Estimating fish stock population parameters from bottom trawl surveys: sources and magnitude of uncertainty

PhD Thesis

Hans D Gerritsen

Supervisors: Dr David McGrath and Dr Pauline King, Galway-Mayo Institute of Technology
Dr Colm Lordan, Marine Institute

Author's contact details

Marine Institute   Galway-Mayo Institute
Rinville   of Technology
Oranmore   Dublin road
Galway   Galway
Ireland   Ireland
Tel: +353 91 387297
Email: hans.gerritsen@marine.ie

Abstract

Fisheries surveys are widely used in the stock assessment process by providing time series of relative abundance and recruitment strength. Surveys also provide biological data that have the potential to improve stock projections by providing alternatives to the use of spawning stock biomass as a measure of stock reproductive potential. The objective of the present work is to illustrate various sources of error in survey data, using examples from bottom trawl surveys and to provide innovative methods for identifying and minimising these errors. In Chapter 1, the current work is framed within the existing literature and an overview is given of the different sources of error. The next six chapters consist of papers that have been published, are in press, or under review with peer-reviewed journals. These papers consist of case-studies, each exploring a different source of error. In Chapter 2, a method for summarising the precision of a length-frequency distribution is presented and recommendations are given on the sample sizes required to achieve a particlar precision level. In Chapter 3, the variability in the assignment of maturity stages of plaice (*Pleuronectes platessa*, L.) and whiting (*Merlangius merlangus,* L.) is examined, using statistical techniques developed in the social sciences. In Chapter 4, the spatial variability in the age-length structure of haddock (*Melanogrammus aeglefinus,* L.) is explored and a new method for comparing age-length keys is developed. In Chapter 5, the spatial structure in the length-weight relationships of whiting and haddock are analysed using geostatistical tools. In Chapter 6, the spatial trend in the sex ratio of megrim (*Lepidorhombus whiffiagonis,* Walbaum) is explored using generalized additive modelling techniques. In Chapter 7, spatial trends in the proportion of mature cod (*Gadus morhua,* L.) are investigated. In the concluding Chapter 8, the scale of various errors is reviewed in the light of the current work and recommendations for sampling design are proposed.

Contents

Chapter 1

General introduction

1.1. Structure of the thesis

This thesis consists of a general introduction followed by five papers that have been published or are in press with peer-reviewed journals and one paper that is under review, the last chapter provides a synthesis. Each of the papers covers a subject related to variability and bias resulting from measurement error or sampling error on bottom trawl surveys. A glossary of terms is provided in Appendix A, page 106.

1.2. Purpose of bottom trawl surveys

Bottom trawl surveys are widely used to collect information on a variety of parameters of fish stocks. Surveys are used in the stock assessment process to calibrate or "tune" age- or size-based population models. Surveys also provide crucial information on the strength of the incoming year class before it enters the fishery. This is particularly important in terms of forecasting stock development in the short term and giving advice on future catch options. The reliability of stock assessment results is greatly influenced by the precision of survey abundance estimates (Smith and Gavaris, 1993; Beare *et al.,* 2003).

Bottom trawl surveys are also increasingly used to provide abundance estimates independently of commercial data (Beare *et al.,* 2005). Due to unknown levels of misreporting and/or discarding, many sources of commercial data are regarded as unreliable. In addition to this, data from commercial sources may lack spatial detail while at the same time commercial fisheries tend to target specific locations such as spawning aggregations. Therefore, data from commercial sources are likely to be significantly biased samples of the population. This can result in biased estimates of the age structure or other biological parameters like maturity, sex ratio or condition. For this reason, commercial sources are often unsuitable for providing reliable biological data. While survey data are sparser than commercial data, survey data are collected under controlled circumstances and can provide spatially detailed biological information.

Time series of biological data have the potential to provide improved estimates of reproductive potential. Historically, the spawning stock biomass (SSB) has been used

as an index of reproductive potential (Beverton and Holt, 1957). However, up-to-date information on age composition, maturity, condition and sexual dimorphism can help to provide more sensitive indices of reproductive potential than SSB estimates alone (Marshall *et al.,* 1998; Marteinsdottir and Thorarinsson, 1998; Scott *et al.,* 1999; Marshall *et al.,* 2003; Tomkiewicz *et al.,* 2003a; Marshall *et al.,* 2006). For example, SSB estimates do not take account of changes in sex ratio or fecundity, while female-only SSB or total egg production estimates are conceptually better proxies for reproductive potential. Additionally, trends in growth rates and maturity can be used as an indication of over-exploitation. For example, the collapse of the Northwest Atlantic cod (*Gadus morhua*) populations was preceded by a significant reduction in age and length at first maturity (Hutchings, 2005).

Including information other than SSB in fisheries advice does not necessarily improve projections, however. The error around estimates of stock parameters will increase as the number of parameters used in this estimate increases. Each of the steps involved in the estimation of stock parameters, has a certain amount of error associated with it (Figure 1.1). The economist John Maynard Keynes is reported to have said that he would rather be approximately right than exactly wrong, but in order to weigh up accuracy versus precision, it is necessary to understand and quantify all sources of error that are involved in the collection and analysis of biological data.

1.3. Types of error

There are two main types of error. Random error is related to the precision of an estimate: it is the variability between replicate experiments. It is not always necessary or possible to perform replicate sampling in order to estimate random error. Analytical (Zar, 1999), Bayesian (Malakoff, 1999), or re-sampling techniques (Efron and Tibshirani, 1993) can be used to estimate this type of error. The second type of error, systematic error or bias, is related to the accuracy of an estimate: it is the difference between the true parameter value and its estimated value. Both random and systematic errors can result either from sampling error or measurement error. Sampling error is determined by the sample size and the variability of the parameter. Measurement error is due to the measuring device or to observer error.

1.4. Sampling error

Gear selectivity - By the time the fish arrive on the deck, the sample is already highly biased (Gulland, 1966). Although not all fish in a population are equally vulnerable to the fishing gear, surveys are generally assumed to assess age groups with equal accuracy from year to year. However, if the size-at-age distribution varies between years, abundance estimates will be affected because trawls are highly length-sensitive tools (Godø and Sunnanå, 1992). In addition to bias in abundance estimates, bias can also occur in estimates of length-at-age, sex ratio-at-age and maturity-at-age for age classes that are not fully selected by the gear. Differences in size and behaviour between the sexes can also result in differences in catchability between the sexes (Beverton, 1964; Rijnsdorp, 1993; Solmundsson *et al.,* 2003), an issue that will be further explored in Chapter 6.

Knowledge on the selectivity of the gear can help to correct for some forms of bias but selectivity itself can be quite variable, particularly for fish that have a low catchability (Godø and Sunnanå, 1992).

Spatial structure – Most fish populations appear to be spatially clustered. Pennington *et al.* (2002) have shown that fish caught together tend to be much more similar in size than can be expected from a random sample. Johnsen (2003) also noted a clustering of fish of similar lengths and used geostatistical techniques to take this spatial covariance into account for the estimation of length frequency distributions. Spatial structure, or clustering, has also been observed for age distribution (Smith *et al.,* 1992), sex ratio (Morgan and Trippel, 1996; Swain, 1997) and the proportion of mature fish (Horn *et al.,* 1998; Bromley, 2000; Rochet, 2000). If trends occur on a large spatial scale, bias may occur if sampling coverage is incomplete. Spatial structures that are small, compared to the scale of sampling, result in increased variability between stations. This between-station variability is often ignored, for example Gulland (1966) states that the precision achieved for length sampling is determined almost entirely by the number of specimens examined; he does not mention the influence of the number of sampling locations on the precision. However, Pennington *et al.* (2002) have shown that the precision of a length sample is almost entirely determined by the number of locations sampled, while the total number of fish measured hardly influences the precision.

Variability between stations or regions will be explored for the following parameters: age-length structure (Chapter 4); condition (Chapter 5); sex ratio (Chapter 6) and maturity (Chapter 7).

Subsampling – As it is often impossible to analyse every fish caught, subsamples are regularly taken from the catch. If the sample is taken without thorough mixing or randomisation, bias might occur. Heales *et al.* (2003) have shown that some catches are not uniformly mixed; they found that that mechanical sorting can take place in seawater hoppers. This potential bias is easily avoided, for example by placing the total catch in fish boxes and repeatedly halving the contents of each box (by tipping the contents of one box into two others) until the desired sample size is achieved. Although it has not been tested whether this results in a truly random sample, it seems reasonable to assume that such subsamples are representative of the catch.

Two-stage sampling is a common way of subsampling a catch (Gulland, 1966; Cochran, 1977). A representative sample is taken to obtain a (presumably) unbiased length frequency distribution of the catch. A smaller sample is then taken for further biological analysis such as individual weight, age, sex and maturity stage. Often, this biological sample is taken on a length-stratified basis; a target sample number is obtained for each length class. An age-length key (or maturity-age-length key) is then constructed for each stratum and applied to the estimated length distribution in that stratum (Fridriksson, 1934). The implicit assumption is that the age-at-length relationship is approximately constant within each stratum (Cotter, 1998). However, Chapter 4 shows that the age-at-length distribution of haddock (*Melanogrammus aeglefinus*) can vary strongly on a small spatial scale, indicating that this assumption might not always be valid.

The precision with which the length distribution of the catch can be determined depends on the size of the subsamples. In Chapter 2, a method is presented that summarises the precision of a length distribution, allowing length data to be collected to a pre-defined precision level. While length measurements can be collected quickly and cheaply, further biological data tend to be much more time-consuming to collect. It is generally assumed that the precision of the estimated age distribution improves rapidly with increasing numbers in the aged sample (Flatman, 1990; Kimura, 1997). Oeberst (2000) has developed a cost function to estimate the optimal number of age readings. Gutreuter and Krzoska (1994) investigated the precision of weight

measurements and Roa *et al.* (1999) evaluated various methods to estimate the precision of maturity estimates. While these studies can give insight into the precision of catch estimates, they give no information on the precision of population estimates, because between-station variability is ignored. Ignoring this covariance between stations can lead to over-fitted models, misleading error distributions and biased estimates (Cotter, 1998).

1.5. Measurement error

Ageing error – There are a number of different types of measurements associated with surveys, each with its own measurement error. The error in age determination has received by far the most attention in the literature, as age determinations are both subjective and central to most stock assessment models. A large number of papers deal with ageing error and quality control (e.g. Bradford, 1991; Kimura and Lyons, 1991; Richards *et al.,* 1992; Campana *et al.,* 1995; Campana, 2001; Reeves, 2003; Clark, 2004; Kimura and Anderl, 2005; Power *et al.,* 2006). In general, random ageing errors result in an underestimate of the numbers of the most common age class in the catch (Richards *et al.,* 1992). This is because a certain percentage of the most common age class will be wrongly assigned to the neighbouring age classes, the reciprocal number of fish that will be wrongly assigned from the neighbouring age classes will be lower because the numbers of fish in these age classes are lower. Therefore, random ageing error tends to smooth out differences in abundance between age classes. It is possible to correct for this type of bias by modelling the age reading errors (Richards *et al.,* 1992). The effects of ageing errors on stock assessments have been investigated by Bradford (1991) and Reeves (2001; 2003).

Systematic error can occur in species where the first growth increment is unclear. If ages are determined from an incorrectly defined starting point, age determinations will be consistently wrong by a constant amount. Alternatively the periodicity of increments might not be annual, also leading to bias. Campana (2001) provides a comprehensive overview of different types of ageing error, their magnitude and quality control standards. Systematic ageing errors can be detected from consistent differences between labs (e.g. Power *et al.,* 2006) or by obtaining known-age material (validation; e.g. Campana, 2001).

Other sources of measurement error – Errors can occur in length measurements (e.g. Gulland, 1966). Unpublished data by the author also show that consistent differences between people repeatedly measuring the same sample of fish can be detected, even when using electronic measuring boards. However, these differences were too small to be significant with the sample size of 30 fish that was used in the experiment (Gerritsen, unpubl.; Appendix B page 110). Gutreuter and Krzoska (1994) state that length can be measured with high precision, while weight measurements are more variable, particularly for small fish. They found that the precision of weight estimates was mostly affected by surface wetness and movements of the fish. Cotter (1998), on the other hand, states that weighing variances can be safely ignored if surveys are equipped with high-quality balances. Unpublished work by the author suggests that weighing errors on Irish groundfish surveys are likely to be smaller than 1% (Gerritsen, unpubl.; Appendix B page 110). A parameter that is not usually associated with measurement error is sex determination. However, it can be difficult to determine the sex of some young fish (Chapters 3 and 6), resulting in another potential source of error. The assignment of maturity stages can also be somewhat subjective. Maturity stages are usually assigned after macroscopic (visual) inspection of the gonads. While this is generally accepted to be less precise and accurate than histological analysis, it is the only practical method that allows large numbers of fish to be analysed (West, 1990). Vitale *et al,* (2006) found consistent differences between the two methods in the estimated proportions mature, but they also found that this bias was negligible in samples taken just before the start of the spawning season. Tomkiewicz *et al.* (2003b) also found the highest agreement between the two methods just before the spawning season. In Chapter 3, the variability in the assignment of maturity stages is further investigated.

The objective of the present work is to illustrate various sources of error in survey data, using case studies from bottom trawl surveys and to provide innovative methods for identifying and minimising these errors.

1.6. Figure
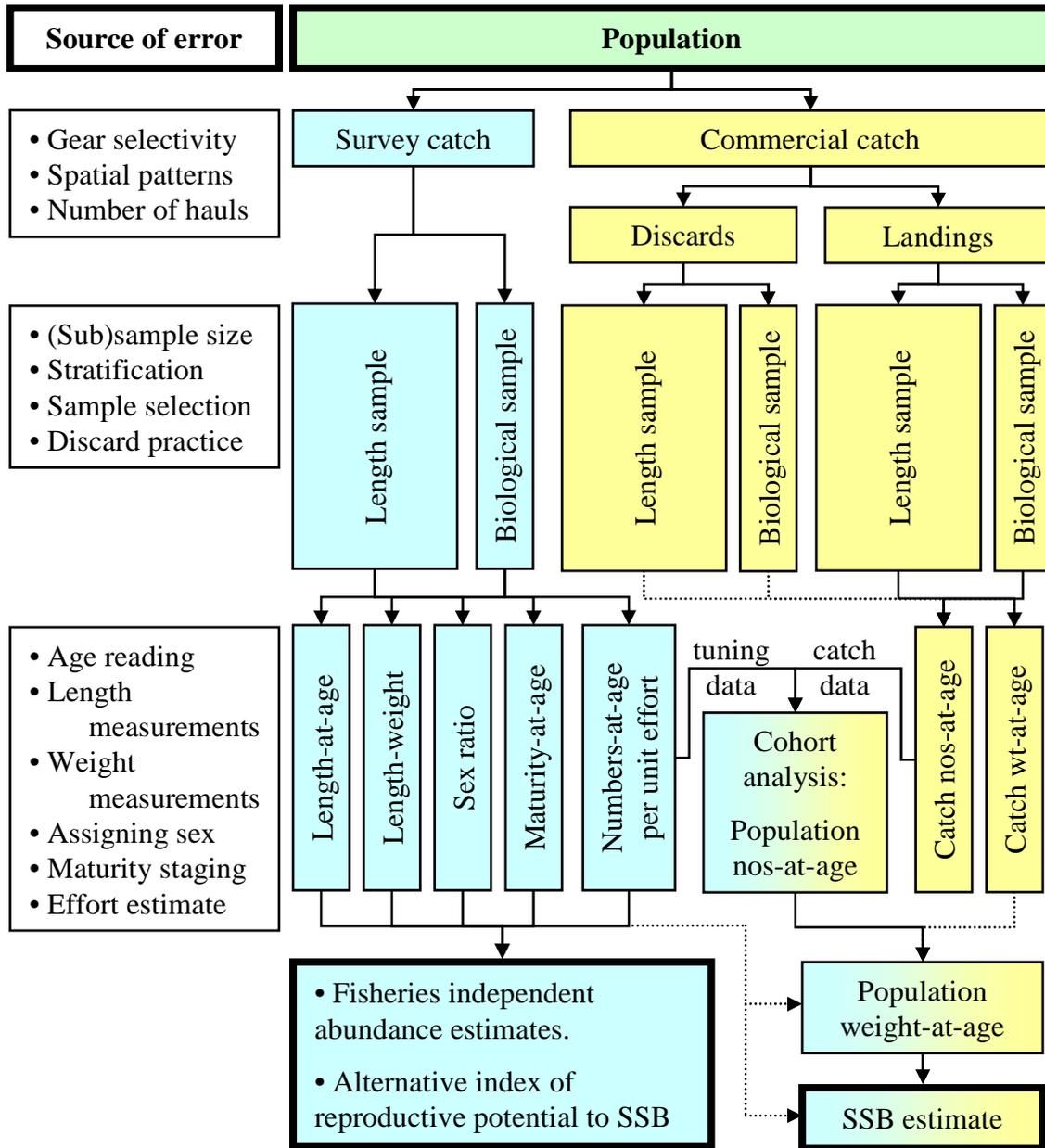
Figure 1.1. Diagram of the various steps involved in data collection and analysis of fisheries data and the sources of error at each step of the process. In age-based stock assessment, one of the most important parameters to be estimated from trawl surveys is an estimate of the numbers-at-age caught per unit effort. This information is used to tune the stock assessment model. Other survey data that can be used in assessments are maturity ogives, length-weight relationships and/or length-at-age data. Survey data can also be used to obtain fisheries-independent abundance estimates and for indices of reproductive potential other than Spawning Stock Biomass (SSB).

Precision estimates and suggested sample sizes for length frequency data

2.1. Abstract

For most fisheries applications, the shape of a length frequency distribution is much more important than its mean length or variance. This makes it difficult to evaluate at which point a sample size is adequate. By estimating the coefficient of variation of the counts in each length class and taking a weighted mean of these, a measure of precision was obtained that takes the precision in all length classes into account. The precision estimates were closely associated with the ratio of the sample size to the number of size classes in each sample. As a rule of thumb, a minimum sample size of 10 times the number of length classes in the sample is suggested, as the precision deteriorates rapidly for smaller sample sizes. In absence of such a rule-of-thumb, samplers have previously under-estimated the required sample size for samples with large fish, while over-sampling small fish of the same species.

Key words: length frequency, precision, sampling design, sample size.

2.2. Introduction

Length measurements are fundamental to many aspects of fisheries science. However, there is little formal guidance on the appropriate size of a length sample. Such guidance is of particular relevance when the number of fish available exceeds the number that can be measured at a reasonable cost, and a sub-sample needs to taken. Clearly, the required precision of a length sample depends on the purpose of sampling. In order to identify modes of individual year classes for a length-based assessment, the precision of the sample needs to be quite high. Sample sizes of more than 1000 are necessary to identify more than half the modes in a typical length distribution (Erzini, 1990). A sample size of at least 100 adult fish was recommended

for age-based stock assessment purposes (Anderson and Neumann, 1996), although the authors do not mention how they arrived at this number.

Regardless of the type of assessment that is used, the shape of the length frequency distribution is of interest, rather than simple summary statistics such as the mean or its variance. For this reason, it has proven difficult to quantify what constitutes a representative or adequately precise length distribution. Some studies have attempted to find minimum or optimum sample sizes by comparing samples to an expected distribution (e.g. Müller, 1996; Gomez-Buckley *et al.,* 1999; Vokoun *et al.,* 2001). However, the true distribution is usually unknown, and dissimilarity from the expected distribution does not necessarily indicate an imprecise sample. In addition, these methods only provide indirect measures of precision that are difficult to evaluate objectively.

Thomson (1987) used the precision of a sample explicitly to establish an appropriate sample size. Thompson proved that a sample size of 510 is sufficient to be 95% confident that all estimated proportions in a multinomial distribution are no more than 0.05 from the true proportion. However, Thomson based this figure on a worst-case scenario, which, in the present case, is a length frequency distribution that is evenly apportioned over three size classes. As this is not the typical shape of a length frequency distribution used in fisheries science, Thomson's measure of precision is too conservative in the vast majority of cases.

For most fisheries applications, it would be more useful to define the precision of a length frequency sample as the mean precision over the entire size range. However, it appears that this approach has not been used to establish an optimum sample size. Such precision estimates might be used to obtain a rule-of-thumb for sample sizes required to obtain a certain precision level of the catch at each location. In the present paper we aim to: 1) Determine a rule-of-thumb for obtaining an appropriate sample size when the number of fish available in a particular sample exceeds the number that can be measured at a reasonable cost. 2) Examine the sample sizes taken in the past, in absence of such guidance.

2.3. Materials and methods

Data were used from the Irish Groundfish survey, which was carried out on R.V. Celtic Explorer in the waters around Ireland during October and November 2005. The catch was sorted into species and, if appropriate, size grades, each of which were

treated as a separate length sample. Length measurements were taken from all fish and squid species that were caught. If the number of individuals in a sample was large, a sub-sample was taken by repeatedly transferring the sample from each fish box into two other boxes and discarding one of these. This method ensures that the entire catch is represented uniformly in the sub-sample. At the time of the survey, the samplers did not have any particular guidance on the appropriate sub-sample size, but used their own judgement to decide on the sample size.

The precision of the number of observations in each length class of a random sample can be estimated by assuming a multinomial distribution (Smith and Maguire, 1983). If the precision in each length class is expressed in the form of a coefficient of variation (CV), an overall measure of precision can be obtained by weighting each CV by the number of fish in each length class. This mean weighted CV (MWCV) provides a description of the precision over the entire range of size classes in a length frequency distribution.

Under the assumption of a multinomial distribution, the standard deviation (σ_i) of the number of fish in a sample that are length category i can be estimated by:

$$(2.1) \quad \sigma_i = \sqrt{np_i(1-p_i)}$$

where n is the total number of fish in the sample and p_i is the proportion of the sample that is length i. The coefficient of variation (CV) of the number of fish at length i, is given by:

$$(2.2) \quad CV_i = \frac{\sigma_i}{np_i}$$

and the mean weighted coefficient of variation (MWCV) is given by:

$$(2.3) \quad MWCV = \sum p_i CV_i = \frac{\sum \sigma_i}{n}$$

The highest possible value of the MWCV results from a length frequency distribution that is evenly distributed over a large number of size classes. The numbers of fish at each length class are then Poisson distributed with a standard deviation that equals the square root of the number at length (Zar, 1999). The theoretical maximum MWCV is therefore given by:

$$(2.4) \quad MWCV = (n/c)^{-0.5}$$

where c is the number of size classes in the sample. The minimum MWCV is zero and would result from a distribution where all observations fall within a single length

category. Therefore, the MWCV estimates will always lie between zero and the curve described by Equation (2.4).

2.4. Results

During the 2005 survey, a total of 2332 length samples were taken for 80 different species of fish and squid. In most cases, the sample size was limited by the number of individuals in the catch. However, 596 samples were deemed too large to measure all individuals and sub-samples were taken. The median sub-sample size was just under a quarter of the total catch (by weight), while 90% of the sub-samples were smaller than half of the total catch. The four most common species that were sub-sampled were poor cod (*Trisopterus minutus*), blue whiting (*Micromesistius poutassou*), haddock (*Melanogrammus aeglefinus*) and Norway pout (*Trisopterus esmarkii*).

The estimated MWCV of the sub-samples was closely associated with the ratio of the number of individuals measured to the number of length classes in the sample (Figure 2.1). The MWCV appeared to follow an exponential curve that was close to the maximum MWCV given by Equation (2.4). The MWCV decreased very rapidly with increasing sample size up to sample sizes of around 10 times the number of length classes in the sample, after which the sample size would need to be increased considerably for a moderate further improvement in precision. If the sample size is taken as 10 times the number of length classes in the distribution, an MWCV of around 0.25 can be expected; a sample size of 48 times the number of length classes would result in an MWCV of 0.10 and a sample size of 155 times the number of length classes would be necessary to reduce the MWCV to 0.05.

The mean sample size in the sub-samples taken on the survey was just under nine times the number of length classes per sample, resulting in a mean MWCV of 0.33. However, there was quite a large spread in the sample sizes (Figure 2.1), so some samples were measured with very low precision, while others had excessively large sample sizes. The range of sample sizes was between 2.2 and 24.7 times the number of length classes (2.5% and 97.5% quantiles), resulting in a range of MWCVs between 0.14 and 0.61. With a minor increase in effort, the sample size might be increased to 10 per length class for each sub-sample, resulting in an MWCV of around 0.25 for all samples. Considering that the precision deteriorates very rapidly for sample sizes of less than 10 per length class, a minimum sample size of 10 times the

number length classes in the sample is suggested as a rule-of-thumb in the present case.

The previous analysis shows that, in order to obtain the same level of precision for all sub-samples, the sample size should be directly proportional to the number of size classes. In absence of specific guidance on the sample size during the 2005 survey, the chosen sample size was only weakly correlated to the number of length classes in the sample for haddock and poor cod, while no significant correlation was found for blue whiting and Norway pout (Figure 2.2). The same figure also shows that the MWCV in sub-samples tended to increase with the mean length of the fish in the sample. This indicates that samples of fish with a large mean length tended to be sampled with lower precision than samples of smaller fish of the same species.

2.5. Discussion

Length distributions that result from combining a number of different samples, exhibit greater variation than expected under the multinomial model given in Equation (2.1) (Smith and Maguire, 1983). Fish populations are usually not uniformly mixed, therefore individual samples are not random samples from the population (Pennington *et al.,* 2002). The simple multinomial model does not take account of the between-sample variability and will therefore underestimate the total variance. However, Equation (2.1) does provide an unbiased estimate of the variability *within* each sample; this is the variability that would occur if one could repeatedly take a random sample at the same location and time and measure these without error. This is the variability that is of interest when deciding whether the sample size is large enough to estimate the length distribution from a particular haul with a certain precision. Therefore, the MWCV is a suitable measure for this exercise.

In order to obtain a precise population estimate, it is important to maximise the number of sampling locations due to the considerable between-sample variability that is usually present (Pennington *et al.,* 2002). These authors suggest maximising the number of sampling locations at the expense of the number of fish measured. However, the number of hauls is often limited by practical considerations, and length measurements can be obtained quickly and cheaply. Therefore, it seems prudent to sample enough fish from each haul to obtain a length distribution that is representative of that catch at that particular location. Detailed information of the length distribution at each station can be valuable for exploratory data analysis, such as investigating the

spatial structure in the data. Nevertheless, this sampling level might not be strictly necessary for a precise population estimate of the length frequency distribution for an age- or length-based assessment.

The samples in Figure 2.1 included a large range of species and size categories of fish, but the variability in the MWCV was small after taking account of the sample sizes. This suggests that the MWCV is not very sensitive to the exact shape of the distribution and can be predicted with high precision, at least within the range of length distributions encountered on the survey. A minimum sample size of 10 times the number of length classes in the sample appears to be a reasonable compromise between effort and precision in the present case.

The current analysis has focussed on sub-sampling during surveys; however the same principles can be applied to any data collection for which the shape of the distribution is of interest. The desired precision level for these cases will depend on a number of factors. For certain species that are of little commercial or scientific interest, but which might span across a large number of length classes, the suggested sample size of 10 per length class might be excessive. Likewise, as the MWCV is directly proportional to the number of length classes in the sample, the choice of the interval of the length classes will determine the precision. Although increasing the size of length intervals will reduce the MWCV, this will result in a loss of information which is undesirable. The cost of sampling, the detail required and the purpose of the data collection need to be considered before the required precision level can be determined for other applications than the present example.

Without formal guidance on the sample size, the sample sizes chosen were, at best, weakly correlated with the number of size classes in the samples. It appears that the samplers underestimated the required sample size for samples with large fish, while samples of smaller fish of the same species were over-sampled. This might be related to the fact that the volume of a sample increases with the cube of its mean length, so a sample size of large fish might appear to be larger than the same number of small fish. In addition, samples with large fish tend to be spread out over a larger number of size classes, thus requiring higher sample numbers.

In practice, it will be difficult for a sampler to estimate both the number of size classes and the number of fish in a sample. Therefore, the Marine Institute in Ireland is developing a software application that allows samplers to examine the length frequencies of the samples directly after they have been measured. The software

estimates the weight of the suggested sample size for each distribution. As size distributions tend to be similar on consecutive hauls, the sampler can get an insight into the required weight of an appropriate sample for each species and size category. The information contained in a length frequency distribution is largely a function of sample size. The present method allows the amount of information to be quantified in terms of precision, allowing samplers to make informed decisions on the sample size that is required to obtain an adequate estimate of the length frequency distribution of a particular catch.

2.6. Acknowledgements

Free R-software has been used for this work and the authors would like to thank the R development core team and all contributors to the R project (http://www.R-project.org). Also thanks to Colm Lordan and two anonymous reviewers for their valuable comments.

2.7. Figures
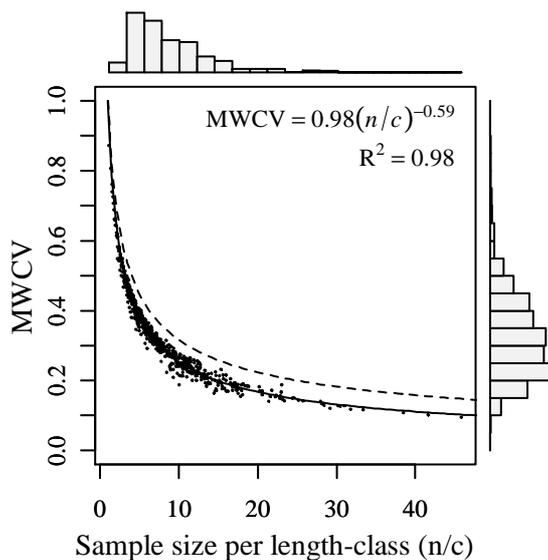
*Figure 2.1. The mean weighted coefficient of variation (MWCV) for 596 sub-samples was closely related to the sample size (*n*) divided by the number per length classes in the sample (*c*). A good fit was obtained for the power function indicated by the solid line; its parameters are given at the top of the plot. The dashed line indicates the theoretical maximum MWCV (Equation 2.4). The histograms show the distribution of the samples on both axes.*
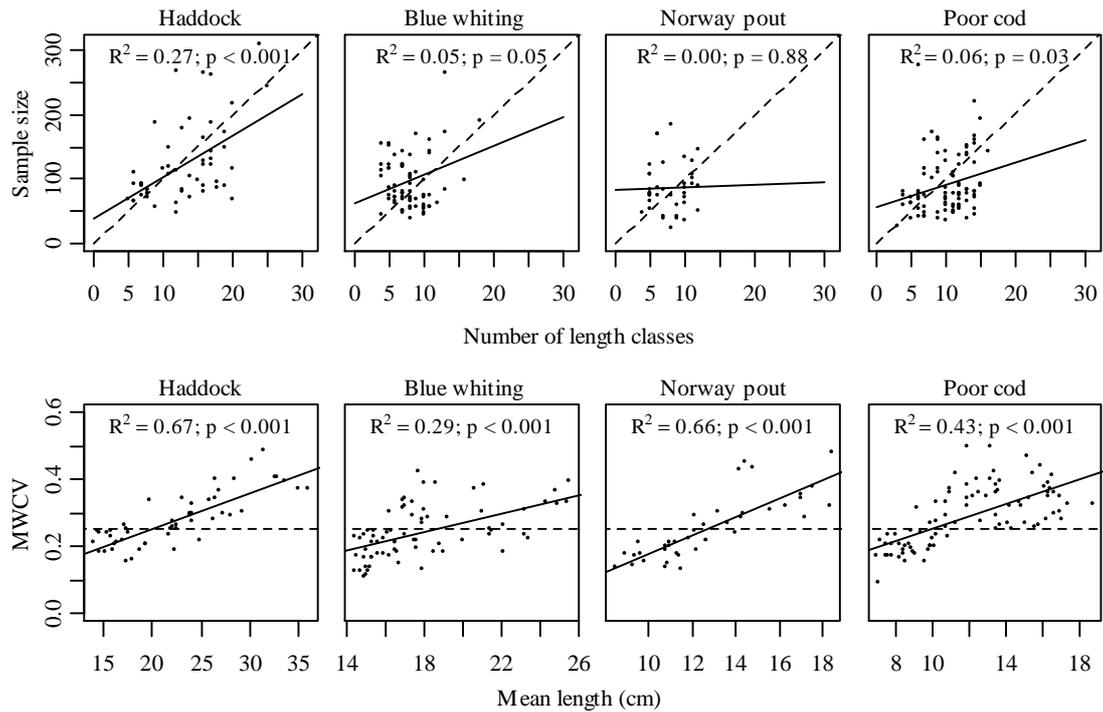
Figure 2.2. The sample sizes of sub-samples taken on the survey were correlated with the number of length classes in the samples for haddock (Melanogrammus aeglefinus) and poor cod (Trisopterus minutus), but not significantly so for blue whiting (Micromesistius poutassou) and Norway pout (Trisopterus esmarkii) (top row). There was considerable variation in the MWCV, which correlated with the mean length of fish in the samples (bottom row). The solid lines represent linear regressions and the dashed lines indicate the sample sizes and MWCV that would have resulted from a sampling scheme where the sample size was chosen to be 10 times the number of length classes in the distribution. The coefficients of determination, R^2, are given together with their p-values.

Variability in the assignment of maturity stages of plaice (*Pleuronectes platessa* L.) and whiting (*Merlangius merlangus* L.) using macroscopic maturity criteria

3.1. Abstract

This study investigates if a macroscopic maturity scale can be applied consistently, by examining the variability between and within ten people who repeatedly assessed the sex and maturity stages of 80 plaice (*Pleuronectes platessa* L.) and 79 whiting (*Merlangius merlangus* L.) gonads. In most cases, agreement within assessors was not significantly higher than agreement between assessors, suggesting that variability was random and not due to differences in interpretation. This finding was supported by the fact that a significant bias was only found for one assessor. Some maturity stages were assigned quite consistently, while other stages were not defined objectively enough to be assigned reliably, even when fish were assessed repeatedly by the same person. For both species, well-defined maturity scales with fewer stages would be preferable over scales that distinguish a larger number of maturity stages. As maturity staging will always contain a form of subjective judgement, it should be subject to continuous quality control measures.

3.2. Introduction

Operator error is potentially a significant factor in many aspects of data collection in fisheries science. While any measurement error should be subject to some sort of quality control mechanism whereby accuracy, precision and bias are being monitored; this is particularly important when subjective criteria are used to quantify a variable. Although even measures such as fish length can have a source of operator error (King, 1995) this is more likely to be a problem for parameters like fish age and maturity stage, where subjective judgements play a large role. There is a large body of literature available on the quality control and quality assurance in the field of age reading (Kimura and Lyons, 1991; Campana *et al.,* 1995; Eltink *et al.,* 2000; Campana, 2001; Kimura and Anderl, 2005). However, the literature on the assignment of maturity stages to fish gonads is more limited. Most work seems to be concerned with routine validation of maturity staging. The most widely used method to determine the maturity stage of fish is visual examination of the gonads, applying macroscopic criteria, a selection of which is usually validated by histological analysis (West, 1990). This work is published mostly as working documents or reports, with a few exceptions (e.g. Tomkiewicz *et al.,* 2003b; Claereboudt *et al.,* 2005). However, apart from validation (a measure of accuracy), any quality control program should also include measures of precision and bias. Measures of precision should include variability between as well as within assessors, which can only be obtained by repeated assessment of the same sample. Bias measures whether disagreement is systematic and differences between workers are the result of differences in interpretation of the maturity scale or whether disagreement is random. This information can be used to improve the design of maturity scales and to monitor training or calibration exercises.

It appears that only one study has been published in which precision of maturity staging was investigated (ICES, 2005a). In this study, 72 images of herring gonads were sent to a number of laboratories to be assessed by various people. The study concluded that agreement was higher for maturity scales that distinguished fewer stages, but no correction was made for the fact that a scale with fewer stages has a higher amount of chance agreement. No comparisons were made within assessors, nor

was it investigated whether disagreement was random or systematic. Additionally, the use of photographs might not be an accurate reflection of assigning fresh material.

The current study investigates whether macroscopic maturity scales can be applied consistently by examining the variability in the assignment of maturity stages to fresh samples of plaice (*Pleuronectes platessa* L.) and whiting (*Merlangius merlangus* L.). The study will address the agreement of individual maturity stages, agreement between and within workers and check whether disagreement is random or systematic. Histological validation will not be addressed in the current study but will be the subject of future work.

3.3. Methods

Two samples of plaice were taken during the Irish Groundfish Survey, carried out in 2004 by the Marine Institute. Additionally, two samples of whiting were taken during the Biological Sampling Survey, which was undertaken in 2005 by the Marine Institute. The first plaice sample was taken on 17 October in ICES area VIa-South (north of Ireland); the second sample was taken on 4 November in ICES area VIIg (Celtic Sea). Both samples were selected from a catch of about 150 plaice, which were sorted into the available maturity categories. Around seven fish were chosen at random from each available sex and maturity category, so that the sample contained approximately equal numbers in each category. Both plaice samples contained a total of 40 fish. At the time of sampling, many plaice had already started gonad development for the next spawning season, which starts around January (Armstrong *et al.,* 2001), so a variety of maturity stages were available. Both samples of whiting were taken in ICES area VIIj (Celtic Sea) on 4 March, just before the peak of spawning for whiting. During that time nearly all maturity stages except spent fish were present. The whiting samples were taken from a catch of around 300 fish and contained 40 and 39 fish respectively from the range of available maturity stages. The fish were tagged with complex codes to ensure the assessors would not remember individual fish. The assessors were not aware that there were approximately equal numbers of fish from each maturity stage.

The maturity of the fish was assessed using the maturity scale given in Table 3.1. In both plaice samples, maturity stages F1, F2 and F4 were identified for females and stages M1, M3 and M4 were found for males. For plaice, maturity stages F3 and M2 are usually not distinguished but the numbering is maintained for consistency with

other species. In the whiting samples, maturity stages F1 to F5 were identified for females and stages M1 to M4 for males.

The plaice samples were assessed by four people each to allow for comparisons between workers. The samples were then stored overnight at 4°C in a sealed plastic bag to keep them as fresh as possible. The next day, the samples were assessed again by the same set of people to allow for comparisons within, as well as between, assessors. The whiting samples were assessed by three people, both samples were assessed twice by each person on the same day; these samples could not be stored overnight as they tended to deteriorate more rapidly than plaice. There was some variation in the experience of the workers concerned. Person A, who had a number of years of experience in the assessment of maturity stages, trained all others except person C, who was trained independently. All others had little previous experience in maturity staging. Training took place while at sea using both fresh and photographic material.

To examine the agreement between two people classifying the same sample, Cohen (1960) proposed a coefficient of agreement, kappa, which takes into account that a certain amount of agreement between two judges might be expected by chance. Cohen's kappa is, in other words, the proportion of agreement after chance agreement has been removed. It is given by:

(3.1) $$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected by chance. For a comparison between persons A and B using n categories ($i = 1, 2, \ldots, n$), the proportion of agreement expected by chance is given by:

(3.2) $$p_e = \sum_{i=n}^{n} p_{a,i} \cdot p_{b,i}$$

where $p_{a,i}$ is the proportion of the sample that person A attributes to category i and $p_{b,i}$ is the proportion of the sample that person B attributes to category i. Kappa ranges from -1 to 1, where $\kappa = -1$ indicates complete disagreement and $\kappa = 1$ indicates complete agreement. Cohen's kappa is widely used in the social and medical sciences where nominal and ordinal scales are used regularly.

The interpretation of kappa can be sensitive to asymmetric or systematic disagreement and to very high or very low prevalence (Schuster, 2004). Systematic disagreement

will be tested for using binomial tests (see below). No cases of extreme prevalence occurred as the samples were selected to have approximately equal numbers in each maturity category. Additionally, the use of kappa requires the assumption of independent ratings (Cohen, 1960). This assumption is met, as each sample only contributes to one paired rating for each comparison and the assessors worked without knowledge of the others' results. The fact that two assessors assess the same samples does not contravene the assumption of independence. Cohen's kappa was originally proposed for the use of nominal scales and another parameter, weighted kappa, was later introduced to deal with ordinal scales (Cohen, 1968). The maturity scale is an ordinal scale as one maturity stage follows the next, however for the current study no distinction will be made between a disagreement over consecutive points on the scale (e.g. F1-F2) or points further apart (F1-F5), therefore unweighted kappa is appropriate in the present case.

Differences between correlated kappa values can be evaluated by bootstrapping the pairs of observations (Efron and Tibshirani, 1993; McKenzie *et al.,* 1996). For each kappa value, 1000 bootstrap replications were performed and confidence intervals (CI) were generated for the differences between two bootstrapped kappa values. The 95% CI were estimated from the percentiles of the differences. If the CI did not include zero, the difference was considered larger than zero and kappa values were significantly different at the 5%. For differences between two mean kappa values, the means were calculated for each bootstrap replication and otherwise the same procedure was followed.

Kappa quantifies agreement but it does not give any information on whether disagreement is random due to bias. One assessor might interpret the maturity scales in a systematically different way from the others or disagreement might be random. If disagreement is random, the number of fish that one assessor assigned to a certain category with which a second assessor did not agree, should not be significantly different from the number of fish that the second assessor assigned to this category with which the first did not agree. If this difference is significant, disagreement could be systematic and might result from a difference in the interpretation of the maturity scale. As the number cases of disagreement in each category is generally low (n<10), a binomial test is appropriate to test for a systematic difference (Zar, 1999). The binomial test was performed separately for each of the available categories. To reduce the risk of type I error (false positive), a Bonferroni correction for repeated tests was

applied to the significance level (Bland and Altman, 1995). With k categories (one test for each category) and an overall significance level of 5%, the significance level for each individual test, α_k, is given by:

(3.3) $\quad \alpha_k = 0.05/(k-1)$

All computations were performed in the R environment (Venables *et al.,* 2005).

3.4. Results

Table 3.2 shows the total of all pair-wise combinations between assessors by maturity stage for the plaice samples. There was some variability in the assignment of male and female plaice at the early maturity stages. This was mainly due to one person, assessor G. Stage F4 was assigned with the highest consistency in plaice, while there was considerable disagreement between stages F1 and F2. For male plaice, most of the disagreement was between stages M3 and M4. Table 3.3 shows the figures for whiting. The most consistent maturity stage for whiting was F5, with high disagreement between F2 and F3 for females and around M1 and M2 for males. In general, most of the disagreement was confined to consecutive maturity stages, indicating that the distinction between these stages was not clearly defined.

Kappa values were calculated for comparisons between and within assessors (Table 3.4 and Table 3.5). To test if there were differences in agreement between the first and second time the samples were assessed (e.g. due to deterioration of the samples), the mean kappa values for all comparisons within the first assessment, were compared with those within the second. For plaice sample I, the mean kappa value was 0.83 for the first time and 0.76 for the second time the samples were assessed. For plaice sample II, mean kappa values were 0.70 for the first and 0.73 for the second assessment. Neither of these differences was significant (bootstrapped difference, p>0.05). Whiting sample I had a higher mean kappa value for the first assessment (0.71) than for the second (0.60) but whiting sample II had a lower mean kappa value for the first assessment (0.65) than for the second (0.80). Again, neither difference was significant at the 5% level. Therefore, there seems to be no reason to assume that there were any differences in agreement between the first and the second time the samples were assessed.

To examine if agreement within assessors was higher than between assessors, the differences in the mean kappa values for comparisons within and between assessors were evaluated. For plaice sample I, the mean kappa value for comparisons within

assessors was significantly higher than the overall mean kappa value between assessors (0.89 and 0.80 respectively; bootstrapped difference: $p<0.05$). For plaice sample II, the mean kappa within assessors was nearly the same as the mean kappa between assessors (0.70 and 0.71); the difference was not significant. As both whiting samples were assessed by the same set of people, the samples were combined; the mean kappa within assessors was higher than between assessors (0.70 and 0.65 respectively) but this difference was not significant.

The most experienced workers, A and C, obtained higher kappa values for within-assessor comparisons than most others, although assessors D and F also scored highly (Table 3.4). Person C scored low kappa values for between-assessor comparisons and appeared to assign some maturity stages in a different way from the others: for the first time plaice sample I was assessed, a significant systematic difference was found between persons A and C (two-tailed binomial test, $p<0.01$), for the second time the sample was assessed systematic differences found between B and C and between C and D (two-tailed binomial test, $p<0.01$). In every case, person C was more likely to assign stage F2 and less likely to assign F1 than the others. Assessor C can therefore be considered to be biased towards the others. No significant systematic differences were found in any of the other comparisons. Another person that deviated from the general pattern was assessor G (plaice sample II). Agreement was very low for all comparisons with this assessor and particularly within person G. This indicates that person G was generally inconsistent in the assignment of maturity stages.

3.5. Discussion

An important assumption of the study is that the first time the samples were assessed, was independent from the second assessment. Although the workers were aware that they were assessing the same sample twice, the experiment was designed to make it difficult for anyone to remember individual fish. The fish were labelled with complex codes and samples contained few very small, very large or otherwise distinctive fish. The two whiting samples were given in a different order than the assessors were told to expect but this was not noticed by anyone so individual codes or fish were not remembered. The first and second assessment of each sample can therefore be considered independent. It is also implicitly assumed that the condition of the samples did not deteriorate over time. There is little reason to assume that they did, as the mean agreement did not decrease significantly for any of the plaice or whiting

samples. It can therefore be assumed that comparisons between and within the assessors are valid.

Variability appeared to be high for certain maturity stages, while others appeared to be quite well defined. It is assumed that fish of either sex at maturity stages 1 and 2 will not spawn in the current season and that all fish of stage 3 and higher are mature. Therefore the most important distinction is between stages 2 and 3. For plaice there was little disagreement between the assignment of mature and immature fish. However, for whiting, there was considerable variability between the immature females (F2) and maturing females (F3). This could potentially have a large effect on the estimate of the proportion of mature fish in the population. In the present case, maturity stage F3 was quite rare in the catches (<4%), so this effect is probably small for the current survey. However, these results highlight the importance of timing of sampling. Size frequency analysis of the oocytes of plaice and whiting (Gerritsen *et al.,* unpublished data; Appendix B page 110) shows that vitellogenic cells in plaice quickly reach a relatively large size. On the other hand, in whiting there is no obvious gap in the size range between pre-vitellogenic and vitellogenic cells. It is therefore more difficult to distinguish females with vitellogenic cells (F3; mature) from those without vitellogenic cells (F2; immature) in whiting.

One person, assessor C, was shown to be biased towards the other assessors. It is interesting to note that C was trained independently from the others. If person C is omitted from plaice sample I, mean kappa within and between assessors is no longer significantly different. So, with the exception of person C, there does not seem to be a higher agreement within assessors than between them. This suggests that the disagreement found was almost entirely due to chance and not due to differences in interpretation of the maturity scale. This has a number of implications, firstly, it is unlikely that the consistency of most assessors can be much improved by further training; some maturity stages simply could not be assigned consistently, even when samples were assessed repeatedly by the same person. On the other hand, it should be possible to address the issues of systematic disagreement (assessor C) and major inconsistency (assessor G) by calibration exercises and additional training.

For the purposes of quality control, it would be very useful to repeatedly assess a number of maturity samples each time they are collected. However comparing more than two assessors or repeatedly assessing the same assessor is quite time consuming and might not be possible on a routine basis. Additionally, unlike age reading, it will

not be possible to return to the maturity samples and assess them again if agreement falls below a certain threshold. On the other hand it would be feasible for two assessors to note maturity stages independently of each other as part of the normal sampling procedure. This information would be valuable to monitor the progress of training or calibration exercises. It would also contribute to finding a realistic value of kappa at which agreement is considered adequate. At the moment this is not possible, as too few samples have been assessed so far.

Hunter and Macewicz (2003) suggested that maturity scales could be improved by reducing the number of classes and focussing on the most reliable characteristics. The present study seems to confirm this; maturity stages that were not unequivocally defined could not be distinguished consistently and therefore the value of a maturity scale with a large number of stages is doubtful. The study on herring gonads (ICES, 2005a) also seems to support this finding. A maturity scale that distinguishes between virgin, maturing, spawning and spent fish might be a good compromise between gleaning detailed information and obtaining consistent results. The current findings will most likely generalise to other species in other areas as most maturity scales are based on similar criteria.

3.6. Acknowledgements

3.7. Tables

Table 3.1. Macroscopic maturity scale for plaice and gadoids. Adapted from Bromley (2000) and Tomkiewicz et al. *(2003b).*

Females		Males	
F1	Ovary small, contents translucent, nearly colourless	M1	Testes tiny, translucent and undeveloped
F2	Contents blurred translucent, pinkish, no oocytes visible	M2	Gadoids: lobules developing but still pinkish and translucent. Plaice: no stage M2
F3	Gadoids: ovary opaque, contents yellow, individual oocytes not yet visible. Plaice: no stage F3	M3	Testes developing, creamy white and opaque. Testes less than 50% of full size
F4	Ovary contains opaque oocytes, clearly visible when ovary is cut	M4	Testes filling but no sperm visible when cut
F5	Ovary contains hydrated cells but does not run	M5	Sperm visible when cut but testes do not run
F6	Like stage 5 but runs under moderate pressure	M6	Testes run under moderate pressure
F7	Ovary slack, contains slime and sometimes remaining oocytes	M7	Testes can be quite small, fluid remaining in sperm duct

Table 3.2. Tally of all pair-wise comparisons between assessors by maturity stage for the combined plaice samples. Percentages of raw agreement are given.

Plaice	M1	M3	M4	F1	F2	F4	Agreement
M1	105	13	1	19	3	0	74%
M3		126	49	10	3	0	63%
M4			150	1	3	0	74%
F1				74	83	0	40%
F2					141	10	58%
F4						169	94%

Table 3.3. Tally of all pair-wise comparisons between assessors by maturity stage for the combined whiting samples. Percentages of raw agreement are given.

Whiting	M1	M2	M3	M4	F1	F2	F3	F4	F5	Agreement
M1	50	15	2	2	0	2	1	0	0	69%
M2		9	9	0	0	0	2	0	0	26%
M3			12	34	0	0	0	1	0	21%
M4				44	0	0	1	1	0	54%
F1					3	4	0	0	0	43%
F2						77	46	2	0	59%
F3							46	8	0	44%
F4								32	10	59%
F5									61	86%

Table 3.4. Kappa values for comparisons between and within assessors A-G for plaice samples I and II. Kappa values above the diagonal are from the first time the samples were assessed; below the diagonal are the values for the second assessment. Comparisons within assessors are shaded.

Plaice I	1st assessment					Plaice II	1st assessment			
2nd assess	A	B	C	D		2nd assess	A	E	F	G
A	0.91	0.97	0.70	0.97		A	0.88	0.84	0.76	0.66
B	0.79	0.79	0.73	0.94		E	0.88	0.67	0.72	0.60
C	0.70	0.66	0.90	0.67		F	0.76	0.79	0.81	0.63
D	0.94	0.73	0.71	0.94		G	0.64	0.64	0.67	0.42

Table 3.5. Kappa values for comparisons between and within assessors I-K whiting samples I and II. See Table 3.4 for further description.

Whiting I	1st assessment				Whiting II	1st assessment		
2nd assess	I	J	K		2nd assess	I	J	K
I	0.54	0.71	0.46		I	0.76	0.59	0.62
J	0.66	0.59	0.65		J	0.59	0.81	0.67
K	0.74	0.74	0.70		K	0.56	0.94	0.76

A simple method for comparing age-length keys reveals significant regional differences within one stock of haddock (*Melanogrammus aeglefinus* L.)

This chapter is reproduced from the following original publication:

> Gerritsen HD, McGrath D, Lordan C (2006) A simple method for comparing age-length keys reveals significant regional differences within a single stock of haddock (*Melanogrammus aeglefinus*). ICES J. Mar. Sci. 63: 1096-1100

An earlier version of the paper has been presented to the ICES workshop WKSDFD:

> Gerritsen H (2005) Modelling age-length keys. ICES Workshop on Sampling Design for Fisheries Data (WKSDFD). Pasajes, Spain, 1-3 February 2005

4.1. Abstract

A multinomial logistic model is presented as a tool for comparing two or more age-length keys. The model provides an objective way to fill in missing values and can be used for estimating uncertainty and visualising age-length keys (ALKs). An example of haddock (*Melanogrammus aeglefinus* L.) in ICES Division VIa (West of Scotland) is used to illustrate that significant regional differences in the proportions of age classes-at-length can exist on a small spatial scale. These differences were caused by regional variation in both length-at-age and relative abundance-at-age. As the length-at-age data are not normally weighted by the local catch rate (abundance), the ALK of the combined age data can result in strongly biased estimates of numbers-at-age. In the present case, the use of unweighted age data would have resulted in an over-estimate of recruitment of nearly 200% and an under-estimate of the spawning stock biomass of 15%. Comparing ALKs using this method will have several applications in fisheries science.

Key words: age-length key; multinomial logistic model; sampling design; haddock.

4.2. Introduction

Most fisheries stock assessments are based on estimates of numbers of fish per age class. Sampling for age data generally takes place on a non-random (length-stratified) basis where sampling targets are set by length class. Additionally, a larger random

sample is taken to obtain the length frequency of the catch or landings. To estimate numbers at age, the aged sample is raised to the total length frequency using an Age-Length Key (ALK), which consists of the proportions at age for each length class (Fridriksson, 1934). The length-stratified sampling strategy ensures that fish from a wide range of sizes are represented in a relatively small aged sample.

All age-at-length data from an entire stock are often combined without weighting under the assumption that differences between gear types or regions can be disregarded (e.g. ICES, 2005c). Differences in size selectivity among gears should not influence the proportions of age classes at a given length, assuming that within each length class the probability of capture is independent of age. However, regional differences in the length-at-age distributions do have the potential to result in a biased ALK. These differences might be caused either by variation in length-at-age distributions or by variation in relative abundance of the age classes. For example, fish of a certain age might have a larger mean length in one area than another, due to differential growth rates or size-specific migration. Additionally, in certain length classes, proportions of young fish might be higher in nursery areas than elsewhere, simply because they are locally more abundant, relative to other age-classes.

Various methods have been applied to test for differences between ALKs. Hayes (1993) and Horbowy (1998) both suggested comparing individual cells of the ALKs using multiple Fisher's or Chi squared tests. Although the application of these tests is straightforward, the interpretation of the results is not, as there are as many p-values as the number of age and length classes that are considered. Additionally, any cells that do not contain enough data, have to be omitted, so the tests can only be applied to large data sets. Dwyer *et al.* (2004) took a different approach and suggested applying a two-dimensional Kolmogorov-Smirnov test. This approach only requires a single test to compare two ALKs. However, the two-dimensional Kolmogorov-Smirnov test is not widely available in statistical packages, nor does it appear to be the most parsimonious solution. Rindorf and Lewy (2001) applied multinomial models of continuation-ratio logits to aged data. This approach has many advantages, however Rindorf and Lewy's model required a polynomial function to be defined to allow every possible type of distribution to be modelled. However if one makes the assumption of normality in the length-at-age distributions, Rindorf and Lewy's method can be greatly simplified by removing the need for arbitrary smooth functions.

The assumption of normality in length-at-age distributions is routinely being made, either with constant variance over the age groups or variance proportional to mean length (e.g. Schnute and Fournier, 1980; Labonté, 1983; Rosenberg and Reddington, 1988). However, in contrast to these studies, the current assumption of normality is a weak one and applies only to the population from which sample was drawn, not the aged sample itself (which is non-random) or even the catch (which is often size selective).

The suggested approach allows for multinomial logistic models to be applied, testing for differences between ALKs. In addition, the models can be used to predict missing values, estimate uncertainty and help visualise ALKs. The method will be illustrated by examining the variability in ALKs of haddock to the west of Scotland (ICES Division VIa) through the application of multinomial models to age-at-length data from the 2004 Irish Groundfish Survey.

4.3. Methods

Logistic models with a binomial error distribution are widely used in fisheries science to describe the relative proportions of two overlapping distributions. Examples include size selection ogives for fishing gear, discarding ogives and maturity ogives. In the case of ALKs, there are mostly more than two overlapping length-at-age distributions and therefore a multinomial logistic model is required to describe the proportions of age-at-length. Multinomial models can be fitted by maximising the product of the conditional binomial trials simultaneously (Beare and McKenzie, 1999; Rindorf and Lewy, 2001). Alternatively the S-PLUS® and R packages provide the function *multinom()* which fits multinomial log-linear models via neural networks (Venables and Ripley, 1994). Some examples of R-code applying multinomial models to age-at-length data are provided in Appendix C, page 114.

Multinomial model selection, testing and estimation can be carried out in a similar way to generalised linear modelling (McCullagh and Nelder, 1989). Model selection allows one to identify which factors contribute significantly to the explanatory power of the model and to test for differences between regions, gear types etc. Model estimation can be used to interpolate missing values. It is a regular occurrence that for certain length classes in the total length frequency, no aged samples are available. These gaps in the data need to be filled in to allocate numbers-at-age for the relevant length classes. The multinomial logistic model provides an objective way to do so.

Here, ALKs of haddock (*Melanogrammus aeglefinus* L.) were obtained from the Irish Groundfish Survey, carried out by the Marine Institute in October and November 2004 on RV "Celtic Explorer". Data from ICES Division VIa (West of Scotland) were selected to illustrate the method. The area was divided into three depth strata: shallow (<75m), medium (75m-125m) and deep (>125m). Sampling targets of five age samples per cm length class were set for each of the strata, so a separate ALK was available for each stratum. Fish ages were determined by sectioning the sagittal otoliths through the nucleus and counting the number of hyaline rings.

Multinomial logistic models of the following form were fitted:

(4.1) $A \sim L + S + L.S$

where A is the predicted age distribution at length L in stratum S. L was fitted as a continuous variable, S as a factor and L.S is an interaction term. The significance of the factor stratum in the model was tested by comparing the initial model to a model without that factor. The difference in residual variance of these nested models was tested against the difference in the model degrees of freedom (υ) using the Chi-square test (Collett, 2003). For the current analysis, age classes of 4-year-olds and older were combined into a single plus-group. As catches of 0-group fish were scarce and did not overlap in size with the other age classes, they were omitted from the analysis.

All haddock from VIa are considered to be a single stock and for the purposes of stock-assessment it is common practice to use a single ALK to obtain numbers-at-age without weighting the age data in any way (ICES, 2005c). For the present study, numbers-at-age in the survey catches were estimated in two ways: firstly by combining all age data without weighting and secondly by weighting the age data by the relative abundance in each stratum. The relative abundance in each stratum was estimated from the catch numbers per unit effort (CPUE), multiplied by the surface area of each stratum. The unit effort is a standard half-hour trawl, towed at 3kn. The length frequency data were expressed as CPUE and weighted by stratum surface area in all cases to obtain an unbiased length frequency for the combined strata.

Standard errors for the numbers-at-age estimates were obtained using a bootstrapping routine (Efron and Tibshirani, 1993). The individual fish in the aged sample were treated as independent sampling units and re-sampled 500 times. This approach, as opposed to re-sampling within length classes, can result in length classes without data, therefore a multinomial model was fitted to the data for each bootstrap iteration. Standard errors were estimated from the standard deviation of the bootstrapped

estimates from the modelled data. The length distributions were assumed to be known without error.

4.4. Results

A very highly significant stratum effect was found for a model that contained data from all three strata (χ^2=133.3; υ=16; p<0.001). When the shallow stratum was omitted from the dataset, the stratum effect was no longer significant (χ^2=9.2; υ=8; p=0.32). However, if either one of the other strata were omitted, the stratum effect remained highly significant. This indicates that the ALK of the shallow stratum was significantly different from the ALKs of two other strata and that the ALKs of the deep and medium strata were not significantly different from each other. Figure 4.1 shows the observed and modelled proportions at age and length distributions. The figure indicates that the main difference between the strata lies in the proportions of one-year-olds in length classes 25-35cm, which were considerably higher in the shallow stratum than in the other strata.

In the medium and deep strata, two-year-olds were by far the most common age class in the catches (Table 4.1). In the shallow stratum, one-year-olds were most abundant, relative to other age classes. In addition, the mean length-at-age appeared to be higher for most age classes in the shallow stratum than in the others (Table 4.2). Combining all aged data into an ALK without weighting, resulted in estimated catch numbers for one-year-olds that were nearly twice as high (88 fish per unit effort) as the estimate using age data weighted by abundance (47 per unit effort; Table 4.2). If the present data were used as an absolute estimate of the spawning stock biomass, the unweighted estimate would have resulted in an under-estimate of the spawning stock biomass by 15%, assuming a knife-edge maturity at age two (ICES, 2005c).

The main reason for the bias in the unweighted ALK, appears to be that fish from the shallow stratum were over-represented in the sampling. Catch rates in the shallow stratum were around 8 times lower than in the medium and deep strata, but the sample numbers for age were actually higher in the shallow stratum (Table 4.2). As the one-year-olds in the shallow stratum were relatively abundant (compared to other age classes) and, on average, about 2cm larger than in the other strata, the proportions of one-year-olds at length were over-estimated in many size classes of the unweighted ALK.

4.5. Discussion

The multinomial model used here, is a special case of the methodology presented by Rindorf and Lewy (2001). It eliminates the need to apply a polynomial function to the length classes, which improves the transparency and simplicity of the model. A model with *A* age classes only requires *2(A-1)* model parameters; the apparently complex shape of the model (e.g. Figure 4.1) results from the added proportions of the various age classes.

The assumption of normality applies not to the aged data but only to the underlying population because the model uses proportions (age-at-length), not length-at-age distributions. This is most clearly demonstrated in the binomial logistic case, for example a discard ogive. The symmetric s-shaped curve that describes a discard ogive, results from the proportions of two overlapping distributions: one length distribution of discards and one of landings. If both distributions were strictly normal (at least in the area of overlap) with equal variance, a logistic binomial curve would describe the proportions-at-length exactly, regardless of any size selection in the sampling. For most binomial applications the assumption of normality cannot be made, however the proportions-at-length still tend to follow an s-shaped curve that is closely described by the logistic curve (McCullagh and Nelder, 1989; Collett, 2003). The multinomial case expands on the binomial model by describing the proportions of more than two overlapping distributions. Unlike many binomial applications, length-at-age distributions do tend to be approximately normally distributed with similar variances (e.g. Schnute and Fournier, 1980; Labonté, 1983; Rosenberg and Reddington, 1988).

Sexual dimorphism in growth could result in bimodal, hence not normal, length-at-age distributions. In this case, it might be advisable to sample the sexes separately, as is feasible for some flatfish that can be sexed without dissection. Alternatively, one can apply an age-sex-length-key, which should restore the normal length-at-age distributions; the factor sex could then be added to the multinomial model.

The model appears to be a useful tool to detect significant differences between ALKs, although the likelihood of finding these differences will, of course, depend on the number of fish sampled. The model is also useful for obtaining confidence limits or variance estimates and it can deal with missing length classes: if no aged data exists for a certain length class, the model can predict the expected proportions of the age

classes for that (or any other) length class. In the future, the model might be expanded to include seasonal changes, for example by fitting smooth curves through a time variable.

The current example shows that there can be a high degree of spatial variability in ALKs, which can result in strongly biased numbers-at-age estimates. This has many implications for the unit-stock and dynamic pool assumptions that underlie many age-based stock assessments. Many stocks are known to have nursery areas or age- or size-specific migration and will therefore have regional differences in the age structure. If the number of age samples is proportional to the local abundance of fish, the estimates will be unbiased, but otherwise the aged samples should be weighted by the abundance in each region before they are combined into an ALK to avoid bias. These considerations apply to survey data, as well as to data from commercial sources, where data from many regions are often combined without weighting.

In the present case, the consequences of using an unweighted ALK would be a large bias in the estimated abundance of one-year-old fish. Many stock assessments use survey indices in a relative sense and this bias might be corrected by a catchability parameter. However if the bias changes from year-to-year due to year class effects, changes in survey design or other mechanisms, there will be implications for the assessment and management advice. If this survey were used in an absolute sense (e.g. Beare *et al.,* 2005) the consequence of the bias would have been a nearly two-fold over-estimate of the 2003 year class and an underestimate of the spawning stock by 15%.

4.6. Acknowledgements

4.7. Figure and Tables
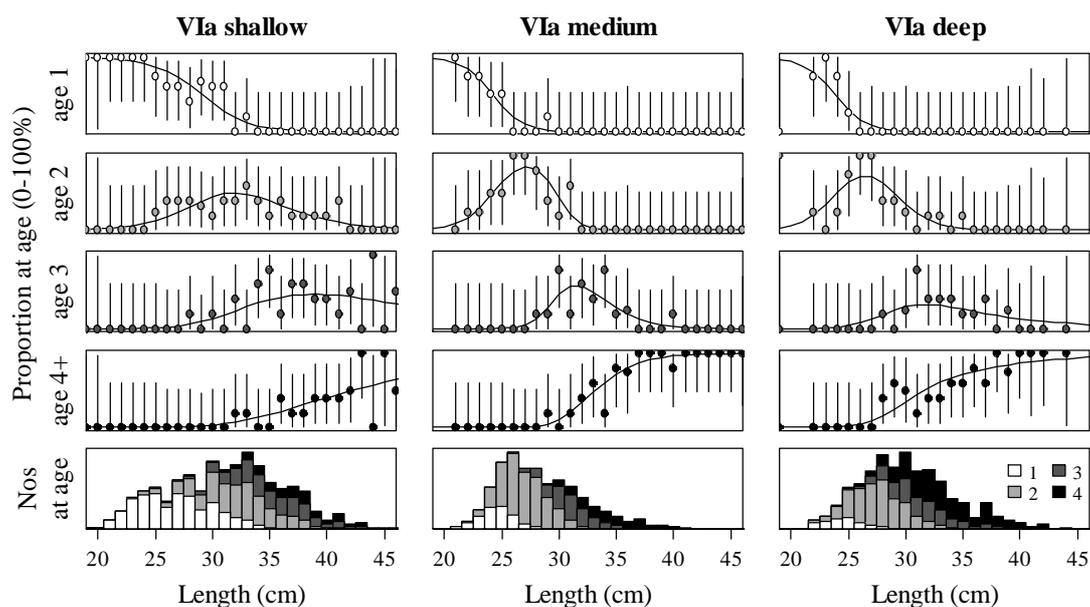
Figure 4.1. Proportions-at-length of age groups 1 to 4+ in the three depth strata and the length-at-age distributions estimated by applying ALK models to the total catch length frequencies. The circles represent the observed proportions with their individual 95% confidence intervals and the curves represent the predicted proportions from the multinomial models. The shades of the stacked bars correspond to the different age classes. The proportions of one-year-olds were higher in the shallow strata than in the other strata for length classes up to 35cm.

Table 4.1. Summary statistics of the three depth strata in ICES area VIa. Given for each stratum are the surface area (in nautical miles); the number of stations; the catch per unit effort (CPUE) and the number of fish sampled for age reading.

	Area (nm)	Stations	CPUE	Nos aged
VIa Shallow	4000	18	59	96
VIa Medium	5400	17	460	62
VIa Deep	2700	6	455	41

Table 4.2. Estimated numbers-at-age and mean lengths-at-age for individual strata and for the combined area after application of weighted and non-weighted ALKs. Weighting the ALKs resulted in considerably lower estimates for the number of one-year-old fish. Standard errors were obtained by bootstrapping and are given in brackets. Numbers-at-age and lengths-at-age were obtained by applying the ALK to the catch length frequency in each stratum.

	Age class					
Stratum	0	1	2	3	4	5
Numbers-at-age (per unit effort)						
VIa Shallow	1 (5)	21 (2)	18 (3)	13 (3)	3 (1)	4 (5)
VIa Medium	3 (0)	71 (19)	226 (24)	91 (17)	40 (10)	30 (7)
VIa Deep	2 (0)	34 (8)	160 (22)	36 (14)	93 (20)	129 (21)
Numbers-at-age for combined strata (per unit effort)						
VIa - not weighted	2 (1)	88 (10)	121 (10)	46 (6)	34 (5)	36 (5)
VIa – weighted	2 (1)	47 (10)	143 (13)	53 (9)	39 (7)	43 (6)
Mean-length-at age (cm)						
VIa Shallow	14.2 (3.1)	26.7 (1.4)	32.1 (0.8)	34.8 (0.7)	36.0 (1.2)	37.4 (3.5)
VIa Medium	15.1 (0.3)	24.6 (0.5)	27.0 (0.2)	31.0 (0.4)	33.3 (0.7)	36.0 (0.9)
VIa Deep	13.8 (0.3)	24.7 (0.7)	27.7 (0.4)	31.7 (0.8)	31.9 (0.6)	33.2 (0.5)

Significant differences in the length-weight relationships of neighbouring stocks can result in biased biomass estimates: examples of haddock (*Melanogrammus aeglefinus*, L.) and whiting (*Merlangius merlangus*, L.)

5.1. Abstract

Length-weight relationships of fish are often used for estimating biomass or to obtain an index of condition. Although large-scale spatial trends are known to exist, it is often assumed that length-weight relationships do not vary significantly within stocks or between neighbouring stocks. The present study examined length-weight relationships of 1334 haddock (*Melanogrammus aeglefinus*) and 1186 whiting (*Merlangius merlangus*) collected on a groundfish survey in the waters around Ireland in 2004. Additionally, condition indices were estimated for individual fish and for length frequency samples to summarise their length-weight relationships. The length-weight regression showed a significant area effect and no differences between the sexes. The condition indices showed a moderate spatial structure for both species: around 25% of the variability could be explained by the location of the samples, the rest of the variability was due to other sources. Length-weight relationships did not appear to vary significantly within stocks, however differences between stocks were significant. In the present case, a bias of up to 10% could occur in biomass estimates as a result of applying length-weight relationships of one stock to length data of a neighbouring stock.

Key words: length-weight; condition; haddock; whiting; biomass estimate.

5.2. Introduction

Length measurements can be obtained quicker and under a larger range of circumstances than weight measurements, therefore a limited number of weight observations is often used to construct a length-weight relationship. This relationship can then be used to convert length distributions into weights for biomass estimates. A widely used relationship between length (L; in cm) and weight (W; in grams) is the power function:

(5.1) $\quad W = aL^b$

This relationship can also be used to estimate a condition index. By keeping parameter b constant for a species or stock, parameter a can be estimated for individual fish and used as a condition index (Anderson and Neumann, 1996).

The parameters a and b can either be estimated by linear regression on the log-transformed variables (Ricker, 1975) or by non-linear regression of non-transformed variables (Hayes *et al.,* 1995). The two methods differ in their assumptions on the error structure. Linear regression on the log-transformed variables assumes that the errors in the observed weights are log-normally distributed and multiplicative, while non-linear regression assumes a normally distributed, additive error structure (Hayes *et al.,* 1995). Linear regression results in a bias due to the logarithmic transformation: the model passes through the geometric mean, rather than the arithmetic mean, but this can be adjusted using a simple correction factor (Sprugel, 1983).

Many biological parameters are known to vary over small geographical ranges (e.g. Armstrong *et al.,* 2004; Gerritsen *et al.,* 2006). Nevertheless, for stock assessment purposes, length-weight relationships are often assumed to be uniform for an entire stock. When data are sparse for a certain stock, length-weight relationships from neighbouring stocks are sometimes applied (e.g. ICES, 2004b). However, regional differences in the length-weight relationships and condition indices of fish are known to exist (Brodziak and Mikus, 2000; Rätz and Lloret, 2003). These differences could potentially bias biomass estimates. It is presently unclear to which extent differences exist within stocks or between neighbouring stocks. The aim of the present study is to investigate the spatial variability in length-weight relationships and condition indices of haddock (*Melanogrammus aeglefinus*) and whiting (*Merlangius merlangus*) in the waters around Ireland and quantify the potential effects this regional variation might have on biomass estimates.

5.3. Methods

Length and weight data were collected on the 2004 Irish Groundfish Survey. This survey is carried out annually on the RV "Celtic Explorer" in the months of October and November. The 2004 survey covered 161 stations around the Irish and Northern Irish coasts in a depth range from 10 to 250m. Trawling took place during daylight hours for 30 minutes at 3kn of speed over the ground using a GOV trawl (ICES, 1999). The survey area was divided into 14 strata, based on ICES Divisions and bottom depth. Each ICES Division was divided into shallow (<75m), medium (75m-125m) and deep (>125m; where present) depth bands (Figure 5.1). The catch was speciated, weighed and samples were taken for length measurements. Additionally, biological samples were taken on a length-stratified basis to obtain the age, round weight, sex and maturity stage of individual fish. A target of five biological samples per cm length class was set for each stratum. The length of the fish was measured to the nearest cm below the total length and round weights were recorded in grams. Haddock and whiting were the most abundant commercially exploited demersal species caught on the survey. Individual weights were recorded for 1334 haddock and 1186 whiting.

Values in the length-weight relationships that were obviously spurious, were removed from the dataset after examination of plots of the raw and log-transformed variables. These spurious values amounted to less than 0.2% of the observations. In addition, all fish under 15cm were omitted, as the precision of the weighing scales was considered too low for fish at those sizes. The residuals from linear models of the log-transformed lengths and weights were approximately normally distributed and their variance appeared constant over the range of the predictor variable. Therefore linear, rather than non-linear modelling techniques were applied. No bias correction was applied, as the length-weight relationships were only evaluated relative to each other and not in absolute terms.

A stepwise linear regression procedure was employed to identify influential variables (Draper and Smith, 1998). The Akaike Information Criterion (AIC) was used to evaluate the improvement of the model when adding or dropping a term (Sakamoto *et al.,* 1986). The predictor variables for *ln(weight)* that were evaluated, were the main effects *ln(length)*; *stratum*; *stock*; *bottom depth*; *age* and *sex* and all possible interactions between them. Age classes of three years and older were collapsed into

one group due to the low catch numbers at these ages. The terms *ln(length)* and *depth* were fitted as continuous variables, the terms *stratum*, *stock, age* and *sex* were fitted as factors. Most gonads were inactive and small at the time of sampling; therefore information on the maturity stage was not included as an explanatory variable. The term *stratum* is likely to be correlated with the terms *bottom depth* and *stock*. Therefore, the terms made available to the models included either *stratum* or *bottom depth* and *stock* but never all three terms.

Due to the relatively low numbers of biological samples at each station, it was not possible to accurately estimate both parameters from Equation (5.1) for each individual haul. Therefore, in order to obtain information on the length-weight relationship on a fine spatial scale, parameter *b* was estimated for the combined data of the entire survey and the condition index (parameter *a*) was then estimated for individual fish by solving Equation (5.1). The condition index for individual fish was used to estimate variability in the condition of individuals within hauls.

The condition index was also estimated for entire length samples for which only a bulk weight was available by solving *a* in the following equation:

$$(5.2) \quad W_{bulk} = a \sum_{i=n}^{n} L_i^b$$

Where W_{bulk} is the bulk weight of a sample of *n* length measurements L_i ($i = 1,2,\ldots,n$). This sample condition index was used to investigate the variability between hauls and the spatial structure in the condition. Spurious values and samples of less than 10 fish were removed, leaving in 108 hauls for haddock and 112 hauls for whiting. At some stations the fish were graded into size classes, resulting in two or more length samples per haul. For the current analysis, these samples were combined, resulting in one sample condition index per location. It was considered that this bulk condition index was a more representative measure of the condition of fish in a particular haul than the average condition index of individual fish taken as part of the length-stratified sampling scheme for biological samples. The reason for this is that the latter is a non-random sample of the catch, while the former consists of a large number of individuals, thus incorporating more of the individual variation in condition.

The spatial structure of the sample condition index was investigated using geostatistical methods (Rivoirard *et al.,* 2000). Station positions were taken as the midpoint of the trawl and projected onto a plane using a transformation of longitude based on the cosine of latitude. Distances between stations were calculated as the

shortest distance between two points, regardless of the presence of landmasses. Experimental variograms were computed using code written in the R environment (R-Development-Core-Team, 2005). A lag spacing of 10nm ±5nm was used; this distance is close to the mean distance of each station to its nearest neighbour. The maximum distance for which the variograms were calculated was 200nm, which was just over half the maximum distance between stations. No weighting was used because each sample condition estimate is derived from a single sample weight and a number of length measurements. The precision of the estimate does therefore not necessarily improve with larger sample numbers. Isotropy was assumed in the geostatistical models, this assumption was tested by investigating the presence of trends in the data by plotting the condition indices against bottom depth, latitude and longitude. Linear and spherical variogram models were fitted and evaluated using the goodness-of-fit statistic, which was weighted by the number of pairs in each distance bin (Rivoirard *et al.,* 2000).

In order to quantify the influence that the regional differences in the length-weight relationships might have on biomass estimates, separate length-weight regressions were obtained for each stratum. These relationships were then used to estimate the weight of a "standard" length distribution of mature fish. This length distribution was the average catch length distribution of mature fish, assuming knife-edge maturation at 25cm for both species. This corresponds roughly to the knife-edge maturation ogive at age two that is applied for stocks of both whiting and haddock around Ireland (ICES, 2004a; b). Confidence limits were estimated from the quantiles of 1000 bootstrap replications from the length-weight data, using the sampling stations as bootstrapping units (Efron and Tibshirani, 1993).

5.4. Results

A simple linear model with *ln(length)* as only predictor variable for *ln(weight)* resulted in a high coefficient of determination, R^2, of around 0.98 for both species (Figure 5.2). The residuals did not show any obvious patterns, nor did the locally weighted running line smoothers that were fitted through the residuals (loess with a span of 25%; Figure 5.2; Hastie and Tibshirani, 1990). This suggests that the linear model provides a good fit.

The stepwise regression procedure identified *ln(length)* as the main explanatory variable for *ln(weight)* for both species, resulting in a very large reduction in the

residual sum of squares (Table 5.1). For both species, the factor *stratum* was the second term in the stepwise procedure to be included into the models; it resulted in a highly significant reduction in the AIC. Inclusion of the term *stratum* resulted in a larger reduction in the AIC than the inclusion of the terms *stock* and/or *bottom depth*, either with or without an interaction term. The terms *stock* and *bottom depth* were therefore omitted from further analysis as they were likely to correlate with the term *stratum*.

The next terms that were added to the model in the stepwise selection procedure were *age* and the interactions *ln(length).age* and *ln(length).stratum* (Table 5.1). Adding these terms reduced the AIC for both species, but the associated reduction in the residual sum of squares was very low. This suggests that these terms have a limited additional explanatory power. No terms were dropped during the stepwise procedure and the term *sex* could not be included at any stage without increasing the AIC, suggesting that this factor does not have significant explanatory power.

The sample condition index varied significantly between strata (ANOVA, $p<0.02$ for haddock and $p<0.001$ for whiting), confirming the highly significant stratum-effect found for the length-weight relationship. The mean length of the samples did not influence the condition index (ANOVA, $p=0.94$ and $p=0.28$ respectively), suggesting that the sample condition index is not influenced by the size composition in the samples.

The variograms for the sample condition index show a large nugget effect for both species of around 75-80% of the sample variance (Figure 5.3). This indicates that 75-80% of the variation is either due to factors other than location or takes place on a smaller scale than the sampling resolution. Nevertheless, the condition indices are not entirely independent of their spatial location; samples taken close together were more similar than samples further apart. The variogram model for haddock increases slowly over the entire range, suggesting a large-scale trend in the data. The whiting variogram stabilises around 150nm after which it does not increase further, indicating a large-scale structure. Data were too sparse to conclusively determine if anisotropy existed. However, no patterns could be discerned when plotting the condition indices against depth, latitude or longitude (data not shown), therefore there was little reason to suspect that the assumption of isotropy was violated.

The condition indices of the individual fish were used to test if the nugget effect observed in Figure 5.3 was due to variation on a smaller scale than the sampling

resolution or due to variation within hauls. For haddock, the mean variance, weighted by sample numbers, within the hauls was 78% of the mean variance of all individual condition indices, for whiting this figure was 73%. This suggests that the nugget effect is not likely to be due to the scale of the sampling resolution, but due to variability that is independent of sampling location. The remaining 22-27% of variability is due to spatial factors.

The ICES working groups for stock assessment (ICES, 2004a; b) distinguish three haddock stocks and four whiting stocks in the waters around Ireland (Figure 5.1). The haddock stocks are allocated as follows: West of Scotland; Irish Sea; Celtic Sea plus West of Ireland. For whiting, the Celtic Sea stock is considered separate from the West of Ireland stock. Figure 5.4 shows biomass estimates obtained by applying separate length-weight relationships for each stratum to the average length distribution of mature fish in the catches. The biomass estimates for haddock were highest using length-weight relationships from the West of Scotland stock. For whiting, the highest biomass estimates were consistently obtained from Celtic Sea length-weight relationships. The lowest biomass estimate was obtained from a length-weight relationship from the West of Ireland. The position of the 95% confidence intervals suggests that differences between some of the stocks are significant, while differences within stocks mostly fall within overlapping confidence intervals.

5.5. Discussion

The linear model appeared to fit the data well and the residuals did not show any patterns. The linear shape of the relationship implies that life-history events like maturation do not influence the length-weight relationship, at least not at the time of sampling. In linear regression, all observations are assumed to be independent. However, sampling took place in a grouped way: each sampling station provided a number of length- weight observations. If the observations within hauls are strongly correlated, this assumption might be violated and the variances might not be estimated correctly. Lai and Helser (2004) suggest using linear mixed-effects models for grouped samples but the sample sizes are often too small to allow separate models to be fitted to data from individual stations. The data on the sample and individual condition indices suggested that there was some correlation within the hauls but that most (around 73-78%) of the variability was due to other factors than location. So for the purpose of variance estimation, the observations could be assumed to be

effectively independent. However, for the purpose of estimating biomass, the spatial structure does appear to be significant.

The analysis showed that, after *ln(length)*, the factor *stratum* appeared the most influential variable. Although the stratification was based on bottom depth, the term *bottom depth* itself did not reduce the residual sum of squares much when added into the model for either species. The factor age might play a small role, suggesting a possible year-class effect. The interaction terms suggest that not only the intercept, but also the slope might differ between strata and age classes. However, these effects are very small compared to the stratum effect. It is also interesting to note that the term sex was not a significant variable at any point in the stepwise selection procedure for either species. This suggests that there are no significant differences between the sexes in their length-weight relationship at the time of sampling.

The variograms of the sample condition index showed that there was a moderate spatial component in the distribution of the condition indices: stations that were close to each other were more similar than stations with large distance between them. However, most of the variability does not depend on location, but is due to other sources of variation. This might explain why no spatial trends could be discerned in the condition of fish of either species: the condition did not appear to vary with depth, longitude or latitude in a straightforward way. The observed variability is therefore possibly due to a more complicated set of parameters than spatial factors alone. Temperature has been linked to variation in condition of cod (Lloret and Rätz, 2000; Yaragina and Marshall, 2000; Chouinard and Swain, 2002; Rätz and Lloret, 2003). However, these authors refer to variation in temperature on a much larger spatial scale than covered by the present study. Local differences in other environmental conditions, food availability and parasites could also play a role (Lambert and Dutil, 1997; Yaragina and Marshall, 2000), as well as individual differences in energy allocation (Chouinard and Swain, 2002).

The length-weight relationships of some neighbouring stocks varied enough to result in significantly different biomass estimates when applied to the same length-frequency distribution. On the other hand, the length-weight relationships within the stocks generally did not appear to vary significantly. In the current study, the largest difference in biomass estimates was between the length-weight relationships obtained from the West of Ireland and Celtic Sea whiting stocks. The estimated weight of the average length distribution of mature fish per haul over the entire survey was 28.4kg

using the West of Ireland length-weight relationship, while the estimated weight of the same length distribution of whiting was 31.3kg using the Celtic Sea relationship. This is a difference of 10% (bootstrapped 95% confidence limits of 6-15%). Although the magnitude of this error is relatively small to other errors involved in stock assessment (e.g. Gerritsen *et al.,* 2006), it is certainly not insignificant. Considering the relatively low cost of obtaining precise length-weight relationships, it seems worthwhile to ensure that this bias is avoided and length-weight data are only applied for areas from which they were obtained.

5.6. Acknowledgements

Free R-software has been used for this work and the authors would like to thank the R development core team and all contributors to the R project (http://www.R-project.org). Also thanks to the crew and scientific staff that participated in the 2004 Irish Groundfish Survey. We are also grateful for the constructive comments from two anonymous referees.

5.7. Figures
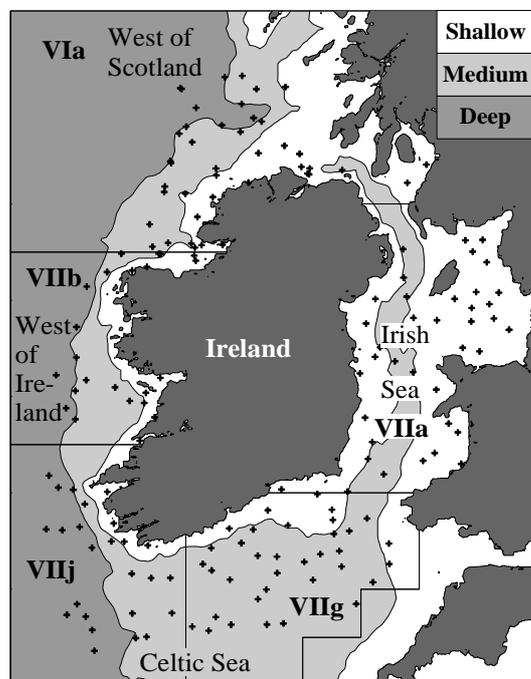
Figure 5.1. The survey area. The crosses represent the sample locations. The survey covers ICES Divisions VIa, VIIa, VIIb, VIIg and VIIj, each of which was stratified into shallow, medium and deep (where present) depth bands.
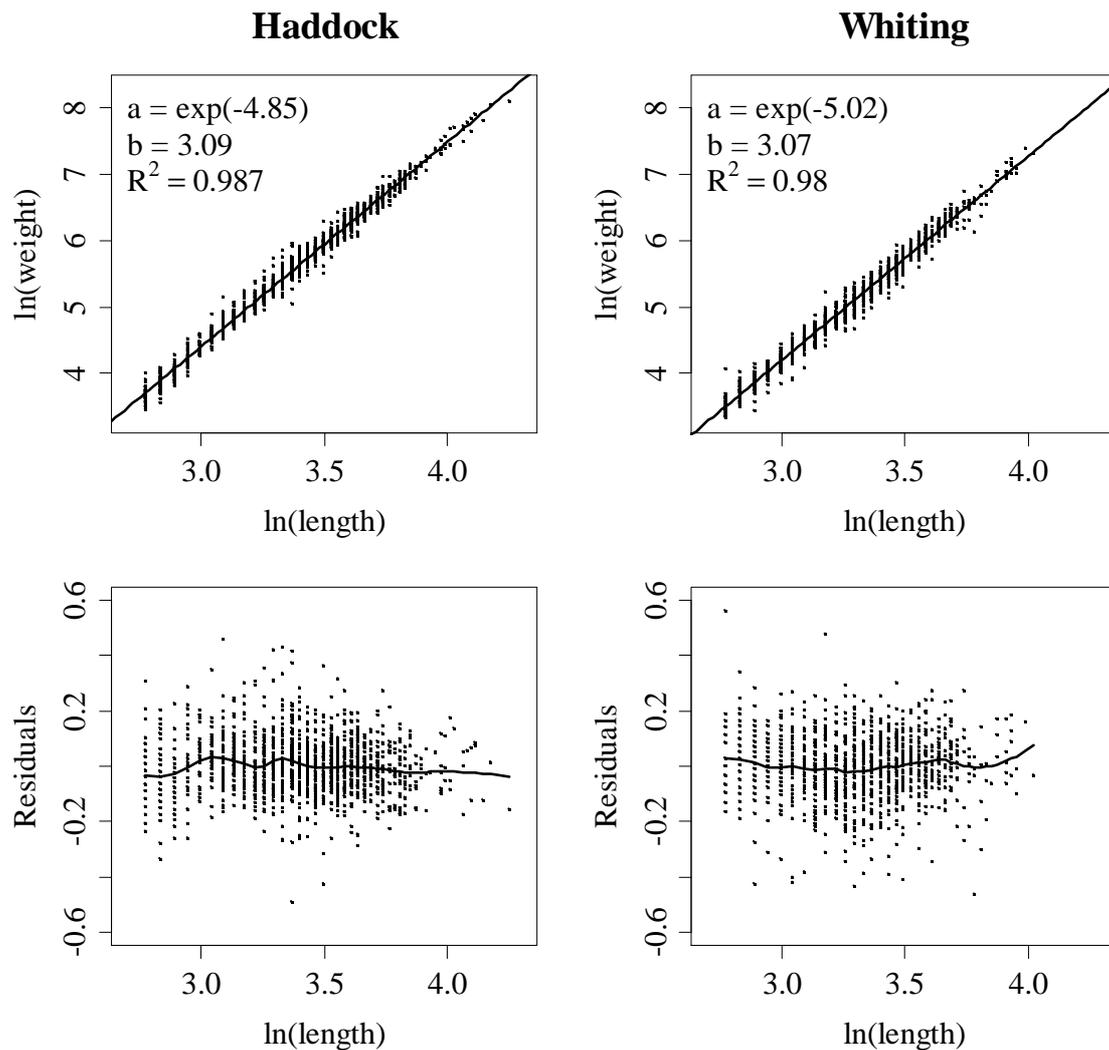
Figure 5.2. Linear models and residuals of the length-weight relationship for haddock and whiting. All data combined. The parameters a *and* b *refer to equation (1), R^2 is the coefficient of determination. A locally weighted running-line smoother (loess) with a span of 25% was fitted through the residuals.*
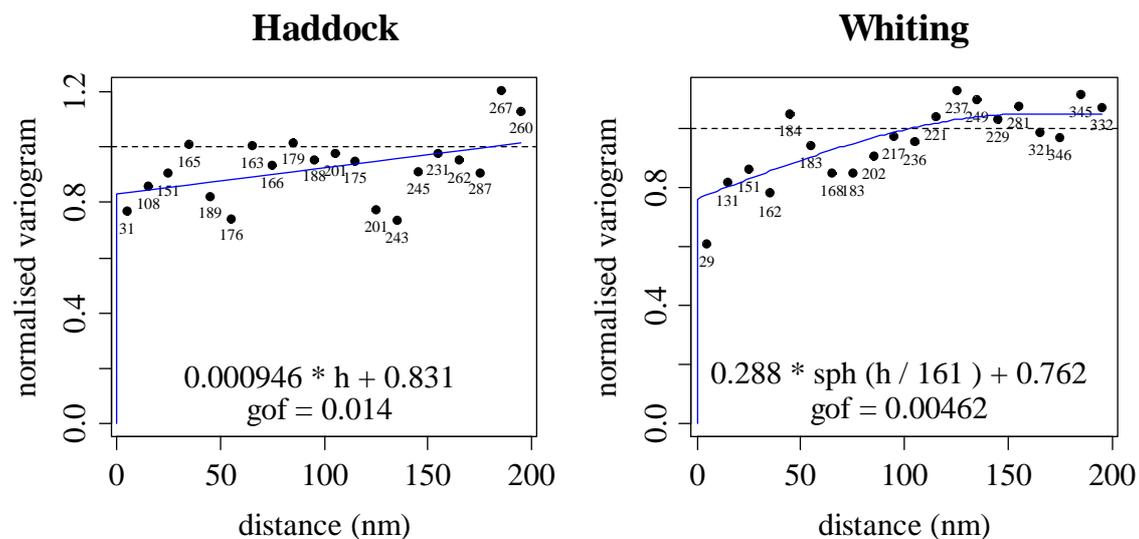
Figure 5.3. Normalised variograms for the sample condition index of haddock and whiting. The dots indicate the values of the experimental variogram and the numbers indicate the number of paired observations in each distance bin. The solid line is the model with the optimum goodness-of-fit (gof). The model parameters and gov are given in the bottom of the plots.
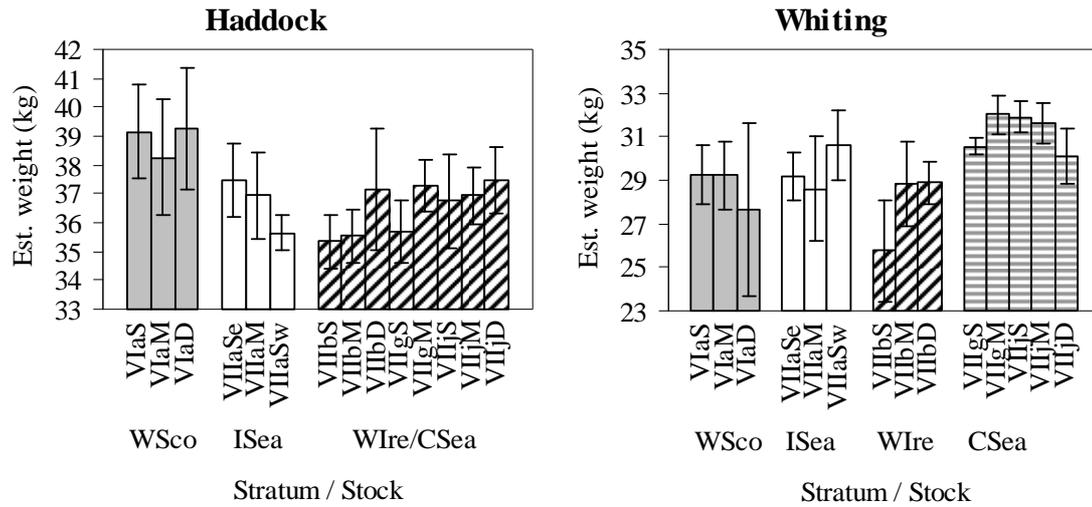
Figure 5.4. Biomass estimates obtained by applying separate length-weight relationships for each stratum to the same length distribution of mature fish. The error bars indicate the 95% confidence limits, obtained by bootstrapping. The strata are indicated by their ICES Division followed by S, M or D for the shallow, medium and deep strata. The shallow stratum in VIIa was further divided into an eastern (e) and western (w) part. The stocks are identified by WSco for West of Scotland; ISea for Irish Sea; WIre for West of Ireland and CSea for the Celtic Sea. Differences within stocks appeared to be limited, but some consistent differences existed between stocks.

Table 5.1. The terms and interactions that were added to the linear model using a stepwise selection procedure.

Haddock

Term	Δdf^a	ΔSS^b	RSS^c	AIC^d	F^e	P^f
null	1 333	973.1				
+*ln(length)*	1	958.6	14.6	-2233	87565	<0.001 ***
+*stratum*	13	1.2	13.4	-2322	9.1516	<0.001 ***
+*age*	3	0.3	13.1	-2347	10.248	<0.001 ***
+*ln(length).age*	3	0.1	12.9	-2355	4.6027	0.003 **
+*ln(length).stratum*	13	0.3	12.6	-2361	2.4407	0.003 **

Whiting

Term	Δdf^a	ΔSS^b	RSS^c	AIC^d	F^e	P^f
null	1 185	806.4				
+*ln(length)*	1	790.6	15.8	-1749	59226	<0.001 ***
+*stratum*	13	2.7	13.1	-1950	18.958	<0.001 ***
+*ln(length).stratum*	13	1.0	12.1	-2016	7.2279	<0.001 ***
+*age*	3	0.1	12.0	-2020	3.0524	0.028 *
+*ln(length).age*	3	0.1	11.9	-2022	2.5532	0.054 .

[a] Change in degrees of freedom when the term is added.

[b] Change is the sum of squares.

[c] Residual sum of squares.

[d] Akaike Information Criterion.

[e] F-statistic.

[f] Result of the F-test. P<0.001 ***; P<0.01 **; P<0.05 *

Chapter 6

Differences in depth distribution between male and female megrim (*Lepidorhombus whiffiagonis*, Walbaum) associated with differences in life-history strategies

6.1. Abstract

The sex ratio in the catches of megrim (*Lepidorhombus whiffiagonis*, Walbaum) on groundfish surveys varied systematically with depth. Female megrim dominated the shallow catches, while males were more common in catches from deeper waters. This pattern was consistent over five surveys that took place off the west coast of Ireland in autumn and spring of the years 2003-5. The observed trend was not likely to be an artefact of sampling, nor was it likely that spawning migrations or differences in preference for temperature, oxygen or salinity could explain the trend. Although size-selective fishing pressure or size-specific habitat preferences could play a role through differences in mean size between the sexes, the patterns in sex ratio remained after size had been taken into account. The most likely explanation, relates to differences in foraging activity between males and females. In order to achieve a large size and high fecundity, females require a higher food intake than males; therefore they have different requirements from their habitat, resulting in differences in the depth distribution between the sexes. An understanding of the mechanisms behind these patterns can improve sampling design and inform fisheries management advice.

Key words: flatfish; megrim; sex ratio; life-history; habitat selection

6.2. Introduction

The sex ratio of fish can be difficult to estimate, as differences in morphology and behaviour between the sexes can influence their catchability (Trippel, 2003). These differences in catchability between males and females are particularly common in flatfish (Beverton, 1964; Rijnsdorp and Witthames, 2005), although they have also been noted for other species (e.g. Armstrong *et al.,* 2004). It is important to understand differences in behaviour and morphology between the sexes if one aims to obtain unbiased samples from a population, for example for the purpose of stock assessment.

Sex ratios close to 1:1 are very common in nature: parents generally invest equally in male and female offspring as neither sex has a selective advantage over the other (Fisher, 1930). If the sex ratio in a population is not even, the average reproductive success of individuals of the minority sex will be higher than that of the majority sex. Natural selection will then favour parents who invest more resources in the minority sex until a balance of equal allocation of resources to male and female offspring is restored. However, if the sexes suffer different mortality rates after parental care has ceased, this will not lead to a selective advantage of parents who invest more in the minority sex because the increased average reproductive success will be offset by the higher mortality. In this way skewed sex ratios can arise.

Skewed sex ratios are not uncommon in survey or commercial catches (e.g. Hannan, 2002; Robson, 2004; King *et al.,* 2006). In some cases, these skewed sex ratios might not reflect the sex ratio in the population, but result from differences in catchability between the sexes (e.g. Armstrong *et al.,* 2004; Rijnsdorp and Witthames, 2005). However, apparent differences in mortality rates are also commonly observed (Beverton, 1964; Fahy and Fannon, 1991; Landa *et al.,* 1996; Sánchez *et al.,* 1998; Landa and Pineiro, 2000) and it is unlikely that such differences are an artefact of sampling. In addition to this, differences in distribution of the sexes have been noted for many species, resulting in spatial tends in their sex ratio (Trippel, 2003). Some of these trends might be the result of differences in migration behaviour during the spawning season (e.g. Warnes and Jones, 1995; Morgan and Trippel, 1996; Stoner *et al.,* 1999). However, trends in sex ratio have also been observed outside the spawning season for a number of flatfish species (Swain and Morin, 1996; Swain, 1997; Poulard

et al., 1999; González and Paz, 2005). In many of these cases, the sex ratio was found to vary with bottom depth. Bottom depth, in turn, is often related to temperature, prey distribution, productivity, dissolved oxygen levels, salinity and sediment type (Swain and Morin, 1997) and is therefore a useful indicator of habitat for groundfish.

Swain (1997) found that differences in temperature preferences between male and female American plaice (*Hippoglossoides platessoides*) determined the differences in their distribution. He suggested that females might prefer warmer waters, as it allows them to have a higher growth rate. Swain and Morin (1996; 1997) also found that females of the same species had a larger stock range than males, possibly reflecting more intensive foraging activity amongst females. The sex ratio of megrim (*Lepidorhombus whiffiagonis*) in the Celtic sea and the Bay of Biscay has also been observed to vary with depth, but no explanation has been proposed as yet (Boon, 1984; Poulard *et al.,* 1999).

Megrim are an valuable by-catch of the Irish mixed fisheries and are caught in large numbers on the groundfish surveys that are undertaken by the Marine Institute in waters around Ireland in the 1st and 4th quarter of each year. The highest densities of megrim are usually found close to the continental break (100-300m; Sánchez *et al.,*1998; Poulard *et al.,* 1999). Spawning of megrim to the west of Scotland takes place between February and April (Anon., 2001). Du Buit (1984) found that megrim to the west of Scotland mainly prey on fish (mostly sprat and small gadoids) and small crustaceans. Similar results were obtained for megrim in the Celtic Sea (Trenkel *et al.,* 2003). Significant differences in growth between sexes appear after the age of two, around the time of first maturation: females grow faster and reach older ages and larger sizes than males (Landa *et al.,* 1996; Landa and Pineiro, 2000).

Spatial trends in the sex ratio of megrim might be observed for a number of reasons. We propose the following hypotheses: (1) Spatial trends in the catches might be an artefact of sampling as males and females could respond differently to changes in the geometry of the gear and the available light at different depths. (2) Males and females might display differences in their migration patterns to and from the spawning grounds. (3) Males and females might respond differently to abiotic variables like temperature, dissolved oxygen or salinity. (4) Large fish might have different depth preferences than small fish or they could be selectively removed from certain areas by the fishery. Due to sexual dimorphism in growth, large fish tend to be females, so the trend in sex ratio with depth could be a consequence, rather than a cause, of the size

distribution of megrim. (5) Females might have different feeding habits to males, resulting in different habitat requirements and therefore different depth distributions. The current paper is aimed at describing the patterns in the sex ratio of megrim in Irish waters and attempting to identify the mechanisms that cause these patterns.

6.3. Materials and methods

Data were collected on three IBTS Irish Groundfish Surveys, carried out on RV "Celtic Explorer" in October and November of 2003, 2004 and 2005 as well as two Biological Sampling Groundfish Surveys, carried out on RV "Celtic Voyager" in February and March of 2004 and 2005. On both survey series, trawling took place during daylight hours using a GOV trawl (ICES, 1999). The gear was towed over the bottom at 3 knots for 30 minutes. The dataset was limited to the area west of Ireland from 51ºN to 54ºN where all surveys had a good spatial coverage (Table 6.1).

The entire catch was generally sexed and measured to the nearest cm below the total length. Occasionally, sub-samples of the catch were sexed and measured if the catch was very large. Megrim were sexed by holding the fish up to the light and checking for the presence of ovaries. Small fish were dissected if their sex was unclear. This method of sexing was verified by dissecting all fish in a sample of 150 megrim after they had been sexed in the normal way (the persons who sexed the fish were not aware that their work would be checked). In the entire sample, only one mistake was found so the error rate in sexing of this sample was less than 1%.

A length-stratified sample of megrim was taken from each catch for the determination of age, weigh, sex and maturity. The maturity stages were determined by visual examination of the gonads. In autumn it can be difficult to distinguish mature-resting gonads from virgin gonads, so maturity data from this period needs to be treated with caution. In addition to this, the assignment of male maturity stages is based on somewhat subjective criteria and immature and spent fish can be confused, even during the spawning season.

Generalised Additive Models (GAM; Hastie and Tibshirani, 1990) with a binomial link function were fitted to the proportion of females, weighted by the numbers of observation in each sample. The following explanatory variables were evaluated: *depth, latitude, longitude, survey* and *season*. As only one of the surveys extended into deep water, the models were fitted to data up to 200m depth. A cubic regression spline with 4 degrees of freedom was used as a smooth function. Separate smooth

functions were fitted for each season or survey if these factors were included in the model. The Un-Biased Risk Estimator (UBRE) score was used to evaluate the trade-off between the model degrees of freedom and the deviance explained by the various models. The mcgv package in R 2.1.1 was used to fit and evaluate the models (R-Development-Core-Team, 2005).

Data on temperature, dissolved oxygen and salinity near the bottom were obtained from a CTD dataset from the International Council for the Exploration of the Sea (ICES) oceanographic database (ICES, 2006a). The dataset was limited to the same geographic area as covered by the survey data. The data available for the period 2003-5 were sparse; therefore all data from 1991 to 2005 were combined. The deepest measurement at any CTD station was taken as a near-bottom measurement.

6.4. Results

The depth distributions of male and female megrim are shown in Figure 6.1. The highest catch rates of females were around 125-150m while the highest catch rates of males were around 200-300m. The female distribution extended further inshore than the male distribution. The males appear to outnumber females from depths of around 200m and more, however any information from depths greater than 250m is based on a single survey only.

The GAMs indicate that inclusion of any of the parameters into the model resulted in a reduction of the UBRE score, compared to the model with an intercept only, indicating that there were significant trends with depth, longitude and latitude, although the latter only reduced the UBRE score marginally (Table 6.2). The model with the lowest UBRE score, despite its large number of degrees of freedom, was the model that contained a separate smooth function for depth for each survey. However, Figure 6.2 shows that the differences between surveys or seasons were slight and no systematic differences between seasons were apparent. In all surveys, the proportion of females in the catches decreased from near unity in the shallow catches to around 50% at 150m. The most recent survey extended into deeper water and it appears that the proportion of females continued to decrease with depth after 200m, although males never fully dominated the catches.

During the autumn surveys, 97% of fish that were identified as mature, were ripening, but not ready to spawn yet: vitellogenesis had started but none of the oocytes were hydrated. During the spring surveys, most mature fish were either in spawning

condition (33%) or recently spent (60%). This suggests that the autumn surveys took place before the spawning season had started and the spring surveys took place during the latter part of the spawning season. Figure 6.3 shows that in the shallow regions, where males are scarce, virtually no ripe females were caught in spring; nearly all mature females in these areas were spent. Spawning seems to occur at depths greater than 100m with the largest proportion of spawning females found around 125m.

Figure 6.4 shows that bottom temperature displayed a strong trend with depth which was reversed between spring and autumn. The same figure shows that in spring the levels of dissolved oxygen near the bottom tended to be highest in shallow areas, but in autumn this trend was not clear. Lastly, the salinity was quite variable at depths up to 200m; in deeper water the salinity was quite stable. The relationship between sex ratio and depth did not show a seasonal pattern, it is therefore unlikely that this relationship is driven by gradients in temperature. The oxygen and salinity gradients were quite variable and it is unlikely that either of these variables can explain the strong and consistent differences in depth distribution between the sexes of megrim.

Females are generally more common in the larger size classes due to differences in growth, so size-specific habitat selection or size-specific fishing mortality might account for the trend in sex ratios. However, within size classes, the proportion of males in the catches still tended to increase with depth (Figure 6.5), so the difference in the depth distribution between the sexes persists after size has been taken into account. This suggests that the trend in sex ratio cannot be fully explained by differences in mean size between the sexes: the sex of the fish is a better predictor of its distribution than its size alone.

6.5. Discussion

The present results allow a number of hypotheses to be rejected. The first hypothesis states that the observed trends in the sex ratio of megrim are an artefact of gear selectivity. It is possible that there are differences in catchability between males and females. However, the sex ratios in the catches were not only biased, they showed a strong trend with depth. For this to occur, males and females would have to respond differently to variations in the gear that occur with depth. The geometry of the net varies with depth and differences in vertical swimming behaviour might exist between males and females, resulting in varying levels of catchability (e.g. Rijnsdorp, 1993). However, although the door- and wingspread are strongly related to depth, the

headline height decreased by less than 5% for each 100m increase in depth in the present study. Therefore, it is unlikely that vertical swimming behaviour is responsible for the observed trends. Another important factor regarding catchability of flatfish is visual avoidance of the net (Gibson, 2005). As the available light will vary strongly with depth. However, there is no evidence to suggest that male and female megrim have vastly different visual capabilities, so the amount of light that penetrates at depth is unlikely to explain the observed trends in sex ratio. Lastly, although the same gear type was used in all studies that identified trends in the sex ratio of megrim (The present study; Boon, 1984; Warnes and Jones, 1995; Poulard *et al.,* 1999) similar trends were observed using a number of different vessels and gear modifications. So, although it cannot be ruled out entirely, it is unlikely that the observed trend in sex ratio is an artefact of gear selectivity.

The second hypothesis, stating that the trends in sex ratio might be explained by differences in the spawning migrations of males and females, can be refuted by the fact that the same pattern was found before (autumn surveys) and during (spring surveys) the spawning season. However, it was surprising to find mature females at depths where males were nearly absent during the spawning season. Figure 6.3 indicates that these females were recently spent, suggesting that they migrated there after spawning. A broader temporal coverage during the entire spawning season would be necessary to resolve the details of migration during the spawning season.

The third hypothesis relates to abiotic variables. In order to explain the consistent differences in the distribution of males and females, the abiotic variables would have to display a seasonally consistent gradient with depth. None of the three variables examined (temperature, oxygen and salinity) displayed strong, consistent trends, so it is unlikely that the trend in sex ratio is driven by any of these variables.

The fourth hypothesis states that the trend in sex ratio is a consequence of differences in growth; size selective fishing could skew the sex ratio in certain areas or, alternatively, the habitat preference of megrim might change with size. The mean length of megrim in the survey catches does tend to decrease with depth (Boon, 1984). However, Figure 6.5 shows that the trend in sex ratio remains apparent after size has been taken into account: both small and large females dominated the catches from shallow areas, while males of all size classes were relatively more common in the deeper catches. So the size difference between the sexes alone cannot explain the differences in distribution, although it might have a confounding effect.

The last hypothesis points towards differences in feeding activities between the sexes as an explanation for the differences in habitat selection. This could not be tested with the available data, but evidence from the literature indicates that food intake of many female flatfish is higher than that of males (Lozán, 1992; Stoner *et al.,* 1999; Villarroel *et al.,* 2001). In many flatfish species, females grow not only larger, but also faster than males, while producing a larger amount of gonadal material (Pauly, 1994; Rijnsdorp and Witthames, 2005). Therefore, one might assume that it is common for female flatfish to have higher energy requirements than males. Differences in feeding behaviour are likely to result in different distributions of male and female megrim, either due to prey availability or higher levels of competition forcing females towards the margins of their distribution.

The cost associated with high growth rates for females might include increased exposure to predation (Milinski and Heller, 1978; Roff, 1982; Abrahams and Dill, 1989; Holtby and Healey, 1990). Flatfish tend to be well camouflaged when they are not actively swimming, therefore it might be particularly advantageous for flatfish to reduce their foraging activity. However, as fecundity is directly related to size (Rijnsdorp and Witthames, 2005), females might have to accept the costs of a high growth rate in order to maximise their reproductive output. The reproductive output of male flatfish does not seem to increase once a critical size has been reached (Rijnsdorp and Witthames, 2005), therefore, males might take advantage of reduced predation risks by reducing their food intake (Rijnsdorp and Witthames, 2005).

In summary, it appears that the observed trend in sex ratio with depth is not likely to be an artefact of sampling. The lack of seasonal differences suggests that the trend in sex ratio is not related to spawning activity either. No abiotic factors were identified as likely causes for the trend in sex ratio. Although localised size-selective fishing pressure or size-specific habitat preferences might play a role through differences in mean size between the sexes, the patterns in sex ratio remain after size has been taken into account. The most likely explanation for these patterns is related to the higher energetic requirements for females, resulting in differences in habitat selection. Although the present data provide no direct evidence for this explanation, it is supported by a large amount of evidence from the literature.

Although the trend in the sex ratio of megrim was particularly pronounced, it is likely that males and females of other flatfish species display similar differences in distribution. Sexual dimorphism is common in flatfish, particularly species attaining a

large size (Rijnsdorp and Witthames, 2005). As differences in growth rate are likely to impact on the distribution of the sexes, one can expect to find spatial trends in the sex ratio of these species.

An understanding of the differences in habitat selection and life-history strategies of males and females can improve sampling design and inform fisheries management advice. Many flatfish stocks are assessed without distinguishing between the sexes, despite well-known differences in weight-at-age (e.g. ICES, 2005b). Any bias in sampling due to differences in distribution between the sexes can therefore have important consequences on the stock assessment. An understanding of the differences between males and females is also required for the use of female-only spawning stock biomass estimates, which might provide an improved index of reproductive potential (Marshall *et al.,* 2003; Marshall *et al.,* 2006). Finally, an understanding of life-history strategies might give an insight into the effect of changes in the fishery or in the ecosystem.

6.6. Acknowledgements

6.7. Tables and figures

Table 6.1. Summary statistics for the five surveys used.

Survey code	Year	Period	Number of Stations	Depth Range	Spatial coverage Latitude	Longitude
IGFS03	2003	Oct/Nov	150	37 - 289	50-56ºN	12-6ºW
BSS04	2004	Feb/Mar	89	36 - 177	51-56ºN	11-6ºW
IGFS04	2004	Oct/Nov	160	35 - 253	50-56ºN	12-6ºW
BSS05	2005	Feb/Mar	36	60 - 185	51-54ºN	11-9ºW
IGFS05	2005	Oct/Nov	140	66 - 549	50-56ºN	13-6ºW

Table 6.2. Generalized Additive Models for the sex ratio (R). The smooth function is indicated by s( ); if the factors Season or Survey were included in the model, separate smooth functions were fitted for each of the seasons or surveys, each with 4 degrees of freedom (df).

Model	Null df.	Model df.	Null deviance	Deviance explained	UBRE score
R~s(Depth)	124	4	681.6	37.9%	2.748
R~s(Depth by Season)	124	8	681.6	43.0%	2.297
R~s(Depth by Survey)	124	20	681.6	47.7%	2.140
R~s(Longitude)	124	4	681.6	18.0%	2.945
R~s(Longitude by Season)	124	8	681.6	22.7%	2.851
R~s(Longitude by Survey)	124	20	681.6	30.7%	2.572
R~s(Latitude)	124	4	681.6	2.1%	3.557
R~s(Latitude by Season)	124	8	681.6	14.2%	3.118
R~s(Latitude by Survey)	124	20	681.6	31.9%	2.506
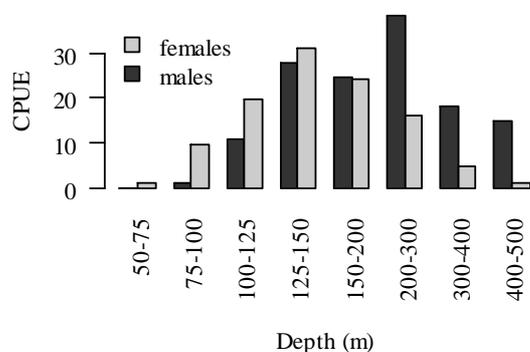
Figure 6.1. The average catch per unit effort (CPUE; numbers per half hour trawled; all surveys combined). Although the sexes generally overlap in their depth distribution, the females are more common in the shallow hauls while males dominate the deeper hauls.
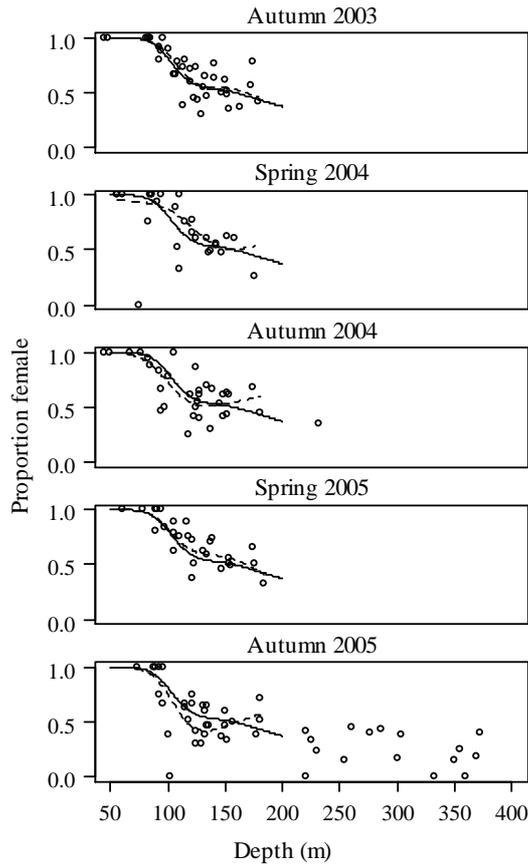
Figure 6.2. The observed sex ratios at depth with cubic regression splines with 4 degrees of freedom. Smooth curves for individual surveys are shown as dashed lines and the solid line represents the common curve for all surveys. Similar trends in the sex ratio were observed in all years and seasons.
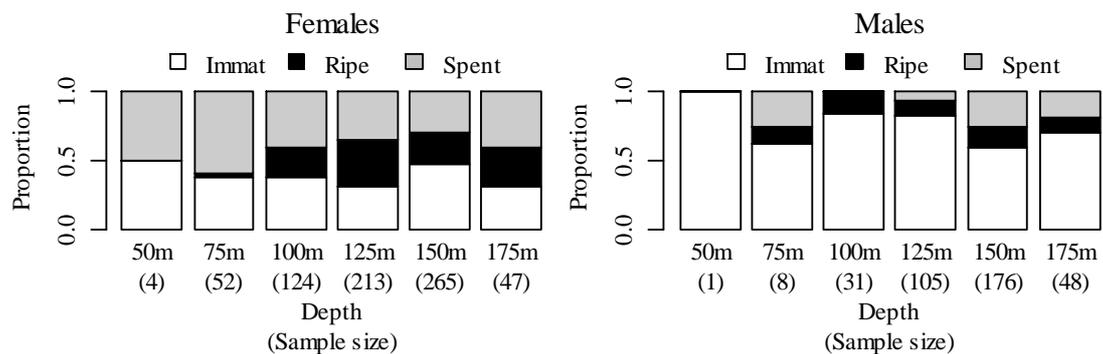
Figure 6.3. The proportions of immature, ripe (or ripening) and spent fish in the samples taken in spring. Sample sizes by depth class are shown between brackets.
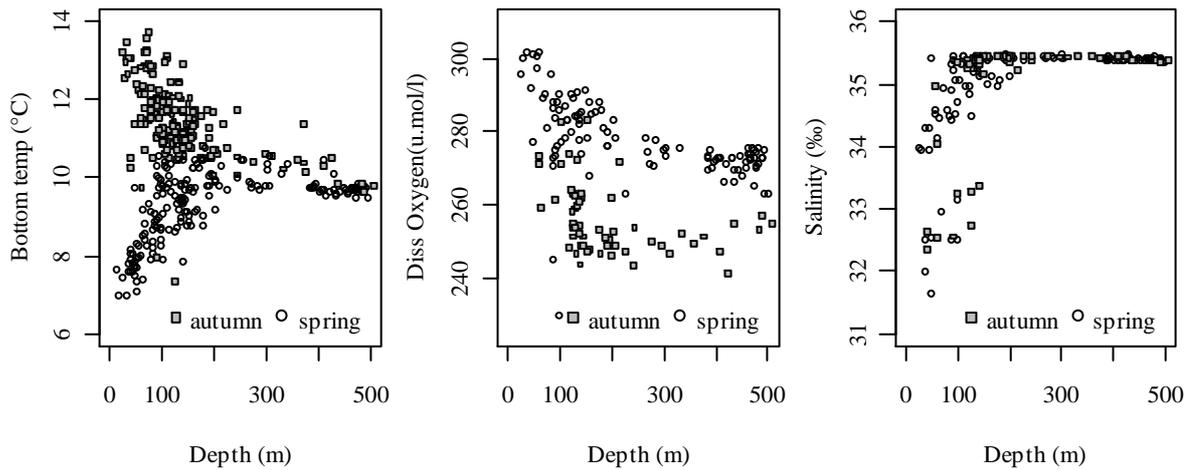
Figure 6.4. *Temperature, dissolved oxygen and salinity near the bottom in the study area during autumn and spring. The trend in bottom temperature was reversed between the two seasons. Oxygen levels showed a trend with depth in spring, but no strong trend in autumn. Salinity levels were variable up to bottom depths of around 200m after which they stabilised.*
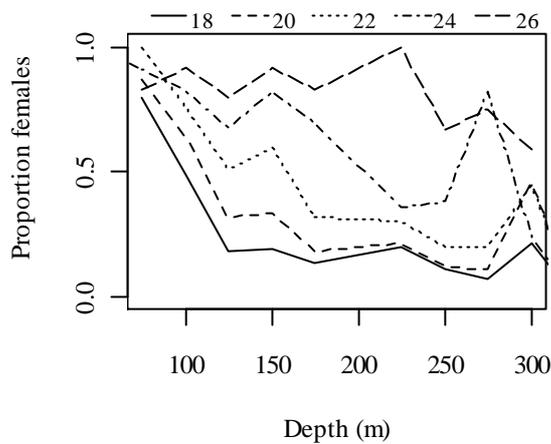
Figure 6.5. *The sex ratio by size class (2cm intervals). The proportion of males increase with depth for all size classes from 18 to 26cm.*

Chapter 7

Spatial patterns in maturity: Cod (*Gadus morhua* L.) in the Irish Sea

This chapter is an extract of the following original publication:

> Armstrong MJ, Gerritsen HD, Allen M, McCurdy WJ, Peel JAD (2004) Variability in maturity and growth in a heavily exploited stock: cod (Gadus morhua L.) in the Irish Sea. ICES J. Mar. Sci. 61: 98-112

See Appendix D, page 116 for details on authorship

7.1. Introduction

Along with the physiological changes that take place when fish mature, their migration patterns also tend to change with maturation. Mature fish might migrate to spawning grounds, while immature fish of the same age class do not, resulting in regional variation in the proportion mature-at-age. If samples are taken from commercial sources, which tend to target spawning aggregations, maturity estimates might be biased. An example of cod in the Irish Sea will be used illustrate and quantify this bias.

7.2. Methods

A time-series of data on length, age, maturity and sex of cod was available from eleven groundfish surveys of the northern Irish Sea in March of the years 1992-2002 using the RV "Lough Foyle" operated by the Department of Agriculture and Rural Development in Northern Ireland. The gear used was a rock-hopper otter trawl with an average headline height of 3m and average door spread of 40m. The surveys comprised of 45 fixed-position stations with a standard tow distance of 3nm at 3 knots of speed over the ground. In most cases, the entire catch of cod was aged and their sex and maturity stage were determined. On relatively few occasions, catches were so large that the most abundant size classes had to be subsampled. Subsampling took place on a length-stratified basis: all fish were measured and up to 25 fish were retained for biological analysis (age, sex and maturity). The subsample was evenly spread out over the size range available. Subsampling affected mainly the smallest size classes: overall, one third of one-year-olds were retained for biological analysis, 85% of two year olds and virtually all older fish were sampled. A total of 3849 cod

were sampled for biological analysis. Three geographic regions were specified based on prior known aspects of cod biology and distribution. Regions 1 and 3 (Figure 7.1) represent areas with high egg production during the spawning season, whilst region 2 is characterised by a low abundance of cod eggs (Nichols *et al.,* 1993; Fox *et al.,* 1997).

Age classes were assigned according to the number of translucent rings in the sagittal otoliths, which were sectioned through the nucleus, embedded in clear resin and viewed by transmitted light under a binocular microscope. Maturity stages were assigned by visual examination of the gonads (Bowers, 1954). Fish were considered mature when they had started producing clearly detectable amounts of milt or yolked oocytes or if they appeared spent. This definition of maturity was considered appropriate for surveys taking place close to the peak of spawning and was supported by histological studies of ovary development in Irish Sea cod in 1995 and 2000 (Armstrong *et al.,* 2001; Anon, 2002)

7.3. Results and discussion

Virtually all cod first matured either at the age of two or three: over the entire time series, less than 0.5% of the one-year-olds in the catches were mature, more than 99.5% of the three-year-olds were mature and a variable proportion of the two-year-olds were mature. Figure 7.1 shows the mean proportions of mature two-year-old cod for each station over the period 1992-2002. For both sexes, the highest proportions of mature two-year-olds were found to northwest and the east of the sampling area, which largely corresponds to regions 1 and 3.

For both sexes, the proportion of mature two-year-olds was higher in the spawning areas (regions 1 and 3) than in region 2 for nearly all length classes (Figure 7.2). Within this age class, length only appeared to play a minor role: a small two-year-old fish was almost as likely to be mature as a large two-year-old fish (Figure 7.2). Similar regional patterns were found over the years (Figure 7.3): in most years the highest proportions of mature two-year-olds were found in regions 1 and 3.

As maturity within the two-year-olds was virtually independent of length, the bias caused by length-stratified subsampling would have been minimal (Morgan and Hoenig, 1997) and the proportions of mature two-year-olds in the samples will accurately reflect those in the catches without the need to raise the samples to the total length distribution.

If the catches are considered to be representative for each area, an index of abundance can be obtained by multiplying the surface area of each region by the mean catch rate per standardised tow. This can be used to obtain estimates of maturity-at-age over the survey area, weighted by the abundance in each area (Table 7.1). If all areas are included, the estimated proportion of mature two-year-olds over all years is 63% for males and 30% for females. However, if sampling had only taken place on the spawning grounds, as might happen when samples are taken from commercial sources, these estimates would have been much higher: 95% for males and 60% for females. This is an extra 30% mature at the age of two for both sexes. As the as two-year-olds contributed around 40% of the total weight of cod in the survey catches, the spawning stock biomass (SSB) estimate might therefore be overestimated by 12% (average over all years). On a yearly basis, this potential bias varied from 6% to 16% of the SSB, is this would be estimated directly from the survey.

Cod in the Irish Sea appear to have fairly well-defined spawning areas e.g. in contrast to whiting in the same area (Gerritsen *et al.*, 2003) and a large proportion of the population is at an age class where both mature an immature fish are found. The example given above is therefore possibly a worst-case scenario. However any species that undertakes a spawning migration and has one or more age classes with both immature and mature fish, would be sensitive to a bias in the proportions mature-at-age if sampling is not carried out representatively or weighted appropriately. As commercial fisheries tend to target spawning aggregations, it might not be appropriate to obtain maturity samples from commercial sources.

7.4. Figures and Table
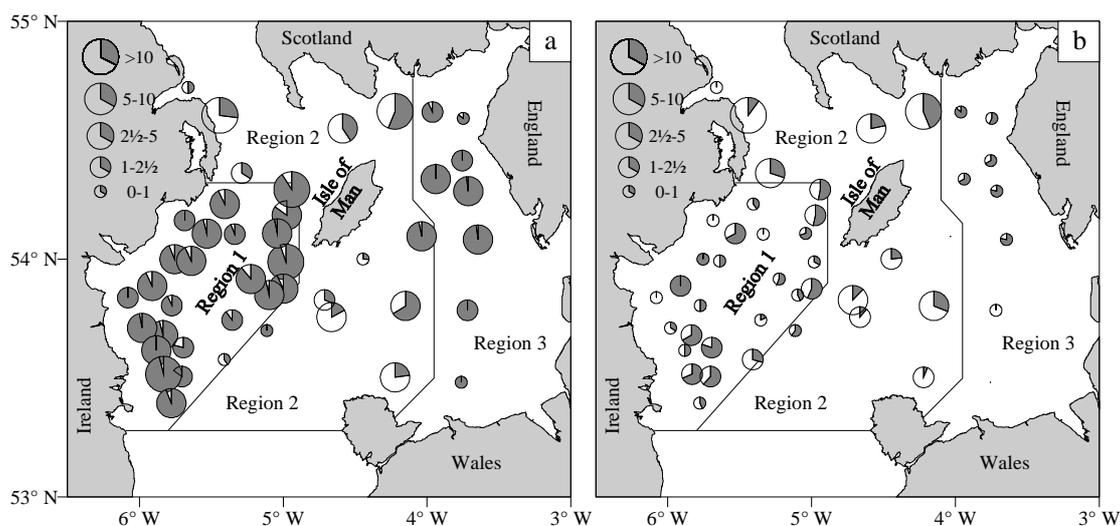
Figure 7.1. Average proportions mature of male (a) and female (b) two-year-old cod over the period 1992-2002, represented by the grey area of the pie charts. The size of the circles represents the average sample numbers per haul.
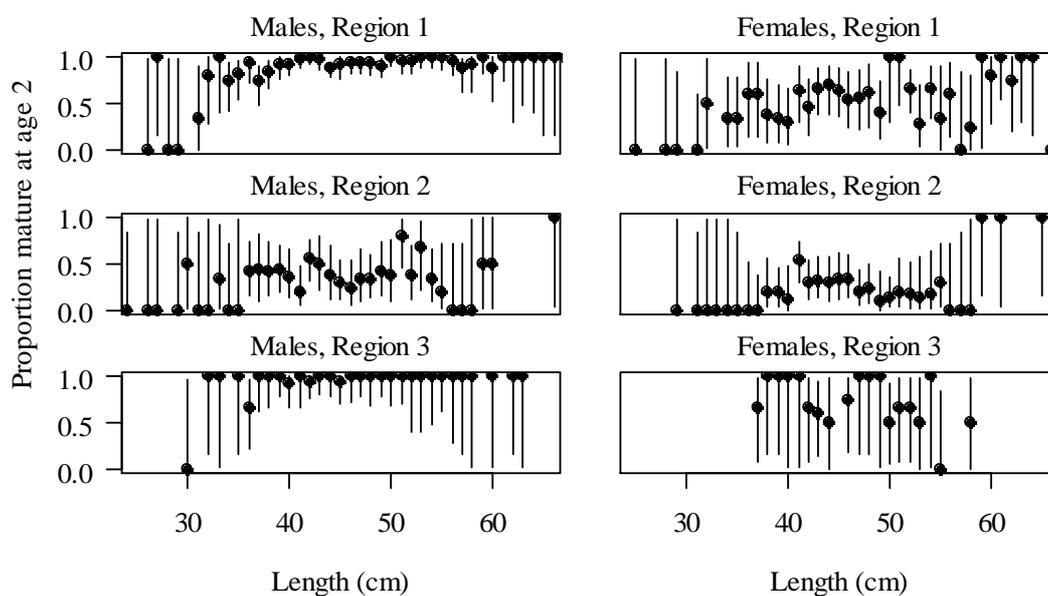
Figure 7.2. Proportions of mature cod at age two in the biological samples for the three geographical regions. The error bars represent the 95% confidence regions for each length class. Data from 1992-2002 combined.
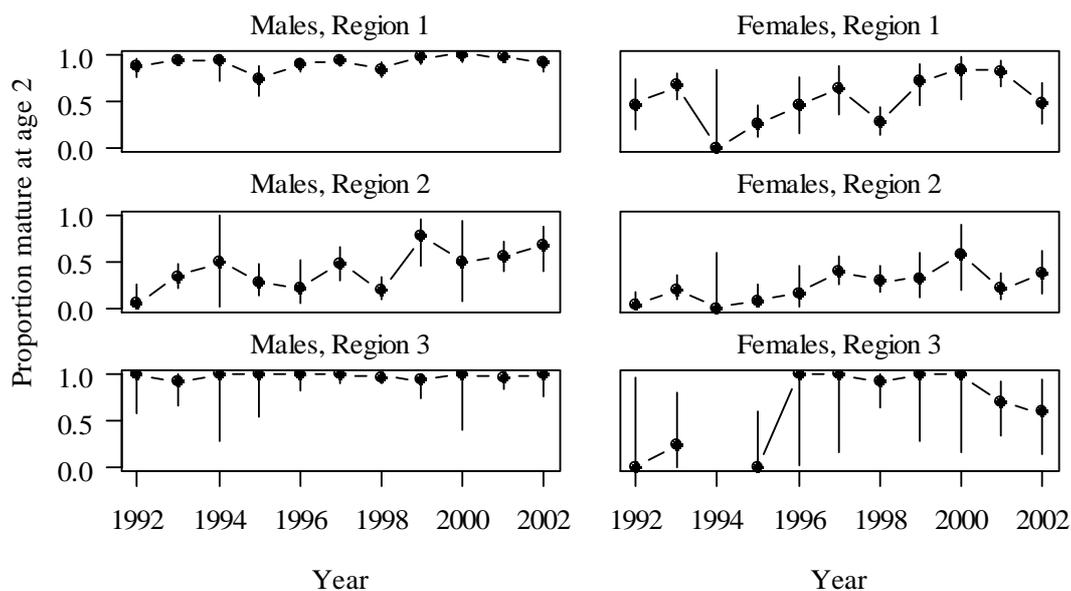
Figure 7.3. Proportions of mature cod at age two in the biological samples for the three geographical regions. The error bars represent the 95% confidence regions for each length class. Data from 1992-2002 combined. No samples of female two-year-olds were available for region 3 in 1994.

Table 7.1 Average proportions of mature two-year-old cod by region and estimates weighted by the surface area and mean catch rate of two-year-olds in each region. Data from 1992-2002 combined.

Region	Surface area (nm2)	Catch rate (nos per ½ hr)		Prop mature at age 2	
		Males	Females	Males	Females
1	2770	3.3	1.0	0.93	0.55
2	5400	3.8	3.6	0.37	0.24
3	2950	2.5	0.5	0.98	0.69
Weighted, regions 1, 2 & 3				0.63	0.30
Weighted, regions 1 & 3 (spawning areas only)				0.95	0.60

Comparison of the magnitude of errors involved in parameter estimates from trawl surveys

An earlier version of the paper has been presented to the ICES Annual Science Conference:

> Gerritsen H, McGrath D (2006) An overview of sources of bias and uncertainty in trawl surveys: a biased overview of errors. ICES Annual Science Conference, Maastricht, 19-23 September 2006. ICES CM 2006/1:19

8.1. Introduction

In Chapter 1, an overview has been given of different sources of errors that influence the parameter estimates produced from trawl surveys (survey outputs). These sources of error were further explored through case studies in the following chapters. The precision of length samples was described in Chapter 2 and the errors in assigning maturity stages were estimated in Chapter 3. The next four chapters concerned spatial variability in the following parameters: the age-length structure (Chapter 4); length-weight relationship (Chapter 5); sex ratio (Chapter 6); and proportion of mature fish (Chapter 7). These case studies provide insight into the nature of various sources of error. However, it is not immediately apparent how these errors will interact in the errors of the survey outputs.

The overall error is the result of a combination of sampling error and measurement error, each of which are determined by a number of factors. Sampling errors are determined by the variability of the parameter and by the sample size (Zar, 1999). Therefore, the number of observations at each sampling station, the number of sampling stations and the variability within and between stations will all influence the sampling error. The interaction of various sources of sampling error is further complicated by the two-stage sampling scheme that is often used (Cochran, 1977). Under this scheme, a sample of the catch is taken to estimate the length distribution and catch numbers. A smaller length-stratified biological sample is taken for the determination of age, weight, sex and maturity of individual fish. The total sampling error is therefore influenced by:

    i)   The variability in the length samples within stations and the sample size per station

ii) The variability in the length samples between stations and the number of stations sampled

iii) The variability in the biological samples within stations and the sample size per station

iv) The variability in the biological samples between stations and the number of stations sampled

The precision associated with these four factors can be assessed separately by obtaining bootstrap estimates with different bootstrapping units. By using observations on individual fish as bootstrapping units, the variability within stations can be assessed. The variability between stations can be assessed by using the sampling stations as bootstrapping units: for each bootstrap iteration, stations numbers are drawn with replacement and the data from the selected stations are used to produce an estimate of the survey outputs. The distribution of these estimates is then used as a proxy for the error distribution in the survey outputs (Efron and Tibshirani, 1993).

The main measurement errors that might occur are:

v) Ageing error

vi) Errors in assigning maturity stages

Measurement errors cannot be assessed using bootstrapping techniques, but they can be simulated using a Monte Carlo approach (Efron and Tibshirani, 1993). A measurement error distribution is estimated and random samples are drawn from this distribution. These random samples are then used to obtain a distribution of parameter estimates of survey outputs from which an uncertainty estimate can be obtained. Bootstrapping and Monte Carlo techniques can be combined to estimate the overall error in the various survey outputs. Additionally, the contribution of each source of error to the survey outputs are investigated by allowing only one source of error to vary and keeping all other sources of error constant.

The influence of the various sources of error will vary between stocks and surveys, so they have to be assessed in each separate case. A single example will be given here for one stock, west of Ireland haddock.

8.2. Methods

Data were collected on the Biological Sampling Survey in March 2005, which was carried out to the west of Ireland. An estimated total of nearly 17.000 haddock were

caught at 38 sampling stations, 4,631 haddock were measured and the age, weight, sex and maturity stages were determined for 573 fish.

Error estimates were obtained for the following survey outputs:

Catch numbers-at-age – Time series of catch numbers per unit effort from surveys are an integral part of the stock assessment models used for this stock (ICES, 2006b). The initial model parameters are adjusted (tuned) until the model outputs become consistent with the survey time series.

Mean weight-at-age – Used in stock assessment to transform numbers-at-age to a biomass estimate (Haddon, 2001).

Mean length-at-age – Used in stock assessment to estimate the mean weight-at-age from a length-weight relationship (Haddon, 2001).

Maturity-at-age – Used to estimate the proportion of the stock that is mature, although maturity data are usually not updated annually (ICES, 2006b).

Sex ratio – Rarely used in stock assessment, even when differences in maturation, growth and mortality between the sexes are known to exist (e.g. ICES, 2005b).

Catch weight of mature females – Analogous to female-only SSB, for stocks where the sex ratio of mature fish is not constant over time, this is likely to be a more sensitive index of reproductive potential than SSB alone (Marshall *et al.,* 2006).

In the present analysis, catch numbers-at-age and mean length-at-age were estimated by applying the total catch length distribution to an Age-Length-Key (ALK) for the whole survey area. Mean weight-at-age was estimated by applying a length-weight relationship to the estimated mean lengths-at-age. Maturity-at-age was estimated using a Maturity-Age-Length-Key (MALK). Maturity was only estimated for one-year-old fish, as virtually all (99.5%) older fish were mature. The sex ratio (by numbers) was estimated by applying the proportion of females at each length class to the total catch length distribution. The catch weight of mature females was estimated by multiplying the weight-at-age by the sex-ratio-at-age and by the proportion mature-at-age. For the purposes of this analysis, no spatial stratification was used as the number of sampling stations in most strata was less than 10, which is too few to perform bootstrapping routines (Efron and Tibshirani, 1993). If, due to re-sampling, gaps occurred in the ALK or MALK, the predicted proportions for the length class(es) concerned were estimated by applying a multinomial logistic model following the methods described in Chapter 4.

8. COMPARISON OF THE MAGNITUDE OF ERRORS

The Marine Institute recently took part in a number of otolith exchanges (Egan *et al.,* 2004; Duarte *et al.,* 2005; Easey *et al.,* 2005; Worsøe Clausen *et al.,* 2005; Woods, unpubl.). Data from these exchanges were used to quantify the measurement error of age readings. This was done by estimating the standard deviations of the distributions of the assigned ages around the modal ages assuming a discrete normal distribution, following Clark (2004). Figure 8.1 indicates that the spread in the error distributions tends to increase linearly with age. The position of the regression line could be regarded as a measure of how consistently a species or stock can be aged. No data were available for haddock, but this species is regarded as relatively easy to age. However, in order to obtain a conservative estimate, the ageing error of haddock was taken to be similar to megrim, a species which is considered to be moderately difficult to age (Figure 8.1). The age reading error of haddock was assumed to follow a discrete normal distribution with a standard deviation (*SD*) of:

(8.1) $SD = 0.5 + 0.05A$

where *A* is the canonical age.

The measurement error distribution of the assignment of maturity stages was also estimated in a conservative way. Results from Chapter 3 suggest that disagreement in the assignment of maturity stages is largely confined to consecutive stages. The measurement error in the distinction between mature and immature fish might therefore be simulated by assuming the worst-case scenario where maturity stage 2 (developing but immature) and maturity stage 3 (early maturing) cannot be distinguished and are assigned at random with equal likelihood.

Other measurement errors occur in the effort estimates, in length measurements, in weight measurements and in sex determination. These errors were not considered in the present analysis as they are likely to contribute very little to the overall error. Pilot experiments on length, weight and sex determination errors suggest that none of these are likely to have CVs of more than 1% (Gerritsen, unpublished).

The errors around the survey outputs were estimated using bootstrapping procedures, Monte Carlo simulation, or a combination of both methods. Estimates of catch numbers-at-age, weight-at-age, length-at-age, sex ratio and the catch weight of mature females were obtained for each bootstrap and Monte Carlo iteration. The distribution of these re-sampled parameter estimates was then used to obtain error estimates for the survey outputs. The uncertainty in the estimates of catch numbers-at-age, mean

weight-at-age and mean length-at-age were expressed as a Mean Weighted Coefficient of Variation (MWCV; Chapter 2). The standard errors at each age class were estimated from the standard deviations of the re-sampled parameter estimates and expressed as CVs by dividing by the means. The mean CV was then estimated by weighting the CV at each age class by the catch numbers in each age class. The CV of the proportions mature and the sex ratio were directly estimated from the standard deviation of the re-sampled parameter estimates. Finally, the CV of the catch weight of mature females was estimated from the standard deviation of the re-sampled parameter estimates, divided by the mean.

The relative contribution of various sources of errors to the overall error in the survey outputs was quantified in the following ways:

Resampling individual length observations – For this approach, individual length measurements were taken as bootstrapping units, simulating the variability that would occur in the survey outputs if all length samples were random samples from the population and all other sources of error were non-existent.

Resampling length samples by station – This approach uses the sampling stations as bootstrapping units, re-sampling the length distributions. This acknowledges the variability in the abundance and length distributions between stations, but ignores all other sources of error.

Resampling individual biological observations – The individual biological samples (age, weight, sex and maturity of individual fish) were used as bootstrapping units. This approach estimates the errors in the survey outputs that would occur if individual biological samples were random samples from the population and ignores all other sources of error.

Resampling biological samples by station – The sampling stations were used as bootstrapping units and the sets of biological samples taken at each station were re-sampled. This estimates the sampling error in the biological samples between stations and ignores all other sources of error.

Simulating ageing error – Random samples were drawn from the simulated ageing error distributions and survey outputs estimated for each iteration. This estimates the error in the survey outputs that are due ageing errors and ignores all other sources of error.

Simulate maturity staging error – Errors in the assignment of maturity stages were simulated and random samples were drawn from this distribution. This

estimates the error in the survey outputs that are due assigning maturity stages and ignores all other sources of error. This source of error is only relevant to estimates of maturity-at-age and estimates of the catch weight of mature females.

Overall error – By simultaneously simulating and re-sampling all the relevant sources of error mentioned above, the overall error in the survey outputs was estimated.

Figure 8.2 summarises the procedure for estimating the errors for the various survey outputs.

8.3. Results

Catch numbers-at-age – The main objective of most trawl surveys is to provide an estimate of numbers-at-age per unit effort. Figure 8.3 indicates that in the present case, the catch numbers-at-age are estimated with a CV of 22%. The figure also indicates that if the variability in the length frequency distributions between stations would be the only source of error, the CV of the catch numbers at age would still be 20%. This suggests that the only way to significantly improve the precision is by reducing this variability between stations. This might be achieved by increasing the number of sampling stations or by spatial stratification of sampling. The variability due to the biological samples was relatively small (CV<10%)

Mean weight-at-age – The total error in this survey output is quite small (Figure 8.3). It appears that no single source of error dominates the total error. This indicates that all of these errors would need to be reduced in order to further improve the precision.

Mean length-at-age – This survey output tends to be estimated with very high precision (Figure 8.3), it is quite robust to any variability between or within stations and to ageing error.

Maturity-at-age-one – The proportion mature was also estimated with high precision (Figure 8.3). Errors in the assignment of maturity stages (staging error) did not seem to contribute significantly to the overall error. The largest source of error appears to be due to the sample size and variability of the biological parameters.

Sex ratio – The proportion of females was estimated with high precision (Figure 8.3). The main source of error appears to result from the sample size and variability of the biological parameters.

Catch weight of mature females – This estimate is the result of combining estimates of catch rate-at-age, length-at-age, length-weight relationship, maturity-at-age and the sex ratio. It is therefore not surprising that many sources of error contribute to the total error. The largest single source of error was the sampling error in biological samples between-stations. The error in assigning maturity stages appeared to be inconsequential.

8.4. Discussion

One of the conclusions that can be drawn from Figure 8.3, is that the total error estimate is often close to the largest single source of error. The reason for this is that errors propagate quadratically (Taylor, 1997), so the largest errors become relatively more important in the total error estimate. This means that the only way to improve the precision of the estimates is to reduce the variability in the main source(s) of error. The smaller errors have a rapidly diminishing effect on the overall error.

For the catch numbers-at-age estimate, the variability due to the length distributions between stations was by far the most important source of error, suggesting that the number of sampling stations needs to be increased to improve the precision. This is in contrast to the findings by Kimura (1997), stating that the precision of the estimated age distribution improves rapidly with increasing numbers in the aged sample. Figure 8.3 suggests that increasing the biological sample numbers would have very little effect on the total error. This might seem surprising, but one has to remember that it not just the shape of the length distribution that varies between stations, but also the absolute numbers. The catch rate can vary quite dramatically between stations; therefore a small number of stations with very high catch numbers can have a very large influence on the catch numbers-at-age estimate for an entire survey. Most of the work done on finding optimum sample sizes, ignores this between-station variability (Flatman, 1990; Gutreuter and Krzoska, 1994; Kimura, 1997; Oeberst, 2000). In the present case, ignoring this source of error would have resulted in an overestimate of the relative importance of the sample size of biological samples to the total error in the catch-numbers-at-age estimate.

Another conclusion that can be drawn from Figure 8.3 is that the variability in the length distributions within stations only contributed a small error for all parameters. This finding is in line with conclusions from Pennington *et al.* (2002) and indicates that the number of haddock measured at each station may be reduced without

significantly affecting the overall error in most parameter estimates. However, reducing the sample sizes might be undesirable, as precise length distributions at each sampling location can be useful for fine-scale spatial analysis (Chapter 2).

Ageing errors seemed to have a small influence on the total precision of most parameters, despite the fact that the simulated ageing errors are likely to be larger than the actual ageing errors. However, ageing errors do tend to result in biased estimates. The catch numbers in the most common age class tended to be underestimated. The mean length-at-age and mean weight-at-age tended to be under-estimated for age classes that were older than the most common age class, while they were over-estimated for younger age classes (data not shown). If the random ageing error is known, this bias can be corrected for by estimating the error matrix (Richards *et al.,* 1992).

Errors in assigning maturity stages appeared to have very little effect on the estimated proportion mature-at-age in the present case. This is probably due to the fact that only 10% of all the maturity stages were assigned as stage 2 or stage 3, all the other samples were either undeveloped or in an advanced state of development, leaving little scope for misclassification. This highlights the importance of the timing of sampling for maturity: in autumn of the same year, 83% of haddock sampled were assigned stages 2 or 3, which would have led to a very imprecise estimate of the proportion of mature fish.

The catch weight of mature females was estimated with a similar precision to the catch numbers-at-age estimate, despite the fact that a larger number of parameters need to be estimated to produce this survey output. The catch rate of mature females is likely to be a more sensitive index of reproductive potential than SSB as it does not require the assumption that the proportion of mature females in a population is constant over time (Marshall *et al.,* 2006). Including more parameters in an index of reproductive potential, for example fecundity estimates, would further reduce the precision of this estimate. However, if these additional parameters are estimated with reasonably good precision, the reduction on precision on the overall estimate might be negligible.

If one aims to improve the precision of any of the parameter estimates, the most efficient way of doing so would be to reduce the main source of error. For catch numbers-at-age, the focus should be on reducing the between-station variability in the length distributions. The simplest way to do this is to increase the number of stations.

However, this would have large resource implications. The alternative is to optimise the spatial stratification around the variability in abundance and mean length, assuming that this variability is related to the sampling location. However, as most surveys are directed at a number of species, a compromise needs to be found between the optimum sampling designs for the various target species.

In the current analysis, the CV of the estimated catch numbers-at-age was in the order of 22%. This is consistent with the results of the EVARES study (Beare *et al.,* 2003), which found sampling CVs of estimated numbers-at-age for a number of surveys to be in the range of 10-30%. The current Irish Groundfish Survey series has not been running for long enough to be used as a time series of abundance indices for the assessment of West of Ireland haddock (ICES, 2006b). However, its predecessor, the Irish West Coast Groundfish survey, has traditionally been used in the assessment. Figure 8.4 shows that the inter-annual variation in the numbers-at-age estimates from this survey series is very large: five-fold increases or decreases from one year to the next were not uncommon. If one assumes that the CV for this survey series is similar to the current estimate of 22%, one would expect the survey to provide a good description of the inter-annual changes in numbers-at-age for West of Ireland haddock. However, data for species with a less variable recruitment and lower catch rates are likely to have a much higher noise-to-signal ratio.

The EVARES study, evaluating the impact of research surveys on management advice (Beare *et al.,* 2003), found that stock assessment outputs for the recent year-classes tend to be very sensitive to survey information, although the impact varies between surveys. Information on recruiting year-classes is particularly important for forecasting and management considerations (Smith and Gavaris, 1993). Therefore the precision of the survey data will have a significant influence on the final assessment estimates. The current results also show that surveys can provide additional biological information with reasonable precision. This type of information is readily available for a large number of stocks (Tomkiewicz *et al.,* 2003a), and management advice would benefit from inclusion of this information as long as the uncertainty levels are explicitly stated.

8.5. Overall conclusions

The previous chapters have highlighted the valuable information that can be obtained from standard fisheries surveys. There is a vast amount of biological data available

from these surveys (e.g. Tomkiewicz *et al.*, 2003a) but these data are mainly used for recruitment and tuning indices. Detailed analysis of the available survey data can lead to increased understanding of the stock structure and life-history of many stocks. These analyses can be very cost-effective as most of the data are already available, while the findings can be great significance to fisheries science. Additionally, it is very easy to conduct small experiments during surveys; these can lead to valuable insights at negligible costs. A particular strength of survey data is the fine spatial resolution at which they are collected. The work presented here has highlighted some of the spatial structures that exist in biological data and the potential bias that might occur if these spatial structures are ignored.

Past survey data can be used to optimise future data collection. However, the optimum sampling strategy is an elusive target that varies between years, areas and stocks. In general, it is nearly always best to sample as representatively as possible. For example, it is inefficient to obtain a large number of otoliths from an area where catch rates are low, as the weighting that will be given to those age data will be very small compared to age data from an area with high catch rates. Unfortunately it is difficult to predict the spatial distribution of fish and a common strategy is to collect a fixed number of otoliths per size class. The advantage of length stratified sampling is that it ensures that the extremes of the distribution are sampled while avoiding over-sampling of the most common length classes. However, it would be very useful to critically evaluate the usefulness of ALKs versus random sampling or a third possibility: random sampling within a small number of size categories (e.g. small-medium-large). The present work has supplied and explored some of the tools necessary to conduct such an analysis.

Incorporation of more detailed biological information into the stock assessment process, will inevitably lead to incorporation of additional sources of error: the rules of error propagation state that the total random error will increase with each variable that is added to a parameter estimate. This argument might be used to argue that an estimate of female-only SSB or Total Egg Production (TEP) will always be less precise than an SSB estimate alone. However, the same rules state that errors propagate quadratically, therefore the largest source(s) of error will contribute disproportionally to the total error estimate. Therefore, adding variables that have a relatively small error, will increase the total error estimate only by a very small amount. Female-only SSB or TEP estimates are conceptually a better index of

reproductive potential than SSB alone and they might not reduce the precision to a great extent as long as the sex ratio of fecundity data are estimated with good precision, relative to the other variables like catch per unit effort.

Biological data can also be used in stock assessment for other purposes than improving stock measures of reproductive potential. Data on the age structure of a stock can be indicative of the resilience of a stock to a number of years of bad recruitment. Changes in mean length-at-age, condition, maturity or sex ratio can all be indicative of over-exploitation and it is essential that these parameters are monitored and analysed.

In summary: surveys are a rich source of data that appear to be under-utilised. Biological data from surveys have to potential to increase our understanding of the stock structure and life-history of stocks and can contribute to the stock assessment process.

8.6. Figures
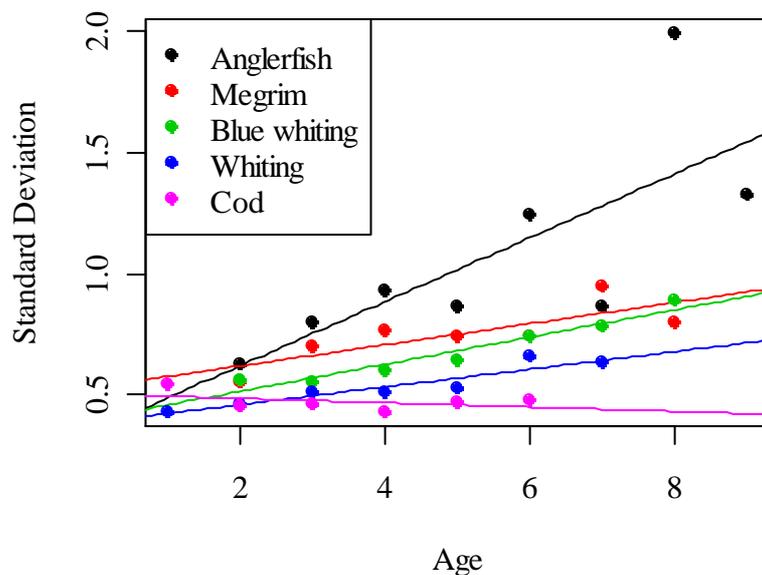
Figure 8.1. Standard deviation of the distribution of assigned ages around the modal age of anglerfish (Duarte et al., 2005), megrim (Egan et al., 2004), blue whiting (Worsøe Clausen et al., 2005), whiting (Easey et al., 2005) and cod (Woods, unpubl.).

Source of error
For each iteration estimate
these survey outputs

Resample individual length measurements	→	Estimate catch numbers-at-age
Resample length frequencies by station	→	Estimate mean weight-at-age
Resample individual biological samples	→	Estimate mean length-at-age
Resample biological data by station	→	Estimate maturity-at-age
Simulate ageing error	→	Estimate sex ratio
Simulate error in assignment of maturity	→	Estimate index of reproductive potential

Estimate errors from distribution of re-sampled parameter estimates

Figure 8.2. Flow diagram of the error estimation process. The total error was estimated by including all sources of error. The relative contribution of each source of error was also estimated by only allowing one source of error to vary and ignoring all other errors.
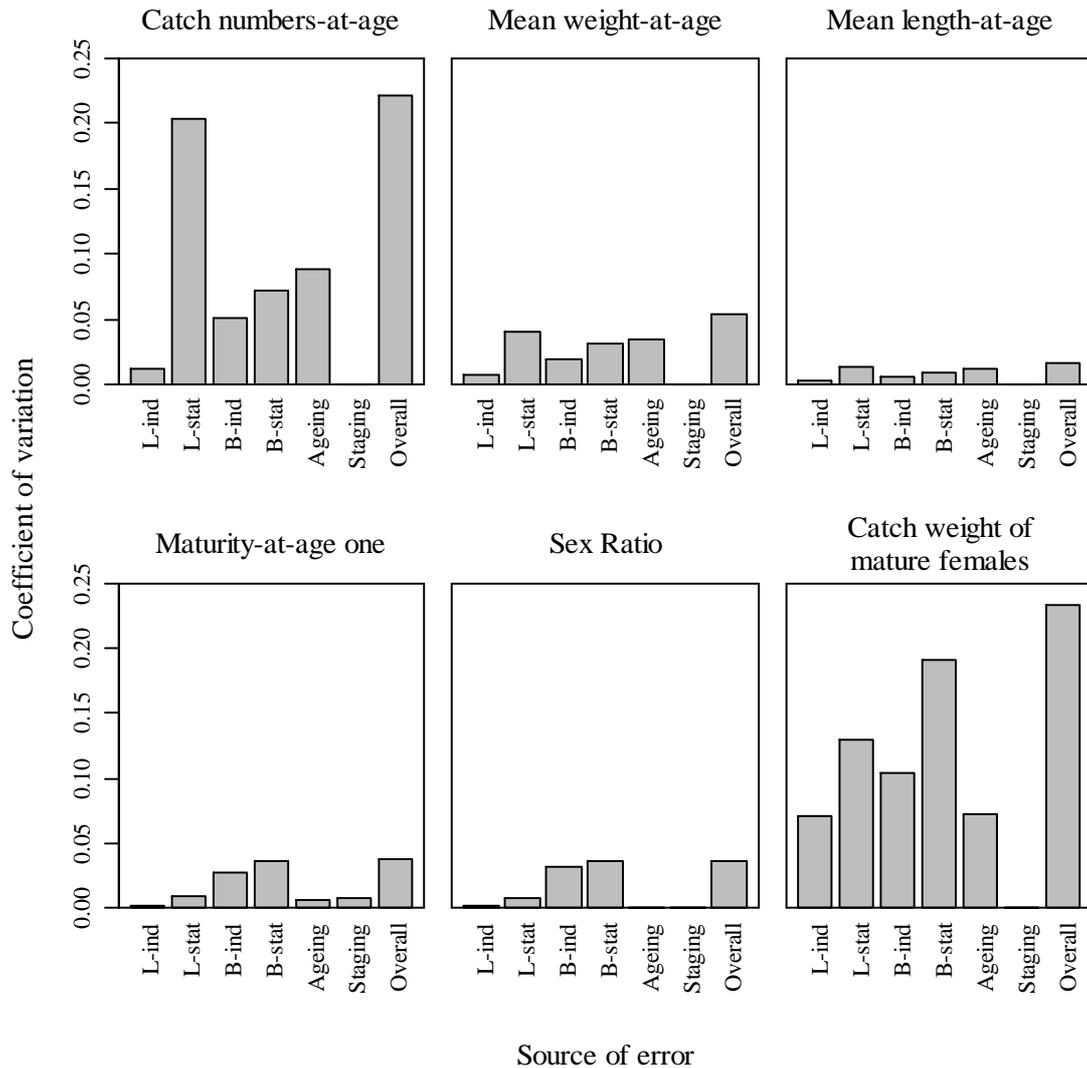
Figure 8.3. The estimated coefficient of variation of a number of survey outputs resulting from a range of error sources. The same scale was used for all plots. L-ind is the error in the survey outputs that is estimated by resampling individual length measurements. L-stat is the error estimate obtained from resampling length distributions by station. B-ind and B-stat are the respective error estimates obtained from resampling individual biological samples and resampling sets of biological samples by station. Ageing and staging refer to the simulated errors in ageing and staging. The overall error is the error estimate that takes all relevant sources of error into account. See methods section for more details.
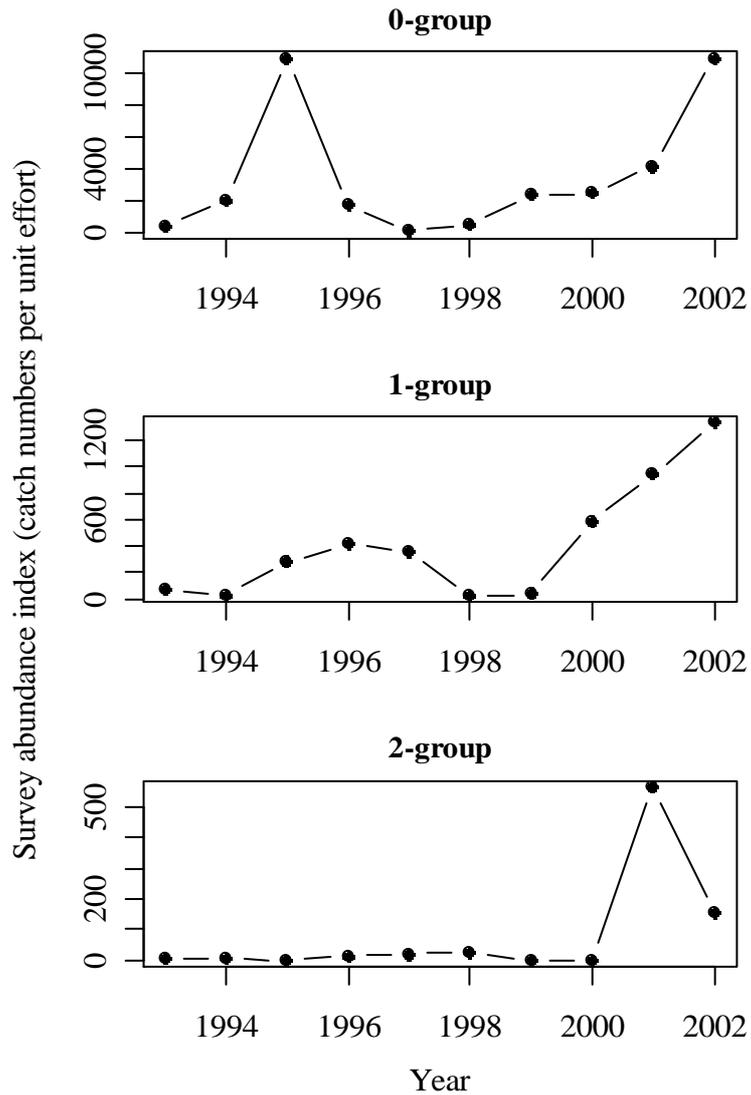
Figure 8.4. Abundance indices (catch numbers-at-age) from the Irish West Coast Groundfish Survey, which was part of the tuning fleet for the assessment of West of Ireland Haddock (ICES, 2006b).

Acknowledgements

References

Abrahams, M.V. and Dill, L.M., 1989. A determination of the energetic equivalence of the risk of predation. Ecology, 70(4): 999-1007.

Anderson, R.O. and Neumann, R.M., 1996. Length, weight and associated structural indices. In: B.R. Murphy and D.W. Willis (Editors), Fisheries Techniques, second edition. American Fisheries Society, Bethesda, Maryland, USA, pp. 447-482.

Anon, 2002. Final report of EU contract 98/090: Development and validation of egg production based estimates, using cod and plaice stocks in the Irish Sea.

Anon., 2001. Distribution and biology of anglerfish and megrim in waters to the west of Scotland. European Commission - DG XIV Study Contract 98/096.

Armstrong, M.J., Connoly, P., Nash, R.D.M., Pawson, M.G., Alesworth, E., Coulahan, P.J., Dickey-Collas, M., Milligan, S.P., O'Neil, M.F., Witthames, P.R. and Woolner, L., 2001. An application of the annual egg production method to estimate the spawning biomass of cod (*Gadus morhua* L.), plaice (*Pleuronectes platessa* L.) and sole (*Solea solea* L.) in the Irish Sea. ICES J. Mar. Sci., 58: 183-203.

Armstrong, M.J., Gerritsen, H.D., Allen, M., McCurdy, W.J. and Peel, J.A.D., 2004. Variability in maturity and growth in a heavily exploited stock: cod (*Gadus morhua* L.) in the Irish Sea. ICES J. Mar. Sci., 61(1): 98-112.

Beare, D., Castro, J., Cotter, J., Van Keeken, O., Kell, L., Laurec, A., Mahé, J.C., Moura, O., Munch-Petersen, S., Nielsen, J.R., Piet, G., Simmonds, J., Skagen, D. and Sparre, P., 2003. Evaluation of research surveys in relation to management advice (EVARES). FISH/2001/02 - Lot 1, European Commission.

Beare, D. and McKenzie, E., 1999. The multinomial logit model: a new tool for exploring Continuous Plankton Recorder data. Fisheries Oceanography, 8(s1): 25-39.

Beare, D.J., Needle, C.L., Burns, F. and Reid, D.G., 2005. Using survey data independently from commercial data in stock assessment: an example using haddock in ICES Division VIa. ICES J. Mar. Sci., 62(5): 996-1005.

Beverton, R.J.H., 1964. Differential catchability of male and female plaice in the North sea and its effect on estimates of stock abundance. Rapp. P.-v. Réun. Cons. int. Explor. Mer., 155: 103-122.

Beverton, R.J.H. and Holt, S.J., 1957. On the dynamics of exploited fish populations. Fisheries Investment Series 2. Vol 19. UK Ministry of Agriculture and Fisheries., London, 553pp. pp.

Bland, J.M. and Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. BMJ, 310: 170.

Boon, M.J., 1984. The UK megrim fishery in ICES sub-area VII: A preliminary analysis. ICES CM 1984/G:17.

Bowers, A.B., 1954. Breeding and growth of whiting (*Gadus merlangus* L.) in the Isle of Man waters. J. Mar. Biol. Ass. UK., 33: 97-122.

Bradford, M.J., 1991. Effects of ageing errors on recruitment time series estimated from sequential population analysis. Can. J. Fish. Aquat. Sci., 48: 555-558.

Brodziak, J. and Mikus, R., 2000. Variation in life history parameters of Dover sole, *Microstomus pacificus,* off the coast of Washington, Oregon and northern California. Fish. Bull., 98(4): 661-673.

REFERENCES

Bromley, P.J., 2000. Growth, sexual maturation and spawning in Central North Sea plaice (*Pleuronectes platessa* L.), and the generation of maturity ogives from commercial catch data. J. Sea Res., 44: 27-43.

Campana, S.E., 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. J. Fish. Biol., 59(2): 197-242.

Campana, S.E., Annand, M.C. and McMillan, J.I., 1995. Graphical and statistical methods for determining the consistency of age determinations. Trans. Am. Fish. Soc., 124: 131-138.

Chouinard, G.A. and Swain, D.P., 2002. Depth-dependent variation in condition and length-at-age of Atlantic cod (*Gadus morhua*) in the southern Gulf of St. Lawrence. Can. J. Fish. Aquat. Sci., 59: 1451-1459.

Claereboudt, M.R., McIlwain, J.L., Al-Oufi, H.S. and Ambu-Ali, A.A., 2005. Patterns of reproduction and spawning of the kingfish (*Scormberomorus commerson*, Lacépède) in the coastal waters of the Sultanate of Oman. Fish. Res., 73: 273-282.

Clark, W.G., 2004. Nonparametric estimates of age misclassification from paired readings. Can. J. Fish. Aquat. Sci., 61: 1881-1889.

Cochran, W.C., 1977. Sampling Techniques. John Wiley & sons, New York, 428pp. pp.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psycol. Meas., 20(1): 37-46.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull., 70: 213-220.

Collett, D., 2003. Modelling binary data. Chapman & Hall/CRC, 387pp. pp.

REFERENCES

Cotter, A.J.R., 1998. Method for estimating variability due to sampling of catches on a trawl survey. Can. J. Fish. Aquat. Sci., 55: 1607-1617.

Draper, N.R. and Smith, H., 1998. Applied regression analysis. Wiley, New York, 706 pp. pp.

Du Buit, M.H., 1984. Prédation de la cardine (*Lepidorhombus whiffiagonis* W.) au nord et à la ouest de l'Ecosse (In French). J. Cons. int. Explor. Mer., 41: 194-198.

Duarte, R., Landa, J., Morgado, C., Marçal, A., Warne, S., Barcala, E., Bilbao, E., Dimeet, J., Djurhuus, H., Jónsson, E., McCormick, H., Ofstad, L.H., Quincoces, I., Rasmussen, H., Thaarup, A., Vidarsson, T. and Walmsley, S., 2005. Report of the Anglerfish Illicia/Otoliths Ageing Workshop., IPIMAR (Lisbon).

Dwyer, K., Koen-Alonso, M. and Walsh, S.J., 2004. Finding the magical minimum sample size: a computer-intensive approach to minimize re-ageing effort to construct age-length keys for yellowtail flounder. NAFO SCR, Doc 04/09(Ser. No. N5002): 1-4.

Easey, M., Henderson, G. and Shanks, A.M., 2005. Report of the whiting (*Merlangius merlangus*, L) otolith exchange scheme 2004 and workshop 2005, CEFAS, Lowestoft, UK.

Efron, B. and Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman & Hall. CRC, 436pp. pp.

Egan, A., Etherton, M., Gomez de Segura, A., Iriondo, A., Marcal, A., Power, A., Quincoces, I., Santurtun, M., Vingaard Larsen, P. and Warne, S., 2004. Workshop on megrim otolith age readings, AZTI, Sukarrieta.

Eltink, A.T.G.W., Newton, A.W., Morgado, C., Santamaria, M.T.G. and Modin, J., 2000. Guidelines and tools for age reading comparisons, European Fish Ageing Network (EFAN), Flødevigen.

Erzini, K., 1990. Sample size and grouping of data for length-frequency analysis. Fish. Res., 9(4): 355-366.

Fahy, E. and Fannon, E., 1991. The exploitation of megrim *Lepidorhombus whiffiagonis* by the demersal fleet. Irish Fish. Invest. Ser. B No. 38, Department of the Marine, Dublin.

Fisher, R.A., 1930. The Genetical Theory of Natural Selection: A Complete Variorum Edition. (Edited by H. Bennet, 2000). Oxford University Press, Oxford, 354pp. pp.

Flatman, S., 1990. The evolution of biological sampling targets and the precision of estimated age compositions with reference to demersal fish sampling in ICES Divisions VIIa, VIIe and VIIf and g. Internal report No. 21, MAFF, Lowestoft.

Fox, C.J., Dickey-Collas, M. and Winpenny, A.J., 1997. Spring plankton surveys of the Irish Sea in 1995: The distribution of fish eggs and larvae. Sci. Ser. Tech. Rep. No. 104, CEFAS, Lowestoft.

Fridriksson, A., 1934. On the calculation of age-distribution within a stock of cod by means of relatively few age-determinations as a key to measurements on a large scale. Rapp. P.-v. Réun. Cons. int. Explor. Mer., 86(6): 1-5.

Gerritsen, H. and McGrath, D., 2005. Improving the consistency in the assignment of maturity stages, ICES Annual Science Conference, Aberdeen 20-24 September 2005.ICES CM 2005/Q:32.

Gerritsen, H.D., Armstrong, M.J., Allen, M., McCurdy, W.J. and Peel, J.A.D., 2003. Variability in maturity and growth in a heavily exploited stock: whiting (*Merlangius merlangus* L.) in the Irish Sea. J. Sea Res., 49(1): 69-82.

Gerritsen, H.D. and McGrath, D., 2006. Variability in the assignment of maturity stages of plaice (Pleuronectes platessa L .) and whiting (Merlangius merlangus L.) using macroscopic criteria. Fish. Res, 77(1): 72-77.

Gerritsen, H.D. and McGrath, D., 2007. Precision estimates and suggested sample sizes for length frequency data. Fish Bull, 105(1): 116-120.

Gerritsen, H.D. and McGrath, D., in press. Significant differences in the length-weight relationships of neighbouring stocks can result in biased biomass estimates: examples of haddock (*Melanogrammus aeglefinus*, L.) and whiting (*Merlangius merlangus*, L.). Fish. Res., XX(X): XX.

Gerritsen, H.D., McGrath, D. and Lordan, C., 2006. A simple method for comparing age-length keys reveals significant regional differences within a single stock of haddock (Melanogrammus aeglefinus). ICES J. Mar. Sci., 63(3): 1096-1100.

Gerritsen, H.D., McGrath, D. and Lordan, C., submitted. Differences in depth distribution of male and female megrim (*Lepidorhombus whiffiagonis*, Walbaum) associated with differences in life-history strategies. J. Fish. Biol., XX: XX-XX.

Gibson, R., 2005. The behaviour of flatfishes. In: R. Gibson (Editor), Flatfishes. Biology and exploitation. Fish and aquatic resources. Blackwell publishing, Oxford, pp. 213-239.

Godø, O.R. and Sunnanå, K., 1992. Size selection during trawl sampling of cod and haddock and its effect on abundance indices at age. Fish. Res., 13(3): 293-310.

REFERENCES

Gomez-Buckley, M., Conquest, L., Zitzer, S. and Miller, B., 1999. Use of statistical bootstrapping for sample size determination to estimate length-frequency distributions for pacific albacore tuna (*Thunnus alalunga*). Final report to National Marine Fisheries Services. FRI UW 9902.

González, C. and Paz, X., 2005. Distribution of American plaice (*Hippoglossoides platessoides*) on the Grand Bank (NAFO Divisions 3NO), 1995-2004. NAFO SCR, 05/30. Ser No. N5118: 16pp.

Gulland, J.A., 1966. Manual of sampling and statistical methods for fisheries biology. Part 1. Sampling methods. FAO Man. Fish. Sci. No. 3: 87pp.

Gutreuter, S. and Krzoska, D.J., 1994. Quantifying precision of in situ length and weight measurements of fish. N. Am. J. Fish. Manage., 14: 318-322.

Haddon, M., 2001. Modelling and quantitative methods in fisheries. Chapman & Hall/CRC, Boca Raton, 406pp. pp.

Hannan, J.F., 2002. Age, growth and population dynamics of lemon sole *Microstomus kitt* (Walbaum 1792) sampled off the west coast of Ireland, Galway-Mayo Institute of Technology, Galway.

Hastie, T.J. and Tibshirani, R.J., 1990. Generalized additive models. Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, 335pp. pp.

Hayes, D.B., 1993. A statistical method for evaluating differences between age-length keys with application to Georges Bank haddock, *Melanogrammus aeglefinus*. Fish. Bull., 91(3): 550-557.

Hayes, D.B., Brodziak, J.K.T. and O'Gorman, J.B., 1995. Efficiency and bias of estimators and sampling designs for determining length-weight relationships of fish. Can. J. Fish. Aquat. Sci., 52(1): 84-92.

REFERENCES

Heales, D.S., Brewer, D.T. and Jones, P.N., 2003. Subsampling trawl catches from vessels using seawater hoppers: are catch composition estimates biased? Fisheries Research, 63(1): 113-120.

Holtby, L.B. and Healey, M.C., 1990. Sex-Specific Life History Tactics and Risk-Taking in Coho Salmon. Ecology, 71(2): 678-690.

Horbowy, J., 1998. Comparison of age-length keys of Baltic cod derived from Polish commercial and research data. Fish. Res., 36: 257-266.

Horn, P.I., Tracey, D.M. and Clark, M.R., 1998. Between-area differences in age and length at first maturity of the orange roughy *Hoplostethus atlanticus.* Mar. Biol., 132: 187-197.

Hunter, J.R. and Macewicz, B.J., 2003. Improving the accuracy and precision of reproductive information used in fisheries. In: O.S. Kjesbu, J.R. Hunter and P. Witthames (Editors), Report of the working group on modern approaches to assess maturity and fecundity of warm- and cold-water fish and squids. Institute for Marine Research, Bergen, pp. 57-68.

Hutchings, J.A., 2005. Life history consequences of overexploitation to population recovery in Northwest Atlantic cod (*Gadus morhua*). Can. J. Fish. Aquat. Sci., 62: 824-832.

ICES, 1999. Manual for the International Bottom Trawl Surveys. Revision VI. ICES CM 1999/D:2, ICES, Lisbon, Portugal.

ICES, 2002. Manual for The International Bottom Trawl Surveys in the Western And Southern Areas. Revision II. ICES CM 2002/D:03, ICES, Dublin, Ireland.

ICES, 2004a. Report of the Working Group on the Assessment of Northern Shelf Demersal Stocks (WGNSDS), Copenhagen.

REFERENCES

ICES, 2004b. Report of the Working Group on the Assessment of Southern Shelf
    Demersal Stocks (WGSSDS), Oostende.

ICES, 2005a. Report for the Planning Group for Herring Surveys (PGHERS), Bergen,
    Norway.

ICES, 2005b. Report of the Working Group on the Assessment of Hake, Monk and
    Megrim (WGHMM), Lisbon, Portugal.

ICES, 2005c. Report of the Working Group on the Assessment of Northern Shelf
    Demersal Stocks (WGNSDS), Murmansk, Russia.

ICES, 2006a. ICES Oceanographic Database.

ICES, 2006b. Report of the Working Group on the Assessment of Southern Shelf
    Demersal Stocks (WGSSDS), Copenhagen, Denmark.

Johnsen, E., 2003. Improving the precision of length frequency distribution estimates
    from trawl surveys by including spatial covariance - using Namibian
    Merluccius capensis as an example. Fish. Res., 62(2003): 7-20.

Kimura, D.K., 1997. Statistical assessment of the Age-Length Key. J. Fish. Res.
    Board Can., 34(3): 317-324.

Kimura, D.K. and Anderl, D.M., 2005. Quality control of age data at the Alaska
    Fisheries Science Center. Mar. Freshwater Res, 56(5): 783-789.

Kimura, D.K. and Lyons, J.J., 1991. Between-reader bias and variability in the age-
    determination process. Fish. Bull., 89: 53-60.

King, M., 1995. Fisheries biology, assessment and management. Fishing News Books.
    Blackwell Science Ltd, Oxford, 431pp. pp.

King, P.A., Hannan, J.F., McGrath, D. and Veldon, M., 2006. Population dynamics,
    age, growth and maturity of lemon sole *Microstomus Kitt* (Walbaum 1792)

sampled between 2000-2002 off the west coast of Ireland. Irish Fish Invest, 16: 1-28.

Labonté, S.S.M., 1983. Aging capelin: enhancement of age-length keys and importance of such enhancement. In: W.G. Doubleday and D. Rivard (Editors), Sampling commercial catches of marine fish and invertebrates, Can. Spec. Publ. Fish. Aquat. Sci. 66, pp. 171-177.

Lai, H.L. and Helser, T., 2004. Linear mixed-effects models for weight-length relationships. Fish. Res., 70: 377-387.

Lambert, Y. and Dutil, J.D., 1997. Condition and energy reserves if Atlantic cod (*Gadus morhua*) during the collapse of the northern Gulf of St. Lawrence stock. Can. J. Fish. Aquat. Sci., 54: 2388-2400.

Landa, J. and Pineiro, C., 2000. Megrim (*Lepidorhombus whiffiagonis*) growth in the North-eastern Atlantic based on back-calculation of otolith rings. ICES Journal of Marine Science, 57(4): 1077-1090.

Landa, J., Pineiro, C. and Perez, N., 1996. Megrim (*Lepidorhombus whiffiagonis*) growth patterns in the northeast Atlantic. Fisheries Research, 26(3-4): 279-294.

Lloret, J. and Rätz, H.J., 2000. Condition of cod (*Gadus morhua*) off Greenland during 1982-1998. Fish. Res., 48: 79-86.

Lozán, J.L., 1992. Sexual differences in food intake, digestive tract size, and growth performance of the dab, *Limanda limanda* L. Neth. J. Sea Res., 29: 223-227.

Malakoff, D., 1999. Bayes offers a 'new' way to make sense of numbers. Science, 286: 1460 - 1464.

Marshall, C.T., Kjesbu, O.S., Yaragina, N.A., Solemdal, P. and Ulltang, Ø. 1998. Is
   spawner biomass a sensitive measure of the reproductive and recruitment
   potential of Northeast Arctic cod? Can. J. Fish. Aquat. Sci., 55: 1766-1783.

Marshall, C.T., Needle, C.L., Thorsen, A., Kjesbu, O.S. and Yaragina, N.A., 2006.
   Systematic bias in estimates of reproductive potential of an Atlantic cod
   (*Gadus morhua*) stock: implications for stock–recruit theory and management.
   Can. J. Fish. Aquat. Sci., 63(5): 980-994.

Marshall, C.T., O'Brien, L., Tomkiewicz, J., Köster, F.W., Kraus, G., Morgan, M.J.,
   Saborido-Rey, F., Blanchard, J.L., Secor, D.H., Wright, P.J., Mukhina, N.V.
   and Björnsson, H., 2003. Developing alternative indices of reproductive
   potential for use in fisheries management: Cases studies for stocks spanning an
   information gradient. J. Northw. Atl. Fish. Sci., 33: 161-190.

Marteinsdottir, G. and Thorarinsson, K., 1998. Improving the stock-recruitment
   relationship in Icelandic cod (*Gadus morhua*) by including age diversity of
   spawners. Can. J. Fish. Aquat. Sci., 55: 1372-1377.

McCullagh, P. and Nelder, J.A., 1989. Generalized linear models. Monographs on
   statistics and applied probability. Chapman and Hall, London, 511pp. pp.

McKenzie, D.P., Mackinnon, A.J., Péladeau, N., Onghena, P., Bruce, P.C., Clarke,
   D.M., Harrigan, S. and McGorry, P.D., 1996. Comparing correlated kappas by
   resampling: is one level of agreement significantly different from another? J.
   Psychiat. Res., 30(6): 483-492.

Milinski, M. and Heller, R., 1978. Influence of a predator on the optimal foraging
   behaviour of sticklebacks (*Gasterosteus aculeatus* L.). Nature, 275: 642-644.

Morgan, M.J. and Hoenig, J.M., 1997. Estimating maturity-at-age from length-
   stratified sampling. J. Northw. Atl. Fish. Sci., 21: 51-63.

REFERENCES

Morgan, M.J. and Trippel, E.A., 1996. Skewed sex ratios in spawning shoals of
Atlantic cod (*Gadus morhua*). ICES J. Mar. Sci., 53: 820-826.

Müller, H., 1996. Minimum sample sizes for length distributions of the catch
estimated by an empirical approach. ICES CM, 1996/J12: 18pp.

Nichols, J.H., Haynes, G.M., Fox, C.J., Milligan, S.P., Brander, K.M. and Chapman,
R.J., 1993. Spring plankton surveys of the Irish Sea in 1982, 1985, 1987, 1988
and 1989: hydrography and the distribution of fish eggs and larvae. Fish. Res.
Tech. Rep. No. 95, MAFF, Lowestoft.

Oeberst, 2000. A universal cost function for the optimization of the number of age
readings and length measurements for Age-Length-Key tables. Arch. Fish.
Mar. Res., 48(1): 43-60.

Pauly, D., 1994. On the sex of fish and the gender of scientists - collected essays in
fisheries science. Fish and fisheries series. Chapman and Hall, London, 272pp.
pp.

Pennington, M., Burmeister, L.M. and Hjellvik, V., 2002. Assessing the precision of
frequency distributions estimated from trawl-survey samples. Fish. Bull.,
100(1): 74-80.

Poulard, J.C., Peronnet, I. and Rivoalen, J.J., 1999. Depth and spatial distributions of
Lepidorhombus whiffiagonis (Walbaum, 1792) by age group in Celtic Sea and
Bay of Biscay. ICES CM, G:43: 11pp.

Power, G.R., King, P.A., Kelly, C.J., McGrath, D., Mullins, E. and Gullaksen, O.,
2006. Precision and bias in the age determination of blue whiting,
Micromesistius poutassou (Risso, 1810), within and between age readers. Fish.
Res., 80: 312-321.

REFERENCES

Rätz, H.J. and Lloret, J., 2003. Variation in fish condition between Atlantic cod (*Gadus morhua*) stocks, the effect on their productivity and management implications. Fish. Res., 60: 369-380.

R-Development-Core-Team, 2005. R: A language and environment for statistical computing. R Foundation for statistical Computing, Vienna, Austria.

Reeves, S.A., 2001. The implications of age-reading errors for stock assessment and management advice: a case study based on Eastern Baltic Cod. ICES CM 2001/P:18.

Reeves, S.A., 2003. A simulation study of the implications of age-reading errors for stock assessment and management advice. ICES J. Mar. Sci., 60: 314-328.

Richards, L.J., Schnute, J., Kronlund, A.R. and Beamish, R.J., 1992. Statistical methods for the analysis of ageing error. Can. J. Fish. Aquat. Sci., 49: 1801-1815.

Ricker, W.E., 1975. Computation and interpretation of biological statistics of fish populations. Bull. Fish. Res. Board Can., 191: 382 pp.

Rijnsdorp, A.D., 1993. Selection differentials in male and female North Sea plaice and changes in maturation and fecundity. In: T.K. Stokes, J.M. McGlade and R. Law (Editors), The exploitation of evolving resources. Springer Verlag, Berlin, pp. 19-36.

Rijnsdorp, A.D. and Witthames, P., 2005. Ecology of reproduction. In: R. Gibson (Editor), Flatfishes. Biology and exploitation. Fish and aquatic resources. Blackwell publishing, Oxford, pp. 68-93.

Rindorf, A. and Lewy, P., 2001. Analysis of length and age distributions using continuation-ratio logits. Can. J. Fish. Aquat. Sci., 58(6): 1141-1152.

REFERENCES

Rivoirard, J., Simmonds, J., Foote, K.G., Fernandes, P. and Bez, N., 2000. Geostatistics for estimating fish abundance. Blackwell Science Ltd., Oxford, 206pp. pp.

Roa, R., Ernst, B. and Tapia, F., 1999. Estimation of size at sexual maturity: an evaluation of analytical and resampling procedures. Fish. Bull., 97: 570-580.

Robson, S.M., 2004. Age, growth reproductive biology and population dynamics of the common megrim (*Lepidorhombus Whiffiagonus*, Waldbaum, 1792) from off the west coast of Ireland. PhD Thesis, Galway-Mayo Institute of Technology, Galway, 231pp pp.

Rochet, M.J., 2000. Spatial and temporal patterns in age and size at maturity and spawning stock biomass of North Sea gadoids. ICES CM, 2000/N:26: 13pp.

Roff, D.A., 1982. Reproductive strategies in Flatfish: a first synthesis. Can. J. Fish. Aquat. Sci., 39: 1686-1698.

Rosenberg, A.A. and Reddington, J.R., 1988. Length-based methods of fish stock assessment. In: J.A. Gulland (Editor), Fish Population Dynamics (Second Edition). John Wiley & Sons Ltd., Chochester, pp. 83-103.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G., 1986. Akaike Information Criterion Statistics. D. Reidel Publishing Co, 312pp. pp.

Sánchez, F., Pérez, N. and Landa, J., 1998. Distribution and abundance of megrim (*Lepidorhombus boscii* and *Lepidorhombus whiffiagonis*) on the northern Spanish shelf. ICES Journal of Marine Science, 55(3): 494-514.

Schnute, J. and Fournier, D., 1980. A new approach to length-frequency analysis: growth structure. Can. J. Fish. Aquat. Sci., 37(9): 1337-1351.

Schuster, C., 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. Educational and Psycological Measurement, 62(2): 243-253.

Scott, B., Marteinsdottir, G. and Wright, P., 1999. Potential effects of maternal factors on spawning stock-recruitment relationships under varying fishing pressure. Can. J. Fish. Aquat. Sci., 56: 1882-1890.

Smith, B.D., McFarlane, G.A. and Saunders, M.W., 1992. Inferring the summer distribution of migratory Pacific hake (*Merluccius productus*) from latitudinal variation in mean lengths-at-age and length frequency distributions. Can. J. Fish. Aquat. Sci., 49: 708-721.

Smith, S.J. and Gavaris, S., 1993. Improving the precision of abundance estimates of eastern Scotian Shelf Atlantic cod from bottom trawl surveys. N. Am. J. Fish. Manage., 13: 35-47.

Smith, S.J. and Maguire, J.J., 1983. Estimating the variance of length composition samples. In: W.G. Doubleday and D. Rivard (Editors), Sampling commercial catches of marine fish and invertebrates. Can. Spec. Pub. Fish. Aquat. Sci. 66, Ottawa, pp. 165-170.

Solmundsson, J., Karlsson, H. and Palsson, J., 2003. Sexual differences in spawning behaviour and catchability of plaice (*Pleuronectes platessa*) west of Iceland. Fish. Res., 61: 57-71.

Sprugel, D.G., 1983. Correcting for bias in log-transformed allometric equations. Ecology, 61(1): 209-210.

Stoner, W.A., Beja, A.J., Manderson, J.P., Phelan, B.A., Stehlik, L.L. and Pessutti, J.P., 1999. Behaviour of winter flounder, *Pseudopleuronectes americanus*,

during the reproductive season: laboratory and field observations on spawning, feeding, and locomotion. Fish Bull, 97(4): 999-1016.

Swain, D.P., 1997. Sex-specific temperature distribution of American plaice (*Hippoglossoides platessoides*) and its relation to age and abundance. Can. J. Fish. Aquat. Sci., 54: 1077-1087.

Swain, D.P. and Morin, R., 1996. Relationships between geographic distribution and abundance of American plaice (*Hippoglossoides platessoides*) in the southern Gulf of St. Lawrence. Can. J. Fish. Aquat. Sci., 53: 106-119.

Swain, D.P. and Morin, R., 1997. Effects of age, sex and abundance on the bathymetric pattern of American plaice in the southern Gulf of St Lawrence. Journal of Fish Biology, 50(1): 181-200.

Taylor, J.R., 1997. Introduction to error analysis: The study of uncertainties in physical measurements. University Science Books, Sausalito, 327pp. pp.

Thompson, S.K., 1987. Sample size for estimating multinomial proportions. The American Statistician, 41(1): 42-46.

Tomkiewicz, J., Morgan, M.J., Burnett, J. and Saborido-Rey, F., 2003a. Available Information for Estimating Reproductive Potential of Northwest Atlantic Groundfish Stocks. J. Northw. Atl. Fish. Sci., 33: 1-21.

Tomkiewicz, J., Tybherg, L. and Jespersen, Å. 2003b. Micro- and macroscopic characteristics to stage gonadal maturation of female Baltic cod. J. Fish. Biol., 62: 253-275.

Trenkel, V.M., Pinnegar, J.K., Dawson, W.A., Du Buit, M.H. and Tidd, A.N., 2003. Does diet in Celtic Sea fishes reflect prey availability? Journal of Fish Biology, 63(s1): 197-212.

REFERENCES

Trippel, E.A., 2003. Estimation of Male Reproductive Success of Marine Fishes. J. Northw. Atl. Fish. Sci., 33: 81-113.

Venables, W.N. and Ripley, B.D., 1994. Modern Applied Statistics with S-PLUS. Statistics and Computing. Springer, New York, 495pp. pp.

Venables, W.N., Smith, D.M. and R-Development-Core-Team, 2005. An Introduction to R. (Notes on R: A programming environment for data analysis and graphics. Version 2.1.1.), 90pp. http://cran.r-project.org/.

Villarroel, J.C., Acuña, E.H. and Andrade, M.J., 2001. Feeding and distribution of the bigeye flounder *Hippoglossina macrops* off Northern Chile. Mar. Freshw. Res, 52(6): 833-841.

Vitale, F., Svedäng, H. and Cardinale, M., 2006. Histological analysis invalidates macroscopically determined maturity ogives of the Kattegat cod *(Gadus morhua)* and suggests new proxies for estimating maturity status of individual fish. ICES J. Mar. Sci., 63(3): 485-492.

Vokoun, J.C., Rabeni, C.F. and Stanovick, J.S., 2001. Sample-size requirements for evaluating population size structure. N. Am. J. Fish. Manage., 21: 660-665.

Warnes, S. and Jones, B.W., 1995. Species distribution from English Celtic Sea groundfish surveys, 1984 to 1991. Fisheries research technical report number 98, MAFF, Lowestoft.

West, G., 1990. Methods of assessing ovarian development in fishes: a review. Aust. J. Mar. Freshwater Res., 41: 199-222.

Woods, F., unpubl. North Sea Cod exchange 2005/6, Marine Institute, Galway (Ireland).

Worsøe Clausen, L., Power, G., Timoshenko, N. and Tangen, Ø. 2005. Report of the Blue Whiting Otolith Ageing Workshop, DIFRES, Hirtshals, Denmark.

REFERENCES

Yaragina, N.A. and Marshall, C.T., 2000. Trophic influences on interannual and

seasonal variation in the liver condition index of northeast Arctic cod (*Gadus*

morhua). ICES J. Mar. Sci., 57: 42-55.

Zar, J.H., 1999. Biostatistical analysis. Prentice-Hall, Inc., Upper Saddle River, New

Jersey, 663 pp.

Appendix A

Glossary of terms

Age-at-length – Age distribution by length class

AIC – Akaike Information Criterion, used to evaluate the fit of a model versus the number of parameters required, mainly used in Generalised Linear Models (Sakamoto *et al.,* 1986)

ALK – Age-Length Key, a matrix with proportions of each age class at each length class

Bayesian statistics – A statistical method that includes prior knowledge in uncertainty estimates and requires likelihood functions to be explicitly defined

Bias – Difference between a parameter estimate and the true value of the parameter

Bonferroni correction – Correction used when a number of independent hypotheses are tested: the significance level for accepting each hypothesis is reduced to avoid false-positive outcomes

Biological data – Demographic data from individual fish like age, weight, sex or maturity stage

Biomass – The sum of weights of individuals in a fish stock

Bootstrap – Re-sampling, with replacement, from a dataset in order to simulate the variability that would result from replicate samples (Efron and Tibshirani, 1993)

Bottom trawl survey – See groundfish survey

Canonical age – Assumed 'true' age of a fish

Catchability – The efficiency with which a certain gear type and configuration catches fish of a certain size, shape and behaviour

Commercial data – Data collected from catches of commercial fisheries vessels

Confidence interval – Range of values within which the true value most likely lies

CV – Coefficient of variation, a measure of precision defined as the standard error divided by the mean

Demersal – Living close to the bottom of the sea

Fecundity – Number of (viable) eggs produced by an individual

GAM – Generalized Additive Model, a combination of multiple regression and additive models (Hastie and Tibshirani, 1990)

Geostatistics - A collection of statistical methods describing spatial autocorrelation among sample data, which can then be used in various types of spatial models

GOV – Grande Overture Verticale, otter trawl used on IBTS surveys (ICES, 2002)

GLM – Generalized Linear Model, generalisation of ordinary least-squares regression (McCullagh and Nelder, 1989)

Groundfish survey – A survey directed at demersal fish.

Histology – A technique to section and stain tissues for microscopic analysis.

Length-at-age – Length distribution by age class

Length-stratified sampling – Non-random sampling whereby a target number of individuals are collected for each length class, usually for the determination of biological parameters

Macroscopic maturity determination – Determination of the maturity stage of a fish by examining the gonads without visual aids

MALK – Maturity-Age-Length Key, a matrix with the proportions of each age class and maturity state at each length class

Maturity-at-age – Proportion of mature fish by age class

Maturity-at-length – Proportion of mature fish by length class

Maturity stage – Classification of development of the gonads. Maturity stages might be assigned either by macroscopic or microscopic examination of the gonads

Maturity state – Distinction between mature and immature fish

Measurement error – Error due to the measuring device or the interpretation of the observer

Observer error – Error due to the interpretation of the observer

Ogive – S-shaped curve, describing expected proportions by age or length class, usually describing maturity or size selection

Oocyte – Female reproductive cell

Otolith – Calcified structure, routinely used for age determination in fish.

Random error – Variability between replicate samples

Recruitment – The addition of young fish to a stock.

Reproductive potential – A measure that is assumed to correlate with the amount of offspring that a stock produces, e.g. SSB or annual egg production.

Re-sampling – Taking random samples from an existing dataset, usually with replacement

Sampling error – Error related to the sampling design

Selectivity – See catchability

Sexual dimorphism – Differences in growth between the sexes.

SSB – Spawning Stock Biomass, the biomass of the proportion of the population that is mature

Spline – A smooth function consisting of a number of polynomial functions

Stock assessment – The process of estimating the retrospective development of a stock and making quantitative predictions about the current and future state of the stock

Stratum – A subdivision of a population. This can be a spatial subdivision or a categorical subdivision like in length-stratified sampling. Within each stratum sampling is usually random.

Subsample – A random sample from the catch (see also length-stratifed sampling)

Systematic error – See bias

Tuning – Use of an index of abundance to calibrate an assessment model. Time series with constant catch efficiency (usually survey data) are used to correct for changes in efficiency in the commercial fleet by adjusting the initial model parameters until it becomes consistent with the time series data.

Two-stage sampling – Obtaining a random sample of the catch for length measurements as well as a (non-random) length-stratified biological sample

Variogram – Geostatistical tool to describe spatial correlation

Vitellogenesis – Yolk formation in an oocyte, giving it an opaque yellowish appearance

Year class – Group of fish of the same age

Appendix B

Unpublished data by the author

There are a number of references in the main text to unpublished work by the author. Below are some brief results of these experiments. The first experiment investigated the variability and bias in length measurements using electronic measuring boards. During the Biological Sampling Survey 2005, three people were asked to repeatedly measure three randomly selected samples of 30 whiting (Figure 1). Electronic measuring boards were used. These have a barcode for each 1cm length class that is read with a light pen (Figure 2). Person 3 consistently estimated the mean length to be lower by around 0.5cm. The difference was not significant using ANOVA or non-parametric tests. Perhaps more interesting is the fact that there were significant differences between the samples, which were selected from a larger sample by haphazardly selecting fish ($P<0.05$, t-test). Further work on subsampling techniques will be carried out.

The second experiment examined the size distribution of oocytes at different developmental stages (Figure 3). These data illustrate that the gap in size between previtellogenic and vittelogenic oocytes is much larger in plaice than it is in whiting. This might explain why there was some confusion between immature and early maturing whiting ovaries in Chapter 3.

The last experiment was conducted to investigate the precision of a number of weighing scales used on the Irish Groundfish Survey 2006. The scales were calibrated at the start of the experiment and calibration weights were placed on the scales at regular intervals to measure the deviation from the 'true' weight (Figure 4). Most scales remained accurate within 1% over a two-day period. It is normal practice to calibrate all the scales at the beginning of each day.
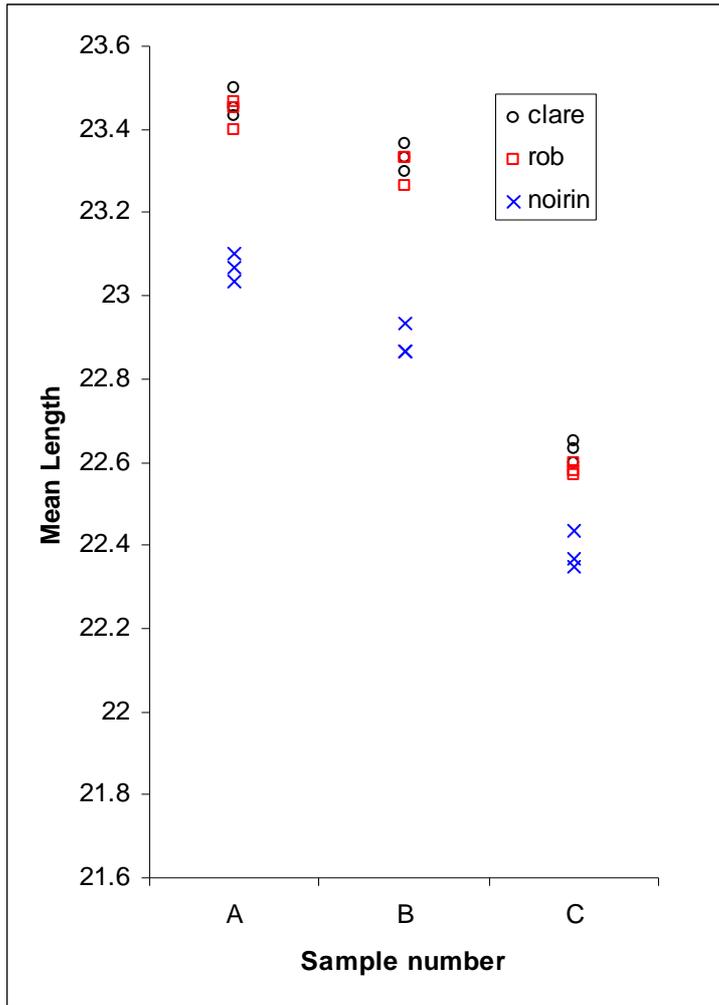
Figure 1. Mean length of three samples of 30 whiting, measured three times by three people. Person 3 consistently estimated the mean length to be around 0.5 cm lower than the others.

Figure 2. Electronic measuring board. Each 1cm length class corresponds to a barcode that is read with a light pen.
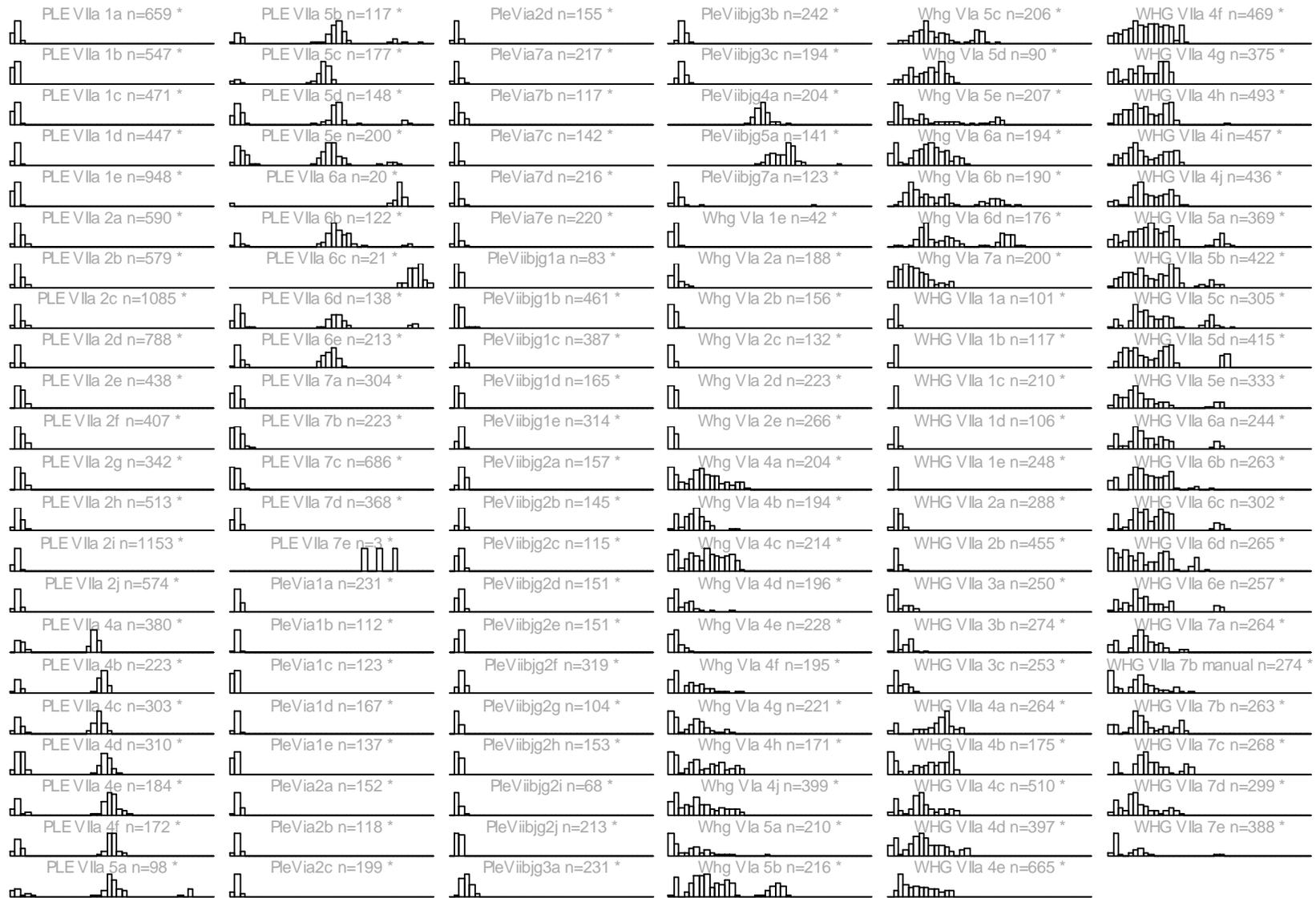
Figure 3. Size frequency analysis of oocytes of plaice (PLE) and whiting (WHG) at various developmental stages (1-7) in ICES divisions VIa, VIIa and VIIbjg.

Figure 4. Deviation of a number of weighing scales over time (Scales A: 45kg capacity; 10kg calibration weight, the other scales have 13kg capacity and a 200g calibration weight was used). Deviation is the difference between the 'true' weight of and the measured weight.

Appendix C

Example R-code

```
# This is an example of the application of a multinomial model to
# age-length data as described in Chapter 4. The R software that
# is required to run this code, can be obtained from:
# www.R-project.org
# This code requires the package 'nnet' to be installed

require(nnet)

# Some age-at-length data
LENGTH=c(8,7,8,14,17,18,14,14,16,15,18,15,14,16,21,17,17,13,20,17,15,
19,14,16,13,16,13,14,15,16,16,19,15,21,25,14,21,20,20,17,19,20,18,14,
21,21,22,24,24,23,16,22,20,22,19,24,20,28,20,23,21,29,24,20,19,19,24,
23,21,15,24,19,17,17,18,21,20,26,20,18,20,20,22,21,22,23,21,27,26,27,
23,20,35,29,25,27,24,23,23,18)
AGE=c(0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3)
AREA=c(1, 2, 2, 2, 1, 2, 1, 1, 2, 1, 1, 2, 1, 1, 1, 2, 2, 2, 2, 1, 1,
2, 1, 1, 1, 2, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1,
2, 1, 1, 2, 2, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 1, 1, 2, 2, 2, 2, 1,
2, 1, 2, 1, 2, 1, 1, 2, 1, 2, 1, 1, 2, 2, 1, 1, 2, 2, 1, 1, 1, 1, 2,
1, 1, 2, 2, 1, 1, 2, 2, 2, 2)

# Length classes for which you want to predict the proportions at age
lengths=0:40

# The model
model=multinom(AGE~LENGTH)

# The predicted values
predicted=predict(model,data.frame(LENGTH=lengths),type='probs')

# Or just predict missing proportions at age for 30cm
round(predict(model,data.frame(LENGTH=30),type='probs'),2)

# Plot data and predicted values
rawdata=table(LENGTH,AGE)/rowSums(table(LENGTH,AGE))
ages=labels(rawdata)$AGE
plot(NA,type='n',xlim=c(min(lengths),max(lengths)),ylim=c(0,1),xlab='
Length',ylab='Proportion-at-age')
for(i in 1:length(ages)){
   lines(lengths,predicted[,i],col=i)
   text(as.numeric(labels(rawdata)$LENGTH),rawdata[,i],ages[i],col=i)
   }

# Compare ALK from 2 regions

model1=multinom(AGE~LENGTH+as.factor(AREA)) #this is a model that
includes the factor area as an explanatory variable
model2=multinom(AGE~LENGTH) #this is the model without AREA
anova(model1,model2,test='Chisq')
# This tests if model1 is 'better' than model2
# No significant difference between the areas.
```

Appendix D

Authorship of Armstrong *et al.* (2004)

To whom it may concern   **Date** 03 October 2006

.

Contribution of Hans Gerritsen to published paper on cod

This is to certify that Hans Gerritsen contributed at least 55% of the work to the following paper:

Armstrong, M.J., Gerritsen, H.D., Allen, M., McCurdy, W.J. and Peel, J.A.D., 2004. Variability in maturity and growth in a heavily exploited stock: cod (*Gadus morhua* L.) in the Irish Sea. ICES J. Mar. Sci., 61(1): 98-112.

Although I was senior author, Hans contributed substantially to the development of the study, completed the bulk of the analysis for the paper, carried out a literature review, prepared the figures and drafted a significant part of the text.

Yours sincerely,

Dr Michael J. Armstrong
Senior Fishery Scientist

Direct line 01502 52 4362
Email mike.armstrong@cefas.co.uk

Declaration

GMIT
GALWAY-MAYO INSTITUTE OF TECHNOLOGY
INSTITIÚID TEICNEOLAÍOCHTA NA GAILLIMHE-MAIGH EO

POSTGRADUATE CANDIDATE STATEMENT -
THESIS PRESENTATION/AWARD

HETAC/GMIT Postgraduate Research Degree Policy and Procedures require that a hardbound copy of the thesis be prepared in accordance with HETAC postgraduate research degree policy and procedures. Accordingly, there is a requirement for *"A Statement, signed by the candidate and the supervisor(s), indicating that the thesis represents the candidate's own work, or, in the case of a thesis based on a group project, indicating the extent of the candidate's individual contribution and making reference to any other theses submitted or material published by each collaborator in the project, should also be submitted at this time."*

Candidate Name: Hans Gerritsen
MRA Ref.
Register/Award   PhD   (MA/MB/MSc/MEng/ /PhD)
Thesis Title: Estimating fish stock population parameters from bottom trawl surveys:
                         sources and magnitude of uncertainty

Supervisor (1):   Dr David McGrath; Dr Pauline King
Supervisor (2):   Dr Colm Lordan

I, the above named, certify that the thesis wholly represents my own work.
I, the above named, certify that the thesis was based on a group project and my contribution and that of each collaborator are detailed on attached.
(Delete as appropriate)

Signature: _____ Date: _____
                (Student)

I/We, the above named Supervisor(s) certify that the thesis represents the candidate's work.

Signature: _____ Date: _____
                (Supervisor 1)

Signature   _____ Date: _____
                (Supervisor 2)