Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species

Rufener, Marie-Christine; Kristensen, Kasper; Nielsen, J. Rasmus; Bastardie, Francois

Published in: Ecological Applications

Link to article, DOI: 10.1002/eap.2453

Publication date: 2021

Document Version Peer reviewed version

Link back to DTU Orbit

Citation (APA): Rufener, M-C., Kristensen, K., Nielsen, J. R., & Bastardie, F. (2021). Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species. Ecological Applications, 31(8), [e02453]. https://doi.org/10.1002/eap.2453
Title:
Bridging the gap between commercial fisheries and survey data to model the spatiotemporal dynamics of marine species

Authors:
Marie-Christine Rufener
Kasper Kristensen
J. Rasmus Nielsen

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/EAP.2453

This article is protected by copyright. All rights reserved
Francois Bastardie

Affiliation:

Technical University of Denmark, Institute for Aquatic Resources, Section for Ecosystem based Marine Management, Kemitorvet, Building 201, 2800 Kgs. Lyngby, Denmark.

*Corresponding author:

mruf@aqua.dtu.dk

Manuscript received 27 April 2020; revised 8 February 2021; accepted 17 May 2021; final version received 10 September 2021.
Abstract
Monitoring and assessment of natural resources often require inputs from multiple data sources. In fisheries science, for example, the inference of a species’ abundance distribution relies on two main data sources, namely: commercial fisheries and scientific survey data. Despite efforts to combine these data into an integrated statistical model, their coupling is frequently hampered due to differences in their sampling designs, which imposes distinct bias sources in the estimator of the abundance distribution. We developed a flexible species distribution model (SDM) that can integrate both data sources while filtering out their relative bias contributions. We applied the model on three different age groups of the western Baltic cod stock. For each age group, we tested the model on (i) survey data and (ii) integrated data (survey + commercial) as a means to compare their differences and investigate how the inclusion of commercial fisheries data improved the spatiotemporal abundance estimator and parameter estimates. Moreover, we proposed a novel validation approach to evaluate whether the inclusion of commercial fisheries data in the integrated model is not in direct contradiction with the survey data. Following our approach, the results indicated that the use of commercial fisheries data is suitable for the integrated model. Across all age groups, our results demonstrated how commercial fisheries supplied additional information on cod’s spatiotemporal abundance dynamics, highlighting sometimes abundance hot-spots that were not detected by the survey model alone. Additionally, the integrated model provided a reduction of up to 20% and 10% in the uncertainty (std. error) of the predicted abundance fields and fixed-effect parameters, respectively. The proposed model represents a valuable benchmark for evaluating spatiotemporal dynamics of fish, and strengthens the science-based advice for marine policymakers.

Key-words: Fishery-dependent data, Fishery-independent data, Hierarchical model, Integrated analysis, Species Distribution Model (SDM), Template model builder (TMB).

Introduction

Understanding the patterns and processes that govern species spatiotemporal dynamics is a central concern in ecology (Fletcher et al. 2019). Such information is used not only for assisting the conservation and management of threatened species, but also for monitoring biodiversity and evaluating impacts of density-dependent and density-independent factors (Franklin 2010). Abundance
distributions are often inferred from empirical data under the assumption that collected data are unbiased. Although survey data are assumed to satisfy this requirement, high costs and laborious sampling leads researchers to use opportunistic data sources that are neither systematic nor randomly stratified, but are cheaper, easily accessible and usually more abundant (Isaac et al. 2019). This is no different in fisheries science, where information from commercial fisheries (fishery-dependent data) are commonly used to monitor commercially valuable fish species, especially when scientific surveys (fishery-independent data) are lacking. The advantages and risks of using these databases differ and can be discussed in terms of their quantity, quality and costs.

Survey data, for example, follow strict protocols to ensure a systematic and/or stratified random sampling design (NRC 2000, Nielsen 2015). This mainly involves the deployment of the same vessel with the same fishing gear over several decades (hence, constant catchability) and a standardized sampling effort (e.g. haul duration; Fig. 1). As such, they are assumed to provide unbiased information on a species’ abundance and its distribution. Yet, due to their costly nature, samplings are collected only a few weeks per year, thereby resulting in a reduced number of samples that imposes two additional restrictions: (i) limited temporal coverage of the species, and (ii) reduced information on the population demographics, such as size and age. Whilst the first can lead to failure to capture the seasonal cycle of a species and its location (Hilborn and Walters 1992), the second can lead to higher uncertainty towards less-frequent size/age groups since a much lower proportion of the population is sampled (Nielsen 2015).

In contrast, commercial fishery data provide information all year round for which a larger proportion of the stock is sampled (but not necessarily the full stock), and form the backbone of most stock assessment models (Hilborn and Walters 1992). The abundance information is derived from catch-and-effort (CPUE) data. However, unlike the survey data, CPUE cannot be assumed to be proportional to the actual abundance since the sampling is commercially driven (Walters and Maguire 1996). Such a poor relationship is likely the result of skippers deliberately selecting fishing grounds that maximize catch of the targeted species and size groups (i.e. preferential sampling). In such cases, the “sampling locations” are prone to aggregate in high-density areas where only a small proportion of the stock and only some life stages may be present (Fig. 1). Moreover, due to fishing restrictions
like the Minimum Conservation Reference Size (MCRS) and mesh-size regulations, only limited size/age groups are sampled.

Regardless of their differences, it is evident that these two data sources provide complementary information on the occurrence and population demographics that could be harmonized to benefit management and conservation advice. This is especially relevant within the Ecosystem-Approach to Fisheries management (EAFM), where gradual shift towards spatial-explicit management measures (e.g. spawning closures) has prompted the need to improve and develop innovative methodological approaches (Plagányi 2007). A focus of recent research has been the development of statistical models that can integrate various types of data sources while providing abundance distribution maps and/or indices that mirror the spatiotemporal dynamics of a given species (e.g. Conn 2009, Pennino et al. 2016, Bourdaud et al. 2017, Ono et al. 2017, Sant’Ana et al. 2017, Pinto et al. 2018, Zhu et al. 2018, Grüss and Thorson 2019, Thorson 2019). These approaches use statistical methods such as Bayesian hierarchical models, multivariate autoregressive state-space models and delta-generalized linear models. Nevertheless, we found some shortcomings that could restrict their applicability into real case studies, such as: (i) reduction of the response-variable into presence/absence (e.g.; Pennino et al. 2016); (ii) comparative analysis between the commercial fishery and survey data via separate models (e.g. Pennino et al. 2016, Bourdaud et al. 2017); (iii) not addressing spatiotemporal dependencies simultaneously (e.g. Bourdaud et al. 2017, Pinto et al. 2018); (iv) not explicitly modelling the differences in terms of fishing catchability and effort (e.g. Pennino et al. 2016, Bourdaud et al. 2017, Pinto et al. 2018, Zhu et al. 2018); (v) omission of the preferential sampling (PS) nature of the commercial fishery data when present (e.g. Sant’Ana et al. 2017, Pinto et al 2018, Thorson 2019), and (vi) omission of the spatial extent of the sampling unit (all aforementioned studies).

We propose a flexible spatiotemporal species distribution model (SDM, Franklin 2010) that can integrate commercial fisheries and survey data while filtering out the data-specific dependencies, and thereby providing a robust framework for mapping abundance distribution. We specifically identified and addressed five main differences between these data, namely: sampling effort, sampling catchability, spatiotemporal sampling coverage, PS, and the spatial extent of the sampling unit. We hypothesized that including commercial fisheries data will improve and shed additional insights into
the spatial-predicted abundance densities (maps) when contrasted to the survey-based model, and reliably fill temporal gaps from the survey data, and hence reconstruct the time-series. To evaluate its broad applicability, we tested the model on three age groups of the western Baltic (WB) cod (*Gadus morhua*) stock that is targeted by the Danish trawl fisheries.

**Materials & Methods**

SDMs usually involves modelling two distinct components: (i) the unobserved process (i.e. latent process) that underlies and explains the species abundance distribution dynamics like biotic (e.g. predation) and abiotic (e.g. temperature) factors, and (ii) the data generating process (i.e. observation process) that arises from the latent process such as sampling effects (e.g. sampling gear). Hierarchical models are natural candidates for modelling both latent and observation processes (Aeberhard et al. 2018, Isaac et al. 2019). These models assume that observed data are conditionally independent given the latent states (Fig. 2), thereby providing a wealth of flexibility for accommodating both types of processes. When the same latent variables describe different observation processes, hierarchical models become especially suitable for data integration. Within the SDM literature, this is best known as integrated distribution models (IDM), where point-process models have been providing untapped potential to combine different data sources (see Miller et al. 2019, Fletcher et al. 2019, Isaac et al. 2019). We rely on a point-process model to integrate commercial fisheries and survey data, under the premise that the underlying latent process is the same between the two data sets (hence, shared). In the following sections, we will outline the step-by-step construction of the generic model and all assumptions therein.

**Latent process**
The latent process $\lambda$, henceforth termed as abundance field, defines a species’ expected numbers as a function of space ($s$) and time ($t$) coordinates and can be modelled as a combination of fixed and random effects:

$$\lambda(s,t) = \exp \left( \sum_{k=1}^{K} \beta_k X_k(s,t) + \xi(s,t) \right)$$

(1)

where $X_k(s,t), k = 1, \ldots, K$ are a set of explanatory variables with corresponding fixed effect parameters $\beta_k$, and $\xi(s,t)$ represents a spatiotemporal structured random effect. $X_k$ are explanatory variables that might influence the underlying abundance field (e.g. temperature, predation, etc.). For the sake of simplicity, we only consider the time-period on a year-quarter basis as an explanatory variable in order to capture both intra- and inter-annual abundance dynamics. Other biotic/abiotic effects were left for future research.

The spatiotemporal dependency term, $\xi(s,t)$, is represented as a Gaussian random field (GRF), and when evaluated at a finite set of locations, it reduces to a multivariate normal distribution with zero mean and covariance matrix $\Sigma$:

$$(\xi(s,t)) \sim \text{MVN}(\mathbf{0}, \Sigma)$$

(2)

where $\Sigma$ is a separable covariance matrix. The variance can be then decomposed into a spatial covariance matrix, $\Sigma_S = (s_{ij}) \in \mathbb{R}^n \times n$, and a temporal covariance matrix, $\Sigma_T$, such that:

$$\Sigma = \Sigma_S \otimes \Sigma_T = \begin{pmatrix} s_{11}\Sigma_T & \cdots & s_{1n}\Sigma_T \\ \vdots & \ddots & \vdots \\ s_{n1}\Sigma_T & \cdots & s_{nn}\Sigma_T \end{pmatrix}$$

(3)

where $\otimes$ denotes the Kronecker product between the two matrices.

For $\Sigma_T$, we used a first-order autoregressive process (AR1) where the covariance depends on an exponential decay function that is expressed in terms of the absolute difference between time points $|t_1 - t_2|$ and the temporal lag-one correlation $\rho$:

$$\Sigma_T(t_1,t_2) = \rho^{|t_1 - t_2|}$$

(4)
To capture the typical seasonal migration of fish and account for the seasonality of the fishing operations, we chose a year-quarter time resolution (Y-Q1, Y-Q2, Y-Q3 and Y-Q4) to express the temporal correlation. This means, for example, that the time points 2014-Q4 and 2015-Q2 would have a correlation $\rho^2$ because they are separated by two quarters. Note, however, that a time resolution other than ours could have been chosen to describe the abundance field.

The spatial covariance matrix $\Sigma_s$, in contrast, was expressed according to Kristensen et al. (2014), which requires the discretization of the study area into a regular grid (hereby on a 5x5 km resolution; see Appendix S1: Fig. S1). The precision matrix $Q$ (inverse of the covariance matrix) in this case is modelled via a first-order conditional autoregressive (CAR) process, which is neither stationary nor isotropic, but accounts for the complex geometry of the spatial area and is computationally more efficient due to the sparseness of the matrix.

**Observation processes**

Conditional on the abundance field $\lambda$, we define the observation process, hereby the catch $Y$ in numbers, through a probability distribution. As this type of data tends to be overdispersed (Lindén and Mäntyniemi 2011), we used a negative binomial distribution to describe the observations, with the variance expressed as a function of the overdispersion parameter $\phi$. This constitutes a log-Gaussian negative binomial point process model, hereafter referred to as the LGNB-SDM.

Since the catch process of each data source is affected by different factors (e.g. catchability; Arreguín-Sánchez 1996), it is reasonable to assume that mean catch $\mu$ should also differ between data. Thus, we must explicitly distinguish how the data-specific mean catches are linked to the abundance field prior to integrating the data. We will henceforth distinguish the catch by a data-specific subscript, such that $Y_{\text{SUR}}$ and $Y_{\text{COM}}$ refer to the survey and commercial fishery catch processes, respectively. We shall describe in more detail the observation processes of both data below, and refer to Table 1 for a summary of all parameters, data, and indices described along with the model construction.

**Survey data**
Survey data rely on the premise that catchabilities remain constant throughout the space-time dimension (e.g. Hilborn and Walters 1992; Nielsen 2015). This is an important assumption, as catches can be assumed to be proportional to the underlying abundance field. Additionally, given that the spatial extent of the survey observations is relatively small compared to those of the commercial fishery data (Appendix S1: Fig. S2), the survey data can be regarded as a single point in space. Thus, for each observation \(i\) (hereby fishing haul), we can link the expected catches \(\mu_{i}^{\text{SUR}} = E(Y_{\text{SUR}}(s_{i}, t_{i}) | \lambda)\) to the abundance field \(\lambda(s, t)\) by adding catchability and effort effects to equation 1:

\[
\log(\mu_{i}^{\text{SUR}}) = \log(\lambda(s_{i}, t_{i})) + \sum_{k=1}^{K_{\text{SUR}}} \beta_{k}^{\text{SUR}} X_{k,i}^{\text{SUR}} + \gamma_{i}
\]

where \(\gamma\) is an offset term for adjusting the catches according to the sampling effort (in our case, the logarithm of the haul duration in minutes). The \(K_{\text{SUR}}\) explanatory variables \(X_{k,i}^{\text{SUR}}\) are indicators (0/1-variables) of the two research vessels (see case study application section), with the corresponding parameters \(\beta_{k}^{\text{SUR}}\) referred to as catchability effects of the survey vessels. Note, however, that additional catchability effects could be included at this stage, such as time of the day, gear type and depth of haul, among others. The variance \(V(Y_{\text{SUR}}(s_{i}, t_{i}) | \lambda)\) is calculated as a function of the mean \(\mu_{\text{SUR}}\) and overdispersion parameter \(\phi_{\text{SUR}}\):

\[
V(Y_{\text{SUR}}(s_{i}, t_{i}) | \lambda) = \mu_{i}^{\text{SUR}} + \frac{(\mu_{i}^{\text{SUR}})^2}{\phi_{\text{SUR}}}
\]

**Commercial fishery data**

Describing the observation process of the commercial fishery data requires at least three additional considerations: (i) accounting for the heterogeneous composition of vessels and fishing gears, (ii) dealing with the larger spatial extent of the observations, and (iii) accounting for the PS nature of the data. One of the main hurdles of commercial fishery data is that the composition and effort of vessels within a fleet may change in time and space; hence, proportionality between catches and underlying stock size cannot be assumed (Conn 2009). This is often sidestepped by including the
fishing vessel as an unstructured random effect in the catch process (Thorson and Ward 2014), an approach we also adopted.

We start by linking expected commercial fishery catch \( E(Y_{\text{COM}}(s_i,t_i)|\lambda) \) to the abundance field as if it had been point-referenced. The equation formulation is similar to the survey data except that the offset term is omitted (explanation follows below):

\[
\log (\mu_i^{\text{COM}}) = \log(\lambda(s_i,t_i)) + \sum_{k=1}^{K_{\text{COM}}} \beta_k^{\text{COM}} X_k^{\text{COM}}
\]  

(7)

As for the survey observations, the added explanatory variables are vessel and gear (métier) indicators; hence, the corresponding parameters are the vessel and gear effects. Our data has two such gear groups (see case study application section), which we consider as fixed effects. All other variation among vessels is considered as random, i.e. a \( N(0,\sigma^2_{\text{vessel}}) \) distribution, with unknown variance parameter that is assigned to more than 80 commercial vessel effects (see case study application section). Likewise to the survey observations, other types of gear and vessel effects could be included at this stage.

The spatial extent of the commercial fishery observations (hauls) is another important aspect to consider. Unlike the survey data, the trawled distance of the commercial fishery hauls can reach extensions of up to 60 km, thereby behaving rather like a line transect (Appendix S1: Fig. S2). Thus, we discretized each haul transect \( j \) into smaller units of fixed length (herein 1 km), such that the expected catch and its variance are calculated as:

\[
E(Y_j^{\text{COM}}|\lambda) = \sum_{i \in I_j} \mu_i^{\text{COM}}
\]  

(8)

\[
V(Y_j^{\text{COM}}|\lambda) = \sum_{i \in I_j} \mu_i^{\text{COM}} + \left( \frac{\sum_{i \in I_j} \mu_i^{\text{COM}}}{\phi_{\text{COM}}} \right)^2
\]  

(9)

By discretizing the commercial hauls, we account for all possible configurations of the abundance field along the transect that could explain the observed catch. Hence, an offset term was omitted from equation 7 since variations in effort are now explicitly accounted for by the haul discretization.
Finally, since the PS of the commercial data can lead to biased estimates and predictions (Diggle et al. 2010, Pennino et al. 2018), we extended the catch process in equation 7 such that it could be described by a combination of counts and sampling positions. In this case, equation 7 remains essentially unchanged except for the addition of an extra parameter ($\alpha$) that relates the sampling position $v$ along the spatial grid $G$ through a probability vector $P$ as follows:

$$P(V_i = v_i|\eta) = \frac{\lambda(s_i t_i)^{\alpha_i}}{\sum_{s \in G_{fi}} \lambda(s_i t_i)^{\alpha_i}} \tag{10}$$

The PS correction method relies on the definition of a sampling support area $f$, which constitutes a subset of the study area where a sampling is likely to occur (refer to Appendix S1: Section S2 for more details). The interpretation of $\alpha$ is such that when $\alpha = 0$, the spatial distribution of the sampling stations is uniform and hence no PS can be inferred. In contrast, when $\alpha > 0$, it indicates that the sampling stations are more concentrated in high-density, resulting in a positive PS. Opposite reasoning applies when $\alpha < 0$ (Appendix S1: Fig. S4).

**Integrating commercial fishery and survey data**

We can now set up the corresponding likelihood functions, whereby the likelihood function based on survey observations alone has the form:

$$L_{SUR}(\theta, \theta_{SUR}) = \int L_{SUR}(\theta_{SUR}|\lambda)P_\theta(\lambda)d\lambda \tag{11}$$

with $P_\theta(\lambda)$ denoting the distribution of the abundance field and $L_{SUR}(\theta_{SUR}|\lambda)$ the likelihood function of the survey data given the abundance field.

To integrate the two data sources, we assume that the spatiotemporal random-effect parameters describing the abundance field are shared between the two data sources. We can then express the joint likelihood of both data sources as:

$$L_{BOTH}(\theta, \theta_{SUR}, \theta_{COM}) = \int L_{COM}(\theta_{COM}|\lambda)L_{SUR}(\theta_{SUR}|\lambda)P_\theta(\lambda)d\lambda \tag{12}$$
where $L_{\text{COM}}(\theta_{\text{COM}} | \lambda)$ is the likelihood function of the commercial fishery data given the abundance field. We refer to $\theta$ and $\lambda$ as the *shared* parameters and random effects, respectively. The parameters specific for the commercial fishery data $\theta_{\text{COM}}$ are catchability parameters and notably year-quarter effects for which only commercial fishery data are available.

**Likelihood-based comparison of data sources**

It is crucial to check that the added information from the commercial fishery data is not in direct contradiction with the survey data. In other words, suppose we start by fitting the model based on the surveys alone using the marginal likelihood (equation 11); we must check that estimates of shared parameters $\theta$ and random effects $\lambda$ do not change significantly when re-fitting the model based on the integrated marginal likelihood function (equation 12).

Standard methods such as AIC, or LRTs, do not apply because the two likelihood functions are based on different data. Instead, we propose to check that parameter and random effect estimates using both data sources (integrated model) are within the confidence regions of the survey-based model. Such tests can be calculated directly from the likelihood functions without explicitly constructing high dimensional confidence regions. For the *parameters*, the proposed procedure is as follows:

1. a. Fit the survey model using equation 11 and denote the estimates $\hat{\theta}^{(1)}$ and $\hat{\theta}_{SUR}^{(1)}$.
2. a. Fit the integrated model using equation 12 and denote the estimates $\hat{\theta}^{(2)}$, $\hat{\theta}_{SUR}^{(2)}$ and $\hat{\theta}_{COM}^{(2)}$.
3. a. Check that the second estimates are within the multivariate confidence region based on the first estimates by reporting the $p$-value $Pr(X \geq x)$ of the statistic

$$ x = 2 \left( \log L_{\text{SUR}}(\hat{\theta}^{(1)}_{\text{SUR}}, \hat{\theta}^{(1)}_{SUR}) - \log L_{\text{SUR}}(\hat{\theta}^{(2)}_{\text{SUR}}, \hat{\theta}^{(2)}_{SUR}) \right) $$

(12)

where $X \sim \chi^2(df)$ and $df = \text{dim}(\theta) + \text{dim}(\theta_{\text{SUR}})$. 

This article is protected by copyright. All rights reserved
A value of $p$ greater than 0.05 implies that the parameters of the integrated model are within the 95% confidence region of the survey-based model. However, note that here $p$ does not have a standard interpretation as it is not uniform under the null.

A parallel approach can be used to check the consistency of the random effects:

1.b Using the parameter estimates from (1a) obtain the most probable random effects $\hat{\lambda}(1)$ by maximizing the integrand of equation 11.

2.b Using the parameter estimates from (2a) obtain the most probable random effects $\hat{\lambda}(2)$ by maximizing the integrand of equation 12.

3.b Using the logarithm of the integrand of equation 10, report $P$ of $\hat{\lambda}(2)$ being inside the confidence region of $\hat{\lambda}(1)$ using a $\chi^2$-distribution with $\text{dim}(\lambda)$ degrees of freedom.

A contradiction between the two data sources occurs if either of the two tests are rejected. In this situation, a natural next step would be to look for misspecification of the commercial likelihood contribution.

**Model selection, estimation and validation**

To investigate whether commercial fisheries data is affected by a PS, we first tested the integrated LGNB-SDM with and without the proposed PS correction method via a LRT. The two models were further evaluated according to the two validation approaches described later below, where we selected the best model as the one meeting both validations.

We conducted all parameter estimation through the Template Model Builder (TMB, Kristensen et al. 2016) that is interfaced with the R programming platform (R Development Core Team 2019). Within TMB, fixed effects are estimated by maximizing the marginal likelihood, whereas random effects are integrated out via the Laplace approximation. We confirmed that the model converged by verifying that the Hessian matrix was positive-definite, and by ensuring that the gradient component of the marginal likelihood was $< 0.001$. 

This article is protected by copyright. All rights reserved
The correctness of the estimation method was checked using a simulation study through the `checkConsistency` function from the TMB R-package. Briefly, this function checks whether the Laplace approximation is suitable for the estimation of the model parameters. In addition, it double-checks, by simulation, that likelihood functions of random effects and data have been correctly implemented. We conducted 500 simulations and considered the estimation method reliable whenever the relative bias ($\hat{\theta}/\theta$) was small (0.05) for the parameters.

To validate the LGNB-SDM performance on real data, we used two different approaches. First, we evaluated the consistency of the fixed and random effect parameters between the two data sources as described in the previous section. Secondly, we assessed the goodness-of-fit by inspecting the residuals for normality both visually (QQ-plot) and quantitatively (Kolmogorov-Smirnov test). In hierarchical models, computing residuals is not a trivial task due to the random effect parameters. We therefore used the simulation-based residuals as proposed by Thygesen et al. (2017). These residuals are based on a sample from the posterior distribution of the latent variables given the data and with the parameters replaced by their MLE. We obtained the sample through the MCMC algorithm implemented in the `tmbstan` R-package (Monnahan and Kristensen 2018), where a single chain was run until convergence and the final iteration $\lambda^*$ extracted. Using $\lambda^*$ as a replacement for $\lambda$ in equations (5), (6), (8), (9), calculation of quantile residuals was straightforward. Under the model assumptions, these residuals are independent standardized normal. For more details on the residual computation, see our Open Research statement.

**Case study application**

To test the LGNB-SDM, we focused on the WB cod stock (*G. morhua*) targeted by the Danish trawl fisheries, which is currently characterized as data-rich (ICES 2019). We applied the model to age groups 2-4+ and left recruits (ages 0-1) out of the analysis, as they were nearly inexistent in the commercial fisheries data (described below; see Appendix S1: Fig. S5) and because common age groups are needed across both data sets. Applying the model separately to each age group also provided means to mimick data sources with different targeting conditions and spatiotemporal dynamics.
To compose the commercial fishery data, we used on-board observer’s data covering the 2005-2016 period. These data provide information on the overall catch (i.e. including discards) of the trawl fisheries, in addition to other relevant biological information such as individual fish length and weight (Storr-Paulsen et al. 2012). The overall sampling scheme is stratified within vessel groups, area and quarter, with vessels randomly selected among all available vessels. However, because on-board sampling is not mandatory, some skippers can decline their presence, resulting in a quasi-random sampling. To retrieve age-disaggregated information on a haul-by-haul level, we coupled the on-board observer’s data to the biological database, which provides information on the catch-in-numbers by age groups (see Appendix S1: Section S3 for more details). The coupled commercial fisheries data included information from 8 different métiers (Appendix S1: Table S1). Nevertheless, we only selected the most representative ones, in this case the otter bottom trawl fleet targeting demersal species (OTB_DEF) with mesh sizes 110-120 (92% of the data), to avoid unnecessary noise from the undersampled métiers. The final commercial fisheries dataset contained information from 432 hauls spanning 86 fishing vessels (Fig. 1; Appendix S1: Table S2). Both data sources were obtained from the Danish National Institute of Aquatic Resources (DTU Aqua).

For the survey data, we used the ICES Baltic International Trawl Survey (BITS) database. The main aim of the BITS is to gather information on the abundance indices to tune the time-series in the ICES fish stock assessments, in addition to the spatial distribution and demographic structure of several demersal fish species from the Baltic Sea, with special focus on cod (ICES 2017). The sampling is randomly stratified within bathymetric layers of each ICES subdivision, and is undertaken twice a year (1st and 4th quarter). The sampling stations are set in such way that 60% of the hauls cover different depth zones, while the remaining 40% are allocated to those areas where cod abundance was highest in the last 5-year (Nielsen et al. 2014, ICES 2017). The standard fishing gear is the TV-3 bottom trawl, and the trawling is standardized to 30 minutes. We extracted the BITS data from the Database of Trawl Surveys (DATRAS, http://www.ices.dk/marine-data/data-portals/Pages/DATRAS.aspx) through its respective R-package (Kristensen and Berg 2018). The selected time-period was the same as for the commercial fisheries data, and only valid hauls were kept. This resulted in a dataset with 1808 hauls derived from two research vessels (R/V Havfisken and R/V Solea) (Fig. 1; Appendix S1: Table S2).
Evaluating the benefits of data integration

For each age group, we fitted the LGNB-SDM to (i) survey data and (ii) integrated data. Fitting the model separately allowed evaluating the extent to which the commercial fishery data bridged the spatial and temporal gaps in the survey data. We quantified the benefit of including commercial data on both parameter estimates and predicted abundance surfaces in terms of uncertainty reduction. To do so, we computed the ratio between the standard errors (on log-scale) of the parameters and abundance fields from the integrated model relative to the survey model. For the abundance fields, we also quantified the difference between the abundance surfaces predicted by the integrated and survey models. As abundance indices are a desired feature for stock assessments, we derived abundance indices for both models to evaluate visually the degree of overlap between their respective confidence intervals. The indices were calculated by summing up the predicted abundance densities over the spatial grid in each given time-period (i.e. year-quarter).

Results

Model selection

The PS correction method suggested that the commercial fishery data is not substantially affected by such bias source. Among all age groups, only age-4+ indicated a positive PS ($\alpha=1.5$), whereas the remaining age groups had a negative and almost absent PS (age-2, $\alpha=-0.08$; age-3, $\alpha=-0.07$). Except for age-4+, the models correcting for PS were only marginally better than those without a PS term (Table 2). Furthermore, when evaluating the models’ consistency (for result interpretation refer to the model validation section below), consistency failed for both fixed and random-effect parameters for age-4+ (Appendix S2: Fig. S1). In this sense, despite favoured by the LRT in the younger age groups, we selected the best model as the one without a PS correction approach for all age groups ($M_A$ in Table 2). Our choice on extending this selection to the younger age groups was primarily due to the fact that the magnitude of the PS was nearly negligible and with almost no impact on the estimated parameters (results not shown here). Moreover, a negative PS as
shown for these age groups implies that the sampling is preferentially conducted on low-density areas, whereas this is implausible within the current context. The following results will be thus explored without the PS framework.

**Model estimation**

Model convergence was confirmed for all age groups and data options. The simulation study indicated small relative biases for the parameters across all age groups and input data (Table 3). The smallest relative biases were generally associated with the model applied to the survey data, whereas marginally higher relative biases occurred for the integrated model. For the latter, some parameters had a greater bias than the specified threshold, namely: the overdispersion ($\phi$) and variance of the vessel effect ($\sigma^2_{\text{vessel}}$) parameters in all age groups, and the spatial scale ($\kappa$) parameter in age group 3 and 4+ (Table 3). Among these, we noted that the highest bias was associated with the variance of the vessel effect parameter and indicated that it was consistently underestimated across all age groups.

We also detected minor differences in the random-effect estimates between the two models for all age groups (Appendix S2: Tables S1 and S2). Overall, the scale parameter ($\kappa$) showed a decreasing trend from the younger to the older age groups. Given that the spatial correlation parameter ($\delta$) remained relatively constant across age groups, we can interpret the scale parameter purely in terms of spatial variability. As such, we can infer that the predicted spatial densities for the older age groups are less variable than the younger ones. The variance parameter of the commercial vessel effect, in turn, dictates the variation in catch that is due to different fishing vessels using different fishing tactics. When evaluating the random effects of the individual vessels (results not shown here), our results revealed that the most efficient vessel can be up to 380 times more efficient than the least efficient vessel (age-2 = 380.6; age-3 = 46.4; age-4 = 86.2). Such catch discrepancy highlights therefore the importance of including the commercial vessels as random effect terms.

The estimated fixed effect parameters had also only marginal differences in common effects between both data sources (the year-quarter time period in Appendix S2: Tables S3-S4 for age-2, S5-S6 for age-3, and S7-S8 for age-4+). Concerning the estimates from the integrated model, our results indicated intra- and inter-annual catch fluctuations across age groups. When inspecting the
catchability effect among the two survey vessels, much smaller differences were detected across age groups when contrasted to those reported for the commercial vessels (see Ship estimates in Appendix S2: Tables S4, S6 and S8). Particularly, our results indicated that the R/V Solea can be, on average, up to ~4 times more efficient than the R/V Havfisken (age-2 = 2; age-3 = 3.7; age-4+ = 3).

For the shared spatiotemporal random effect parameters that were estimated by the integrated model, our results highlighted that including commercial fisheries data could reduce up to 19% of the parameter estimates (Appendix S2: Fig. S2). Nevertheless, this reduction was not consistent across age groups and parameters, where for younger age groups (age-2 and age-3) we noticed an increase in the uncertainty for all parameters, with highest increase (31%) detected for the time correlation parameter in age-2 (Appendix S2: Fig. S2). Contrasting findings applied to the uncertainty reduction of the shared fixed-effect parameters (i.e. year-quarter effect; Appendix S2: Fig. S3). For these parameters, our results depicted that the integrated model could reduce, on average over the time-periods, up to 4.7% and 9.6% for age groups 2 and 3, respectively, whereas for the oldest age group the average uncertainty increased by 3.8% (Appendix S2: Fig. S3). For those time periods not covered by the survey model (2nd and 3rd quarters), we further noted that the integrated model was able to supply estimates within an acceptable range of uncertainty given their reasonably small average standard errors (age-2=1.9, age-3=1.6, age-4=1.2; see Appendix S2: Tables S4, S6, S8).

**Model validation**

For all age groups, the estimated shared parameters were compared between the survey and integrated models (Fig. 3). If the integrated model is valid, we should expect a tight linear relationship between the shared parameters in both fixed and random effects as this would indicate a high goodness of fit to a linear regression. The comparison shows that the shared parameters of the fixed effects are similar, following a linear trend (Fig. 3). Similar results are also observed for the random effects, but with much more variability around the linear trend (Fig. 3). A key question is whether these differences are a sign of the commercial fisheries data supporting or contradicting the survey data. Following our suggested test (see Likelihood based comparison of data sources section), no inconsistencies were detected for neither the fixed nor random effects (Table 4). This means that the vector of shared parameter estimates obtained by the integrated model was within the permissible
range (confidence region) of the survey model. As such, we could not detect any contradiction between the two data sources.

There was also a reasonable fit for both survey and integrated models (Appendix S2: Fig. S4). The residual pattern in the Q-Q plots followed a fairly normal distribution, which was further confirmed by the Kolmogorov-Smirnov’s test ($p > 0.05$; Appendix S2: Fig. S4). From the integrated model, we could further decompose the residuals into a survey and commercial component, where residual normality was also confirmed for both cases (Appendix S2: Fig. S5).

**Spatiotemporal dynamics of the western Baltic cod**

Overall, our results showed that the spatial abundance dynamics of all age groups varied between years, with some occasional seasonal migration detected mainly in the latter years (2011-2016; Appendix S2: Figs. S6-S11). Some degree of spatial segregation between age groups could be also detected, especially when contrasting the spatiotemporal abundance dynamics of the youngest (age-2) with the oldest (age-4+) age group (Appendix S2: Figs. S6/S7 and S10/S11).

Besides, our results indicated that the commercial fisheries data in the integrated model was able to supply additional information on cod’s abundance dynamics when compared to the survey-based predicted densities. Despite the predicted abundance surfaces were mostly driven by the survey-based model, we could clearly distinguish the contribution from the commercial fisheries data in almost all time-periods (Appendix S2: Figs. S6-S11). For illustration, we depict a snapshot of the age-specific abundance densities predicted by each data source in Figure 4.

For age-2, for example, the survey model predicted higher abundances around the Arkona Basin (see ICES area 24 in Fig. 1) and the Sound (area 23), whereas the integrated model depicted an additional abundance hot-spot in the vicinities of Fehmarn Belt (area 22). Likewise, for age-3, the integrated model highlighted more strongly and in more detail the two abundance hot-spots in the surroundings of the Kiel Bay (area 22). In the case of age-4+, the commercial fishery data supplied more detailed information especially west of Bornholm, where higher abundance densities were found more closely to Bornholm (area 24). All these abundance hot-spots were predicted at low uncertainty (Appendix S2: Fig. S12), with higher uncertainties mostly associated to the edges of the study area.
The difference in the spatial predictions could be as high as 40% in some areas, with the highest differences usually associated to those hot-spots that were depicted exclusively by the commercial fisheries data (Fig. 4). In addition, the integrated model showed that the uncertainty of the abundance surfaces could be reduced by 20% (Fig. 4). This reduction was particularly notable for age groups 2 and 3, where the highest reductions were located in the Fehmarn Belt and vicinities of Kiel Bay and Little Belt (Fig. 4). When evaluated across the considered time-period, our results indicated that for the same age groups the uncertainty could be reduced by 40%, with average reductions between 5 and 10% (Appendix S2: Fig. S13). In contrast, for age group 4+ the uncertainty tended to increase by an average of ~4%, although some areas could display a reduction of nearly 21% (Appendix S2: Fig. S13).

The abundance indices from the integrated model followed the same seasonal trend as those of the survey model (Fig. 5). This was notably true for age groups 3 and 4+, where the indices were nearly identical and with an almost full overlap between their respective 95% confidence intervals. For some time-periods we noticed that uncertainty increased slightly (e.g. Age-2 in 2013-Q1 and 2016-Q1); this was, nevertheless, often negligible, especially concerning age groups 3 and 4+. Most importantly, we noted a gain of information from the integrated model in relation to the time-periods that were not sampled by the survey data (2nd and 3rd quarters of each year). The relative low uncertainty associated with these time-periods demonstrates that we can reliably use the commercial fishery data to fill in the temporal survey gaps, and as such reconstruct the time-series to estimate and predict better a species’ seasonal abundance fluctuations.

Discussion

We relied on the features of point-process models to integrate commercial fisheries and scientific survey data while filtering out their relative bias contributions in the abundance distribution estimator. To our knowledge, this represents the first attempt to integrate these two very different, yet complementary, data sources while addressing several data-related pitfalls in a unifying framework.

We covered three different age groups of the same species having different selectivity and distribution patterns, whereby different descriptors were also used to describe the catch process of
each data source. The latter aspect was especially valuable to address the catchability issue, given that it is a main factor hampering fisheries data integration (e.g. Bourdaud et al. 2017, Pinto et al. 2018). We accounted for the hidden catchability effect by including the vessel as an unstructured random term whenever the commercial fishery data was considered; hence, it was essentially regarded as a nuisance parameter. This is a simple attempt to describe this process and we note that no strict consensus exists on the best way to account for catchability differences, since its effect arises from a set of complex and poorly understood interactions between fishers, vessel type and size, technological equipment, gear selectivity, fish species and size (life stage), and ecosystems (Hilborn and Walters 1992, Quinn and Deriso 1999). Parallel to Thorson and Ward (2014), we found that including a random vessel effect provided a straightforward way to account for differences in catchability. The simulation study indicated a high bias for the variance parameter of the vessel effect, and as such requires careful interpretation. However, bias in variance parameters is a common problem and future applications could possibly sidestep this limitation by relying on the bias-correction method proposed by Thorson and Kristensen (2016).

Integrated distribution models (IDM) also tend to increase the precision and accuracy of parameter estimates, and consequently outperform models based on a single data source (see Fletcher et al. 2019, Isaac et al. 2019, Miller et al. 2019, and references therein). Our results were in general agreement with state-of-the-art literature, where the abundance fields and shared fixed-effect parameters (year-quarter effect) yielded the best model improvement. Indeed, depending on the considered age groups, our results have shown that the uncertainty (std. error) could be reduced by up to 20% and 10% for the abundance fields and fixed-effect parameters, respectively. Yet, for the seasonal abundance indices we detected only a very subtle improvement, where for some time-periods the uncertainty slightly increased. Similar outcomes were also found by Grüss and Thorson (2019), who reported a greater improvement in the spatial distribution predictions than in the abundance index. From this work, we can also infer that an increase in precision is not necessarily warranted when relying on integrated models, and from our case study it seems to depend not only on the age structure of the population under concern, but also on the time-period being analyzed. This further suggests that the integrated model is currently of more use for predicting abundance distributions than for deriving an exact, overall abundance index. However, we note that the integrated model still
provides a safe ground to reconstruct the survey’s time-series. Yet, this is likely to change for other stocks and case studies that rely on different fishing gears and catchabilities, in which case we recommend future applications to accommodate more complex functions to describe the catchability (e.g. through gear selectivity ogives or gear saturation effects), especially if the aim is to derive an abundance index to inform stock assessments.

Spatiotemporal scale is also an important aspect to consider when dealing with these two data sources. Survey data typically have a low temporal and wide spatial coverage, whereas the opposite occurs for the commercial fisheries data. These differences can have several non-mutually exclusive implications when the aim is to reconstruct the abundance trend surface or the time-series. Here we assume that the underlying abundance density field is fundamentally the same irrespective of the data sources. That said, the spatiotemporal correlation parameters describing variation in abundance can be used to predict density in un-sampled areas and time-periods by borrowing information across nearby sites and time-periods, and represents therefore a valuable tool for data imputation (Thorson et al. 2015). Moreover, by widening the spatial frame and extending the time series, we found that the integrated LGNB-SDM captured a more refined and precise description of cod’s spatiotemporal dynamics.

It is well established from the literature that the WB cod stock exhibits a complex and not yet fully understood migration pattern along the Kattegat and western Baltic sea, where the stock progressively moves from the Kattegat (ICES area 21) and Øresund (area 23) towards the most southerly (area 22) and easterly (area 24) areas (Hüssy 2011). We identified a similar migration pattern along the investigated age groups, where we also noted that most of the abundance hot-spots identified in the first quarter of each year corresponded to recurring spawning areas (Hüssy 2011, and references therein). Our results have also shown that the abundance densities in the Belts, an important spawning area for cod, were nearly extinguished after 2007. Korpinnen et al. (2012) reported that this area experienced high levels of local overfishing that might have depleted this particular spawning site (Börjesson et al. 2013). Our results could likely support the latter authors’ claims. Nevertheless, given that we only considered a subset of the available fishery data (i.e. trawl fisheries) and spawning age groups, we could have missed additional abundance hot-spots and therefore more research on this particular aspect is warranted.
We should note that in the current context, no environmental predictors were used by the LGNB-SDM to supplement the abundance field predictions, and hence the observed abundance densities were purely a reflection of the spatiotemporal random field. This might be a critical simplification, as there are several indications that the distribution of WB cod stock is influenced by environmental forcing conditions such as water temperature, salinity, oxygen concentration, depth and seabed substrate (Peuchet et al. 2015). Thus, future applications would benefit from exploring the effect of environmental predictors on the WB cod abundance dynamics, including investigating the extent to which these variables could increase the precision and accuracy of the spatial predictions given that they often help to reduce the spatial bias (Simmonds et al. 2020).

We did not detect any significant issue regarding the PS nature of the commercial fishery data. Albeit our approach indicated a very weak PS for all age groups, we did not find substantial model improvement nor significant differences in the parameter estimates by accounting for a PS correction term. These findings contrast those of Pennino et al. (2018), who reported improved parameter estimates and spatial predictions of the blue and red shrimp when accounting for the PS nature of the evaluated fishing fleet. Three hypotheses are proposed to explain these patterns: (i) The on-board observers’ data used herein is characterized by a quasi-random sampling scheme where only a subset of the fishing fleets are sampled. In this sense, the effect of targeted sampling can become diluted; (ii) Although cod is a targeted species, only older age groups (> Age 3) are effectively targeted. This would align with our results, since a positive PS started to appear only for the oldest age group (Age-4+); (iii) Following Conn et al. (2017) and Thorson et al. (2015), a PS arises only when the sampling intensity is directly correlated to the underlying abundance field. Whenever the species’ abundance is predominantly explained by environmental predictors, the spatial targeting on high density areas will not cause bias since in this case the sampling intensity would be primarily correlated to the environmental predictor.

We highlight that the model provides a set of flexibilities that render it applicable to a wide range of case studies. Reasons for this include: (i) Any type of count-related data can be modelled (e.g. number of by-caught/discardcd individuals); (ii) Additional hidden effects on the response variable can be accounted for through unstructured random-effect terms. For instance, one could imagine a situation where the catchability changes over time or area, in which case a vessel-time (e.g.
year or year-quarter) or vessel-area effect could be included as an unstructured term. Or, if more than one survey data source is available, survey catchability effects will need to be accounted for as for the commercial fisheries data; (iii) Abiotic and/or biotic covariates can be included in both abundance and observation processes; (iv) The model can be switched to a single data source whenever the other is missing, therefore extending its use beyond the integrated framework; (v) Differences in the spatial extent of the sampling unit can be accounted for, and could include cases where both data source would behave either as a point-referenced (e.g. small-scale commercial fisheries) or a line transect (e.g. acoustic surveys) representation; And, ultimately, (vi) situations in which a PS behavior is known to occur can be addressed. We envision that our PS approach could be potentially used to validate sampling designs of scientific surveys, i.e., to evaluate whether they are not deliberately targeting a particular species or certain size/age groups as this could jeopardize the actual abundance estimates.

These multiple win-win combinations also open several avenues for future improvement and expansion of the LGNB-SDM. For instance, multiple survey data could be used whilst accounting for their distinct catchabilities and sampling efforts, as demonstrated for the commercial fisheries data. This approach could also be extended to accommodate data sources that rely on other sampling strategies (e.g. longline fisheries and acoustic surveys), although we do not recommend integrating data sources (i.e. gears) with very distinct selectivities/catchabilities, an aspect also endorsed by Conn (2009). The current model could also be extended by allowing other probability distributions to describe the observation process, including continuous and/or binary data like biomass and presence/absence data, respectively. For biomass data, in particular, we note that it could be applied in the current framework if the user rounds the data to the nearest integer (Maunder and Punt, 2004). Furthermore, we envision that the outputs from the LGNB-SDM could be applied to calibrate spatial bio-economic simulation tools such as DISPLACE (Bastardie et al. 2014), and to test the effect of different sampling designs on the abundance estimator. In fact, a simulation tool is currently embedded within the LGNB-SDM routine, which would allow conducting a full simulation study to address a wide-range of interesting questions. For instance, one might explore effects of integrating the two data sources when one is strongly affected by a PS, or when more commercial fisheries data is available than survey data. Furthermore, one could investigate potential drawbacks of using
commercial fisheries data, and under what conditions do these incur. Studies of this nature are only now starting to emerge (see Simmonds et al. 2020), and extending them to the fisheries context would be highly beneficial for fisheries conservation and management.

Acknowledgments

This work was part of MSPTOOLS (Grant Agreement No. 33113-B-16-064), COFASP ECOAST (Grant Agreement No. 321553) and NORDFO (Grant Agreement No. 33113-B-19-128) projects that were funded by the European Maritime and Fisheries Fund (EMFF) and the Danish Fisheries Agency during the period of 2017-2021. We thank Josefine Egekvist, Kirsten Håkansson and Line Pinna from DTU Aqua for providing all fisheries-related commercial fisheries data. We also thank all researchers and staff involved in the commercial and survey data collection/sampling processing under DTU Aqua. Finally, we warmly thank Anders Nielsen from DTU Aqua and the anonymous referees for their thorough comments that greatly improved the manuscript.

Supporting Information

Additional supporting information may be found online at: [link to be added in production]

Open Research

Data and all R and C++ codes, with a practical tutorial of the model application, (Rufener 2021) are available from the Zenodo repository at https://doi.org/10.5281/zenodo.4506948. Note commercial fisheries data are provided on an aggregated level due to sensitivity information, and survey data are fully disclosed. DTU Aqua has a data agreement with the Danish Ministry for Food, Agriculture and Fisheries where DTU receive commercial fisheries data as part of an agreement on science based advice, to be used for obligations under the EU Data Collection Framework (EU 2017/1004), advice and research; DTU Aqua does not have permission to forward these data un-aggregated to third part...
due to data sensitivity under the GDPR regulation, however further and more detailed information on the commercial fisheries data can be requested from the DTU Aqua data specialists.

References


Table 1: Summary of the parameters used to describe the latent and observation processes of the LGNB model.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K$</td>
<td>Number of explanatory variables of the abundance field (number of year-quarters for which we have data)</td>
</tr>
<tr>
<td>$K_{SUR}$</td>
<td>Number of explanatory variables of survey catchability (number of survey gears)</td>
</tr>
<tr>
<td>$K_{COM}$</td>
<td>Number of explanatory variables of commercial catchability (number of vessels + number of métiers)</td>
</tr>
<tr>
<td>$γ$</td>
<td>Logarithm of the sampling effort (Survey data only)</td>
</tr>
<tr>
<td>$Y_{SUR}$</td>
<td>Survey catch in numbers</td>
</tr>
<tr>
<td>$Y_{COM}$</td>
<td>Commercial catch in numbers</td>
</tr>
<tr>
<td>$β_k$</td>
<td>Effect of the $k^{th}$ covariate in the abundance field</td>
</tr>
<tr>
<td>$β_{SUR}^k$</td>
<td>Effect of the $k^{th}$ survey catchability covariate</td>
</tr>
<tr>
<td>$β_{COM}^k$</td>
<td>Effect of the $k^{th}$ commercial catchability covariate</td>
</tr>
<tr>
<td>$φ_{SUR}$</td>
<td>Overdispersion parameter for survey observations</td>
</tr>
<tr>
<td>$φ_{COM}$</td>
<td>Overdispersion parameter for commercial observations</td>
</tr>
<tr>
<td>$κ$</td>
<td>Random field scale parameter (see Kristensen et al. 2014)</td>
</tr>
<tr>
<td>$δ$</td>
<td>Random field parameter controlling the degree of spatial correlation (see Kristensen et al. 2014)</td>
</tr>
<tr>
<td>$ρ$</td>
<td>Abundance field lag-one temporal correlation</td>
</tr>
<tr>
<td>$σ_v^2$</td>
<td>Variance of the vessel random effect</td>
</tr>
<tr>
<td>$θ$</td>
<td>Vector of shared parameters $ρ, δ, β$ etc. from the abundance field. Only includes year-quarter effects for which survey observations are available.</td>
</tr>
<tr>
<td>$θ_{SUR}$</td>
<td>Vector of survey data parameters $φ_{SUR}, β_{SUR}$ etc. from the survey observation process. Only $num_{gears} - 1$ parameters included because one is used as a reference.</td>
</tr>
<tr>
<td>$θ_{COM}$</td>
<td>Vector of commercial fisheries data parameters $φ_{COM}, β_{COM}$ etc. from the commercial observation process. Includes remaining year-quarter effects i.e., those for which only commercial observations are available. Here $num_{métiers}$ parameters are used</td>
</tr>
</tbody>
</table>
\( s \) Spatial location (longitude and latitude)
\( t \) Time period (year-quarter)
\( i \) Survey observation index (point referenced fishing haul)
\( j \) Commercial observation index (segment referenced fishing haul)
\( \lambda \) Intensity of abundance
Table 2: Comparison of the integrated LGNB-SDM fitted without (M_A) and with (M_B) a preferential sampling (PS) correction term. PS (α)=estimated PS parameter, NLL=negative log-likelihood, Npar=number of parameters.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Model</th>
<th>PS (α)</th>
<th>NLL</th>
<th>Npar</th>
<th>$\chi^2$</th>
<th>Pr &gt; $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>M_A</td>
<td>-</td>
<td>-22570</td>
<td>110</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M_B</td>
<td>-0.08</td>
<td>-22566</td>
<td>111</td>
<td>7.87</td>
<td>0.005</td>
</tr>
<tr>
<td>A3</td>
<td>M_A</td>
<td>-</td>
<td>-21587</td>
<td>110</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M_B</td>
<td>-0.07</td>
<td>-21585</td>
<td>111</td>
<td>4.59</td>
<td>0.032</td>
</tr>
<tr>
<td>A4+</td>
<td>M_A</td>
<td>-</td>
<td>-20221</td>
<td>110</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>M_B</td>
<td>1.5</td>
<td>-20080</td>
<td>111</td>
<td>281.41</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 3: Summary of the simulation study based on the LGNB model applied to both survey and integrated. The *value* column stands for the transformed parameter values that were used in the simulation, whereas the *bias* column stands for the *relative* bias associated with the given parameter. Asterisks denote cases that were not applicable.

<table>
<thead>
<tr>
<th>Age group</th>
<th>Parameters</th>
<th>Survey</th>
<th></th>
<th>Integrated</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Value</td>
<td>Bias</td>
<td>Value</td>
<td>Bias</td>
</tr>
<tr>
<td>A2</td>
<td>log (δ)</td>
<td>-7.65</td>
<td>-0.049</td>
<td>-7.56</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>log (κ)</td>
<td>0.58</td>
<td>-0.008</td>
<td>0.57</td>
<td>-0.011</td>
</tr>
<tr>
<td></td>
<td>ρ/√(1 + ρ^2)</td>
<td>1.12</td>
<td>-0.002</td>
<td>1.20</td>
<td>-0.001</td>
</tr>
<tr>
<td></td>
<td>log (σ^2_vessel)</td>
<td>*</td>
<td>*</td>
<td>0.44</td>
<td>-0.505</td>
</tr>
<tr>
<td></td>
<td>log (φ_{SUR})</td>
<td>1.50</td>
<td>0.113</td>
<td>1.20</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>log (φ_{COM})</td>
<td>*</td>
<td>*</td>
<td>-0.04</td>
<td>0.12</td>
</tr>
<tr>
<td>A3</td>
<td>log (δ)</td>
<td>-7.63</td>
<td>0.016</td>
<td>-7.45</td>
<td>-0.165</td>
</tr>
<tr>
<td></td>
<td>log (κ)</td>
<td>0.46</td>
<td>-0.005</td>
<td>0.45</td>
<td>-0.028</td>
</tr>
<tr>
<td></td>
<td>ρ/√(1 + ρ^2)</td>
<td>1.09</td>
<td>-0.007</td>
<td>1.16</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>log (σ^2_vessel)</td>
<td>*</td>
<td>*</td>
<td>0.00</td>
<td>-0.572</td>
</tr>
<tr>
<td></td>
<td>log (φ_{SUR})</td>
<td>1.67</td>
<td>0.104</td>
<td>1.46</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>log (φ_{COM})</td>
<td>*</td>
<td>*</td>
<td>0.16</td>
<td>0.038</td>
</tr>
<tr>
<td>A4+</td>
<td>log (δ)</td>
<td>-7.23</td>
<td>-0.092</td>
<td>-7.25</td>
<td>-0.102</td>
</tr>
<tr>
<td></td>
<td>log (κ)</td>
<td>0.20</td>
<td>-0.031</td>
<td>0.25</td>
<td>-0.039</td>
</tr>
<tr>
<td></td>
<td>ρ/√(1 + ρ^2)</td>
<td>4.14</td>
<td>0.176</td>
<td>3.71</td>
<td>0.253</td>
</tr>
<tr>
<td></td>
<td>log (σ^2_vessel)</td>
<td>*</td>
<td>*</td>
<td>0.14</td>
<td>-0.451</td>
</tr>
<tr>
<td></td>
<td>log (φ_{SUR})</td>
<td>0.41</td>
<td>-0.027</td>
<td>0.45</td>
<td>-0.021</td>
</tr>
<tr>
<td></td>
<td>log (φ_{COM})</td>
<td>*</td>
<td>*</td>
<td>-0.17</td>
<td>0.029</td>
</tr>
</tbody>
</table>
Table 4: Summary of the test statistics applied to the fixed and random effect parameters. Note that because $p$ does not have a standard interpretation (not uniform under the null), we express its value as True whenever $p>0.05$.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Age group</th>
<th>$\chi^2$</th>
<th>df</th>
<th>Pr &gt; $\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-effect</strong></td>
<td>A2</td>
<td>6.37</td>
<td>30</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>2.60</td>
<td>30</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>A4+</td>
<td>4.42</td>
<td>30</td>
<td>True</td>
</tr>
<tr>
<td><strong>Random-effect</strong></td>
<td>A2</td>
<td>15971</td>
<td>147678</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>12048</td>
<td>147678</td>
<td>True</td>
</tr>
<tr>
<td></td>
<td>A4+</td>
<td>44443</td>
<td>147678</td>
<td>True</td>
</tr>
</tbody>
</table>
Figure 1: Example from the Danish cod commercial trawl fisheries and survey data in the western Baltic Sea to illustrate the main differences between both data sources. Panel (a) contrasts the spatial distribution of the data-specific sampling stations, with numbers indicating the ICES subdivisions. Panels (b) and (c) highlights the difference in fishing catchability (catches standardized to the haul duration) and fishing effort, respectively. Acronyms in panel (b) stands for the different métiers (A = OTB_DEF_>=105_1_110; B = OTB_DEF_>=105_1_120; H=TVS).

Figure 2: Schematic illustration of the conditional independence, where the dashed circle represents the latent states $\eta$, with $\eta = log(\lambda)$, and the gray boxes the data-specific observations processes. The integrated LGNB-SDM (survey + commercial fishery data) states that the two data sources are conditionally independent given the latent field $\eta$. By removing each data source separately, it follows that the model must still hold for the remaining data source.

Figure 3: Validation of the integrated LGNB model by contrasting the shared parameters of the fixed effects (upper panels) and random effects (lower panels) between the survey and integrated data.

Figure 4: Snapshot of the spatial abundance distribution predicted by the two model options for age groups 2-4+ of the WB cod (first and second row panels). Mid panels highlight the difference (%) in the spatial predictions between the integrated and survey model, whereas the lower panels display the reduction in the uncertainty of the abundance fields $\left(\sigma^2_{\text{Integrated}}/\sigma^2_{\text{survey}}\right)$. For visualization purposes, abundances were standardized to a 0-1 interval.

Figure 5: Integrated (blue) and survey-specific (yellow) inter-annual abundance indices for age groups 2-4+ of WB cod, with shaded areas and range bars denoting the 95% confidence intervals.
Figure 2

Latent process

Survey

Commercial

Observation process

$\eta$

$\theta_s/\theta_c$
Figure 3

Age 2

Random Effects (survey data)

Random Effects (integrated data)

Fixed Effects (survey data)

Fixed Effects (integrated data)

Age 3

Age 4