ICES Journal of
Marine Science

International Council for the Exploration of the Sea
Conseil International pour l'Exploration de la Mer

ICES Journal of Marine Science (2019), 76(4), 1189–1199. doi:10.1093/icesjms/fsy191

Original Article

# Intercalibration of survey methods using paired fishing operations and log-Gaussian Cox processes

Uffe Høgsbro Thygesen[1]*, Kasper Kristensen[2], Teunis Jansen[2,3], and Jan E. Beyer[2]

[1]*Department of Applied Mathematics and Computer Science, Technical University of Denmark, Building 303, 2800 Lyngby, Denmark*
[2]*National Institute of Aquatic Resources (DTU Aqua), Technical University of Denmark, Building 201, 2800 Lyngby, Denmark*
[3]*Greenland Institute of Natural Resources, Kivioq 2, Nuuk, Greenland*

**Corresponding author: tel: (+45) 45 25 30 60; e-mail: uhth@dtu.dk.*

Thygesen, U. H., Kristensen, K., Jansen, T., and Beyer, J. E. Intercalibration of survey methods using paired fishing operations and log-Gaussian Cox processes. – ICES Journal of Marine Science, 76: 1189–1199.

Received 6 February 2018; revised 13 November 2018; accepted 15 November 2018; advance access publication 8 January 2019.

We present a statistical method for intercalibration of fishery surveys methods, i.e. determining the difference in catchability and size selectivity of two methods, such as trawl gears or vessels, based on data from paired fishing operations. The model estimates the selectivity ratios in each length class by modelling the size distribution of the underlying population at each station and the size-structured clustering of fish at small temporal and spatial scales. The model allows for overdispersion and correlation between catch counts in neighbouring size classes. This is obtained by assuming Poisson-distributed catch numbers conditional on unobserved log-Gaussian variables, i.e. the catch is modelled using log-Gaussian Cox processes. We apply the method to catches of hake (*Merluccius paradoxus* and *M. capensis*) in 341 paired trawl hauls performed by two different vessels, viz. the RV Dr Fridtjof Nansen and the FV Blue Sea, operating off the coast of Namibia. The results demonstrate that it is feasible to estimate the selectivity ratio in each size class, and to test statistically the hypothesis that the selectivity is independent of size or species. For the specific case, we find that differences between size classes and species are statistically significant.

**Keywords:** intercalibration, log-Gaussian Cox processes, mixed-effects models, selectivity

## Introduction

Fishery independent surveys are of pivotal importance for fish stock assessments, where they provide a relative abundance index, as well as for basic biological research (Millar, 1992). While the objective of a survey is to assess the abundance of the underlying population, it only provides a filtered view, specified by the selectivity of the operation. The vessels, riggings, and gears applied in these surveys often develop or shift over time, as do fishing methods by captains (Weinberg and Kotwicki, 2008), leading to changes in size selectivity and overall catch efficiency (Miller, 2013; Thorson and Ward, 2014). To maintain as long-time series as possible, it is often desireable to combine information from different operations. However, differences in selectivity of vessel–gear combinations must be accounted for before time series and spatial distribution data can be combined and synthesized, which can be problematic (Axelsen and Johnsen, 2015). To this end, dedicated experiments may be performed, involving two or more vessel–gear combinations, with the objective of calibrating these combinations against each other, i.e. intercalibration. Here, the difference in catch rates are investigated by performing pairwise near-simultaneous hauls in the same area, so as to minimize the time-space variation of the fished population between the hauls. With such data, the selectivity ratios, which measure the efficiency of the two vessel–gear combinations against each other, can be estimated for each species and each size class. Then, these selectivity ratios can be used as calibration factors by adjusting catches from one type of operation so that they are comparable with the catches from the other operation (Kotwicki *et al.*, 2017).

Multiple calibration procedures have been proposed and applied over time, in particular differing in how the size dependency in selectivity ratios is modelled and estimated. When considering the selectivity curve of a single gear, a common choice is to restrict attention to a parametric family of curves; for example logistic functions for towed gear and Gaussian functions for gill nets (Millar and Fryer, 1999). When comparing two gears, a typical choice has been to use polynomials in length to describe the ratio between the two selectivity curves (Lewy *et al.*, 2004; Millar *et al.*, 2004; Holst and Revill, 2009; Kotwicki *et al.*, 2017). The coefficients in these polynomials may be estimated in a GLM framework, but a point of particular importance is to allow for overdispersion relative to Poisson counts (Lewy *et al.*, 2004). This overdispersion arises for many reasons, including between-haul variation in the selectivity (Millar, 1993). If this effect is ignored, and catches from different hauls are pooled, it will lead to overconfidence in the accuracy of estimates; a remedy is to use a double bootstrap to assess the accuracy of estimates (Millar, 1993; Sistiaga *et al.*, 2016). An alternative is a GLMM approach where the relative selectivity curves are allowed to vary between hauls; either non-parametrically using autoregressive processes (Cadigan et al., 2006) or parametrically in terms of shifting and scaling slope base curves (Cadigan and Dowden, 2010). Alternatives to fixed polynomials include orthogonal polynomials, GAMs, or smooth-curve mixed models (Fryer *et al.*, 2003; Miller, 2013). A typical problem of these data is the large number of zero catches; therefore Thorson and Ward (2014) considered delta-GLMM's, where the probability of zero catch is explicitly modelled. Kotwicki *et al.* (2017) compared three models, two of which included polynomials to account for the dependence on length, and one which used GAM's to this effect, and advocated cross-validation techniques to select the best fitting model for a given data set.

When the original assumption is that the catch in each size class and in each haul is Poisson-distributed conditional on the abundance, a common approach is to condition on the total catch in each size class. Then, the catch in the individual haul is binomially distributed (Millar, 1992). Conceptually, a related approach is the beta regression, in which a ratio of catches per unit effort in each size class is assumed to be beta distributed (Kotwicki *et al.*, 2017).

A common phenomenon for size structures in catches is that not only are the numbers in each length group overdispersed, but there is also strong tendency for positive correlations between nearby size classes in the same haul (Pennington and Vølstad, 1994; Kristensen *et al.*, 2014). If not taken into account, this phenomenon means that fluctuations across size classes in raw selectivity ratios will be over-interpreted. Pragmatically, the consequence of this is that estimated selectivity ratio curves should be smoothed, but preferrably, the size correlations should be included in the statistical model structure. This ensures that the model describes the fluctuations in data adequately which is a prerequisite for the statistical analysis to be valid.

Overdispersion and correlation in count data are, in general, conveniently modelled using compound Poisson distributions. These are hierarchical models, where it is assumed that the random data are generated through a two-stage procedure: In the first stage, a random *intensity* is generated for each data point. In the second stage, this intensity is used as the mean value for Poisson variables which constitute the count data. With this construction, the variance of the random intensity yields overdispersion relative to Poisson data, while the correlation structure of the intensity cascades to the count data. A recent example of such a model structure is Miller et al. (2018). A particular framework of interest is that of log-Gaussian Cox processes (Diggle *et al.*, 2013), where the log-intensity is a Gaussian process. Since a Gaussian process is fully described by its mean and covariance, this framework is highly operational and lends itself readily to computations. Log-Gaussian Cox processes have previously been applied to the spatio-temporal modelling of size-structured populations, where it has elucidated distributions of cod (*Gadus morhua*) in the North Sea (Lewy and Kristensen, 2009; Kristensen *et al.*, 2014), of whiting (*Merlangius merlangus*) in the Baltic (Nielsen *et al.*, 2014), of the larvae and juveniles of mackerel (*Scomber scombrus*) in the North Sea (Jansen *et al.*, 2012, 2015), and of shallow-water hake (*Merluccius capensis*) (Jansen *et al.*, 2016) and deep-water hake (*M. paradoxus*) (Jansen et al., 2017) in the Benguela current system.

Since log-Gaussian Cox processes proved suitable for these applications, it is natural to ask if the framework is also suitable for the problem of estimating selectivity ratios. The article addresses this question. When applying the framework of log-Gaussian Cox processes to the selectivity ratios, the unobserved size-dependent phenomena include the selectivity ratios, which is the primary object of inference, but also the local abundance present for each pair of operations, as well as aggregations that are specific to the individual operation. Each of these phenomena is characterized by a covariance structure, which describes both the magnitude of fluctuations and their persistence across size ranges. The construction is a fairly simple application of the log-Gaussian Cox framework, and has the appeal that we can specify the properties of the various processes affecting the catch, from which the properties of the log-intensity follow automatically.

In this article, we describe the framework and the resulting method. We demonstrate the method using data from a case where the objective was to investigate differences between two vessels which used gear with the same specifications: The RV Dr Fridtjof Nansen and the FV Blue Sea, which have been used for surveying the stocks of hake in Namibian waters. The objective of the analysis is to estimate the selectivity ratios between the two vessels, including confidence intervals, and to test if the ratios depend on size and the particular hake species. In addition, we perform a simulation experiment to verify the model, test for significance of certain specific model components, and compare the full model with a simplified model where inference is conditional on total catch at length for each station.

# Methods

## Statistical model

Our method for intercalibration is based on a statistical model for the selectivity ratios which explains the size composition of the catch in survey operations, and in particular differences in this composition between operations conducted differently on the same fish population. For ease of reference, we refer to these operations as "hauls", whether the gear involved is, e.g. trawls, longlines, or gill nets. Similarly, we refer to differences between "gear", even if the actual differences between operations could also involve different vessels, personel, or procedures. The model is a non-linear mixed effect model involving both fixed effects parameters and random effects. We conduct inference in the model using numerical maximum likelihood estimation, employing the Laplace approximation (Kristensen *et al.*, 2016) to integrate out random effects.

The observed quantities are count data, $N_{ijk}$, which represents number of individuals caught at station $i = 1, \ldots, n_s$, with gear $j = 1, 2$, and in length group $k = 1, \ldots, n_l$. Thus, at each station $i$, two operations have been performed; one with each gear $j$, and the size distribution of the catch has been measured.

We assume that these catches depend on swept area $A_{ij}$ (or a similar measure of effort) and three sets of random variables, which all depend on the size class $k$: First, $\Phi_{ik}$ which for a given station $i$ characterizes the distribution across size of the population encountered by both hauls $j$. Second, haul-specific fluctuations $R_{ijk}$ in the size composition which we will term the "nugget effect" with a reference to geostatistics (Cressie, 1993; Petitgas, 2001) and elaborate on the following. Third, the relative selectivity $S_{jk}$ which is specific to the gear. Given these random variables $\Phi$, $R$, $S$, we assume that the count data are Poisson distributed:

$$N_{ijk}|\Phi, R, S \sim \text{Poisson}(A_{ij} \cdot \exp(S_{jk} + \Phi_{ik} + R_{ijk}))$$

The swept area $A_{ij}$ is a known input to the model. This is Cox model of catches, also referred to as a doubly stochastic Poisson model, in that the mean values for the Poisson variates are themselves random. The joint distribution of the processes $S$, $\Phi$, and $R$ is Gaussian, so that the entire model is a log-Gaussian Cox process (Møller *et al.*, 1998; Diggle *et al.*, 2013). We now describe the details of the processes $S$, $\Phi$, and $R$ (Figure 1).

First, the selectivity (on the log scale) $S_{jk}$ of gear $j$ in size group $k$ is the main object of interest. Since we do not know the actual size distribution of the stock, we cannot estimate the absolute selectivities $S_{1k}$ and $S_{2k}$ of the two types of gear, but only the relative selectivity, i.e. $S_{1k} - S_{2k}$. We therefore require

$$S_{1k} = -S_{2k}, \tag{1}$$

which allows us to focus on $S_{1k}$. This symmetric choice ensures

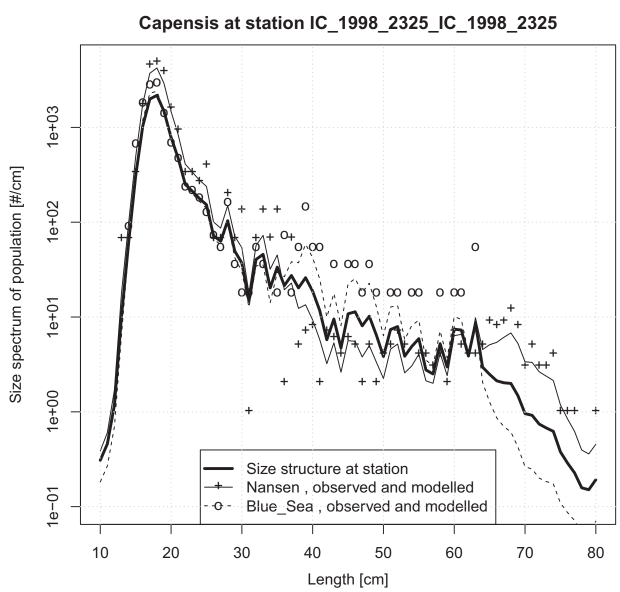


**Figure 1.** Example of data and model components. Estimated density $\exp(\Phi)$ of the size distribution at one particular station (thick solid lines). Different nugget effects $R$ apply to the two hauls and results in different size structures encountered by the two hauls (thin solid and dashed lines). The relative selectivity $S$ modifies the expected catch in each size group and for each haul (not shown). Observed counts $N$ in each size group and in each haul are shown with "o" and "+", respectively. Note log scale on the count axis; zero catches are not shown.

that $N_{i1k}$ and $N_{i2k}$ are identically distributed, which ultimately implies that the estimated selectivities $S_{jk}$ simply change sign if the gears are relabelled.

We note an alternative would be to enforce $S_{1k} = 0$ and estimate $S_{2k}$. This would be reasonable when the first gear is a reference gear that we measure the second gear against. In that case the variance on $N_{i1k}$ would then be smaller than that on $N_{i2k}$, since $N_{i2k}$ would contain the extra variance component $S_{2k}$. This asymmetry would cascade to the estimates, so that the estimated relative selectivities depend on which gear is considered the reference gear. In the present study, we have no reason to consider the one gear a reference, and therefore we prefer the symmetric choice $S_{1k} = -S_{2k}$.

To interpret the selectivities $S_{jk}$, it is useful to momentarily disregard the nugget effect $R$. Then, conditional on $\Phi$ and $S$, the expected catches at station $i$ and in size class $k$ with the two types of gear are $A_{i1}\exp(\Phi_{ik} + S_{1k})$ and $A_{i2}\exp(\Phi_{ik} - S_{1k})$, respectively. Thus, $\exp(2S_{1k})$ is the ratio between the expected catch per unit effort with the two types of gear:

$$\exp(2S_{1k}) = \frac{\mathbb{E}\{N_{i1k}/A_{i1}|\Phi, S\}}{\mathbb{E}\{N_{i2k}/A_{i2}|\Phi, S\}}. \tag{2}$$

This ratio is termed the selectivity ratio (Kotwicki *et al.*, 2017). Since this ratio must be positive, and since we do not assume a particular parametric form, it is convenient to represent it on the log scale, i.e. in terms of the process $S$. We model $S_{1k}$ as a random walk in size $k$, i.e.

$$S_{1(k+1)} - S_{1k} \sim N(0, \sigma_S^2) \quad \text{for} \quad k = 1, \ldots, n_l - 1$$

and assume independence between increments. To ensure that the log-selectivity ratio $S$ is a well-defined stochastic process, we complement this recursion with initial conditions $S_{j1} \sim N(0, \sigma_1^2)$ where $\sigma_1$ is fixed at a "large" value 10, which from a practical point of view implies that the level of the estimated log-sensitivity ratio $S$ is not dictated by the prior model but rather by data.

Next, $\Phi_{ik}$ is a log-density which describes the size distribution of the fish caught at station $i$. Specifically, $A_{ij}\ \exp(\Phi_{ik})$ is the expected number of fish caught in size group $k$ at station $i$ with a hypothetical gear which averages the two gears $j = 1$ and $j = 2$, in absence of nuggets ($R = 0$).

We assume independence of size distributions at different stations, i.e. $\Phi_{ik}$ and $\Phi_{i'k'}$ are independent for $i \neq i'$. At each station $i$, we assume that the log-density of the size distribution is a random walk over size groups, i.e.

$$\Phi_{i(k+1)} - \Phi_{ik} \sim N(0, \sigma_\Phi^2) \quad \text{for} \quad k = 1, \ldots, n_l - 1,$$

and that these increments are independent. Thus, the prior on the log-density $\Phi$ is a standard random walk which enforces continuity; the most probable density is flat. We add initial conditions

$$\Phi_{i1} \sim N(0, \sigma_1^2)$$

with the same "large" standard deviation $\sigma_1 = 10$, so that the overall level of $\Phi$ is not dictated by the prior model but rather by the total catch. The parameter $\sigma_\Phi^2$ is estimated. Since we assume independence between stations, we do not attempt to model any large-scale spatiotemporal structure of the population. We note that this is the main difference between this model and the GeoPop model (Kristensen *et al.*, 2014), where emphasis is exactly on this spatiotemporal structure.

Finally, the haul-specific fluctuations $R_{ijk}$ are akin to the nugget effect in spatial statistics; i.e. they describe variability in the catch data on very small spatial and temporal scales. While the term "nugget" originates in applications to mining, where repeated measurements on the same location may hit or miss a nugget, the envisioned mechanism in survey operations is that the gear may hit or miss aggregations of fish such as schools or shoals, that have limited range in space and quickly form, move, dissolve, and regroup. Since the two hauls at one station have been performed at slightly different locations and times, they will encounter different aggregations, and therefore $R_{ijk}$ and $R_{i'j'k'}$ are independent unless $(i, j) = (i', j')$, i.e. the same haul. Thus, at a given station $i$ and in a given size class $k$, $\Phi_{ik}$ models the population that is common to the two hauls, while $R_{ijk}$ models independent components which are distinct to each haul. We think of the aggregations giving rise to the nugget effect $R_{ijk}$ as size structured, and therefore, for a given haul $(i, j)$ and as a function of size $k$, the nugget effect arises as the sum of a white noise process and a zero-mean first-order autoregressive process. Specifically,

$$R_{ijk} = R_{ijk}^{\mathrm{WN}} + R_{ijk}^{\mathrm{AR}}$$

where $R_{ijk}^{\mathrm{WN}} \sim N(0, \sigma_{\mathrm{WN}}^2)$ and are independent. In turn $R_{ijk}^{\mathrm{AR}} \sim N(0, \sigma_{\mathrm{AR}}^2)$ and are independent for different stations $i$ or gear $j$, but correlated between size classes at a given station $i$ and gear $j$ so that $\mathbb{E}(R_{ijk}^{\mathrm{AR}} R_{ijk'}^{\mathrm{AR}}) = \sigma_{\mathrm{AR}}^2 \rho^{|k-k'|}$. The white noise component allows overdispersion relative to Poisson without correlation, while the autoregressive component models the size-specific clustering: If a particular size group is more abundant in the haul than expected, we would expect the same to apply to nearby size groups but not necessarily to very different size groups. We note that this same model structure was used by Cadigan et al. (2006) with the same motivation, but also that the effect could equally well represent other differences between the individuals hauls, e.g. differences in the way the gear is deployed, or combinations of such differences.

The model has five fixed effects parameters which are estimated, viz. the variance parameters $\sigma_{\mathrm{S}}^2$, $\sigma_{\Phi}^2$ $\sigma_{\mathrm{WN}}^2$, $\sigma_{\mathrm{AR}}^2$, and the correlation $\rho$. In addition there are a large number of random effects: $\Phi$ has $n_\mathrm{s} n_\mathrm{l}$ variables, $S$ has $n_\mathrm{l}$, and $R$ has $n_\mathrm{s} 2 n_\mathrm{l}$.

## Implementation

The statistical model in the previous section defines the joint distribution of the count data, $N$, and the unobserved random variables $\Phi$, $R$, $S$, for given parameters $\sigma_{\mathrm{S}}$, $\sigma_{\Phi}$, $\sigma_{\mathrm{WN}}$, $\sigma_{\mathrm{AR}}$, and $\rho$. The unobserved $\Phi$, $R$ and $S$ are integrated out using the Laplace approximation, to yield the likelihood as a function of the five parameters. The likelihood function is maximized to yield estimates of the five parameters, after which the posterior modes of the random effects $\Phi$, $R$, and in particular $S$ are reported.

The computations are performed in R version 3.1.2; we use the Template Model Builder (TMB) package (Kristensen *et al.*, 2016) for evaluating the likelihood function and its derivatives, and in particular for integrating out unobserved random variables using the Laplace approximation. Typical run-times for the models considered in this article, where there are 77 680 random effects, are 25 s on a standard laptop computer. The code is available at GitHub in package github.com/Uffe-H-Thygesen/Intercalibration.

The code and the statistical model are verified by simulation. Briefly, we simulate 1000 realizations of random effects and data sets, adjusting the mean of the size distributions $\Phi$ so that the total catch in the simulated data sets are approximately 17 000 fish, which corresponds to the total catch in the case described in the following. For each realization, we re-estimate the parameters in the model and the log-selectivity ratios. The variance parameters $\sigma_{\mathrm{S}}^2$, $\sigma_{\Phi}^2$, $\sigma_{\mathrm{WN}}^2$, and $\sigma_{\mathrm{AR}}^2$ are estimated on the log scale. We construct $1\sigma$ confidence intervals for each of the five parameters using the estimated standard deviation as computed from the Hessian of the log-likelihood. Theoretically, these confidence intervals should contain the true parameters for 68% of the simulated data sets; we find that they do so for between 66% and 71% of the simulated data sets, except for the parameter $\log \sigma_{\Phi}^2$, where the coverage is only 48%. For this parameter, the low coverage is explained by a bias in the estimates: The mean estimate is 0.07 smaller than the true value, which should be compared with an estimated standard deviation which is also 0.07. While negative bias is not uncommon for maximum likelihood estimates of variance parameters, it could possibly be reduced with restricted maximum likelihood (REML) (Pawitan, 2001). We also constructed $2\sigma$ confidence limits, which should contain the true value in 95% of the runs, and find that they do so for between 86% and 96% of the simulated data sets. The relative uncertainties on the variance parameters $\sigma_{\mathrm{S}}^2$ and $\sigma_{\mathrm{AR}}^2$ (measured from the standard deviation on estimates) are 13% and 7%, respectively, with a bias which is an order of magnitude smaller. The relative uncertainty on $\rho$ is 2% with a bias of 0.2%. In roughly half the simulations, the model cannot identify the white noise component in the residuals and consequently estimates $\sigma_{\mathrm{WN}}^2$ to be very low ($\sigma_{\mathrm{WN}}^2/\sigma_{\mathrm{AR}}^2 < 10^{-5}$); in these cases, also the estimated variance on $\log \sigma_{\mathrm{WN}}^2$ is very large (i.e. $>10$) so that the confidence intervals still cover the true value. While the white noise component is effectively removed from the model through the estimation for these simulated data sets, the reduced model is estimated well. We note that such problems of estimating separate variance components in hierarchical models are not uncommon (Auger-Méthé *et al.*, 2016). With this caveat, the simulation experiments verifies the code and the model.

## Data

We apply the method to a case study involving two vessels, the Norwegian fisheries research vessel Dr Fridtjof Nansen and the commercial trawler F/V Blue Sea, conducting hake surveys in Namibian waters.

Following independence of Namibia in 1990, abundance of Namibia's hake stocks was monitored by trawl surveys conducted by the R/V Dr Fridtjof Nansen. From 2000 the Ministry of Fisheries and Marine Resources in Namibia (MFMR) conducted the surveys using the F/V Blue Sea. In 1998 and 1999, before the shift, extensive experiments were performed by completing the entire annual survey in parallel with both vessels. The two vessels used Gisund fishing gear and rigging following the same specifications; nevertheless, some difference in the performance of the gear must be anticipated (Weinberg and Kotwicki, 2008). The stations are mapped in Figure 2.

Catch data collected from these surveys were extracted from the NAN-SIS database in November 2014 (Strømme, 1992).

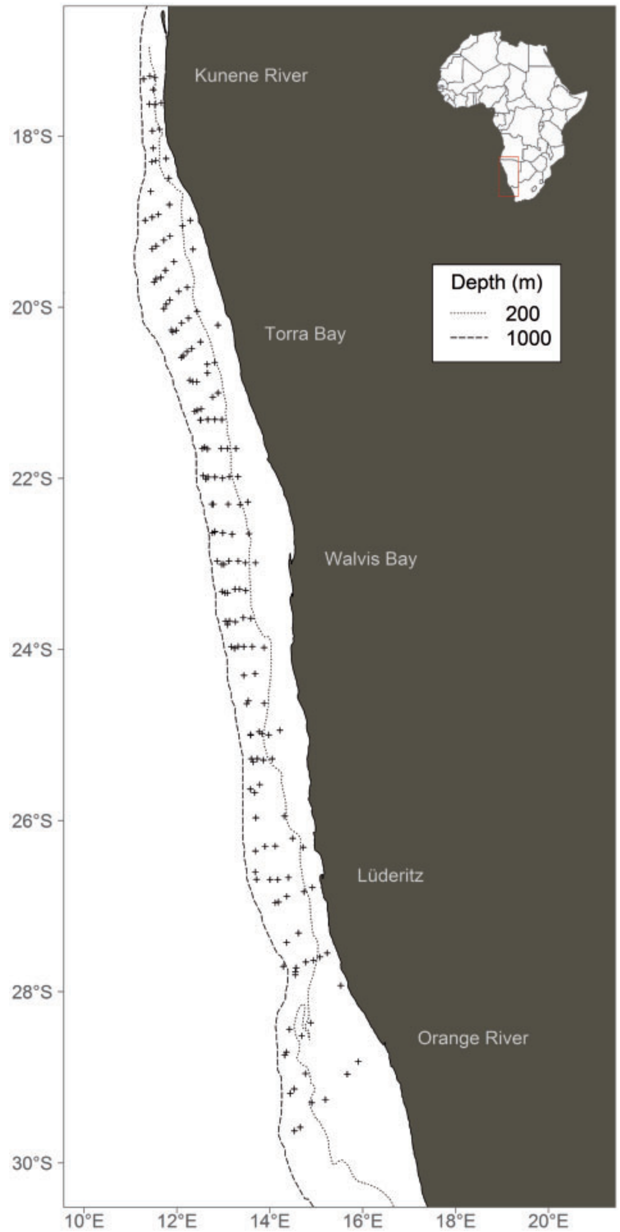


**Figure 2.** Map of the study area.

The analysis was based on 341 of the 365 pairs of trawl hauls. A total of 24 pairs were excluded because the trawl durations were less than 15 min and/or the difference in trawl durations exceeded 10 min.

Catch in numbers per length group and the hauling distance were available for each haul. Figure 3 shows all catches, summed over all stations, for the two species *M. paradoxus* (deep-water hake) and *M. capensis* (shallow-water hake). Since the two species have different preferred habitats but are morphologically very similar (Jansen *et al.*, 2016, 2017), a question of particular relevance is if the two species have the same selectivity.

## Results

Figure 4 shows the selectivity ratio from Equation (2), i.e. $\exp(2S_{1k})$, between the RV Dr Fridtjof Nansen and the FV Blue Sea. Index 1 corresponding to FV Blue Sea, so that a ratio above 1 indicates that the FV Blue Sea has higher expected catch than the RV Dr Fridtjof Nansen. Estimated parameters, including standard errors derived from the Hessian of the log-likelihood function, are shown in Table 1. Since the gears used on the two vessels have the same specifications, a reasonable hypothesis is that there is no size structure in these calibration factors. This hypothesis could be accepted for *M. capensis* (a likelihood ratio test of the hypothesis $\sigma_S = 0$ has critical significance level $p \sim 0.08$) but is rejected strongly for *M. paradoxus* ($p < 10^{-9}$). These *p*-values have been computed with the standard asymptotic $\chi^2$-distribution of the log-likelihood ratio, which does not strictly apply since the null hypothesis $\sigma_S = 0$ is on the boundary of the parameter space, so that the correct *p*-values may be somewhat smaller. It holds for both species that the FV Blue Sea is more efficient at catching larger hakes than the RV Dr Fridtjof Nansen. The size dependency is more pronounced for *M. paradoxus*, where the FV Blue Sea is less efficient in the small size classes. The selection of small *M. capensis* is similar for the two vessels. The estimated relative selectivity appears to fluctuate more between neighboring size classes for *M. paradoxus* than for *M. capensis*. This may be because the smaller catches of *M. capensis* imply less statistical certainty, so that the smooth prior is more visible in the estimates. It could also be connected to the observation that the estimated correlation $\rho$ is closer to 1 for *M. paradoxus* than for *M. capensis* so that small-scale fluctuations in the data are attributed to the nugget effect for *M. capensis* but, to a larger degree, to fluctuations in the relative selectivity for *M. paradoxus*.

Since there is no clear prior explanation why the selectivity curves for the two species would differ, a reasonable hypothesis is that they are identical. This hypothesis appears to be strengthened by the qualitative similarity between the estimated curves in Figure 4. This suggests to estimate a combined selectivity ratios for the two species, see Figure 5. In this combined model, we assume that the size distribution and the nugget effect applies to the two species separately, i.e. the small-scale clustering of fish is species specific. Since each fit yields a likelihood, it is possible to select between the two models (i.e. the two species have the same relative selectivity curve, or two different curves) using an information criterion such as that of Akaike, the AIC. The log-likelihood of the combined model is 258 less than that of the original model; this decrease results from the reduction of the number of parameters (fixed effects) from 10 to 5. Thus, the AIC will prefer strongly the model where the two species have separate selectivity ratios; for a likelihood ratio test, the critical *p*-value would be $10^{-108}$. We note that since the primary objective of inference is on the relative selectivity curves, which are random effects in the model, one could argue that model selection should be performed with the *conditional* AIC (Vaida and Blanchard, 2005). While the computation of the conditional AIC is a non-trivial task in our settings, a bound can be obtained by including the random effects in the degrees of freedom; this holds because each random effect in the cAIC framework is associated with a non-integer degree of freedom between 0 and 1. Then, the difference in log-likelihood should be compared with a maximum difference of 76 in the degrees of freedom, which would still favour strongly separate selectivity ratios for the two species. We conclude that the differences between the two species are statistically significant, even if the relative selectivity curves for the two species show similar qualitative features.

To illustrate the importance of the correlation between the different size classes, we fit a new model to this combined data set,

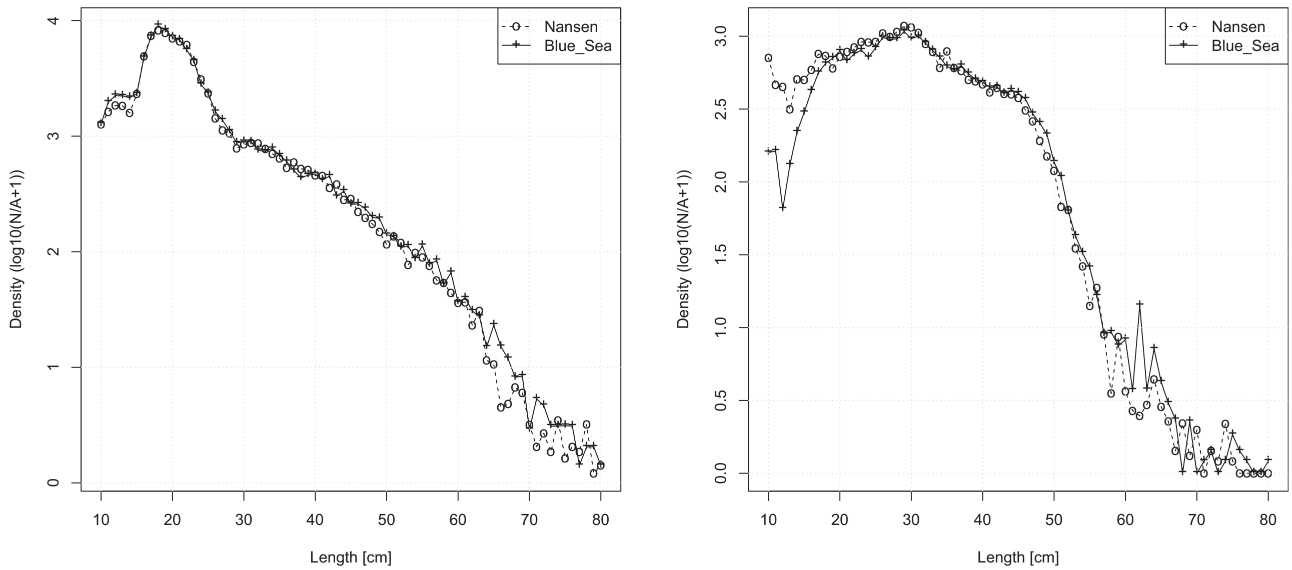


**Figure 3.** Density (total catch divided by swept area) by size, summed over all hauls. Left panel: *M. capensis*. Right panel: *M. paradoxus*.

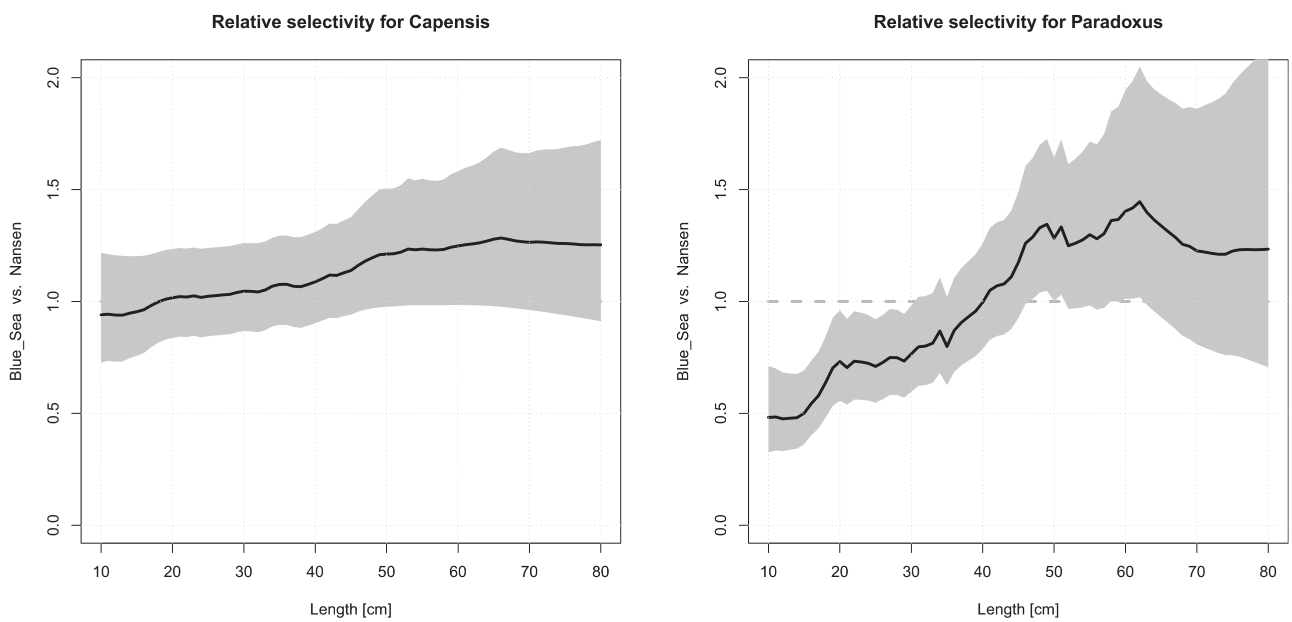


**Figure 4.** Relative selectivity (vessel calibration factor), comparing catches of *M. capensis* (left) and *M. paradoxus* (right) with Gisund gear on RV Dr Fridtjof Nansen and FV Blue Sea. Large values indicate that the FV Blue Sea has higher selectivity. Solid curve: Estimated relative selectivity (posterior mode). Grey region: Marginal 95% confidence intervals for the relative selectivity, computed as 1.96-$\sigma$-intervals on the log scale.

**Table 1.** Parameter estimate for the two species separately and combined, with estimated standard deviations. Included is also the negative log-likelihood and the number of parameters (fixed effects) of the model.

| Species | log $\sigma_\Phi$ | $\rho$ | log $\sigma_{WN}$ | log $\sigma_{AR}$ | log $\sigma_S$ | $-\log L$ | DF |
|---|---|---|---|---|---|---|---|
| *M. capensis* | $-0.15\pm0.02$ | $0.95\pm0.01$ | $-0.41\pm0.02$ | $-0.05\pm0.04$ | $-4.17\pm0.50$ | 44 940 | 5 |
| *M. paradoxus* | $-0.25\pm0.02$ | $0.98\pm0.01$ | $-0.95\pm0.03$ | $0.06\pm0.05$ | $-3.29\pm0.24$ | 34 607 | 5 |
| Sum | | | | | | 79 547 | 10 |
| Combined | $-0.19\pm0.01$ | $0.96\pm0.01$ | $-0.61\pm0.01$ | $0.01\pm0.03$ | $-3.68\pm0.24$ | 79 805 | 5 |
| Combined w/o $\rho$ | $-0.18\pm0.01$ | | $-0.10\pm0.01$ | | $-3.87\pm0.27$ | 82 417 | 3 |

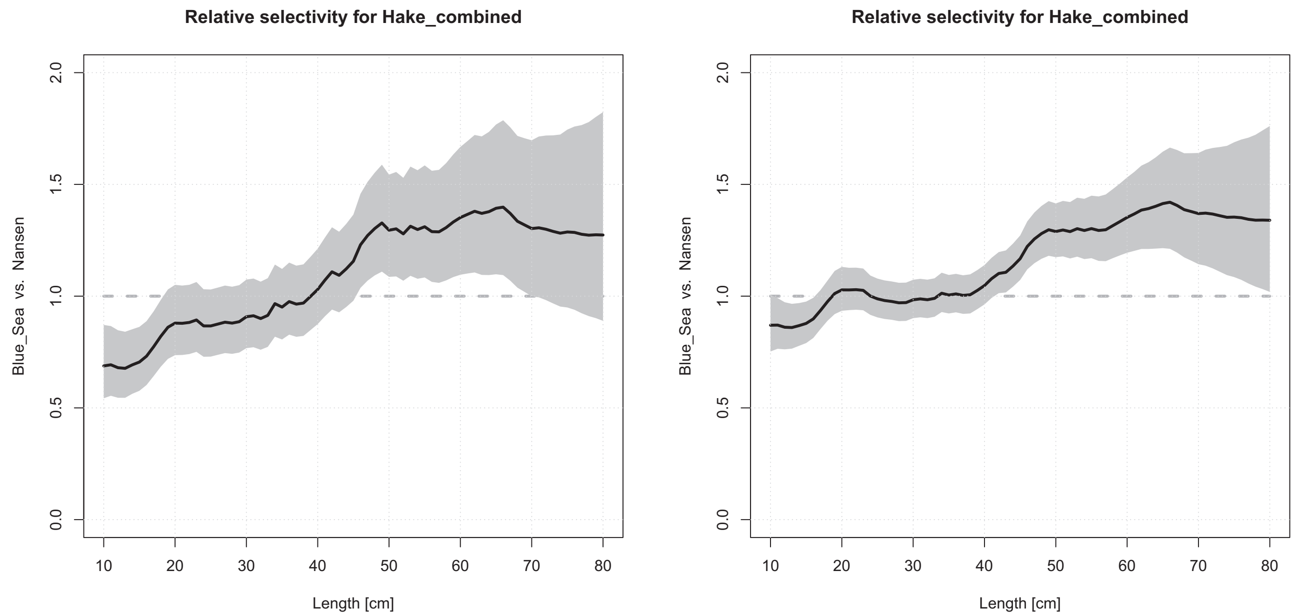


**Figure 5.** Left panel: Relative selectivity, as in Figure 4, for the two species combined. Right panel: Same, but without the autoregressive component in the nugget effect.

in which the autoregressive component of the nugget effect has been removed, so that the nugget effect acts independently at each size class (Figure 5, right panel). Removing this component from the model results in a decrease in the maximum log-likelihood of 2612 while decreasing the numbers of parameters by 2; thus this autoregressive component is extremely significant ($p \approx 10^{-1134}$). Nevertheless, the estimates from this reduced model agree qualitatively with the those from the model that includes autocorrelation in the nugget effect (compare Figure 5, left panel), although some minor differences are noticeable. Moreover, omitting the autocorrelation decreases the estimated variance associated with the selectivity ratio curves, so that the simpler model indicates higher accuracy than warranted.

Since several previous studies including Millar (1992), Lewy *et al.* (2004), and Cadigan and Dowden (2010) have considered a conditional approach, where inference is conditional on the total catch at length at each station, Appendix 1 compares such a conditional model with the model as described in "Statistical model" section. The two models give qualitatively similar results, but the estimated selectivity ratios from the unconditional model are generally closer to 1. The unconditional model has slightly narrower confidence intervals and is slightly more demanding in terms of computing time.

## Discussion

We developed a statistical method for intercalibrating survey gear and vessels, based on estimating the selectivity ratios from paired hauls. The method is directly available through an R package on GitHub. The envisioned application of our method is to adjust data obtained from multiple surveys, thus allowing them to be combined to yield a longer time series which may enter into a stock assessment. The adjustment would take place by multiplying the one series with the estimated selectivity ratios. The uncertainties on the estimated selectivity ratios would then propagate to the adjusted time series, for example using the delta method as implemented in TMB (Kristensen *et al.*, 2016). While one could envision integrated stock assessment models that use multiple raw survey indices as well as data from paired fishing operations, the preliminary step of adjusting and combining surveys appears to be preferable at least in the foreseeable future.

Our model is based on log-Gaussian Cox processes, which have been used earlier in the context of fisheries surveys to map spatiotemporal dynamics of stocks (Kristensen *et al.*, 2014; Jansen *et al.*, 2016), but not in the present way for comparing selectivities. The framework uses a non-parametric model for the relative selectivity and allows for overdispersion relative to the Poisson distribution, as well as correlations between size groups in paired trawl catches. These features all contribute to larger variability in data, and the Gaussian structure of the components simplifies analysis and computations. If the statistical analysis is based on models which fail to include such variance contributions, there is a risk that the confidence in the results are inflated, e.g. in the sense that confidence intervals appear narrower than justified. Such phenomena of overconfidence are well known, both in general statistics and in the specific context of selectivity studies (Fryer, 1991). They can be seen as a manifestation of the general bias-variance trade-off. Previous methods to address between-haul and within-haul variation include bootstrap (Millar, 1993; Sistiaga *et al.*, 2016) in addition to mixed effects models (Cadigan *et al.*, 2006). In the present study, an example of such overconfidence is seen in Figure 5, comparing the two panels, where the right panel is based on a simplified model in which the autoregressive component of the nugget effect has been removed. Recalling that a hypothesis test rejected this simplification, and noticing that the reduced model produces estimated confidence intervals which are considerably narrower, we can conclude that these confidence intervals give an overoptimistic view on the accuracy of estimates. This overoptimism can be attributed to the omission of an important variance component.

As another example of possible overconfidence, selectivity ratios can be modelled as constants which apply to all size classes, as size-dependent functions using parametric forms, or non-parametrically as we have done here. While specific parametric families of functions are convenient in the analysis, it is difficult to hypothesize a reasonable functional form prior to seeing the data. If a specific functional form is postulated, then it is likely that parameters in this form can be estimated with seemingly high accuracy. However, the sensitivity of the results to misspecification of the functional form needs to be taken into consideration which is not straightforward. As a result, we would be prone to overestimate our confidence in estimated selectivity ratio curves, by the same reasoning as in the previous paragraph. Thus non-parametric curves, such as the ones we provide in this study, involve the smallest number of assumptions and are the most conservative choice in the sense of not risking overinterpretation of data. For some applications it is convenient to report parametric forms. This would be a minor extension, technically, but a subsequent step of model validation needs to ensure that the parametric family is suitable. On the other hand, non-parametric estimates require some regularization to avoid erratic fluctuations in the estimated curves. Here, we have obtained this smoothness by using a random walk prior to the relative selectivity curve, which is a minimal way of enforcing continuity. An alternative is to use smooth basis functions or smoothing splines (Miller, 2013).

The core of our approach is to take into consideration the covariance between different size classes, both in the selectivity ratio curves that we aim to estimate, and in the catch data. Neglecting this covariance would require that data are binned into large size bins with sufficiently high catch numbers, so that we can estimate the selectivity in each bin without borrowing information from neighbouring bins. If the true selectivity ratios vary with size, this would lead to a classical trade-off between bias and variance of the estimates. Specifying the fluctuations between size classes, as we have done, bypasses this trade-off and will give consistent results regardless of how small-size bins are chosen. The crux of this approach is the correct specification of the covariance structure. Here, we have taken a conservative approach in that we model the log-densities $\Phi$ and the relative selectivity $S$ as random walks across size, which amounts to enforcing continuous dependency on size. In turn, the nugget effect is an autoregressive process. The effect of this structure is that large catches across size groups in a specific haul is attributed to high selectivity ($S$) or to high density at the station ($\Phi$), whereas an isolated peak in catch numbers at a given size range in a specific haul is attributed to size-specific shoaling aggregations, i.e. the nugget effect $R$.

In our model, the random walks have unbiased and identically distributed steps. One would expect that the selectivity ratios fluctuate more in those size classes, where the selectivity curve of each gear changes the most, and less for the large-size classes where both gears have full selectivity. Similarly, we would expect that the size distributions are skewed toward the smaller size classes. Thus, our model structure relies on simplifying assumptions, and we do not expect the model to fully describe all variability in the data. Nevertheless, our simulation study indicates that the model structure allows estimation of the selectivity ratios which is the objective of the model.

Inspecting the appearance of the nugget effect in the model, we see that it could equally well be interpreted as a factor that modifies the selectivity of the gear in the operation, although we interpret it as a factor affecting the local abundance. Such random fluctuations in selectivity have been considered previously (Fryer, 1991; Miller, 2013). Based on the information in data sets such as the present, the two effects are confounded (Cadigan and Dowden, 2010): It is not possible to tell if a high catch in one particular operation was because the gear encountered an aggregation, or because the gear functioned better than average in that operation. In both cases, the net effect is a larger variability between repeated hauls.

A key question that the model aims to answer is if the gear (or vessel) effect can be assumed to be identical for all size classes, and it is interesting to notice that this does not appear to be the case for *M. paradoxus*. Similarly, it is interesting that the two species appear to have different selectivity ratios. Although there is no single clear biological explanation for this, there will always be several minor differences in the nets, the rigging, and the way the hauls are performed, which can contribute to such differences (Weinberg and Kotwicki, 2008), keeping in mind the numerous processes that interact and influence the catchability. At the same time caution most be exercised: The results indicate that the size structure in the catches would be extremely improbable if the gear effect acted identically on all size classes, or identically to the two species, under the assumptions in the model. The result therefore hinges on the model representing the variability in catches correctly. While informal model checks suggest that this is the case, we have not performed a stringent model validation using, e.g. the techniques in (Thygesen *et al.*, 2017), as the computations would be prohibative. Thus, there is a risk that some overdispersion in the data is not included in the model, and that the apparent differences between size classes and species are artifacts of this overdispersion.

While our main motivation for investigating the relative selectivity is scientific surveys, another important area of application is the selectivity of commercial gear. Here, trade-offs between efficiency and environmental impact is one concern that motivates comparative studies of the selectivity of different gear (Sistiaga *et al.*, 2015; Vogel *et al.*, 2017).

An underlying assumption behind our analysis is that the two operations at a given station do not affect each other. This assumption conflicts somewhat with the requirement that the two operations are performed close to each other, both in space and time, so that it is plausible that they encounter the same population. In contrast, Lewy *et al.* (2004) focused on the disturbance effect that a first haul has on the local fished population, and the implications for the second haul. In the present study, none of the pairs in the available data set are exceedingly close, so it would be superfluous to include such effects. Nevertheless, when applying the method to other data sets, it would be possible to parametrize such an effect and include it in the model. A logical extension would be to let the variance on the nugget effect increase with the distance between the two operations in space and time; however, it may be difficult to identify such structures reliably. The limiting case of unpaired fishing operations (Sistiaga *et al.*, 2016) is straightforward to analyze with our present framework but we have not investigated the quality of the resulting estimates.

Several previous similar studies have used a conditional approach along the lines in Appendix 1. In the present study, we found that the estimates from the conditional and unconditional model differed somewhat with estimates from the unconditional model generally being closer to 1. The conditional model has fewer random effects, but computing times are becoming less important thanks to the efficiency of Template Model Builder. The unconditional model has the advantage that it is applicable also to data sets with unpaired, or partially paired, hauls, but it is conceivable that the prior model for the size distribution in the population (Φ) is more critical in such situations and would require further scrutiny.

## Conclusion

We have demonstrated the feasibility of estimating size-specific selectivity ratios from paired fishing operations, using conditional Poisson distributions while overdispersion and the covariance structure is modelled using unobserved random fields. These fields represent stock size composition, small-scale size-structured clustering, and gear selectivity. The Laplace approximation, implemented in TMB, allows us to integrate out the many unobserved random variables so that the model is computationally feasible. The model allows testing of various hypotheses using the likelihood ratio principle, and model selection using for example AIC. The model, of which an R implementation is publically available, yields non-parametric selectivity ratios, including confidence regions, which can be used to integrate survey catches obtained with different vessels or gear configurations.

## Acknowledgement

This work was supported by EuropeAid through the EcoFish project (CRIS Number C-222387).

## References

Auger-Méthé, M., Field, C., Albertsen, C. M., Derocher, A. E., Lewis, M. A., Jonsen, I. D., and Flemming, J. M. 2016. State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. Scientific Reports, 6: 26677.

Axelsen, B. E., and Johnsen, E. 2015. An evaluation of the bottom trawl surveys in the Benguela Current Large Marine Ecosystem. Fisheries Oceanography, 24: 74–87.

Cadigan, N. G., and Dowden, J. J. 2010. Statistical inference about the relative efficiency of a new survey protocol, based on paired-tow survey calibration data. Fishery Bulletin, 108: 15–30.

Cadigan, N. G., Walsh, S. J., and Brodie, W. B. 2006. Relative efficiency of the Wilfred Templeman and Alfred Needler research vessels using a Campelen 1800 Shrimp Trawl in NAFO Subdivision 3Ps and Divisions 3LN. Department of Fisheries and Oceans Canadian Science Advisory Secretariat Research Document 2006/085, St. John's, Newfoundland.

Cressie, N. A. C. 1993. Statistics for Spatial Data, 2nd edn. Wiley Series in Probability and Statistics. Wiley, New York.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. 2013. Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. Statistical Science, 28: 542–563.

Fryer, R. J. 1991. A model of between-haul variation in selectivity. ICES Journal of Marine Science, 48: 281–290.

Fryer, R. J., Zuur, A. F., and Graham, N. 2003. Using mixed models to combine smooth size-selection and catch-comparison curves over hauls. Canadian Journal of Fisheries and Aquatic Sciences, 60: 448–459.

Holst, R., and Revill, A. 2009. A simple statistical method for catch comparison studies. Fisheries Research, 95: 254–259.

Jansen, T., Kainge, P., Singh, L., Wilhelm, M., Durholtz, D., Strømme, T., Kathena, J., *et al.* 2015. Spawning patterns of shallow-water hake *(Merluccius capensis)* and deep-water hake *(M. paradoxus)* in the Benguela Current Large Marine Ecosystem inferred from gonadosomatic indices. Fisheries Research, 172: 168–180.

Jansen, T., Kristensen, K., Fairweather, T. P., Kainge, P., Kathena, J. N., Durholtz, M. D., Beyer, J. E., *et al.* 2017. Geostatistical modelling of the spatial life history of post-larval deepwater hake *Merluccius paradoxus* in the benguela current large marine ecosystem. African Journal of Marine Science, 39: 349–361.

Jansen, T., Kristensen, K., Kainge, P., Durholtz, D., Strømme, T., Thygesen, U. H., Wilhelm, M. R, *et al.* 2016. Migration, distribution and population (stock) structure of shallow-water hake *(Merluccius capensis)* in the Benguela Current Large Marine Ecosystem inferred using a geostatistical population model. Fisheries Research, 179: 156–167.

Jansen, T., Kristensen, K., Payne, M., Edwards, M., Schrum, C., and Pitois, S. 2012. Long-term retrospective analysis of mackerel spawning in the North Sea: a new time series and modeling approach to CPR data. PLoS One, 7: e38758.

Kotwicki, S., Lauth, R. R., Williams, K., and Goodman, S. E. 2017. Selectivity ratio: a useful tool for comparing size selectivity of multiple survey gears. Fisheries Research, 191: 76–86.

Kristensen, K., Nielsen, A., Berg, C. W., and Skaug, H. 2016. TMB: automatic differentiation and Laplace approximation. Journal of Statistical Software, 70.

Kristensen, K., Thygesen, U. H., Andersen, K. H., and Beyer, J. E. 2014. Estimating spatio-temporal dynamics of size-structured populations. Canadian Journal of Fisheries and Aquatic Sciences, 71: 326–336.

Lewy, P., and Kristensen, K. 2009. Modelling the distribution of fish accounting for spatial correlation and overdispersion. Canadian Journal of Fisheries and Aquatic Sciences, 66: 1809–1820.

Lewy, P., Nielsen, J. R., and Hovgård, H. 2004. Survey gear calibration independent of spatial fish distribution. Canadian Journal of Fisheries and Aquatic Sciences, 61: 636–647.

Millar, R. B. 1992. Estimating the size-selectivity of fishing gear by conditioning on the total catch. Journal of the American Statistical Association, 87: 962–968.

Millar, R. B. 1993. Incorporation of between-haul variation using bootstrapping and nonparametric estimation of selection curves. Fishery Bulletin, 91: 564–572.

Millar, R. B., Broadhurst, M. K., and Macbeth, W. G. 2004. Modelling between-haul variability in the size selectivity of trawls. Fisheries Research, 67: 171–181.

Millar, R. B., and Fryer, R. J. 1999. Estimating the size-selection curves of towed gears, traps, nets and hooks. Reviews in Fish Biology and Fisheries, 9: 89–116.

Miller, T. J. 2013. A comparison of hierarchical models for relative catch efficiency based on paired-gear data for US Northwest Atlantic fish stocks. Canadian Journal of Fisheries and Aquatic Sciences, 70: 1306–1316.

Miller. T. J., Deborah, R. H., Karen, H., Norman, H. V., Richard, T., Amber, D. Y., and Scott, M. G. 2018. Estimation of the capture efficiency and abundance of Atlantic sea scallops *Placopecten magellanicus* from paired photographic-dredge tows using hierarchical models. Canadian Journal of Fisheries and Aquatic Sciences (in press).

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. 1998. Log Gaussian Cox processes. Scandinavian Journal of Statistics, 25: 451–482.

Nielsen, J. R., Kristensen, K., Lewy, P., and Bastardie, F. 2014. A statistical model for estimation of fish density including correlation in size, space, time and between species from research survey data. PLoS One, 9: e99151.

Pawitan, Y. 2001. In All Likelihood: Statistical Modelling and Inference Using Likelihood. Oxford University Press, Oxford, UK.

Pennington, M., and Vølstad, J. H. 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics, 50: 725–732.

Petitgas, P. 2001. Geostatistics in fisheries survey design and stock assessment: models, variances and applications. Fish and Fisheries, 2: 231–249.

Sistiaga, M., Herrmann, B., Grimaldo, E., Larsen, R. B., and Tatone, I. 2015. Effect of lifting the sweeps on bottom trawling catch efficiency: a study based on the Northeast arctic cod *(Gadus morhua)* trawl fishery. Fisheries Research, 167: 164–173.

Sistiaga, M., Herrmann, B., Grimaldo, E., and O'Neill, F. 2016. Estimating the selectivity of unpaired trawl data: a case study with a pelagic gear. Scientia Marina, 80: 321–327.

Strømme, T. 1992. Software for Fishery Survey Data Logging and Analysis: User's Manual, vol. 4. Food and Agriculture Organization, Rome.

Thorson, J. T., and Ward, E. J. 2014. Accounting for vessel effects when standardizing catch rates from cooperative surveys. Fisheries Research, 155: 168–176.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kasper, K., and Anders, N. 2017. Validation of ecological state space models using the Laplace approximation. Environmental and Ecological Statistics, 24: 317–339.

Vaida, F., and Blanchard, S. 2005. Conditional akaike information for mixed-effects models. Biometrika, 92: 351–370.

Vogel, C., Kopp, D., Morandeau, F., Morfin, M., and Méhault, S. 2017. Improving gear selectivity of whiting (*Merlangius merlangus*) on board French demersal trawlers in the English Channel and North Sea. Fisheries Research, 193: 207–216.

Weinberg, K. L., and Kotwicki, S. 2008. Factors influencing net width and sea floor contact of a survey bottom trawl. Fisheries Research, 93: 265–279.

*Handling editor: Michael Pol*

# Appendix

## Conditioning on the total catch at length and station

We compare the model as described in "Statistical model" section with a variant where we condition on the total catch at length and station. Specifically, let $N_{i\cdot k} = N_{i1k} + N_{i2k}$ be the total catch at station i in length group $k$. Then the conditional distribution of the catch in the first haul, $N_{i1k}$ given this total catch $N_{i\cdot k}$ is binomial:

$$N_{i1k}|\Phi, R, S, N_{i\cdot k} \sim \text{Binom} \left( N_{i\cdot k}, \frac{A_{i1}\exp(S_{1k} + R_{i1k})}{A_{i1}\exp(S_{1k} + R_{i1k}) + A_{i2}\exp(S_{2k} + R_{i2k})} \right) \quad (3)$$

In turn, the probabilities of the total catches $N_{i\cdot k}$ are

$$N_{i\cdot k}|\Phi, R, S \sim \text{Poisson}(A_{i1}\exp(\Phi_{ik} + S_{1k} + R_{i1k}) + A_{i2}\exp(\Phi_{ik} + S_{2k} + R_{i2k})) \quad (4)$$

The joint density as developed in "Statistical model" section could therefore alternatively be written as a product of these binomial probabilities [Equation (3)], the Poisson probabilities [Equation (4)], and the prior density of the Gaussian processes $\Phi$, $R$, $S$. We may now condition the inference on the total catch $N_{i\cdot k}$ and thus remove the term in the joint density that originates from the total catches $N_{i\cdot k}$, i.e. the terms [Equation (4)]. Since the size distributions $\Phi$ do not enter into the conditional probabilities [Equation (3)], they only appear in the joint density through their prior distribution. Thus, the size distributions $\Phi$ vanish after integration, so they can be removed from the model.

Figure 6 shows the result from this modified model. The figure should be compared with Figure 4, which shows the corresponding results for the original model. Notice that the estimates of the selectivity curves are not completely identical, since the omitted term [Equation (4)] does depend on the selectivities, but still fairly similar. The estimates from the unconditional model are, in general, closer to 1, and the marginal confidence intervals are somewhat wider when conditioning on the total catches. This is not surprising, since this model excludes the information in the total catches. The conditional models has 20% less random effects, which allows faster computations, although our code does not fully exploit this.

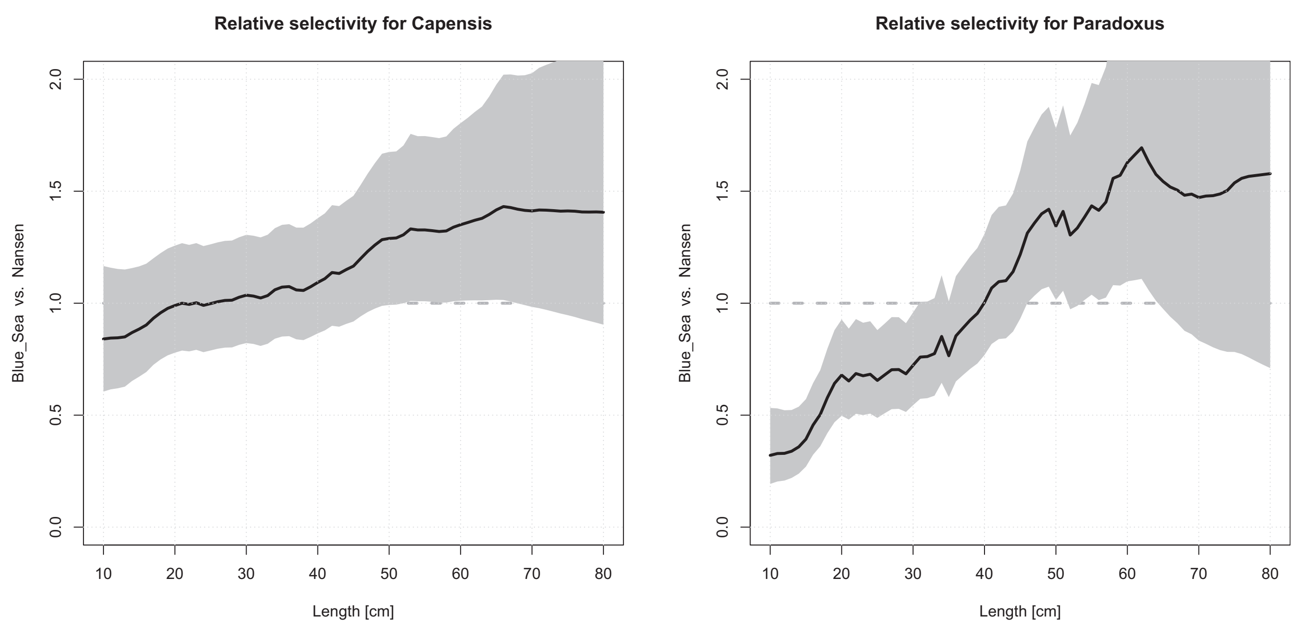


**Figure 6.** As Figure 4, but based on the model where we condition on the total catch in each length group, i.e. without the terms [Equation (4)] in the likelihood.