval Research Logistics (NRL), 48: 1–17.

Lindén, A., and Mäntyniemi, S. 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology, 92: 1414–1421.

Madsen, R. E., Kauchak, D., and Elkan, C. 2005. Modeling word burstiness using the dirichlet distribution. *In* Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, pp. 545–552.

Magurran, A. E. 1990. The adaptive significance of schooling as an anti-predator defence in fish. Annales Zoologici Fennici, 27: 51–66.

Mäntyniemi, S., Haapasaari, P., Kuikka, S., Parmanne, R., Lehtiniemi, M., and Kaitaranta, J. 2013b. Incorporating stakeholders' knowledge to stock assessment: Central Baltic herring. Canadian Journal of Fisheries and Aquatic Sciences, 70: 591–599.

Mäntyniemi, S., and Romakkaniemi, A. 2002. Bayesian mark-recapture estimation with an application to a salmonid smolt population. Canadian Journal of Fisheries and Aquatic Sciences, 59: 1748–1758.

Mäntyniemi, S., Romakkaniemi, A., and Arjas, E. 2005. Bayesian removal estimation of a population size under unequal catchability. Canadian Journal of Fisheries and Aquatic Sciences, 62: 291–300.

Mäntyniemi, S., Uusitalo, L., Peltonen, H., Haapasaari, P., and Kuikka, S. 2013a. Integrated age-structured length-based stock assessment model with uncertain process variances, structural uncertainty and environmental covariates: Case of central Baltic herring. Canadian Journal of Fisheries and Aquatic Sciences, 70: 1317–1326.

Maunder, M. N., Deriso, R. B., and Hanson, C. H. 2015. Use of state-space population dynamics models in hypothesis testing: Advantages over simple log-linear regressions for modeling survival, illustrated with application to longfin smelt (*Spirinchus thaleichthys*). Fisheries Research, 164: 102–111.

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.

Millar, R. B., and Meyer, R. 2000. Bayesian state-space modeling of age-structured data: Fitting a model is just the beginning. Canadian Journal of Fisheries and Aquatic Sciences, 57: 43–50.

Newman, K. B., Buckland, S. T., Lindley, S. T., Thomas, L., and Fernández, C. 2006. Hidden process models for animal population dynamics. Ecological Applications, 16: 74–86.

Newman, K. B., Fernández, C., Thomas, L., and Buckland, S. T. 2009. Monte Carlo inference for state-space models of wild animal populations. Biometrics, 65: 572–583.

Pulkkinen, H., and Mäntyniemi, S. 2013. Maximum survival of eggs as the key parameter of stock-recruit meta-analysis: Accounting for parameter and structural uncertainty. Canadian Journal of Fisheries and Aquatic Sciences, 70: 527–533.

Richards, S. A. 2008. Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology, 45: 218–227.

Taylor, I. G., and Methot, R. D. 2013. Hiding or dead? A computationally efficient model of selective fisheries mortality. Fisheries Research, 142: 75–85.

Tzu-Tsung, W. 1998. Generalized Dirichlet distribution in Bayesian analysis. Applied Mathematics and Computation, 97: 165–181.

Warton, D. I., and Hui, F. K. C. 2011. The arcsine is asinine: The analysis of proportions in ecology. Ecology, 92: 3–10.

Xiao, Y. 2005. Catch equations: Restoring the missing terms in the nominally generalized Baranov catch equation. Ecological Modelling, 181: 535–556.

Xu, Z., and Akella, R. 2008. A New Probabilistic Retrieval Model Based on the Dirichlet Compound Multinomial Distribution. Proceedings of the 31st Annual International ACM SIGIR Conference, Singapore.

Appendix 1

Properties of the DM distribution

The DM distribution can be derived in number of ways (Madsen *et al.*, 2005; Bi, 2006). One way is to consider a hierarchical model where a proportion vector $(p_1, \ldots, p_k)$ follows a Dirichlet distribution

$$(p_1, \ldots, p_k) \sim \mathrm{Dir}(\eta(\mu_1, \ldots, \mu_k)),$$

where μ_i is the expected proportion of class $i = 1, \ldots, k$ and parameter η controls the covariance matrix of the distribution. The proportions are then used as parameters for a multinomial distribution

$$(x_1, \ldots, x_k) \sim \mathrm{Multi}(N, (p_1, \ldots, p_k)).$$

Integrating over the proportions p_i gives the DM distribution as the marginal distribution of the count vector, denoted as

$$(x_1, \ldots, x_k) \sim \mathrm{DM}(N, \eta(\mu_1, \ldots, \mu_k)).$$

The probability mass function of the DM distribution is

$$P((x_1, \ldots, x_k) | \eta, (\mu_1, \ldots, \mu_k)) = \frac{N!}{\prod_{i=1}^{k}(x_i!)} \frac{\Gamma(\eta)}{\Gamma(\eta + N)} \prod_{i=1}^{k} \frac{\Gamma(x_i + \eta\mu_i)}{\Gamma(\eta\mu_i)}.$$

The mean is $E(x_i) = \mu_i N$. The variance is $V(x_i) = \mu_i(1 + \mu_i)N(N + \eta)/(1 + \eta)$, which approaches the variance of the multinomial distribution as $\eta \to \infty$. The covariance is $\mathrm{COV}(x_i, x_j) = -\mu_i\mu_j N(N + \eta)/(1 + \eta)$, which also reduces to multinomial covariance in the limit.

The marginal and conditional distributions of elements and subvectors of the count vector are analytically available. The marginal

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

distributions of the elements are Beta-Binomial (Bi, 2006)

$$x_i \sim \text{Beta} - \text{Bin}\left(N, \eta\mu_i, \eta\left(\sum_{j=1}^{k}\mu_j - \mu_i\right)\right).$$

Vector elements can be combined, and the distribution of the resulting vector is DM. For example:

$$(x_1 + x_2, x_3 + x_4, x_5, \ldots, x_k) \sim \text{DM}(N, \eta(\mu_1 + \mu_2, \mu_3 + \mu_4, \mu_5, \ldots, \mu_k)).$$

Given the sum of a subvector, the counts in the subvector follow a DM distribution. For example:

$$(x_5, \ldots, x_k) \sim \text{DM}\left(\sum_{i=5}^{k} x_i, \eta(\mu_5, \ldots, \mu_k)\right).$$

Appendix 2

Growth matrix from the von Bertalanffy growth model

The probabilities $g_{t,i,j}$ are determined as follows. Denote by $L^{(i)}$ the length of individuals at time t, which were in length class i at time $t-1$. Assume then that $L^{(i)}$ is distributed as

$$L^{(i)} \sim N(\mu_L^{(i)}, \sigma_L^2),$$

for all $t = 1, 2, \ldots$. The parameters of the distribution are obtained from the von Bertalanffy (VB) growth model as

$$\mu_L^{(i)} = (L_\infty - l_i)(1 - e^{-k}) + l_i$$

and

$$\sigma_L^2 = \sigma_{L_\infty}^2(1 - e^{-2k}),$$

where L_∞ and $\sigma_{L_\infty}^2$ are the asymptotic expected value and variance, respectively, for the length distribution of old individuals, and $k > 0$ is a parameter controlling the growth rate. The Equation for the expected value is a straightforward application of the VB growth model, whereas the variance can be motivated by interpreting the VB model as an AR(1) process, see derivation below for details. Finally, denoting the cumulative distribution function of the normal distribution by Φ, the probabilities $g_{t,i,j}$ are obtained as

$$g_{t,i,j} = \frac{\Phi((I_{j+1} - \mu_L^{(i)})/\sigma_L) - \Phi((I_j - \mu_L^{(i)})/\sigma_L)}{\Phi((I_{m+1} - \mu_L^{(i)})/\sigma_L) - \Phi((I_1 - \mu_L^{(i)})/\sigma_L)}.$$

For the derivation of the transition variance σ_L^2, we consider the VB growth model

$$L_\tau = (L_\infty - L_{\tau-1})(1 - e^{-k}) + L_{\tau-1} + \varepsilon_\tau,$$

with an additive error term $\varepsilon_\tau \sim N(0, \sigma_L^2)$, for all $\tau = 1, 2, \ldots$.

Here, the time index τ refers to the growth history of an individual, so that L_τ is the length of the individual at age τ.

Assume now that $\tau > M$ for some constant M. By setting $c = (1 - e^{-k})L_\infty$ and $\varphi = e^{-k}$, the model can then be interpreted as an AR(1) process

$$L_\tau = c + \varphi L_{\tau-1} + \varepsilon_\tau,$$

for which $E(L_\tau) = L_\infty$ and $\text{Var}(L_\tau) = \sigma_{L_\infty}^2$ as $M \to \infty$. Finally, by standard properties of AR(1) processes, it follows that

$$\sigma_L^2 = \sigma_{L_\infty}^2(1 - e^{-2k}).$$

Appendix 3

Dirichlet approximation of the DM distribution

In this section, we show that the Dirichlet multinomial distribution can be approximated using a single Dirichlet distribution. Such an approximation may be necessary for computational reasons. The Dirichlet density is faster to evaluate than DM because it includes fewer calls to the gamma function and statistical software commonly have built in functions and sampling routines for the Dirichlet distribution while the implementation of DM is more rare. The idea is to express the count vector as the product of a proportion vector and total count, then find a Dirichlet distribution for the proportion vector using moment matching, so that the mean and covariance matrix of the resulting count vector matches the corresponding statistics of the original DM distribution.

First, we derive the mean vector and covariance matrix of proportions based on the DM distribution. The mean is given by

$$E\left(\frac{x_i}{N}\right) = \frac{1}{N}N\mu_i = \mu_i,$$

the variance is

$$V\left(\frac{x_i}{N}\right) = \frac{1}{N^2}N\frac{N+\eta}{1+\eta}\mu_i(1-\mu_i) = \frac{N+\eta}{N(1+\eta)}\mu_i(1-\mu_i),$$

and the covariance can be obtained as

$$\text{COV}\left(\frac{x_i}{N}, \frac{x_j}{N}\right) = -\frac{1}{N^2}N\frac{N+\eta}{1+\eta}\mu_i\mu_j = -\frac{N+\eta}{N(1+\eta)}\mu_i\mu_j.$$

Next, we define a proportion vector $(\phi_1, \ldots, \phi_k) \sim D(\eta^*(\mu_1, \ldots, \mu_k))$, where the mean vector is $(\mu_1, \ldots, \mu_k)$, the variance is given by $V(\phi_i) = \mu_i(1-\mu_i)/(\eta^*+1)$, and the covariance is $\text{COV}(\phi_i, \phi_j) = -\mu_i\mu_j/(\eta^*+1)$. Now, it is easy to see that setting $\eta^* = N(1+\eta)/(N+\eta) - 1$ provides exactly the same covariance matrix as is obtained from the DM distribution. As a summary, the DM count vector can be approximated by

$$(x_1, \ldots x_k) \approx N(\phi_1, \ldots, \phi_k).$$

The obvious drawback of the approximation is that 0 and N are not included in the support of the distribution. In predictive simulation, this can be mitigated to some extent by rounding. When calculating the likelihood for the parameters given a count vector, values close to 0 and N can be used for approximation.

Appendix 4

Simulation study

The data-generating and estimation models for the simulated data are age-structured with a maximum age of 12, 20 time-steps, and two fleets. The models account for process error in survival and recruitment. The general population dynamics model (GPDM) for

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

simulated data has observation models for the total catch, the proportion of the total catch accounted for by the landings and discards of each of the two fleets, and the age structure in the landed catch of the first fleet. Observation models for total catch and proportions of the catch from each fleet can be found in unpublished material (R. Whitlock *et al.*), while the observation model for the catch age composition is provided below. Observations were simulated assuming Beverton–Holt stock–recruitment dynamics, while uncertainty about the functional form of the stock–recruitment curve was admitted in the estimation model.

Data were simulated in two steps: first, a vector of parameters was simulated from the priors for fishing mortality and biological parameters. Annual total fishing mortality rates were given independent lognormal priors with a prior coefficient of variation (CV) of 0.15 (Figure A1a); maximum fishing mortality rates for each of the two fleets $F\max_t$ were then obtained by multiplying total annual fishing mortality by a vector of values for the proportion of the total fishing mortality accounted for by the first fleet, and its complement. The proportion of the total fishing mortality accounted for by the first fleet was assumed to decline approximately linearly over the 20-year period, from 0.66 in the first year to 0.20 in the final year (Figure A1b). The model is parameterized for a hake-like species: priors for biological parameters in simulations were the same as those used for the ECOKNOWS Cyclades mixed fishery case study, except stock–recruitment parameters, which were given priors corresponding to demographic equilibrium (i.e. steady-state population dynamics in the absence of fishing).

In the second step, population dynamics were projected forwards with the selected parameter vector and vectors of observations were sampled from their prior predictive distributions in the same way as was used for the parameters. The CV of the total catch in biomass was determined within the model for generation of observations.

We present results from two GPDM estimation runs with the same simulated dataset; one with and one without an observation model for the age composition of the landed catch of the first fleet. In both estimation model runs, the steepness of the logistic curves describing length-specific discard probabilities was fixed to −1 for both fleets, while lengths corresponding to a 50% discarding probability were estimated. Mechanical selectivity parameters were fixed to the values used to simulate observations in the run without age composition data, and estimated for the first fleet in the run with age composition data. The full model specification used for the

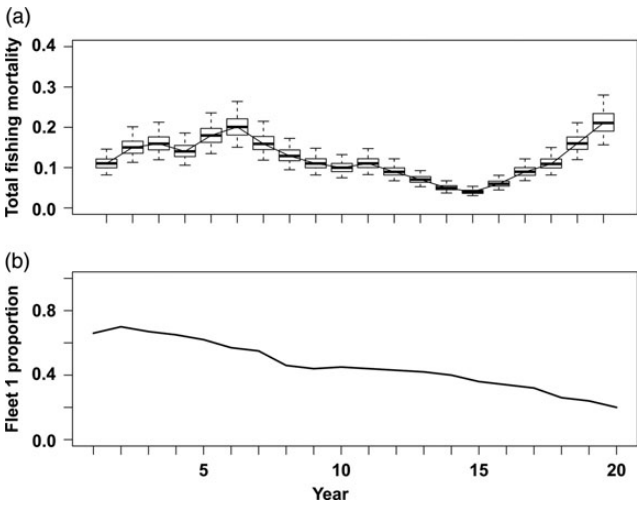

Figure A1. Prior distribution of fishing mortality used in the data generation (a) and the assumed proportion of fleet1 (b).

Table A1. Summary of posterior distributions for biological and selectivity parameters from runs of the GPDM using simulated data, with and without an observation model for the age composition of the first fleet's catch.

Parameter name	Parameter	Simulated data value	Prior	Posterior without age composition observation model	Posterior with age composition observation model
Natural mortality	M	0.25	0.18 (3.6×10^{-02})	0.21 (3.2×10^{-02})	0.24 (3.2×10^{-02})
SR slope	α	3.5×10^{-6}	2.1×10^{-06} (2.0×10^{-6})	2.8×10^{-06} (3.6×10^{-07})	3.2×10^{-06} (4.4×10^{-07})
SR carrying capacity	K	0.58	1.3 (2.1)	1.5 (0.87)	0.77 (0.15)
Asymptotic length	L_∞	78	79 (4.2)	77 (3.4)	78 (2.1)
von Bertalanffy growth coefficient	k	0.13	0.12 (8.8×10^{-03})	0.12 (4.7×10^{-03})	0.13 (4.6×10^{-03})
Length–weight parameter a	$\log(a_w)$	−5.4	−5.4 (4.4×10^{-02})	−5.5 (4.2×10^{-02})	−5.4 (2.4×10^{-02})
Length–weight parameter b	b_w	3.2	3.2 (2.4×10^{-02})	3.2 (2.1×10^{-02})	3.2 (2.0×10^{-02})
Length at 50% selectivity	L_1^{50}	10.1	11 (7.0)	–	11 (1.5)
Logistic selectivity slope	ν_1	0.34	0.46 (0.28)	–	0.40 (0.26)
Length at 50% discarding probability, fleet 1	DL_1^{50}	29	25 (15)	27 (1.1)	29 (1.0)
Length at 50% discarding probability, fleet 2	DL_2^{50}	15	25 (15)	15 (0.78)	15 (1.0)
Process error inverse variance parameter	η	38	20 (3.3×10^{02})	24 (5.7)	23 (4.8)
Initial population size	N_1	4.2×10^{07}	5.1×10^{07} (3.6×10^{07})	5.1×10^{07} (1.8×10^{07})	4.8×10^{07} (1.7×10^{07})

Posteriors are summarized as median and standard deviation in parentheses.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

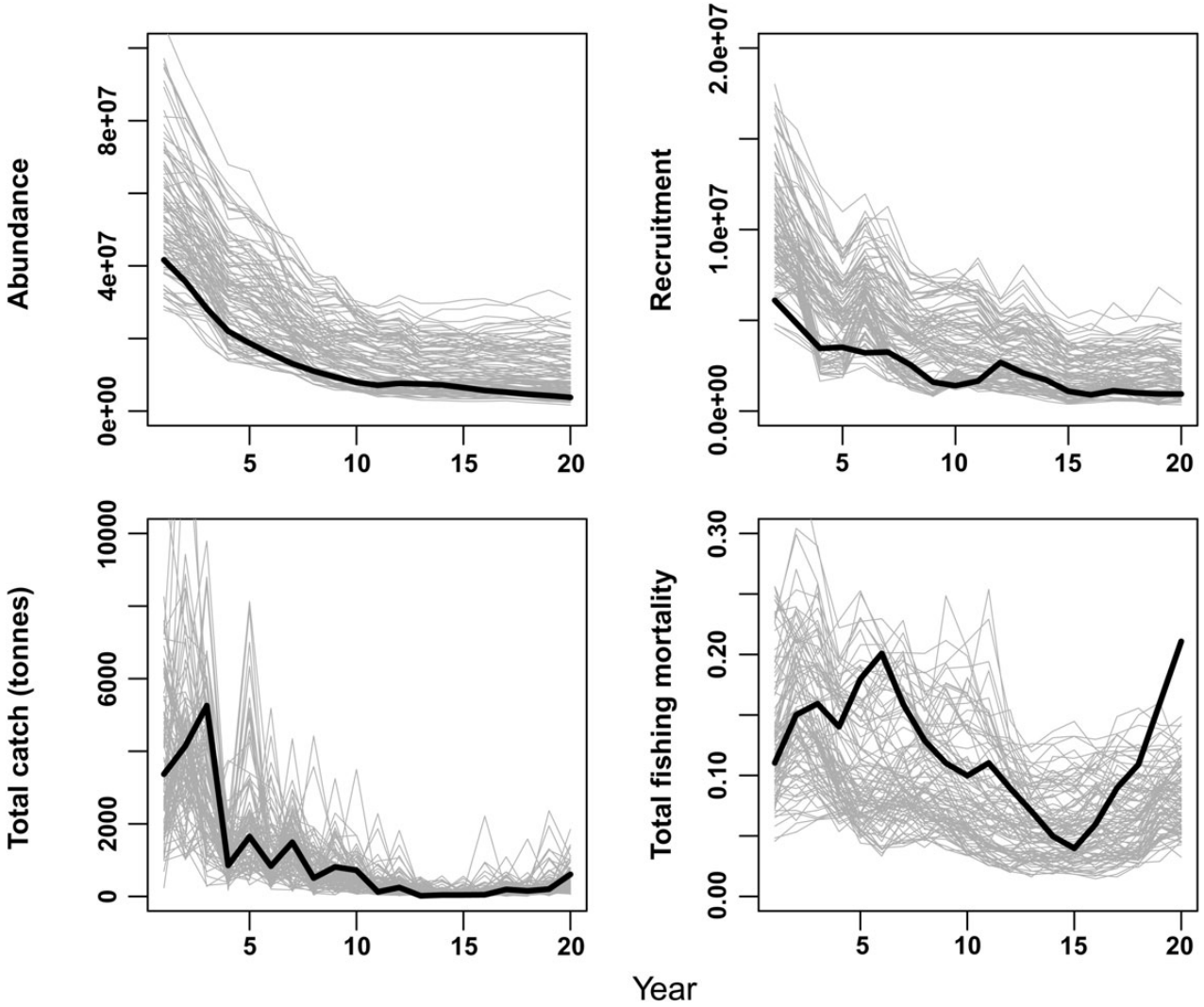

Figure A2. Posterior distributions of total abundance, recruitment, catch, and fishing mortality from a simulation study with an observation model for the total catch in biomass only. Each panel shows 200 trajectories randomly drawn from the posterior distribution (grey lines); the thick black line shows the true simulated values.

simulation study is provided in JAGS language (Supplementary data).

DM observation model for the landed catch age structure

The expected age distribution in the catch (retained fish) of fleet j is given by:

$$\delta_{t,a,j} = \frac{\gamma_{t,a,j}(1 - Dsel_{a,j})\phi_{t,a}/(1 - p_t)}{\sum_{a=1}^{A} \gamma_{t,a,j}(1 - Dsel_{a,j})\phi_{t,a}/(1 - p_t)},$$

where $\delta_{t,a,j}$ is the proportion of fish of age a in the landed catch of fleet j at time t and $\mathrm{Dsel}_{a,j}$ is an age-specific discard probability for fleet j. The proportions of fish of age a in the landed catch from fleet j, $\omega_{t,a,j}$ were assumed to follow a Dirichlet distribution:

$$\omega_{t,1:12,j} \sim \mathrm{Dirichlet}(a^{\omega}_{t,1:12,j}),$$

with shape parameters $a^{\omega}_{t,1:12,j}$ given by: $a^{\omega}_{t,a,j} = 100\delta_{t,a,j}$. Observed numbers of fish of different ages in the landed catch of fleet j were then assumed to follow a Multinomial distribution: $x_{t,1:12,j} \sim \mathrm{Multinomial}(\omega_{t,1:12,j}N_{t,j})$, where $x_{t,a,j}$ is the observed number of fish of age a in the landed catch of fleet j at time t, and $N_{t,j}$is the effective sample size of aged fish in the landed catch of fleet j at time t.

When the GPDM was fitted to simulated time-series of total catch and fleet proportion data, updating of priors for population dynamics and selectivity parameters occurred (Table A1). Addition of an observation model for the simulated age composition of one of the fleet's catches resulted in more precise and less biased estimates of abundance and recruitment (Figures A2 and A3); posteriors for biological parameters (e.g. natural mortality, stock recruitment carrying capacity (K), VB growth parameters and length–weight parameters) were also more precise and/or less biased (medians closer to the "true" simulated data value) when both catch volume and composition data were available (Table A1). In both data scenarios, the "true" stock recruitment functional form (Beverton–Holt) was given a posterior probability of 1, indicating that the catch data alone were informative about the form of the stock–recruitment relationship.

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018

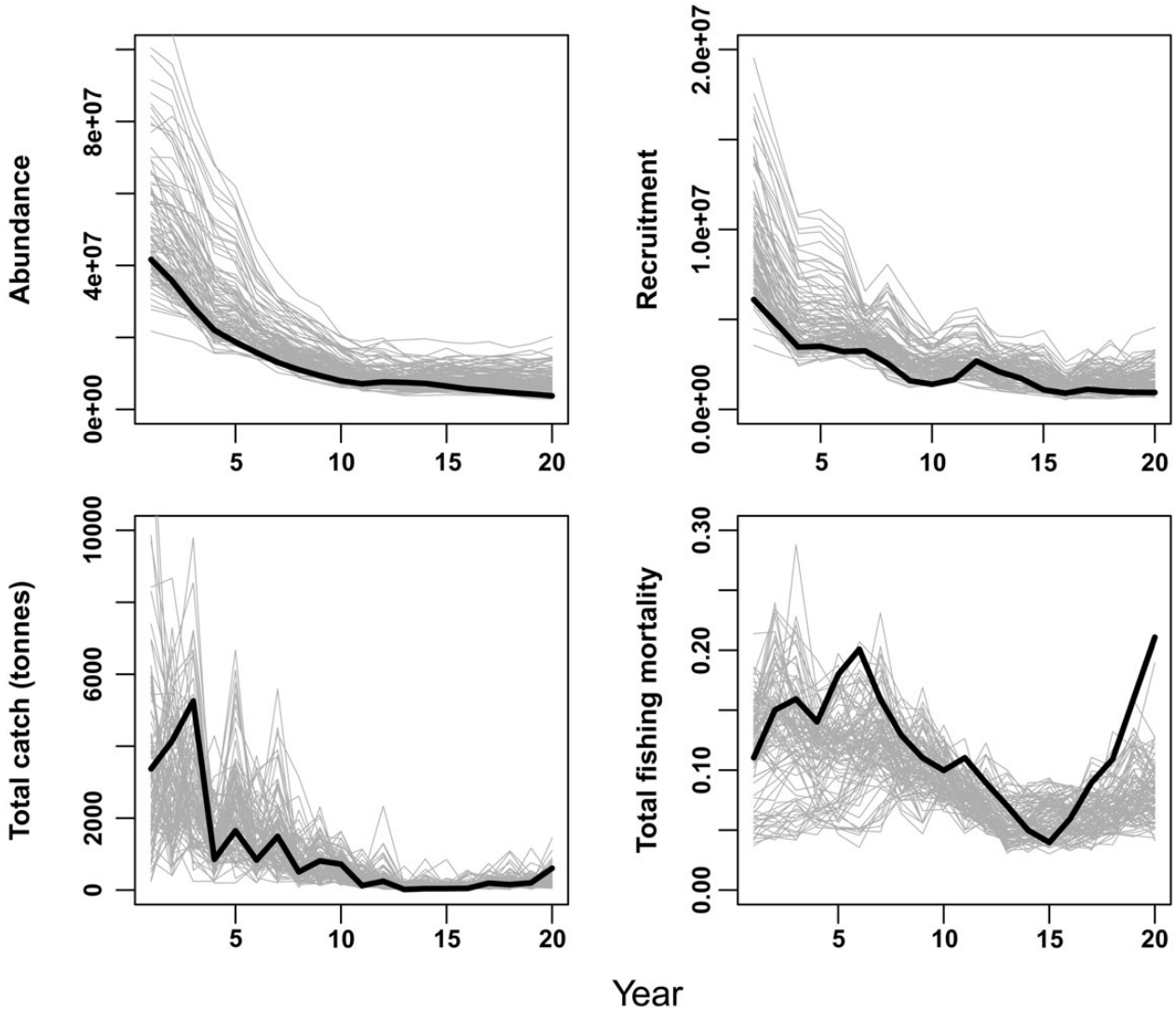

Figure A3. Posterior distributions of total abundance, recruitment, catch, and fishing mortality from a simulation study with observation models for total catch in biomass and age distribution in the catch. Each panel shows 200 trajectories randomly drawn from the posterior distribution (grey lines); the thick black line shows the true simulated values.

Handling editor: Shijie Zhou

Downloaded from https://academic.oup.com/icesjms/article-abstract/72/8/2209/2459021 by Biblioteca Tomás Navarro Tomás user on 11 October 2018