Informative Data and Uncertainty in Fisheries Stock Assessment

Arni Magnusson

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Ray Hilborn, Chair

André E. Punt

James N. Ianelli

Program Authorized to Offer Degree:
School of Aquatic and Fishery Sciences

University of Washington

Abstract

Informative Data and Uncertainty
in Fisheries Stock Assessment

Arni Magnusson

Chair of the Supervisory Committee:
Professor Ray Hilborn
School of Aquatic and Fishery Sciences

Uncertainty is an integral part of fisheries stock assessment. Successful resource management requires scientific analysis to evaluate the uncertainty about the status of each stock and related quantities of interest. A failure to incorporate uncertainty into management advice increases the risk of suboptimal yields and can lead to a fishery collapse. In practice, it is not always clear which features of stock assessment data make them informative or uninformative, and it is also unclear how well different statistical methods are likely to perform when evaluating uncertainty.

This study uses simulation analysis to measure the performance of alternative methods, based on a large number of simulated datasets where the underlying true values are known. The methods are then applied to data from an actual fishery, and the overall inference takes into account the performance of the methods in the simulations.

The results show that the historical levels of stock size and harvest rate greatly affect how informative the data are about the current stock status. The key parameters natural mortality M and stock-recruitment steepness h pose challenges when it comes to statistical estimation, and long-term management advice is likely to depend strongly on the estimated or assumed values of M and h. The most informative fishing history is one where the data include years of high and low stock size, which is informative about h, as well as high and

low harvest rates, which is informative about M. The results also indicate that confidence intervals describing the uncertainty about the stock status and other quantities of interest are likely to be too narrow in general. Benchmark analysis indicates that the delta method, Markov chain Monte Carlo (MCMC), and profile likelihood approaches are likely to perform better than the bootstrap for quantifying uncertainty. A bias correction algorithm for the bootstrap improved its performance, but not enough to match the performance of the other methods. Additional approaches to evaluate the estimation uncertainty include retrospective analysis and bivariate confidence regions for the current stock status. The use of harvest control rules to incorporate uncertainty into management advice is also discussed.

The main value of this study is to present a comprehensive overview and evaluation of methods to analyze uncertainty. The study concludes with a checklist of recommendations for confronting uncertainty in stock assessment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It has truly been a privilege to have Ray Hilborn and André Punt as my main supervisors. After I completed the stock assessment coursework at UW, Ray invited me to work closely with him for three winters in New Zealand (summertime in the southern hemisphere) doing stock assessment work for the Seafood Industry Council. During these trips, we assessed a variety of stocks and presented management advice to the Ministry of Fisheries. For the eager apprentice, every stock presented a new set of challenges, puzzles to solve, and I cannot think of a better way to learn about all aspects of the stock assessment process, with the grand master at my side. Thank you, Ray, for offering me this opportunity.

At this point I had decided that my primary interest was in the statistical aspects of stock assessment. So how incredibly fortunate that André Punt, known as the world's foremost expert in that field, largely took over my supervision at that stage. He made sure my research ideas were implemented in the best way possible, and challenged me with statistical techniques that were at the very edge of my capability. During our meetings, sometimes late in the evening, he would provide me with just enough hints so I would continue learning and pushing myself. André also hired me as a co-instructor to help him put together and teach the first incarnation of Fish 507 on statistical methods in fisheries science, an experience that motivated me greatly. When it comes to the all-important process of converting a first draft into a high-quality manuscript, his dedication and abilities are just beyond belief. I'm grateful for all the knowledge, encouragement, and time that André has given me.

My supervisory committee members Jim Ianelli and John Skalski have had a distinct influence on my budding scientific career. Jim introduced me to AD Model Builder, a statistical software platform which completely captured my interest from then on. Meanwhile,

John handed me an elegant weapon for a more civilized age in statistics, the delta method. ADMB and the delta method are recurring themes in this dissertation and have shaped my research interests in general. I'm thankful to Jim and John for broadening my horizon.

I also want to thank all of the committee members, including the Graduate School representative, Christine Ingebritsen, for making it possible for me to return to defend my Ph.D. dissertation after many years away from school. Amy Fox, the graduate advisor at the School of Aquatic and Fishery Sciences, and Machelle Allman at the International Student Services offered their generous support, which I greatly appreciate.

With my interest and frequent usage of AD Model Builder, it was with great pride and enthusiasm that I accepted an offer to join the ADMB Core Team, as the software project was going open source. I was lucky to join at a point early enough to coauthor the ADMB paper, and I'm grateful to Dave Fournier, Hans Skaug, Johnoel Ancheta, Jim Ianelli, Mark Maunder, Anders Nielsen, and John Sibert for that opportunity. Working side-by-side with these gurus has elevated my skills in statistical computing and directly benefited the research presented in this dissertation. They are enjoyable companions and great mentors.

It makes me happy every time that I meet the good friends from my student years at UW, who are now successful scientists in different places around the world. Mi amigos Billy Ernst and Juan Valero soon convinced me that I was a Latino at heart, and Carolina Minte-Vera, Nicolas Gutierrez, and Julian Burgos confirmed this observation. Ana Parma and the late (and legendary) Lobo Orensanz blew me away from the first encounter, with their charisma, positive influence, and wise outlook. They will always be my role models. Allan Hicks, Trevor Branch, Vivian Haist, Brandon Chasco, and Juan (again) made the work trips to New Zealand unforgettable, thanks to magnificent scenic hikes, Maori parties under the starry sky, and culinary feasts organized by Paul Starr. Ulrike Hilborn's visit to New Zealand brought in an air of elegance, when she invited me to the symphony and gave me good advice on how to enjoy life to the fullest.

When I first arrived in Seattle, I was pleased to discover that I had relatives there, the Thorstensons and Schonbergs. They welcomed their newfound nth-generation cousin with open arms, picking him up at the doorstep of his house every Thanksgiving and Christmas. That doorstep was at the Scandinavian House in the University District, inhabited by international students famous for hosting large dinners and even larger parties. These fortunate living conditions meant that I was always feeling upbeat, ready to take on the scientific tasks of the day.

None of the above would have happened if it wasn't for Gunnar Stefansson and Jakob Jakobsson, who first introduced me to fisheries science and stock assessment at the University of Iceland. Likewise, I'm indebted to Bjorn Steinarsson at the Marine Research Institute for allowing me to combine my work on the Icelandic saithe with completing my Ph.D. studies.

Last, but not least, I express my deepest appreciation and gratitude to my family, including my in-laws. Their immeasurable help and encouragement has now brought us to the top. Let us enjoy the view.

DEDICATION

to

Steffí and Emma

for their patience and love

INTRODUCTION

Fisheries management relies on stock assessment models to provide estimates of population abundance, and to shed light on the underlying dynamics of the resources being managed. It is necessary to quantify and understand the uncertainty about model parameters and reference points to evaluate the consequences of alternative management actions.

The uncertainty about estimated quantities reflects the information contained in the available data, but also depends on model choice and implicit assumptions that are made when the assessment is conducted. Informative data in fisheries stock assessment are those that lead to accurate estimates of abundance and reference points. In practice, the estimation accuracy is unknown and it is often unclear which features of the data make them informative or uninformative. Despite theoretical and practical advances in the field of stock assessment, our ability to answer some key questions remains limited. Chapter 1 of this dissertation focuses on one such question: What kinds of data are particularly informative in stock assessments, and how is this influenced by model assumptions? This simulation study compares four fishing histories, which represent different contrast in the stock size and harvest rates, and the effect of the fishing history on the estimation of three key parameters: natural mortality rate, stock-recruitment steepness, and asymptotic vs. dome-shaped selectivity.

A common approach to quantify uncertainty is to calculate confidence intervals, and a variety of statistical methods exist for this purpose. However, different methods give different intervals, and fisheries scientists are likely to choose the statistical method they are most familiar with, or one that has become traditional for a particular stock. It is not obvious which method to recommend in stock assessment, where model complexity and non-linearity are likely to degrade the performance of standard methods. Ideally, the method should generate

intervals that are neither too narrow nor too wide, in order to cover the true value of estimated quantities with a probability matching the claimed confidence level. Chapter 2 presents a benchmark analysis that compares the performance of the delta method, bootstrap, and Markov chain Monte Carlo (MCMC) to evaluate the uncertainty about stock status and reference points from simulated datasets.

The objective of Chapter 3 is to provide a broader and more complete overview and demonstration of techniques to confront uncertainty in stock assessment. Using Icelandic saithe data as a case study, it revisits the approaches and findings from the first two chapters to examine four main issues: (1) the overall fishing history and whether it is likely to be informative about the stock status and key parameters, (2) the effects of different assumptions about the natural mortality rate, stock-recruitment steepness, and the shape of the selectivity curve for the oldest fish, (3) the amount of information contained in the survey data about the stock status, and (4) whether the delta method, profile likelihood, bootstrap, and MCMC lead to similar conclusions. Additional methods used to evaluate the estimation uncertainty in this chapter include retrospective analysis and bivariate confidence regions for the current stock status. This chapter also highlights the use of harvest control rules to incorporate uncertainty into management advice.

Finally, Chapter 4 presents a brief overview of statistical software that was developed for the analysis in previous chapters.

Chapter 1

WHAT MAKES FISHERIES DATA INFORMATIVE?

Abstract

Informative data in fisheries stock assessment are those that lead to accurate estimates of abundance and reference points. In practice, the accuracy of estimated abundance is unknown and it is often unclear which features of the data make them informative or uninformative. Neither is it obvious which model assumptions will improve estimation performance, given a particular data set. In this simulation study, 10 hypotheses are addressed using multiple scenarios, estimation models, and reference points. The simulated data scenarios all share the same biological and fleet characteristics, but vary in terms of the fishing history. The estimation models are based on a common statistical catch-at-age framework, but estimate different parameters and have different parts of the data available to them. Among the findings is that a 'one-way trip' scenario, where harvest rate gradually increases while abundance decreases, proved no less informative than a contrasted catch history. Models that excluded either abundance index or catch at age performed surprisingly well, compared to models that included both data types. Natural mortality rate, M, was estimated with some reliability when age-composition data were available from before major catches were removed. Stock-recruitment steepness, h, was estimated with some reliability when abundance-index or age-composition data were available from years of very low abundance. Understanding what makes fisheries data informative or uninformative enables scientists to identify fisheries for which stock assessment models are likely to be biased or imprecise. Managers can also benefit from guidelines on how to distribute funding and manpower among different data collection programmes to gather the most information.

1.1 Introduction

1.1.1 Stock assessment and informative data

Fisheries management relies on stock assessment models to provide estimates of population abundance, and to shed light on the underlying dynamics of the resources being managed. It is necessary to quantify and understand the uncertainty about model parameters and reference points to evaluate the consequences of alternative management actions. The uncertainty about estimated quantities reflects the information contained in the available data, but also depends on the choice of model and implicit assumptions that are made when the assessment is conducted. Despite theoretical and practical advances in the field of stock assessment, our ability to answer some key questions remains limited. This study focuses on one such question: What kinds of data are particularly informative in stock assessments, and how is this influenced by model assumptions?

Understanding what makes fisheries data informative or uninformative has obvious value for fisheries management, enabling us to identify fisheries for which stock assessment models are likely to be biased or imprecise. Managers can also benefit from guidelines on how to distribute funding and manpower among different data collection programmes to gather the most information. Moreover, adaptive management decisions can be taken today to make future data as informative as possible (Ludwig and Hilborn 1983, Walters 1986, Walters 2007).

Shepherd (1984) ranked types of fisheries data in terms of potential information provided by each type of data in isolation. Annual landings and age-specific abundance indices were ranked the highest, for example, while age-composition data alone was assigned a low score. Such statements are of course highly generalized, but nevertheless provide useful guidelines for planning data collection programmes. Shepherd (1984) also points out how different data types complement each other: landings provide information about the absolute scale of the fishery, age-composition data about the relative cohort size, and abundance-index data about the relative changes in abundance over time. Changes in growth or maturity

can provide some information about changes in population density, confounded with other ecological and evolutionary factors (Rose et al. 2001). Less commonly used data types that provide information about stock status include tag recoveries and egg/larval surveys. In a Bayesian context, any prior information about estimated or derived parameters can also be seen as a type of data source (Gelman et al. 2004), where information from previous studies is expressed in the form of a probability distribution for estimated or derived parameters. Many non-Bayesian estimation methods assume that specific parameters are known without error; the effect of such assumptions is similar to that of a highly informative Bayesian prior.

The general rule in statistical inference is that more data leads to less uncertainty. But other features of the data also play a role, e.g. the range of observed values and temporal patterns in time-series data. This is easy to show analytically for simple stock assessment techniques, such as depletion models and catch-curve analysis, as outlined below. When more complex models are used, it becomes less concrete what is meant when stock assessment modellers discuss the 'informative' data in a particular assessment, or perhaps more often, 'uninformative' data.

Before taking a closer look at what kinds of data are informative for a given model, it is helpful to begin with an overview of some commonly used models.

1.1.2 Models and assumptions in stock assessment

A variety of stock assessment models have been developed, as reviewed by Megrey (1989), Hilborn and Walters (1992), Quinn and Deriso (1999), Quinn (2003) and Smith and Addison (2003). This variety of models reflects the diversity of fisheries to which stock assessment techniques need to be applied, the data available for assessment purposes, and what is known or assumed about the fishery dynamics and stocks. There has been a move away from simple and restrictive assumptions (Schaefer 1954, Chapman and Robson 1960, Gulland 1965) towards more flexible models that incorporate all of the available data in a likelihood-based statistical framework.

A depletion model (Leslie and Davis 1939) can be used to estimate absolute abundance

from a time series of landings and an index of relative abundance when a stock is fished down. Assuming a closed population where the impact of fishing mortality is much greater than those of recruitment, growth, and natural mortality, the biomass at the start of time step $t+1$ equals the biomass at the start of time step t less the catch during time step t, that is $B_{t+1} = B_t - Y_t$, and the abundance index is proportional to the stock biomass, $I_t = qB_t$. The catchability coefficient, q, is assumed to be constant, although empirical data suggest that may not always be a justifiable assumption (Ricker 1958). The Schaefer (1954) biomass-dynamic model adds two parameters representing the maximum growth rate and carrying capacity, $B_{t+1} = B_t + rB_t(1 - B_t/k) - Y_t$. When $r = 0$, it is identical to the depletion model. A slightly different approach is taken in the Kimura and Tagart (1982) stock-reduction model, where two parameters represent the natural mortality rate and recruitment, $B_{t+1} = B_t e^{-(F_t+M)} + R$. The fishing mortality rate F_t is evaluated from the landings, $Y_t = B_t(1-e^{-F_t-M})F_t/(F_t+M)$. As the above models do not distinguish between age groups, they can be formulated either in terms of numbers or biomass.

In catch-curve analysis (Chapman and Robson 1960), the total mortality rate of fully recruited fish in a given year can be estimated using the catch-at-age composition, assuming that recruitment variability is inconsequential. Catch curves can also be applied to individual cohorts (Hilborn and Walters 1992), relaxing the assumption that recruitment is the same across cohorts, $N_{t+1,a+1} = N_{t,a}e^{-(F_t+M)}$. In practice, M is often assumed to be known, and F can in turn be used to estimate absolute abundance if the annual landings are known. Related models include virtual population analysis (Gulland 1965), cohort analysis (Pope 1972), adaptive framework (Gavaris 1988) and extended survivors analysis (Shepherd 1999). The assumption of a known constant M is frequently challenged (Cotter et al. 2004), but restrictive assumptions about M and recruitment are often necessary to evaluate the consequences of alternative catch levels (Punt and Hilborn 1997). Even in fisheries where large quantities of data have been collected for decades, an age-structured assessment model can fit the data equally well when M is fixed at a very low value or a high value (Gavaris and Ianelli 2002).

The forward-projecting statistical catch-at-age model (Fournier and Archibald 1982, Deriso et al. 1985, Methot 1989) can be described conceptually as a stock-reduction model with variable recruitment, combined with catch-curve analysis of multiple cohorts. The basic framework can be tailored fairly easily to the specifics of the fishery to be modelled. The data types most commonly included in statistical catch-at-age analysis are: annual landed catch, catch at age and an index of relative abundance. The landed catch is often assumed to be measured without error, and model parameters are estimated by minimizing the difference between model predictions and the observed catch at age and abundance index. A statistical catch-at-age model can be fitted without any age data (Hilborn 1990) or without an index of abundance. However, Deriso et al. (1985) concluded that all three data types are required to estimate abundance and reference points reliably. Hilborn et al. (2003) describe how statistical catch-at-age models can incorporate sex-specific data from multiple fisheries with ageing error, catch-at-length data, and allow certain parameters to vary over time.

Parameters that are almost always estimated when fitting a statistical catch-at-age model include the age-structure of the population in the first year considered in the model, the selectivity curve for each fishery, the catchability coefficient for each abundance index and the annual recruitment. There remain, however, many decisions to fully specify a statistical catch-at-age model. For example, whether to estimate or fix the natural mortality rate M, how to model the relationship between spawning stock size and recruitment, whether to parametrize selectivity using an asymptotic or dome-shaped curve, and how to choose which parameters vary over time (Patterson et al. 2001, Gavaris and Ianelli 2002). M can be accurately estimated if the data include the catch age composition from a nearly unfished population, or if fishing effort is kept very low for some years (Beverton and Holt 1957), while the shape of the stock-recruitment curve can be estimated only if there is considerable contrast in stock size (Ricker 1958). Analyzing a model that estimates M and right-hand selectivity, i.e. a shape parameter determining the selectivity of older age classes, Thompson (1994) noted that those parameters were confounded and recommended fixing either M or the selectivity shape parameter at an assumed value.

The simpler models (depletion, catch curves) can be emulated fairly adequately using statistical catch-at-age models, by fixing parameters, selecting specific functional forms for biological relationships, or excluding likelihood components from the objective function. The main difference between the depletion, biomass-dynamic, stock-reduction and delay-difference models is how recruitment, somatic growth and natural mortalities are handled. Schnute (1985) showed how the Schaefer (1954) biomass-dynamic model, the Deriso (1980) delay-difference model and Kimura and Tagart (1982) stock-reduction model are special cases of a generalized catch-effort model. Xiao (2000) showed further that the above models, along with the Leslie and Davis (1939) depletion model, Gulland's (1965) virtual population analysis and Fournier and Archibald's (1982) statistical catch-at-age models are all special cases of a generalized age-structured model. In light of their flexibility, superior performance (NRC 1998, Punt et al. 2002), and increasing usage, statistical catch-at-age models are used in this study to address the questions of interest.

1.1.3 Simulation studies and informative data scenarios

The term data scenario is used here to denote temporal patterns found in the data, regardless of the amount and types of data. These temporal patterns are impacted by how the fishery has been conducted historically, e.g. whether fishing effort and landings have been increasing or decreasing, held relatively steady, or perhaps the stock might be rebuilding after heavy depletion. One of the questions confronted in this study is which data scenarios are more informative than others.

A data scenario is informative when it enables a given model to estimate the status of a fishery with greater accuracy than most other data scenarios would. In the real world, a data scenario can be said to be informative if it resembles a scenario that has been shown to be informative, either analytically or in a simulation study. Analytical demonstration is only viable for the simplest of models, such as linear regression, but for more complex models, simulations are used to evaluate estimation accuracy. Simulation studies use an 'operating' model to generate artificial data similar to those used in stock assessment, except the true

population parameters are known.

The depletion model can be expressed as a linear regression with the abundance index as the response variable and accumulated catch as the predictor, $\hat{I}_t = qB_{\text{init}} - q\sum_{i \leq t} Y_i$. As with any simple linear regression model, the uncertainty about the slope and intercept depends on (i) how closely the data points are aligned in a straight line, as residuals will be smaller when model assumptions are not violated substantially and when measurements are reasonably accurate, (ii) the range of values on the x-axis and (iii) the number of data points. The biomass before catches were removed, B_{init}, corresponds to the x-intercept. This intercept can be predicted more accurately when the y-values, relative abundance, are observed both at high and quite low values. Ricker (1958) noted that intense fishing effort that reduces the abundance considerably leads to informative data for the depletion model, and Pope (1972) found the same to be true for cohort analysis. Many fisheries have undergone a period of rapid removals and can therefore be expected to yield informative data, if scientific data were being collected at the time.

Hilborn (1979) demonstrated why and how certain data scenarios are informative, using a simplified Schaefer biomass-dynamic model that has a closed-form solution. He concluded that contrast is needed in both abundance and harvest rate to obtain unbiased and precise parameter estimates. Specifically, Hilborn (1979) identified the most informative data scenario as one that includes a period of quite heavy exploitation, followed by a period where the stock is allowed to rebuild to an intermediate level, after which the exploitation rate increases again.

The parameter of main interest in catch-curve analysis is the fishing mortality rate in each year, F_t, frequently used in fisheries management. This parameter is confounded with the natural mortality rate, as cohorts decline at an exponential rate $Z_t = F_t + M$. If little or no fishing has taken place in previous years, Z corresponds to the rate of natural mortality M, which is otherwise a very difficult parameter to estimate. More generally, catch-at-age data that contain years with high and low fishing effort are informative to bound possible values of M, and therefore F_t (Beverton and Holt 1957). Variation in fishing effort has also been

found to be informative for more complex age-structured models to separate natural and fishing mortalities (Hilborn and Walters 1992). Of course, uncertainty about the estimated parameters will also decrease if large numbers of fish are sampled at random and measured with negligible ageing error, and if M varies only slightly between years, without a consistent increasing or decreasing trend (Beverton and Holt 1957).

Several simulation studies have explored the behaviour of statistical catch-at-age models. Bence et al. (1993) found that current abundance is estimated more reliably when harvest rate has been high, and when the true survey selectivity curve is asymptotic rather than dome-shaped. The study of Sampson and Yin (1998), later updated by Yin and Sampson (2004), showed how low natural mortality M, high recruitment variability and small changes in the harvest rate all lead to unreliable estimates. They also concluded that for the U.S. West Coast groundfish fishery, it would be more cost-effective to gain information by increasing sampling for age composition than by improving the precision of the survey on which abundance indices are based, at least from a single-species perspective. Ianelli (2002) found that reference points are overestimated when the true steepness h of the stock-recruitment curve is low, and underestimated when the true value of h is high. In their simulation study, Punt et al. (2002) showed how depletion level, defined as current abundance compared with average virgin abundance, was estimated more reliably than other reference points. They also found that the statistical catch-at-age model performed substantially worse when age-composition data were not available.

1.1.4 This study

The goal of this study is to improve our understanding of how uncertainty about the status of a fishery resource depends on data, models and assumptions. An 'informative' data scenario is one that enables a given model to estimate the status of a fishery with greater accuracy than most other data scenarios would. The hypotheses that will be addressed are:

H_1 Fisheries data are most informative when they span a period where the population was fished down to a low level.

H_2 Fisheries data are most informative when they span a period where the population was fished down to a low level and then allowed to rebuild for some time.

H_3 The level of stock depletion is estimated more reliably than other reference points.

H_4 A data set that includes both an index of relative abundance and catch-at-age data is much more informative than a data set that includes only one of these two types of data.

H_5 Not knowing M, h and right-hand selectivity leads to inaccurate estimates of stock abundance and reference points.

H_6 Models estimating M perform about as well as models estimating h.

H_7 M can be estimated reliably if age-composition data are available from when the population was unfished.

H_8 M can be estimated reliably from the rate of population increase if the stock is allowed to rebuild from a low level.

H_9 h can be estimated reliably from catch-at-age data and an index of relative abundance when the data cover a period in which abundance varies substantially.

H_{10} Right-hand selectivity can only be estimated reliably when M is known.

1.2 Methods

First, we define four fishing history scenarios and generate stochastic data sets using an 'operating model,' based on an age-structured population dynamics model. The performance of a suite of estimation models is then evaluated, with respect to how well they estimate the values of six reference points. The simulation procedure, outlined in Figure 1.1, is repeated for

each scenario, random seed and estimation model. A scenario consists of chosen parameter values and a harvest rate schedule, described in more detail below. The operating model first applies stochastic recruitment and outputs the resulting reference point values. It then applies random observation noise and outputs the assessment data that are used as input for the estimation models. Finally, the estimated reference points are derived from the parameter estimates, and compared with the 'true' reference points that were not subject to observation noise.

1.2.1 Scenarios

Four fishing history scenarios are simulated in the analysis: (A) *one-way* trip where harvest rate is gradually increased while the abundance decreases, (B) *no change* where abundance is steady at a constant and somewhat low harvest rate, (C) *good contrast* where the stock is fished down to less than half its initial size and then allowed to rebuild and (D) *rebuild only* where the stock begins at a very low abundance and is allowed to rebuild under low fishing pressure. The fishing history scenarios are designed specifically to address hypotheses 1–2 and 7–9, in terms of harvest rate and the expected value of the abundance index (Figure 1.2). Time trajectories offer a more traditional view of the same data (Figure 1.3).

1.2.2 Operating model

The biological component

The operating model is a statistical catch-at-age model (Fournier and Archibald 1982) with biological characteristics (Table 1.1) and parameter values (Table 1.2) based on Atlantic cod (*Gadus morhua*, Gadidae). It follows the parametrization of the Coleraine statistical catch-at-age software (Hilborn et al. 2003), which is used to implement the estimation models.

The population dynamics are governed by the equation

$$N_{t+1,a+1} = N_{t,a}e^{-M}(1 - {}_{C}S_a u_t) \tag{1.1}$$

where $N_{t,a}$ is population size at time t and age a, M is the rate of natural mortality, ${}_{C}S$ is the selectivity of the commercial fishery and u is harvest rate. The oldest age group, age A, is treated as a plus group:

$$N_{t+1,A} \;=\; N_{t,A-1}\, e^{-M}(1-{}_{C}S_{A-1}u_t) \;+\; N_{t,A}\, e^{-M}(1-{}_{C}S_{A}u_t) \tag{1.2}$$

Selectivity is an asymmetric normal curve determined by three shape parameters,

$$S_a \;=\; \begin{cases} \exp\!\left(\dfrac{-(a-S_{\text{full}})^2}{\exp(S_{\text{left}})}\right), & a \le S_{\text{full}} \\[2ex] \exp\!\left(\dfrac{-(a-S_{\text{full}})^2}{\exp(S_{\text{right}})}\right), & a > S_{\text{full}} \end{cases} \tag{1.3}$$

where S_{full} is the age at full selectivity, S_{left} describes the left-hand slope and S_{right} the right hand slope of the curve. The survey selectivity curve has a high ${}_{S}S_{\text{right}} = 15$ (Table 1.2) so the oldest fish are fully selected, but the commercial selectivity has an intermediate ${}_{C}S_{\text{right}} = 6$, resulting in a slightly dome-shaped curve (Figure 1.4). Harvest rate is defined as the fraction removed from the vulnerable biomass in the middle of the fishing year, $u_t = Y_t \,/\, \sum_a ({}_{C}S_a N_{t,a} w_a) e^{-M/2}$, where Y is catch and w is body weight.

The population size at the start of the first year is

$$\begin{aligned} N_{1,1} &= R_0\, R_{\text{init}} \times \exp({}_{R}\varepsilon_{1,1} - \sigma_R^2/2) \\ N_{1,a} &= R_0\, R_{\text{init}}\, e^{-(a-1)M} \prod_{i=1}^{a-1} (1-{}_{C}S_i u_{\text{init}}) \times \exp({}_{R}\varepsilon_{1,a} - \sigma_R^2/2) \\ N_{1,A} &= R_0\, R_{\text{init}}\, e^{-(A-1)M} \prod_{i=1}^{A-1} (1-{}_{C}S_i u_{\text{init}}) \,/\, [1-e^{-M}(1-{}_{C}S_A u_{\text{init}})] \times R_{\text{plus}} \end{aligned} \tag{1.4}$$

for 1-year-olds, intermediate ages, and the plus group. R_0 is average virgin recruitment, R_{init} scales the initial population size across all ages, and u_{init} is the initial harvest rate.

The ${}_R\varepsilon$ elements are random recruitment deviates generated from the normal distribution, ${}_R\varepsilon \sim \mathrm{N}(0, \sigma_R^2)$, where σ_R is recruitment variability. The R_{plus} term scales the initial plus group and is not drawn from the same distribution as the ${}_R\varepsilon$ recruitment deviates for the younger ages. Instead, a large number of initial ages are generated, up to 100 years old, and ages 10 and over are aggregated in a plus group.

Recruitment is stochastic around a Beverton-Holt stock-recruitment function, reparametrized according to Francis (1992),

$$N_{t+1,1} \;=\; \frac{4hR_0(B_t/B_0)}{1-h+(5h-1)(B_t/B_0)} \;\times\; \exp({}_R\varepsilon_{t+1,1} - \sigma_R^2/2) \tag{1.5}$$

where $B_t = \sum_a N_{t,a}\, \Phi_a w_a$ is spawning biomass,

$$B_0 \;=\; \sum_{a=1}^{A-1} R_0\, e^{-(a-1)M}\, \Phi_a w_a \;+\; R_0\, e^{-(A-1)M}\, \Phi_A w_A/(1-e^{-M}) \tag{1.6}$$

is average virgin spawning biomass, h is steepness of the stock-recruitment curve, and Φ is proportion mature.

Generating the simulated data sets

One hundred data sets are generated for each fishing history scenario. These data sets vary in terms of landings, survey abundance index and commercial catch-at-age. The harvest rate is always the same in each scenario, but the resulting landings change as population size changes with stochastic recruitment. There are 10 age classes and 20 years of data, nominally referred to as 1985–2004. The landings are assumed to be known exactly, but the catch at age and abundance index are subject to random observation error. When stochastic recruitment and observation noise is added to the original templates from Figure 1.2, the observed abundance index shows random fluctuations, but the overall fishing history is still recognizable (Figure 1.5).

Even though the harvest rates in Table 1.2 are followed precisely, the resulting landings

vary among the data sets because of stochastic recruitment. The level of recruitment variability ($\sigma_R = 0.6$), observation noise for the abundance index ($\sigma_I = 0.2$) and observation noise for the commercial catch at age ($n = 50$) are similar to those used in recent assessments of the Icelandic cod stock (ICES 2003).

The survey abundance index is proportional to the biomass vulnerable to the survey in the middle of the fishing year,

$$I_t = q \sum_a {}_S S_a N_{t,a} w_a \, e^{-M/2} \times \exp({}_I\varepsilon_t) \tag{1.7}$$

where I is the observed abundance index, q is the catchability coefficient, ${}_S S$ is survey selectivity, and ${}_I\varepsilon_t \sim \mathrm{N}(0, \sigma_I^2)$ is random observation noise. The commercial catch-at-age data are provided to the assessment model in the form of proportions at age. These proportions are generated assuming that the sampling is multinomial,

$$P_{t,a} \sim \mathrm{Multinom}\left(n, \frac{{}_C S_a N_{t,a}}{\sum_a {}_C S_a N_{t,a}}\right) \Big/ n \tag{1.8}$$

where P is the observed catch at age and n is the sample size used to generate observation noise.

Survey catch-at-age data are not used in this study, to keep the analysis and interpretation as simple as possible. The survey abundance index and the commercial catch at age are independent sources of information, one about changes in relative abundance, the other about relative cohort sizes and mortality rates. Data are assumed to be available for each year and the landings are output without observation error.

1.2.3 Estimation models

Thirteen estimation models are fitted to the simulated data. They have the same parametrization as the operating model (Equations 1.1–1.8) and are implemented with the Coleraine statistical catch-at-age software (Hilborn et al. 2003). The models differ in terms of which

data types are included in the objective function and which parameters are estimated (Table 1.3). The models are designed specifically to address hypotheses 4–10.

The 13 models consist of three 'families,' indicated by the first digit of the abbreviation used to identify the model: family 1 uses only landings and abundance index, family 2 uses only landings and catch at age, family 3 uses all three data types. Thus, models from family 1 are akin to biomass-dynamic models (with R_0 scaling the absolute size of the population instead of K, and h and M determining the intrinsic growth rate instead of r), models from family 2 resemble catch-curve analysis based on multiple cohorts, and those from family 3 are several variants of statistical catch-at-age analysis.

Although it is possible to examine the implications of estimating every combination of parameters, the focus of this study is on three key parameters: the steepness of the stock-recruitment relationship (h), the natural mortality rate (M) and the right-hand selectivity shape parameter (${}_C S_{\text{right}}$) for the commercial fishery. Models that estimate these parameters have 'h,' 'M' or 'r' in their abbreviations. When parameters are not estimated, they are fixed at the true value, as is done for the survey selectivity parameters.

The Coleraine software requires that all estimated parameters be bounded. Wide bounds (Table 1.4) are assigned to all parameters so as not to impose any major constraints on the values for the parameters.

The objective function for the estimation models is the sum of three components. The first two relate to data included in the analysis and the last is a penalty on recruitment deviations from the stock-recruitment relationship:

$$f \;=\; -\log L_I - \log L_C + \text{Pen} \tag{1.9}$$

The abundance-index likelihood component is lognormal,

$$-\log L_I \;=\; \sum_t \frac{(\log I_t - \log \hat{I}_t)^2}{2\sigma_I^2} \tag{1.10}$$

where I and $\hat{I}$ are the observed and model-predicted abundance indices. The robust normal

likelihood for proportions (Fournier et al. 1990) is assumed for the catch-at-age data,

$$-\log L_C = -\sum_t \sum_a \log\left[\exp\left(\frac{-(P_{t,a}-\hat{P}_{t,a})^2}{2\big[P_{t,a}(1-P_{t,a})+0.1/A\big]n^{-1}}\right)+0.01\right] \tag{1.11}$$

where P and $\hat{P}$ are observed and the model-predicted catch proportions at age. Finally, recruitment deviates are penalized under the assumption of lognormality,

$$\text{Pen} = \sum_{a=2}^{A-1} \frac{{}_R\varepsilon_{1,a}^2}{2\sigma_R^2} + \sum_{t=2}^{t_{\max}-1} \frac{{}_R\varepsilon_{t,1}^2}{2\sigma_R^2} \tag{1.12}$$

where ${}_R\varepsilon_{1,a}$ and ${}_R\varepsilon_{t,1}$ are recruitment deviates in the initial year and subsequent years, and σ_R is a measure of the extent of recruitment variability. The estimation models are given the correct (i.e. the operating model) values for $\sigma_I = 0.2$, the effective sample size $n = 50$ for the catch-at-age data, and recruitment variability $\sigma_R = 0.6$.

1.2.4 Reference points

Six reference points are evaluated as potential management quantities of interest: B_{current} (current biomass), u_{current} (current harvest rate), Depletion (current depletion level), MSY (maximum sustainable yield), $B_{\text{current}}/B_{\text{MSY}}$ (current biomass relative to B_{MSY}, and Surplus (current surplus production). These reference points are chosen because they are commonly used in fisheries management. B_{current}, u_{current}, and Depletion are calculated using the equations:

$$B_{\text{current}} = \sum_a N_{2005,a}\, \Phi_a w_a \tag{1.13}$$

$$u_{\text{current}} = Y_{2004} \,/\, \sum_a ({}_C S_a\, N_{2004,a}\, w_a) e^{-M/2} \tag{1.14}$$

$$\text{Depletion} = B_{\text{current}} \,/ B_0 \tag{1.15}$$

The maximum sustainable yield, MSY, is defined as the long-term average catch when the harvest rate is set to an optimal value, u_{MSY}. The average catch at a given harvest rate can

be calculated in closed form, by combining methods from Lawson and Hilborn (1985) and Francis (1992). First, the equilibrium age composition is standardized so that the number of 1-year olds equals 1:

$$n_a^\star = \begin{cases} e^{-(a-1)M} \prod_{i=1}^{a-1} (1 - {}_C S_i u), & a < A \\ \dfrac{e^{-(A-1)M} \prod_{i=1}^{A-1} (1 - {}_C S_i u)}{1 - e^{-M}(1 - {}_C S_A u)}, & a = A \end{cases} \tag{1.16}$$

At this harvest rate, the average recruitment is $R^\star = (\mathrm{SBPR}^\star - \alpha)/(\beta \mathrm{SBPR}^\star)$, where $\alpha = \mathrm{SBPR}_0(1-h)/(4h)$, $\beta = (5h-1)/(4hR_0)$, $\mathrm{SBPR}^\star = \sum_a n_a^\star \Phi_a w_a$, and SBPR_0 is calculated in the same way as $\mathrm{SBPR}^\star$ except that $u = 0$. The average long-term catch for a given harvest rate is

$$Y^\star = u R^\star e^{-M/2} \sum_a n_{a\,C}^\star S_a w_a \tag{1.17}$$

and the corresponding spawning biomass is:

$$B^\star = R^\star \times \mathrm{SBPR}^\star \tag{1.18}$$

MSY and B_{MSY} are calculated by searching iteratively for the u that maximizes $Y^\star$. Finally, current surplus production is defined as the last year's catch, plus the resulting change in vulnerable biomass:

$$\mathrm{Surplus} = Y_{2004} + \sum_a {}_C S_a w_a (N_{2005,a} - N_{2004,a}) e^{-M/2} \tag{1.19}$$

The true reference point values from the operating model vary due to stochastic recruitment, except u_{current} which is pre-defined in each scenario (Table 1.2) and MSY which depends only on R_0, h, M and commercial selectivity. The true MSY value is in all cases

203 thousand tonnes, with harvest rate $u_{\mathrm{MSY}} = 0.154$ and spawning biomass $B_{\mathrm{MSY}} = 1270$ thousand tonnes. Table 1.5 gives an idea of the approximate values of the reference points, using the special case of deterministic recruitment (all ${}_R\varepsilon_{t,a} = 0$ and $\sigma_R = 0$) as an example.

1.2.5 Performance measures

The performance of an estimation model is quantified by comparing the estimates from the 100 data sets with the true values from the operating model, using two performance measures. One performance indicator is the bias of estimators,

$$\text{Median bias} \;=\; \text{median}\left(\frac{\hat{\theta}-\theta}{\theta}\right) \tag{1.20}$$

where $\hat{\theta}$ is the estimated value of a reference point and θ is the true value. The median bias is used rather than the mean, to make the performance indicator more robust to outlying estimates of the management quantities. The other performance indicator is the proportion of estimates that are less than half or greater than twice the true value,

$$\text{Failure rate} \;=\; \Pr(\hat{\theta}/\theta < 0.5 \cup \hat{\theta}/\theta > 2) \;\times\; 100 \tag{1.21}$$

where 0.5 and 2 bound an arbitrarily chosen range of 'acceptable' error. The failure rate is a robust measure of accuracy, capturing both bias and imprecision, while the median bias is better at detecting relatively small but consistent bias. Median bias has a possible range from -1 to ∞, and failure rate is between 0 and 100. An estimation model that performs well has median bias close to 0 and failure rate close to 0. The performance is also presented graphically using Tukey's boxplots, where a solid box shows the inner quartiles, and whiskers extend from the box to the outermost datapoint within 1.5 times the interquartile range (Tukey 1977).

1.3 Results

A total of 5200 model runs are analyzed: 100 data sets for each of the four scenarios and 13 estimation models. In the first part of the results, we look at how well the models estimate the reference points, and the second part focuses on selected model parameters.

1.3.1 Reference points

To facilitate comparison, the distribution of the estimated reference points (Figure 1.6) are expressed as ratios of the true values known from the operating model. The multipanel boxplot allows one to visually evaluate the estimation performance for each reference point across data scenarios and estimation models. For example, the top left panel shows how well each model estimates current spawning biomass when the data are simulated based on scenario A (one-way trip). In this panel the boxplot medians are not far from 1, indicating that the models estimate current abundance with relatively small bias. However, the uncertainty of the estimates is considerably greater for model families 1 and 2 than for model family 3. This is understandable, because model families 1 and 2 ignore the catch-at-age and abundance-index information, respectively, while model family 3 uses all of the available data. The two performance measures, median bias and failure rate (Tables 1.6 and 1.7), summarize the information in Figure 1.6.

B_{current}

When estimating current abundance (Figure 1.6, top row of panels), the models exhibit only a small bias in data scenario A (one-way trip), with models 2 and 2m exhibiting a negative bias of -0.2. The failure rate is also relatively low in scenario A, ranging from 0 for model 3, to 24 for models 1h and 2h. Most of the boxplots are wider in scenario B (no change), indicating that the data in this scenario are less informative about current abundance. Models 3 and 3h are exceptions from this general pattern, as their performance is comparable to scenario A. The considerably higher failure rate of models 3m, 3r, 3mr and 3hmr in scenario B shows

how the uncertainty increases when the natural mortality rate and/or right-hand selectivity are unknown. The performance of the estimation models in scenario C (good contrast) is better than in scenario B and about as good as scenario A. The greatest bias in scenario C is −0.4 for model 2m, which also has a relatively high failure rate of 51, while models 3 and 3h have 0. Scenario D (rebuild only) is the least informative about current abundance. The lowest failure rates are 6 and 9 for models 3 and 3h, but the estimates from these models have a bias of −0.2 and −0.3. The other models have much higher failure rates in scenario D, including the highest of all cases, 92 for model 2m.

u_{current}

The current harvest rate (Figure 1.6, second row of panels) is never greatly overestimated in scenario A. This is because the estimated fraction of the biomass caught in a year cannot be many times higher than the true value of 0.408 in this scenario (Table 1.5). Nevertheless, a small but consistent positive bias of c. +0.1 is shown by model family 3, but failure rates are quite low, 0 for models 3, 3h and 3r, up to 20 for model 1h. In scenario B, all models have high failure rates except for models 3, 3h and 3r, with 0, 2 and 10, respectively. Model 3r is unbiased, but 3 and 3h are positively biased by +0.1. The models that estimate natural mortality, 1m, 2m, 3m, 3mr and 3hmr all show high failure rates, between 37 and 70. Failure rates in scenario C are lower than in scenario B, but higher than in scenario A. The median bias ranges from 0 for models 3 and 3h, to +0.8 for model 2m, and failure rates are lower for model family 3 than the simpler models. Model performance in scenario D is considerably worse than in the other scenarios. Models 3 and 3h have low failure rates of 5 and 6, but consistently overestimate the harvest rate with a bias of +0.3 and +0.4. Model 3r has a smaller bias of −0.1, but a failure rate of 25, while model 2m shows an extreme −1.0 bias and a failure rate of 90.

Depletion

Many of the models estimate current depletion (Figure 1.6, third row of panels) in scenario A about as well as current abundance, but there are noteworthy exceptions. Specifically, the failure rate is consistently higher for current depletion compared to current abundance when the natural mortality rate is unknown, in models 1m, 2m, 3m, 3mr and 3hmr. Scenario B is again less informative than scenario A about current depletion, with median bias ranging from -0.8 for models 1m and 3hmr, to $+0.3$ for model 3m, and failure rate from 3 for model 3 to 81 for models 1h and 3hmr. Models 2, 2r, 3, 3h and 3r perform quite well in scenario C, with failure rates below 10, while model 2m is negatively biased and has a high failure rate of 62. Scenario D is clearly more informative about depletion than absolute abundance, with the models showing less extreme biases and failure rates. Nevertheless, models 1h, 1m and 3hmr provide negatively biased and inaccurate estimates of current depletion in scenario D.

MSY

All models in scenario A estimate the maximum sustainable yield (Figure 1.6, fourth row of panels) with quite low failure rates, although the estimates are often slightly biased towards overestimation. Model 1h has the highest failure rate, 29, but the failure rates of the remaining models are between 0 and 12. Models 2, 2r, 3, 3h and 3r perform relatively well in scenario B, but the other models overestimate MSY considerably. Model performance in scenario C is again similar to that in scenario A, except that models 1m and 2m have larger bias and higher failure rates. Scenario D is highly uninformative about MSY for all models except 3 and 3h, which have failure rates 3 and 13, respectively. Other models in this scenario are positively biased with failure rates between 26 and 88.

$B_{\text{current}}/B_{\text{MSY}}$

The ability to estimate the ratio $B_{\text{current}}/B_{\text{MSY}}$ (Figure 1.6, fifth row of panels) is largely similar to that for current depletion, the other ratio reference point, reflecting a strong

correlation between the B_0, B_{current} and B_{MSY} parameter estimates. The failure rates in scenario A range from 2 for model 3, to 45 for model 2h. Scenario B is less informative about the stock status relative to B_{MSY}, although models 2, 3 and 3r perform relatively well. The estimation models in scenario C are subject to rather small biases, with the exception of -0.6 for model 2m. In scenario D, most of the models have failure rates below 50, but model 3hmr is strongly biased downwards, with median bias -1.0.

Surplus

All models estimate current surplus production (Figure 1.6, bottom row of panels) quite accurately in scenario A, with failure rates ranging from 0 for model 3r, to 20 for model 2h, which also shows the greatest bias of $+0.2$. Scenario B is much more informative about surplus production than about other reference points, with a bias of -0.2 to $+0.1$, and failure rates from 8 for model 3, to 45 for model 2m. The estimation performance is also good in scenario C, with the greatest bias being -0.3 for model 2r, and failure rates ranging from 4 for model 3, to 39 for model 2m. In scenario D, failure rates are generally high, over 50 for all models in families 1 and 2. The only models that perform well here are 3 and 3h, with failure rates of 13 and 15, and a bias of -0.1. Thus, estimating surplus production reliably in scenario D requires that both catch-at-age data and an index of abundance are available, as well as perfect knowledge about the true natural mortality rate, recruitment steepness and right-hand selectivity.

1.3.2 Parameters

Steepness h and natural mortality M are estimated directly in the model, while selectivity at oldest age S_{10} is a derived parameter from Equation 1.3. In Figure 1.7, the estimated parameter values are divided by the true values from the operating model, which are $h=0.7$, $M=0.2$, and $S_{10}=0.94$.

Steepness is overestimated by all models in all scenarios, but relatively accurate estimates are seen in scenario D using models 2h and 3h. By definition, steepness has an upper bound

of 1 (Francis 1992) and many estimates in the top row of panels in Figure 1.7 run into this bound, where $\hat{h}/h = 1.0/0.7 = 1.43$, but less frequently in scenario D. Estimates of natural mortality rate are generally unreliable, especially using the 1m or 3hmr models, but relatively accurate estimates are seen in scenario A using the 3m model. When right-hand selectivity is estimated as well, in models 3mr and 3hmr, M becomes biased towards underestimation. In other words, when the estimated selectivity does not fully target older fish, the relatively high frequency of older fish in the catch can be fitted by increasing the natural mortality rate. Selectivity at oldest age is consistently underestimated (Figure 1.7, bottom row of panels). This bias is partly due to the true value of $S_{10} = 0.94$ being so near the theoretical upper bound of 1, but the estimates are also inaccurate, in many cases less than half the true value. The performance does not differ much between models 2r, 3r, 3mr and 3hmr, but scenario A is slightly more informative than the others about selectivity at oldest age.

1.4 Discussion

Below, the hypotheses are reviewed in light of the results, using average failure rate as a summary statistic. This is followed by a general discussion about implications and the strengths and weaknesses of the experimental design.

1.4.1 Hypotheses

H_1 Fisheries data are most informative when they span a period where the population was fished down to a low level.

H_2 Fisheries data are most informative when they span a period where the population was fished down to a low level and then allowed to rebuild for some time.

Table 1.8 shows the estimation performance in each scenario, where each average is based on 65 failure rates from Table 1.7, across models and reference points. There is a clear division, also noticeable in Figure 1.6, where fishing histories A (one-way trip) and C (good contrast) provide more reliable data to estimate the reference points than B (no change) and

D (rebuild only). The results provide slightly more support to hypothesis 1 (average failure rate in scenario A = 12.7) than hypothesis 2 (C = 21.0), but both of those scenarios are much more informative than B or D.

The results from the 'one-way trip' scenario imply that fisheries data spanning an early period of high abundance followed by low abundance are likely to be informative in age-structured stock assessment, even if the fishing history does not include subsequent rebuilding. This is in contrast to findings from simulation studies of biomass-dynamic models, where a rebuilding phase provides necessary information to estimate the population growth parameters (Hilborn 1979, Hilborn and Walters 1992). It is worth noting that those studies looked at how well the parameters of the Schaefer model were estimated, not just reference points.

H_3 The level of stock depletion is estimated more reliably than other reference points.

The reference point with the lowest overall failure rate is not the current depletion level, but surplus production (Table 1.9). But from Figure 1.6 it is clear that Depletion is a more robust reference point across all scenarios. The reference points that are in absolute biomass units (B_{current}, MSY and Surplus) become highly unreliable in scenario D (rebuild only), particularly when natural mortality is unknown. The relative biomass estimates (Depletion and $B_{\text{current}}/B_{\text{MSY}}$) perform much better in those cases.

Punt et al. (2002) and other studies have shown that depletion is generally estimated more reliably than other reference points. The greater the correlation is between the B_{current} and B_0 parameter estimates, the smaller the variance around the estimated ratio of the two. The results from scenarios A through D indicate that depletion may be subject to slightly higher failure rates than some other reference points when the data are informative, but is a robust quantity to estimate in worst-case uninformative scenarios.

An interesting exception is how accurately current surplus production is estimated in scenario B. This is understandable, since if the abundance index and catch is constant over time, then the surplus production must be roughly equal to the catch.

H_4 A data set that includes both an index of relative abundance and catch-at-age data is much more informative than a data set that includes only one of these two types of data.

Estimation models of family 3 (all data types) perform better than family 1 (no age data) and family 2 (no abundance index), as shown in Table 1.10. This comes as no surprise and is in agreement with the recommendations by Deriso et al. (1985). Models similar to those of family 1 have been used by Hilborn (1990) and others, and are seen by many as a preferable alternative to traditional biomass-dynamic models (Maunder 2003). Their argument is that traditional biomass-dynamic models make implicit assumptions that offer limited freedom to explore different hypotheses about the fishery dynamics.

Model family 2 performs surprisingly well. Even in the absence of abundance-index data, the landings and age-composition data provide considerable information to estimate both absolute abundance and relative depletion. The common view is that estimation of these quantities requires either an abundance index or highly restrictive assumptions (Shepherd 1984, Deriso et al. 1985, Hilborn and Walters 1992), but here the assumptions of model family 2 are similar to families 1 and 3. Furthermore, convergence diagnostics indicated that models of family 2 were no less estimable than the other models. This behaviour may be due to the simplified nature of the simulation environment, but it is worth remembering that a precursor of the statistical catch-at-age model (Doubleday 1976) did not include abundance-index data. Statistical catch-at-age models combine the landings and the commercial catch-at-age data in a framework that provides more insight than analyzing each cohort separately (Fournier and Archibald 1982). An additional element of information for model families 1–3 is the penalty (Equations 1.9 and 1.12) that allows the estimated recruitment to vary considerably (c. 20-fold difference between largest and smallest recruitment in scenario B with $\sigma_R = 0.6$) but not by many orders of magnitude, whereas recruitment is completely free in the model used by Deriso et al. (1985).

H_5 Not knowing M, h, and right-hand selectivity leads to inaccurate estimates of stock

abundance and reference points.

As expected, the models perform better when parameters are fixed at the true value, than when they are estimated. Even so, the 3hmr model performs quite well in the informative scenarios A and C, especially estimating current abundance, harvest rate, MSY and surplus production (Figure 1.6, Tables 1.6 and 1.7).

By admitting uncertainty about M, h, and right-hand selectivity, model 3hmr represents the real task facing stock assessment scientists. These parameters are highly confounded, so model 3hmr cannot be expected to perform reliably when fitted to real fisheries data, that come from a much more complex system than the operating model used in this study. In practice, some or all of these parameters would be fixed at an assumed value, or be assigned an informative Bayesian prior probability distribution. The effect of fixing these parameters at values that are very different from the true dynamics has been explored by Thompson (1994), Clark (1999), Ianelli (2002), and others.

H_6 Models estimating M perform about as well as models estimating h.

Models estimating h perform better on the average than those estimating M, especially when the data include both catch at age and an index of abundance (Table 1.11). This means that uncertainty about the natural mortality rate is more important than the uncertainty about the shape of the stock-recruitment curve, when estimating the stock status. Model 3m shows particularly bad performance in scenarios B and D, so in those scenarios any external information about M would be valuable. This information could be used to construct a Bayesian prior for M, or to fix the parameter, which is analogous to an extremely narrow Bayesian prior (Gelman et al. 2004).

The highest overall failure rates were shown by model 2m in scenario D. As expected (Beverton and Holt 1957), it is simply not feasible to estimate harvest rate and natural mortality from catch-at-age, when harvest rate has been steady and low.

H_7 M can be estimated reliably if age-composition data are available from when the population was unfished.

H_8 M can be estimated reliably from the rate of population increase if the stock is allowed to rebuild from a low level.

Model 3m estimates M with greater accuracy in scenarios A and C than in the other scenarios (Figure 1.7). This was expected, since the age structure in the first few years of the fishery carries information about the natural mortality rate (Beverton and Holt 1957). After taking the individual cohort sizes into account, using data from all years, M can be inferred from the age composition in the first years. Importantly, the model was not 'told' that the stock was unfished in the first year in scenarios A and C, as the parameters R_{init} and u_{init} were estimated in all cases.

The estimation of M is less accurate in scenario D. One might have expected this scenario to be informative about the value of M, as the rate at which the stock rebuilds is dependent on this parameter. The other main factor determining the rebuilding rate is recruitment, so the variable recruitment ($\sigma_R = 0.6$) might explain why scenario D is not informative about M. Another reason could be that the observation noise ($\sigma_I = 0.2$) causes the observed abundance index to suggest random fluctuations instead of a steady growth.

H_9 h can be estimated reliably from catch-at-age data and an index of relative abundance when the data cover a period in which abundance varies substantially.

Although scenario A involves the widest range of abundance (Figure 1.2), it is only in scenario D that model 3h estimates h reliably (Figure 1.7). Recruitment success at low spawning stock levels is informative about the shape of the stock-recruitment curve (Ricker 1958), and scenario D includes a large number of cohorts spawned by a small parent stock. The initial stock status in scenario D (c. 3% of B_0) is also considerably lower than the last year's stock status in scenario A (c. 10% of B_0). When the data do not include years of very low abundance, the models tend to overestimate the steepness parameter (Figure 1.7).

H_{10} Right-hand selectivity can only be estimated reliably when M is known.

The results from this study are not conclusive about the estimation of right-hand selectivity, as the only case considered is when the true selectivity curve is nearly asymptotic. Nonetheless, the results do suggest that when the true selectivity is nearly asymptotic, the reliability of estimating right-hand selectivity (Figure 1.7, bottom panel) depends more on the scenario than on whether M is estimated or fixed. Thompson (1994) performed a more thorough analysis of the relationship between these parameters, concluding that right-hand selectivity can only be estimated reliably when M is known.

1.4.2 Implications

The results presented here show how the perceived uncertainty about stock status is not only affected by the available data, but also by the assumptions made in the estimation process.

The features of different fishing history scenarios determine how informative the data are about management quantities. The main feature of an informative fishing history is a large decrease in abundance, while other features, such as contrast in harvest rate, seem to be of secondary importance. Although strong depletion is to be avoided due to the ecological risk and economic cost, it does provide informative data. In the words of John G. Pope (personal communication), 'the more fish you catch, the better you know how many there were.'

An uninformative fishing history, commonly seen in practice, is when a relative index of abundance and age data are not available from the early years of the fishery. In these cases, depletion level tends to be more robust than other commonly used reference points, although surplus production can also be estimated accurately when the abundance remains stable over a long period. When the data are informative, other reference points can be expected to perform just as well, or better. Despite regular criticism, MSY remains a key concept in fisheries management, if not as a goal, then as an upper limit of a precautionary approach (Mace 2001, Punt and Smith 2001). MSY is independent of the current stock status, being a function of R_0, M, h, commercial selectivity, weight, and maturity at age.

When the estimation models are given the true value of most of those quantities, MSY will be estimated quite accurately. This can be seen from the performance of our models 1, 2 and 3, which generally estimated MSY with a lower failure rate than the other reference points. However, it is also important to note that MSY was more often overestimated than underestimated.

Catch-at-age data can provide information about the current stock status, even without a relative abundance index. When the true value of M is known, the total mortality rate of cohorts leads to an accurate estimate of annual harvest rate, which combined with known annual catches leads to an accurate estimate of vulnerable biomass. The assumption of a known constant M plays a central role here. This assumption is commonly made in practice, and the effects of its violations are largely understood (Mertz and Myers 1997, Clark 1999). It is an unrealistic assumption (Cotter et al. 2004), but a time-constant M, estimated or fixed, is seen as necessary to evaluate the consequences of alternative catch levels (Punt and Hilborn 1997), which is the central purpose of fisheries stock assessment.

The statistical catch-at-age model yields more information from catch-at-age data than earlier catch-curve methods, given that the added assumptions about recruitment are justifiable. Catch-at-age and abundance-index data become particularly informative when used together, as they provide complementary information about different aspects of the population dynamics, and are subject to different assumptions. It is known that the sampled and processed catch-at-age data do not necessarily reflect the population age-structure very well (Pope 1988), and empirical evidence also undermines the assumption about a constant linear relationship between the abundance index and population abundance (Harley et al. 2001). When these two data types tell a consistent story about the population trends, it indicates that the model assumptions are likely to be justifiable. This can be checked by fitting models that exclude data components, as was carried out in this study, or by changing the likelihood weights via the catch-at-age sample size and observation uncertainty σ_I about the abundance index. When the two data types provide contradicting information about the stock status, the validity of each data source needs to be examined (Schnute and Hilborn 1993).

When evaluating confidence bounds around estimated quantities, one should strive to incorporate all major sources of uncertainty. This means estimating parameters instead of fixing them, but this is not always statistically feasible. Confounded parameters like natural mortality rate M, stock-recruitment steepness h and declining right-hand selectivity can be estimated when the data carry information about these quantities. For M, this means catch-at-age data from the early years of the fishery, or at least a contrasted history of harvest rates (Beverton and Holt 1957) and for h it means catch-at-age data from a period of very low abundance (Ricker 1958), as verified in this study. With the constant harvest rates in scenarios B and D, there is no information to separate the total mortality rate between natural mortalities and fishing mortalities. The selectivity of older fish can be estimated when M is a fixed parameter (Thompson 1994). In a Bayesian model, one or more of these parameters can be assigned an informative prior distribution, perhaps from a meta-analysis of many related stocks (Myers et al. 1999), instead of estimating as a free parameter or fixing completely.

Overall, the estimation models showed considerably high failure rates, where management quantities were underestimated or overestimated by a factor of two or more. Bearing in mind the simple 'laboratory conditions' of this simulation study, stock assessment models can only be expected to have higher failure rates when fitted to real fisheries data. A retrospective look at fisheries assessments around the world shows that management quantities are not estimated as accurately as statistical theory suggests, due to violated assumptions and ignored sources of uncertainty (NRC 1998, Walters and Martell 2004).

1.4.3 Strengths and weaknesses

This study advances our understanding of fisheries stock assessment models, with respect to what kinds of data are informative or uninformative, and highlights the role of assumptions. Based on the experimental design and findings of Hilborn (1979), Hilborn and Walters (1992), NRC (1998), Gavaris and Ianelli (2002) and Punt et al. (2002), this simulation study uses up-to-date statistical methods that take advantage of the computing power available today.

The scope is wide, addressing a variety of questions, and the conclusions can be used to support various decisions made in any fisheries stock assessment.

Compared to real fisheries, with their complex interaction between biological and human systems, the simulation approach is a simplified abstraction. Apart from stochastic recruitment, the parameters in the operating model are constant over time (natural mortality rate, catchability and selectivity), and the estimation models are specified without model error and given the true survey selectivity. These decisions were made deliberately to make the results as easy to interpret as possible. Excluding survey catch-at-age data from the study allowed a clear separation between two kinds of information: commercial catch at age reflecting the age distribution of the population over time and a survey abundance index reflecting relative changes in biomass over time.

The wide scope of this study comes at a cost, as the experimental design is not optimal for any one of the 10 hypotheses. Each of them could be tested more rigorously with a simulation study specifically designed for that purpose. Similarly, there are many more hypotheses that could be addressed using the same simulation framework, but applying other treatments than was done here. For example, examining the effect of fixing parameters at values that are substantially lower or higher than the true value. Model errors and violated assumptions are inevitable in stock assessments, but the combined experience from real fisheries and simulation studies will help making fisheries data as informative as possible.

Acknowledgements

This research was supported by the Icelandic Centre for Research and the New Zealand Seafood Industry Council. We thank André Punt, Jim Ianelli, Rick Methot, John Pope, and two anonymous reviewers for insightful comments, and Pamela Nelle for a thorough review of the manuscript.

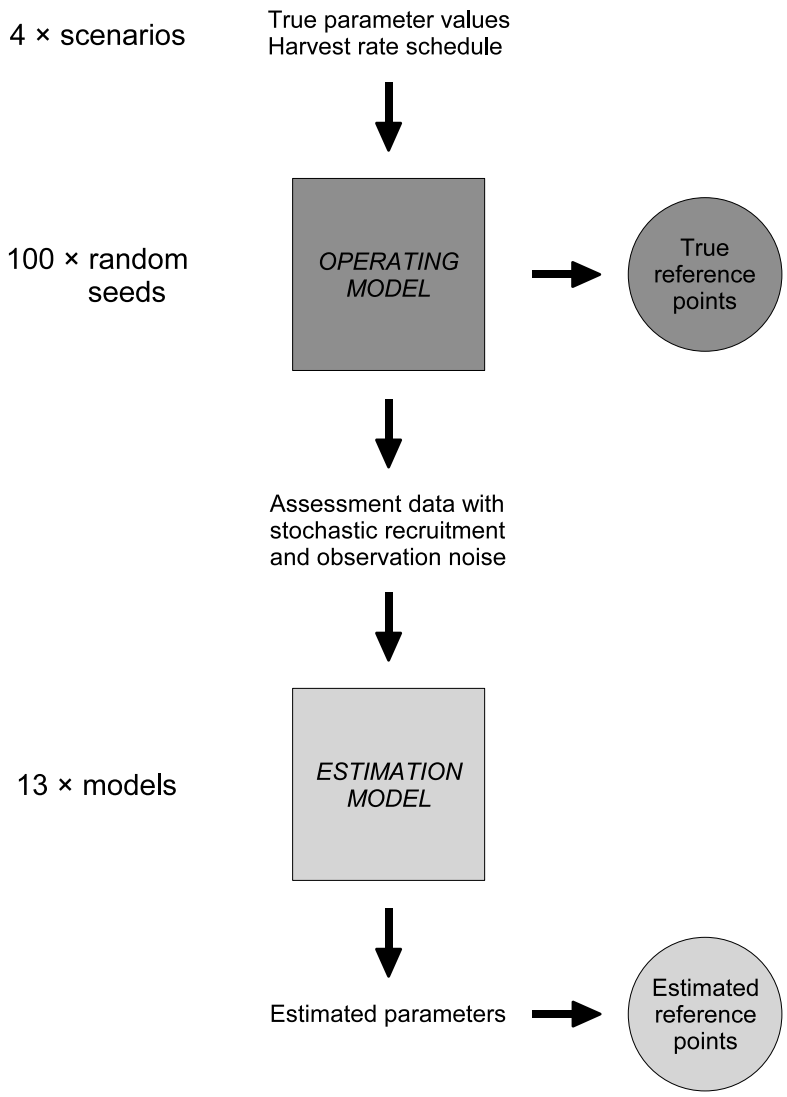

Figure 1.1. The simulation procedure. Arrows and boxes indicate the workflow for a single run, and multiplications describe how the study consists of multiple runs.

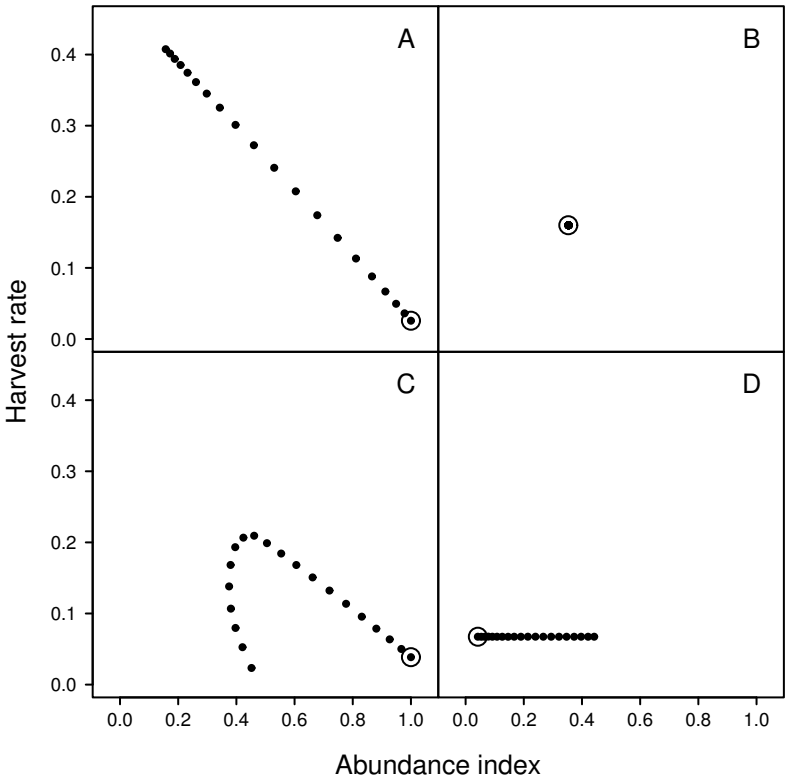

Figure 1.2. The four fishing history scenarios considered in this study, in terms of the relationship between the harvest rate and the expected value of the abundance index. Circles represent the status of the fishery in the first year. (A) One-way trip, (B) no change, (C) good contrast, (D) rebuild only.

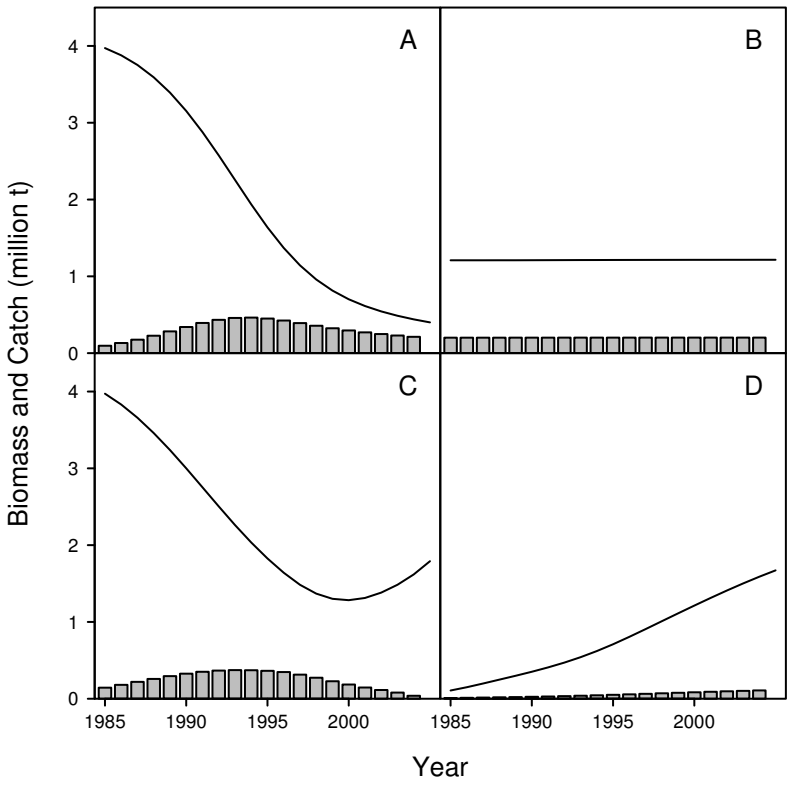

Figure 1.3. The four fishing history scenarios considered in this study, in terms of spawning biomass (line) and landed catch (bars). (A) One-way trip, (B) no change, (C) good contrast, (D) rebuild only.

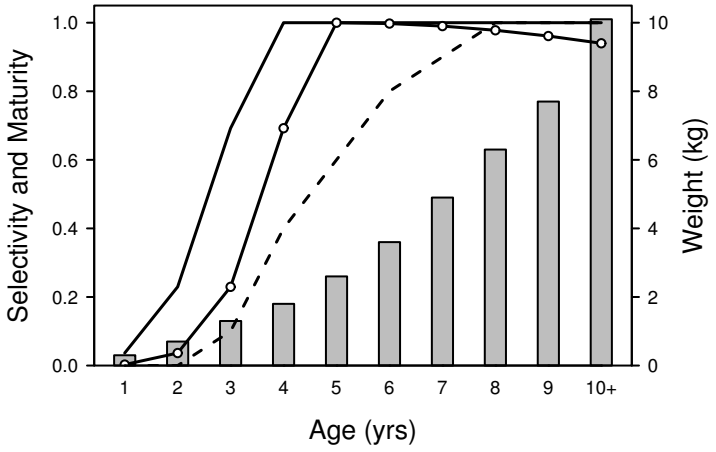

Figure 1.4. Age-specific characteristics of the operating model: survey selectivity (plain line), commercial selectivity (line with circles), maturity (dashed line), and weight (bars).

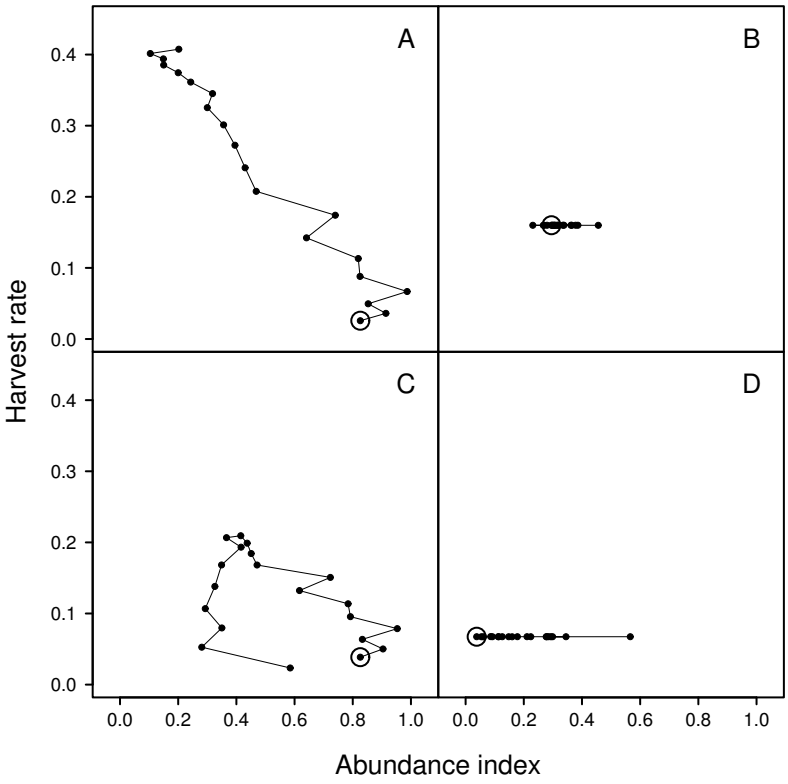

Figure 1.5. Examples of stochastic data sets (random seed = 100), in terms of the relationship between the harvest rate and the expected value of the abundance index. Circles represent the status of the fishery in the first year. (A) One-way trip, (B) no change, (C) good contrast, (D) rebuild only.

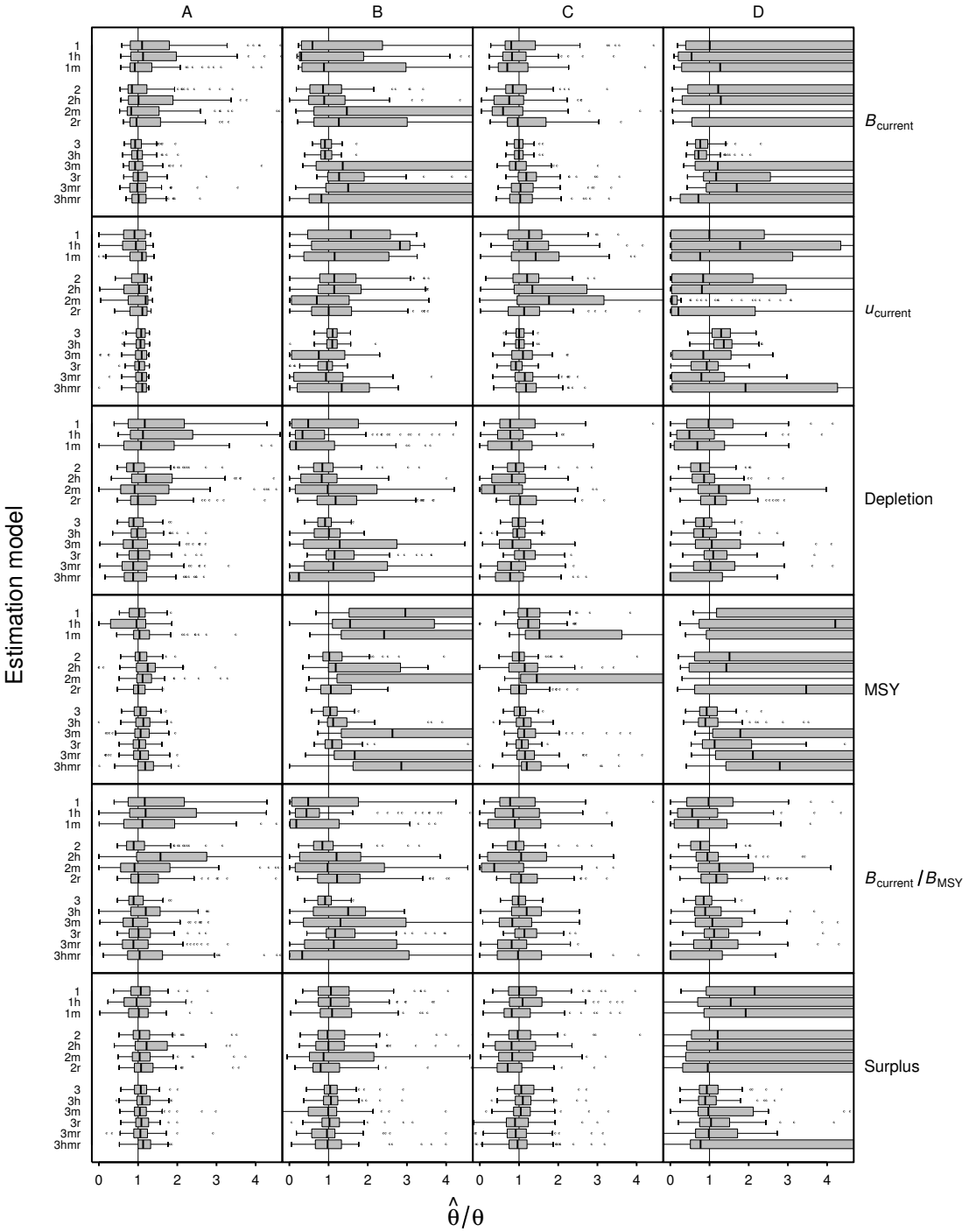

Figure 1.6. Distribution of estimated reference points. Panel columns correspond to fishing history scenarios A–D and panel rows are the six different reference points. Each Tukey boxplot shows the distribution of 100 estimates, divided by the true value of the reference point from the operating model. The x-axis is truncated to avoid loss of detail.

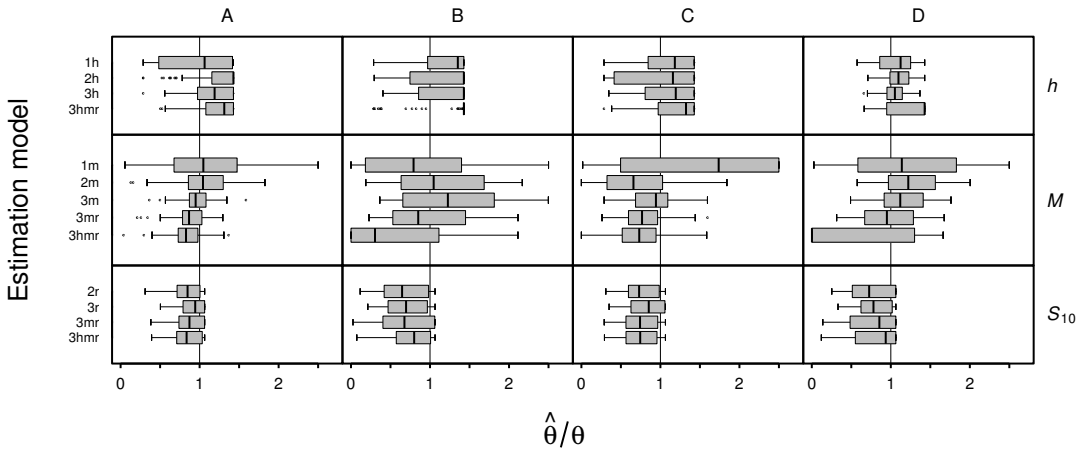

Figure 1.7. Distribution of estimated parameters h (steepness), M (natural mortality), and derived parameter S_{10} (selectivity at oldest age). Panel columns correspond to fishing history scenarios A–D and panel rows are the different parameters. Each Tukey boxplot shows the distribution of 100 estimates, divided by the true value of the parameter from the operating model.

Table 1.1. Age-specific weight (kg) and maturity (proportion) used in the operating and estimation model.

Age	1	2	3	4	5	6	7	8	9	10+
Weight (kg)	0.3	0.7	1.3	1.8	2.6	3.6	4.9	6.3	7.7	10.1
Maturity	0.0	0.0	0.1	0.4	0.6	0.8	0.9	1.0	1.0	1.0

Table 1.2. Parameter values and harvest rate schedules for the four fishing history scenarios.

Scenario	A One way	B No change	C Contrast	D Rebuild
Parameters				
R_0	250 000	*	*	*
h	0.7	*	*	*
M	0.2	*	*	*
R_{init}	1	0.8	1	0.2
u_{init}	0	0.16	0	0.4
R_{plus}	1	*	*	*
${}_{C}S_{\text{full}}$	5	*	*	*
${}_{C}S_{\text{left}}$	1	*	*	*
${}_{C}S_{\text{right}}$	6	*	*	*
${}_{S}S_{\text{full}}$	4	*	*	*
${}_{S}S_{\text{left}}$	1	*	*	*
${}_{S}S_{\text{right}}$	15	*	*	*
q	2.5×10^{-7}	*	*	*
Harvest rate				
1985	0.026	0.160	0.039	0.067
1986	0.036	0.160	0.050	0.067
1987	0.050	0.160	0.064	0.067
1988	0.067	0.160	0.079	0.067
1989	0.088	0.160	0.096	0.067
1990	0.113	0.160	0.114	0.067
1991	0.142	0.160	0.132	0.067
1992	0.174	0.160	0.151	0.067
1993	0.208	0.160	0.168	0.067
1994	0.241	0.160	0.184	0.067
1995	0.273	0.160	0.199	0.067
1996	0.301	0.160	0.209	0.067
1997	0.325	0.160	0.207	0.067
1998	0.345	0.160	0.193	0.067
1999	0.361	0.160	0.168	0.067
2000	0.374	0.160	0.138	0.067
2001	0.385	0.160	0.107	0.067
2002	0.394	0.160	0.080	0.067
2003	0.401	0.160	0.053	0.067
2004	0.408	0.160	0.023	0.067

An asterisk indicates that the same parameter value applies across all scenarios.

Table 1.3. The 13 estimation models in terms of data types and parameters estimated.

Model	1	1h	1m	2	2h	2m	2r	3	3h	3m	3r	3mr	3hmr
Data													
Catch	x	x	x	x	x	x	x	x	x	x	x	x	x
Index	x	x	x					x	x	x	x	x	x
CA				x	x	x	x	x	x	x	x	x	x
Estimated													
R_0	x	x	x	x	x	x	x	x	x	x	x	x	x
h		x			x				x				x
M			x			x				x		x	x
R_{init}	x	x	x	x	x	x	x	x	x	x	x	x	x
u_{init}	x	x	x	x	x	x	x	x	x	x	x	x	x
R_{plus}				x	x	x	x	x	x	x	x	x	x
${}_{C}S_{\text{full}}$				x	x	x	x	x	x	x	x	x	x
${}_{C}S_{\text{left}}$				x	x	x	x	x	x	x	x	x	x
${}_{C}S_{\text{right}}$							x				x	x	x
q	x	x	x					x	x	x	x	x	x
${}_{R}\varepsilon$				x	x	x	x	x	x	x	x	x	x

Catch stands for landings, Index for survey abundance index, and CA for commercial catch at age.

Table 1.4. Bounds on estimated parameters, along with the true values from the operating model.

Parameter	True value	Lower bound	Upper bound
R_0	250 000	1 000	10 000 000
h	0.7	0.2	1
M	0.2	0	0.5
R_{init}	0.2−1	0	5
u_{init}	0−0.4	0	1
R_{plus}	1	0	2
${}_{C}S_{\text{full}}$	5	3	10
${}_{C}S_{\text{left}}$	1	−2	5
${}_{C}S_{\text{right}}$	6	−2	15
$\log q$	−15.2	−30	0
${}_{R}\varepsilon$	*	−15	15

The true value of R_{init} and u_{init} varies among scenarios (see Table 1.2).

*: Initial age structure and annual recruitment varies varies between the simulated data sets.

Table 1.5. True reference point values from the operating model, given deterministic recruitment.

Scenario	A One way	B No change	C Contrast	D Rebuild
B_{current}	400	1216	1791	1672
u_{current}	0.408	0.160	0.023	0.067
Depletion	0.101	0.306	0.451	0.421
MSY	203	203	203	203
$B_{\text{current}}/B_{\text{MSY}}$	0.315	0.957	1.410	1.316
Surplus	170	203	195	185

Table 1.6. Bias of estimated reference points, by scenario and model.

	B_{current}	u_{current}	Depletion	MSY	$\frac{B_{\text{current}}}{B_{\text{MSY}}}$	Surplus
A1	+0.1	−0.1	+0.2		+0.2	+0.1
A1h	+0.1	−0.1	+0.1		+0.2	
A1m	−0.1	+0.1	+0.1		+0.1	
A2	−0.2	+0.2	−0.1		−0.1	
A2h			+0.2	+0.2	+0.6	+0.2
A2m	−0.2	+0.2	−0.1	+0.1	−0.1	
A2r		+0.1				+0.1
A3	−0.1	+0.1	−0.1	+0.1	−0.1	+0.1
A3h		+0.1		+0.1	+0.2	+0.1
A3m	−0.1	+0.1	−0.1	+0.1	−0.1	
A3r						+0.1
A3mr		+0.1	−0.1	+0.1	−0.1	+0.1
A3hmr		+0.1	−0.1	+0.2		+0.1
B1	−0.4	+0.6	−0.5	+2.0	−0.5	+0.1
B1h	−0.7	+1.8	−0.7	+0.5	−0.6	+0.1
B1m	−0.1	+0.2	−0.8	+1.4	−0.8	+0.1
B2	−0.1	+0.1	−0.2		−0.2	
B2h	−0.1	+0.1	−0.2	+0.2	+0.2	
B2m	+0.5	−0.3		+4.3		−0.1
B2r	+0.3		+0.2	+0.1	+0.2	−0.2
B3	−0.1	+0.1	−0.1		−0.1	
B3h	−0.1	+0.1		+0.1	+0.5	+0.1
B3m	+0.4	−0.2	+0.3	+1.6	+0.3	
B3r	+0.3		+0.2	+0.1	+0.2	
B3mr	+0.5	−0.1	+0.1	+0.7	+0.1	
B3hmr	−0.2	+0.3	−0.8	+1.9	−0.7	
C1	−0.2	+0.3	−0.2	+0.2	−0.2	
C1h	−0.2	+0.2	−0.2	+0.2	−0.1	+0.1
C1m	−0.3	+0.4	−0.2	+0.5	−0.1	−0.2
C2	−0.2	+0.2	−0.1		−0.1	
C2h	−0.3	+0.3	−0.2	+0.1	+0.1	−0.2
C2m	−0.4	+0.8	−0.6	+0.5	−0.6	−0.2
C2r		+0.1			+0.1	−0.3
C3						+0.1
C3h				+0.1	+0.2	+0.1
C3m	−0.1	+0.1	−0.2	+0.1	−0.2	
C3r	+0.2	−0.1	+0.1	+0.1	+0.1	−0.1
C3mr		+0.1	−0.2	+0.2	−0.2	−0.1
C3hmr		+0.2	−0.2	+0.2		
D1				+12.8		+1.2
D1h	−0.5	+0.8	−0.5	+3.2	−0.4	+0.5
D1m	+0.3	−0.2	−0.3	+12.1	−0.3	+0.9
D2	+0.2	−0.2	−0.2	+0.5	−0.2	+0.2
D2h	+0.3	−0.2	−0.1	+0.4	−0.1	+0.2
D2m	+26.9	−1.0	+0.2	+23.8	+0.2	+4.8
D2r	+4.3	−0.8	+0.1	+2.5	+0.2	
D3	−0.2	+0.3	−0.1	−0.1	−0.1	−0.1
D3h	−0.3	+0.4	−0.2	−0.1	−0.1	−0.1
D3m	+0.2	−0.2	+0.1	+0.8	+0.1	
D3r	+0.2	−0.1	+0.1	+0.1	+0.1	
D3mr	+0.7	−0.2		+1.1		
D3hmr	−0.3	+0.9	−1.0	+1.8	−1.0	−0.2

Blank entries denote negligible bias, between −0.05 and +0.05.

Table 1.7. Failure rates of estimated reference points, by scenario and model.

	B_{current}	u_{current}	Depletion	MSY	$\frac{B_{\text{current}}}{B_{\text{MSY}}}$	Surplus
A1	20	13	34	1	34	7
A1h	24	20	29	29	39	17
A1m	15	11	39	12	41	12
A2	9	3	10	0	10	5
A2h	24	15	25	6	45	20
A2m	16	10	39	7	40	7
A2r	19	4	16	1	17	5
A3	0	0	2	0	2	1
A3h	1	0	5	1	17	1
A3m	5	4	28	7	28	4
A3r	1	0	4	0	4	0
A3mr	2	1	26	7	26	4
A3hmr	2	1	25	4	27	1
B1	74	70	73	63	73	24
B1h	83	82	81	46	75	25
B1m	74	70	74	58	77	25
B2	34	27	22	12	22	21
B2h	45	38	41	28	49	28
B2m	65	64	67	61	69	50
B2r	45	39	29	21	30	30
B3	0	0	3	0	3	4
B3h	2	2	20	20	43	5
B3m	49	39	68	56	69	36
B3r	21	10	18	5	20	9
B3mr	45	37	62	45	65	34
B3hmr	60	58	81	69	88	29
C1	32	32	35	13	35	24
C1h	21	22	32	17	36	25
C1m	41	40	46	39	52	23
C2	16	14	8	4	8	11
C2h	36	36	33	31	47	38
C2m	51	52	62	35	62	39
C2r	21	16	8	6	9	37
C3	0	0	0	0	0	4
C3h	0	0	7	5	19	7
C3m	3	3	28	7	30	12
C3r	9	1	3	0	3	17
C3mr	8	8	30	8	31	22
C3hmr	8	8	35	15	42	22
D1	74	73	46	59	46	57
D1h	91	90	56	59	56	59
D1m	84	83	52	69	49	58
D2	69	66	26	54	26	61
D2h	77	76	24	73	25	77
D2m	92	90	44	88	48	87
D2r	83	85	21	78	23	77
D3	6	5	10	3	10	13
D3h	9	6	17	13	21	15
D3m	49	47	36	47	38	42
D3r	31	25	9	26	11	32
D3mr	53	46	33	52	35	41
D3hmr	85	89	64	57	65	55

Table 1.8. Average failure rate in each scenario, across all reference points and models.

Scenario A	Scenario B	Scenario C	Scenario D
12.7	41.8	21.0	49.1

Table 1.9. Average failure rate for each reference point, across all models and scenarios.

B_{current}	u_{current}	Depletion	MSY	$\frac{B_{\text{current}}}{B_{\text{MSY}}}$	Surplus
34.3	31.4	32.4	27.2	35.4	26.1

Table 1.10. Average failure rate for each estimation model family, across all reference points and scenarios.

Family 1	Family 2	Family 3
45.4	35.8	20.9

Table 1.11. Average failure rate for estimation models 1h, 1m, 2h, 2m, 3h, and 3m, across all reference points and scenarios.

	h	m
Model family 1	46.4	47.7
Model family 2	39.0	51.9
Model family 3	9.8	30.6

Chapter 2

MEASURING UNCERTAINTY IN FISHERIES STOCK ASSESSMENT: THE DELTA METHOD, BOOTSTRAP, AND MCMC

Abstract

Fisheries management depends on reliable quantification of uncertainty for decision-making. We evaluate which uncertainty method can be expected to perform best for fisheries stock assessment. The method should generate confidence intervals that are neither too narrow nor too wide, in order to cover the true value of estimated quantities with a probability matching the claimed confidence level. This simulation study compares the performance of the delta method, the bootstrap, and Markov chain Monte Carlo (MCMC). A statistical catch-at-age model is fitted to 1000 simulated datasets, with varying recruitment and observation noise. Six reference points are estimated, and confidence intervals are constructed across a range of significance levels. Overall, the delta method and MCMC performed considerably better than the bootstrap, and MCMC was the most reliable method in terms of worst-case performance, for our relatively data-rich scenario and catch-at-age model, which was not subject to substantial model misspecification. All three methods generated too narrow confidence intervals, underestimating the true uncertainty. Bias correction improved the bootstrap performance, but not enough to match the performance of the delta method and MCMC. We recommend using MCMC as the default method for quantifying uncertainty in fisheries stock assessment, although the delta method is the fastest to apply, and the bootstrap is useful to diagnose estimator bias.

2.1 Introduction

2.1.1 Which method performs best?

Fisheries management relies not only on point estimates of key quantities, such as biomass and harvest rate, but also on the uncertainty about these estimates. The uncertainty can be used to convey likely outcomes resulting from different management decisions, or incorporated into management strategy evaluation to find a long-term harvest strategy that performs well in face of uncertainty.

When estimating measurement uncertainty, fisheries scientists generally choose the statistical method they are most familiar with, or one that has become traditional for a particular stock. Three commonly used methods that will be evaluated here are the delta method, the bootstrap, and Markov chain Monte Carlo (MCMC) simulation. These methods have been shown to perform well with simple models, when all assumptions are met (Oehlert 1992; Efron and Tibshirani 1993; Gelman et al. 2004). In this study, we ask the question: given a typical age-structured stock assessment model and simulated datasets, which method performs best?

Patterson et al. (2001) provide a thorough review of uncertainty methods and describe three paradigms for evaluating uncertainty in stock assessment: frequentist, likelihood, and Bayesian inference. For the purposes of fisheries stock assessment, the theoretical difference between these paradigms is often ignored in practice (Restrepo et al. 2000; Patterson et al. 2001; Gavaris and Ianelli 2002; Hilborn 2003), and the methods are all used to express the plausible range of estimated quantities. In the strict frequentist sense, a confidence interval is a probabilistic statement about the proportion of such intervals that would cover the true parameter value in repeated experiments (Neyman 1937; Casella and Berger 2002). This frequentist statement treats the interval limits as random and the parameter as fixed, in the context of repeated experimental trials, and is therefore quite meaningful in a simulation study like this one, but it does not directly answer the relevant questions for environmental decision-making (Ellison 1996; Punt and Hilborn 1997; Ascough et al. 2008). Bayesian

inference, on the other hand, treats the interval limits as fixed and the parameter as random, leading to an intuitive statement about the probability that the true parameter value lies in the interval. The Bayesian interval is sometimes called a 'credible interval' (Casella and Berger 2002), a 'posterior interval' (Gelman et al. 2004), or simply a 'confidence interval' (Hilborn and Mangel 1997; Clark 2005) when the theoretical difference is considered of secondary importance, as is the case in this study.

For the purposes of this study, an uncertainty method is considered to perform well when it generates $x\%$ confidence intervals for estimated quantities that contain the true value approximately $x\%$ of the time. The method should generate neither too narrow intervals that underestimate uncertainty, nor too wide intervals that overestimate uncertainty.

The delta method was introduced by Cramér (1946) and popularized in ecological modelling by Seber (1973). Most applications of the delta method in stock assessment (e.g. Booth and Quinn 2006; Trzcinski et al. 2006; McGarvey et al. 2007) use the AD Model Builder programming framework to automate the computation of the required partial derivatives (Schnute et al. 1998; Fournier et al. 2012). The bootstrap was introduced by Efron (1979) and popularized by Efron and Tibshirani (1993). Early applications of the bootstrap in stock assessment include Mohn (1993) and Punt and Butterworth (1993). Variations of the bootstrap are outlined by Patterson et al. (2001), citing Gavaris and Van Eeckhaute (1998) as the current recommended bootstrap method for stock assessment. MCMC simulation of probability distributions was introduced by Metropolis et al. (1953) and Hastings (1970), and popularized in fisheries circles by Gelman et al. (1995). The potential usefulness of MCMC in stock assessment was described by McAllister and Ianelli (1997) and Punt and Hilborn (1997), with early applications including Punt and Kennedy (1997), Virtala et al. (1998), and Patterson (1999).

Patterson et al. (2001) list five desirable properties of methods quantifying uncertainty. They should be (i) based on statistical distributions derived from data rather than arbitrarily chosen distributions, (ii) unbiased, (iii) accurate, (iv) use few distributional assumptions and be robust to misspecifications of such assumptions, and least importantly (v) easy to

understand and implement. They mention that the bootstrap and MCMC have become more common than the delta method in fisheries stock assessment to avoid restrictive distributional assumptions. Hilborn (2003) noted that the use of the bootstrap has faded in recent years, as Bayesian methods have grown in popularity, because of their intuitive probability statements and theoretical and technical progress in this field of computational statistics. The bootstrap has been described as an automatic processor for frequentist inference, with MCMC as its Bayesian counterpart (Efron 2000).

2.1.2 Previous comparison studies

There are mainly two approaches to compare the performance of uncertainty methods, either using real stock assessment data or using simulated data. With real data, one can compare the estimated uncertainty for each method and speculate why differences occur. With simulated data, one knows the true value of the estimated quantities and can therefore quantitatively judge the performance of each method. A simulation study can use a relatively complex operating model to generate the simulated datasets and a simpler assessment model to fit those datasets, or use the same model to violate fewer assumptions.

Mohn (1993) compared the delta method and bootstrap, fitting an age-structured model to actual cod data. Retrospective analysis was used to approximate the true estimated values, showing that the delta method tended to underestimate uncertainty. Gavaris (1999) also compared the delta method and the bootstrap, fitting an age-structured model to haddock data. The bootstrap distribution indicated skewed uncertainty about stock abundance, implying that the delta method with a symmetric Gaussian distribution would be inappropriate for statistical inference. Patterson (1999) compared the bootstrap and MCMC, fitting an age-structured model to herring data and noted that MCMC generated wider confidence intervals than the bootstrap. Gavaris et al. (2000) compared the delta method, the bootstrap, and MCMC and analyzing data from three stocks using two age-structured models. The uncertainty methods gave somewhat different results, but no clear or consistent trends emerged. Booth and Quinn (2006) compared the delta method and MCMC, fitting a

simple age-structured model to monkfish data. The two methods gave similar results when non-informative Bayesian priors were used for MCMC, and the study highlighted how prior information can be incorporated to decrease uncertainty when using MCMC. Mohn (2009) compared the delta method, bootstrap, and MCMC, fitting an age-structured model to cod data. The bootstrap generated considerably wider confidence intervals than the delta method and MCMC, and the author pointed out that the bootstrap might be overestimating measurement uncertainty.

Fewer studies have used simulated data to compare the performance of uncertainty methods. Punt and Butterworth (1993) compared the delta method and the bootstrap, using an age-structured operating model and a simpler biomass-dynamic assessment model. The methods worked equally well, as long as some bootstrap pitfalls were avoided. Restrepo et al. (2000) compared the delta method, bootstrap, and MCMC, fitting age-structured assessment models to a simulated dataset. The delta method and bootstrap performed marginally better than MCMC in their study, and bias-correction methods proved beneficial.

2.1.3 This study

Overall, previous comparison studies have not identified which uncertainty method performs best. They have highlighted the strengths and weaknesses of each method and provided useful recommendations regarding their implementation. This study revisits the question with previous recommendations in mind, using a modern statistical catch-at-age model both to simulate and to analyze data that are known to be informative (Magnusson and Hilborn 2007). The study also benefits from greater computing power than was available a decade ago, allowing a more rigorous experimental design that involves a larger number of simulated datasets and population trajectories.

The working hypothesis is that all three methods work perfectly, for example, that 90% confidence intervals for a reference point contain the true value 90% of the time. This hypothesis is not going to be accepted or rejected, but the delta method, bootstrap, and MCMC will be rated in terms of how accurate the probabilistic statement is.

2.2 Methods

First, we define a set of true population parameters and generate stochastic datasets, using an operating model based on age-structured population dynamics. The performance of three uncertainty methods is then evaluated, with respect to how accurately they report the uncertainty about reference points. The simulation procedure (Figure 2.1) is repeated 1000 times, using 10 different recruitment scenarios so the results do not depend on a particular population trajectory. The operating model first outputs the resulting reference point values, and then applies random observation noise to the assessment data that are used as input for the estimation model. Finally, the confidence interval for each reference point is evaluated using the delta method, bootstrap, and MCMC, and compared with the 'true' reference points.

2.2.1 Operating model

The operating model is age-structured and follows the parametrization of the Coleraine generalized population model (Hilborn et al. 2003). The population dynamics of this operating model are described in detail by Magnusson and Hilborn (2007). There are 10 age classes, including a plus group, and 20 years of data, nominally referred to as 1985–2004, and the biology and fishery characteristics (see Appendix, Figure A.1, Tables A.1 and A.2) are based on Atlantic cod (*Gadus morhua*, Gadidae).

Each dataset includes landings, a survey abundance index, commercial catch at age and survey catch at age. The landings are assumed to be known exactly, but the commercial and survey catch-at-age data and the abundance index are subject to random observation error. The datasets are based on 10 recruitment scenarios that are generated randomly (Table A.3), and within each scenario there are 100 stochastic datasets with different realizations of observation noise. The level of recruitment variability (lognormal $\sigma_R = 0.6$), observation noise for the abundance index (lognormal $\sigma_I = 0.2$), and observation noise for the commercial (multinomial ${}_Cn = 50$) and survey catch-at-age (multinomial ${}_Sn = 50$) are similar to those used in assessments of Icelandic cod (ICES 2003). All the scenarios follow the same harvest rate

schedule, but the recruitment pattern leads to 10 different landings and biomass trajectories (Figure 2.2).

The survey abundance index is proportional to the biomass vulnerable to the survey in the middle of the fishing year,

$$I_t = q \sum_a {}_S S_a N_{t,a} w_a \, e^{-M/2} \times \exp({}_I\varepsilon_t) \tag{2.1}$$

where I_t is the observed abundance index at time t, q is the catchability coefficient, ${}_S S_a$ is survey selectivity at age a, $N_{t,a}$ is population size, w_a is body weight, M is the natural mortality rate, and ${}_I\varepsilon_t \sim \mathrm{N}(0, \sigma_I^2)$ is observation noise. The commercial catch-at-age data are provided to the assessment model in the form of proportions at age. These proportions are generated assuming that sampling is multinomial,

$${}_C P_{t,a} \sim \mathrm{Multinom}\left({}_C n, \frac{{}_C S_a N_{t,a}}{\sum_a {}_C S_a N_{t,a}}\right) \Big/ {}_C n \tag{2.2}$$

where ${}_C P_{t,a}$ is the observed catch at age and ${}_C n$ is the multinomial sample size. Survey catch-at-age data are generated in the same way.

2.2.2 Estimation model

The estimation model is a statistical catch-at-age model (Fournier and Archibald 1982) implemented using Coleraine and has the same parametrization as the operating model. It would therefore fit the data perfectly, if it was not for the observation noise both in the survey abundance index and in the commercial and survey catch-at-age data. The parametrization allows the commercial selectivity curve to decline at the oldest ages, but the survey selectivity curve is correctly assumed to be asymptotic. Some of the estimated parameters, including natural mortality rate M, stock-recruitment steepness h, and declining right-hand commercial selectivity are known to be correlated and problematic to estimate (Magnusson and Hilborn 2007). Wide bounds (Table A.2) are assigned to all estimated parameters so as

not to impose any major constraints on the parameter values. The estimation model is given the correct (i.e. operating model) value for recruitment variability, $\sigma_R = 0.6$.

The objective function for the estimation model is the sum of four components. The first three relate to observed data, and the last component is a penalty on deviations from Beverton-Holt recruitment. The abundance index is assumed to be lognormally distributed, the robust normal likelihood for proportions (Fournier et al. 1990) is assumed for the commercial and survey catch-at-age data, while the recruitment deviates are assumed to be lognormal. The magnitude of observation error for the abundance index is estimated using maximum likelihood, while the effective sample sizes for the commercial and survey catch-at-age data are estimated using the approach of McAllister and Ianelli (1997). The same age-composition sample size is assumed for all years, calculated as the median of estimated annual effective sample sizes.

2.2.3 Reference points

Six reference points are evaluated as potential management quantities of interest: B_{current} (current spawning biomass), u_{current} (current harvest rate), Depletion (depletion level, B_{current} relative to virgin spawning biomass), maximum sustainable yield (MSY), $B_{\text{current}}/B_{\text{MSY}}$ (B_{current} relative to B_{MSY}) and Surplus (current surplus production). These reference points describe the current stock status and potential yield, and are described in detail by Magnusson and Hilborn (2007). MSY and B_{MSY} are defined as the long-term average catch and spawning biomass when the harvest rate is set to an optimal value, u_{MSY}. Surplus production is defined as the last year's catch, plus the resulting change in vulnerable biomass.

The true reference point values from the operating model vary between recruitment scenarios (Table 2.1), except u_{current} that is predefined (Figure 2.2, Table A.3), and MSY that depends only on R_0, h, M, and commercial selectivity. The true MSY value is in all cases 203 thousand t, with harvest rate $u_{\text{MSY}} = 0.154$ and spawning biomass $B_{\text{MSY}} = 1270$ thousand t.

2.2.4 Evaluating uncertainty

The three methods used to quantify uncertainty start with the same input, the simulated datasets. Equation (2.3) summarizes how each method generates a probability distribution that is used to construct confidence intervals,

$$
\begin{aligned}
y &\xrightarrow[\text{delta}]{\text{model}} \hat{\theta},\, \widehat{\text{SE}}_{\hat{\theta}} \xrightarrow{\text{Norm}} p(y|\theta) \\
y &\xrightarrow{\text{model}} \hat{\theta} \xrightarrow{\text{bootstrap}} y_1^\star, y_2^\star, \ldots, y_B^\star \xrightarrow{\text{model}} \hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star \xrightarrow[\text{bias corr}]{\text{density}} p(y|\theta) \\
y &\xrightarrow[\text{MCMC}]{\text{model}} \theta_1, \theta_2, \ldots, \theta_T \xrightarrow{\text{density}} p(\theta|y)
\end{aligned}
\tag{2.3}
$$

where y denotes the observed data, θ is a vector of parameters (and derived quantities), the $\hat{}$ symbol indicates an estimate of a parameter or derived quantity, $\widehat{\text{SE}}_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$, $y_b^\star$ is a bootstrap dataset, $\hat{\theta}_b^\star$ is a bootstrap estimate, and θ_t is an MCMC iteration. The sampling distribution $p(y|\theta)$ and posterior distribution $p(\theta|y)$ are used to generate confidence intervals at any given confidence level.

Delta method

The estimation model uses automatic differentiation (Griewank and Corliss 1991; Fournier et al. 2012) to evaluate the Hessian matrix and hence the approximate variance-covariance matrix for the estimated parameters. The delta method (Seber 1973), which assumes that both estimation bias and the quadratic terms of the Taylor series are negligible, is then used to estimate the variance of each derived quantity,

$$
\widehat{\text{SE}}_{\hat{g}} = \sqrt{\sum_i \sum_j \widehat{\text{Cov}}\left(\hat{\theta}_i, \hat{\theta}_j\right)\left(\frac{\partial g}{\partial \theta_i}\right)\left(\frac{\partial g}{\partial \theta_j}\right)}
\tag{2.4}
$$

where g is a derived quantity, such as a reference point, that is a function of some estimated parameters $\theta_1, \theta_2, \ldots, \theta_n$. The symmetric confidence interval for g is then

$$\left[\hat{g} - z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}}, \quad \hat{g} + z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}}\right] \tag{2.5}$$

where z is the standard normal quantile.

The reference points B_{current} and MSY are log-transformed for the purpose of applying the delta method, because the uncertainty about these quantities can be expected to be closer to lognormal than normal (Mohn 1993; Patterson et al. 2001), and exploratory bootstrap and MCMC runs indicated that this was the case. Although surplus production is also measured in biomass units, it is not log-transformed, as exploratory results showed fairly symmetric distributions, and because surplus production can be negative when weak cohorts are entering the fishable stock.

Bootstrap

A parametric model-conditioned approach is used to generate 1000 bootstrap datasets for each simulated dataset. In their simulation study, Punt and Butterworth (1993) found that 100 bootstrap datasets was adequate for variance estimation, but 1000 bootstrap datasets are used here, because more replicates are needed to estimate quantiles than variance. The bootstrap is parametric with residuals sampled from estimated probability distributions, and model-conditioned in that the residuals are not applied to the observed data but to predictions from the model fit to the original data (Efron and Tibshirani 1993). The parametric bootstrap was chosen because it is probably what would be used in practice for this particular assessment model, as there is no straightforward way to resample residuals for the catch-at-age data when they are proportions. The bootstrap survey abundance index is

$$I_t^\star \;=\; \hat{I}_t \times \exp({}_I\varepsilon_t^\star), \quad {}_I\varepsilon_t^\star \sim \mathrm{N}(0, \hat{\sigma}_I{}^2) \tag{2.6}$$

where $I_t^\star$ is the bootstrap datum for year t, $\hat{I}_t$ is the predicted index for year t from the model

fit to the original dataset, ${}_I\varepsilon_t^\star$ are bootstrap residuals, and $\hat{\sigma}_I$ is the estimated magnitude of observation error. The bootstrap commercial catch at age is

$${}_CP_{t,a}^\star \sim \text{Multinom}\left({}_C\hat{n},\ {}_C\hat{P}_{t,a}\right)\Big/{}_C\hat{n} \tag{2.7}$$

where ${}_CP_{t,a}^\star$ are the bootstrap data, ${}_C\hat{n}$ is the estimated effective sample size, and ${}_C\hat{P}_{t,a}$ is the model-predicted commercial catch at age for year t.

The estimation model is fitted to each of the 1000 bootstrap datasets, resulting in 1000 bootstrap estimates for each parameter and derived quantity. A bias-correction factor is then applied, which has been shown to lead to more accurate confidence intervals (Efron 2003). In fisheries stock assessment, Gavaris and Van Eeckhaute (1998) and others have used the BC_a algorithm (bias correction and acceleration, Efron and Tibshirani 1993) with the acceleration coefficient set to zero. Acceleration relates to the rate of change of the standard error of $\hat{\theta}$ with respect to the true parameter value θ, so zero acceleration implies that the standard error of $\hat{\theta}$ is the same for all θ. The algorithm then simplifies to

$${}_{\text{BC}}\vec{\hat{\theta}}^\star = \hat{\Omega}^{-1}\Big[\Phi\Big(2\Phi^{-1}\big[\hat{\Omega}(\hat{\theta})\big] + \Phi^{-1}(\vec{\alpha})\Big)\Big] \tag{2.8}$$

where ${}_{\text{BC}}\vec{\hat{\theta}}^\star$ is a vector of bias-corrected bootstrap estimates in ascending order, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{\Omega}(x) = \#\{\hat{\theta}^\star < x\}/B$ is the empirical cumulative distribution function of the bootstrap estimates $\hat{\theta}^\star$, while $\Phi^{-1}(\cdot)$ and $\hat{\Omega}^{-1}(\cdot)$ are the corresponding inverse functions, B is the number of bootstraps, and $\vec{\alpha}$ is a vector of probability levels $1/B$, $2/B$, $\ldots$, B/B.

The bias-correction algorithm compares the bootstrap estimates of a given quantity to the original point estimate. If the median of the bootstrap estimates is above or below the original point estimate, it is seen as an indication of a biased estimator. As the original point estimate was subject to the same bias, the algorithm corrects for the bias by transforming the bootstrap estimates (Figure 2.3). The algorithm performs no transformation if the median of the bootstrap estimates is the same as the original point estimate. It is also worth noting that

the bias-corrected bootstrap estimates are always within the range of the 'raw' bootstrap estimates. The resulting confidence interval may be narrower or wider.

The algorithm fails in the rare case when the bias is so extreme that all the bootstrap estimates are above or below the original point estimate. In these cases, the $\hat{\Omega}(\hat{\theta})$ term, the proportion of bootstrap estimates that are below the original point estimate, is 0 or 1, resulting in expressions such as $\hat{\Omega}^{-1}[\Phi(-\infty+\infty)]$, which is not mathematically defined. To avoid this problem, a robust algorithm is used, where the $\hat{\Omega}(\hat{\theta})$ term is bounded between 0.1 and 0.9:

$$_{\text{BC}}\vec{\hat{\theta}}^{\star} = \hat{\Omega}^{-1}\Big[\Phi\Big(2\Phi^{-1}\big[\max\{0.1, \min\{0.9, \hat{\Omega}(\hat{\theta})\}\}\big] + \Phi^{-1}(\vec{\alpha})\Big)\Big] \tag{2.9}$$

These safety bounds guarantee a valid interval, but without the safety bounds, 5 of 6000 bias-corrected bootstrap intervals were undefined. The bias demonstrated in Figure 2.3 corresponds to $\hat{\Omega}(\hat{\theta}) = 0.84$, so the safety bounds at 0.1 and 0.9 are irrelevant for that example. The computer code for the robust bias-correction algorithm is provided in Section 4.3. The bias-corrected bootstrap confidence interval is calculated as:

$$\left[\frac{\alpha}{2} \text{ quantile from } {}_{\text{BC}}\vec{\hat{\theta}}^{\star}, \quad \left(1-\frac{\alpha}{2}\right) \text{ quantile from } {}_{\text{BC}}\vec{\hat{\theta}}^{\star}\right] \tag{2.10}$$

MCMC

Markov chain Monte Carlo simulation is used to approximate the posterior distribution of estimated parameters and reference points. The simulation method is Metropolis-Hastings with an adaptive multivariate normal jumping distribution (Gelman et al. 2004; Fournier et al. 2012).

All model parameters are assigned uniform priors based on their bounds (Table A.2, except the deviations about Beverton-Holt stock-recruitment relationship have a lognormal prior. The MCMC simulation is run for 1 million iterations and then thinned, keeping every 1000th iteration. Convergence of the estimated reference points is diagnosed using the

'coda' package (Plummer et al. 2006), adopting an autocorrelation threshold of 0.1, Geweke threshold of 1.96, and Heidelberger-Welch threshold of 0.05. If any criteria are not met, the MCMC chain is extended to a maximum of 10 million iterations, still thinning to 1000 iterations, to reduce autocorrelation and stabilize the distribution quantiles. This proved to be necessary for a few hundred model runs, owing to unstable model convergence as can be expected when simultaneously estimating correlated parameters such as natural mortality rate M, stock-recruitment steepness h, and declining right-hand selectivity (Magnusson and Hilborn 2007).

The MCMC confidence interval is calculated as

$$\left[\frac{\alpha}{2} \text{ quantile from } \theta_1, \theta_2, \ldots, \theta_T, \quad \left(1-\frac{\alpha}{2}\right) \text{ quantile from } \theta_1, \theta_2, \ldots, \theta_T\right] \tag{2.11}$$

where $\theta_1, \theta_2, \ldots, \theta_T$ are the iterations retained from the MCMC chain.

2.3 Results

We first examine the performance of confidence intervals at the 90% confidence level, then broaden the analysis to all confidence levels, and finally examine the sensitivity of the results to changes to assumptions. Results are shown for both bias-corrected and 'raw' (non-bias-corrected) bootstrap confidence intervals to evaluate whether and how much the bias correction improves the bootstrap performance.

2.3.1 90% confidence level

A total of 24 000 confidence intervals are analyzed at the 90% confidence level (four uncertainty methods, six reference points, 10 recruitment scenarios, and 100 stochastic datasets for each recruitment scenario). Before summarizing, it is useful to look at an example set of confidence intervals (Figure 2.4), where the uncertainty method is MCMC, the reference point is current surplus production, and the recruitment scenario is 10. For a 90% confidence level, one would expect around 90% of the confidence intervals to cover the true value (227

thousand t), so the coverage probability in this example, 93 of 100, is slightly higher than the nominal value of 90.

Looking across all uncertainty methods, reference points, and recruitment scenarios, the coverage probability is usually lower than the target (Table 2.2), with 216 of 240 combinations having coverage probabilities below 90. The example described previously, with a coverage probability of 93, can be found in the lower right-hand corner of Table 2.2. The coverage probabilities vary considerably between recruitment scenarios and the purpose of including ten scenarios instead of only one is to prevent the results from depending on a particular recruitment history.

The overall trends emerge after averaging over recruitment scenarios (Table 2.3), with the delta method, bootstrap, and MCMC all showing coverage probabilities < 90, that is, the methods lead to 90% confidence intervals that cover the true value $< 90\%$ of the time. Overall, the delta method and MCMC perform better than the bias-corrected bootstrap, with mean coverage probabilities of 73.0, 72.5 and 64.1, respectively. The performance of the bootstrap is considerably poorer before bias correction, with a mean coverage probability of 57.4. The delta method outperforms the other methods in evaluating the uncertainty about the current biomass, current harvest rate, depletion, and surplus production, but performs poorly for $B_{\text{current}}/B_{\text{MSY}}$. MCMC performs better than the delta method and bootstrap for MSY and $B_{\text{current}}/B_{\text{MSY}}$, and its mean coverage probability is above 60 for all reference points. The bias-corrected bootstrap has similar or lower coverage probabilities than the delta method and MCMC, including a particularly low coverage probability of 45.6 for MSY. Bias correction generally improves the bootstrap performance, although it reduces the coverage probability from 71.2 to 65.6 for $B_{\text{current}}/B_{\text{MSY}}$. On the other hand, bias correction leads to a substantial increase in coverage probability for current harvest rate, from 44.9 to 66.5.

2.3.2 All confidence levels

When the analysis is repeated for different confidence levels (Figure 2.5, Table A.4), the results confirm the trends for the 90% confidence level. The delta method, bootstrap, and

MCMC show coverage probabilities that are consistently lower than expected at all confidence levels. The general pattern is that the delta method and MCMC perform better than the bootstrap, the main exception being $B_{\text{current}}/B_{\text{MSY}}$, where the delta method performs considerably worse than MCMC and the bootstrap. The delta method performs slightly better than MCMC for current biomass, current harvest rate, and depletion at confidence levels higher than 50%, but the two methods perform equally well for MSY and surplus production. On the whole, the bias-corrected bootstrap performs poorer than the delta method and MCMC, particularly for MSY. Bias correction leads to improved performance of the bootstrap, with the exception of $B_{\text{current}}/B_{\text{MSY}}$, and the improvement is especially noticeable for current harvest rate. MCMC has the most consistent performance for the various reference points (Figure 2.5). Its coverage probability is always close to that for the best-performing method, and it shows no conspicuous failures, unlike the bootstrap for MSY and the delta method for $B_{\text{current}}/B_{\text{MSY}}$.

When the results are averaged across the six reference points (Fig. 2.6), the delta method and MCMC show similar performance, substantially better than the bias-corrected bootstrap. At the 50% confidence level, MCMC has a mean coverage probability of 38, the delta method has 35 and the bootstrap 30. At the 95% confidence level, the delta method and MCMC have a mean coverage probability of 80, while the bootstrap has 71 (Figure 2.6, Table A.4).

2.3.3 Sensitivity analysis

Four analyses are used to examine what factors may lead to the low coverage probabilities (Figure 2.6). The first analysis assumes that the estimation method 'knows' the true magnitude of observation error, the second analysis uses a multinomial catch-at-age likelihood, the third assumes that the estimation method 'knows' the bias of estimated reference points, and the fourth combines all of the above. These analyses are only conducted for the delta method owing to computational demands. Finally, a supplementary sensitivity analysis (Table A.5) indicates that the overall results would not change very much if more recruitment scenarios

would be included in the study.

Known magnitude of observation error

The observation noise in the simulated datasets is generated using lognormal $\sigma_I = 0.2$ and multinomial ${}_Cn = 50$ and ${}_Sn = 50$, but this magnitude of observation error is often underestimated by the estimation model. The iteratively estimated $\hat{\sigma}_I$ is often too low, the median estimate being 0.186, while ${}_C\hat{n}$ and ${}_S\hat{n}$ are often too high, with median estimates 52 and 53. This leads to narrower confidence intervals, which could explain the low coverage probabilities. When the estimation model uses the true values for σ_I, ${}_Cn$ and ${}_Sn$, the coverage probability of the delta method improves only marginally, from 72.9 to 74.5 at the 90% confidence level (Figure 2.7, left panel).

Multinomial catch-at-age likelihood

The operating model generates catch-at-age data under the assumption of multinomial sampling, but the estimation model uses the Fournier robust normal likelihood for proportions. This introduces a model misspecification, which could explain the low coverage probabilities. When the estimation model assumes a multinomial catch-at-age likelihood instead of the robust normal likelihood for proportions, the coverage probability of the delta method improves noticeably, from 72.9 to 78.4 at the 90% confidence level (Figure 2.7, center panel).

Known bias

Each reference point is estimated with some bias. The median of the 1000 point estimates compared with the true value, $\text{median}\,(\hat{\theta} - \theta)/\theta$, is -0.13 for current biomass, $+0.22$ for current harvest rate, -0.20 for current depletion level, $+0.20$ for MSY, $+0.05$ for $B_{\text{current}}/B_{\text{MSY}}$, and $+0.14$ for current surplus production. When the delta-method confidence intervals are shifted to correct for the median bias of each reference point, the coverage probability improves noticeably, from 72.9 to 78.8 at the 90% confidence level (Figure 2.7, right panel).

Combined effect

When the estimation model assumes a multinomial catch-at-age likelihood, given the true values for σ_I, ${}_C n$ and ${}_S n$, and the confidence intervals are then shifted to correct for the median bias of each reference point, the coverage probability improves considerably, from 72.9 to 82.6 at the 90% confidence level.

2.4 Discussion

2.4.1 Confidence intervals are too narrow

The delta method, bootstrap, and MCMC all produced confidence intervals that did not cover the true value as often as the nominal confidence level implies (Figure 2.6). The three methods are well established in the statistical literature, widely used, and have been shown to perform well for simple models, when all assumptions are met (Seber 1973; Efron and Tibshirani 1993; Gelman et al. 2004). The purpose of this study was to examine how well they perform with a typical stock assessment model of medium complexity, when most assumptions are met. Generally, the performance of all statistical methods degrades with increased model complexity, as non-linearity and correlated parameter estimates undermine the assumptions and lead to estimation bias (Seber and Wild 1989). The optimization method also becomes less likely to find the global minimum. In this study, sensitivity analysis showed that even after correcting for estimation bias, the delta-method 90% confidence intervals still covered the true value $<80\%$ of the time.

The expectation was that the methods would show some inaccuracy, because of model complexity, but not necessarily that the confidence intervals would be too narrow. It would seem, a priori, just as possible that some of the methods might generate confidence intervals that covered the true value more often than the nominal confidence level implies.

This study is based on a scenario that is known to be informative (Magnusson and Hilborn 2007), where the stock is first fished down and then allowed to rebuild (Figure 2.2), with standardized surveys and age data from the start of the fishery. Furthermore, the

data are generated using the same dynamics as the estimation model, and landings, body weight, maturity, and recruitment variability are known without error. When analyzing real data, we can expect model error and process variability to lead to considerably more bias, and therefore, confidence intervals that are even less likely to cover the true value. The notion that statistical methods in stock assessment tend to underestimate the real extent of uncertainty is reflected in the literature (Hilborn and Mangel 1997; Punt and Hilborn 1997; Patterson et al. 2001; Gavaris and Ianelli 2002) and demonstrated here in a setting where one would expect the methods to perform well. In a recent meta-analysis of multiple assessment models fitted to 17 stocks off the United States West Coast, Ralston et al. (2011) found that model specification error can be expected to be considerably greater than the estimation error.

2.4.2 Delta method and MCMC perform better than bootstrap

The delta method and MCMC provided better confidence intervals than the bootstrap on average (Figure 2.6). For example, at the 90% confidence level, the delta-method intervals covered the true reference point value 73.0% of the time, MCMC 72.5% and bias-corrected bootstrap 64.1%. Although the intervals from all three methods were generally too narrow, the delta method and MCMC were considerably closer to attaining the nominal confidence level.

It is somewhat surprising to see how well the delta method performed, compared with the bootstrap and MCMC. Automatic differentiation (Griewank and Corliss 1991; Fournier et al. 2012) facilitates the use of the delta method with complex models, where derived quantities are not simple functions of estimated parameters, by applying automated algorithms to compute the partial derivatives. In application, the delta method is orders of magnitude faster than the computationally intensive bootstrap and MCMC methods, which can be a major advantage for iterative simulations, complex models, and/or large datasets (Maunder et al. 2009).

The delta method has been shown to perform about as well as the bootstrap for stock

assessment (Punt and Butterworth 1993; Restrepo et al. 2000), or slightly worse (Mohn 1993; Gavaris 1999). A simulation study comparing the delta method, bootstrap, and MCMC (Restrepo et al. 2000) found that the delta method and bootstrap performed about as well, but MCMC performed poorer. The present study's ranking of the uncertainty methods is therefore quite different from the results of previous simulation studies. The contradictory results are most likely due to model differences; the previous studies used relatively simple biomass-dynamic models and ADAPT, with fewer estimated parameters, fewer objective function components, and more restrictive assumptions than the statistical catch-at-age model used here. Variations in the implementation of the delta method, bootstrap and MCMC can also affect their performance (Patterson et al. 2001; Gelman et al. 2004; Givens and Hoeting 2005). Finally, the studies vary in terms of whether they compare confidence intervals or variance estimates, and whether they focus on the uncertainty about model parameters or reference points.

2.4.3 Bias correction improves bootstrap performance

Overall, the bootstrap performed considerably better with bias correction than without it (Figure 2.6), with the mean coverage probability at the 90% confidence level increasing from 57.4 to 64.1. This shift of 6.7, compared with a shift of 32.6 that would bring it to ideal performance, amounts to around 20% improvement. This performance improvement did not, however, apply uniformly across all reference points (Figure 2.5), ranging from particularly beneficial for current harvest rate and depletion, to slightly detrimental for MSY and $B_{\text{current}}/B_{\text{MSY}}$.

The estimation of current harvest rate was subject to greater bias than the other reference points. It is therefore reassuring to see that bias correction was most beneficial for that reference point, effectively correcting for the positive bias. It is also reassuring to see a similar performance gain for negatively biased reference points, such as current depletion. When bias correction does not lead to improved performance, it is because the perceived bias, that is, the difference between the bootstrap estimates and the point estimate, does

not reflect the true estimation bias.

This study evaluates the performance of the BC_a bias-correction method for the bootstrap using zero acceleration. Alternative approaches include ABC (approximate bootstrap confidence), ABCq, and various ways to estimate the BC_a acceleration coefficient (Efron and Tibshirani 1993, DiCiccio and Efron 1996). The implementation used here is recommended in the current fisheries stock assessment literature (Patterson et al. 2001; Gavaris and Ianelli 2002), and the results from this study support that recommendation.

2.4.4 Other findings

Why did the bootstrap perform so poorly for MSY? This reference point was positively biased, mainly due to a positive bias in the estimated stock-recruitment steepness parameter h. This bias can be expected in statistical catch-at-age models when the stock-recruitment steepness is estimated, unless the data include years of extremely low abundance, and the natural mortality rate and selectivity of older fish are known (Magnusson and Hilborn 2007; Conn et al. 2010). Despite this bias, the delta method and MCMC performed very well for MSY, providing confidence intervals of appropriate width at any given confidence level (Figure 2.5).

The delta method showed unusually low coverage probability for $B_{\text{current}}/B_{\text{MSY}}$, compared with the other reference points. The most likely reason for this is that the assumption of symmetric Gaussian uncertainty is not appropriate for this ratio statistic. There are many transformations that could be used for each reference point, and it is beyond the scope of this study to explore all possibilities. Logarithm and square-root transformation are ruled out if the original quantity can be negative, such as, surplus production, as are logit and probit transformation when the quantity can exceed 1.0, such as depletion level and $B_{\text{current}}/B_{\text{MSY}}$. Transforming reference points has an important effect on the performance of the delta method, but transforming model parameters can improve the performance of the bootstrap and MCMC as well (Efron and Tibshirani 1993; Gelman et al. 2004). The use of statistical transformations in stock assessment models is a topic worthy of further

investigations.

The sensitivity analysis showed that the estimation model performed noticeably better when multinomial likelihood was used for catch at age, instead of the default Fournier robust normal likelihood for proportions. As the operating model uses multinomial random draws to generate the catch at age data, this sensitivity test quantifies the model error introduced by likelihood misspecification. The Fournier likelihood is designed to be more robust than the multinomial likelihood when observed data are subject to greater variability than statistical theory predicts (Fournier et al. 1990) and has been shown to perform well when that is the case (Ernst 2002). This study, on the other hand, shows that the Fournier likelihood does not perform as well as the multinomial likelihood when the data are random draws from the multinomial distribution. The Fournier likelihood is not a generalization of the multinomial that allows greater variance, but rather a hybrid between normal and multinomial that explicitly downweights two kinds of outliers: ages with few observations and predictions that are far from the observations. We recommend using the Fournier likelihood to analyze real fisheries data, and use it in this study to represent a typical estimation model in stock assessment.

The additional analyses also examined the impact of biased reference points, and how much of the total error is because of bias, as opposed to too narrow confidence intervals. Magnusson and Hilborn (2007) described what kinds of biases can be expected when estimating reference points, depending on the fishing history, model assumptions, and available data. As the fishing history and estimation model analyzed here were selected from that study, the biases were known beforehand. When these biases were corrected for each reference point, the performance of the delta method improved about one-third towards ideal performance. When analyzing real fisheries data, the total error cannot be partitioned in this way, because bias can only be evaluated when the true value is known.

2.4.5 Recommendations

The overall performance trends suggest that MCMC is the most reliable of the three uncertainty methods, given the dataset and catch-at-age assessment method. Both the delta method and the bootstrap performed poorly for one or more reference points, while MCMC was always close to the best-performing method. When time and resources allow, we recommend using more than one method to evaluate uncertainty, to see whether they lead to markedly different conclusions. If only one method is to be used, it seems that MCMC is the least likely to severely misrepresent uncertainty. All three methods, however, have a strong tendency to underestimate the uncertainty.

On the average, the delta method performed well compared with the computationally intensive bootstrap and MCMC methods and can be recommended for quick evaluation of uncertainty while exploring a variety of modelling options, before applying the bootstrap and/or MCMC to selected model runs. The delta method may also be useful when confidence intervals are required in a large number of simulations. In this study, the delta-method calculations were around 1000 times faster than the bootstrap and MCMC. Management strategy evaluation (Butterworth and Punt 1999; De Oliveira et al. 2008) is a common application where this can be relevant. Possible transformations of model parameters and reference points should be explored when using the delta method.

One advantage of the bootstrap is that it can detect bias in the estimation model, thus providing valuable diagnostic information for the modeller (Haddon 2003). We recommend applying bias correction when using the bootstrap, having seen around 20% overall performance improvement in this study. That said, the bootstrap was generally outperformed by the delta method and MCMC, in spite of bias. It would be interesting to see a similar performance comparison of uncertainty methods where the estimators are more biased than here. Another potential advantage of the bootstrap is that computations can be split into parallel threads, thus taking less time than computing a very long MCMC chain.

Each method for evaluating uncertainty is based on a particular set of assumptions. It

appears that the choice between frequentist methods, such as the delta method and bootstrap, and Bayesian methods, such as MCMC, is not the most important decision for the modeller. In this study, for example, the overall performance of the delta method and MCMC was more similar than that of the bootstrap. It is at least as important that the modeller considers, and preferably tests, the sensitivity of the results to specific assumptions within a method. The effects of different transformations for the delta method have been discussed earlier, and Patterson et al. (2001) describe several bias-correction methods. Choices within the bootstrap include parametric vs. nonparametric, model-conditioned vs. non-conditioned, and a variety of bias-correction methods (Efron and Tibshirani 1993). In Bayesian inference, the choice of prior distributions can be important, and different algorithms to approximate the posterior probability have their strengths and weaknesses (Gelman et al. 2004). The same estimation model can often be expressed as frequentist or Bayesian with few or no modifications, as is done in this study. The trend in the current statistical literature (e.g. Kass 2011) has been to deflate the frequentist-Bayesian debate, focusing instead on the assumptions that relate models to data. When frequentist and Bayesian procedures do lead to very different conclusions, the choice should primarily be based on their performance with simulated data.

Although we have limited the analysis to the delta method, bootstrap, and MCMC, other methods can also be used to evaluate uncertainty in stock assessment. Sampling-importance resampling (SIR) can be used to simulate Bayesian posterior distributions instead of MCMC, but both methods should lead to the same distribution if run long enough (Gelman et al. 2004). When stock assessment models include more than a dozen parameters, MCMC is more computationally efficient than SIR (McAllister et al. 1994; Punt and Hilborn 1997). Profile likelihood (Edwards 1992; Hilborn and Mangel 1997) is a straightforward method to evaluate the uncertainty about estimated parameters, but it is problematic to generate the profile likelihoods for derived quantities such as reference points and future projections. Finally, adjunct Monte Carlo can be used to diagnose the consequences of changing the value of a fixed parameter, such as natural mortality rate or the shape of a stock-recruitment function (Patterson et al. 2001).

The main limitation of a study such as this one is that the conclusions are based on one particular estimation model and one artificial suite of data. Our goal in the experimental design was to use a typical stock assessment model of medium complexity, with generic groundfish data scenarios that are known to be rather informative (Magnusson and Hilborn 2007). Many stock assessments use simpler or more complex models that are conceptually and analytically related to the statistical catch-at-age model used here. In cases where the data and models are fundamentally different from what we used, perhaps involving species interaction or migration between areas, we can recommend using this study's simulation framework to investigate the performance of candidate uncertainty methods.

Acknowledgements

This research was supported by the Fulbright Program and the American-Scandinavian Foundation. We thank Jim Bence, Jim Ianelli, Steve Martell, Mark Maunder, Anders Nielsen, Jon Schnute, John Skalski, and two anonymous reviewers for insightful discussions and comments that improved the manuscript.

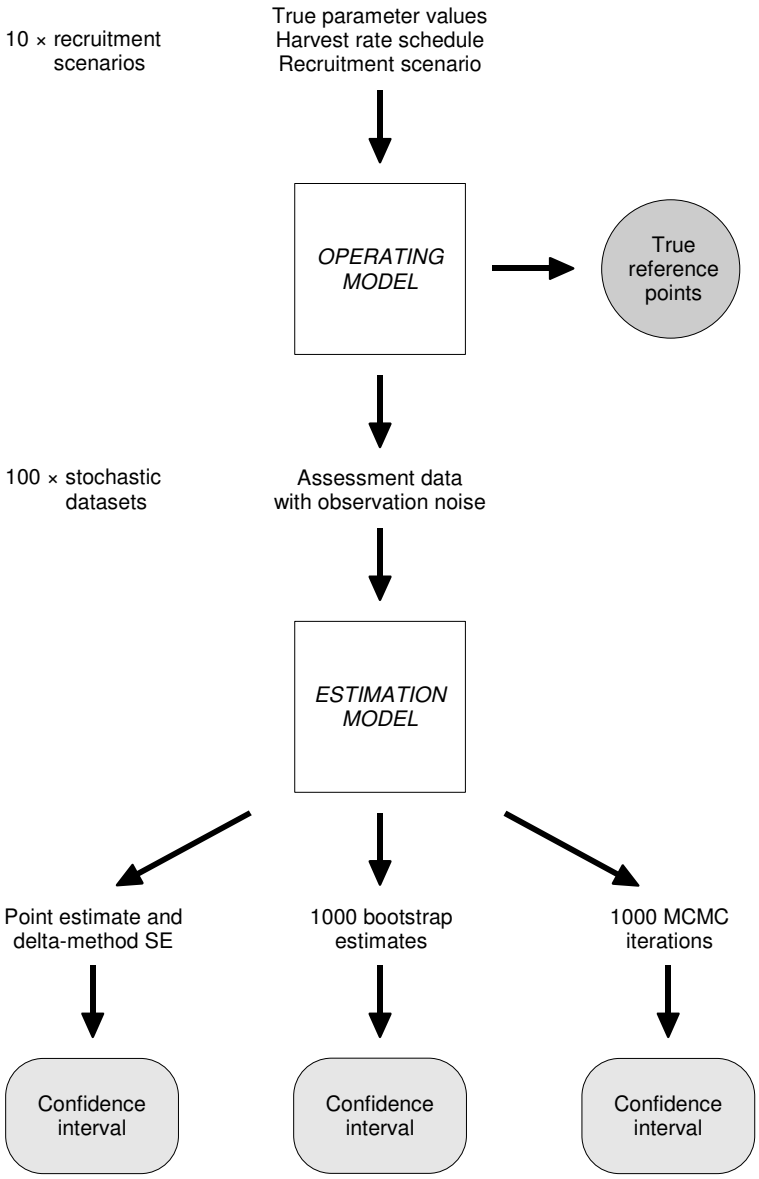

Figure 2.1. The simulation procedure. Arrows indicate the process for a single run, and replications indicate how the study consists of multiple runs.

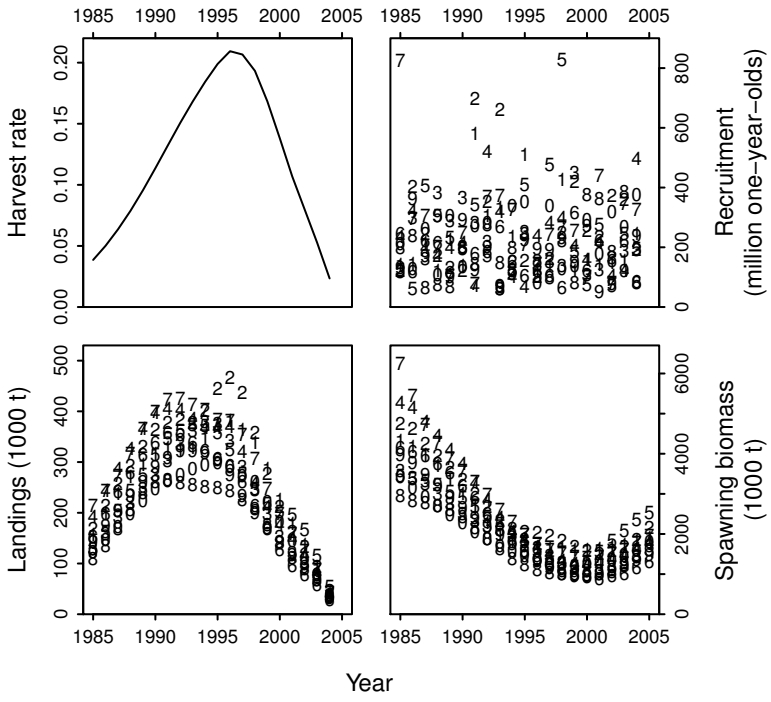

Figure 2.2. Harvest rate, recruitment, landings, and biomass in the operating model. The plotting symbols identify recruitment scenarios 1–10.

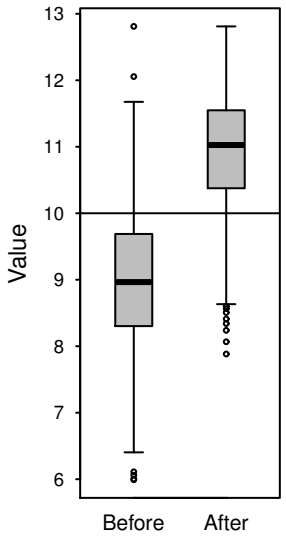

Figure 2.3. Effect of bias correction on bootstrap estimates. In this hypothetical example, the bootstrap estimates (left boxplot) are lower than the point estimate from the original data (horizontal line). The resulting bias-corrected bootstrap estimates (right boxplot) take into account that the original point estimate was subject to the same bias.

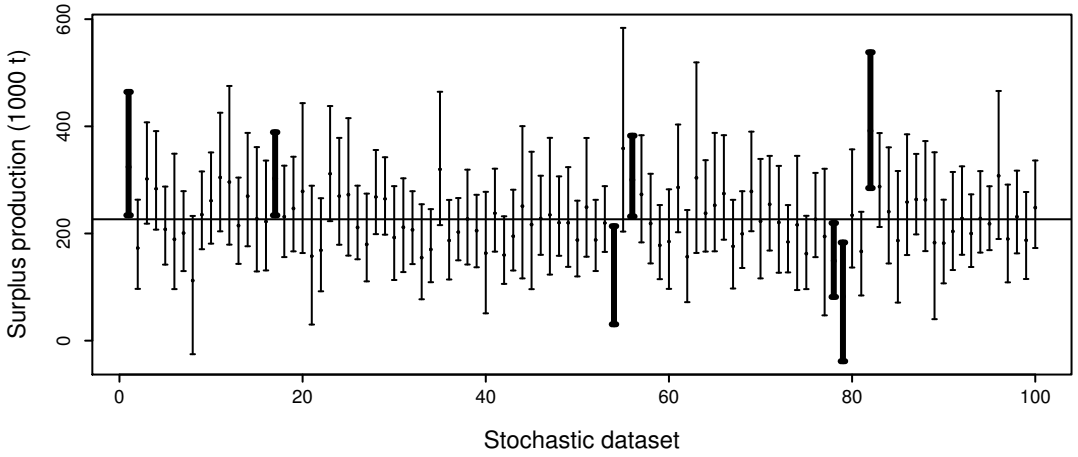

Figure 2.4. Example results, showing 90% confidence intervals for surplus production, from Markov chain Monte Carlo analysis of 100 stochastic datasets for recruitment scenario 10. Seven confidence intervals (thick lines) of one hundred do not cover the true value (horizontal line). In this example, the coverage probability is 93.

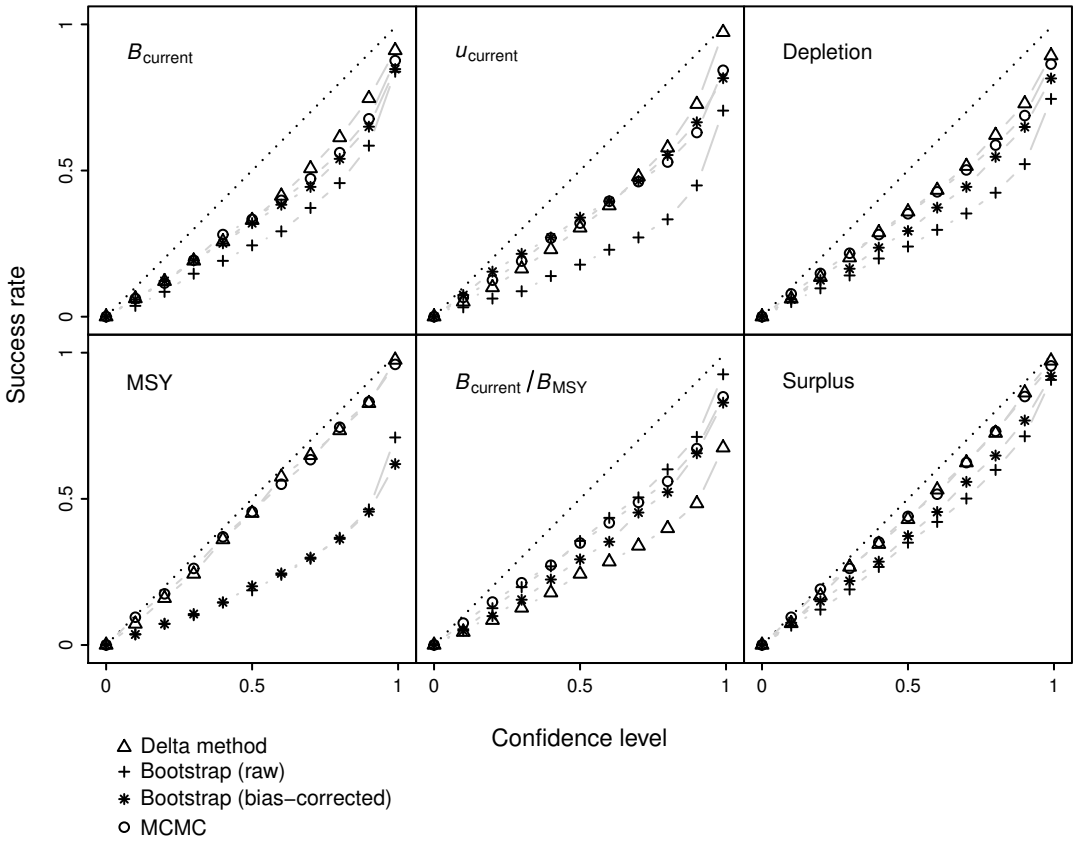

Figure 2.5. Coverage probability for confidence intervals by uncertainty method and reference point, evaluated at several confidence levels (0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 99%).

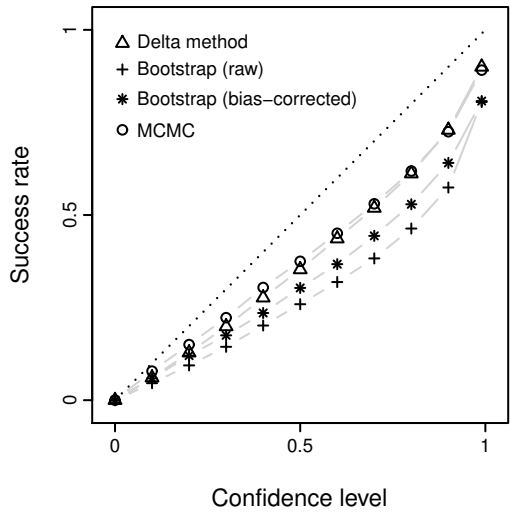

Figure 2.6. Coverage probability for confidence intervals for each uncertainty method averaged across all six reference points, evaluated at several confidence levels (0, 10, 20, 30, 40, 50, 60, 70, 80, 90 and 99%).

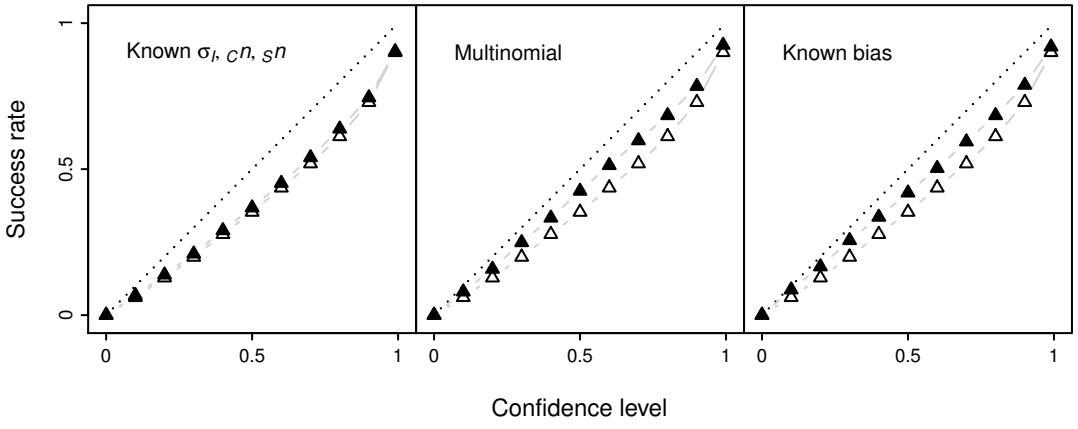

Figure 2.7. Coverage probabilities for the sensitivity tests. White triangles indicate the base case delta method (same as Figure 2.6), and black triangles indicate the outcomes of each sensitivity test.

Table 2.1. True reference point values from the operating model for each recruitment scenario. B_{current}, MSY, and Surplus are expressed in thousands of tonnes. u_{current} and MSY are 0.023 and 203, respectively, for all 10 recruitment scenarios.

	Recruitment scenario									
Reference point	1	2	3	4	5	6	7	8	9	10
B_{current}	1904	2156	1611	1793	2537	1318	1960	1704	1484	1802
Depletion	0.479	0.543	0.405	0.451	0.639	0.332	0.493	0.429	0.374	0.454
$B_{\text{current}}/B_{\text{MSY}}$	1.499	1.697	1.268	1.411	1.997	1.038	1.543	1.341	1.168	1.418
Surplus	83	315	158	164	166	198	245	300	84	227

Table 2.2. Coverage probability for 90% confidence intervals for each uncertainty method, reference point, and recruitment scenario. The non-bias-corrected bootstrap is referred to as 'raw' and bias-corrected bootstrap as 'bootstrap'. Ideally, the coverage probability at this confidence level should be 90.

	Reference	Recruitment scenario									
Method	point	1	2	3	4	5	6	7	8	9	10
Delta	B_{current}	80	70	75	77	73	80	78	70	72	72
	u_{current}	78	71	73	74	68	78	73	70	71	71
	Depletion	79	54	66	74	77	79	67	83	70	80
	MSY	72	19	100	96	94	99	61	99	100	87
	$B_{\text{current}}/B_{\text{MSY}}$	37	54	51	47	42	50	63	39	57	44
	Surplus	70	95	90	86	81	94	89	93	71	94
Raw	B_{current}	71	45	65	58	57	78	42	59	49	61
	u_{current}	52	28	53	44	46	71	18	52	36	49
	Depletion	61	17	46	58	62	79	5	76	50	68
	MSY	20	1	49	47	28	100	6	98	78	37
	$B_{\text{current}}/B_{\text{MSY}}$	85	62	49	75	72	70	78	73	65	83
	Surplus	44	94	82	67	67	86	75	83	20	96
Bootstrap	B_{current}	62	66	66	66	61	68	74	69	61	57
	u_{current}	65	65	69	72	60	78	62	77	59	58
	Depletion	67	63	71	73	64	69	39	71	73	59
	MSY	30	6	73	56	42	58	27	76	59	29
	$B_{\text{current}}/B_{\text{MSY}}$	61	71	74	66	63	63	56	57	87	58
	Surplus	68	77	81	78	70	80	79	85	67	83
MCMC	B_{current}	66	72	61	71	66	69	68	66	69	69
	u_{current}	63	64	55	66	59	67	66	62	65	63
	Depletion	73	72	66	74	71	64	53	61	71	83
	MSY	81	30	89	98	79	98	78	93	99	87
	$B_{\text{current}}/B_{\text{MSY}}$	70	84	66	78	62	53	76	36	74	73
	Surplus	75	94	90	88	78	90	85	91	66	93

Table 2.3. Coverage probability for 90% confidence intervals for each uncertainty method and reference point, averaged across recruitment scenarios. The non-bias-corrected bootstrap is referred to as 'raw' and bias-corrected bootstrap as 'bootstrap'. Ideally, the coverage probability at this confidence level should be 90.

	Delta	Raw	Bootstrap	MCMC
B_{current}	74.7	58.5	65.0	67.7
u_{current}	72.7	44.9	66.5	63.0
Depletion	72.9	52.2	64.9	68.8
MSY	82.7	46.4	45.6	83.2
$B_{\text{current}}/B_{\text{MSY}}$	48.4	71.2	65.6	67.2
Surplus	86.3	71.4	76.8	85.0
Average	73.0	57.4	64.1	72.5

Chapter 3

CONFRONTING UNCERTAINTY IN FISHERIES STOCK ASSESSMENT

Abstract

Fisheries stock assessment and the resulting management advice is subject to considerable uncertainty, and it is important to incorporate this uncertainty in the advice to quantify the risk of undesired outcomes. A variety of statistical methods and approaches exist to evaluate uncertainty. The objective of this study is to give an overview of commonly used methods and to summarize their performance, based on simulations and actual fisheries data. The benchmark comparison of uncertainty methods involves the delta method, profile likelihood, bootstrap, and Bayesian Markov chain Monte Carlo (MCMC) analysis. Approaches to evaluate uncertainty also include comparing alternative estimation models, retrospective analysis, quantifying the amount of information in each data component, and evaluating whether a dataset is likely to be informative about the stock status and key parameters. Following the review of analytical techniques and their performance, a list of general recommendations is provided for confronting uncertainty in stock assessments.

3.1 Introduction

3.1.1 Stock assessment and uncertainty

The general objectives in stock assessment are to understand the history of abundance, exploitation and productivity of a fish stock. The assessment is then used to formulate management advice that is robust to violated assumptions. The exact format of the advice depends on the management goals and procedures in place. In some cases, the advice is in the form of short-term projections of the stock response to alternative levels of harvest, but

in other cases point estimates are used as the basis of advice, e.g., as input to a harvest control rule.

A key part of stock assessment is to evaluate estimation uncertainty, in the form of probabilistic statements about quantities of interest, as well as model uncertainty, which involves exploring various assumptions about the underlying dynamics. The basic principle for incorporating uncertainty into management is the precautionary approach, to harvest conservatively when faced with a high level of uncertainty. A formal statistical approach to incorporate uncertainty into management is to adopt a harvest control rule that has been shown to perform well in long-term stochastic simulations across the many dimensions of uncertainty (Jakobsson and Stefánsson 1998, Butterworth and Punt 1999).

Fisheries stock assessment involves making technical decisions that have a direct impact on the management advice. These decisions include choosing a model family, fixing or estimating model parameters, deciding which data components to include when fitting the model, and choosing methods to evaluate uncertainty. The available data are the main basis for such decisions. Ideally, the technical decisions are backed by simulation studies where the relative performance of alternative methods is evaluated across a variety of simulated datasets.

Fisheries data are informative when they lead to accurate and precise estimates of model parameters and derived quantities, such as stock size and reference points. The quantity and resolution of the data are not the only factors determining whether data are informative or not, but also features such as the fishing history (Hilborn 1979, Magnusson and Hilborn 2007) and what kinds of data are available. In some cases, the only data component is the estimated annual landed catch, which contains important but limited information about the stock status, and methods based only on catch generally do not provide reliable information to guide management (Hilborn et al. 2003, Pauly et al. 2013). Adding a survey biomass index and/or age composition makes a dataset considerably more informative (Shepherd 1984, Chen et al. 2003, Magnusson and Hilborn 2007).

The most commonly used methods to quantify uncertainty in stock assessment are the

delta method, profile likelihood, bootstrap, and Bayesian analysis using Markov chain Monte Carlo (MCMC). Each method has its strengths and weaknesses, both theoretical and practical (Patterson et al. 2001, Bolker et al. 2013). The delta method is orders of magnitude faster to compute, profile likelihood can isolate selected quantities of interest, bootstrap simulation can detect estimation bias, and Bayesian MCMC analysis has been shown to have the most reliable performance across a range of simulated stock assessment scenarios (Magnusson et al. 2013).

3.1.2 Example: Icelandic saithe

Saithe (*Pollachius virens*) is a gadoid species that forms the basis for one of Iceland's major fisheries, with an average annual catch of 60 kt (ICES 2015b). Since 2013, the fishery has been managed with a 20% harvest control rule that sets the TAC each year as the average of two numbers: 20% of the current biomass of ages 4 and older, and the preceding year's TAC. Moreover, if the spawning biomass is estimated below $SSB_{trigger}$, defined as 65 kt, the target harvest rate is decreased (Hjörleifsson and Björnsson 2013, Magnusson 2013, ICES 2015b).

Stock assessment of Icelandic saithe is based on similar data as the other two major gadoid stocks, cod (*Gadus morhua*) and haddock (*Melanogrammus aeglefinus*), but is subject to greater uncertainty. This is apparent from large residuals between the data and model fits, as well as retrospective estimation errors (Hjörleifsson and Björnsson 2013, ICES 2015b). Factors contributing to the uncertainty include fluctuations in the survey data, as well as irregular changes in fleet selectivity. Although the annual groundfish bottom trawl survey (Pálsson et al. 1989) is data-rich in terms of numbers of stations and otoliths sampled, it does not specifically target the saithe, a species that is partly pelagic, schooling, and relatively widely migrating (ICES 2015b).

3.1.3 This study

The objective of this study is to give an overview of commonly used methods to confront uncertainty in stock assessment and to summarize their performance, based on previous sim-

ulations and analysis of the Icelandic saithe data. Four issues will be examined in particular: (1) the overall fishing history and whether it is likely to be informative about the stock status and key parameters, (2) the effects of different assumptions about the natural mortality rate M, stock-recruitment steepness h, and the shape of the selectivity curve for the oldest fish, (3) the amount of information contained in the survey data about the stock status, and (4) whether the delta method, profile likelihood, bootstrap, and MCMC lead to similar conclusions about the uncertainty of estimated quantities. The results from analyzing the saithe data are interpreted in light of previous studies, which evaluated the estimation performance of the same methods using simulated data (Magnusson and Hilborn 2007, Magnusson et al. 2013).

The goal is not to challenge the official assessment, which takes into account features specific to the saithe stock and fishery, such as migration events and time-varying selectivity (ICES 2015b). Rather, the models used in this analysis can be seen as diagnostic tools with relatively well known properties from previous studies. Based on simulations (Magnusson and Hilborn 2007) the expectation is that the data are informative about the stock status and M if they include years of varying fishing intensity, and informative about h if they include years of very small as well as moderate stock size. Furthermore, the delta method and MCMC can be expected to evaluate the uncertainty more accurately than the bootstrap (Magnusson et al. 2013), but confidence intervals can be expected to be too narrow in general. Neither of those studies involved data from actual fisheries, which gave rise to the current study. Following the analysis, some general findings and recommendations are summarized for stock assessment and uncertainty analysis.

3.2 Methods

3.2.1 Data

There is a relatively large amount of data available from the saithe fishery: catch-at-age data from commercial landings go back to 1980 and the annual groundfish survey started

in 1985. As a result, there are mainly four data components that contain information about the status of the saithe stock: annual landed catch in tonnes since 1980, a survey biomass index since 1985, and the age distributions from the fishery and the survey (Figure 3.1). Saithe is primarily caught in a mixed bottom trawl fishery around Iceland, and the annual otolith sample sizes are around 3000 from the fishery and 1000 from the survey. Discards are estimated to be negligible, around 0.1% of the catch in numbers (ICES 2015b). The complete dataset, including mean weight and maturity by year and age, can be found in the annual assessment reports (ICES 2015b, MRI 2015) and on the Marine Research Institute data server, `http://data.hafro.is`. For the purpose of the current analysis, ages 11 and older are pooled together in a plus group that comprises 3% of the catch data.

3.2.2 Models

Eight assessment models are used in the analysis and they differ in terms of which data types are included and which parameters are estimated (Table 3.1). The model naming scheme is based on Magnusson and Hilborn (2007), and a detailed description of the population dynamics and likelihood components is given in Appendix B.1.

The first three models use only a subset of the available data. Model 1 uses only landings and the survey index, model 2 uses landings and commercial catch at age, model 3 uses all of the above, and model 4 adds the survey catch at age. Model 4 is termed the basic model and the parameters that are not estimated in models 1–3 take their values from the basic model estimates. The rest of the models estimate three key parameters separately or together, where 'h' stands for stock-recruitment steepness, 'm' natural mortality rate, and 'r' right-hand selectivity. When these parameters are not estimated, they are fixed at $h = 0.9$, $M = 0.2$, and asymptotic selectivity.

The purpose of the simpler models (1–3) is to diagnose the effect of each data component on the estimated stock status and the uncertainty. The purpose of the more complex models (4h, 4m, 4r, 4hmr) is to diagnose the effects of those parameters on the estimated stock status and uncertainty, as well as evaluating how informative the data are about those parameters.

3.2.3 Bayesian priors / likelihood penalties

Diffuse Bayesian priors are used in the 4h, 4r, and 4hmr models to prevent h and right-hand-selectivity from hitting the lower or upper bounds during the estimation. Such likelihood penalties can also be applied in non-Bayesian modelling context to implement bounds in constrained optimization (Bard 1974). The prior on h follows the approach of Dorn (2002), defining a dummy parameter β as a shifted logit of h,

$$\beta_h \;=\; \log\!\left(\frac{h-0.2}{1-h}\right) \tag{3.1}$$

and the inverse relationship is:

$$h \;=\; \frac{0.2+\exp(\beta_h)}{1+\exp(\beta_h)} \tag{3.2}$$

For h in the interval $(0.2, 1)$, the logit β ranges from $-\infty$ to ∞. The diffuse prior

$$\beta_h \;\sim\; \mathrm{N}(2, 10^2) \tag{3.3}$$

is relatively flat for h between 0.25 and 0.999, but drops towards zero as h tends to the bounds at 0.2 and 1. The choice of $\mu = 2$ places the peak of the prior at $h = 0.90$, the same value as used in the basic model, while $\sigma = 10$ makes the prior sufficiently wide to allow high and low values of h.

The prior on right-hand selectivity of the commercial fleet is applied to the shape parameter ${}_C S_{\mathrm{right}}$ (see Appendix, Equation B.3). The broad normal distribution for this parameter,

$${}_C S_{\mathrm{right}} \;\sim\; \mathrm{N}(6, 10^2) \tag{3.4}$$

places the peak of the prior at $S_{11} = 0.94$, where the oldest age class is almost fully selected by the commercial fleet, but allows the estimated selectivity to be fully asymptotic or fully declining, as the prior is relatively flat for S_{11} between 0.002 and 0.999.

The diffuse priors are designed to be quite flat over a wide range of values and then curve

down at the predefined bounds. In this way, they improve model convergence and evaluation of confidence intervals, while having a negligible effect on the estimated values.

3.2.4 Uncertainty methods

Four methods are used to evaluate the uncertainty about the estimated quantities of interest: the delta method, profile likelihood, bootstrap, and MCMC. The procedures are described in detail in Appendix B.2. The delta method (Seber 1973) is implemented using 'sdreport' variables in AD Model Builder (Fournier et al. 2012). Profile likelihood (Venzon and Moolgavkar 1988) is applied directly to model parameters, but a penalty approach is used for derived quantities such as biomass and harvest rate. The parametric bootstrap is conditioned on the model fit to the data, and applied with and without BC_a bias correction with zero acceleration (Efron and Tibshirani 1993). The MCMC simulation (Metropolis et al. 1953, Hastings 1970) is run for 10 million iterations and then thinned, keeping every 10,000th iteration. For the sake of convenience, the term 'confidence interval' is used in this study to refer collectively to frequentist confidence intervals and Bayesian posterior intervals. Bivariate confidence regions are calculated using the R package 'r2d2' (Magnusson and Burgos 2014), combining two-dimensional kernel smoothing (Wand and Jones 1995) and polygon overlay algorithms (Bivand et al. 2013).

3.3 Results

3.3.1 Basic model fit to data

Model 4 is termed the basic model, being the simplest model that uses all the available data. The observed survey biomass index (Figure 3.2, panel A) fluctuates greatly between years in the early period of 1985–1994, and in later years the index shows medium-term fluctuations that the model does not fit very well. Overall, the residuals are positive when the fitted biomass is high and negative when the fitted biomass is low. The magnitude of the survey index observation noise is estimated iteratively as $\sigma_I = 0.43$.

The model fits the age distribution in the commercial catch and survey in most years, but with some exceptions, such as 1984 in the commercial catch and 1993 in the survey (Figure 3.2, panels B and C). The effective sample size is of similar magnitude for the commercial and survey catch at age, ${}_{C}n = 46$ and ${}_{S}n = 56$, estimated iteratively across all years. The estimated recruitment variability is $\sigma_R = 0.51$.

3.3.2 Model estimates and fishing history

For the purpose of managing the Icelandic saithe fishery, there are three quantities of main interest: the spawning biomass, the reference biomass (ages 4 and older), and the harvest rate (landings divided by the reference biomass). Results from the basic model indicate that the spawning biomass was at the reference point $\mathrm{SSB}_{\mathrm{trigger}} = 65$ kt in 1997, but above it in other years (Figure 3.3). The reference biomass declined from 451 to 150 kt between 1988 and 1995 and has increased in recent years, from 200 to 286 kt between 2009 and 2015. The estimated harvest rate exceeded 30% in 1994–1995 and again in 2008–2009, but has declined in recent years and is estimated at 16.1% for 2014. The average reference biomass since 1980 has been 272 kt with a 23% harvest rate.

Recruitment has varied between 11 and 111 million recruits at age 3 (Figure 3.4, panel A) and shows a weak relationship with the spawning biomass (panel B). Surplus production (the catch plus the subsequent change in reference biomass, panel C) is mainly driven by recruitment, with a correlation of 0.93 between panels A and C at a lag of three years, when cohorts are about to enter the reference biomass. The relationship between stock size and surplus production is weak, but a loess line indicates an optimum around 330 kt (panel D).

Based on historical average recruitment, maturity, and weights, the virgin spawning biomass (SSB_0) is estimated at 696 kt. From this, the relative SSB can be calculated as $\mathrm{SSB}/\mathrm{SSB}_0$. When the historical harvest rates are plotted against the relative SSB, the development of the fishing history shows a repeated anticlockwise pattern (Figure 3.5. During the period 1980–2012, the relative SSB has varied from 9 to 26 percent of SSB_0.

3.3.3 Retrospective analysis

When the basic model is fitted to truncated datasets, sequentially removing the last year of data, the stock size trajectories diverge at the terminal end (Figure 3.6). The stock size estimates are relatively high when the datasets end in 2013 and 2014, corresponding to high survey biomass indices in those years (Figure 3.1, panel B). The opposite trend can be seen around 2007–2009, when a sequence of low survey biomass indices cause the stock size estimate to decline several years in a row. Overall, the retrospective pattern indicates considerable estimation errors, but not a consistent bias towards under- or overestimation of the stock size.

3.3.4 Uncertainty and sensitivity analysis

The uncertainty about the estimated reference biomass in 2015 (B_{current}) and harvest rate in 2014 (u_{current}) depends on both the estimation model and the method used to construct the intervals (Figure 3.7). For example, the estimates and 90% confidence intervals for the basic model using the delta method are $B_{\text{current}} = 286$ kt (199–410 kt) and $u_{\text{current}} = 16\%$ (11–21%). For the delta method and profile likelihood analysis, the point estimate is the maximum likelihood estimate, but for the bootstrap and MCMC analysis the point estimate is the median of the bootstrap estimates and MCMC posterior. Profile likelihood and MCMC produce quite similar intervals as the delta method, while the 'raw' (non-bias-corrected) bootstrap intervals indicate smaller biomass and higher harvest rate. Applying bias correction to the raw bootstrap estimates produces intervals around a larger biomass and lower harvest rate than the other methods.

Compared to the basic model, the uncertainty increases slightly when the survey catch-at-age data are excluded from the analysis (model 3), and the intervals become almost twice as wide when the survey biomass index is also excluded in model 2 (Figure 3.7). With the annual landings and survey biomass index as the only input data, model 1 tends towards an infinite stock size with zero harvest rate. Estimating steepness h or right-hand selectivity

in models 4h and 4r has almost no effect on the intervals, but when natural mortality M is estimated, models 4m and 4hmr tend towards an infinite stock size with zero harvest rate (Figure 3.7). The profile likelihood and MCMC intervals are not restricted to have any predetermined shape, but they show a similar lognormal shape as the delta method intervals for both B_{current} and u_{current}.

The likelihood improvement in models 4h and 4r is negligible compared to the basic model 4 (Table 3.2). Models 4m and 4hmr, that tend to an infinite stock size, fit the data considerably better, specifically the commercial catch at age, although the improved fit to this data component is not visually apparent. The likelihood values from models 1, 2, and 3 are not comparable, since they include fewer likelihood components than models in family 4. The point estimates from model 2 indicate a slightly smaller current biomass than model 4, and model 3 a slightly larger current biomass, but the confidence intervals overlap to a great extent.

3.3.5 Steepness, natural mortality, and right-hand selectivity

When stock-recruitment steepness h is estimated in model 4h instead of fixing it at 0.9, the point estimates and intervals are close to the upper bound of 1.0 for the delta method and bootstrap, but the profile likelihood and MCMC 90% intervals extend slightly below 0.8 (Figure 3.8, upper panel). For model 4hmr, that tends to an infinite stock size, the 90% intervals extend to steepness values as low as 0.4.

Natural mortality rate M is estimated at 0.57 yr^{-1} in models 4m and 4hmr, almost three times higher than the fixed value of $M = 0.2$ yr^{-1} that is assumed in the other models (Figure 3.8, middle panel). The delta method and MCMC intervals extend from 0.55 to 0.60, the profile likelihood intervals are slightly wider, 0.49 to 0.60 for both models 4m and 4hmr, and the bootstrap intervals are much wider.

The right-hand side of the fleet selectivity curve in models 4r and 4hmr is estimated as asymptotic, with $S_{11} = 0.99$ as the selectivity for the oldest age group (Figure 3.8, bottom panel). The delta method and MCMC intervals are relatively narrow, from 0.92 to 1.00,

but the profile likelihood intervals are slightly wider, 0.82 to 0.99 for model 4hmr. The raw bootstrap intervals are much wider, 0.56 to 1.00 for model 4hmr, with most of the bootstrap estimates below the maximum-likelihood estimate, resulting in a strong BC_a bias-correction effect that leads to bias-corrected bootstrap intervals between 0.99 and 1.00 for models 4r and 4hmr.

3.3.6 Maximum sustainable yield

The optimal harvest rate u_{MSY} is estimated at 0.23, based on models 1, 2, 3, and 4 (Figure 3.9, upper panel). In these models there is effectively no perceived uncertainty about the estimated u_{MSY} where h and M are fixed, as u_{MSY} depends strongly on those parameters. When h is estimated in model 4h, the point estimate of u_{MSY} is 0.27, with the delta method and bootstrap producing very narrow intervals, but the 90% intervals from profile likelihood and MCMC are between 0.17 and 0.28. In models 4m and 4hmr, the point estimate of u_{MSY} is 0.37 with relatively wide confidence intervals, especially for the 4hmr model.

The estimated B_{MSY} (ages 4+) is 278 kt in model 4, with the delta method, profile likelihood, and MCMC intervals between 230 and 330 kt (Figure 3.9, middle panel). MSY is estimated at 63 kt, with 90% intervals between 53 and 75 kt, but the bootstrap intervals are narrower than the other methods. Models 2, 3, 4h, and 4r give similar results as model 4, except the estimate of B_{MSY} is lower in model 4h. In models 4m and 4hmr, the estimates of B_{MSY} and MSY tend to infinity.

Overall, the status of the Icelandic saithe stock is estimated to be close to B_{MSY} with a current harvest rate slightly below the target harvest rate (Figures 3.5 and 3.10). To evaluate the stock status, models 2, 3, 4, and 4h are of main interest, as models 1, 4m, and 4hmr tend to an infinite stock size, while estimates from model 4r are identical to model 4. The uncertainty about the stock status decreases considerably between models 2 and 3, and also between models 3 and 4, reflecting the amount of information contained in the survey data. The relative degree of uncertainty for each model can be quantified by calculating the unitless area of the 90% bivariate confidence region with $B_{\mathrm{current}}/B_{\mathrm{MSY}}$ and $u_{\mathrm{current}}/u_{\mathrm{MSY}}$ on the x

and y axes (Figure 3.10). This area is 0.497 for model 2, but 0.261 for model 3 and 0.200 for model 4. Based on this comparison between models 2 and 4, the survey data (catch at age and biomass index) decrease the uncertainty about the stock status by 60%. When h is estimated, the perceived uncertainty increases again, to 0.287 for model 4h.

3.4 Discussion

3.4.1 Summary of findings

Fishing history

The Icelandic saithe assessment demonstrates that a data-rich fishery, where age data have been sampled intensively for decades from the fishery and annual surveys, does not necessarily mean informative data. The narrow range of historical harvest rates results in a dataset that is not informative about M, and the measurement noise in the survey data lead to greater uncertainty about the stock status than for the related species cod and haddock (MRI 2015).

Besides having a relatively narrow range of harvest rates and stock size, the fishing history shows a clear circular pattern (Figure 3.5). The underlying causes are likely to be an interaction of dynamics that can be seen as a predator-prey cycle. The fishery could be considered the driving force: when the harvest rate increases the stock size goes down, then harvest rate is reduced, the stock recuperates, and the system is ready for another circle. Alternatively, the fish stock could be the driving force: when the stock size goes down due to natural fluctuations, humans react slowly so the harvest rate increases for some years, and when both the stock size and harvest rates have reached a low level, a strong cohort or two increase the stock size again. Charles (2001) describes similar dynamics between the natural and human components of fisheries as systems. Management actions are applied to the human component, and with the harvest control rule in place it is likely that the harvest rate will become more stable than in the past.

Sensitivity analysis

The sensitivity analysis comparing models 2, 3, and 4 (Figure 3.7) shows confidence intervals that overlap to a great extent. The effect of removing entire data components (survey catch at age, survey biomass index) from the basic model has the effect of increasing the overall uncertainty, but the point estimates of current biomass and harvest rate do not shift greatly. The general agreement between models 2, 3, and 4 indicates that the various data sources are consistent and not contradictory (Richards 1991, Schnute and Hilborn 1993).

The stock-recruitment steepness h cannot be estimated reliably from the saithe data, but appears to be high. Diagnostic model runs (Figure 3.8) give a point estimate close to 1.0 for h, suggesting that when the spawning stock size is near 20% of the virgin stock size, the expected recruitment is about the same as when the stock is much larger. The data include years when the relative stock size is between 9% and 26%, which should be an informative range for determining the shape of the curve, but the stock-recruitment scatter (Figure 3.4) does not indicate that recruitment is greatly affected at the lower range of spawning stock size. The lower bound of 90% confidence intervals for h is around 0.8, with profile likelihood and MCMC indicating greater uncertainty than the delta method and bootstrap. The shape of the stock-recruitment curve is important for long-term management advice, and basing the management on a maximum likelihood estimate of $h = 1.0$ would be neither precautionary nor biologically plausible (Mangel et al. 2010). The Icelandic saithe harvest control rule defines a lower threshold of stock size, below which recruitment is expected to be impaired and harvest rate should be decreased. In the harvest control rule evaluation for saithe (Hjörleifsson and Björnsson 2013) it is noted that such a threshold stock size is not well defined by the data, but $\mathrm{SSB}_{\mathrm{trigger}} = 65$ kt is set at the lowest historical spawning stock size, implicitly assuming a high steepness.

Likewise, the natural mortality rate M cannot be estimated reliably from the saithe data. Diagnostic model runs give a point estimate of M at 0.57 yr^{-1}, almost three times higher than the traditionally assumed value of 0.20 yr^{-1}. The diagnostic models estimating

M also estimate the stock size as being infinitely large, with a harvest rate of 0.00. In other words, these diagnostic models successfully estimate the average total mortality rate, but fail to partition the total mortality rate of 0.57 yr^{-1} between natural and fishing mortalities. Estimating M is an ongoing challenge in current stock assessment research (Lee et al. 2011, Hamel 2015, Johnson et al. 2015, Maunder and Piner 2015, Then et al. 2015), and the reliability of the estimation is largely determined by the fishing history. To estimate M reliably, one would require a dataset with age data from years with very low harvest rates, as well as years when harvest rate was several times higher (Beverton and Holt 1957, Magnusson and Hilborn 2007). The value of M is important to evaluate the optimal long-term harvest rate, and the harvest control rule evaluation for saithe is based on $M = 0.2$ yr^{-1}. Estimation of right-hand selectivity is strongly confounded with M (Thompson 1994, Magnusson and Hilborn 2007). When M is fixed at 0.2 yr^{-1}, the fleet selectivity is estimated fully or nearly asymptotic, but the uncertainty increases about the selectivity of older fish when M is estimated (Figure 3.8, bottom panel).

The delta method does not perform well evaluating the uncertainty about u_{MSY} when h is estimated, generating a much narrower interval than profile likelihood and MCMC. This may be a result of how u_{MSY} is estimated in the model, using bisection optimization within each function evaluation to search for the harvest rate that maximizes the sustainable yield (Magnusson and Hilborn 2007). Since u_{MSY} is not directly derived from the estimated parameters, profile likelihood and MCMC seem better suited to evaluate the uncertainty about u_{MSY}.

Overall, the bootstrap estimates are often biased compared to the point estimates, the bias being positive for current harvest rate but negative for current biomass and selectivity of the oldest fish (Figures 3.7 and 3.8). This could reflect a real estimator bias, which can be expected with challenging estimation problems in stock assessment (Magnusson and Hilborn 2007). The problem of too narrow bootstrap intervals when estimating MSY (Figure 3.9) was also identified by Magnusson et al. (2013), when an estimation model similar to the 4hmr model was applied to well-behaved simulated data. The BC_a bias correction improves

the reliability of bootstrap intervals, but the delta method and MCMC tend to produce more reliable intervals, even when stock assessment model estimates are biased (Magnusson et al. 2013). It is worth noting that parametric model-conditioned bootstrap is a form of 'self-testing' an assessment model, which can be useful to diagnose estimation biases and structural inconsistencies (Deroba et al. 2015).

Comparison with predictions

It is in agreement with expectations that the data are not informative about M, as the saithe fishery has been subject to rather steady levels of fishing intensity. It has also been noted prior to this analysis that the saithe stock has never been fished down to a critical level where recruitment is affected (Hjörleifsson and Björnsson 2013), so it is not surprising that there is insufficient information to distinguish whether h is high or very high. The estimated optimal harvest rate increases with h, but the resulting MSY changes much less.

Although the bottom trawl survey is known to be an unreliable indicator of changes in the saithe stock size (ICES 2015b), the analysis presented here suggests that the overall effect of the survey data is to decrease uncertainty about the stock status by around 60% (Figure 3.10). Nevertheless, the model fit to the survey biomass index has large residuals, similar to the official assessment model (ICES 2015b).

From previous simulations, it was expected that the parametric model-conditioned bootstrap would not perform well for evaluating the uncertainty about estimated quantities using this assessment model (Magnusson et al. 2013). Profile likelihood was not part of the earlier simulation analysis, so it is reassuring to see that this method generates intervals that are generally comparable to the delta method and MCMC intervals.

Comparison with official assessment results

The point estimates $B_{2015} = 286$ kt and $u_{2014} = 16.1\%$ are relatively close to the point estimates of the official assessment, $B_{2015} = 255$ kt and $u_{2014} = 17.5\%$ (ICES 2015b). In the current study, the optimal harvest rate is estimated at 23% using the basic model 4, but the

simulations behind the 20% harvest control rule adopted for the Icelandic saithe fishery take into account that each year's assessment is subject to considerable error when estimating the current stock size. Based on historical assessments, the harvest control rule simulations used an autocorrelated assessment error of $\sigma = 0.20$ and $\rho = 0.45$ (Hjörleifsson and Björnsson 2013).

3.4.2 General recommendations

The delta method, profile likelihood, bootstrap, and MCMC are commonly used to evaluate uncertainty in stock assessments. These methods have different strengths and weaknesses, and several variations are available for each method (Patterson et al. 2001, Magnusson et al. 2013). Profile likelihood is well-suited to evaluate the uncertainty about key parameters such as h and M, but slightly cumbersome to analyze derived quantities such as biomass and harvest rate. In the analysis presented here, the profile likelihood intervals are comparable to the delta method and MCMC intervals, but the bootstrap intervals seem biased and too narrow. For challenging estimation problems, model convergence can become a greater issue for the bootstrap than the other uncertainty methods. A more sophisticated bootstrap approach, applied to non-aggregated data at the sampling level, may capture the underlying spatial and temporal correlations and perform better than the bootstrap approached used here (Elvarsson et al. 2014).

For a given stock assessment, it is better to compare the results from more than one method of estimating uncertainty, rather than just using one arbitrarily chosen method. Uncertainty analysis is not only about evaluating probabilities and confidence intervals, but iterative methods such as profile likelihood, bootstrap, and MCMC can also indicate lack of model convergence, find a new global optimum, identify highly correlated or ill-defined parameters, and suggest parameter transformation.

When analyzing complex data, it is important to run a variety of models to explore model uncertainty and test the effects of different assumptions (Tukey 1997). Comparative model runs should therefore involve both variations of one model, as done in this study to exam-

ine the effects of specific assumptions, and also compare fundamentally different models to approximate the true range of uncertainty. For example, the ICES (2015b) assessment compared ADAPT, state-space, and statistical catch-at-age models that represent a continuum in how variability in the data can be partitioned between process variability and observation error. Model comparison can also help identify how informative the data are, separate the information contained in each data component (Magnusson and Hilborn 2007), and examine contradictory data sources (Schnute and Hilborn 1993) by downweighting or excluding data components. Thus, model comparison is not only about selecting the best model, but forms an essential part of analyzing stock assessment data.

Several parameters describing fish population dynamics are known to be problematic for statistical estimation, which can have a large effect on the resulting management advice. These include h, M, and right-hand selectivity (Magnusson and Hilborn 2007). In the best case, informative data allow the estimation of these parameters. One approach is to use Bayesian priors to borrow information from similar stocks (Punt and Hilborn 1997), but in many assessments the parameters are simply fixed at a somewhat arbitrary value.

It is worth considering that a harvest control rule (HCR), whose aim is to keep the harvest rate as constant as possible, may produce the least informative data for future stock assessments, as it minimizes the potential contrast in harvest rates and stock size. This can still be the optimal harvest strategy, if earlier data have already provided sufficient information about the stock dynamics. In some cases, it might be sensible to probe a stock with very high and very low harvest rates (Ludwig and Hilborn 1983, Walters 1986) before adopting a HCR, to get more reliable estimates of what harvest rate is optimal and what level of spawning biomass should be chosen as the lower threshold.

Before a HCR is adopted, it is evaluated against the main sources of uncertainty, but after the adoption many HCRs only require point estimates as input, with limited or no uncertainty analysis performed annually. Examples include Icelandic cod (ICES 2009), haddock (Björnsson 2013) and saithe (Hjörleifsson and Björnsson 2013), where the required input is the current reference biomass and spawning stock biomass (SSB). For those stocks, the man-

agement goals are to achieve maximum sustainable yield and to reduce the fishing mortality rate when SSB is estimated below a defined limit. Other HCRs explicitly incorporate annual uncertainty, which is expected to vary between years. For example, the Icelandic capelin HCR (ICES 2015a) is based on the lower confidence limit of the latest acoustic survey estimate of biomass, where the management goal is to ensure with 95% probability that SSB is above a defined limit, and to fish the excess.

To summarize, our general recommendations are:

- Use more than one method to evaluate uncertainty.
- Keep in mind that the real uncertainty is greater than the analytical confidence intervals indicate.
- Use more than one model and variations of models to evaluate how sensitive the main conclusions are to alternative assumptions.
- Use retrospective analysis to evaluate uncertainty from an empirical viewpoint.
- Use simulation analysis to evaluate the performance of the estimation model, which parameters can be estimated reliably, and which uncertainty methods work best.
- Examine the fishing history to evaluate whether the data are likely to be informative about the stock status and key parameters like h and M.
- Consider ways to reduce uncertainty by generating informative data via management (e.g., applying different fishing mortalities between years) and research (e.g., design a dedicated survey for a given stock, sample age data).
- Harvest control rules can be a practical way to incorporate uncertainty into management advice.

The overall conclusion regarding uncertainty in fisheries stock assessment is simple: we know we will always be wrong, but if we're smart we can avoid being terribly wrong.

Acknowledgements

We thank Jan Jaap Poos for insightful comments that improved the manuscript, and the crew members and scientists at the Marine Research Institute for collecting and analyzing the Icelandic saithe data over the years.

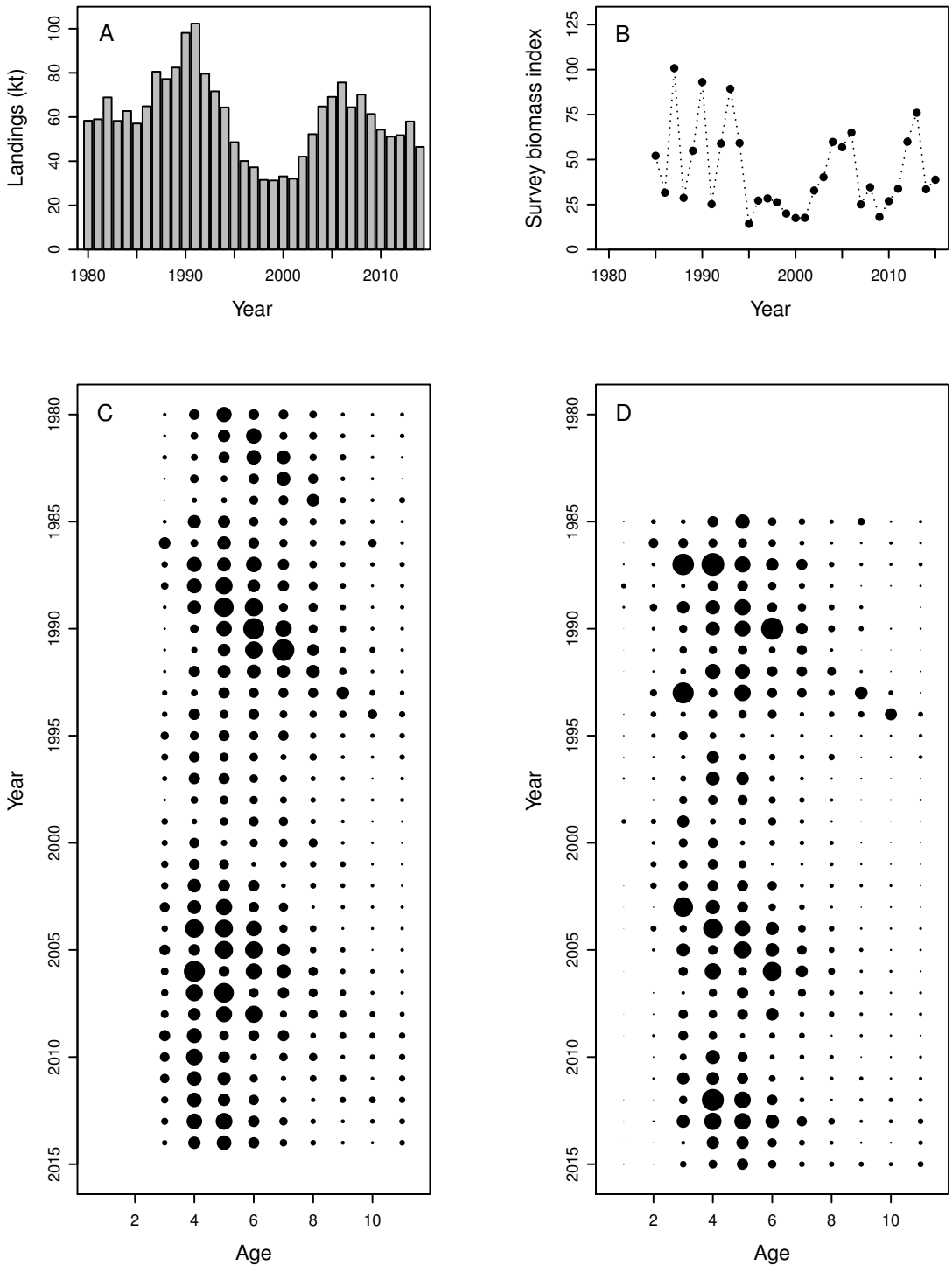

Figure 3.1. Icelandic saithe data: (A) landings 1980–2014, (B) survey biomass index 1985–2015, (C) commercial catch at age 1980–2014, and (D) survey catch at age 1985–2015. The area of each dot in panels C and D reflects absolute numbers of fish, with age 11 as a plus group.

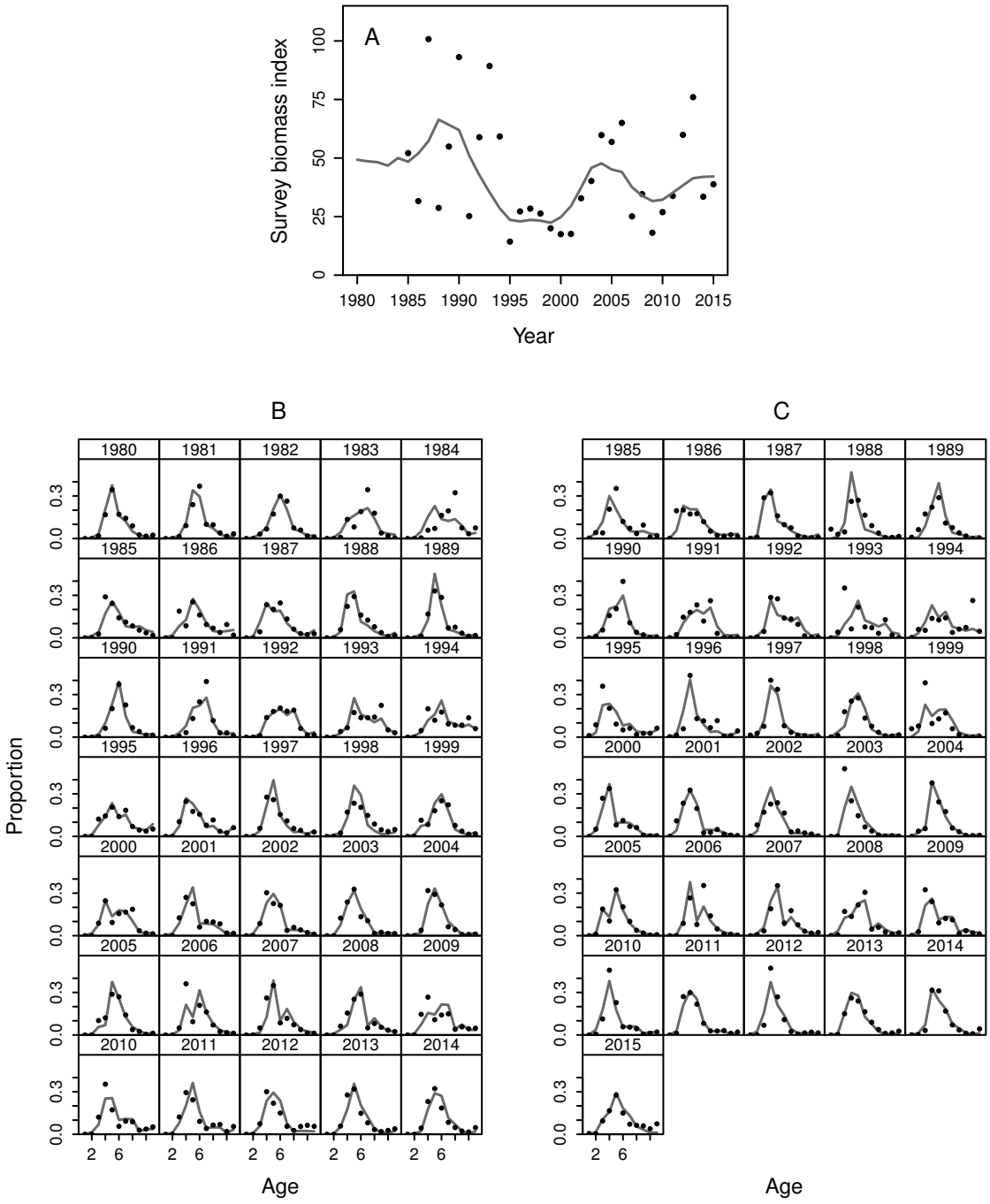

Figure 3.2. Basic model fit to the data: (A) survey biomass index, (B) commercial catch at age, and (C) survey catch at age. Observed data are shown as dots and model fit as lines.

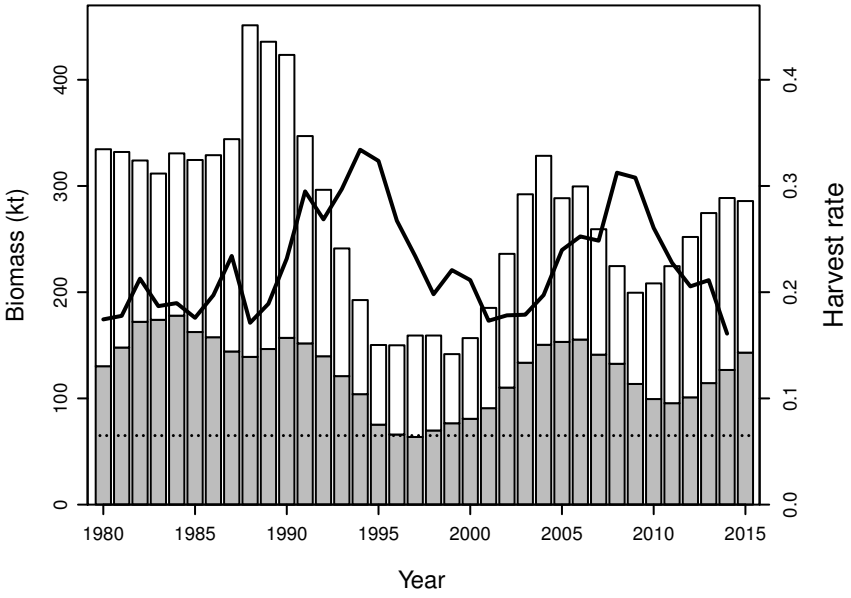

Figure 3.3. Estimated biomass and harvest rate for the basic model. The reference biomass (B_{4+}) is shown as columns with the mature part (SSB) in gray, the $\mathrm{SSB}_{\mathrm{trigger}} = 65$ kt reference point is shown as a horizontal dotted line, and the harvest rate as a solid line.

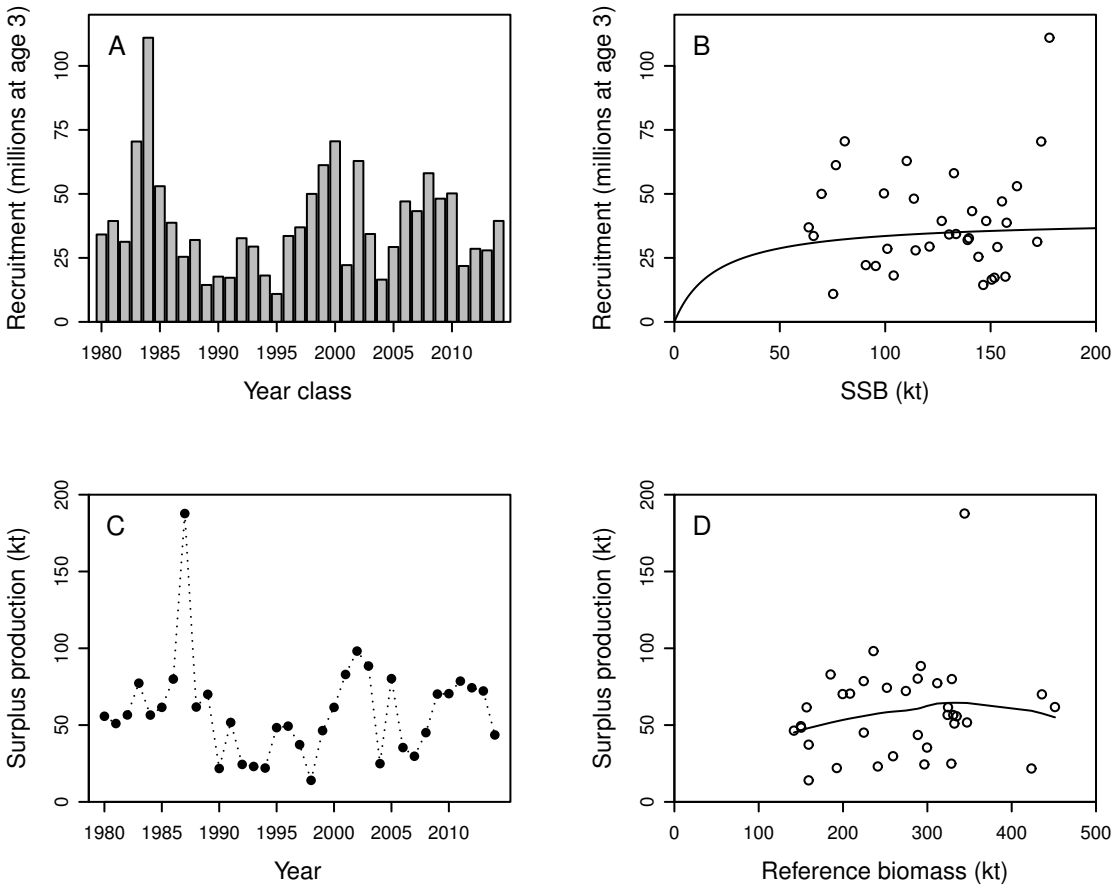

Figure 3.4. Estimated recruitment and surplus production for the basic model: (A) cohorts, (B) stock-recruitment with Beverton-Holt line, (C) surplus production by year, and (D) surplus production vs. stock size with loess line.

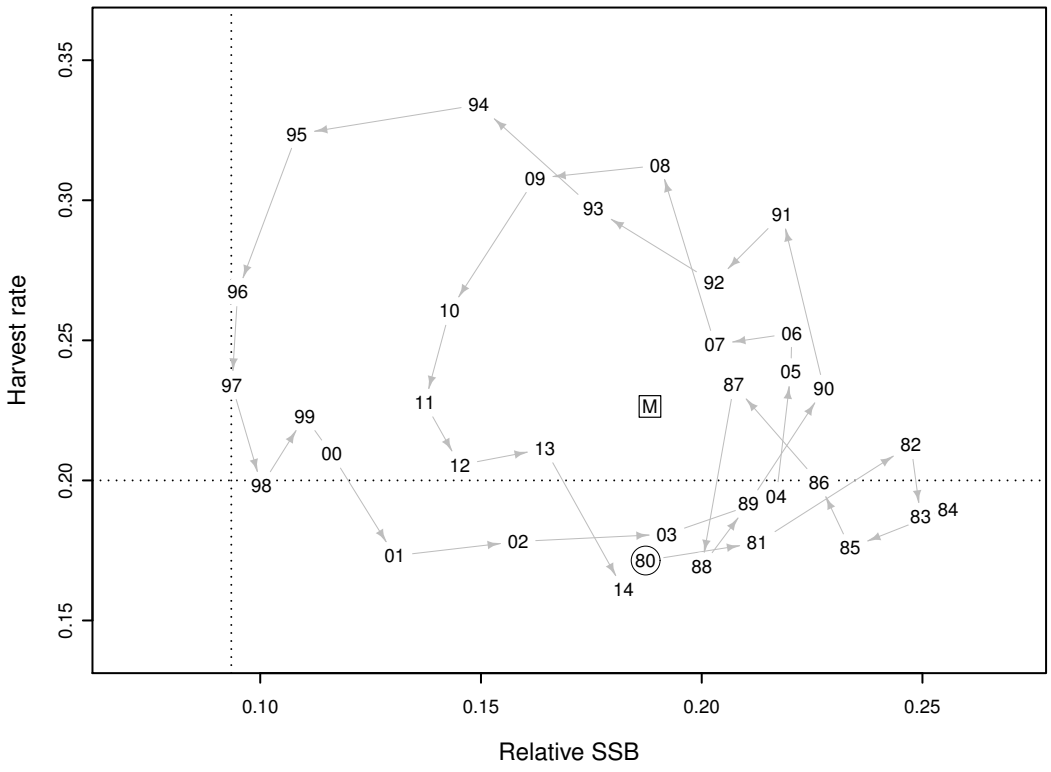

Figure 3.5. Estimated fishing history for the basic model. Labels show the years 1980–2014, with a circle around the initial year and arrows indicating the development of the fishing history. The dotted lines show the 20% target harvest rate and $\mathrm{SSB}_{\mathrm{trigger}} = 65$ kt reference point set by the harvest control rule. The M marker shows the estimated optimal harvest rate u_{MSY} and the long-term average spawning stock size at that harvest rate.

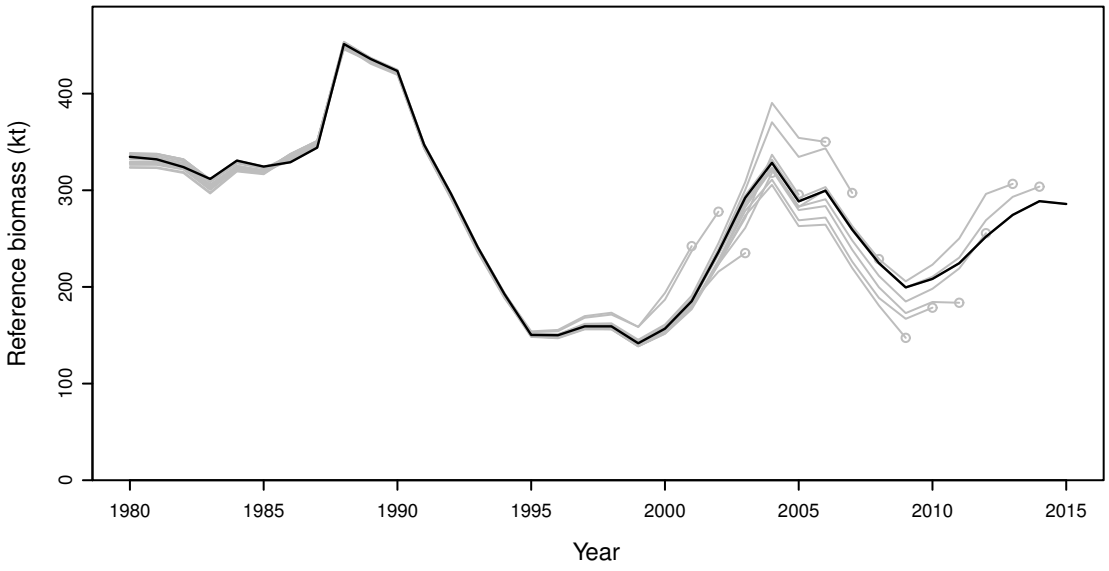

Figure 3.6. Retrospective analysis of reference biomass (B_{4+}) for the basic model. The black line shows the biomass estimated from the full dataset, but the gray lines are estimates from sequentially truncated datasets.

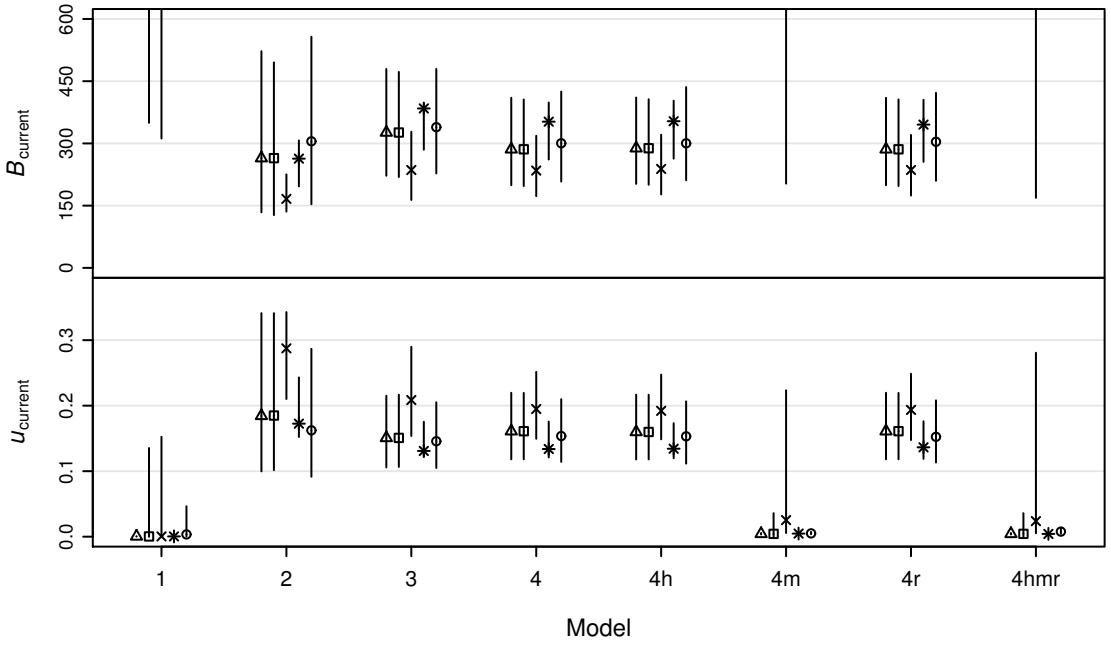

Figure 3.7. 90% confidence intervals for current reference biomass (B_{current}) and harvest rate (u_{current}). The intervals are constructed using the delta method (triangle), profile likelihood (square), raw bootstrap (x), bias-corrected bootstrap (asterisk), and MCMC (circle).

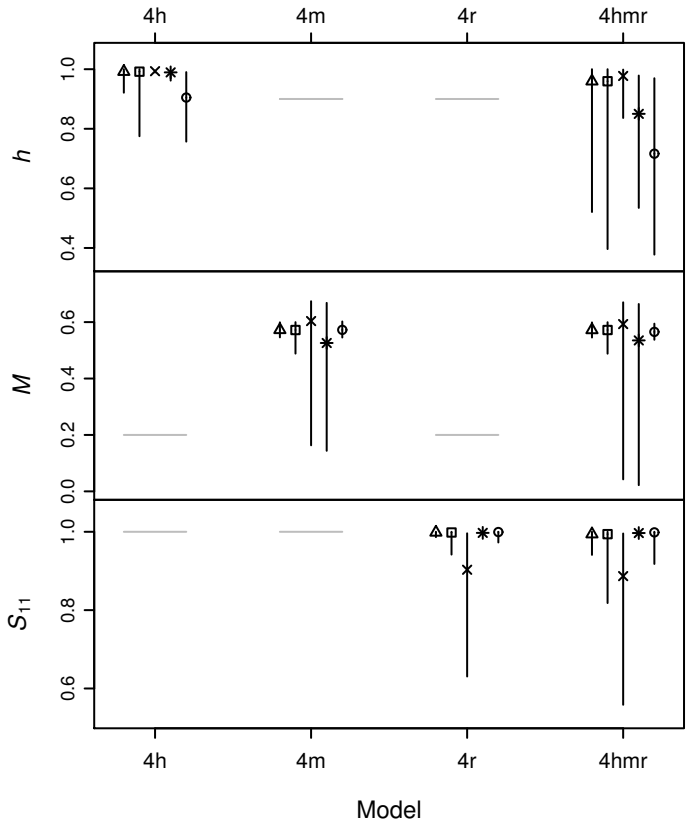

Figure 3.8. 90% confidence intervals for stock-recuitment steepness h, natural mortality rate M, and selectivity of the oldest fish S_{11}. The intervals are constructed using the delta method (triangle), profile likelihood (square), raw bootstrap (x), bias-corrected bootstrap (asterisk), and MCMC (circle). When parameters are not estimated in a model, the fixed value is shown as a horizontal gray line.

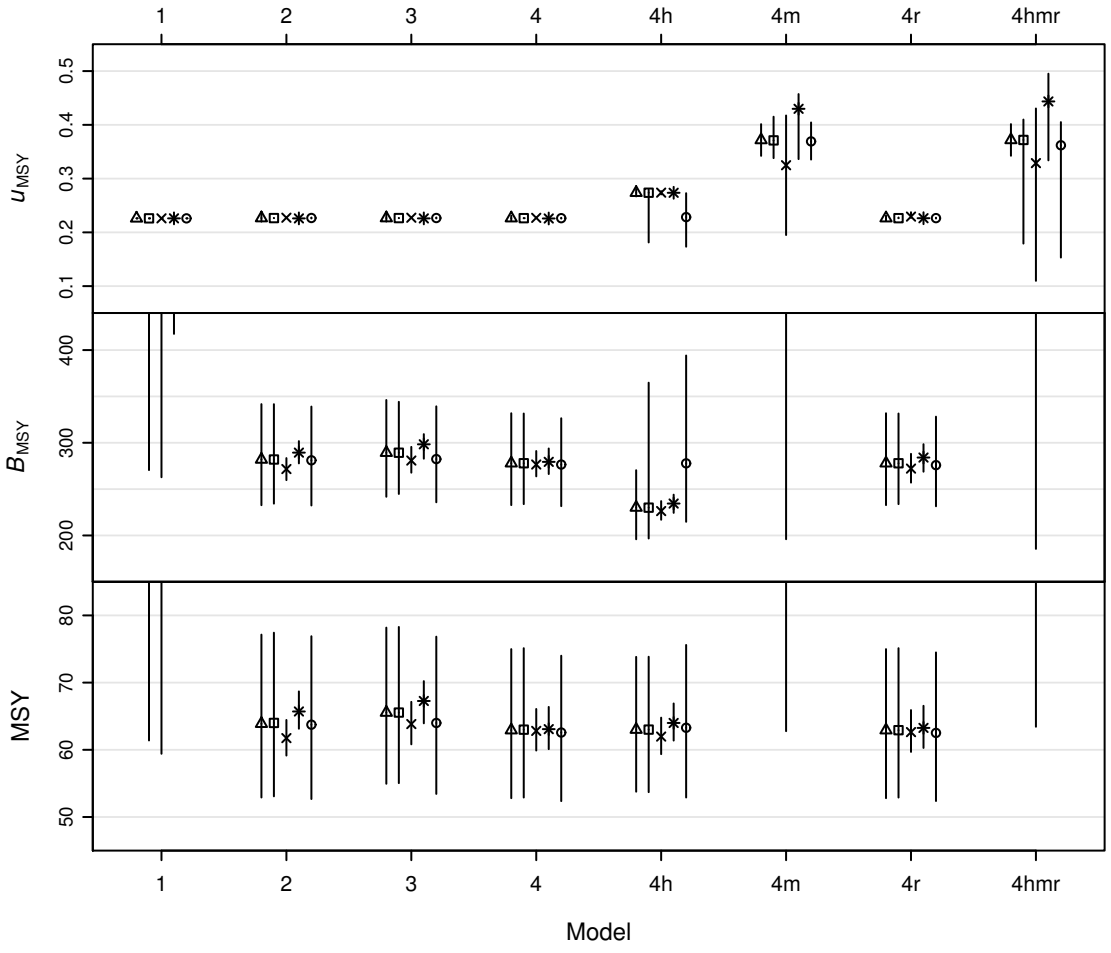

Figure 3.9. 90% confidence intervals for MSY-related quantities: optimal harvest rate (u_{MSY}), average long-term biomass (B_{MSY}, ages 4+), and average long-term yield (MSY). The intervals are constructed using the delta method (triangle), profile likelihood (square), raw bootstrap (x), bias-corrected bootstrap (asterisk), and MCMC (circle).

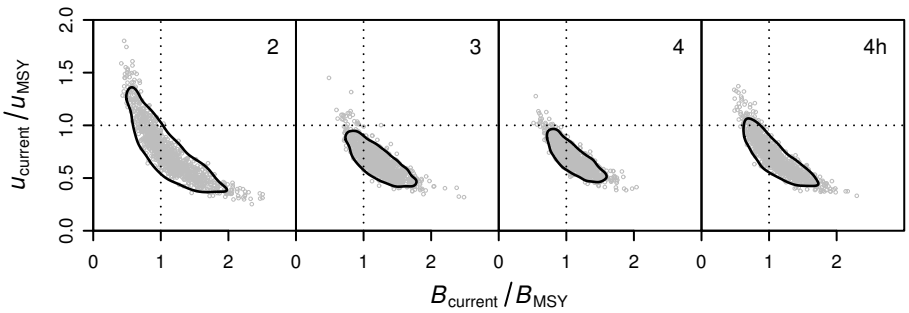

Figure 3.10. Bivariate 90% confidence regions for current biomass relative to B_{MSY} (x-axis) and current harvest rate relative to u_{MSY} (y-axis). For each model (panel), the region (black) contains 90% of the MCMC iterations (gray dots).

Table 3.1. Assessment models used in this study, in terms of data types used and parameters estimated. Catch stands for landings, Index for survey biomass index, cCA for commercial catch at age, and sCA for survey catch at age.

	1	2	3	4	4h	4m	4r	4hmr
Data								
Catch	×	×	×	×	×	×	×	×
Index			×	×	×	×	×	×
cCA		×	×	×	×	×	×	×
sCA				×	×	×	×	×
Estimated								
R_0	×	×	×	×	×	×	×	×
h					×			×
M						×		×
R_{init}	×	×	×	×	×	×	×	×
u_{init}	×	×	×	×	×	×	×	×
R_{plus}		×	×	×	×	×	×	×
${}_{C}S_{\text{full}}$		×	×	×	×	×	×	×
${}_{C}S_{\text{left}}$		×	×	×	×	×	×	×
${}_{C}S_{\text{right}}$							×	×
${}_{S}S_{\text{full}}$				×	×	×	×	×
${}_{S}S_{\text{left}}$				×	×	×	×	×
q	×		×	×	×	×	×	×
${}_{R}\varepsilon$		×	×	×	×	×	×	×

Table 3.2. Summary of diagnostic model runs. The column 'npar' shows the number of parameters estimated in each model and Δf is the objective function improvement when h, M, and ${}_{C}S_{\text{right}}$ parameters are estimated, compared to the basic model 4. B_{current} and u_{current} are the estimated current biomass and harvest rate.

Model	npar	Δf	B_{current}	u_{current}
1	4	n/a	∞	0.0%
2	50	n/a	264	18.4%
3	51	n/a	326	15.1%
4	53	0.0	286	16.1%
4h	54	−0.4	288	16.0%
4m	54	−6.3	∞	0.0%
4r	54	0.0	286	16.1%
4hmr	56	−6.4	∞	0.0%

Chapter 4

SOFTWARE PACKAGES DEVELOPED FOR SIMULATION ANALYSIS

Introduction

Simulation studies can be considered a third mode of science, complementing and adding to experimental/observational studies (inference from data) and theoretical studies (method development and logical relationships). Common objectives in simulation studies include:

- Evaluate uncertainty about estimated quantities, e.g., using bootstrap or MCMC.
- Evaluate estimation performance, often comparing different methods in terms of bias, precision, robustness, power, coverage probability, etc.
- Evaluate performance of policies, such as alternative harvest control rules, in face of uncertainty.
- Test whether an estimation model has software bugs, by fitting to datasets that were simulated with an operating model written in another programming language.

The idea of using long sequences of random numbers to support statistical analysis, also known as the Monte Carlo method, is not entirely new. Early probabilists used simple dice to validate methods and by the late 19th century, tools had been devised to generate random normal deviates as a foundation for simulation analysis (Stigler 1991). Computers ushered in new possibilities in applying simulations to iteratively approximate solutions that are unobtainable in closed form (Metropolis and Ulam 1949).

Simulating a large number of realistic datasets as model input was first used to evaluate uncertainty about estimated quantities (Efron 1979) and later to evaluate the performance of alternative models (Sacks et al. 1989). The role of simulations within the scentific method

is a current challenge in the philosophy of science (Peck 2004, Winsberg 2010) and technical references on the design and analysis of simulation studies include those by Banks (1998), Santner et al. (2003), Asmussen and Glynn (2007), and Robert and Casella (2010).

The research presented in this dissertation is largely based on simulations. It is the nature of such analysis that computational tasks are repeated, often thousands of times, and then re-run with modifications. Any improvements in the efficiency and capabilities of software tools are therefore valuable.

Statistical simulations require the analyst to work with large amounts of data and results, to modify and apply methods, examine the effects of alternative options, diagnose statistical issues that arise, and make informed decisions on the final experimental design based on intermediate results. This is achieved with a combination of scripts and functions to perform calculations, along with interactive tools for exploratory analysis.

Some of the software written to perform the analysis of Chapters 1–3 was implemented in a general way so that other statistical modellers might find it useful for their work. The functions and packages listed below have been actively maintained and developed for a while, made available on the ADMB and R Project websites (with the exception of `BCboot` which was published in Fish and Fisheries, as part of Chapter 2), and some of the software is now widely used. The packages and functions can be broadly categorized as follows:

Software to implement and modify methods

1. ADMB additions: `ADMB-IDE` and compilation scripts
2. Bivariate confidence regions: R package `r2d2`
3. Bootstrap bias correction: R function `BCboot`

Software to aggregate and diagnose results

4. CODA addition: R function `cumuplot`
5. MCMC diagnostic plots: R package `plotMCMC`
6. R-Core additions: R functions `boxplot/bxp` and `aggregate` (coauthor)
7. Statistical catch-at-age plotting environment: R package `scape`

4.1 ADMB additions

AD Model Builder (ADMB) is a programming framework based on automatic differentiation, aimed at highly nonlinear models with a large number of parameters (Fournier et al. 2012). The main advantages of ADMB are flexibility, speed, precision, stability, and built-in methods to quantify uncertainty. Some of the challenging issues for both new and experienced ADMB users used to be:

1. Shell commands to build models were numerous and inconsistently named between different combinations of operating systems and compilers.
2. Editing code was cumbersome, especially with models consisting of several thousand lines of code.
3. Installation was difficult, requiring the user to modify environment variables and setting up additional components, most importantly a C++ compiler.

My first contribution as a member of the ADMB Core Team in 2009 was a redesigned model compilation pathway, based on a single shell command `admb` that works consistently to build models on all platforms (cf. issue 1 above). The new compilation scripts were introduced in ADMB 9.1 released that same year and have been the default since then.

The improvement in the user interface at the shell level made it possible to design an efficient integrated development environment (IDE), providing syntax highlighting, menu commands to compile and debug, as well as commands to quickly navigate between sections of the code (cf. issue 2 above). Users can choose between two ways to set up the IDE: (1) experienced Emacs users can download a lightweight `admb-mode` and use it without changing their keyboard settings, while (2) non-Emacs users can download `ADMB-IDE` which automatically installs and configures ADMB along with the additional components of a full-featured IDE (compiler, debugger, and a customized editor for ADMB with simple keyboard settings). In this way, ADMB-IDE (Magnusson 2009, Magnusson 2015) addresses both issues 2 and 3 above (Figure 4.1).

4.2 Bivariate confidence regions

In Chapter 3, an algorithm was needed to quantify the two-dimensional uncertainty about stock status, in terms of relative stock size and harvest rate (see Figure 3.10). A search through the literature and existing packages revealed that no out-of-the-box solution was available. The problem is a general one, and multiple theoretical solutions exist to draw an arbitrary shape that includes a specified proportion of a two-dimensional cloud of points. In challenging cases, no method can guarantee that a contiguous shape can be found that includes exactly the right number of points.

After some experimentation, the solution was to combine two-dimensional kernel smoothing (Wand and Jones 1995) and polygon overlay algorithms (Bivand et al. 2013), and the algorithm was implemented as a function `conf2d` released in an R package 'r2d2' (Magnusson and Burgos 2014).

The function constructs a large number of smooth polygons, and then chooses the polygon that comes closest to containing a given proportion of the total points. The user can select between two kernel smoothers and specifies the confidence level, along with several shape parameters. Figure 4.2 shows the default output that contains 950 out of 1000 points in an example dataset that comes with the package.

My coauthor, Julian Burgos, has a strong background in spatial statistics and contributed to the overall approach and the choice of algorithms.

4.3 Bootstrap bias correction

The following R function was implemented for this study to apply BC_a bootstrap bias correction with zero acceleration, robust to extremely biased cases (Equation 2.9). The effect of BC_a bias correction is demonstrated in Figure 2.3. This function was published in the journal Fish and Fisheries (Magnusson et al. 2013, p. 342).

```
BCboot <- function(thetastar, thetahat, bounds=c(0.1,0.9))
#############################################################################
###                                                                       #
### Function: BCboot                                                      #
###                                                                       #
### Purpose:  Apply bias correction to bootstrap estimates                #
###                                                                       #
### Args:     thetastar is a vector of bootstrap estimates                #
###           thetahat is a point estimate from original data             #
###           bounds is a vector of lower and upper limits to handle extremely #
###             biased cases                                              #
###                                                                       #
### Notes:    BCa with zero acceleration                                  #
###           Based on bcanon() in package 'bootstrap' by Tibshirani      #
###           See Efron and Tibshirani (1993, pp. 184-186), Gavaris and Van   #
###             Eeckhaute (1998, p.10), Gavaris (1999, p. 47), and Magnusson  #
###             et al. (2013, p. 342)                                     #
###                                                                       #
### Returns: Vector of bias-corrected bootstrap estimates                 #
###                                                                       #
#############################################################################
{
  B <- length(thetastar)
  alpha <- (1:B) / B
  lower <- bounds[1]
  upper <- bounds[2]
  z0 <- qnorm(max(lower, min(upper, sum(thetastar<thetahat)/B)))
  zalpha <- qnorm(alpha)
  newalpha <- pnorm(2*z0 + zalpha)
  Omegainv <- approx(alpha, sort(thetastar), newalpha, rule=2)$y
  bias.corrected <- Omegainv[rank(thetastar)]

  return(bias.corrected)
}
```

4.4 CODA addition

CODA is an R package for MCMC analysis and diagnostics, providing a suite of plots and tests for this purpose. In the past, it was missing a plot to diagnose visually whether the MCMC chain has been run long enough for the quantiles to stabilize.

Posterior quantiles, e.g., the 5th and 95th percentiles, are the standard approach for constructing Bayesian intervals, which is an important output from Bayesian analysis. It is therefore highly relevant to diagnose whether increasing the number of MCMC iterations would be likely to result in changes in the Bayesian interval.

The `cumuplot` function has been a part of the main diagnostic plots of the CODA package since version 0.6-1. This contribution makes me a coauthor of the CODA package (Plummer et al. 2015), and the usage of `cumuplot` as an MCMC diagnostic is recommended by authorities in the field such as Robert and Casella (2010, pp. 243–268).

4.5 MCMC diagnostic plots

The main shortcoming of the CODA package is the limited support for multipanel plots. As computer monitors have increased in size and resolution, multipanel plots offer an efficient way to visually diagnose MCMC chains for many estimated quantities simultaneously.

The 'lattice' package provides a flexible framework to produce multipanel plots (Sarkar 2008). The purpose of the 'plotMCMC' package is to combine the features of CODA and 'lattice' to produce multipanel plots for MCMC analysis and diagnostics (Magnusson and Stewart 2014). An example plot is shown in Figure 4.4. The package consists of six user functions, whose purpose is listed below.

Diagnostic plots

`plotTrace`	look for unwanted trends or patterns in MCMC traces
`plotAuto`	calculate autocorrelation to decide if further thinning is required
`plotCumu`	check whether quantiles have converged
`plotSplom`	evaluate confounding of parameters

Posterior plots

`plotDens`	plot posterior distribution and quantiles
`plotQuant`	plot multiple posteriors on a common y-axis

My coauthor, Ian Stewart, introduced me to formal MCMC diagnostics and contributed to the overall approach and the design of each plot.

4.6 R-Core additions

R is a popular statistical software platform (R Core Team 2015, Tippmann 2015), particularly effective for visual and interactive data analysis. The packages that make up the minimal installation of R and provide the basic functionality are called the core packages, and to maximize the stability and long-term maintainability of R, the Core Team is very reluctant to add new features to these packages. Instead, they encourage users to fulfill their own 'wishlist' features by contributing new packages to the CRAN central package repository, or to add new functions to existing user-contributed packages. For these reasons, many of my proposed changes to R-Core have been rejected, but over the years I have contributed a few dozen lines of code that have made it into the R codebase. My implementation of these features was guided by R core developers Martin Mächler and Kurt Hornik.

Boxplot graphical parameters

The formatting options for boxplots in R used to be quite limited, in terms of controlling line widths, line types, point symbols, and colors. The graphical parameters that specify these visual aspects are defined in the `bxp` function, which is called by the user function `boxplot` to render the plot. Since R version 2.0, every graphical aspect of the boxplot can be specified by the user. See R help page on `bxp` and Figure 4.5.

Aggregate formula interface

The `aggregate` function in R splits data into subsets and computes summary statistics for each subset. The statistic can be the sum, mean, median, count, minimum, maximum, standard deviation, or in fact any function, while the data grouping variables can be a combination of factors such as year, age, species, area, etc.

The user interface to specify the main variables and grouping variables for `aggregate` used to be rather cumbersome. R 2.11 introduced a more convenient formula interface, similar to the formula interface used in standard plots and models. See R help page on `aggregate`.

4.7 Statistical catch-at-age plotting environment

For the analysis in Chapters 1–3, the Coleraine assessment model was fitted to thousands of simulated datasets. It became clear at an early stage that an efficient suite of plotting functions was needed to quickly visualize and diagnose any model run of interest. This was implemented as an R package called 'scape' (Statistical Catch-at-Age Plotting Environment, Magnusson 2005, Magnusson 2014). Besides plotting, the package provides functions to iteratively estimate lognormal sigmas and multinomial effective sample sizes, and to import MCMC results that can be plotted with the 'plotMCMC' package (Section 4.5).

Import model results

`importCol`	import Coleraine model results

Plot model fit to data

`plotCA`	plot catch at age
`plotCL`	plot catch at length
`plotIndex`	plot abundance index
`plotLA`	plot length at age

Plot derived quantities

`plotB`	plot biomass, recruitment, and landings
`plotN`	plot numbers at age
`plotSel`	plot selectivity and maturity

Sigmas and sample sizes

`getN`, `getSigmaI`, `getSigmaR`	extract sigmas and sample sizes
`estN`, `estSigmaI`, `estSigmaR`	estimate sigmas and sample sizes
`iterate`	iteratively estimate all sigmas and sample sizes

Import MCMC results

`importMCMC`	import MCMC traces of likelihoods, parameters, biomass, and recruitment
`importProj`	import MCMC future projections of biomass and catch

The 'scape' package comes with a function to import results from Coleraine output files, but users can write their own import function for other statistical catch-at-age models. An example of this is an in-house function at the Marine Research Institute called `importADCAM`, tailored for the model used to assess the Icelandic cod and saithe stocks (Bjornsson and Magnusson 2009).

Each plotting function can display the data and model results in various ways. Examples of three ways to plot catch at age with the `plotCA` function are shown in Figures 4.6–4.8.

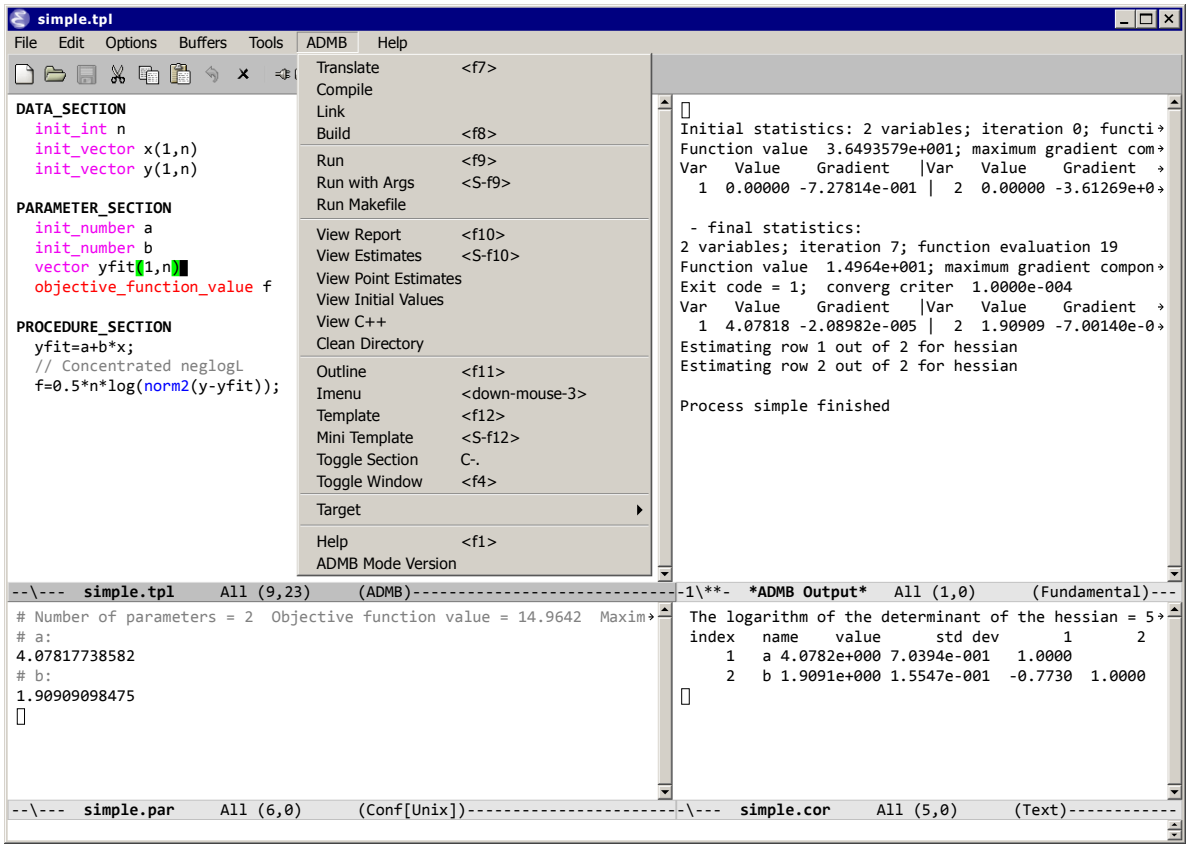

Figure 4.1. Typical AD Model Builder session, using ADMB-IDE. The top-left window shows the model code, demonstrating some of the sections and classes recognized by the ADMB-to-C++ translator. The menu includes commands to build a model, run, and view the output. Other windows show point estimates, standard errors, and correlation of estimated quantities. [From Fournier et al. 2012, p. 239.]

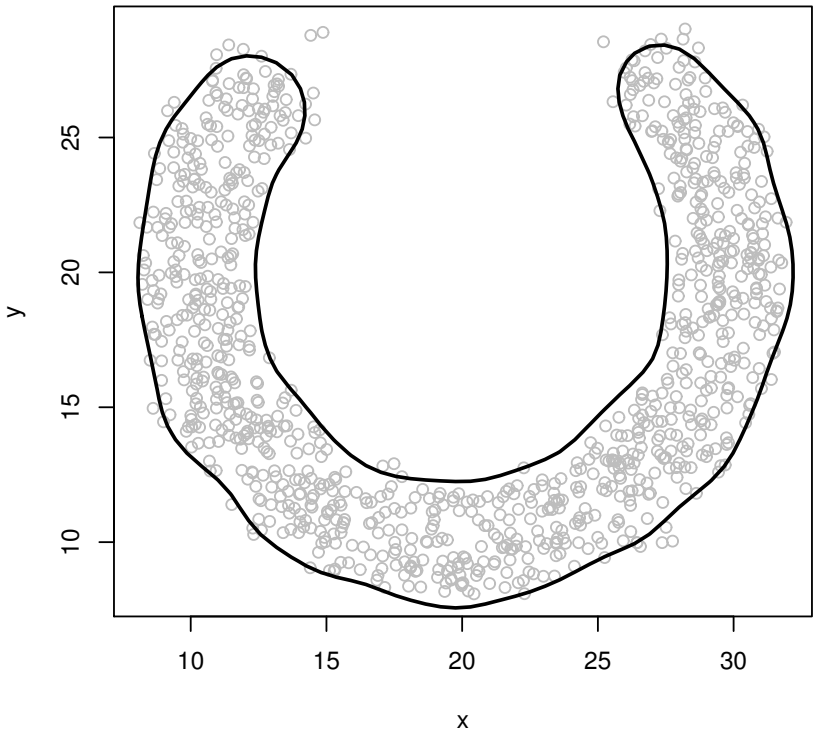

Figure 4.2. 95% bivariate confidence region for a random U-shape scatter. This corresponds to the output of the first example on the `conf2d` help page.

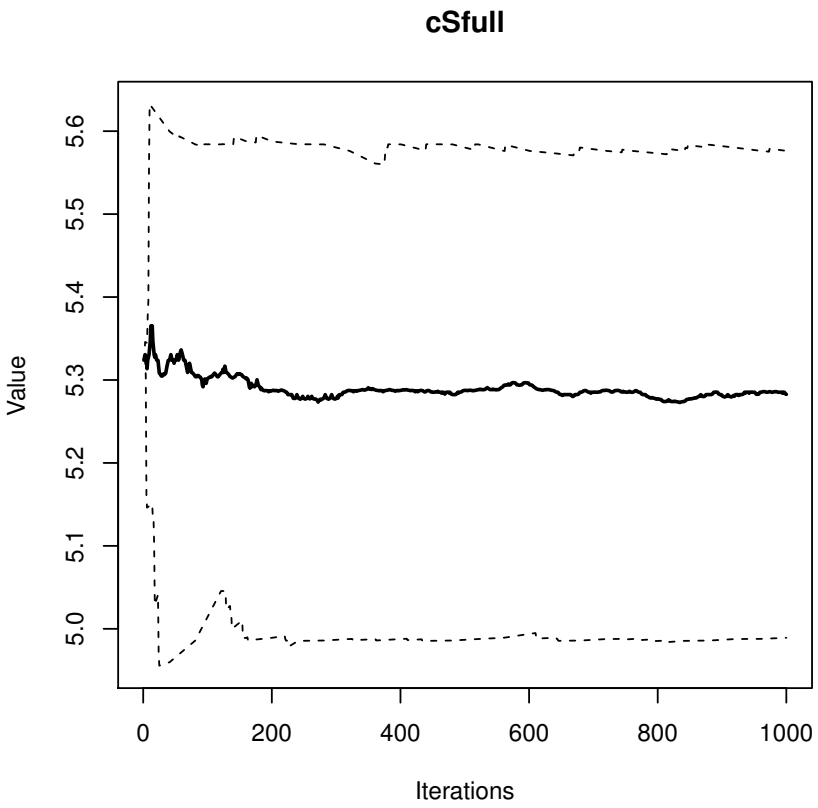

Figure 4.3. Cumulative median and quantiles (2.5%, 97.5%) of an MCMC chain for the model parameter ${}_{C}S_{\text{full}}$. This example dataset is from the 'plotMCMC' package, which calls `cumuplot` to render the plot.

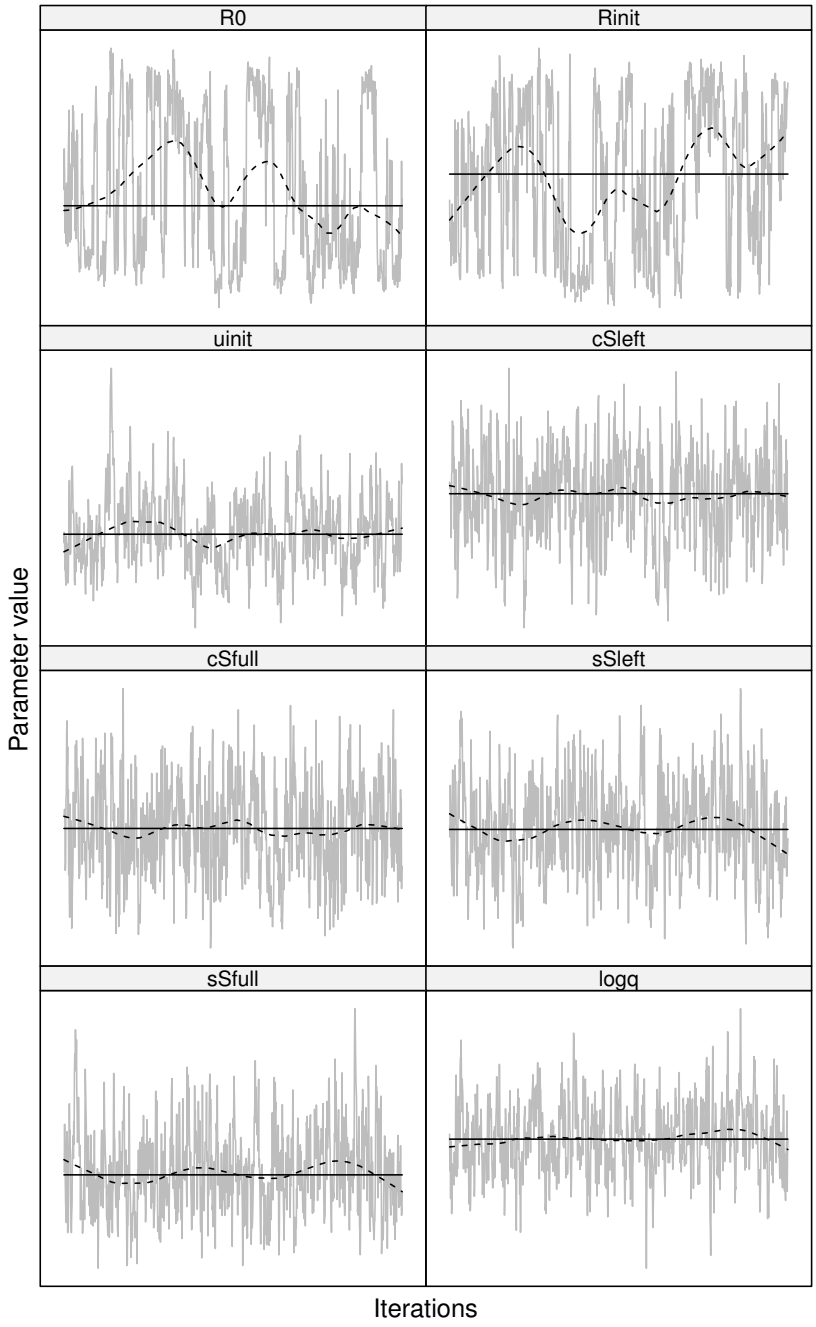

Figure 4.4. MCMC traces for parameters R_0, R_{init}, u_{init}, ${}_cS_{\text{left}}$, ${}_cS_{\text{full}}$, ${}_sS_{\text{left}}$, ${}_sS_{\text{left}}$, and $\log q$. This corresponds to the output of the first example on the `plotTrace` help page.

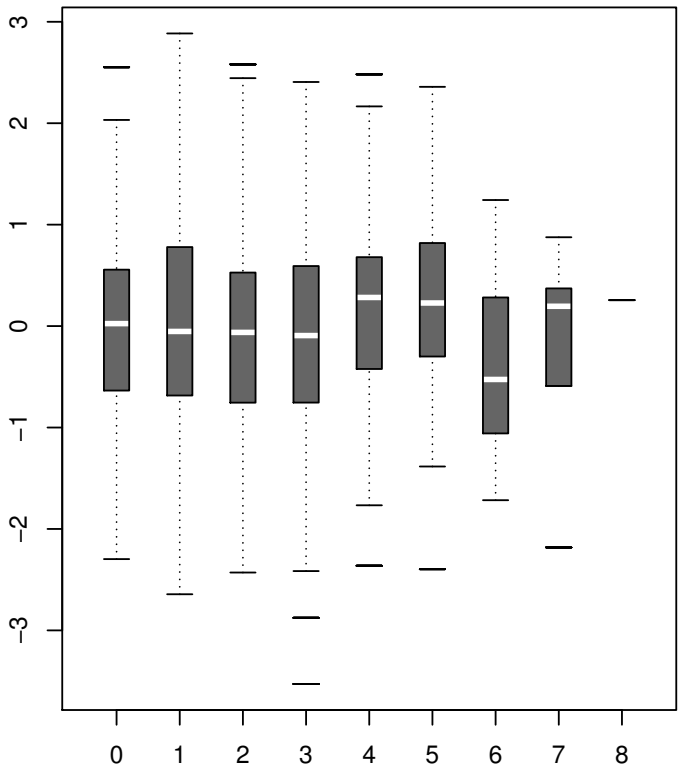

Figure 4.5. Boxplot with non-default format, as a result of user-specified graphical parameters. This corresponds to the output of an example on the `bxp` help page.

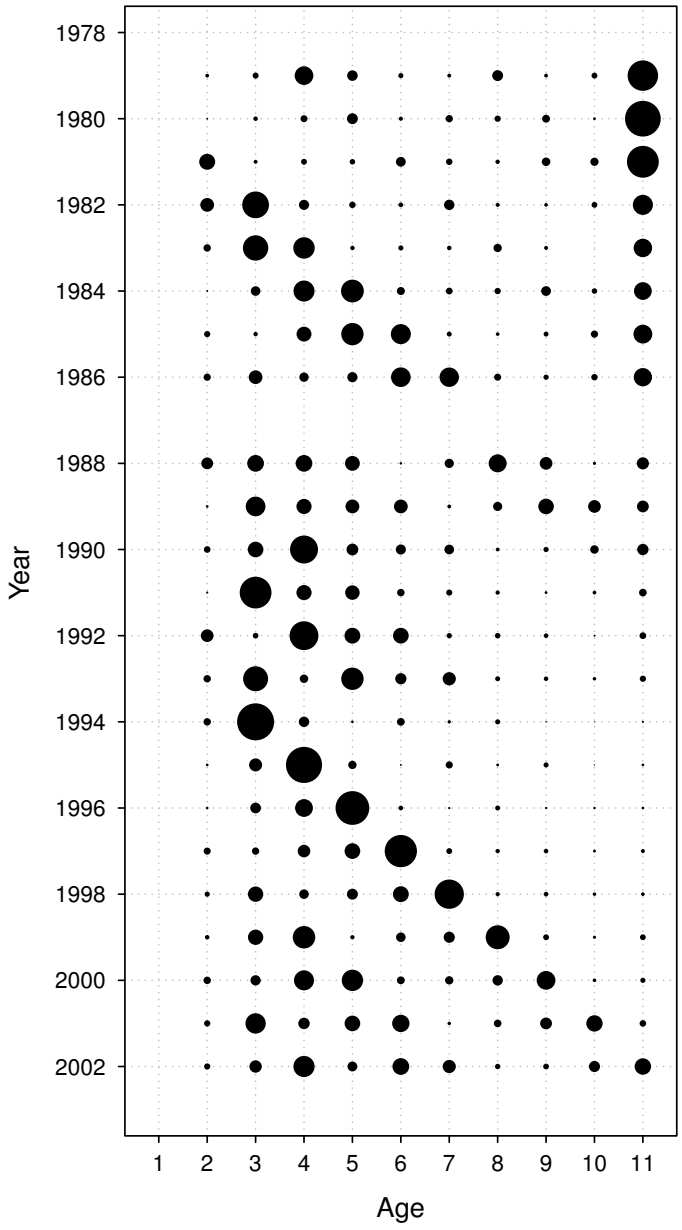

Figure 4.6. Bubble plot showing observed catch at age. The area of each bubble reflects the proportion of fish at that age in the catch, with no data available from 1987. This corresponds to the output of an example on the `plotCA` help page.

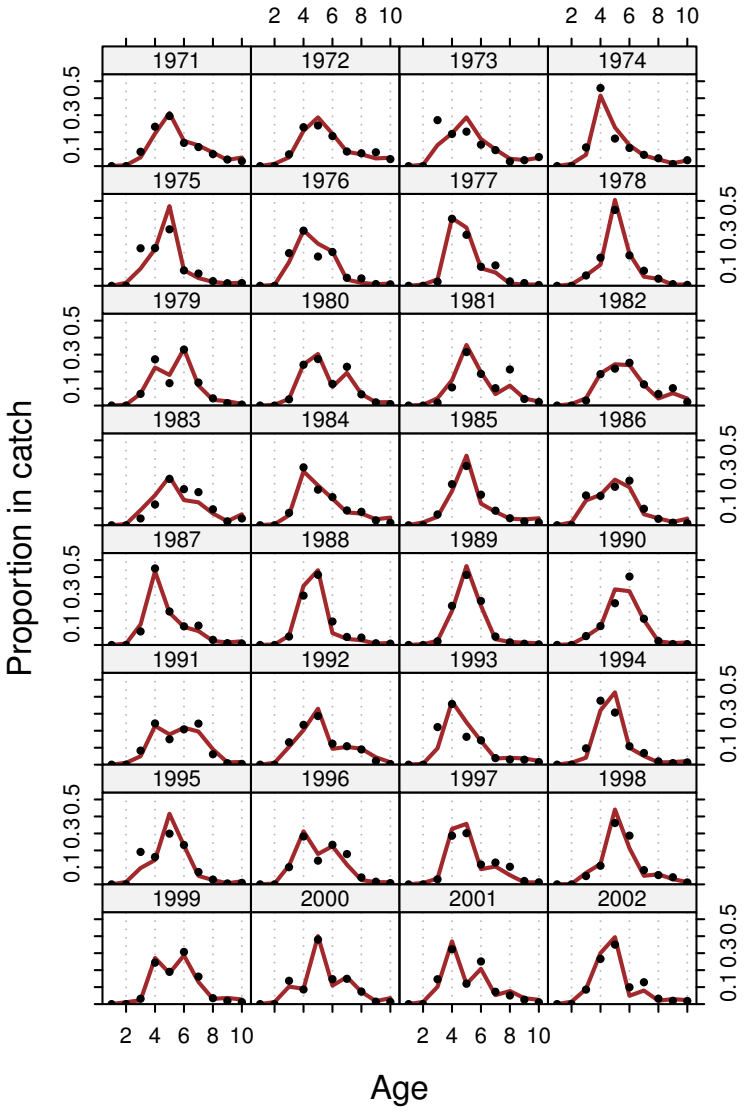

Figure 4.7. Multipanel plot showing model fit to catch at age. This corresponds to the output of an example on the `plotCA` help page.

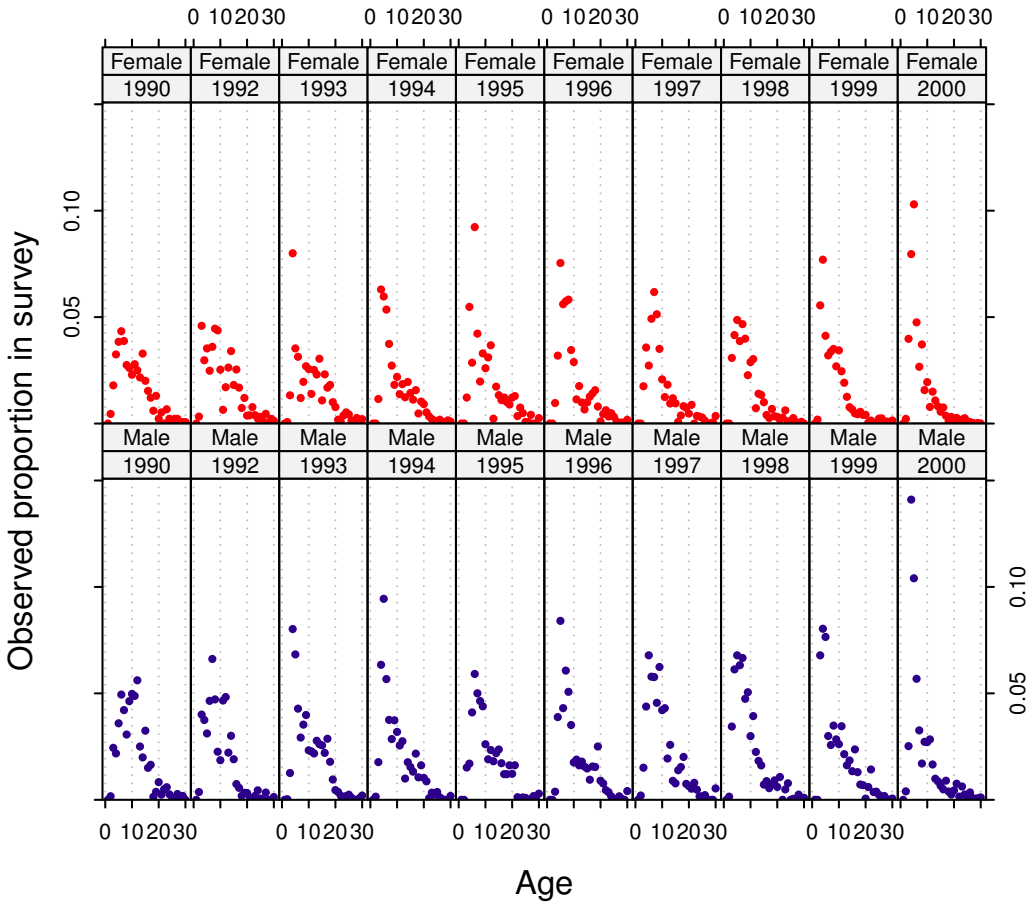

Figure 4.8. Multipanel plot showing observed survey catch at age of females and males. This corresponds to the output of an example on the `plotCA` help page.

CONCLUSIONS

The chapters in this dissertation share a common theme, which is uncertainty in fisheries stock assessment. As a whole, the dissertation provides guidelines for stock assessment practitioners, as a compendium of methods, recommendations, and caveats. The analysis in Chapter 1 sheds light on what kind of fisheries data are informative about stock status and other quantities of interest. Chapter 2 tests the performance of different methods for evaluating uncertainty, to identify which approaches can be recommended in stock assessment. Chapter 3 acts as a synthesis of the previous chapters, as well as broadening the scope to cover additional methods to confront uncertainty, using the Icelandic saithe fishery as a case study.

Among the findings in Chapter 1 is that a 'one-way trip' scenario, where harvest rate gradually increases while abundance decreases, proved no less informative than a contrasted catch history. Although strong depletion is to be avoided due to the economic cost and risk of collapse, it does provide informative data. This observation is encapsulated in the words of John Pope, that 'the more fish you catch, the better you know how many there were'. Data from a well-managed stock, where fluctuations in stock size and harvest rate are avoided, are therefore not informative for stock assessment unless they include an earlier period with greater contrast. The fishing history also affects the ability to estimate key parameters. Data can be expected to be informative about the stock status and M if they include years of varying fishing intensity, and informative about h if they include years of very small as well as moderate stock size.

The benchmark results from Chapter 2 suggest that the delta method and Markov chain Monte Carlo (MCMC) can be expected to evaluate uncertainty in stock assessment more accurately than the bootstrap. All the methods, however, can be expected to give intervals

that are too narrow in general. Bias correction improved the bootstrap performance, but not enough to match the performance of the delta method and MCMC, which were clear 'winners' in this benchmark. It is important to keep in mind, though, that the analysis is based on one particular estimation model applied to an artificial suite of data, using parametric model-conditioned bootstrap. For different estimation models, other variations of the bootstrap have been reported to perform well in simulation studies. The general recommendation, therefore, is to use simulations to analyze the expected performance of different uncertainty methods given a specific model and data scenarios of interest.

The case study in Chapter 3 presents a fishing history that is not informative about natural mortality rate M, as a result of too little contrast in fishing intensity. Diagnostic model runs give a point estimate of M at 0.57 yr^{-1}, almost three times higher than the traditionally assumed value of 0.20 yr^{-1}. The diagnostic models estimating M also estimate the stock size as being infinitely large, with a harvest rate of 0.00. In other words, these diagnostic models successfully estimate the average total mortality rate, but fail to partition the total mortality rate of 0.57 yr^{-1} between natural and fishing mortalities. The data suggest a high value of steepness h, but are not informative enough for reliable estimation, with the diagnostic model run essentially running into the upper bound of $h = 1.0$. The key parameters M and h have a direct effect on the estimated optimal harvest rate, so the basis of the long-term advice becomes subjective when these parameters cannot be estimated. The survey data in the Chapter 3 case study are characterized by a high level of measurement noise, as reflected in the interannual variability. Nevertheless, the analysis shows that these survey data reduce the overall uncertainty about the stock status by around 60% when compared to no survey data. As for the results from the uncertainty analysis, the bootstrap is clearly the odd one out, while the delta method, MCMC, and profile likelihood intervals are comparable in most cases. Profile likelihood is well-suited to evaluate the uncertainty about key parameters such as M and h, but slightly cumbersome to analyze derived quantities such as biomass and harvest rate. In the end, uncertainty analysis is not only about evaluating probabilities and confidence intervals, but iterative methods such as the bootstrap, MCMC, and profile

likelihood can also indicate lack of model convergence, find a new global optimum, identify highly correlated or ill-defined parameters, and suggest parameter transformation.

To summarize, a checklist of general recommendations emerging from this dissertation are:

- Use more than one method to evaluate uncertainty.
- Keep in mind that the real uncertainty is greater than the analytical confidence intervals indicate.
- Use more than one model and variations of models to evaluate how sensitive the main conclusions are to alternative assumptions.
- Use retrospective analysis to evaluate uncertainty from an empirical viewpoint.
- Use simulation analysis to evaluate the performance of the estimation model, which parameters can be estimated reliably, and which uncertainty methods work best.
- Examine the fishing history to evaluate whether the data are likely to be informative about the stock status and key parameters like h and M.
- Consider ways to reduce uncertainty by generating informative data via management (e.g., applying different fishing mortalities between years) and research (e.g., design a dedicated survey for a given stock, sample age data).
- Harvest control rules can be a practical way to incorporate uncertainty into management advice.

BIBLIOGRAPHY

Ascough, J.C., II, H.R. Maier, J.K. Ravalico, and M.W. Strudley. 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecological Modelling 219:383–399.

Asmussen, S. and P.W. Glynn. 2007. Stochastic simulation: Algorithms and analysis. New York: Springer.

Banks, J. (ed.). 1998. Handbook of simulation: Principles, methodology, advances, applications, and practice. New York: Wiley.

Bard, Y. 1974. Nonlinear parameter estimation. New York: Academic.

Bence, J.R., A. Gordoa, and J.E. Hightower. 1993. Influence of age-selective surveys in the reliability of stock synthesis assessments. Canadian Journal of Fisheries and Aquatic Sciences 50:827–840.

Beverton, R.J.H. and S.J. Holt. 1957. On the dynamics of exploited fish populations. London: Her Majesty's Stationery Office.

Bivand, R.S., E.J. Pebesma, and V. Gomez-Rubio. 2013. Applied spatial data analysis with R. 2nd ed. New York: Springer.

Björnsson, H. 2013. Evaluation of the Icelandic haddock management plan. ICES CM 2013/ACOM:59.

Bjornsson, H. and A. Magnusson. 2009. ADCAM user manual (draft version). ICES CM 2009/ACOM:56, Annex 6.

Bolker, B., B. Gardner, M. Maunder, C. Berg, M. Brooks, L. Comita, E. Crone, S. Cubaynes, T. Davies, P. de Valpine, J. Ford, O. Gimenez, M. Kéry, E. Kim, C. Lennert-Cody, A. Magnusson, S. Martell, J. Nash, A. Nielsen, J. Regetz, H. Skaug, and E. Zipkin. 2013. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods in Ecology and Evolution 4:501–512.

Booth, A.J. and T.J. Quinn II. 2006. Maximum likelihood and Bayesian approaches to stock assessment when data are questionable. Fisheries Research 80:169–181.

Butterworth, D.S. and A.E. Punt. 1999. Experiences in the evaluation and implementation of management procedures. ICES Journal of Marine Science 56:985–998.

Casella, G. and R.L. Berger. 2002. Statistical inference. 2nd ed. Pacific Grove: Duxbury.

Chapman, D.G. and D.S. Robson. 1960. The analysis of a catch curve. Biometrics 16:354–368.

Charles, A. 2001. Sustainable fishery systems. Oxford: Blackwell.

Chen, Y., L. Chen, and K.I. Stergiou. 2003. Impacts of data quantity on fisheries stock assessment. Aquatic Sciences 65:92–98.

Clark, J.S. 2005. Why environmental scientists are becoming Bayesians. Ecology Letters 8:2–14.

Clark, W.G. 1999. Effects of an erroneous natural mortality rate on a simple age-structured stock assessment. Canadian Journal of Fisheries and Aquatic Sciences 56:1721–1731.

Conn, P.B., E.H. Williams, and K.W. Shertzer. 2010. When can we reliably estimate the productivity of fish stocks? Canadian Journal of Fisheries and Aquatic Sciences 67:511–523.

Cotter, A.J.R., L. Burt, C.G.M. Paxton, C. Fernandez, S.T. Buckland, and J.X. Pan. 2004. Are stock assessment methods too complicated? Fish and Fisheries 5:235–254.

Cramér, H. 1946. Mathematical methods of statistics. Princeton: Princeton University Press.

De Oliveira, J.A.A., L.T. Kell, L.T., A.E. Punt, B.A. Roel, and D.S. Butterworth. 2008. Managing without best predictions: The Management Strategy Evaluation framework. In: A. Payne et al. (eds.) Advances in fisheries science: 50 years on from Beverton and Holt. Oxford: Blackwell, pp. 104–134.

Deriso, R.B. 1980. Harvesting strategies and parameter estimation for an age-structured model. Canadian Journal of Fisheries and Aquatic Sciences 37:268–282.

Deriso, R.B., T.J. Quinn II, and P.R. Neal. 1985. Catch-age analysis with auxiliary information. Canadian Journal of Fisheries and Aquatic Sciences 42:815–824.

DiCiccio, T.J. and B. Efron. 1996. Bootstrap confidence intervals. Statistical Science 11:189–212.

Dorn, M.W. 2002. Advice on West Coast rockfish harvest rates from Bayesian meta-analysis of stock-recruit relationships. North American Journal of Fisheries Management 22:280–300.

Doubleday, W.G. 1976. A least squares approach to analysing catch at age data. ICNAF Research Bulletin 12:69–81.

Edwards, A.W.F. 1992. Likelihood. 2nd ed. Baltimore: Johns Hopkins University Press.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. Annals of Statistics 7:1–26.

———. 2000. The bootstrap and modern statistics. Journal of the American Statistical Association 95:1293–1296.

———. 2003. Second thoughts on the bootstrap. Statistical Science 18:135–140.

Efron, B. and R. Tibshirani. 1993. An introduction to the bootstrap. New York: Chapman and Hall.

Ellison, A.M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6:1036–1041.

Elvarsson, B.T., L. Taylor, V. Kupca, and G. Stefansson. 2014. A bootstrap method for estimating bias and variance in statistical fisheries modelling frameworks using highly disparate datasets. African Journal of Marine Science 36:99–110.

Ernst, B. 2002. An investigation on length-based models used in quantitative population modeling. Ph.D. thesis, University of Washington.

Fournier, D. and C.P. Archibald. 1982. A general theory for analyzing catch at age data. Canadian Journal of Fisheries and Aquatic Sciences 39:1195–1207.

Fournier, D.A., H.J. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M.N. Maunder, A. Nielsen, and J. Sibert. 2012. AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software 27:233–249.

Fournier, D.A., J.R. Sibert, J. Majkowski, and J. Hampton. 1990. MULTIFAN: A likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). Canadian Journal of Fisheries and Aquatic Sciences 47:301–317.

Francis, R.I.C.C. 1992. Use of risk analysis to assess fishery management strategies: A case study using orange roughy (*Hoplostethus atlanticus*) on the Chatham Rise, New Zealand. Canadian Journal of Fisheries and Aquatic Sciences 49:922–930.

Gavaris, S. 1988. An adaptive framework for the estimation of population size. Canadian Atlantic Fisheries Scientific Advisory Committee Research Document 88/29.

———. 1999. Dealing with bias in estimating uncertainty and risk. NOAA Technical Memorandum NMFS-F/SPO-40:46–50.

Gavaris, S. and J.N. Ianelli. 2002. Statistical issues in fisheries' stock assessments. Scandinavian Journal of Statistics 29:245–271.

Gavaris, S. and L. Van Eeckhaute. 1998. Assessment of haddock on eastern Georges Bank. CSAS Research Document 98/66.

Gavaris, S., K.R. Patterson, C.D. Darby, P. Lewy, B. Mesnil, A.E. Punt, R.M. Cook, L.T. Kell, C.M. O'Brien, V.R. Restrepo, D.W. Skagen, and G. Stefánsson. 2000. Comparison of uncertainty estimates in the short term using real data. ICES CM 2000/V:03.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin. 1995. Bayesian data analysis. London: Chapman and Hall.

———. 2004. Bayesian Data Analysis. 2nd ed. Boca Raton: Chapman and Hall.

Givens, G.H. and J.A. Hoeting. 2005. Computational statistics. Hoboken: Wiley.

Griewank, A. and G.F. Corliss (eds.) 1991. Automatic differentiation of algorithms: Theory, implementation, and application. Philadelphia: SIAM.

Gulland, J.A. 1965. Estimation of mortality rates. ICES CM 1965 Gadoid Fisheries Committee Document 3.

Haddon, M. 2003. To be Bayesian or to bootstrap: What is the risk? In: S.J. Newman et al. (eds.) Towards sustainability of data-limited multi-sector fisheries. Perth: Department of Fisheries, pp. 98–104.

Hamel, O.S. 2015. A method for calculating a meta-analytical prior for the natural mortality rate using multiple life history correlates. ICES Journal of Marine Science 72:62–69.

Harley, S.J., R.A. Myers, and A. Dunn. 2001. Is catch-per-unit-effort proportional to abundance? Canadian Journal of Fisheries and Aquatic Sciences 58:1760–1772.

Hastings, W.K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Hilborn, R. 1979. Comparison of fisheries control systems that utilize catch and effort data. Journal of the Fisheries Research Board of Canada 36:1477–1489.

———. 1990. Estimating the parameters of full age-structured models from catch and abundance data. INPFC Bulletin 50:207–213.

———. 2003. The state of the art in stock assessment: Where we are and where we are going. Scientia Marina 67(S1):15–20.

Hilborn, R. and C.J. Walters. 1992. Quantitative fisheries stock assessment: Choice, dynamics and uncertainty. New York: Chapman and Hall.

Hilborn, R. and M. Mangel. 1997. The ecological detective: Confronting models with data. Princeton: Princeton University Press.

Hilborn, R., M. Maunder, A. Parma, B. Ernst, J. Payne, and P. Starr. 2003. Coleraine: A generalized age-structured stock assessment model. User's manual version 2.0. University of Washington Report SAFS-UW 0116.

Hilborn, R., T.A. Branch, B. Ernst, A. Magnusson, C.V. Minte-Vera, M.D. Scheuerell, and J.L. Valero. 2003. State of the world's fisheries. Annual Review of Environment and Resources 28:359–399.

Hjörleifsson, E. and H. Björnsson. 2013. Evaluation of the Icelandic saithe management plan. ICES CM 2013/ACOM:60.

Ianelli, J.N. 2002. Simulation analyses testing the robustness of productivity determinations from West Coast Pacific Ocean perch stock assessment data. North American Journal of Fisheries Management 22:301–310.

ICES (International Council for the Exploration of the Sea). 2003. Report of the North Western Working Group (NWWG). ICES CM 2003/ACFM 24:144–227.

———. 2009. Icelandic cod HCR evaluation (AGICOD). ICES CM 2009/ACOM:56.

———. 2015a. Report of the benchmark workshop on Icelandic stocks (WKICE). ICES CM 2015/ACOM:31.

———. 2015b. Report of the North Western Working Group (NWWG). ICES CM 2015/ACOM:07.

Jakobsson, J. and G. Stefánsson. 1998. Rational harvesting of the cod-capelin-shrimp complex in the Icelandic marine ecosystem. Fisheries Research 37:7–21.

Johnson, K.F., C.C. Monnahan, C.R. McGilliard, K.A. Vert-pre, S.C. Anderson, C.J. Cunningham, F. Hurtado-Ferro, R.R. Licandeo, M.L. Muradian, K. Ono, C.S. Szuwalski, J.L. Valero, A.R. Whitten, and A.E. Punt. 2015. Time-varying natural mortality in fisheries stock assessment models: Identifying a default approach. ICES Journal of Marine Science 72:137–150.

Kass, R.E. 2011. Statistical inference: The big picture. Statistical Science 26:1–9.

Kimura, D.K. and J.V. Tagart. 1982. Stock reduction analysis: Another solution to the catch equations. Canadian Journal of Fisheries and Aquatic Sciences 39:1467–1472.

Lawson, T.A. and R. Hilborn. 1985. Equilibrium yields and yield isopleths from a general age-structured model of harvested populations. Canadian Journal of Fisheries and Aquatic Sciences 42:1766–1771.

Lee, H.H., M.N. Maunder, K.R. Piner, and R.D. Methot. 2011. Estimating natural mortality within a fisheries stock assessment model: An evaluation using simulation analysis based on twelve stock assessments. Fisheries Research 109:89–94.

Leslie, P.H. and D.H.S. Davis. 1939. An attempt to determine the absolute number of rats on a given area. Journal of Animal Ecology 8:94–113.

Ludwig, D. and R. Hilborn. 1983. Adaptive probing strategies for age-structured fish stocks. Canadian Journal of Fisheries and Aquatic Sciences 40:559–569.

Mace, P.M. 2001. A new role for MSY in single-species and ecosystem approaches to fisheries stock assessment and management. Fish and Fisheries 2:2–32.

Magnusson, A. 2005. R goes fishing: Analyzing fisheries data using AD Model Builder and R. Proceedings of the 4th International Workshop on Directions in Statistical Computing, Seattle 12–14 August 2005.

———. 2009. ADMB-IDE: Easy and efficient user interface. ADMB Foundation Newsletter 1(3):1–2.

———. 2013. Mathematical properties of the Icelandic saithe HCR. ICES North Western Working Group (NWWG) Working Document 31.

———. 2014. scape: Statistical catch-at-age plotting environment. R package version 2.2-0. `https://cran.r-project.org/package=scape`

———. 2015. AD Model Builder IDE: Emacs `admb-mode` without the Emacs. Version 11.2. `http://admb-project.org/tools/admb-ide`

Magnusson, A. and I. Stewart. 2014. plotMCMC: MCMC diagnostic plots. R package version 2.0-0. `https://cran.r-project.org/package=plotMCMC`

Magnusson, A. and J. Burgos. 2014. r2d2: Bivariate confidence region and frequency distribution. R package version 1.0-0. `https://cran.r-project.org/package=r2d2`.

Magnusson, A. and R. Hilborn. 2007. What makes fisheries data informative? Fish and Fisheries 8:337–358.

Magnusson, A., A.E. Punt, and R. Hilborn. 2013. Measuring uncertainty in fisheries stock assessment: The delta method, bootstrap, and MCMC. Fish and Fisheries 14:325–342.

Mangel, M., J. Brodziak, and G. DiNardo. 2010. Reproductive ecology and scientific inference of steepness: A fundamental metric of population dynamics and strategic fisheries management. Fish and Fisheries 11:89–104.

Maunder, M.N. 2003. Is it time to discard the Schaefer model from the stock assessment scientist's toolbox? Fisheries Research 61:145–149.

Maunder, M.N. and K.R. Piner. 2015. Contemporary fisheries stock assessment: Many issues still remain. ICES Journal of Marine Science 72:7–18.

Maunder, M.N., J.T. Schnute, and J.N. Ianelli. 2009. Computers in fisheries population dynamics. In: B.A. Megrey and E. Moksness (eds.) Computers in fisheries research. 2nd ed. New York: Springer, pp. 337–372.

McAllister, M.K. and J.N. Ianelli. 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences 54:284–300.

McAllister, M.K., E.K. Pikitch, A.E. Punt, and R. Hilborn. 1994. A Bayesian approach to stock assessment and harvest decisions using the sampling/importance resampling algorithm. Canadian Journal of Fisheries and Aquatic Sciences 51:2673–2687.

McGarvey, R., J.E. Feenstra, and Q. Ye. 2007. Modeling fish numbers dynamically by age and length: Partitioning cohorts into "slices". Canadian Journal of Fisheries and Aquatic Sciences 64:1157–1173.

Megrey, B.A. 1989. Review and comparison of age-structured stock assessment models from theoretical and applied points of view. In: E.F. Edwards and B.A. Megry (eds.) Mathematical analysis of fish stock dynamics. Bethesda: American Fisheries Society, pp. 8–48.

Mertz, G. and R.A. Myers. 1997. Influence of errors in natural mortality estimates in cohort analysis. Canadian Journal of Fisheries and Aquatic Sciences 54:1608–1612.

Methot, R.D. 1989. Synthetic estimates of historical abundance and mortality for northern anchovy. In: E.F. Edwards and B.A. Megrey (eds.) Mathematical analysis of fish stock dynamics. Bethesda: American Fisheries Society, pp. 66–82.

Metropolis, N. and S. Ulam. 1949. The Monte Carlo method. J. Am. Stat. Assoc. 44:335–341.

Millar, R.B. 2011. Maximum likelihood estimation and inference: With examples in R, SAS and ADMB. Chichester: Wiley.

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. Journal of Chemical Physics 21:1087–1092.

Mohn, R.K. 1993. Bootstrap estimates of ADAPT parameters, their projection in risk analysis and their retrospective patterns. Canadian Special Publication in Fisheries and Aquatic Sciences 120:173–184.

———. The uncertain future of assessment uncertainty. In: R.J. Beamish and B.J. Rothschild (eds.) The future of fisheries science in North America. New York: Springer, pp. 495–504.

MRI (Marine Research Institute). 2015. State of marine stocks in Icelandic waters 2014/2015 and prospects for the quota year 2015/2016. Marine Research in Iceland 182.

Myers, R.A., K.G. Bowen, and N.J. Barrowman. 1999. Maximum reproductive rate of fish at low population sizes. Canadian Journal of Fisheries and Aquatic Sciences 56:2404–2419.

Neyman, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society A 236:333–380.

NRC (National Research Council). 1998. Improving fish stock assessments. Washington: National Academy.

Oehlert, G.W. 1992. A note on the delta method. The American Statistician 46:27–29.

Pálsson, O.K., E. Jonsson, S.A. Schopka, G. Stefansson, and B.A. Steinarsson. 1989. Icelandic groundfish survey data used to improve precision in stock assessments. Journal of Northwest Atlantic Fishery Science 9:53–72.

Patterson, K., R. Cook, C. Darby, S. Gavaris, L. Kell, P. Lewy, B. Mesnil, A. Punt, V. Restrepo, D.W. Skagen, and G. Stefánsson. 2001. Estimating uncertainty in fish stock assessment and forecasting. Fish and Fisheries 2:125–157.

Patterson, K.R. 1999. Evaluating uncertainty in harvest control law catches using Bayesian Markov chain Monte Carlo virtual population analysis with adaptive rejection sampling and including structural uncertainty. Canadian Journal of Fisheries and Aquatic Sciences 56:208–221.

Pauly, D., R. Hilborn, and T.A. Branch. 2013. Does catch reflect abundance? Nature 494:303–306.

Peck, S.L. 2004. Simulation as experiment: A philosophical reassessment for biological modeling. Trends Ecol. Evol. 19:530–534.

Plummer, M., N. Best, K. Cowles, and K. Vines. 2006. CODA: Convergence diagnosis and output analysis for MCMC. R News 6(1):7–11.

Plummer, M., N. Best, K. Cowles, K. Vines, D. Sarkar, D. Bates, R. Almond, and A. Magnusson. 2015. coda: Output analysis and diagnostics for MCMC. R package version 0.18-1. `https://cran.r-project.org/package=coda`

Pope, J.G. 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. ICNAF Research Bulletin 9:65–74.

———. 1988. Collecting fisheries assessment data. In: J.A. Gulland (ed.) Fish population dynamics. 2nd ed. Chichester: Wiley, pp. 63–82.

Punt, A.E. and R. Hilborn. 1997. Fisheries stock assessment and decision analysis: The Bayesian approach. Reviews in Fish Biology and Fisheries 7:35–63.

Punt, A.E. and A.D.M. Smith. 2001. The gospel of maximum sustainable yield in fisheries management: Birth, crucifixion and reincarnation. In: J.D. Reynolds et al. (eds.) Conservation of exploited species. Cambridge: Cambridge University Press, pp. 41–66.

Punt, A.E. and D.S. Butterworth. 1993. Variance estimates for fisheries assessment: Their importance and how best to evaluate them. Canadian Special Publication in Fisheries and Aquatic Sciences 120:145–162.

Punt, A.E. and R.B. Kennedy. 1997. Population modeling of Tasmanian rock lobster, *Jasus edwardsii*, resources. Marine and Freshwater Research 48:967–980.

Punt, A.E., A.D.M. Smith, and G. Cui. 2002. Evaluation of management tools for Australia's South East Fishery 2: How well can management quantities be estimated? Marine and Freshwater Research 53:631–644.

Quinn, T.J., II. 2003. Ruminations on the development and future of population dynamics models in fisheries. Natural Resource Modeling 16:341–392.

Quinn, T.J., II and R.B. Deriso. 1999. Quantitative Fish Dynamics. New York: Oxford University Press.

R Core Team. 2015. R: A language and environment for statistical computing. Vienna: R Foundation.

Ralston, S., A.E. Punt, O.S. Hamel, J.D. DeVore, and R.J. Conser. 2011. A meta-analytic approach to quantifying scientific uncertainty in stock assessments. Fishery Bulletin 109:217–231.

Restrepo, V.R., K.R. Patterson, C.D. Darby, S. Gavaris, L.T. Kell, P. Lewy, B. Mesnil, A.E. Punt, R.M. Cook, C.M. O'Brien, D.W. Skagen, and G. Stefánsson. 2000. Do different methods provide accurate probability statements in the short term? ICES CM 2000/V:08.

Ricker, W.E. 1958. Handbook of computations for biological statistics of fish populations. Bulletin of the Fisheries Research Board of Canada 119.

Robert, C.P. and G. Casella. 2010. Introducing Monte Carlo methods with R. New York: Springer.

Rose, K.A., J.H. Cowan Jr., K.O. Winemiller, R.A. Myers, and R. Hilborn. 2001. Compensatory density dependence in fish populations: Importance, controversy, understanding and prognosis. Fish and Fisheries 2:293–327.

Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn. 1989. Design and analysis of computer experiments. Stat. Sci. 4:409–423.

Sampson, D.B. and Y. Yin. 1998. A Monte Carlo evaluation of the stock synthesis assessment program. In: F. Funk et al. (eds.) Fishery stock assessment models. Fairbanks: Sea Grant Program, pp. 315–338.

Santner, T.J., B.J. Williams, and W.I. Notz. 2003. The design and analysis of computer experiments. New York: Springer.

Sarkar, D. 2008. Lattice: Multivariate data visualization with R. New York: Springer.

Schaefer, M.B. 1954. Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. IATTC Bulletin 1:27–56.

Schnute, J.T. 1985. A general theory for analysis of catch and effort data. Canadian Journal of Fisheries and Aquatic Sciences 42:414–429.

Schnute, J.T. and R. Hilborn. 1993. Analysis of contradictory data sources in fish stock assessments. Canadian Journal of Fisheries and Aquatic Sciences 50:1916–1923.

Schnute, J.T., L.J. Richards, and N. Olsen. 1998. Statistics, software, and fish stock assessment. In: F. Funk et al. (eds.) Fishery stock assessment models. Fairbanks: Sea Grant Program, pp. 171–184.

Seber, G.A.F. 1973. The estimation of animal abundance and related parameters. London: Griffin.

Seber, G.A.F. and C.J. Wild. 1989. Nonlinear regression. Hoboken: Wiley.

Shepherd, J.G. 1984. The availability and information content of fisheries data. In: R.M. May (ed.) Exploitation of marine communities. Berlin: Springer, pp. 95–109.

———. 1999. Extended survivors analysis: An improved method for the analysis of catch-at-age data and abundance indices. ICES Journal of Marine Science 56:584–591.

Smith, M.T. and J.T. Addison. 2003. Methods for stock assessment of crustacean fisheries. Fisheries Research 65:231–256.

Stewart, I.J., A.C. Hicks, I.G. Taylor, J.T. Thorson, C. Wetzel, and S. Kupschus. 2013. A comparison of stock assessment uncertainty estimates using maximum likelihood and Bayesian methods implemented with the same model framework. Fish. Res. 142:37–46.

Stigler, S.M. 1991. Stochastic simulation in the nineteenth century. Stat. Sci. 6:89–97.

Then, A.Y., J.M. Hoenig, N.G. Hall, and D.A. Hewitt. 2015. Evaluating the predictive performance of empirical estimators of natural mortality rate using information on over 200 fish species. ICES Journal of Marine Science 72:82–92.

Thompson, G.G. 1994. Confounding of gear selectivity and the natural mortality rate in cases where the former is a nonmonotone function of age. Canadian Journal of Fisheries and Aquatic Sciences 51:2654–2664.

Tippmann, S. 2015. Programming tools: Adventures with R. Nature 517:109–110.

Trzcinski, M.K., R. Mohn, and W.D. Bowen. 2006. Continued decline of an Atlantic cod population: How important is gray seal predation? Ecological Applications 16:2276–2292.

Tukey, J.W. 1977. Exploratory data analysis. Reading, MA: Addison-Wesley.

———. 1997. More honest foundations for data analysis. Journal of Statistical Planning and Inference 57:21–28.

Venzon, D.J. and S.H. Moolgavkar. 1988. A method for computing profile-likelihood-based confidence intervals. Applied Statistics 37:87–94.

Virtala, M., S. Kuikka, and E. Arjas. 1998. Stochastic virtual population analysis. ICES Journal of Marine Science 55:892–904.

Walters, C. 1986. Adaptive management of renewable resources. New York: Macmillan.

———. 2007. Is adaptive management helping to solve fisheries problems? Ambio 36:304–307.

Walters, C.J. and S.J.D. Martell. 2004. Fisheries ecology and management. Princeton: Princeton University Press.

Wand, M.P. and M.C. Jones. 1995. Kernel smoothing. London: Chapman and Hall.

Winsberg, E.B. 2010. Science in the age of computer simulation. Chicago: University of Chicago Press.

Xiao, Y. 2000. A general theory of fish stock assessment models. Ecological Modelling 128:165–180.

Yin, Y. and D.B. Sampson. 2004. Bias and precision of estimates from an age-structured stock assessment program in relation to stock and data characteristics. North American Journal of Fisheries Management 24:865–879.

Appendix A

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

A.1 Supplementary figures and tables

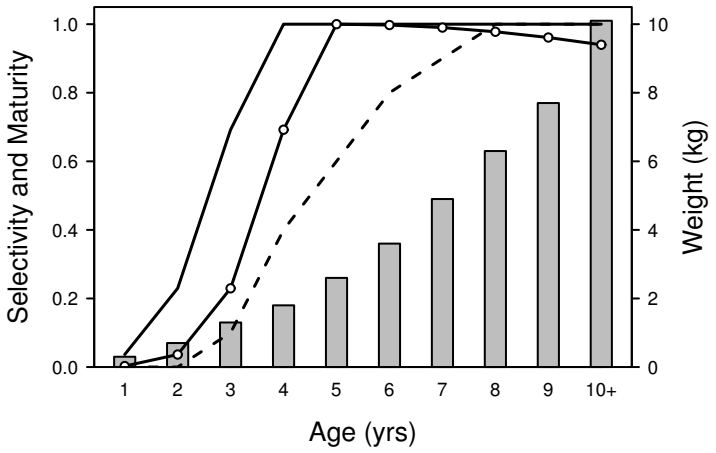

Figure A.1. Age-specific characteristics of the operating model: survey selectivity (plain line), commercial selectivity (line with circles), maturity (dashed line), and weight (bars).

Table A.1. Age-specific weight (kg) and maturity (proportion) used in the operating and estimation model.

Age	1	2	3	4	5	6	7	8	9	10+
Weight (kg)	0.3	0.7	1.3	1.8	2.6	3.6	4.9	6.3	7.7	10.1
Maturity	0.0	0.0	0.1	0.4	0.6	0.8	0.9	1.0	1.0	1.0

Table A.2. Parameter values used in the operating model, along with bounds used in the estimation model.

Parameter	Meaning	True value	Lower bound	Upper bound
R_0	Average virgin recruitment	250 000	1 000	10 000 000
h	Recruitment steepness	0.7	0.2	1
M	Natural mortality rate	0.2	0	0.5
R_{init}	Initial population scaler	1	0	5
u_{init}	Initial harvest rate	0	0	1
R_{plus}	Initial plus group scaler	1	0	2
${}_{C}S_{\text{full}}$	Age at full selectivity (fleet)	5	3	10
${}_{C}S_{\text{left}}$	Selectivity left curve (fleet)	1	−2	5
${}_{C}S_{\text{right}}$	Selectivity right curve (fleet)	6	−2	15
${}_{S}S_{\text{full}}$	Age at full selectivity (survey)	4	2	10
${}_{S}S_{\text{left}}$	Selectivity left curve (survey)	1	−2	5
$\log q$	Catchability coefficient	−15.2	−30	0
${}_{R}\varepsilon$	Recruitment deviates	*	−15	15

*: Initial age structure and annual recruitment varies between scenarios.

Table A.3. Annual harvest rate and recruitment used in the operating model.

	Harvest	Recruitment in scenario number									
Year	rate	1	2	3	4	5	6	7	8	9	10
1985	0.039	143	122	117	238	126	245	824	198	132	211
1986	0.050	145	405	299	324	62	119	297	241	363	130
1987	0.064	215	248	161	200	405	221	306	64	172	261
1988	0.079	121	174	384	172	307	217	203	74	306	108
1989	0.096	228	133	290	198	231	115	104	68	103	305
1990	0.114	138	123	367	215	195	197	251	202	293	134
1991	0.132	583	700	283	74	344	178	79	145	125	271
1992	0.151	305	355	220	519	173	184	370	276	203	282
1993	0.168	334	663	62	319	60	271	373	150	70	69
1994	0.184	238	151	127	101	123	152	327	200	111	339
1995	0.199	510	155	252	67	408	107	208	245	225	354
1996	0.209	125	103	141	240	104	122	146	154	197	82
1997	0.207	138	163	154	285	476	104	242	99	191	340
1998	0.193	428	250	137	299	826	64	266	222	231	129
1999	0.168	123	420	451	207	184	317	256	83	152	140
2000	0.138	159	256	108	158	74	122	103	378	271	293
2001	0.107	143	237	124	217	277	222	441	362	50	178
2002	0.080	150	373	164	106	71	157	83	187	78	319
2003	0.053	154	359	190	121	245	221	345	385	120	270
2004	0.023	247	194	191	496	196	86	326	82	238	375

Table A.4. Coverage probability for confidence intervals by uncertainty method and reference point, evaluated at several confidence levels. The non-bias-corrected bootstrap is referred to as "raw", and bias-corrected bootstrap as "bootstrap". Ideally, the coverage probability should equal the confidence level. The bottom part of the table averages across reference points.

Reference point	Conf. level	Delta	Raw	Bootstrap	MCMC
B_{current}	50%	33.0	24.4	32.0	33.3
	75%	56.3	40.9	48.8	51.3
	90%	74.7	58.5	65.0	67.7
	95%	82.7	69.7	73.4	76.2
u_{current}	50%	30.4	17.8	33.9	32.0
	75%	52.3	29.2	51.0	49.7
	90%	72.7	44.9	66.5	63.0
	95%	83.6	54.7	73.0	72.6
Depletion	50%	35.9	24.0	29.3	35.2
	75%	56.9	38.4	49.2	54.6
	90%	72.9	52.2	64.9	68.8
	95%	80.5	60.7	72.7	76.5
MSY	50%	45.1	18.7	20.1	45.6
	75%	69.0	33.3	32.5	68.8
	90%	82.7	46.4	45.6	83.2
	95%	89.4	55.4	52.7	88.4
$B_{\text{current}}/B_{\text{MSY}}$	50%	24.3	35.6	29.3	34.9
	75%	37.0	54.8	49.0	52.5
	90%	48.4	71.2	65.6	67.2
	95%	55.4	79.0	72.3	74.4
Surplus	50%	43.0	35.0	37.3	44.0
	75%	68.0	55.3	59.8	67.2
	90%	86.3	71.4	76.8	85.0
	95%	91.4	80.2	84.0	90.1
Average	50%	35.3	25.9	30.3	37.5
	75%	56.6	42.0	48.4	57.4
	90%	73.0	57.4	64.1	72.5
	95%	80.5	66.6	71.4	79.7

Table A.5. Coverage probability for 90% confidence intervals (cf. bottom line in Table 2.3), when the computations are repeated while leaving out one recruitment scenario at a time. The best-performing method is shown in boldface. Adding one or more recruitment scenarios might alter the relative rank of the delta method and MCMC, but not the ranking of the two bootstrap methods.

	Delta	Raw	Bootstrap	MCMC
Include all	**73.0**	57.4	64.1	72.5
Leave out 1	**73.4**	57.6	64.6	72.6
Leave out 2	**74.3**	59.2	64.7	72.8
Leave out 3	72.6	57.4	63.1	**72.6**
Leave out 4	**72.6**	57.4	63.6	71.7
Leave out 5	**73.0**	57.7	64.5	72.9
Leave out 6	72.2	54.9	63.5	**72.4**
Leave out 7	**73.1**	59.7	64.9	72.6
Leave out 8	72.6	55.6	63.1	**73.0**
Leave out 9	**72.9**	58.3	63.7	72.3
Leave out 10	**72.8**	56.5	64.8	71.9

Appendix B

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

B.1 Estimation model

The population dynamics are governed by the equation,

$$N_{t+1,a+1} = N_{t,a}e^{-M}(1 - {}_{C}S_a u_t^{\star}) \tag{B.1}$$

where $N_{t,a}$ is population size at time t and age a, M is the rate of natural mortality, ${}_{C}S$ is the selectivity of the commercial fishery and $u^{\star}$ is harvest rate. The oldest age group, age A, is treated as a plus group:

$$N_{t+1,A} \;=\; N_{t,A-1}\,e^{-M}(1-{}_{C}S_{A-1}u_t^{\star}) \;+\; N_{t,A}\,e^{-M}(1-{}_{C}S_A u_t^{\star}) \tag{B.2}$$

Selectivity is an asymmetric normal curve determined by three shape parameters,

$$S_a \;=\; \begin{cases} \exp\left(\dfrac{-(a - S_{\text{full}})^2}{\exp(S_{\text{left}})}\right), & a \leq S_{\text{full}} \\[2ex] \exp\left(\dfrac{-(a - S_{\text{full}})^2}{\exp(S_{\text{right}})}\right), & a > S_{\text{full}} \end{cases} \tag{B.3}$$

where S_{full} is the age at full selectivity, S_{left} describes the left-hand slope and S_{right} the right hand slope of the curve. Harvest rate is defined as the fraction removed from the vulnerable biomass in the middle of the fishing year, $u_t^{\star} = Y_t \,/\, \sum_a ({}_{C}S_a N_{t,a}\, w_{t,a}) e^{-M/2}$, where Y is catch and w is body weight. For the purposes of the harvest control rule, however, the term harvest rate is used in the context of the 20% harvest control rule, which defines harvest rate as the annual catch divided by the reference biomass of ages 4 and older at the beginning of the year, $u_t = Y_t / \sum_{a=4}^{A} (N_{t,a} w_{t,a})$.

The population size at the start of the first year is

$$
\begin{aligned}
N_{1,1} &= R_0\, R_{\text{init}} \times \exp({}_R\varepsilon_{1,1} - \sigma_R^2/2) \\
N_{1,a} &= R_0\, R_{\text{init}}\, e^{-(a-1)M} \prod_{i=1}^{a-1} (1 - {}_C S_i u^\star_{\text{init}}) \times \exp({}_R\varepsilon_{1,a} - \sigma_R^2/2) \\
N_{1,A} &= R_0\, R_{\text{init}}\, e^{-(A-1)M} \prod_{i=1}^{A-1} (1 - {}_C S_i u^\star_{\text{init}}) \,/\, [1 - e^{-M}(1 - {}_C S_A u^\star_{\text{init}})] \times R_{\text{plus}}
\end{aligned}
\tag{B.4}
$$

for 1-year-olds, intermediate ages, and the plus group. R_0 is average virgin recruitment, R_{init} scales the initial population size across all ages, $u^\star_{\text{init}}$ is the initial harvest rate, and R_{plus} scales the initial plus group.

Recruitment is stochastic around a Beverton-Holt stock-recruitment function, reparametrized according to Francis (1992),

$$
N_{t+1,1} = \frac{4hR_0(\text{SSB}_t/\text{SSB}_0)}{1-h+(5h-1)(\text{SSB}_t/\text{SSB}_0)} \times \exp({}_R\varepsilon_{t+1,1} - \sigma_R^2/2) \tag{B.5}
$$

where $\text{SSB}_t = \sum_a N_{t,a}\, \Phi_{t,a} w_{t,a}$ is spawning biomass,

$$
\text{SSB}_0 = \sum_{a=1}^{A-1} R_0 e^{-(a-1)M}\, \bar{\Phi}_a \bar{w}_a \;+\; R_0\, e^{-(A-1)M}\, \bar{\Phi}_A \bar{w}_A / (1 - e^{-M}) \tag{B.6}
$$

is average virgin spawning biomass, h is steepness of the stock-recruitment curve, and Φ is proportion mature, $\bar{\Phi}$ and $\bar{w}$ are the average maturity and weights over all years.

Maximum sustainable yield (MSY) and related reference points (u_{MSY}, B_{MSY}) are evaluated using an inner optimization routine (Magnusson and Hilborn 2007), taking advantage of the following relationships,

$$
\begin{aligned}
R^{\star} &= \frac{\mathrm{SBPR}^{\star} - \alpha}{\beta \mathrm{SBPR}^{\star}} \\
\alpha &= \frac{\mathrm{SBPR}_0 (1-h)}{4h} \\
\beta &= (5h-1)/(4hR_0)
\end{aligned}
\tag{B.7}
$$

where $R^{\star}$ is the average recruitment at a given harvest rate, $\mathrm{SBPR}^{\star}$ is spawning biomass per recruit at that harvest rate, SBPR_0 is virgin spawning biomass per recruit, and α and β are Beverton-Holt dummy parameters to simplify the first equation.

The model is fitted to three data components: an annual survey biomass index, commercial catch at age, and survey catch at age. The predicted survey index is proportional to the biomass vulnerable to the survey in the middle of the fishing year,

$$
\hat{I}_t = q \sum_a {}_S S_a N_{t,a} w_{t,a} \, e^{-M/2} \tag{B.8}
$$

where $\hat{I}$ is the predicted survey index, q is the catchability coefficient, and ${}_S S$ is survey selectivity. The catch-at-age predictions are in the form of proportions that sum to one within each year,

$$
{}_C \hat{P}_{t,a} = \frac{{}_C S_a N_{t,a}}{\sum_a {}_C S_a N_{t,a}} \tag{B.9}
$$

$$
{}_S \hat{P}_{t,a} = \frac{{}_S S_a N_{t,a}}{\sum_a {}_S S_a N_{t,a}} \tag{B.10}
$$

where ${}_C \hat{P}_{t,a}$ and ${}_S \hat{P}_{t,a}$ are the predicted commercial and survey catch at age.

The objective function for fitting the model relates to the three data components, as well as a penalty on recruitment deviations from the stock-recruitment relationship:

$$f \;=\; -\log L_I - \log L_C - \log L_S + \text{Pen} \tag{B.11}$$

The survey index likelihood component is lognormal,

$$-\log L_I \;=\; \sum_t \frac{(\log I_t - \log \hat{I}_t)^2}{2\sigma_I^2} \tag{B.12}$$

where I and $\hat{I}$ are the observed and model-predicted survey indices, and σ_I is the observation noise for the survey index. The robust normal likelihood for proportions (Fournier et al. 1990) is assumed for the catch-at-age data,

$$-\log L_C \;=\; -\sum_t \sum_a \log\left[\exp\left(\frac{-({}_C P_{t,a} - {}_C\hat{P}_{t,a})^2}{2\left[{}_C P_{t,a}(1-{}_C P_{t,a}) + 0.1/A\right]{}_C n^{-1}}\right) + 0.01\right] \tag{B.13}$$

$$-\log L_S \;=\; -\sum_t \sum_a \log\left[\exp\left(\frac{-({}_S P_{t,a} - {}_S\hat{P}_{t,a})^2}{2\left[{}_S P_{t,a}(1-{}_S P_{t,a}) + 0.1/A\right]{}_S n^{-1}}\right) + 0.01\right] \tag{B.14}$$

where ${}_C P$ and ${}_C\hat{P}$ are the observed and the model-predicted catch proportions at age, ${}_C n$ is the effective sample size for the commercial catch, while ${}_S P$, ${}_S\hat{P}$, and ${}_S n$ are the corresponding quantities for the survey catch at age. Finally, recruitment deviates are penalized under the assumption of lognormality,

$$\text{Pen} \;=\sum_{a=2}^{A-1} \frac{{}_R\varepsilon_{1,a}^2}{2\sigma_R^2} \;\;+\;\; \sum_{t=2}^{t_{\max}-1} \frac{{}_R\varepsilon_{t,1}^2}{2\sigma_R^2} \tag{B.15}$$

where ${}_R\varepsilon_{1,a}$ and ${}_R\varepsilon_{t,1}$ are recruitment deviates in the initial year and subsequent years, and σ_R is a measure of the extent of recruitment variability.

The level of observation noise (σ_I, ${}_Cn$, ${}_Sn$) in the data and recruitment variability of the stock (σ_R) is estimated iteratively (McAllister and Ianelli 1997) from the empirical residuals and deviates of the base model,

$$\hat{\sigma_I} = \sqrt{\frac{\sum(\log I_t - \log \hat{I}_t)^2}{T-1}} \tag{B.16}$$

$$\hat{\sigma_R} = \sqrt{\frac{\sum \varepsilon_d^2}{D}} \tag{B.17}$$

$${}_C\hat{n} = \text{median}({}_Cn_t), \quad {}_Cn_t = \frac{\sum_a \left({}_C\hat{P}_{t,a}[1 - {}_C\hat{P}_{t,a}] \right)}{\sum_a \left({}_CP_{t,a} - {}_C\hat{P}_{t,a} \right)^2} \tag{B.18}$$

$${}_S\hat{n} = \text{median}({}_Sn_t), \quad {}_Sn_t = \frac{\sum_a \left({}_S\hat{P}_{t,a}[1 - {}_S\hat{P}_{t,a}] \right)}{\sum_a \left({}_SP_{t,a} - {}_S\hat{P}_{t,a} \right)^2} \tag{B.19}$$

where T is the number of survey index datapoints and D is the number of recruitment deviates.

B.2 Uncertainty methods

Four methods are used to quantify uncertainty: the delta method, profile likelihood, bootstrap, and MCMC. The computations are the same as in Magnusson et al. (2013), with the addition of profile likelihood. Equation B.20 summarizes how each method generates a distribution that is used to construct confidence intervals,

$$
\begin{aligned}
y &\xrightarrow[\text{delta}]{\text{model}} \hat{\theta},\, \widehat{\text{SE}}_{\hat{\theta}} \xrightarrow{\text{Norm}} p(y|\theta) \\
y &\xrightarrow{\text{model}} \max_{\eta} L(\eta|\theta_p, y) \xrightarrow{\text{profile}} L(\theta|y) \\
y &\xrightarrow{\text{model}} \hat{\theta} \xrightarrow{\text{bootstrap}} y_1^\star, y_2^\star, \ldots, y_B^\star \xrightarrow{\text{model}} \hat{\theta}_1^\star, \hat{\theta}_2^\star, \ldots, \hat{\theta}_B^\star \xrightarrow[\text{bias corr}]{\text{density}} p(y|\theta) \\
y &\xrightarrow[\text{MCMC}]{\text{model}} \theta_1, \theta_2, \ldots, \theta_T \xrightarrow{\text{density}} p(\theta|y)
\end{aligned}
\tag{B.20}
$$

where y denotes the observed data, θ is a vector of parameters (and derived quantities), the $\hat{}$ symbol indicates an estimate of a parameter or derived quantity, $\widehat{\text{SE}}_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$, η is a vector of model parameters other than the θ_p parameter of interest, $y_b^\star$ is a bootstrap dataset, $\hat{\theta}_b^\star$ is a bootstrap estimate, and θ_t is an MCMC iteration. The sampling distribution $p(y|\theta)$, profile likelihood $L(\theta|y)$, and posterior distribution $p(\theta|y)$ are then used to generate intervals expressing the uncertainty about estimated parameters and derived quantities.

Delta method

The estimation uses automatic differentiation (Fournier et al. 2012) to evaluate the Hessian matrix and hence the approximate variance-covariance matrix for the estimated parameters. The delta method (Seber 1973), which assumes that both estimation bias and the quadratic terms of the Taylor series are negligible, is then used to estimate the variance of each derived quantity,

$$
\widehat{\text{SE}}_{\hat{g}} = \sqrt{\sum_i \sum_j \widehat{\text{Cov}}\left(\hat{\theta}_i, \hat{\theta}_j\right)\left(\frac{\partial g}{\partial \theta_i}\right)\left(\frac{\partial g}{\partial \theta_j}\right)}
\tag{B.21}
$$

where g is a derived quantity, such as a reference point, that is a function of some estimated parameters $\theta_1, \theta_2, \ldots, \theta_n$. The symmetric confidence interval for g is then

$$\left[\hat{g} - z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}}, \quad \hat{g} + z_{1-\alpha/2}\widehat{\mathrm{SE}}_{\hat{g}}\right] \tag{B.22}$$

where z is the standard normal quantile.

The quantities B_{current}, B_{MSY}, and MSY are log-transformed for the purpose of applying the delta method, because the uncertainty about these quantities can be expected to be closer to lognormal than normal (Magnusson et al. 2013, Stewart et al. 2013). The current harvest rate u_{current} (landings/biomass in 2014) is also log-transformed, since a constant divided by a lognormal random quantity is also lognormal.

Profile likelihood

It is straightforward to calculate the profile likelihood (Venzon and Moolgavkar 1988, Hilborn and Mangel 1997, Millar 2011) for model parameters such as h, M, and ${}_{C}S_{\mathrm{right}}$. An iterative procedure is applied to a parameter of interest, fixing it at a different value in each iteration, while maximizing the total likelihood (Equation B.11) over all other (nuisance) parameters,

$$\max_{\eta} L(\eta|\theta_p, y) \tag{B.23}$$

where L is the likelihood, η is the vector of nuisance parameters, θ_p is the parameter of interest, and y is the data. The resulting confidence interval,

$$[\,\theta_p \mid \log L = \log L_{\max} - 0.5\chi^2_{df=1,\alpha}] \tag{B.24}$$

contains all values of θ_p where the log-likelihood is less than $0.5\chi^2_{df=1,\alpha}$ away from the global maximum log-likelihood.

A variation of this procedure is used to calculate the profile likelihood for derived quantities such as B_{current}, u_{current}, u_{MSY}, B_{MSY}, and MSY. These quantities cannot be fixed at a certain value during the optimization, so an objective function penalty $\lambda_p(\hat{\theta}_p - \theta_{\mathrm{target}})^2$ is

introduced to force the estimated quantity to be close to a given value in each iteration. The penalty weight λ_p is chosen so the penalty ends small in each iteration, and the penalty is excluded from the objective function when storing the profile likelihood (Equation B.23) at each value of θ_p.

Bootstrap

A parametric model-conditioned approach is used to generate 1000 bootstrap datasets, with residuals sampled from probability distributions and applied to the model fit to the original data (Efron and Tibshirani 1993). The bootstrap survey abundance index is

$$I_t^\star \;=\; \hat{I}_t \times \exp({}_I\varepsilon_t^\star), \quad {}_I\varepsilon_t^\star \sim \mathrm{N}(0, \hat{\sigma}_I{}^2) \tag{B.25}$$

where $I_t^\star$ is the bootstrap datum for year t, $\hat{I}_t$ is the predicted index for year t from the model fit to the original dataset, ${}_I\varepsilon_t^\star$ are bootstrap residuals, and $\hat{\sigma}_I$ is the estimated magnitude of observation error. The bootstrap commercial catch at age is

$${}_CP_{t,a}^\star \sim \text{Multinom}\left({}_C\hat{n},\, {}_C\hat{P}_{t,a}\right)\Big/ {}_C\hat{n} \tag{B.26}$$

where ${}_CP_{t,a}^\star$ are the bootstrap data, ${}_C\hat{n}$ is the estimated effective sample size, and ${}_C\hat{P}_{t,a}$ is the model-predicted commercial catch at age for year t. Similarly, the bootstrap survey catch at age is:

$${}_SP_{t,a}^\star \sim \text{Multinom}\left({}_S\hat{n},\, {}_S\hat{P}_{t,a}\right)\Big/ {}_S\hat{n} \tag{B.27}$$

Each estimation model is fitted to the bootstrap datasets, yielding bootstrap estimates for each parameter and derived quantity. A bias-correction factor is then applied, which has been shown to lead to more accurate confidence intervals (Magnusson et al. 2013),

$${}_{\mathrm{BC}}\vec{\hat{\theta}}^\star \;=\; \hat{\Omega}^{-1}\Big[\Phi\Big(2\Phi^{-1}\big[\max\{0.1, \min\{0.9, \hat{\Omega}(\hat{\theta})\}\}\big] \;+\; \Phi^{-1}(\vec{\alpha})\Big)\Big] \tag{B.28}$$

where ${}_{\text{BC}}\vec{\hat{\theta}}^{\star}$ is a vector of bias-corrected bootstrap estimates in ascending order, $\Phi(\cdot)$ is the standard normal cumulative distribution function, $\hat{\Omega}(x) = \#\{\hat{\theta}^{\star} < x\}/B$ is the empirical cumulative distribution function of the bootstrap estimates $\hat{\theta}^{\star}$, while $\Phi^{-1}(\cdot)$ and $\hat{\Omega}^{-1}(\cdot)$ are the corresponding inverse functions, B is the number of bootstraps, and $\vec{\alpha}$ is a vector of probability levels $1/B$, $2/B$, ..., B/B. The bias-corrected bootstrap confidence interval is calculated as:

$$\left[\frac{\alpha}{2} \text{ quantile from } {}_{\text{BC}}\vec{\hat{\theta}}^{\star}, \quad \left(1-\frac{\alpha}{2}\right) \text{ quantile from } {}_{\text{BC}}\vec{\hat{\theta}}^{\star}\right] \tag{B.29}$$

Bayesian MCMC analysis

Markov chain Monte Carlo (MCMC) simulation is used to approximate the posterior distribution of estimated parameters and derived quantities. The simulation method is Metropolis-Hastings with an adaptive multivariate normal jumping distribution (Gelman et al. 2004, Fournier et al. 2012).

All model parameters are assigned uniform priors with wide bounds, except the recruitment deviates (Equation B.15), which have a lognormal prior in all estimation models, and h and ${}_{C}S_{\text{right}}$, which have diffuse priors (Equations 3.3–3.4). The MCMC simulation is run for 1 million iterations and then thinned, keeping every 1000th iteration. The simulation starts at the best model fit, so no burn-in period is required. Convergence of the estimated quantities is diagnosed using the 'coda' package (Plummer et al. 2006), adopting an autocorrelation threshold of 0.1, Geweke threshold of 1.96, and Heidelberger-Welch threshold of 0.05. If any criteria are not met, the MCMC chain is extended to a maximum of 10 million iterations, keeping every 10,000th iteration, to reduce autocorrelation and stabilize the distribution quantiles. The MCMC confidence interval is calculated as

$$\left[\frac{\alpha}{2} \text{ quantile from } \theta_1, \theta_2, \ldots, \theta_T, \quad \left(1-\frac{\alpha}{2}\right) \text{ quantile from } \theta_1, \theta_2, \ldots, \theta_T\right] \tag{B.30}$$

where $\theta_1, \theta_2, \ldots, \theta_T$ are the iterations retained from the MCMC chain.

VITA

Arni Magnusson was born in Stuttgart, Germany. He grew up in Reykjavik, Iceland, his country of citizenship. He obtained his B.Sc. in Fisheries Biology from the University of Iceland and was awarded a Fulbright scholarship to study at the University of Washington, where he obtained his M.Sc. in Fisheries Science under the supervision of Ray Hilborn. After working for several years at the Marine Research Institute and the United Nations University in Iceland, Arni returned to the University of Washington to obtain his Doctorate in Fisheries Science, supervised by Ray Hilborn and André Punt. This year, Arni will join the ICES Secretariat in Copenhagen, Denmark.