Inge Christoffer Olsen

The Analysis of Continuous Mark-Recapture Data

Doktoravhandling
for graden doktor ingeniør

Trondheim, mai 2006

Norges teknisk-naturvitenskapelige universitet
Fakultet for informasjonsteknologi, matematikk og elektroteknikk
Institutt for matematiske fag

NTNU
Det skapende universitet

NTNU
Norges teknisk-naturvitenskapelige universitet

Doktoravhandling
for graden doktor ingeniør

Fakultet for informasjonsteknologi, matematikk og elektroteknikk
Institutt for matematiske fag

ISBN 82-471-7844-3 (trykt utg.)
ISBN 82-471-7843-5 (elektr utg.)
ISSN 1503-8181

Doktoravhandlinger ved NTNU,

Trykt av Tapir Uttrykk

Preface

As I sit here under the shady trees of the Pines after an early climb in the wonderful Mt. Arapiles, it is hard to imagine that I have spent the last five years contemplating on statistics. But I have, and this thesis, submitted as partial fulfilment of the requirements for the degree of "Doctor Ingeniør" (Dr.Ing.) at the Norwegian University of Science and Technology (NTNU), is proof thereof. The research was financed by The Research Council of Norway under the BeMatA program, and was carried out at the Department of Mathematical Sciences.

In connection with this thesis, I would like to thank my supervisor Professor Håvard Rue and co-supervisor Professor Arnoldo Frigessi for trusting me with this assignment, and for guiding me safely through it. During my stay at the Norwegian Computing Centre I enjoyed the collaboration of Øyvind Skare and Christian Brink, and the advices on biological matters from Professor Nils Christian Stenseth.

I owe my colleagues at the group of statistics for their questions, suggestions and laughs. The administration staff I thank for their support, especially for providing me with shelter during the movement period. The computer management team I reward with a gold medal for providing the most reliable computer systems I have ever known, and for being helpful whenever help was needed.

I thank my mother and father for the initial phases of this thesis (the first 25 years), my sister for the stair exercise, and the rest of my family for being nice to me. A most warm hug I give my grandmother for her waffles, her "kringler" and every little story of the war and Fritjof Nansen. My gratitude also goes to my friends for providing me with an alternative to the office, and to the "12-kaffe" for good company during lunch.

Finally, to my wife Kjersti: Thank you for the laughs, the trips, the climbs and the love.

Mt. Arapiles, Australia
January 2006

Inge Christoffer Olsen

Thesis Outline

The thesis consists of the following parts:

Part I: A Review of the Mark-Recapture Methodology Report.

Part II: Analysing the Risør Coastal Cod Mark-Recapture Experiment Report.

Part III: Analysing Continuous Mark-Recapture Data on Open Populations Report.

Part IV: Assessing the Effects of Lengthy Sampling Periods on Mark-Recapture Methodology Submitted.

Appendix: Analysing Continuous Mark-Recapture Data on Open Populations Submitted.

Background

The mark-recapture census has been one of the major experimental methods for gaining information on wildlife populations over the last 100 years. In short, mark-recapture methods generally consist of marking a subset of the population in question, and then registering marked individuals during subsequent recaptures. The term "recapture" is not to be taken literary in this context, as the registering is often done by resighting only. In the following, the term "mark-recapture" is used in connection with any experiment utilising marks and registering of marked individuals of wildlife populations.

Mark-recapture methods have evolved as the statistical methods for inference have been developed. The first models (the Petersen and the Schnabel models) were developed during the first half of the twentieth century to estimate the size of wildlife populations, and are dependent of several limiting assumptions. Firstly, the population must be closed, meaning that there may be no additions or subtractions (such as births or deaths) to the population during the experiment. Secondly, every individual of the experiment must have the same probability of capture at each sampling occasion. Lastly, marked individuals cannot loose their marks and all recaptured marks must be reported. These assumptions are very strict, and very few experiments are able to meet them. Much attention has therefore been given to relax them. In fact, it is possible to claim that all development of mark-recapture models is based on relaxing the assumptions of the first models.

The problem with mark loss is usually dealt with by double-tagging studies, in which some animals are marked with two two tags. The tag-loss probability is estimated by the proportion of double-tagged individuals who have lost one tag. This probability is then included in the models.

Capture heterogeneity between individuals and sampling occasions has been given much attention. Three main reasons for heterogeneity has been recognised: temporal variation; behavioural response; and heterogeneity among individuals.

Temporal variation indicates that there are differences in capture probability between the sampling occasions. This may be due to unequal sampling effort or changes in the environmental conditions. When the capture probability changes as a result of previous capture, it is known as behavioural response. The response may be either trap-happiness (increased probability of capture after initial capture) or trap-shyness (decreased probability of capture after initial capture). Heterogeneity between individuals apply when the capture probability is expected to differ between each individual of the population. The models associated with this conception are often parsimonious, and restrictions are usually applied. For a thorough consideration of the models developed as a response to capture heterogeneity, see Williams et al. (2002).

The closure assumption was relaxed independently by Jolly (1965) and Seber (1965), and the resulting model is known as the Jolly-Seber model. The development of this model was a major achievement, and much attention was given in the further development of open population models. The Jolly-Seber model includes temporal variation, but assumes equal capture and survival probabilities between the individuals of the experiment. In addition, the model assumes no tag loss, no emigration from the sample area, independency between the marked animals, and instantaneous sampling periods. Of the assumptions, all but the last two has received more or less attention. For a review of the development, again refer to Williams et al. (2002).

The assumption of instantaneous sampling periods has not yet been addressed in a general matter. A solution has been suggested using so-called robust design (see e.g. Schwarz and Stobo (1997)), but this solution still assumes instantaneous sampling to some extent. Models for continuously sampled closed populations exists, and good abundance estimators are given by e.g. Wang and Yip (2003). Models for continuously sampled open populations are non-existent, and are the subject of this thesis.

Summary

The purpose of this thesis is to develop a general methodology for the analysis of continuously sampled data from open populations. The thesis consists of four parts. and one appendix. The four parts may be read independently of each other. The appendix consists of a submitted article based on the developments in Part III.

The aim of Part I is to present the major contributions to the mark-recapture methodology over the last century. It is not a complete reference to the subject, but gives the reader an introduction to mark-recapture models. In addition to the classic models, some recent development of interest is included. The introduction of Bayesian statistics to mark-recapture models is covered, in addition to the analysis of continuously sampled data from closed populations. With the ever increasing amount of models available, the need for model selection tools is grave. We look at some of the most used methods such as the Likelihood Ratio Test (LRT) and the Akaike Information Criterion in connection with mark-recapture models. For the Bayesian models, we review some of the latest developments in model selection such as the Deviance Information Criterion.

Part II addresses the challenge of analysing a set of continuously sampled mark-recapture data from a Norwegian coastal cod population (the Risør data set). These data are already analysed with discrete models by Julliard et al. (2001), but it is suggested that this approach may lead to biased estimates of the survival. The approach taken in Part II is to renounce the discrete models for mark-recapture data, and to develop an new, continuous model. The analysis of this model is preformed by a Monte Carlo EM algorithm. The survival and capture estimates are presented and compared with the estimates of Julliard et al. (2001). The Risør data set contains much auxilliary information such as size at release and capture, release location and capture gear. A semi-parametric multiplicative hazard model is implemented to include some of this information.

The EM algorithm from Part II is developed further in Part III. Instead of a Monte Carlo simulation in the E step of the algorithm, an analytic solution is presented. The algorithm is developed for the discrete Cormack-Jolly-Seber(CJS) model, and shows nice convergence and stability properties for this model. An argument is then given for applying the CJS model directly to continuous data. The stability of the EM algorithm makes it suitable for such analysis. When covariate information is to be included into the analysis of continuous data, the CJS model is no longer appropriate. A semi-parametric multiplicative model for the survial and capture processes is presented as an answer to this challenge. The EM algorithm is still used for analysis. The methodology is applied to a set of continuously sampled data on the Eurpean Dipper, gathered by G. Marzolin (Marzolin (2002)). These data have been analysed discretely by Lebreton et al. (1992) (frequentistic) and by Brooks et al. (2000) (Bayesian).

One of the assumptions of the CJS model is that the sampling is done instantaneously. The aim of Part IV is to assess the consequences of violating this assumption. A simulation analysis showes that there is potential for substantial bias of the survival estimates when the sampling periods are long.

References

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, 15(4):357–76.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52:225–47.

Julliard, R., Stenseth, N. C., Gjøsæter, J., Lekve, K., Fromentin, J.-M., and Danielssen, D. S. (2001). Natural mortality and fishing mortality in a costal cod population: a release-recapture experiment. *Ecological Applications*, 11(2).

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Marzolin, G. (2002). Influence of the mating system of the eurasian dipper on sexspesific local survival rates. *Journal of wildlife management*, 66(4):1023–30.

Schwarz, C. J. and Stobo, W. T. (1997). Estimating temporary migration using the robust design. *Biometrics*, 53:178–94.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, pages 249–59.

Wang, Y. and Yip, P. S. F. (2003). A semiparametric model for recapture experiments. *Scandinavian journal of statistics*, 30(4):667–76.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic press.

situations. The population size estimator for the different models are denoted accordingly, such that $\hat{N}_b$ is the population size estimator for the M_b model and so on.

We begin by addressing the time dependent model M_t. This is natural, since this is the extension of the Peterson model to more than two sample periods. The model was first introduced by Schnabel (1938), and received much attention in the following decades. It is known as the Schnabel model, and uses the basic assumptions of mark-recapture, most notably that all individuals have the same probability of capture at any given capturing occasion. For the Schnabel model, each individual in the population is treated as a trial which may result in k^2 possible outcomes, defined by the capture history ω. Remember that a_ω is the number of individuals with capture history ω. If the trials are independent, the joint probability function of the random variables $\{a_\omega\}$ is multinomial

$$f(\{a_\omega\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{s+1})!} Q^{N-M_{s+1}} \prod_\omega P_\omega^{a_\omega},$$

where $Q = 1 - \sum_\omega P_\omega$ (the probability of never being captured) and P_ω is the probability of sampling history ω. If there is independence between samples, P_ω is given by a product of the sample capture probabilities. The set of parameters in this model is thus the population size N and p_j, $j = 1, 2, \ldots, s$ where p_j is the probability of being captured in the jth occasion.

The reason for using the multinomial distribution is two-folded. First, it gives a model with parameters which are biologically interpretable (the capture probabilities $p_1, p_2, \ldots, p_k$). Second, the sample size can be treated as random instead of fixed, as in the hyper-geometric case.

Following Darroch (1958) the resulting joint probability distribution of $\{a_\omega\}$ is given by

$$f(\{a_\omega\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{k+1})!} \prod_{i=1}^{k} p_i^{n_i}(1 - p_i)^{N-n_i} \tag{2}$$

From the assumption that each individual has the same probability of being caught, and that they are independent, the $\{n_i\}$ may be regarded as independent binomial variables with distribution

$$f(\{n_i\}) = \prod_{i=1}^{k} \binom{N}{n_i} p_i^{n_i}(1 - p_i)_i^{N-n_i}.$$

The equation (2) is also the distribution of $\{a_\omega\} \cap \{n_i\}$ since all information on $\{n_i\}$ is contained in $\{a_\omega\}$. This means that the joint probability distribution of $\{a_\omega\}$ conditional on fixed sample sizes $\{n_i\}$ is

$$f(\{a_\omega\}|\{n_i\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{k+1})!} \prod_{i=1}^{k} \binom{N}{n_i}^{-1}.$$

For this expression Darroch (1958) computes the maximum likelihood estimate $\hat{N}$ of N as the unique root greater than M_{k+1} of the $(k-1)$th degree polynomial given by

A Review of the Mark-Recapture Methodology

Inge Christoffer Olsen

Abstract

This review aims to present the major statistical contributions to the analysis of mark-recapture experiments. It presents the evolution of mark-recapture methods from the simple Petersen estimator to the state-of-the-art solutions of today. The models and sampling schemes are presented in a unified manner, pointing out similarities and differences when these occur.

1 Basic concepts

A population is a group of individuals, either animals or human beings, that have some common features. Usually we mean a specific species living on a limited area, such as a group of bears living in a forest region or trout residing in a certain lake. There is a number of potentially interesting features connected to a population. The population size is may often be of interest for several reasons. For fish populations, the abundance is important in order to regulate the commercial catch rates, while monitoring the population size for endangered species may prevent distinction. Other interesting features include immigration and emigration in addition to mortality and recruitment. These are characteristics that are essential for understanding the dynamics and biology of the population.

When we study a population, we must decide under what conditions the study is performed. A population living in a laboratory is easy to monitor, but may give limited information on populations of the same species living outside the laboratory. However, with surveys performed on natural outdoor living populations, there is an issue of interference. Almost all survey will disturb the population in some sense. The aim is therefore to gather as much information as possible, without inflicting any major disturbance. In cases where disturbance is unavoidable, this should be taken into consideration in the analysis of the resulting data.

Populations are either closed or open. With closed populations we understand that there are no emigration or immigration, no births or deaths or any other additions or removals. No population is closed over time, but for a relative short time scope, a population may often be regarded as closed. Many vertebrae populations and other large animal populations with low birth and death rates are usually regarded as closed if the survey is done within reasonable time. A slight generalisation is to regard known losses (trap death or deliberate removals) as no violation of the closure assumption.

In cases when the closure assumption is violated, the population is regarded as open. A survey that lasts for a year or more is usually regarded as an open population survey. With open populations, birth/death rates and immigration/emigration rates are important features.

2 Mark-recapture methods

There are three major sampling methods for wildlife populations: distance sampling, mark-recapture sampling and harvest sampling. In short, distance sampling is mainly sighting and counting members of the population while the researcher is moving along a preset course, while harvest sampling is mainly non-scientific captures of the population such as commercial fisheries. We do not pursuit these sampling methods further here.

The mark-recapture method generally consists of marking a subset of the population such that it is distinguishable from the unmarked members of the population. Subsequent samples from the population are made, the marked and possibly unmarked individuals are identified, and form the data from which inference on the population is drawn. In this review we concentrate on mark-recapture methods.

There are differences between mark-recapture experiments. Some of them use anonymous tags, some use identification tags, some use tags that disappear. The different marking methods determines the amount of information provided by the experiment. We list some of the usual choices, following Seber (1982):

Paints and dyes This mark is generally a change of colour on some part of the animal body. This may be applied manually by e.g. catching and painting, or automatically, e.g. by setting up a trap that marks the animal instead of killing it. Marking with paint or dye usually means that the mark is anonymous. The advantage of this method is that it is usually fast, easy and does not disturb or harm the individual. It is also suitable for sampling by sighting instead of catching. The drawback is that the individual is not uniquely identifiable, and that the marking usually disappears after some time. The paint may also change the animals attributes if it is not done properly. Painting a winter-coloured rabbit with red paint will most likely introduce bias in the survival rate estimation.

Tags By tags, we mean a tag, a band or a ring that is attached externally. These are mainly identification marks, and are similar to those used by farmers to keep track of their livestock. Many bird studies are performed by ring-marking. The advantage is that tags are permanent and identify the individual uniquely. The disadvantage is that the animal must be captured for the tag to be applied, and a further capture is usually needed for registration.

Mutilation This marking method consists of a change in the physical appearance of the animal such as fin clipping on fish or toe clipping on small mammals. The advantage is that it is permanent, the drawback is that it may change

behaviour and survival rate of the animal. This is also usually an anonymous marking method.

Radioactive isotopes Radioactive isotopes are applied to the animal, either external or internal through food. This marking is related to paints and dyes, but the registration process is done by radioactivity registrants. It may cause behavioural differences, and can also be dangerous to animals and personnel.

Parasites Parasites and symbiosis organisms living together with the individual of interest may act as natural tags.

Genetic marking DNA sampling is a relatively new method of identifying animals and humans, but demands some effort and equipment.

Radio telemetry Radio telemetry devices may be used to monitor a population continuously, as it is done to some individuals of the Norwegian bear population. The advantage is that "sightings" are much likelier, since the distance from the tag to the observer may be much greater than in the passive tag case. The drawbacks are that tag loss is likely to appear due to battery capacity, and that the method only identify tags, without knowing that the tag is still attached to the animal, or if the animal is dead. Another issue is that the equipment is rather expensive.

Which marking strategy to use, depends on the importance of the study. An experiment with anonymous tags are usually easier and faster, but provides less information than identification marks. For small but important populations (such as bear populations), radio tracking may be suitable. Generally, the amount of information from the mark is reflected by the amount of information on the mark.

The extent of mark loss is important. If this issue is not considered, bias may occur in the analysis. Except in the case of mutilation, mark loss is likely to happen. Paint may disappear, tags fall off and batteries run low. By applying two marks on some of the animals in the study, mark loss may be estimated.

There are two main sampling methods. The first method is sampling by recapture. With this method, it is assumed that the individual is alive up to the point of capture. Usually, the individual is released back into the population, and is prone to subsequent recaptures. The most gentle method is resighting, where the mark is registered without touching the individual. Sometimes the individual is killed by the capture event, or some other issue prevents the individual to be released back into the population. In this case the individual is considered as censored. The other method is sampling by recovery. In this case, the individuals are only sampled after their death. This sampling method may only be used for populations where there is a certain probability that a dead individual is found. Thus, fish populations are seldom sampled by recovery. The advantage of this sampling method, is that you gain direct information on the time of death. For the recapture method, we only get indirect knowledge of this time. These two methods may be combined.

When individuals are recaptured or recovered, auxiliary information may be registered. In addition to the tag number, covariate information such as location,

size, weight and sex may be interesting to the analysis. It may be of interest to see if covariates affect the capture or survival probabilities. Note, however, that some covariate information such as weight and size may only be registered at capture.

Traditionally, the sampling is done discretely. The reason for this is that the available models assume instantaneous sampling occasions, and thus are discrete. The sampling is denoted continuous when there are no restrictions on the time of sampling.

The last issue we present here, is whether the marked animals originally belong to the population or not. The normal procedure is to mark a subset of the population and then release it back into the population, hoping that the marking has not altered the animal. Another method is to release individuals that come from another population for example animals that are bred in captivity. This may be sensible when the capture of live animals prove difficult. A problem with this procedure is that we need to be sure that a marked animals possess the same qualities as a original member of the population.

3 Closed population models

We begin this chapter by presenting the basic capture data, some general notation, and assumptions that has to be made in order to obtain useful models for the closed population case. We generally follow Otis et al. (1978).

Due to the fact that we are dealing with closed populations, the population size N is constant over time and is a central parameter. The capture occasions are numbered $1, \ldots, k$, and the individuals are numbered $1, \ldots N$. The basic capture data may then be given by the the matrix

$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nk} \end{pmatrix}$$

where

$$X_{ij} = \begin{cases} 1 & \text{if the } i\text{th individual is caught on the } j\text{th occasion} \\ 0 & \text{otherwise.} \end{cases}$$

Generally, X_{ij} may be vectors with auxiliary information on the individuals such as weight, location, sex etc. in addition to capture information. Some of these auxiliary data will be missing, since they are only registered on capture.

From this data matrix, we may extract some essential statistics:

$n_j =$ the number of individuals captured in the jth sample $j = 1, 2, \ldots, k$

$n_{\cdot} =$ the total number of captures during the study

$m_j =$ The number of marked individuals in the jth sample

$a_\omega =$ the number of individuals with a particular sampling history ω, where ω is any ordered subset of $\{1, 2, \ldots k\}$. For example, a_{367} means the number of individuals captured in the third, sixth and seventh sample,

$M_j =$ the number of marked individuals in the population at the time of the jth sample, $j = 1, \ldots, k$. Note that $M_0 = 0$ and that M_{k+1} is the total number of distinct individuals caught during the experiment,

$M_\cdot = \sum_{j=1}^{k} M_j$,

$m_\cdot = \sum_{j=1}^{k} m_j$.

For the majority of mark-recapture models, the capture probability is an essential parameter. We define

$p_{ij} =$ the capture probability of the ith individual in the population on the jth trapping occasion.

3.1 The Petersen model

The simplest and most basic mark-recapture model for closed population is the Petersen model (also known as the Lincoln-Petersen model), and it requires two sampling occasions. At the first occasion, n_1 individuals are caught, marked and released. Then, a second sample of size n_2 is captured some time later, of which m_2 is marked. If the individuals behave independently, identically and randomly, the conditional distribution of m_2 given n_1, n_2 and N is hyper-geometric with expression

$$f(m_2|n_1, n_2, N) = \frac{\binom{n_1}{m_2}\binom{N-n_1}{n_2-m_2}}{\binom{N}{n_2}}. \tag{1}$$

A natural estimator for the population size is $\hat{N} = n_1 \cdot n_2/m_2$, and is called the Peterson estimate. The Peterson estimate is not a very good estimator, mainly because it is biased, and the bias may be large for small samples (see Chapman (1951)). An alternative possibly unbiased estimator was suggested by Robson and Regier (1964).

For the equation (1) to be true, some assumptions have to be made:

a) The population must be closed

b) All individuals in the population must have the same probability of being caught in each sample

c) Individuals do not loose their marks, and all marks are reported

The closure assumption may be weakened. If additions occur, they are unmarked and the Petersen estimator is valid for the population size at the time of the second sample. If removals occur with equal probability for the marked and unmarked individuals, the Petersen estimate is valid for the population size at the time of the

first sample. But even with this weakening of the closure assumption, the assumptions are very strict, and the number of directly applicable situations are few. The assumption of equal catchability (assumption **b)**) is often violated due to individual differences. In general, the literature agrees that there are two major reasons for violation:

Heterogeneity Qualities of the individual such as sex, age, size and social status may effect the capture probability.

Trap response The event of capture may alter the individuals behaviour. Marked individuals may have a higher capture probability (trap happiness) or a lower capture rate (trap shyness).

Experiment design may soften the effect of unequal catchability, e.g. by altering the capture methods to prevent trap responses, but it is very hard to totally eliminate it.

The effect of heterogeneity may result in that individuals with high capture probability more likely is captured in both the first and second sample. This makes the Peterson estimator underestimate the true population size, since the proportion m_2/n_2 will be greater than without the heterogeneity effect. The effect of trap happiness will impose a similar result, while trap shyness will overestimate the size.

If the individuals loose their marks or tags, the observed recaptures will be smaller than expected, and $\hat{N}$ overestimates N. A possible method for testing whether individuals loose their tag or not, is to apply two or more tags to the individuals in the first sample. Then the probability of tag loss is estimated by how many individuals there are in the second sample with tag loss (only one tag).

When there are incomplete tag returns, the observed value of m_2 will be to small, and $\hat{N}$ will again overestimate N. This problem arises when tags are returned by non-professionals such as volunteers.

3.2 Several sampling occasions

As mentioned, the bias of the Peterson estimate for small samples may be considerable. The variance of the estimates is also depending on the sample size. A way of addressing these two problems is to extend the number of sample occasions. By doing this, we hope to gather more information about the population. However, this also introduces a time effect. There may be a difference in capture probability for each capture occasion. The differences may be due to environmental effects such as weather, or unequal sampling effort. The literature usually lists three sources of which the capture probabilities are effected: heterogeneity (h), trap response (b) and time dependence (t). These sources lead to eight different models, each model assuming the presence or absence of the effects listed above. The models are denoted by $M_0, M_t, M_b, M_h, M_{tb}, M_{th}, M_{bh}$ and M_{thb}. This notation was introduced by Pollock (1974). The assumption that these are the only available models for closed population situations is not very strict, and applies to most mark-recapture

$$\left(1 - \frac{M_{k+1}}{N}\right) = \prod_{i=1}^{k} \left(1 - \frac{n_i}{N}\right).$$

This expression can be solved for $k = 2, 3$ but for $k > 3$, numerical algorithms have to be used, see Otis et al. (1978). Note that the left side of this equation is the probability of not being captured during the experiment. The right side is the product of not being captured at each capture occasion.

We see that the Petersen and the Schnabel models only use information on the number of captures at each sample and the number of different individuals caught during the experiment. Any individual information is excluded from the analysis. Thus, for the Petersen and Schnabel models it is sufficient to mark the individuals with anonymous marks.

Simulation results performed by Otis et al. (1978) indicated that the population size $\hat{N}_t$ under the model M_t is non-robust to additional heterogeneity and trapping effects. The bias is also unpredictable, and further knowledge is required to know if the population size is overestimated or underestimated. The same applies to the Peterson model, as these two models obtain mainly the same features. This non-robustness implies that the model is unsuitable unless we are sure that the assumptions are met.

3.3 M_0: equal capture probability

This is the simplest of all mark-recapture models, and assumes that every member of the population has the same probability of capture at each capture occasion. This model involves only two parameters, the population size N and the overall capture probability p. The probability distribution for the set of capture histories $\{a_\omega\}$ is given by

$$f(\{a_\omega\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{s+1})!} p^{n_\cdot}(1-p)^{Ns-n_\cdot}.$$

The maximum likelihood (ML) estimator for N and p is not analytically obtainable, but Otis et al. (1978) presented a numerical solution to the problem.

The underlying assumptions of the M_0 model makes it unsuitable to almost any natural animal population. It is non-robust to behavioural response and heterogeneity, and little is gained by using it instead of M_t. It is usually included for pedagogic reasons, and to serve as a reference model for testing sources of variance.

3.4 M_b: unequal capture probability due to behavioural response to capture

A slightly more sophisticated model than M_0 is the M_b model (Pollock (1974),Otis et al. (1978)). This model does not take into account heterogeneity or time dependence, but allows for trap response. It divides the capture probability into p and c, where p is the capture probability for unmarked individuals, while c is the capture

probability for marked individuals. This means that the individual changes catchability after a capturing event, modelling a trap shyness or a trap happiness effect. The capture history distribution is given by

$$f(\{a_\omega\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{k+1})!} p^{M_{k+1}}(1-p)^{kN - M_{k+1} - M_\cdot} \cdot c^{m_\cdot}(1-c)^{M_\cdot - m_\cdot}$$

The ML estimation of N is again not analytically tractable, but can be obtained numerically.

An interesting feature with this model is that the estimation of c is independent of the estimation of N and p. This means that once an individual is caught, it does not bring any more information to the N and p estimators. This means that if N is the only interesting parameter, we may use results obtained by removal models. These are models where individuals are removed from the experiment after capture. Simulations suggest that the model performs well when the assumptions are met. The model is non-robust to heterogeneity effects.

3.5 M_h: unequal capture probability due to individual differences

Model M_h assumes that all individuals has its own unique capture probability $p_i, i = 1, \ldots, N$, constant through the survey. It would be impossible to estimate the $N + 1$ parameters, due to identification problems. The solution given by Otis et al. (1978) follows Burnham and Overton (1978) and uses the jackknife to estimate the population size. Assume that the p_i's are a random sample of size N from some probability distribution $F(\{p_i\})$ (an effort in modelling this distribution as a beta distribution was made by Burnham (1972), but proved unsuccessful). With this assumption, the capture frequencies $\{f_1, f_2, \ldots f_k\}$ are a sufficient statistic for the M_h model. The jackknife estimator is a linear combination of the capture probabilities. This estimator proves robust and is much used.

3.6 $M_{hb}, M_{ht}, M_{bt}, M_{tbh}$: the other models

The heterogeneity and trap response model (M_{hb}) is included in Otis et al. (1978) using an estimation procedure they call the generalised removal method. Pollock and Otto (1983) concludes that their jackknife estimator performs better than the estimator from the generalised removal method. The population size for the other models proves impossible to estimate without further assumptions. Lee and Chao (1994) presents an approach to the problem using sample coverage. They show that this may be used on all the eight models. For the M_{bt} model, Chao et al. (2000) assume that the recapture probability c_j is proportional to the initial capture probability p_j, that is $c_j/p_j =$ constant for all $j = 2, 3, \ldots, s$ and thus reducing the parameter space. The population size is then estimated using techniques based on maximum likelihood.

Pledger (2000) models the heterogeneity, time dependence and trap response by assuming that the capture probabilities are distributed according to a finite mixture

distribution. This model is given as follows: Assume that the individuals may be divided into G different groups. Within each of these groups the capture probability is assumed equal. Which group each animal belongs to is not known, but it is assumed that each individual belongs to group $g, g = 1, 2, \ldots, G$ with probability π_g. The capture probability is now denoted by p_{jbg}, the probability that an individual of group g with a trap response indicator b is captured at the jth capture occasion. Using the logit transformation, the capture probability is given by a linear expression

$$\log\left(\frac{p_{jbg}}{1-p_{jbg}}\right) = \mu + \tau_j + \eta_g + (\tau\beta)_{jb} + (\tau\eta)_{jg} + (\beta\eta)_{bg} + (\tau\beta\eta)_{jbg}$$

where μ is a constant unknown factor, τ_j is a fixed main time dependent effect depending on the sample j, β_b is a fixed main trap response effect (such that $b = b_{ij} = 1$ if individual i was caught before sample j, else $b = 0$), and η_g is an individual random main effect depending on which group g the individual belongs to. The later terms are two- and three-way interactions between the main effects.

The distribution for the capture histories is then given by

$$f(\{a_\omega\}) = \frac{N!}{\prod_\omega a_\omega!(N - M_{s+1})!} Q^{N-M_{s+1}} \sum_{g=1}^{G} \pi_g P_\omega^{a_\omega},$$

where the capture histories ω depends on the trap response and the group, and Q is again the probability of never being caught. This is the general likelihood for the restricted M_{thb} model. We see that by excluding appropriate terms from the capture probability formulation, all the eight models in Otis et al. (1978) can be treated. In addition, more sophisticated models such as M_{t+b+h_A}, where the two- and three-ways interactions are excluded, is easily defined. The model parameters are estimated by maximising the log likelihood numerically. Pledger (2000) uses the EM algorithm.

Another useful feature for these models is that additional covariate information is easily incorporated. For example, suppose there is heterogeneity and a sex effect may be suspected. Let $e_j = 1$ if the individual is a male and $e_j = 0$ otherwise. Then the capture probability may be modelled as

$$\log\left(\frac{p_{ja}}{1-p_{ja}}\right) = \mu + \alpha e_j + \eta_a + \gamma_a e_j,$$

where α is a pure sex-effect while γ_a is a group-sex-interaction. With fixed covariate information, the summation over the groups in the likelihood is omitted.

3.7 Models for recaptures in continuous time

So far, the models we have dealt with have assumed discrete sampling. This means that the sampling is done at predefined sampling occasions, and that no individuals are captured at any other time. There may, however, be sampling situations where the sampling is done continuously. The development of mark-recapture models for

continuous time is largely underdeveloped compared to the massive amount of research on discrete time mark-recapture models. Most of the development has been based on a martingale approach presented by Becker (1984), including Yip et al. (1996), Lin and Yip (1999), Yip and Wang (2002) and Wang and Yip (2003). The last paper presents a semi-parametric model for continuous mark-recapture studies. In the following, the main ideas and results of this paper is presented.

3.7.1 The approach of Wang and Yip (2003)

Let ν be the population size, and τ the length of the study. Let in addition $N_i(t)$ be the number of times individual i has been captured up to time t. The capture intensity is assumed to be on the multiplicative form $\alpha_p(t;\boldsymbol{\beta},\mathbf{Z}_i) = \alpha_{0p}(t)\exp\{\boldsymbol{\beta}^\intercal\mathbf{Z}_i\}$ where $\boldsymbol{\beta}$ is a regression coefficient vector, $\mathbf{Z}_i$ is the time-independent covariate vector of individual i, and $\alpha_{0p}(t)$ is the baseline capture hazard rate. In this paper, no censoring at capture is assumed. Furthermore, let δ_i be an indicator variable to whether individual i has been captured during the experiment or not.

Let $\theta = (\boldsymbol{\beta}, A_{0p}(t))$, where $A_{0p}(t)$ is the cumulative baseline capture rate such that

$$A_{0p}(t) = \int_0^p \alpha_{0p}(s)\mathrm{d}s.$$

Since we have no information about the covariates for the uncaptured individuals, the conditional likelihood on the captured individuals is used for inference. This conditional likelihood is given by

$$L(\theta) = \prod_{i=1}^{\nu}\left\{\frac{\prod_{0\le t\le\tau}\alpha_p(t;\boldsymbol{\beta},\mathbf{Z}_i)^{\mathrm{d}N_i(t)}\cdot\exp\{-\int_0^\tau\alpha_p(s;\boldsymbol{\beta},\mathbf{Z}_i)\mathrm{d}s\}}{\pi_i}\right\}^{\delta_i}$$

where $\pi_i = 1-\exp\{-\int_0^\tau\alpha_p(s;\boldsymbol{\beta},\mathbf{Z}_i)\mathrm{d}s\}$, the prop ability that a random individual of the population is captured during the experiment, and $\mathrm{d}N_i(t) = 1$ if individual i is captured at time t and zero otherwise . By this likelihood, the Breslow estimator for the cumulative capture rate is given by

$$\widehat{A}_{0p}(t) = \int_0^t \frac{\mathrm{d}N_\cdot(s)}{\sum_{i=1}^{\nu}\delta_i\exp\{\boldsymbol{\beta}^\intercal\mathbf{Z}_i\}/\pi_i}$$

is given. The partial likelihood score function is given by

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{\nu}\int_0^\tau\left\{Z_i - \frac{\sum_{j=1}^{\nu}\delta_j\mathbf{Z}_i\exp\{\boldsymbol{\beta}^\intercal\mathbf{Z}_i\}/\pi_i}{\sum_{j=1}^{\nu}\delta_j\exp\{\boldsymbol{\beta}^\intercal\mathbf{Z}_i\}/\pi_i}\right\}\mathrm{d}N_j(t).$$

Then Wang and Yip (2003) presents a simple iterative procedure to compute the maximum likelihood estimates for the baseline capture rate $A_{0p}(t)$ and the regression coefficients $\boldsymbol{\beta}$. However, they do not try to justify the procedure or show that it converges in all cases. Using standard likelihood theory, variance estimators are produced.

Finally, they present a Horviz-Thompson type estimator

$$\hat{\nu} = \sum_{i=1}^{\nu} \frac{\delta_i}{\hat{\pi}_i}$$

for the population size. Here,

$$\hat{\pi}_i = 1 - \exp\{-\exp\{\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Z}_i\}\widehat{A}_{0p}(\tau)\}.$$

3.8 The Bayesian approach

The introduction of Bayesian statistics to closed population mark-recapture models opened for the introduction of prior knowledge to the population size estimation. Castledine (1981) was one of the first to apply the Bayesian approach, and his results are presented in the following.

It is natural to use the M_t model as a reference, since this model has received a lot of attention in the frequentistic case. The likelihood of this model is given by equation (2). We have $k+1$ unknown parameters, the k capture probabilities $p_1, \ldots p_k$ and the population size N for which we need to define priors. Castledine (1981) proposed

$$p_i \sim \text{beta}(a, b) \quad \text{(independent)},$$

for $i = 1, \ldots, k$. The beta distribution is used because it is flexible and is defined on $[0, 1]$. This prior results in the posterior

$$\pi(N, p_1, \ldots, p_k|data) \propto \frac{N!}{(N - M_{k+1})!} \prod_{i=1}^{k} p_i^{n_i+a-1}(1-p_i)^{N-n_i+b-1}\pi(N). \qquad (3)$$

We see that the information given by a_ω is left out. The marginal posterior of N is now easily obtainable by integrating out the capture probabilities $p_1, \ldots p_k$,

$$\pi(N|data) \propto \frac{N!}{(N - M_{k+1})!} \left\{ \prod_{i=1}^{k} \frac{\Gamma(N - n_i + b)}{\Gamma(N + a + b)} \right\} \pi(N), \qquad (4)$$

since the full posterior is a beta distribution with regards to the capture probabilities. We note that depending on the choice of $\pi(N)$, the marginal posterior is far from standard. As with many Bayesian models, numerical algorithms must be applied to analyse the posterior. A Simpson rule quadrature approximation for $\pi(N|data)$ is suggested by Castledine (1981).

Similar posterior distributions for models M_0 and M_b are easily obtainable by imposing beta prior distributions on respectfully p and (p, c). The analysis of the Bayesian M_0 model was done by Castledine (1981). Other developments within the Bayesian approach to closed population mark-recapture data include Lee and Chen (1998) who regards the M_{tb} model and Basu and Ebrahimi (2001) in which a rather constrained M_{th} model is applied to software error estimation.

The posterior distribution obtains all the information available, both the prior information and the information given by the data. This information must be used

for inference on the population size. Since the parameter information is presented as a distribution, usual characteristics such as mean, variance, median, mode etc. may be presented. But what is a good estimator for the population size? To answer this question, we need to define what we mean by "good". This is done by defining a loss function. The loss function $L(N, \hat{N})$ describes how we penalise deviations of the estimate $\hat{N}$ from the true population size N. The loss function should reflect the decision makers relative fear of purposing a wrong estimate. For example, if an overestimate of the population size is critical, this should be weighted in the loss function. The Bayes estimator of the population size given the prior distribution $\pi(N)$ and the loss function L is defined as the estimator $\hat{N}$ minimising the posterior expected loss $E^{\pi}[L(N, \hat{N})|x]$.

The determination of consistent, subjective loss functions is often difficult, prompting the statistician to use generic loss functions such as quadratic loss ($L(N, \hat{N}) = (N - \hat{N})^2$), absolute error loss ($L(N, \hat{N}) = |N - \hat{N}|$) or "0-1" loss ($L(N, \hat{N}) = 1$ if $\hat{N} = N$, 0 else). For these loss functions, the Bayes estimate is the posterior mean, the posterior median and the posterior mode, respective.

The credibility interval is the Bayesian answer to the frequentistic confidence interval, and defines a range of plausible values for N given the posterior. A confidence region C_x at level α is such that $P(N \in C_x) = 1 - \alpha$.

4 Open populations

With open population models, we understand models for mark-recapture experiment data where births and deaths are expected during the experiment period. Up to 1965 there had been done very little to relax the assumption of closure on mark-recapture models. Some approaches using deterministic addition and removal were proposed, but it was not until Jolly (1965) and Seber (1965) independently of each other applied the Schnabel census to open population mark-recapture experiments, that a comprehensive solution appeared.

4.1 The Jolly-Seber model

The assumptions for the Jolly-Seber model are as follows:

a) All individuals in the population have the same probability of being caught in each sample

b) All individuals have the same probability of surviving every inter-sample period

c) Individuals do not loose their marks, and all marks are reported

We see that the assumptions are very similar to the assumptions for the Schnabel model, relaxing the closure assumption, and including an assumption on survival. Before we continue with the likelihood, we present some additional notation appropriate for the situation. Note that much of the notation from the closed models is

retained. Let

$N_i =$ the size of the population at the time of the ith sample

$\phi_i =$ the probability of surviving the interval between the ith and the $(i+1)$th sample,

$q_i =$ The probability of not being captured at sample i, such that $q_i = 1 - p_i$. This is used to simplify the notation.

$\chi_i =$ the probability that an individual is not caught after the ith sample given that it is alive at the time of the ith sample. Observe that χ_i is given recursively by $\chi_i = 1 - \phi_i(1 - q_{i+1}\chi_{i+1})$, with $\chi_k = 1$,

$n_{<i} =$ the number of different animals captured before the ith sample

$\omega =$ capture history. For example, $\omega = 367$ means capture at samples 3, 6 and 7.

Notice that $n_{<i}$ is not equal to M_i in the closed case. To be able to divide the capture histories properly, we need to define $\omega \supset \{i\}/\{1, 2, \ldots, i-1\}, i = 2, 3, \ldots, k$ as the capture histories which includes capture occasion i but not capture occasions $1, 2, \ldots, i-1$. For example if $k = 5$,

$$\omega \supset \{3\}/\{1, 2\} = \{3\}, \{3, 4\}, \{3, 5\}, \{3, 4, 5\}.$$

$\omega \supset \{1\}$ equals all capture histories which includes the first capture event.

The idea is that all individuals that has been a part of the population, may be categorised within a finite number of categories depending on their capture histories, according to a multinomial distribution. The likelihood can be presented as the product of s multi-nominal distributions, $L[\{u_\omega\}] = L_1 \cdot L_2 \cdots L_k$. Let us start with L_1: In this first likelihood we categorise all individuals that are alive at the first sample event. That is, we have to place N_1 individuals. This leads to the expression

$$L_1 = \frac{N_1!}{(N_1 - n_{<2})! \prod_{\omega \supset \{1\}} a_\omega!} q_1^{N_1 - n_{<2}} \prod_{\omega \supset \{1\}} p_\omega^{a_\omega},$$

where p_ω is the probability that an individual experience the capture history ω. This probability is a function of the sample probabilities and the survival probabilities. For example, $p_{12} = (p_1\phi_1 p_2\chi_2)$. $(N_1 - n_{<2})$ is the number of individuals not caught in the first sample.

The next likelihood categorises all individuals that were caught in the second sample except for those that are already included in the first likelihood. We have to place $N_2 - n_{<2}$ individuals:

$$L_2 = \frac{(N_2 - n_{<2})!}{(N_2 - n_{<3})! \prod_{\omega \supset \{2\}/\{1\}} a_\omega!} q_2^{N_2 - n_{<3}} \prod_{\omega \supset \{2\}/\{1\}} p_\omega^{a_\omega}$$

$(N_2 - n_{<3})$ is the number out of $(N_1 - n_{<2})$ that was not captured in this sample.

The general expression for any sample is

$$L_i = \frac{(N_i - n_{<i})!}{(N_i - n_{<i+1})! \prod_{\omega \supset \{i\}/\{1,2,\ldots,i-1\}} a_\omega!} q_i^{N_i - n_{<i+1}} \prod_{\omega \supset \{i\}/\{1,2,\ldots i-1\}} p_\omega^{a_\omega}$$

These expressions are enough to produce a maximum likelihood estimate of the population sizes N_i, the survival rates ϕ_i and the capture rates p_i. However, we shall follow the direction of Seber (1982) and present intuitive argument estimators for the wanted parameters. As earlier, let M_i be the total number of marked individuals in the population at time t_i. In the case of open populations this variable is stochastic. An estimate of this variable can be obtained by noting that M_i consist of the number m_i captured in sample i and the number $M_i - m_i$ not captured. The probability of subsequent captures for the individuals of the two groups are per assumption equal, so we may expect

$$\frac{z_i}{M_i - m_i} = \frac{r_i}{m_i}$$

where z_i is the number of subsequently recaptured individuals given that they were not caught in sample i, and r_i is the number of the m_i released individuals that were subsequently caught. This leads to the estimator

$$\hat{M}_i = \frac{m_i z_i}{r_i} + m_i, i = 2, 3, \ldots s - 1.$$

Following this estimate, the rest is straight forward. The population size is estimated by the proportion of marked individuals in the sample, i.e.

$$\hat{N}_i = \frac{\hat{M}_i n_i}{m_i}, i = 2, 3, \ldots s - 1,$$

similar to the Petersen estimate. The survival probability estimate is given naturally by

$$\hat{\phi}_i = \frac{\hat{m_{i+1}}}{\hat{M}_i}, i = 2, 3, \ldots s - 2,$$

and

$$\hat{\phi}_1 = \frac{\hat{M}_2}{n_1}.$$

The intuitive estimator of the capture probability p_i is given by

$$\hat{p}_i = \frac{n_i}{\hat{N}_i} = \frac{m_i}{\hat{M}_i}, \qquad i = 2, 3, \ldots s - 1,$$

It should be noted that of all these estimators, only $\hat{p}_i$ and $\hat{\phi}_i$ are maximum likelihood estimators. All these estimators are biased, and Seber (1982) purposes some similar approximately unbiased estimators instead. It should also be noted that because $M_1 = 0$, the parameters N_1 and p_1 cannot be estimated by these means. This is also the case for N_k, p_k and ϕ_{k-1} because z_k and r_k are both zero.

Approximations of the variances and covariances are obtained via the delta method (see Jolly (1965)).

Generally, the two parameters p_k and ϕ_{k-1} are confounded, and can only be estimated as a product, $p_k\phi_{k-1}$. Let us clarify this with an example. If there are no marked individuals in the last recapture event, this could be the result of low recapture rate as well as low survival rate. It is not possible to separate the two effects without making further assumptions, for example that $p_k = p_{k-1}$.

A slightly more general case is when not every individual caught in a sample is returned. This could be due to accidents during handling, or it could be deliberate, for example by commercial fishing. This case is easily implemented into the Jolly-Seber model, see Seber (1982).

We see that although the capture and survival probabilities are estimated, the population size is still the central parameter. Many applications of the Jolly-Seber model do not share this interest, but emphasises the estimation of the survival probability. The likelihood is then simplified not to include the N_i's and the $n_{<i}$'s. We return to this later.

4.2 Special cases and generalisations of the Jolly-Seber model

The Jolly-Seber model is known as a time-dependent model because both the capture probabilities and the survival probabilities are dependent on the sampling times only. After 1965, a lot of work was done, both in generalising and specialising the Jolly-Seber model. Specialising, because one wanted better estimates of the more general parameters; generalising because one wanted better models to fit the underlying biological structure.

Jolly (1965) discussed a case where the the effects of births and immigration were negligible. This may be the cases for spacial constrained populations outside of the breeding period. The population size can then be estimated intuitively by the ratio comparison

$$\frac{z_i}{N_i - n_i} = \frac{r_i}{n_i}, i = 1, 2, \ldots s-1.$$

The left side of the equation is the fraction of individuals that were not caught in the ith sample to the number of these that was later recaptured. The right side is the fraction of individuals that was caught in the ith sample to the number of these that was later recaptured. The straight forward estimator of the survival probability is then

$$\hat{\phi}_i = \frac{\hat{N}_{i+1}}{\hat{N}_i}.$$

The rare case of no death or emigration is also discussed by Jolly (1965), but we do not pursue pursue this case here because of the limited natural applications.

Three natural restrictions to the Jolly-Seber model is to assume that the survival and/or capture probabilities are constant. When these assumptions can be made, models with a reduced number of parameters can be formulated, and presumably better estimates can be made. This situation is discussed in Jolly (1982). The three models are denoted B (constant survival rate), C (constant capture rate) and D (both

constant survival and capture rate). The basic Jolly-Seber model is denoted A. The major disadvantage of these three restriction models is that maximum likelihood estimates are not analytically tractable, and numerical algorithms must be applied.

The most significant change from the closed population is of course the introduction of survival rate. With this parameter, heterogeneity is almost inevitable. In many natural populations, the survival rate is dependent on age. The life starts with very small survival rate, then increases, and when the individual is getting old, the survival rate drops again. Other factors that may contribute to the survival rate are sex, size and habitat.

The first major generalisation of the Jolly-Seber model was done by Pollock (1981), where the capture and survival probabilities were allowed to be age dependent. The maximum likelihood estimators of the population size is similar to the Jolly-Seber estimators, with the exception that the population at the sampling time is assumed to be divided into several age groups. The estimates are given by

$$\begin{aligned} \hat{N}_i^{(\nu)} &= \frac{n_i^{(\nu)} \hat{M}_i^{(\nu)}}{m_i^{(\nu)}}, \quad \nu = 1, 2, \ldots, l-1, \\ \hat{M}_i^{(\nu)} &= m_i^{(\nu)} + \frac{n_i^{(\nu)} z_i^{(\nu)}}{r_i^{(\nu)}}, \quad \nu = 1, 2, \ldots, l-1, \end{aligned}$$

where the superscript (ν) denotes age group. The estimates of the survival and capture probabilities follows straight forward from the Jolly-Seber estimates. Pollock (1981) also presents a method for testing whether the survival and capture rates are different between the age groups. The number of age groups should be kept as small as it is biologically reasonable for precision reasons.

Brownie et al. (1986) presented the age-dependent versions of model B and D with the inclusion of two age groups. Numerical algorithms are needed to carry out the analysis.

4.3 The Cormack-Jolly-Seber model

With the Jolly-Seber model, the main interest is the population size. Recently, the interest has shifted towards estimation of the survival probabilities. The reason for this is according to Lebreton et al. (1992):

> ... because survival estimates are substantially more robust to the partial failure of assumptions than are estimators of population size (e.g., heterogeneity, of individuals to capture or recapture is common and violates a fundamental assumption of capture- recapture theory, Carothers 1973).

When focus is on the survival rate, only the marked individuals in the survey are regarded. We split the population in two, the ones that at one point in the survey is marked, and the rest. We then assume that the marked part of the population possesses the same survival skills as the rest of the population. If there is heterogeneity in the capture rate, it is assumable that the difference within the two groups are less

t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
R_1	m_{12}	m_{13}	$\cdots$	$m_{1,k-1}$	m_{1k}
	R_2	m_{23}	$\cdots$	$m_{2,k-1}$	m_{2k}
		R_3	$\cdots$	$m_{3,k-1}$	m_{3k}
			$\ddots$	$\cdots$	$\vdots$
				R_{k-1}	$m_{k-1,k}$

Table 1: Table of mark-recapture data

than between the groups. For example, the population may consist of both catchable and non-catchable individuals. This could be the case for bird, where only breeding birds are catchable. Then an uncompensated estimator $\hat{N}$ would only estimate the size of the catchable part of the population. An estimator of the survival rate would not possess such bias, under the assumption that the survival rate is equal for the two sub-populations. A result of this point of view is that it is impossible to estimate the recruitment rate of the population, since we only regard a subset of the population with no births (no individual are born with a mark).

With the emphasis on survival, the data are usually presented as an array of releases and recaptures. Table 1 shows such an array, where $R_i,\ i = 1, \ldots, k-1$ is the number of released, marked individuals at time t_i, and $m_{i,j},\ i = 1, \ldots, k-1, j = i, \ldots k$ is the number of marked individuals that are released at capture occasion t_i and not recaptured until capture occasion t_j. Note that an individual may be included in several of these numbers. If an individual is captured and released on all capture occasions between t_1 and t_k, it is included in R_1, $m_{1,2}, R_2, m_{2,3}, R_3, m_{3,4} \ldots R_{k-1}$ and up to $m_{k-1,k}$.

The likelihood with respect to the data as presented in Table 1 is given by

$$L(\boldsymbol{\phi}, \mathbf{p}; \mathbf{R}, \mathbf{a}) \propto \Delta \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} \left(\phi_i p_j \prod_{l=i+1}^{j-1} \phi_l q_l \right)^{m_{ij}}, \tag{5}$$

where $\Delta = \prod_{i=1}^{k} \chi_i^{R_i - m_{i,\cdot}}$ is the term considering the marked individuals never recaptured after their release. Here, $\chi_i, i = 1, \ldots k$ is computed recursively as described for the Jolly-Seber model, and $m_{i\cdot} = \sum_{j=i+1}^{k} m_{ij}$. This model is known as the Cormack-Jolly-Seber model, due to Cormack (1964). The confounding of p_k and ϕ_{k-1} still exist, as for the Jolly-Seber model.

Constraints on the survival and capture rates p_i and ϕ_i may be applied. The easiest method is to assume equality between some or all of the capture or survival probabilities. Other options is to include covariate information such as environmental data and capture effort into the model using generalised linear models. Using the logit transformation, the capture probability p_i may be expressed by

$$\text{logit}(p_i) = \beta_0 + \boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}$$

where $\boldsymbol{\beta}$ is the generalised regression parameter vector, and $\mathbf{Z}$ is the covariate vector. The survival probabilities may be modelled accordingly. Other transformations may

be used, but the logit transform assures us that the probabilities are kept between 0 and 1.

The inclusion of constraints and covariate information introduces the need for model selection. Which constraints are the most appropriate according to the data, and which covariates do we include? We address these questions in Section 5.

In order to estimate the parameters in the CJS model (with or without constraints), we need to apply numerical methods. Standard optimising algorithms is used to compute the maximum likelihood estimates.

4.4 Mark-recovery models

A much used study design for the understanding of bird populations is the so-called ring-recovery method. This design differs from the usual mark-recapture situation in that only tags recovered from dead individuals are recorded. The models associated with these studies are also known as tag-recovery, ring-recovery or band-recovery models. We shall use the name mark-recovery models in accordance with the mark-recapture name. For a thorough presentation of mark-recovery models, see Brownie et al. (1985).

Let $\lambda_j, j = 1, 2, \ldots k$ be the probability that an animal is recovered at time j given that it died in the interval between the $j-1$th and jth sample. The likelihood of a time-dependent model is then given by

$$L(\phi, \lambda, \mathbf{R}, \mathbf{a}) \propto \Delta \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} \theta_{ij}^{m_{ij}},$$

where

$$\theta_{ij} = \lambda_j (1 - \phi_j) \prod_{k=i}^{j-1} \phi_k,$$

the probability that an animal marked at time i is recovered at time j. The Δ term denotes the likelihood associated with not recovered animals, $\Delta = \prod_{i=1}^{k-1} \chi_i^{R_i - m_{i\cdot}}$, where $\chi_i = (1 - \sum_{j=i+1}^{k} \theta_{ij})$. The maximum likelihood estimates of the parameters λ_i and ϕ_i is computed using numerical methods.

The most thorough reference to mark-recovery models are given by Brownie et al. (1985). They allow for age-dependence which is important. If the study include juvenile animals, the survival rate of the juveniles tend to differ from the adult animals. An analysis of the situation when we have both live recaptures and dead recoveries is done by Catchpole et al. (1998), where the Cormack-Jolly-Seber model is integrated with the mark-recovery models of Brownie et al. (1985). A Bayesian analysis of the mark-recapture experiment is done by Brooks et al. (2000).

4.5 Spatial movement

The migration patterns of certain populations may be of interest to many biologists. For example, the movement of a migrating bird population may be an important factor in understanding the population. The mark-recapture method has been

t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
$\mathbf{R}_1$	$\mathbf{m}_{12}$	$\mathbf{m}_{13}$	$\cdots$	$\mathbf{m}_{1(k-1)}$	$\mathbf{m}_{1k}$
	$\mathbf{R}_2$	$\mathbf{m}_{23}$	$\cdots$	$\mathbf{m}_{2(k-1)}$	$\mathbf{m}_{2k}$
		$\mathbf{R}_3$	$\cdots$	$\mathbf{m}_{3(k-1)}$	$\mathbf{m}_{3k}$
			$\ddots$	$\cdots$	$\vdots$
				$\mathbf{R}_{k-1}$	$\mathbf{m}_{(k-1)k}$

Table 2: Table of mark-recapture data

much used in pursuing these questions, see Chapman and Junge (1958) and Darroch (1961). Lately, radio telemetry devices has made the tracking of animals easier, and the movement may be watched continuously. However, radio tracking is expensive and is usually only applied to a few animals. Thus it lacks the large scale properties of the standard mark-recapture experiment.

Movement and migration patterns may be analysed using so-called multi-state mark-recapture models. The general assumption of multi-state mark-recapture models is that the individuals move between a finite number of states, such as geographical areas. To be able to make inference on movement between states, the state at each recapture must be recorded. The data of multi-state mark-recapture experiments are usually presented in a similar way to the data array of the single-state mark-recapture experiment (see Table 1). A multi-state mark-recapture experiment data array is presented in Table 2. Here,

$$\mathbf{R}_i = \begin{bmatrix} R_i^1 \\ R_i^2 \\ \vdots \\ R_i^K \end{bmatrix}$$

where R_i^j is the number of individuals released in state j at time i, and

$$\mathbf{m}_{ij} = \begin{bmatrix} m_{ij}^{11} & m_{ij}^{12} & \cdots & m_{ij}^{1K} \\ m_{ij}^{21} & m_{ij}^{22} & \cdots & m_{ij}^{1K} \\ \vdots & \vdots & \ddots & \vdots \\ m_{ij}^{K1} & m_{ij}^{K2} & \cdots & m_{ij}^{KK} \end{bmatrix}$$

where m_{ij}^{rs} is the number of animals captured in state s at time j that were last captured in state r at time i. We use K as the total number of states.

The parameters of the Arnason-Schwarz multi-state mark-recapture model (first introduced in Schwarz et al. (1993)) is given by

$\phi_i^{rs} =$ the probability of being alive and in stratum s at time $i+1$, for an individual alive in stratum r at time i,

$p_i^s =$ the probability of capture at time i given that the individual is in stratum s at time i.

The usual assumption of independence between individuals still applies. It is also assumed that ϕ_i^{rs} and p_i^s is independent of the stratum occupied at time $i-1$ or earlier.

The likelihood function is again formed from a multinomial distribution as with the Cormack-Jolly-Seber model. The closed-form expression is rather complex, and we shall not present it here. It is , however, given in King and Brooks (2003).

The maximum likelihood estimates are only obtainable by numerical algorithms, and a discussion of these are given in Brownie et al. (1993). As commented in Schwarz et al. (1993), the number of parameters is substantial and parameter estimation may be imprecise unless there is gathered a lot of data. Due to this, there is an interest in reducing the number of parameter to increase the precision. This is discussed in Brownie et al. (1993).

In multi-strata situations, there is generally a large number of parameters to be estimated. With a relative small amount of data to estimate these data, these situations are suitable for Bayesian analysis. The initial contribution was done by Dupuis (1995), and an extension of this work to include model selection was presented by King and Brooks (2002b). Both these papers use a missing data approach, such that the later paper is an extension of the method used in the first paper. The extension mainly consists of introducing inference on the different models that results from constraining the Arnason-Schwarz model.

4.6 The Bayesian approach to open populations

The applications of Bayesian statistics to open populations are mainly concerned with the estimation of the survival rate. We present here the paper of Brooks et al. (2000).

The likelihood is given by equation (5) and the experiment design and model assumptions are as described for that situation. We recall it as

$$L(\boldsymbol{\phi}, \mathbf{p}; \mathbf{R}, \mathbf{a}) \propto \Delta \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} \left(\phi_i p_j \prod_{l=i+1}^{j-1} \phi_l \tilde{p}_l \right)^{m_{ij}}.$$

We now present the prior distributions and the resulting posterior distribution. As usual when priors for probability-parameters are considered, the beta distribution is used. Let thus

$$\begin{aligned} \phi_l &\sim \text{beta}(\alpha_\phi, \beta_\phi), \qquad l = 1, \ldots, k-1, \\ p_l &\sim \text{beta}(\alpha_p, \beta_p), \qquad l = 2, \ldots, k. \end{aligned}$$

The resulting posterior joint distribution of $\mathbf{p}$ and $\boldsymbol{\phi}$ is highly nonstandard. A much used representation of the posterior distribution is the full conditional posterior. The full conditional posterior for ϕ_l is the posterior distribution of ϕ_l conditional on all the other parameters. That is

$$\pi(\phi_l | \boldsymbol{\phi}_{(l)}, \mathbf{p}, \mathbf{R}, \mathbf{a}) \propto \phi_l^{\alpha_\phi - 1} (1 - \phi_l)^{\beta_\phi - 1} \Delta \phi_l^r,$$

where r is the number of recaptures where ϕ_i is involved, i.e. $r = \sum_{i=1}^{l} \sum_{j=l+1}^{k} m_{ij}$. The subscript of $\boldsymbol{\phi}_{(l)}$ means all the survival parameters except for ϕ_l. We see that the full posterior distribution is a product between Δ and a beta distribution. It is proved (by the Hammersley-Clifford theorem, see p. 298 Robert and Casella (1999)) that the full conditional distribution contains the same amount of information as the joint distribution. The full conditional for the capture probability is given similarly. The reason for presenting the posterior in this way, is that the Gibbs algorithm may be applied directly.

For biologists, the most interesting parameter is usually the survival rate. This means that the capture probability may be regarded as an auxiliary variable. Bayesian inference on the survival rate is made with loss functions, Bayesian estimators and credibility intervals, as described for the Bayesian analysis on closed populations.

5 Model selection

With the steady growth of available models, a natural question arises: Which model do we use? This problem applies to many fields within statistics, and is known as the problem of model selection. Model selection normally means deciding how many, and which parameters to be included in the model. It is based on the general knowledge that with increasing number of parameters, the bias of the estimators decreases, but the uncertainty increases. The right trade-off between these two effects is the mission of model selection.

One of the most general statistical tools for model selection is the Likelihood Ratio Test (LRT). This test requires two models, where one is a constrained version of the other. The full Jolly-Seber model tested against an constrained model of constant survival probability may be an example. Let Θ_1 be the parameter space of the unconstrained model, and Θ_0 be the parameter space of the constrained model. The test is based on the likelihood functions on the two parameter-spaces. If $\hat{\theta}_1$ and $\hat{\theta}_0$ is the maximum likelihood estimates for the two models, then the so-called log-likelihood ratio is given by

$$\text{LRT} = -2\text{log}L(\hat{\theta}_0|\mathbf{x}) - (-2\text{log}L(\hat{\theta}_1|\mathbf{x})$$

where $\mathbf{x}$ is the observed data. The classical result in likelihood theory is that the log-likelihood ratio is asymptotically chi-square distributed . The hypothesis $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ can be tested with this distribution. When the LRT is large enough, the hypothesis is rejected. In the usual hypothesis test situation, the emphasis is on control over the Type I error (reject a true H_0), at the risk of making a Type II error (failing to reject a false H_0). This is usually because a Type I error is more severe than a Type II error. This is not necessary the case in model selection. Our goal is not to prove anything, just to select between two possible models. So the term "large enough" is not necessary a straight forward choice. The power of the test is just as important as the significance level.

A problem with the LRT is that it only tests two models at a time. With a large number of available models, as is often the case with mark-recapture models, the task of testing all interesting models against each other becomes infeasible.

Another much used tool within model selection, is the Akaike's Information Criterion (AIC), usually defined as

$$AIC = -2\log L(\hat{\theta}|\mathbf{x}) + 2d$$

where d is the dimension of the parameter space, and $L(\hat{\theta}|\mathbf{x})$ is the likelihood evaluated at the maximum (see Akaike (1973)). A low AIC value indicates a good fit of the model to the data compared to a model with a high AIC value. The idea is that the AIC value of several models are computed, and that the model with the lowest AIC value is preferred. From the expression we see that this criterion favours models with high likelihood and few parameters. The main drawback is again that with increasing numbers of models, the time to compute the likelihoods may be considerable.

Both the LRT and the AIC were used by Lebreton et al. (1992) in order to test and decide which is the most appropriate mark-recapture model for several data sets. For example, analysing the survival and capture rates of European Dippers concluded with that there were no time dependence except for an effect due to floods.

Within the Bayesian paradigm, the different models may be seen as part of the parameter space, and appropriate priors may be associated with each model. Let $M_1, \dots M_k$ be a set of plausible models with prior probabilities $p_1, \dots p_k$. With each model M_i, a likelihood $L_i(\theta|\mathbf{x})$ and a prior distribution $\pi_i(\theta)$ may given. The posterior probabilities associated with model M_i is then given by

$$\pi(M_i|\mathbf{x}) \propto f(\mathbf{x}|M_i) \cdot p_i, \quad i = 1, \dots k,$$

where

$$f(\mathbf{x}|M_i) = \int L_i(\theta|\mathbf{x})\pi_i(\theta)d\theta.$$

This model posterior distribution may then be used for model selection. For applications of this model selection method, see King and Brooks (2002b), King and Brooks (2003) or King and Brooks (2002a). All these papers computes the posterior model distribution and interpret the result biologically, increasing the understanding of the population.

A Bayesian equivalence to the AIC is the so-called deviance information criterion, DIC, developed by Spigelhalter et al. (2002). The DIC is defined as

$$\text{DIC} = E_{\theta|\mathbf{x}}[\underbrace{-2\log\{L(\theta|\mathbf{x})\} + 2\log\{g(\mathbf{x})\}}_{D(\theta)}] + p_D \tag{6}$$

where $g(\mathbf{x})$ is some fully specified standardising term that is a function of the data alone, and p_D is an estimator of the effective number of parameters in the model. The term of which the posterior expectation is taken is known as the Bayesian deviance, and is denoted $D(\theta)$. Note that the p_D is different from the given number of parameters used in the frequentistic AIC, since parameters may be dependent in the Bayesian case. By Spigelhalter et al. (2002) this measure is given by

$$p_D = \overline{D(\theta)} - D(\tilde{\theta}). \tag{7}$$

where $\overline{D(\theta)} = E_{\theta|\mathbf{x}}[D(\theta)]$ and $\tilde{\theta}$ is some estimator of θ depending on $\mathbf{x}$, for example the posterior mean. Note that an alternative representation of the DIC is

$$\text{DIC} = D(\tilde{\theta}) + 2\, p_D$$

which reveals the close connection to the AIC.

The properties of the p_D estimator is examined by Spigelhalter et al. (2002), and the perception of p_D as an estimator of the effective number of parameters is confirmed by some examples. They also show that the estimator is not invariant to apparently straight forward transformations of the parameters, which is clearly unwanted. Another unwanted property is that negative values of p_D may occur for non-log-concave likelihoods. Spigelhalter et al. (2002) argues that a negative p_D indicates substantial conflict between prior and data. A last issue is that the estimation of the complexity of the model is based on the assumption that the model is correct. Thus, the estimation depends on how accurate the model describes the data, and a poor models may produce poor complexity estimates.

A central term in this method is the Bayesian deviance. This is simply a transformation of the likelihood, and contains all information about the parameters given by the data. The interesting part is what the posterior expectation of this term translates to. When the model is correct, the expectation of the posterior expected Bayesian deviance with regard to data distribution is given by

$$\begin{aligned} E_{\mathbf{X}}[\overline{D(\theta)}] &= E_{\mathbf{X}}[E_{\theta|\mathbf{X}}[D(\theta)]] \\ &= E_{\theta}[E_{\mathbf{X}|\theta}[-2\,\log\{L(\theta|\mathbf{X})\} + 2\,\log\{g(\mathbf{X}\}]] \end{aligned}$$

where the transition is made by changing the conditioning between $\mathbf{X}$ and θ. If we now let $g(\mathbf{X}) = L(\hat{\theta}(\mathbf{X})|\mathbf{X})$, where $\hat{\theta}(\mathbf{X})$ is the maximum likelihood estimate, then

$$E_{\mathbf{X}|\theta}\left[-2\,\log\left\{\frac{L(\theta|\mathbf{X})}{L(\hat{\theta}(\mathbf{X})|\mathbf{X})}\right\}\right]$$

is the expectation of the likelihood ratio between the true null model and the model with the fitted parameter $\hat{\theta}(\mathbf{X})$. Under standard conditions, this is the expectation of an approximated chi-squared distribution with p degrees of freedom. p is in this case the dimensionality of θ. The posterior expected Bayesian deviance, standardised by the maximised log-likelihood, is thus expected to be the number of free parameters in θ.

Let us now turn our attention to equation (7) with $\tilde{\theta} = \overline{\theta}$, the posterior mean of the parameters, such that

$$D(\overline{\theta}) = -2\text{log}L(\overline{\theta}|\mathbf{x}) + 2\text{log}L(\hat{\theta}|\mathbf{x}).$$

In the asymptotic case, when the likelihood dominates the prior, the posterior mean approaches the maximum likelihood estimate, and $D(\overline{\theta})$ approaches zero. Thus p_D approaches $\overline{D(\theta)}$ whose mean is the dimensionality of θ. Since the dependence between the parameters is introduced through the prior, and the prior is suppressed by the likelihood, the effective number of parameters should approach $\dim(\theta)$.

Another interesting issue is the close resemblance between $\overline{D(\theta)}$ standardised by the maximised log likelihood and p_D. We see that in this case,

$$\overline{D(\theta)} = p_D + D(\overline{\theta}) - D(\hat{\theta}).$$

The difference is determined by how much the posterior mean of the parameters differs from the maximum likelihood estimate.

So, what are the advantages and disadvantages of using the DIC method for model selection instead of the posterior distribution method? The main advantage is that the DIC includes a nice and controlled handling of model complexity. The influence of complexity on the posterior model is less understood. The drawback of the DIC is clearly the invariance to transformations. This is an unwanted property, and further research should be performed to deal with this problem. In Spigelhalter et al. (2002), the invariance properties of the posterior mode is explored, and this proved somewhat successful. Also, the property that poor models produce poor complexity estimates, as mentioned above, is not wanted.

An optimal strategy would be to first compute the posterior model distribution, and then estimate the complexity of the models with highest posterior probability. Using this approach, the best of both methods is utilised.

Both the posterior model method and the DIC method relies on the assumption that there is a feasible number of interesting models, and that the chosen model space does not exclude potentially good models. This we have no guarantee for. Another approach to model selection is to start off with one model, and then perform model criticism as described in O'Hagan (2001) to see which parameters to improve and which assumptions that are valid.

6 Mark-recapture programs

Since inference is mostly done by likelihood maximisation, and the likelihood of mark-recapture and mark-recovery models usually are quite complex, it is of no surprise that most of the computations is be done numerically. There exists several programs for this purpose. The most flexible and most used computer program for mark-recapture models is without doubt the MARK program. The idea of MARK is to unify the different algorithms developed in mark-recapture research over the years, and to expand this collection where it is weak or non-existing. The result is a program with a nice, intuitive interface, flexible enough to analyse a large variety of mark-recapture experiments. Of the methods and strategies presented in this paper, the only ones not included in MARK are continuous models and the Bayesian approach.

For Bayesian mark-recapture models, there exists no overall program. Often, an ad hoc implemented Gibbs sampler (Dupuis (1995), Lee and Chen (1998), Barry et al. (2003) is used. There exists a powerful and flexible computer package called BUGS which implements the Gibbs sampler. This package was used successfully to analyse mark-recapture data by Brooks et al. (2000). The implementation of the posterior model distribution (5) is a bit to demanding for the Gibbs sampler, and

the reversible jump Markov chain Monte Carlo method (Green (1995)) has been used (see King and Brooks (2002b), King and Brooks (2003)).

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihhood principle. In Petran, B. N. and Csaki, F., editors, *International Symposium on Information Theory, second edition*. Akademiai Kiadi, Budapest, Hungary.

Barry, S. C., Brooks, S. P., Cathcpole, E., and Morgan, B. J. T. (2003). The analysis of ring-recovery data using random effects. *Biometrics*, 59(1):54–65.

Basu, S. and Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation for population size: Heterogenity and dependence. *Biometrika*, 88(1):269–79.

Becker, N. G. (1984). Estimating population size from capture-recapture experiments in continous time. *Australian journal of statistics*, 26:1–7.

Brooks, S. P., Catchpole, E. A., and Morgan, B. J. T. (2000). Bayesian animal survival estimation. *Statistical Science*, 15(4):357–76.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). *Statistical inference from band-recovery data: a handbook*. Fish and wildlife service, U.S. Department of the Interior, second edition.

Brownie, C., Hines, J. E., and Nichols, J. D. (1986). Constant-parameter capture-recapture models. *Biometrics*, 42:561–74.

Brownie, C., Hines, J. E., Nichols, J. D., Pollock, K. H., and Hestbeck, J. B. (1993). Capture-recapture studies for multiple strata includeing non-markovian transitions. *Biometrics*, 49:1173–87.

Burnham, K. P. (1972). *Estimation of population size in multiple capture-recapture studies when capture probabilities vary among animals*. PhD thesis, Oregon State Univeristy.

Burnham, K. P. and Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3):625–33.

Castledine, B. J. (1981). A bayesian analysis of multiple-recapure sampling for a closed population. *Biometrika*, 67:197–210.

Catchpole, E. A., Freeman, S. N., Morgan, B. J. T., and Harris, M. P. (1998). Integrated recovery/recapture data analysis. *Biometrics*, 54:33–46.

Chao, A., Chu, W., and Hsu, C. H. (2000). Capture-recapture when time and behavioral response affect capture probabilities. *Biometrics*, 56(2):427–33.

Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses . *Univ. Calif. Public. Stat.*, 1:131–60.

Chapman, D. G. and Junge, C. O. (1958). The estimation of the size of a stratified animal population. *Annals of mathematical statistics*, 27:375–89.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika*, 51:429–38.

Darroch, J. N. (1958). The multiple-recapture census. 1. Estimation of a closed population. *Biometrika*, 45:343–59.

Darroch, J. N. (1961). The two-sample caputer-recapture census when tagging and sampling are stratified. *Biometrica*, 48:241–60.

Dupuis, J. A. (1995). Bayesian estimation of movement and survival probabilities from capture-recapture data. *Biometrika*, 82:761–72.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–32.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52:225–47.

Jolly, G. M. (1982). Mark-recapture models with parameters constant in time. *Biometrics*, 38(2):301–23.

King, R. and Brooks, S. (2002a). Model selection for integrated recovery/recapture data. *Biometrics*, 58(4):841–51.

King, R. and Brooks, S. (2003). Survival and spatial fidelity of mouflon: the effect of lovation,age and sex. *Journal of agricultural, biological and environmental statistics*, 8(4):486–513.

King, R. and Brooks, S. P. (2002b). Bayesian model discrimination for multiple strata capture-recapture data. *Biometrika*.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Lee, S. M. and Chao, A. (1994). Estimating population size via sample coverage for closed capture-recapture models. *Biometrics*, 50(1):88–97.

Lee, S. M. and Chen, C. W. S. (1998). Bayesian inference of population size for behabioral response models. *Statistica sinica*, 8:1233–47.

Lin, D. Y. and Yip, P. S. F. (1999). Parametric regression models for continous time removal and recapture studies. *Journal of the Royal Statistical Society, Series B* , 61:401–411.

O'Hagan, A. (2001). Hss model criticism. Technical report, Department of probability and statistics, University of Sheffild. to appear in a book.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, 62.

Pledger, S. (2000). Unified maximum likelihood estimates for closed capture-recapture models using mixtures. *Biometrics*, 56(2):434–442.

Pollock, K. H. (1974). *The assumption of equal catchability of animals in tag-recapture experiments*. Ph.d. dissertation, Cornell University, Ithacha, New York.

Pollock, K. H. (1981). Capture-Recapture Models Allowing for Age-Dependent Survival and Capture Rates. *Biometrics*, 37:521–29.

Pollock, K. H. and Otto, M. C. (1983). Robust estimation of population size in closed animal populations from capture-recapture experiments. *Biometrics*, 39:1035–49.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer-Verlag New York.

Robson, D. S. and Regier, H. A. (1964). Sample size in Petersen mark-recapture experiments. *Trans. Amer. Fish. Soc.*

Schnabel, Z. E. (1938). The estimation of the total fish population of a lake. *Amer. Math. Mon.*, 45:348–52.

Schwarz, C. J., Schweigert, J. F., and Arnason, A. N. (1993). Estimating migration rates using tag-recovery data. *Biometrics*, 49:177–93.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, pages 249–59.

Seber, G. A. F. (1982). *The estimation of animal abundance*. Charles Griffin, London.

Spigelhalter, D. J., Best, N. G., Carlin, B. C., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *J. R. Statist. Soc. B.* to appear.

Wang, Y. and Yip, P. S. F. (2003). A semiparametric model for recapture experiments. *Scandinavian journal of statistics*, 30(4):667–76.

Yip, P. S. F., Higgins, R. M., and Lin, D. Y. (1996). Inference for capture-recapture experiments in continous time with variable captues rates. *Biometrika*, 83:477–83.

Yip, P. S. F. and Wang, Y. (2002). A unified parametric regression model for recapture studies with random removals in continous time. *Biometrics*, 58(1):192–99.

Analysing the Risør Coastal Cod Mark-Recapture Experiment

Inge Christoffer Olsen

Abstract

From 1986 to 1989, almost 40 000 cod were marked and released into the fjords of the Risør area in the southern part of the Norwegian coast. The ultimate goal of this experiment was to determine whether coastal cod was susceptible to sea farming. The result was negative, but the data that followed the experiment is valuable as a source to better understand the nature of the Norwegian coastal cod population.

This paper considers the mortality of the Norwegian coastal cod population, particularly how it is affected by age, time and other covariate information. A Cox proportional hazard model is used for the capture and mortality processes, and the Monte Carlo EM algorithm is used to analyse it.

1 Introduction

> Global production from capture fisheries and aquaculture and the food fish supply is currently the highest on record and remains very significant for global food security, providing more than 15 percent of total animal protein supplies(...) (SOFIA (2002)).

The importance of fish as a source of food and employment is firmly established by the Food and Agriculture Organization of the United States (FAO) in their biannual publication "The State of World Fisheries and Aquaculture" (SOFIA). The latest issues of 2002 states that 35 million workers were employed in capture fisheries and aquaculture production sectors. This represents 2.6% of the 1.3 billion people active in agriculture worldwide at that time. If we exclude China (due to poorly based statistics), the global per capita fish supply was 13.1 kg, with an annual global capture of about 77 to 70 million tonnes. The revenues of international trade in fish products was reported to be US$55.2, with a substantial part contributed by developing countries (US$18 billion).

1.1 The importance of research for management of fisheries

We understand that fishing is an important human enterprise. However, the fact that the normal fish stock is wild makes the management of fisheries a challenging task. The global view has evolved from the naive, but at the time (1884) probably

correct statement of Thomas Huxley that "probably all great sea-fisheries are inexhaustible" (Smith (1994)), to todays understanding that human influence is a mayor factor on certain marine ecosystems. There are numerous examples where fishing has been either the direct or indirect cause of degeneration and alteration of marine ecosystems. The collapse of several abundant fish stocks such as the sardine stocks of California and Japan in the late 1940s, the anchovy off Peru and Chile in 1972 and the Newfoundland cod stock in 1992 are all assumed caused or partly caused by overfishing (Botsford et al. (1997)). The effects of such collapses are overwhelming. Not only for the marine ecosystems, who are severely altered, but also for the communities dependent on the fisheries. In Newfoundland, 20,000 people were put out of work and the Canadian government payed over Can$1 billion per year to support unemployed fishermen.

It is estimated that almost half of the individual fish stocks of the world are fully exploited, and another 22% are overexploited (Botsford et al. (1997)). The term "overexploited" or over-fished means that there are "serious risks of reduced recruitment and possible stock collapse" (Hilborn et al. (2003)). That is, although the stock seems stable, unforeseen events such as a toxic algae bloom or disease may cause a collapse.

We understand that the need for effective management of the fisheries is definitely present. But there are difficulties: The decision makers, usually politicians, are pressured by the short-term benefits to society (jobs and profits) to increase the harvest. When scientists are not able to predict the precise effect of the harvest increase, the managers give in for the pressure. This is known as the "ratchet effect" (Ludwig et al. (1993)). There are two ways to meet this challenge. The first challenge is to increase the biological pressure by reaching consensus in the biological questions such as natural mortality rate of the exploited stock (Hilborn et al. (2003)), and the broader impact of enhanced harvest on the marine system (Botsford et al. (1997)). The second one is to release the political pressure on the management by ex. presenting alternative sources of jobs and profits for the coastal fisheries and their communities. This is a motivating factor for research in stock enhancement and sea ranching.

1.2 The enhancement of cod stocks

The idea of enhancing fish stocks has existed since the 19th century. Shad fry was released in New England in 1867, and chum salmon in Japan in 1876 (see Liao and Leaño (2003), and the references therein). In Norway, cod enhancement programs were initiated in 1884 by the former sea captain Gunder Mathiesen Dannevig. He built a hatchery for cod eggs in Flødevigen near Arendal on the Norwegian Skagerrak coast. In these early stages, the hypothesis was that by releasing huge amounts of larvae into the fjords of the Skagerrak coast, the number of recruits to the cod population would increase. In the end, the fishermen would experience larger catches, or at least the natural fluctuations in catches between years would be damped. This hypothesis was disputed by the fishery biologist Johan Hjort, who believed that the larvae releases had little effect on the local cod population. To settle this dispute,

Dannevig initiated a series of beach seine hauls in 1903 that would not end until 1971. By this time, after billions of larvae released, the releases came to an end without any evidence of benefit (Tveite (1971)).

In 1983, scientists in Austevoll of Western Norway, managed to produce large numbers of viable juvenile cod (Øyestad et al. (1985)). This led to a renewed interest in stock enhancement, and new experiments were planned to see whether release of juvenile cod into the fjords would prove profitable in forms of increased harvest. A huge interdisciplinary research programme on sea ranching of cod was in 1985 initiated by the Norwegian Fisheries Research Council, called the "Cod in Fjords" programme. Marked, reared 10-30 cm large juvenile cods were released into seven main areas along the Norwegian coast. The aim was to understand the biology and ecology of the targeted population, and through this understanding enhance the stocks for harvesting and develop profitable sea ranches for cod. The conclusion of this programme was that releases of juvenile cod did not significantly increase cod production and catches (Svåsand et al. (2000)).

2 The Risør mark-recapture experiment

Although the "Cod in Fjord" program was labelled a failure with respect to profitable sea ranching, the huge amount of data collected is an important source of basic biological understanding of the Atlantic cod,*Gadus morhua L.* For a deeper analysis, we look at the mark-recapture experiment performed in the Risør area. The recapture rates from this experiment was throughout reasonably high, the area was rather well studied with regard to environment, and the population is considered stationary (Danielssen and Gjøsæter (1994)). In addition, the population dynamics of the Skagerrak coastal cod stock has recently been given considerably attention (see ex. Fromentin et al. (2001) and references therein).

The experiment area consists of the Søndeledsfjord and the Skerries, a collection of small islands, narrow inlets and shallow archipelago, separating the fjord from the Skagerrak and making it a naturally, semi-closed area (see Figure 1).

The experiment was performed during the years from 1986 to 1989. Almost 40 000 fishes were marked and released, and the last recapture was reported in 1994. There were two categories of released fish. The reared fish were raised from the coastal cod stock on the western coast of Norway, and transported to the Risør area in large tanks on boats or trucks. At the arrival, the fish were kept in the tanks for a few days to assure that they were not injured by the transportation. They were tagged with the FD-67 Floy tag attached under the first dorsal fin (see Figure 2) and remained in the tank for some days to see how they adapted to the tag. Only the fishes who seemed fit were released.

The other category of marked individuals consisted of wild-caught fish from commercial pot catches in the Risør area. They were mainly 18 months or older, and only seemingly healthy fishes were used. According to Julliard et al. (2001): "...the size distribution of wild fish caught in December appeared to be very close to the size distribution of one-year old reared fish recaptured at that time..." Marked fish were released in small groups of not more than a couple of hundred cod, to ensure

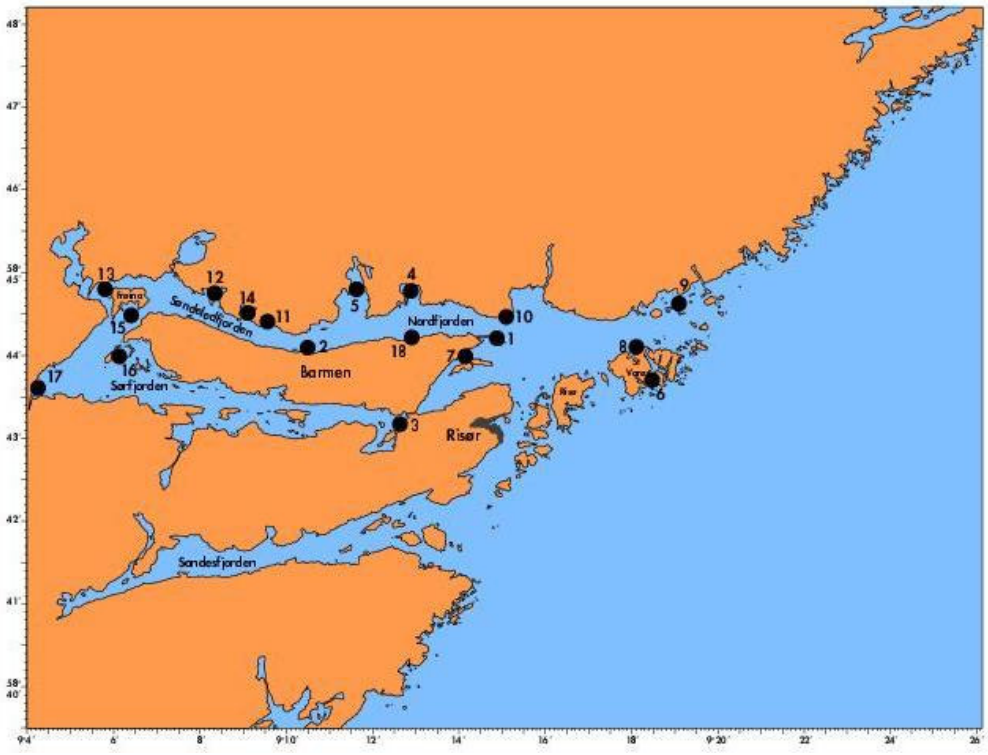


Figure 1: The Risør area with the release sites marked

a good distribution. The release sites are marked in Figure 1. An overview of the releases is given in Table 1.

The release experiment was announced in newspapers, radio and posters along the Norwegian Skagerrak coast and a reward of NOK 25 (approx. $7) for each recaptured marked cod was offered. With each fish captured, a number of of data was reported. The time of recapture, the place of recapture, the size of the fish, the depth at wich the fish was recaptured and the gear used to catch the fish were the most essential data gathered. There existed an arrangement with the eel fishermen of the area to release the recaptured fish, after recording the data.

The marked fishes were mainly fished with four different fishing equipments: gillnets, traps, angling and jig. The gill net is a mesh of line in which the fish entangles (see Figure 3a). There is a size selection on gill nets, large fish do not entangle when the mesh size is relative small, and small fish swim through the net when the mesh size is relative large. The fishes caught by traps were by-cathces of eel-fishermen using eel traps (see Figure 3b). For eel traps there is also a size selection. Small fish swim back through, while large fish are unable to enter the trap. Jigging is the setting of a line, with baited hooks or lures, that is continually jerked. The motion, achieved by hand or with a jigging machine, induces fish to take the hook. There is a size selection due to the bait used. The feeding habit of cod changes with age, and small fish takes other baits than larger fish. Jigging is done from a boat. Angling is usually done from the shore with a fishing pole. The size selection is similar to jigging. The denominator of all these fishing methods is

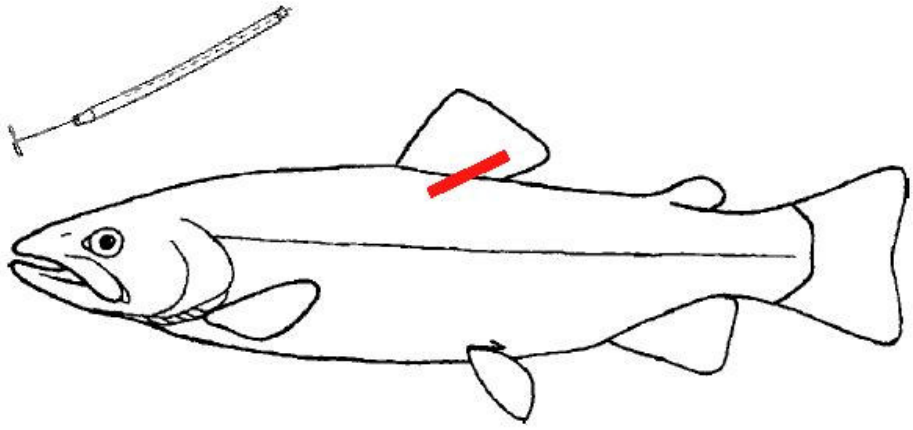


Figure 2: The Floy tag and the tag placement on the cod

Cohort	Date	Type	# released	Mean length	% recaptured
R86	Oct. 1986	Reared	5894	15.9 cm	3.5
R87	Oct. 1987	Reared	6701	15.2 cm	2.0
R88	Oct. 1988	Reared	11408	17.4 cm	15.9
R89	Oct. 1989	Reared	12725	17.3 cm	9.3
W86	Dec. 1986	Wild-caught	791	34.6 cm	28.4
W88	Dec. 1988	Wild-caught	1387	35.1 cm	26.8
W89	Dec. 1989	Wild-caught	237	38.0 cm	20.7

Table 1: Main figures of the seven cohorts (from Julliard et. al. (2001) with corrections)

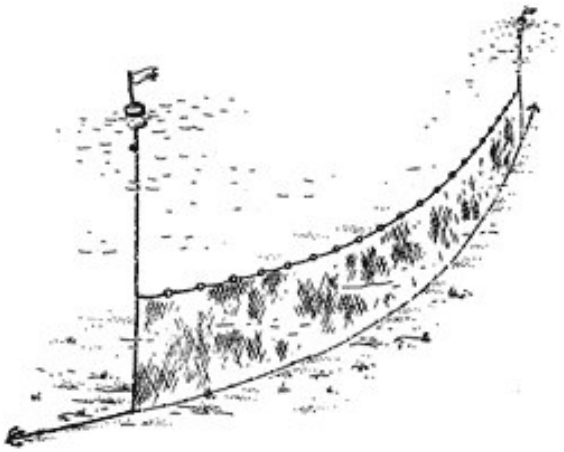
that fishing for small fish is usually unwanted, and thus avoided.

There were total 39,243 cod released in this experiment of which 4030 were recaptured. 973 of these were released after recapture. A total of 4623 recaptures were registered, the last one in December 1993.

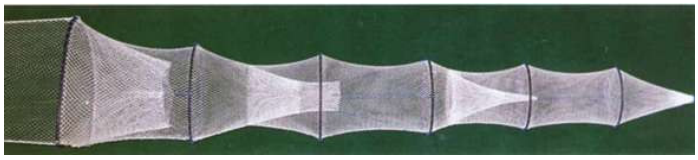
For convenience, and in accordance with the existing literature, we classify the age groups. From hatching (usually during March-April) until first of April the next year, the cod are denoted 0-group cod. The next year they constitute the 1-group cod, and so on.

2.1 The toxic algae bloom of 1988

In May-June of 1988, a bloom of the toxic algae *Chrysochromulina polylepis* killed a number of marine species in the upper 20 meter of the Skagerrak coast, see Figure 4. The bloom led to an estimated 60% reduction in the 0-group cod (approx. 3-months old at the time of the bloom) (Chan et al. (2003)), while the older cod assumingly survived by ascending to deeper waters. Due to the population dynamics, there seems to be a long-term effect of the bloom. The 0-group cod of 1989 does not seem to be affected negatively of the bloom, but for the cod hatched in 1990, there is an



(a) The Gill-net



(b) The Eel Trap

Figure 3: Two illustrations of the fishing equipment used during the captures.

estimated 40% reduction. This is explained by the population dynamics. Although the 1-group cod of 1988 survived, they experienced a lack of pray due to the algae bloom. Thus the 1-group was indirectly weakened. The 0-group of 1989 experienced low predator pressure from the older cod, and became strong. In 1990, this strong cohort preyed on the 0-group cod, explaining the reduction in stock (see Chan et al. (2003) for a thorough analysis).

Note that the reared 0-group cod of the experiment were not yet released during the bloom. Thus, the bloom only possibly affected the R86, R87 and W86 cohorts. Since all three of these cohorts consisted of 1-group or older individuals, we may assume that the algae bloom does not directly affect our analysis. There may be an indirect affect, as indicated by Chan et al. (2003).

2.2 Presentation of the data

To get a feeling of the data material, let us look at some basic figures. From the histogram of recaptures from October 1st 1986 to December 1st 1993 in Figure 2.2, we might notice a time effect in that the recapture probability in the two first years seems notably smaller than in the following three years. The number of captures was very small before and after the algae bloom. When we look at the recapture histogram with respect to age (Figure 2.2) we notice that there is a jump in the recaptures when the fish are approximately one year. After the fish turn two years, there is a notable decline in recapture. The seasonal recapture histogram (Figure 2.2) shows that the recapture rate is highest in May and June, and lowest in August and December. The fish size is an important factor with respect to recapture. This

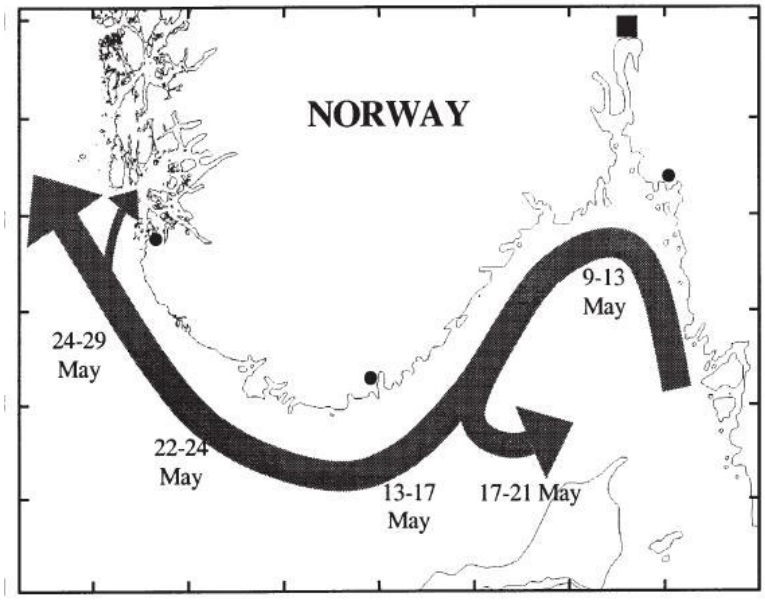


Figure 4: The development of the toxic algae bloom along the Skagerrak coast during May-June 1988

is clearly shown in the length at recapture histogram of Figure 8. Fish below 20 cm are very rarely captured.

The different fishing gear has an effect on different recapture histograms. In Figure 9 we see example of the size selection of different fishing gear. We notice that there is a significant difference in capture size between gear 10 (trap) and gear 20 (gill-net) ($t = 5.65$). There is also a significant difference in the age at recapture (Figure 10). These two results are connected owing to the natural correlation between age and size. The fishing season for the different gears is apparent in Figure 11. We see that the season for trap fishing is in summer and autumn, while the gill-net fishing is in winter-spring.

From earlier research results on the Norwegian coastal cod, we know that it is quite stationary. If we look at the histogram of the distance covered from one capture to the next (see Figure 12), we see that this is also the case for the cod participating in this experiment. When the distance covered against the time between captures is plotted in Figure 13, we see that the time of freedom does not affect the distance covered to much degree.

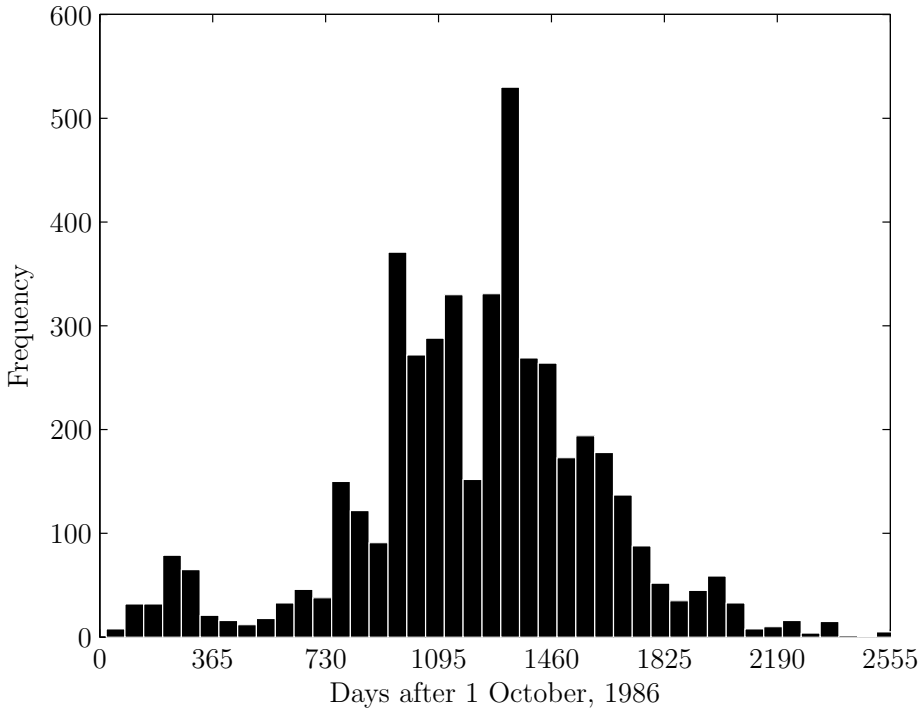


Figure 5: Histogram of the recapture time

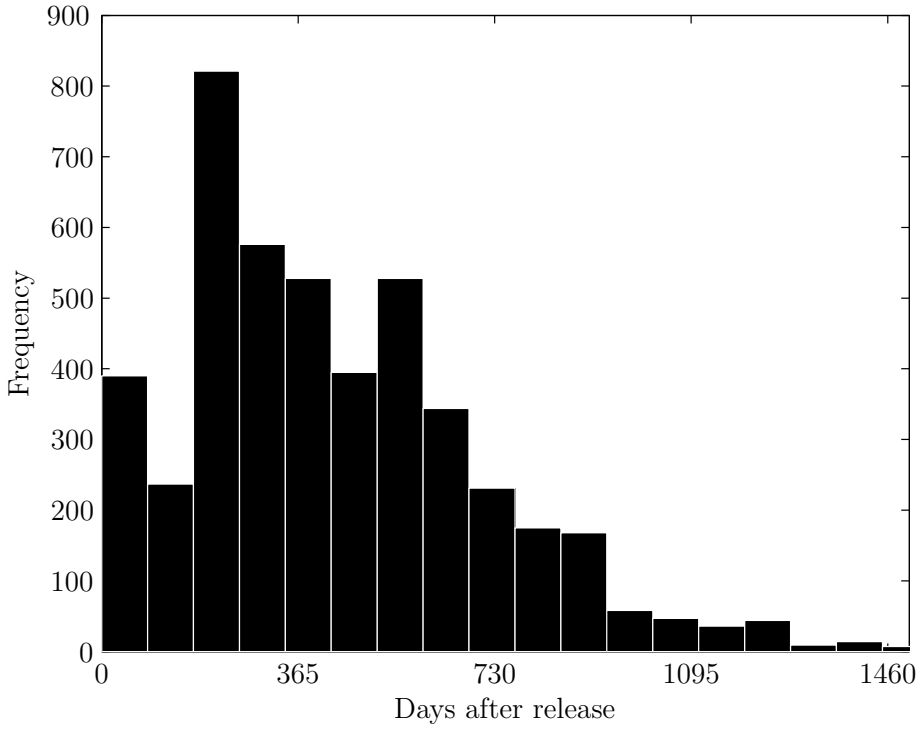


Figure 6: Histogram of the days after release

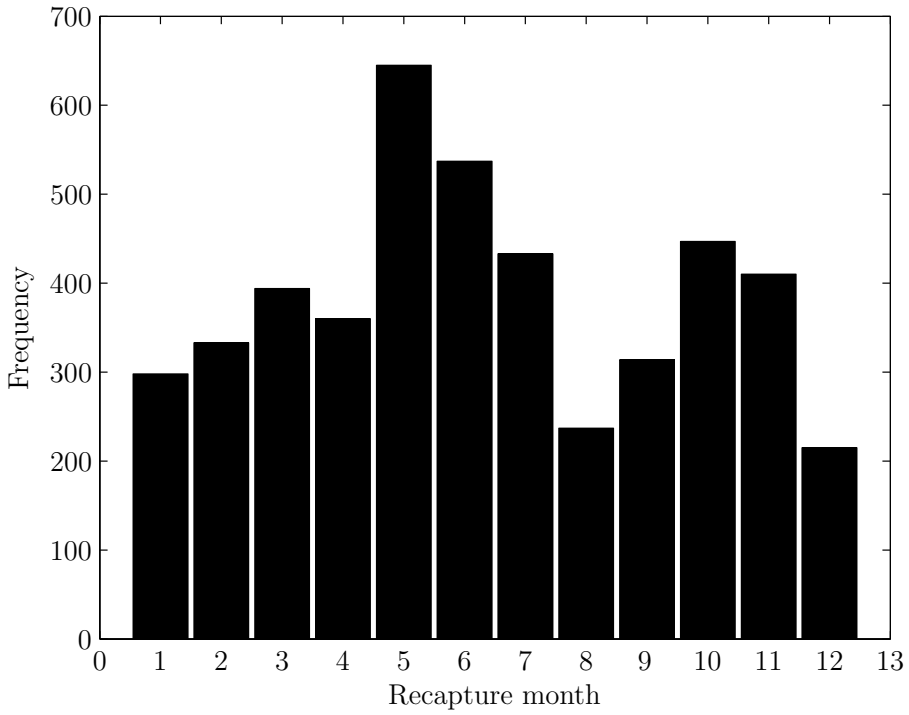


Figure 7: Histogram of the recapture month

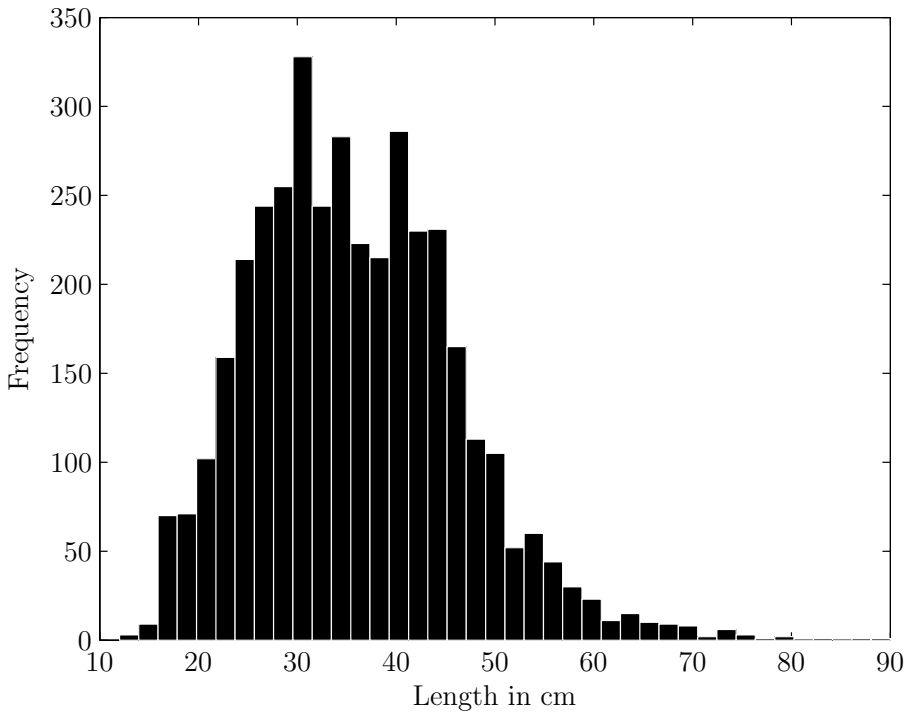
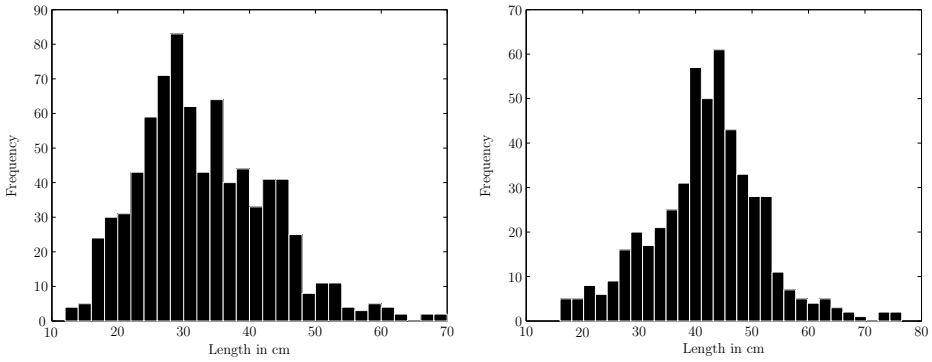
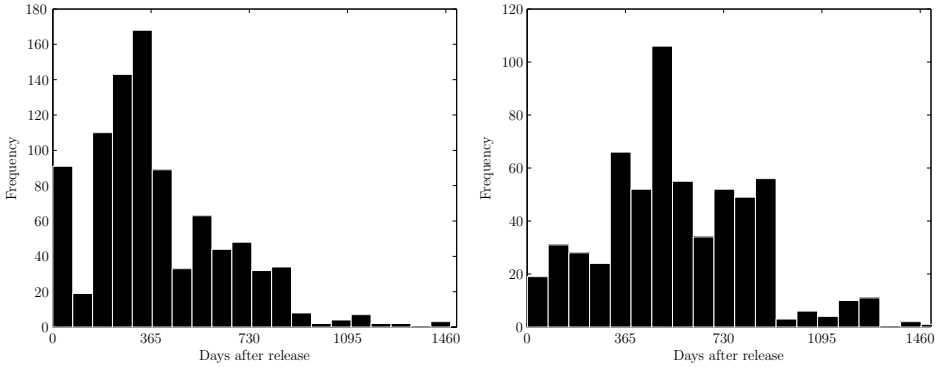


Figure 8: Histogram of length at recapture



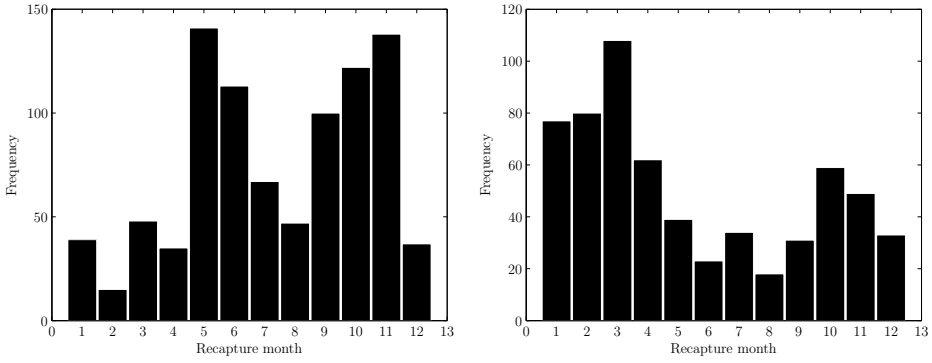
(a) Trap (b) Gill-net

Figure 9: Histogram of length at recapture for equipment trap and gill-net



(a) Trap (b) Gill-net

Figure 10: Histogram of recapture in days after release for equipment trap and gill-net



(a) Trap (b) Gill-net

Figure 11: Histogram of the recapture month for equipment trap and gill-net

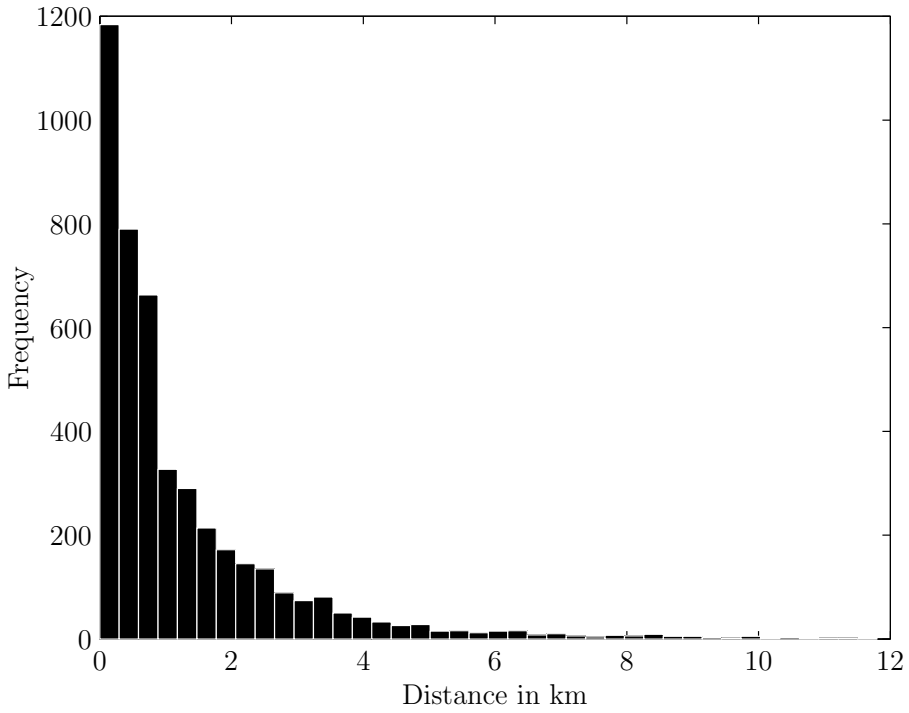


Figure 12: Histogram of distance in kilometre between captures

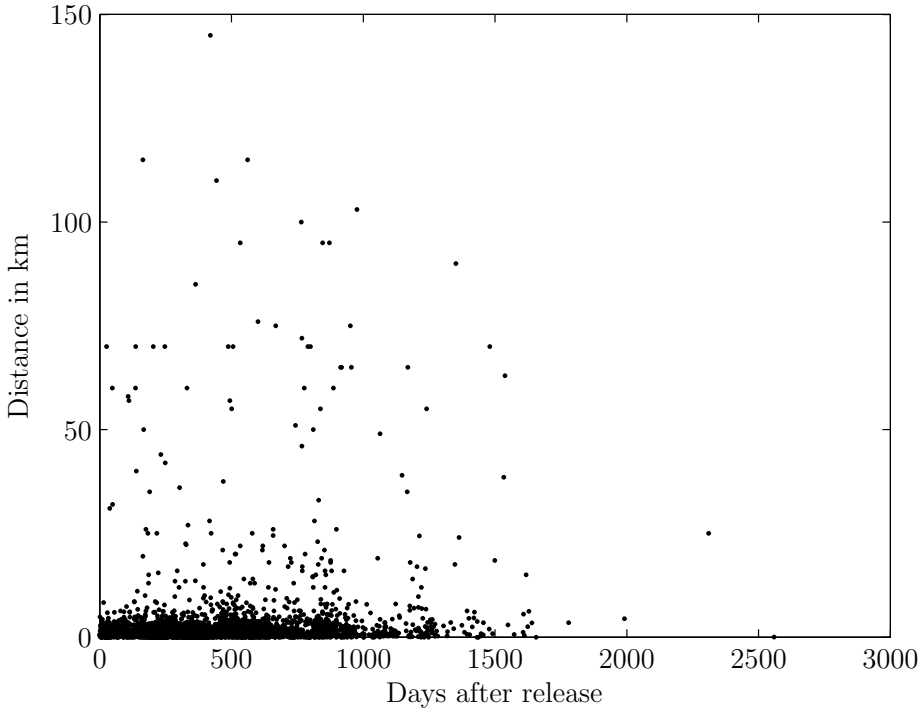


Figure 13: Scatter-plot of days of freedom vs. distance in kilometer

3 Discussion of the data

The Risør capture-recapture data set contains a lot of information, and there are potentially many biological questions that may be answered through proper analysis. Note however, that the reared cod need not necessary be representative for the natural hatched and raised cod. The individuals used in these data consist partly of a different stock than the original stock. Most of the information is based on the recaptured fish (there is some information from fish that was released and never recaptured). Thus, the results applies only to the part of the cod population which is catchable and registrable. This excludes possibly cod migrated away from the area, cod situated in non-fishing areas and cod prone to be caught by equipment where the fishermen do not return the tag for some reason. It is, however, believed that the issues presented above do not represent a major effect on the results, and that the data gives us valuable information on the life of the coastal cod stock.

The most frequent aim for mark-recapture experiments is mortality analysis. In particular, natural mortality is an important parameter of any animal population. For fish populations, however, the fishing mortality is also very important. The impact of the fisheries on the population is high, and must be taken into consideration when analysing fish populations. Interesting issues would be if there are any age or time dependencies on the mortality and furthermore if there exists any non-homogeneity effects such as size at release, age at release or release place.

From the plots in the previous chapter, we know that different gear capture cod of different size. It would be very interesting to analyse the different components of the fishing pressure to see how they affect the fishing mortality. For instance, does enhanced fishing on young cod represent a bigger danger to the stock than enhanced fishing on older cod?

The movement pattern of the cod is also a very interesting subject. The coastal cod population of the Skagerrak coast is known to be non-migrating, and the data material shows that above 90% of the recaptured fish was caught less than 10 nautical miles from the release site. But this figure does not tell the whole truth, and there are biologist who believe that the cod changes its movement pattern at maturation. Also, the vertical movement may play an important role in the understanding of the coastal cod.

3.1 What are we going to study?

In this work, the emphasis is on the estimation of mortality. As pointed out, both the natural and the fishing mortality are important measures of the cod population, and it is important to have control of the basic mortality before we possibly continue with movement analysis.

We have seen that the Risør data set contains a huge amount of data. Not only the number of recaptures, but with each capture there is a large number of attributes recorded. We need to decide what part of the capture attributes we would like to incorporate in the analysis. Naturally, the time of release and recapture must be included. Note that the cohort information is included in the time of first release (reared fish were released in October, while the wild-caught fish were released in

December). That is, the age at release and recapture is known and should be included to see if there exists an age effect. The size at release is shown to have a significant effect on the recapture probability, and should be included in the analysis. Note that we have only length at release information for the recaptured fish. For the fish never recaptured, there only exist mean and standard deviation calculations for each cohort. Release and recapture position, as well as depth recordings are excluded from the analysis at this stage, due to the complexity of the model.

The sample space spans from 10th of October 1986 to 1st of December 1993, and is discrete. However, since the number of days is large (ca. 2600) compared to the number of recaptures an individual experiences, an approximation to continuous time is plausible and useful.

It is important to address the interpretation of the capture data. This experiment differs from other mark-recapture experiment in that the sampling is not done by experimentalists, but by fishermen that would have captured the fish independent of the experiment. This means that a reported captured fish is in fact a reported death event, if it is not released again. If it is released again, the question arises: "Would it have been released again if it was not tagged?". If the answer is "yes", then the event must be considered a true capture event. If the answer is "no", then the event must be considered a death event, and the released fish must be considered as a new fish. This is because the only reason that it was released was the fact that it had a tag connected to its body. The exception is the captures by eel pots. These are generally released again, because they are not big enough to be commercially interesting. Thus, we have a segregation of the captures. Those that represent a death-by-fishing event, and those that represent a capture-and-release event. Approximately 35% of the captures were done by eel pots.

3.2 Trap happiness

Trap happiness is normally connected to semi-intellectual species who seek the traps for food or other baits. The opposite effect is trap shyness, where the act of trapping is so traumatic, that the animal avoids similar traps. In statistical terms, trap happiness indicates an increased capture probability for previously captured individuals. Trap shyness indicates a decreased capture probability for previous captured individuals.

Since cod is considered to have no memory of earlier events, we may assume that there is no trap happiness effect in a biological sense. But we may still experience a certain increase in capture probability after a re-release. We know that the fishing pressure in the area is uneven. This apply especially to eel fishermen, who place their traps at regular places. It is also known that cod is non-migrating, meaning that they move very little. Thus, it is reasonable to believe that a cod caught by an eel fisherman, who is responsible for most of the re-releases, is released back in the same area as it was caught. Since this is an area with presumably higher fishing pressure, we may assume that the fish has a larger probability of capture than a similar fish released at a random location.

Addressing this issue, let us compare the wild-caught cods released in 1989 with

Cohort	# released	% recaptured	Z under H_0
W89	237	20.7	
R88 re-released	510	44.5	6.95 (p<0.001)
W89 above 34 cm	163	22.1	
R88 re-released above 34 cm	132	45.5	4.31 (p<0.001)

Table 2: Trap happiness effect

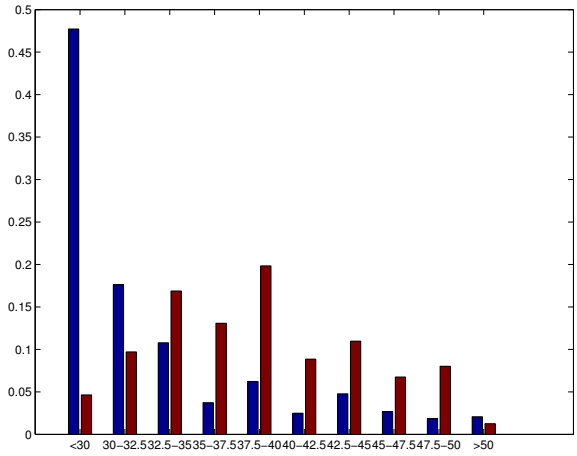


Figure 14: Size histogram of the W89 group (red) vs. the R88 group (blue)

artificially reared cods released in 1988 who were recaptured and released by eel fishermen (i.e. we only consider the subset of the R88 cohort which was captured and then released). These two groups are very similar. They are both captured by eel fishermen, and the age is approximately the same. The main issue that distinguishes them, is that the W89 cohort was released independently of where they were captured. A test was performed to see if the recapture probability was equal for the two groups. The result is listed in table 2. We see that the hypothesis of equal recapture probability is rejected at a very high level.

The size distribution for the two groups is plotted in Figure 14. We notice a clear difference in release size. To see if this difference has any effect on the test, we also performed the test on a subset of the two groups, namely the cods with release length of 35 or larger. The size distributions in this case is fairly similar (see figure 15). The hypothesis of equal recapture probability was rejected again.

The relative low recapture probability for the W89 cohort may also be explained by a high tag mortality (i.e. some amount of the cods died immediately after release due to tagging), since the W89 cohort experienced a tagging event. Measures were taken to avoid this, but there may still be an effect. However, in order for the difference to be insignificant, the tag mortality would have to be very high (above 25%).

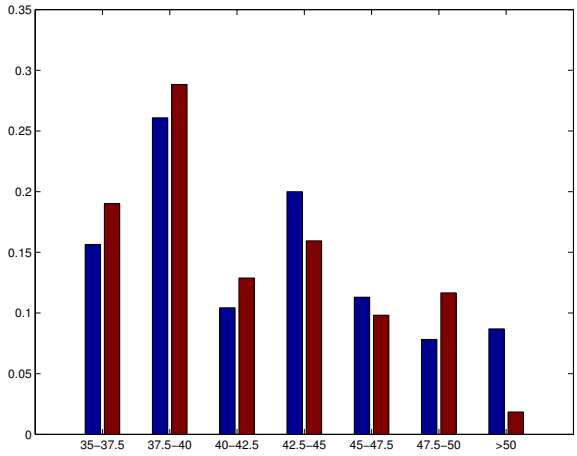


Figure 15: Size histogram of the W89 group (red) vs. the R88 group (blue) above 34 cm

3.3 Sources of bias

We have chosen to concentrate on mortality estimation, and we want to include information on time, age and cohort in the analysis. We have also shown that a trap happiness effect should be included. When we now continue, note that there are several sources of bias in the data. The tag reporting probability is very dependent on the equipment and on time (announcing was only done over a certain period of time). If the fish was re-released, there is no guaranty that the fish was unhurt. In addition, the question whether the reared cod is representable for the "natural" cod population is unresolved.

The toxic algae bloom of 1988 discussed in Section 2.1 is an effect that cannot be overlooked. As pointed out by Chan et al. (2003), it affects the mortality of the cod population not only during the bloom, but possibly also the following generations. This could be the reason for the high recovery of the R88 cohort and the low recovery of the R87 cohort. When the 1988 0-group died in the algae bloom, the 1-group cod of 1988 (R87) experienced a lack of prey (assuming cannibalism). A possibly delayed increased mortality of the R87 cohort may be due to this lack of pray. When the R88 cohort was released, the 1-group predators were weakened, leading to a decreased mortality in the vital first winter of the R88 cohort.

The difference in recapture between the R87 and the R88 cohort (see Table 1) may also have other explanations, as suggested by Danielssen and Gjøsæter (1994) and Julliard et al. (2001). The size at release is generally very important in release experiments. The size at release were on average smaller for the R87 cohort compared to the R88 cohort. The level of awareness of the experiment among the fishermen, and thus the tag return rate, may not be assumed to be constant, and could explain some of the difference. Other suggested explanation variables were transportation stress and environmental factors such as salinity of the water.

4 Relation to the existing literature

There are two previous papers on the Risør mark-recapture experiment, Danielssen and Gjøsæter (1994) and Julliard et al. (2001). The first is an overview of the experiment with some explanatory analysis, while the last is a more thorough analysis with emphasis on mortality analysis.

4.1 Danielssen and Gjøsæter (1994)

This is a descriptive report of the experiment, with no deeper analysis. Materials and methods are presented, and some results regarding recapture, migration, vertical migration and type of fishermen and gear are given. In the discussion, some explanation to the difference in recapture between the cohorts are given, along with the confirmation of earlier studies that the Norwegian Skagerrak coast cod population is non-migrating.

4.2 Julliard et al. (2001)

This paper is by far the most advanced and thorough of the papers on the Risør mark-recapture data. It addresses several central properties of the Risør coastal cod population such as natural and fishing mortality, and seasonal pattern of fishing. Several issues concerning the experiment is considered, such as the size at release for the reared juveniles, the recapture probabilities and the tag return rate. The effect of the algae bloom is also analysed.

In order to analyse these issues, they mainly use the mark-recapture methodology built around the Cormack-Jolly-Seber (CJS) model. The authors seem to use the software SURGE to compute the maximum likelihood estimates of the parameters. For hypothesis testing and comparison, they use likelihood ratio tests (LRT) and the Akaike information-theoretic criterion (AIC). Let us present the analysis and the result of this paper.

The mark-recapture data of the Risør experiment is, as we have already stated, almost continuous. Since the CJS model assumes that the capture and mortality periods are disjoint, some adjustments have to be done. In their paper, the authors discretise the capture data in monthly intervals. That is, they assume that the captures periods last one month. This implies that the mortality periods are squeezed between capture periods, and are assumed instantaneous. There are made no attempts to explain how the violation of this assumption affects the results.

It is assumed that the monthly capture and mortality probabilities are time and age dependent. However, some constraints are introduced to make the parameters independently identifiable. For survival, a time and age structure with four age classes is assumed, each interacting with year. This structure demands 19 parameters. The capture probability is modelled more detailedly with different capture probability each month, each interacting with age class (180 parameters). Note that the number of parameters for the capture probability is much larger than the number of parameters for the survival probability. This is reasonable because there is much more information on capture than on survival in the data. This they denote as

the panel model, and more constrained models are compared with this model using likelihood ratio tests and AIC.

In addition to CMR analysis, an analysis of the size effect on recovery rate is performed. It shows that the size at release greatly influences the recapture rate. This is due to the low capture probability of small individuals, together with a high mortality for young and small cods.

The main results of their analysis are as follows: The size effect analysis concluded that the size at release had a great impact on the recovery rate, at least for the reared cod. For the wild-caught cod, this impact was less pronounced, and disappeared when the smallest individuals were excluded from the analysis. An analysis of the disappearance right after release indicated that this had no effect on the reared cohorts, but that approximately 60% or more of the wild-caught cod died or migrated just after tagging. Also, the tag return rate was estimated to be about 50%-60%. They admit that this analysis is very crude, though. The mortality analysis stated that a high natural mortality rate was found during the first 6 months after release for the reared cohorts. For the older age-classes, the natural mortality was reported low, and most of the mortality was then due to fishing.

4.3 Weaknesses in Julliard et al. (2001)

In the paper, the authors use the same data both for model selection and for inference. This is generally problematic in a decision-theoretic view, and tend to underestimate the width of the parameter confidence intervals. The exact magnitude of this underestimation is hard to compute, and differs by parameter and model. An analysis of this effect is difficult, and we do not go down that road. Refer to Lebreton et al. (1992) who states that: "...our preliminary feeling is that the effect on measures of precision is not a major problem in capture-recapture.".

The paper states that the algae bloom had a great impact on the mortality of all age classes. They prove this by comparing a model of the mortality where the bloom effect was included, with a model where the bloom effect was excluded. The analysis gave a clear rejection on the latter model, while the model including the bloom effect was not rejected. This seems like a sound analysis, but be aware. We know that there are only three cohorts released at the time of the bloom; R86, R87 and W86. We know also that the recapture rate of the R86 and R87 cohorts were very low compared to the other cohorts. There are some evidence that this low recapture rate relates to the relative small average size of these two cohorts. This size effect is not taken account for in the CJS model analysis, and the algae bloom is possibly confounded with the low recapture probability. The analysis of Chan et al. (2003) together with the discussion in Section 2.1 also indicates that the algae bloom did not affect the released cohorts greatly, at least not during or directly after the bloom.

The model used to analyse the Risør experiment, assumes a normal capture-recapture experiment. This means that the sampling is done by experimenters, and that any interference by the experimenters on the population must be accounted for in the model. For instance, censoring due to damage or death during trapping

must be included in the model. In the Risør experiment, the sampling is done by fishermen which interfere with the population independent of the experiment, as discussed in chapter 3.1. That is, when a cod is captured, and the tag is returned, this is both a capture event and an observed death event. This is because the fishing used to gather data is part of the total mortality of the stock. This difference in interpretation of the capture events is not discussed in the paper.

With high dimensional parametric models such as the Cormack-Jolly-Seber model, model selection routines are used to choose between the different parameterisations. This is done thoroughly in the paper. However, when the parameter space is as large as in the Cormack-Jolly-Seber model of Julliard et al. (2001), it is impossible to go through every available model. A selection has to be made subjectively, and there are no way to be certain that this selection contains the model with highest likelihood.

By their own analysis, the authors report that the size at release is an important covariate for the recovery rate, at least for the reared cohorts. They fail, however, to include this information in the CJS model. This exclusion may lead to bias, especially on the survival probabilities of the R86 and R87 cohorts for which the size at release is significantly different from the other cohorts.

The last weakness we shall discuss, is perhaps the most interesting. We have described how the authors discretise the data to fit them into the CJS model. We would like to know how this discretation affects the results. In the next section, we see how this adjustment of the data possibly leads to bias.

4.4 Using CJS models on continuous data

When a discrete model is used to analyse continuous data such as the Risør mark-recapture data, a natural question arises; What is the cost of using this sort of simplification? Let us try to analyse the situation when continuous mark-recapture data is analysed by the Cormack-Jolly-Seber model.

First we take a closer look at the CJS model. The model is built around an assumption that the experiment consists of alternating capture and mortality periods, with no mortality during the capture periods, and no capture during the mortality periods. This assumption is usually met by limiting the capture period such that deaths during the period is unlikely. For continuous data, the capture and mortality processes run simultaneously, and we have no purely capture or mortality periods. If we want to analyse this situation using the CJS model, we need to adjust the data. A natural method is to divide the whole experiment time span into smaller intervals, and then pool the captures and releases within each interval. This approach implies that the capture occasion is considered to lie in the middle of the interval, and that some events are moved forward in time and some are moved backward in time. If the individual is not captured again during a later interval, this movement influences the estimates of the survival probability. Consider an individual that is captured during the first half of an interval, released and never seen again. By the movement forward in time due to the discretation process, this individual has it known lifespan prolonged. It thus contributes to an overestimation of the survival.

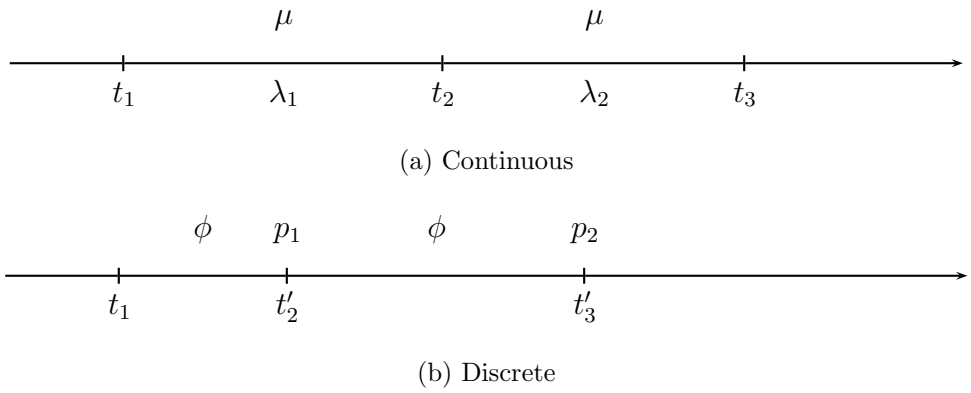
t_1	t'_2	t'_3
R_1	m_{12}	m_{13}
	R_2	m_{23}

Table 3: The discretised recapture data, in standard mark-recapture notation.

Individuals that are captured during the last half of the interval, released and never seen again, contributes to an underestimation of the survival. It has its known lifespan shortened. If there for some reason is more captures during the first half, the survival probability of this interval is overestimated, otherwise it is underestimated.

The bias in survival probability may also affect the capture probabilities. If the survival probabilities are overestimated, more individuals are assumed alive at the capture occasions. The number of captures are the same, and the capture probabilities are underestimated. If the survival probabilities are underestimated, the capture probabilities are overestimated.

Let us present an example to illustrate. We consider a situation where we have continuous captures and removals. At time t_1, R_1 marked individuals are released. In the first time period after release, they are captured with capture rate λ_1, and in the second period with capture rate λ_2. After a capture, the individual is released. The mortality rate for both periods is μ. After the second time period, no more individuals are recaptured. For each individual, the recapture events are recorded. See Figure 16a for illustration.



(a) Continuous

(b) Discrete

Figure 16: Illustration of the mark-recapture example situation

We now analyze this situation using the Cormack-Jolly-Seber model, which is a discrete model. The discretation is done as described, with the new capture occasions t'_2 and t'_3 in the middle of the interval. Using standard mark-recapture notation, the discretised data may be summarised as in Table 3. Recall that m_{12} is the number of individuals released at t_1 and recaptured at t'_1, m_{13} is the number of individuals released at t_1 and not recaptured again until at t'_3, and m_{23} is the number of individuals released at t'_2, and recaptured at t'_3. Since all recaptured individuals are released, R_2 equals m_{12}. The parameters included in the model are p_1 and p_2, the capture probability at capture occasion t'_1 and t'_2, and ϕ_1 and ϕ_2, the

probability of surviving respectively the first and second time period. Since we have more parameters than data, we need to restrain the parameter space. This is usually done by letting $\phi_1 = \phi_2$. See Figure 16b for illustration. Note that this is the same setup as the model approximation presented in Julliard et al. (2001). The initial release event happens right at the middle of an interval in the discrete sense, and does not contribute to neither a overestimation or underestimation of the survival. Since the number of individuals alive is declining throughout the experiment, the number of captures is also declining. This means that there is a majority of captures at the first half of the interval. Since these captures contribute to the overestimation of the survival probability, we may suspect an overestimation of the overall survival probability.

Let us consider a numerical example. A set of data is simulated with $\lambda_1 = 0.2, \lambda_2 = 0.3$ and $\mu = 0.2$. 100,000 individuals are simulated, which is a large enough number that any asymptotic properties should apply. The resulting simulated data are presented in Table 4. The probability of being captured in the first period is

t_1	t_2'	t_3'
$R_1 = 100000$	$m_{12} = 16590$	$m_{13} = 15672$
	$R_2 = 16590$	$m_{23} = 3586$

Table 4: The simulated data

given by $p_1 = 1 - \exp(-\lambda_1) = 0.181$, $p_2 = 1 - \exp(-\lambda_2) = 0.259$ and the probability of surviving a period is $\phi = \exp(-\mu) = 0.818$.

An analysis was performed using the recognized program MARK (White and Burnham (1999)) and the result is presented in Table 5. We see that the survival probability is overestimated, and that not even the 95% confidence interval coveres the true value by far. There is a one-to-one conversion between survival probability and mortality rate, $\mu = -\log(\phi)$. When we apply this conversion to our situation, we see that the estimated mortality rate $\hat{\mu}$ is $-\log(0.8909) = 0.1155$ which is almost half the true parameter value. The first capture probability is overestimated (within the 95% confidence interval) while the last is underestimated (outside the 95% confidence interval). With the exception of the first capture probability, this example confirms what we had expected from the discussion earlier.

The bias due to discretation is pronounced in this example. Note however, that this is not an example of a typical mark-recapture situation. It has only two capture intervals, and the capture rate is relatively high. In addition, there is a very large number of individuals simulated. But the example illustrates the point.

We wanted originally to know if the discretation of the Risør mark-recapture experiment implies a major bias to the survival estimates. The answer is probably no. The reason is that the discretation interval of one month is to short that we may suspect a significant change in capture from the start of the interval to the end. The capture or mortality intensity is not high enough with respect to the interval length.

Parameter	Estimate	Standard Error	Lower	Upper
1:Phi	0.8909379	0.0119235	0.8652788	0.9122058
2:p	0.1862083	0.0028051	0.1807727	0.1917690
3:p	0.2426143	0.0066596	0.2298012	0.2559044

Table 5: Results from the discrete analysis of the simulated data

4.5 Previous analysis of similar data elsewhere

The mark-recovery method, as presented by Brownie et al. (1985), should be considered when analysing maritime mark-recapture situations such as the Risør experiment. The basis for mark-recovery experiments is that marked individuals may only be registered after their death. That is, a death event is required in order for a recovery event to take place. This requirement is partly fulfilled in the Risør experiment, since a fish is only registered by a fishing event, which in fact usually is a death event (see remark at the end of Section 3.1). By the mark-recovery method, no distinction is made between a natural death and a fishing death. This is in contrast to the mark-recapture method, where $(1 - \phi_i)$ is the probability of dying a natural death or being fished without registering, while p_i is the probability of dying and being registered. The result is that we utilise more of the information in the data to survival estimation. That is, we use the information that a capture event is not only that, but also a registered death event. The drawback of this approach is that we cannot say anything about the natural mortality compared to the fishing mortality. Relying on assumptions on the tag report probability, this is possible in a mark-recapture analysis.

Bias due to discretation of continuous data also apply to the mark-recovery analysis, and a discussion similar to the one in Section 4.4 may be made.

The capture-release behaviour of the eel fishermen may prove troublesome for this model. Since they are known to release small individuals, these captures may not be viewed as recoveries. To cope with this, every capture-release event made by an eel fisherman may be considered as a resighting event as described by Catchpole et al. (1998).

Other ideas to consider is the use of Bayesian methods as described in Brooks et al. (2000). This approach does not solve the basic problem with bias due to discretation of continuous data, but it may used to include auxiliary information about the Risør cod stock that otherwise could not have be included.

4.6 Summary

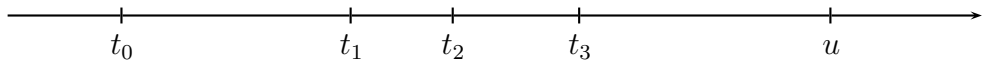
We have seen that none of the existing literature on analysis of the Risør mark-recapture experiment model the data consistently. The Danielssen and Gjøsæter (1994) paper is superficial, and the Julliard et al. (2001) use a model that does not take into account important aspects of the data. In addition, it leaves out important information such as the release size and trapping effects. Thus, there is a strong need for a reanalysis of the data, with other and better models.

5 Using counting processes to model mark-recapture data

As we have seen, the use of discrete models on continuous mark-recapture data may lead to biased survival probability estimates. This chapter presents a new, continuous model to mark-recapture data, and thereby establishes a link between mark-recapture methodology and classical survival analysis.

5.1 The basic model

Let us begin by proposing a general continuous mark-recapture model. The essential events for an individual in a mark-recapture experiment, is the release time t_0, the recapture times $t_1, \ldots t_k$, and the time of death, u (see figure 5.1). Note that the time variable $t.$ may as well be age as calendar time. We assume independence between the individuals. For now, we assume that the time of death is observed. The notation and theory are based on Andersen et al. (1993).



Let $N_i^p(t)$ be the number of capture events and $N_i^\phi(t)$ be the number of death events for individual i on the time interval $[t_0, t]$. Clearly, $N_i^\phi(t) \leq 1$, while $N_i^p(t) \in \{0, 1, 2, \ldots\}$. Given the capture intensity $\alpha_p(t)$ and the mortality intensity $\alpha_\phi(t)$, the intensity processes for the two counting processes $N_i^p(t)$ and N_i^ϕ are given by

$$\lambda^p(t) = \alpha_p(t) Y_i(t)$$
$$\lambda^\phi(t) = \alpha_\phi(t) Y_i(t)$$

where $Y_i(t)$ equals one if individual i is at risk of capture or death at time t and else zero. Since $\alpha_p(t)$ and $\alpha_\phi(t)$ are deterministic, and $Y_i(t)$ is a predicable and observable process, this is a multiplicative intensity counting model as defined by Andersen et al. (1993)

Now, let

$$N_\cdot^p(t) = \sum_{i=1}^{n} N_i^p(t)$$
$$N_\cdot^\phi(t) = \sum_{i=1}^{n} N_i^\phi(t),$$

the total number of captures and deaths before time t, and

$$Y_\cdot(t) = \sum_{i=1}^{n} Y_i(t),$$

the total number of individuals at risk of capture or death at time t. $N_{\cdot}^{p}(t)$ and $N_{\cdot}^{\phi}(t)$ are then counting processes with intensity processes

$$\lambda_{\cdot}^{p}(t) = \alpha_p(t)Y_{\cdot}(t)$$
$$\lambda_{\cdot}^{\phi}(t) = \alpha_\phi(t)Y_{\cdot}(t).$$

Note that the only restrictions on $\alpha_p(t)$ and $\alpha_\phi(t)$ are that

$$\int_{t_0}^{t} \alpha_p(u)\, du < \infty$$

and

$$\int_{t_0}^{t} \alpha_\phi(u)\, du < \infty$$

for all t in the sample space.

5.2 Estimation

We have defined a model that suits our data, and now we are interested in estimating the capture and mortality intensities. We use the nonparametric Nelson-Aalen estimator, proposed independently by Nelson (1969, 1972) and Altshuler (1970), and generalised to use for counting processes by Aalen (1975, 1978). Let

$$\widehat{A}_p(t) = \sum_{\{j,i:t_{ji}\leq t\}} Y_{\cdot}(t_{ji})^{-1} \tag{1}$$

and

$$\widehat{A}_\phi(t) = \sum_{\{i:u_i\leq t\}} Y_{\cdot}(u_i)^{-1} \tag{2}$$

where t_{ji} is the j'th capture event of individual i, and u_i is the time of death for individual i. That is, the estimators are step functions starting in zero, and for each event (capture or death) the respective estimator is increased by one over the number of individuals at risk for that event.

Note that the Nelson-Aalen estimator is an estimator of the cumulative intensities

$$A_p(t) = \int_{t_p}^{t} \alpha_p(u)du,$$

and

$$A_\phi(t) = \int_{t_0}^{t} \alpha_\phi(u)du.$$

So far, our discussion has assumed that we know the time of death for all individuals. This is generally not the case in mark-recapture experiments. In the Risør data set, we observe dead recaptures which we regard as death events, but a substantial part of the death events are never reported. There are two reasons for this. The individuals may be captured dead but not reported, or they may have

died naturally. We are thus dealing with a missing data situation, and a natural next step is to introduce the EM (expectation-maximisation) algorithm introduced by Dempster et al. (1977) (more precisely we apply the Monte Carlo EM algorithm proposed by Wei and Tanner (1990)).

To ease the notation, let $\theta = (A_p(\cdot), A_\phi(\cdot))$, the cumulative capture and mortality rates of our model, and let $\mathbf{t}$ be the data in the model. This means that $\mathbf{t}$ encapsulates all our information about capture and reported deaths. Let $L(\theta|\mathbf{t})$ be the total likelihood of the cumulative intensities. The ultimate goal is to maximise this likelihood. Let $\mathbf{u}$ be the unknown death events of the released individuals. We know that the Nelson-Aalen estimators (1) and (2) maximises the augmented likelihood $L(\theta|\mathbf{t}, \mathbf{u})$, and the idea is to maximise $L(\theta|\mathbf{t})$ with the help of this knowledge. Note that it is not necessary to know the explicit expression for this likelihood, it is enough to know that the Nelson-Aalen estimator maximises it.

Using common EM-notation, we let for any value θ_0,

$$Q(\theta|\theta_0, \mathbf{t}) = \mathbb{E}_{\theta_0}[\log L(\theta|\mathbf{t}, \mathbf{u})|\theta_0, \mathbf{t}], \tag{3}$$

where the expectation is with respect to $f(\mathbf{u}|\mathbf{t}; \theta_0)$, the conditional distribution of the missing data $\mathbf{U}$ given the observed data $\mathbf{t}$. This conditional distribution is given by

$$f(\mathbf{u}|\mathbf{t}; \theta_0) = \prod_{i=1}^{n} \left(\alpha_\phi(u_i) \ \exp\left\{ - \int_{t_{ki}}^{u_i} (\alpha_p(v) + \alpha_\phi(v)) \, dv \right\} \right) \tag{4}$$

where t_{ki} is the last release event of individual i. The first term of the expression may be seen as the probability density of death at time u_i, while the last term is the probability of no death or capture events between t_{ik} and u_i. A sequence $\widehat{\theta}_{(1)}, \widehat{\theta}_{(2)}, \ldots$ is now defined by the identity

$$Q(\widehat{\theta}_{(j+1)}|\widehat{\theta}_{(j)}, \mathbf{t}) = \max_\theta Q(\theta|\widehat{\theta}_{(j)}, \mathbf{t}).$$

Standard EM theory states that this sequence satisfies

$$L(\widehat{\theta}_{(j+1)}|\mathbf{t}) \geq L(\widehat{\theta}_{(j)}|\mathbf{t})$$

with equality holding if and only if

$$Q(\widehat{\theta}_{(j+1)}|\widehat{\theta}_{(j)}, \mathbf{t}) = Q(\widehat{\theta}_{(j)}|\widehat{\theta}_{(j)}, \mathbf{t}).$$

This means that the likelihood we seek to optimise, increases with each step of the sequence.

Computing the expectation in equation (3) proves hard, and the following maximisation even harder and somewhat futile, because we are not able to utilise the maximum likelihood properties of the Nelson-Aalen estimators. Instead of computing the expectation analytically, it is estimated by simulation. Let

$$\widehat{Q}(\theta|\theta_0, \mathbf{t}) = \frac{1}{l} \sum_{i=1}^{l} \log L(\theta|\mathbf{t}, \mathbf{u_i}),$$

where the vector $\mathbf{u}_i$ are simulated times of death from the conditional distribution $f(\mathbf{u}|\mathbf{t};\theta_0)$. When l goes to infinity,

$$\widehat{Q}(\theta|\theta_0,\mathbf{t}) \to Q(\theta|\theta_0,\mathbf{t}),$$

by the law of large numbers, and the limiting form of the Monte Carlo EM algorithm is the regular EM algorithm. Note that instead of maximising $\widehat{Q}(\theta|\theta_0,\mathbf{t})$, we might as well maximise $l \cdot \widehat{Q}(\theta|\theta_0,\mathbf{t})$, such that

$$\widehat{Q}(\widehat{\theta}_{(j+1)}|\widehat{\theta}_{(j)},\mathbf{t}) = \max_{\theta} \sum_{i=1}^{l} \log L(\theta|\mathbf{t},\mathbf{u_i})$$

where $\mathbf{u}_i$ again is simulated from $f(\mathbf{u}|\mathbf{t};\widehat{\theta}_{(j)})$. The sum in the above expression may now be seen as the likelihood of a complete counting process, and the estimator maximising this process is the Nelson-Aalen estimator. More explicit, let the capture times of a typical individual i be $\mathbf{t}_i$, and assume that it is released after the last capture. Now, we simulate l "new" individuals, each with a unique simulated time of death u_{ij}, but with equal capture times $\mathbf{t}_i$, such that we get the augmented dataset $(\mathbf{t},u_{i1}),\ldots,(\mathbf{t},u_{il})$. This augmentation is repeated for each individual, and the resulting dataset forms the basis for the Nelson-Aalen estimator. Note that there are true ties in the augmented dataset, and the estimator for $A_p(t)$ is changed to

$$\widehat{A}_p(t) = \sum_{\{ji\,:\,T_{ji}\le t\}} \frac{l}{Y_{\cdot}(T_{ji})},$$

where the simulated times of death are included in $Y_{k\cdot}(t)$.

The simulation scheme for the time of death u_{ij} is a simple rejection sampler. For every individual captured k times and released but never recaptured, sample l times of death using the following scheme:

Algorithm 1 The time of death sampler

for $j = 1$ to l **do**
  repeat
    Draw $V, W \sim \text{Unif}[0,1]$;
    Let X be such that $\widehat{A}_\phi(X) = \widehat{A}_\phi(t_k) - \log(1-V)$;
  until $\exp\{-(\widehat{A}_p(X) - \widehat{A}_p(t_k))\} \ge W$
  $U_j = X$;
end for

This algorithm produces l samples for the conditional density of the time of death given in equation (4).

Proof.

$$
\begin{aligned}
P(U \leq u) &= P\left(X \leq u | \exp\{-(A_p(X) - A_p(t_k))\} \leq W\right) \\
&= \frac{P\left(X \leq u \cap \exp\{-(A_p(X) - A_p(t_k))\} \leq W\right)}{P\left(\exp\{-(A_p(X) - A_p(t_k))\} \leq W\right)} \\
&= \frac{\int_{t_k}^{u} \int_{0}^{\exp\{-(A_p(x) - A_{k \wedge K}(t_k))\}} dw \exp\{-(A_\phi(x) - A_\phi(t_k))\} \alpha_\phi(x) dx}{\int_{t_k}^{\infty} \int_{0}^{\exp\{-(A_p(x) - A_p(t_k))\}} dw \exp\{-(A_\phi(x) - A_\phi(t_k))\} \alpha_\phi(x) dx} \\
&= \int_{t_k}^{u} \exp\{-(A_p(x) - A_{k \wedge K}(t_k))\} \exp\{-(A_\phi(x) - A_\phi(t_k))\} \alpha_\phi(x) dx \\
&= \int_{t_k}^{u} \alpha_\phi(x) \ \exp\left\{ - \int_{t_k}^{x} (\alpha_p(v) + \alpha_\phi(v)) \ dv \right\} \ dx \\
&= \int_{t_k}^{u} f(x|t_k) \ dx
\end{aligned}
$$

□

A nice property of the sampler is that it uses the Nelson-Aalen estimates directly at each step, and never needs to obtain estimates of the intensity rates $\alpha_p(t)$ or $\alpha_\phi(t)$.

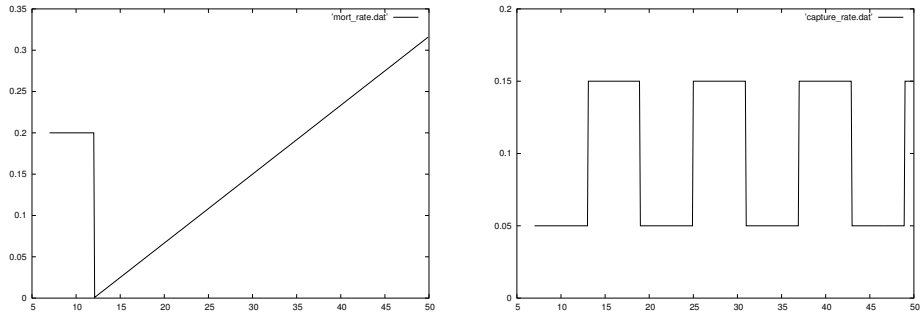
The explicit scheme of the Nelson-Aalen EM algorithm is given in Algorithm 2. The number of iterations N should be large enough for the Nelson-Aalen estimates for the cumulative capture and mortality rates to have converged.

Algorithm 2 The Nelson-Aalen EM algorithm

```
Initialise Â_p(t) and Â_φ(t)
for i = 1 to N do
  for j = 1 to n do
    if individual j is released after recapture or never recaptured again then
      for k = 1 to l do
        Draw a death event u_jk with respect to Â_p^(i)(t) and Â_φ^(i)(t) using Algorithm 1
      end for
    end if
  end for
  Compute the new Nelson-Aalen estimates Â_p^(i+1)(t) and Â_φ^(i+1)(t) using the simulated death times and the observed capture times.
end for
```

6 Simulation Study

We would like to test this new estimator, which is hereby denoted the EM Nelson-Aalen (EM-NA) estimator, against the existing discrete Cormack-Jolly-Seber (CJS) estimator, on simulated continuous data. Two data sets are simulated. The first



(a) Capture Rate (b) Mortality Rate

Figure 17: The capture rate and the mortality rate with respect to age in months, of the simulation study

tries to mimic the attributes of the Risør data set, while the second contains more information in form of increased reporting and re-release probabilities.

The capture and mortality rate of the two simulated data sets are given in Figure 17, and follows our beliefs on cod mortality and capture rates. The capture rate is set to follow a seasonal pattern, with a period of one year. The suggested mortality rate follows from a belief that the cod experiences high mortality for the first year of his life (the first five months for the cod released at seven months age). After this period, the mortality drops to a low level from which it steadily rises as the cod grow older.

In the first simulated data set, the reporting probability is 0.5, and the re-release probability is 0.24. This is partly in accordance with the Risør data. In the second simulated data set, the reporting probability is 0.9 , and the re-release probability is 0.9. These choices imply that the second data set contains much more information than the first data set.

The number of simulated cod is equal in the two data sets; 37 000 released at the age of 7 months, and 2400 released at age 19 months. The simulation follows algorithm 3

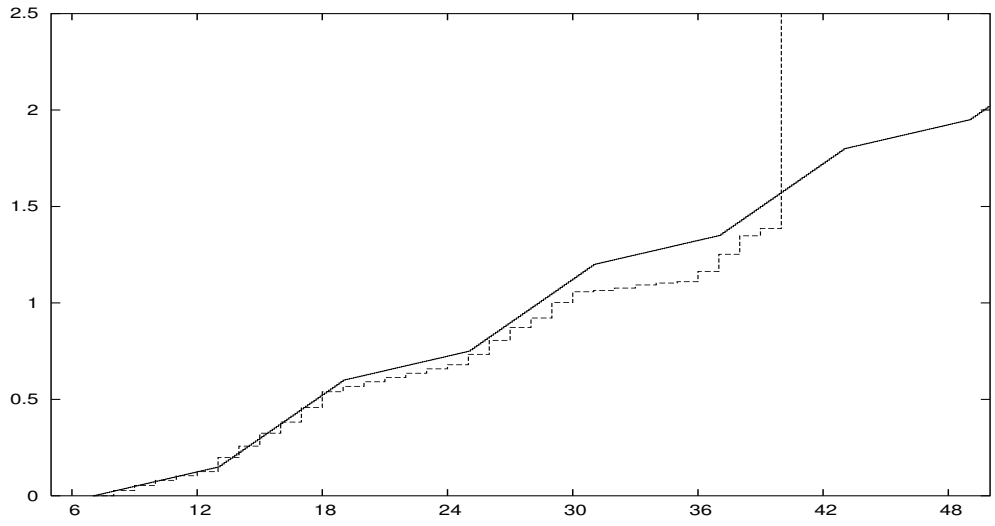
6.1 Results from the first simulated data set

The estimates for the cumulative capture and mortality rate, using the Cormack-Jolly-Seber and the EM-NA estimators are presented in Figure 18 and 19.

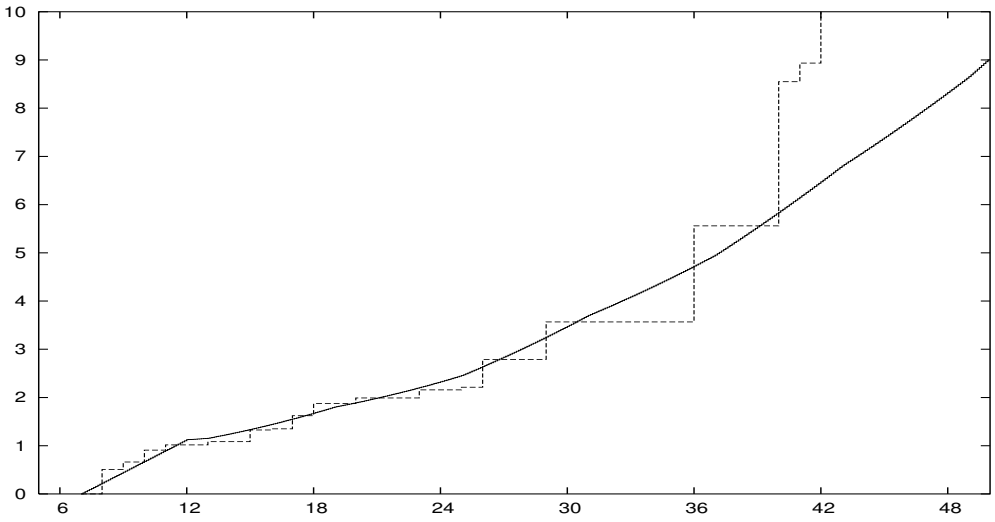
We see that the capture rate is underestimated by the CJS estimator from age 18 months and onwards. The mortality rate is also underestimated between between 20 and 38 months, otherwise it is fairly well estimated.

For the EM Nelson-Aalen estimator, estimation of the capture rate is accurate up to an age of 27 months. After this, the estimator underestimates the rate. The mortality rate is relatively well estimated up to age 23, after which it drifts off into underestimation.

Both the EM-NA estimator and the CJM estimator of the capture rate performs quite well. When estimating the mortality rate, the EM-NA estimator performs better up to the age of 27, after which the EM-NA estimate drifts off. The reason

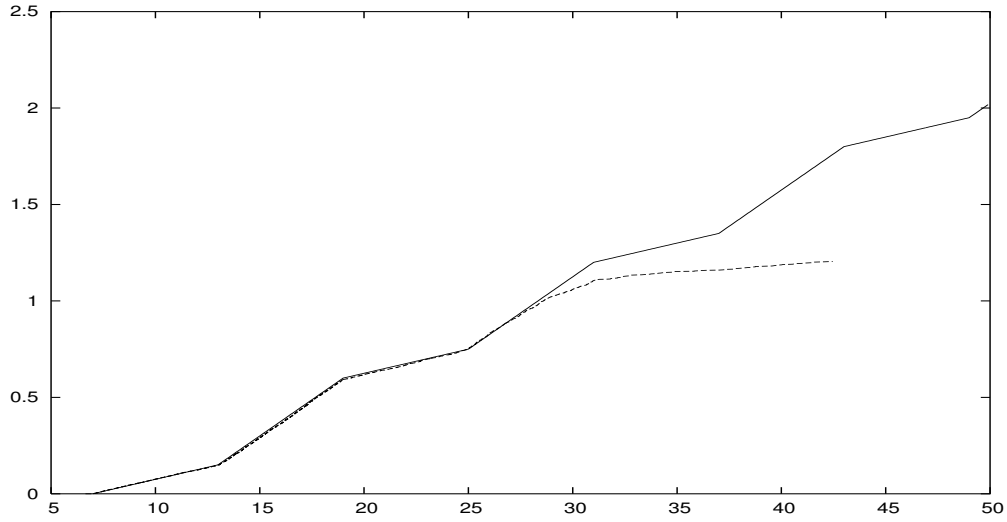


(a) Cumulative Capture Rate

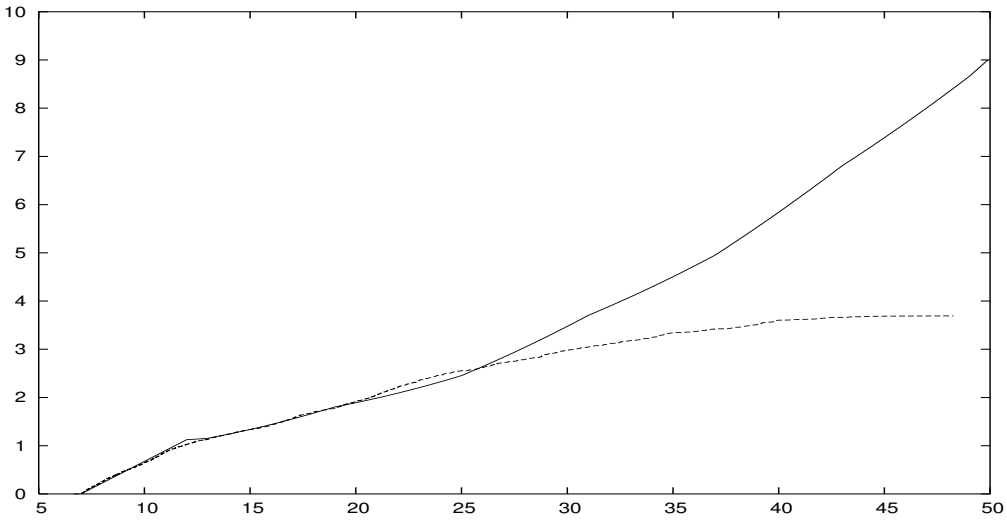


(b) Cumulative Mortality Rate

Figure 18: Cormack-Jolly-Seber estimates of the cumulative capture and mortality rate with respect to age in months. The true rate is given by the whole drawn line, while the estimated rate is given by the dotted line.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 19: EM-NA estimates of the integrated capture and mortality rate with respect to age in months. The true integrated rate is given by the whole drawn line, while the Nelson-Aalen estimate is given by the dotted line.

Algorithm 3 Simulation scheme

```
for j = 1 to N do
    Draw a death event X according to the mortality rate
    Draw a capture event Y according to the capture rate
    while Y < X do
        Draw U_1 ~ Unif[0, 1]
        if U < p_reg then
            Register Y
            Draw U_2 ~ Unif[0, 1]
            if U < p_rel then
                Draw a new capture event Y according to the capture rate
            end if
        else
            Set Y > X to end the while loop
        end if
    end while
end for
```

for this drift, is probably some badly defined boarder conditions. We notice that the CJS cumulative capture rate estimate displays a similar drift behaviour around age 30, but it is forced back on track.

A discretation effect is possibly present, and the underestimation of the capture rate from age 18 and onward may be due to this.

6.2 Results from the second simulated data set

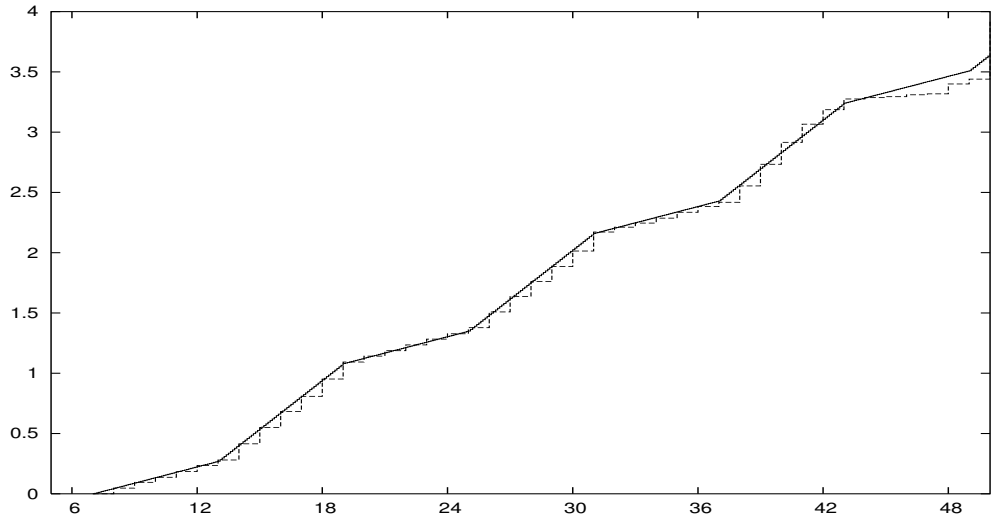
The estimates of the cumulative capture and mortality rates, using the Cormack-Jolly-Seber and the EM-NA estimators, are presented in Figure 20 and 21

The Cormack-Jolly-Seber estimator performs very well. There is again a slight underestimation of the true capture rate, but this is less explicit than for the first simulated data set. We notice the oscillating behaviour of the estimator around the true mortality rate. The mortality rate is also nicely estimated, with some underestimation during the first 6 months.

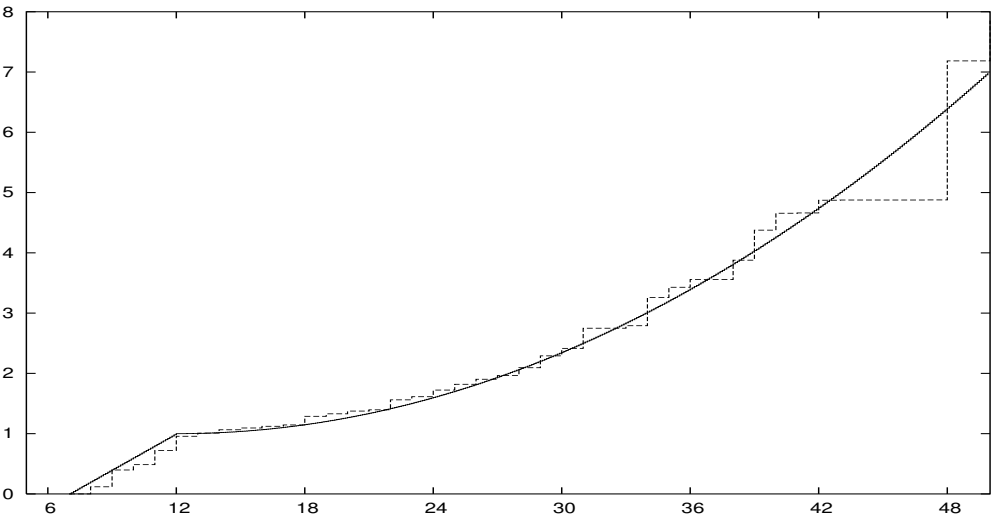
For the EM-NA estimator, the capture rate estimation is very good. For the mortality rate, the estimator performs even better than the CJS estimator with less oscillation.

Performance on the capture rate estimation is comparable between the two estimation methods. The mortality rate, however, is better estimated by the EM-NA estimator. There is less oscillation around the true rate. It is interesting to notice the similarities between the two CJS estimators and the two EM-NA estimators. They seem to follow each other, with smaller amplitude for the EM-NA estimators.

Again, the slight underestimation of the capture rate with the CJS method might be due to the discretation of the data.

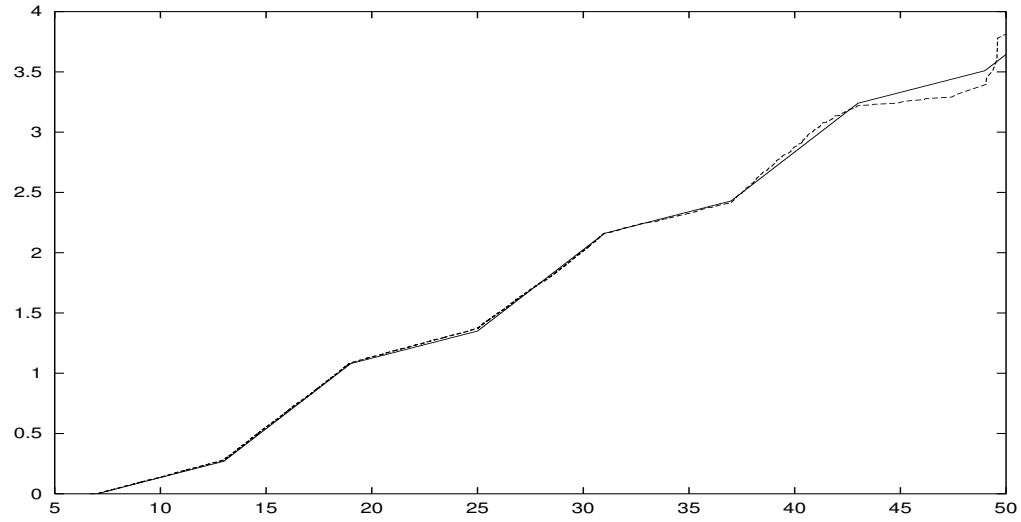


(a) Cumulative Capture Rate

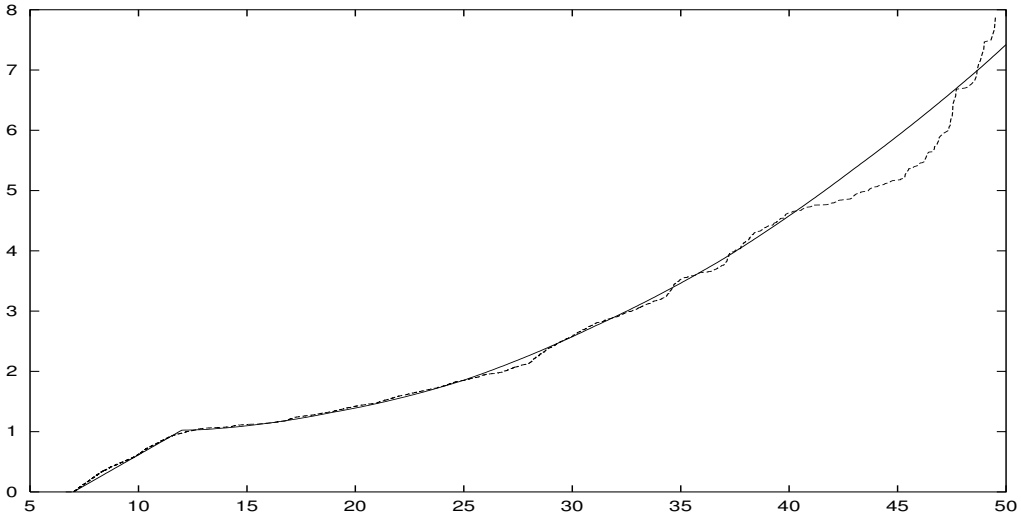


(b) Cumulative Mortality Rate

Figure 20: Cormack-Jolly-Seber estimates of the cumulative capture and mortality rate with respect to age in months. The true rate is given by the whole drawn line, while the estimated rate is given by the dotted line.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 21: EM-NA estimates of the integrated capture and mortality rate with respect to age in months. The true rate is given by the whole drawn line, while the estimate is given by the dotted line.

7 Applying the EM-NA algorithm to the Risør data

Let us apply the EM-NA algorithm to the Risør mark-recapture data set. Initially, we wish to implement a purely age dependent model, and thereby average out the time effects. This assumption is probably wrong for the capture rate, since the report probability is most likely time dependent. For the mortality rate, the assumption is plausible on the condition that the algae bloom did not affect the marked individuals. We have argumented against the direct effects of the algae bloom (see Section 2.1).

Note that the approach here is purely age dependent, as opposed to Julliard et al. (2001) which apply an age-time dependent model. That is, the time effects are averaged out.

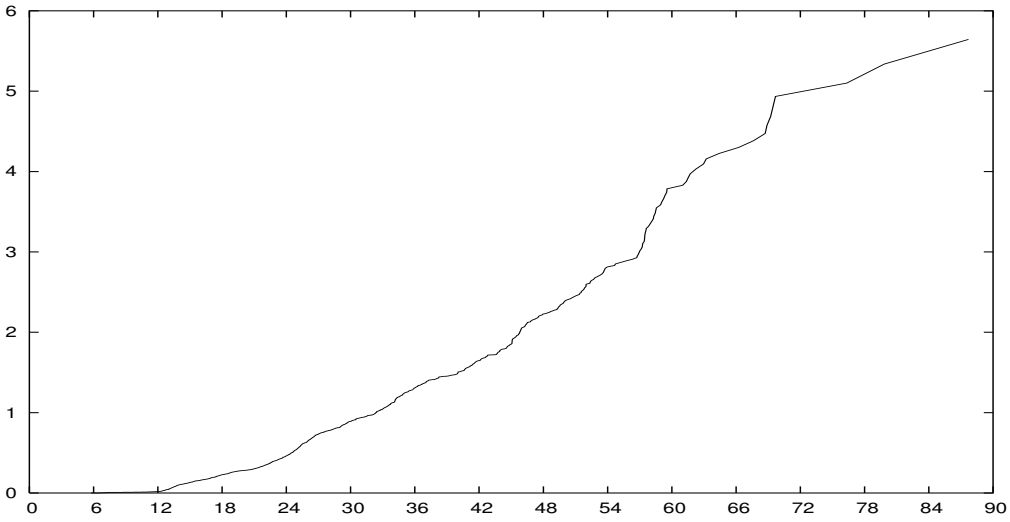
The age at release for the reared cohorts is set to 190 days. This is fairly correct, since they were hatched in the beginning of April. The age of the wild caught fish was determined to 620 days. This is in accordance with Julliard et al. (2001) which states that *"... the size distribution of wild fish caught in December appeared to be very close to the size distribution of one-year-old reared fish recaptured at that time"*.

The resulting EM-NA estimates of the cumulated capture and mortality rates are presented in Figure 22. We see that up to about one year of age, the cod is practically not captured. This means we have very little information to work with on this interval. After this age, the capture rate increases suddenly. This shift in capture rate forces the simulated death events to take place before the shift. With no simulated death events after the shift, the estimated mortality is zero up to the point where the wild-caught cods are released. After the release, the number of recaptures does not rise correspondingly to the number released, and it seems like a large part of the released wild-caught cod dies within four months. Note that this high mortality period works as a collection bin for the released reared cods that survived the first 6 months.

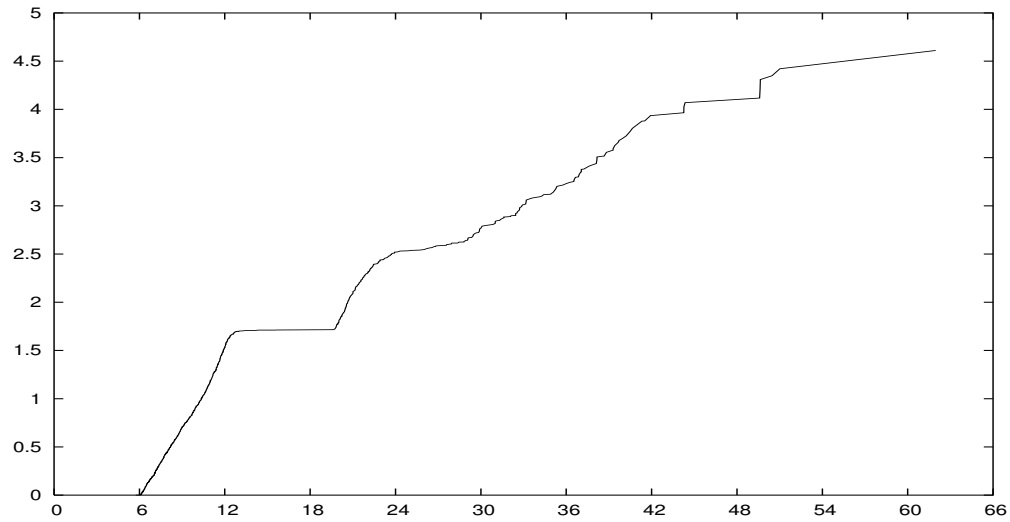
8 Including covariate information

So far, we have assumed a simple age-dependent model for the capture and mortality rates, and we have used only the age at release and recapture to estimate these rates. The Risør release-recapture data set, however, contains much more information, and more elaborate models should be considered.

There is a number of auxiliary information that would be interesting to include in an enhanced model. Time dependency was included in the Julliard et al. (2001) paper, and it could be interesting to see how this affects our estimates. We have information about the gear used at each capture. We know the length at release for all released individuals, and the length at recapture for the recaptured individuals. As we have seen in Chapter 3.2, a recaptured and re-released individual has seemingly a much higher probability of being captured again, compared to non-recaptured individuals of the same age. We begin by including the length at release.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 22: The EM-NA estimates of the cumulative capture and mortality rates with respect to age in months.

8.1 Including length at release

The length at release has generally a large impact on the recovery rate in fish mark-recapture experiments (see e.g. Svåsand and Kristiansen (1990)). It is assumed that this mainly is due to low fishing probability on small fish. From the Risør data set, we know that the fishing rate is negligible for fish smaller than 25 cm. We also notice that there are hardly any recaptures between release in October (reared fish) and April the following year, when the cod are approx. one year old.

Previously, we have assumed that all the reared cod were of the same age at release (reared), namely 6 months. This is a somewhat artificial and hard to control statement which probably is wrong for most of the reared cod. Is there another way of estimating the age at release for the reared cod at release? We know the length at release, and we could use this as an indication of physical development, i.e. age. That is, let us state that a large fish is older than a small fish. This implies that a small fish spends a longer period of time in the low fishing rate interval than a larger fish which seems reasonable. Julliard et al. (2001) estimated the growth rate to be 1.5 cm/month. The data imply that the catchable size is reached at approx. 12 months of age. We now state that the catchable size is reached at 12 months of age, and we compute the age at release from the length at release using the following relationship:

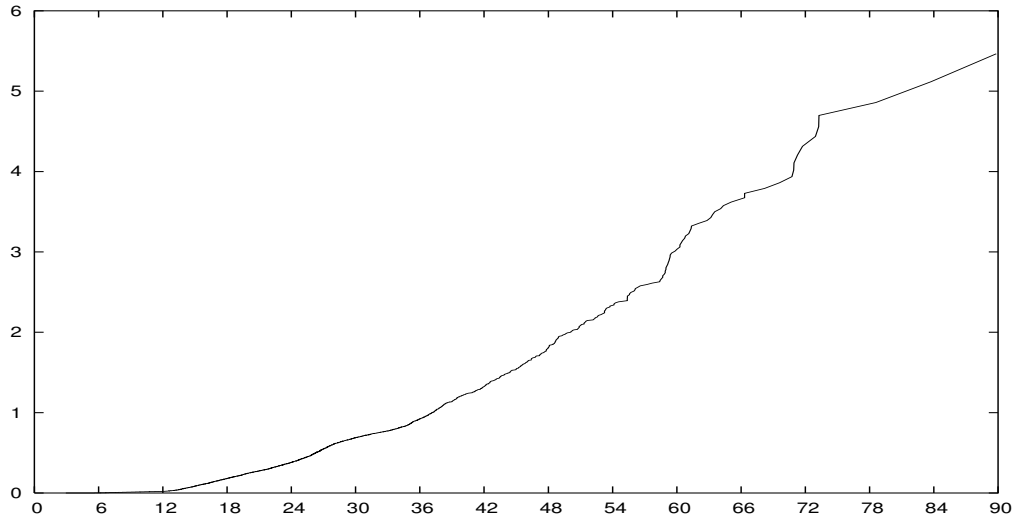
$$\text{age at release} = 12 - \frac{25 - \text{length at release}}{1.5}.$$

The wild-caught cod we have previously assumed to be 19 months (one and a half year) on release. This assumption is obviously wrong, and should be relaxed. On the basis of a length-age analysis, we have set a cut-off value of 40 cm at release. Wild-caught cod below this length we have assumed to be 19 months, those above this length we have assumed to be 31 months (two and a half year). A more elaborate model could be used to determine the age of the wild-caught cod, but we feel that this crude differentiation is sufficient for our purpose.

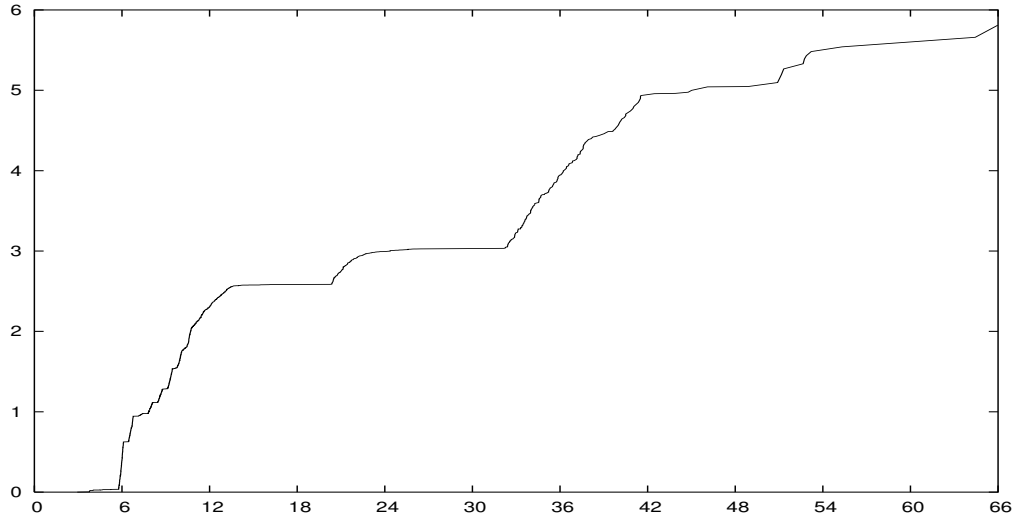
The EM-NA estimates for the cumulated capture and mortality rates are given in Fig. 23 We see that the capture rate is unchanged compared to the previous analysis, but the mortality rate estimate is quite changed. We notice that the cumulative mortality rate increases from 0 to 1 right before 6 months, and after this a steady increase occurs up until the age of 12 when it reaches catchable age. An estimated age of 6 months correspond to a release length of 16 cm. This indicates that an individual released with length below 16 cm has approx. $\exp\{-2.6\} = 0.074$ probability of surviving up to catchable age, while individuals above this length has estimated over $\exp\{-1.6\} = 0.20$ probability of surviving the low capability period, increasing with length. We see the two steps just after 19 months and 31 months, due to the segregation of the wild-caught cod.

8.2 Semi-parametric multiplicative hazard models

In the following analysis, let us use the release length information as described above, with semi-parametric multiplicative models for the capture and mortality rates. Again, the presentation and notation of Andersen et al. (1993) is followed.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 23: The EM-NA estimates of the length-adjusted cumulative capture and mortality rates with respect to age in months.

We consider models for the capture and mortality processes, given by the capture and mortality rates

$$\alpha_{pi}(t; \mathbf{Z}_{pi}(t)) = \alpha_p(t) \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_{pi}(t)\}$$

and

$$\alpha_{\phi i}(t; \mathbf{Z}_{\phi i}(t)) = \alpha_\phi(t) \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_{\phi i}(t)\}$$

where $\mathbf{Z}_{pi}(t)$ is the covariate vector for individual i on the capture process, and $\mathbf{Z}_{\phi i}(t)$ is the covariate vector for individual i on the mortality process. We notice that both covariate vectors may be time-dependent. The regression coefficient vectors $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_\phi$ are of the same dimension as $\mathbf{Z}_{pi}(t)$ and $\mathbf{Z}_{\phi i}(t)$ for all i. Note also that the ratio between the mortality and capture rates of two individuals depends only on the covariates for these individuals, and the regression coefficients. E.g. for individual 1 and 2,

$$\frac{\alpha_{p1}(t; \mathbf{Z}_{p1}(t))}{\alpha_{p2}(t; \mathbf{Z}_{p2}(t))} = \exp\{\boldsymbol{\beta}_p^\intercal (\mathbf{Z}_{p1}(t) - \mathbf{Z}_{p2}(t)\}.$$

For fixed values of $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_\phi$, the Nelson-Aalen estimates are given by

$$\widehat{A}_p(t; \boldsymbol{\beta}_p) = \sum_{\{i,j: T_{ji} \leq t\}} \frac{1}{S_p(t; \boldsymbol{\beta}_p)} \tag{5}$$

and

$$\widehat{A}_\phi(t; \boldsymbol{\beta}_\phi) = \sum_{\{i: U_i \leq t\}} \frac{1}{S_\phi(t; \boldsymbol{\beta}_\phi)} \tag{6}$$

where

$$S_p(t; \boldsymbol{\beta}_p) = \sum_{i=1}^{n} \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_{pi}(t)\} Y_i(t)$$

and

$$S_\phi(t; \boldsymbol{\beta}_\phi) = \sum_{i=1}^{n} \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_{\phi i}(t)\} Y_i(t).$$

We remember from earlier that T_{ji} is the j'th capture of individual i, and that U_i is the death time of individual i. Note that $S_p(t; \boldsymbol{\beta}_p)$ and $S_\phi(t; \boldsymbol{\beta}_\phi)$ may be viewed as weighted numbers at risk for the capture and mortality processes.

The Cox's partial likelihoods (Cox (1975)) of $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_\phi$ are given by

$$\mathrm{L}(\boldsymbol{\beta}_p) = \prod_{i,j} \frac{\exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_{pi}(t)\}}{S_p(T_{ji}; \boldsymbol{\beta}_p)} \tag{7}$$

and

$$\mathrm{L}(\boldsymbol{\beta}_\phi) = \prod_{i,j} \frac{\exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_{\phi i}(t)\}}{S_\phi(T_{ji}; \boldsymbol{\beta}_\phi)}. \tag{8}$$

We denote the values of $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_\phi$ which maximises (7) and (8) respectively as $\widehat{\boldsymbol{\beta}_p}$ and $\widehat{\boldsymbol{\beta}_\phi}$.

Estimates of $A_p(t;\boldsymbol{\beta}_p)$ and $A_\phi(t;\boldsymbol{\beta}_\phi)$ may now be given by $\widehat{A}_p(t;\widehat{\boldsymbol{\beta}_p})$ and $\widehat{A}_\phi(t;\widehat{\boldsymbol{\beta}_\phi})$ [cf. (5) and (6)]. These are known as Breslow estimators, and are nonparametric maximum likelihood estimators.

These estimators are now included in the NA-EM algorithm. Since the Breslow estimators are maximum likelihood estimators, this is straight forward.

Algorithm 4 The semi-parametric NA-EM algorithm

Initialise $\widehat{A}(t;\widehat{\boldsymbol{\beta}_p})$ and $\widehat{A}_\phi(t;\widehat{\boldsymbol{\beta}_\phi})$
Draw new death events with respect to $\widehat{A}(t;\widehat{\boldsymbol{\beta}_p})$ and $\widehat{A}_\phi(t;\widehat{\boldsymbol{\beta}_\phi})$
Estimate $\widehat{\boldsymbol{\beta}_p}$ and $\widehat{\boldsymbol{\beta}_\phi}$ by maximising the partial likelihoods $\mathrm{L}(\boldsymbol{\beta}_p)$ and $\mathrm{L}(\boldsymbol{\beta}_\phi)$
Estimate new $\widehat{A}(t;\widehat{\boldsymbol{\beta}_p})$ and $\widehat{A}_\phi(t;\widehat{\boldsymbol{\beta}_\phi})$ using the capture events and the drawn death events.
Repeat until convergence

8.3 Including re-release information

From the analysis in Chapter 3.2, we may presume that a re-released individual has a larger probability of recapture than an individual of the same age which has not been recaptured after the initial release. We would like to incorporate this presumption in the model, using a semi-parametric hazard model.

Let the covariate $Z_{pi}^{rel}(t)$ be such that

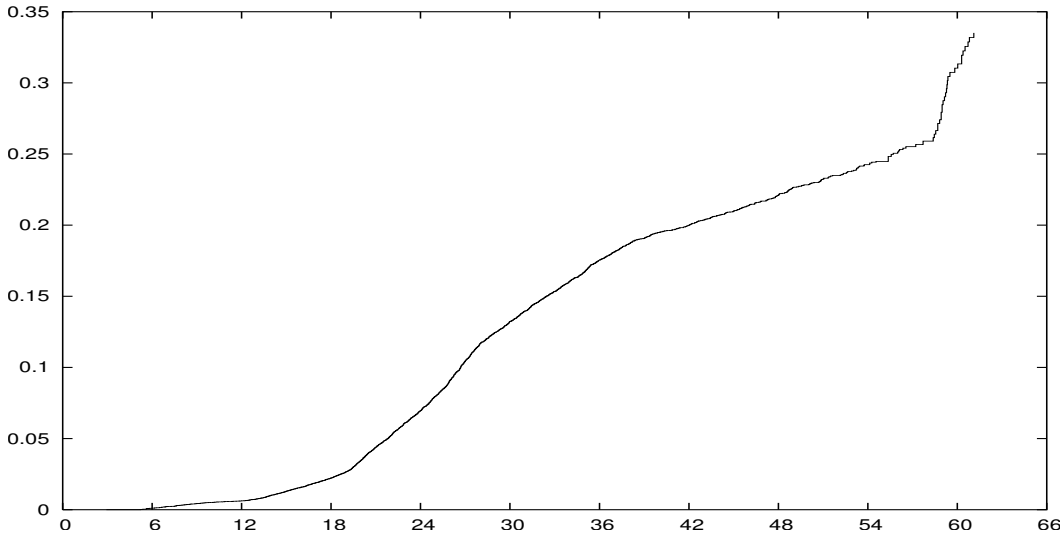
$$Z_{pi}^{rel}(t) = \begin{cases} 1 & \text{if individual } i \text{ has been re-released at time } t, \\ 0 & \text{else} \end{cases}$$

We consider a model with the covariate $Z_{pi}^{release}(t)$ included, in addition to the length at release adjustment on age described in Chapter 8.1. The capture rate for this case is then given by

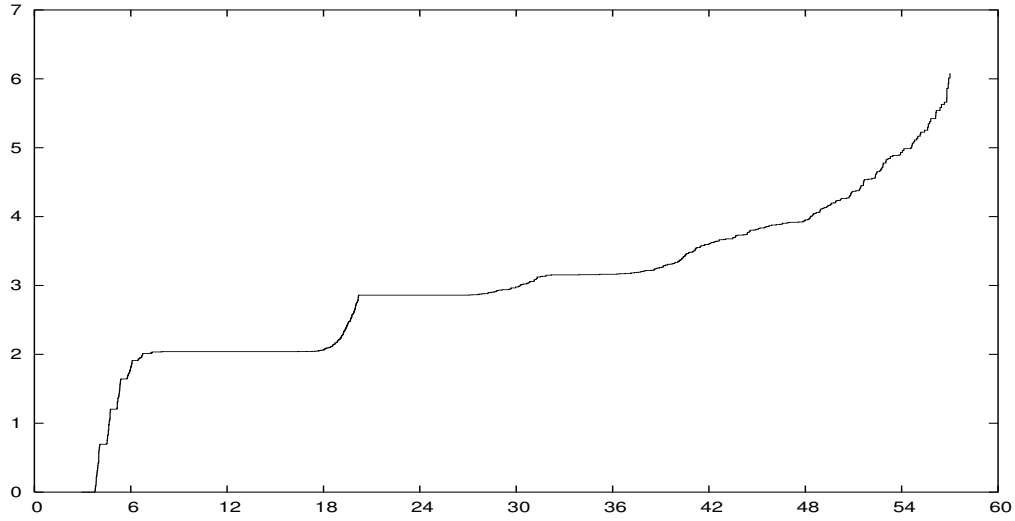
$$\alpha_{pi}(t) = \alpha_p(t)\exp\{\beta_{rel}Z_{pi}^{rel}(t)\}.$$

The results show that $\widehat{\beta}_{rel} = 1.95$ which indicates that the re-release of an individual greatly affects the recapture rate. The resulting baseline NA-EM estimates for the cumulative capture and mortality rates are given in Fig. 24

By introducing the covariate information on re-release, we see some major changes compared to the previous analysis. The capture rate increases by a factor of 7.02 after a recapture-release event. This affects both the capture and mortality rate. The baseline capture rate drops significantly compared to earlier analysis, and the mortality rate assumes a bathtub shape. Note that the baseline cumulative capture rate reaches a maximum of ca. 0.35. If we assume that the reporting probability is ca. 0.5, the true cumulative capture rate reaches 0.7. This means that a cod that otherwise would have survived five year, has a $1 - \exp\{-0.7\} = 0.5$ probability of being captured.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 24: EM-NA Estimates of the length-adjusted integrated capture and mortality rates with covariate information on re-release. The time scale is age in months.

8.4 Including time dependent covariates

Finally, an analysis with time dependent covariates is presented. For the Risør data, it is believed that the capture rate is time dependent. The reporting awareness among the fishermen is prone to change during the experiment period, and we would like to include this effect in the model. Let us assume that the capture rate changes from year to year, with the shifts placed at 1st of October 1987, 1988, 1989, 1990 and 1991. That is, the first interval is from 1st of October 1986 to 1st of October 1987, the second interval is from 1st of October 1987 to 1st of October 1988 and so on. The last interval spans from 1st of October 1991 and until the experiment end.

The re-release covariate as described in Chapter 8.3 is kept, and the resulting covariate for the time dependency is given by

$$\mathbf{Z}_{pi}^{time}(t) = \begin{cases} (1,0,0,0,0,0)' & \text{if } t \text{ corresponds to the first interval} \\ (0,1,0,0,0,0)' & \text{if } t \text{ corresponds to the second interval} \\ \vdots, \end{cases}$$

The resulting capture rate for individual i is then given by

$$\alpha_{pi}(t) = \alpha_p(t) \exp\{\beta_{rel} Z_{pi}^{rel} + \boldsymbol{\beta}_{time}^{\intercal} \mathbf{Z}_{pi}^{time}(t)\},$$

Note that

$$\boldsymbol{\beta}_{time} = (1, \beta_1^{time}, \beta_2^{time}, \beta_3^{time}, \beta_4^{time}, \beta_5^{time})'$$

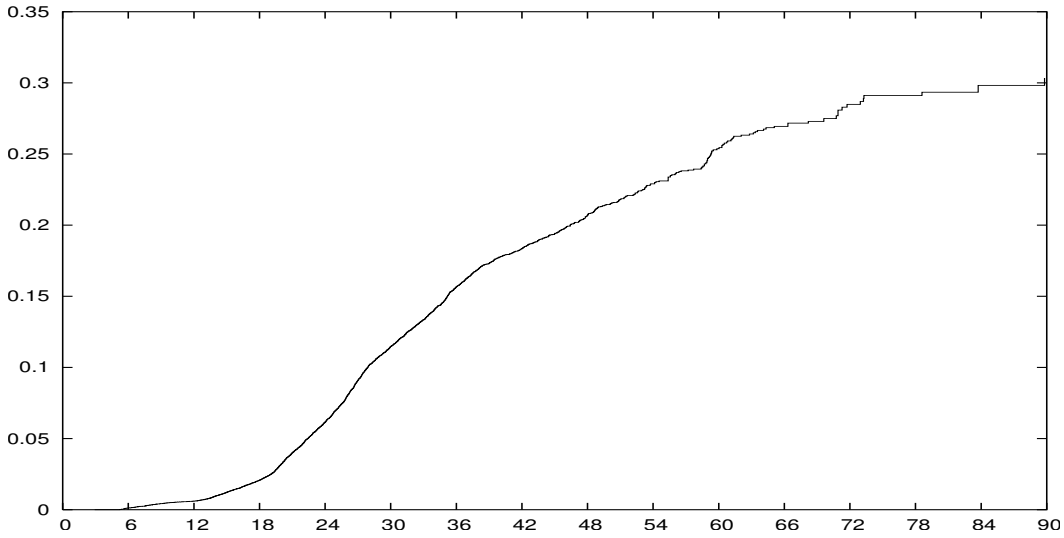
such that the baseline capture rate $\alpha_p(t)$ refers to a not re-released individual in the first interval.

The estimated regression coefficients read

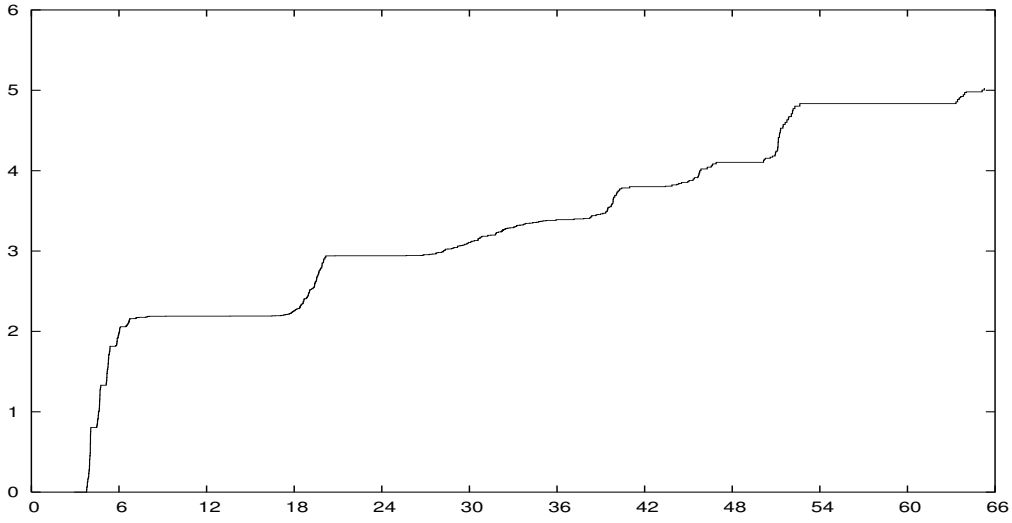
$$\begin{aligned} \widehat{\beta}_1^{time} &= -1.09 \\ \widehat{\beta}_2^{time} &= 0.35 \\ \widehat{\beta}_3^{time} &= 0.38 \\ \widehat{\beta}_4^{time} &= 0.07 \\ \widehat{\beta}_5^{time} &= -0.45 \\ \widehat{\beta}_{rel} &= 1.92 \end{aligned}$$

If we compare with the histogram of recaptures in Fig.2.2, we see that the estimated regression coefficient are reasonable, with few captures reported in the second period, many captures in the two next periods, and then lower capture rate in the two last periods.

The cumulative capture and mortality rates are given in Fig. 25. We see that they are comparable with the results from the analysis in the previous chapter, with some small changes. It seems like the introduction of time-dependent covariates does not affect the analysis greatly, except for a less steep curve towards the end for both the cumulative capture and mortality rate.



(a) Cumulative Capture Rate



(b) Cumulative Mortality Rate

Figure 25: EM-NA estimates of the length-adjusted cumulative capture and mortality rates with covariate information on re-release and time-dependency

(a) Yearly mortality rate

Age class	CJS	EM-NA 1	EM-NA 2	EM-NA 3	EM-NA 4
0-age class	3.60	3.38	4.05	2.97	3.19
1-age class	0.30	0.76	0.51	0.82	0.75
2-age class	0.55	0.67	1.13	0.30	0.43
$\geq$ 3-age class	0.58	0.64	0.69	1.65	0.66

(b) Yearly capture rate

Age class	CJS	EM-NA 1	EM-NA 2	EM-NA 3	EM-NA 4
0-age class	0.015	0.032	0.035	0.003	0.003
1-age class	0.25	0.39	0.35	0.05	0.06
2-age class	0.51	0.81	0.54	0.11	0.10
$\geq$ 3-age class	0.51	0.79	1.01	0.05	0.05

Table 6: The average yearly mortality and capture rates for the Cormack-Jolly-Seber analysis presented in Julliard et al. (2001) compared to our EM-NA estimates.
EM-NA 1: The basic EM-NA estimate as presented in Fig. 22
EM-NA 2: The EM-NA estimate with length-adjusted age at release as presented in Fig. 23
EM-NA 3: The EM-NA estimate with length-adjusted age at release and covariate information on re-release as presented in Fig. 24
EM-NA 4: The EM-NA estimate with length-adjusted age at release and covariate information on re-release and time-dependency as presented in Fig. 25

8.5 Comparison with Julliard et al. (2001)

We have now analysed the Risør mark-recapture experiment data using the EM-NA estimator with different covariate information. The cumulative capture and mortality rates have been the natural method of presenting the results, but especially the cumulative mortality rate estimate behaves unexpected with more or less justifiable jumps. In the literature, yearly rates or probabilities are more common. The estimated yearly rates for this analysis are presented together with the results from Julliard et al. (2001) in Table 6.

When we compare the CJS estimator with the EM-NA estimator, remember that the CJS estimator is included an algae bloom effect. This means that some of the mortality of the 1-age, 2-age and 3-age classes are incorporated in the algae bloom parameter. We have argued against the inclusion of this effect (see Section 2.1), and the mortality estimates of the CJS estimator may therefore be biased.

For the average yearly mortality rates (Table 6a), we see that it is the first continuous model estimate (EM-NA 1) that corresponds most to the discrete model estimate. This is plausible, since this is the model that is most similar to the discrete model (no length adjustment or re-release covariate information).

Maybe the most interesting result of our analysis is the major drop in recapture rate when the covariate information on re-release is included. The EM-NA 3 and

EM-NA 4 capture rate estimates are roughly one fifth of the CJS estimate, and from the analysis we know that the capture rate of a re-released individual is 7.2 times the capture rate of an individual that is not re-released. The interpretation of the re-released capture rate is the capture rate of a fish that we know is released in a capture prone area. The basic capture rate is the capture rate of a fish that is released randomly in the fjord. Since we may assume that a larger part of the population resides at areas with high fishing pressure, we may assume that the true capture rate lies somewhere between these capture rates. Be careful to interpret the capture rate as fishing rate. Remember that a very large part of the recaptures were done by eel pots, and that these captures were generally released again.

Note the drop in capture rate between the 2-age class and the $\geq$ 3-age class for the EM-NA 3 and EM-NA 4 estimates. This may be explained by the size selection of the eel traps. The older fish are not captured by the eel fishermen, and the re-release effect is possibly lost. The eel trap recaptures represent one third of the total recaptures, and a size selection of these recaptures should influence the data.

Since the EM-NA 4 mortality estimator includes more information than the other estimators, we trust these results the most. We notice that we have a high mortality rate for the 0-age class, followed by a considerably lower mortality rate for the 1-age class. The decline continues for the 2-age class, while for the 3-age class and onward, there is an increase in mortality rate. This fits with the perception that the natural capture rate is high during the first year and then declining, while the fishing rate is small for the first few years, and then increasing.

Note that the 1-age class mortality estimates are generally higher for the CJS-estimates than for the EM-NA estimates. This could be explained by the algae bloom effect, which we have not included in any of our models.

Let us review some of the results from the Julliard et al. (2001) paper in lieu of the analysis presented here. The "Synopsis of findings" from the discussion chapter in Julliard et al. (2001) is listed, together with a comment based on our analysis afterwards.

1. *There was a low disappearance probability immediately after release for reared 6-mo-old fish, whereas* $\leq 60\%$ *of older fish disappeared just after tagging.*

   Since the information during the first 6 months (123 captures) is very sparse, it is hard to say how much of the estimated mortality correspond to natural mortality, and how much correspond to disappearing (tag loss, tagging mortality and emigration). The disappearing just after release for the of the wild-caught cod correspond to the steep region around 20 months of the estimated cumulative mortality (see Figure 25b). This region also serves as a collection bin for the reared cod that survived the initial high mortality period, and thus the disappearance probability of the wild caught cods is most likely much less than 60%.

2. *The tag return was estimated to be* $\sim 50 - 60\%$.

   It is not possible from our analysis to confirm this probability without making strong assumptions.

3. *The cumulative recovery rate was clearly dependent on the size of the fish at release.*

   We found a clear difference between the analysis with and without length-adjusted age at release, so there is support for this statement in our analysis.

4. *A high natural-mortality rate was found during 6-12 mo of age (i.e., first 6 mo after release).*

   This statement is supported by our analysis.

5. *The natural mortality of the older age classes was low; most of the mortality was then due to fishing.*

   It is hard to support this result as it is based on several weakly founded assumptions. The assumption that the tag-return probability is 50%-60% is reasonable, but not trustworthy. The capture rate is considered to represent the fishing mortality rate, although a large portion of the captures are done by eel trappers which release the fish again, and should not be included in the total mortality rate. If we were forced to make a statement, it would be that natural mortality is prevailing up to and including the 2-age class. After this, the fishing mortality is most prominent.

6. *The seasonal pattern of fishing appeared to vary greatly between age classes.*

   This issue is not covered by our analysis.

7. *During several months after the algae bloom of 1988, fish disappeared at a high rate, whatever their age. No emigration was detectable in relation to the algae bloom.*

   Our analysis, in addition to the analysis of Chan et al. (2003) does not indicate that the mortality rate was increased during the algae bloom. Note that a high mortality rate due to the algae bloom may be confounded with a cohort effect on the low recovery rate for the R86 and R87 cohorts.

9 Discussion

In this paper, a new, semi-parametric model for analysing continuous mark-recapture experiment data is presented. To estimate the parameters of this model, the EM algorithm is applied. The approach represents an improvement compared to previous approaches, in that the model accounts for the continuous nature of the data. In addition, the semi-parametric model is less parsimonious than previous models, and almost no model selection is necessary. The model is general, and may be used on any mark-recapture experiment data, also those with discrete capture occasions. The main benefit of this approach is that it connects the analysis of capture-recapture data with standard counting processes, and in particular survival analysis. Survival analysis has received substantial attention the last 20-30 years, and much work has been done in this field. As stated by one of the recent reference articles in the field of capture-recapture, Lebreton et al. (1992):

> ..., Cox's proportional hazard model (see, e.g., Cox and Oakes 1984) allows a straightforward capture-recapture generalisation. This link with models used in human biology will probably be of benefit to both fields (epidemiology and ecology).

We see that the hereby established link has been long anticipated.

The NA-EM algorithm is not bound to mark-recapture data only. Any survival data set with unknown censoring times of this kind may be analysed using our approach.

So far, it has not been possible to incorporate variance estimates, goodness-of-fit tests or any other of the tools available to us through the huge amount of literature on counting processes. This is not straightforward, as we are not dealing solely with true events, but mainly with simulated events. The estimation of variance for either of the capture or mortality rates depends on the other, and no method has yet been devised to estimate the variances simultaneous.

We have analysed a mark-recapture experiment in the Risør area of the Norwegian Skagerrak coast. Compared to previous analysis on the data, more covariate information is included such as the size at release and if the individual has been re-released. In addition, the continuous nature of the data is taken into consideration.

The results of the analysis are ambiguous, and several central questions such as the ratio between fishing and natural mortality for the different age groups, remain open. The central findings are summarised below:

- The capture rate is about seven times as high for a re-released individual compared to an individual that has never been recaptured after the initial release. This may be explained by the immobile nature of the coastal cod, together with the agreement made with the eel fishermen to register and release marked individuals.

- The size at release is important to the survival up to catchable size (25cm). We have seen that fish larger than 16 cm at release has three times higher probability of surviving until catchable size as fish smaller than 16 cm. This may explain the low recapture rate of the R86 and R87 cohorts.

- The total mortality seems to be very high for cod aged 6-12 months. During this period, the mortality due to fishing is presumably low. During the next year the mortality is significantly lower, while the 2-age class experiences the lowest mortality rate. After this, the mortality rate increases again. An interpretation may be that the natural mortality steadily decreases during as the cod grows older, while the fishing mortality increases. The capture rate may be misleading as an estimate of the fishing mortality rate because of the tendency by the eel fishermen to release small cod.

References

Aalen, O. (1975). *Statistical inference for a family of counting processes.* PhD thesis, University of California, Berkeley.

Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.

Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments. *Mathematical Bioscience*, 6:1–11.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag.

Botsford, L. W., Castilla, J. C., and Peterson, C. H. (1997). The management of fisheries and marine ecosystems. *Science*, 277:509–15.

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, 15(4):357–76.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). *Statistical inference from band-recovery data: a handbook*. Fish and wildlife service, U.S. Department of the Interior, second edition.

Catchpole, E. A., Freeman, S. N., Morgan, B. J. T., and Harris, M. P. (1998). Integrated recovery/recapture data analysis. *Biometrics*, 54(1):33–46.

Chan, K.-S., Stenseth, N. C., Lekve, K., and Gjøsæter, J. (2003). Modeling pulse disturbance impact on cod population dynamics: The 1988 algal bloom of Skagerrak Norway. *Ecological Monographs*.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62.

Danielssen, D. S. and Gjøsæter, J. (1994). Release of 0-group cod, *Gadus morhua* L., on the southern coast of norway in the years 1986-1989. *Aquaculture and Fisheries Management*, 25:129–42.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistics Society Series B*, 39:1–38.

Fromentin, J.-M., Myers, R. A., Bjørnstad, O. N., Stenseth, N. C., Gjøsæter, J., and Christie, H. (2001). Effects of density-dependent and stochastic processes on the regulation of cod populations. *Ecology*, 82:567–79.

Hilborn, R., Branch, T. A., Ernst, B., Magnusson, A., Minte-Vera, C. V., Scheuerell, M. D., and Valero, J. L. (2003). State of the world's fisheries. *Annual Review of Environment and Resources*, 28:359–99.

Julliard, R., Stenseth, N. C., Gjøsæter, J., Lekve, K., Fromentin, J.-M., and Danielssen, D. S. (2001). Natural mortality and fishing mortality in a costal cod population: a release-recapture experiment. *Ecological Applications*, 11(2).

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Liao, I. and Leaño, E. M. (2003). Status of research in stock enhancement and sea ranching. *Reviews in Fish Biology and Fisheries*, 13:151–63.

Ludwig, D., Hilborn, R., and Walters, C. (1993). Uncertainty, resource exploitation, and conservation: Lessens from history. *Science*, 260:17.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1:27–52.

Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14:945–965.

Smith, T. D. (1994). *Scaling Fisheries*. Cambridge University Press, Cambridge.

SOFIA (2002). *The State of World Fisheries and Aquaculture*, chapter 1. FAO Fisheries Department.

Svåsand, T. and Kristiansen, T. S. (1990). Enhancement studies of costal cod in western norway, part iv. mortality of reared cod after release. *Journal du Conseil International pour l'Exploration de la Mer*, 47:30–39.

Svåsand, T., Kristiansen, T. S., Pedersen, T., Salvanes, A. G. V., Engelsen, R., Nævdal, G., and Nødtvedt, M. (2000). The enhancement of cod stocks. *Fish and Fisheries*, 1:173–205.

Tveite, S. (1971). Fluctuations in year-class strength of cod and pollack in southeastern norwegian coastal waters during 1920-69. *Fiskeridirektoratets Skrifter Serie Havundersøkelser*, 16:65–76.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704.

White, G. C. and Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study 46 Supplement*, pages 120–38.

Øyestad, V., Pedersen, T., and Folkvord, A. (1985). Mass production of atlantic cod juveniles (*Gadus morhua*) in a norwegian saltwater pond. *Transactions of the American Fisheries Society*, 114:590–95.

Analysing Continuous Mark-Recapture Data on Open Populations

Inge Christoffer Olsen

Abstract

This part of the thesis presents a fully analytical EM algorithm for the analysis of mark-recapture models. The E step of the algorithm is analytical, and thus presents an improvement on the Monte Carlo EM algorithm of Part II. Nice criteria of convergence are presented, and it is showed that the covariance matrix is quite easily estimated. The Cormack-Jolly-Seber (CJS) model has been used extensively on discrete mark-recapture data. We see in this paper that the CJS model may be used directly on continuous mark-recapture data, much in the same way as the Kaplan-Meier estimator is used for survival data. To include covariate information, semi-parametric multiplicative hazard models are defined for the capture and mortality process. The EM algorithm is still used for the analysis. The developed methodology is then applied on a set of mark-recapture data from a population of European Dippers in southern France.

1 Introduction

In Part II of this thesis, a Monte Carlo EM algorithm was used to analyse continuous mark-recapture data. It was showed that the estimator worked well for simulated data, and that it produced nice interpretable results for the Risør mark-recapture data. There are some drawbacks to this method, though. Firstly, there is a problem with convergence. Because it uses simulated mortality events, we do not reach convergence in the normal sense. This could be solved by using the same random numbers at each iteration. Second, we have not found any method to estimate the variance.

With these considerations in mind, a new estimation method is presented, where the EM algorithm is applied directly. With this new method, a convergence criterion is available, and variance estimation becomes a manageable task. Note that in the following discussion, the notation alternates between a general EM notation and a specific mark-recapture notation. The difference will be clear from the context.

2 The EM Algorithm

To begin with, the main idea and notation of the general EM algorithm as defined by Dempster et al. (1977) is presented. Let $\mathbf{y}$ be the original data, $\mathbf{x}$ be the data

augmented with the missing data, and $\boldsymbol{\theta}$ be the parameters in the model. Using standard EM-algorithm notation,

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0) = \mathbb{E}_{\boldsymbol{\theta}_0}[\log L(\boldsymbol{\theta}|\mathbf{x})|\boldsymbol{\theta}_0, \mathbf{y}], \tag{1}$$

where the expectation is with respect to $f(\mathbf{x}|\mathbf{y}; \boldsymbol{\theta}_0)$, the conditional distribution of the missing data given the observed data. An EM iteration $\boldsymbol{\theta}^{(k)} \rightarrow \boldsymbol{\theta}^{(k+1)}$ is now given by

E step Compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$

M step Choose $\boldsymbol{\theta}^{(k+1)}$ to be a value of $\boldsymbol{\theta}$ which maximises $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.

These steps are repeated until a stopping criterion has been met. A stopping criterion is possibly that

$$\left(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\right)^{\intercal} \left(\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(k)}\right)$$

drops below a certain level. This scheme increases the likelihood at each iteration.

2.1 Variance estimation by the observed information matrix

An usual way to estimate the covariance matrix of the maximum likelihood estimator, is to assume asymptotic normality and then invert the observed information matrix (the negative Hessian matrix of the log-likelihood). By using a numerical optimiser, the Hessian is often computed as a by-product. For the EM algorithm, this is not the case. There exists, however, several methods for estimation of the Hessian matrix when the EM algorithm is used. Two methods are implemented here: the SEM algorithm introduced by Meng and Rubin (1991) and numerical differentiation of the Fisher score vector (considered by Jamshidian and Jennrich (2000)). The latter method is denoted the NDS approach. Both methods work well when the number of capture occasions are small (<10), but when the capture occasion number rise, only the NDS approach remain stable (according to our experience). Because the NDS approach is easy to implement and seems stable, this is the preferred method and the one described here.

In the following, the general notation of Jamshidian and Jennrich (2000) is used. Let $\mathbb{S}[\boldsymbol{\theta}|\mathbf{x}]$ and $\mathbb{H}[\boldsymbol{\theta}|\mathbf{x}]$ denote the gradient and the Hessian of the log-likelihood. Note that $\mathbb{S}[\boldsymbol{\theta}|\mathbf{x}]$ is also known as the score function of $\boldsymbol{\theta}$. It can be shown quite easily (see Dempster et al. (1977)) that

$$\mathbb{S}[\boldsymbol{\theta}|\mathbf{x}] = \left.\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{2}$$

To obtain an estimate of $\mathbb{H}[\boldsymbol{\theta}|\mathbf{x}]$, the score function $\mathbb{S}[\boldsymbol{\theta}|\mathbf{x}]$ is numerically differentiated by a first-order Richardson extrapolation of the central difference, as described by Jamshidian and Jennrich (2000).

In addition, still following Jamshidian and Jennrich (2000), the precision of the covariance estimate is computed by regarding the skewness of the estimate. Let $C = (\widehat{V} + \widehat{V}^{\intercal})/2$ and $K = (\widehat{V} - \widehat{V}^{\intercal})/2$ denote the symmetric and skew symmetric parts

t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
R_1	$m_{1,2}$	$m_{1,3}$	$\cdots$	$m_{1,k-1}$	$m_{1,k}$
	R_2	$m_{2,3}$	$\cdots$	$m_{2,k-1}$	$m_{2,k}$
		R_3	$\cdots$	$m_{3,k-1}$	$m_{3,k}$
			$\ddots$	$\cdots$	$\vdots$
				R_{k-1}	$m_{k-1,k}$

Table 1: Table of mark-recapture data. $R_i,\ i = 1, \ldots, k-1$ are the number of marked and released individuals at time t_i, and $m_{i,j},\ i = 1, \ldots, k-1, j = i, \ldots k$ are the marked individuals that are released at capture occasion t_i and not recaptured until capture occasion t_j.

of the covariance matrix estimate $\widehat{V}$. Then a measure of the algorithm performance is given by the precision estimate

$$\widehat{\mathrm{PRE}}(\widehat{V}) = -\log_{10} \|C^{-1/2} K C^{-1/2}\|$$

where $\|A\|$ denotes the spectral norm of A. Roughly, the precision measures the number of leading digits of the estimate covariance matrix that agree with the true covariance matrix.

2.2 Variance estimation by the bootstrap

The implementation of the bootstrap to the EM algorithm is easy in its simplest form. Let $x_1, \ldots x_n$ be the i.i.d. observed data. Then the procedure would be as follows:

1. Calculate the maximum likelihood estimate $\widehat{\boldsymbol{\theta}}$ using the EM algorithm.
2. Sample pseudo-data $X_1^*, \ldots, X_n^*$ with replacement from the observed data. Compute the maximum likelihood estimate of $\boldsymbol{\theta}$ with respect to the pseudo-data using the same EM algorithm as in step 1.
3. Repeat step 2 until enough bootstrap estimates of $\boldsymbol{\theta}$ have been reached. Use the bootstrap estimates to estimate the covariance matrix of the maximum likelihood estimator.

For a thorough introduction to the bootstrap, refer to Efron and Tibshirani (1993).

3 EM algorithm estimation

Let us begin by analysing the standard discrete capture-recapture experiment using the EM algorithm. The data of such experiments are usually presented as in Table 1. For such data, a time-dependent Cormack-Jolly-Seber (CJS) model with survival probabilities $\boldsymbol{\phi} = (\phi_1, \ldots \phi_{k-1})$ and capture probabilities $\mathbf{p} = (p_2, \ldots p_k)$ may be fitted. Here, ϕ_i is the probability of surviving the interval (t_i, t_{i+1}), and p_i

is the probability of being captured at time t_i. The likelihood for these parameters with respect to the data presented in Table 1 is quite easy to define and maximum likelihood estimates may be computed using numerical optimisation. Variance is estimated by the inverse observed information matrix.

We now want to establish an EM algorithm to compute the maximum likelihood estimates. Let us begin by augmenting the data with the mortality data. That is, we assume that we know the number of individuals that did not survive a given interval. The augmented data matrix with notation is then given in Table 2 The

t_1		t_2		t_3	$\cdots$	t_{k-1}		t_k	
R_1	$n_{1,1}$	$m_{1,2}$	$n_{1,2}$	$m_{1,3}$	$\cdots$	$m_{1,k-1}$	$n_{1,k-1}$	$m_{1,k}$	$n_{1,k}$
		R_2	$n_{2,2}$	$m_{2,3}$	$\cdots$	$m_{2,k-1}$	$n_{2,k-1}$	$m_{2,k}$	$n_{2,k}$
				R_3	$\cdots$	$m_{3,k-1}$	$n_{3,k-1}$	$m_{3,k}$	$n_{3,k}$
					$\ddots$	$\vdots$	$\vdots$	$\vdots$	
						R_{k-1}	$n_{k-1,k-1}$	$m_{k-1,k}$	$n_{k-1,k}$

Table 2: Table of augmented capture-recapture data. Here, $n_{i,j}$ is the number of individuals released at time i which did not survive the interval (t_j, t_{j+1}). Note that $n_{i,k}$ is the number of individuals released at time i who survived the whole experiment. The remaining notation is as described for Table 1

.

augmented data is then rearranged, and some new notation is introduced such that

$$\begin{aligned} d_i^\phi &= \sum_{j=1}^{i} n_{j,i}, \quad i = 1, \ldots, k \\ d_i^p &= \sum_{j=1}^{i-1} m_{j,i+1}, \quad i = 2, \ldots, k \\ r_i^\phi &= \sum_{j=1}^{i} R_i - \sum_{j=1}^{i-1} d_i^\phi - \sum_{j=2}^{i} d_i^p, \quad i = 1, \ldots, k \\ r_i^p &= r_i^\phi - d_i^\phi, \quad i = 1, \ldots, k-1 \end{aligned} \tag{3}$$

where r_i^ϕ is the number of individuals at risk of dying during the interval (t_i, t_{i+1}), d_i^ϕ is the total number of individuals that did not survive the interval (t_i, t_{i+1}), r_i^p is the number of individuals at risk of being captured at capture occasion t_{i+1} and d_i^p is the total number of individuals that was captured on capture occasion t_i. The rearranged, augmented data are presented chronologically in Table 3

Using a sequence of conditional binomial distributions, the likelihood given the augmented data may be expressed as

$$L(\boldsymbol{\phi}, \mathbf{p}) = \prod_{i=1}^{k-1} (1-\phi_i)^{d_i^\phi} \phi_i^{(r_i^\phi - d_i^\phi)} \cdot \prod_{i=2}^{k} p_i^{d_i^p} (1-p_i)^{(r_{i-1}^p - d_i^p)}, \tag{4}$$

t_1		t_2		t_3	$\cdots$	t_{k-1}		t_k	
r_1^ϕ	d_1^ϕ	r_2^ϕ	d_2^ϕ	r_3^ϕ	$\cdots$	r_{k-1}^ϕ	d_{k-1}^ϕ	r_k^ϕ	d_k^ϕ
	r_1^p	d_2^p	r_2^p	d_3^p	$\cdots$	d_{k-1}^p	r_{k-1}^p	d_k^p	

Table 3: Table of the rearranged, augmented data.

and the log-likelihood is given by

$$l(\boldsymbol{\phi}, \mathbf{p}) = \sum_{i=1}^{k-1} d_i^\phi \log(1-\phi_i) + (r_i^\phi - d_i^\phi)\log(\phi_i) + \sum_{i=2}^{k} d_i^p \log(p_i) + (r_{i-1}^p - d_i^p)\log(1-p_i). \quad (5)$$

The maximum likelihood estimates of the capture and mortality probabilities are then given by

$$\begin{aligned} \widehat{\phi}_i &= \frac{r_i^\phi - d_i^\phi}{r_i^\phi}, \quad i = 1, \ldots, k-1 \\ \widehat{p}_i &= \frac{d_i^p}{r_{i-1}^p}, \quad i = 2, \ldots, k \end{aligned} \quad (6)$$

We notice that these estimates are reasonable. The estimated survival probability is the number that survived the interval divided by the number at risk of dying just before the interval. The estimated capture probability is the number of captures divided by the number at risk of being captured. We would now like to use these identities in an EM algorithm.

For the E step of the EM algorithm, we need to take the expectation of the log-likelihood of the augmented data, conditional on the observed recapture data.

$$\begin{aligned} Q(\boldsymbol{\phi}, \mathbf{p}|\boldsymbol{\phi}_0, \mathbf{p}_0) &= \mathbb{E}[l(\boldsymbol{\phi}, \mathbf{p}|\mathbf{m}, \mathbf{n})|\boldsymbol{\phi}_0, \mathbf{p}_0, \mathbf{m}] \\ &= \sum_{i=1}^{k-1} \mathbb{E}[d_i^\phi]\log(1-\phi_i) + (\mathbb{E}[r_i^\phi] - \mathbb{E}[d_i^\phi])\log(\phi_i) + \\ &\quad + \sum_{i=2}^{k} d_i^p \log(p_i) + (\mathbb{E}[r_{i-1}^p] - d_i^p)\log(1-p_i). \end{aligned} \quad (7)$$

where all the expectations are conditional on $\boldsymbol{\phi}_0, \mathbf{p}_0$ and $\mathbf{m}$. When we compare with the expressions in 3, we see that we only need to compute $\mathbb{E}[n_{i,j}], i = 1, \ldots k-1,\ j = i, \ldots, k$ to be able to compute the expectation terms above. It is easy to see that each line in the augmented data matrix (Table 2) is multinomial distributed with cell probabilities

$$(1-\phi_i, \phi_i p_{i+1}, \phi_i(1-p_{i+1})(1-\phi_{i+1}), \phi_i(1-p_{i+1})\phi_{i+1}p_{i+2}, \ldots)$$

and trial number R_i. The cell probabilities are simply: the probability of dying in the first interval after release; the probability of surviving the first interval and being

captured on the first capture occasion; the probability of surviving the first interval, not being captured on the first capture occasion and then dying during the second interval; and so on.

The conditional joint distribution of the unknown $n_{i,i}, n_{i,i+1}, \dots n_{i,k}$, given the capture data $m_{i,i+1}, m_{i,i+2}, \dots m_{i,k}$ is also multinomial with parameters

$$\left(R_i - m_i, \frac{1-\phi_i}{C_i}, \frac{\phi_i(1-p_{i+1})(1-\phi_{i+1})}{C_i}, \dots, \prod_{j=i}^{k-1} \phi_j(1-p_{j+1}) \cdot \frac{(1-\phi_k)}{C_i} \right),$$

where

$$m_i = \sum_{j=i+1}^{k} m_{i,j}$$

and C_i is the normalising constant. The expected number of deceased in each interval is then given by the expectation in the multinomial distribution,

$$\begin{aligned} \mathbb{E}[n_{i,i}] &= (R_i - m_i)\frac{(1-\phi_i)}{C_i} \\ \mathbb{E}[n_{i,j}] &= (R_i - m_i)\prod_{k=i}^{j-1} \phi_k(1-p_{k+1}) \cdot \frac{(1-\phi_j)}{C_i} \quad j = i+1 \dots k. \end{aligned} \tag{8}$$

Following this pattern, the expected number of deaths at each interval is easily computed, and the expectation of the log-likelihood is given. The M step is then given straight forward by

$$\begin{aligned} \phi_i^{(k+1)} &= \frac{\mathbb{E}[r_i^{\phi}] - \mathbb{E}[d_i^{\phi}]}{\mathbb{E}[r_i^{\phi}]}, \quad i = 1, \dots, k-1 \\ p_i^{(k+1)} &= \frac{d_i^{p}}{\mathbb{E}[r_{i-1}^{p}]}, \quad i = 2, \dots, k \end{aligned} \tag{9}$$

The resulting EM algorithm is then as follows:

E step Compute the expected mortality data given the observed capture data and the current capture and survival parameters $\boldsymbol{\phi}^{(k)}$ and $\mathbf{p}^{(k)}$ using the expressions in Eq. (8)

M step Set $\boldsymbol{\phi}^{(k+1)}$ and $\mathbf{p}^{(k+1)}$ according to Eq. (9).

Repeat these steps until a stopping criterion has been met.

A classic result of the CJS model is that the parameters p_k and ϕ_{k-1} may only be estimated as the product $\phi_{k-1}p_k$. By the EM algorithm, p_k is estimated to one, and ϕ_{k-1} is estimated to the proportion of the remaining individuals not captured at the last capture occasion.

3.1 Variance

To compute the score function, we see that the log-likelihood is given by equation (5), and the element of the score vector corresponding to ϕ_i and p_i is then according to (2) given by

$$\mathbb{S}_{\phi_i}[\boldsymbol{\phi}, \mathbf{p}] = \frac{\mathbb{E}[d_i^{\phi}]}{1-\phi_i} - \frac{\mathbb{E}[r_i^{\phi}] - \mathbb{E}[d_i^{\phi}]}{\phi_i}$$

and

$$\mathbb{S}_{p_i}[\boldsymbol{\phi}, \mathbf{p}] = \frac{\mathbb{E}[d_i^{p}]}{p_i} - \frac{\mathbb{E}[r_{i-1}^{p}] - d_i^{p}}{1-p_i}$$

respectively. The expectations is with respect to $\boldsymbol{\phi}, \mathbf{p}$ and the data $\mathbf{m}$. To estimate the Hessian matrix $\mathbb{H}[\boldsymbol{\phi}, \mathbf{p}]$, we need to differentiate the score functions numerically. The first-order Richardson extrapolation of the central difference is used for the differentiation. This is given by

$$\frac{f(x-2h) - 8f(x-h) + 8f(x+h) - f(x+2h)}{12h}$$

where h is a small positive number. the negative of this is then inverted to compute the estimated covariance matrix.

Note that the EM algorithm produces the same estimates and covariance estimates as the traditional likelihood maximisation algorithm. A nice bi-product is however that we get the number at risk for each capture and mortality period.

3.2 Continuous data

If we look at the Cormack-Jolly-Seber model, we see that there are no formal restrictions on the interval length. Equally spaced capture times are only used because the result is easier to interpret, and that it is convenient for experimenters to space the sampling at regular intervals. There are, however, situations where it is inconvenient to sample at equally time-spaced occasions. Sometimes the sampling has been done almost continuous, as in the case of the previously analysed Risør data. In these cases, the analysts have usually discretised the data into equally spaced capture occasions (see ex. Julliard et al. (2001)).

Another possibility is to let each single capture serve as a capture occasion. That is, for each line i in Table 1, no more than one of the $m_{i,j}$ are equal to one, with the possibility that all are zero (released and never recaptured). In this case, the data matrix would have as many columns as recaptured individuals, and as many lines as released individuals. The sampling intervals would be highly irregular.

4 Kaplan-Meier EM algorithm estimation

When dealing with continuous data, the interval capture and survival probabilities $\boldsymbol{\phi}$ and $\mathbf{p}$ are of less interest since they are clearly dependent on the interval length. It would be better to estimate the survival and capture functions $S_{\phi}(t)$ and $S_p(t)$, where $S_{\phi}(t)$ is the probability of surviving time t and $S_p(t)$ is the probability of

not being captured up to and including time t. Using the survival and capture probabilities,

$$S_\phi(t) = \prod_{j:t_{j+1}\leq t} \phi_j \qquad \text{and} \qquad S_p(t) = \prod_{j:t_{j+1}<t} (1-p_j).$$

With the augmented data of Table 2 and the rearrangement of Table 3, the Kaplan-Meier estimates of these two functions would be

$$\widehat{S}_\phi(t) = \prod_{j:t_{j+1}\leq t} \widehat{\phi}_j = \prod_{j:t_{j+1}\leq t} \left(\frac{r_j^\phi - d_j^\phi}{r_j^\phi} \right) \tag{10}$$

and

$$\widehat{S}_p(t) = \prod_{j:t_{j+1}<t} (1-\widehat{p}_j) = \prod_{j:t_{j+1}<t} \left(\frac{r_j^p - d_{j+1}^p}{r_j^p} \right) \tag{11}$$

It is a well known fact that the Kaplan-Meier estimate is a maximum likelihood estimate of the survival function (see e.g. Kalbfleisch and Prentice (1980)), and may be used as the M-step of an EM algorithm.

The E-step in the Kaplan-Meier EM algorithm is again a question of computing the expectation of the unknown death numbers $n_{i,j}, i = 1, \ldots, k-1, j = i, \ldots, k$. Since we have established a connection between the Kaplan-Meier estimates and the maximum likelihood estimates in Equation (5) by Equation (10) and (11), the E-step of the Kaplan-Meier algorithm should be equivalent to (8). That is,

$$\begin{aligned} \mathbb{E}[n_{i,i}] &= (R_i - m_i)\frac{(1 - S_\phi^*(t_{i+1}))}{C_i} \\ \mathbb{E}[n_{i,j}] &= (R_i - m_i)\frac{S_p^*(t_{j+1})(S_\phi^*(t_j) - S_\phi^*(t_{j+1}))}{C_i} \quad j = i, \ldots k \end{aligned} \tag{12}$$

where

$$\begin{aligned} S_\phi^*(t_{i+j}) &= S_\phi(t_{i+j})/S_\phi(t_i), \quad j = 1, \ldots, k-i \\ S_p^*(t_{i+j}) &= S_p(t_{i+j})/S_p(t_i), \quad j = 1, \ldots, k-i \end{aligned}$$

and C_i is still the normalising constant. The M step is trivially

$$\begin{aligned} S_\phi^{(k+1)}(t) &= \prod_{j:t_{j+1}\leq t} \left(\frac{\mathbb{E}[r_j^\phi] - \mathbb{E}[d_j^\phi]}{\mathbb{E}[r_j^\phi]} \right) \\ S_p^{(k+1)}(t) &= \prod_{j:t_{j+1}<t} \left(\frac{\mathbb{E}[r_j^p] - d_{j+1}^p}{\mathbb{E}[r_j^p]} \right) \end{aligned} \tag{13}$$

The resulting Kaplan-Meier EM algorithm is then as follows:

E step Compute the expected mortality data given the observed capture data and the current capture and survival functions $S_\phi^{(k)}$ and $S_p^{(k)}$ using the expressions in Eq. (12)

M step Set $S_\phi^{(k+1)}(t)$ and $S_p^{(k+1)}(t)$ according to Eq. (13).

Repeat these steps until a stopping criterion has been met.

4.1 Variance

To estimate the covariance matrix of the Kaplan-Meier estimates, note that by the chain rule,

$$\begin{aligned}\mathbb{S}[S_\phi, S_p] &= \frac{\partial Q(S_\phi, S_p | S_\phi, S_p)}{\partial(S_\phi, S_p)} \\ &= \frac{\partial Q(\boldsymbol{\phi}, \mathbf{p} | \boldsymbol{\phi}, \mathbf{p})}{\partial(\boldsymbol{\phi}, \mathbf{p})} \cdot \frac{\partial(\boldsymbol{\phi}, \mathbf{p})}{\partial(S_\phi, S_p)} \\ &= \mathbb{S}[\boldsymbol{\phi}, \mathbf{p}] \cdot \left(\frac{\partial(S_\phi, S_p)}{\partial(\boldsymbol{\phi}, \mathbf{p})} \right)^{-1}\end{aligned}$$

The first transition is possible because there is a one-to-one connection between (S_ϕ, S_p) and $(\boldsymbol{\phi}, \mathbf{p})$ The first term of this expression is given by the score vector of the interval capture and survival case, computed earlier. The element of the second term corresponding to p_i is then $S_p(t_i)^{-1}$, while the element corresponding to ϕ_i is $S_\phi(t_{i-1})^{-1}$. The first-order Richardson extrapolation of the central difference is used to estimate $\mathbb{H}[[S_\phi, S_p]$, and then invert to compute the estimated covariance matrix. Now, the diagonal elements of this covariance matrix are the variance of the Kaplan-Meier estimates. The standard asymptotic 95% pointwise confidence interval may then be computed for both S_ϕ and S_p as

$$\widehat{S}(t) \pm 1.96\widehat{\sigma}(t),$$

where $\widehat{\sigma}(t)$ is the diagonal element of of the estimated covariance matrix that correspond to t. One issue of this method is that the Kaplan-Meier estimator is restricted between 0 and 1, while this method may result in intervals that partly lies outside this boundary. In order to cope with this, the log-log-transformation $(g(x) = \log(-\log(x))$ may be applied as a variance stabilising transformation. Using this transform, we arrive at the following 95% asymptotic point wise confidence interval

$$\widehat{S}(t)^{\exp\left\{\pm \frac{1.96\widehat{\sigma}(t)}{(\widehat{S}(t)\log(\widehat{S}(t))}\right\}}.$$

4.2 Bootstrapping

Note that the point wise confidence intervals of the Kaplan-Meier estimates may give a wrong impression of the estimates uncertainty. The confidence interval estimates are based on the marginal distributions, and thus does not take into account the dependencies between the estimated probabilities. We know there is a strong dependency between the capture and mortality estimates for a given time, but this does not show in the estimated confidence intervals. We would like to use the bootstrap method to better picture the uncertainty of the Kaplan-Meier estimates. See Efron and Tibshirani (1993) for a thorough review of the bootstrap method, and Buckland and Garthwaite (1991) for an implementation of the bootstrap on mark-recapture models.

The implementation to this situation is quite straight forward. The data consists of a set of release and recapture events, and a set of release and never recapture

events. Two bootstrap pseudo-data subsets may then be drawn for each iteration, one from the empirical distribution of the release and recapture events, and one from the empirical distribution of release and never recapture events. Then the EM algorithm is used to compute bootstrap Kaplan-Meier estimates of the mortality and capture functions.

One of the fundamental requirements of bootstrapping is that the data to be re-sampled must have originated as an i.i.d sample. It could be argued that since one individual may contribute to several release-recapture events, there is dependence in the sample. But in this model we assume no dependency between events, and thus this dependency is neglected.

5 Nelson-Aalen EM algorithm estimation

The cumulative survival hazard function $A_\phi(t)$ is connected to the survival function through the identity

$$A_\phi(t) = -\log(S_\phi(t)) = \sum_{j:t_{j+1}\leq t} \log \phi_j.$$

When the ϕ_j are close to one,

$$A_\phi(t) \approx \sum_{j:t_{j+1}\leq t} (1-\phi_j).$$

For the cumulative capture hazard function,

$$A_p(t) \approx \sum_{j:t_{j+1}\leq t} p_j$$

when the p_j are small. The cumulative hazard function gives another impression of the process than the survival function and is easily compared to a homogeneous Poisson process, whose cumulative hazard function is linear. Although minus the logarithm of the Kaplan-Meier estimator could be used to estimate the cumulative hazard function, it is more convenient to use the so-called Nelson-Aalen estimator. With the augmented data of Table 2 and the rearrangement of Table 3, the Nelson-Aalen estimates of the mortality and capture cumulative intensities are given by

$$\widehat{A}_\phi(t) = \sum_{j:t_{j+1}\leq t} \frac{d_j^\phi}{r_j^\phi} \qquad \text{and} \qquad \widehat{A}_p(t) = \sum_{j:t_{j+1}<t} \frac{d_j^p}{r_j^p}. \tag{14}$$

Note that the Nelson-Aalen estimator may be viewed as a nonparametric maximum likelihood estimate of the cumulative hazard rate (see Andersen et al. (1993)) A Nelson-Aalen estimate of the capture and survival function is then given by

$$\widehat{S}_\phi(t) = \exp\{-\widehat{A}_\phi(t)\} \quad \text{and} \quad \widehat{S}_p(t) = \exp\{-\widehat{A}_\phi(t)\}$$

The Nelson-Aalen EM algorithm would then schematic look like this:

E-step Estimate the unknown mortality numbers as described in (12) using the Nelson-Aalen estimates of the capture and survival function.

M-step Compute the Nelson-Aalen estimate of the cumulative hazard function of the mortality and capture process (Eq. 14).

5.1 Variance

Using the same argument as for the Kaplan-Meier covariance estimate,

$$\begin{aligned}\mathbb{S}[A_\phi, A_p] &= \frac{\partial Q(A_\phi, A_p | A_\phi, A_p)}{\partial(A_\phi, A_p)} \\ &= \frac{\partial Q(\boldsymbol{\phi}, \mathbf{p} | \boldsymbol{\phi}, \mathbf{p})}{\partial(\boldsymbol{\phi}, \mathbf{p})} \cdot \frac{\partial(\boldsymbol{\phi}, \mathbf{p})}{\partial(A_\phi, A_p)} \\ &= \mathbb{S}[\boldsymbol{\phi}, \mathbf{p}] \cdot \left(\frac{\partial(A_\phi, A_p)}{\partial(\boldsymbol{\phi}, \mathbf{p})} \right)^{-1}\end{aligned}$$

where

$$\frac{\partial A_\phi}{\partial \phi_i} = -1 \qquad \text{and} \qquad \frac{\partial A_p}{\partial p_i} = 1.$$

The first-order Richardson extrapolation of the central difference is again used to estimate $\mathbb{H}[A_\phi, A_p]$. From the resulting variance estimates, confidence intervals may be computed. However, the same objections may be made against these confidence intervals, as for the intervals of the Kaplan-Meier estimates. Bootstrap estimates equivalent to those described in Section 4.2 may therefore be preferable.

6 Semi-parametric Multiplicative Hazard Models

Within the field of survival analysis, semi-parametric multiplicative hazard models are used extensively. It would be interesting to incorporate these models in the mark-recapture methodology, in order to include covariate information. To do this, a reconsideration of the likelihood is needed.

The semi-parametric multiplicative hazard model is defined by the multiplicative hazard function

$$\alpha(t; \mathbf{Z}_i(t)) = \alpha_0(t) \exp\{\boldsymbol{\beta}^\intercal \mathbf{Z}_i(t)\}$$

where $\mathbf{Z}_i(t)$ is the covariate vector of individual i at time t, and $\alpha_0(t)$ is the underlying, unspecified hazard function. The covariate vector must be predictable, that is the value of $\mathbf{Z}_i(t)$ should be known just before time t. The elements of the parameter vector $\boldsymbol{\beta}$ are usually called regression coefficients. Note that we here have defined a model with possibly time-dependent covariates. This representation also covers the case where the covariates do not change with time, by letting $\mathbf{Z}_i(t) = \mathbf{Z}_i \ \forall \ t$. For a stringent consideration of semi-parametric multiplicative hazard models, see Andersen et al. (1993). We shall here use a more heuristic approach.

Let us look at the mark-recapture situation once more, as presented by the augmented data of Table 2. Let us assume that all death events of a mortality

n	The total number of individuals participating in the experiment
k	The total number of capture occasions during the experiment.
t_i	The capture times of the experiment, $i = 2, \ldots k$. See Section 3
$\boldsymbol{\beta}_\phi+$	The regression coefficients for the mortality covariates
$\boldsymbol{\beta}_p$	The regression coefficients for the capture covariates
$\alpha_{0\phi}(t)$	The underlying, unspecified, mortality hazard function
$\alpha_{0p}(t)$	The underlying, unspecified, capture hazard function
$\Delta N_{\phi j}(t_i)$	An indicator function which is one if individual j dies at time t_i, zero else. Note that $\sum_{j=1}^{n} \Delta N_{\phi j}(t_i) = d_{i-1}^{\phi}$ (refer to Table 3)
$\Delta N_{pj}(t)$	An indicator function which is one if individual j is captured at time t_i, zero else. Note that $\sum_{j=1}^{n} \Delta N_{pj}(t_i) = d_i^p$ (refer to Table 3)

Table 4: Notation used in the likelihood (15)

period happen at the very end of the period. This is just a convenience assumption. We have no information on when the events occur, and may thus just as well place them at the end. In addition, we assume that the captures are instantaneous. With these assumptions, a likelihood may be stated. The notation (consistent with the notation of Andersen et al. (1993)) is explained in Table 6.

$$
\begin{aligned}
&L(\alpha_\phi(\cdot), \alpha_\phi(\cdot), \boldsymbol{\beta}_\phi, \boldsymbol{\beta}_\phi | \mathbf{m}, \mathbf{n}) = \\
&\prod_{i=2}^{k} \exp\left\{ - \int_{t_{i-1}}^{t_i} \mathcal{S}_\phi(\boldsymbol{\beta}_\phi, u) \alpha_{0\phi}(u) \mathrm{d}u \right\} \prod_{j=1}^{n} \left[\alpha_{0\phi}(t_i) \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t_i)\} \right]^{\Delta N_{\phi j}(t_i)} \\
&\cdot \prod_{i=2}^{k} \exp\left\{ - \int_{t_{i-1}}^{t_i} \mathcal{S}_p(\boldsymbol{\beta}_p, u) \alpha_{0p}(u) \mathrm{d}u \right\} \prod_{j=1}^{n} \left[\alpha_{0p}(t_i) \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_{+j}(t_i)\} \right]^{\Delta N_{pj}(t_i)}
\end{aligned}
\tag{15}
$$

In addition to the parameters defined in Table 6,

$$\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t) = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t)\} Y_j(t)$$

and

$$\mathcal{S}_p(\boldsymbol{\beta}_p, t) = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t)\} Y_j(t)$$

where $Y_i(t)$ is an indicator for individual i being at risk for capture or death at time t. Thus, $\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t)$ and $\mathcal{S}_p(\boldsymbol{\beta}_p, t)$ are the mortality and capture risk sets, weighted with the covariate information. The likelihood (Eq. 15) is explained as follows: The first term of the first line expresses the probability of not being killed during the interval (t_{i-1}, t_i), while the second term corresponds to the killed individual of this interval. Remember the assumption that all death events occur at the end of the

interval. The first term of the second line is the probability of not being captured during the interval, and the second term represents the captured individuals at t_i.

Now, for fixed value of $\boldsymbol{\beta}_\phi$, maximisation of the likelihood with respect to $\alpha_{0\phi}(t)$ leads to

$$\widehat{\alpha}_{0\phi}(t, \boldsymbol{\beta}_\phi) = \frac{\sum_{i=1}^n \Delta N_{\phi i}(t)}{\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t)}. \tag{16}$$

Note that for $t_{i-1} < t \leq t_i$

$$\sum_{i=1}^n \Delta N_{\phi i}(t) = d_{i-1}^\phi$$

and the Nelson-Aalen estimate of $A_{0\phi}(t)$ becomes

$$\widehat{A}_{0\phi}(t, \boldsymbol{\beta}_\phi) = \sum_{j:t_j \leq t} \frac{d_{j-1}^\phi}{\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t_j)}$$

A corresponding result applies to $\widehat{\alpha}_{0p}(t, \boldsymbol{\beta}_p)$ and $\widehat{A}_{0p}(t, \boldsymbol{\beta}_p)$ as well. It should be noted that this is an approximation, since no more than a single death or capture is supposed to happen at a time. But as long as $d_\cdot^\phi$ and $d_\cdot^p$ are not to large compared to the corresponding risk set, this approximation is good. See Cox and Oakes (1984) Ch. 7.6 for a discussion of ties in the data.

Inserting (16) into (15) the following partially maximised likelihood is obtained, depending only on $\boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$:

$$L(\boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p) \exp\{-\sum_{j=2}^k (d_{j-1}^\phi + d_j^p)\} \prod_{i=2}^k (d_{i-1}^\phi)^{d_{i-1}^\phi} (d_i^p)^{d_i^p}$$

Here,

$$L(\boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p) = \prod_{i=2}^k \prod_{j=1}^n \left(\frac{\exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t_i)\}}{\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t_i)} \right)^{\Delta N_{\phi j}(t_i)} \left(\frac{\exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t_i)\}}{\mathcal{S}_p(\boldsymbol{\beta}_p, t_i)} \right)^{\Delta N_{pj}(t_i)}. \tag{17}$$

This representation is known as the Cox partial likelihood (Cox (1975)). We denote The maximising values of $\boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$ is denoted with $\widehat{\boldsymbol{\beta}}_\phi$ and $\widehat{\boldsymbol{\beta}}_p$. Estimates of $A_{0\phi}(t, \boldsymbol{\beta}_\phi)$ and $A_{0p}(t, \boldsymbol{\beta}_p)$ is now given by inserting $\widehat{\boldsymbol{\beta}}_\phi$ and $\widehat{\boldsymbol{\beta}}_p$ into their respective Nelson-Aalen estimates. This is known as the Breslow estimator, and is a nonparametric maximum likelihood estimator.

The M-step of the EM algorithm is thus established. To establish the E-step, we need to take the expectation of the logarithm of the likelihood (15) with respect to the missing data. The log-likelihood is given by

$$\begin{aligned} &l(\alpha_\phi(\cdot), \alpha_p(\cdot), \boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p | \mathbf{m}, \mathbf{n}) = \\ &\sum_{i=2}^k - \int_{t_{i-1}}^{t_i} \mathcal{S}_\phi(\boldsymbol{\beta}_\phi, u) \alpha_{0\phi}(u) \mathrm{d}u + \sum_{j=1}^n \Delta N_{\phi j}(t_i) [\log \alpha_{0\phi}(t_i) + \boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t_i)] \\ &+ \sum_{i=2}^k - \int_{t_{i-1}}^{t_i} \mathcal{S}_p(\boldsymbol{\beta}_p, u) \alpha_{0p}(u) \mathrm{d}u + \sum_{j=1}^n \Delta N_{pj}(t_i) [\log \alpha_{0p}(t_i) + \boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t_i)] \end{aligned} \tag{18}$$

To compute the expectation of this expression is equivalent to compute the expectation of $\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t), \mathcal{S}_p(\boldsymbol{\beta}_p, t)$ and $\Delta N_{\phi j}(t_i)$. Now,

$$\mathbb{E}[\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t)] = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_{\phi j}(t)\} \mathbb{E}[Y_j(t)],$$

and

$$\mathbb{E}[\mathcal{S}_p(\boldsymbol{\beta}_p, t)] = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_{pj}(t)\} \mathbb{E}[Y_j(t)].$$

That is, we need to compute $\mathbb{E}[Y_i(t)]$ and $\mathbb{E}[\Delta N_{\phi j}(t_i)]$ for each individual $j = 1, \ldots, n$ and every time $t_i, i = 2, \ldots, k$ given the estimated regression coefficients $\widehat{\boldsymbol{\beta}_\phi}$ and $\widehat{\boldsymbol{\beta}_p}$, and the estimated Breslow estimates $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}_\phi})$ and $\widehat{A}_{0p}(t, \widehat{\boldsymbol{\beta}_p})$ from the previous EM iteration. When an individual j is recaptured after release, we know that $\mathbb{E}[Y_j(t)] = Y_j(t) = 1$ for the whole period between release and recapture. Thus, we need only compute $\mathbb{E}[Y_j(t)]$ and $\mathbb{E}[\Delta N_{\phi j}(t)]$ for the individuals that are released, but never recaptured again. Let individual j be released at time t_{i-1} and never seen again. That is, we need to compute $\mathbb{E}[Y_j(t)]$ and $\mathbb{E}[\Delta N_{\phi j}(t)]$ for all the following capture times $t_i, t_{i+1}, \ldots, t_k$. Note that

$$Y_j(t_l) = \sum_{m=i}^{l} \Delta N_{\phi j}(t_m), \quad l = i, \ldots k$$

thus we only need to compute $\mathbb{E}[\Delta N_{\phi j}(t)]$. Equivalent with the discussion of the unknown mortality data in Section 3 it may be argued that $\Delta N_{\phi j}(t_l), l = i, \ldots k$ are multi-nominal distributed. Then the expectation is computed as

$$\begin{aligned}
\mathbb{E}[\Delta N_{\phi j}(t_i)] &= \frac{(1 - \widehat{S}^*_{\phi j}(t_i))}{C_i} \\
\mathbb{E}[\Delta N_{\phi j}(t_{i+1})] &= \frac{\widehat{S}^*_{pj}(t_{i+1})(\widehat{S}^*_{\phi j}(t_i) - \widehat{S}^*_{\phi j}(t_{i+1}))}{C_i} \\
&\vdots \\
\mathbb{E}[\Delta N_{\phi j}(t_k)] &= \frac{\widehat{S}^*_{pj}(t_k)(\widehat{S}^*_{\phi j}(t_{k-1}) - \widehat{S}^*_{\phi j}(t_k))}{C_i}
\end{aligned} \tag{19}$$

where

$$\begin{aligned}
\widehat{S}^*_{\phi j}(t_l) &= \widehat{S}_{\phi j}(t_l)/\widehat{S}_{\phi j}(t_{i-1}), \quad l = i, \ldots, k \\
\widehat{S}^*_{pj}(t_l) &= \widehat{S}_{pj}(t_l)/\widehat{S}_{pj}(t_{i-1}), \quad l = i, \ldots, k
\end{aligned}$$

and

$$\begin{aligned}
\widehat{S}_{\phi j}(t_l) &= \exp\{-\widehat{A}_{\phi j}(t_l, \widehat{\boldsymbol{\beta}}_\phi)\} = \exp\{-\widehat{A}_{\phi 0}(t_l, \widehat{\boldsymbol{\beta}}_\phi) \exp\{\widehat{\boldsymbol{\beta}}_\phi^\intercal \mathbf{Z}_{\phi j}(t)\}\} \\
\widehat{S}_{pj}(t_l) &= \exp\{-\widehat{A}_{pj}(t_l, \widehat{\boldsymbol{\beta}}_p)\} = \exp\{-\widehat{A}_{p0}(t_l, , \widehat{\boldsymbol{\beta}}_p) \exp\{\widehat{\boldsymbol{\beta}}_p^\intercal \mathbf{Z}_j(t)\}\}.
\end{aligned}$$

C_i is still the normalising constant. The Breslow estimates $\widehat{A}_{\phi 0}(t_l, \widehat{\boldsymbol{\beta}}_\phi)$ and $\widehat{A}_{p0}(t_l, \widehat{\boldsymbol{\beta}}_p)$ are from the previous EM iteration.

We see that the EM algorithm for the semi-parametric multiplicative hazard model is principally equal to the Kaplan-Meier and the Nelson-Aalen EM algorithms. The only difference is that there is more emphasis on the individual.

6.1 Variance

Following Andersen et al. (1993), the covariance matrix of the regression coefficients estimator $\widehat{\boldsymbol{\beta}}$ is estimated by the inverse observed information matrix $\mathcal{I}(\widehat{\boldsymbol{\beta}})$. The inverse observed information matrix is given by the Hessian of the log Cox partial likelihood (17). To compute this Hessian, the NDS approach is again used.

We need to numerically differentiate the score vector at the maximum likelihood estimate. The score vector is given by

$$\mathbb{S}[\boldsymbol{\beta}^\dagger | x] = \left. \frac{\partial Q(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger)}{\partial \boldsymbol{\beta}} \right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^\dagger},$$

where

$$Q(\boldsymbol{\beta}|\boldsymbol{\beta}^\dagger) = \mathbb{E}[l(\boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p)|\boldsymbol{\beta}_\phi^\dagger, \boldsymbol{\beta}_p^\dagger],$$

the expectation of the log Cox likelihood with respect to $\boldsymbol{\beta}_\phi^\dagger$ and $\boldsymbol{\beta}_p^\dagger$. The score function for $\boldsymbol{\beta}$ is then given by

$$\begin{aligned} \mathbb{S}[\boldsymbol{\beta}_\phi | x] &= \mathbb{E}\left[\sum_{i=2}^{k} \sum_{j=1}^{n} \Delta N_{\phi j}(t_i) \left\{ \mathbf{Z}_j - \frac{\mathcal{S}'(\boldsymbol{\beta}_\phi, t_i)}{\mathcal{S}(\boldsymbol{\beta}_\phi, t_i)} \right\} \right] \\ &\approx \sum_{i=2}^{k} \sum_{j=1}^{n} \mathbb{E}[\Delta N_{\phi j}(t_i)] \left\{ \mathbf{Z}_j - \frac{\mathbb{E}[\mathcal{S}'_\phi(\boldsymbol{\beta}_\phi, t_i)]}{\mathbb{E}[\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t_i)]} \right\} \end{aligned} \tag{20}$$

where

$$\mathcal{S}'_\phi(\boldsymbol{\beta}_\phi, t_i) = \frac{\partial \mathcal{S}(\boldsymbol{\beta}_\phi, t_i)}{\partial \boldsymbol{\beta}_\phi} = \sum_{j=1}^{n} \mathbf{Z}_j \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j\} Y_j(t_i).$$

The approximation in equation (20) is due to a Taylor expansion. The score function of $\boldsymbol{\beta}_p$ is computed correspondingly.

The score functions are now numerically differentiated with respect to $\boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$ at the maximum likelihood. Let us look at this procedure more closely. For the first element of $\boldsymbol{\beta}_\phi$ a small perturbation ϵ is added and the underlying hazard function $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}}_\phi + \epsilon)$ is computed using the existing times of death. Then $\mathbb{E}[\Delta N_{\phi j}(t_i)]$ and $\mathbb{E}[\Delta N_{pj}(t_i)]$ is computed for all i and j using $(\widehat{\boldsymbol{\beta}}_\phi + \epsilon)$, $\widehat{\boldsymbol{\beta}}_p$, $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}}_\phi + \epsilon)$ and $\widehat{A}_{0p}(t, \widehat{\boldsymbol{\beta}}_p)$. Finally, the score functions for $\boldsymbol{\beta}_\phi$ is computed using (20) (and correspondingly for $\boldsymbol{\beta}_p$). This procedure is then repeated for all elements of $\boldsymbol{\beta}_\phi$ and of $\boldsymbol{\beta}_p$.

Asymptotically, the distribution of the regression coefficient estimator $\boldsymbol{\beta}$ is normal, and the Wald test statistic

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\intercal \mathcal{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$$

for the simple hypothesis $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ is asymptotically χ^2 distributed with $p = \dim(\boldsymbol{\beta})$ degrees of freedom.

7 Example: European Dipper

The methods are illustrated by applying them to a set of mark-recapture data on the European Dipper (*Cinclus cinclus*), gathered by G. Marzolin in Eastern France (Marzolin (2002)). The data were collected from 1981 and up to 1998, but we only consider the subset 1981-1987, since these data have been analysed thoroughly using discrete mark-recapture methodology by Lebreton et al. (1992) (frequentistic) and Brooks et al. (2000) (Bayesian). The sampling was done quite continuously between September and June each year (see Figure 1 for a cumulative counting plot of captures in 1982).

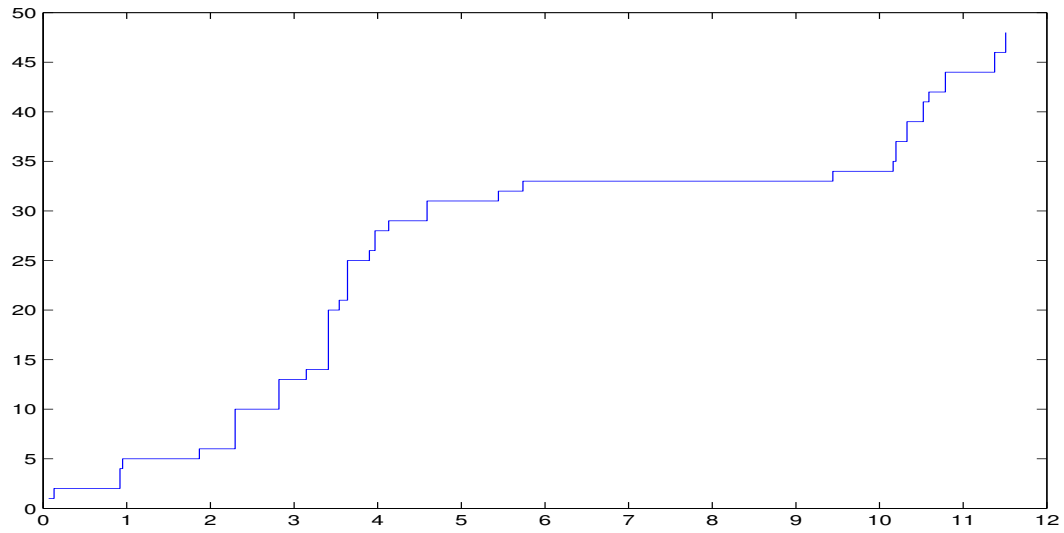


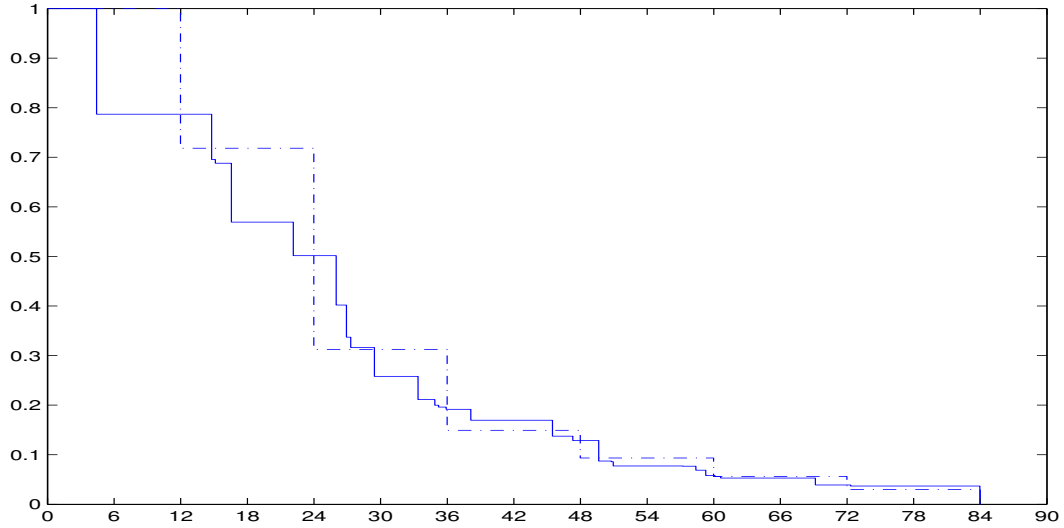
Figure 1: A cumulative counting plot of the capture events in 1982. The time scale is in months after January 1st, 1982

We consider both the discrete and the continuous data, and see how the discretation influences the analysis. The discrete data used in the established literature is given in Table 5. The discretation performed by Lebreton et al. (1992) was done such that only breeding adults captured during the breeding period each year from early March to early June was included. However, a substantial part of the recaptures was done during the autumn as we see in the cumulative counting plot of 1982 (Figure 1). For 1982, we see from Table 5 that only 11 recaptures was included in the discrete analysis, while the total number was 48. Thus, a large part of the information is neglected in the discrete analysis. The resulting maximum likelihood estimates of the parameters in the Cormack-Jolly-Seber model is given in Table 6. Note that we reach the exact same result by applying the EM algorithm to the discrete data.

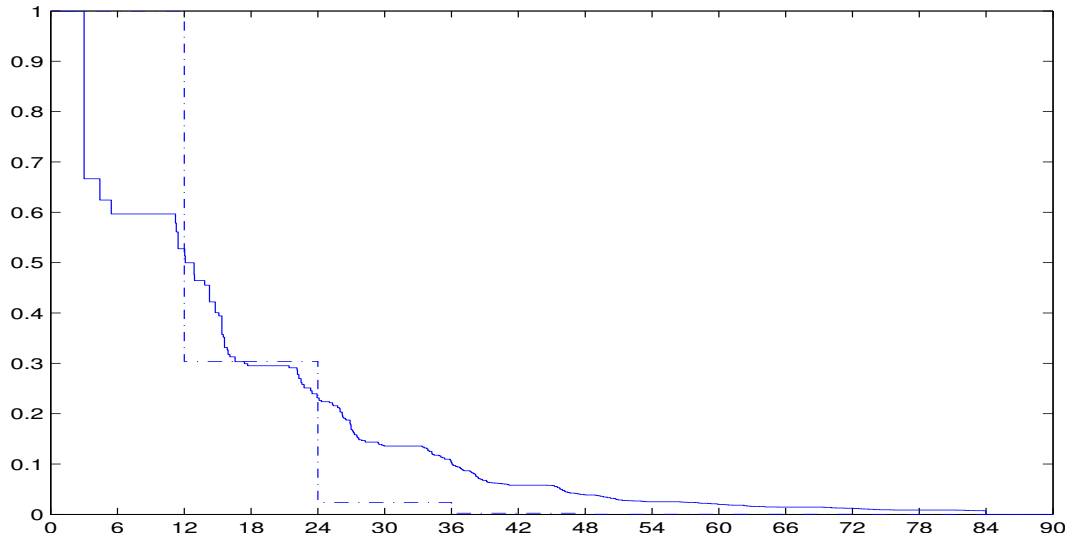
The Kaplan-Meier EM algorithm is now applied to the continuous European Dipper data, and the estimated survival and capture functions is plotted in Figure

Year	i	R_i	t_2	t_3	t_4	t_5	t_6	t_7
1981	1	22	11	2	0	0	0	0
1982	2	60		24	1	0	0	0
1983	3	78			34	1	0	0
1984	4	80				45	1	2
1985	5	88					51	0
1986	6	98						52

Table 5: Capture-recapture data for pooled male and female European Dippers (from Marzolin (1988))



(a) Estimate of the Survival Function



(b) Estimate of the Capture Function

Figure 2: The Kaplan-Meier EM estimates of the survival and capture function for the discrete ($- \cdot -$) and the continuous (—) data. The time scale is in months after January 1st 1981

Year	$\widehat{\phi}$	$\widehat{SE[\widehat{\phi}]}$	$\widehat{p}$	$\widehat{SE[\widehat{p}]}$
1981-1982	0.718	0.155		
1982-1983	0.435	0.069	0.696	0.166
1983-1984	0.478	0.060	0.923	0.073
1984-1985	0.626	0.059	0.913	0.058
1985-1986	0.599	0.056	0.901	0.054
1986-1987			0.932	0.046

Table 6: Summary of estimates of annual survival (ϕ) and capture (p) probabilities for the pooled data on male and female European Dipper, reproduced from Lebreton et al. (1992).

Year	$\widehat{\phi}_{\text{cont}}$	$\widehat{\phi}_{\text{disc}}$
1981-1982	0.724	0.718
1982-1983	0.453	0.435
1983-1984	0.657	0.478
1984-1985	0.456	0.626
1985-1986	0.685	0.599
1986-1987	0.697	0.531

Table 7: Estimates of the annual survival probabilities based on the continuous and the discrete data estimates

2 together with the corresponding discrete estimates. We see from the estimates of the survival function (see Figure 2a), that the continuous and the discrete estimates are quite similar. Note however, the steep slope of the curve between month 24 and 30. During this period, a major flood occurred, and washed away a number of nests. This effect we see quite clearly in the estimate of the continuous data, but is not revealed in the estimate of the discrete data. To assess the difference between the continuous and the discrete models, the yearly continuous survival probabilities were estimated. The results are presented in Table 7. We see that the difference between the two estimates are quite large. The flood of spring 1983 effects both the 1982-1983 and the 1983-1984 estimates based for the discrete data, while it only affects the 1982-1983 continuous data estimate. We also see that the survival estimate for year 1984-1985 is low in the continuous case. We have no definite explanation for this increase in mortality, but we know that there was a period of very low temperatures during January 1985.

When we look at the estimates for the capture function (see Figure 2b), the difference between the discrete and the continuous case is more pronounced. We see that the yearly capture probability is estimated much higher for the discrete data than for the continuous data. This may seem strange, since the mortality estimates were almost equal, but an explanation may be given: The sampling of the Dipper experiment was done such that the experimenter moved between sampling sites, and seldom visited the same site close in time. Since the European Dipper is quite

stationary, the actual capture rate for an individual right after a recapture event was quite low. The yearly capture probability however, was quite high. This effect covers much of the difference between the discrete and continuous analysis of the capture process.

The uncertainty of the estimates is computed using the NDS algorithm described in section 2.1. The log-log-transform is used to estimate point-wise confidence limits. The results are presented in Figure 3 for the mortality estimates, and in Figure 4 for the capture estimates. We see that the results look rather strange in the continuous case. It seems like the capture function estimate is very precise, while the survival function estimate is very uncertain. This behaviour could be explained by the correlation between the parameters, as discussed in Section 4.2. In addition, 500 bootstrap replicas of the mortality and capture functions are computed and presented in Figure 5. We see from the results that the uncertainty of the survival function estimate is probably smaller than indicated by the confidence limits computed from the covariance matrix. For the capture function estimate, the uncertainty is probably larger than indicated by the confidence limits computed from the covariance matrix.

The Nelson-Aalen estimates of the mortality and capture cumulative intensities are also estimated and presented in Figure 6 together with 500 bootstrap replicas.

The final analysis presented in connection with the Dipper data, is the analysis of a Cox proportional hazard model with sex as covariate. That is,

$$Z_i = \begin{cases} 0 & \text{if individual } i \text{ is a male} \\ 1 & \text{if individual } i \text{ is a female} \end{cases}$$

The resulting Breslow estimates of the baseline mortality and capture cumulative intensities are given in Figure 7. The analysis resulted in $\widehat{\beta}_\phi = 0.103$ and $\widehat{\beta}_p = 0.180$, in addition to

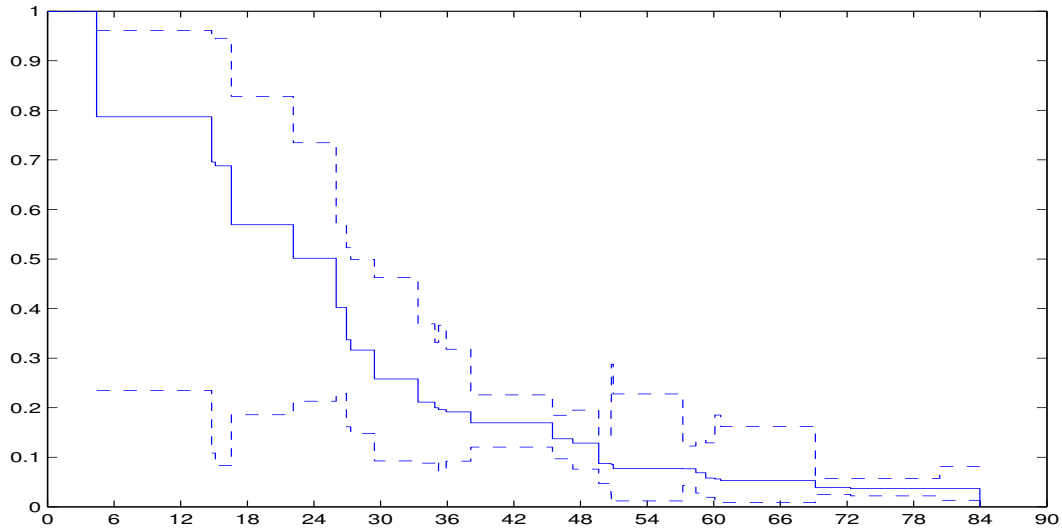
$$\widehat{\mathrm{Cov}}(\widehat{\beta}_\phi, \widehat{\beta}_p) = \begin{bmatrix} 0.0130 & 0.0047 \\ 0.0043 & 0.0142 \end{bmatrix}.$$

We see that the estimate of the covariance matrix is not symmetric. This is probably due to the Tylor approximation that we did according to equation (20). The Wald test statistic for the hypothesis that $(\beta_\phi, \beta_p) = (0, 0)$ equals 2.625, which means that we have no reason to reject this hypothesis. This is in accordance with the analysis in Lebreton et al. (1992).

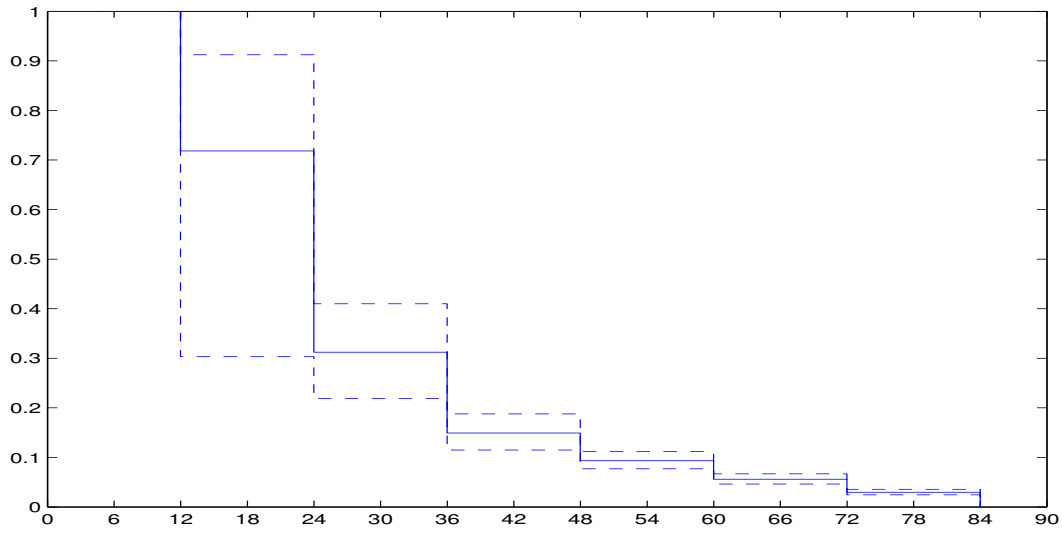
8 Summary and Discussion

In this paper, an EM algorithm to analyse the Cormack-Jolly-Seber model for mark-recapture data is presented. Since the EM algorithm returns the maximum likelihood estimates, this algorithm is completely equivalent to the classic method of maximising the likelihood numerically. In addition, equivalent variance estimates, based on the observed information matrix, is given.

An argument was then presented that the Cormack-Jolly-Seber model also could be used on continuous data. In this case, the survival and capture functions $S_\phi(t)$

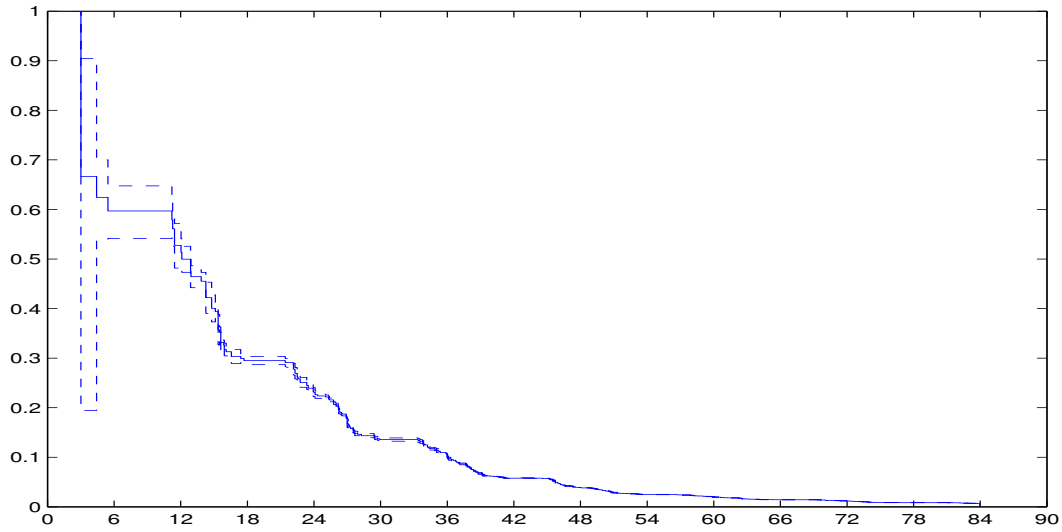


(a) Continuous Data

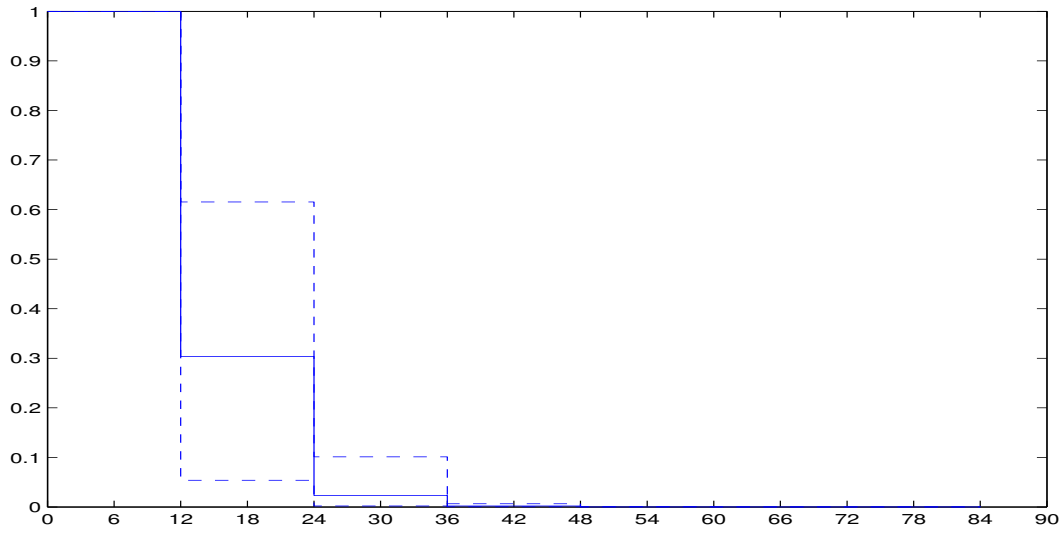


(b) Discrete Data

Figure 3: The EM estimates of the survival function for the continuous and discrete data with 95% pointwise confidence limits using the log-log-transform. The time scale is in months after January 1st 1981

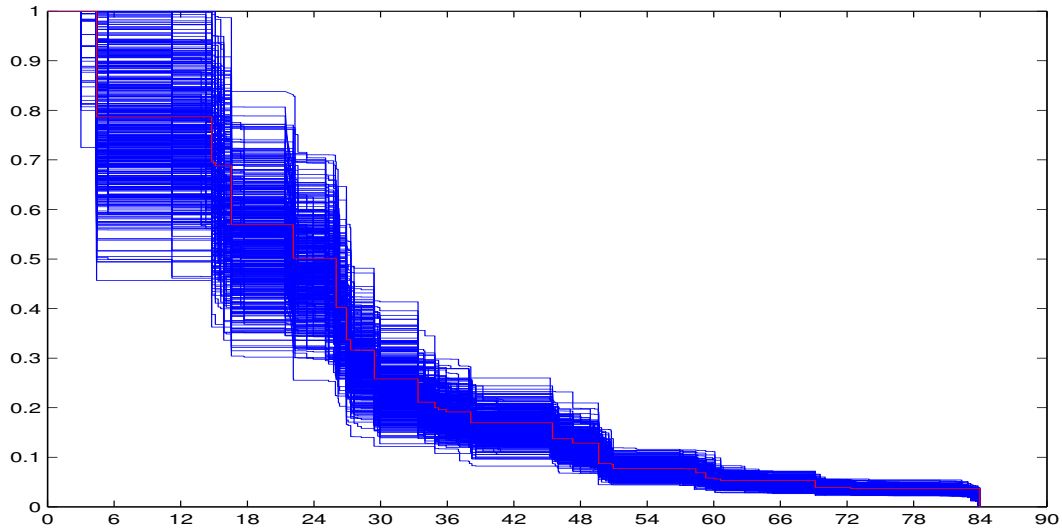


(a) Continuous Data

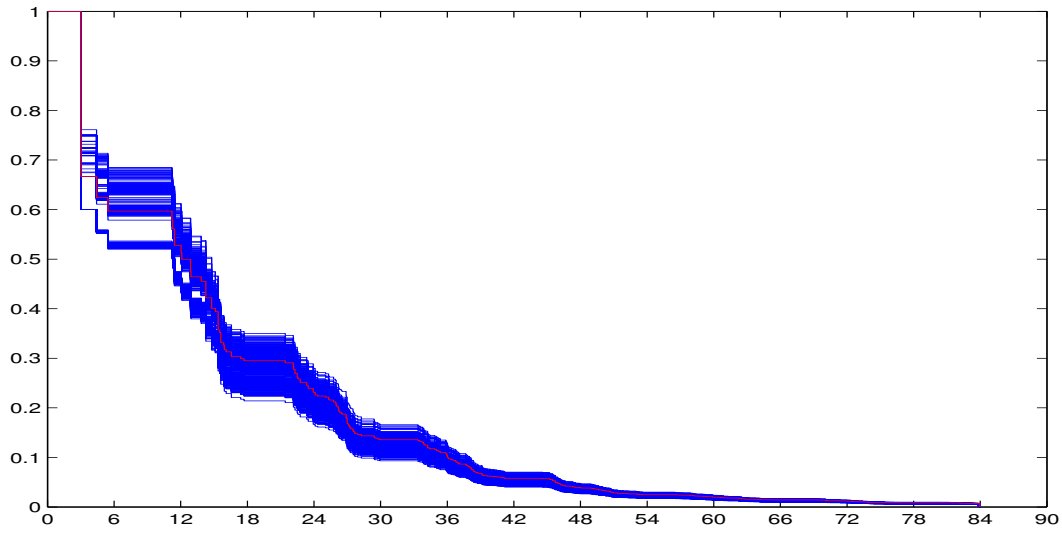


(b) Discrete Data

Figure 4: The EM estimates of the capture function for the continuous and discrete data with 95% pointwise confidence limits using the log-log-transform. The time scale is in months after January 1st 1981

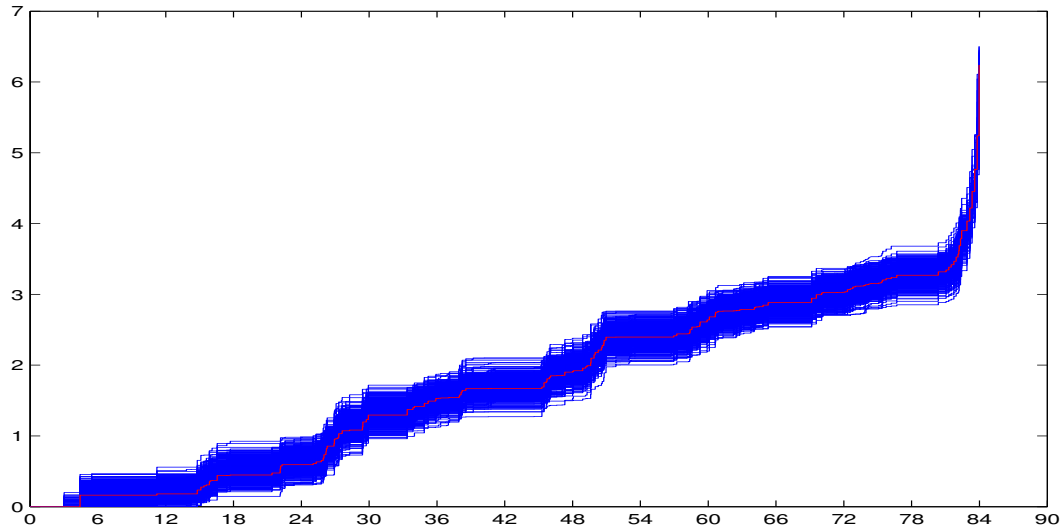


(a) Survival Function

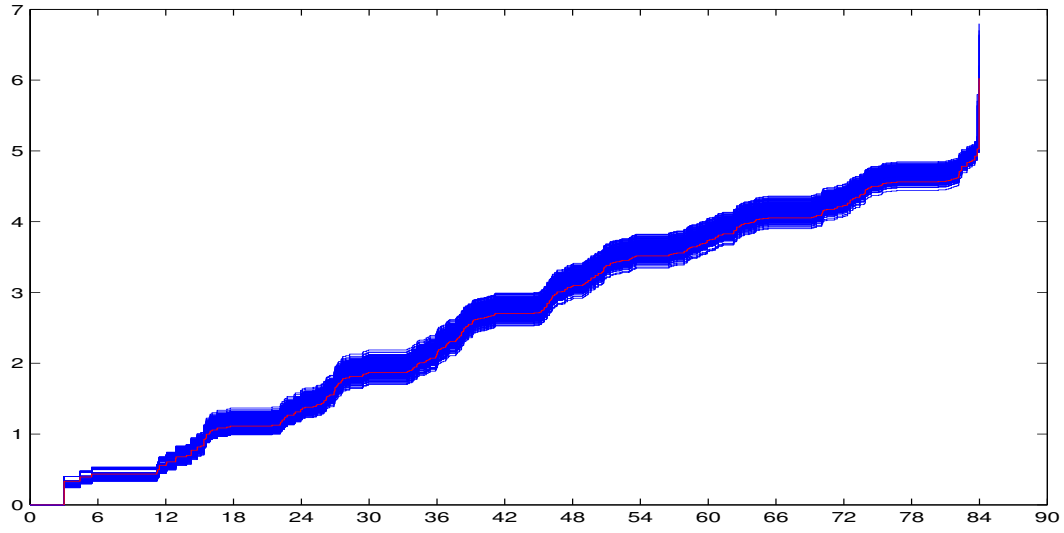


(b) Capture Function

Figure 5: The Kaplan-Meier EM estimates of the survival and capture function for the continuous data from 500 bootstrap replications. The original estimates are shown in red. The time scale is in months after January 1st 1981

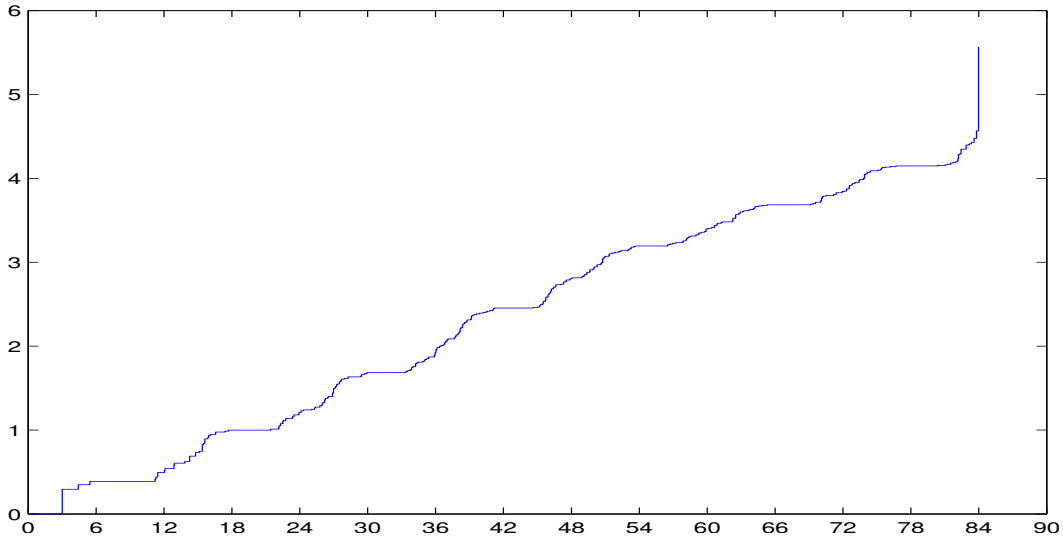


(a) Cumulative Mortality Rate

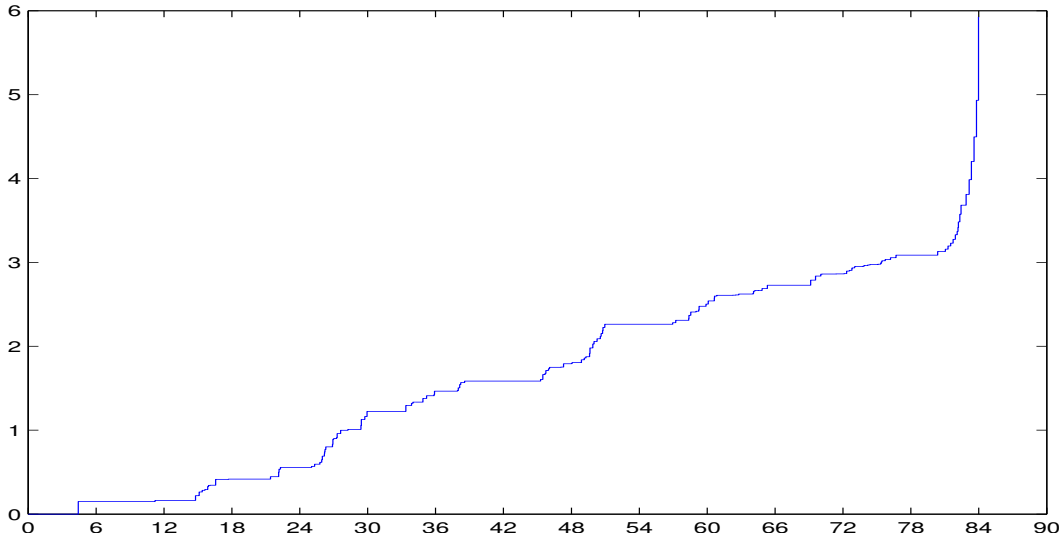


(b) Cumulative Capture Rate

Figure 6: The Nelson-Aalen EM estimates of the cumulative mortality and capture cumulative rates with 500 bootstrap replications. The original estimates are shown in red. The time scale is in months after January 1st 1981



(a) Cumulative Mortality Rate



(b) Cumulative Capture Rate

Figure 7: The Breslow EM estimates of the baseline cumulative mortality and capture rates. The time scale is in months after January 1st 1981

and $S_p(t)$ are preferable to the interval survival and capture probabilities. An EM algorithm to compute the Kaplan-Meier estimates of the survival and capture functions was described, in addition to an EM algorithm to compute the Nelson-Aalen estimates of the cumulative mortality and capture hazards. Variance estimates for both the Kaplan-Meier and the Nelson-Aalen estimators were presented, using both the observed information matrix and the bootstrap.

Lastly, the Cormack-Jolly-Seber model was abandoned altogether and replaced by the semi-parametric multiplicative hazard model. It was showed how this model could be analysed using the EM algorithm.

The developments were implemented on a much used data set on European Dippers, previously analysed classically. It was showed that the continuous analysis gave a more detailed account for the mortality process, and revealed effects which were hidden in the discrete analysis. Then a semi-parametric multiplicative hazard model with sex as covariate was analysed. The conclusion was that there were no significant differences between males and females with respect to capture or survival. This is in accordance with the classic analysis.

8.1 Contributions

Looking back, three distinct new contributions is recognised.

EM algorithm The EM algorithm for capture-recapture experiments is quite easy to implement, and does not rely on an optimising algorithm. The drawback is that the convergence may be slow, especially when there is a lot of missing data (low recapture probability).

Continuous data Although the Cormack-Jolly-Seber model is suitable for continuous data, this has never been implemented. Instead, the data has been discretised and then analysed. It is believed that applying a continuous model to the continuous data directly is better than the present method of discretising the data and using the CJS model. When analysing continuous data, it is preferable to describe the results using Kaplan-Meier or Nelson-Aalen estimates.

Semi-parametric multiplicative hazard models Semi-parametric multiplicative hazard models are used extensively in epidemiology and biostatistics. Introduce these models to continuous mark-recapture data seems sensible, and they allow us to include covariates to the analysis.

8.2 Further developments

A natural next step, would be to apply Bayesian methodology to continuous data, and the Cox proportional hazard model. In this case, one would no longer use the EM algorithm, but apply some prior to the missing data and then compute the posterior distribution using an appropriate algorithm. Another expected development is to apply parametric models to the data, such as Weibull regression models. It is also

believed that there may be possible to use some of the methods developed here on epidemiological problems.

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag.

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, 15(4):357–76.

Buckland, S. T. and Garthwaite, P. H. (1991). Quatifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, 47:255–68.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62.

Cox, D. R. and Oakes, D. (1984). *Analysis of surivival data*. Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistics Society Series B*, 39:1–38.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, 62:257–70.

Julliard, R., Stenseth, N. C., Gjøsæter, J., Lekve, K., Fromentin, J.-M., and Danielssen, D. S. (2001). Natural mortality and fishing mortality in a costal cod population: a release-recapture experiment. *Ecological Applications*, 11(2).

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure data*. Wiley series in probability and mathematical statistics.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Marzolin, G. (1988). Polygynie du cincle plongeur (*Cinclus cinclus*) dans les côtes de lorraine. *L'Oiseau et la Revue Francaise d'Ornithologie*, 58:277–286.

Marzolin, G. (2002). Influence of the mating system of the eurasian dipper on sex-spesific local survival rates. *Journal of wildlife management*, 66(4):1023–30.

Meng, X.-L. and Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86:899–909.

Assessing the Effects of Lengthy Sampling Periods on Mark-Recapture Methodology

Inge Christoffer Olsen
Department of Mathematical Sciences,
Norwegian University of Science and Technology

Summary The Cormack-Jolly-Seber (CJS) model is fundamental to the analysis of mark-recapture experiments. One of the assumptions for this model is that the sampling is done instantaneously. Little has been done to assess the consequences of violating this assumption. In this paper, the potential risks of using lengthy sampling periods are discussed, and it is shown by simulated data that this use may lead to substantial bias of the CJS survival estimator.

Keywords: Assessing bias; Capture-recapture; Cormack-Jolly-Seber model; Instantaneous sampling; Mark-recapture; Violating model assumption;

1 Introduction

One assumption of the Cormack-Jolly-Seber (CJS) model for mark-recapture experiment data, is that the sampling is done instantaneously. This is usually not possible, and the sampling has to be done during a certain period of time for each sample occasion. The duration of a sampling period should of course be as short as possible. However, due to practical and technical limitations, the sampling period may stretch out considerably. We want to assess the bias of the survival probability estimates connected to this violation of the CJS model assumptions.

To motivate the interest in this matter, two mark-recapture experiments with quite long sampling periods are presented. The first experiment to be considered, is on the European Dipper (*Cinclus cinclus*) performed in eastern France from 1981 to 1998 (see Marzolin (2002) and Lebreton et al. (1992)). For this experiment, the sampling was done during the breeding period of the adult Dippers, which lasted from the beginning of March to the beginning of June (three months of the year).

The second one is a survey on the Norwegian coastal cod population performed on the southern coast of Norway during 1986-1993. In this experiment, reared and wild-caught cod were marked and batch released by experimentalists. The recaptures, however, were done by commercial and recreational fishermen. The sampling was thus done almost continuously around the year. This experiment is analysed by the CJS model in Julliard et al. (2001). To cope with the continuous data, the data

t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
R_1	$m_{1,2}$	$m_{1,3}$	$\cdots$	$m_{1,k-1}$	$m_{1,k}$
	R_2	$m_{2,3}$	$\cdots$	$m_{2,k-1}$	$m_{2,k}$
		R_3	$\cdots$	$m_{3,k-1}$	$m_{3,k}$
			$\ddots$	$\cdots$	$\vdots$
				R_{k-1}	$m_{k-1,k}$

Table 1: Table of mark-recapture data

are discretised and each month is considered as a sampling occasion. No attempt is done to determine the effect of this CJS model assumption violation.

The issue of lengthy sampling intervals has already been considered. A general discussion is made in Williams et al. (2002) p. 435, and a more specific analysis is done in Smith and Anderson (1987). This analysis however, is done only for tag-recovery models. It concludes that the survival estimator for tag-recovery models, as presented by Brownie et al. (1985), is robust to long ringing periods. While long capture periods only affect the marking phase of tag-recovery models, they also affect the recapture phase of mark-recapture models. Thus, the effect of long sampling intervals may be different for tag-recovery and mark-recapture analysis.

This paper is organised as follows: In Section 2, the standard mark-recapture experiment is presented, and the time-dependent CJS model (see Cormack (1964), Jolly (1965) and Seber (1965)) for analysing such experiments is defined. In addition, we will regard the assumptions for this model. Some of the expected effects of using lengthy sampling periods are also considered in this section. In Section 3, several sets of mark-recapture data are simulated. This is done to examine the effects of the model violation on the survival estimates. In the last section, an account of the results and a discussion is given.

2 The Model

Let $t_1, t_2, \ldots, t_k$ be the capture occasion times of a standard mark-recapture experiment. The experiment data are usually presented as an array with the number of releases and recaptures. Such an array is presented in Table 1, where R_i is the number of released individuals at time t_i, and m_{ij} is the number of individuals that are released at time t_i and not recaptured until time t_j. Note that the same individual may occur on several lines as a result of recurring recaptures and releases.

The parameter space of the CJS model consists of the survival parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{k-1})$ and the capture parameters $\mathbf{p} = (p_2, \ldots, p_k)$. Here, ϕ_i is the probability of surviving the interval (t_i, t_{i+1}), and p_i is the probability of being captured at time t_i given that it is alive. A likelihood on these parameters conditional on the data presented in Table 1 is then given by

$$L(\boldsymbol{\phi}, \mathbf{p}|\mathbf{R}, \mathbf{m}) \propto \Delta \prod_{i=1}^{k-1} \prod_{j=i+1}^{k} \left(\phi_i p_j \prod_{l=i+1}^{j-1} \phi_l (1 - p_l) \right), \tag{1}$$

where

$$\Delta = \prod_{i=1}^{k} \chi_i^{R_i - m_{i\cdot}},$$

Δ considers the individuals that are never recaptured after their last release, and χ_i is the probability that an individual alive at time t_i is not seen again during the remains of the experiment. This probability may be computed recursively by

$$\chi_i = (1 - \phi_i) + \phi_i(1 - p_{i+1})\chi_{i+1}$$

where $\chi_k = 1$. Maximum likelihood estimates are computed numerically using this likelihood.

2.1 Assumptions of the model

Following Williams et al. (2002), there are six typically listed assumptions for the CJS model:

1. Each marked individual has the same probability of being recaptured at a given sampling occasion.
2. Each marked individual has the same probability of surviving a given interval between two sampling occasions.
3. No marks are lost or ignored during capture.
4. The sampling occasions are instantaneous, and recaptured individuals are released immediately.
5. Any emigration from the population is permanent.
6. There is independence between the individuals with respect to capture and survival.

We shall here emphasise on the effects of violations of assumption 4: Instantaneous sampling periods.

Consider an experiment where the sampling is not done instantaneously, but during an interval of length T_r (see Figure 1). We may presume that there is at least

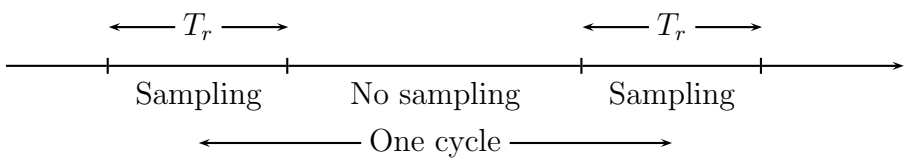


Figure 1: Illustration of the mark-recapture situation

some mortality during the sampling periods. As stated, the CJS model assumes instantaneous sampling periods at times $t_1, t_2, \ldots, t_k$. We thus need to adjust the data

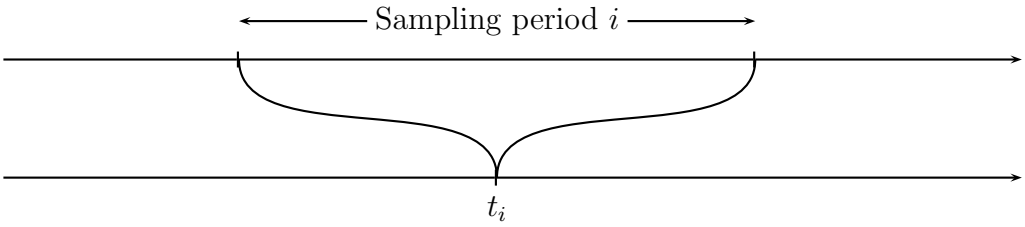


Figure 2: Illustration of the pooling procedure

to comply with this assumption. The adjustment is done by pooling the captures within one sampling period and then allocating the pooled captures to a sampling occasion time (see Figure 2). The exact location of the sampling occasion times $t_1, t_2, \ldots, t_k$ is implicitly determined by the likelihood maximisation, but they may be expected to lie near the median of the capture times for each sampling period.

By this reallocation of the captures, some captures are moved forward in time and some are moved backward. This movement influences the estimated survival probabilities. Consider an individual which is captured early in the interval, then released and never seen again. By the adjustment to instantaneous sampling, this individual has its known lifespan shortened. It thus contributes to an underestimation of the survival. Individuals that are recaptured late in the interval, then released and never seen again, contribute to an overestimation of the survival. They have their known lifespan prolonged. These effects are especially important with regard to newly marked individuals. Releasing them in one batch at the beginning of the sample interval may lead to an underestimation of the survival. Batches released at the end of the interval may contribute to an overestimation of the survival.

3 Simulation study

In order to assess the effects of the discretation discussed in the previous section, mark-recapture data for different scenarios of mortality, capture rates, length of sampling periods and release strategies of the newly marked individuals are simulated. We want to estimate the bias related to the CJS analysis for each of these scenarios.

Assume that the mortality rate is constantly equal to μ throughout the experiment period, and that the capture rate λ is constant during the sampling periods. Let T_r be the relative length of the sampling intervals. That is, $T_r = 0$ if the the sampling is instantaneous, and $T_r = 1$ if the sampling is continuous and the sampling intervals coincide. In addition, let T_m be the relative length of the mortality period. Note that $T_r + T_m = 1$.

With these figures defined, we may compute some interesting probabilities:

- $e^{-\mu T_m}$ is the probability of surviving a mortality period.
- $e^{-(\lambda+\mu)T_r}$ is the probability of surviving a sampling period and not being captured.

- $e^{-\mu T_r}(1 - e^{-\lambda T_r})$ is the probability of surviving a sampling period and being captured.
- $\frac{\mu}{\lambda+\mu}(1 - e^{-(\lambda+\mu)T_r})$ is the probability of dying without being captured during a sampling period
- $\frac{\lambda}{\lambda+\mu}(1-e^{-(\lambda+\mu)T_r})-e^{-\mu T_r}(1-e^{-\lambda T_r})$ is the probability of dying after a recapture event but within the sampling period.

In addition, the following probabilities for surviving the release interval for newly marked individuals is given:

$$\text{Prob} = \begin{cases} e^{-\mu T_r} & \text{when releasing the batch at the beginning of the interval} \\ e^{-\frac{\mu}{2}T_r} & \text{when releasing the individuals at a constant rate during the interval} \\ 1 & \text{when releasing the batch at the end of the interval} \end{cases}$$

We may now compute the expected mark-recapture data to fit into the CJS model. Note that the true survival probability of one cycle is given by $e^{-\mu}$. We are thus able to compute the bias of the CJS survival estimates.

When we compute the data according to the listed probabilities, four factors may be varied upon: (1) λ, (2) μ, (3) the length of T_r and (4) the release strategy of the newly marked individuals. These factors are varied according to:

1. For the mortality rate μ,three levels $\mu \in \{1, 1/2, 1/3\}$ is applied. These levels correspond to survival probabilities of 0.36, 0,60 and 0.71.
2. For the length T_r of the capture period,regularly spaced values between 0.05 and 1 is used.
3. For the capture rate λ,values such that $\lambda T_r \in \{1, 1/2, 1/3\}$ is applied. This choice corresponds to capture probabilities (with no mortality) of 0.64, 0.4 and 0.29 regardless of the length of the interval.
4. For the newly marked individuals, three different strategies is applied: released at the beginning of the interval, released at the end of the interval, and released at at constant rate.

To summarise, we are using simulated data which meet all the standard CJS model assumptions, except the instantaneous sampling assumption. Note that the data are simulated such that the assumption of immediate release of recaptured individuals is met. A time-dependent CJS model with ten capture occasions is used for analysis. The mean of the nine survival estimates is computed to assess the bias.

3.1 Results

We now consider three figures which illustrate the CJS survival estimates. Let us begin with Figure 3. It shows results from the situation where the newly marked individuals are released at the beginning of the sampling interval. As anticipated,

the bias increases with the mortality rate and the sampling interval length. Less expected is the correlation between the bias and the sampling rate. An explanation of this behaviour is given by the amount of information in the data. Less information (less recaptures) gives the model more space to adjust for the induced bias. Note in addition the general underestimation of the survival. This is in accordance with the discussion in Section 2.

Figure 4 illustrates the case where the newly marked individuals are all released at the end of the sampling interval. We notice that the bias is more pronounced for this situation compared to the situation illustrated by Figure 3. As expected from the discussion in Section 2, the survival is overestimated.

The results from the situation where the newly marked individuals are released with a constant rate during the sampling period, are illustrated by Figure 5. The bias is still clearly present, but it is much smaller than for the two other situations. In addition, the CJS model overestimates the survival for this release strategy. This behaviour may not be readily understood, and may need some explanation: Notice that the strategy of releasing individuals at a constant rate corresponds to a batch release at the middle of the sampling interval with regard to the CJS model. With constant capture and mortality rates during the sampling, the median of the capture times lies somewhere in the first half of the sampling interval. Since we expect the sampling occasion time for the CJS model to lie near the median, the corresponding batch release time is situated after the sampling occasion time. According to the discussion in Section 2, this situation leads to an overestimation. Note that the bias is less pronounced than for the situation corresponding to Figure 3, because the batch release time is closer to the sampling occasion time.

4 Discussion

The results in Section 3.1 demonstrate that the CJS survival estimator may be severely biased when the assumption of instantaneous sampling occasions is violated. As anticipated, the mortality rate and the length of the sampling intervals influence the bias. The most interesting result, perhaps, is the influence of the release strategy. When the newly marked individuals are batch released at the beginning or end of the sampling intervals, the bias may be substantial even for low mortality and short sampling intervals.

For the two experiments presented in the introduction, any severely biased estimates of the survival is not anticipated. For the Dipper experiment, the mortality was presumably quite low during the breeding period, and the sampling period was relatively short ($T_r \approx 0.25$). In addition, the release process of the newly marked individuals resembled the recapture process. Note that adverse events took place during some of the sampling periods. Floods and cold weather periods may have affected the mortality, and it is not known how these events influence the estimates.

For the Norwegian coastal cod experiment, some bias may be suspected. The sampling periods were long (they corresponds to $T_r = 1$), and the mortality was high (at least for the young individuals). In addition the newly marked individuals were batch released, but usually around the middle of the sampling intervals. The

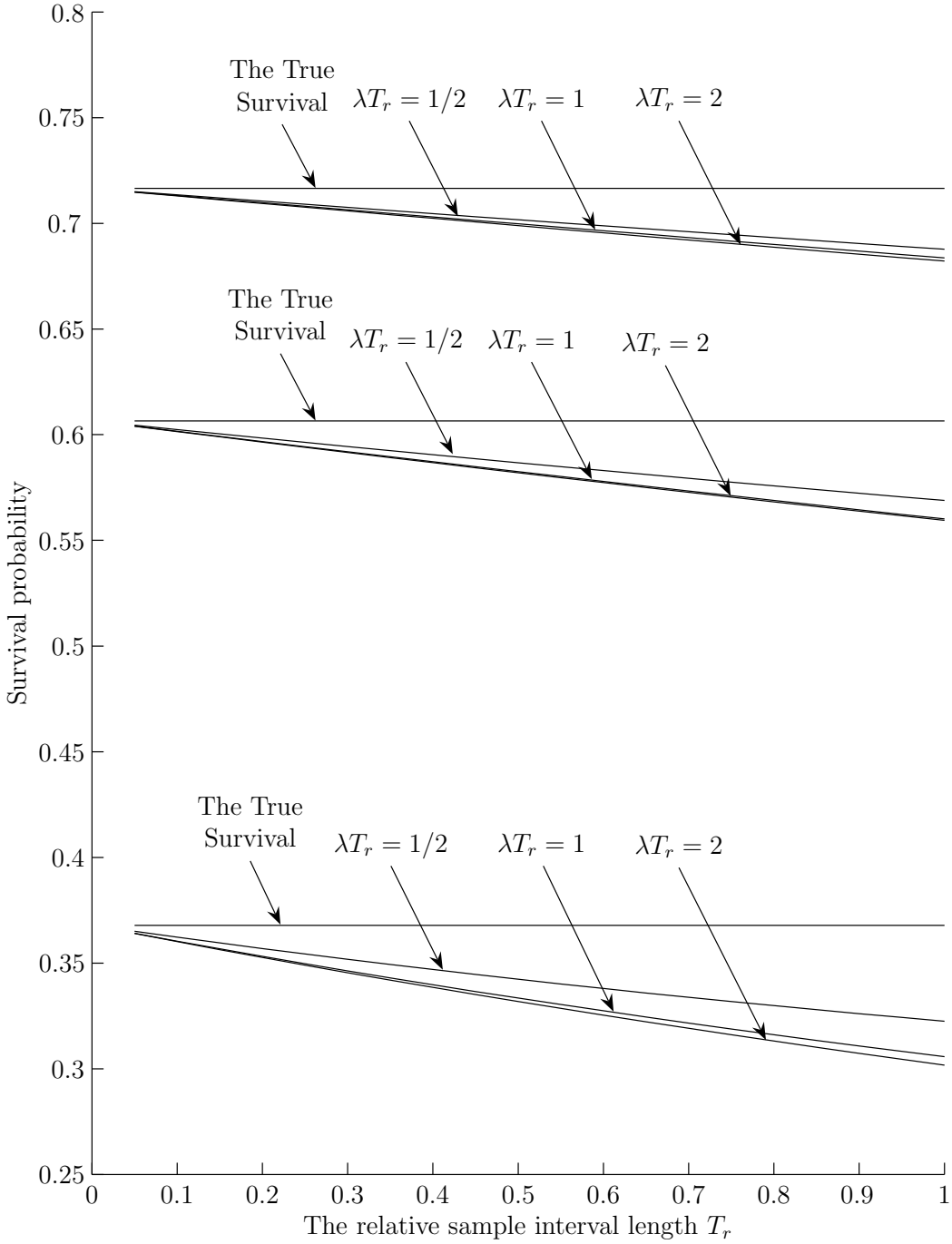


Figure 3: The CJS survival estimates for the situation where the newly marked individuals are batch released at the beginning of the sampling interval

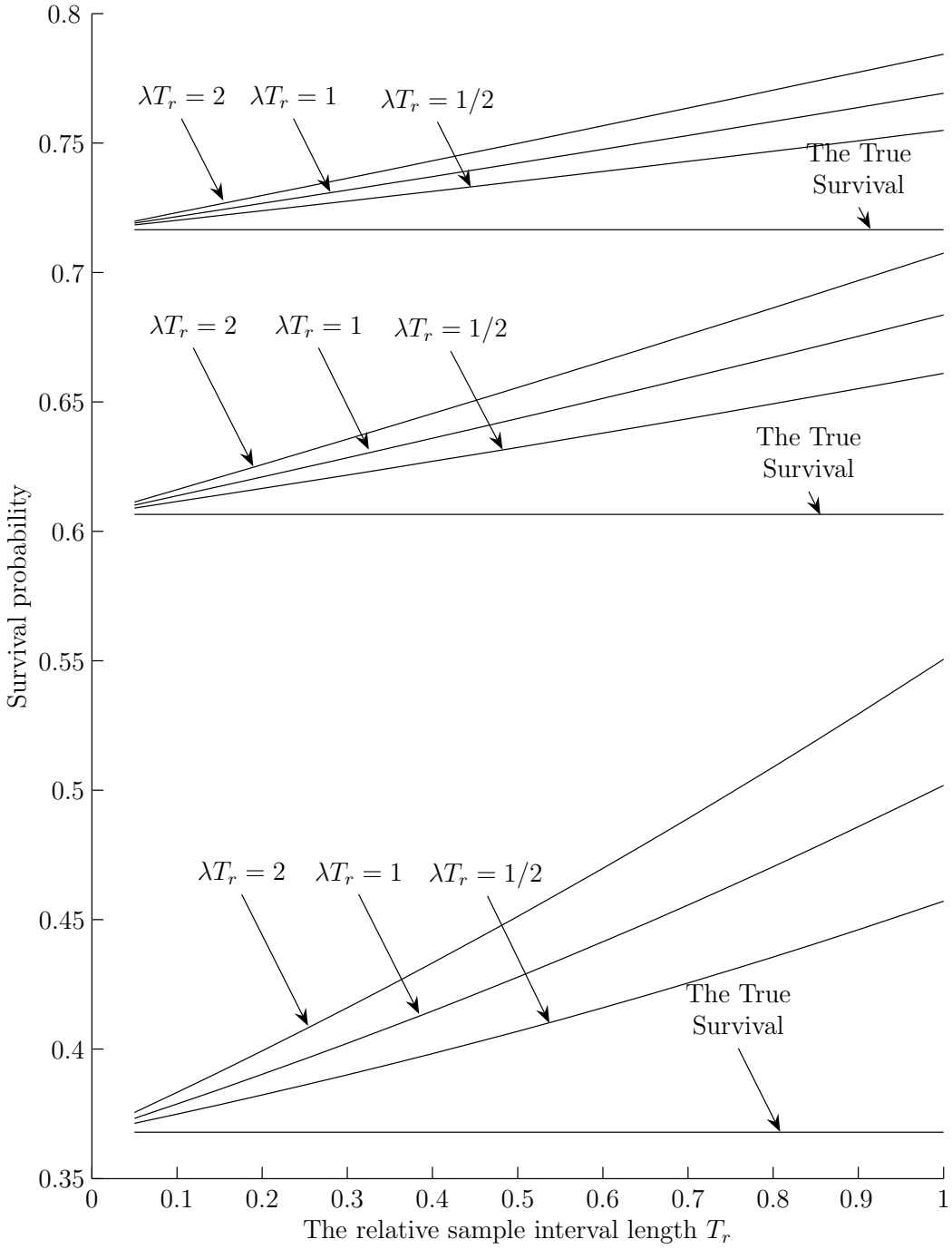


Figure 4: The CJS survival estimates for the situation where the newly marked individuals are batch released at the end of the sampling interval

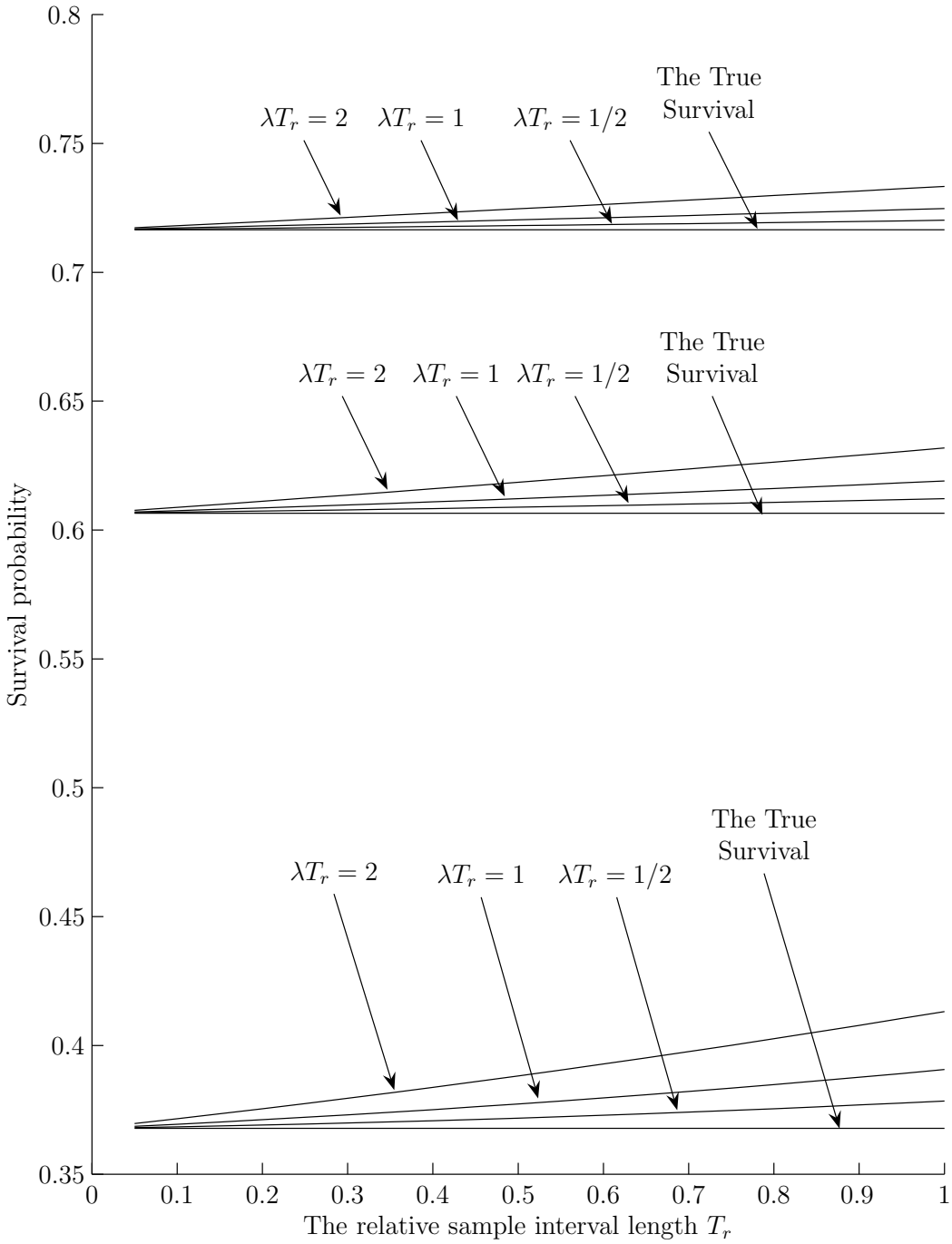


Figure 5: The CJS survival estimates for the situation where the newly marked individuals are released with a constant rate during the sampling period

releases were done only once a year, and not every cycle (month) as in the simulated data analysed here.

Following the analysis performed in this paper, the sampling period of mark-recapture experiments should be as short as possible. When this is not obtainable, and there is considerable mortality during the sampling intervals, it is important that the release process of the newly marked individuals mimics the recapture process. Batch releases at the beginning or end of the sampling intervals are not recommendable.

References

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson, D. S. (1985). *Statistical inference from band-recovery data: a handbook*. Fish and wildlife service, U.S. Department of the Interior, second edition.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika*, 51:429–38.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52:225–47.

Julliard, R., Stenseth, N. C., Gjøsæter, J., Lekve, K., Fromentin, J.-M., and Danielssen, D. S. (2001). Natural mortality and fishing mortality in a costal cod population: a release-recapture experiment. *Ecological Applications*, 11(2).

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Marzolin, G. (2002). Influence of the mating system of the eurasian dipper on sex-spesific local survival rates. *Journal of wildlife management*, 66(4):1023–30.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, pages 249–59.

Smith, D. R. and Anderson, D. R. (1987). Effects of lenghty ringing periods on estimators of annual survival. *Acta Ornithol.*, 23:69–76.

Williams, B. K., Nichols, J. D., and Conroy, M. J. (2002). *Analysis and management of animal populations*. Academic press.

Analysing Continuous Mark-Recapture Data on Open Populations

Inge Christoffer Olsen
Department of Mathematical Sciences,
Norwegian University of Science and Technology

Summary This paper presents a fully analytical EM algorithm for the analysis of the Cormack-Jolly-Seber (CJS) model for mark-recapture data. The algorithm represents an improvement over the existing algorithm (van Deusen (2002)) in that the E step of the provided algorithm is analytical. Nice criteria of convergence are presented, and it is showed how the covariance matrix may be estimated quite easily. The CJS model has been used extensively on discrete mark-recapture data. We see in this paper that the CJS model may be used directly on continuous mark-recapture data, much in the same way as the Kaplan-Meier estimator is used for survival data. To include covariate information, semi-parametric multiplicative hazard models are defined for the capture and mortality process. The EM algorithm is still used for the analysis. The developed methodology is then applied on a set of mark-recapture data from a population of European Dippers in southern France.

Keywords: Capture-recapture; Continuous sampling; Cormack-Jolly-Seber model; Mark-Recapture; EM; Proportional hazard; Survival analysis

1 Introduction

The estimation of survival abilities of animals living in their natural habitat has been given considerable attention over the last century. A central procedure for analysing populations of wild animals is the mark-recapture methodology, by which animals are marked, released and recaptured. The sampling is usually done as follows: At a fixed set of times $t_1, t_2, \ldots, t_k$, captures are made from the population at interest. Each of the captures consists of marked and unmarked individuals. The marked individuals are registered and usually released, while the unmarked individuals are normally marked and released.

There exist numerous models for mark-recapture data. Most of them are discrete, with product-multinomial likelihoods such as the Cormack-Jolly-Seber (CJS) model (Cormack (1964),Jolly (1965),Seber (1965)). There are models for continuously sampled mark-recapture data, but only for closed populations (no births, deaths, emigrations or immigrations); see Wang and Yip (2003) and the references therein.

For some experiments, the discrete model assumption of instantaneous sampling is violated. This may be due to long sampling periods, or that the sampling is done continuously. For the mark-recapture experiment on the European Dipper described in Marzolin (2002), the sampling was performed during the breeding period, which lasted three months. In a survey on the southern coast of Norway, the sampling was done by commercial and recreational fishermen. Thus, the marked individuals were recaptured continuously. When discrete models are used to analyse continuously or semi-continuously sampled data, the survival estimates may be substantially biased (see Olsen (2005)).

In this paper, a model for continuously sampled data on open populations is presented. First, an EM algorithm for the discrete CJS model is presented. This has been done previously (see van Deusen (2002)), but the algorithm presented here is analytical in both the E and the M step, while the algorithm by van Deusen (2002) uses a Newton-Raphson method for the M step. Then, a continuous model for continuously sampled mark-recapture data is presented in Section 3. We show how the EM algorithm developed in the previous section may be used to estimate the capture and mortality functions. In Section 4, semi-parametric multiplicative hazard models are used to include covariate information. Again, the EM algorithm is used in the analysis of the model. The approach is then applied to a set of mark-recapture data on the European Dipper (*Cinclus cinclus*), gathered in eastern France from 1981-1987. These data are analysed discretely in Lebreton et al. (1992) and Marzolin (2002). A Bayesian analysis is performed in Brooks et al. (2000). We compare the results from the discrete analysis with the results from the continuous analysis. The paper is concluded with a discussion in Section 6.

2 The EM algorithm

The aim of this section is to present an EM algorithm for the discrete CJS model. The EM algorithm was developed by Dempster et al. (1977), and we use the general notation used there. The EM algorithm is generally used when missing data complicates the likelihood in terms of estimation. In the case of mark-recapture data, the missing data is the unregistered deaths between the sampling occasions. If we knew these numbers, the estimation of the capture and survival probabilities would be simple. The algorithm consist of two steps. During the E-step, the expected values of the missing data with respect to the observed data and the parameter values from the previous EM iteration are computed. Then the log-likelihood is maximised with respect to the model parameters, conditional on the augmented data from the E-step. This is known as the M-step.

Following Lebreton et al. (1992), the data of a standard mark-recapture experiments may be presented as a table of releases and recaptures as in Table 1. Note that an individual appears at most twice on each line. If an individual is captured and released several times, it appears on several lines. For these data, a time-dependent CJS model may be fitted (Cormack (1964), Jolly (1965), Seber (1965)). This model consists of survival probabilities $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{k-1})$ and capture probabilities $\mathbf{p} = (p_2, \ldots, p_k)$, where ϕ_i is the probability of surviving the interval (t_i, t_{i+1})

t_1	t_2	t_3	$\cdots$	t_{k-1}	t_k
R_1	$m_{1,2}$	$m_{1,3}$	$\cdots$	$m_{1,k-1}$	$m_{1,k}$
	R_2	$m_{2,3}$	$\cdots$	$m_{2,k-1}$	$m_{2,k}$
		R_3	$\cdots$	$m_{3,k-1}$	$m_{3,k}$
			$\ddots$	$\vdots$	$\vdots$
				R_{k-1}	$m_{k-1,k}$

Table 1: Table of mark-recapture data, notation inspired by Lebreton et al. (1992). $R_i,\ i = 1, \ldots, k-1$ is the number of marked and released individuals at time t_i, and $m_{i,j},\ i = 1, \ldots, k-1, j = i, \ldots k$ is the number of marked individuals that are released at capture occasion t_i and not recaptured until capture occasion t_j.

t_1		t_2		t_3	$\cdots$	t_{k-1}		t_k	
R_1	$n_{1,1}$	$m_{1,2}$	$n_{1,2}$	$m_{1,3}$	$\cdots$	$m_{1,k-1}$	$n_{1,k-1}$	$m_{1,k}$	$n_{1,k}$
		R_2	$n_{2,2}$	$m_{2,3}$	$\cdots$	$m_{2,k-1}$	$n_{2,k-1}$	$m_{2,k}$	$n_{2,k}$
				R_3	$\cdots$	$m_{3,k-1}$	$n_{3,k-1}$	$m_{3,k}$	$n_{3,k}$
					$\ddots$	$\vdots$	$\vdots$	$\vdots$	
						R_{k-1}	$n_{k-1,k-1}$	$m_{k-1,k}$	$n_{k-1,k}$

Table 2: Table of augmented mark-recapture data. Here, $n_{i,j}$ is the number of individuals released at time i that did not survive the interval (t_j, t_{j+1}). The remaining notation is as described for Table 1

and p_i is the probability of capture at time t_i. A likelihood based on mark-recapture data is quite easily defined, and traditionally the capture and survival probabilities have been estimated by maximising this likelihood numerically.

We want to establish an EM algorithm to compute the maximum likelihood estimates of the capture and survival probabilities. Let us begin by augmenting the mark-recapture data by the number of individuals that die between each capture occasion (see Table 2). The augmented data may be rearranged such that

$$\begin{aligned} d_i^\phi &= \sum_{j=1}^{i} n_{j,i}, \quad i = 1, \ldots, k \\ d_i^p &= \sum_{j=1}^{i-1} m_{j,i+1}, \quad i = 2, \ldots, k \\ r_i^\phi &= \sum_{j=1}^{i} R_i - \sum_{j=1}^{i-1} d_i^\phi - \sum_{j=2}^{i} d_i^p, \quad i = 1, \ldots, k \\ r_i^p &= r_i^\phi - d_i^\phi, \quad i = 1, \ldots, k-1 \end{aligned} \tag{1}$$

where r_i^ϕ is the number of individuals at risk of dying during the interval (t_i, t_{i+1}), d_i^ϕ is the total number of individuals that did not survive the interval (t_i, t_{i+1}), r_i^p is the number of individuals at risk of being captured at capture occasion t_{i+1} and d_i^p

t_1		t_2		t_3	$\cdots$	t_{k-1}		t_k	
r_1^ϕ	d_1^ϕ	r_2^ϕ	d_2^ϕ	r_3^ϕ	$\cdots$	r_{k-1}^ϕ	d_{k-1}^ϕ	r_k^ϕ	d_k^ϕ
	r_1^p	d_2^p	r_2^p	d_3^p	$\cdots$	d_{k-1}^p	r_{k-1}^p	d_k^p	

Table 3: Table of the rearranged, augmented data.

is the total number of individuals that were captured on capture occasion t_i. With this notation, an augmented mark-recapture data array may be created (see Table 3).

Using sequential conditional binomial distributions, the likelihood given the augmented data reads

$$L(\boldsymbol{\phi}, \mathbf{p}) \propto \prod_{i=1}^{k-1} (1-\phi_i)^{d_i^\phi} \phi_i^{(r_i^\phi - d_i^\phi)} \cdot \prod_{i=2}^{k} p_i^{d_i^p} (1-p_i)^{(r_{i-1}^p - d_i^p)} \tag{2}$$

The maximum likelihood estimates of the capture and mortality probabilities, conditional on the augmented data, are then given by

$$\begin{aligned} \widehat{\phi}_i &= \frac{r_i^\phi - d_i^\phi}{r_i^\phi}, \quad i = 1, \ldots, k-1 \\ \widehat{p}_i &= \frac{d_i^p}{r_{i-1}^p}, \quad i = 2, \ldots, k \end{aligned} \tag{3}$$

These expressions form the M step of our EM algorithm.

For the E step, we need to take the expectation of the log-likelihood of the augmented data, conditional on the observed data $\mathbf{m}$, and the values of the capture and survival probabilities from the previous EM iteration ($\boldsymbol{\phi}_0$ and $\mathbf{p}_0$). That is,

$$\begin{aligned} \mathbb{E}[l(\boldsymbol{\phi}, \mathbf{p}|\mathbf{m}, \mathbf{n})|\boldsymbol{\phi}_0, \mathbf{p}_0, \mathbf{m}] = &\sum_{i=1}^{k-1} \mathbb{E}[d_i^\phi] \log(1-\phi_i) + (\mathbb{E}[r_i^\phi] - \mathbb{E}[d_i^\phi]) \log(\phi_i) \\ &+ \sum_{i=2}^{k} d_i^p \log(p_i) + (\mathbb{E}[r_{i-1}^p] - d_i^p) \log(1-p_i). \end{aligned} \tag{4}$$

Comparing (4) with the expressions in Eq. (1), we see that only $\mathbb{E}[n_{i,j}]$, $i = 1, \ldots k-1$, $j = i, \ldots, k$ needs to be computed. By studying the augmented data matrix of Table 2, we notice that the elements of each line i are multinomial distributed with trial number R_i and cell probabilities depending on the capture and survival probabilities. The conditional joint distribution of the unknown mortality numbers $n_{i,i}, n_{i,i+1}, \ldots n_{i,k}$, given the capture data $m_{i,i+1}, m_{i,i+2}, \ldots m_{i,k}$ are also multinomial with parameters

$$\left(R_i - m_i; \frac{1-\phi_i}{C_i}, \frac{\phi_i(1-p_{i+1})(1-\phi_{i+1})}{C_i}, \ldots, \prod_{j=i}^{k-1} \phi_j (1-p_{j+1}) \cdot \frac{(1-\phi_k)}{C_i} \right),$$

where

$$m_i = \sum_{j=i+1}^{k} m_{i,j}$$

and C_i is the normalising constant (which is easily computed as the sum of the nominators). The expected number of deceased individuals in each interval is then given by the expectation in the multinomial distribution,

$$\begin{aligned} \mathbb{E}[n_{i,i}] &= (R_i - m_i)\frac{(1-\phi_i)}{C_i} \\ \mathbb{E}[n_{i,j}] &= (R_i - m_i)\prod_{k=i}^{j-1}\phi_k(1-p_{k+1})\cdot\frac{(1-\phi_j)}{C_i} \quad j = i+1\ldots k. \end{aligned} \tag{5}$$

One iteration of the EM algorithm is then as given by the following:

E step Compute the expected mortality data given the observed capture data and the current capture and survival parameters $\boldsymbol{\phi}^{(k)}$ and $\mathbf{p}^{(k)}$ using the expressions in Eq. (5)

M step Compute $\boldsymbol{\phi}^{(k+1)}$ and $\mathbf{p}^{(k+1)}$ according to Eq. 3.

This procedure is repeated until some convergence criterion is reached. Such a criterion could be that the difference in parameter value, measured by some norm, drops below a certain level.

The major difference between our EM algorithm, and the one presented by van Deusen (2002) is that we use an analytical approach for the maximisation step, while van Deusen (2002) uses a standard numerical Newton-Raphson method.

2.1 Variance

We estimate the covariance matrix of $\widehat{\boldsymbol{\phi}}$ and $\widehat{\mathbf{p}}$ by the observed information matrix. For the EM algorithm there exists several methods for computing the information matrix. We have implemented two such methods, the SEM algorithm by Meng and Rubin (1991) and the numerical differentiation of the Fisher score vector (denoted NDS) by Jamshidian and Jennrich (2000). Both of these methods are fairly easily implemented, and they both give nice results when the number of parameters is low relative to the amount of data. When the number of parameters is high compared to the amount of data, the SEM algorithm has convergence problems convergence while the NDS algorithm performs well.

3 Continuous data

In this section, a model for mark-recapture data which does not assume sampling is developed. The sample space no longer consists of counts of capture histories, but of release and possibly recapture times of marked individuals. Since it is assumed that

capture may occur at any time between release and death, the model is considered continuous.

First, let the experiment be constricted on the interval $[0, \tau]$. Thus, no marked individuals are released before time zero or registered after time τ. For an individual of the experiment, there are two interesting processes: the capture and the mortality process. Let $\alpha_\phi(t)$ be the mortality rate at time t, and $\alpha_p(t)$ be the capture rate at time t. Note that the mortality process acts as a censoring process on the capture process (a dead individual may not be captured). In case of removal after capture, the capture process acts as a censoring process for the mortality process. Applying standard non-homogeneous Poisson process assumptions, the survival and capture functions for individuals of the experiment are given by

$$S_\phi(t) = \exp\left\{-\int_0^t \alpha_\phi(u)\mathrm{d}u\right\} \quad \text{and} \quad S_p(t) = \exp\left\{-\int_0^t \alpha_p(u)\mathrm{d}u\right\},$$

where $S_\phi(t)$ is the probability of surviving the interval $[0, t]$ and $S_p(t)$ is the probability of not being captured during the interval $[0, t]$. In order to estimate the survival and capture functions, we use a survival analysis approach (see e.g. Cox and Oakes (1984)).

Regarding the CJS model, register that there are no restriction on the number of lines in the corresponding data array Table 1. Ultimately, there may be one line for each release and one column for each recapture. The CJS model likelihood is still defined, and the capture and survival probabilities may still be estimated by the maximum likelihood estimators $\widehat{p}_{i+1}$ and $\widehat{\phi}_i$, $i = 1, \ldots, k-1$. The survival and capture functions may be estimated by

$$\widehat{S}_\phi(t) = \prod_{j:t_{j+1}<t} \widehat{\phi}_j \quad \text{and} \quad \widehat{S}_p(t) = \prod_{j:t_{j+1}<t} \widehat{p}_j$$

The survival and capture functions are more suitable than the probability estimates, since the probability estimates tend to be very small and irregular. Note that the argumentation presented here follows the argumentation for the Kaplan-Meyer estimator of survival data as given by Cox and Oakes (1984).

Note that there ultimately are two parameters for each recapture event. Thus, the number of parameters tend to be very large and the numerical optimisation of the likelihood may become unstable. The EM algorithm presented in Section 2 utilises the information we have on the complete likelihood, and may therefore be more stable. The European Dipper mark-recapture experiment presented in Section 5 had a total of 451 recaptures. In the analysis, the EM algorithm converged quite quickly, while an attempt to optimise the likelihood by the program MARK (see White and Burnham (1999) was unsuccessful.

3.1 Variance

We would like to obtain pointwise confidence limits for the survival and capture function estimates. Kaplan-Meier estimates. In survival analysis the confidence

limits are based on the variance estimates, which is the diagonal vector of the estimated covariance matrix. This is not a good idea in our case, because there is a high correlation structure between the capture and the mortality function estimates. A better approach is to use the bootstrap method (see Efron and Tibshirani (1993)) to assess the uncertainty of our estimates. This is in accordance with the analysis and discussion of Buckland and Garthwaite (1991).

4 Semi-parametric multiplicative hazard models

When analysing wildlife populations, it is often necessary to include covariate information to account for variation in mortality and/or capture rates among the population. Both individual covariates (such as age, weight and size), and environmental covariates (such as weather conditions and food supply) may have a strong influence.

For the CJS model, covariate information is usually included using generalised linear models (see e.g. Lebreton et al. (1992)). This approach is not so well adapted to continuous data. Within survival analysis, semi-parametric multiplicative hazard models are used extensively, and this is the approach we take here.

Let $Z_i(t)$ be the covariate vector of individual i at time t. Then the mortality and capture rates may be defined by

$$\alpha_\phi(t) = \alpha_{0\phi}(t)\exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_i(t)\} \quad \text{and} \quad \alpha_p(t) = \alpha_{0p}(t)\exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_i(t)\},$$

where $\alpha_{0\phi}(t)$ and $\alpha_{0p}(t)$ are the underlying mortality and capture rates, while $\boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$ are known as regression coefficients. In addition, let $A_{0\phi}(t)$ and $A_{0p}(t)$ be the underlying cumulative mortality and capture rates such that

$$A_{0\phi}(t) = \int_0^t \alpha_{0\phi}(u)\mathrm{d}u \quad \text{and} \quad A_{0p}(t) = \exp\int_0^t \alpha_{0p}(u)\mathrm{d}u.$$

The covariate vector must be predictable, meaning that the value of $\mathbf{Z}_i(t)$ must be known just before time t. This restricts the model, since some variable individual information is only known at capture (e.g. weight). Thus, only permanent or slow-changing covariates such as sex or fur colour may be used.

Now, we assume that we know the time of death for each of the n marked individual of the experiment. Let $t_1, t_2, \ldots, t_k$ be the ordered capture times of the experiment. Likewise, let $u_1, u_2, \ldots, u_n$ be the ordered mortality times. Then the complete likelihood is given by

$$\begin{aligned} &L(\alpha_{0\phi}(\cdot), \alpha_{0p}(\cdot), \boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p) = \\ &\prod_{i=1}^{n} \exp\left\{-\int_{u_{i-1}}^{u_i} \mathcal{S}_\phi(\boldsymbol{\beta}_\phi, v)\alpha_{0\phi}(v)\mathrm{d}v\right\} \prod_{j=1}^{n}\Big[\alpha_{0\phi}(u_i)\exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(u_i)\}\Big]^{\Delta N_{\phi j}(u_i)} \\ &\cdot \prod_{i=2}^{k} \exp\left\{-\int_{t_{i-1}}^{t_i} \mathcal{S}_p(\boldsymbol{\beta}_p, v)\alpha_{0p}(v)\mathrm{d}v\right\} \prod_{j=1}^{n}\Big[\alpha_{0p}(t_i)\exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t_i)\}\Big]^{\Delta N_{pj}(t_i)} \end{aligned} \tag{6}$$

where $\Delta N_{\phi j}(t)$ ($\Delta N_{pj}(t)$) is an indicator function which is one if individual j dies (is captured) at time t, and zero otherwise. In addition,

$$\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t) = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t)\} Y_j(t)$$

and

$$\mathcal{S}_p(\boldsymbol{\beta}_p, t) = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_p(t)\} Y_j(t),$$

where $Y_i(t)$ is an indicator for individual i being at risk of capture or death at time t. Thus, $\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t)$ and $\mathcal{S}_p(\boldsymbol{\beta}_p, t)$ are mortality and capture risk sets, weighted with the covariate information. The likelihood and the notation are consistent with standard counting process analysis notation, as defined by ex. Andersen et al. (1993).

Now, $\alpha_{0\phi}(t), \alpha_{0p}(t), \boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$ are estimated by maximum likelihood. For simplicity the procedure is showed only for $\alpha_{\phi 0}(t)$ and $\boldsymbol{\beta}_\phi$, but it is equivalent for $\alpha_{0p}(t)$ and $\boldsymbol{\beta}_p$. For fixed value of $\boldsymbol{\beta}_\phi$, maximisation of the complete likelihood with respect to $\alpha_{0\phi}(u_i)$ leads to

$$\widehat{\alpha}_{0\phi}(u_i, \boldsymbol{\beta}_\phi) = \frac{\sum_{j=1}^{n} \Delta N_{\phi j}(u_i)}{\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, u_i)}. \tag{7}$$

The corresponding estimator of the underlying cumulative hazard rate is given by

$$\widehat{A}_{0\phi}(t, \boldsymbol{\beta}_\phi) = \sum_{i:u_i \le t} \widehat{\alpha}_{0\phi}(u_i, \boldsymbol{\beta}_\phi) \tag{8}$$

and is known as the Nelson-Aalen estimator. The Cox partial likelihood of the mortality regression coefficients is given by

$$L(\boldsymbol{\beta}_\phi) \propto \prod_{i=1}^{n} \prod_{j=1}^{n} \left(\frac{\exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(u_i)\}}{\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, u_i)} \right)^{\Delta N_{\phi j}(u_i)}, \tag{9}$$

and is the result when Eq. (7) is inserted into the complete likelihood (Eq. 6). The value of $\boldsymbol{\beta}_\phi$ which maximises Eq. (9) is denoted $\widehat{\boldsymbol{\beta}}_\phi$. An estimate of $A_{0\phi}(t)$ is given by $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}}_\phi)$. This estimator is known as the Breslow estimator, and is a nonparametric maximum likelihood estimator of $A_{0\phi}(t)$.

We now use the EM algorithm to estimate $A_{0\phi}(t)$, $A_{0p}(t)$, $\boldsymbol{\beta}_\phi$ and $\boldsymbol{\beta}_p$ from the capture events of a continuous mark-recapture experiment. The M step of the algorithm is established by the Breslow estimator and the Cox partial likelihood. For the E step, we adapt the procedure of Chapter 2 and estimate the unknown mortality numbers based on the estimated capture and mortality functions. The details are presented in the Appendix.

The m'th iteration of the resulting EM algorithm is then given by:

E step Compute the expected mortality data given the observed capture events and the current estimates $\widehat{\boldsymbol{\beta}}_\phi^{(m)}$, $\widehat{\boldsymbol{\beta}}_p^{(m)}$, $\widehat{A}_{0\phi}^{(m)}(t, \widehat{\boldsymbol{\beta}}_\phi^{(m)})$ and $\widehat{A}_{0p}^{(m)}(t, \widehat{\boldsymbol{\beta}}_p^{(m)})$.

M step Compute the estimated regression coefficients $\widehat{\boldsymbol{\beta}}_{\phi}^{(m+1)}$ and $\widehat{\boldsymbol{\beta}}_{p}^{(m+1)}$ by maximisation of the Cox partial likelihood using some numerical optimiser. Then compute the Breslow estimates $\widehat{A}_{0\phi}^{(m+1)}(t, \widehat{\boldsymbol{\beta}}_{\phi}^{(m+1)})$ and $\widehat{A}_{0p}^{(m+1)}(t, \widehat{\boldsymbol{\beta}}_{p}^{(m+1)})$.

4.1 Variance

The covariance matrix of the regression coefficients estimator $\widehat{\boldsymbol{\beta}}$ is estimated by the inverse observed information matrix. This is given by the Hessian of the partial likelihood. We use the NDS algorithm of Jamshidian and Jennrich (2000) to compute the Hessian. For confidence intervals of the Breslow estimators $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}}_{\phi})$ and $\widehat{A}_{0p}(t, \widehat{\boldsymbol{\beta}}_{p})$ we use the bootstrap method.

5 Example: European Dipper

We illustrate the methods presented in this paper by applying them to a set of mark-recapture data on the European Dipper (*Cinclus cinclus*), gathered by G. Marzolin in Eastern France (see Marzolin (2002)). The data were gathered in 1981 to 1987, and are analysed discretely in Lebreton et al. (1992) (frequentistic) and Brooks et al. (2000) (Bayesian). The sampling of this population was done quite continuously between September and June each year of the study.

The estimates of the survival and capture functions $S_{\phi}(t)$ and $S_{p}(t)$ based on both the discrete and the continuous data are presented in Figure 1. We see that the discrete and continuous estimates of the survival function are quite similar. Note, however, the steep slope of the continuous curve between month 24 and 30. During this period, a major flood occurred and washed away a number of nests. This effect we see quite clearly in the estimate of the continuous data, but it is not revealed in the discrete estimate. Further, we find a period of high mortality between month 44 and 50. During this interval there was a period of very low temperatures, which may be an explanation.

The difference between the discrete and the continuous case is more pronounced for the capture function estimate. The yearly capture probability is estimated much higher for the discrete data than for the continuous data. At first sight, this seems strange since the mortality estimates are almost equal. But this result is in fact reasonable: The Dipper experiment was sampled such that the field worker moved between sampling sites, and seldom visited the same site close in time. Since the European Dipper is quite stationary, recaptures of the same individual close in time were quite rare. When the capture rate is estimated, these "low capture rate"-periods contribute to an underestimation. The discrete estimates of yearly capture probability avoids this problem. The solution is better study design, e.g. by randomisation.

In a final analysis, we want to assess the effect of sex on the mortality and capture rate. For this purpose, a proportional hazard model with sex as a covariate is applied. We use

$$Z_i = \begin{cases} 0 & \text{if individual } i \text{ is a male} \\ 1 & \text{if individual } i \text{ is a female} \end{cases}$$

The resulting Breslow estimates of the baseline mortality and capture cumulative intensities with 500 bootstrap replications are given in Figure 2. Note that the two high mortality periods are still present. We find $\widehat{\beta}_\phi = 0.103$ and $\widehat{\beta}_p = 0.180$, in addition to

$$\widehat{\mathrm{Cov}}(\widehat{\beta}_\phi, \widehat{\beta}_p) = \begin{bmatrix} 0.0130 & 0.0047 \\ 0.0043 & 0.0142 \end{bmatrix}.$$

A Wald test statistic for the hypothesis that $(\beta_\phi, \beta_p) = (0, 0)$ equals 2.625, which means that we have no reason to reject this hypothesis. This is in accordance with the analysis in Lebreton et al. (1992).

6 Discussion

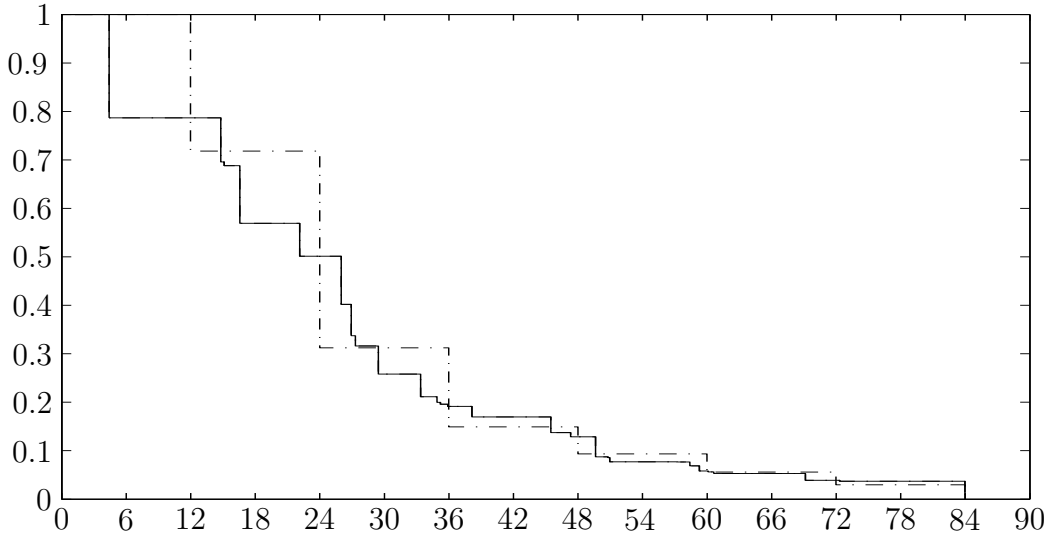
In this paper, three new contributions is identified: the development of an EM algorithm for the CJS model; the extension of the CJS model to continuous data; and the use Cox proportional hazard models on mark-recapture data through the EM algorithm. We discuss these contributions separately.

EM algorithm The proposed EM algorithm is analytical in both the E step and the M step. This is an improvement compared to the EM algorithm of van Deusen (2002), who uses a Newton-Raphson method in the M step. Compared to the standard likelihood optimising algorithm, we have seen behaviour suggesting that the EM algorithm may be more stable when the number of parameters is large.

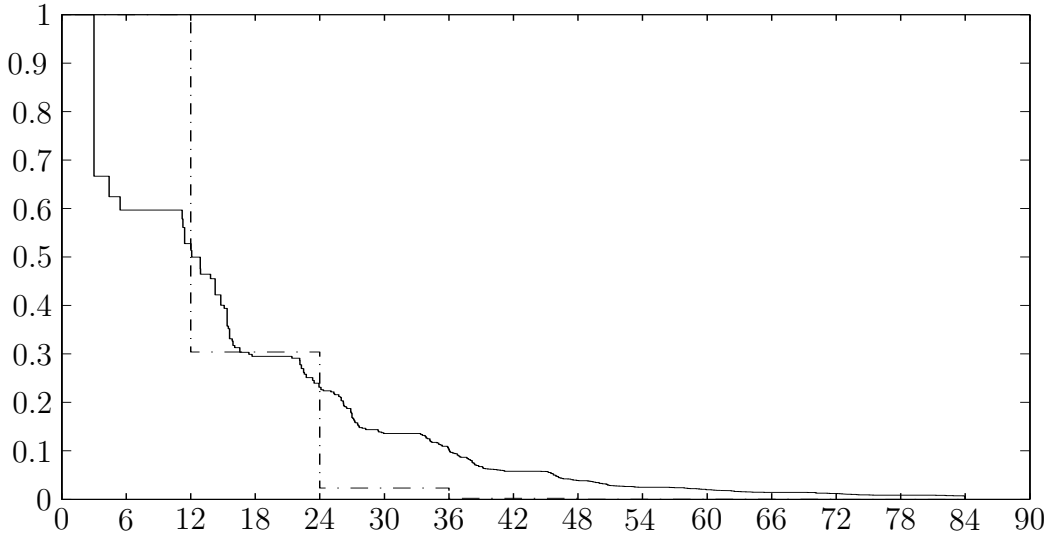
Continuous data When mark-recapture data is continuous, the usual approach has been to discretise the data to fit them into a standard CJS model. We have shown that this is unnecessary, since the CJS model may be used directly on continuous data. In this case, it is better to use an estimator of the capture and mortality functions..We show by example that this approach may reveal time effects more precisely than the discrete approach (i.e. the flood effect of the Dipper experiment).

It is believed that this approach may prove beneficial both in a retrospective and prospective sense. Retrospective, because already continuously sampled data may be reanalysed using continuous models. Prospective, because sampling may now be done continuously or semi-continuously when this is appropriate.

Semi-parametric multiplicative hazard models Semi-parametric multiplicative hazard models are much used in survival analysis to include covariate information. These models are introduced to continuous mark-recapture models through the EM algorithm. Knowing the extensive amount of models related to the proportional hazard model (such as frailty models and nonhomogeneous Markov processes), it is believed that the established link may be of importance to mark-recapture analysis.

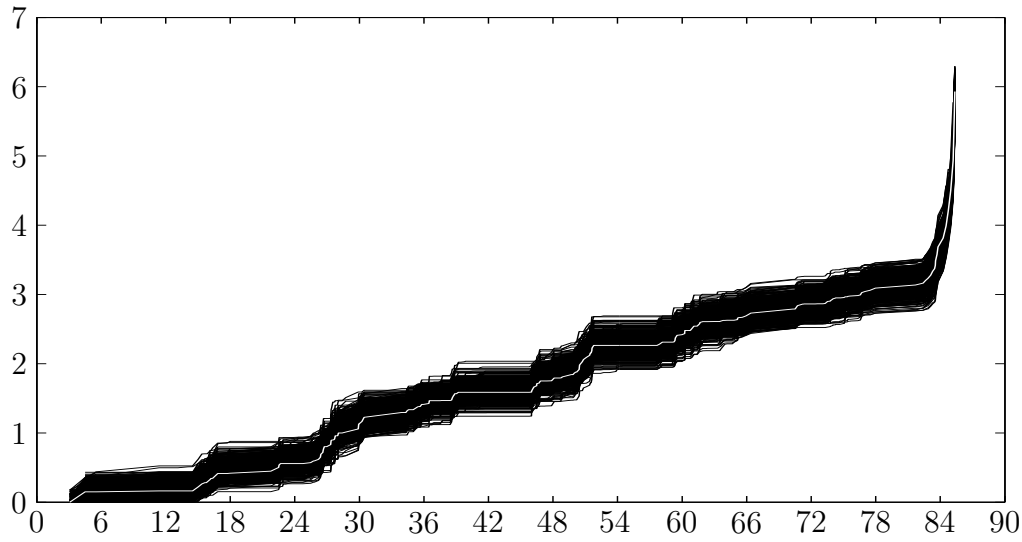


(a) Cumulative Mortality Rate

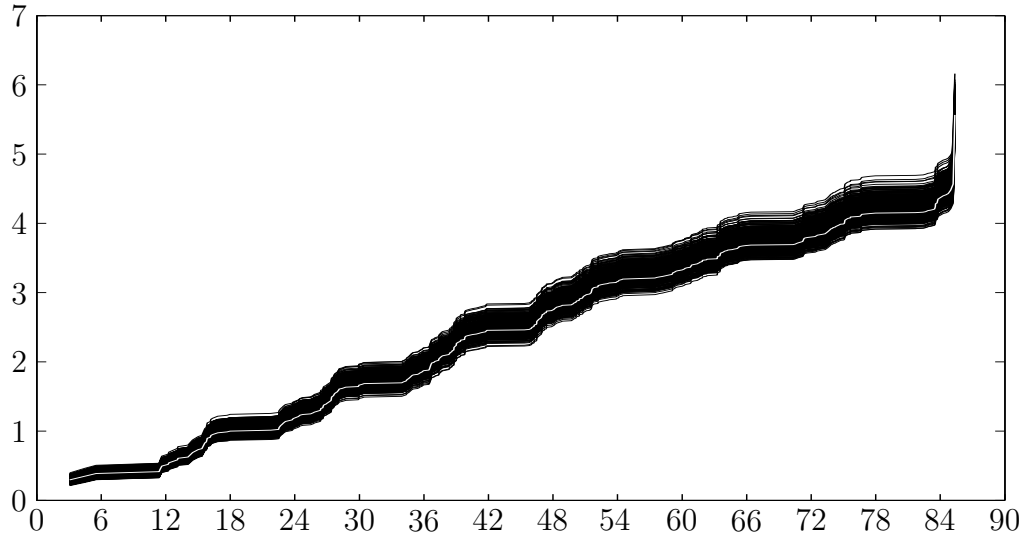


(b) Cumulative Capture Rate

Figure 1: The Kaplan-Meier estimates of the survival and capture functions based on the discrete ($- \cdot -$) and the continuous (—) data. The time scale is in months after 1 January, 1981



(a) Cumulative Mortality Rate



(b) Cumulative Capture Rate

Figure 2: The Breslow EM estimates of the baseline cumulative mortality and capture rates (white), with 500 bootstrap replications. The time scale is in months after 1 January, 1981

References

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer-Verlag.

Brooks, S., Catchpole, E., and Morgan, B. (2000). Bayesian animal survival estimation. *Statistical Science*, 15(4):357–76.

Buckland, S. T. and Garthwaite, P. H. (1991). Quatifying precision of mark-recapture estimates using the bootstrap and related methods. *Biometrics*, 47:255–68.

Cormack, R. M. (1964). Estimates of survival from the sighting of marked animals. *Biometrika*, 51:429–38.

Cox, D. R. and Oakes, D. (1984). *Analysis of surivival data*. Chapman and Hall.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistics Society Series B*, 39:1–38.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Jamshidian, M. and Jennrich, R. I. (2000). Standard errors for EM estimation. *Journal of the Royal Statistical Society, Series B*, 62:257–70.

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, 52:225–47.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992). Model survival and testing biological hypothesis using marked animals: a unified approach with case studies. *Ecological Monographs*, 62:67–118.

Marzolin, G. (2002). Influence of the mating system of the eurasian dipper on sexspesific local survival rates. *Journal of wildlife management*, 66(4):1023–30.

Meng, X.-L. and Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of the American Statistical Association*, 86:899–909.

Olsen, I. C. (2005). Assessing the effects of lengthy sampling periods on mark-recapture models. Preprint in Statistics 11/05 ,Department of Mathematics, Norwegian University of Technology and Science,Trondheim, Norway.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, pages 249–59.

van Deusen, P. C. (2002). An EM algorithm for capture-recapture estimation. *Environmental and Ecological Statistics*, 9:151–65.

Wang, Y. and Yip, P. S. F. (2003). A semiparametric model for recapture experiments. *Scandinavian journal of statistics*, 30(4):667–76.

White, G. C. and Burnham, K. P. (1999). Program MARK: Survival estimation from populations of marked animals. *Bird Study 46 Supplement*, pages 120–38.

7 Appendix

For the E step, we need to take the expectation of the complete log-likelihood conditional on the observed capture data and the parameters from the previous EM iteration. The complete log-likelihood is given by

$$\begin{aligned} &l(\alpha_{0\phi}(\cdot)\alpha_{0p}(\cdot), \boldsymbol{\beta}_\phi, \boldsymbol{\beta}_p) = \\ &\sum_{i=1}^{n} - \int_{u_{i-1}}^{u_i} \mathcal{S}_\phi(\boldsymbol{\beta}_\phi, v)\alpha_{0\phi}(v)\mathrm{d}v + \sum_{j=1}^{n} \Delta N_{\phi j}(u_i)[\log \alpha_{0\phi}(u_i) + \boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(u_i)] \\ &+ \sum_{i=2}^{k} - \int_{t_{i-1}}^{t_i} \mathcal{S}_p(\boldsymbol{\beta}_p, v)\alpha_{0p}(v)\mathrm{d}v + \sum_{j=1}^{n} \Delta N_{pj}(t_i)[\log \alpha_{0p}(t_i) + \boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t_i)] \end{aligned} \tag{10}$$

Computing the expectation of this expression is equivalent with computing the expectation of $\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t), \mathcal{S}_p(\boldsymbol{\beta}_p, u)$ and $\Delta N_{\phi j}(u_i)$. Now,

$$\mathbb{E}[\mathcal{S}_\phi(\boldsymbol{\beta}_\phi, t)] = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_\phi^\intercal \mathbf{Z}_j(t)\}\mathbb{E}[Y_j(t)],$$

and

$$\mathbb{E}[\mathcal{S}_p(\boldsymbol{\beta}_p, t)] = \sum_{j=1}^{n} \exp\{\boldsymbol{\beta}_p^\intercal \mathbf{Z}_j(t)\}\mathbb{E}[Y_j(t)].$$

$Y_j(t)$ is given linearly by the mortality process, and we only need to compute $\mathbb{E}[\Delta N_{\phi j}(t)]$. We compare this situation with the situation for the discrete mark-recapture experiment. Let individual j be released at time t_{i-1} and never seen again. This individual die in either interval (t_{i-1}, t_i), in interval (t_i, t_{i+1}) or in any subsequent interval until the end of the experiment. Thus, we argue that the interval where the individual dies is multinomial distributed. There are no information on the exact time of death within an interval, and we assume that all deaths occur at the end of the intervals.

We denote the assumed mortality time of interval (t_{i-1}, t_i) by t_i^-, $i = 2, \ldots, k$.

The expectation of $\mathbb{E}[\Delta N_{\phi j}(t)]$ is then given by

$$\begin{aligned}
\mathbb{E}[\Delta N_{\phi j}(t_i^-)] &= \frac{(1-\widehat{S}_{\phi j}^{(i-1)}(t_i))}{C_i} \\
\mathbb{E}[\Delta N_{\phi j}(t_{i+1}^-)] &= \frac{\widehat{S}_{pj}^{(i-1)}(t_{i+1})(\widehat{S}_{\phi j}^{(i-1)}(t_i) - \widehat{S}_{\phi j}^{(i-1)}(t_{i+1}))}{C_i} \\
&\vdots \\
\mathbb{E}[\Delta N_{\phi j}(t_k^-)] &= \frac{\widehat{S}_{pj}^{(i-1)}(t_k)(\widehat{S}_{\phi j}^{(i-1)}(t_{k-1}) - \widehat{S}_{\phi j}^{(i-1)}(t_k))}{C_i}
\end{aligned} \tag{11}$$

where $\widehat{S}_{\phi j}^{(i-1)}(t))$ is the estimated probability that individual j released at time t_{i-1} survives the time t and $\widehat{S}_{pj}^{(i-1)}(t))$ is the estimated probability that individual j released at time t_{i-1} is not recaptured by the time t. C_i is the normalising constant. Note that

$$\begin{aligned}
\widehat{S}_{\phi j}^{(i-1)}(t) &= \widehat{S}_{\phi j}(t)/\widehat{S}_{\phi j}(t_{i-1}), \quad t > t_{i-1} \\
\widehat{S}_{pj}^{(i-1)}(t) &= \widehat{S}_{pj}(t)/\widehat{S}_{pj}(t_{i-1}), \quad t > t_{i-1}
\end{aligned}$$

and that

$$\begin{aligned}
\widehat{S}_{\phi j}(t) &= \exp\{-\widehat{A}_{\phi j}(t, \widehat{\boldsymbol{\beta}}_\phi)\}, \\
\widehat{S}_{pj}(t) &= \exp\{-\widehat{A}_{pj}(t, \widehat{\boldsymbol{\beta}}_p)\}.
\end{aligned}$$

Here $\widehat{A}_{0\phi}(t, \widehat{\boldsymbol{\beta}}_\phi)$ and $\widehat{A}_{0p}(t, \widehat{\boldsymbol{\beta}}_p)$ are the Breslow estimates from the previous EM iteration.

The assumption that all death events happen at the same time contradicts the basic assumption of counting processes. Thus, when the estimated number of deaths are high compared to the number at risk, the resulting survival estimate may be biased. This may happen when the mortality rate is high and the capture rate is low.