STATISTICAL MODELLING OF FISH STOCKS

Trine Kvist

LYNGBY 1999
IMM-PHD-1999-64

IMM

Trykt af IMM, DTU

Preface

This thesis has been prepared in the section for statistics at the Department of Mathematical Modelling at DTU, the Technical University of Denmark, in partial fulfillment of the requirements for the degree of PhD within the Mathematical Phd Program at DTU.

In this thesis uncertainty associated with stock assessment has been considered, in particular uncertainty associated with the input data to the model. The thesis provides new approaches to analyse the sources of variation in the input data and their magnitude, and an alternative approach for modelling the dynamics of a fish population is suggested.

The project has been directed towards the North Sea sandeel fishery. However, the methods developed may easily be transferred to other fisheries and areas.

Lyngby, July 1999

Trine Kvist

Acknowledgements

First of all I would like to thank my two supervisors, Poul Thyregod, Department of Mathematical Modelling, the Technical University of Denmark, and Henrik Gislason, University of Copenhagen, c/o Danish Institute for Fisheries Research, for their help and encouragement and for inspiring discussions during this work.

I also would like to thank the Danish Institute for Fisheries Research and the leader of the project, Peter Lewy, for the excellent collaboration.

My colleagues at IMM are thanked for their readiness to help and discuss the statistical matters of the project. Especially my room-mate through many years, Helle Andersen is thanked for her humour, encouragement and rational approach to statistical problems. The time-series group and Uffe Thygesen are thanked for their kind help on the matters of stochastic differential equations.

My colleagues at DFU, especially Anna Rindorf, are thanked for their help and discussions on matters related to the biological aspect of the project.

My husband Henrik, family and friends are thanked for their help, patience and encouragement during the hard parts of this work.

Summary

In this thesis uncertainty associated with stock assessment has been considered, especially uncertainty associated with the input data to the model. The thesis provides new approaches to analyse the sources of variation in the input data and their magnitude, and an alternative approach for modelling the dynamics of a fish population is suggested.

A new approach is introduced to analyse the sources of variation in age composition data, which is one of the most important sources of information in the cohort based models for estimation of stock abundancies and mortalities. The approach combines the continuation-ratio logits, which can take the ordinal and multinomial characteristics of the response into account, and the generalized linear mixed models, which allow for fixed as well as random effects to be analysed.

Catch at age data and the associated uncertainties have been estimated, by separating the statistical analysis into separate analyses of the various data sources. The results were combined into estimates of the catch at age data and the associated uncertainties for the sandeel landings from the North Sea in 1989 and 1991.

An overview of age-structured stock assessment models is given and it is argued that an approach utilising stochastic differential equations might be advantagous in fish stoch assessments.

Resumé

Denne PhD afhandling vedrørerer usikkerhed i modellering af fiskebestande, især usikkerhed i datagrundlaget. Afhandlingen beskriver en ny metode til analyse af variationskilder og deres omfang i datagrundlaget, og en alternativ metode for modellering af populationsdynamiken i en fiskebestand fremlægges.

Afhandlingen beskriver en ny metode til analyse af variationskilderne i alderssammensætningsdata, som er en af de vigtigste informationskilder i kohortebaserede modeller for estimation af bestandsstørrelser og dødeligheder. Metoden kombinerer teorier for fortsættelses-logiter, som tager højde for ordningen af responset såvel som de multinomiale karakteristika af responset, og de generaliserede lineære mixed modeller, som tillader analyse af både tilfældige og systematiske effekter.

Estimater af fangst per aldersgruppe og tilhørende usikkerheder er estimeret ved at opdele den statistiske analyse i særskilte analyser af de forskellige datakilder. Resultaterne kombineres til estimater af fangst per aldersgruppe samt usikkerheder, for tobisfangster fra Nordsøen i 1989 og 1991.

Et overblik over alders-strukturerede bestandsmodeller gives og det argumenteres for at en metode som benytter stokastiske differentialligninger kan være fordelagtig i modellering af fiskebestande.

Contents

Chapter 1

Introduction

1.1 Background

Exploited fish stocks are modelled in order to optimise the yield, make sure that it is sustainable and assess the impact of the fishery on the ecosystem. Such models are presumably rather inaccurate and model errors of a certain magnitude must be expected. In addition, observing the system in an ocean is difficult; some important information may not be available at all, inducing further uncertainties in the model and the observations might be prone to errors. Thus, in order to obtain reliable estimates and assess the associated uncertainties, statistical modelling of the fish stocks are certainly needed. Although much work has already been done in this area, lack of computer capacity has limited the development of the models.

The background of this particular project is that doubts have been raised by environmental organisations about the sustainability of the Danish industrial fishery in the North Sea. Although the present assessment of the impact of the fishery suggests that the fishery is sustainable (ICES, 1996), environmental organisations argue that the uncertainties are so large that it is reasonable to fear that the fishery might lead inadvertely to a stock collapse. They also fear that such a collapse could have detrimental consequences for the North Sea ecosystem at large. Thus it is important to assess the uncertainties of the relevant quantities in order to evaluate the legitimacy of the critisism.

The project has been directed towards the North Sea sandeel fishery, because it is the main target of the Danish industrial fishery. In addition, the investigations can be performed on Danish data alone as the sandeel fishery in the North Sea is completely dominated by the Danish fishery, except for a few areas outside Norway which is excluded from the investigations. Gaining access to fishery data can be both difficult and time-consuming.

Although directed towards a single species, the methods developed can easily be transferred to other fisheries and areas.

The sandeel fishery actually covers a few variants of sandeel, but the fishery is completely dominated by the lesser sandeel(*Ammodytes marinus* Raitt) and therefore in the following the term sandeel refers to *Ammodytes marinus*. It is one of the most abundant fish species in the North Sea (Sparholt, 1990). The name is apt because of its burrowing behaviour and physical appearance. It is a small slender fish, which feed on plankton, with a maximum length of approximately 25 cm. Sandeels occur in shoals and tend to be concentrated in well-defined areas where there is coarse well-oxygenated sand (Macer, 1966). The sandeel constitutes an important prey for many species of fish, seabirds and marine mammals (Daan *et al.*, 1990 and Wright, 1996).

The industrial fishery in the North Sea began in the early 1950s and has since developed into an important fishery accounting for approximately two thirds of the total landings of fish from the North Sea. The landings are processed to fish meal and oil or used directly as animal foodstuff. In the early years herring made up the bulk of the industrial landings, but in the 1970s the sandeel fishery increased rapidly (refer to figure 1.1) (Kirkegaard and Gislason, 1996). In the last 20 years appr. 700 000 tonnes of sandeel have been landed every year.

Since its start the industrial fishery has been subject to intense debate and discussion. On one hand it has been argued that the fishery provides a good way to utilise a resource that otherwise would remain untapped. On the other hand it has been argued that the large amount of small fish caught may deplete the food supplies of human consumption fish stocks and other predators such as seabirds, seals, cetaceans and salmonids. Another possible consequence is that industrial fishing because of the by-catch of species such as haddock, whiting and herring, remove fish which would become available to human consumption fisheries if they were left in the sea (Kirkegaard and Gislason, 1996). However, the by-catch in the sandeel

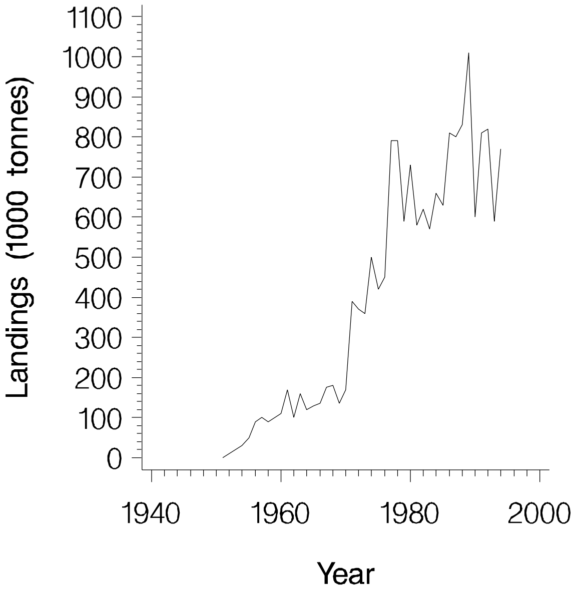

Figure 1.1: Sandeel landings in the Danish industrial fishery

fishery is small and hence that particular aspect is not a problem for this fishery.

The stock size and the impact of the fishery are assessed regularly by the International Council of Exploration of the Sea (e.g. ICES, 1996). The basic information in most fish stock assessments, including the assessment of the sandeel stock, is the catch at age data. The data consists of the estimated number of individuals caught for a given species, age, area and time period. In addition, information of the catch rate is also utilised as a measure of the abundance of the population. One often uses a standardised unity called catch per unit of effort (CPUE) and assumes that this is proportional to the abundance of the species. Information on CPUE can often be obtained from data on the fishery, but often fishery-independent information is desirable and surveys are performed regularly by the authorities. The advantages of the surveys are that they can be more controlled regarding to fishing position, equipment etc. On the other hand they are expensive to perform and the amount of data is much smaller than from the fishery, although probably less prone to errors. Unfortunately, survey data has not been available for the sandeel fishery, because sandeel is not caught by the standard equipment on the survey vessels. The main characteristics of the assessment model for sandeel are that the rate of removals from the population is proportional to the abundance of the population, that the mortality caused by other reasons than fishery has to be established outside the model and that the CPUE is assumed to be proportional to the abundance of the population. Even this short overview of the data sources and main assumptions in the model has adumbrated that considerable uncertainties may be associated with the data as well as the model. It is the aim of this project to contribute to the detection and assessment of the uncertainties and to improve the assessment model such that the sources of uncertainties may be more realistically modelled and thus give improved estimates of parameters and uncertainties. The obtained information on sources of uncertainties may also be utilised to find efficient ways of reducing the uncertainties. Focus has been on the uncertainties associated with the age composition and catch at age data, but approaches to improve upon the assessment model has also been investigated.

1.2 Outline of the thesis

In the present chapter the background and motivation for the work is given, and the organisation of the thesis is outlined.

Chapter 2 gives an overview of age-structured stock assessment models and the model used at present for assessment of the sandeel stock in the North Sea.

In chapter 3 the uncertainty associated with one of the most important sources of information in age-structered assessments is investigated, viz. the age composition estimates of the catches. The importance of various possible sources of uncertainty is evaluated and estimators of uncertainties of age compositions are provided. Age composition estimates are derived from samples taken at random from the catch or by a stratified sampling scheme. An analysis of the importance of various factors is impeded by the structure of the response, which may be considered ordered categorical. A new method to analyse such data is presented.

In chapter 4 a method for analysing the uncertainties associated with the age composition under stratification on length groups is presented. Stratifying on length groups is a common approach to reduce the number of age determinations, as age determinations often are time-consuming and expensive to perform. Instead simple measurements of the lengths are made and the correlation between age and length is utilised.

In chapter 5 the accumulated uncertainty of catch at age is assessed utilising the results from chapter 3. Besides uncertainty of the age composition, catch at age also is influenced by uncertainty of the catch per area and the species composition of the catches, and uncertainties associated with the transformation of the unit of measurement of the size of the catch from tonnes to numbers.

In chapter 6 it is argued that an approach utilising stochastic differential equations might be advantegous in fish stoch assessments.

Chapter 7 contains conclusions.

Appendix A to C contain the papers 'Using continuation-ratio logits to analyse the variation of the age-composition of fish catches' (Kvist *et al.*, 1998), 'Sources of variation in the age composition of sandeel landings' (Kvist *et al.*, 1999a), and 'Uncertainty of Catch at Age Data for Sandeel' (Kvist *et al.*, 1999b).

Appendix D contains plots referred to in chapter 4: 'Analysis of age composition stratified by length groups'.

Chapter 2

Age-structured Stock Assessment Models

2.1 Introduction and overview

Age-structured stock assessment methods constitute the primary basis for providing management advice in many world fisheries, because the population dynamics of exploited fish stocks may be reconstructed and vital mortality rates and absolute abundances may be provided (Megrey, 1989). A historical overview of the age-structured methods is given by Megrey (1989). The first part of the overview presented here, relies to a large extent on his work.

Age-structured stock assessment methods can be traced back to the beginning of the 19th century (Ricker, 1971). The basic idea was to consider a stock as consisting of cohorts. A cohort is constituted by fish of the same species, spawned in the same year and area. By use of catch per age group and year, the size of a cohort at the time the cohort enters the exploitable phase may be reconstructed by simply adding the catches removed from that cohort during the years it has contributed to the fishery. This provides an estimate of the population that must have been alive in order to generate the catches observed. The estimated stock size from these calculations is the minimum stock size, and the quantity is often referred to as the 'utilised stock' because it does not include fish that die for other reasons than fishery. In order to get a more realistic estimate of the stock abundance, models were developed to include "mortality" caused by other reasons (natural mortality) (Beverton and Holt, 1957; Paloheimo, 1958). They also included effort data to model the fishing mortality as a product of fishing effort and catchability.

In models incorporating natural mortality, the main assumption is that removals from a cohort is proportional to the number of alive individuals from that cohort:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = -zN(t) \tag{2.1}$$

with the solution

$$N(t) = N(0)\exp(-zt) \tag{2.2}$$

where $N(t)$ denotes the number of individuals in the cohort at time t and $z = f + m$ denotes the mortality constituted by two components; f and m. f stands for the fishing mortality, comprehending all deaths caused by fishing and m stands for the natural mortality comprehending all other deaths, such as deaths caused by predation, disease, old age etc.. The natural mortality is difficult to estimate due to lack of data. It mostly has to be inferred from investigations on similar species elsewhere and it is often assumed to be constant through the years. However, natural mortality has been shown to vary with age, density, disease, parasites, food supply, predator abundance, water temperature, fishing pressure, sex and size. Attempts are made to estimate the natural mortality by eg. mark-recapture data and stomach-content analyses or by deriving analytical relationships with quantities such as maximum age, length and weight, growth rate and age at sexual maturity (an overview is given by Vetter (1988)).

The number of individuals fished from the cohort constitutes the basic observation for estimation of the stock size, $C(t)$:

$$C(t) = \frac{f}{z}N(0)(1 - \exp(-zt)) \tag{2.3}$$

The parameters f and m are assumed to be constant within a time period, often a year and therefore equation (2.1) is broken down into intervals within which the parameters are assumed to be constant. The equation to connect the number of individuals in two subsequent intervals is called the stock equation (here the length of the interval is a year):

$$N_{a+1,y+1} = \exp(-z_{a,y})N_{a,y} \tag{2.4}$$

$N_{a,y}$ denotes the number of individuals of age a at the beginning of year y and $z_{a,y} = f_{a,y} + m_{a,y}$. The corresponding equation for the number of individuals of age a fished in year y, $C_{a,y}$ is:

$$C_{a,y} = \frac{f_{a,y}}{z_{a,y}} N_{a,y}(1 - \exp(-z_{a,y})) \tag{2.5}$$

The two equations, the stock size equation (2.4), and the catch equation (2.5) are fundamental in the age structured assessments. By using these equations, the historical stock abundancies may be reconstructed. However, some additional information is needed. Gulland (1965) suggested a backwards solution of the equations. At first, the fishing mortality for the oldest fish and the last year, Y, are needed. Thereafter, $N_{a,Y}$, the number of individuals of age a at the beginning of the last year where catch at age data exist, Y, may be calculated by the catch equation, (2.5), as all other quantities are known; the catch at age, $C_{a,Y}$, natural mortality, $m_{a,Y}$, and the fishing mortality, $f_{a,Y}$. The fishing mortality in the previous year, $f_{a-1,Y-1}$, may hereafter be estimated from the catch equation for $C_{a-1,y-1}$, by substituting $N_{a-1,y-1}$ with $N_{a,y}\exp(z_{a-1,y-1})$ from the stock equation, (2.4):

$$C_{a-1,y-1} = \frac{f_{a-1,y-1}}{z_{a-1,y-1}} N_{a,y}\exp(z_{a-1,y-1})(1 - \exp(-z_{a-1,y-1})) \tag{2.6}$$

However, an iterative procedure is required. At this point the algorithm starts over again at the previous step and calculates the number of individuals at the beginning of the second last year using the stock equation (2.4), etc.. This approach and its similarities are often referred to as virtual population analysis (VPA), although the term first was used by Fry (1957) to describe the utilised stock, i.e. corresponding to setting the natural mortality to zero in the catch and stock equation. Murphy (1965) suggested a catch-ratio model, where the catches were defined as in (2.5). The ratios of catches from the same cohort in two successive periods were considered, such that a density-independent model was obtained. The linked system of equations was solved by an iterative procedure which is similar to the one which was used in Gulland's (1965) model. Equation (2.6) is time-consuming and difficult to solve without proper computer facilities and therefore Pope (1972) suggested an approximation that greatly simplified the computations and became very widespread. The approximation was based on the assumption that all fish caught in any age group are taken exactly half way through the year.

The robustness of Pope's approximation (Pope, 1972) and other VPA-like models (Gulland, 1965; Murphy, 1965) have been investigated by eg. Jones (1981), Pope (1972), Agger (1973), and Ulltang (1977). They found that the methods were relatively robust towards errors in the starting guesses of the fishing mortalities and seasonal trends in the mortalities, but that the bias of the fishing mortality would be appr. 25% if the natural mortality is known with a mean error of 0.1.

The obvious drawbacks of the deterministic models are that they are heavily parametrised; they contain more parameters than observations. Thus the estimates are extremely dependent on the data and no uncertainties can be estimated. In addition, cohorts are not linked, i.e. each cohort is analysed separately. Parameter values estimated from one cohort are in no way related to those from other cohorts in the population. In order to reduce the number of parameters and utilise a presumed common structure of the fishing mortality for the cohorts, a separability assumption was introduced. The idea is that the fishing mortality, $f_{a,y}$, of a-year-olds in year y may be described by the product of two factors; a time-dependent factor describing the variation in fishing effort between years, and an age-dependent factor describing the selectivity of age groups (Agger *et al.*, 1971). This reduces the number of parameters dramatically, and the parameters are statistically estimated in a simultaneous manner rather than sequentially, by minimizing the squared difference between observed and predicted catch observations. At the same time, a separability assumption simultaneously link data from several cohorts. Introduction of the separable formulation of fishing mortality was an important conceptual advance because it moved the study of stock assessment methodology into the realm of more generalised mathematical models and went a long way toward promoting statistical analysis of catch at age data (Megrey, 1989). However, the separability assumption does not always hold. The selectivity of the age groups may change with time because of technological developments of the gear. The assumption may also be violated in cases where the stocks are exploited by more than one fleet, using fishing gear with different selectivity, if the relative proportion caught by each fleet changes. The problem can sometimes be avoided by working with catch at age data disaggregated by fleets.

Despite the advantages of the separability assumption, it did not overcome the problem that catch at age data alone do not contain enough information to estimate fishing mortality in the most recent fishing year with acceptable precision. In addition, stock sizes and fishing mortality parameters become highly negatively correlated when based on catch at age data alone (Doubleday, 1976; Pope, 1977). Therefore various approaches often referred to as 'tuning virtual population analyses' or 'integrated analysis' have been attempted, where auxiliary information in terms of additional data or assumed relationships which restrict the model has been introduced in the stock assessments. Such information could be catch per unit effort (CPUE) data assumed to be an index of abundance either estimated on the basis of data from the fishing vessels (Pope and Shepherd, 1985) or from research vessels (Doubleday, 1981), relationships between spawners and recruits (eg. Ricker, 1954; Beverton and Holt, 1957) or more or less complicated models for the catchability, under the assumption that the fishing mortality may be described as a product of the effort and the catchability. Lewy (1988) utilised the following model for catchability in the assessment of the North Sea whiting stock:

$$q_{\varphi,a,y} = s_{\varphi,a,y}\ q_{1,\varphi,y}\ q_{2,\varphi,a} \tag{2.7}$$

where one relationship is determined for every fleet, φ. $q_{1,\varphi,y}$ is a technological factor accounting for development of fishing power. $s_{\varphi,a,y}$ is the selectivity defined as the proportion of fish retained in the trawl modelled as a function of age or length of the fish and the mesh size. $q_{2,\varphi,y}$ is an age factor dependent on availability and behaviour of the fish. The performance of some tuning methods are compared by Pope and Shepherd (1985).

Fournier and Archibald (1982) and Deriso *et al.* (1985) proposed very generalised mathematical models incorporating the separability assumption. The models allow incorporation of fishery-independent data directly into the simultaneous parameter estimation procedure. Unfortunately, neither of those models can encompass both measurement error of catch at age data and model errors at the same time. However, introducing time series models one obtains the capability to account for model errors with their cumulative properties as well as observation errors. The models may be estimated by the use of Kalman filter techniques. Age structured time series models for fish stocks have been described in eg. Kettunen (1983) and Schnute (1994), and applied by eg. Mendelssohn (1988), Gudmundsson (1994), Fargo and Richards (1998). The models are discretised into intervals of years before a term accounting for model errors is entered. Gudmundsson (1994) discretises (2.2) and (2.3) into (2.4) and (2.5) before terms are introduced to take the uncertainties into account. The fishing mortality is modelled as a state variable by a separable model allowing for four sources of random variation having transcient and permanent influence on the fishing mortality. Gudmundsson (1994) assumes that the natural mortality is known and that the recruitment is varying around a constant level, described by a Gaussian distribution and thus is independent of the amount of sexually mature fish. Recently Bayesian approaches have been applied (McAllister and Ianelli, 1997; Punt and Hilborn, 1997). The approaches are attractive because rather complicated relationships may be fairly easy to describe and because prior knowledge of the distribution of the parameters may be provided, either by 'expert' knowledge, by historical data or by results from assessments from other stocks. The output is a distribution of the model parameters. However, care must be taken when interpreting the results. In particular, selection of priors designed to be noninformative with respect to quantities of interest is problematic (Punt and Hilborn, 1997).

The overview above has focused on age-structured models for stock assessment as these are most relevant to the project. Another main approach to stock assessment is models based on the length composition of the catch instead of the age composition. E.g. Ralston and Ianelli (1998) give an example of a species (Bocaccio), where the age determination is so difficult that the estimates of the age-composition is too uncertain to be of any use. Instead length composition data was used. Quinn *et al.* (1998) and Matsuishi (1998) also discuss length-based population analyses. Sullivan (1992) presents a state-space model of a length structured population under commercial harvest. A Kalman filter is used for estimation. Approaches for stock assessment based on catch-effort data alone, i.e. where the data consists of annual aggregated catches and annual aggregated fishing effort, could be another supplement to the catch at age based sandeel stock assessments when age determinations are too uncertain. Such models are presented by eg. Chen and Paloheimo (1994) and Reed and Simons (1996).

2.2 Stock assessment of sandeel

The sandeel stock in the North Sea is assessed every year by the International Council of the Exploration of the Sea (ICES). The method used is called Seasonal Extended Survivors Analysis (SXSA) (Skagen, 1994), which is a modification of Extended survivors analysis (XSA) (Doubleday, 1981). The name is apt because the method focuses on the estimation of the abundance of the survivors at the end of the period covered by the catch data, for each cohort. Most other VPA-like methods estimate the stock size at the beginning of the years where catch at age data is available. Thus, the stock size at the end of the last year, which often is of great importance for the fishery management, is not assessed in the algorithm but derived from the fishing mortality for the last year. The term 'seasonal' in SXSA refers to the fact that constant fishing mortality is assumed in periods of half years, in stead of whole years, due to the seasonal characteristic of the fishery. In general, the fishery peaks during spring and early summer.

The method is a 'tuning' of Pope's approximation to VPA (Pope, 1972) by additional measures of relative stock abundance, CPUE data. In the approximation it is assumed that the entire catch is taken exactly midway through the period. Thus, the number of survivors at the end of a period, which is the same as the the number of individuals at the beginning of the following period, $N_{a'+1,y'+1}$ is:

$$N_{a'+1,y'+1} = N_{a',y'} \exp(-m_{a'}) - C_{a',y'} \exp(-m_{a'}/2) \tag{2.8}$$

where a' and y' denotes the age and year, counted in half years; $C_{a',y'}$ is catch at age data and $m_{a',y'}$ is the natural mortality, which is estimated to be higher for the youngest fish and generally lower in the second half of the year, because the sandeel hides in the sediment in the winter period. The natural mortality for $\frac{1}{2} - 1$-year-olds is assumed to be 0.8 and for 1-$1\frac{1}{2}$-year-olds it is assumed to be 1.0. For older fish the natural mortality is assumed to be 0.4 in the first half of the year and 0.2 in the second half of the year. From equation (2.8) the survivors each period may be expressed easily as a function of the survivors from the previous period.

The 'tuning' is done on the basis of CPUE data, $u_{a',y'}$. It is assumed that $u_{a',y'}$ is proportional to the mean number of individuals $\overline{N}$ of that age group in the period:

$$u_{a',y'} = q_{a'}\overline{N} \tag{2.9}$$

where the catchability, $q_{a'}$, is dependent on the age only. Thus, the CPUE data is based also on information on the population size. Estimates of catchability and survivors are obtained by a trade-off between the two sources of information (the CPUE and the catch at age data), by means of a least square approach. The estimates associated with older fish and with the second half of the year are assumed to be more uncertain than other estimates. To give the observations an influence in accordance with these expectations a manual weighting in the estimation has been introduced.

2.3 Sources of uncertainties in the assessment of sandeel

2.3.1 Uncertainties associated with the model

The models of fish stocks are of course only crude approximations of the actual population dynamics. When the system is complex and informative observations are difficult to obtain it cannot be otherwise. However, the characteristics of sandeel make the results even less reliable compared to most other species. The natural mortality of sandeel is much higher than for most other commercially exploited fish species. As the sandeel is a short-lived species only a few observations are obtained per cohort. A reliable model for recruitment is difficult to obtain, as no clear dependency can be recognized between the spawning stock biomass and the number of 0-year-olds (figure 2.1). The spawning stock is the weight of the individuals in the stock that is at least two years old; the estimated age of maturity of sandeel (Macer, 1966).

Fishery independent data such as survey data are not available to improve the assessments, because sandeel is not caught by the standard equipment on the survey vessels. Adult sandeels bury themselves in the sediment at night and during winter and are mostly found in areas of coarse well-oxygenated sand, (Macer, 1966). Presumably, there is little migration of

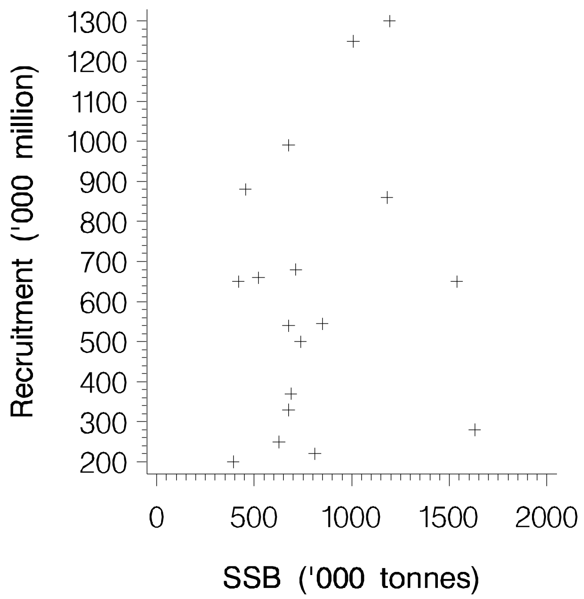

Figure 2.1: Abundance of 0-year-olds versus spawning stock biomass.

adult sandeel between the various sandeel grounds in the North Sea, and regional differences in age-composition can therefore be expected. In addition, there are indications that the fishery can direct their fishery towards particular cohorts. Thus, the separability assumption is questionable. Due to the burrowing behaviour of the adults the catch rates vary between different age groups, with season and during the day (Reeves, 1994). The burrowing behaviour may also depend on tide and weather.

2.3.2 Uncertainties associated with the estimation procedure

The estimation procedure is of course closely connected to the model. The scope for development of models is often restricted by the estimation procedure and the evaluation of a model depends on the properties of the procedure. Therefore, it has been considered irrelevant to evaluate the sources of errors in connection with the estimation of the SXSA in this project, as the aim is to improve the model. This subject will be addressed whenever relevant in the thesis, i.e. in connection with estimates of the age composition of catches (chapter 3, section 3.3.1) and discussion of fish stock models (chapter 6, section 6.1).

2.3.3 Uncertainties associated with the data

The data utilised in the assessment of sandeel is catch at age data and CPUE data. The data are far from direct observations even though they are referred to as such. They are the result of a combination of information from several different data sources. The first hand buyers report the weight of each industrial landing that is bought. The fishermen on vessels with an overall length of 17 m or more report daily in their logbooks information on what, where, when and how much they have caught. In addition information on the vessel size and gear is reported. The authorities collect samples to estimate the catch compositions with regard to the species composition, the age composition and the distributions of weight and length. Of course all those observations presumably have some kind of observational error associated. The uncertainties are addressed in connection with the assessment of the uncertainties in the catch at age data (Kvist *et al.*, 1999b) (Appendix C).

The dominating sources of uncertainties in catch at age data are associated with the age composition, discussed in chapter 3. The procedure for estimation of catch at age data and their associated uncertainties is described in chapter 5.

The CPUE data is estimated on the basis of information from the logbooks. The number of fishing days by vessel category are estimated by counting the number of days where the logbook indicates that sandeel constituted more than 70% of the total daily catch. Seven categories of vessel sizes are used. The corresponding total catch of sandel for each vessel category is estimated as the sum of the logbook estimate of the sandeel catch. In each vessel category the mean catch per fishing day, season and year is estimated. In order to account for differences in fishing power between year, season and vessel size, the following model was fitted to the data:

$$\mathrm{CPUE}_{y,seas,cat} = a_{y,seas}\ \mathrm{V}_{cat}{}^{b_{y,seas}} \tag{2.10}$$

where y denotes year, $seas$, indicates winter and summer season, V_{cat} is the mean vessel size of the category cat and $a_{y,seas}$ and $b_{y,seas}$ are the parameters to be estimated. By this procedure an estimate of the CPUE of a vessel of standard size is provided.

The fishing effort, i.e. the number of standardised fishing days, per season and year may be obtained by dividing the catch of sandeel by the standardised CPUE. However, the information on the amount caught recorded by the fishermen is more impresise than the information from the first hand buyer, where the catch has actually been weighed and not appreciated by eye, an eye which might be prone to underestimate. Thus, an improved estimate is obtained by using the catch per species estimated on the basis of a combination of the information from the logbooks, the first hand buyer and the samples taken by the authorities. How this is obtained is described in Kvist *et al.* (1999b) (Appendix C).

An analysis of the possible sources of uncertainties of the CPUE data has not been performed in this project. This could be done by choosing the catch of sandeel per fishing day where sandeel indeed was the target species, as response. Thus, the raw data is used and not an average of all fishing days in a season per vessel group. By this approach no uncertainties are aggregated and therefore sources of uncertainties and their magnitudes are easier to assess. The response could then be modelled as a function of various possible factors or functions of those; vessel size, year, time of year, geographical position, gear, mesh size. Presumably, the distribution of the response may be satisfactorily described by a continuous distribution belonging to an exponential family or perhaps a log-normal distribution, i.e. that the logarithm to the response is normally distributed. It is, however, not trivial to determine in each case whether the target species was sandeel or not because it may happen that the catch is different from the target. Currently, the definition of a fishing day where sandeel was the target species is a day where more than 70% of the catch is constituted by sandeel.

Richards and Schnute (1992) also suggest some methods for analysing CPUE data. However, they focus on transformation parameters for normalising the response instead of utilising the theory of generalised linear models (McCullagh and Nelder, 1989), which can handle other distributions than the normal distribution.

Chapter 3

Uncertainties Associated with the Estimated Age Composition

3.1 Introduction

The age composition of the catch provides vital information for age-structured stock assessments. In addition, the age composition of the catch gives a picture of the age composition of the part of the stock that is available to the fishery, although the picture presumably is biased. Estimates of the age composition are derived from samples taken at random from the catch or by a stratified sampling scheme. The catch samples are sorted into species, the number of individuals of each species is counted, and the individuals are measured and their age determined by counting the number of growth rings in hard parts such as otoliths. The age composition may vary from sample to sample due to a multitude of factors including spatial or temporal differences in catch composition and errors in the age determination itself. By breaking down the variation into its original sources, improved estimates of the age compositions and their uncertainties, and valuable information concerning the stock dynamics may be obtained. Furthermore, the gained knowledge of the sources may be used to optimise the sampling scheme under stratified sampling. Thus, it is certainly useful to detect the sources and magnitudes of variation in the age composition. However, this subject has seldom been addressed. This may partly be due to a lack of suitable methods. The distribution of the number of individuals in different age groups in a sample may be described by a multinomial distribution and no standard methods are available for evaluating the significance of factors influencing such a distribution. A new method for analysing age composition data has been presented in Kvist *et al.* (1998) (Appendix A). The method combines continuation-ratio logits (Agresti, 1990) and the theory for generalised linear mixed models (Breslow and Clayton, 1993; Wolfinger and O'Connel, 1993). It transforms the probability of the multinomial response into a product of binomial probabilities for which generalised linear mixed models can be directly applied to study the possible sources of variation. It is particularly suitable for age composition data because it allows individual cohorts to be followed and compared over time.

3.2 Transforming the multinomial response probability into a product of binomial probabilities

An important issue of the assessment of the uncertainties associated with the age composition is to establish factors of importance for the age composition. This may be done by modelling the age composition as a function of various possible factors and testing the significance of these. The response is the index of individuals in each age group, $\mathbf{X}_s = (X_{Rs}, \ldots, X_{As})$, where s denotes the sample number, and the age groups are $R, \ldots, A$. The number of the age group usually corresponds to the age it covers, except for age group A, which most often covers ages A and above. R stands for recruitment age, which is the youngest age group that appears in the landings. If we assume that the age composition of the species of interest in a particular sample does not depend on the occurence of other species in the sample and that the samples are representative for the age composition in the catch then the response may be modelled by a multinomial distribution:

$$\mathbf{X}_s \in \text{Mult}(n_s, p_{Rs}, \ldots, p_{As}) \tag{3.1}$$

where n_s denotes the sample size and p_{js} denotes the proportion of individuals in the catch classified as belonging to age group j, $j = R, \ldots, A$. With $A-R+1$ age groups present $A-R$ probabilities will be needed to describe the distribution. The p_{js}'s describe the real age composition of the catches if the age determination is unbiased. If a bias exists, the proportion p_{js} describes the proportion of the species in the catch that would be classified into age group j. In order to analyse the age composition, the multinomial probability is factorised into a product of binomial probabilities. This is done by considering the conditional distributions of $X_{Rs}, \ldots, X_{(A-1)s}$, where the distribution of X_{js} is conditioned on the event that the age is j or higher:

$$X_{js}|X_{js} + \ldots + X_{As} = sum_{js} \in \mathrm{Bin}(sum_{js}, \pi_{js}) \tag{3.2}$$

where $j = R, \ldots, A-1$; X_{js} is the number of j-year-olds and π_{js} is the probability of age j given that the age is at least j:

$$\pi_{js} = \frac{p_{js}}{p_{js} + \ldots + p_{As}} \tag{3.3}$$

Thus, the probability of the multinomial response, $\mathbf{X}_s$, of dimension $A-R$ is transformed into a product of $A-R$ binomial probabilities. The ordinary logits associated with the π_{js}'s:

$$L_{js} = \log \frac{\pi_{js}}{1 - \pi_{js}} \tag{3.4}$$

are called continuation-ratio logits for $\mathbf{X}_s$, because such a logit compares the proportion of an age group to the proportion of older age groups, which becomes obvious if the conditional probabilities in (3.4) are substituted by the unconditional probabilities, i.e. (3.4) equals:

$$L_{js} = \log \frac{p_{js}}{p_{(j+1)s} + \ldots + p_{As}} \tag{3.5}$$

3.3 Analysis of the transformed probability

The factorisation of the multinomial probability has the advantage that the conditioned probabilities may be modelled separately by means of generalised linear mixed models as long as they do not have any parameters in common. Thus, the distribution of the multinomial response, $\mathbf{X}_s$, of dimension $A - R$ is modelled by $A - R$ continuation-ratio logits of the form:

$$L_j = \mathbf{b}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{u}_j \tag{3.6}$$

where $j = R, \ldots, A-1$. $\mathbf{b}_j$ denotes the explanatory variables associated with the fixed parameters $\boldsymbol{\beta}_j$ and $\mathbf{Z}_j$ the explanatory variables associated with the random parameters $\mathbf{u}_j$. The random parameters are assumed to be normally distributed on the logit scale. If the random parameters are omitted the model is a generalised linear model, described in McCullagh and Nelder (1989).

A dispersion parameter, ϕ, is included to account for the variance that could not be attributed to the binomial variance or the explanatory variables. The dispersion parameter enters as a simple multiplicative factor on the binomial variance, and must therefore be greater than zero. $\phi = 1$ indicates that the variance of the response is in accordance with the nominal binomial variance. $\phi < 1$ indicates that the data is underdispersed, and that the variance of the response is less than the nominal binomial variance. $\phi > 1$ indicates overdispersion, where the variance of the responce exceeds the nominal binomial variance. Introducing a dispersion parameter means that the conditional distributions are no longer exactly binomial:

$$X_{js}|x_{js} + \ldots + x_{As} \in \widetilde{\mathrm{Bin}}(X_{js} + \ldots + X_{As}, \pi_{js}, \phi_{js}) \tag{3.7}$$

The dispersion parameter has been described in more detail in e.g. McCullagh and Nelder (1989).

The excess random variation of the model is thus modelled partly by a dispersion parameter and partly by variance components (from random effects). A variance component describes variation between observations with different probabilities and the dispersion parameter describes variation between observations with the same probabilites. The magnitudes of the two are difficult to compare as they are measured on different scales, but the interpretation of the dispersion parameter and the variance components may be illustrated further by considering the following simple example.

Example

Assume X is binomially distributed with an associated dispersion parameter, ϕ:

$$X \in \widetilde{\mathrm{Bin}}(n, p, \phi) \tag{3.8}$$

where $\mathrm{E}[p] = p_0$, $l = \log(p/(1-p))$, and $\mathrm{V}[l] = \sigma^2$.

The variance of the observation X/n can then approximately be expressed as:

$$\mathrm{V}\left[\frac{X}{n}\right] \approx p_0(1-p_0)\left[p_0(1-p_0)\sigma^2 + \frac{\phi}{n}\left(1 - p_0(1-p_0)\sigma^2\right)\right] \tag{3.9}$$

The first factor of the expression describes the basic binomial variance structure. The first term within the square brackets describes the variation between observations with different p's (transformed from the logit scale to the probability scale), and the last term describes the average variation between observations with the same p (because of the convexity of $p(1-p)$ this average variation will be less than $\phi p_0(1-p_0)$). Note that if the variance component,σ, is zero the variance reduces to the variance, $p_0(1-p_0)\phi/n$, corresponding to a binomial distribution with a dispersion parameter. Note also that according to (3.9), an increase of the sample size will reduce the contribution from the dispersion parameter, but not the contribution from the random effect.

The choice between modelling an effect as fixed or random depends on the purpose of the model and the nature of the effect. In the three examples presented in the papers Kvist *et al.*, (1998) (Appendix A); Kvist *et al.*, (1999a) (Appendix B) and Kvist *et al.*, (1999b) (Appendix C), three different models have been introduced for the same response, where the structure depends on the purpose of the model. Each model has been motivated in the respective paper. In particular it has been found advantageous, to model geographical effects as random, because modelling an effect as random implies a dependency between the observations, in the present cases between age compositions within the same area. Thus, age compositions in subareas within the same area are correlated. The dependency also has the advantage that the age composition of areas with no samples and the associated variance can be estimated, simply by estimating it to be the average age composition within the larger geographical area it belongs to. Another important advantage is that even though there was a significant variation of the smallest possible areas available in the analysis, the significance of larger geographical areas could be evaluated.

The primary interest of a random effect is often the magnitude of the associated variance component. However, estimates of the effects of the separate levels of the normally distributed variable, may also be of interest. An estimate of the effect at a particular level of a random effect is determined as a compromise between the specific observations associated with that level and the average effect. Usually this estimate is chosen as the Best Linear Unbiased Predictor (BLUP) (e.g. Robinson, 1991). There is some confusion in the terminology regarding to whether an estimate of the effect at a level of a random effect should be called estimate or predictor (Robinson, 1991).

Continuation-ratio logits are particularly suitable for analysing age composition data because they allow individual cohorts to be followed and compared over time. By considering the difference between logits associated with the same cohort from two succesive years, $L_{y,a,c}$ and $L_{y+1,a+1,c}$ the relative development of the cohorts may be studied. A different indexing than (3.5) has been used in order to emphasize that it is the same cohorts that are compared; y denotes year, a denotes age and $c = y - a$ denotes the cohort. I.e. the logits to be compared are

$$L_{y,a,c} = \log \frac{p_{y,a,c}}{p_{y,a+1,c-1} + p_{y,a+2,c-2} + \dots} \tag{3.10}$$

and

$$L_{y+1,a+1,c} = \log \frac{p_{y+1,a+1,c}}{p_{y+1,a+2,c-1} + p_{y+1,a+3,c-2} + \dots} \tag{3.11}$$

As the indices indicate, the same cohorts are compared to each other. Such a comparison has been performed based on age composition data from the sandeel fishery (Kvist *et al.*, 1999a) (Appendix B). The analyses showed that the age proportion of a cohort changed through the years, even though compared to the same cohorts, viz. older. Apparently, the fishery for 0-year-olds at the present fishing intensity does not influence the fishing possibilities of 1-year-olds the year after. Another conclusion was that there is indication of that the fishery has been attracted to 1-year-old fish in years where they were abundant. In addition, a pattern could be recognized to be utilised for inference and prediction.

The transformed probability of the multinomial response was modelled by generalized linear mixed models, where the random effects were modelled as normally distributed on the logit scale. Another promising approach has been suggested by Lee and Nelder (1996). They present models even more generalized, which may handle random components of other distributions than the normal distribution. Thus, more natural distributions of the random components may be applied. In the application presented here, where the multinomial probability is transformed into a product of binomial probabilities, the beta-distribution may be applied. This alternative approach may very well result in improved estimates of the age composition and its uncertainties, as the beta-distribution presumably describes the random variation of probabilities better than the normal distribution on the logit scale. If however, the variance components are small the two will result in approximately the same results. Unfortunately, the random components are large in the actual case studied. Therefore, the new approach presented by Lee and Nelder (1996), might be beneficial.

3.3.1 Estimation of generalised linear mixed models

At present, there is not any procedures available in the statistical software packages to fit the generalised linear mixed models using exact maximum likelihood. The estimation procedure utilised in this project is a procedure suggested by Wolfinger and O'Connel (1993) and implemented in the macro glimmix in SAS 6.12. It is an approximate method combining two analytical and one probabilistic approximation. First, consider the generalised linear mixed models formulated as:

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e} \tag{3.12}$$

where $\mathbf{y}$ is a vector containing n observations, $\boldsymbol{\mu}$ is the mean defined by a link function which should be monotonic and differentiable:

$$g(\boldsymbol{\mu}) = \mathbf{b}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \tag{3.13}$$

Thus the mean is a nonlinear function of the explanatory variables. $\boldsymbol{\beta}$ is a vector of unknown fixed effects with known model matrix $\mathbf{b}$ and $\mathbf{u}$ is a vector of unknown random effects with known model matrix $\mathbf{Z}$. $\mathrm{E}\{\mathbf{u}\} = \mathbf{0}$ and $\mathrm{Cov}\{\mathbf{u}\} = \mathbf{G}$, where $\mathbf{G}$ is unknown. $\boldsymbol{e}$ is a vector of unobserved errors with

$$\mathrm{E}\{\mathrm{e}|\boldsymbol{\mu}\} = \mathbf{0} \tag{3.14}$$

and

$$\mathrm{Cov}\{\mathrm{e}|\boldsymbol{\mu}\} = \mathbf{R}_{\boldsymbol{\mu}}^{1/2}\mathbf{R}\mathbf{R}_{\boldsymbol{\mu}}^{1/2} \tag{3.15}$$

where $\mathbf{R}_{\boldsymbol{\mu}}$ is a diagonal matrix containing evaluations at $\boldsymbol{\mu}$ of a known variance function for the model under consideration and $\mathbf{R}$ is unknown. The first analytical approximation is to approximate $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}$ by a first order Taylor series approximation expanding about $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{u}}$:

$$\tilde{\mathrm{e}} = \mathbf{y} - \widehat{\boldsymbol{\mu}} - (\mathrm{g}^{-1})'(\mathbf{b}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}})(\mathbf{b}\boldsymbol{\beta} - \mathbf{b}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\mathbf{u} - \mathbf{Z}\widehat{\mathbf{u}}) \tag{3.16}$$

where $\widehat{\boldsymbol{\mu}} = \mathrm{g}^{-1}(\mathbf{b}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{u}})$ is a diagonal matrix with elements consisting of evaluations of the first derivative of g^{-1}. Hereafter the conditional distribution of $\tilde{\mathrm{e}}$ given $\boldsymbol{\beta}$ and $\mathbf{u}$ is approximated with a Gaussian distribution:

$$\tilde{\mathrm{e}}|\boldsymbol{\beta}, \mathbf{u} \in \mathrm{N}(\mathbf{0}, \mathbf{R}_{\boldsymbol{\mu}}^{1/2}\mathbf{R}\mathbf{R}_{\boldsymbol{\mu}}^{1/2}) \tag{3.17}$$

At last $\boldsymbol{\mu}$ is substituted by $\widehat{\boldsymbol{\mu}}$ in the variance matrix. The approximations result in model equations similar to those of ordinary mixed models. The mixed model equations are:

$$\begin{bmatrix} \mathbf{b}'\mathbf{R}^{-1}\mathbf{b} & \mathbf{b}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{b} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{b}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3.18}$$

and the generalised linear mixed model equations are:

$$\begin{bmatrix} \mathbf{b}'\mathbf{W}\mathbf{b} & \mathbf{b}'\mathbf{W}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}\mathbf{b} & \mathbf{Z}'\mathbf{W}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{b}'\mathbf{W}\mathbf{y}^* \\ \mathbf{Z}'\mathbf{W}\mathbf{y}^* \end{bmatrix} \tag{3.19}$$

where

$\mathbf{W} = \mathbf{D}\mathbf{R}^{-1}\mathbf{D}$
$\mathbf{y}^* = \widehat{\boldsymbol{\eta}} + (\mathbf{y} - \widehat{\boldsymbol{\mu}})\mathbf{D}^{-1}$
$\mathbf{D} = [\partial\boldsymbol{\mu}/\partial\boldsymbol{\eta}]$
$\mathbf{R} = \mathrm{Var}\{\mathbf{e}\}$
$\boldsymbol{\eta} = \mathbf{b}\beta + \mathbf{Z}\mathbf{u}$

The differences between the ordinary mixed model equations and the generalised linear mixed model equations are two transformations; one of the observations, $\mathbf{y}$, and one of the weighting matrix, $\mathbf{R}^{-1}$. The transformations are necessary because whereas observations and paramaters from an ordinary mixed model are measured on the same scale, observations and parameters from generalised linear mixed models are not necessarily measured on the same scale. Thus, to obtain a solution for the fixed and random parameters, $\boldsymbol{\beta}$ and $\mathbf{u}$, the observations utilised in the generalised linear mixed models are transformed into observations, $\mathbf{y}^*$, on the scale where the fixed and random parameters are measured. The weighting matrix $\mathbf{R}^{-1}$ is replaced by $\mathbf{D}\mathbf{R}^{-1}\mathbf{D}$, again to make a transformation into the scale of the fixed and random parameters. The procedure is presented in detail in Wolfinger and O'Connel (1993). Unfortunately, approximate maximum likelihood estimates of this kind have some unsatisfactory properties. In particular, the variance of the predictions of the separate levels of a random effect (reffered to as Best Linear Unbiased Predictor (BLUP)) is biased and underestimated under standard (small domain) asymptotic assumptions especially if the variance components are not small (Kuk (1995), Lin and Breslow (1996), Breslow and Lin (1995) and Booth and Hobert (1998)).

Attempts are made for developing procedures for finding exact maximum likelihood estimates in the generalised linear mixed models setting, e.g. Booth and Hobert (1999). They suggest two methods based on the Monte Carlo EM algorithm (Wei and Tanner, 1990). However, the methods break down when the intractable integrals in the likelihood function are of high dimension. Booth and Hobert (1999) suggest that approximate methods such as those implemented in the macro glimmix should be used for model selection until the exact methods have been improved.

3.4 Estimation of the age composition and the associated uncertainties

By application of generalised linear mixed models, estimates of the conditional probabilities, $\widehat{\boldsymbol{\pi}}_s = (\widehat{\pi}_{Rs}, \ldots, \widehat{\pi}_{As})$, and variances and covariances between conditional probabilities for age groups i and i', and samples s and s', $\widehat{\mathrm{Cov}}\{\widehat{\pi}_{is}, \widehat{\pi}_{i's'}\}$, may be obtained. The unconditional probabilities may be obtained by the following equation:

$$p_i = \pi_i(1 - \sum_{j=R}^{i-1} p_j) \tag{3.20}$$

where $i = R, \ldots, A$. Note that the conditional probability for the oldest age group, π_A, equals 1. The variances and covariances are estimated by using a first order Taylor approximation for a product of independent variables:

$$\mathrm{V}\{\prod_{i=1}^{n} \widehat{\pi}_i\} \approx \sum_{i=1}^{n} \left[\mathrm{V}\{\widehat{\pi}_i\} \prod_{j=1}^{i-1} \widehat{\pi}_j^2 \prod_{k=i+1}^{n} \widehat{\pi}_k^2 \right] \tag{3.21}$$

3.5 Uncertainties associated with the age composition of the sandeel landings

The approach described above has been applied to age composition data collected from the Danish sandeel fishery in the North Sea in 1993 in order to illustrate the method (Kvist *et al.*, 1998) (Appendix A). The model was formulated, the significance of effects were tested and estimates of the unconditioned probabilities as well as their variances and covariances were provided (the first order Taylor approximation (3.21) for the case of the sandeel fishery. In the example (Appendix A, section A.3.6), the resulting covariances had similar characteristics as covariances for multinomial data, in the sence that the largest uncertainties were observed for age groups represented in proportions close to $\frac{1}{2}$. However, the uncertainties in the middle age groups varied more smoothly than suggested by a crude multinomial model. Another discrepancy from the covariance matrix of ordinary multinomial data, is that positive covariances were observed.

The sources of variation in the age composition of the sandeel landings taken in the North Sea between 1984 and 1993, have been analysed in order to evaluate the significance of spatial and temporal differences in the age composition of the sandeel samples as well as to study the importance of differences in age readings between laboratories. The analyses and results are presented in Kvist *et al.* (1999a) (Appendix B). The analyses show that the proportion of older sandeel in the catches is significantly lower in the start and end of the fishing season and that the age composition differ between laboratories. There is considerable variation in the age composition within small areas, as well as considerable undetected sources of variation resulting in a large and significant overdispersion.

When the purpose of the model for age composition data changed from detecting important sources of variation to estimating the resulting uncertainties of the age composition, changes in the models were made. Although the logits for the different levels are modelled independently and might have different sources of variation and common sources of variation might be of different magnitude (refer to Kvist *et al.* (1998), Appendix A, section A.2), an overall evaluation of the significant effects are needed in order to evaluate the relevance of a model. E.g. it is difficult to see the sense in a geographical effect of importance for ages 0, 1 and 3, but not for age 2 (refer to Kvist *et al.* (1999a) (Appendix B), table 1). Therefore, when the purpose of the model changed, the model structure was also changed (refer to Kvist *et al.* (1999b) (Appendix C), table 1).

Whether or not predictions/estimates of the effects of the individual levels of the random effects are relevant depends on the purpose of the model. For illustrative purposes it is assumed that the following model applies:

$$L_{ijk} = \log \frac{\pi_{ijk}}{1 - \pi_{ijk}} = o + a_i + s(a)_{j(i)} + l_k \qquad (3.22)$$

where L_{ijk} is the logit for the conditioned probability π_{ijk}; a is a geographical effect covering a larger area, with the parameters a_i, $i = 1, \ldots, 4$, describing the difference between the effect of area i and the overall effect

o. The area a_i is constituted by a set of smaller areas, s_j, $j = 1, \ldots, 4$, where $s(a)_{j(i)}$ describes the difference between the effect of area i and the smaller area j within area i (refer to figure 3.1).

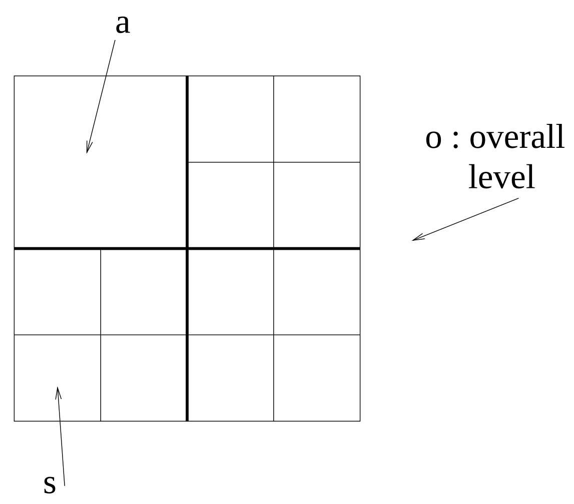

Figure 3.1: An illustrative example of estimation and prediction

l_k, $k = 1, \ldots, K$, is a laboratory effect describing age determination errors. $a_i, s(a)_{j(i)}$ and l_k are modelled as random, with the variance components, $\sigma_a^2, \sigma_{s(a)}^2$, and σ_l^2. It is assumed that $\mathrm{E}\{l_k\} = 0$ corresponds to an overall unbiased classification.

The best estimate of the effect, a_i, in an area i, without observations, is $\widehat{o}$. The variance of this parameter is $\widehat{\mathrm{V}\{\widehat{o}\}}$. However, the variance of the estimate/prediction of the effect of such an area is $\widehat{\mathrm{V}\{\widehat{o}\}} + \widehat{\sigma}_a^2$. The best estimate/prediction of the effect of a small area within that larger area without observations is still $\widehat{o}$, whereas the prediction variance becomes $\widehat{\mathrm{V}\{\widehat{o}\}} + \widehat{\sigma}_a^2 + \widehat{\sigma}_{s(a)}^2$. If observations exist from the larger area, a_i, but not from the smaller area, $s(a)_{(j(i))}$, then the best estimate/prediction of the effect of the smaller area would be $\widehat{o} + \widehat{a}_i$, and the associated variance would be $\widehat{\mathrm{V}\{\widehat{o}\}} + \widehat{\mathrm{V}\{\widehat{a}_i\}} + 2\widehat{\mathrm{Cov}\{\widehat{o}, \widehat{a}_i\}} + \widehat{\sigma}_{s(a)}^2$. Thus, in this example, estimates of the effects of the separate laboratories, l_k, $k = 1, \ldots, K$, were not relevant, whereas the magnitude of the variance component, as well as estimates/predictions of the separate effects of the random effects were relevant for the a and $s(a)$ effects.

To examplify, the predictions in the illustrative example in Kvist *et al.* (1998) (Appendix A, A.3.6), are predictions of the effect in a small area, viz. a square, SQ, without observations, within a larger area, viz. the southern part of the North Sea. The estimates of the catch at age data for years 1989 and 1991 (refer to table 5 in Kvist *et al.* (1999b) (Appendix C)) are based on predictions of the separate effects in each square (i.e. ICES rectangle). In the cases where observations exist from the square, they are utilised to improve the predictions, in cases where they are not, the effect of the level is predicted by the mean effect of the larger area, A, the square belongs to.

3.6 Link between age composition and stock dynamics

In the analysis performed on the age composition data, no assumption of the population dynamics such as (2.1) have been modelled. If such an assumption is made, additional analyses may be performed. To illustrate the analyses, it is assumed that (2.1) holds. Assume further that the age composition in the sample is representative of the age composition in the catch. Then the observed age composition in the sample may be used as an estimate of the age composition in the catch, $p(t)_R, \ldots, p(t)_A$, taken in the instantanous period from t to t':

$$p(t)_i = \frac{\Delta C(t)_i}{\sum_{j=R}^{A} \Delta C(t)_j} \tag{3.23}$$

where $i = R, \ldots, A$ and $\Delta C(t)_i$ is the catch of i-year-olds in the period from t to t':

$$\Delta C(t)_a = C(t')_a - C(t)_a = \frac{f_a}{z_a} N(t)_a (1 - \exp(-z_a(t' - t))) \tag{3.24}$$

The nonlinear description of the population dynamics has the consequence that the age composition will change through the year, unless the age groups have identic population sizes and mortalities. However, if the pattern of removals may be assumed constant through a year (i.e. no change in quantities such as mortalities and availability), this change may be ignored in most cases, as in the example shown in figure 3.2 and figure 3.3. The first example is based on data from the cod fishery in the Icelandic waters (data from Gudmundsson, 1994) and the second example is a constructed situation, where the fishing mortalities are larger and differs more from each other than in the example from the cod fishery.

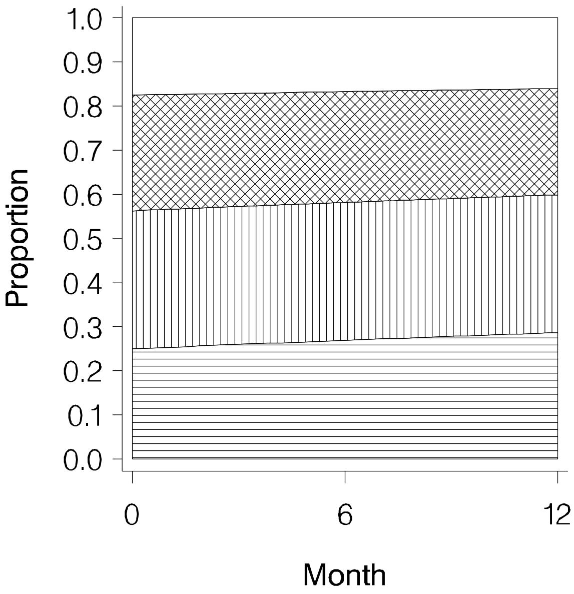

Figure 3.2: Changes of the proportions of the age groups among 4 to 7-year-old cod through a year. Typical fishing mortalities and stock numbers (N) are used for 4- to 7-year-olds in the cod fishery in Icelandic waters. Stock numbers as estimated for the whole North Sea. The horizontal lines corresponds to the proportion of 4-year-olds ($N = 200$, $z = 0.4$). Vertical lines corresponds to proportion of 5-year-olds ($N = 100$, $z = 0.7$). Crosshatched lines corresponds to proportion of 6-year-olds ($N = 60$, $z = 0.9$). White area corresponds to proportion of 7-year-olds ($N = 40$, $z = 0.9$). Natural mortality is 0.2.

Because the changes of the age composition through a year are small in those cases, they may not be utilised to assess the mortalities. However,

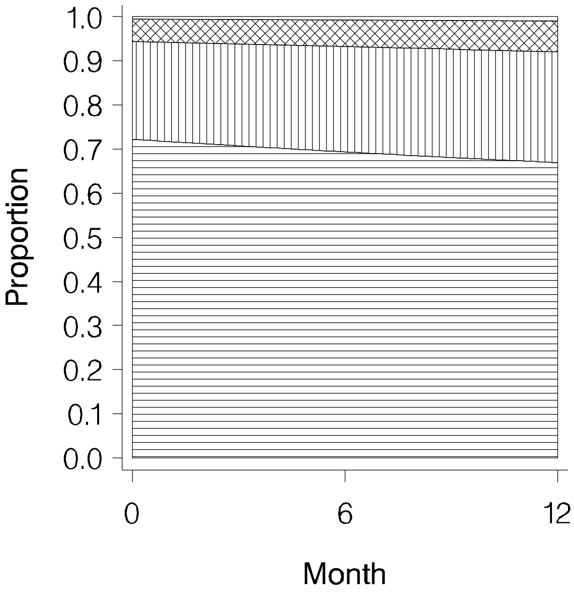

Figure 3.3: Changes of the proportions of four age groups through a year for constructed data. The horizontal lines corresponds to the proportion for an age group with stock numbers at the beginning of the year, $N = 200$, and total mortality, $z = 1.5$. Vertical lines corresponds to $N = 100$ and $z = 1.0$. Crosshatched lines corresponds to $N = 60$ and $f = 0.5$. White area corresponds to $N = 40$ and $f = 0.3$. The natural mortality is assumed to be 0.2.

the approximately constant level of the age composition (3.23) indicates that a linear approximation of the term $(1 - \exp(-z_a(t' - t)))$ from (3.24) is reasonable. A first order Taylor series approximation of the catch taken in a short period, (3.24), gives

$$\Delta C(t)_a \approx f_a N(t)_a \Delta t \tag{3.25}$$

where $\Delta t = t' - t$. This is the approximation that is used when assuming that CPUE is proportional to the abundance. The approximation is inserted into the equation for a continuation-ratio logit (refer to (3.5)):

$$L(t)_i = \log \frac{p(t)_i}{\sum_{j=i+1}^{A} p(t)_j} \approx \log \frac{f_i N(t)_i}{\sum_{j=i+1}^{A} f_j N(t)_j} \tag{3.26}$$

For simplicity it is assumed that age group A only covers age A. Letting (3.26) cover a whole year, the continuation-ratio logit becomes:

$$L_{y,a} = \log \frac{f_{y,a} N_{y,a}}{\sum_{i=a+1}^{A} f_{y,i} N_{y,i}} \tag{3.27}$$

where y denotes year and a age group. Thus, instead of the noninformative comparison of logits within a year, logits for succesive years covering the same cohort are compared. The difference between the logits is:

$$L_{y+1,a+1} - L_{y,a} = \log \frac{f_{y+1,a+1} N_{y+1,a+1} \sum_{i=a+1}^{A} f_{y,i} N_{y,i}}{f_{y,a} N_{y,a} \sum_{i=a+1}^{A} f_{y+1,i+1} N_{y+1,i+1}} \tag{3.28}$$

$N_{y+1,a+1} = N_{y,a} \exp(-(f_{y,a} + m_a))$ is inserted:

$$L_{y+1,a+1} - L_{y,a} = \log \frac{f_{y+1,a+1} \ \exp(-f_{y,a} - m_a) \ \sum_{i=a+1}^{A} f_{y,i} N_{y,i}}{f_{y,a} \ \sum_{i=a+1}^{A} f_{y+1,i+1} N_{y,i} \exp(-(f_{y,i} + m_i))} \tag{3.29}$$

The expression is complicated. Even applying a separability assumption such as $f_{y,a} = u_a v_y$, does not result in sufficient simplicity. However, if another logit called adjacent-ratio logit is used instead of the continuation-ratio logit, the difference between succesive logits becomes less complex.

The adjacent-ratio logit, L^*_{js}, compares an age group only to the subsequent age group instead of all subsequent age groups. The definition is:

$$L^*_{js} = \log \frac{p_{js}}{p_{(j+1)s}} \tag{3.30}$$

where the same notation is used as in (3.5). Using the approximation (3.25) as before, the approximate adjacent-ratio logit becomes:

$$L^*_{y,a} = \log \frac{f_{y,a} N_{y,a}}{f_{y,a+1} N_{y,a+1}} \tag{3.31}$$

The difference between two succesive logits is:

$$L^*_{y+1,a+1} - L^*_{y,a} = \log \frac{f_{y+1,a+1}\ f_{y,a+1}\ N_{y+1,a+1}\ N_{y,a+1}}{f_{y+1,a+2}\ f_{y,a}\ N_{y+1,a+2}\ N_{y,a}} \tag{3.32}$$

Inserting the expression $N_{y+1,a+1} = N_{y,a} \exp(-m_a - f_{y,a})$ and corresponding expression for $N_{y+1,a+2}$ the difference becomes:

$$L^*_{y+1,a+1} - L^*_{y,a} = \log \frac{f_{y+1,a+1}\ f_{y,a+1}\ \exp(-m_a - f_{y,a})}{f_{y+1,a+2}\ f_{y,a}\ \exp(-m_{a+1} - f_{y,a+1})} \tag{3.33}$$

Thus, the difference is not dependent on the actual stock numbers. If a separability assumption $f_{y,a} = v_y u_a$ is applied, the expression is further reduced:

$$\begin{aligned} L^*_{y+1,a+1} - L^*_{y,a} &= 2\log u_{a+1} - \log u_{a+2} - \log u_a + m_{a+1} - m_a + \\ &\quad v_y(u_{a+1} - u_a) \end{aligned} \tag{3.34}$$

Thus, based on the age composition data only, the fishing mortalities may be estimated.

However, the assumption of constant mortalities is crucial. In the case of the sandeel fishery, we have seen that the assumption does not hold, as the proportion of older sandeel in the catches is significantly lower in the start and end of the fishing season. The phenomenon is probably caused by a change of the proportion of the stock that is available to the fishery rather than a change in the fishing mortality. However, a change in the fishing mortality might cause the same change. Therefore, for simplicity, the availability has been included in the fishing mortality. In the case of changing mortalities within a year, shorter periods may be considered. Comparisons of the age proportions of the same cohort within the same year are then for adjacent-ratio logits:

$$L^*_{a,t_2} - L^*_{a,t1} = \log \frac{p_{a,t_2}}{p_{a+1,t_2}} - \log \frac{p_{a,t_1}}{p_{a+1,t_1}} \tag{3.35}$$

Inserting the approximate expression for the catch (equation (3.23) and (3.25)) one obtains:

$$L^*_{a,t_2} - L^*_{a,t1} = \log \frac{f_{a,t_2} N(t_2)_a}{f_{a+1,t_2} N(t_2)_{a+1}} - \log \frac{f_{a,t_1} N(t_1)_a}{f_{a+1,t_1} N(t_1)_{a+1}} \tag{3.36}$$

Assuming that the interval between t_1 and t_2 is short enough to assume that the fishing mortality f_{a,t_1} applies in the interval, relationship $N(t_2)_a = N(t_1)_a \exp(-z_{a,t_1}(t_2 - t_1))$ and corresponding relationship for $N(t_2)_{a+1}$ may be inserted. One obtains:

$$L^*_{a,t_2} - L^*_{a,t_1} = \log \left(\frac{f_{a,t_2}}{f_{a+1,t_2}} \frac{f_{a+1,t_1}}{f_{a,t_1}} \frac{\exp(-z_{a,t_1}(t_2 - t_1))}{\exp(-z_{a+1,t_1}(t_2 - t_1))} \right) \tag{3.37}$$

which might be simplified into:

$$\begin{aligned} L^*_{a,t_2} - L^*_{a,t_1} \quad &= \quad \log f_{a,t_2} - \log f_{a+1,t_2} + \log f_{a+1,t_1} - f_{a,t_1} \\ &\quad -z_{a,t_1}(t_2 - t_1) + z_{a+1,t_1}(t_2 - t_1) \end{aligned} \tag{3.38}$$

Further simplification may be obtained if a structure, such as separability may be assumed. (However, presumably the separability assumption does not apply to the sandeel fishery.) Thus, even in the case of varying mortalities through the year, the approach may be applied when the year may be divided into several intervals within each of which one may assume that the mortalities are constant.

The approach of comparing cohorts between two succesive years has also been considered by Pope and Shepherd (1982). Although the approaches have the similarity of comparing cohorts between succesive years, the approaches are different. Pope and Shepherd (1982) compare the catch from a cohort for two succesive years, $D_{y,a}$:

$$D_{y,a} = \log \frac{C_{y+1,a+1}}{C_{y,a}} \tag{3.39}$$

whereas I compare the *relative* catch from a cohort for two succesive years (rewriting of equation (3.28)):

$$L_{y+1,a+1} - L_{y,a} = \log \frac{C_{y+1,a+1}/\sum_{i=a+2}^{A} C_{y+1,i}}{C_{y,a}/\sum_{i=a+1}^{A} C_{y,i}} \tag{3.40}$$

and in the case of continuation-ratio logits (rewriting of equation (3.33)):

$$L^*_{y+1,a+1} - L^*_{y,a} = \log \frac{C_{y+1,a+1}/C_{y+1,a+2}}{C_{y,a}/C_{y,a+1}} \tag{3.41}$$

in the case of adjacent-ratio logits. The main difference between the approach presented here and the approach of Pope and Shepherd (1982) is that whereas the approach of Pope and Shepherd (1982) requires catch at age data, the approach suggested here requires only age composition data. Thus, information on the position of the fishery, the actual amount caught, the species composition and the mean weight of the species is not needed in the approach presented here. Other differences are that the approach of Pope and Shepherd (1982) is based on a least squares estimation procedure, which does not take into account the special characteristics of the variance structure of catch at age data, induced by the age composition data, which is one of the main data sources in estimation of catch at age data. Furthermore, the approach presented here facilitates an investigation of the sources of variation.

The adjacent-ratio logits have the advantage compared to continuation-ratio logits of not being dependent on the stock numbers. However, they do have the drawback of being based on fewer individuals, in that one age group is compared to the subsequent age group, instead of all subsequent age groups. Secondly, the influence from the age determination errors is reduced by utilising continuation-ratio logits instead of adjacent-ratio logits, as continuation-ratio logits require a classification of the sandeels into at most three age groups, whereas adjacent-ratio logits require an additional age group (except for the comparison of the two oldest age groups). The classifications needed for the continuation-ratio logit for age group a, L_a, are a classification of the youngest sandeels not encompassed by the logit (not needed for the logit for age group R), and a classification into the two groups two be compared, i.e. a group consisting of a-year olds and a group consisting of older fish. Adjacent-ratio logits require an additional age group; the oldest sandeels not encompassed by the comparison have to be sorted out. This extra uncertainty which is added if the adjacent-ratio logits are utilised instead of continuation-ratio logits, may be of relatively considerable magnitude, as the difficulties of distinguishing between the ages are presumed to increase as the sandeels grow older. At last, continuation-ratio logits are independent as long as they do not have any parameters in common, whereas adjacent-ratio logits are not. Independence between the logits of different age groups are a necessary condition for analysing the logits for the different age groups separately. Which circumstances that are of most importance may be difficult to judge. Preliminary analyses have not shown great differences between the models of adjacent-ratio logits and continuation-ratio logits.

Chapter 4

Analysis of Age Composition Stratified by Length Groups

Age determinations of fish are often time-consuming and expensive to perform, and therefore reduction of the number of age determinations is an important subject in the designing of sampling schemes (e.g. Ketchen, 1950; Schweigert and Sibert, 1983; Horppila and Peltonen, 1992). The number of age determinations may be reduced by utilising the correlation between age and length. Often large samples are collected for estimation of the length composition of the catch, and then the length groups are subsampled for estimation of the age composition. Because length determinations are much easier to perform than age determinations, length determinations are prefered to age determinations. The age composition may be estimated by e.g. a Bayesian approach (Nandram *et al.*, 1997) or by a maximum likelihood approach (e.g. Kimura and Chikuni, 1987; Martin and Cook, 1990; and Hoenig *et al.*, 1993). The basic idea of (Kimura and Chikuni, 1987; Martin and Cook, 1990; and Hoenig *et al.*, 1993) is to assume that the length distribution in the catch is compounded by overlapping normal distributions; one for each age group i, $f(l)_i$, where $i = R, \ldots, A$:

$$f(l)_i = \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{(l-\mu_i)^2}{2\sigma_i^2}) \tag{4.1}$$

μ_i and σ_i^2 are the mean and variance of age group i. The proportion of age group i among individuals of length l in the catch, $p(l)_i$, is then

$$p(l)_i = \frac{p_i f(l)_i}{\sum_{j=R}^{A} p_j f(l)_j} \tag{4.2}$$

where p_i denotes the proportion of i-year-olds of the catch, $i = R, \ldots, A$. The likelihood function may then be formulated from the data, which consists of a set of observations of individuals which have been age determined as well as length determined and a set of observations of individuals which have been length determined only. Because estimation is based on maximum likelihood, significance of effects may be evaluated by likelihood ratio tests (e.g. Hoenig *et al.*, 1993). However, models of handling structures of the age composition which encompass fixed as well as random effects (such as the structures in chapter 3) have not been investigated in this work. Instead an approach to analyse the age composition data for given length separately, excluding the length composition data, has been investigated. The purpose of the analysis is to identify sources of variation and determine their magnitudes, i.e. the same as in chapter 3 and 5. The approach is described below.

4.1 Analysing sources of variation in age composition for given length

Assume that the length distribution in the catch, for age group i may be described by a continuous function $f(l)_i$. The corresponding age distribution for length l is then as shown in equation (4.2). The continuation-ratio logit (defined in equation (3.5)) for given length l, for age group i, $L(l)_i$, is:

$$L(l)_i = \log \frac{p(l)_i}{\sum_{j=i+1}^{A} p(l)_j} = \log \frac{p_i f(l)_i}{\sum_{j=i+1}^{A} p_j f(l)_j} \tag{4.3}$$

The length distributions for the various age groups for sandeel are shown in Appendix D for the years 1984-1993. Distributions such as a normal distribution or a gamma distribution may presumably describe the length distribution satisfactorily. Note that it is likely that the variation may be resolved into various components associated with the sources of variation, such as the period of the year, geographical areas, etc.. If these sources of variation are taken into account, the overlap between the length distributions for the various age groups will be reduced. Thereby the age groups will be easier to distinguish from each other on the basis of the length distribution. The continuation-ratio logit for age group i, assuming normal distribution for the length, $L^N(l)_i$, is:

$$L^N(l)_i = \log p_i - \log \sigma_i - \frac{1}{2}\frac{(l-\mu_i)^2}{\sigma_i^2} - \log(\sum_{j=i+1}^{A} \frac{p_j}{\sigma_j}\exp(-\frac{(l-\mu_j)^2}{2\sigma_j^2})) \quad (4.4)$$

where equation (4.1) has been inserted into (4.3). Unfortunately, the logit is not a simple function of the length, l. If however, it is realistic to assume that the length distribution for $i+2$-year-olds does not overlap with the length distribution for individuals which are two years younger, i.e. i-year-olds, then the continuation-ratio logit may be approximated by the adjacent-ratio logit, $L^{N*}(l)_i$:

$$L^{N*}(l)_i = \log p_i - \log \sigma_i - \frac{(l-\mu_i)^2}{2\sigma_i^2} - \log p_{i+1} + \log \sigma_{i+1} + \frac{(l-\mu_{i+1})^2}{2\sigma_{i+1}^2} \quad (4.5)$$

Thus, the continuation-ratio logit may approximately be described by a second degree polynomium of the length, provided the length distribution for an age group is approximately normally distributed. The dependency of the length is even simpler if the standard deviations of the length distributions of age group i and $i+1$, may be assumed approximately equal, i.e. if $\sigma_i = \sigma_{i+1} = \sigma$. In this case (4.5) simplifies to:

$$L^{N*}(l)_i = \log p_i - \log p_{i+1} - \frac{1}{2\sigma^2}(2l(\mu_{i+1}-\mu_i) - \mu_{i+1}^2 + \mu_i^2) \quad (4.6)$$

i.e., a linear function of the logit. If the length distribution for age group i, $i = R, \ldots, A$, is assumed to be gamma distributed with mean $k_i\beta_i$ and coefficient of variation $1/\sqrt{k_i}$, instead of normally distributed, the adjacent-ratio logit approximation of the continuation-ratio logit (equation (4.5)), becomes even closer to a linear relation:

$$\begin{aligned} L^{G*}(l)_i &= \log p_i - \log \Gamma(k_i) - k_i \log \beta_i + (k_i - 1) \log l - \frac{l}{\beta_i} \\ &- \log p_{i+1} + \log \Gamma(k_{i+1}) + k_{i+1} \log \beta_{i+1} - \\ &(k_{i+1} - 1) \log l + \frac{l}{\beta_{i+1}} \end{aligned} \tag{4.7}$$

where $\Gamma(k)$ is:

$$\Gamma(k) = \int_0^\infty t^{k-1} \exp(-t) dt \tag{4.8}$$

and $k > 0$.

In order to assess which of the above approximations of the continuation-ratio logit that is most appropriate for the sandeel landings, four scenarios considered to cover most cases of length compositions for sandeel, are analysed.

4.1.1 Four scenarios of the length composition

In the four scenarios, two cases of length distributions are considered; the Gaussian or normal distribution and the Gamma distribution. The parameters are chosen, so that the mean and standard deviation are equal to the averages of the mean and standard deviation of the observed distributions for the years 1984-1993 (Appendix D). Thus, the distributions considered are conservative regarding the assumed standard deviation, because it has not been attempted to be resolve the variation into smaller components. The averages of the means and standard deviations of the years 1984-1993 are:

Age group	Mean (scm=$\frac{1}{2}$ cm)	Std. (scm)
0	18.2	2.4
1	24.3	3.6
2	30.3	3.2
3	33.7	3.1
4	36.7	3.1

As regards the catch at age data, two scenarios for each distribution are examplified to cover the most common cases; a factor 10 in difference between the catch at age data for two subsequent age groups, and equal sized catch at age data for all age groups. The relative frequencies are shown in figures 4.1 and 4.2 for the Gaussian case and 4.3 and 4.4 for the Gamma case.

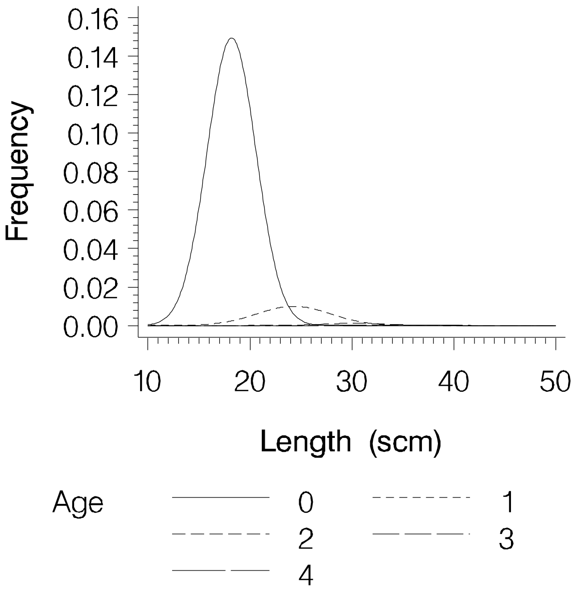

Figure 4.1: Length distribution for the various age groups for the Gaussian case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed.

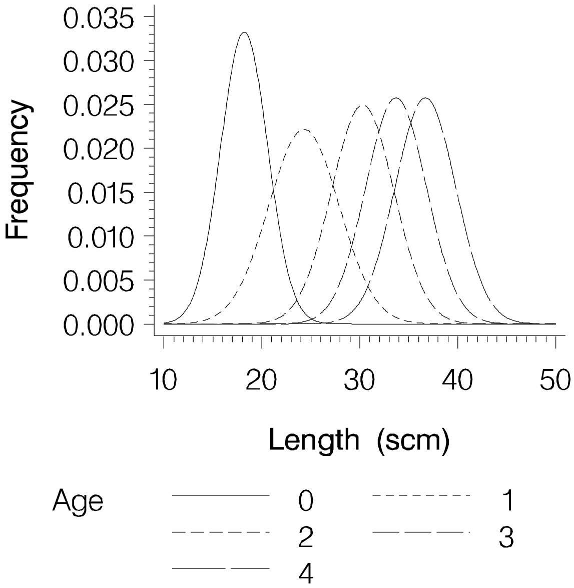

Figure 4.2: Length distribution for the age groups for the Gaussian case where equal catches are assumed for all age groups.

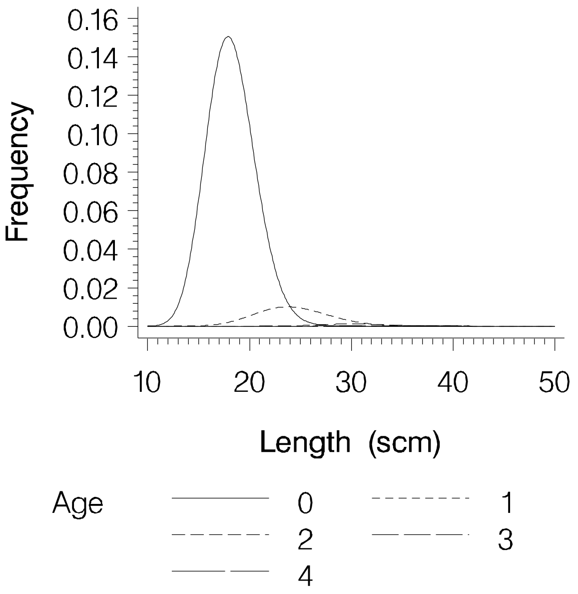

Figure 4.3: Length distribution for the various age groups for the Gamma case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed.

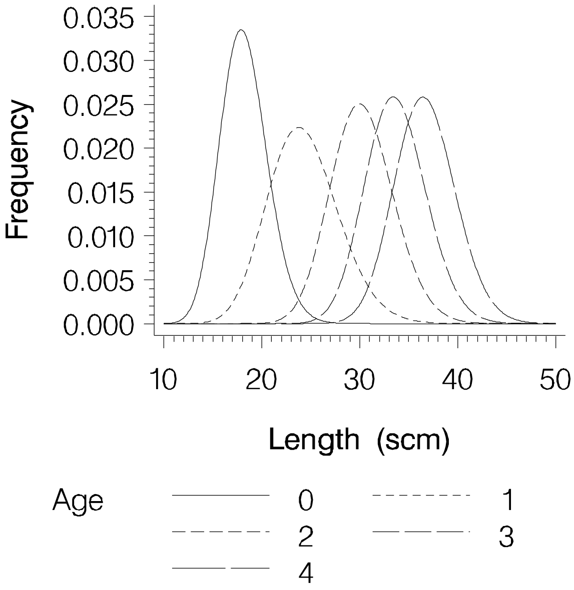

Figure 4.4: Length distribution for the age groups for the Gamma case where equal catches are assumed for all age groups.

The continuation-ratio logits are shown in figures 4.5 and 4.6 (Gaussian case) and 4.7 and 4.8 (Gamma case).

A linear relationship and a second degree polynomium between the continuation-ratio logit and the length are fitted based on the range from $]-5;5[$, which covers appr. 99% of the observations. Thereby heavy extra polations into the far tails are avoided. The polynomium of second degree almost coincides with the true relationship. In the Gaussian case, the linear approximation of the relationship between the contination-ratio logit and the length for age group 0 is certainly crude, which is caused by a relative large difference between the standard deviation of the length of 0-year-olds and the neighbour length distribution, which is that of the 1-year-olds. For the gamma case, the non-linearities are largest for age group 1 in the case where the difference between the sizes of the age groups is large. Thus, in some cases, it might be relevant to estimate a second degree polynomium, however, in most cases the linear approximation holds. The estimated age compositions based on the fitted lines are shown together with the true age compositions in figures 4.9 and 4.10 for the Gaussian case, and in figures 4.11 and 4.12 for the Gamma case. The largest bias occurs in the case where there are large differences between the catch numbers for the various age groups.

4.1.2 Discussion

The four scenarios indicate that the relationship between the logit and the length may be approximated with a linear relationship or perhaps a polynomial of second degree. Thus the length stratified data may be analysed analogously to the data where no stratification has been performed on length groups. The only difference is that all effects in the models must contain the length and perhaps even the squared length as a regression variable. Thus, the continuation-ratio logits, $L_{R,\ldots,A}$, may be modelled as:

$$L_i = \mathsf{Y} * l + \mathsf{Y} * l^2 + \mathsf{A} * l + \mathsf{A} * l^2 + \ldots \tag{4.9}$$

where $i = R, \ldots, A$, Y and A are examples of possible sources of variation. However, the interpretaion of the parameters refers to age composition as well as growth. Thus, the analysis does not only concern the age composition, but also growth.

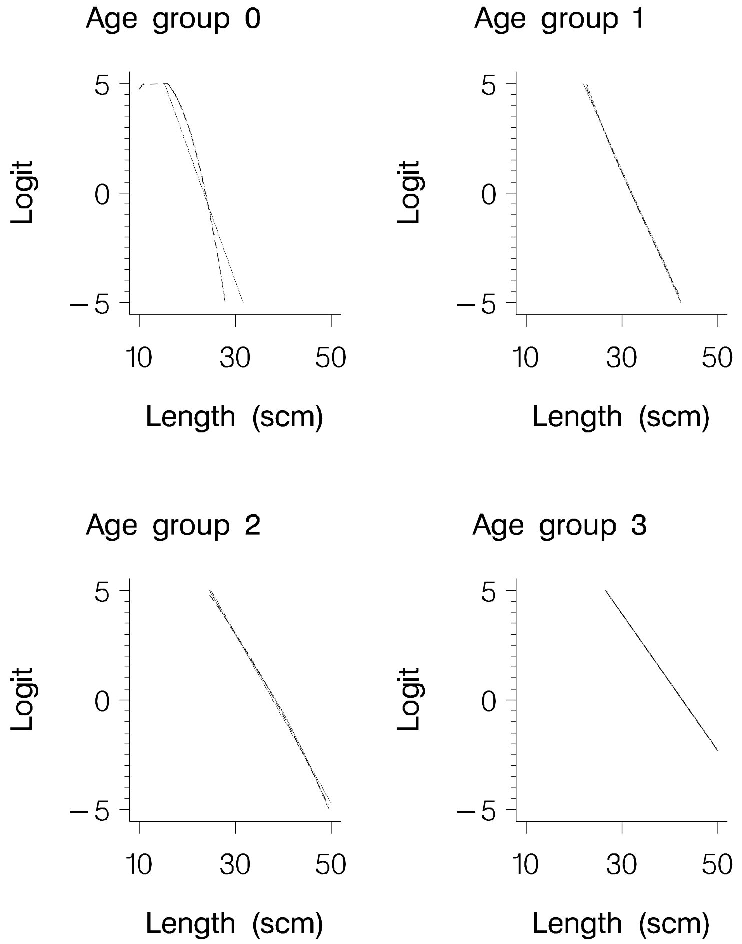

Figure 4.5: Continuation-ratio logits versus length for the Gaussian case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed. A straight line and a polynomium of second degree are fitted.

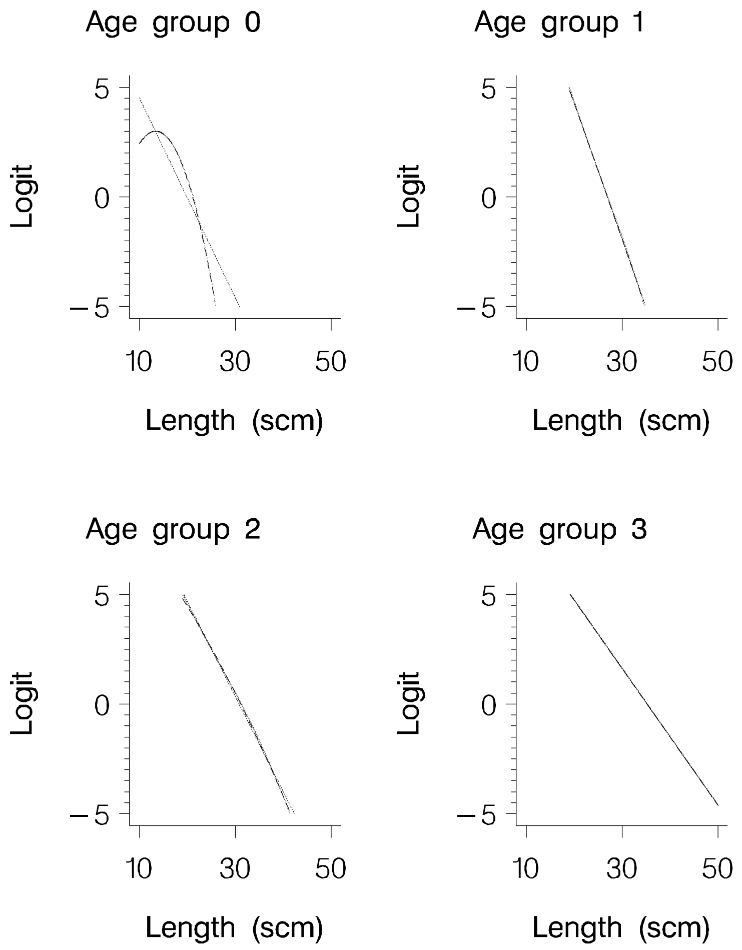

Figure 4.6: Continuation-ratio logits versus length for the Gaussian case where equal catches are assumed for all age groups. A straight line and a polynomium are fitted.

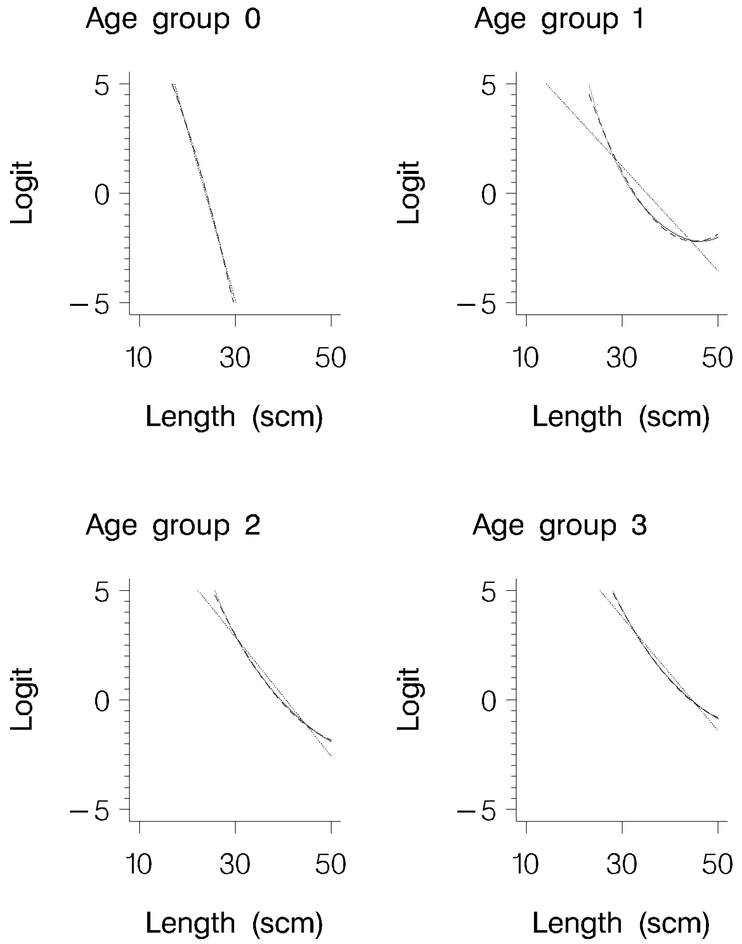

Figure 4.7: Continuation-ratio logits versus length for the Gamma case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed. A straight line and a polynomium of second degree are fitted.

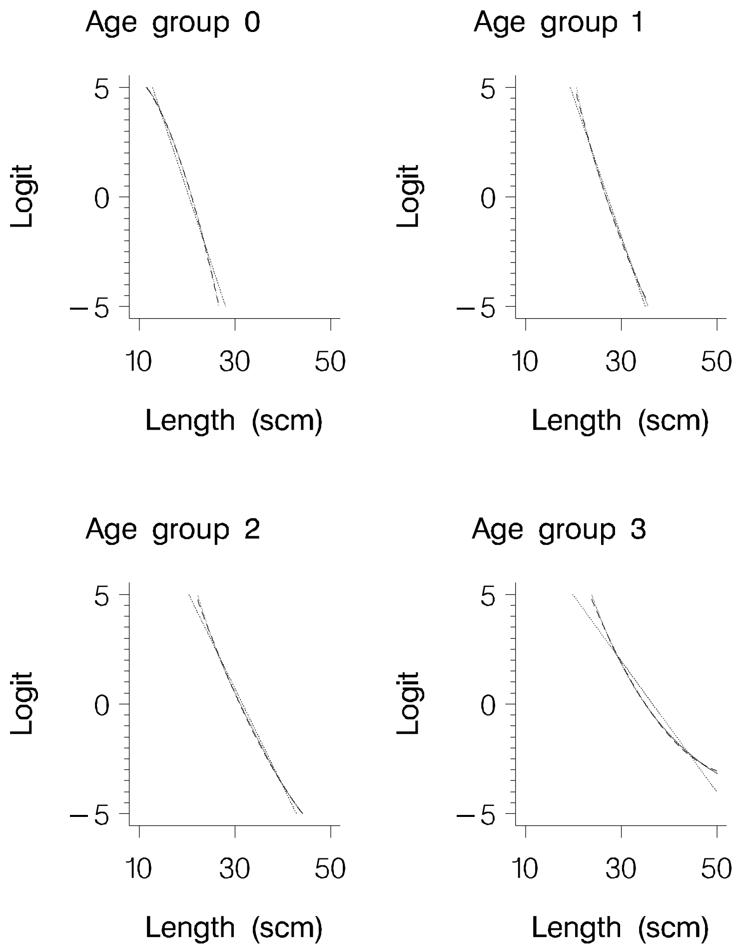

Figure 4.8: Continuation-ratio logits versus length for the Gamma case where equal catches are assumed for all age groups. A straight line and a polynomium of second degree are fitted.

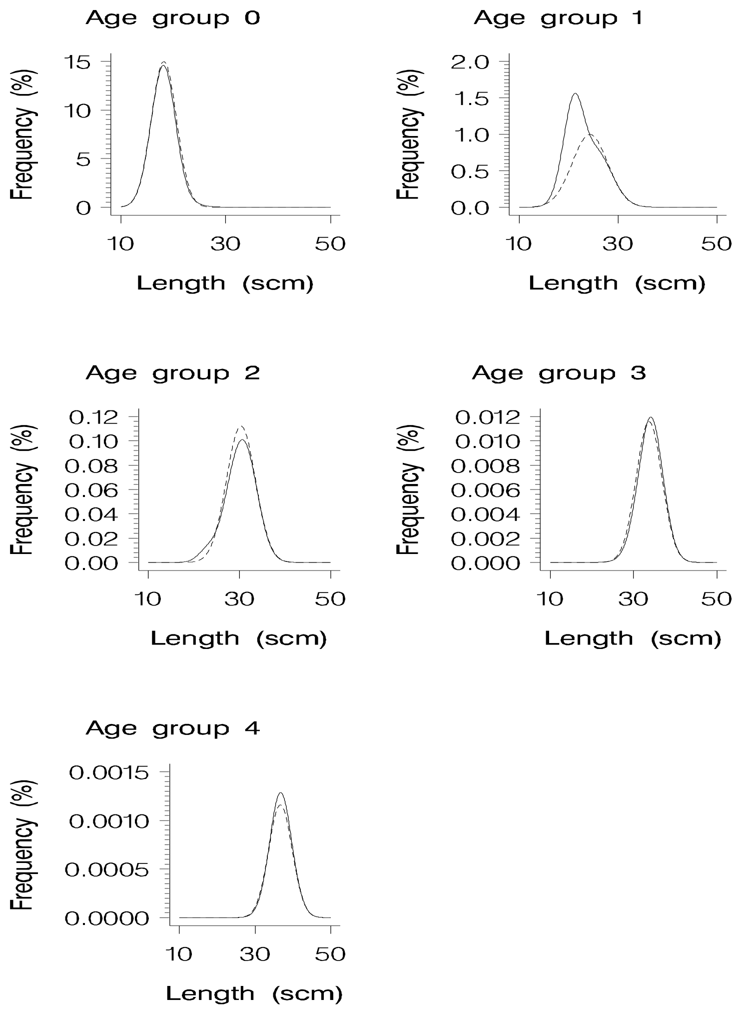

Figure 4.9: Comparison of true (dashed line) and estimated (solid line) length distributions for the age groups for the Gaussian case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed.

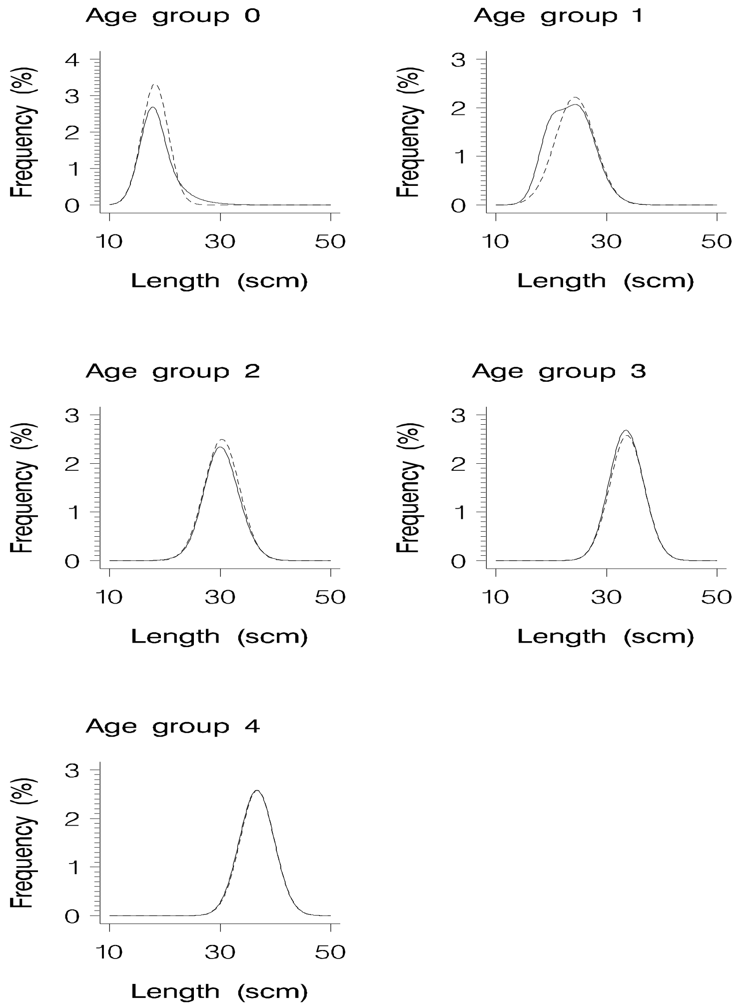

Figure 4.10: Comparison of true (dashed line) and estimated (solid line) length distributions for the age groups for the Gaussian case where equal catches are assumed for all age groups.

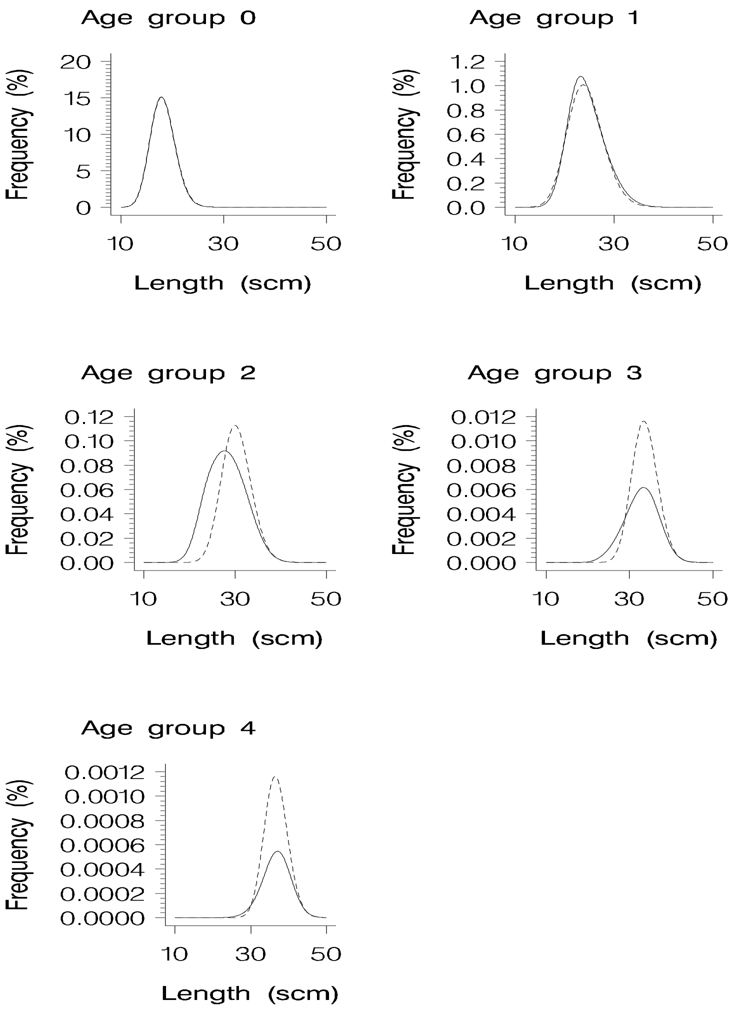

Figure 4.11: Comparison of true (dashed line) and estimated (solid line) length distributions for the age groups for the Gamma case where a factor 10 in difference between catch numbers for two subsequent age groups is assumed.

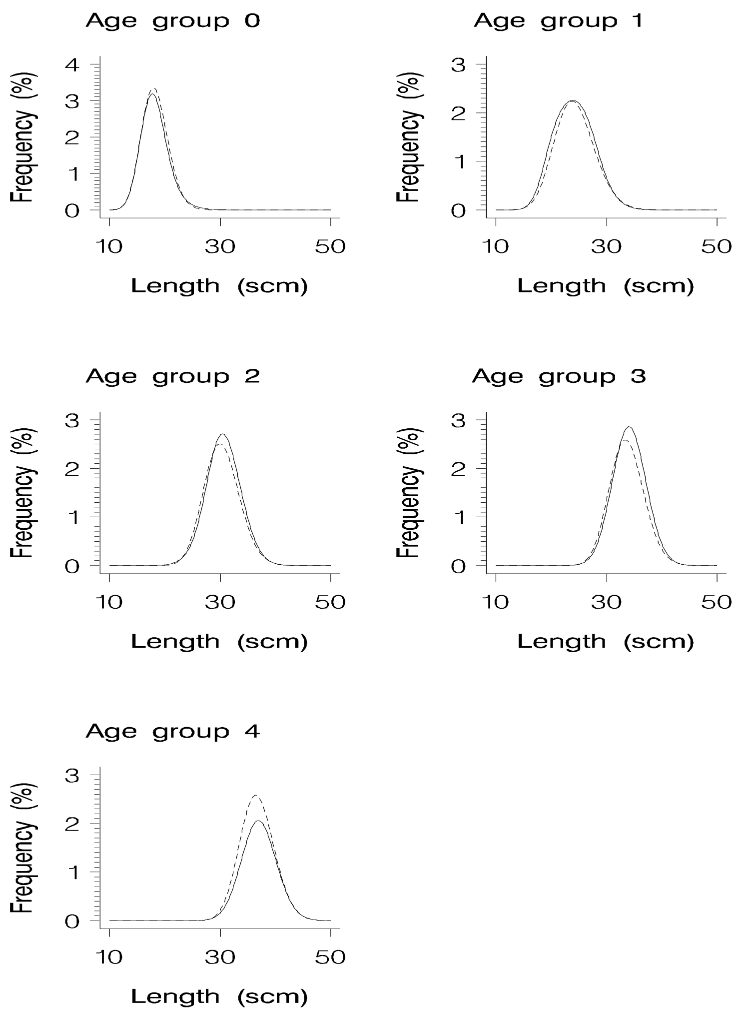

Figure 4.12: Comparison of true (dashed line) and estimated (solid line) length distributions for the age groups for the Gamma case where equal catches are assumed for all age groups.

If estimates of the age compositions are desired from the length composition data, an extra step compared to the simple case without stratification on length groups is needed (chapter 3). The relevant results from the analysis of the continuation-ratio logits are estimates of the age composition for given length group, ℓ, $\mathbf{p}_\ell = (p_{R\ell}, \ldots, p_{A\ell})$, where $\ell = 1, \ldots, n$, and the associated covariance matrix, $\Sigma_{p\ell}$, which describes the covariances between any two age groups from the same or different length groups. Before the age composition for given length is combined with length composition estimates, the length distribution has to be analysed, in order to determine geographical areas and periods with similar length composition of the cathes. Other factors such as the mesh size may also influence the length composition of the catch. This analysis is impeded by the structure of the response, which is compounded by several continous functions, i.e. the length distribution for each age group. The analysis should provide estimates of the length composition, $\boldsymbol{\nu}=(\nu_1, \ldots, \nu_n)$, and the associated uncertainties, Σ_ν, where n denotes the number of length groups and ν_ℓ the proportion of sandeel in length group ℓ, $\ell = 1, \ldots, n$. Finally, the length composition and age composition for given length are combined, and the age composition is obtained from:

$$p_i = \sum_{\ell=1}^{n} p_{i\ell}\nu_\ell \tag{4.10}$$

where $i = R, \ldots, A$. The corresponding variance structure may be obtained by a Taylor approximation of (4.10).

The drawback of this approach to estimate the age composition is that information on the shape of the length distribution is not utilised in the analysis of the age composition to resolve the length distribution into separate length distributions for the age groups. It is likely that the length distribution for an age group may be described by a uni-modal distribution, such as a normal distribution. However, such a restriction is not utilised in the suggested approach; here there is no restriction at all on the shape of the length distribution for given age, as may be seen on the estimated length distributions (figures 4.9 and 4.10 for the Gaussian case and figures 4.11 and 4.12 for the Gamma case). An extension of the method to also utilise the presumed shape of the length distribution for given age would improve the estimates of the age composition.

Chapter 5

Uncertainties of Catch at Age Data for Sandeel

Catch at age data constitutes the main source of information in age-structured stock assessment models, and thus assessment of the associated uncertainties are important. The uncertainties are assessed in this chapter, using the sandeel fishery in the North Sea as a case study. However, the analyses may be applied to other fisheries as well, with a similar structure of the data sources.

The perhaps most important sources of variation are associated with the age composition estimates, considered in chapter 3. However, other sources may also be of importance, such as uncertainty of the catch per area and the species composition of the catches, and uncertainties associated with the transformation of the unit of measurement of the size of the catch from tonnes to counts. In order to obtain an assessment of the accumulated uncertainty of catch at age data, the various sources of information utilised in the estimation of catch at age data are analysed separately and thereafter combined into estimates of the catch at age data. At first, the information on the weight of the industrial catch is combined with estimates of the species composition in weight giving the catch weight of sandeel. Secondly, the mean weight of sandeels is utilised to transform the weight of the catch into the number of sandeel caught. At last the age composition of the catches is utilised to get estimates of the number of sandeel caught from each age group. The materials, methods, results and discussion have been presented in Kvist *et al.* (1999b) (Appendix C), and is not repeated here.

Chapter 6

Modelling Fish Stocks by Means of Stochastic Differential Equations

The development of computer facilities has widened the possibilities of using modelling approaches that have not been considered previously within this field because of limitations in computer power. Stochastic differential equations are an example of such a computer intensive tool which facilitates more realistic models. The stochastic differential equations are an extension of the ordinary deterministic differential equations in continuous time to handle also uncertainties in the system dynamics, as well as uncertainties in the data. An introduction to stochastic differential equations is given by Øksendal (1995). The tool has become widely used within different fields where the fundamental dynamics are described by one or more differential equations. Madsen and Holst (1995) give an example of modelling the heat dynamics of a building and a good introduction into the subject of stochastic differential equations. Melgaard (1994) considers the general problems of identification of physical models within the framework of stochastic differential equations, and gives several examples of areas of application. One of those examples is estimation of parameters in a multi-species system with simulated data. Gard (1988) discusses various models of population dynamics by means of stochastic differential equations. Lungu and Øksendal (1997) consider a model for population growth and discuss optimal harvest strategies. Dennis *et al.* (1991) consider the modelling of endangered species in order to estimate quantities related to growth rates and extinction probabilities. An example of the use of stochastic differential equations to model an extremely non-linear system is given within finance, by Nielsen *et al.* (1999).

Stochastic differential equations seem to be a promising tool for modelling fish stocks, as the fundamental dynamics are considered to be described by one or several differential equations, e.g. in the single-species model the dynamic is described by equation (2.1), which is repeated here:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = -(f(t) + m(t))N(t) \tag{6.1}$$

where $N(t)$ denotes the number of individuals in the cohort at time t. $f(t)$ is the fishing mortality, and $m(t)$ is the natural mortality where the time-dependency of the mortalities is indicated explicitly. In the current time series models which do take the cumulative properties as well as observation errors into account, such as e.g. Gudmundsson (1994), the mortalities are assumed to be constant through a period, often a year, and therefore the equation (6.1) is solved for such a period, before any term to account for uncertainties is entered. References to other time series models are given in chapter 2, section 2.1.

However, the mortalities are likely to vary during that period and do not remain constant e.g. through a year, as demonstrated for sandeel in chapter 3. The random variation of those quantities is presumably also varying through the year. Thus, an intuitively more realistic approach would be to include a term accounting for the variation in continuous time of the mortalities:

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = -\left[(f(t) + m(t))N(t) + \sigma(t, N(t)) \times \text{``}noise''\right] \tag{6.2}$$

i.e. a stochastic differential equation. $\sigma(t, N(t))$ is a function of the time, t, and the cohort size, $N(t)$. For most practical purposes, it is desired that the noise term has certain basic properties, such as the noise at two different time points are independent and the proces is stationary. However, there does not exist a function with continuous paths that fulfills those two basic assumptions (see e.g. Øksendal (1995)) and therefore the equation (6.2) is rewritten in the form:

$$dN(t) = -\left[(f(t) + m(t))N(t)dt + \sigma(t, N(t))dw(t)\right] \tag{6.3}$$

where $w(t)$ is the standard Wiener process, representing the source of noise in the system. The standard Wiener process has the appealing properties that the increments of the process in two non-overlapping periods, are independent of each other, and that these increments are Gaussian with mean zero and a variance which is proportional to the length of the interval.

It is reasonable to assume that the uncertainty in model (6.3) is proportional the abundance of the cohort, $N(t)$, i.e.:

$$dN(t) = -(f(t) + m(t) + \sigma dw(t))N(t)dt \tag{6.4}$$

so that the uncertainty is associated with the mortalities and thus the uncertainty of the cohort size is dependent on the abundance of the cohort; the greater abundance, the greater variance. The uncertainty $w(t)$ accounts for variation of the fishing mortality, $f(t)$, as well as the natural mortality, $m(t)$. It would be convenient to split the uncertainty into two additive terms associated with each of the mortalities, i.e. $w(t) = w^f(t) + w^m(t)$. Inserting this relationship into equation (6.4) one obtains:

$$dN(t) = -(f(t) + \sigma_f dw^f(t) + m(t)\sigma_m dw^m(t))N(t)dt \tag{6.5}$$

The catch is then:

$$dC(t) = (f(t) + \sigma_f dw^f(t))N(t)dt \tag{6.6}$$

The observations are the catch in a certain period:

$$O_{t_j} = C(t_j) - C(t_{j-1})) + \epsilon_{t_j} \tag{6.7}$$

where ϵ_{t_j} accounts for the uncertainty of the observation O_{t_j}, $j = 1, \ldots, k$. These uncertainties may be estimated outside the model, on the basis of the data sources utilised for estimation of catch at age data, as shown in Kvist *et al.* (1999b) (Appendix C). The length of the period could be a year or shorter, such as a month. As regards the fishing mortality, it often is modelled by means of the fishing effort, E, which is a standardised number of fishing days, taking into account the effectiveness of the vessels (refer to chapter 2, section 2.1). The fishing mortality is assumed proportional to the effort, i.e. $f(t) = q(t) \times E(t)$, where q is called the catchability. The catchability is associated with the effectiveness of the vessels and might be dependent on the age of the cohort (and/or perhaps length dependent). Development of the techniques and equipment may be reflected in an increase of the catchablity over time, i.e. a possible model for the catchability might be

$$dq(t) = \Theta dt + dw^q(t) \tag{6.8}$$

where Θ represents the trend and $w^q(t)$ the uncertainty, described by a Wiener process. The equations described here are valid for a single cohort only. Therefore a term representing the age is not necessary.

Another issue of importance for the modelling of the sandeel stock is that the population available to the fishery varies through a year as the sandeel buries in the sediment during winter. In addition, the availability varies differently, for different ages (refer to Kvist *et al.* (1999a) (Appendix B)). These characteristics may be modelled by introducing an availability coefficent which varies through the year and is different for different age groups. Fournier and Doonan (1987) defined the availability as the proportion of individuals in an age group with a positive probability of being caught. The concept was first used by (Widrig, 1954), who referred to the availability as the accessibility of the fish in the population to the fishing gear. Alternatively one may say that the availability is the proportion of the stock that is present on the fishing grounds. The availability is different from the catchability in the sense that the availability refers to the individuals being available to the fishery, whereas the catchability refers to the probability that, once available, a fish will be caught by a unit of effort. However, other definitions of the terms might be used, e.g. one of the reasons that fishing mortality is different for the first and second half of the year is that a smaller part of the stock is available to the fishery in the second half of the year. For the part of the population which is not avaliable to the fishery the fishing mortality is zero, whereas the natural mortality might be the same as for the same age group during the winter period, because the difference in availability between two age groups presumably is caused by a difference in the length of the periods they are buried in the sediment. Introducing the catchability and separability into (6.5) one obtains:

$$\begin{aligned} dN(t) &= -(q(t)E(t) + \sigma_f dw^f(t) + \tilde{m}(t)\sigma_m dw^{\tilde{m}}(t))a(t)N(t) - \\ & (\breve{m}\sigma_{m^w} dw^{\breve{m}}(t))(1 - a(t))N(t)dt \end{aligned} \tag{6.9}$$

where $a(t)$ denotes the availability, which may be estimated from the age composition analyses. $\tilde{m}$ is the natural mortality of that part of the stock that is available to the fishery, and $\breve{m}$ is the natural mortality of the part of the stock that is assumed to be hiding in the sediment. $w^{\tilde{m}}(t)$ and $w^{\breve{m}}(t)$ are the uncertaintes associated with the natural mortalities described by a Wiener process.

The equations above all refer to a single cohort only. When they are extended to cover several cohorts common structures may be utilised, e.g. the catchability and the natural mortality could be age and time dependent only. However, it is not certain if an approximation such as a separability assumption may be utilised, as there has been indications that the fishery might be directed towards specific cohorts (Kvist *et al.*, 1999a) (Appendix B).

6.1 Estimation

The stochastic differential equations have seldom analytical solutions. One exception is the case where the noise is proportional to the state variable, $N(t)$, e.g. the solution of (6.5) is, assuming constant mortalities, f and m:

$$N(t) = N(0)\exp -(ft + \sigma_f e_f(t) + mt + \sigma_m e_m(t) + \frac{\sigma_f^2 + \sigma_m^2 + 2\sigma_f\sigma_m}{2}t) \tag{6.10}$$

where $e(t) = \int_0 tdw(t')$, i.e. normally distributed, which means that $N(t)$ is log-normal distributed (i.e. $\log N(t)$ is normally distributed). However the equation for the accumulated catch (6.6) does not have an analytical solution:

$$C(t) = \int_0^t (fN(t')dt' + \sigma_f N(t')dw(t')) \tag{6.11}$$

Instead numerical methods have to be applied. As an example of a program for estimation of stochastic differential equations one could mention CTLSM (Madsen and Melgaard, 1991). CTLSM is a program for maximum likelihood estimation in stochastic, continuous time dynamical models. The program has been compared to other software in a system identification competion, where it proved to be the best tool for estimation of stochastic differential equations (Madsen *et al.*, 1996).

However, a program such as CTLSM cannot be applied directly to the data associated with fish stock assessments. An investigation of the possible use of CTLSM for estimation of the parameters in the stochastic model (6.9) for a fish stock with several cohorts, showed that certain minor changes of CTLSM are needed. The observations might be the catch at age for a period as described by equation (6.7), or the accumulated catch at age, $O^*_{t_j}$:

$$O^*_{t_j} = C(t_j) + \xi_{t_j} \tag{6.12}$$

where ξ_{t_j}, $j = 1, \ldots, k$, is the accumulated uncertainty of the catch at age data. The covariance matrix of the observations may be estimated outside the model (refer to Kvist *et al.* (1999b) Appendix C). The technical problem is that the program CTLSM was built to handle constant covariance matrix. This is not a realistic assumption for the accumulated catches (equation 6.12), but possibly for the observations described by equation (6.7). However, the latter case requires that the accumulated catch at the previous measurement time, t_{j-1}, is available in the estimation routine, which is not possible in the current version of CTLSM.

Another problem is that of handling cohorts. The natural approach is to let one state variable, $N(t)$, correspond to one cohort. All state variables in the system have to be present in the description of the system from the beginning. Thus, the state variable may not be entered at the time the cohort is born. A practical solution to this problem could be to keep the fishing mortality and natural mortality at zero for the cohort until birth, and then at the time of birth assign a recruitment population.

It might also be relevant to include CPUE data, which provide information on the catch per unit of effort, i.e. a more detailed information on the catch per time unit, where the vessel size is included as an explanatory variable.

However, these problems only seem to be minor technical problems. When these problems have been solved, the approach might provide an alternative to current stock assessment models with the advantage that variations of mortalities and availability through the year may be more realistically modelled.

Chapter 7

Conclusion

In this thesis uncertainty associated with stock assessment has been considered, especially uncertainties associated with the input data. The thesis provides new approaches to analyse the sources of variation and their magnitude in the input data, and an alternative approach for modelling the dynamics of a fish population is suggested.

The main results of the thesis are that the combination of continuation-ratio logits and the generalized linear mixed models is a powerful tool for analysing sources of variation and their magnitude in age composition data. By combining the continuation-ratio logits and the generalized linear mixed models, the ordinal and multinomial characteristics of the response may be taken into account at the same time as fixed as well as random effects may be analysed. The analysis provides improved estimates of the age composition and the associated variances and covariances, information which is important in the assessment of the stock abundance and mortalities and their uncertainties. Knowledge of the sources of variation may also be utilised to improve the efficiency of the sampling. In addition valuable information on the stock dynamics are obtained, which may be utilised to improve models describing the stock dynamics of sandeel.

The method was used to quantify the importance of the various sources of variation in the age composition of sandeel landings from the North Sea caught in the years 1984-1993. The main conclusions were that larger fish presumably emerge from the sediment later in the season and re-enters the sediment earlier. This implies that the variation of the availability of sandeel through the year depends on the age, a circumstance important to consider in the modelling of the population dynamics of the sandeel stock. Other information of importance for the structure of such a model is that data seemed to indicate that the fishery has been attracted to 1-year-old fish in years where they were abundant. If this is correct, the often used separability assumption of the fishing mortality in the stock assessment model is not valid. Data suggested that the age proportion of a cohort might depend on the age proportion of the cohort in the previous year. This dependency might be utilised for prediction of age composition of catches. The influence of gear and mesh size was found to be negligible and therefore stratifying the sampling effort by gear and mesh size is unlikely to result in a lower overall variation. The effect of the laboratory performing the age determinations was found to be significant and suggests that perhaps intercalibration of the age readings should have been performed more frequently. It was also found that there is considerable variation in the age composition even within small areas. Three geographical stratifications of the North Sea were compared. The age composition data supports a stratification based on the distribution of the fishery, rather than stratifications based on biological reasoning, although such reasoning is believed to better reflect the sub-structure of the North Sea sandeel population. The analysis also indicated that there are considerable undetected sources of variation resulting in a large and significant overdispersion.

A model resulting from combining the model for the continuation-ratio logits and the adjacent-ratio logits with the often used deterministic differential equation to describe the population dynamics, has been discussed. It was shown that the mortalities may be estimated from the age composition data alone.

Catch at age numbers and the associated uncertainties have been estimated, by separating the statistical analysis into analyses of the separate data sources. The results were combined into estimates of the catch at age numbers and the associated uncertainties for the sandeel landings from the North Sea in 1989 and 1991. Besides uncertainty of the age composition, catch at age numbers also is influenced by uncertainty of the catch per area and the species composition of the catches, and uncertainties associated with the transformation of the unit of measurement of the size of the catch from tonnes to numbers. By establishing the significance of factors that might influence the catch composition, common structures may be recognised and utilised, and when e.g. geographical or temporal differences in the catch compositions are of importance, they may be taken into account in the model. Thereby improved estimates of the catch composition and the associated uncertainty may be obtained. In addition, the identification of the common structures has the advantage that qualified estimates may be provided if some data are missing. Also more reliable predictions may be performed.

The major source of uncertainty in the catch at age is caused by uncertainties in the estimation of the age composition. The estimation is particularly difficult because of large variations in the age composition between small areas. The species composition was estimated using a compound distribution to account as well for the inaccurate definition of the sandeel fishery as for by-catches. The analysis of the species composition of the landings showed that the most important factor to explain misclassifications within the sandeel fishery is the mesh size, an information not utilised today.

For the case where the age composition data was stratified on length groups, a method for analysing sources of variation and their magnitude was presented. However, the method has the drawback that differences in growth may influence the analysis and that information on the shape of the length distribution is not utilised in the analysis to resolve the length distribution into separate length distributions for the age groups.

Finally, modelling the stock dynamics of sandeel by means of stochastic differential equations has been discussed. The approach extends the classical dynamical modelling by means of deterministic differential equations that is believed to describe the main dynamics of the stock. The discussion indicates that it may be possible to estimate quantities such as fishing mortalities and stock abundances by means of contemporary statistical methods, thereby modelling the variation of availability and fishing mortalities through the year.

Appendix A

Using Continuation-ratio Logits to Analyse the Variation of the Age-composition of Fish Catches

Trine Kvist, Henrik Gislason, Poul Thyregod

Keywords: generalised linear mixed models, continuation-ratio logits

Abstract

Major sources of information for the estimation of the size of the fish stocks and the rate of their exploitation are samples from which the age-composition of catches may be determined. However, the age-composition in the catches often varies as a result of several factors. Stratification of the sampling is desirable because it leads to better estimates of the age-composition and the corresponding variances and covariances. The analysis is impeded by the fact that the response is ordered categorical. The paper introduces an easily applicable method to analyse such data. The method combines continuation-ratio logits and the theory for generalised linear

mixed models. Continuation-ratio logits are designed for ordered multinomial response and have the feature that the associated likelihood splits into separate terms for each category level. Thus, the continuation-ratio logits may be modelled as if they were associated binomial distributions which were independent of each other. Thus, generalised linear mixed models can be applied separately to each level of the logits. The method is illustrated by the analysis of age-composition data collected from the Danish sandeel fishery in the North Sea in 1993. The significance of possible sources of variations is evaluated and formulae for estimating the proportions of each age group, and their variance-covariance matrix, are derived.

A.1 Introduction

Most of the methods used to assess the size of fish stocks and the rate of their exploitation are based on regular estimates of the age composition of the catch. These estimates are derived from samples taken at random from the catch or by a stratified random sampling scheme. The catch samples are sorted into species, the numbers of individuals of each species are counted and the individuals are measured and their age determined by counting the number of growth rings in hard parts such as otoliths. The age-composition will vary from sample to sample. This variation can be caused by a multitude of factors, including spatial or temporal differences in catch composition and errors in the age-determination itself. Modelling the age proportions by means of the explanatory variables improves the estimates of the age proportions and reduces the uncertainty associated with the estimation. Another advantage is that such a model of the age-composition and the sources and magnitude of sampling variation can be used to optimise the sampling scheme under stratified sampling.

The data collected for determining the age-composition may be considered as an ordered categorical response, consisting of the number of individuals in each age group. In the rare case where only two age groups are present, the response is binary. In this case, the proportion of the fish in each age group can be transformed by a logit transformation and the sources of variation analysed by standard tools such as generalised linear models (McCullagh and Nelder, 1989) or by extended generalised linear mixed models (Breslow and Clayton, 1993, Wolfinger and O'Connel, 1993). However, with more than two age groups present, the response is multinomial. In this case, a standard logit transformation cannot be used.

In this paper, we show how continuation-ratio logits (Agresti, 1990) can be utilised to analyse the variation of multinomial age-composition samples. In the multinomial case, a response probability is described by several logits. Continuation-ratio logits have the particular feature that the different logits for a response can be regarded as logits for independent binomially distributed data. Each of the logits can then be analysed separately by means of a generalised linear mixed model.

The generalised linear mixed models described by Breslow and Clayton (1993) and Wolfinger and O'Connel (1993) assume the random effect to be normally distributed on the transformed scale. Thus, in the case of binomially distributed data, and using a logit link, the random effect is normally distributed on the logit scale.

In order to illustrate the method, we apply it to age-composition data collected from the Danish sandeel fishery in the North Sea in 1993. The significance of possible sources of variations is evaluated and formulae for estimating the proportions of each age group, and their variance-covariance matrix, are derived.

A.2 Model

The response variable is the number of fish of the species of interest in each age group observed in the sample, $\mathbf{X}_s = (X_{Rs}, \ldots, X_{As})$, where s denotes the sample number, R denotes the youngest age group represented in the catches and A the oldest. An age group is comprised of fish spawned in the same year, except for age group A which often consists of fish of age A and older. Assuming that a sample is representative of the age-composition in the catch, and that the species composition does not influence the age-composition of a particular species, the number of individuals of that species in each age group in a sample, $\mathbf{X}_s$, is distributed according to a multinomial distribution:

$$\mathbf{X}_s \in \mathrm{Mult}(n_s, p_{Rs}, \ldots, p_{As}) \tag{A.1}$$

where n_s denotes the sample size and p_j denotes the proportion of individuals in the catch classified as belonging to age group j, $j = R, \ldots, A$. $A-R$ probabilities are necessary to describe the distribution, since $\sum_{j=R}^{A} p_j = 1$.

The p_j's describe the age-composition of the catches if the age-determination is unbiased. If such a bias exists, the proportion p_j describes the proportion of fish in the catch *classified* into age group j.

A set of explanatory variables is associated with each sample. An explanatory variable could be the position of the fishery, the time of fishery or information on the age-determination, such as the laboratory technician or laboratory performing the analyses.

The age-composition in the samples is modelled by means of continuation-ratio logits. The number of continuation-ratio logits necessary to describe the distribution of $\mathbf{X}_s$ is equal to the number of probabilities, $A - R$. The first logit describes the odds of age R of a sampled fish given that the age is at least R. The second logit describes the odds of age $R+1$ of a sampled fish, given that the age is at least $R+1$, etc..

The definition is (Agresti, 1990):

$$L_j = \log\left(\frac{\pi_j}{1-\pi_j}\right) \tag{A.2}$$

where $j = R, \ldots, A-1$. j denotes the age group and π_j is the conditional probability of age j given that the age is at least j:

$$\pi_j = \frac{p_j}{p_j + \ldots + p_A} \tag{A.3}$$

The continuation-ratio logit can also be described as the log of the ratio between the probability of age j of a sampled fish and the probability of an older age: (A.2) can also be expressed as:

$$L_j = \log\left(\frac{p_j}{p_{j+1} + \ldots + p_A}\right) \tag{A.4}$$

By analogy with the theory for generalised linear mixed models (Breslow and Clayton, 1993, Wolfinger and O'Connel, 1993), the logits were modelled as a linear function of explanatory variables:

$$L_j = \mathbf{b}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{u}_j \tag{A.5}$$

where $\mathbf{b}$ denotes the explanatory variables associated with the fixed parameters $\boldsymbol{\beta}$ and $\mathbf{Z}$ the explanatory variables associated with the random parameters $\mathbf{u}$. The random parameters are assumed to be normally distributed. If the random parameters are omitted, the model is a generalised linear model, as described in McCullagh and Nelder (1989).

A model for L_R describes the ratio between the proportion of R-year-olds and the proportion of older fish in the catches by means of the explanatory variables. By analysing the estimated effects, significant sources of variation in the relative number of recruits can be identified. For instance, time periods and geographical areas with similar relative recruitment may be identified and the magnitude of variation between geographical areas or time periods that have different relative recruitment may be estimated (provided appropriate explanatory variables). As regards possible errors in the age-determination, only errors in the distinction between recruits and older fish influence the model for the first logit.

A model for L_j, $j = R+1, \ldots, A-1$, describes the ratio between the proportion of j-year-olds and the proportion of older fish. A model for L_j only concerns fish of age j and above and is thus unaffected by the proportion of younger fish. Analogous to the logit of the first age level, the significance of effects can be evaluated and geographical areas and time periods with similar ratio between proportions of age group j and older age groups and the magnitude of important sources of variation can be determined. Regarding age-determination errors, only errors in distinguishing between j-year-olds and younger fish and j-year-olds and older fish influence the model for logit L_j.

Logits of different age levels might have different sources of variation and common sources of variation might be of different magnitude. A hypothetical situation where this could occur could be where the age-determination of younger ages is easy to perform and only seldomly subject to error, but, as the fish get older, the age-determination gets more uncertain, difficult and subjective. In this case, the laboratory or laboratory technician effects may be insignificant for L_R and increase as the age level of the logit increase. Similarly, if the recruits are inhomogenously geographically distributed, but more homogenously distributed as they grow older because of migration, the geographical variation will decrease as the age level of the logit increases.

The continuation-ratio logits are estimated independently of each other. The logits can be considered as logits of probabilities connected to $A - R$

independent binomially distributed variables:

$$X_j|X_R = x_R, \ldots, X_{j-1} = x_{j-1} \in \mathrm{Bin}(n_j, \pi_j) \tag{A.6}$$

where $j = R, \ldots, A-1$, $\pi_R, \ldots, \pi_{A-1}$ are defined in (A.3) and $n_R, \ldots, n_{A-1}$ are:

$$n_j = x_j + \ldots + x_A \tag{A.7}$$

$j = R, \ldots, A-1$.

The factorisation of the likelihood can be proved by showing that the simultaneous frequency distribution for $X_R, \ldots, X_{A-1}$ can be written as a product of the frequency distribution of each of the conditioned variables in (A.6):

$$\begin{aligned} & f(x_R, \ldots, x_A) = \\ & f(X_{A-1} = x_{A-1}|X_{A-2} = x_{A-2}, \ldots, X_R = x_R) \\ & \qquad \times \ldots \times \\ & \qquad f(X_{R+1} = x_{R+1}|X_R = x_R) \times \\ & \qquad f(X_R = x_R) \end{aligned} \tag{A.8}$$

The factorisation does not apply if dependency between the parameters $\pi_R, \ldots, \pi_{A-1}$ is imposed upon them through the model specification, eg. by a common parameter.

The factorisation is very useful in model fitting and testing. As long as the parameters in the models for different levels of categories are distinct from each other, the models can be fitted separately using methods for binomially distributed variables, e.g. generalised linear mixed models.

The estimates of the unconditioned probabilities, $p_R, \ldots, p_A$, can be obtained from $\pi_R, \ldots, \pi_A$, and the variances and covariances can be obtained by considering the Taylor approximation of $\widehat{p}_j = f(\widehat{\pi}_R, \ldots, \widehat{\pi}_j)$.

The application of continuation-ratio logits to age-composition data is illustrated by applying the model to age-composition data for sandeels in the North Sea.

A.3 Example

A.3.1 Background

The lesser sandeel(*Ammodytes marinus* Raitt) is one of the most abundant fish species in the North Sea (Sparholt, 1990). It is a small slender fish with a maximum length of approximately 25 cm, and constitutes an important prey for many species of fish, seabirds and marine mammals (Daan *et al.*, 1990) and (Wright, 1996). In addition, it forms the main target of the Danish industrial fishery (Gislason and Kirkegaard, 1998). From 1977 onwards, the annual landings of sandeels from the North Sea have fluctuated between 0.5 and one million tonnes. The landings are processed as fishmeal and oil or used directly as an animal foodstuff.

The sustainability of the sandeel fishery has been subject to intense debate and discussion. On one hand, the present assessment of the impact of the fishery suggests that the fishery is sustainable (ICES, 1996). On the other hand, environmental organisations argue that the uncertainties in the assessment are so large that it is reasonable to fear that the fishery could lead inadvertently to a stock collapse. However, so far, nobody has estimated these uncertainties.

The size of the sandeel stock and the impact of the fishery are assessed regularly by the International Council of Exploration of the Sea (e.g. ICES, 1996). The assessment relies heavily on the estimated age-composition of the landings. Preliminary investigations by Lewy (1995), suggested that the coefficient of variation of the estimated numbers caught at age is low, but his analysis did not include estimates of the variation between samples. It is therefore interesting to use continuation-ratio logits to investigate the magnitude of variation and its possible sources.

A.3.2 Data

The age-composition of the sandeel landings is estimated from samples collected by the Danish fisheries inspection. Samples are taken at random from the landings by lowering a 10-litre pail into the hold of the vessel. Some of these samples are analysed at the Danish Institute for Fisheries Research where the age-composition of the sandeels is determined from age reading of otoliths (Lewy, 1995). Data are available for several years. We have restricted ourselves to an analysis of the data collected in 1993.

Several factors may influence the age-composition in the samples (Gislason and Kirkegaard, 1998). The distribution of the sandeel fishery in the North Sea is patchy. Adult sandeels bury themselves in the sediment at night and during winter and are therefore mostly found in areas of coarse well-oxygenated sand. Presumably there is little migration of adult sandeels between the various sandeel grounds in the North Sea, and regional differences in age-composition can therefore be expected. The fishery is highly seasonal. In general, it peaks during spring and early summer. Because of the burrowing behaviour of the adults, the catch rates vary between different age groups, with season and during the day (Reeves, 1994). In addition, the catch rates will be influenced by changes in tide and weather.

Insufficient information is available to investigate all of these potential sources of variation. The date and the approximate position of the catch is recorded by 30*30 square nautical miles rectangles, but no information about time of day, sediment type and position of individual hauls is available. The primary explanatory variables selected in this analysis were therefore month, MON, and rectangle, SQ. To investigate differences in age-composition between larger geographical areas, the rectangles were assigned to seven sandeel areas, AR, and into the northern and southern North Sea, NS. The size and location of the sandeel areas were based on the overall distribution of the fishery. Months with similar age-composition were furthermore grouped into periods, PER, based on the results of the analysis.

In all, the data used in the analysis consist of 70 samples, representing 36 rectangles and the 11 months from February to November. The following number of samples were collected in each month:

Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec
1	7	9	13	9	14	7	3	2	3	2

The geographical distribution of the samples is shown in figure A.1. The number of sandeels in each sample generally varies between 80 and 1000.

A.3.3 Model

The sandeels in the sample have been grouped into five age groups, $0, \ldots, 4+$, where group 4+ includes sandeels classified as being four years

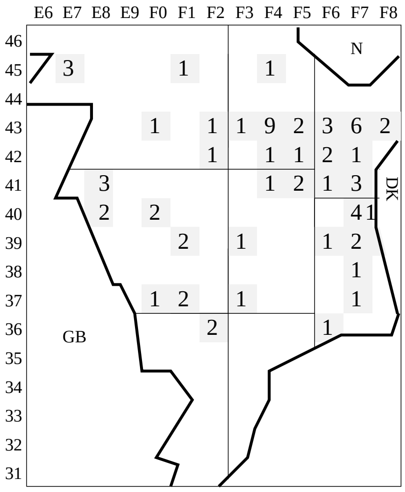

Figure A.1: *The geographical distribution of samples into the seven areas. The number in the shaded squares indicates the number of samples from the square included in the analysis.*

old or older. It is desirable that the samples provide an unbiased picture of the age-composition of the sandeel in the catch. The collection of a fixed volume of sandeels, rather than a fixed number, might introduce a bias, i.e. an expected age-composition in the sample differing from that in the catch (see Schaeffer, 1969). This bias could occur if the volume of one sandeel was large relative to the sampling volume. However, because the volume of one sandeel is small relative to the total sample volume, we do not expect such a bias to influence our results. Similarly, the amount of other species in the sandeel catches is small and we do not expect that the species composition of the sample will influence the age-composition of the sandeel population in the catch. Age-determination errors may result in biases, but the current dataset does not allow for this problem to be addressed. Therefore, it is assumed that the proportions $p_{0,s}, \ldots, p_{4+,s}$, represent the age-composition in the catch.

The continuation-ratio logits for the response are:

$$L_{0s} = \log \frac{p_{0s}}{p_{1s} + \ldots + p_{4+s}} \tag{A.9}$$

$$L_{1s} = \log \frac{p_{1s}}{p_{2s} + \ldots + p_{4+s}} \tag{A.10}$$

$$L_{2s} = \log \frac{p_{2s}}{p_{3s} + p_{4+s}} \tag{A.11}$$

$$L_{3s} = \log \frac{p_{3s}}{p_{4+s}} \tag{A.12}$$

(A.13)

The models for each age level are analysed separately by means of generalised linear mixed models. The explanatory variables are NS, AR, SQ, PER and MON, as described above in section A.3.2.

The full models for the logits are:

$$\begin{aligned} L_{ijklmo} = {} & \mu_i + \mathsf{PER}_{ij} + \mathsf{NS}_{ik} + \mathsf{MON(PER)}_{il(j)} + \mathsf{AR(NS)}_{im(k)} + \\ & \mathsf{SQ(AR*NS)}_{io(mk)} + \mathsf{PER*NS}_{ijk} + \mathsf{PER*AR(NS)}_{ijm(k)} + \\ & \mathsf{PER*SQ(AR*NS)}_{ijo(mk)} + \mathsf{MON*NS(PER)}_{ilk(j)} + \\ & \mathsf{MON*AR(PER*NS)}_{ilm(jk)} + \\ & \mathsf{MON*SQ(AR*NS*PER)}_{ilo(mkj)} \end{aligned} \tag{A.14}$$

where

i	$=$	$0,\ldots,4+$	age level of logit
j	$=$	$1,\ldots,e_i$	number of period
e_i			number of periods for age level i
k	$=$	$1,2$	part of the North Sea
l	$=$	$2,\ldots,12$	number of month
m	$=$	$1,\ldots,7$	number of area
o	$=$	$1,\ldots,f_m$	square number within area m
f_m			number of squares in area m

The effects might be interpreted as:

μ_i Overall mean of the logit.

$\mathsf{PER}_{ij}, \mathsf{MON(PER)}_{il(j)}$ Period and month effect describing how the age-composition at the same geographical position may vary through the year.

NS_{ik} The effect of differences in the overall age-composition between the northern and southern part of the North Sea.

$\mathsf{AR(NS)}_{im(k)}$ The effect of differences in the overall age composition between areas within the northern and the southern part of the North Sea. AR is nested within NS.

$\mathsf{SQ(AR*NS)}_{io(mk)}$ The effect of differences in the overall age-composition between squares within areas. SQ is nested within AR.

$\mathsf{PER*NS}_{ijk}, \mathsf{PER*AR(NS)}_{ijm(k)}, \mathsf{PER*SQ(AR*NS)}_{ijo(mk)}$ Interactions between period of the year and geographical parts of the North Sea. The interaction effects indicate that the period effects may vary between geographical positions.

$\mathsf{MON*NS(PER)}_{ilk(j)}, \mathsf{MON*AR(PER*NS)}_{ilm(jk)}$,
$\mathsf{MON*SQ(AR*NS*PER)}_{ilo(mkj)}$ Interactions between months within period and geographical parts of the North Sea. The interpretation of these interaction effects is analogous to the interactions described above.

The model is heavily parameterised. The variation between squares is assumed to be random, because a model with one or more parameters per square will contain too many parameters to be of any practical use. Therefore, the square parameters and parameters describing interactions with the square effect are modelled as being normally distributed whenever they are found to be significant.

$$
\begin{aligned}
\mathsf{SQ}(\mathsf{AR} * \mathsf{NS})_{in(mk)} &\in \mathrm{NID}(0, \sigma^2_{\mathsf{SQ},i}) \\
\mathsf{PER} * \mathsf{SQ}(\mathsf{AR} * \mathsf{NS})_{ijo(mk)} &\in \mathrm{NID}(0, \sigma^2_{\mathsf{PER*SQ},i}) \\
\mathsf{MON} * \mathsf{SQ}(\mathsf{AR} * \mathsf{NS} * \mathsf{PER})_{ilo(mkj)} &\in \mathrm{NID}(0, \sigma^2_{\mathsf{MON*SQ},i})
\end{aligned}
$$

where the indices are defined as in (A.14).

Besides the explanatory variables, a dispersion parameter, ϕ, is included in the model to account for the variance that cannot be attributed to the binomial variance or the explanatory variables. The dispersion parameter enters as a simple multiplicative parameter on the binomial variance, and hence it must be greater than zero. $\phi = 1$ indicates that the variance is in accordance with the assumed distribution. $\phi < 1$ indicates that there is underdispersion and $\phi > 1$ indicates overdispersion. The dispersion parameter has been described, e.g. in McCullagh, 1989.

The mixed models were fitted using restricted pseudo-likelihood, REPL (Wolfinger and O'Connel, 1993). The procedure is implemented in a SAS version 6.12 macro called GLIMMIX described by Littell. *et al.* (1996). The significance of effects are tested by approximate F-tests based on the Wald statistics, described by Littell *et al.* (1996), p. 437. The significance level is chosen to be 5%.

A.3.4 Results

The approximate F-test for the various effects indicated that the period effect, PER, is significant for all age levels. However, the periods are defined differently for different age levels with different parameter estimates. The northern-southern effect, NS, is significant for age levels zero and one, but not for ages two and three. As for the period effect, the parameter estimates are different for different age levels. The square effect, SQ(NS), for age level zero and one, and SQ for age level two and three is significant. These random effects are modelled by normal distributions. For age levels zero and one, the square parameters from the northern and the southern part are normally distributed with mean zero (and the overall levels in the two parts are modelled by the fixed effects) and the same variance for the northern and southern part, $\sigma^2_{\mathsf{SQ(NS)}i}$. i denotes the age level. For age level two and three, which have the same overall level for the whole North

Sea, the square effect is modelled by a normal distribution with zero mean applied to the whole North Sea. None of the area effect, AR, or interaction effects between temporal effects and geographical effects is significant. As an example of the significance of the different effects, the approximate F-tests for age level two are shown (table 1). Insignificant effects have been removed from below and upwards. Effects above the square effect have been tested in a model where the square effect is assumed to be random. Reading the table, it becomes apparent that the square effect is by far the most significant effect.

Table 1. *Approximate F-tests for age level two.*

Effect	NDF	DDF	Type III F	p-value
PER	1	37	4.83	0.03
NS	1	22	0.47	0.50
NS $*$ PER	1	41	3.34	0.08
AR(NS)	5	14	0.98	0.46
AR $*$ PER(NS)	4	30	2.19	0.09
M(PER)	4	10	2.17	0.15
M $*$ NS1(TID)	2	9	1.43	0.29
S(A1 $*$ NS1)	18	10	7.68	0.00
S $*$ TID(A1 $*$ NS1)	.	.	.	.
M $*$ A1(NS1 $*$ TID)	.	.	.	.
M $*$ S(AR $*$ NS $*$ PER)	1	8	3.28	0.11

The final models for the logits suggested by the test are:

$$L_{0jko} = \mathsf{PER0}_j + \mathsf{NS}_{0k} + \mathsf{SQ(NS)}_{0o(k)} \tag{A.15}$$
$$L_{1jko} = \mathsf{PER1}_j + \mathsf{NS}_{1k} + \mathsf{SQ(NS)}_{1o(k)} \tag{A.16}$$
$$L_{2jo} = \mathsf{PER2}_j + \mathsf{SQ}_o \tag{A.17}$$
$$L_{3jo} = \mathsf{PER3}_j + \mathsf{SQ}_o \tag{A.18}$$

where

$i = 0, \ldots, 3$	age level
$k = 1, 2$	1: Northern part, 2: Southern part
$o = 1, \ldots, 18$	square number for $i = 0$ and $k = 1$
$o = 1, \ldots, 18$	square number for $i = 0$ and $k = 2$
$o = 1, \ldots, 14$	square number for $i = 1$ and $k = 1$
$o = 1, \ldots, 17$	square number for $i = 1$ and $k = 2$
$o = 1, \ldots, 26$	square number for $i = 2$
$o = 1, \ldots, 23$	square number for $i = 3$

$\mathsf{PER0}_1$:	Months	2, ..., 5
$\mathsf{PER0}_2$:	Month	6
$\mathsf{PER0}_3$:	Months	7, ..., 12
$\mathsf{PER1}_1$:	Month	2, 3
$\mathsf{PER1}_2$:	Months	4, ..., 12
$\mathsf{PER2}_1$:	Months	3, ..., 6
$\mathsf{PER2}_2$:	Months	7, ..., 12
$\mathsf{PER3}_1$:	Month	2, 3, 4
$\mathsf{PER3}_2$:	Months	5, 6
$\mathsf{PER3}_3$:	Months	7, ..., 12

and

$$\begin{aligned} \mathsf{SQ(NS)}_{io(k)} &\in \mathrm{N}(0, \sigma^2_{\mathsf{SQ(NS)}i}) \quad i = 0, 1 \\ \mathsf{SQ}_{io} &\in \mathrm{N}(0, \sigma^2_{\mathsf{SQ}i}) \quad i = 2, 3 \end{aligned}$$

Parameter estimates of the fixed effects including standard errors are shown in table 2, using the standard parameterising method as the default for procedure GLM, Mixed in SAS (SAS, 1996). The method solves the often occuring problem of overparameterising by selecting one of the levels of an effect as a reference level.

Table 2. *Estimates of fixed effect parameters, including std (in parantheses).*

Age Level	Intercept	N.-S. Effect		Period Effect		
		Northern	Southern	1	2	3
0	-1.32	5.02	0	$-\infty$	-3.59	0
	(0.95)	(1.23)	(.)	(.)	1.32	(.)
1	-1.80	2.43	0	4.91	0	-
	(0.57)	0.86	(.)	(1.63)	(.)	-
2	0.15	-	-	0.89	0	-
	(0.41)	-	-	(0.40)	(.)	-
3	0.48	-	-	-1.22	2.06	0
	(0.43)	-	-	(0.43)	(0.47)	(.)

The estimate of the first period of the proportion of 0-year-olds is set to $-\infty$ because this age group does not occur in the catches at this early period of the year.

The estimates of the dispersion parameter and the variance components are shown in table 3.

Table 3. *Estimates of variance component and dispersion parameter.*

Age Level	Var. Comp.	Disp. Par
0	5.6	15
1	3.6	26
2	1.4	4
3	2.1	0.2

A.3.5 Discussion of results

The significance of the period effect, PER, indicates that the age-composition of the catches does in fact vary through the year. Months with similar effects could be identified, but the similarities identified for one age group did not apply to other age groups. The classification into periods with similar effects for the various age groups is illustrated in figure A.2.

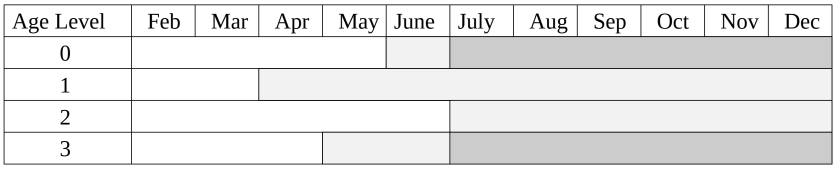

Figure A.2: *Amalgamation of the months into periods with similar effects.*

The estimates of the fixed effect parameters, including 95% confidence intervals, are shown in figure A.3. Here, the changes of the conditional proportions on the logit scale are illustrated for the different age levels. The proportion of 0-year-olds is zero until late summer and autumn when it increases (the parameter estimates for the first period of the year are $-\infty$, because there are no occurrences of 0-year-olds in the catches in this period). The conditional proportions of 1- and 2-year-olds decrease as the proportion of older sandeels increases through the year. The relative proportion of 3-year-olds compared to older sandeels seems to be somewhat more complicated to interpret. The model indicates that the relative proportion of 3-year-olds during the year increases and then decreases again.

The northern-southern effect, NS, is significant for age level 0 and 1, but not for age level 2 and 3, a result which indicates that there are differences between the northern and southern parts of the North Sea regarding the proportions of young sandeels, but not regarding older sandeels. The conditional proportions are larger for the northern part than the southern part (refer to figure A.3).

The insignificance of the area effect, AR, indicates that there are no substantial differences between areas within neither the northern nor the southern parts of the North Sea for either of the age levels.

There are differences between squares for all age levels. The square effect, SQ/SQ(NS), is modelled as random, following a normal distribution on the logit scale. The estimates of the variance components for the different age levels are shown in figure A.4, together with the dispersion parameter, ϕ. The variance component tends to decrease as the age level increases. Thus, there is a tendency towards larger variation between squares the smaller the age level. The same tendency is seen for the dispersion parameter,

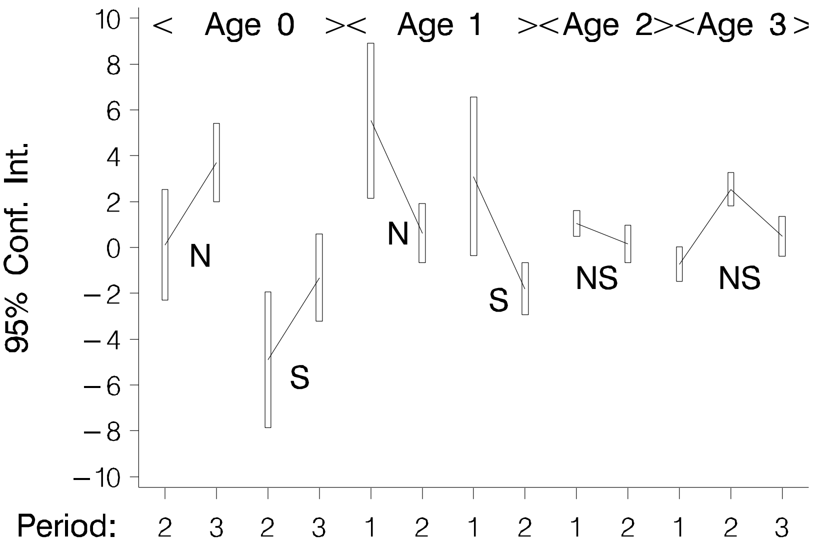

Figure A.3: *95% confidence interval for the fixed effects on the logit scale. 'N' denotes the Northern Part, 'S' denotes the Southern part and 'NS' denotes both parts. Subsequent periods are joined by a line.*

ϕ. The dispersion parameter gives an indication of the variation which cannot be explained by the binomial sampling variance and the explanatory variables available. Thus, there are additional factors that influence the age-composition of the catches. Possible factors could be those mentioned in section A.3.2, such as uncertainties in age-determination, differences between the age-compositions on the different fishing grounds within a square, fishery on different times of the day, tide and weather.

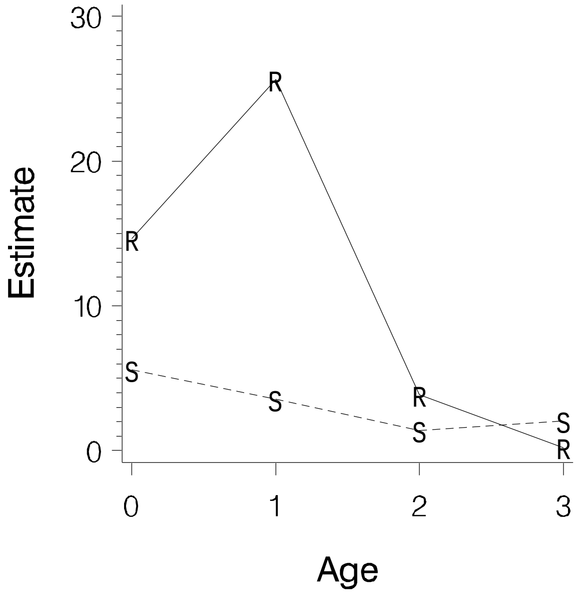

Figure A.4: *Covariance parameters. 'S' denotes Square effect. 'R' denotes dispersion parameter.*

A.3.6 Estimation of proportions of each age group

Although it has been advantageous to perform the assessment of the importance of the various effects on the continuation ratio logit logit-scale, the quantities of practical interest are the proportions of the catch in the various age groups and the uncertainties associated with these estimated proportions.

Thus, the linear mixed model for the continuation ratio logits L_0, L_1, L_2 and L_3 has to be translated to the resulting model for the proportions p_0, p_1, p_2, p_3 and p_{4+} in the various age groups.

First, the continuation-ratio logits are transformed to the conditional proportions π_0, π_1, π_2, and π_3. The approximate mean and variance of $\pi_0, \ldots, \pi_3$ are:

$$\mathrm{E}\{\widehat{\pi}_i\} \approx \frac{\exp(\widehat{L}_i)}{1+\exp(\widehat{L}_i)} \tag{A.19}$$

and

$$\mathrm{V}\{\widehat{\pi}_i\} \approx \frac{\exp(2\widehat{L}_i)}{(1+\exp(\widehat{L}_i))^4}\mathrm{V}\{\widehat{L}_i\} \tag{A.20}$$

It should be noted that, although the logit transformation is non-linear, the effect from the random parameters does not affect the first order approximation to the mean value of the π's, but only the first order approximation to the variance of the π's, viz. Furthermore, the different $\widehat{\pi}_i$'s are independent of each other, as long as they do not have any parameters in common (shown in (A.8)).

The transformation from the conditioned probabilities, $\pi_0, \ldots, \pi_3$ to proportions in age groups is given by the following relations:

$$\begin{aligned}
p_0 &= \pi_0 \\
p_1 &= \pi_1(1-p_0) \\
p_2 &= \pi_2(1-(p_0+p_1)) \\
p_3 &= \pi_3(1-(p_0+p_1+p_2)) \\
p_{4+} &= 1-p_0-p_1-p_2-p_3
\end{aligned} \tag{A.21}$$

The variances and covariances are estimated by using the Taylor approximation for a product of independent variables:

$$\mathrm{V}\{\prod_{i=1}^{n}\widehat{\pi}_i\} \approx \sum_{i=1}^{n}\left[\mathrm{V}\{\widehat{\pi}_i\}\prod_{j=1}^{i-1}\widehat{\pi}_j^2\prod_{k=i+1}^{n}\widehat{\pi}_k^2\right] \tag{A.22}$$

The proportions of each age group for the different periods of the year and parts of the North Sea and the variances and covariances can be estimated by using (A.21) and (A.22). The procedure is illustrated for the first few variances and covariances.

$$
\begin{aligned}
\mathrm{V}\{\widehat{p}_0\} &= \mathrm{V}\{\widehat{\pi}_0\} && \text{(A.23)}\\
\mathrm{V}\{\widehat{p}_1\} &\approx (1-\widehat{p}_0)^2\mathrm{V}\{\widehat{\pi}_1\} + \widehat{\pi}_1^2\mathrm{V}\{\widehat{p}_0\} && \text{(A.24)}\\
\mathrm{Cov}\{\widehat{p}_0,\widehat{p}_1\} &\approx -\widehat{\pi}_1\mathrm{V}\{\widehat{\pi}_0\} && \text{(A.25)}\\
\mathrm{V}\{\widehat{p}_2\} &\approx (1-\widehat{p}_0-\widehat{p}_1)^2\mathrm{V}\{\widehat{\pi}_2\} + && \text{(A.26)}\\
& \quad \widehat{\pi}_2^2(\mathrm{V}\{\widehat{p}_0\} + \mathrm{V}\{\widehat{p}_1\} + 2\mathrm{Cov}\{\widehat{p}_0,\widehat{p}_1\}) && \text{(A.27)}\\
\mathrm{Cov}\{\widehat{p}_0,\widehat{p}_2\} &\approx -\widehat{\pi}_2(\mathrm{V}\{\widehat{p}_0\} + \mathrm{Cov}\{\widehat{p}_0,\widehat{p}_1\}) && \text{(A.28)}\\
\mathrm{Cov}\{\widehat{p}_1,\widehat{p}_2\} &\approx -\widehat{\pi}_2(\mathrm{V}\{\widehat{p}_1\} + \mathrm{Cov}\{\widehat{p}_0,\widehat{p}_1\}) && \text{(A.29)}
\end{aligned}
$$

The estimated age-compositions of the catches through the year for the northern and southern parts are shown in figure A.5 and A.6.

As an example, the estimated proportions and the prediction error variances, covariances and correlations for the southern part of the North Sea in May are shown:

$$
\widehat{\mathbf{p}}_{May,Southern} = \begin{pmatrix} 0.00 \\ 0.14 \\ 0.63 \\ 0.21 \\ 0.02 \end{pmatrix} \tag{A.30}
$$

$$
\widehat{\mathbf{\Sigma}}_{\mathbf{May},\mathbf{Southern}} = \begin{pmatrix} 0 & & & & \\ 0 & 0.06 & & & \\ 0 & -0.04 & 0.07 & & \\ 0 & -0.01 & -0.03 & 0.04 & \\ 0 & -0.001 & -0.002 & 0.002 & 0.00 \end{pmatrix} \tag{A.31}
$$

$$
\widehat{\boldsymbol{\rho}}_{\mathbf{May},\mathbf{Southern}} = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \\ 0 & -0.66 & 1 & & \\ 0 & -0.29 & -0.52 & 1 & \\ 0 & -0.17 & -0.30 & 0.47 & 1 \end{pmatrix} \tag{A.32}
$$

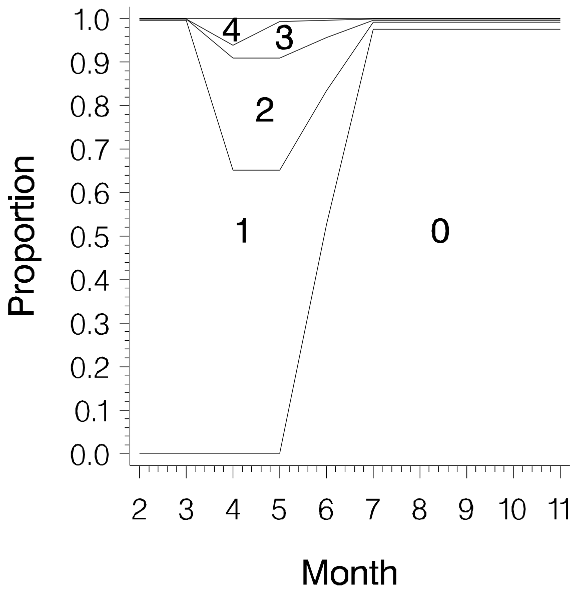

Figure A.5: *Estimated Age-Composition of the catch from the Northern part. The proportion each age group constitutes of the catch is represented by an area. The numbers indicate the age groups.*

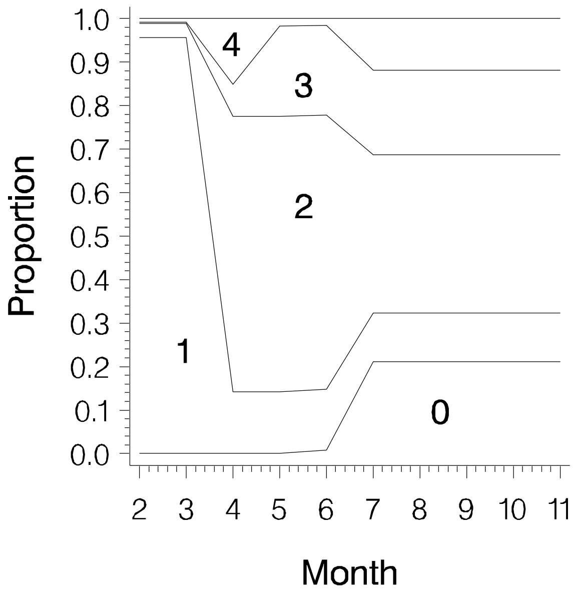

Figure A.6: *Estimated Age-Composition of the catch from the Southern part. The proportion each age group constitutes of the catch is represented by an area. The numbers indicate the age groups.*

The covariance matrix reflects the covariance matrix for multinomial data. The uncertainty is largest for age groups represented in proportions close to 0.50. The uncertainty decreases towards zero for the 4+-year olds that are represented in a proportion close to zero. It should, however, be noted that despite the fact that the estimated proportions in age groups 1 and 3 are far less than the proportion in age group 2, the uncertainties in the middle age groups vary more smoothly than suggested by a crude multinomial model.

The overall feature of the correlation matrix indicates the usual negative correlations between multinomially distributed variables imposed by the constraint that the proportions shall add to 1. Thus, if the proportion in one of the age groups 1, .., 4+ happens to be overestimated, then the proportions in the other age groups will tend to be underestimated. The only exception is for the 3- and 4+ year olds. The estimates for these two age groups are positively correlated. This result may be explained by the fact that 3- and 4+ year olds tend to occur at the same time in the catches.

A.4 Summary and discussion

The size of the fish stocks and the rate of their exploitation are issues of great importance and concern, and therefore the estimates of those quantities and their uncertainties are constantly undergoing attempts at improvement. A major source of information utilised in the estimation is samples from catches to determine the age-composition, their variances and covariances. The age-composition often varies considerably because of a number of factors and an explanation of the variation by means of those factors is desirable to achieve better estimates of the age-composition. However, the stratification is subject to certain difficulties, because the quantities to be estimated are the proportions of each age group and the response the number of individuals in each age group. It is well-known that utilising ordinary linear predictors for rates and proportions may lead to predictions outside the natural parameter space. Moreover, the usual least squares predictors do not account for the fact that the variance depends on the proportion being modelled (the mean value). Generalised linear models for binomial data take those properties into account by utilising a link function and a method of estimation that incorporates the dependency of the variance upon the mean. However, in most cases of age-composition data, the response is not binomial but multinomial. Because the proportions are correlated, it is not valid to model each of the proportions separately, using their corresponding ordinary logits. Instead, continuation-ratio logits are more practical, because the associated likelihood splits into separate terms for each category level, and hence the continuation-ratio logits can be modelled separately by means of generalised linear models. Furthermore, the continuation-ratio logits have the appealing feature that a logit for an age group only concerns fish of that age and above, and thus systematic features for younger age groups do not enter into the model for that age group.

If some of the effects are random by nature, or require a level of detail that is not practical to handle, they may be modelled by utilising the extended generalised linear mixed models. Modelling those effects as random makes it possible to obtain a model of practical use, and to test the significance of higher level effects.

Another strategy of modelling the random effect could be by using the conjugated prior (Consonni and Veronese, 1992). In the binomial case, this corresponds to modelling the proportion by means of a beta-distribution. Use of the conjugated prior has the advantage that the proportion is modelled by a well-known distribution on the response scale itself and hence its parameters may be more easily interpreted. The approach of modelling the effect by a normal distribution on a relevant scale has the advantage of often being more flexible, and the methods for estimation and testing are implemented in commonly used software.

The variation of the age-composition of catches from the Danish sandeel fishery in the North Sea in 1993 has been analysed in order to illustrate the applicability of the method. The significance of possible temporal and geographical sources of variations has been evaluated on the basis of a generalised linear mixed model. The proportion of each age group and the corresponding variance-covariance matrix have been estimated. It was estiblished that the variance-covariance matrix was different from that of a multinomial distribution: the variance and covariance were not directly determined by the size of the proportion. In fact, a positive correlation even occurred between two age groups (age group 3 and 4+) which would never have been the case for the multinomial distribution.

Estimation of the age-composition is often based upon an age-length key. In those cases, the suggested method can still be used. The length or a function of the length then enters as a regression variable. However, if the length distributions for the different age groups exhibit large overlaps, the estimation of the age-length relation becomes complicated. In those cases, methods based on suggestions by Kimura and Chikuni (1987), might be more practical.

A.5 References

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9-25.

Consonni, G. and Veronese, P. (1992). Conjugate priors for exponential families having quadratic variance functions. *Journal of the American Statistical Association*, 87:1123-1127.

Daan, N., Bromley, P., Hislop, J. and Nielsen, N. (1990). Ecology of North Sea fish. *Netherlands Journal of Sea Research*, 26:343-386.

Gislason, H. and Kirkegaard, E. (1998). Is the industrial fishery in the North Sea sustainable? In Symes, D., editor, *Northern Waters: Management Issues and Practice.* Fishing News Books. London.

ICES (1996). *Report of the ICES Advisory Committee on Fishery Management, 1995.* International Council for the Exploration of the Sea.

Kimura, D. K. and Chikuni, S. (1987). Mixtures of empirical distributions: An iterative application of the age-length key. *Biometrics*, 43:23-35.

Lewy, P. (1995). Sampling methods and errors in the Danish North Sea industrial fishery. *Dana*, 11:39-64.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). Sas system for mixed models. Technical Report, SAS Institute Inc.

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models (second edition).* Chapman and Hall (London; New York).

Reeves, S.A (1994). Seasonal and annual variation in catchability of sandeels at Shetland. Technical report, International Council of the Exploration of the Sea, CM 1994/D:19.

SAS (1996). *SAS/STAT Software: Changes and enhancements through Release 6.11.* SAS Institute Inc.

Scheaffer, R.L. (1969). Sampling mixtures of multi-sized particles: an application of renewal theory. *Technometrics* 11:285-298.

Sparholt, H. (1990). An estimate of the total biomass of fish in the North Sea. *Journal du Conseil International pour l'Exploration de la Mer*, 46:200-210.

Wolfinger, R. and O'Connel, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48:233-243.

Wright P.J. (1996). Is there a conflict between sandeel fisheries and seabirds? A case study at Shetland. In Greenstreet, S.P.R. and Tasker, M.L., editors, *Aquatic predators and their prey.* Fishing News Books, Blackweell Science Ltd. Oxford.

Appendix B

Sources of Variation in the Age Composition of Sandeel landings

T. Kvist, H. Gislason and P. Thyregod

Key words: age composition data, continuation-ratio logits, generalised linear mixed models.

Abstract

A new method of analysing age composition data is applied to the lesser Sandeel fishery in the North Sea. The method provides improved estimates of the age composition and the associated uncertainty. The estimates may be utilised in common statistical stock assessment methods and thereby improve the assessments. Furthermore, valuable information concerning the stock dynamics may be obtained by careful analysis of the data, as the age composition of the catches provides information of the age composition of the part of the stock that is available to the fishery. The analyses show that the proportion of older sandeel in the catches is significantly lower in the start and end of the fishing season and that the age composition differ

between laboratories. There is considerable variation in the age composition within small areas, as well as important undetected sources of variation resulting in a large and significant overdispersion.

B.1 Introduction

Catch in numbers at age constitutes the primary input to age-structured assessment models (Megrey, 1989). Sampling errors are known to influence estimates of catch and hence of stock size and fishing mortalities (Rivard, 1989). However, the sources and magnitude of errors in the age composition data have seldomly been studied and quantified. This may partly be due to a lack of suitable methods. The distribution of the number of individuals in different age groups in a sample may be described by a multinomial distribution and no standard methods are available for evaluating the significance of factors influencing the distribution.

In this paper a new method for analysing age composition samples is applied to catch at age data for the lesser sandeel (*Ammodytes marinus* Raitt) in the North Sea. The lesser sandeel is one of the most abundant fish species in this area (Sparholt, 1990). It constitutes an important prey for many species of fish, seabirds and marine mammals (Daan *et al.*, 1990) and forms the main target of the Danish industrial fishery. The sustainability of this fishery has been subject to intense debate and discussions (Wright, 1996; Gislason and Kirkegaard 1998). On one hand, the present assessment of the sandeel stock suggests that the fishery is sustainable (ICES, 1996). On the other hand, environmental organisations argue that the uncertainty in the assessment is so large that the fishery inadvertently could lead to a collapse of the stock. So far the uncertainty has not been quantified.

The method combines the so-called generalised linear mixed models (Breslow and Clayton, 1993; Wolfinger and O'Connel, 1993) with the theory for ordered categorical responses (Agresti, 1990). It transforms the probability of the multinomial response into a product of binomial probabilities for which generalised linear mixed models can be directly applied to study the possible sources of variation. It is particularly suitable for age composition data because it allows individual cohorts to be followed and compared over time. We use the model to evaluate the significance of spatial and temporal differences in the age composition of the sandeel samples as well as to study the importance of differences in age readings between laboratories.

B.2 Methods

The analysis of age composition data is impeded by the categorical structure of the response, which is the number of sandeels classified as belonging to each age group, $\mathbf{X}_s = (X_{0s}, \ldots, X_{4s})$. s denotes the sample number and the age groups are $0, \ldots, 4$, where group 4 includes ages 4 and above. If we assume that the age composition of the sandeel in a particular sample does not depend on the occurence of other species in the sample and that the samples are representative for the age composition in the catch then the response can be modelled by a multinomial distribution:

$$\mathbf{X}_s \in \mathrm{Mult}(n_s, p_{0s}, \ldots, p_{4s}) \tag{B.1}$$

where n_s denotes the sample size and p_{js} denotes the proportion of individuals in the catch classified as belonging to age group j, $j = 0, \ldots, 4$. With five age groups presents four probabilities will be needed to describe the distribution. The p_{js}'s describe the real age composition of the catches if the age determination is unbiased. If a bias exists, the proportion p_{js} describes the proportion of fish in the catch that would be classified into age group j.

A new method for analysing the influence of various factors on age composition data has been presented by Kvist *et al.* (1998, *submitted*). The idea is to split the probability of the multidimensional response into binomial probabilities. This is done by considering the conditional distributions of $X_{0s}, \ldots, X_{3s}$, where the distribution of X_{js} is conditioned on the event that the age is j or higher.

$$X_{js}|x_{js} + \ldots + x_{4s} \in \mathrm{Bin}(X_{js} + \ldots + X_{4s}, \pi_{js}) \tag{B.2}$$

where $j = 0, \ldots, 3$, X_{js} is the number of j-year-olds and π_{js} is the probability of age j given that the age is at least j:

$$\pi_{js} = \frac{p_{js}}{p_{js} + \ldots + p_{4s}} \tag{B.3}$$

The continuation-ratio logits for $\mathbf{X}_s$ are the ordinary logits for the conditional distribution of the variables $X_{0s}, \ldots, X_{3s}$, i.e.

$$L_{js} = \log \frac{\pi_{js}}{1-\pi_{js}} = \log \frac{p_{js}}{p_{(j+1)s} + \ldots + p_{4s}} \tag{B.4}$$

Thus by a logit transformation the interval [0,1] of a probability is conveniently transformed to the interval $]-\infty, \infty[$, which is more practical during estimation. The first logit describes the odds of age 0 of a sampled fish. The second logit describes the odds of age 1 of a sampled fish, given that the age is at least 1, etc..

Because the continuation-ratio logits may be estimated independent of each other as long as they do not have any parameters in common they can be modelled separately by means of generalised linear mixed models (Breslow and Clayton, 1993; Wolfinger and O'Connel, 1993). Thus $\mathbf{X}_s$ is modelled by four continuation-ratio logits of the form:

$$L_j = \mathbf{b}_j \boldsymbol{\beta}_j + \mathbf{Z}_j \mathbf{u}_j \tag{B.5}$$

where $j = 0, \ldots, 3$, $\mathbf{b}_j$ denotes the explanatory variables associated with the fixed parameters $\boldsymbol{\beta}_j$ and $\mathbf{Z}_j$ the explanatory variables associated with the random parameters $\mathbf{u}_j$. The random parameters are assumed to be normally distributed on the logit scale. If the random parameters are omitted the model is a generalised linear model, described in McCullagh and Nelder (1989).

We also included a dispersion parameter, ϕ, to account for the variance that could not be attributed to the binomial variance or the explanatory variables. The dispersion parameter enters as a simple multiplicative factor on the binomial variance, and must therefore be greater than zero. $\phi = 1$ indicates that the variance of the response is in accordance with the nominal binomial variance. $\phi < 1$ indicates that the data is underdispersed, and that the variance of the response is less than the nominal binomial variance. $\phi > 1$ indicates overdispersion, where the variance of the responce exceeds the nominal binomial variance. Introducing a dispersion parameter means that the conditional distributions are no longer exactly binomial:

$$X_{js} | x_{js} + \ldots + x_{4s} \in \widetilde{\mathrm{Bin}}(X_{js} + \ldots + X_{4s}, \pi_{js}, \phi_{js}) \tag{B.6}$$

The dispersion parameter has been described in more detail in e.g. McCullagh and Nelder (1989).

The random variation of the model is thus modelled partly by a dispersion parameter and partly by variance components (from random effects). A variance component describes variation between observations with different probabilities and the dispersion parameter describes variation between observations with the same probabilites. The magnitudes of the two are difficult to compare as they are measured on different scales, but the interpretation of the dispersion parameter and the variance components can be illustrated further by considering the following simple example.

Assume X is binomially distributed with an associated dispersion parameter, ϕ:

$$X \in \widetilde{\mathrm{Bin}}(n, p, \phi) \tag{B.7}$$

where $\mathrm{E}[p] = p_0$, $l = \log(p/(1-p))$, and $\mathrm{V}[l] = \sigma^2$.

The variance of the observation X/n can then approximately be expressed as:

$$\mathrm{V}\left[\frac{X}{n}\right] \approx p_0(1-p_0)\left[p_0(1-p_0)\sigma^2 + \frac{\phi}{n}\left(1 - p_0(1-p_0)\sigma^2\right)\right] \tag{B.8}$$

The first factor of the expression describes the basic binomial variance structure. The first term within the square brackets describes the variation between observations with different p's (transformed from the logit scale to the probability scale), and the last term describes the average variation between observations with the same p (because of the convexity of $p(1-p)$ this average variation will be less than $\phi p_0(1-p_0)$). Note that if the variance component is zero the variance reduces to the variance, $p_0(1-p_0)\phi/n$, corresponding to a binomial distribution with a dispersion parameter. Note also that according to (B.8), an increase of the sample size will reduce the contribution from the dispersion parameter, but not the contribution from the random effect.

B.3 Materials

The Danish fisheries inspection collects samples from the sandeel landings in the major landing ports. Samples are taken at random by lowering a

10-litre pail into the hold of the vessels. The samples are sorted into species and the age composition of sandeel determined by reading the age of the otoliths. At present the method requires that a random subsample of fish are aged. This was the case before 1993. After 1993 the procedure changed and from then on only a fixed limited number of fish from each length group was aged. The analysis was therefore restricted to samples from the period between 1984 and 1993. In this period a total of 700 samples were collected from the fishery, each sample containing between 30 and 400 sandeels. Most of the samples were taken during the main fishing season in spring and early summer, figure B.1. The number of samples collected decreased over the years, figure B.2, and as very few samples were collected in 1990 this year was excluded from the analyses. The geographical distribution of the samples reflects the distribution of the fishery with most samples being collected in the eastern and southern North Sea, figure B.3.

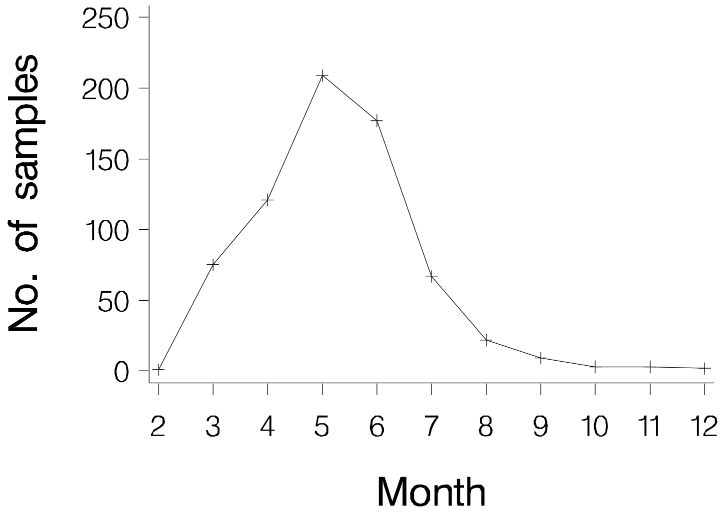

Figure B.1: *Number of samples distributed on months.*

Several factors are likely to influence the age composition of the samples (Gislason and Kirkegaard, 1998). Adult sandeels bury themselves in the sediment at night and outside the fishing season and are mostly found in areas of coarse well-oxygenated sand. Due to the burrowing behaviour the catch rates vary between different age groups, with season and during the

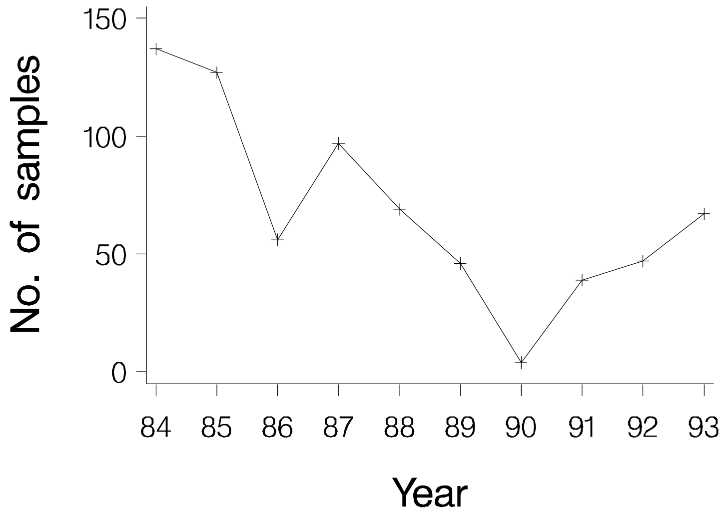

Figure B.2: *Number of samples collected per year.*

day (Reeves, 1994). Differences in the time of emergence of small and large sandeels will influence the age composition. It has thus been proposed that the larger and older individuals will emerge from the sediment later in the season than the younger and smaller individuals, and that they will re-enter the sediment earlier at the end of the season. A special relationship applies for the 0-year-olds. This age group does not appear in the samples from the first half of the year. Presumably there is little migration of adult sandeel between the various sandeel grounds in the North Sea, and regional differences in age-composition can therefore be expected. Further variation will be added by differences in trawl design and mesh size used by individual vessels as well as by errors in the reading of the otoliths.

There is insufficient information to investigate all of these potential sources of variation. Information about the laboratory, L, performing the age reading, the type of fishing gear used, G, and its mesh size, E, has been recorded. The date and the approximate position where the catch was taken on an ICES rectangle (30*30 square nautical miles) basis is also available, but information about time of day, sediment type and position of individual hauls is not available. The primary temporal and geographical variables in

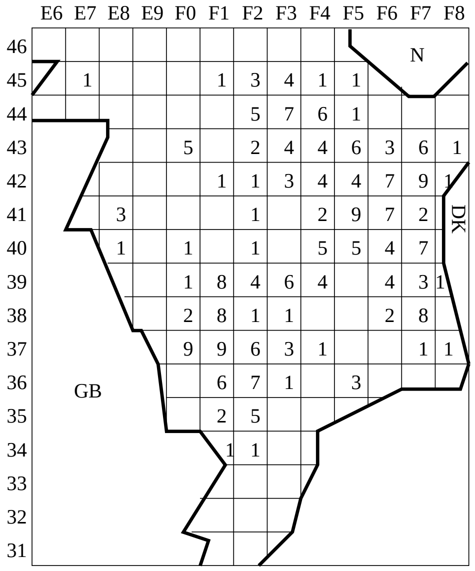

Figure B.3: *Number of years for which samples are available for a particular square in the period from 1984 to 1993.*

the analysis were therefore year, Y, month, M and rectangle, S.

The age determination took place in laboratories in the main fishing harbors. Differences between the age determinations performed by individual laboratories can be used to estimate the likely bias caused by age determination errors. Differences in age composition between larger geographical areas can be studied by subdividing the North Sea into sub-areas, A. Three different subdivisions were considered, figure B.4. Area stratification 1 has been utilised in previous assessments and is based upon the overall distribution of the fishery (Lewy, 1995). Area stratification 2 was proposed by EU project 94/071 (Wright *et al.*, 1998) based on tracking sandeel larvae in a two-dimensional sea circulation model (Proctor *et al.*, 1998). Area stratification 3 is a modification of the latter based on an overall evaluation of the present data (Pedersen *et al.*, 1998). In addition to the sub-areas a variable, R, was used to characterise samples from the northern and southern part of the North Sea. Because a distinct set of squares constitutes an area and a distinct set of areas constitutes the northern or southern part of the North Sea, S is nested within A, which again is subordinate to R.

The variation between squares within areas was modelled as random, thus assuming that the effects of squares within the same area vary around the same mean. An estimate of the age composition for a square is a compromise between the specific samples from that square and the samples from the whole area. The more imprecise an estimate of the age composition in the square, the more weight on the average for the whole area and the more variation between squares, the more weight on the samples from the individual squares. Besides utilising information from samples taken from squares with effects of roughly the same size, reducing the variance, the effect in a square without samples may be estimated, simply by assuming it to be the average effect in the area. Furthermore, modelling an effect as random also has the advantage that the significance of the effect that it is nested within may be tested, i.e. the area effect A(R). If the area effect is found to be non-significant the partition of the North Sea into a Northern and Southern part, R, may be tested.

If the older sandeel becomes available to the fishery later in the season than the younger a decrease in the continuation ratio logit in spring or early summer will be followed by an increase in late summer or early autumn. These changes in availability were accounted for by introducing a polynomial of second degree. The polynomial allowed availability to increase, decrease or remain unchanged over the season. However, the 0-year-olds are not caught

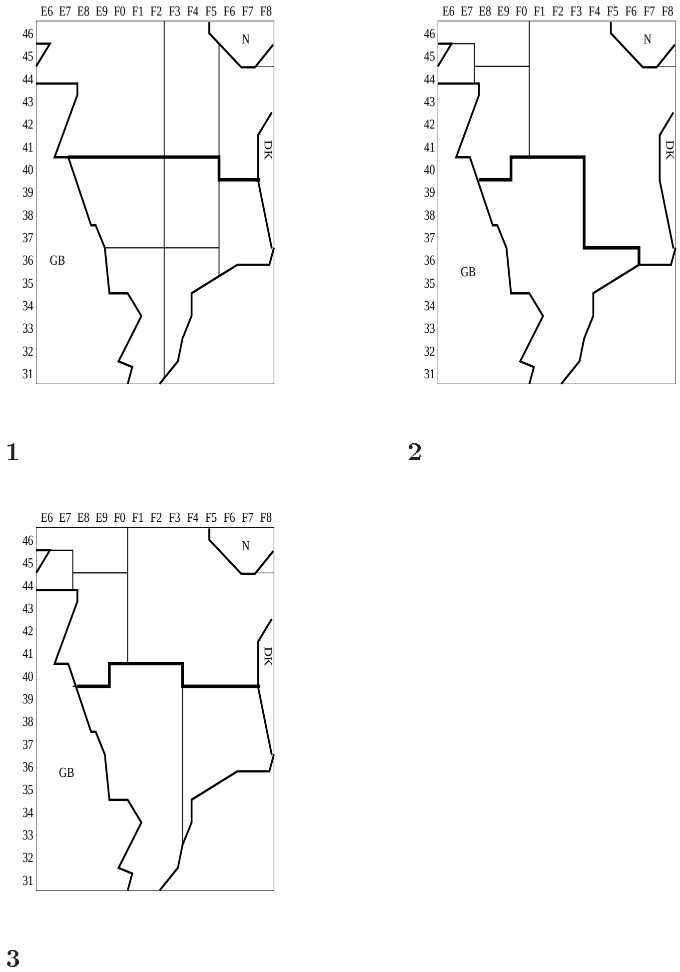

Figure B.4: *Different area stratifications. The bold line in the figures shows the border between the northern and southern areas.*

in the fishery in the first half of the year, and for this age group the logit was set to $-\infty$ in the first half of the year.

B.4 Results

The full model, with main effects and interactions all present, was too heavily parametrised for the estimation proces to converge. It was therefore necessary to reduce it, by using separate screening models for each year to identify effects that were never or only seldom significant. The initial screening models were (Wilkinson-Rogers notation):

$$L_{ayr} = \mathsf{M}+\mathsf{A}+\mathsf{S(A)}+\mathsf{L}+\mathsf{E}+\mathsf{T}+\mathsf{M}*\mathsf{A}+\mathsf{M}*\mathsf{S(A)}+\mathsf{M}*\mathsf{L}+\mathsf{A}*\mathsf{L}+\mathsf{M}*\mathsf{A}*\mathsf{L} \tag{B.9}$$

where $a = 0,\ldots,3$ denotes the age group, y=1984,...,1989,1991...,1993 denotes the year and $r = 1, 2, 3$ denotes the area stratification. In these preliminary models the month effect, M, was modelled as a class effect. A possible assignment of the areas to a northern and southern part (variable R) was not considered in these preliminary models. Interactions involving gear, G, mesh size, E, or the interaction between laboratory and square, $\mathsf{L}*\mathsf{S(A)}$, were not considered relevant and were therefore omitted. All tests were performed on a 5%-level.

The $\mathsf{E}, \mathsf{G}, \mathsf{M}*\mathsf{A}, \mathsf{M}*\mathsf{S(A)}, \mathsf{M}*\mathsf{L}, \mathsf{A}*\mathsf{L}, \mathsf{M}*\mathsf{A}*\mathsf{L}$ effects were never or rarely significant in the screening models and were therefore excluded from the model covering to all years. The rest of the effects, M, A, $\mathsf{S(A)}$ and L were included.

The initial overall models with all years included were:

$$L_{ar} = \mathsf{Y}+\mathsf{R}+\mathsf{Y}*\mathsf{R}+\mathsf{A(R)}+\mathsf{Y}*\mathsf{A(R)}+\mathsf{M}+\mathsf{MM}+\mathsf{L}+\mathsf{Y}*\mathsf{L}+\mathsf{Y}*\mathsf{S(A)} \tag{B.10}$$

where a and r were defined as before and where M and MM are regression variables. The value of M is the number of the month and the value of MM the corresponding square. Inclusion of the interaction effects $\mathsf{Y}*\mathsf{M}$ and $\mathsf{Y}*\mathsf{MM}$ was not practical due to the patchy data. The highly significant square effect was chosen to be modelled as a random effect, $\mathsf{Y*S(A)}$. Thus each square has different levels in different years, but the parameters are normally distributed with the same variance, σ^2, and the mean determined by the fixed effects.

An overview of significant effects is given for all age groups and area stratifications in table 1. In addition, the estimates of the variance component and dispersion parameter are shown in table 2. Note that the observations, particularly for age group 1, are highly overdispersed, suggesting that additional sources of variation remain to be included in the model.

Comparison of the results for the various age groups show that the age composition can be described by a relative simple model with six fixed main effects and one two-factor interaction effect, a random effect and a dispersion parameter. The dispersion parameters and variance components are similar to each other for the three different area stratifications considered. The area effect, $\mathsf{A(R)}$, for stratification 1 is significant for all age groups except for 2-year-olds (only significant on a 6% level for age group 3). Stratification 2 is not significant for any age group and stratification 3 is significant for age group 1 only. There are only small differences between the dispersion parameters and the variance components for the three stratifications. Since stratification 1 results in significant area effects, this stratification is chosen in the further presentation. The most relevant parameter estimates of the fixed effects are shown in figures B.5 to B.10. The principle by which the model has been parametrised is the standard principle used in GLM, Mixed in SAS (SAS Institute Inc. 1996). The method solves the often occuring problem of overparametrisation by selecting one of the levels of an effect as a reference level.

In the following each effect will be interpreted. A summary of the most important conclusions are given subsequently.

Table 1. *Significant fixed effects in the model described by (B.10).*

Age	Area Strat.	Y	R	Y*R	A(R)	Y*A(R)	M	MM	L	Y*L
0	1	Y	R		A(R)		M			
	2	Y	R				M			
	3	Y	R				M			
1	1	Y	R	Y*R	A(R)		M	MM	L	
	2	Y	R				M	MM	L	
	3	Y	R		A(R)		M	MM	L	
2	1	Y	R	Y*R					L	
	2	Y	R	Y*R					L	
	3	Y	R	Y*R					L	
3	1	Y	R		A(R)				L	
	2	Y								
	3	Y								

Table 2. *Random components in the model described by (B.10).*

		Random components:	
		ϕ	$\mathsf{Y} * \mathsf{S(A)}$
Age	Area Strat.	$\widehat{\phi}$	$\widehat{\sigma^2}$
0	1	12.8	5.2
	2	12.2	6.4
	3	11.9	6.2
1	1	31.1	1.0
	2	30.5	1.5
	3	30.2	1.2
2	1	7.3	0.9
	2	7.1	1.2
	3	7.1	1.0
3	1	2.3	1.4
	2	2.3	1.5
	3	2.3	1.5

B.4.1 Importance of year, Y

The year effect is significant for all age groups. The parameter estimates for age group 1 are shown in figure B.5. The graph provides an indication of the relative year class strength of the cohorts.

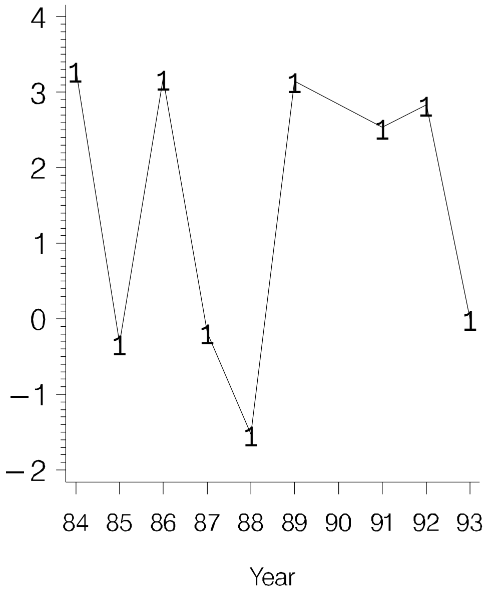

Figure B.5: *Estimated parameters for age group 1 for* Y.

In figure B.6, each cohort is followed through time and the proportion of a particular cohort compared to the proportion of the same cohorts the following year. For 0-year-olds a large year effect in one year is followed by a relatively large effect for the same cohort the following year. For older ages the tendency is the opposite. A relatively big catch of a cohort is followed by a realatively small catch of the same cohort the following year. This pattern for the 1- to 3-year-olds has formally been evaluated by an approximate test and found significant. The change in the logits, $D_{a,y} = par_{a+1,y+1} - par_{a,y}$, where $par_{a,y}$ denotes the parameter estimate of the year effect for age group a and year y, was modelled by a general linear model (Wilkinson-Rogers notation):

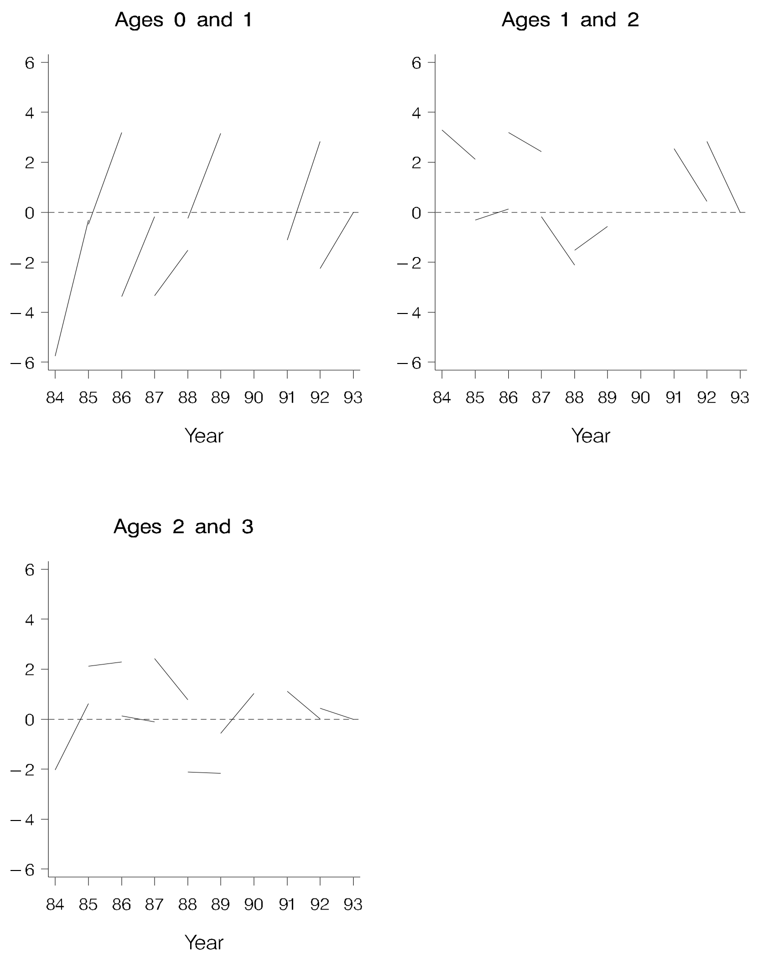

Figure B.6: *Estimated parameters for* Y *for subsequent ages. Estimates for the same cohorts are joined by a line.*

$$D_{a,y} = Age + Startlevel + Age * Startlevel \tag{B.11}$$

Age denotes the age and *Startlevel* indicates if the proportion was high (> 0.5) or low (≤ 0.5). $D_{a,y}$ is tabulated in table 3 together with its variance which was utilised as weights in the fitting and testing. The two main effects were significant on a 5% level, but not the interaction effect. The estimated changes are shown in table 4. The change is largest for the 1-year-olds. Years where 1-year-old sandeel consitutes a large part of the catch are followed by years where 2-year-olds constitutes a smaller part. Another relationship exists between 0-year-olds and 1-year-olds; the proportion of 1-year-olds can be predicted by the proportion of 0-year-olds the year before by merely adding a constant.

Table 3. *Change in parameter estimates between subsequent years.*

Year, y	*Age*, a	*Startlevel*	Change, D_{ay}	Variance
1985	2	low	2.7	0.48
1985	1	high	-1.2	0.42
1986	2	high	0.2	1.70
1986	1	low	0.4	0.55
1987	2	high	-0.3	0.64
1987	1	high	-0.8	0.56
1988	2	high	-1.5	0.50
1988	1	low	-1.9	0.53
1989	2	low	0.0	0.58
1989	1	low	1.0	0.92
1992	2	high	-1.0	0.65
1992	1	high	-2.1	0.55
1993	2	high	-0.4	0.19
1993	1	high	-2.8	0.24

Table 4. *Estimated changes in the proportion of a cohort compared to older cohorts for 1- and 2- year-olds.*

Age	*Startlevel*	Estimated Change on logit scale
1	Low	-0.20
1	High	-2.04
2	Low	1.26
2	High	-0.58

B.4.2 Importance of geographical differences in the catches, A(R), R, Y*A(R), Y*R and Y*S(A)

The area effect is significant for all age groups except for age group 2 (only significant on a 6% level for age group 3). The interaction effect between year and area, Y*A(R), is not significant for any of the age groups, i.e. the same parameter estimate for the area effect applies to an area through all years.

The R effect has been included for all age groups, either because it was significant, or because the nested effect A(R) is significant. The interaction effect between year and region, Y*R, was significant for age groups 1 and 2. This result shows that there are differences between the age composition in northern and southern part of the North Sea, and that the magnitude of this difference varies from year to year. The significance of A(R) and the insignificance of Y*A(R), indicates that there are consistent differences between the age compositions in areas within the northern and the southern North Sea, but the relative compositions in the two regions vary through the years.

The annual levels for the two regions are shown in figure B.7 as the sum of the parameters for the Y, R and Y*R effects. For age groups 0 and 3, Y*R is insignificant and the two lines are parallel. For 1-year-olds the pattern of changes in the two regions differ a little, but for the 2-year-olds the two parts of the North Sea are completely different. There are large differences between the average levels in the two regions for the 0 and 1-year-old sandeel, but not for older fish. The proportion of 0-year-olds among all and the proportion of 1-year-olds in the 1+ group is much higher in the Northern than in the Southern part.

There is large variation between squares within areas indicated by Y*S(A). For age groups 1 to 3, the variation is approximately the same. For age group 0 the variation is five times larger.

B.4.3 Importance of laboratory, L and Y*L

Although the laboratories to some extent cover different geographical areas, there is sufficient overlap to allow an evaluation of possible differences in age determinations. The laboratory effect, L, is significant for all age groups except age group 0, figure B.8. Differences in the interpretation of the otoliths seem therefore to have biased the age determinations. It is probably easier to determine the age of fish in the 0 group, because they only appear in the second half of the year and because their length distribution usually is well separated from that of the older fish. The interaction effect between year and laboratory, Y*L is not significant for any age group. This means that the laboratory effect does not change over the years.

B.4.4 Importance of variation through the year, M and MM

If the larger sandeel remained buried for a longer time period than the smaller the logits for age group 1 to 3 would be expected to decrease in the beginning of the year and increase in the end of the year. This pattern was significant for age group 1 only. For age groups 2 and 3 the month effects were insignificant, figure B.9.

For 0-year-olds the variation through the second half of the year may be described by a straight line on the logit scale, indicating an increase of the proportion of 0-year-olds caught. Thus the pattern for 0-year-olds also supports the hypothesis that larger sandeel remain buried for a longer time period than the smaller.

B.4.5 Comparison of the importance of the sources

To illustrate the importance of the different sources of variation the parameter estimates for each effect are plotted in figure B.10. The magnitude of

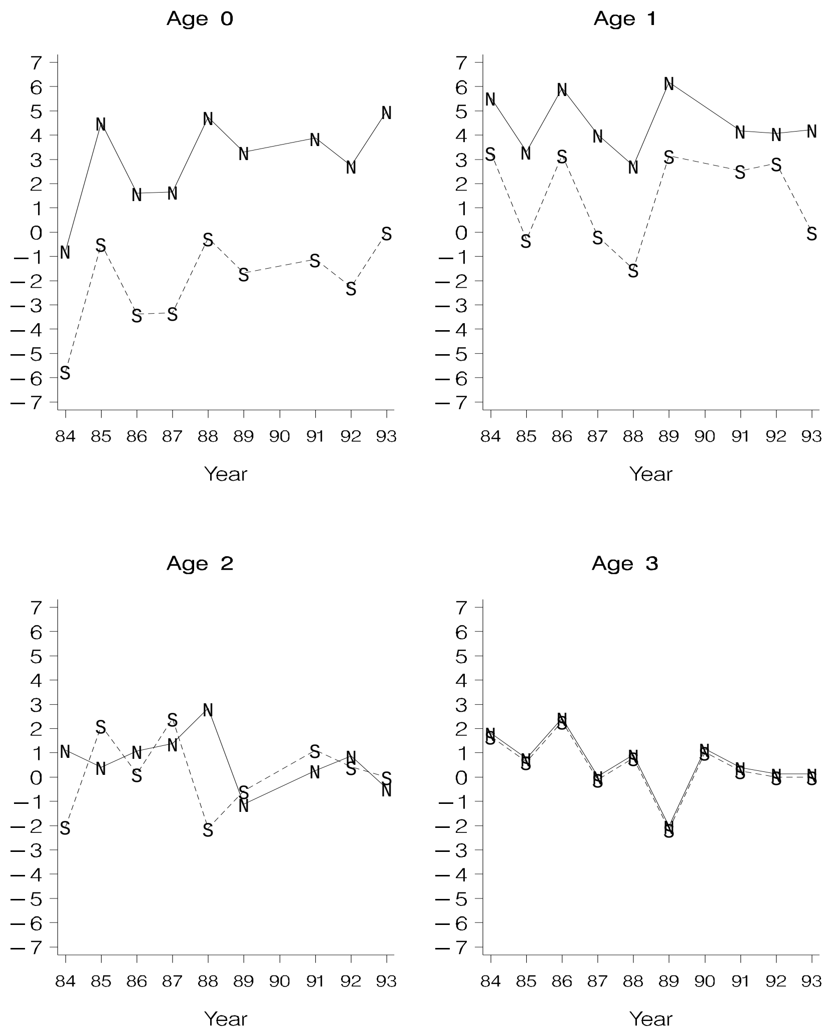

Figure B.7: *Estimated parameters for* Y+R+Y*R *progress of the logit through the years for the different age groups.*

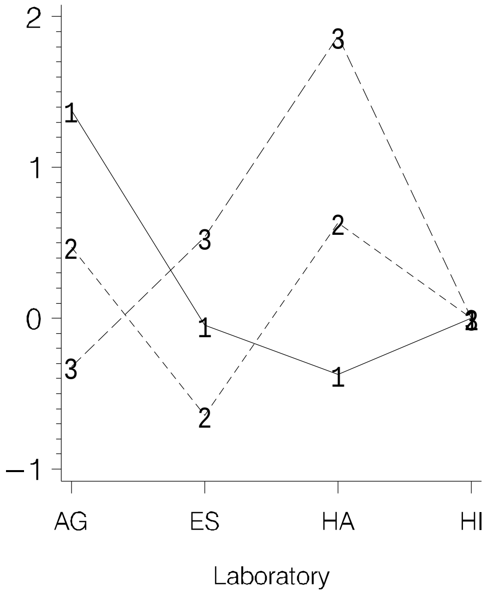

Figure B.8: *Estimates of parameters of the laboratory effect,* L. *The number indicates the age group. Estimates of the same age group are joined.*

the different sources appears to be of approximately same size. It is noteworthy that the variation between squares within areas (effect Y*S(A)) is as great as the variation between years (effect Y) and that the variation caused by the laboratory effect, L, is approximately half as large.

In order to evaluate the benefits of further investigations of the unknown sources, the magnitude of the uncertainty caused by ϕ is compared to the other random effect, $\mathsf{Y} * \mathsf{S(A)}$. It is difficult to compare those two sources because they are measured on different scales and because their influence on the variance is different for different probabilities and sample sizes (refer to equation B.8). In table 5, we have shown the approximate percentage of the total variation that is caused by the variance between squares (Y*S(A) effect), σ^2, for various probabilities and typical sample sizes for the age groups. The equation for the percentage is:

$$f_{\sigma^2} \approx \frac{\left(p_0(1-p_0)\right)^2 \sigma^2}{p_0(1-p_0)\left[p_0(1-p_0)\sigma^2 + \frac{\phi}{n}\left(1 - p_0(1-p_0)\sigma^2\right)\right]} \tag{B.12}$$

Table 5. *Percentage of the total variation that is caused by the variance between squares in model (B.10).*

Age group	0	1	2	3
$\widehat{\phi}$	12.8	31.1	7.3	2.3
$\widehat{\sigma}^2 = 5.2$	5.2	1.0	0.9	1.3
Sample size	140	130	60	20
E[p]=0.1 or 0.9	91%	29%	42%	54%
E[p]=0.5	100%	58%	70%	81%

For age group 0, the variation between squares are the dominant source of variation. For the other age groups both sources are of approximately equal important.

B.5 Discussion

We have used the model to quantify the importance of the various sources of variation in the age composition of sandeel landings in the North Sea. Knowing the sources and magnitude of variation it is furthermore possible to improve the sampling strategy. For instance, if samples from different months indicate the same age composition in the catch they may be pooled without loss of information, but with a smaller variance as a result. At last, valuable information concerning the stock dynamics may be obtained by careful analysis of the data, as the age composition of the catches provides information of the age composition of the part of the stock that is available to the fishery.

Our results show that variation between squares is considerable and therefore efforts to obtain samples from all relevant squares should be encouraged. The use of area stratification 1, which is based on the distribution of the fishery, does lead to significant area effects, while stratification 2 and 3, do not lead to significant reductions in unexplained variance. This is somewhat surprising as the latter stratifications are based on biological reasoning and believed better to reflect the sub-structure of the North Sea sandeel population. Using area stratification 1, the analysis shows that there are consistent differences in the age composition between and within the northern and southern North Sea over the years. However, except for age group 2, the changes occur in parallel, suggesting that relative year class strength changes little within the North Sea.

The changes in the logits over the year show that the older fish are available to the fishery for a shorter time period than the 1-year-olds, probably because they emerge from the sediment later in the season and re-enters the sediment earlier. The influence of gear and mesh size is negligible and suggests that stratifying the sampling effort by gear and mesh size is unlikely to result in a lower overall variation. The laboratory effect is significant and suggests perhaps that comparative age readings should have been performed more frequently.

The model was also used to study the link between the age composition in subsequent years. Our analyses showed that the proportion of 1-year-olds can be predicted from the proportion of 0-year-olds by merely adding a constant to the logit and apparetly, the fishery for 0-year-olds at the present fishing intensity does not influence the fishing possibilities of 1-year-olds the year after. However, this pattern does not apply to older fish; years with a large proportion of 1-year-olds in the catch, are followed by years where the proportion of the same cohort has decreased. This might indicate that the fishery has been attracted to 1-year-old fish in years where they were abundant. This dependency may be utilised to generate predictions of future age compositions. However, the pattern for the 1 and 2-year-olds might also be caused by bias due to age determination errors (see Fournier and Archibald 1982; Rivard, 1989; Kimura and Lyons, 1991).

The overdispersion might be caused by differences in age composition between adjacent sandeel grounds within the same square, by time of day and by weather. Although one source of variation probably caused by age determination errors has been detected, viz. the laboratory performing the age determination, several sources might exist, such as the differences between laboratory technicians or dependency between age determinations from the same samples. If the age readings of the otoliths have not been performed independently but rather by a common analysis, an overdispersion may arise due to the correlation indicating that the effective number of observations are smaller than the sample size. Another possibility is that the samples are labelled incorrectly. The samples are taken at port while the vessel is unloading the fish. If the fisherman has trawled in several squares, part of the unexplained variation should probably be attributed to variation between squares within areas. Evaluation of the magnitude of those unknown sources should be done cautiously, since the model of those probably is not valid. By modelling the unknown sources by a dispersion parameter; one supposes that each individual has a different level of the unknown sources.

The estimates of the uncertainty of the age composition differs from results previously presented by taking into account the binomial variance structure of the proportions, e.g. Cochran (1977), Schweigert and Sibert (1983) and Horppila and Peltonen (1992) calculate empirical variances and thus ignore the relation between the variance of a proportion and its expected value.

The estimated uncertainty of the age composition may be utilised to estimate the uncertainty of the catch at age data. This estimate can be directly utilised by statistical stock assessment methods to improve the estimates of uncertainty about the current situation. For instance the information may be used to provide an informative prior distributions in Bayesian stock assessment models (McAllister and Ianelli, 1997), or as a known variance of the observations used in Time Series Models (e.g. Gudmundsson, 1994 and Schnute, 1994;). At present external estimates of observation error are not fully utilised. Most often catch at age data is aggregated into average numbers per year and geographical unit before it is entered into the assessment model.

In the models presented here a rough spatial correlation is introduced through the random effect Y*S(A); age compositions for squares within the same area are more correlated than squares from different areas. Another approach of modelling the age composition is to make a more detailed structure of the spatial correlation such as modelling the correlation between age compositions of two squares as a function of the distance between them. However, preliminary analyses in terms of variograms (Isaaks, 1989) of the parameters of squares within the same area do not indicate that such a correlation structure exist. In a more detailed spatial model it would also be relevant to evaluate factors describing the environmental conditions such as temperature, oxygen content and bottom conditions.

A drawback of the method presented is that the method requires age determination of all sandeels in a sample or of a random subsample. For this reason only data from 1984-1993 were considered. The method needs to be extended to cases where only a proportion of the sandeels in each length group is age determined.

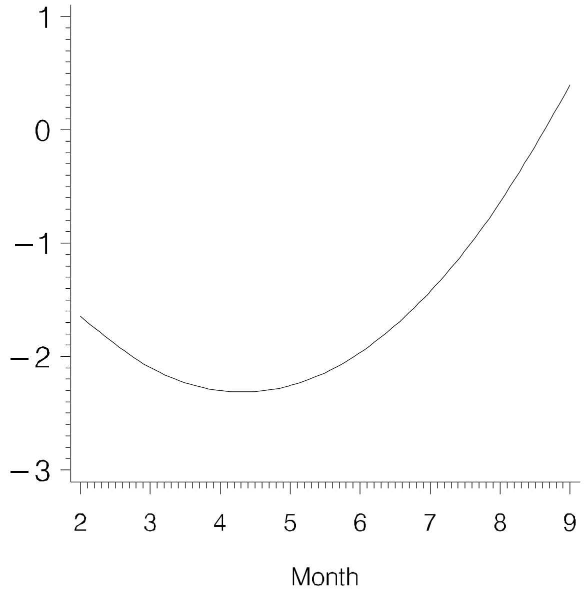

Figure B.9: *The estimated progress of the logit through the year for age group 1.*

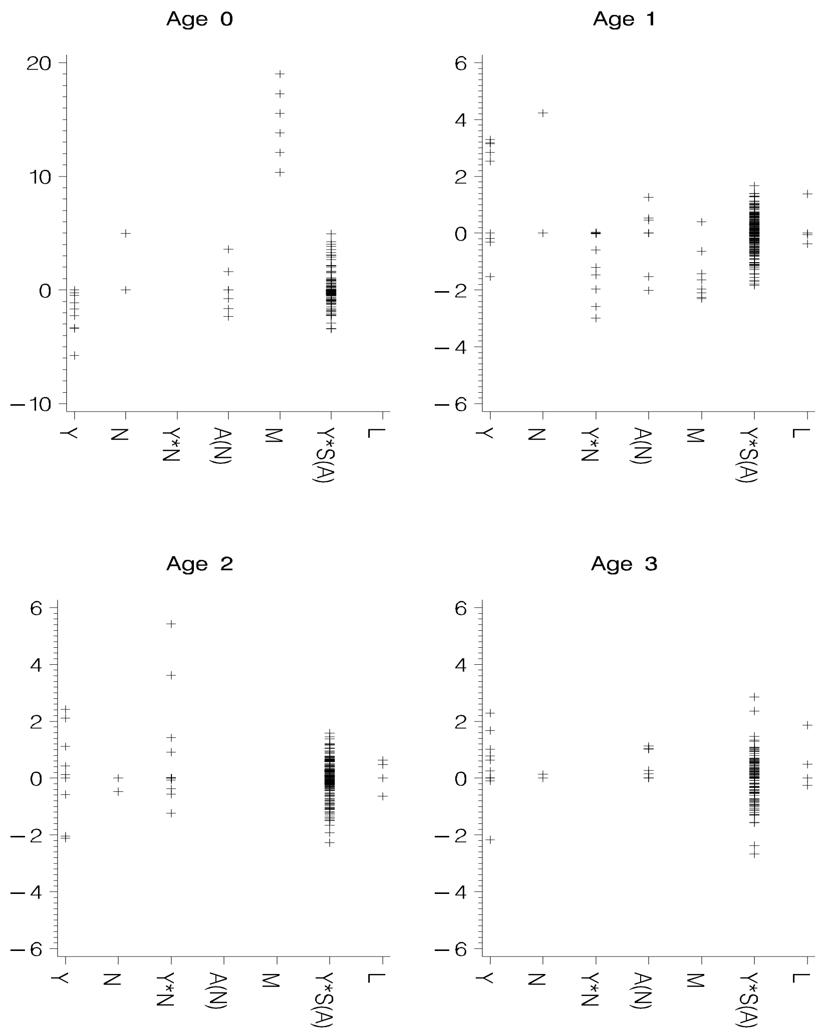

Figure B.10: *Estimates on the logit scale for the significant effects. Note that the scale for age group 0 is different from the others.*

B.6 References

Agresti, A. 1990. Categorical data analysis. John Wiley and Sons, Inc.
Breslow, N.E. and Clayton, D.G. 1993. Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88:9-25.

Cochran, W. G. 1977. Sampling techniques. Third edition. John Wiley and Sons, Inc., New York, NY. 428 p.

Daan, N., Bromley, P.J., Hislop, J.R.G and Nielsen, N.A. 1990. Ecology of North Sea fish. Netherlands Journal of Sea Research, 26:343-386.

Fournier, D. and Archibald, C. P. 1982. A General Theory for Analyzing Catch at Age Data. Canadian Journal of Fisheries Aquatic Science, 39:1195-1208.

Gislason, H. and Kirkegaard, E. 1998. Is the industrial fishery in the North Sea sustainable?. *In* Northern Waters: Management Issues and Practice. *Edited by* D. Symes. Fishing News Books. London.

Gudmundsson, G. 1994. Time Series Analysis of Catch-at-age Observations. Applied Statistics, 43:117-126.

Horppila, J. and Peltonen, H. 1992. Optimizing sampling from trawl catches: contemporaneous multistage sampling for age and length structures. Canadian Journal of Fisheries Aquatic Science, 49:1555-1559.

ICES 1996. Report of the ICES Advisory Committee on Fishery Management, 1995. ICES Cooperative Research Report no. 214.

Isaaks, E. H. and Mohan, R. S. 1989. An introduction to applied geostatistics. Oxford University Press.

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

Kvist, T., Gislason, H. and Thyregod, P. 1998. Using continuation-ratio logits to analyse the variation of the age-composition of fish catches. Submitted for publication, 1998.

Lewy, P. 1995. Sampling Methods and Errors in the Danish North Sea Industrial Fishery. Dana, 11:39-64.

McAllister, M. K. and Ianelli, J. N. 1997. Bayesian Stock Assessment using Catch-age Data and the Sampling - Importance Resampling Algorithm. Canadian Journal of Fisheries Aquatic Science, 54:284-300.

McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models (second edition). Chapman and Hall (London; New York).

Megrey, B.A. 1989. Review and Comparison of Age-structures Stock Assessment Models from Theoretical and Applied Points of View. American Fisheries Society Symposium 6:8-48.

Pedersen, S. A., Lewy, P. and Wright, P. 1998. Assessments of the lesser sandeel (Ammodytes marinus), and its relevance to fishing pressure in the North Sea, ICES C. M: 1998/AA:7.

Proctor, R., Wright, P.J. and Everitt, A. 1998. Modelling the transport of larval sandeels on the north-west European shelf. Fisheries Oceanography, 7(3/4):347-354.

Reeves, S.A 1994. Seasonal and annual variation in catchability of sandeels at Shetland, ICES CM 1994/D:19 (mimeo.).

Rivard, D. 1989. Overview of the Systematic, Structural , and Sampling Errors in Cohort Analysis. American Fisheries Society Symposium 6:49-65.

SAS Institute Inc. 1996. SAS/STAT Software: Changes and enhancements through Release 6.11.

Schnute, J. T. 1994. A General Framework for Developing Sequential Fisheries Models. Canadian Journal of Fisheries Aquatic Science, 51:1676-1688.

Schweigert, J. F. and Sibert, J. R. 1983. Optimizing survey design for determining age structure of fish stocks: an example from British Columbia Pacific herring (*Clupea harengus pallasi*). Canadian Journal of Fisheries Aquatic Science, 40:588-597.

Sparholt, H. 1990. An estimate of the total biomass of fish in the North Sea. Journal du Conseil International pour l'Exploration de la Mer, 46:200-210.

Wolfinger, R. and O'Connel, M. 1993. Generalized Linear Mixed Models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48:233-243.

Wright, P.J. 1996. Is there a conflict between sandeel fisheries and seabirds? A case study at Shetland. *In* Aquatic predators and their prey. *Edited by* S.P.R. Greenstreet and M.L. Tasker. Fishing News Books, Blackweell Science Ltd. Oxford.

Wright, P., Verspoor, E., Andersen, C., Donald, L., Kennedy, F., Mitchell, A., Munk, P., Pedersen, P.A., Jensen, H., Gislason, H., Lewy, P. 1998. Population structure in the lesser sandeel (Ammodytes marinus) and its implications for fishery-predator interactions. DG XIV Contract No. 94/071. Final Report October 1998.

Appendix C

Uncertainty of Catch at Age Data for Sandeel

T. Kvist, H. Gislason and P. Thyregod

Keywords: catch at age data, generalized linear mixed models, continuation-ratio logits, compound distribution

Abstract

The uncertainties associated with catch at age data for sandeel landings in the North Sea have been assessed. The uncertainties may be associated with the estimated age composition, the catch figures per area, the estimated species composition of the catches, and the transformation of the unit of measurement of the magnitude of the catch from tonnes to numbers. The various sources of uncertainty were analysed separately and thereafter combined into estimates of the catch at age data. Continuation-ratio logits and generalised linear mixed models were utilised for estimation of the age composition and the associated uncertainties. A novel approach based on a compound distribution for analysing the species composition was used. The results showed that the uncertainties associated with the catch at age data are huge; occasionaly the coefficient of variation is larger than 50%,

and that the uncertainties associated with the age composition data are the most dominating.

C.1 Introduction

Most cohort based stock assessment methods, such as Virtual Population Analysis (VPA) (Gulland, 1965 and Pope, 1972) and similar methods rely heavily on catch at age data and the data has great influence on the resulting quantities such as population sizes (Lai and Gunderson (1987), Tyler *et al.* (1989), Fournier and Archibald (1982), Pelletier (1990) and Bradford (1991)). Therefore estimation of the uncertainty of catch at age data is necessary for estimation of the uncertainties of those quantities. Various statistical models have been developed with the purpose of assessing the uncertainties, e.g. those presented in Fournier and Archibald (1982), Deriso *et al.* (1985), Gudmundsson (1994), McAllister and Ianelli (1997) and Punt and Hilborn (1997), Fargo and Richards (1998). Such models might be improved by incorporating explicit estimates of the uncertainties of the catch at age data.

The uncertainty of catch at age data can be estimated by utilising the information from the biological samples taken from the landings. There are several possible sources of uncertainty in the catch at age data. Those might be selectivity and age determination errors, temporal and geographical variations in the age composition of the part of the stock that is available to the fishery, the species composition in the sea and growth.

Age determination errors may arise due to systematic or random errors. However, both sources will cause bias in the ageing due to the truncation of the age distribution. In general, ageing error makes strong year classes appear weaker and weak year classes appear stronger (Fournier and Archibald 1982, Rivard 1989, Kimura and Lyons, 1991). Richards *et al.* (1992) presents statistical models for the analysis of ageing error. However, those methods require multiple independent age readings, and can therefore not be applied to historical sample data which do not contain multiple age readings.

Many authors have addressed the problem of estimating the uncertainty of catch at age data, e.g. Gavaris and Gavaris (1983), Rivard (1983), Schweigert and Sibert (1983), Pelletier and Gros (1991), Lewy (1995). However, recent development of the theory of Generalised linear mixed models

(McCullagh and Nelder (1989), Breslow and Clayton (1993) and Wolfinger and O'Connel (1993)) and computer facilities makes it possible to take into account as well the special characteristics of the variance of a proportion as various sources of random variation. Pelletier and Gros (1991) attempts to avoid the problem of dependency between a proportion and its variance by choosing the number of individuals in an age group in a sample as the variable of interest, rather than the proportion. However, this approach requires that the samples consist of a fixed number of individuals and even then the variance of the number of individuals in an age group is not constant; the variance is smallest for small and large numbers of individuals in an age group analogous to the characteristics of proportions. Recently, Crone and Sampson (1998) have investigated the stochastic properties of catch at age data from five groundfish species commersially landed at Oregon ports. Crone and Sampson (1998) reached the conclusion that a multinomial probability error structure, included in models that are based on maximum likelihood estimation, more closely follows the variability associated with the sampled landing data than does a log-normal error structure used in models based on least squares estimation, and thus support the modelling approach of age composition data applied here.

Another important issue has not been addressed in papers concerning the estimation of catch at age data and its uncertainties, viz. an evaluation of factors which may have importance for the age composition. Such a method has been presented in Kvist *et al.* (1998).

The aim of this paper is twofold: to utilise the benifits of the new method (Kvist *et al.* (1998) and Kvist *et al.* (1999)) in the estimation of the uncertainty of the catch at age data for the sandeel fishery in the North Sea and to present a novel approach in estimation of the species composition and its uncertainty. The lesser sandeel(*Ammodytes marinus* Raitt) is one of the most abundant fish species in the North Sea (Sparholt, 1990) and it is the main target of the Danish industrial fishery. The fishery is characterized by not having information on the actual weight of the catch of sandeel, as only the weight of the total industrial fishery is recorded. Therefore the biological samples from the landings are utilised also for estimation of the species composition. This characteristic may also apply for other fisheries where the catches are composed of several species and where the catch by species is not recorded, e.g. for certain fisheries in the tropics (Sparre and Venema, 1992). Therefore the novel approach for estimation of the species composition may be relevant for other fisheries as well.

C.2 Materials

The materials used for estimation of catch at age data for sandeel consist of information on the total industrial catch together with biological samples taken from the landings with the purpose of determining species composition, age composition and weight. The analyses presented here are based on data from 1984 to 1991.

The information on the appr. catch by ICES rectangle comes from fishermens logbooks and covers nearly all the industrial fishery in the North Sea, namely vessels with an overall length of 17 m or more. The logbooks contain information on the species, the appr. amount caught, the date of fishery, appr. position in terms of the ICES rectangle, and information on the vessel size and gear. This information is combined with information from the first hand buyers who are obliged to report for each landing, the quantity of industrial species and in which ICES Division the landing was taken. The information from the logbook database is utilised to estimate the relative industrial catch in each month and ICES rectangle, whereas the more accurate information from the first hand buyer on the actual weight of the catch is utilised to get an estimation of the absolute industrial catch in every ICES rectangle and month. Since the logbooks cover almost all industrial fishery and the information from the first hand buyers covers all industrial landings, the contrinution of uncertainty from sampling errors is assumed to be negligible. However, the information recorded in the logbooks may be prone to errors, due either to human mistakes, or to attempts to cover up illegal fishing. The magnitude of these errors could possibly be evaluated by comparing the control samples taken by the Danish Fisheries Inspection with the logbook data. However, such an evaluation would only disclose some of the errors, such as the species caught. Other errors such as the position of the fishery cannot be checked. In addition, interpretation of the results of such an evaluation, as eg. the proportion of errors would be impeded by the sampling procedure because the samples are not taken at random but directed towards vessels under suspicion. In the present analyses it was assumed that the uncertainty in the estimate of the total weight of the industrial landings caught in every ICES rectangle and month could be neglected. This assumption is partly supported by the fact that the sandeel fishery was subject to few regulations in the period from which the samples were obtained.

The biological samples are taken from the landings in the harbours by lowering a 10-litre pail into the hold and taking out a random sample (Lewy, 1995). It is assumed that the samples are collected randomly among fish, i.e. that every fish is equally likely to be sampled. However, the documentation of the sampling design for the period studied is incomplete and therefore it is uncertain whether or not the assumption holds. If the samples are taken at random among vessels in stead of among fish, the samples should be weighed with the size of the landing. The samples are sorted into species and the species composition by weight is determined. A random subsample is collected for determining the length and age composition of each species. The length and age of the individuals and the weight of each length group are recorded. The age determination is performed mainly by counting the number of growth rings in the otoliths. For each sample the date of fishery, ICES rectangle, vessel size, gear and mesh size are recorded as well.

The collection of a fixed volume of sandeel, rather than a fixed number, might introduce a bias, i.e. an expected age-composition in the sample differing from that in the catch (Scheaffer, 1969, Buslik, 1950). This could happen if the volume of one sandeel was large relative to the sampling volume. However, because the volume of one sandeel is small relative to the total sample volume we do not expect this to influence our results. Similarly, the by-catch in the sandeel fishery is small and we do not expect that the species composition of the sample will influence the age-composition of sandeel.

C.3 Methods

The various sources of information utilised in the estimation of catch at age data are analysed separately and thereafter combined into estimates of the catch at age data. At first, the information on the weight of the industrial catch is combined with estimates of the species composition in weight giving the catch weight of sandeel (Lewy, 1995). Secondly, the mean weight of sandeels is utilised to transform the weight of the catch into the number of sandeel caught. At last the age composition of the catches is utilised to get estimates of the number of sandeel caught from each age group. The various analyses are presented in the sections below.

C.4 Species composition

Estimation of the weight-proportion of sandeel in the industrial fishery is done on the basis of the biological samples described above. An overview of those samples is given in figure C.1, where the weight-proportion of sandeel in the samples is shown. The figure shows that there is a clear stratification of the samples; those which are dominated by sandeel with generally more than 75 weight-percentage sandeel and those which hardly contain any sandeel. The two types of samples are utilized to categorize the catches into sandeel catches and other catches. A catch is defined as a *sandeel catch* if the sample contains more than 50 weight-percentage sandeel and as an *other catch* otherwise.

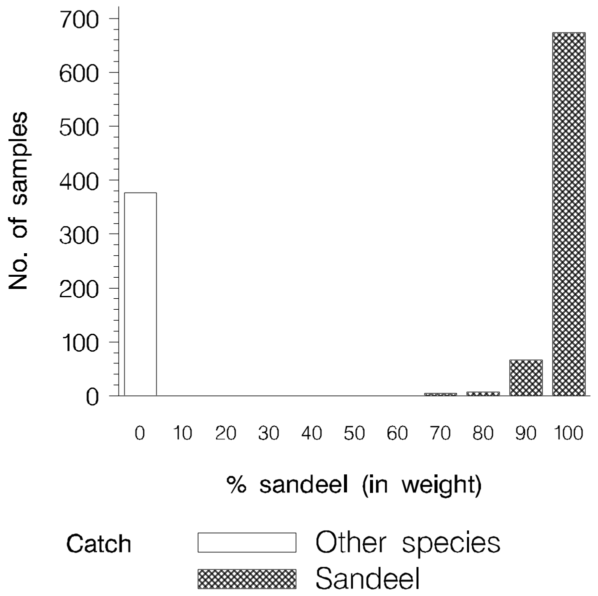

Figure C.1: *Histogram of the % of sandeel (in weight) in samples collected from 1984 to 1991.*

A classification of fishery into sandeel fishery and remaining fishery (other fishery) based on gear, mesh size, period of fishing and position of fishing has been described by Lewy (1995). Thus, the classification distinguishes between sandeel fishery and remaining industrial fishery, in the following referred to as other fishery. Because the definition is based on information available from the logbook database, which contains information on almost every trip, the sandeel fishery may be defined with very high accuracy provided the uncertainty caused by errors in the logbooks is considered negligible.

The distribution of the weight-percentage of sandeel in samples from the two types of fishery (sandeel/other) is shown in figure C.2.

The plots show, that the classification results in only a few misclassifications of sandeel catches into other fishery, whereas the misclassifications of other catches into sandeel fishery are more comprehensive. As a result of this the misclassification of sandeel catches into other catches are considered negligible and thus ignored, whereas the misclassification of other catches into sandeel catches is taken into account in the estimation. The objective of estimating the weight-proportion of sandeel in the industrial catches has thus been reduced to estimating the weight-proportion of sandeel in the *sandeel* fishery, $\boldsymbol{\rho}$, and its variance $\boldsymbol{\Sigma}_{\boldsymbol{\rho}}$.

As can be seen from the plot, there are two major sources of variation in the weight-proportion of sandeel in the sandeel fishery. Firstly, there is a classification error resulting from errors in the definition of the sandeel fishery; probably there is catches included in the sandeel fishery where the target fish was not sandeel. Secondly, there is by-catches occuring in some of the sandeel catches, indicated by weight-percentages of sandeel a little smaller than 100. By-catches in the other fishery is very small and assumed negligible. Thus, the distribution of the weight-proportion of sandeel is compounded by a distribution describing misclassification of other catches and a distribution which describes by-catches in sandeel catches. The two sources of variation is analysed separately and then combined into estimates of the weight-proportions of sandeel and their variances and covariances.

C.4.1 Classification of catches within sandeel fishery

The definition of the sandeel fishery based on the gear, mesh size, time of year and position of fishery gives an overall selection of sandeel catches, but as mentioned above misclassifications may occur. The classification of a catch may be described by a Bernoulli distribution; either the classification is correct or it is false. It is considered correct if the catch is a sandeel

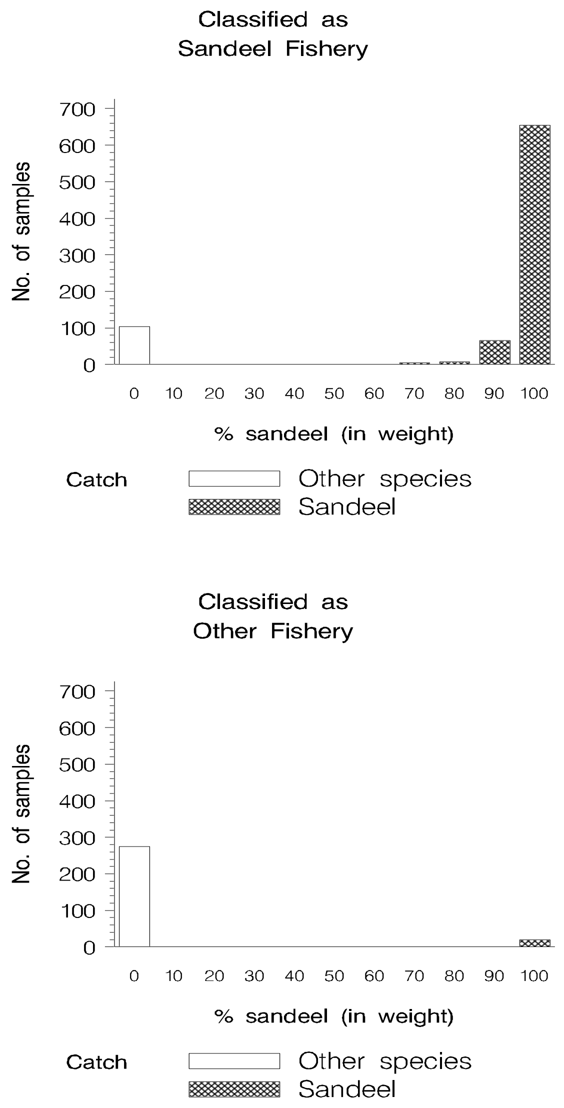

Figure C.2: *Histogram of the weight-% of sandeel in samples collected from 1984 to 1991.*

catch, i.e. if there is more than 50 weight-percentage of sandeel in the sample. Let Λ indicate whether the catch is a sandeel catch or not:

$$\Lambda = \begin{cases} 1 & \text{Sandeel catch} \\ 0 & \text{Other catch} \end{cases} \tag{C.1}$$

then Λ is Bernoulli distributed:

$$\Lambda \in \mathrm{Ber}(\boldsymbol{\lambda}) \tag{C.2}$$

where $\boldsymbol{\lambda}$ is the proportion of correctly classified samples.

Since the response is Bernoulli distributed, the proportion of correctly classified samples may be analysed by means of a generalised linear model (McCullagh and Nelder, 1989). By this method factors which may be of significance for the proportion of misclassifications may be evaluated. Such factors could be temporal, geographical or describing characteristics of the vessels or equipment. The year and month effect describe differences in the proportion of misclassifications between and within the years and the area effect geographical differences (refer to figure C.3 for the definition of areas). Significance of total catch might indicate that the general level of the catch capacities for vessels in the sandeel fishery is different from that of other industrial fisheries. Analogously, significance of the mesh size might indicate that there is more information in the mesh size concerning the target species than that already utilized to distinguish between sandeel fishery and other fishery.

The analysis results in estimates of the proportion of correctly classified samples within the sandeel fishery for each ICES rectangle and month, $\widehat{\boldsymbol{\lambda}}$, and of the variance and covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}}$.

C.4.2 By-catches in sandeel catches

In this section we will address the problem of estimation of the by-catches in the sandeel catches within the sandeel fishery, i.e. the weight-proportion of sandeel given that the catch is a sandeel catch, $\rho|\Lambda = 1$. The weight-proportion is modelled by a beta-distribution because the beta distribution is a flexible distribution especially useful for describing proportions (Lewy, 1996). It has not been attempted to model different by-catch distributions

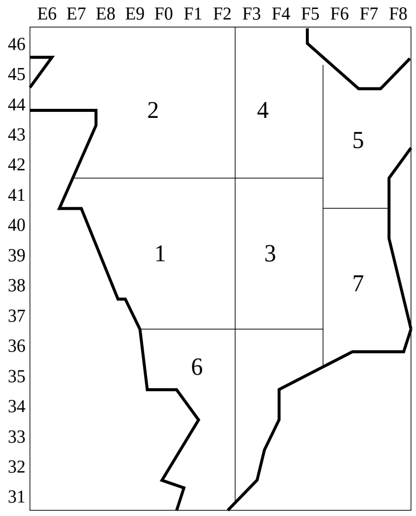

Figure C.3: *The seven sandeel areas.*

for different explanatory variables, because the by-catch is rather small, and thus bias and uncertainty due to inadequate modelling of by-catches are considered negligible. The estimated beta-distribution is here described by its estimated mean, μ_1, and the variance of this estimate, σ_1^2.

C.4.3 Combining the distributions into estimates of species composition

In order to obtain estimates of the mean weight-proportion of sandeel in the sandeel fishery, $\widehat{\boldsymbol{\rho}}$ and its variance, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}}$, the two distributions are combined.

The weight-proportion, $\boldsymbol{\rho}$, for each ICES rectangle and month is estimated by:

$$\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\lambda}}\widehat{\mu}_1 \tag{C.3}$$

and the variance and covariance matrix, $\boldsymbol{\Sigma}_\rho$ of $\widehat{\boldsymbol{\rho}}$ is estimated by:

$$\widehat{\boldsymbol{\Sigma}}_\rho = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\lambda}} \times \widehat{\mu}_1^2 + \widehat{\sigma}_1^2 \times \widehat{\boldsymbol{\lambda}} \times \widehat{\boldsymbol{\lambda}}^{\text{‘}} \tag{C.4}$$

using first order Taylor approximation.

C.5 Estimation of the mean weight of sandeels

The mean weight of sandeels is needed in order to transform the weight of the sandeel catch into a number of sandeel.

It is anticipated that the mean weight will be different for different parts of the North Sea and different time periods due to differences in age composition and growth. The variation of the age composition of sandeel has been analysed and the results are presented in Kvist *et al.* (1999). The variation of growth of the sandeel has been illustrated by Petersen *et al.* (1999).

The particular aim of the present analysis is the mean weight and its variance only. From the analyses above we know that factors of significance could be: year, Y, month, M, area A and ICES rectangle, $\mathsf{S}(\mathsf{A})$. Assuming that the weight of sandeels may be approximated by a normal distribution, the variation in the weight can be analysed by an ordinary analysis of variance.

C.6 Estimation of age composition

The age composition is estimated on the basis of biological samples as in the determined species composition. However, the number of samples is smaller because it is more time consuming and resource demanding to determine the age.

It is desirable to detect which factors that influence the age composition so that they can be taken into account in the estimation and thus reduce the variation. And factors that do not influence the age composition may be ignored. For instance if samples from different months have the same age composition they can be pooled without loss of information, but with a smaller variance as a result. The analysis is impeded by the fact that the response is ordered categorical. An easily applicable method to analyse such data has been presented in Kvist *et al.* (1998). The idea is to split the probability of the multidimensional response into binomial probabilities. This is done by comparing each age group to the union of older age groups. The age composition data utilised here has already been analysed (Kvist *et al.*, 1999). The results of that analysis is utilised to choose a relevant model. The factors that were found to be of significance for the age composition are year, Y, area, A, ICES rectangle, S, month, M, and the laboratory, L, which has performed the age determination (refer to Kvist *et al.* (1999)). Furthermore there is interaction between ICES rectangle and year. Because of the strong indication of large geographical differencies in the age composition separate parameters are estimated for each area and ICES rectangle every year. They are, however, modelled as random with the effect that the estimates in a certain ICES rectangle or area to be a compromise between the data in that particular position and the mean of the superior geographical unit, e.g. the estimate for an ICES rectangle is a compromise between its own data and the mean of the data in the area it belongs to (Robinson, 1991). Furthermore, a qualified estimate of the level in an ICES rectangle or area without observations may be provided. A month effect common for all years is estimated for age groups 0 and 1 and likewise a laboratory effect covering all years is estimated for age groups 1 to 3 in accordance with the results presented in Kvist *et al.* (1999).

The effects included in the models for the age groups are presented in table 1.

Table 1. *Effects included in the age composition models.*

Age Group	Fixed effects		Random effects		
0	Y	M_0	$Y * A$	$Y * S(A)$	
1	Y	M_1	$Y * A$	$Y * S(A)$	L
2	Y		$Y * A$	$Y * S(A)$	L
3	Y		$Y * A$	$Y * S(A)$	L

For age group 0, the proportion is assumed to be 0 between January and May, and 1 between October and December. The month effect is modelled as a class effect, each month representing a separate level. For 1-year-olds, the month effect is also modelled as a class effect, letting each month represent a separate level. However, the months April to June are assumed to have the same level (in accordance with data).

The approach to estimate the proportions of each age group, $\mathbf{p}_{a_1}$, and the variance/covariance matrix of the estimate of age a_1 and a_2, $\mathbf{\Sigma}_{\mathbf{p},a_1,a_2}$ has been presented in Kvist *et al.* (1998). a_1 and a_2 denote age groups.

C.7 Combining all sources into an estimate of catch at age data

Catch at age data and their variances and covariances are estimated by combining the estimates provided by the analyses given above. Those are estimates of the weight of the catch in the sandeel fishery, $\widehat{\mathbf{w}}_{\mathsf{SF}}$, the weight-proportion of sandeel, $\widehat{\boldsymbol{\rho}}$, the mean weight of sandeels, $\widehat{\mathbf{v}}$, and the proportion of each age group, $\widehat{\mathbf{p}}_0, \ldots, \widehat{\mathbf{p}}_4$, and their variance and covariance matrices, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{v}}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{p},0,0}$, $\widehat{\boldsymbol{\Sigma}}_{\mathbf{p},0,1}, \ldots, \widehat{\boldsymbol{\Sigma}}_{\mathbf{p},4,4}$. The vector and matrices contain estimates for each combination of year, month and ICES rectangle. The estimates of catch at age data are obtained in successive steps. First the weight of sandeel caught, $\mathbf{w}_{\mathsf{S}}$, is estimated by:

$$\widehat{\mathbf{w}}_{\mathsf{S}} = \mathrm{diag}(\widehat{\mathbf{w}}_{\mathsf{SF}}) \times \widehat{\boldsymbol{\rho}} \tag{C.5}$$

where diag(.) transforms the argument, which must be a vector, into a diagonal matrix.

The estimate of the variance and covariance matrix is:

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{w}_{\mathsf{S}}} = \mathrm{diag}(\widehat{\mathbf{w}}_{\mathsf{SF}}) \times \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}} \times \mathrm{diag}(\widehat{\mathbf{w}}_{\mathsf{SF}}) \tag{C.6}$$

$\widehat{\mathbf{w}}_{\mathsf{SF}}$ is treated as a constant because it is considered to have negligible uncertainty.

Secondly, the number of sandeel caught, $\mathbf{s}$, is estimated by dividing the weight of the sandeel catch by the mean weight of sandeels:

$$\widehat{\mathbf{s}} = \widehat{\mathbf{w}}_{\mathsf{S}} \times [\mathrm{diag}(\widehat{\mathbf{v}})]^{-1} \tag{C.7}$$

Using first order Taylor approximation it is found that the appr. corresponding variance and covariance matrix is:

$$\widehat{\boldsymbol{\Sigma}}_s = [\mathrm{diag}(\widehat{\mathbf{v}})]^{-2} \times \widehat{\boldsymbol{\Sigma}}_{\mathbf{w}_{\mathsf{S}}} + [\mathrm{diag}(\widehat{\mathbf{v}})]^{-2} \times \mathrm{diag}(\widehat{\mathbf{w}}_{\mathsf{S}}) \times \widehat{\boldsymbol{\Sigma}}_v \times [\mathrm{diag}(\widehat{\mathbf{v}})]^{-2} \times \mathrm{diag}(\widehat{\mathbf{w}}_{\mathsf{S}}) \tag{C.8}$$

Thirdly, the number of sandeel caught in a given rectangle and time period is allocated to the various age groups by multiplying with the estimated proportions in each age group. The number of sandeel caught per year, month, ICES rectangle and age group a is:

$\widehat{\mathbf{a}}_a = \mathrm{diag}(\widehat{\mathbf{s}}) \times \widehat{\mathbf{p}}_a$

and the appr. variance/covariance matrix between age groups a_1 and a_2 is:

$\widehat{\mathbf{\Sigma}}_{\mathbf{a},a_1,a_2} = \mathrm{diag}(\widehat{\mathbf{s}}) \times \widehat{\mathbf{\Sigma}}_{\mathbf{p},a_1,a_2} \times \mathrm{diag}(\widehat{\mathbf{s}}) + \mathrm{diag}(\widehat{\mathbf{p}}_{a_1}) \times \widehat{\mathbf{\Sigma}}_{\mathbf{s}} \times \mathrm{diag}(\widehat{\mathbf{p}}_{a_2})$.

Again Taylor approximation has been used.

At last, the estimates for each year, month and ICES rectangle are added into the number of sandeel per age group and year, i.e. the catch at age data:

$$\widehat{\mathbf{C}}_a = \widehat{\mathbf{a}}_a \times \mathbf{1} \tag{C.9}$$

where the matrix symbolised by **1** contains one column for each year with the figure 1 in each position corresponding to the year the column represents, and the figure 0 in the other positions.

The variance and covariance matrix of the catch at age data is estimated by:

$\widehat{\mathbf{\Sigma}}_{\mathbf{C},a_1,a_2} = \mathbf{1} \times \widehat{\mathbf{\Sigma}}_{\mathbf{a},a_1,a_2} \times \mathbf{1}'$

C.8 Results

The results of the subanalyses are presented in the sections below. The results are hereafter utilised to estimate the catch at age data and its uncertainty for the sandeel fishery in the sandeel areas for the years 1989 and 1991.

C.8.1 Species Composition

Classification of catches within the sandeel fishery

All the potential explanatory variables are included in a model of the proportion of correctly classified catches:

$$\eta = \mathsf{Y} + \mathsf{M} + \mathsf{A} + \mathsf{T} + \mathsf{ME} \tag{C.10}$$

where η is

$$\eta = \log \frac{\boldsymbol{\lambda}}{1 - \boldsymbol{\lambda}} \tag{C.11}$$

and the symbols Y, M, A, T and ME correspond to the effects year, month, area, total catch and mesh size. The model is based on data from samples taken in the period 1984-1996.

The model has been analysed by means of PROC GENMOD in SAS (SAS Institute Inc., Cary, NC, USA. Release 6.12). Unfortunately convergence problems occur when interactions are included and therefore only main effects have been tested. The likelihood ratio statistics are shown in table 2.

Table 2. *Likelihood ratio statistics for type 3 analysis of model C.10.*

Effect	DF	ChiSquare	Chisquare/DF	Pr>Chi
M	5	50	10	0.0001
ME	2	403	201	0.0001
Y	12	97	8	0.0001
A	6	11	2	0.0808
T	3	8	3	0.0551

Although all factors were significant, it is obvious that the overall dominating effect is the mesh size. The month effect and year effect explain only a fraction of what the mesh size effect explain. The area effect and effect from the size of the total catch is even smaller. Those two effects are removed from the model.

Thus the final model becomes:

$$\eta = \mathsf{Y} + \mathsf{M} + \mathsf{ME} \tag{C.12}$$

Thus to estimate the species composition of the catches, information on year, month and mesh size from the fishermens logbooks is desirable. At present, a different stratification is applied, viz. a stratification on year, area and month. Thus information on mesh size is not utilised. Unfortunately the information about mesh size was missing in the available dataset on information from the fishermens logbooks and the first hand buyers. Therefore this source of information was omitted from the present analysis.

The estimated weight-proportions of sandeel is shown in figure C.4.

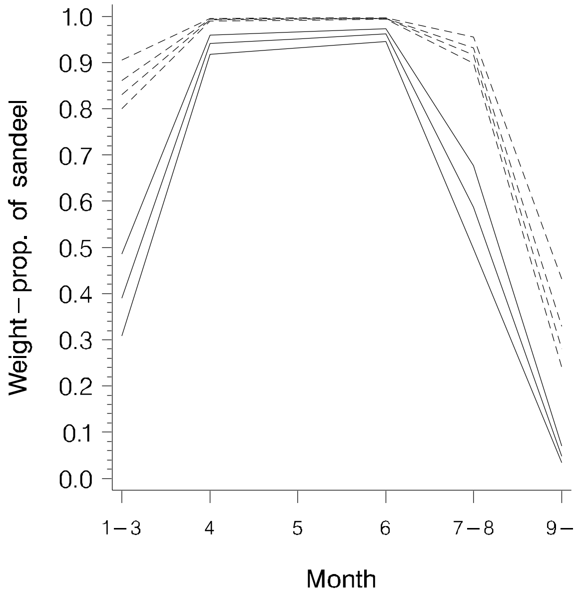

Figure C.4: *Estimated proportion of sandeel through the year for the years 1984-1991. The dashed line represents the years 1986-1989, the other the years 1984, 1985 and 1989.*

By-catches in sandeel catches

The proportion of sandeel in sandeel catches within the sandeel fishery (refer to figure C.2) is modelled by a beta-distribution. In figure C.5 the weight-percentage of sandeel is shown, together with the fitted beta-distribution. The mean proportion, μ_1:

$$\mu_1 = \mathrm{E}\{\rho|\Lambda = 1\} \tag{C.13}$$

is estimated to 0.982, and the standard deviation of this estimate, σ_1:

$$\sigma_1 = \sqrt{\mathrm{V}\{\widehat{\mu_1}\}} \tag{C.14}$$

is estimated to 0.0014. The estimates are based on 732 samples taken from sandeel catches.

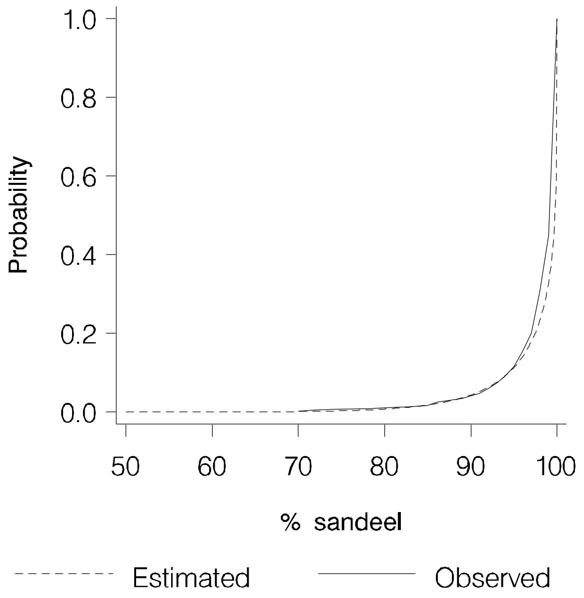

Figure C.5: *Estimated and fitted distribution of by-catches in sandeel catches based on samples collected from 1984 to 1991.*

The estimates of the proportion of correctly classified samples, $\widehat{\boldsymbol{\lambda}}$, and the weight-proportion of sandeel in the sandeel catches, $\widehat{\mu}_1$, are combined into estimates of the weight-proportion of sandeel for the whole of sandeel fishery, $\widehat{\boldsymbol{\rho}}$, and its variance and covariance matrix, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}}$.

C.8.2 Mean weight of sandeels

The mean weight of sandeels regardless of its age is analysed by an analysis of variance assuming that the distribution of the weight of a sandeel may be approximated by the normal distribution. The approximation is rather crude because the distribution is likely to be multi-modal due to the mixture of age groups. However, the estimate needed is the mean value and its variance and therefore utilising the argument of the central limit theorem, the approximation is considered to be satisfactory.

Unfortunately all effects are highly significant in the full model including all combinations of the effects $\mathsf{Y}, \mathsf{M}, \mathsf{A}$ and $\mathsf{S(A)}$. This might be caused by a very large number of DF of the residual, 91 000. With such a large number of DF small discrepancies from the normal distribution or weak confounding with latent variables may cause significance. Therefore, instead of choosing a significance level, the relative sizes of the type 3 test statistics are compared. The aim is to end up with a model which consists of only a few effects which explain a great part of the variation.

The final model was chosen to:

$$v = \mathsf{Y} + \mathsf{M} + \mathsf{A} \qquad \text{(C.15)}$$

The test statistics for the final model are shown in table 3.

Table 3. *Test statistics for fixed effects in model C.15.*

Source	NDF	DDF	Type III F	Pr > F
M	7	91E3	2172.60	0.0001
Y	6	91E3	1097.97	0.0001
A1	6	91E3	338.32	0.0001
Residual				20.81

Comparing the test statistics one can see that the month and year effect accounts for the main part of the variation, whereas the permanent geographical differences accounts for a relatively small part of the variation.

C.8.3 Combining the results of the subanalyses into estimates of catch at age and its variance

Unfortunately the industrial catch is not recorded with information on the factors that are needed for dividing the industrial fishery into the sandeel fishery and other fishery in the data available at present. Therefore previous estimates for the weight of the sandeel catch is used instead. The previous estimates and the estimates resulting from the method presented here are expected to be of the same magnitude and therefore the uncertainty estimates will only be slightly different. The weight of the catch within the sandeel fishery utilised in the estimation of the variance is estimated by

$$\widehat{\mathbf{w}}_{\mathsf{SF}} = \mathrm{diag}(\widehat{\boldsymbol{\rho}})^{-1} \times \widehat{\mathbf{w}}_{\mathsf{S}} \tag{C.16}$$

The estimated catch at age data in the sandeel areas and the coefficients of variation for these estimates for 1989 and 1991 are shown in table 4 and the correlation matrix in table 5.

Table 4. *Estimated number of sandeel of each age group (in '000). Coefficient of variation in % is shown in paranthesis.*

Age group	1989	1991
0	16 818 (27)	11 252 (76)
1	89 811 (5)	49 855 (16)
2	7 349 (42)	16 619 (29)
3	2 319 (58)	2 383 (46)
4+	2 786 (49)	462 (69)

Table 5. *Estimated correlation matrix for catch at age data.*

Age group and Year	0, 89	0, 91	1, 89	1, 91	2, 89	2, 91	3, 89	3, 91	4+, 89	4+, 91
0, 89	1	0	0.02	0	3E-6	0	3E-6	0	2E-6	0
0, 91	0	1	0	-0.76	0	-0.45	0	-0.28	0	-0.18
1, 89	0.02	0	1	0.34	-0.90	-0.46	-0.70	-0.32	-0.73	-0.23
1, 91	0	-0.76	0.34	1	-0.31	-0.21	-0.24	-0.15	-0.26	-0.12
2, 89	3E-6	0	-0.90	-0.31	1	0.45	0.44	0.18	0.47	0.14
2, 91	0	-0.45	-0.46	-0.21	0.45	1	0.29	0.48	0.31	0.34
3, 89	3E-6	0	-0.70	-0.24	0.44	0.29	1	0.39	0.52	0.04
3, 91	0	-0.28	-0.32	-0.15	0.18	0.48	0.39	1	0.33	0.46
4+, 89	2E-6	0	-0.73	-0.26	0.47	0.31	0.52	0.33	1	0.47
4+, 91	0	-0.18	-0.23	-0.12	0.14	0.34	0.04	0.46	0.47	1

One can see that the uncertainty is considerable. There is also great correlation between the estimates of the number of sandeel in the various age groups between years. This is due to the utilisation of the common structures in the data.

In order to investigate the origin of the large uncertainties the results on the intermediate stages are calculated. The estimated total sandeel catch in 1000 tons are 818 for 1989 and 699 for 1991. The coefficients of variation (std/mean) are 3% and 5% respectively. Thus the uncertainty of the species composition causes only little uncertainty of the estimated catch in weight. The estimated number of sandeel i millions are 119 for 1989 and 81 for 1991. The coefficients of variation are 11% and 6% respectively. Thus the contribution from the uncertainty of the mean weight of sandeels is small. The conclusions are that the uncertainty of the age composition contributes the most to the uncertainty.

The causes of variation in the age composition data for sandeel has been analysed in Kvist *et al.* (1999). Unfortunately, the estimated variances of the estimates of the age compositions of the rectangles do not stand up to scrutiny, because they are prone to bias and underestimation (Kuk (1995), Lin and Breslow (1996), Breslow and Lin (1995) and Booth and Hobert (1998)). Figure C.6 illustrates that the estimates of the variances of the estimates of the age compositions in the rectangles, called BLUPs (Best Linear Unbiased Predictors, (there is some disagreement in the literature on whether they should be called predictors or estimators, see Robinson (1991))) are inconsistent under basic sensible assumptions, such as that the information is greater and thus the variance smaller of an estimate for a rectangle with samples collected compared to a rectangle without samples collected.

It clearly shows that estimates for rectangles where samples have been collected have greater standard deviation than rectangles without samples taken. This is of course not reasonable for a model designed with the purpose of estimating the variance of estimates and evaluate the significance of collecting samples from each rectangle. The discrepancy is caused by an approximation of the variance that is too crude (Booth and Hobert (1998)). Booth and Hobert (1998) suggest an improved estimate of the variance, using a bootstrap estimate of the bias. However, there the implementation is time-consuming and has therefore not been attempted. The overall estimates of the uncertainties of the catch at age data are used merely as an indication of the order of magnitude.

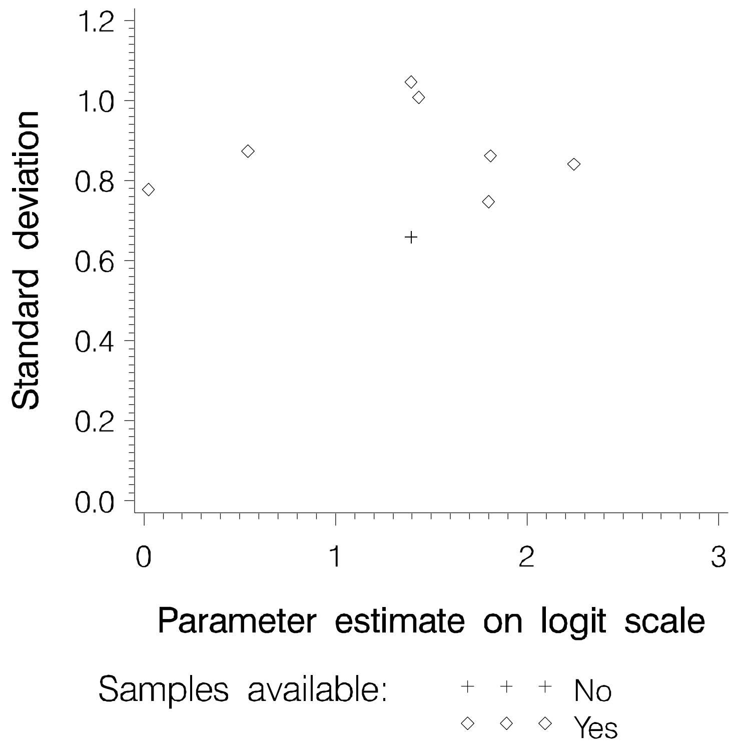

Figure C.6: *Std of BLUPs for rectangles in area 1, July 1991.*

C.9 Discussion

Despite the recent interest in quantifying risk and uncertainty in fish stock assessment (e.g. Smith *et al.*, 1993; Francis and Shotton, 1997; Flaaten *et al.*, 1998) surprisingly little effort has been spent on quantifying the uncertainty and error structure in the basic assessment data. However, in stochastic fish stock assessment models, a realistic observation error structure is necessary to avoid biased parameter estimates (Virtala *et al.*, 1998; Chen and Andrew, 1998; Chen and Paloheimo, 1998). Furthermore, knowing observation error will greatly enhance the possibilities for estimating process error (Schnute, 1987). Estimates of the catch at age data and their uncertainties are more reliable when the total variation is resolved into its components. Knowing the sources of variation sampling schemes can be improved. We found that presently some insignificant factors were utilised for stratification, whereas factors containing important information were overlooked. By establishing the significance of factors that might influence the catch composition, common structures can be recognised and utilised, and when for instance geographical or temporal differences in the catch compositions are of importance, they can be taken into account. Improving the stratification makes the estimates less prone to errors in data and reduces their variation. In addition, the identification of the common structures has the advantage that qualified estimates can be provided if observations are missing. Also more reliable predictions can be performed.

The statistical evaluation was separated into analyses of the separate data sources and combined into estimates of the catch at age data for 1989 and 1991. The catch at age data for 1990 was not estimated because of the low number of samples collected this year. The present model for the age composition may be utilised to estimate the catch at age data and its uncertainty in 1990. However, an estimate of the overall level of the year effect for 1990 would be required.

The species composition was estimated using a compound distribution to as well account for the inaccurate definition of the sandeel fishery as to account for by-catches. We found that the most important factor to explain misclassifications within the sandeel fishery is the mesh size, information not utilised today. Lewy (1996) developed a delta-Dirichlet distribution for fitting singularaties at 0 and 1. He applied it to Danish North Sea fishery data for 1993, but found that the distribution did not fit the sandeel fishery.

The mean weight of sandeels was estimated by only utilising the information of the weight from the biological samples. Improved estimates will probably be obtained if the combination of the estimated age composition and an estimated relationship of the age and weight had been utilised, i.e. the estimate of the mean weight of sandeels in a given rectangle and month, $\tilde{v}$, is

$$\tilde{v} = \sum_{a=0}^{4} \widehat{p}_a \tilde{v}_a \qquad \text{(C.17)}$$

where $\widehat{p}_a$ is the estimated proportion of age group a and $\tilde{v}_a$ is the estimated mean weight of a-year-olds. In this case a dependency between the two is present and thus the first order Taylor approximation of the variance/covariance matrix of the estimated number of sandeel caught per age group, $\widehat{\mathbf{\Sigma}}_{\mathbf{a},a_1,a_2}$, encompasses the covariance between the estimate of the mean weight and the age composition. A drawback of this approach is however that age determination errors are introduced in the estimates of the mean weight of sandeels. If such errors are of considerable magnitude, the benefits of this approach is limited.

We found that the major source of uncertainty in the catch at age data is caused by uncertainties in the estimation of the age composition. The estimation is in particular difficult because of large variations in the age composition between small areas.

The estimates presented in this paper are based on separate age distributions for each 30*30 square nautical miles ICES rectangle, because previous analyses have shown that there is large variations even between such small areas (Kvist *et al.* 1999). The estimates provided for each rectangle utilise both common structures of the age distribution and the specific observations in each rectangle. The model would however become much simpler and especially from a sampling perspective more attractive if the same age distribution could be assumed for the whole of a sandeel area. But as is illustrated in the following, this would cause bias of about 10%.

Let $\widehat{E}_S$ denote the estimate of the number of sandeels for an age group in an area, ar, within a particular month, mo, based on estimates for each ICES rectangle. $\widehat{E}_S$ is thus:

$$\widehat{E}_{S,ar,mo} = \sum_{i=1}^{k_{ar}} \widehat{p}_{ar,mo,i} \times \widehat{n}_{ar,mo,i} \qquad \text{(C.18)}$$

where $p_{ar,mo,i}$ is the proportion of sandeels of the age group of interest and $n_{ar,mo,i}$ is the number of sandeel caught. Both are stated on area ar, ICES rectangle i and month mo. k_{ar} is the number of ICES rectangles.

If we instead used an overall estimate per sandeel area of the proportion of that age group, $p_{ar,mo}$, we would obtain the following estimate of the number of sandeel in that age group:

$$\widehat{E}_{A,ar,mo} = \widehat{p}_{ar,mo} \times (\sum_{i=1}^{k_{ar}} \widehat{n}_{i,ar,mo}) \tag{C.19}$$

The bias of such an estimate is calculated as

$$bias_{ar,mo} = \frac{|\widehat{E}_{S,ar,mo} - \widehat{E}_{A,ar,mo}|}{\widehat{E}_{S,ar,mo}} \tag{C.20}$$

The average bias in % resulting from such an approach for the 1989 and 1991 data is shown in table 6.

Table 6. *Bias introduced as a consequence of using area-specific age composition estimates instead of rectangle-specific estimates.*

Age group	Mean bias (%)	Number of estimated biases	Max. bias (%)
0	9	47	57
1	4	85	38
2	8	79	39
3	10	79	61
4+	10	76	56

Although it has been documented that the variances of the BLUPs for the age composition are underestimated and biased, the uncertainties estimated for the catch at age data give an indication of the level of the uncertainty. The method as such is under all circumstances recommendable, although detailed analyses require improved estimates which can be obtained by e.g. using the corrections suggested by Booth and Hobert (1998). Attempts are made at present to improve the methods of fitting the generalised linear models. Booth and Hobert (1999) present two methods based on the Monte Carlo EM algorithm (Wei and Tanner, 1990) for finding exact maximum likelihood estimates. However, the methods break down when the intractable integrals in the likelihood function are of high dimension. Booth and Hobert (1998) suggest that approximate methods such as those suggested by e.g. Wolfinger and O'Connel (1993) should be used for model selection until the exact methods have been improved.

C.10 References

Booth, J.G. and Hobert, J.P., 1998. Standard errors of prediction in generalized linear mixed models. J. Am. Statist. Ass. 93, 262-272.

Booth, J.G. and Hobert, J.P., 1999. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. J. R. Statist. Soc. B 61, Part 1, 265-285.

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

Breslow, N.E. and Clayton, D.G., 1993. Approximate Inference in Generalized Linear Mixed Models. JASA 88, 9-25.

Breslow, N.E. and Lin, X., 1995. Bias correction in generalized linear mixed models with a single component of dispersion. Biometrika 82, 81-91.

Buslik, D., 1950. Mixing and sampling with special reference to multi-sized granular material. ASTM Bull. 165, 166.

Chen, Y. and Andrew, N., 1998. Parameter estimation in modelling the dynamics of fish stock biomass: are current used observation-error estimators reliable? Can. J. Fish. Aquat. Sci. 55, 749-760.

Chen, Y. and Paloheimo, J. E., 1998. Can a more realistic model error structure improve the parameter estimation in modelling the dynamics of fish populations? Fisheries Research 38, 9-17.

Crone, P.R. and Sampson, D. B., 1998. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition. In: Funk, F., Quinn II, T.J., Heifetz, J., Ianelli, J.N., Powers, J.E., Schweigert, J.F., Sullivan, P.J., and Zhang., C.I. (Eds.), Fishery stock assessment models, Alaska Sea Grant Program Report No. AK-SG-98-01, University of Alaska Fairbanks.

Deriso, R.B., Quinn, T.J. II and Neal, P.R., 1985. Catch-age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. 42, 815-824.

Fargo, J. and Richards, L.J., 1998. A modern approach to catch-age analysis for Hecate Strait rock sole (*Pleuronectes bilineatus*). Journal of Sea Research 39, 57-67.

Flaaten, O. Salvanes, A.G.V., Schweder, T. and Ulltang, Ø., 1998. Fisheries management under uncertainty - an overview. Fisheries Research 37, 1-6.

Fournier, D., and Archibald, C.P., 1982. A general theory for analyzing catch at age data. Can. J. Fish. Aquat. Sci. 39, 1195-1207.

Francis, R.I.C.C. and Shotton, R., 1997. "Risk" in fisheries management: a review. Can. J. Fish. Aquatic. Sci. 54, 1699-1715.

Gavaris, S. and Gavaris, C. A, 1983. Estimation of catch at age and its variance for groundfish stocks in the Newfoundland region, p. 178-182. In: Doubleday, W.G. and Rivard, D. (Eds.), Sampling commercial catches of marine fish and invertebrates. Can. Spec. Publ. Fish. Aquat. Sci. 66.

Gudmundsson, G., 1994. Time series analysis of catch-at-age observations. Appl. Stat. 43, 117-126.

Gulland, J.A., 1965. Estimation of mortality rates. Annex to Arctic Fisheries Working Group Report. ICES C.M. 1965. Doc. No. 3.

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

Kuk, A.Y.C., 1995. Asymptotically unbiased estimation in generalized linear models with random effects. J. R. Statist. Soc. B 57, 395-407.

Kvist, T., Gislason, H. and Thyregod, P., 1998. Using continuation-ratio logits to analyse the variation of the age-composition of fish catches. Submitted for publication, 1998.

Kvist, T., Gislason, H. and Thyregod, P., 1999. Analysing age-composition of sandeel landings by means of continuation-ratio logits. Submitted for publication, 1999.

Lai, H.L. and Gunderson, D.R., 1987. Effects of ageing error on estimates of growth, mortality and yield per recruit for walleye pollock (*Theragra chalcogramma*). Fish. Res. 5, 287-302.

Lewy, P., 1995. Sampling Methods and Errors in the Danish North Sea Industrial Fishery. Dana 11, 39-64.

Lewy, P., 1996. A Generalized Dirichlet Distribution Accounting for Singularities of the Variables. Biometrics 52, 336-351.

Lin, X. and Breslow, N.E., 1996. Bias correction in generalized linear mixed models with multiple components of dispersion. J. Am. Statist. Ass. 91, 1007-1016.

McAllister, M. and Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54, 284-300.

McCullagh, P. and Nelder, J.A., 1989. Generalized Linear Models (second edition). Chapman and Hall (London; New York).

Pelletier, D., 1990. Sensetivity and variance estimators for virtual population analysis and the equilibrium yield per recruit model. Aquat. Living Resour. 3, 1-12.

Pelletier, D. and Gros, P., 1991. Assessing the impact of sampling error on model-based management advice: comparison of equilibrium yield per recruit variance estimators. Can. J. Fish. Aquat. Sci. 48, 2129-2139.

Petersen, S.A., Lewy, P. and Wright, P., 1999. Assessments of the lesser sandeel (*Ammodytes marinus*) in the North Sea based on revised stock divisions. Fisheries Research (in press).

Pope, J.G., 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. Int. Comm. Northwest Atl. Fish. Res. Bull. 9, 65-74.

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

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

Rivard, D., 1983. Effects of systematic, analytical, and sampling errors on catch estimates: a sensitivity analysis, p. 114-129. In: Doubleday, W.G. and Rivard, D. (Eds.), Sampling commercial catches of marine fish and invertebrates. Can. Spec. Publ. Fish. Aquat. Sci. 66.

Rivard, D., 1989. Overview of the systematic, structural, and sampling errors in cohort analysis. Am. Fish. Soc. Symp. 6, 49-65.

Robinson, G. K., 1991. "That BLUP is a Good Thing: the estimation of random effects". Statistical Science 6, 15-32.

Scheaffer, R.L., 1969. Sampling mixtures of multi-sized particles: an application of renewal theory. Technometrics 11, no. 2, 285-298.

Schnute, J. T., 1987. Data uncertainty, model ambiguity, and model identification. Natural Resource Modelling, vol. 2(2), 159-212.

Schweigert, J.F. and Sibert, J.R., 1983. Optimizing survey design for determining age structure of fish stocks: an example from British Columbia

Pacific herring (*Clupea harengus pallasi*). Can. J. Fish. Aquat. Sci. 40, 588-597.

Smith, S.J., Hunt, J.J., Rivard, D., 1993 (Eds.). Risk evaluation and Biological Reference Points for Fisheries Management. Canadian Special Publication of Fisheries and Aquatic Sciences 120: viii + 442 p.

Sparholt, H., 1990. An estimate of the total biomass of fish in the North Sea. J. Cons. int. Explor. Mer 46, 200-210.

Sparre, P. and Venema, S.C., 1992. Introduction to tropical fish stock assessment. Part 1. Manual.-FAO Fisheries technical paper no. 306.1. Rome, FAO. 376 pp.

Tyler, A.V., Beamish, R.J. and McFarlane, G.A., 1989. Implications of age determination errors to yield estimates. In: Beamish, R.J. and McFaralane, G.A. (Eds.), Effects of ocean variability on recruitment and an evaluation of parameters used in stock assessment models. Can. Spec. Publ. Fish. Aquat. Sci. 108, pp. 27-35.

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

Wei, G. C. G. and Tanner, M. A., 1990. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. J. Am. Statist. Ass. 85, 699-704.

Wolfinger, R. and O'Connel, M., 1993. Generalized Linear Mixed Models: A pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48, 233-243.

Appendix D

Length Distributions for Age Groups

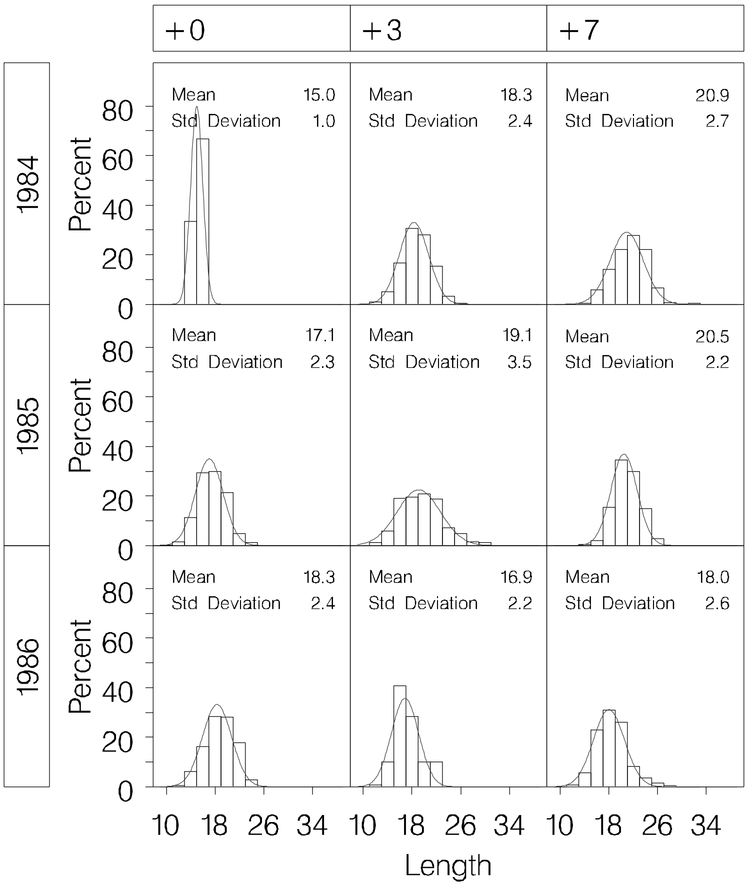

Figure D.1: Length distribution for 0-year-olds for the years 1984-1993. A normal distribution is fitted to the data.

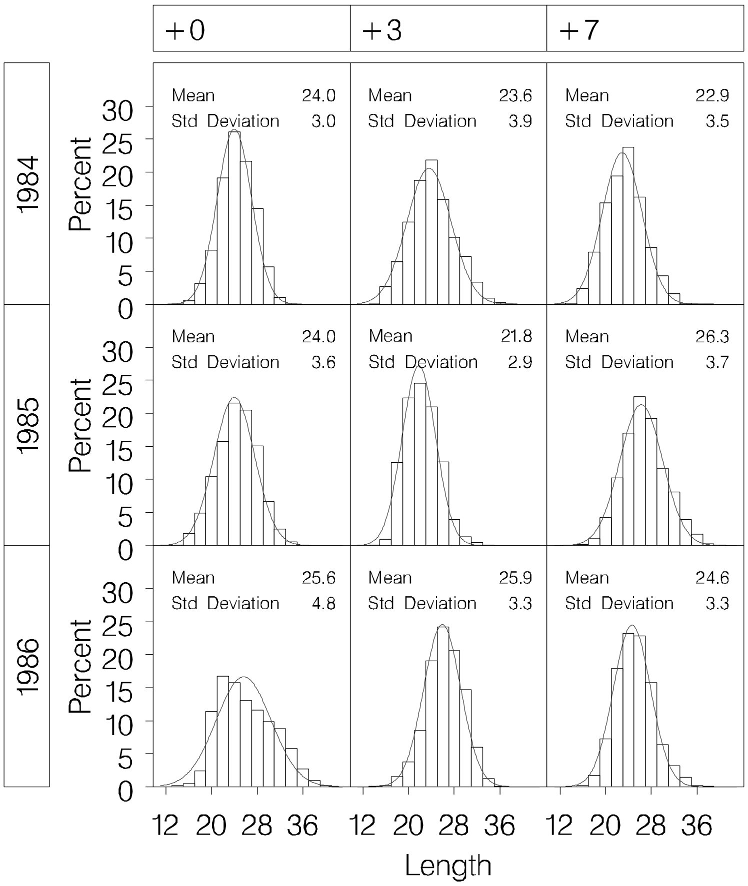

Figure D.2: Length distribution for 1-year-olds for the years 1984-1993. A normal distribution is fitted to the data.

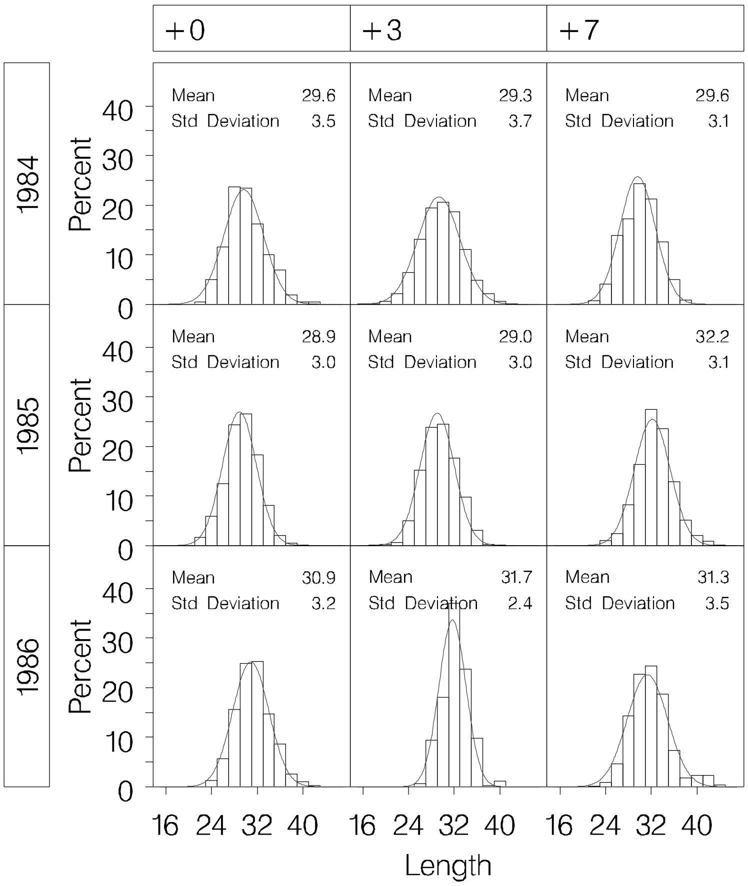

Figure D.3: Length distribution for 2-year-olds for the years 1984-1993. A normal distribution is fitted to the data.

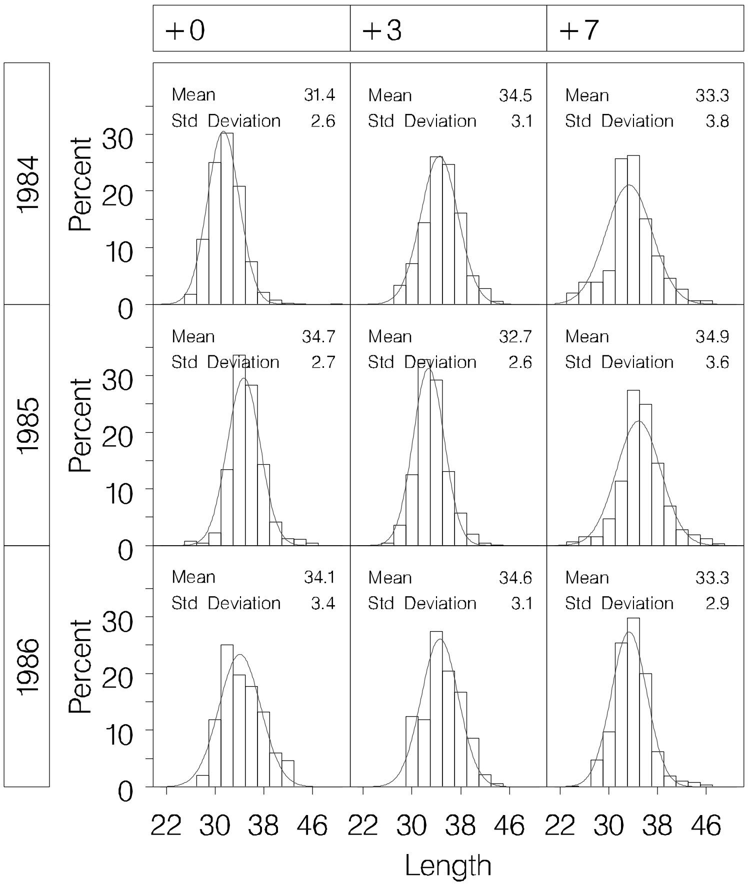

Figure D.4: Length distribution for 3-year-olds for the years 1984-1993. A normal distribution is fitted to the data.

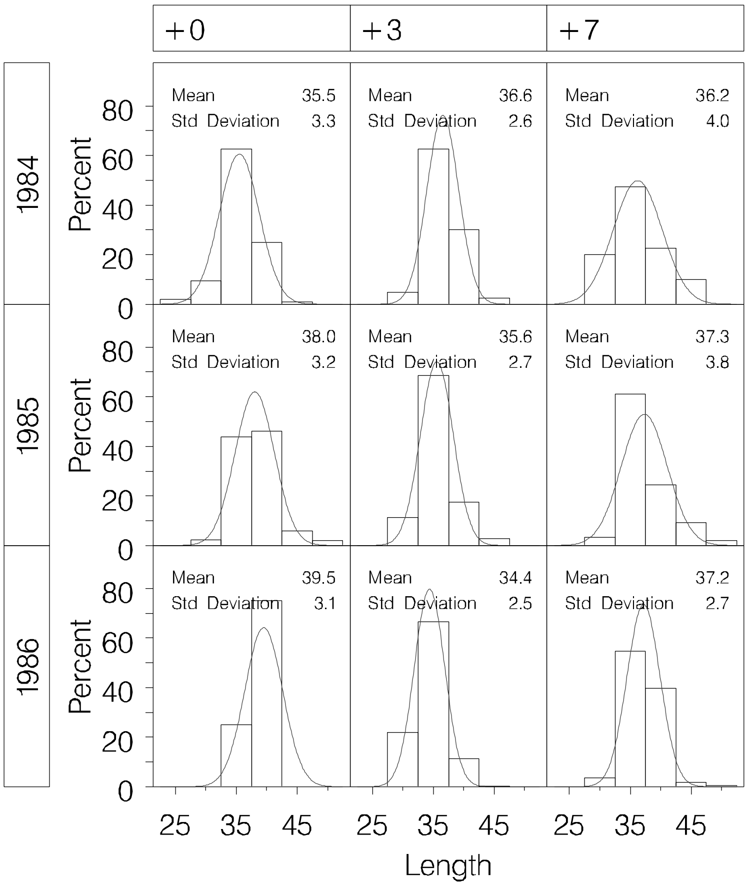

Figure D.5: Length distribution for 4-year-olds and older sandeel for the years 1984-1993. A normal distribution is fitted to the data.

Bibliography

Agger, P., Boetius, I. and Lassen, H. (1971) On errors in the virtual population analysis. International Council of the Exploration of the Sea, C.M. 1971/H:16, Copenhagen.

Agger, P., Boetius, I. and Lassen, H. (1973) Error in the virtual population: the effect of uncertaintes in the natural mortality coefficient. *Journal du Conseil, Conseil International pour l'Exploration de la Mer*, 35:93.

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons.

Beverton, R.J.H., and Holt, S.J. (1957) *On the dynamics of exploited fish populations.* Fishery Investigations series 2, Marine fisheries, Great Britain Ministry of Agriculture, Fisheries and Food, 19:533p.

Booth, J.G. and Hobert, J.P. (1998) Standard errors of prediction in generalized linear mixed models. *J. Am. Statist. Ass.*, 93: 262-272.

Booth, J.G. and Hobert, J.P. (1999) Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc.* B, 61, Part 1, pp. 265-285.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88: 9-25.

Breslow, N.E. and Lin, X. (1995) Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, 82: 81-91.

Chen, Y. and Paloheimo, J.E. (1994) Robust regression approach to estimating fish mortality rates with a cohort-based model. *Transactions of the American Fisheries Society*, 123: 508-518.

Daan, N., Bromley, P.J., Hislop, J.R.G and Nielsen, N.A. (1990) Ecology of North Sea fish. *Netherlands Journal of Sea Research*, 26: 343-386.

Dennis, B., Munholland, P.L. and Scott, J.M. (1991) Estimation of growth and extinction parameters for endangered species. *Ecological Monographs*, 61:(2) 115-143.

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

Doubleday, W.G. (1976) A least squares approach to analysing catch at age data. *Res. Bull. Int. Comm. Northw. Atl. Fish.*, 12: 69-81.

Doubleday, W.G. (1981) A method of estimating the abundance of survivors of an exploited fish population using commercial fishing catch at age and research vessel abundance indeces. *Canadian Special Publication of Fisheries and Aquatic Sciences*, 58: 164-178.

Fargo, J. and Richards, L.J. (1998) A modern approach to catch-age analysis for Hecate Strait rock sole (*Pleuronectes bilineatus*). *Journal of Sea Research*, 39: 57-67.

Fournier, D. and Archibald, C.P. (1982) A general theory for analyzing catch at age data. *Canadian Journal of Fisheries and Aquatic Science*, 39: 1195-1207.

Fournier, D. and Doonan, I.J. (1987) A length-based stock assessment method utilizing a generalized delay-difference model. *Canadian Journal of Fisheries and Aquatic Sciences*, 44: 422-437.

Fry, F. E. J. (1957) Assessment of mortalities by use of the virtual population. Proceedings of the joint scientific meeting of ICNAF (International Commission for Northwest Atlantic Fisheries), ICES (International Council for Exploration of the Sea, and FAO (Food and Agriculture Organization of the United Nations)) on fishing effort, the effects of fishing on resources and the selectivity of fishing gear, contribution P15 (mimeo), Lisbon.

Gard, T.C. (1988) *Introduction to stochastic differential equations.* Marcel Dekker, New York.

Gudmundsson, G. (1994) Time series analysis of catch-at-age observations. *Applied Statistics*, 43: (1) 117-126.

Gulland, J.A. (1965) Estimation of mortality rates. Annex to Arctic Fisheries Working Group Report (meeting in Hamburg, January, 1965). International Council of the Exploration of the Sea, C.M. 1965, document 3 (mimeo), Copenhagen.

Hoenig, J.M., Heisey, D.M. and Hanumara, R.C. (1993). Using prior and current information to estimate age composition: a new kind of age-length key. International Council for the Exploration of the Sea. Report: ICES C.M. 1993/D:52.

Horppila, J. and Peltonen, H. (1992) Optimizing sampling from trawl catches: contemporaneous multistage sampling for age and length structures. *Canadian Journal of Fisheries and Aquatic Science*, 49: 1555-1559.

ICES (1996). *Report of the ICES Advisory Committee on Fishery Management, 1995.* International Council for the Exploration of the Sea.

Jones, R. (1981) The use of length composition data in fish stock assessments (with notes on VPA and cohort analysis). FAO (Food and Agriculture Organization of the United Nations) Fisheries Circular 734.

Ketchen, K.S. (1950). Stratified subsampling for determining age distributions. *Trans. Am. Fish. Soc.*, 79: 205-212.

Kettunen, J. (1983) The extended Kalman filter approach to VPA. International Council for Exploration of the Sea, C.M. 1983/D:17.

Kimura, D.K. and Chikuni, S. (1987) Mixtures of empirical distributions: an iterative application of the age-length key. *Biometrics*, 43: 23-35.

Kirkegaard, E. and Gislason, H. (1996) *The industrial Fisheries in the North Sea.* Danish Institute for Fisheries Research.

Kuk, A.Y.C. (1995) Asymptotically unbiased estimation in generalized linear models with random effects. *J. R. Statist. Soc.*, B, 57, 395-407.

Kvist, T., Gislason, H. and Thyregod, P. (1998) Using continuation-ratio logits to analyse the variation of the age-composition of fish catches. Submitted for publication in *Journal of Applied Statistics*, 1998.

Kvist, T., Gislason, H. and Thyregod, P. (1999a) Sources of variation in the age composition of sandeel landings. Submitted for publication in *ICES Journal of Marine Science*, 1999.

Kvist, T., Gislason, H. and Thyregod, P. (1999b) Uncertainty of Catch at Age Data for Sandeel. Submitted for publication in *Fisheries Research*, 1999.

Lee, Y. and Nelder, J.A. (1996). Hierarchical generalized linear models. *J. R. Statist. Soc. B*, 58:(4) 619-678.

Lewy, P. (1988) Integrated stochastic virtual population analysis: estimates and their precision of fishing mortalities and stock sizes for the North Sea whiting stock. *J. Cons. int. Explor. Mer*, 44: 217-228.

Lin, X. and Breslow, N.E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Statist. Ass.*, 91, 1007-1016.

Lungu, E.M. and Øksendal, B. (1997) Optimal harvesting from a population in a stochastic crowded environment. *Mathematical Biosciences*, 145:(1) 47-75.

Macer, C.T. (1966) Sand eels (*Ammodytidae*) in the south-western North Sea; their biology and fishery. Fishery investigations, series 2, vol. 24, no. 6. Ministry of agriculture, fisheries and food, G.B.

Madsen, H., van Dijk, D., Hansen, L., Graaf, M., Söderström, T., Wouters, P. Norlen, U., Anderlind, G., Gutschker, O., Neirac, F., Kreider, J., Guy, A., Händel, P., Nielsen, J.N., Versluis, R., Bloem, H. (1996). *System identification competition.* Joint research centre, European Commission. Ed. Bloem, J.J. ISBN 92-827-6348-X. ECSC-EC-EAEC Brussels, Luxembourg.

Madsen, H. and Holst, J. (1995) Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22: 67-79.

Madsen, H. and Melgaard, H. (1991) *The mathematical and numerical methods used in CTLSM - a program for ML-estimation in stochastic, continuous time dynamical models.* Research report no. 7/1991, Department of Mathematical Modelling, Technical University of Denmark.

Martin, I. and Cook, R.M. (1990) Combined analysis of length and age-at-length data. *J. Cons. int. Explor. Mer.*, 46: 178-186.

Matsuishi, T. (1998) Examination of a length-based population analysis. *In* Fishery stock assessment models, edited by F. Funk, T.J. Quinn II, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.I. Zhang. Alaska Sea Grant Program Report No. AK-SG-98-01, University of Alaska Fairbanks, 1998.

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models (second edition).* Chapman and Hall (London;New York).

Megrey, B.A. (1989) Review and comparison of age-structured stock assessment models from theoretical and applied points of view. American Fisheries Society Symposium, 6: 8-48.

Melgaard, H. (1994) *Identification of physical models.* PhD IMM-PHD-1994-4, Department of Mathematical Modelling, Technical University of Denmark. ISSN 0909-3192.

Mendelssohn, R. (1988) Some problems in estimating population sizes from catch-at-age data. *Fishery Bulletin*, 86:(4) 617-630.

Mohn, R.K. and Cook, R. (1993) *Introduction to sequential population analysis.* Scientific council studies, no. 17. Northwest Atlantic fisheries organization.

Murphy, G.I. (1965) A solution of the catch equation. *Journal of the Fisheries Research Board of Canada*, 22: 191-202.

Nandram, B., Sedransk, J. and Smith, S.J. (1997) Order-restricted Bayesian Estimation of the Age Composition of a population of Atlantic cod. *Journal of the American Statistical Association*, 92:(437) 33-40.

Nielsen, J.N., Vestergaard, M. and Madsen, H. (1999) Estimation in continuous-time stochastic volatility models using nonlinear filters. *International Journal of Theoretical and Applied Finance* (To appear).

Paloheimo, J.E. (1958) A method of estimating natural and fishing mortalities. *Journal of the fisheries research board of Canada*, 15: 749-758.

Pope, J.G. (1972) An investigation of the accuracy of virtual population analyses using cohort analysis. *International Commission for the Northwest Atlantic Fisheries Research Bulletin*, 9: 65-74.

Pope, J.G. (1977) Estimation of fishing mortality, its precision and implications for the management of fisheries, p. 63-76. *In* J.H. Steele (ed.), *Fisheries mathematics.* Academic Press, New York, NY.

Pope, J.G. and Shepherd, J.G. (1982) A simple method for the consistent interpretaion of catch-at-age data. *J. Cons. Explor. Mer*, 40: 176-184.

Pope, J.G. and Shepherd, J.G. (1985) A comparison of the performance of various methods for tuning VPAs using effort data. *Journal du Conseil, Conseil International pour l'Exploration de la Mer*, 42: 129-151.

Punt, A.E. and Hilborn, R. (1997) Fisheries stock assessment and decision analysis: the Bayesian approach. *Reviews in Fish Biology and Fisheries*, 7: 35-63.

Quinn II, T.J., Turnbull, C.T. and Fu, C. (1998) A length-based population model for hard-to-age invertebrate populations. *In* Fishery stock assessment models, edited by F. Funk, T.J. Quinn II, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.I. Zhang. Alaska Sea Grant Program Report No. AK-SG-98-01, University of Alaska Fairbanks, 1998.

Ralston, s. and Ianelli, J.N. (1998) When lengths are better than ages: the complex case of Bocaccio. *In* Fishery stock assessment models, edited by F. Funk, T.J. Quinn II, J. Heifetz, J.N. Ianelli, J.E. Powers, J.F. Schweigert, P.J. Sullivan, and C.I. Zhang. Alaska Sea Grant Program Report No. AK-SG-98-01, University of Alaska Fairbanks, 1998.

Reed, W.J. and Simons, C.M. (1996) Analyzing catch-effort data by means of the Kalman filter. *Canadian Journal of Fisheries and Aquatic Science*, 53: 2157-2166.

Reeves, S.A (1994). Seasonal and annual variation in catchability of sandeels at Shetland, ICES CM 1994/D:19 (mimeo.).

Richards, L.J. and Schnute, J.T. (1992) Statistical models for estimating CPUE from catch and effort data. *Canadian Journal of Fisheries and Aquatic Science*, 49: 1315-1327.

Ricker, W.E. (1954) Stock and recruitment. *J. Fish. Res. Board. Can.* 11: 559-623.

Ricker, W.E. (1971) Derzhavin's biostatistical method of population analysis. *Journal of the Fisheries Research Board of Canada*, 28: 1666-1672.

Ricker, W.E. (1975) Computation and interpretation of biological statistics of fish populations. *Fisheries Research Board of Canada Bulletin 191.*

Robinson, G.K. (1991) "That BLUP is a Good Thing: the estimation of random effects". *Statistical Science*, 6: 15-32.

Schnute, J.T. (1994) A general framework for developing sequential fisheries models. *Canadian Journal of Fisheries and Aquatic Science*, 51: 1676-1688.

Schweigert, J.F. and Sibert, J.R. (1983) Optimizing survey design for determining age structure of fish stocks: an example from British Columbia Pacific herring (*Clupea harengus pallasi*). *Canadian Journal of Fisheries and Aquatic Science*, 40: 588-597.

Skagen, D.W. (1994) Revision and extension of the Seasonal Extended Survivors Analysis (SXSA). Working Document, Norway Pout and Sandeel Assessment WG.

Sparholt, H. (1990) An estimate of the total biomass of fish in the North Sea. *Journal du Conseil International pour l'Exploration de la Mer*, 46: 200-210.

Sullivan, P.J. (1992) A Kalman filter approach to catch-at-length analysis. *Biometrics*, 48: 237-257.

Ulltang, O. (1976) Catch per unit effort in the Norwegian purse seine fishery for Atlanto-Scandian (Norweigian spring spawning) herring. FAO (Food and Agriculture Organization of the United Nations) Fisheries Technical Paper, 155: 91-101.

Vetter, E.F. (1988). Estimation of natural mortality in fish stocks: A review. *Fishery Bulletin*, 86:(1) 25-43.

Wei, G.C.G. and Tanner, M.A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Am. Statist. Ass.*, 85: 699-704.

Widrig, T. (1954) Method of estimating fish populations, with applications to Pacific sardines. *U.S. Fish and Wildlife Fishery Bulletin*, 56: 141-166.

Wolfinger, R. and O'Connel, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation*, 48: 233-243.

Wright, P.J. (1996) Is there a conflict between sandeel fisheries and seabirds? A case study at Shetland. *In* Aquatic predators and their prey. *Edited by* S.P.R. Greenstreet and M.L. Tasker. Fishing News Books, Blackweell Science Ltd. Oxford.

Øksendal, B. (1995) *Stochastic differential equations, An introduction with applications, fourth edition.* Springer.

Ph. D. theses from IMM

1. **Larsen, Rasmus.** (1994). *Estimation of visual motion in image sequences.* *xiv* + 143 pp.
2. **Rygard, Jens Moberg.** (1994). *Design and optimization of flexible manufacturing systems.* *xiii* + 232 pp.
3. **Lassen, Niels Christian Krieger.** (1994). *Automated determination of crystal orientations from electron backscattering patterns.* *xv* + 136 pp.
4. **Melgaard, Henrik.** (1994). *Identification of physical models.* *xvii* + 246 pp.
5. **Wang, Chunyan.** (1994). *Stochastic differential equations and a biological system.* *xxii* + 153 pp.
6. **Nielsen, Allan Aasbjerg.** (1994). *Analysis of regularly and irregularly sampled spatial, multivariate, and multi-temporal data.* *xxiv* + 213 pp.
7. **Ersbøll, Annette Kjær.** (1994). *On the spatial and temporal correlations in experimentation with agricultural applications.* *xviii* + 345 pp.
8. **Møller, Dorte.** (1994). *Methods for analysis and design of heterogeneous telecommunication networks.* Volume 1-2, *xxxviii* + 282 pp., 283-569 pp.
9. **Jensen, Jens Christian.** (1995). *Teoretiske og eksperimentelle dynamiske undersøgelser af jernbanekøretøjer.* ATV Erhvervsforskerprojekt EF 435. *viii* + 174 pp.
10. **Kuhlmann, Lionel.** (1995). *On automatic visual inspection of reflective surfaces.* ATV Erhvervsforskerprojekt EF 385. Volume 1, *xviii* + 220 pp., (Volume 2, *vi* + 54 pp., fortrolig).
11. **Lazarides, Nikolaos.** (1995). *Nonlinearity in superconductivity and Josephson Junctions.* *iv* + 154 pp.
12. **Rostgaard, Morten.** (1995). *Modelling, estimation and control of fast sampled dynamical systems.* *xiv* + 348 pp.
13. **Schultz, Nette.** (1995). *Segmentation and classification of biological objects.* *xiv* + 194 pp.
14. **Jørgensen, Michael Finn.** (1995). *Nonlinear Hamiltonian systems.* *xiv* + 120 pp.
15. **Balle, Susanne M.** (1995). *Distributed-memory matrix computations.* *iii* + 101 pp.
16. **Kohl, Niklas.** (1995). *Exact methods for time constrained routing and related scheduling problems.* *xviii* + 234 pp.

17. **Rogon, Thomas.** (1995). *Porous media: Analysis, reconstruction and percolation. xiv* + 165 pp.
18. **Andersen, Allan Theodor.** (1995). *Modelling of packet traffic with matrix analytic methods. xvi* + 242 pp.
19. **Hesthaven, Jan.** (1995). *Numerical studies of unsteady coherent structures and transport in two-dimensional flows.* Risø-R-835(EN) 203 pp.
20. **Slivsgaard, Eva Charlotte.** (l995). *On the interaction between wheels and rails in railway dynamics. viii* + 196 pp.
21. **Hartelius, Karsten.** (1996). *Analysis of irregularly distributed points. xvi* + 260 pp.
22. **Hansen, Anca Daniela.** (1996). *Predictive control and identification - Applications to steering dynamics. xviii* + 307 pp.
23. **Sadegh, Payman.** (1996). *Experiment design and optimization in complex systems. xiv* + 162 pp.
24. **Skands, Ulrik.** (1996). *Quantitative methods for the analysis of electron microscope images. xvi* + 198 pp.
25. **Bro-Nielsen, Morten.** (1996). *Medical image registration and surgery simulation. xxvii* + 274 pp.
26. **Bendtsen, Claus.** (1996). *Parallel numerical algorithms for the solution of systems of ordinary differential equations. viii* + 79 pp.
27. **Lauritsen, Morten Bach.** (1997). *Delta-domain predictive control and identification for control. xxii* + 292 pp.
28. **Bischoff, Svend.** (1997). *Modelling colliding-pulse mode-locked semiconductor lasers. xxii* + 217 pp.
29. **Arnbjerg-Nielsen, Karsten.** (1997). *Statistical analysis of urban hydrology with special emphasis on rainfall modelling.* Institut for Miljøteknik, DTU. *xiv* + 161 pp.
30. **Jacobsen, Judith L.** (1997). *Dynamic modelling of processes in rivers affected by precipitation runoff. xix* + 213 pp.
31. **Sommer, Helle Mølgaard.** (1997). *Variability in microbiological degradation experiments - Analysis and case study. xiv* + 211 pp.
32. **Ma, Xin.** (1997). *Adaptive extremum control and wind turbine control. xix* + 293 pp.
33. **Rasmussen, Kim Ørskov.** (1997). *Nonlinear and stochastic dynamics of coherent structures. x* + 215 pp.
34. **Hansen, Lars Henrik.** (1997). *Stochastic modelling of central heating systems. xxii* + 301 pp.

35. **Jørgensen, Claus.** (1997). *Driftsoptimering på kraftvarmesystemer.* 290 pp.
36. **Stauning, Ole.** (1997). *Automatic validation of numerical solutions.* *viii* + 116 pp.
37. **Pedersen, Morten With.** (1997). *Optimization of recurrent neural networks for time series modeling.* *x* + 322 pp.
38. **Thorsen, Rune.** (1997). *Restoration of hand function in tetraplegics using myoelectrically controlled functional electrical stimulation of the controlling muscle.* *x* + 154 pp. + Appendix.
39. **Rosholm, Anders.** (1997). *Statistical methods for segmentation and classification of images.* *xvi* + 183 pp.
40. **Petersen, Kim Tilgaard.** (1997). *Estimation of speech quality in telecommunication systems.* *x* + 259 pp.
41. **Jensen, Carsten Nordstrøm.** (1997). *Nonlinear systems with discrete and continuous elements.* 195 pp.
42. **Hansen, Peter S.K.** (1997). *Signal subspace methods for speech enhancement.* *x* + 214 pp.
43. **Nielsen, Ole Møller.** (1998). *Wavelets in scientific computing.* *xiv* + 232 pp.
44. **Kjems, Ulrik.** (1998). *Bayesian signal processing and interpretation of brain scans.* *iv* + 129 pp.
45. **Hansen, Michael Pilegaard.** (1998). *Metaheuristics for multiple objective combinatorial optimization.* *x* + 163 pp.
46. **Riis, Søren Kamaric.** (1998). *Hidden markov models and neural networks for speech recognition.* *x* + 223 pp.
47. **Mørch, Niels Jacob Sand.** (1998). *A multivariate approach to functional neuro modeling.* *xvi* + 147 pp.
48. **Frydendal, Ib.** (1998.) *Quality inspection of sugar beets using vision.* *iv* + 97 pp. + app.
49. **Lundin, Lars Kristian.** (1998). *Parallel computation of rotating flows.* *viii* + 106 pp.
50. **Borges, Pedro.** (1998). *Multicriteria planning and optimization. - Heuristic approaches.* *xiv* + 219 pp.
51. **Nielsen, Jakob Birkedal.** (1998). *New developments in the theory of wheel/rail contact mechanics.* *xviii* + 223 pp.
52. **Fog, Torben.** (1998). *Condition monitoring and fault diagnosis in marine diesel engines.* *xii* + 178 pp.
53. **Knudsen, Ole.** (1998). *Industrial vision.* *xii* + 129 pp.
54. **Andersen, Jens Strodl.** (1998). *Statistical analysis of biotests. -*

Applied to complex polluted samples. *xx* + 207 pp.
55. **Philipsen, Peter Alshede.** (1998). *Reconstruction and restoration of PET images.* *vi* + 134 pp.
56. **Thygesen, Uffe Høgsbro.** (1998). *Robust performance and dissipation of stochastic control systems.* 185 pp.
57. **Hintz-Madsen, Mads.** (1998). *A probabilistic framework for classification of dermatoscopic images.* *xi* + 153 pp.
58. **Schramm-Nielsen, Karina.** (1998). *Environmental reference materials methods and case studies.* *xxvi* + 261 pp.
59. **Skyggebjerg, Ole.** (1999). *Acquisition and analysis of complex dynamic intra- and intercellular signaling events.* 83 pp.
60. **Jensen, Kaare Jean.** (1990). *Signal processing for distribution network monitoring.* *x* + 140 pp.
61. **Folm-Hansen, Jørgen.** (1999). *On chromatic and geometrical calibration.* *xiv* + 241 pp.
62. **Larsen, Jesper.** (1999). *Parallelization of the vehicle routing problem with time windows.* *viii* + 241 pp.
63. **Clausen, Carl Balslev.** (1999). *Spatial solitons in quasi-phase matched struktures.* *vi* + (flere pag.)
64. **Kvist, Trine.** (1999). *Statistical modelling of fish stocks.* *xiv* + 175 pp.