Estimating cetacean bycatch from non-representative samples (I): a simulation study with regularized multilevel regression with post-stratification

Matthieu Authier[1,2,3], Etienne Rouby[4], and Kelly Macleod[5]

[1]Observatoire PELAGIS, UMS 3462, CNRS-La Rochelle Université, Pôle Analytique, allée de l'Océan, 17000 La Rochelle, France
[2]ADERA,162 avenue Albert Schweitzer CS 60040, 33608 Pessac Cedex, France
[3]mauthier@univ-lr.fr
[4]Centre d'Études Biologiques de Chizé, UMR 7372 CNRS-La Rochelle Université, 405 route de Prissé-la-Charrière, 79360, Villiers-en-Bois, France
[5]Joint Nature Conservation Committee, Inverdee House, baxter Street, Aberdeen, Scotland

May 28, 2021

Abstract and Aims

Bycatch, the non-intentional capture or killing of non-target species in commercial or recreational fisheries, is a world wide threat to protected, endangered or threatened species (PETS) of marine megafauna. Obtaining accurate bycatch estimates of PETS is challenging: the only data available may come from non-dedicated schemes, and may not be representative of the whole fisheries effort. We investigated, with simulated data, a model-based approach for estimating PETS bycatch from non-representative samples. We leveraged recent development in the statistical analysis of surveys, namely regularized multilevel regression with post-stratification, to infer total bycatch under realistic scenarios of data sampling such as under-sampling or over-sampling when PETS bycatch risk is high. Post-stratification is a survey technique to re-align the sample with the population and addresses the problem of non-representative samples. Post-stratification requires to sub-divide a population of interest into potentially hundreds of cells corresponding to the cross-classification of important attributes. Multilevel regression accommodate this data structure, and the statistical technique of regularization can be used to predict for each of these hundreds of cells. We illustrated these statistical ideas by modelling bycatch risk for each week within a year with as few as a handful of observed PETS bycatch events. The model-based approach led to improvements, under mild assumptions, both in terms of accuracy and precision of estimates and was more robust to non-representative samples compared to more design-based methods currently in use. In our simulations, there was no detrimental effects of using the model-based even when sampling was representative. Estimating PETS bycatch ideally requires dedicated observer schemes and adequate coverage of fisheries effort. We showed how a model-based approach combining sparse data typical of PETS bycatch and recent methodological developments can help when both dedicated observer schemes and adequate coverage are challenging to implement.

Introduction

Bycatch, the non-intentional capture or killing of non-target species in commercial or recreational fisheries, is a world wide threat to protected, endangered or threatened species (PETS) of marine megafauna (Gray & Kennelly, 2018), including seabirds (Martin *et al.*, 2019), elasmobranchs (Pacoureau *et al.*, 2021) and cetaceans (Avila *et al.*, 2018). Bycatch in fishing gears, such as gillnets, is currently driving some small cetacean species to extinction (Jaramillo-Legorreta *et al.*, 2019; Brownell *et al.*, 2019). The European Commission recently issued infringement procedures against several Members States for failing to correctly transpose some provisions of European environmental law (the Habitats Directive, Council Directive 92/43/EEC), in particular the obligations related to the establishment of a coherent monitoring scheme of cetacean bycatch (`https://ec.europa.eu/info/news/july-infringements-package-commission-moves-against-member-states` `en`). The Data Collection Framework (DCF) provides a common framework in the European Union (EU) to collect, manage, and share data within the fisheries sector (Anonymous, 2019a). The Framework indicates that the Commission shall establish a Multi-Annual Union Programme (EU-MAP) for the collection and mangement of fisheries data which should be inclusive of data that allows the assessment of fisheries' impact on marine ecosystems. With respect to PETS (including cetaceans), the collection of high quality data usually requires a dedicated sampling scheme and methodology, and is generally different from those applied under the DCF (Stransky & Sala, 2019): "EU MAP remains not well suited for the dedicated monitoring of rare and protected bycatch in high-risk fisheries since its main focus is the statistically-sound random sampling of all commercial fisheries (Ulrich & Doerner, 2021; page 126)". In practice, the introduction of any programme on PETS bycatch under the DCF may be met with caution because of its perceived potential to disrupt data collection for fisheries management (Stransky & Sala, 2019). This perception implicitly relegates PETS bycatch as a side issue for fishery management rather than an integral part of it. It may explain the usually poor quality of bycatch data on PETS (ICES, 2020a).

Recent EU legislation (Regulation 2019/1241), refered to as the Technical Measures Regulation (TMR), requires Members States to collect *scientific data* on cetacean bycatch for the following métiers: pelagic trawls (single and pair), bottom-set gillnets and entangling nets; and high-opening trawls (Anonymous, 2019b). Unlike its predecessor (Council Regulation EC No. 812/2004), this Regulation does not require the establishment of dedicated observer schemes for cetacean bycatch data collection (Dolman *et al.*, 2020). Furthermore, only vessels of an overall length of 15 metres or more are to be monitored, but these represent a small fraction of the European fleet (less than 10% in 2019; `https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=fish_fleet_alt&lang=en`). This vessel length criterion introduces bias in the bycatch monitoring data as the sample of vessels larger than 15 metres is unlikely to be similar to smaller vessels. Even within the sample of vessels that are monitored, pragmatic considerations can complicate sampling. For example, in the United States, observer sampling trips are allocated first by region, port and month, then randomly to vessels of particular categories within those monthly and spatial strata (ICES, 2009). Random allocation of observers to vessels follows sound statistical methodology and increases the likelihood of collecting unbiased data. In France, observer days are allocated by port and by month for each fishery, but the exact vessel allocation is then negotiated and left at the discretion of skippers (ICES, 2009). Allocation is no longer random as skippers may only accept observers when cetacean bycatch risk is low (Benoît & Allard, 2009). Non-random allocation means potential bias in the collected data for monitoring bycatch as the sub-sample of skippers accepting an observer may be very different from skippers refusing to do so.

One pragmatic solution bypassing observers is to mandate skippers to self-declare the non-intentional capture or killing of any PETS, as already required under the DCF (Anonymous, 2019a). In France, a national law from 2011 mandate fisheries to declare (without fear of prosecution) the bycatch of any cetacean species, but this law remained largely unknown to French fishermen until late 2019 (Cloâtre, 2020). In general, self-reported PETS bycatch data are sub-optimal as they may be heavily biased, non-representative (ICES, 2009) and typically provide poor information on which to base management decisions (National Marine Fisheries Service, 2004). Once again, the set of skippers who choose to declare bycatch may differ markedly from those who do not: for example the former take the extra time required to fill logbooks and thus provide accurate data while the latter do not. If this behaviour is correlated to other attributes, *e.g.* a more acute awareness of threats to PETS resulting in practices that tend to minimize impact on PETS, data collected from skippers reporting bycatch may not be representative. There may also be an element of skippers simply "forgetting" to log "PETS" bycatch in the bustle of the fishing operation but this is random and unlikely to introduce bias. In addition, ground-truthing, for example with remote-electronic monitoring (REM; Course *et al.*, 2020), would be required in order to ensure data quality before their statistical analyses.

Another hurdle, of the statistical kind, with cetacean bycatch is the low frequency of these events. Assuming that implementing a representative sampling program were feasible, if bycatch is a rare event (Komoroske & Lewison, 2015), then few events would be observed for realistic sampling effort (ICES, 2009). This paucity of observed event means a large uncertainty in statistical estimates: with a bycatch rate of the order of 0.01 event per fishing operation, a sample size of $1,000$ bycatch events would be required to obtain, for example, the US recommended coefficient of variation of 30% (National Marine Fisheries Service, 2005; ICES, 2009; Carretta & Moore, 2014; National Marine Fisheries Service, 2016). This is in the best case scenario (no bias, statistical independence *etc.*) since, in practice, the sampling error depends on the overall design of the survey, of which the sample size is only one factor (*e.g.* in pratice the sample size needed would be larger as the same vessels would contribute fishing operations, and these would not be statistically independent). With a small sample size, uncertainty may be so large as to prevent using estimates altogether, even if one were to assume no bias in the data. Given this challenge and the lack of uptake of dedicated monitoring programmes of cetacean bycatch in Europe over the last decade or more (Sala *et al.*, 2019), it would appear prudent to seek methods of analysis that can handle the few and non-representative data available to robustly estimate bycatch rates.

The problem of having non-representative samples to carry out statistical analyses is ancient (Hansen & Hurwitz, 1946) and widespread: it pops up in many applied disciplines, including election forecasting (Wang *et al.*, 2015; Kiewiet De Jonge *et al.*, 2018), political sciences (Lax & Phillips, 2009; Zahorski, 2020), social sciences (Halsny, 2020), addiction studies (Rhem *et al.*, 2020) or epidemiology (Zhang *et al.*, 2014; Downes *et al.*, 2018). In these disciplines, there are also intrinsic limits on improving the representativeness of sampling. For example, in polling, non-response rates can be above 90% (Forsberg, 2020). In other cases, some populations of interest may be hard to reach (Rhem *et al.*, 2020), or answers may not be honest (St. John *et al.*, 2014). Challenges lie in the accurate estimation of quantities of scientific interest (*e.g.* the true magnitude of bycatch in a fishery) with the construction of statistical weights that can calibrate a non-representative survey sample to the population targets. Such weights are implicit with simple random sampling where each unit in a population has the same, non-nil, probability of being included in the sample. When inclusion probabilities differ between units, weights inversely proportional to the former can be used to adjust the sample. However, constructing survey weights is in general more elaborate than using inverse probabilities of selection in the sample (Gelman, 2007). Model-based approaches, and multilevel regression modelling with post-stratification in particular, has become an attractive alternative to weighting to adjust non-representative samples (Gelman, 2007).

Multilevel regression modelling allows to summarize how predictions of an outcome of scientific interest vary across statistical units defined by a set of attributes or covariates (Gelman *et al.*, 2021 page 4): for example bycatch events are a binary outcome at the fishing operation level (a unit) associated with attributes, such as date-time, location, gears and vessels (*e.g.* Palka & Rossman, 2001). Post-stratification is a standard technique to generalize inferences from a sample to the population by adjusting for known discrepancies between the former and the latter. The key insight of combining multilevel regression modelling with post-stratification is that, even if observations are not a representative sample of the population of interest, it may be possible to construct a regression model to first *predict* unobserved cases, and then *post-stratify* to average the fitted regression model's predictions over the population of interest (Gelman *et al.*, 2021, page 313). Lennert *et al.* (1994) provided an early example of model-based estimates of bycatch with post-stratification. Their model however was not a multilevel regression, a technique which can handle complex data structure for better inferences or predictions (Bolker *et al.*, 2009). In particular, adequate post-stratification may require to handle hundreds of cells (the crossing of several attributes; *e.g.* week by statistical area by gears). Some predictions for each cell may be too noisy, especially if there are sparse or no data for that particular combination of attributes. Multilevel regression allows to borrow strength from similar units to improve and stabilize (*i.e.* regularize) these predictions (Cam, 2012). In other words, multilevel regression allows an efficient use of the sample to estimate the outcome of interest within each cell, even if these cells are very numerous (*e.g.* several hundreds).

Technically, when data arise as signal plus noise, overfitting occurs when a regression model captures too much of the noise compared to the signal; that is in using an ill-conditioned (unstable) model that will provide an excellent in-sample fit but make poor out-of-sample predictions (George & Ročková, 2021; Authier *et al.*, 2017). Overfitting may result when using richly parametrized models without using adequate estimation methods such as regularization to stabilize parameter estimates and buffer them against noise (Gelman *et al.*, 2021, pages 459-460). Weakly-informative priors in a Bayesian framework regularize the estimation of the large number of parameters that may be present in a multilevel model. Multilevel modelling allows to take into account complex data structures with structured prior models for batches of parameters; the simplest example are so-called 'random effects' whereby a common (Gaussian) distribution centered on zero and with an unknown variance to be estimated for data is assumed for a group of parameters; for example years or sites (Cam, 2012). This common distribution for the parameters is a prior model, and this model for parameters means that the latter are not independently estimated but in tandem according to the postulated prior model. For example, Sims *et al.* (2008) used a model-based approach to obtain spatially smoothed estimates of bycatch in a gillnet fishery. Spatial-smoothing (also known as 'small-area estimation'; Fay & Herriot, 1979) was used to stabilize estimated bycatch rates by using a Conditional Autoregressive prior model that leverages information from spatial neighbours to improve the prediction at a specific location. Prior models add some soft constraints to the overall model and these constraints are very useful in data sparse settings to mitigate variance and bias in predictions. In other words, these prior models represent additional assumptions about the data, assumptions, which if approximatively correct, add information in the analyses and increase the precision and stability of predictions at the cost of a usually small estimation bias. Introducing bias to reduce variance is a common statistical technique known as shrinkage or regularization (George & Ročková, 2021).

Regularized multilevel regression with post-stratification is thus the combination of several important ideas to obtain accurate predictions (Gao *et al.*, 2019). First, post-stratification is a survey technique to re-align the sample with the population and addresses the problem of non-representative samples. In practice, post-stratification requires to sub-divide the population of interest into many cells corresponding to the combination of important attributes. Multilevel regression can be used to accommodate all these cells in a single model, but the problem has now moved to how to obtain useful estimates for all these cells, which can number in the several hundreds. Regularization solves this estimation problem: it introduces model-driven bias in statistical estimates in order to stabilize them. These new developments in the statistical analysis of non-representative samples may help in obtaining a better quantification of bycatch rates and numbers. Our aim is to assess with simulations, the potential of regularized multilevel regression with post-stratification for analyzing already collected bycatch data, with the full knowledge that these data are non-representative and biased in several respects. These bias in sampling are manifold (see above): bias may be due to regulation exempting certain vessels (*e.g.* no monitoring for vessels smaller than 15 metres); to non-dedicated observers or because sampling is driven for other purposes than bycatch monitoring of PETS (commercial discards, stock assessment); or in the case of dedicated schemes, to over-sampling a few "cooperative" skippers or focussing sampling in métiers with the highest or lowest bycatch risk. Our focus will be narrower, honing in on specific sampling scenarios whereby observer coverage is correlated to bycatch risk. In other words, we will assess the potential of regularized multilevel regression with post-stratification to estimate accurately bycatch numbers with samples preferentially collected either during low- or high-bycatch risk periods. Our investigation is largely framed from our knowledge on small cetacean bycatch in European waters, such as short-beaked common dolphin (*Delphinus delphis*) in the Bay of Biscay (Peltier *et al.*, 2021) or harbour porpoises (*Phocoena phocoena*) in the Celtic Seas (Tregenza *et al.*, 1997). In the remainder, we first introduce methods and notations to detail the proposed model to perform multilevel regression with post-stratification with bycatch data. Next, we explain our data simulation scenarios and how we emulate non-representative sampling. We then compare the results (*i.e.* estimates of bycatch) from the proposed modelling approach with those from the method currently used by the working group on bycatch of protected species from the International Council for the Exploration of the Sea (ICES WGBYC) before concluding on some recommendations for future investigations.

Material and Methods

We carried out Monte Carlo simulations to assess the ability of regularized multilevel regression with post-stratification to estimate bycatch risk and bycatch rates from representative and non-representative samples. ICES WGBYC collate data through an annual call from dedicated and DCF surveys collecting data on the bycatch of PETS through onboard observers or REM. These surveys may be qualified as "design-based" in the sense that, ideally, a representative coverage of fisheries would be sought in order to scale up the observed sample to the whole population using ratio-estimators. There are many caveats around the use of these ratio-estimators as EU MAP is not well suited for monitoring PETS bycatch (Ulrich & Doerner, 2021). Given these shortcomings in the collection of bycatch data under EU MAP, the data available to ICES WGBYC are unlikely to be representative of fisheries of interest but nevertheless, ratio-estimators are used as part of a Bycatch Risk Approach (BRA) to identify relative risk of bycatch across species and metiers. Cetacean bycatch observer programmes may aim at achieving a pre-specified precision for bycatch rates (with a coefficient of variations less than 30%; National Marine Fisheries Service, 2005; ICES, 2009; Carretta & Moore, 2014; National Marine Fisheries Service, 2016). Achieving this is very difficult in practice, and a given coverage of effort deployed by the total fleet is, instead, aimed at: for example 10% (5%) for pair-trawlers (level-3 métier PTM) larger (smaller) than 15 metres in France. Data from onboard observer programmes are then used to estimate total bycatch using ratio estimators (Lennert *et al.*, 1994; Julian & Beeson, 1998; Amandè *et al.*, 2012) and the bootstrap or a classical approach (Clopper-Pearson) for uncertainty quantification (ICES, 2018, page 57). We used an approach similar to that of WGBYC (hereafter refered to as a "design-based" approach) as a benchmark to compare against results from regularized multilevel regression with post-stratification. We honed in on the accurate estimation of the number of bycatch events for a complete fleet. We assume that information on the total effort deployed by a fleet operating in a spatial domain are available and measured without error. This assumption is necessary to scale estimates from the sample to the population. We also assumed that there are no false-negatives in the sample, that is no bycatch event went unrecorded by onboard observers (assuming thereby a dedicated observer programme). These two assumptions are customary with ratio estimators, whether design- or model-based, and do not deviate from current norms. We assume however that these population data on total effort can be disaggregated at a finer temporal scale in order to post-stratify on calendar weeks. This assumption of accurate measurement of effort at the week-level is crucial for post-stratification.

Notations

Let y_{ijkl} denotes the i^{th} fishing operation of vessel j in week k of year l, with $y_{ijkl} = 1$ if a bycatch event occurs and 0 otherwise:

$$y_{ijkl} \sim \text{Bernoulli}\left(\text{p}_{ijkl} = \text{logit}\left(\mu + \beta_{kl} + \alpha_j\right)\right) \tag{1}$$

where p_{ijkl} is the product of the probability of a bycatch event occurring and the probability of dolphin presence. This unconditional probability is denoted 'bycatch risk' hereafter. Once estimated, bycatch risk may be multiplied by the average number of animals involved in a bycatch event to recover a bycatch rate. Bycatch risk is a function of several parameters (on a logit scale): μ is the intercept (overall risk), $\alpha_j \sim \mathcal{N}(0, \sigma_{\text{vessel}})$ are (unstructured, normal random effects) vessel-effects accounting for heterogeneity (*e.g.* 'fishing style' of skippers); and β_{kl} are cyclical time effects, modelled with a Gaussian Process ($\mathcal{GP}()$):

$$\boldsymbol{\beta}_l \sim \mathcal{GP}\left(\boldsymbol{\varepsilon}, \boldsymbol{\Omega}\right) \tag{2}$$

where $\boldsymbol{\varepsilon}$ is a mean vector of week effects and $\boldsymbol{\Omega} = \Omega\left(t_k, t_{k'}\right)$ is the covariance between week t_k and week $t_{k'}$ within a year l. A Matérn covariance function of order $\frac{3}{2}$ and range parameter fixed to $\frac{3}{2}$

was assumed: $\Omega\left(t_k, t_{k'}\right) = \sigma^2_{\text{year}} \times \left(1 + \frac{2\sqrt{3}\times|t_k - t_{k'}|}{3}\right) \times \exp -\frac{2\sqrt{3}\times|t_k - t_{k'}|}{2}$. This choice corresponds to a temporal correlation of 0.05 after 4 weeks (i.e. temporal independence after a month). The choice of the covariance function also translate an assumption of smoothness in the temporal profile of bycatch risk: bycatch risk is assumed to change gradually across weeks, with no abrupt increase or decrease. The vector ε of mean weekly effects (on a logit scale) was modelled with a first order random walk:

$$\begin{cases} \varepsilon_t \sim \mathcal{N}(0, \sigma_{\text{week}}) & t = 1 \\ \varepsilon_{t+1} \sim \mathcal{N}(\varepsilon_t, \sigma_{\text{week}}) & t > 1 \end{cases} \tag{3}$$

The model in equation 1 is a decomposition of bycatch risk into a time-varying component (at the week-scale, equation 3; and with an interaction with year, equation 2) and time-invariant component which can be interpreted as fishing-style effects whereby some skippers may have consistent practices that increase or decrease bycatch risk. The time-varying component is assumed to be cyclical to reflect, for example, a year-round pattern in the distribution of dolphins and fisheries. Importantly, bycatch risk is modelled here with no attempt to model dolphin presence directly as relevant data to do so may be missing in the general case. Bycatch risk is thus to be estimated for each week of a year, and each of these weeks represent *de facto* a stratum. In any applied case, additional factors, such as statistical area, may need to be included in Eq.1 for improved realism. For simplicity, we did not consider space in simulations, and solely focussed on time.

Data Simulation

To test the ability of model 1 to estimate bycatch risk, data were simulated thusly (Figure 1):

1. bycatch probability conditional on dolphin presence was constant and set to 0.3, that is roughly one fishing operation out of 3 generates a bycatch event when dolphins are present;

2. dolphin presence is seasonal: it peaks at the beginning and end of the year, but quickly drops to 0 for roughly 2 thirds of a year; and

3. a fishery of 20 vessels is operating all year round, with an overall activity rate of 80% each week (that is, for any week, $20 \times \frac{80}{100} = 16$ vessels are fishing). Each fishing day (5 days per week), on average 2.3 fishing operations are carried out. The expected total number of fishing operations for a year is $5 \times 52 \times 2.3 \times 16 \approx 10,000$. During each of these operations, a bycatch event may occur depending on dolphin presence at the time.

4. Observers are accepted onboard vessels either with a constant probability of 0.05 corresponding to a coverage of 5% of all fishing operations (unbiased sampling scenario) or with a probability that covaries with dolphin presence (biased sampling scenarios). In the latter case, realized coverage is a random variable. With under-sampling, the bulk of the observer data is collected when bycatch risk (the product of dolphin presence and bycatch probability) is nil (Figure 1). With over-sampling, the bulk of the observer data is collected when bycatch risk is high but no data are collected when the risk is nil (Figure 1).

5. In a year, the number of fishing operations is $\approx 10,000$, and the number of bycatch events ≈ 300, which yields a rate of $\approx 3\%$. This rate is not large, but is not extremely rare either.

Bycatch events were simulated for each fishing operations during a day when an observer was present from a Bernoulli distribution according to the product of bycatch probability given dolphin presence and dolphin presence probability for that day. If no observer was present, no data were recorded. For each sampling scenario, 100 datasets were generated for 1, 5, 10 or 15 years. All data simulations were carried out in `R` v.4.0.1 (R Core Team, 2020). When simulating only one year of

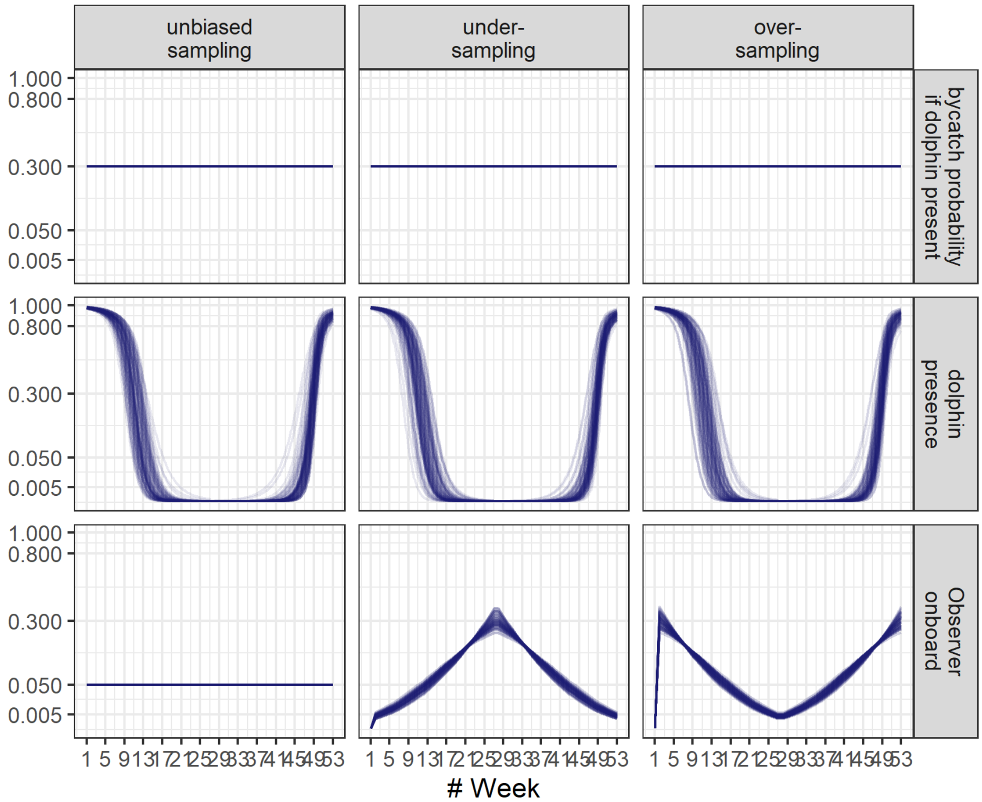

Figure 1: Inputs for data simulation. Top row: bycatch probability if dolphins are present during a fishing operation. Middle row: dolphin presence during a year. Bottom row: Probability for a skipper to accept an observer onboard. Left column: sampling is unbiased; Middle colum: sampling is biased downwards (under-sampling). Right column: sampling is biased upwards (over-sampling). Each line corresponds to one of the 100 data simulations that were carried out. The y-axis is on a square-root scale to better visualize small values.

data, equation 2 is not necessary as there is no between-year variation to estimate. Our Monte Carlo study had a comprehensive factorial design crossing (a) sampling regime (either unbiased or not) and (b) sample size as controlled with the number of years for which the observer programme was assumed to have been in operation.

Estimation

Estimation of the parameters of model 1 from simulated data was carried out in a Bayesian framework using programming language `Stan` (Carpenter *et al.*, 2017) called from `R` v.4.0.1 (R Core Team, 2020) with library `Rstan` (Stan Development Team, 2020). Weakly-informative priors were used for regularization:

$$\begin{cases} \mu \sim \mathcal{N}(0, \frac{3}{2}) \\ \text{prop} \sim \mathcal{D}(1, 1, 1) \\ \sigma_{\text{total}} \sim \mathcal{GG}(\frac{1}{2}, \frac{1}{2}, \frac{\log 2}{10}) \end{cases}$$

where $\mathcal{D}()$ denotes the Dirichlet distribution for modelling proportions (such that $\sum_{i=1}^{3} \text{prop}_i = 1$) and $\mathcal{GG}()$ the Gamma-Gamma distribution for scale parameters (Griffin & Brown, 2017; Pérez *et al.*, 2017). With this parametrization, the several variance components of the model were:

$$\begin{cases}\sigma^2_{\text{vessel}} = \sigma^2_{\text{total}} \times \text{prop}_1\\ \sigma^2_{\text{week}} = \sigma^2_{\text{total}} \times \text{prop}_2\\ \sigma^2_{\text{year}} = \sigma^2_{\text{total}} \times \text{prop}_3\end{cases}$$

These priors are weakly-informative (Gabry *et al.*, 2019): the prior for the intercept allows to cover the whole interval between 0 and 1 on the probability but is informative on the logit scale. The prior for the scale (square-root of the variance) is heavy tailed and has a median set to $\frac{\log 10}{2}$ (Griffin & Brown, 2017; Pérez *et al.*, 2017), which translate an assumption about the plausible range of variations in bycatch risk spanning *a priori* two full order of magnitude from one tenth to a ten-fold increase compared to the mean bycatch rate. Thirty random realisations from our choice of priors are depicted on Figure 2: the whole interval between 0 and 1 is covered, and between-week variations can be large or small.

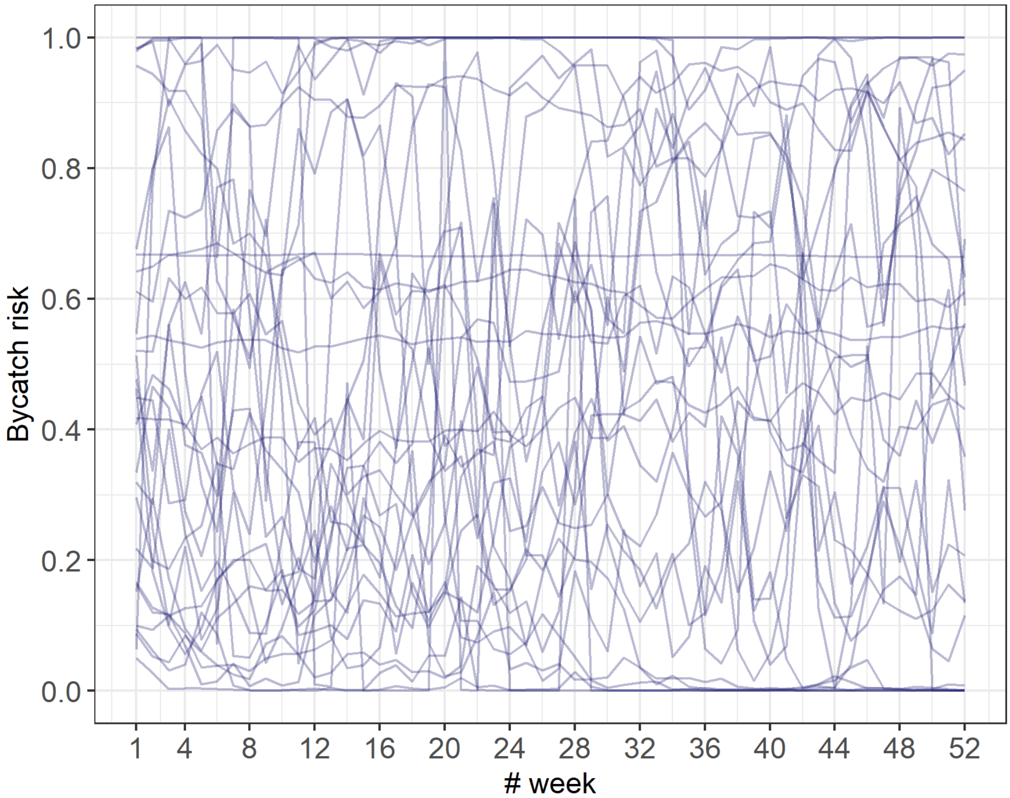

Figure 2: Prior predictive checks *sensu* (Gabry *et al.*, 2019). Bycatch risk (p_{ijkl} in Eq. 1) is depicted: 30 random realizations from the priors are depicted.

For each simulated dataset, four chains were initialized and run for a total of $1,000$ iterations, discarding the first 500 as warm-up. Parameter convergence was assessed using the $\hat{R}$ statistics (Vehtari *et al.*, 2019) and assumed if $\hat{R} < 1.025$.

Using the posterior distribution, bycatch risk $\hat{\mathrm{p}}_{j^*kl}$ for a randomly chosen vessel j^* operating in week k of year l was computed as: $\hat{\mathrm{p}}_{j^*kl} = \text{logit}\left(\hat{\mu} + \hat{\beta}_{kl} + \hat{\alpha}^*_j\right)$ with $\hat{\alpha}^*_j \sim \mathcal{N}(0, \hat{\sigma}_{\text{vessel}})$. This quantity incorporates between-vessel variability, that is it takes into account the fishing style of skippers. The total number of bycatch events, N_{bycatch} was estimated as:

$$\hat{N}^{\text{model-based}}_{\text{bycatch}} = \sum_{l=1}^{n_{\text{year}}} \sum_{k=1}^{n_{\text{week}}} \hat{\mathrm{p}}_{j^*kl} \times N_{kl} \tag{4}$$

where N_{kl} is the total number of fishing operations that took place is week k of year l. Highest

Posterior Density credibility intervals at the 80% level were computed with package `coda` (Plummer *et al.*, 2006) for uncertainty evaluation. Equation 4 is an instance of a ratio-estimator with post-stratification, except that it uses model-based estimates of bycatch risk. This model-based approach allows to regularize estimates with partial pooling (Gelman & Shalizi, 2013): the variance of estimates is greatly reduced by introducing some bias with structured priors (Gao *et al.*, 2019). Our results were benchmarked against an approach similar to that of ICES WGBYC whereby total number of bycatch events was estimated as:

$$\hat{N}_{\text{bycatch}}^{\text{designed}-\text{based}} = \sum_{l=1}^{n_{\text{year}}} \bar{p}_l \times \sum_{k=1}^{n_{\text{week}}} N_{kl} \tag{5}$$

where $\bar{p}_l$ is the average bycatch risk estimated as the mean from the sample of observed bycatch events in year l. Confidence intervals at the 95% level were computed using either the bootstrap or the Clopper-Pearson approach as customary in ICES WGBYC. In practice, ICES WGBYC often pooled several years to stabilize the estimate of $\bar{p}$ (*e.g.* ICES 2018 pages 57-58; Carretta & Moore, 2014): equation 5 translate an ideal case that is rarely met. The total number of strata for post-stratification was $n_{\text{year}} \times n_{\text{week}}$, with a maximum of $15 \times 52 = 780$ cells. ICES WGBYC usually works on bycatch rates (in number of PETS per unit effort), not bycatch risk. We focused on risk for simplicity, but scaling bycatch risk to a rate is straightforward by multiplying with the average number of PETS bycaught in a bycatch event.

Reproducibility

`R` codes to reproduce the results are available at `https://gitlab.univ-lr.fr/mauthier/regularized_bycatch`.

Results

Design- *vs* model-based approach

Comparing the design- and model-based approach was done with simulating one year of data. When data sampling was unbiased, both the design- and model-based approach were able to recover the true number of bycatch events (Figure 3, Table 1). Estimates of bycatch events were statistically unbiased but their precision low with a (frequentist 95%) confidence or (Bayesian 80%) credibility interval (CI) as large as 100% of the point estimate (Table 1). This was unsurprising as only 15 bycatch events were recorded on average by onboard observers (Table 1). With under-sampling, design-based estimates were negatively biased (that is, they were under-estimates) whereas model-based estimates were still unbiased on average (Figure 3, Table 1). With over-sampling, design-based estimates were positively biased (that is, they were over-estimates) but so were model-based estimates, although bias was 5 times smaller (Figure 3, Table 1). In all cases, coverage was 100% but largely as a result of low precision: precision was very low with CI spanning some 200% of the point estimate for the unbiased and under-sampling scenarios. This low precision was the result of having to work with as few as 5 observed bycatch events on average (Table 1). Precision improved with over-sampling, but was still as high as 50% of the point (over-)estimate. The model-based approach was well calibrated in both the unbiased and under-sampling scenarios (Figure 4): model-based estimates were on average equal to the truth whereas this was only the case with design-based estimates when sampling was unbiased. In addition, the model-based approach was able to recover the temporal profile of bycatch risk (Figure 5) in these two scenarios, but with an increased accuracy and precision if sampling was unbiased. In the over-sampling scenario, both the design- and model-based approaches were not well calibrated (Figure 4) and the model-based approach over-estimated bycatch risk when no data were collected (Figures 1 and 5).

Method	Uncertainty	Data sampling	n_{years}	bias (%)	coverage (%)	Width of CI (%)	n_{obs}
Design-based	Bootstrap	unbiased	1	3.5	100.0	102.5	15
Design-based	Clopper-Pearson	unbiased	1	3.5	100.0	115.0	15
Model-based	Bayesian	unbiased	1	3.6	100.0	120.4	15
Design-based	Bootstrap	under-	1	-83.5	100.0	195.0	5
Design-based	Clopper-Pearson	under-	1	-83.5	100.0	259.6	5
Model-based	Bayesian	under-	1	3.0	100.0	204.3	5
Design-based	Bootstrap	over-	1	121.0	100.0	46.1	63
Design-based	Clopper-Pearson	over-	1	121.0	100.0	50.1	63
Model-based	Bayesian	over-	1	22.1	100.0	78.6	63

Table 1: Statistical properties of estimates from the design- and model-based approach. One year of data was simulated a 100 times. Bias of point estimate, coverage of (frequentist 95%) confidence or (Bayesian 80%) credibility interval (CI) and precision (as CI width relative to the point estimate) are reported. The last column indicates the average number of bycatch events ($n_{\text{obs}} = \mathbb{E}\left[\sum_{ijk} y_{ijk}\right]$) that were recorded by onboard observers during data sampling.

Model-based approach with several years of data

With several years of data, the model-based approach was able to yield nearly unbiased estimates: the bias was smaller than 3 bycatch events when sampling was unbiased, but as large as 10 (on average) with biased sampling and three years of data. The precision of estimates improved with several years of data, as expected with larger sample size. Precision of model-based estimates with over-sampling

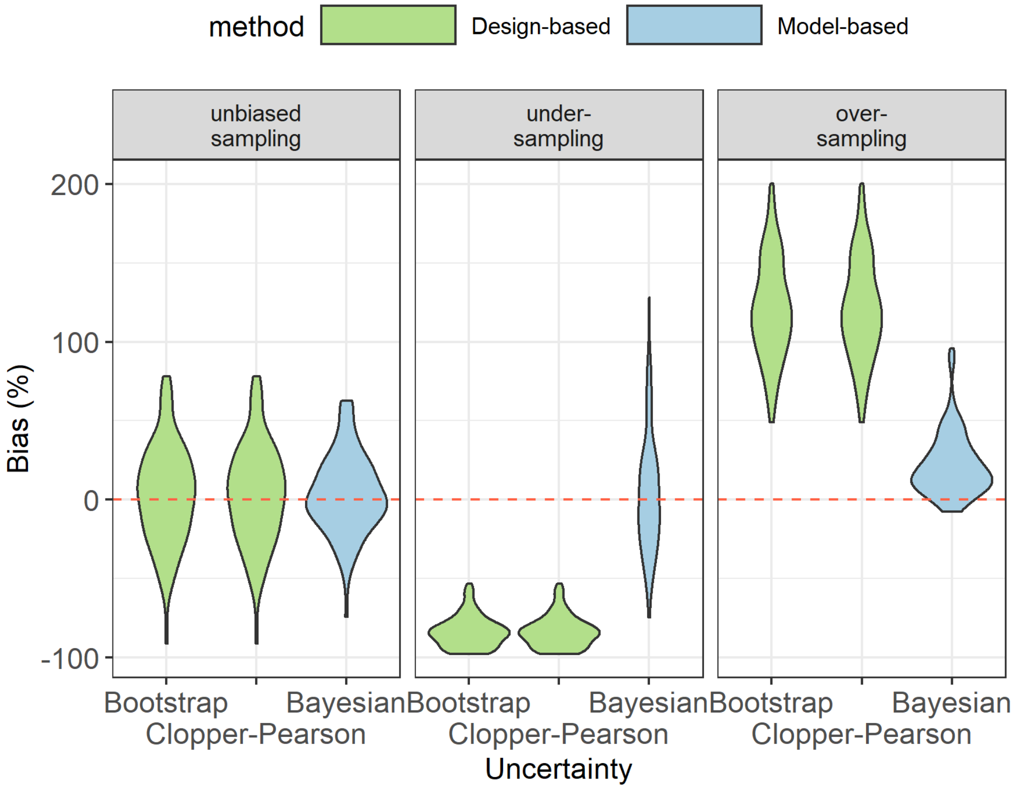

Figure 3: Violin plot of point estimates of total bycatch events. Left column: data sampling was unbiased and all methods yielded statistically unbiased estimates. Middle column: Under-sampling scenario: only the model-based approach was accurate. Right column: Over-sampling scenario: both the design- and model-based approaches were biased upwards. Violin plots are based on 100 simulations.

were already acceptable with 3 years of data: an 80% credibility interval width of 50% corresponds to a coefficient of variation of $\frac{50}{2.5} \approx 20\%$ assuming a normal distribution for the posterior. The model-based approach allowed to obtain estimates at the weekly scale (Figure 6): these estimates were approximately unbiased in the unbiased and over-sampling scenarios, but were biased for the under-sampling scenario. In that latter case, the bias was correlated with the temporal pattern used to simulate dolphin presence (Figure 1): it was the largest when dolphin presence was at its highest but positive at the beginning of a year and negative at the end of the same year. Both biases were greatly attenuated with increased sample size.

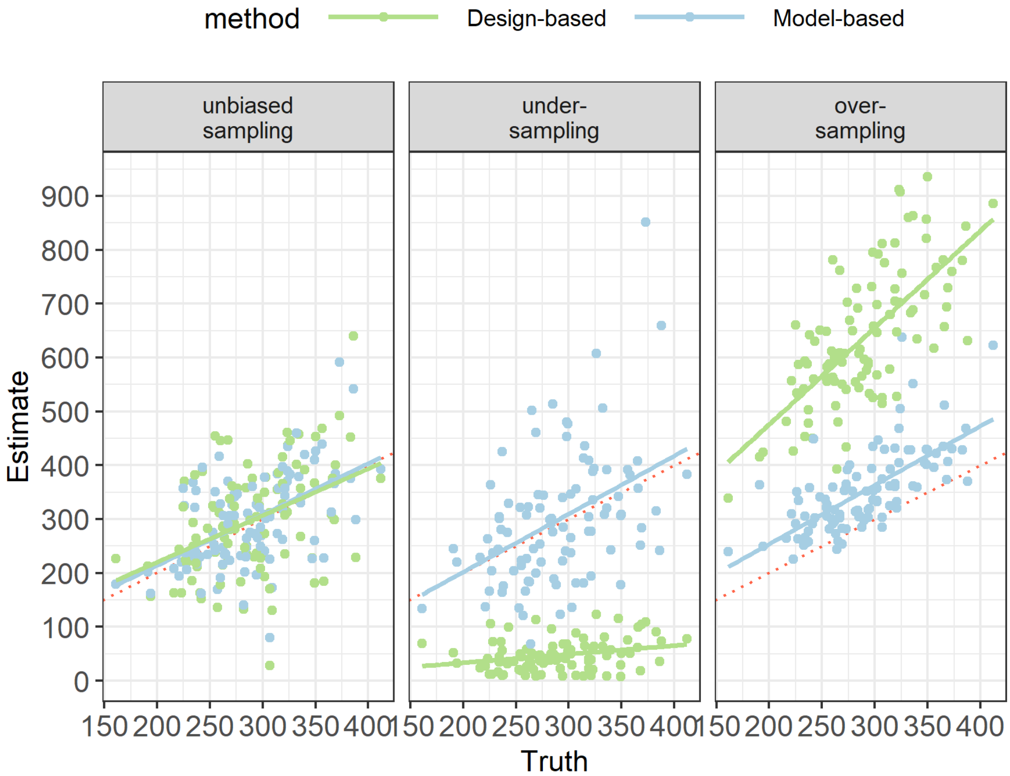

Figure 4: Regression lines of point estimates against the true number of bycatch events, showing the calibration of the design- and model-based approach. The x-axis shows the true number of bycatch events across 100 simulations, spanning between 150 and 400 events. Left: data sampling was unbiased and all methods yielded statistically unbiased estimates. Middle column: Under-sampling scenario: only the model-based approach was well calibrated. Right column: Over-sampling scenario: both the design- and model-based approaches were not calibrated to the truth.

Method	Uncertainty	Data sampling	n_{years}	bias (bycatch events)	coverage (%)	Width of CI (%)	n_{obs}
Model-based	Bayesian	unbiased	3	3.0	100.0	91.1	45
Model-based	Bayesian	unbiased	5	2.1	100.0	76.3	75
Model-based	Bayesian	unbiased	10	1.1	100.0	59.1	150
Model-based	Bayesian	unbiased	15	1.9	100.0	50.9	225
Model-based	Bayesian	under-	3	10.0	100.0	164.6	15
Model-based	Bayesian	under-	5	6.4	100.0	142.0	25
Model-based	Bayesian	under-	10	8.3	100.0	112.9	50
Model-based	Bayesian	under-	15	5.3	100.0	97.8	75
Model-based	Bayesian	over-	3	7.4	100.0	53.2	63
Model-based	Bayesian	over-	5	4.8	100.0	42.6	126
Model-based	Bayesian	over-	10	3.5	100.0	32.6	630
Model-based	Bayesian	over-	15	3.3	100.0	27.7	756

Table 2: Statistical properties of estimates from the model-based approach. Several years of data were simulated a 100 times. Bias of point estimate (in number of bycatch events), coverage of (Bayesian 80%) credibility interval (CI) and precision (as CI width relative to the point estimate) are reported. The last column indicates the average number of bycatch events ($n_{\text{obs}} = \mathbb{E}\left[\sum_{ijkl} y_{ijkl}\right]$) that were recorded by onboard observers during data sampling.

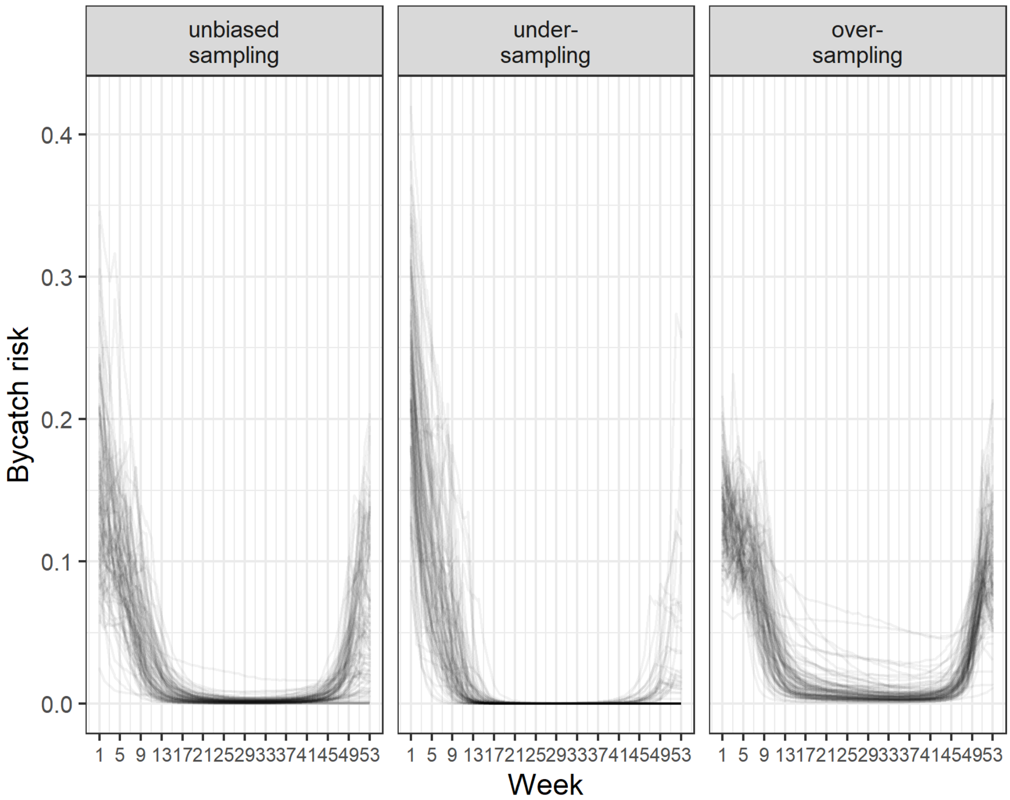

Figure 5: Estimated temporal pattern in mean bycatch risk from the model-based approach. Left column: data sampling was unbiased. Middle column: Under-sampling. Right column: Over-sampling. The model-based approach recovered the correct pattern overall, but overestimated risk in the over-sampling scenarios when risk was, in fact, nil but no data were collected.

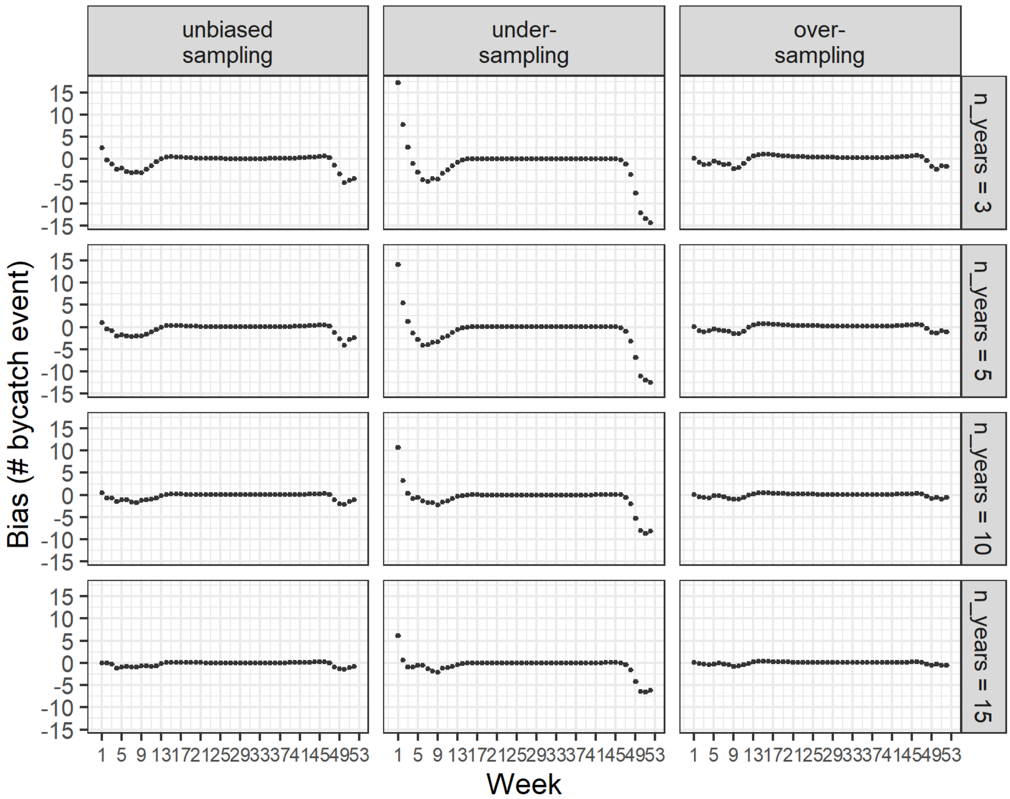

Figure 6: Box plots of bias (in number of estimated bycatch events compared to the truth) in the weekly model-based estimates of bycatch events. Left column: data sampling was unbiased. Middle column: Under-sampling. Right column: Over-sampling. Each row corresponds to data simulated for a different number of years.

Discussion

Using Monte-Carlo simulations, we investigated the statistical properties of a model-based approach, regularized multilevel regression with post-stratification, to estimate the total number of bycatch events in a fishery operating year-round. Simulations were broadly informed from the case of common dolphins and pair-trawlers in the Bay of Biscay and from harbour porpoises and set-gillnets in Celtic Seas. A salient feature of simulations was biased sampling with observers being preferentially accepted onboard when bycatch risk was either high or low. Data simulations in that latter case, which is the most realistic one in the Bay of Biscay (Peltier *et al.*, 2016), resulted in as few as 5 observed bycatch events per year on average (Tables 1, 2). This aligns with the ubiquitous description of small cetacean bycatch being a rarely observed event. It was nevertheless possible to fit a regularized multilevel regression model on these data. Importantly, estimates from this model-based approach were statistically less biased than the design-based estimates when sampling was biased. Model-based estimates were, however, imprecise but this is largely to be expected (Amandè *et al.*, 2012), especially with as few as 5 observed bycatch events per year. The design-based approach was also imprecise, even in the unbiased data sampling scenario of 5% coverage of the fleet, which is not reached in practice (Anonymous, 2016; ICES, 2020b). The design-based approach was very sensitive to how data were collected: this approach severely under- or over-estimated bycatch when sampling was biased, whereas the model-based approach was still well calibrated with under-sampling, but not with over-sampling (Figure 4).

Biases in onboard observer data are pervasive and widely acknowledged (Benoît & Allard, 2009; Peltier *et al.*, 2016). Enforcing coverage as required to achieve a pre-specified precision in estimates can be challenging in practice. For example, in 2016, France only achieved a coverage rate less than 2% for most métiers and concluded on the impossibility of scaling-up observed bycatch rates to the whole fleet (Anonymous, 2016; page 24). There were, however, 9 bycatch events of common dolphins in pair-trawlers targeting European hake (*Merluccius merluccius*). From these numbers, bycatch was described a 'rare' event (Anonymous 2016, page 23). Such a conclusion would be warranted if sampling were representative, in which case the design-based estimate could be used, even though its precision would still be very low. On the other hand, with under-sampling, this conclusion is misleading as our simulations further illustrated: although only 5 bycatch events were observed on average (Table 1), the true number of bycatch events was on average 60 times larger (Figure 4). In our simulations, the true bycatch rate was on average $\approx 3\%$ over a year, which is not rare, but not frequent either. Moreover, interviews with French skippers deploying trawls or gillnets in the Bay of Biscay revealed that more than 80% of respondents declared to having experienced at least one small cetacean bycatch event in a year (Cloâtre, 2020). Such a large proportion contradicts the idea of common dolphin bycatch being a rare event in the Bay of Biscay, but rather suggest severe biases in onboard observer data that result in the rare reporting of bycatch events, rather than a rarity of events *per se*. The common dolphin in the Bay of Biscay illustrates how under-sampling may distort the perception of bycatch as a very rare event when it can, in fact, be widespread. This is a catch-22 situation whereby cetacean bycatch is described as a rare event because it is rarely reported, and this perceived rarity may serve to argue against ambitious dedicated monitoring programmes out of cost-effective considerations, thereby preventing to dispel the initial misconception.

Finding an optimal sampling plan for fisheries with rare bycatch events is long standing problem (ICES, 2009). Several strategies have been attempted: for example in the United States, one strategy is 'pulsed sampling' whereby a particular fishery or métier is very heavily sampled for a short period of time in order to maximize the chance for observers to record any bycatch that might occur (ICES, 2009). This pulsed sampling strategy corresponds to our over-sampling scenario wherein monitoring effort is positively correlated with bycatch risk. Under this scenario, the absence of any sampling at all when bycatch risk was low was detrimental to the accurate estimation of bycatch events with our model. Model-based estimates were, however, less biased than design-based estimates. Arguably, this comparison is somewhat artificial as a correct comparison would use all the available information and exclude the period when bycatch risk is low if such a period is known to the investigator. Notwithstanding this shortcoming, model-based estimates represented an improvement and allowed to infer the bycatch risk profile accurately, especially with several years of data.

We showed with our Monte-Carlo simulations that regularized multilevel regression with post-stratification can nevertheless be used to analyze bycatch data despite concerns about non-representative sampling. Model-based approaches (Palka & Rossman, 2001), with post-stratification (Lennert *et al.*, 1994), or machine learning (Carretta *et al.*, 2017), or multilevel regression (Sims *et al.*, 2008; Martin *et al.*, 2015) have previously been used to estimate bycatch rates. Traditional, design-based, ratio estimates are biased if sampling is biased; imprecise if observer coverage is low (as is the usual case in the North East Atlantic; see for example Figure 14 page 114 in ICES, 2020b); and volatile if bycatch events are only observed occasionally (Carretta *et al.*, 2017). The traditional remedy to stabilize estimates and improve precision is to bypass year-specific estimation and pool several years together (ICES, 2018; Carretta & Moore, 2014). This pragmatic solution improves precision but does not address the problem of biased sampling. It also introduces estimation bias for any year-specific estimates by pooling completely several years in order to stabilize the variance of estimates (ICES, 2009; page 36): any between-year differences are thus ignored in order to obtain a better precision of estimates. It is a reasonable approach in practice, but one that can be improved. Model-based approaches offer a trade-off between no-pooling (keeping all years separate) and complete-pooling with a third option: partial pooling or regularization (Gelman & Shalizi, 2013). Regularization is a general term for statistical procedures that give more stable estimates. Our model-based approach achieves regularization by leveraging, via a structured prior model (equations 2 and 3, see **Methods**), the within-year information at the weekly scale. The result were more stable and accurate annual bycatch estimates at the cost of some modelling assumptions and weakly-informative priors. Importantly, weekly estimates could also be obtained with our model-based approach.

Our model-based approach is semi-parametric as it uses a random walk prior to learn from the data the weekly pattern in bycatch risk. This prior is also ensuring some smoothness in the temporal risk profile as it translates an assumption on the correlation between two consecutive weeks. This random walk model remains simple as the correlation is fixed to 1 and not estimated. We further complexified this model to allow for between-years variation in the weekly risk profile with a Gaussian Process prior (Neal, 1998; Goldin & Purse, 2016). Importantly, these two prior choices (a random walk and a Gaussian Process prior) add structure to the model and help in leveraging the information present in the sparse data typical of onboard observer programmes. Even when with over-sampling, these choices were not detrimental as model-based estimates were statistically unbiased and precise with 3 years of data (Table 2). The explicit consideration of time effects is key to mitigate bias in sampling. In our simulations, dolphin presence was caricaturally seasonal, and observers could be preferentially allowed on fishing vessels when dolphins were less or more likely to be present (Figure 1). Our model was still able to provide statistically unbiased estimates of bycatch in those scenarios, although these estimates were very imprecise with under-sampling. However, they were not more imprecise than the traditional (but biased) design-based estimates (Table 1) if 80% credibility interval were used. In addition to being unbiased, these estimates could also reveal with accuracy the temporal risk profile (Figure 5). It is important to keep in mind here that our model is different from the data-generating model used in simulating data: our results were not simply an instance of using a true model, which is impossible in practice as a model is by definition a simplification used to capture the salient features of a phenomenon. Our model had some shortcomings: for example, bias increased with 3 years of data compared to 1 year for the under-sampling scenario (contrast Tables 1 & 2). This increased bias was the result of partial pooling but came with a gain in precision as evidenced in the width of credibility intervals. The bias progressively wore off with more years of data, illustrating thereby the attractiviness of partial pooling and structured priors to regularize estimates (Gelman & Shalizi, 2013; Gao *et al.*, 2019). The gain in reducing bias in estimates and increasing their precision was most evident with over-sampling (Table 1 & 2).

Our model could also provide weekly bycatch estimates which were largely unbiased except in the under-sampling scenario where a positive and negative bias remained at the beginning and end of a year respectively, even with 15 years of data (Figure 6). With under-sampling, few observed bycatch events can be collected by design because observers are very unlikely to be accepted on board by skippers. Weekly estimates were too high at the beginning of a year but too low at the end, but this somewhat cancelled out at the year-level. There was still a slight overestimation bias resulting from our choice of a non-symmetric pattern for dolphin presence and a symmetric pattern for biased coverage: observing bycatch events at the end of a year was comparatively more difficult than at the beginning of a end because overlap between a non-nil coverage and dolphin presence was smaller at the end of year (Figure 1). These shortcomings illustrate that a model-based approach should be tailored to the context of the study, and we designed our simulations largely from our knowledge on the common dolphin in the Bay of Biscay. However, the framework of regularized multilevel regression with post-stratification is very flexible and we believe our proposed model has large potential for generality as it simply translates a decomposition of bycatch risk into a smooth time-varying and (unstructured) time-invariant effects. The model can easily be made more complex, data permitting, to accommodate spatial effects with, for example, a Besag-type prior (Sims *et al.*, 2008; Morris *et al.*, 2019).

An important limitation of our model is that it is phenomenological, *i.e.* it is agnostic of the causes behind the temporal variations in bycatch risk. Bycatch risk is the product of dolphin presence and bycatch probability given presence (the latter was constant in our simulations). The model only estimates this product of two probabilities and thus cannot distentangle them without other sources of data. This limitation is inconsequential for the aim of accurate estimation of the total number of bycatch events as interest lies in the effects of causes (how much bycatch?) rather than in the causes of effects (why bycatch occurred?). An important assumption underlying accurate estimation is that the information on the total effort must also be accurate and available at the scale of weeks for post-stratification. This assumption is crucial to scale-up estimates from the (potentially biased) sample to the population, but it does not necessarily hold with fisheries effort as the latter is more often estimated rather than measured directly (Julian & Beeson, 1998; ICES, 2018, 2020b). Here we assumed that the total number of fishing operations (*e.g.* number of tows for trawls; Tremblay-Boyer & Berkenbusch, 2020) are available as auxiliary information for post-stratification. This assumption about the availability of disaggregated data stems from the explicit consideration of time as an important predictor of variations in bycatch risk. This assumption is necessary for using post-stratification to align the sample with the population targets.

This assumption on the availability of accurate effort data at fine temporal scale may be difficult to meet in pratice. Currently, ICES WGBYC uses in its BRA a coarse, but admittedly comparable proxy across fisheries and countries to quantify fishing effort, namely days at sea (ICES, 2019). A day at sea is any continuous period of 24 hours (or part thereof) during which a vessel is present within an area and absent from port (Anonymous, 2019a). Importantly, this definition is not at the level of a fishing operation, and effort thus quantified is already aggregated at a level above that at which bycatch data are collected. This coarsening of fisheries effort data is fundamentally a measurement problem, and one that modelling should not be expected to remedy easily. BRA uses an estimate of total fishing effort for the fisheries of concern in a specific region, together with some estimate of likely or possible bycatch rates that might apply for the species of concern, in order to evaluate whether or not the total bycatch in that area might be a conservation issue. A regularized multilevel regression model could be used to obtain estimates of bycatch rates to be used in BRA. Post-stratification could also be attempted using the coarse days at sea proxy for effort, and thus our framework could well be adapted to match the requirements of ICES WGBYC.

Conclusion

We investigated with simulations the ability of multilevel regularized regression with post-stratification to estimate cetacean bycatch under various sampling scenarios. These scenarios were all caricatural in some respect. The unbiased sampling case appears unrealistic in practice: biased sampling, either under-sampling or over-sampling (ICES, 2009), may be the general case. We considered both cases, under quite extreme scenarios whereby data collection was highly correlated with bycatch risk, resulting in either very few observed events with under-sampling, and a large number of observed events with over-sampling. In both cases, multilevel regularized regression with post-stratification was able to produce nearly unbiased bycatch estimates with as few as 5 observed events data. With only one year of data, precision was low, especially with under-sampling, and there was some estimation bias with over-sampling one. These results stemmed from the extreme scenarios we considered but illustrate nevertheless that a model cannot be expected to solve all the deficiencies of data collection and measurement. Good measurement is key for accurate estimation and our results actually re-emphasize the importance of design. However, they also show that a good data collection design *and* an adequate modelling framework are synergistic and allow to extract a lot of information for sparse data. Assuming a normal distribution for the bycatch estimates, a 80% Bayesian CI width divided by 2.5 gives an idea of the associated coefficient of variation: the model-based approach can yield a coefficient of variation of 50% with as few as 15 observed events if sampling is unbiased. With under-sampling, one would need 10 years of data (under our data simulation schemes) to obtain the same precision. This re-iterates the need to (i) have dedicated observer schemes, (ii) ensure adequate coverage and (iii) use a model-based approach tailored to extract as much information as possible from sparse data, as the first two points are very difficult to live up to in practice.

The key assumptions behind regularized multilevel regression with post-stratification in our simulations are that bycatch risk changes smoothly through time and that accurate data on the number of fishing operations at the same temporal scale are available (*e.g.* number of tows for trawls; Tremblay-Boyer & Berkenbusch, 2020). When both assumptions can be reasonably entertained, we showed how a model-based approach using recent methodological developments is efficient, irrespective of how data were collected. A further asset of the explicit consideration of a temporal scale is that it may help in pinpointing more precisely windows of heightened risk in order to target adequate mitigation measures (*e.g.* spatio-temporal closures). The framework of multilevel modelling is very flexible and can accommodate spatial effects and other complexifications, data permitting. Regularization will, in general, be needed to mitigate data sparsity and leverage partial pooling in order to obtain stable estimates of bycatch. Given the satisfactory performance of regularized multilevel regression with post-stratification in our simulations, we recommend further investigations using this technique to estimate bycatch rate and numbers from both representative or non-representative samples. A re-analysis of > 15 years of observer data on common dolphin bycatch in pair trawlers flying a French flag is currently underway (Rouby et al. *in preparation*) in order to obtain better bycatch estimates that could be further used to estimate conservation reference points in order to better manage this fishery (Cooke, 1999; Punt *et al.*, 2021).

Acknowledgments

References

Amandè, M.J., Chassot, E., Chavance, P., Murua, H., Molina, A.D. & de Bez, N. (2012) Precision in Bycatch Estimates: the Case of Tuna Purse-Seine Fisheries in the Indian Ocean. *ICES Journal of Marine Science*, **69**, 1501–1510.

Anonymous (2016) Rapport annuel sur la mise en øeuvre du règlement européen (CE) 812/2004 établissant les mesures relatives aux captures accidentelles de cétacés dans les pêcheries. Rapport annuel réglementaire. techreport, Direction des pêches maritimes et de l'aquaculture (DPMA).

Anonymous (2019a) Commission Implementing Decision (EU) 2019/910 of 13 March 2019 establishing the multiannual Union programme for the collection and management of biological, environmental, technical and socioeconomic data in the fisheries and aquaculture sectors.

Anonymous (2019b) Regulation (EU) 2019/1241 of the European Parliament and of the Council of 20 June 2019 on the conservation of fisheries resources and the protection of marine ecosystems through technical measures.

Authier, M., Saraux, C. & Péron, C. (2017) Variable Selection and Accurate Predictions in Habitat Modelling: a Shrinkage Approach. *Ecography*, **40**, 549–560.

Avila, I.C., Kaschner, K. & Dormann, C.F. (2018) Current Global Risks to Marine Mammals: Taking Stock of the Threats. *Biological Conservation*, **221**, 44–58.

Benoît, H. & Allard, J. (2009) Can the Data from at-Sea Observer Surveys be Used to Make General Inference about Catch Composition and Discards? *Canadian Journal of Fisheries and Aquatic Sciences*, **66**, 2025–2039.

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M. & White, J.S. (2009) Generalized Linear Mixed Models: a Practical Guide for Ecology and Evolution. *Trends in Ecology and Evolution*, **24**, 127–135.

Brownell, R.J., Reeves, R., Read, A., Smith, B., Thomas, P., Ralls, K., Amano, A., Berggren, P., Chit, A.M., Currey, R., Dolar, M.L.L., Genov, T., Hobbs, R.C., Kreb, D., Marsh, H., Zhigang, M., Perrin, W.F., Phay, S., Rojas-Bracho, L., Ryan, G.E., Shelden, K.E.W., Slooten, E., Taylor, B.L., Vidal, O., Ding, W., Whitty, T.S. & Wang, J.Y. (2019) Bycatch in Gillnet Fisheries Threatens Critically Endangered Small Cetaceans and Other Aquatic Megafauna. *Endangered Species Research*, **40**, 285–296.

Cam, E. (2012) "each site has its own survival probability, but information is borrowed across sites to tell us about survival in each site": Random Effects Models as Means of Borrowing Strength in Survival Studies of Wild Vertebrates. *Animal Conservation*, **15**, 129–132.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017) `Stan`: A Probabilistic Programming Language. *Journal of Statistical Software*, **76**.

Carretta, J.V. & Moore, J.E. (2014) Recommendations for Pooling Annual Bycatch Estimates when Events are Rare. techreport NOAA-TMNMFS-SWFSC-528, NOAA. NOAA Technical Memorandum.

Carretta, J.V., Moore, J.E. & Forney, K.A. (2017) Regression Tree and Ratio Estimates of Marine Mammal, Sea Turtle, and Seabird Bycatch in the California Drift Fishery: 1990-2015. Technical Report NOAA-TM-NMFS-SWFSC-568, NOAA. NOAA Technical Memorandum, 83 p.

Cloâtre, T. (2020) Projet LICADO - Limitation des captures accidentelles de dauphins communs dans le golfe de Gascogne. Bilan des enquêtes auprès des professionnels du golfe de Gascogne. Version 4.

Cooke, J.G. (1999) Improvement of Fishery-Management Advice Through Simulation Testing of Harvest Algorithms. *ICES Journal of Marine Science*, pp. 797–810.

Course, G.P., Pierre, J. & Howell, B.K. (2020) What's in the Net? Using Camera Technology to Monitor, and Support Mitigation of, Wildlife Bycatch in Fisheries. Technical report, Worldwide Wildlife Fund. 104 pp.

Dolman, S., Evans, P., Ritter, F., Simmonds, M. & Swabe, J. (2020) Implications of New Technical Measures Regulation for Cetacean Bycatch in European Waters. *Marine Policy*, **124**, 104320.

Downes, M., Gurrin, L.C., English, D.R., Pirkis, J., Currier, D., Spittal, M.J. & Carlin, J.B. (2018) Multilevel Regression and Poststratification: A Modeling Approach to EstimatingPopulation Quantities From Highly Selected Survey Samples. *American Journal of Epidemiology*, **187**, 1780–1790.

Fay, R. & Herriot, R. (1979) Estimates of Income for Small Place: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 269–277.

Forsberg, O.J. (2020) Polls and the US Presidential Election: Real or Fake? *Significance*, **17**, 6–7.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. & Gelman, A. (2019) Visualization in Bayesian Workflow. *Journal of the Royal Statistical Society Series A*, **182**, 389–402.

Gao, Y., Kennedy, L., Simpson, D. & Gelman, A. (2019) Improving Multilevel Regression and Poststratification with Structured Priors. resreport, University of Toronto & Columbia University.

Gelman, A. (2007) Struggles with Survey Weighting and Regression Modelling. *Statistical Science*, **22**, 153–164.

Gelman, A., Hill, J. & Vehtari, A. (2021) *Regression and Other Stories*. Cambridge University Press, 1st edition edition.

Gelman, A. & Shalizi, C. (2013) Philosophy and the Practice of Bayesian Statistics. *British Journal of Mathematical and Statistical Psychology*, **66**, 8–38.

George, E.I. & Ročková (2021) Comment: Regularization via Bayesian Penalty Mixing. *Technometrics*.

Goldin, N. & Purse, B.V. (2016) Fast and Flexible Bayesian Species Distribution Modelling Using Gaussian processes. *Methods in Ecology and Evolution*, **7**, 598–608.

Gray, C.A. & Kennelly, S.J. (2018) Bycatches of Endangered, Threatened and Protected Species in Marine Fisheries. *Reviews in Fish Biology and Fisheries*, **28**, 521–541.

Griffin, J. & Brown, P. (2017) Hierarchical Shrinkage Priors for Regression Models. *Bayesian Analysis*, **12**, 135–159.

Halsny, V. (2020) Nonresponse Bias in Inequality Measurement: Cross-Country Analysis Using Luxembourg Income Study Surveys. *Social Science Quaterly*, **101**, 712–731.

Hansen, M.H. & Hurwitz, W.N. (1946) The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association*, **41**, 517–529.

ICES (2009) Report of the Study Group for Bycatch of Protected Species (SGBYC), 19-22 January 2009, Copenhagen, Denmark. ICES CM 2009/ACOM:22. 117 pp. Technical report.

ICES (2018) Report from the Working Group on Bycatch of Protected Species, (WGBYC). 1-4 May, Reykjavik, Iceland. 128 pp. . Technical Report No. ICES CM 2018/ACOM:25, International Council for the Exploration of the Sea.

ICES (2019) Working Group on Bycatch of Protected Species (WGBYC). ICES Scientific Reports. 1:51. 163 pp. Faro, Portugal. Technical report, International Council for the Exploration of the Sea.

ICES (2020a) Bycatch of Protected and Potentially Vulnerable Marine Vertebrates - Review of National Reports under Council Regulation (EC) No. 812/2004 and Other Information. ICES Advice : Ecoregions in the Northeast Atlantic and adjacent seas. Technical report. 16 pp.

ICES (2020b) Report from the Working Group on Bycatch of Protected Species, (WGBYC). techreport 2:81, International Council for the Exploration of the Sea. 209 pp.

Jaramillo-Legorreta, A.M., Cardenas-Hinojosa, G., Nieto-Garcia, E., Rojas-Bracho, L., Thomas, L., Ver Hoef, J.M., Moore, J., Taylor, B., Barlow, J. & Tregenza, N. (2019) Decline Towards Extinction of Mexico's Vaquita Porpoise (*Phocoena sinus*). *Royal Society Open Science*, **6**.

Julian, F. & Beeson, M. (1998) Estimates of Marine Mammal, Turtle, and Seabird Mortality for Two California Gillnet Fisheries: 1990-1995. *Fishery Bulletin*, **96**, 271–284.

Kiewiet De Jonge, C.D., Langer, G. & Sinozich, S. (2018) Predicting State Presidential Election Results Using National Tracking Polls and Multilevel Regression With Poststratification (MRP). *Public Opinion Quarterly*.

Komoroske, L.M. & Lewison, R.L. (2015) Addressing Fisheries Bycatch in a Changing World. *Frontiers in Marine Science*, **2**.

Lax, J. & Phillips, J. (2009) How Should We Estimate Public Opinion in the States? *American Journal of Political Science*, **53**, 107–121.

Lennert, C., Kruse, S., Beeson, M. & Barlow, J. (1994) Estimates of incidental marine mammal bycatch in California gillnet fisheries for July through December, 1990. *Report of the International Whaling Commission*, volume Special Issue 15, pp. 449–463.

Martin, R., Pearmain, E.J., Burfield, I.J., Small, C., Phillips, R.A., Yates, O., Lascelles, B., Garcia Borboroglu, P. & Croxall, J.P. (2019) Threats to Seabirds: a Global Assessment. *Biological Conservation*, **237**, 525–537.

Martin, S.L., Stohs, S.M. & Moore, J.E. (2015) Bayesian Inference and Assessment for Rare-Event Bycatch in Marine Fisheries: a Drift Gillnett Fishery Case Study. *Ecological Applications*, **25**, 416–429.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S.J., Gelman, A. & DiMaggio, C. (2019) Bayesian Hierarchical Spatial Models: Implementing the Besag York Mollié Model in Stan. *Spatial and Spatio-Temporal Epidemiology*, **31**, 100301.

National Marine Fisheries Service (2004) Evaluating Bycatch: a National Approach to Standardized Bycatch Monitoring Programs. Technical Report NOAA Tech. Memo. NMFSF/SPO-66, U.S. Department of Commerce, NOAA. 108 pp.

National Marine Fisheries Service (2005) Revisions to Guidelines for Assessing Marine Mammal Stocks. Technical report, NOAA. 24 pp.

National Marine Fisheries Service (2016) Guidelines for Preparing Stock Assessment Reports Pursuant to the 1994 Amendments to the MMPA. Technical report, NOAA. 24 pp.

Neal, R. (1998) *Bayesian Statistics*, chapter Regression and Classification Using Gaussian Process Priors, pp. 000–000. Oxford University Press.

Pacoureau, N., Rigby, C.L., Kyne, P.M., Sherley, R.B., Winker, H., Carlson, J.K., Fordham, S.V., Barreto, R., Fernando, D., Francis, M.P., Jabado, R.W., Herman, K.B., Liu, K., Marshall, A.D., Pollom, R.A., Romanov, E.V., Simpfendorfer, C.A., Yin, J.S., Kindsvater, H.K. & Dulvy, N.K. (2021) Half a Century of Global Decline in Oceanic Sharks and Rays. *Nature*, **589**, 567–571.

Palka, D.L. & Rossman, M.C. (2001) Bycatch estimates of coastal bottlenose dolphin (*Tursiops truncatus*) in US Mid-Atlantic gillnet fisheries for 1996 to 2000. Technical Report Ref. Doc. 01-15, Northeast Fisheries Science Center. 77 pp.

Peltier, H., Authier, M., Caurant, F., Dabin, W., Daniel, P., Dars, C., Demaret, F., Meheust, E., Van Canneyt, O., Spitz, J. & Ridoux, V. (2021) In the Wrong Place at the Wrong Time: Identifying Spatiotemporal Co-occurrence of Bycaught Common Dolphins and Fisheries in the Bay of Biscay (NE Atlantic) from 2010 to 2019. *Frontiers in Marine Science*.

Peltier, H., Authier, M., Deaville, R., Dabin, W., Jepson, P., van Canneyt, O., Daniel, P. & Ridoux, V. (2016) Small Cetacean Bycatch as Estimated from Stranding Schemes: the Common Dolphin Case in the Northeast Atlantic. *Environmental Science & Policy*, **63**, 7–18.

Pérez, M., Perrichi, L.R. & Ramírez, I. (2017) The Scaled Beta2 Distribution as a Robust Prior for Scales. *Bayesian Analysis*, **12**, 615–637.

Plummer, M., Best, N., Cowles, K. & Vines, K. (2006) CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, **6**, 7–11.

Punt, A.E., Siple, M., Francis, T.B., Hammond, P.S., Heinemann, D., Long, K.J., Moore, J.E., Sepúlveda, M., Reeves, R.R., SigurÃ°sson, G.M., Víkingsson, G., Wade, P.R., Williams, R. & Zerbini, A.N. (2021) Can We Manage Marine Mammal Bycatch Effectively in Low-Data Environments ? *Journal of Applied Ecology*.

R Core Team (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rhem, J., Kilian, C., Rovira, P., Shield, K.D. & Manthey, J. (2020) The Elusiveness of Representativeness in General Population Surveys for Alcohol. *Drug and Alcohol*, **40**, 161–165.

Sala, A., Konrad, C. & Doerner, H., eds. (2019) *Review of the Implementation of the EU Regulation on the Incidental Catches of Cetaceans (STECF-19-07)*. JRC117515. Publications Office of the European Union, Luxembourg.

Sims, M., Cox, T. & Lewison, R. (2008) Modeling Spatial Patterns in Fisheries Bycatch: Improving Bycatch Maps to Aid Fisheries Management. *Ecological Applications*, **18**, 649–661.

St. John, F., Keane, A., Jones, J. & Milner-Guiland, E. (2014) Robust Study Design Is as Important on the Social as It Is on the Ecological Side of Applied Ecological Research. *Journal of Applied Ecology*, **51**, 1479–1485.

Stan Development Team (2020) RStan: the R interface to Stan. R package version 2.21.2.

Stransky, C. & Sala, A. (2019) Scientific, Technical and Economic Committee for Fisheries (STECF) - Revision of the EU-MAP and Work Plan template (STECF-19-12). Technical report, Publications Office of the European Union, Luxembourg. ISBN 978-92-76-11255-6.

Tregenza, N.J.C., Berrow, S.D., Hammond, P.S. & Leaper, R. (1997) Harbour porpoise (*Phocoena phocoena* L.) By-Catch in Set Gillnets in the Celtic Sea. *ICES Journal of Marine Science*, **54**, 896–904.

Tremblay-Boyer, L. & Berkenbusch, K. (2020) Characterisation of Marine Mammal Interactions with Fisheries & Bycatch Mitigation. Report for project INT2019-03 prepared for Department of Conservation. Technical report, Dragonfly Data Science. 64 pages.

Ulrich, C. & Doerner, H., eds. (2021) *Scientific, Technical and Economic Committee for Fisheries (STECF) - 66th Plenary Report (PLEN-21-01).* Publications Office of the European Union, Luxembourg.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. & Bürkner, P. (2019) Rank-Normalization, Folding, and Localization: an Improved $\hat{R}$ for Assessing Convergence of McMC. resreport, Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, Finland.

Wang, W., Rothschild, D., Goel, S. & Gelman, A. (2015) Forecasting Elections with Non-Representative Polls. *International Journal of Forecasting*, **31**, 980–991.

Zahorski, A. (2020) Multilevel Regression with Post-Stratification for the National Level `VIBER/Street` Poll on the 2020 Presidential Election in Belarus. Technical report.

Zhang, X., Holt, J.B., Lu, H., Wheaton, A.G., Ford, E.S., Greenlund, K.J. & Croft, J.B. (2014) Multilevel Regression and Poststratification for Small-Area Estimation of Population Health Outcomes: a Case Study of Chronic Obstructive Pulmonary Disease Prevalence Using the Behavioural Risk Factor Surveillance System. *American Journal of Epidemiology*, **179**, 1025–1033.