Economists counting fish: modeling profit-maximizing behavior to improve stock assessments

Y. Allen Chena\*, Cole C. Monnahanb, Owen S. Hamela, Alan C. Hayniec

a Fishery Resource Analysis and Monitoring Division, Northwest Fisheries Science Center, NOAA Fisheries, 2725 Montlake Blvd. E., Seattle, WA 98112

b Resource Ecology and Fisheries Management Division, Alaska Fisheries Science Center, NOAA Fisheries, 7600 Sand Point Way NE, Bldg. 4, Seattle, WA 98115

c International Council for the Exploration of the Sea, H. C. Andersens Boulevard 44-46, 1553 Copenhagen V, Denmark

\* Corresponding author, allen.chen@noaa.gov

# Abstract

Stock assessments are an integral part of contemporary fisheries management, estimating the health of targeted populations and ensuring sustainable marine resource management. Fishery-independent data (e.g., scientific surveys) are preferred as input data to assessments because they usually more reliably reflect changes in the population, despite fishery-dependent data being more widely and consistently available, and potentially less expensive to compile. One reason the use of fishery-dependent data in assessments is treated with care is because the data do not account for the spatial site selection and incentives of fishers. By understanding and explicitly modeling how fishers make tradeoffs, economic models can correct for selection if fishers systematically choose areas with greater expected catches. We develop and test an approach to correct abundance indices taking into account the sampling process of fishers in a simulation framework. Corrected indices can supplement fishery-independent surveys and improve the accuracy of spawning biomass estimates in some scenarios, but are unreliable in others, e.g., when there is poor spatial coverage coupled with a shifting spatial trend. We demonstrate the potential for economic models as a tool to standardize indices of abundance, improve stock assessment estimates, and the management based upon them.

# Keywords

Selection bias, economics, fishery-dependent data, CPUE index standardization

# Introduction

Stock assessments estimate the status and sustainable levels of catch in exploited fisheries and are a vital tool for sustainable marine resource management (Hilborn & Walters 1992). When stock assessments use fishery-dependent observations to construct relative abundance indices, it is important to account for effects other than true changes in the population trend, such as gear or spatial changes, or the site selection and incentives of fishers (Eales & Wilen 1986), through a process known as index standardization (e.g., Maunder & Punt 2004; Walters 2003). Fishers do not sample randomly across fish habitat, but instead choose areas to target different species, densities, and sizes of fish. Standardizing fishery-dependent indices to correct for these factors is an ongoing challenge to producing accurate and reliable stock assessments and the associated advice used by fisheries managers.

There have been a number of fisheries studies that examine standardization of catch per unit effort (CPUE) indices, for example by allowing correlation between observations (Bishop et al. 2004), taking into account changes in fishing power (Allen & Punsly 1984), or incorporating targeting strategies that vary spatially (Hoyle & Okamoto 2013). More recently, geostatistical approaches were shown to have improved properties and this class of models has proliferated (Thorson et al. 2015, 2020). However, little attention has been devoted to index standardization methods that use economic drivers, which offer a mechanistic understanding into the sampling process that generated the fishery-dependent data. This information could potentially be used to calculate more accurate relative indices.

Economic models that explicitly characterize how fishers make tradeoffs can correct for selection that occurs if, for example, fishers choose to systematically fish at areas with greater expected catches or size-driven price premiums. The study of self-selection in economics is well-developed and robust, the seminal example being the work of Heckman (1979). An example of correcting for selection bias in a polychotomous choice setting is provided by Dahl (2002), examining how migrants choose where to live based on their expected wages. In a fisheries extension, Chen et al. (2023) modelled how fishers choose areas in the Bering Sea pollock fishery, suggesting a method that estimates average catches at areas, corrected for selection bias.

Chen et al. (2023) were primarily interested in understanding the spatial tradeoffs and preferences of fishers, which allows policymakers and researchers to evaluate policies such as spatial management strategies. An intermediate step necessary for developing that inference was to obtain unbiased estimates of fisher catches across space: a researcher cannot understand how fishers value catch if estimates of catch are incorrect. We hypothesize that the same process to correct for fisher-induced selection could potentially be used to create corrected fishery-dependent indices of abundance that better reflect changes in stock abundance.

Our methods use a structural model of behavior where fishers optimize over a given set of preferences, are statistically testable for selection, and can incorporate any number of relevant factors or covariates into the fisher’s profit function. This provides an opportunity to improve existing index standardization methods by explicitly incorporating structural models of fisher behavior (and not just economic covariates), and provide a mechanistic understanding of their site selection process, which to our knowledge is novel in this literature. We detail the model and test it in a simulation framework using scenarios that vary in how fish are distributed spatially and how fisher decisions are made. Then, we apply the economic model to recover corrected fishery-dependent indices of abundance. Finally, we supply the corrected indices to a stock assessment to quantify the improvement to management advice.

# Methods

To investigate the potential of using fishery-dependent indices of abundance corrected for selection bias, we conduct simulations using a biological operating model. This operating model generates age-structured population dynamics for each simulation replicate. Given a population trajectory we then simulate annual spatial sampling patterns with a fishing model that encodes fisher behavior. The fishing model and simulated sampling patterns generate fishery-dependent data, which can then be corrected with an econometric model of fisher behavior that recognizes that fishers trade off the expected catch and revenue in an area with the costs of fishing in different areas. The corrected data are then used to construct indices of relative abundance, which are included in an estimating model to evaluate if they improve stock assessment estimates.

We compare assessments with the addition of uncorrected versus corrected indices of abundance, as well as indices from a fishery constructed with random rather than preferential fishing area selection (emulating an additional annual fishery-independent survey). Note that we do not complete the loop in the simulation (Figure 1), as the estimating model is not used to dictate catch limits or fishing mortality rates in the operating model, but rather the fishing mortality rate *F* is fixed for each year within the operating model. This still allows us to evaluate how correcting the fishery-dependent abundance index improves the assessment, without conducting a full closed-loop simulation analysis.

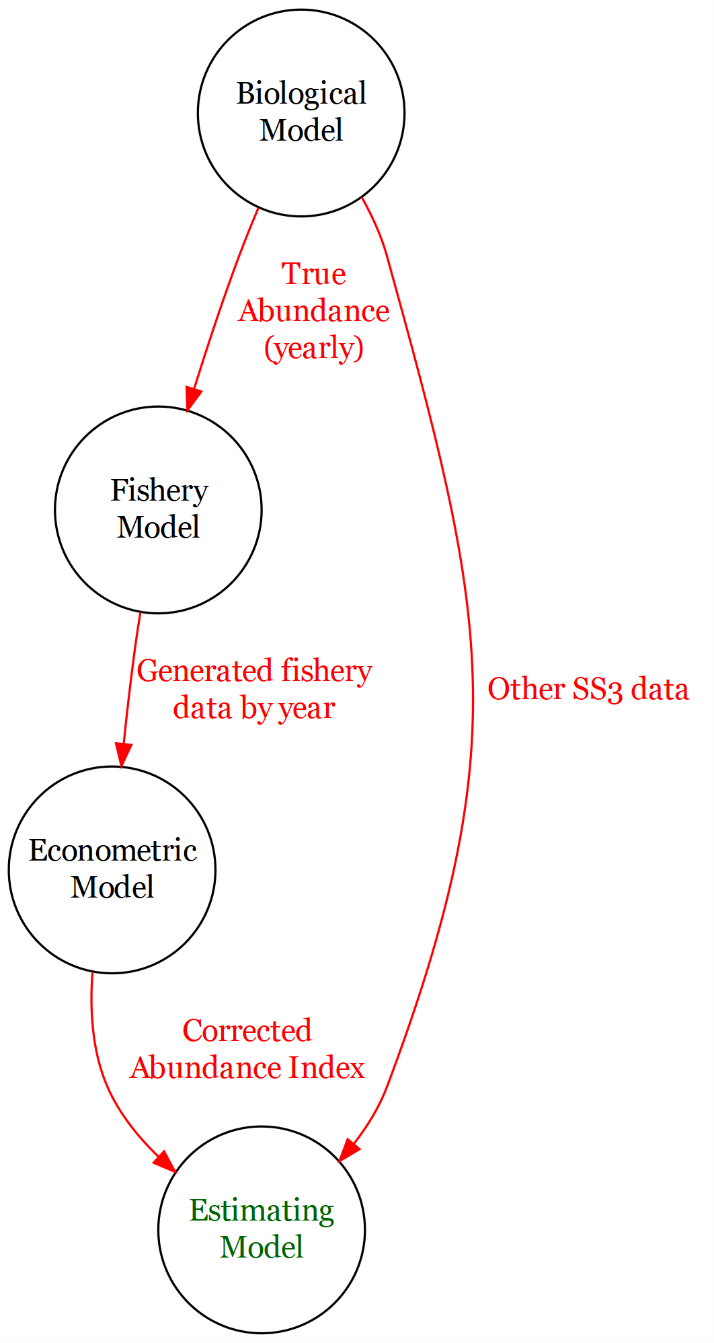


Figure 1: Conceptual flowchart of the simulation approach. The Operating and Estimating models are matching Stock Synthesis (SS3) assessment models, and other SS3 data include survey indices and compositions.

## 2.1 The “true” biological operating model

We use the R package ss3sim (Anderson et al. 2014) to configure a single-sex, age-structured fishery based on a “slow-growing and long-lived” cod-like species, calibrated using North Sea cod. This package allows us to simulate a “true” abundance trend in its operating model, and then test the effect of different fishery sampling patterns in the estimating model using the assessment software Stock Synthesis (Methot and Wetzel 2013). The cod-like species is a default model included in the ss3sim package, and is a similar model to those investigated in other fisheries applications (e.g., Johnson et al. 2015, Ono et al. 2015). Key life-history and modeling parameter calibrations are described in Table 1, and we refer the reader to previous studies cited above for more information. The modified package, including the fishery model, econometric correction, and code to reproduce the results in this paper, is available on the corresponding author’s Github.[[1]](#footnote-2)

Table 1. Key parameters and quantities used in the operating model and estimating model (which match structurally). If parameters are estimated the initial values are the true values, and if not estimated the estimating model uses the true value (there is no misspecification).

|  |  |  |  |
| --- | --- | --- | --- |
| Operating model parameters and structure | Symbol | Value | Estimated |
| Natural mortality |  | 0.2 | No |
| Reference year |  | 1 | No |
| Mean length-at-age |  | 20 | Yes |
| Mean asymptotic size |  | 132 | Yes |
| Growth rate |  | 0.2 | Yes |
| Coefficient of variation of |  | 0.1 | Yes |
| Coefficient of variation of |  | 0.1 | Yes |
| Length-weight scaling |  | 6.8e-6 | No |
| Allometric factor |  | 3.1 | No |
| Maturity slope |  | -0.27 | No |
| Length at 50% maturity |  | 38.2 | No |
| Number of length bins | -- | 45 | No |
| Fishery selectivity beginning size for plateau |  | 50.8 | Yes |
| Fishery selectivity width of plateau (*ln(width)*) |  | -3 | No |
| Fishery selectivity ascending width (*ln(width)*) |  | 5.1 | Yes |
| Fishery selectivity descending width (*ln(width))* |  | 15 | No |
| Survey selectivity beginning size for plateau |  | 41.8 | Yes |
| Survey selectivity width of plateau |  | -4 | No |
| Survey selectivity ascending width (*ln(width)*) |  | 5.2 | Yes |
| Survey selectivity descending width (*ln(width))* |  | 14 | No |
| Log-catchability (fishery-dependent CPUE index) |  | 0 | Yes |
| Log-catchability (survey) |  | 0 | No |
| Log mean virgin recruitment |  | 18.7 | Yes |
| Steepness |  | 0.65 | No |
| Log recruitment deviation |  | 0.4 | No |
| Survey observation error standard deviation |  | 0.2 | No |

We explore a linearly increasing fishing mortality rate starting in year 62 and ramping up to FMSY (~0.17; Figure 1). Natural mortality does not vary with time nor age and is fixed at a constant 0.2. Fishery and survey size selectivity is specified as double normal but is calibrated to be logistic (see the parameters in Table 1). Growth uses the von Bertalanffy equation, while spawner-recruitment is specified with Beverton-Holt, both as parameterized in Table 1. Finally, our process error takes the form of independent lognormal random deviations around median recruitment. The choices in the operating model produce the true abundance trend and total catches, from which simulated data are generated (described below) and used in the estimation and fishing model.

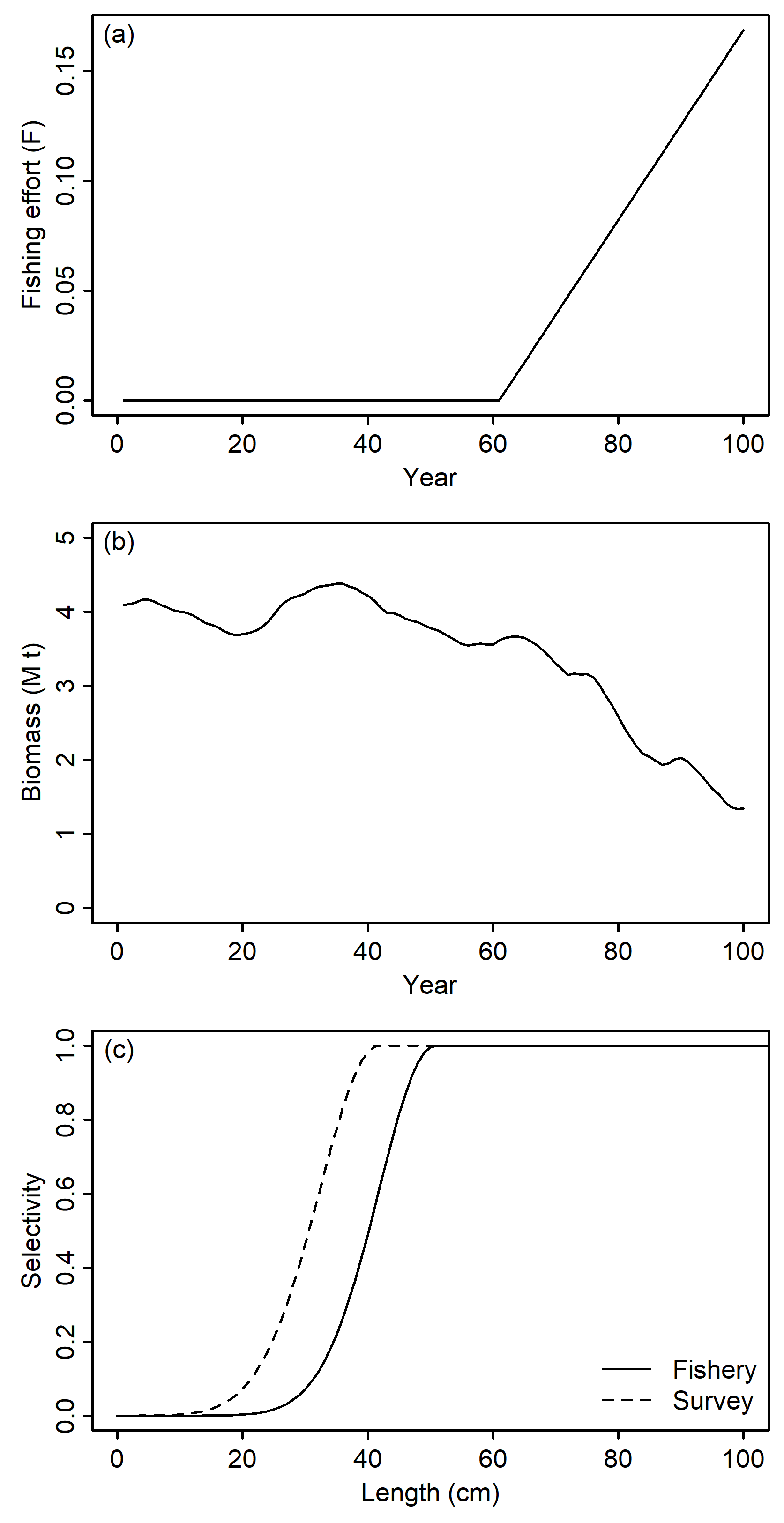


Figure 2. The assumed trend in fishing mortality in the operating model (a), and an example population trend from one replication under that mortality (b). The first 60 years without fishing are used to equilibrate the simulated population. Example fishery selectivity and survey selectivity (c), which are double-normal parameterized to be logistic-shaped.

## 2.2 Simulated fishing model

Notably, the ss3sim framework is not spatially explicit. To characterize a fishing model that allows fishers to choose different areas, we must distribute total abundance in year *y* available to the fishery across a defined spatial grid. We assign each grid cell *j* a relative abundance, parameterized by whose sum is equal to a scalar with constant value over all years, and we multiply these by the total “true abundance” for that year. Thus, the only factor changing relative abundance across years is a factor scaling total operating model abundance in year *y*,from all *J* areas, relative to the true mean across years :

|  |  |  |
| --- | --- | --- |
|  |  | (1a)   (1b) |

Relative abundance among areas might be static over time, or we can calibrate to change across *y* (for example representing fish movement); however, the sum of relative abundances are always constant across *y*, such that the only change in summed relative abundance across years comes from operating model abundance. While this stylized model represents a simple case, and does not take into account important considerations such as area-specific fishing intensity or localized depletion, for example, this simplification allows us to take a first step at delineating ideal conditions where the correction works well, and begin to identify conditions where the correction struggles.

After biomass from the operating model is distributed spatially according to , tThus the fishing model simulates artificial catch data comprising *N* independent haul occasions each year given the true spatial abundance for a single year. Similar to fisheries observers that record catches on fishing vessels, we then assume data for *Y* total years that have observer coverage, in each replicate of the simulation. The population is simulated for 100 years total, allowing burn-in, but only the last *Y* years have fishing activity and observer coverage. Note there could be years with fishing activity but no observer coverage, depending on parameterization.

We take fishing mortality from the operating model as exogenous, simulating haul level data until total catches equal what is dictated by the operating model. Therefore the fishing mortality from the biological operating model dictates total catches, while the fishing model only simulates how fishers would make spatial decisions and potentially choose areas with greater realized CPUE. This is akin to a fishery where, for example, management sets a total allowable catch (TAC), and the TAC is always fulfilled in our fishery. An extension for future research could be to estimate the effects from declines in effort. Also note that fishery size selectivity is assumed to be independent of catch area, and we do not estimate an impact on size/age structure.

### 2.2.1 Modeling fishing behavior

For each simulated haul, we first specify the catch at area *j* on the *ith* haul occasion as a linear combination of covariates with dimension *1xn* where *n* equals the number of vessel-specific covariates, scaled by area-specific coefficients with dimension *nx1*, an area-specific constant, plus some unobservable variation (Chen et al. 2023), where represents vector multiplication. Vessel-specific covariates could include vessel characteristics or previous catches in an area, for example. We allow part of this variation to be known by the fisher, while the remainder is stochastic: the specific form of our selection bias results from private information known to the fisher that is unobserved by the analyst.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

We assume the errors are non-correlated and normally distributed (the specified parameters of these errors are listed in Table 2). The assumed shape of the distribution is the choice of the analyst and can be accounted for in the final likelihood; i.e. the joint likelihood can be parameterized for different distributional assumptions, however, an assumption must be made.

Our fishing model assumes a fisher starting at area *k* chooses among *M* discrete areas on haul *i*, for *i = 1 … N.* such that the fisher moves to the *jth* area as long as the utility at that area is greater than at all other areas: by maximizing utility fishers will choose to fish where they derive the greatest net benefit, and this utility describes the objective function that fishers optimize over. In the case where all utilities are equal the fisher would randomly choose a location; this is not an outcome that has ever occurred in our simulations (due to the precision of floating-point random numbers in *R*). We assume all areas are potentially accessible to the fisher from the current area, where *M* is the total number of alternatives. However, an area may not practicably be accessible depending on its distance, or a fisher’s disutility for travel, et cetera.

|  |  |  |
| --- | --- | --- |
|  | Choose area *j* on *ith* haul *if and only if* | (3) |

Part of the fisher’s utility, denoted , is allowed to depend on any number of individual- or area-specific variables known to the analyst and is therefore the deterministic portion of the fisher’s utility. is therefore the deterministic portion of the fisher’s utility. Notably, the fisher’s utility depends on their expectations of catches at an area,, where is the marginal utility from catch, their starting area *k*, and a vector of parameters that scales area- and individual-specific variables , where and are estimated. For the fisher’s expectation , the fisher’s information set includes the realization of , as well as average catches . Other covariates such as lagged catches, capturing historical experience, could also be included in (and is generally common in the fisheries economics literature). Therefore,

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

where varies in an unobservable way and is assumed to be independently and identically distributed generalized extreme value type I (Gumbel).

Note that while our stylized model only includes one species in the fisher’s utility function, it is relatively straightforward to include multiple species, as long as the analyst has catch data for those species. Our approach allows analysts to incorporate any number of relevant factors or covariates into the fisher’s utility function. Our vessels vary based on their vessel-specific covariate , but additional factors representing vessel heterogeneity or differences in catchability could also be included.

Expected catch of a different species could also be modeled in the same fashion as above, and appended as an additional term (similar to). The marginal utility from catch is an estimated parameter and is agnostic as to sign; bycatch for example might have a negative coefficient. Finally, if the analyst also believes the catch of the other species might suffer from selection bias, additional correction functions can be added for that species. Catch of the two different species might also be correlated, which would require parametric assumptions as to the joint distribution of catch for the two species. As multiple species are added, the likelihood of encountering the curse of dimensionality also increases, as each species requires an additional polynomial correction at each area.

Common steps between all four specifications in how we model fisher behavior and produce different realized catches are outlined below. Steps 4 and 5 describe the econometric model and

1. The catch data are generated for each area on a *3x3* or *2x4* grid from the data-generating process in equation (2), depending on the scenario, where the elements of are stochastic, and are fixed. We choose these grid sizes as a practical matter, to simplify the model with the goal of comparing asymptotic behavior. Catches at areas are scaled over years to match the exogenous abundance trend from the operating model; i.e. if abundance decreases, the deterministic part of catch) must decrease to match. The variable can be thought of as a vessel-specific covariate that affects catchability (e.g. vessel power). Note that any number of covariates affecting catch could be included, depending on the data the analyst has on hand. We only generate one covariate that affects catch. We assume haul duration here is constant for all hauls, normalizing catches to catch per unit effort.

1. Fishers start at some given area and make their decision - representing a single choice occasion - according to the information available to them. The starting area may represent a port, the area of the last haul, or may potentially abstract from how the fisher arrived at that area altogether, depending on the scenario specification. We allow a fisher to choose a single area according to the selection criteria in equation (3), given stochastic, fixed and, where the fisher knows,, and the realization of , but not. Notably, this implies the fisher’s expectation of catch is a part of the selection criteria, allowing them to choose sites with greater relative abundance, if they desire. The fisher then moves to the site with the largest of its expected utility values. The vessel fishes at this new site and the realized catch (including both ) at the chosen area is included in the data available in the assessment.

Steps 1 and 2 are repeated for *N* independent haul occasions until the sum of catches from all hauls equal total catches from the operating model. In practice the sum of catches from N independent hauls will exceed total catch in the OM slightly; fishing ceases when catches from the last haul simulated meets or exceeds the total catch. This comprises the catch data for a single year *y*, and this process is repeated for all *Y* years. This provides a simulated yearly time series of hauls and haul-level data for all years in a single replicate, based on economic factors driving fishery behavior. Note that we simulate hauls, not individual fishers. For each haul, realizations of and are drawn. To the extent and represent vessel-specific covariates, there are a minimum number of vessels equal to the number of unique permutations of and. However, different vessels could also have similar vessel characteristics (two vessels may both be 100 meters long). We simulate individual hauls until catches bind, but the number of vessels is not directly modeled.

Our choice of a 3x3 or 2x4 grid is unlikely to represent the true spatial extent of many fisheries, but illustrates proof of concept for our methods. Computationally, a model with a larger number of areas might take a few days to converge. This is manageable for practical use, where an analyst may only need to update their fishery-dependent abundance index once a year, perhaps by running a few different candidate models (investigating the inclusion of various covariates or distributional assumptions). However, because we are interested in asymptotic behavior, our workflow requires running a model for each year fishing occurs, for 200 replicates, which requires thousands of model runs. A smaller grid size allows us to report asymptotic results.

### 2.2.2 Spatial abundance and fishing behavior scenarios

We compare four scenarios describing different spatial fishing patterns and spatial abundances in Table 2, which represent conditions where the statistical correction is hypothesized to be more or less successful. Column I lists the name of each scenario. Column II broadly describes the scenario and what we believe might result from that specific fisher behavior and abundance pattern. Column III describes fisher behavior and spatial coverage resulting from that sampling. For example, in Scenario 1, vessels start (are seeded) at a random area. They can then choose to move to another area, or stay in the seeded area. Or, in Scenario 2, vessels make multiple hauls in a trip. The subsequent haul uses the last chosen area as the current area. Then the vessel can choose to move to another area, or stay at the current area. Finally Column IV describes the spatial abundance pattern of the fishery stock.

An area’s distance from the vessel can be important, depending on the relative abundance at more distant areas as well as a fisher’s disutility for travel. By scaling these factors, we can control fisher spatial coverage which then affects the variation in vessel types and trip types we see in our data, and correspondingly the efficacy of a statistical correction for selection bias in catch data. We investigate what happens if there is good fisher coverage (Scenario 1), if some areas are not sampled by fishers (Scenarios 3 and 4), and when the abundance at areas follows different temporal patterns (Scenario 4).

Table 2. Fisher behavior and spatial abundance scenarios, along with their hypothesis and justification.[[2]](#footnote-8)

|  |  |  |  |
| --- | --- | --- | --- |
| I. Scenario name | II. Description and hypothesis | III. Fisher behavior and coverage | IV. Spatial abundance of the stock |
| 1: Baseline | All areas are well-sampled, and fishers of different characteristics are well distributed across different starting areas. We hypothesize the corrected index completely purge the selection bias in observed catches. | There is an equal probability of starting in each area: vessels only make one haul from the starting area, choosing to stay or go to another area. | Relative abundance is fixed per area and constant over time:  *.* |
| 2: Partial fisher coverage (PFC) | Fishers always start in Area 1 corresponding to the port, but are allowed to make multiple hauls in a trip. Some areas may not be well sampled or omitted entirely, and not all vessels visit all areas. We hypothesize accuracy of the corrected index may vary year by year. | Vessels always start at *j=1*. Then, they make a haul as in the baseline scenario, and continue recursively making hauls with the starting area as the last fished area. The probability of continuing fishing decreases with each subsequent haul, where fishing is continued until a random draw , where is the cumulative number of hauls number and approximating the average number of hauls in a trip. | Same parameterization as **1: Baseline**. |
| 3: PFC, random spatial abundance | The spatial distribution of abundance changes from year to year. However, we hypothesize the corrected index may still perform well on average if sampling is not consistently skewed to areas with larger or smaller catches. | Same parametrization as **2: Partial fisher coverage (PFC)**, except with a habitat barrier that impedes fisher coverage in some areas | The abundance at an area is random in every year: |
| 4: PFC, trending spatial abundance | Spatial abundance follows an eastwardly trend over time. With fewer samples in distant areas, we hypothesize the corrected index may miss the spatial trend over time, performing poorly. | Same parametrization as **2: Partial fisher coverage (PFC)**, except with a habitat barrier that impedes fisher coverage in some areas | Abundance follows an eastwardly trend as *y* increases. |

Our four scenarios therefore characterize different spatial fishing patterns and spatial abundances, so we can examine when a statistical correction for selection bias would perform better or worse. Figure 3 demonstrates an example where fishing occurred in a sample replicate-year for Baseline Scenario 1 (left panel) and Partial Fisher Coverage Scenario 2 (right panel). In both of these scenarios, the spatial distribution of abundance is static across all years, shaded in blue. In Scenario 1 however, each “trip” (an example trip is illustrated in red) consists of only a single haul, and a vessel’s starting area is randomized. The example trip in red, for instance, has the vessel randomly placed in Area 9 to start, and then the vessel examines its alternatives, and chooses to move to Area 6, where it makes a single haul.

This might correspond to a world where fishers have preferred starting areas, based on their past experience, and the fisher steams to that area to begin their fishing trip, but this data-generating process abstracts from how fishers might make multiple hauls in a trip. Then, the baseline scenario observes a large number of observations starting in each area (labeled in green). These are catch observations as if there were a hypothetical observer on board; note length observations are not included in the fishery model but rather passed directly from the operating model. Because the correction function uses probabilities of moving to an area as covariates, variation in the types of vessels and different starting areas is also important to the performance of the correction. For example the starting area determines the distance to each area, and therefore the probability a vessel travels to each area, where we might expect greater distance to decrease the probability.

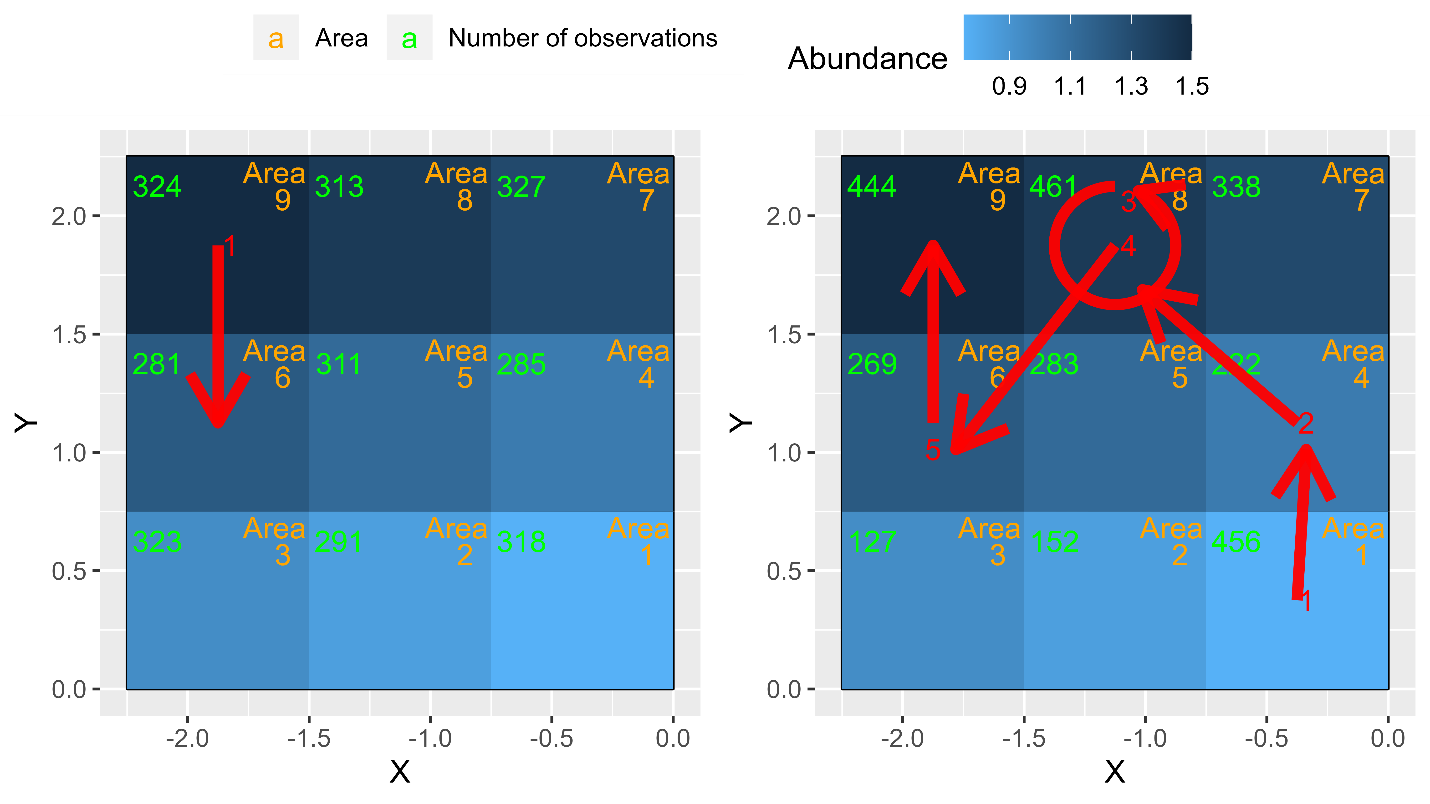


Figure 3: Scenarios 1 (left panel) and 2 (right panel). Areas are shaded by abundance, while each sequence of arrows represents a representative (randomly chosen) trip. Red numbers represent a choice occasion, and the arrow where the vessel moved to.



As an alternative, in Scenario 2, vessels always start at Area 1. Then they take multiple hauls in a trip, where the red arrows illustrate five choice occasions in a single hypothetical trip. For example, for choice occasion 1 (in red), they start in Area 1, and choose to move to Area 4 and make a haul. Then for choice occasion 2, they are already in Area 4, and choose to move to Area 8, and make a haul. For choice occasion 3, they choose to stay in Area 8 and make another haul. For choice occasion 4, they move to Area 6, make a haul, and finally for a total of 5 hauls in this trip, during their last choice occasion they move to Area 9 and make their final haul.

When all vessels start in Area 1, they tend to migrate towards the areas with greater abundance. Then the probability of moving to an area with low abundance is relatively small. Note that areas with low abundance in the southwest have fewer observations, and depending on the choice of polynomial approximation the number of observations may need to be larger in order to have sufficient variation in the probability covariates to identify the free parameters in the correction function.

In Scenario 3 there is a buffer in the form of habitat (e.g. reefs or rocks) that prevents fishing (Figure 3). We investigate the effect of not being able to estimate abundance in some areas, including what happens if the westward area contains most of the abundance in a year. Fishers start in Area 1, and make multiple hauls in a trip. Abundance across space is not static across years, but random across space as described in Column IV of Scenario 3, where each panel is one realization of the random distribution (in one year). For example in one year on the left panel, a vessel starts in Area 1, moves to Area 4 to make a haul, and then chooses to stay in Area 4 and make a second haul. The right panel then represents the randomly distributed abundance in a different year, where a vessel starts in Area 1, and stays there two make two hauls.

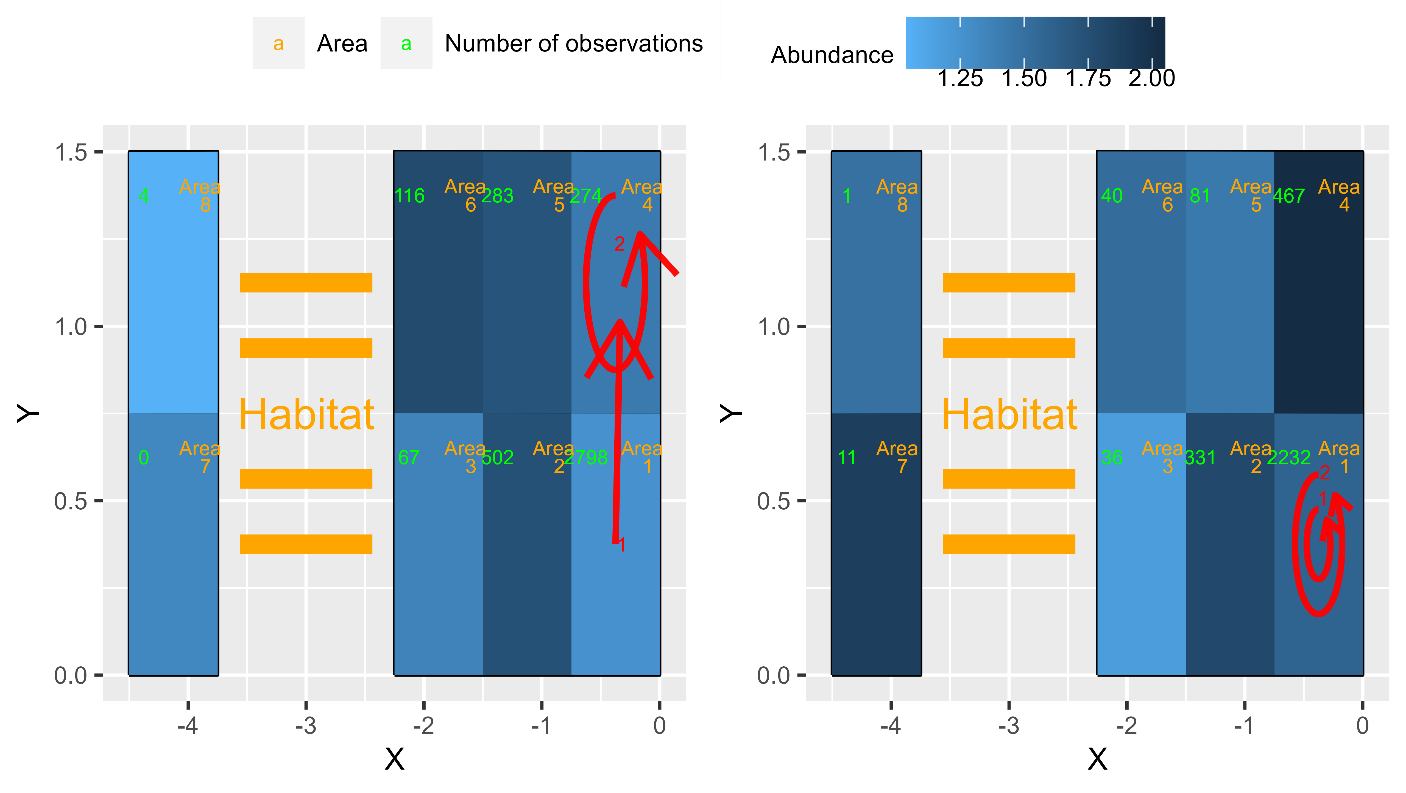


Figure 4: Scenario 3 Random Abundance. Areas are shaded by abundance, while each sequence of arrows represents a representative (randomly chosen) trip. Each panel represents the spatial abundance in a different year. Habitat and cost of travel makes some areas untenable for fishing.

Therefore, in some years the unfished area might be high abundance, while in other years there might be low abundance in the unfished area. There are also not enough observations to estimate catches at the most westward areas. To be just identified there needs to be a number of observations equal to the number of location-specific parameters, equal to the number of covariates in and the number of parameters in the polynomial approximation. However, in practice, a larger sample is required for good performance because the correction relies on a good amount of variation in the probabilities used in the correction functions.

Finally in Scenario 4 there is an eastward trend in spatial abundance. The panels in Figure 5 correspond to years 1 (upper left), 5 (upper right), 9 (lower left), and 13 (lower right) respectively, for an example replicate. Fishers start in Area 1, and make multiple hauls in a trip. As each year passes, the spatial distribution of the population in terms of abundance shifts eastward, while fishing vessels still do not frequent the sites furthest away.

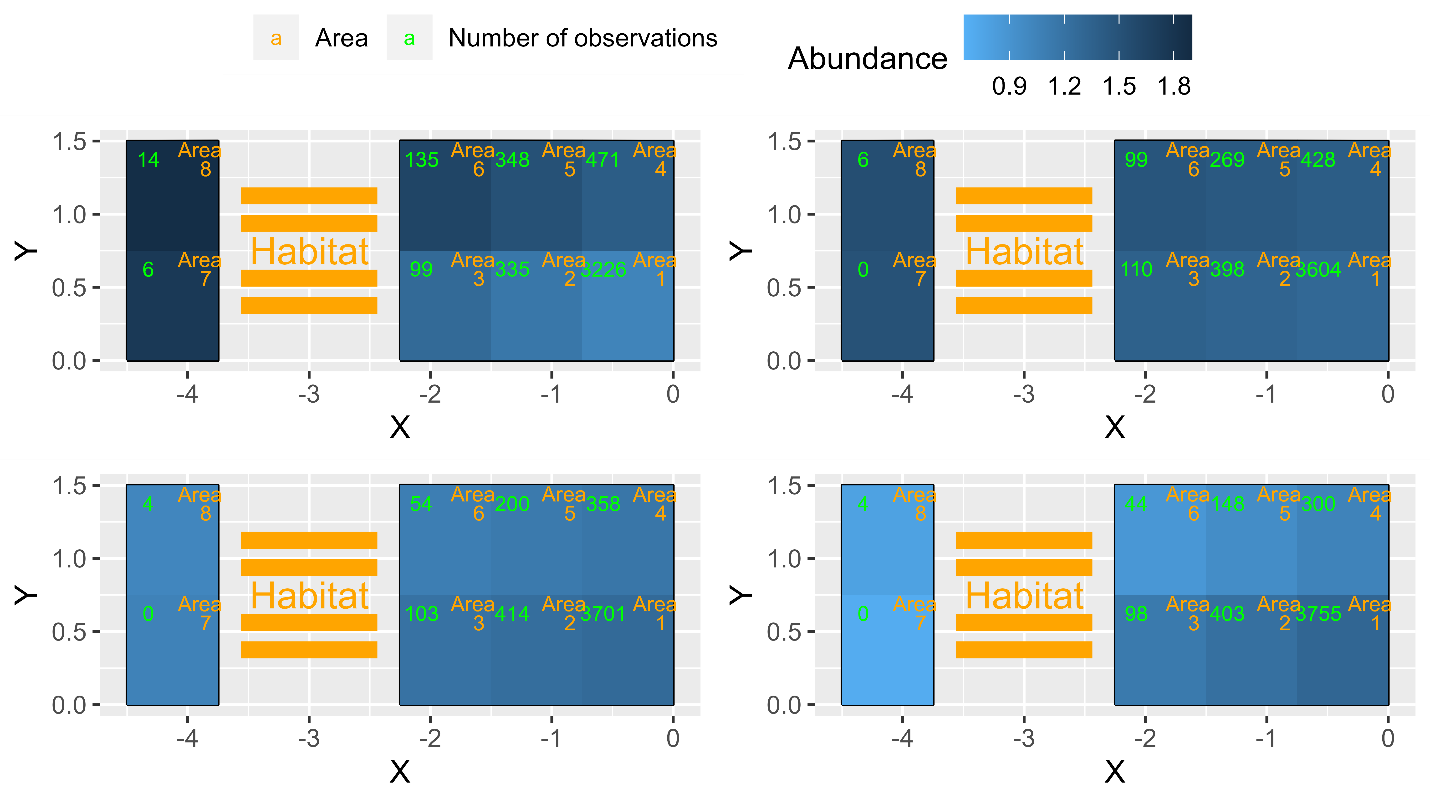


Figure 5: Scenario 4 Spatial Trend. Habitat and cost of travel makes some areas untenable for fishing. The panels correspond to years 1 (upper left), 5 (upper right), 9 (lower left), and 13 (lower right) respectively.

The fishing model characterizes a non-random sampling process induced by fisher behavior and extends the biological model to produce true spatial catches, given fisher behavior and spatial abundance. These catches are then passed to an econometric model.

## 2.3 Econometric model of fisher behavior and corrected abundance indices

The role of the econometric model is to try to recover the true spatial abundance from the observed catches (i.e., decisions) and limited knowledge about fisher behavior. We investigate three different constructions of an abundance index: an uncorrected index, a corrected index, and a randomly sampled index. While the first two differ as to whether a statistical correction is applied, the randomly sampled index does not use the selection criteria in equation (3), but rather randomly chooses a site independent of its characteristics for each haul. This is meant to mimic the effect of including an additional (more frequent and ongoing) fishery-independent survey which provides an abundance index rather than an absolute abundance estimate. The resulting abundance indices are then passed to each of three respective estimating models, and used with a fishery-independent abundance index from a survey which occurs every four years.

The assumption that the error distribution is extreme value type I allows a convenient expression of the probability of choosing area *j*, where *m* denotes all other areas besides area *j*.

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

This is the common conditional logit model (McFadden 1974). Note that parameters and are only identified up to a scale parameter ; that is, we can identify the extent fishers value catch relative to distance , for example, but not their absolute values (McFadden 1974). Then, when we substitute in the fisher’s expectation of catch the probability of choosing the *jth* area is a function of the private information available to the fisher. For instance, when the probability of choosing a site increases with catch, the analyst who is creating an abundance index is more likely to observe an area when the fisher knows catches are larger at an area, and would more likely observe larger catches. Conversely if the fisher desires to avoid bycatch, for example, the opposite intuition would hold true.

It follows that catch data, *conditional on observation*, will also incorporate this private information, which is correlated with the probability of choosing an area. This is because the fisher observes the realization of in equation 2 (but not). As a result, when fishers choose areas (and catches) that result in the greatest expected utility, and visit areas when they have private information fishing is good at the area. Specifically, note that enters directly into the probability of choosing an area in equation 5.

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

While the mean of the unconditional expectation of is zero, the conditional mean is not. If we interpret as an area-specific average, observed catches will deviate from the average at an area equal to a selection bias term. The resulting fishery-dependent data are more likely to be characterized by observations with larger catches, relative to a randomized site sampling (i.e., a scientific survey or uninformed fisher). Scientific surveys usually follow some semi-deterministic sampling plan, but with a design to avoid statistical bias.

If an analyst could derive unbiased estimates of, they could predict catches and create an abundance index using some weighted sum over areas, for example. However, in this case, if analysts regressed observed covariates on observed catches, they risk obtaining as the sample of observed catches is biased (equation 6). This would bias estimates of as well. A correction function approach (Dahl 2002) allows us to approximate the conditional error as a polynomial function. This allows consistent estimation of the remaining parameters in the catch equation. Specifically, the polynomial is a function of the probability of visiting an area and observing the catch (, where is a vector of coefficients to be estimated that corresponds to each term of the polynomial. For example, a second-order polynomial for a fisher who moved to area *j* from area *k* might be written as, which is then included in the regression as additional covariates. Then, we estimate the catch function simultaneously with the fisher’s choice problem, where we can simultaneously obtain the probabilities.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

Note that the probabilities are equal to equation (5) and are equal to the second term of the full likelihood (equation 7). Then, the first term of equation 7 is the likelihood of observing the catch, given the assumption of normality (, the second term the probability of selection (equation 5), and the full likelihood is the probability of observing the catch data, conditional on the observed area being chosen. In addition, note that the existence of selection bias is empirically testable, to the extent that the polynomial terms for an area enter statistically significantly during estimation. For example, we can jointly test the significance of a correction function at the median values of the data (probabilities), using the delta method. Statistical significance implies the existence of a non-zero selection bias term.

The term is a kernel function that shrinks as the probability of observation decreases. We need a kernel function to estimate both levels in the catch equation as well as a constant in the polynomial correction, which would otherwise be perfectly collinear. A constant appears in both the catch equation and the polynomial correction. The specific kernel function we use is equal to, but alternative forms are feasible under specific assumptions (derived in Andrews & Schafgans 1998). Intuitively, we expect the selection bias term to be non-zero when the probability of observation is low yet the area was chosen anyway, implying the existence of private information the analyst has not accounted for. Chen et al. (2023) describe additional simulations and an empirical example of this model.

Note that the correction depends on estimation of the probability of selection, and assumes the analyst has consistent estimates of those probabilities. There are some advantages and disadvantages to our specific method, which uses a conditional logit estimation, requiring a parametric assumption of the distribution of the error term (Type 1 extreme value). This allows us to jointly estimate the catch and choice functions, and simulations suggest this provides more accurate estimation of the fisher’s preference parameters (for example the marginal utility of catch, Chen et al. 2023). However, other more flexible methods are available as well; for example, Dahl (2002) suggests a semiparametric method that groups individuals into “cells” based on their individual characteristics (in our example, perhaps vessel gross tonnage, horsepower, etc.). Then the probability of selection is the fraction of individuals in a cell who moved from *j* to *k*, and is distribution-free. Simulations in Chen et al. (2023) showed significant differences in the estimation of fisher preference parameters; however, estimates of the catch parameters were somewhat similar. A two-stage approach with cell probabilities may also be more computationally efficient.

The assumed distribution of the error term might also be incorrect in our method, although robustness checks examining an assumption of normality suggests good performance with alternative distributions (Chen et al. 2023). Dahl (2002) also notes that the estimation method performs well under a variety of distributional assumptions for the error terms in the “catch” equation. As we note in describing Scenario 2, consistency of the correction function coefficients necessitates sufficient variation in the probabilities used as covariates (Dahl 2002). Our approach also allows us to use multiple continuous covariates in estimating the probabilities, mimicking continuous covariates for the correction functions. Any number of covariates can be included in the fisher’s utility function at the analyst’s discretion, including but not limited to bycatch species, management actions, time in season (to account for depletion for example), herding of vessels, or state dependence. These can be statistically tested by nesting models and performing likelihood ratio tests, for example, or by comparing information criterion. Finally, additional correction functions can be added for movers versus stayers, or additional probabilities (of second-, third-, etc.-best choices). A Wald-type test can then examine if the catch coefficients change as additional probabilities are included.

For each year *y*, the abundance index is calculated as the area-weighted sum of abundance of areas with catch data, where scales proportionally to the size of area *j*. As all our areas are the same size, In addition, we estimate, which is the predicted catch estimated for a vessel characteristic of unity (recall we have a single vessel characteristic distributed ).

|  |  |  |
| --- | --- | --- |
|  |  | (8) |

Constructed estimates of abundance in all years are then used as an index in the estimating model (assessment).

## 2.4 Investigate performance of assessment with estimating model

Finally we quantify the assessment impacts when using a corrected index in an estimating model that matches the structure and parameterization of the operating model (Table 1). This indicates when the use of a properly corrected index could improve stock assessment estimates, while an uncorrected index would bias results. We use simulations of the common assessment software Stock Synthesis through the R package ss3sim, to examine whether and when fishery data corrected for economic endogeneity can be valuable.

We assume quadrennial fishery-independent surveys with biomass index CVs of 0.2 and 500 length composition samples. Length composition samples are collected every ten years starting in year 26, and every four years starting year 60. The fishery indices are annual from the start of the fishery in year 88. While fishing occurs starting in year 62, our “observer coverage” does not start until year 88. For each replicate, the uncorrected, corrected and randomly sampled (survey-like) indices are passed to the estimation model and spawning stock biomass (SSB) are compared to the truth from the operating model. Steps 2.1-2.4 are repeated for 200 replicates of independent process errors and stochastic data simulation.

# Results

## 3.1 Correction model performance

Figure 6 below illustrates mean abundance indices from our stochastic data simulation from scenarios 1 and 2, plotted as differences from the true abundance. Unsurprisingly, when we randomly sample one of the *M* areas on each haul occasion, instead of following the selection criteria in equation (3), the resulting abundance index closely mirrors the true abundance over time. A randomly-sampled index can be thought of mimicking an additional fishery-independent survey.

In contrast, an uncorrected index tends to overestimate abundance when mortality is increasing. This is consistent with observed hyperstable patterns seen in many fisheries (and their associated fishery-dependent abundance indices). Fishers can perform well even when the population is less abundant, due to targeting and private information. Meanwhile, an index corrected for site selection of fishers mirrors the true abundance pattern well in baseline scenarios, and gives similar results to a randomly sampled population.

In our Baseline Scenario 1, the corrected index performs well both on average and in terms of uncertainty, when compared to a randomly-sampled index meant to mimic an additional fishery-independent survey. However, even when fisher coverage is not evenly dispersed in Partial Fisher Coverage Scenario 2, the correction still performs relatively well, however, estimates are more uncertain across replicates.

When the spatial abundance distribution is random in Random Spatial Abundance Scenario 3, again the correction performs relatively well, especially in comparison to the uncorrected index. This is because even though the fisher does not sample all areas due to cost of travel and habitat, missing observations are not systematically biased in any way – fishers are just as likely to miss high-abundance areas as low-abundance areas, when travel is costly enough. However the uncertainty around the estimates is again larger compared to our baseline randomly-sampled index.

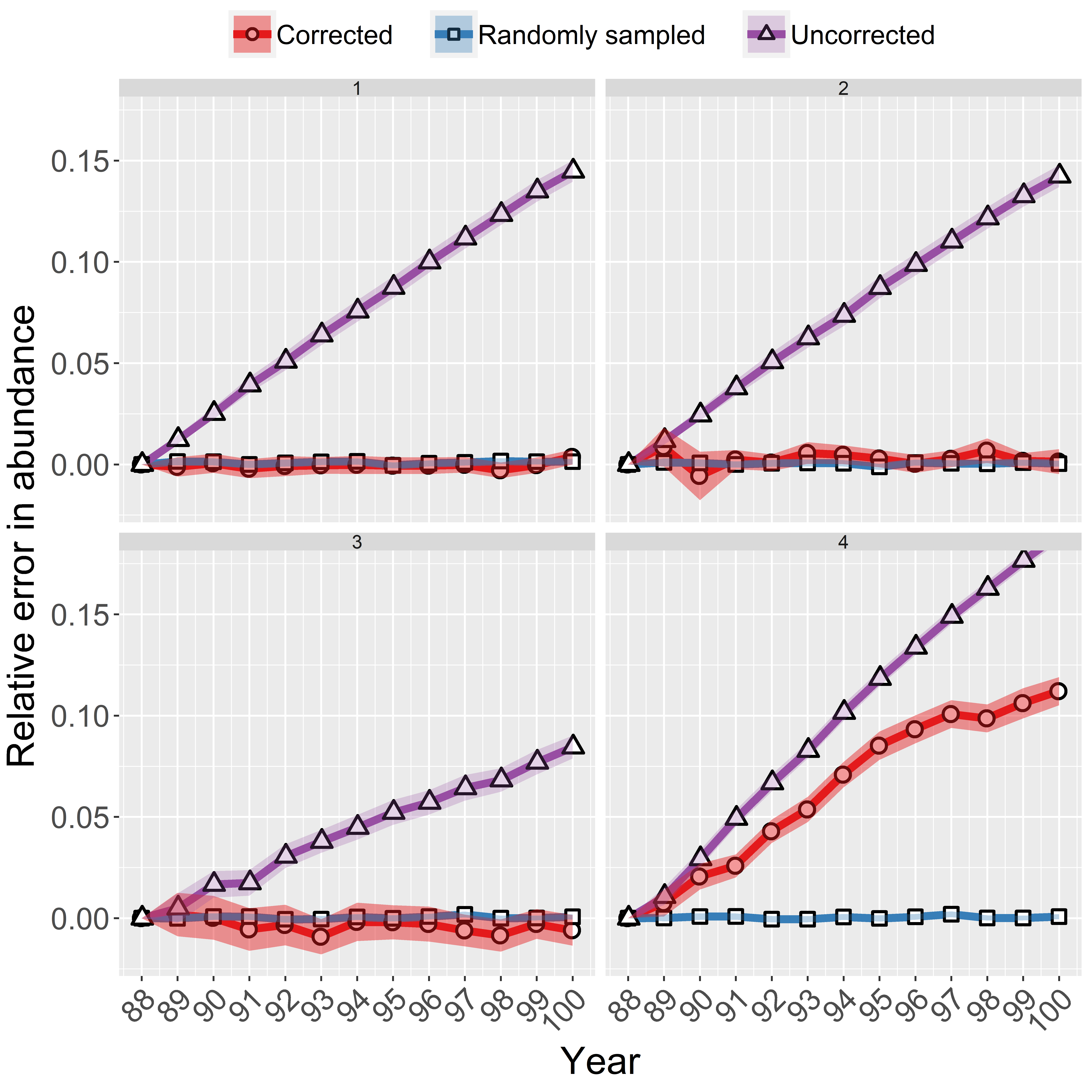


Figure 6: Scenario 1 Baseline (upper left), Scenario 2 Partial Fisher Coverage (upper right), Scenario 3 Random Abundance (lower left), and Scenario 4 Trending Abundance (lower right). Points are mean values in each year, and shaded regions are two standard errors where the standard deviation in a year is divided by the square root of the number of iterations.

Finally, when there is a trend, both the uncorrected and corrected indices perform poorly (Trending Spatial Abundance Scenario 4). Fishers are no longer missing high-abundance areas randomly, but rather systematically fishing on more and more abundant areas due to the eastward spatial trend. Without the ability to sample or extrapolate to areas not visited by the fisher, for example with fishery-independent surveys, the analyst is unable to correct for data they do not observe. However, Figure 6 suggests that as the majority of the biomass moves into the sampled area, the corrected estimates still exhibit less error than the uncorrected. Compared to Scenario 3, this implies non-static abundance over years is not necessarily a problem for the corrected estimates, rather it is the existence of a spatial trend causes more of an issue in estimation.

## 3.2 Assessment performance with corrected indices

In Figure 7 relative error in spawning biomass is plotted under partial fisher coverage (Scenario 2). In Figure 7, for all four panels a fishery-independent survey is conducted every four years, ending 12 years prior to the end of the scenario, with an increasing fishing mortality pattern, and the base case with no additional data is plotted in the upper left panel.

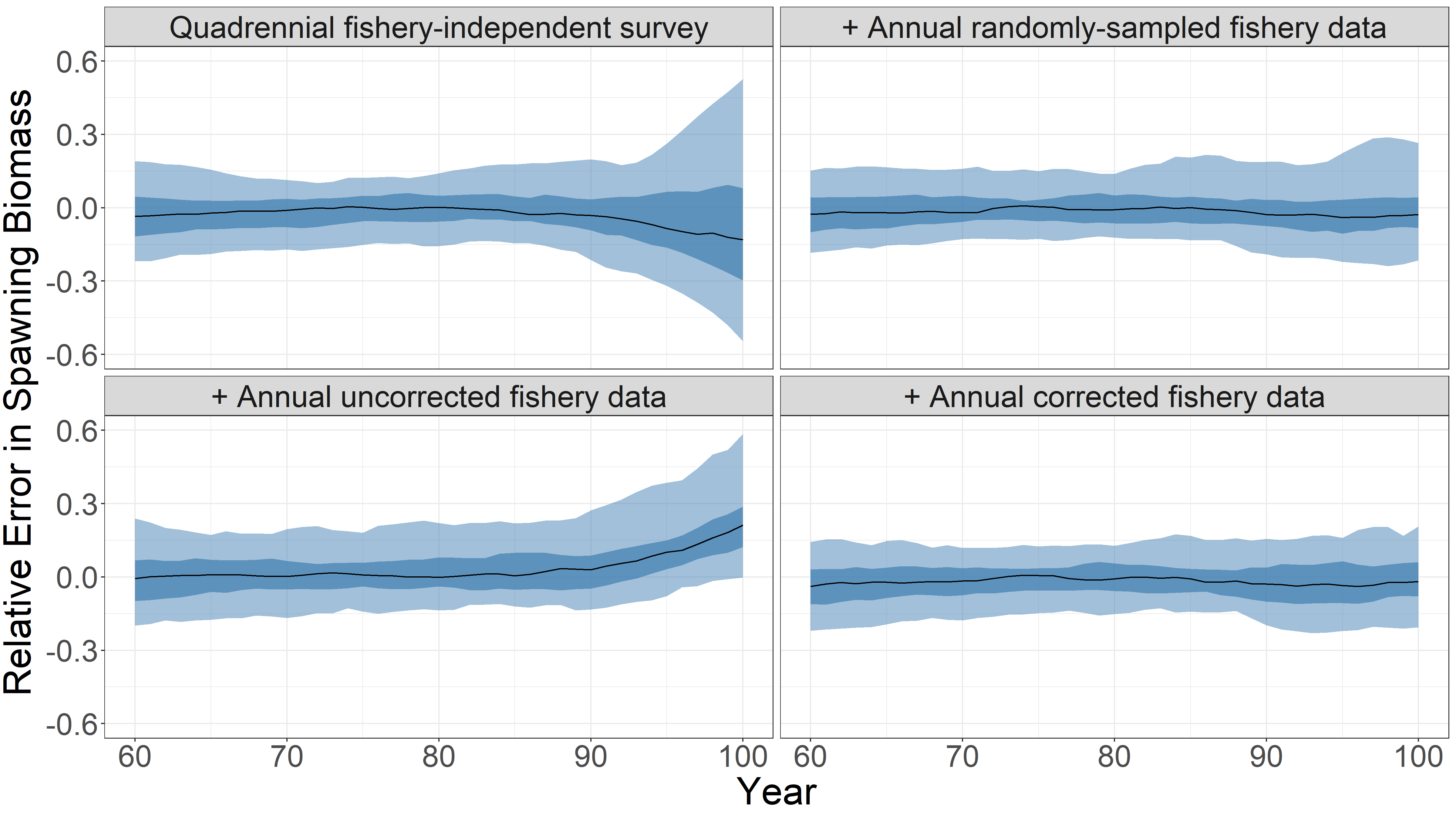


Figure 7: Scenario 2 spawning biomass error relative to the operating model truth under 4 different data scenarios. The first incorporates only a quadrennial fishery-independent survey (upper left), while each of the other data scenarios includes one of three additional indices: the second adds a randomly-sampled fishery-dependent index (upper right), the third an uncorrected fishery-dependent index (lower left), and the last includes a corrected fishery-dependent index (lower right). The shaded regions represent the fiftieth and ninety-fifth percentiles of the replications respectively. Fishery-independent surveys occurred in years 76-88.

The lower left panel also includes a yearly nominal uncorrected abundance index: this results in biased estimates of spawning biomass. If, instead, an additional yearly randomly sampled index is included (upper right), estimates are unbiased and also more accurate, especially at the end of the time series. This is because an additional randomly sampled index is akin to including another fishery-independent survey. However, the lower right panel shows that a similar increase in efficiency can be achieved by included a yearly corrected fishery-dependent abundance index instead.

Figure 7 illustrates that when there are fishery-independent surveys (here, four surveys in years 76-88 providing absolute estimates of abundance), but these surveys end before the end of our time series, the corrected index can add value to the stock assessment. In addition, using an uncorrected index will introduce bias and worsen performance. Meanwhile the corrected index tightens up uncertainty around the estimates, especially near the end of the time series. An assessment that relies only on a fishery-independent survey that is limited in years sampled exhibits considerably more uncertainty. These results are largely intuitive given the abundance indices in Figure 6, but test and validate that there were no unforeseen interactions or issues in the estimating model.

# Discussion

We investigate using a structural economic model to correct for selection bias in fishery-dependent data. Our specific form of statistical selection bias occurs when fishers have some private information about catches that allows them to systematically select where to fish in a way the analyst cannot observe, including self-selecting based on their economic and vessel characteristics (Chen et al. 2023). By explicitly modeling their decision-making behavior, we can estimate the probability a fisher visits an area and use that knowledge to produce unbiased abundance indices from fishery-dependent data for use in stock assessments.

While our proposed methods can perform relatively well, especially with sufficient richness of sample variation and number of observations, they do not represent a panacea for all index standardization issues. We outline one scenario where the correction performs particularly poorly in this paper – a spatial trend in abundance and poor fisher coverage – and we leave additional investigation for future research. Non-static, random abundance over years is not necessarily a problem for our correction method, but the existence of a spatial trend is. Intuitively, this is because even if areas are poorly sampled due to habitat or travel costs, when abundance is random, the analyst is just as likely to miss large positive shocks, as negative ones.

Notably, our results illustrate that the models cannot correct for data they do not observe. This can occur e.g. for trans-boundary stocks, when there is a modeled change in density and distribution related to climate change, or where there are Marine Protected Areas (MPAs) in which densities increase or stay steady, while fishing impacts result in a different relative trend outside the MPAs. In these cases, the proposed methods should be used in conjunction with existing index standardization best practices, as well as techniques to extrapolate over less-sampled areas (Maunder & Punt 2004), such as coupling these methods with spatiotemporal smoothers (Thorson et al. 2020).

This study represents an important step forward to applying these methods to real data, but there are some important caveats and next steps to highlight. Building toward real applications of our method is an important next step, and we suggest future research to expand our simulation framework to more directly investigate the utility to fisheries. Our results are based on a single simulated fish stock (cod like) and exploitation pattern (one-way trip). The four fishing model scenarios (Table 2) tested do not capture the full richness of fisher behavior nor spatial distribution of fish, such as competition across fishers, or depletion over the season or localized areas (Walter et al. 2007). We do not anticipate the biology to greatly influence the method, but alternative fishing histories, such as a severely depleted stock, could interact with the fishing and econometric models. Such a scenario would require including some time-varying variable in the fishing model to account for changes within-season of the desirability of sites.

Due to computational limitations we used a smaller number of areas to illustrate proof of concept and to report asymptotic results (which require thousands of model runs). For practical use, an analyst could increase the number of areas and explore different candidate models and their robustness to alternative model assumptions such as different error distributions or polynomial orders. This would be the case if the analyst only needs to update their fishery-dependent abundance index once a year, running a handful of candidate models.

Targeting is also often a multispecies process, but we only examined a simplified scenario with a single stock. It is possible to extend our method to include multiple species, but that is beyond the scope of this current study. Additional species could be included as additional covariates in the fisher’s utility function, and the marginal utility from catch is an estimated parameter and is agnostic as to sign; bycatch for example might have a negative coefficient. Assumptions as to the joint distribution of catch for the two species may need to be made. If the analyst believes the catch of the other species might also suffer from selection bias, additional correction functions could also be added for that species, although the number of covariates would quickly increase.

In addition, if model of site selection is misspecified, the probabilities that enter our polynomial correction may be misspecified. Broadly, the polynomial only requires consistent estimates of the probabilities of selection, and conditional logit provides consistent estimates under specific conditions (McFadden 1974). Some of these (such as independence from irrelevant alternatives) may be strong. However, there are many alternative methods, such as mixed logit, to estimate the probabilities that relax these assumptions as well (Train 2009). Furthermore, while we suggest a full information approach in this paper, Dahl (2002) also suggests estimating “cell probabilities”. These would be constructed by, for example, grouping vessels with similar characteristics together (e.g. vessel tons or horsepower), creating cells, and calculating the proportion of individuals in each cell that move from one location to another. These are distribution-free, do not assume a parametric framework, and are computationally straightforward to include in a linear regression.

We also assume that the error terms in the fisher’s catch and utility are independent, but they could be correlated. To the extent they are positively correlated, we would expect this to exacerbate the selection bias described, and attenuate the bias if they are not correlated. All our various caveats above emphasize that the process of selecting a final model from competing choices can be time-consuming and subjective, however, due to the variety of modeling assumptions and specifications this can be a necessary and common undertaking. The researcher must check the robustness of different catch and utility functions, as well varying assumptions of the catch and fisher error terms, and decide both the fineness and size of the spatial extent, etc.

While we do find that our correction improves estimates of spawning biomass, the magnitude of this improvement is lower when an existing fishery-independent survey can be used. Thus, we expect our methods to be more useful for situations where fishery-dependent data play a more prominent or exclusive role; an example might be tuna fishery management. Another avenue for future research is to understand how survey frequency impacts the value of our method. An advantage of our approach is that fishing generates data every year across many months, instead of once every few years as is typical for fishery-independent surveys. As a result, fishing data generally covers more months of the year for more years.

The proposed methods therefore are useful for fisheries where fishery-independent surveys do not occur, or to take better advantage of fishery-dependent data that is collected for regulatory purposes. In addition, we note that because a polynomial approximation is relatively straightforward to implement and is statistically testable for whether or not selection bias occurs, we emphasize these methods do not represent an either-or proposition. Rather, they are an additional tool allowing stock assessors to potentially better use the wealth of data generated by observer coverage and the fishing process itself.

# Acknowledgements

The authors thank Kiva Oken, Leif Anderson, Ian Taylor, and Lisa Pfeiffer for valuable comments and assistance. In addition, we thank participants at the North American Association of Fisheries Economists Forum. Financial assistance was provided by FishSET, the Pacific States Marine Fisheries Commission, and NOAA Fisheries Office of Science and Technology.

This research did not receive any additional specific grants from funding agencies in the public, commercial, or not-for-profit sectors.

A repository containing replication code for this paper can be found at: https://github.com/allen-chen-noaa-gov/ss3simfork.

# References

Anderson, S.C., Monnahan, C.C., Johnson, K.F., Ono, K., & Valero, J.L. (2014). ss3sim: An R package for fisheries stock assessment simulation with Stock Synthesis. *PLOS ONE*. 9(4): e92725. http://doi.org/10.1371/journal.pone.0092725.

Andrews, D., & M. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. The Review of Economic Studies, 65(3), 497-517.

Bishop, J., Venables, W. N., & Wang, Y.-G. (2004). Analysing commercial catch and effort data from a Penaeid trawl fishery: A comparison of linear models, mixed models, and generalised estimating equations approaches. *Fisheries Research*, *70*(2-3 SPEC. ISS.), 179–193.

Chen, Y.A., A.C. Haynie, & C.M. Anderson. (2023). Full information selection bias correction for discrete choice models with observation-conditional regressors. *Journal of the Association of Environmental and Resource Economists*, volume 10, number 1, January.

Dahl, G. B. (2002). Mobility and the Return to Education: Testing a Roy Model with Multiple Markets. *Econometrica*, *70*(6), 2367–2420.

Eales, J. & Wilen, J.E. (1986). An Examination of Fishing Location Choice in the Pink Shrimp Fishery. *Marine Resource Economics*, 2(4), 331–351.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Hilborn, R. & Walters, C. J. (1992). Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty. Chapman and Hall, New York. 570 p.

Hoyle, S. D., & Okamoto, H. (2013). Target changes in the tropical WCPO Japanese longline fishery, and their effects on species composition. *Western and Central Pacific Fisheries Commission Scientific Committee Ninth Regular Session*, *August*.

Johnson, K.F., Anderson, S.C., Doering, K., Monnahan, C.C., Stawitz, C.C., & Taylor, I.G. (2019). ss3sim: Fisheries Stock Assessment Simulation Testing with Stock Synthesis. R package version 1.0.3.

Johnson, K.F., Monnahan, C.C., McGilliard, C.R., Vert-pre, K.A., Anderson, S.C., Cunningham, C.J., Hurtado-Ferro, F., Licandeo, R.R., Muradian, M.L., Ono, K., Szuwalski, C.S., Valero, J.L., Whitten, A.R., & Punt, A.E. (2015). Time-varying natural mortality in fisheries stock assessment models: identifying a default approach. *ICES Journal of Marine Science* 72, 137–150. <https://doi.org/10.1093/icesjms/fsu055>.

Maunder, M. & Punt, A.E. (2004). Standardizing catch and effort data: A review of recent approaches. *Fisheries Research*. 70. 141-159. 10.1016/j.fishres.2004.08.002.

McFadden, D. L. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*. Academic Press.

Methot, R.D., & Wetzel, C.R. (2013). Stock synthesis: A biological and statistical framework for fish stock assessment and fishery management. *Fisheries Research*, 142, 86-99.

Ono, K., Licandeo, R., Muradian, M.L., Cunningham, C.J., Anderson, S.C., Hurtado-Ferro, F., Johnson, K.F., McGilliard, C.R., Monnahan, C.C., Szuwalski, C.S., Valero, J.L., Vert-Pre, K.A., Whitten, A.R., & Punt, A.E. (2015). The importance of length and age composition data in statistical age-structured models for marine species. *ICES Journal of Marine Science* 72, 31–43. https://doi.org/10.1093/icesjms/fsu007.

Thorson, J.T., Maunder, M.N., & Punt, A.E. (2020). The development of spatio-temporal models of fishery catch-per-unit-effort data to derive indices of relative abundance. *Fisheries Research*, Volume 230, 105611, https://doi.org/10.1016/j.fishres.2020.105611.

Thorson, J. T., Shelton, A. O., Ward, E.J., & Skaug, H.J. (2015). Geostatistical delta-generalized linear mixed models improve precision for estimated abundance indices for West Coast groundfishes. *ICES Journal of Marine Science*, Volume 72, Issue 5, May/June, 1297–1310, https://doi.org/10.1093/icesjms/fsu243.

Train, Kenneth E. (2009). Discrete choice methods with simulation. Cambridge University Press, New York, (2nd), 1148. https://doi.org/10.1016/S0898-1221(04)90100-9.

Walters, C. (2003). Folly and fantasy in the analysis of spatial catch data. *Canadian Journal of Fisheries and Aquatic Sciences* 60. 1433-1436. 10.1139/f03-152.

Walter, J. F. III, Hoenig, J. M., & Gedamke, T. (2007). Correcting for effective area fished in

fishery-dependent depletion estimates of abundance and capture efficiency. ICES Journal of

Marine Science Volume 64, Issue 9, December 2007, Pages 1760–1771, https://doi.org/10.1093/icesjms/fsm147

1. https://github.com/allen-chen-noaa-gov/ss3simfork [↑](#footnote-ref-2)
2. For all scenarios For scenarios 1 and 2, , and for scenarios 3 and 4, . [↑](#footnote-ref-8)