Time to the first medical treatment of farmed salmon to control salmon lice, a survival analysis approach.

by

Ekaterina Jakobsen

Thesis for the degree of

MASTER OF SCIENCE

Statistics and Data Analysis (MOD5960)

Department of Mathematics
Faculty of Mathematics and Natural Sciences
University of Oslo

May 2016

Preface

I would like to thank my supervisors, Anja Bråthen Kristoffersen and Ingrid Kristine Glad, for an interesting and enjoyable topic. They have always been willing to take the time to answer my questions, but have still made me work independently. They have given me the right pushes throughout the process to dig deeper into theory and learn more.

I have had a desk at the Veterinary Institute for the last year and a half. I appreciate how the institute has included me in different events, such as meetings, Christmas lunch and cakes on Fridays. I also want to thank Peder Jansen at the Veterinary Institute for reading through my thesis several times during the process and providing invaluable help with salmon lice theory and English grammar.

The most important person in my life throughout this period has been my fiancé, Emil. I would like to thank him for being so supportive, patient and comforting when my work has been tough. I appreciate that he has shown interest and understanding, has made good dinners, and has been helpful when I wondered about anything, always.

Lastly I would like to thank my friends and family, for always being there for me.

Oslo, May 2016
Ekaterina Jakobsen

Abstract

The spread of salmon lice has long been and continues to be an increasing problem for the fish farming industry, causing economic and environmental problems. In this thesis, in cooperation with the Norwegian Veterinary Institute, I wished to increase the understanding of this problem. The main interest was analysing what factors that cause lice to appear and require fish to be treated against lice, and whether those factors can be controlled by fish farmers. By using data from the Aquaculture Registry and the Aquaculture Database I was able to construct a dataset including many of the explanatory variables that were essential, and that were worth consideration in regards to appearance of salmon lice.

The modelling approach used in this study has been multiple Cox regression. This method is widely used in survival analyses. I analysed the hazard rates related to first bath treatments against salmon lice at the farms along the Norwegian coast, and used them to model the "survival" times, i.e. time till treatment.

A general Cox model is constructed by using proportional hazard assumptions, and therefore the covariates in a Cox analysis must remain fixed throughout study. However, in real life, many potential explanatory variables that can be obtained from the Aquaculture Database vary with time since the reported measurements are delivered each week. For this reason I first constructed a general Cox model with all covariate values collected at time of stocking, and then extended the constructed model to include time-dependent covariates to explore how time-dependency in the covariates affect the treatment hazard rate and whether the results change.

The results showed that factors mostly addressing to the neighbor's situation (distances, neighbor's lice amount and the infection spread from them) are most significant for time till first bath treatment. Seawater temperature and amount of fish at the farm of interest also appeared to be significant in the analysis, but only when they were allowed to vary each month. I could also see that the lice situations are different in different parts of Norway, and vary from one year to the next. Both constructed models produced similar results, and a validation of the Cox model indicated that the model is proper for the dataset.

Contents

Chapter 1

Introduction

1.1 Background

The fish farming industry has grappled with the increasing problem of sea lice for a long time. With a growing demand for salmon, regarded by many as a delicacy, comes a need for larger and more efficient production.

In Norway, a country with unique access to the sea, the fish farming industry has become one of the big players in the economy with several companies listed on the stock exchange grossing billions of Kroner every year from salmon farming. At the moment Norway is in fact the world's largest producer of bred Atlantic salmon [Fisheries.no 2014]. However, this intensification of salmon farming leads to an increased prevalence of different infectious diseases and parasites, that in turn causes economic and environmental problems. One of these parasites is the salmon louse (*Lepeophtheirus salmonis*), an ecto-parasite living on and off of the surface tissues of salmonid fishes.

1.1.1 What are salmon lice?

Salmon lice are in fact not lice, but rather copepods that attach to the fish and use their mouthparts to feed off of mucus and skin. The damage that the copepods cause their salmonid hosts makes the fish more susceptible to infections, and may disturb the hosts osmotic balance. As a result the salmon lose their appetite and become emaciated [Forskning.no 2005]. Should a salmon get too many wounds from the parasite, it might die. Smolts are considered more susceptible to damage from salmon lice. Since smolts are relatively small (<250 grams), they are harmed more severely than adult fish. According to the Institute of Marine Research the tolerance for wild fish is set to 0.1 lice per gram of fish weight for smolts, and to 0.025 lice per gram for fish that weigh more than 250 grams. This corresponds to a maximum of 20 lice for a fish that weighs 200 grams, and to a maximum of 50 lice for a fish that weighs 2 kilos [Helland et al. 2012]. Should the limit be exceeded, the fish is likely to die.

Lice in larval stages drift freely in the sea causing the spread of lice from farmed fish to wild fish and vice versa. Larval stages last up to several weeks depending on seawater temperatures [Stien et al. 2005]. The lice go through non-parasitic and planktonic nauplistages, and attach themselves to a fish host when they reach their copepodid stage. Once attached, the parasites go through two sessile stages followed by two pre-adult and mobile stages, and finally develop into adult males or females. The time it takes for the lice to develop depends on seawater temperatures. Adult females will start producing egg strings, which look like two tails trailing behind the genital complex. Egg strings may be longer than the lice themselves and may contain between 120 and 190 eggs per string. One female produces between 6 and 11 pair of egg strings in a lifetime [Heuch, Nordhagen, and Schram 2000].

1.1.2 Parasite hosts

As already mentioned salmon lice depend on a host to stay alive and produce eggs. They attach themselves to salmon of any age and size, but they are likely to cause more harm to younger fish.

The marine phase of fish farming starts by releasing juvenile smolts into net-pens. In this study I distinguish between spring and fall cohorts. Spring cohorts are hereafter defined as smolts stocked between February and July, and fall cohorts are stocked between August and January. The smallest smolts weigh as little as 40 grams when they are released. The average weight of the smolts in the dataset used in this thesis is 140 grams. After stocking the fish are left to grow until they reach between four and six kilos. This usually takes between 18 and 24 months. If the fish are meant to be broodstock, farmers may keep them up to three and a half years. They may then reach a weight of up to 20 kilos. Some incidents can prevent farmers from keeping the fish in the net-pens as long as they want, meaning that fish may be slaughtered prematurely, or relocated. After the salmonids are harvested, the net-pens are required to be fallowed for a period of time before a new cohort of smolts may be stocked [Bernhoft and Fardal 2007].

1.1.3 Regulatory requirements

To prevent spread of farm-bred parasites into the wild, it is important to minimise the number of female lice. The Food Safety Authority in Norway (FSA) requires that the number of female lice on a farm at any time is less than 0.5 per fish [*Forskrift om bekjempelse av lakselus i akvakulturanlegg* 2013]. Treatments are required if the average approaches this limit. In practice farmers are reluctant since treatments are expensive. The fish are also expected to be more susceptible to other diseases right after a treatment. Repeated treatments may also lead to the lice becoming resistant to treatment [Jansen et al. 2016]. All of this leads to farmers pushing the legal boundaries. Delousing is also required by law if the total number of pre-adult and adult male[1] and female salmon lice at a farm exceeds 0.1 in spring time [*Forskrift om bekjempelse av lakselus i akvakulturanlegg* 2013]. This is because wild salmon smolts move from rivers out into the sea in that time-period. Salmon farmers are responsible for the number of salmon lice not exceeding the limit. If the limit is exceeded greatly and farmers take no action, the FSA may require a premature harvest, which is very costly for farmers.

[1] Pre-adult and adult male is abbreviated PAAM

In order to maintain these requirements, the FSA monitors fish farms. The regulation *Forskrift om bekjempelse av lakselus i akvakulturanlegg* 2013 states that farmers must report several statistics every week, such as:

- seawater temperature at depth of 3 m
- use of lice treatment (bath treatment, in-feed treatment etc.)
- mean number of female lice
- mean number of PAAM (pre-adult and adult male stages of lice are counted together)
- mean number of larvae

According to the rules, each farm has to report weekly the average number of lice in half of their cages. All cages are therefore counted once every two weeks. The reported numbers must be an average of a minimum of ten fish from each cage. Each month the report also contains number of fish in the farm and average weight of the salmon.

1.1.4 Bath treatment

To keep the number of lice low, farmers use different treatments. One type of treatment is bath submergence. Such a treatment implies adding a medicament to the water to remove lice from the fish. This is done in one of two ways. A watertight cloth is fitted around the net-pens before treatment to separate the water in the treatment unit and the sea, or all the fish are moved into a big tank in a well boat where medicaments are administered [Lusedata 2012].

Bath treatments are time-consuming, and fish farmers sometimes need more than one week to treat their cages. Later when they have to report whether or not they have treated their farm, the same round of treatment might be reported in consecutive weeks. I have omitted this problem by analysing only the time till the first treatment that is applied to the cohort of fish at a farm.

There are two different reasons for initiating treatment. The first is to control lice numbers within maximum legal levels. The second reason is regionally synchronized treatments during spring.

1.2 Problem to be addressed

In this study I will look at time until first bath treatment against salmon lice. I will look only at the time that fish spend in net-pens, meaning the time from release until they are removed for any reason, e.g. for slaughtering. The main problem will be to find out and understand what factors cause salmon lice to appear in the first place, and therefore what factors lead to the need for a first bath treatment after fish stocking. Since treatments are expensive, it is in the farmers' interest to treat as rarely as possible. At the same time they do not want to get a high number of lice and expose themselves to the risk of having to harvest the whole farm.

A survival analysis approach has never been applied to this particular problem, and therefore I am interested in analysing the data by using Cox regression as a first choice. The Norwegian Veterinary Institute has also shown great interest in the results from such an analysis to inform the fish farming industry about risk factors associated with control of salmon lice infections. Another motivation behind calling for a survival analysis tools, is that since I was free to choose how to construct the dataset and look at the time till first treatments, I could also choose how to define time intervals related to the salmon lice treatments. Since not all farms treat their fish (due to relocation, harvest etc.), the observation time related to the treatments is incomplete. A survival analysis approach allows me to model censored observations and thus avoid this problem.

By performing a multiple Cox regression, one gets an indication of how the hazard rates related to time till first bath treatment are affected by the various explanatory variables. The results indicate what explanatory variables help prolong time till first bath treatment, and what cause an early treatment.

1.2.1 Overview of this thesis

The statistical background material that the analyses are based on, will be reviewed in chapter two. I will present theory material about Cox regression models, as well as other topics that should be covered in regards to this study. My modelling approach will be presented at the end of chapter two.

In chapter three I will present a general overview of the data material that might be used in the analysis, as well as some data preparations that had to be made before the analysis, and the information that was obtained when the dataset was constructed.

The final results and the steps performed to obtain an optimal model will be presented in chapter four. Since some data exploration needs to be done before a regression model is constructed, the data examination for the particular model and results related to the data will also be presented in this chapter. The reported results will be presented in relative detail, such that the reader will be able to understand how the final model is constructed. At the end of the chapter, survival curves for different levels of the significant covariates will be presented to help the reader understand better the influence of the covariates on the hazard rater.

The constructed model will be validated in chapter five. The material about model validation methods will be presented first, and then possible validation data will be considered and investigated. Further the validation steps and results will be explained and demonstrated, and finally a conclusion about validation will be made.

Chapter six will contain a discussion and concluding remarks. I will summarise the findings, report some challenges that occurred during the process of the model construction, discuss the results and the advantages and disadvantages of including time-dependency in the analysis, and draw a conclusion. I will also discuss possible improvements and extensions of the constructed model, as well as some alternative models that could be used to perform a similar analysis based on the same data.

The statistical software R will be used in all analyses, and a relevant part of the computer code and output can be found in Appendix. Furthermore an overview of some veterinary terms, that may be useful to know for improved understanding, will be presented in Appendix.

Chapter 2

Methods

The theory material in this chapter is adapted from the book ”Survival and Event History Analysis: A Process Point of View” by Aalen, Borgan, and Gjessing 2008, and from lectures held in the course STK4080: Survival and Event History Analysis at the University of Oslo. Further references will be made subsequently.

2.1 Survival analysis

Survival analysis is used to examine survival data, i.e. to model the time it takes for an event to take place. Usually this event is death, and an individual is observed until its death occurs. This is the origin of the term ”survival analysis”.

In survival analysis, studytime is usually limited, which leads to some incomplete observations referred to as censoring. There are different types of censoring, and the most common one is right-censoring, which occurs when the event of interest is not observed during the studytime, or when the individual is removed from the study and cannot be monitored. To understand censoring better let T_i^* be the time until the event of interest occurs for individual i and let C_i be the right-censoring time, i.e. time at the end of the study. Observed lifetime for individual i will then be $T_i = min(T_i^*, C_i)$, which is time until either the event occurs or the study ends, depending on what happens first. If one let δ_i be an indicator variable for censoring, one get

$$\delta_i = \begin{cases} 1 & \text{if } T_i^* \leq C_i \\ 0 & \text{if } T_i^* > C_i \end{cases}$$

The indicator variable will be 1 in cases where the event actually happens (for example when the individual dies before the censoring time) and 0 when the event is censored.

Another common case in survival analysis is left-truncation, meaning that the individual has already been at risk before entering the study. One example of when left-truncation is needed, is when the time between birth of an individual and occurrence of an event is studied, and the studytime begins on a fixed date, such that some of the individuals entering the study have already been born. To express left-truncation in a mathematical way let $Y_i(t) = \mathrm{I}\{v_i < t \leq T_i\}$ be the risk indicator, which takes value 1 when the individual is at risk just before time t. This means that individual i, which is entering the study at left-truncation time v_i, has not experienced the event of interest before time t.

Regression models in survival analysis specify how hazard rate related to the event of interest depends on explanatory variables. This means that coefficients in a survival regression model relate to hazard. The intensity process, denoting frequency of occurrences of an event for individual i as a function of studytime t, is given by:

$$\lambda_i(t) = Y_i(t)\alpha(t|\mathbf{x}_i), \tag{2.1}$$

where $Y_i(t)$ is the risk indicator as mentioned above, $\alpha(t|\mathbf{x}_i)$ denotes hazard rate and $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})^T$ is a vector of covariates for individual i. Covariates may be time-fixed ($\mathbf{x}_i$) or time-dependent ($\mathbf{x}_i(t)$). To simplify the discussion about covariates, time-dependency will be omitted in the first place. In this study I will consider one special case of the hazard rate:

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i),$$

where $\alpha_0(t)$ is the baseline hazard and $r(\boldsymbol{\beta}, \mathbf{x}_i)$ is called relative risk function, with $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ being regression coefficients. The baseline hazard corresponds to the hazard rate for an individual with $\mathbf{x}_i$ set to reference points, such that the hazard rate equals baseline hazard when $r(\boldsymbol{\beta}, \mathbf{x}_i) = 1$.

2.2 Cox regression

Cox regression is a class of survival models with a specific choice of the relative risk function:

$$\alpha(t|\mathbf{x}_i) = \alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i) = \alpha_0(t)\exp(\boldsymbol{\beta}^T\mathbf{x}_i) = \alpha_0(t)\exp\left(\beta_1 x_{i1} + ... + \beta_p x_{ip}\right). \tag{2.2}$$

This model is also called proportional hazards model. (I will explain the fundamental assumptions for this model later in this chapter.) The hazard function here consists of two parts: non-parametric and parametric. The non-parametric part is the baseline hazard $\alpha_0(t)$, which vary with time, may take any form, and does not depend on covariates. Therefore it is unnecessary to specify it. The parametric part, $\exp(\boldsymbol{\beta}^T\mathbf{x}_i)$, depends on covariates and shows how hazard rate changes as a function of covariates.

To see how a change in one covariate affects the hazard, one can use the hazard ratio, which is calculated easily by considering two individuals with equal components in $\mathbf{x}_1$ and $\mathbf{x}_2$, except for the j-th component, where $x_{2j} = x_{1j} + \Delta$, where $\Delta = 1$ (one unit as defined in the model). One unit can be 1°C, or one factor (male=0, female=1), or when dealing with large numbers and a number with value 1 offers too little information (the difference between 100000 and 100001 is not informative enough) one unit can be defined as for example 10^5 in the model. The change in hazard affected by one units change in the j-th component is then

$$\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \frac{\alpha_0(t)\exp\left(\boldsymbol{\beta}^T(\mathbf{x}_2)\right)}{\alpha_0(t)\exp\left(\boldsymbol{\beta}^T(\mathbf{x}_1)\right)} = \exp\left(\boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1)\right) = \exp(\beta_j).$$

This means that $\exp(\beta_j)$ is the time-independent hazard ratio for the increase of one unit in the j-th covariate, when all other covariates are constant.

A possible extension of the above model is to introduce time-dependency in covariates. The equation 2.2 can then be expressed as:

$$\alpha(t|\mathbf{x}_i(t)) = \alpha_0(t) \exp \left(\sum_{k=1}^{p_1} \beta_k x_{ik} + \sum_{j=1}^{p_2} \delta_j x_{ij}(t) \right).$$

This model is called extended Cox model where p_1 is amount of time-fixed covariates and p_2 is amount of time-dependent covariates. The entire collection of covariates for individual i at time t is denoted by $\mathbf{x}_i(t)$, and hazard rate at time t depends on the value of time-dependent covariates at the same time. Although a covariate can change its value over time, the corresponding coefficient value is constant.

2.2.1 Covariates in Cox regression

As mentioned, covariates can be treated as time-fixed or time-dependent. There are some requirements for covariates used in Cox regression. The main requirement for time-fixed covariates is that the values must be measured in advance (either at time zero, or as history up to time zero), and that they will remain fixed throughout the whole study. Sometimes the covariates are measured multiple times and their values change over time (for example every week or every month). In that cases time-dependency can be introduced for those covariates. Time-dependent covariates too must be known in advance, but since there is time-dependency, the value of a covariate needs to be known just before time t. There are some points one needs to be aware of when time-dependency in the covariates is introduced. Those points will be presented below.

When using time-dependent covariates, it is vital to distinguish between internal and external variables. Internal covariates are associated directly with the individual under study and are related to the event of interest. Changes in their value are generated by the individual itself. External covariates are not related directly to the event of interest. Changes in their values are not caused by the individual itself, since the development is based on external influence. External covariates can either be defined, where their development is given at the beginning of the study (which can be implemented from a time-fixed covariate). Two examples of a defined covariate are the age of an individual which increases linearly as years pass, and a fixed covariate multiplied by a given function of time $x \cdot g(t)$, where the function can be defined as for example $g(t) = t$ or $g(t) = \log(t)$. Another type of external covariates are termed ancillary: "An ancillary time-dependent covariate is the observed path of a stochastic process whose development over time is not influenced by the occurrences of the event being studied. An example of an ancillary time-dependent covariate is the observed level of air pollution" [Aalen, Borgan, and Gjessing 2008].

The main difference between internal and external time-dependent covariates lies in the survival function. In general, the survival function given covariate history is defined by

$$S(t|\mathbf{X}) = P(T > t|\mathbf{x}(t)),$$

where $\mathbf{x}(t)$ is the value of covariates at time t and $\mathbf{X} = \{\mathbf{x}(s) : 0 \leq s \leq t\}$ is the history of the covariate up to time t. For external covariates the survival function for individual i

becomes

$$S(t|\mathbf{X}_i) = \exp\left(-\int_0^t \alpha(s|\mathbf{x}_i(s))ds\right) = \exp\left(-\int_0^t \alpha_0(s)\exp(\boldsymbol{\beta}^T\mathbf{x}_i(s))ds\right),$$

and the usual relationship between survival and hazard function remains.

Internal covariates do not have a direct relation between the hazard function $\alpha(..)$ and survival function $S(..)$. Furthermore, the survival of an individual is required for existence of the corresponding internal covariate. This implies that the survival function for individual i will be

$$S(t|\mathbf{X}_i) = 1,$$

when dealing with internal covariates and provided that $\mathbf{x}(t^-) \neq 0$. It is important to consider that internal covariates may be affected by the event of interest, and that an existing value of an internal covariate may contain information about the failure time [Fisher and Lin 1999; Kalbfleisch and Prentice 2002]. For these reasons, covariates must be chosen carefully to perform a successful analysis.

2.2.2 Estimation

The non-parametric baseline in Cox regression models prevents one from ordinary maximum likelihood estimation. To estimate parameters ($\boldsymbol{\beta}$) it is vital to maximize the partial likelihood. Covariates will be denoted by $\mathbf{x}(t)$ when performing the estimation. A similar approach is used for time-fixed covariates.

Let $N_i(t)$ be a counting process, denoting number of occurrences of an event for individual i as a function of studytime t with intensity process $\lambda_i(t)$, as defined in equation (2.1). The aggregated counting process is then given by $N_\bullet(t) = \sum_{k=1}^n N_k(t)$, which is the total number of occurrences of an event for all individuals as a function of studytime t, and the corresponding intensity process is

$$\lambda_\bullet(t) = \sum_{k=1}^n \lambda_k(t) = \sum_{k=1}^n Y_k(t)\alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_k(t)). \tag{2.3}$$

Considering that the past is given (let $\mathscr{F}_{t-}$ denote the "history" just before time t), and that it is known that an event is observed at time t, one can look at a conditional probability of observing an event at time t for the individual i.

$$\begin{aligned}
\pi(i \mid t) &= P(dN_i(t) = 1 \mid dN_\bullet(t) = 1, \mathscr{F}_{t-}) \\
&= \frac{P(dN_i(t) = 1 \mid \mathscr{F}_{t-})}{P(dN_\bullet(t) = 1 \mid \mathscr{F}_{t-})} \\
&= \frac{\lambda_i(t)}{\lambda_\bullet(t)} \\
&= \frac{Y_i(t)\alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{k=1}^n Y_k(t)\alpha_0(t)r(\boldsymbol{\beta}, \mathbf{x}_k(t))} \\
&= \frac{Y_i(t)r(\boldsymbol{\beta}, \mathbf{x}_i(t))}{\sum_{k=1}^n Y_k(t)r(\boldsymbol{\beta}, \mathbf{x}_k(t))}
\end{aligned}$$

One can now see that the baseline is eliminated from the fraction, and the probability does not depend on it. To compute the partial likelihood function of the number of individuals who have experienced the event (where an individual that has experienced an event at T_j is denoted by i_j), one multiply $\pi(i \mid t)$ over all observed event times $T_1 < T_2 < ...$

$$L(\boldsymbol{\beta}) = \prod_{T_j} \pi(i_j \mid T_j) = \prod_{T_j} \frac{Y_{i_j}(T_j) r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{k=1}^{n} Y_k(T_j) r(\boldsymbol{\beta}, \mathbf{x}_k(T_j))} = \prod_{T_j} \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{k \in \mathscr{R}_j} r(\boldsymbol{\beta}, \mathbf{x}_k(T_j))}, \qquad (2.4)$$

where $\mathscr{R}_j = \{k \mid Y_k(T_j) = 1\}$ is risk set at time T_j. Hence, this partial likelihood is a product of the conditional probabilities of observing an event for individual i_j, who experiences an event at T_j, given the past and given that an event is observed at time T_j.

Further one can continue as in ordinary maximum likelihood estimation, by maximizing the partial likelihood function $L(\boldsymbol{\beta})$ to find $\widehat{\boldsymbol{\beta}}$. Because this function is independent of the baseline hazard, one can find the estimates of the regression coefficients without specifying the baseline. The log-likelihood function becomes

$$\begin{aligned}
\log[L(\boldsymbol{\beta})] &= \sum_{T_j} \log \left[\frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_j))}{\sum_{k \in \mathscr{R}_j} r(\boldsymbol{\beta}, \mathbf{x}_k(T_j))} \right] \\
&= \sum_{T_j} \log \left[\frac{\exp(\boldsymbol{\beta}^T \mathbf{x}_{i_j}(T_j))}{\sum_{k \in \mathscr{R}_j} \exp(\boldsymbol{\beta}^T \mathbf{x}_k(T_j))} \right] \\
&= \sum_{T_j} \boldsymbol{\beta}^T \mathbf{x}_{i_j}(T_j) - \sum_{T_j} \log \left[\sum_{k \in \mathscr{R}_j} \exp(\boldsymbol{\beta}^T \mathbf{x}_k(T_j)) \right]
\end{aligned}$$

Maximizing the log-likelihood function makes it possible to find the estimates of $\boldsymbol{\beta}$. Deriving the function with respect to $\boldsymbol{\beta}$ gives a vector with score functions that can be used to find $\widehat{\boldsymbol{\beta}}$:

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log[L(\boldsymbol{\beta})]$$

The covariance matrix of $\boldsymbol{\beta}$ can be found by using the observation matrix:

$$\mathbf{I}(\boldsymbol{\beta}) = -U'(\boldsymbol{\beta}) = \frac{\partial^2}{\partial \beta_i \beta_j} \log[L(\boldsymbol{\beta})]$$

Maximum partial likelihood estimators has the same properties as ML-estimators. Which means that they are normally distributed around the true value, $\mathrm{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$, with the estimated covariance matrix $\mathbf{I}(\boldsymbol{\beta})^{-1}$.

2.2.3 Stratified Cox-model

In the Cox model 2.2 it is assumed a common baseline for all individuals. Whenever a common baseline is unrealistic, or whenever one do not wish to specify a particular model for some components of $\mathbf{x}$, the population can be grouped into l strata. The hazard rate for individual i in stratum s will then be

$$\alpha_s(t|\mathbf{x}_i(t)) = \alpha_{s0}(t) \exp(\boldsymbol{\beta}^T \mathbf{x}_i(t)),$$

such that effect of covariates $\boldsymbol{\beta}$ are assumed to be the same across strata, while baseline hazard may vary. The baseline hazard may also develop independently over time for each group.

The partial likelihood function for stratum s, which is a part of 2.4, is now as following:

$$L_s(\boldsymbol{\beta}) = \prod_{T_{sj}} \frac{r(\boldsymbol{\beta}, \mathbf{x}_{i_j}(T_{sj}))}{\sum_{k \in \mathscr{R}_{sj}} r(\boldsymbol{\beta}, \mathbf{x}_k(T_{sj}))},$$

where $T_{s1} < T_{s2} < ...$ are observed event times in stratum s, and $\mathscr{R}_{sj}$ is risk set in this stratum at time T_{sj}. The product of partial likelihood functions over each stratum is then:

$$L(\boldsymbol{\beta}) = \prod_{s=1}^{l} L_s(\boldsymbol{\beta}),$$

and $\hat{\boldsymbol{\beta}}$ has the same properties as in a model without stratification.

Stratification variables need to be categorical in the first place or made categorical by grouping. This is because they are used to divide the subjects or observations into a disjoint set of groups. In Fox and Weisberg 2011 the following is written about stratification:

> "Each stratum is permitted to have a different baseline hazard function, while the coefficients of the remaining covariates are assumed to be constant across strata. An advantage of this approach is that we do not have to assume a particular form of interaction between the stratifying covariates and time. A disadvantage is the resulting inability to examine the effects of the stratifying covariates. Stratification is most natural when a covariate takes on only a few distinct values, and when the effect of the stratifying variable is not of direct interest."

2.2.4 Model assumptions

In the Cox regression model one makes no assumptions on the baseline hazard, but there are restrictions about the parametric part when constructing a model with time-fixed covariates. A regular Cox model is also called Cox proportional hazards (PH) model, that is because the following assumptions need to be made before a Cox analysis with time-fixed covariates is carried out:

1. Log-linearity: $\log(\alpha(t|\mathbf{x})) = \log(\alpha_0(t)) + \boldsymbol{\beta}^T\mathbf{x}$
2. Proportional hazards (independent of time): $\frac{\alpha(t|\mathbf{x}_2)}{\alpha(t|\mathbf{x}_1)} = \exp(\boldsymbol{\beta}^T(\mathbf{x}_2 - \mathbf{x}_1))$

The first assumption is about log-linearity in each time-fixed numeric covariate. One way to check for linearity of the covariate i is to fit a penalized smoothing spline for this covariate, which is a smooth curve fitted to the set of noisy observations. The spline is denoted by

$$s(x_i) = \sum_{j=1}^{n} \gamma_j f_j(x_i),$$

where n is number of functions $f(x_i)$ and γ's are corresponding coefficients. The fit of $s(x_i)$ can be performed either by looking at the covariate alone, or by including remaining covariates, assuming that they are log-linear:

$$\alpha(t|\mathbf{x}) = \alpha_0(t)\exp\left(s(x_i) + \boldsymbol{\beta}_{(-i)}^T \mathbf{x}_{(-i)}\right),$$

where $\mathbf{x}_{(-i)}$ are all covariates except the i-th covariate and $\boldsymbol{\beta}_{(-i)}^T$ is the corresponding regression coefficients. The spline estimate should then appear approximately linear. Should it not, the functional form in the parametric part of the model is specified incorrectly, and the covariate needs to be transformed or grouped to be used in the Cox analysis. Looking at the spline will suggest a number of patterns that can indicate how the variable may be transformed, and provides an insight into possible undue influence of outliers [Bellera et al. 2010].

The second assumption is about proportional hazard rates. This assumption focuses on the need for the hazard function to be proportional over time. Meaning that the baseline can be time-dependent, but the explanatory variables can not, such that the hazard function will remain unchanged over time for individual i when x_i changes. One way to check for proportional hazard rates is to add a known and time-dependent function $g(t)$ to each of the explanatory variables and check the null-hypothesis that $\beta_{j2} = 0$, for $j = 1, 2, ...p$, in the following model

$$\alpha(t|\mathbf{x}) = \alpha_0(t)\exp(\beta_{11}x_{i1} + \beta_{12}x_{i1}g(t) + ... + \beta_{p1}x_{ip} + \beta_{p2}x_{ip}g(t)).$$

There are several ways of treating non-proportional hazards in covariates. Two particular methods are to stratify the covariates, or to make them time-dependent for example trough an interaction with time [Fox and Weisberg 2011].

If time-dependent covariates are introduced in the model, Cox regression can still be used. The linearity assumption must still be satisfied, because the model is still on the same form: $\log(\alpha(t|\mathbf{x}(t))) = \log(\alpha_0(t)) + \boldsymbol{\beta}^T\mathbf{x}(t)$. On the other hand the proportional hazard assumption is no longer satisfied for time-dependent covariates because the covariates vary with time, and the model is then called extended Cox model.

2.2.5 Testing for significance of covariates

To test whether the covariate x_{ij} is significant for the Cox model for all i, a Wald test can be used. Because $\hat{\beta}'s$ are approximately normally distributed around the true value, as mentioned in section 2.2.2, the null hypothesis: $H_0 : \beta_j = 0$ can be tested by using the Wald test statistics:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

which is approximately standard normally distributed under the null hypothesis. Extending the testing to significance of all covariates can be done by testing the following null hypothesis: $H_0 : \boldsymbol{\beta} = \mathbf{0}$ using

$$\chi^2 = \hat{\boldsymbol{\beta}}^T \mathbf{I}(\boldsymbol{\beta})^{-1}\hat{\boldsymbol{\beta}},$$

which is approximately chi-squared distributed with p degrees of freedom under the null hypothesis, where p is the length of $\boldsymbol{\beta}$.

2.2.6 Nelson-Aalen estimator

In the cases where the data is incomplete - truncated or censored, Nelson-Aalen estimator can be used to generate a cumulative survival hazard rate function. Cumulative hazard rate can be expressed as following: $A(t) = \int_0^t \alpha(u)du$, where $\alpha(u)$ is the hazard rate at time u.

Note that the decomposition for the counting process $N(t)$, where $dN(t)$ is number of jumps of the intensity process 2.3 in $[t, t + dt)$, is assumed to be 0 or 1.

$$\underbrace{dN(t)}_{observation} = \underbrace{\lambda(t)dt}_{signal} + \underbrace{dM(t)}_{noise} = Y(t)dA(t) + dM(t)$$

Now one estimate the equation for cases when $Y(t) > 0$

$$\begin{aligned} dN(t) &= Y(t)d\hat{A}(t) \\ d\hat{A}(t) &= \frac{dN(s)}{Y(t)} \\ \hat{A}(t) &= \int_0^t \frac{dN(s)}{Y(s)}ds, \end{aligned}$$

such that $\hat{A}(t)$ is Nelson-Aalen estimator, which is non-parametric and estimates the hazard at each distinct time of event as a ratio between number of events and number of individuals at risk. It can be interpreted as expected number of treatments in time-interval $[0, t)$ per individual at risk.

To manage the estimation in cases where $Y(t) = 0$, one can include an indicator $J(t) = I\{Y(t) > 0\}$ and interpret $\frac{0}{0} = 0$. The result is then

$$\hat{A}(t) = \int_0^t \frac{J(s)}{Y(s)}dN(s)ds$$

A plot of the Nelson-Aalen estimator will show trends in the hazard shape. When constructing those plots, a stratification of the covariates of interest is needed such that they are a part of the baseline, since $\hat{A}(t)$ is non-parametric. The plots will then show a graph of how the cumulative hazard rate for different levels of the covariates changes over time.

In this study I will look at the plots of Nelson-Aalen estimator, to investigate how the cumulative hazard rate increases as a function for a covariate divided in different groups. The Nelson-Aalen plot of that covariate will indicate whether there is any difference between the event-times at different levels of the covariate. If the graphs are clearly separated, the difference between the groups is clear. If the graphs cross each other and follow the same trend, there is little difference in how the groups affect hazard.

2.3 Other topics worth mentioning

2.3.1 Variance inflation factors

Variance inflation factor is a measurement of collinearity, i.e. high correlation between the explanatory variables. This factor can be calculated for each explanatory variable x_i in a model by $\frac{1}{1-R_i^2}$. Here R_i^2 is the coefficient of determination in a model where explanatory variable x_i is explained by linear regression of all the other explanatory variables. For example if $i = 1$, R^2 is computed from the following model

$$x_1 = \alpha_2 x_2 + \alpha_3 x_3 + ... + \alpha_n x_n + c_0 + \epsilon$$

where c_0 is a constant term and ϵ is the error term. If all the other explanatory variables contain enough information about x_i, then R^2 will be close to one, leading VIF towards infinity. In cases where VIF is higher than 3 or 5, the variable is considered as collinear [Zuur et al. 2009].

When a variable is transformed, truncated or grouped, VIF is calculated in the same way with x_i containing the adjusted values.

2.3.2 AIC and BIC

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are criteria used for model selection. The two measures estimate the quality of a model relative to other models. AIC and BIC estimate a relative value only and do not suggest whether the quality of a model is high or low. To be able to compare the quality of two models, they need to be constructed from the same dataset. Measurements are calculated using the following formulas:

$$\begin{aligned} \text{AIC} &= 2p - 2\log(L), \\ \text{BIC} &= \log(n) \cdot p - 2\log(L), \end{aligned}$$

where p is number of parameters in the fitted model, n is number of observations and $\log(L)$ is model's log-likelihood. The lowest values of AIC and BIC are preferred.

BIC penalizes model complexity more heavily than AIC, because it takes into account how many observations the dataset contains. Thus BIC will not imply that an overfitted model is the best model, whereas AIC might do exactly that. Lower AIC and BIC indicate on either fewer explanatory variables, a better fit, or both.

2.4 Modelling approach

Before starting the regression analysis, the dataset needs to be constructed and careful data exploration needs to be done. Because the data is obtained from the Aquaculture Database, I have access to a lot of data and many potential covariates, but not all pieces of information may be appropriate for this analysis. I will use the data to construct a Cox proportional hazards (PH) model, and therefore all covariate values need to be known at time $t = 0$. An alternative is to use their mean, max or min value over a known time-interval before time $t = 0$. Because covariates in a Cox PH model are time-fixed, they have to be known in advance and remain fixed throughout the study.

I will examine the relationships between all potential explanatory covariates by checking for correlation and collinearity between the them. The collinearity check is performed by obtaining variance inflation factors (VIF), as described in section 2.3.1. I will use the `corvif` function from Highland Statistics library files in R to see which covariates should be omitted from the analysis before starting a multivariate regression. The function prints a number for each covariate, and this number is the variance inflation factor. I chose to use a cut-off value of 3 to remove collinear variables one at a time to find a set of covariates that do not contain collinearity, i.e. until all VIF values are smaller than 3.

Once I have decided which covariates are appropriate for this analysis, I will examine each explanatory variable by looking at max and min values, mean, median, range interval for 50% of observations for numerical variables, and distribution of categorical variables. This will allow me to see whether any variables have outliers that I would need to take into account, to gain insight into each of the covariates, and to define one unit of each covariate.

Now I can perform a Cox PH analysis. I will start by looking at the null residual plots for each numeric variable. The plot will produce the residuals in a null model (baseline residuals) plotted against the observed values of a covariate and a smoothing line fitted to the observations. Looking at the plot will help me gain insight into how possible outliers in each covariate affect hazard rate and how treatment hazard rate is affected by an increase of each covariate in a univariate model. An interesting thing to keep in mind when observing plots is to see whether the effect of a covariate becomes constant at any time. Should the smoothing line be horizontal before or after reaching a certain covariate value, the influence of those variable-values will neither increase nor decrease. This means that the influence will remain, and I may decide to truncate the variable-values at the point where the effect is constant. Should I observe that outliers pull the curve in a direction opposite to the majority of the observations, I may also consider truncation of a covariate that contain outliers.

Truncation of a variable is limiting the values of a variable above or below, and is done by taking min or max value of each data point x_k and the truncating value a, by $\min(a, x_k)$ or $\max(a, x_k)$.

I will also check whether the numeric variables have the correct functional form, as described in section 2.2.4. Possible truncation of any outliers seen from null residual plot may improve the linearity in a covariate. However I still need to test for the linearity assumption. I will do this by looking at the univariate analysis with one covariate at a time and check the linearity by using `psplines` function from the `survival` library in R. The output will then contain p-values after a Wald test for significance of the smoothing spline term for each covariate, split into linear and non-linear portions. Covariates can be used in the Cox analysis if the non-linear component of the fit is non-significant at the same time as the linear component of the fit is significant. In this analysis I will use significance level $\alpha = 5\%$ to decide whether the covariate is important. Therefore, for acceptance of linearity the p-values after Wald test need to be higher than 5% for the non-linear part of the fit and lower than 5% for the linear part of the fit for each covariate. Should that not be the case, the variable will have to be made more linear by transforming or grouping.

Transformation of a variable is an application of a mathematical function to each point of the variable, meaning that each data point x_{ik} is replaced by a transformed value $v_{ik} = f(x_{ik})$, where f is a function. In this analysis I will apply the following transformations on x_{ik}, for all i, in the equation 2.2, to check whether linearity can be improved:

- Logarithmic: $\log(x_{ik})$
- Exponential: $\exp(x_{ik})$
- Square root: $\sqrt{x_{ik}}$
- Squared: $(x_{ik})^2$
- Cube: $(x_{ik})^3$
- One over: $\frac{1}{x_{ik}}$

Here too I will apply `psplines` function at the univariate model with a transformed covariate and look at the p-values of linear and non-linear parts of the smoothing spline fit. The results will be summarized in a table to facilitate analysis. In the cases where I find that a transformation has improved the variable enough, I will only report those results, excluding the transformations that have not improved the model.

I can also decide to transform a truncated variable if this improves the linearity in the variable. Should all of the above fail, my choice is to group the variable. Grouping of the variable can be performed by looking at the smoothing line and finding the places where the line shows breaks, and then divide the covariate into several groups. The next step is to look at the Nelson-Aalen plots to discover how hazard rates differ in the different groups. I will also test for different hazard rates in the groups by using log-rank test. Should the p-value be lower than 5%, I will assume that the groups differ significantly. By reviewing the results from the log-rank tests and Nelson-Aalen plots, I will adjust the grouping until I have achieved an acceptable result.

Performing a univariate analysis will help me decide which values of the covariates might be more suitable for this analysis. Should I be unable to decide whether to use a value at time $t = 0$ or rather mean, max or min value of the "history" of a covariate in the first place, I will determine which of the above choices in the univariate analysis would be more significant, before I merge the covariates into a multiple analysis. I will also analyse how the hazard rate for treatment is affected by each covariate in the univariate analysis, and examine the p-values to determine whether a covariate is significantly important in the univariate model. I have decided to report univariate results not for every single covariate, but only for a selection of them, as the procedure is much the same.

Forward selection will be used to choose in which order the covariates should enter the model. I will start by looking at the AIC values of every univariate fit. The model with the lowest AIC value will be my first choice, meaning that I have decided to consider the covariate as the most important in a Cox PH model. Next I will select the most important covariate and add the remaining covariates one by one, well aware that linearity may change when covariates are added to the model. Linearity check is required throughout this process to determine whether the covariates remain linear. Should changes in linearity occur, the variables need to be adjusted to become linear. I will continue by choosing, according to AIC, the best model with two covariates, and carry on with the multivariate analysis by including remaining covariates one by one. Note that I have decided not to follow AIC strictly, and I will not exclude covariates from the model based solely on the AIC results.

After including all possible explanatory variables, I will confirm that transformed and truncated values do not correlate with the other explanatory variables, by using VIF one more time. The motivation behind this procedure is that the correlation and VIF are linear dependencies, and if a covariate is transformed or grouped, the collinearity needs to be reconsidered.

Further, I will check for interactions. BIC penalizes the model complexity more heavily, and since the dataset is huge, I will use BIC as a criterion for adding interactions to avoid ending up with unnecessary many.

In the final model I will check whether the proportional hazard assumption is satisfied, by using `cox.zph` function from `survival` library in R. The output provided will show results after a tests of proportionality of the explanatory variables in the constructed model. The test is performed by creating interactions with time using a specified transformation of time, as described in section 2.2.4.

Once I have performed the above analysis, I will expand the model to the extended Cox model to deepen insights further. Many of the variables are measured once per week or once per month, meaning that they change their value over time. I will consider whether to expand time-varying covariates as ancillary time-dependent. Additionally I will decide the time-points at which their values should be obtained: just before time t, or their mean, max or min values over a known time-interval just before time t. I will also distinguish carefully between internal and external covariates. Results from the analysis that includes time-dependent covariates will be compared to results from the analysis that includes time-fixed covariates.

Chapter 3

Material

3.1 General introduction to the data material

The data for this study are obtained either from the Aquaculture register (Norwegian Directorate for Fisheries), containing fixed farm concession data, or the Aquaculture Database, containing farm reported data.

The Aquaculture Database contains huge amounts of historical data stored in different files. Every file contains information about approximately 1550 salmon farms. Each farm has a unique identifier, and reports date back to the year of 2002. In this study I will use data recorded between January 2012 and December 2014 since a new reporting system was introduced in 2012 and farmers started reporting lice data on a weekly basis. Not all 1550 farms mentioned above were active in the period of interest, and therefore I will only use reports from farms that contain at least one fish between January 2012 and December 2014. This makes 823 active Norwegian marine fish farms, and gives 80 658 actual reports for each factor reported on a weekly basis when a farm has at least one fish. It also makes 18 623 observations for each factor reported on a monthly basis.

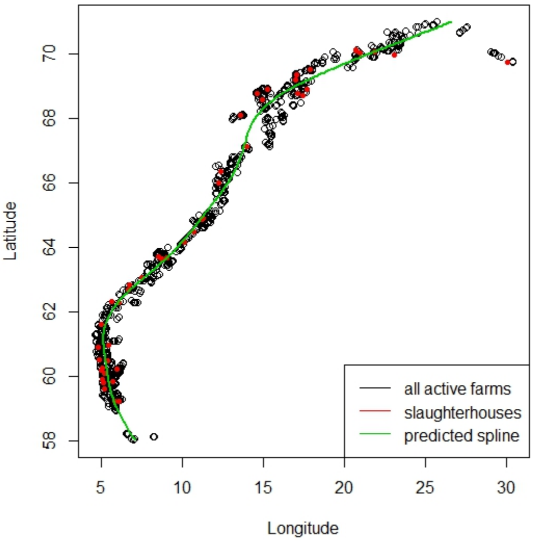
I will use both reported data and modelled data in this study. By modelled data I mean quantities that are calculated from observations. For example local biomass was calculated as the number of fish multiplied by mean weight, meaning the weight of the entire population of fish at the farm. Because number of fish and their weight were reported monthly, biomass is also stored on a monthly basis. Neighbor biomass density is likewise calculated on a monthly level in Jansen et al. 2012, as "a kernel density of stock biomasses within 40 km seaway distances of given farms". This variable examines biomass of surrounding localities, excluding the farm of interest and by weighting the distance to all neighbor farms. Furthermore location density (farm density) was also calculated on a monthly basis, as weighted number of active localities within a 40 km distance. Should the location density equal 0, there are no other farms within a 40 km distance of the given farm, and when the density increases, the amount of other farms nearby increases. However this variable is also a weighted variable, meaning that one neighboring farm very close to the farm of interest counts more than one that is further away.

One of the files contains seawater temperatures, which are measured at a depth of 3 meters every week. In Jansen et al. 2012 temperature was discovered to be a highly significant factor for the spread of lice, and therefore it is important to include seawater temperature in this study. In fact low water temperatures have proven to inhibit lice reproduction, whereas higher temperatures have proven the opposite: female lice produce eggstrings more rapidly which in turn leads to more eggs and a greater spread.

Farmers practice different types of treatments on their fish, and the response variable in this study will be time from stocking to first bath treatment. However there are other types of treatments that might be appropriate to consider here. In-feed treatment might be applicable as an explanatory variable in a regression analysis. In cases where a farm reports that they have in-feed treated their fish, it means that they have been adding medicaments to the fish food. Farmers perform in-feed treatment because they have experienced a propensity for lice, or because their neighbors have a lice problem. They do it in hope of preventing lice problems in the future. Ideally farmers start in-feed treatments before they have lice, or in the least before the lice situation turns critical. There are different medicaments, and their effect differ. However, I am not interested in distinguishing between different medicaments, and if using in-feed treatments in the study, my only concern will be whether or not the fish have been in-feed treated in the study period.

Seaway distances will also be used in the following analysis. I will look at two types of seaway distances: to the nearest neighbor and to the nearest slaughterhouse, which is a net-pen where fish are stored before they are harvested. Along the coast of Norway there are about 50 slaughterhouses, where the fish may spread the parasites in the neighborhood. By knowing the distances, I am also able to find out who the nearest neighbor is. This allows me to include some information about the nearest neighboring farm in the analysis.

Figure 3.1: Locations of 823 active salmon farms along the coast of Norway, 50 slaughterhouses, and the fitted spline used to derive geographical index.



To capture the effect of where the farms are located, I will use geographical index. An additive model is fitted to the longitude of a locality, with corresponding latitude as a explanatory smoothing term, by using the `gam` function from the `mgcv` package in R. Further, predictions of longitude are produced given the existing localities, creating a fitted spline along the coast of Norway. Latitudes are then ordered from southern to northern, and geographical index for the the southernmost farms is set to 0. For each farm further north, geographical index is computed as previous geographical index plus the following error term ξ_i:

$$\xi_i = \sqrt{(\text{lat}_{i-1} - \text{lat}_i)^2 + (\widehat{\text{lon}}_{i-1} - \widehat{\text{lon}}_i)^2},$$

where lat_i is latitude of the farm i and $\widehat{\text{lon}}_i$ is corresponding predicted longitude. Geographical index is then increasing with increased latitude and instead of depending on east and west, it depends on the neighbors along the coastline.

Another modelled data that I will use in this study is infection pressure, compiled in Kristoffersen et al. 2014. Infection pressure was modelled from abundance of reproductive adult female salmon lice, number of fish at the given farm and the neighboring farms during the given months, as well as seawater temperatures based on reports. It was also used a simplified version of models in Stien et al. 2005 to estimate the production and development of lice on active salmon farms along the Norwegian coast. Infection pressure estimates were divided into internal and external. Internal infection pressure (IIP) refers to infections within the farm of interest, whereas external infection pressure (EIP) refers to infections originating from neighboring farms. Before examining the formula used in computations of infection pressure, let us review a few underlying terms and elements. Total amount of female lice at a farm is computed by $n_{AF} = A_{AF} \cdot n_{\text{fish}}$, where n_{fish} is amount of fish at the farm, and A_{AF} is reported amount of female lice. Daily production of newly hatched larvae from one female lice is computed by $F = 300\text{eggs}/\{41.98/[T - 10 - 41.98 \cdot 0.338]\}^2$, where T is seawater temperature and the denominator refers to development time of eggstrings [Stien et al. 2005]. The nominator is the approximate amount of eggs per female lice. Consequently total daily production of newly hatched larvae at a farm is $F_{\text{tot}} = F \cdot n_{AF}$. Further, larvae need to develop themselves into PAAM. This process takes time and not every single larvae survive. Development was divided into several stages: (i) from eggs hatching to infective stage, set to 35 degree-days (which is the sum of seawater temperatures for each of the 35 days), (ii) from infective stage to successful attachment to a host, set to 4 days (independent of seawater temperature), and (iii) development into the PAAM stage, set to 155 degree-days. Daily mortality rate is then assumed to be 0.17 in development stage (i) and 0.05 in stage (iii). Therefore from Stien et al. 2005, the proportion of hatched eggs that survive the development period (i) is $S_{PI} = (1 - 0.17)^{\Delta t_{PI}}$, and similar for survival of development period (iii) which is $S_{CH} = (1 - 0.05)^{\Delta t_{CH}}$, where Δt_{PI} and Δt_{CH} are amount of days it takes to accumulate 35 degree-days and 155 degree-days respectively. The formula for the estimate of IIP at farm i is then:

$$\begin{aligned}\text{IIP}_{i,\text{day}} = \sum_{\Delta t^*} & n_{AF,i,(\text{day}-\Delta t_{PI,i}-\Delta t_{CH,i}-4)} \cdot F_{i,(\text{day}-\Delta t_{PI,i}-\Delta t_{CH,i}-4)} \cdot \\ & S_{PI,\Delta t_{PI,i}} \cdot S_{CH,\Delta t_{CH,i}},\end{aligned}$$

where Δt^* refers to all timepoints $\Delta t_{PI,i} + \Delta t_{CH,i} + 4$ that contribute with lice on the given day. IIP for one week is then $\text{IIP}_{i,t} = \sum_{\text{day}\in\text{week}} \text{IIP}_{i,\text{day}}$. Computing external infection pressure, requires a variable that captures infection from neighbors. Relative risk of infection of salmon lice between farms i and j is assumed to be

$$\text{RR}_{ij} = \frac{\exp(-1.44 - (d_{ij}^{0.57} - 1)/0.57)}{\exp(-1.44 - (d_{jj}^{0.57} - 1)/0.57)},$$

where d_{ij} is the seaway distance between farms and $d_{jj} = 0$. The total infection pressure at all farms within a distance of 100 km is then

$$\mathrm{IP}_{j,t} = \sum_{\forall i} \mathrm{IIP}_{i,t} \cdot \mathrm{RR}_{i,j}.$$

External infection pressure is then determined by $\mathrm{EIP}_{j,t} = \mathrm{IP}_{j,t} - \mathrm{IIP}_{j,t}$, see Kristoffersen et al. 2014.

3.2 Preparation for the analysis

In this analysis I will use Cox regression to study the time, measured in number of weeks, between stocking of smolts and the first bath treatment. In general survival models it is common to look at the individual i, but here I will look at farm location with fish cohort i. Had all farms stocked their fish only once during this study, I would have denoted farm locations with i, but because some farms stock new cohorts of fish at succeeding times, where the cohorts are independent of each other, I will denote cohorts with i. I have assumed independence between cohorts, even in cases where several cohorts appear on the same locality. This is because the net-pens are required to be fallowed for some time before a new cohort of smolts is stocked, and I assume that this time is enough to prevent lice from surviving. I also assume that there are no factors, other than the ones included in the dataset, that affect survival time. This is probably a crude assumption, but investigation into whether this is the case, is outside the scope of this thesis. Some of the data are recorded monthly (number of fish, their weight, biomass and location density). Because I will analyse time in weeks, I chose to convert all monthly values into weekly values. I have done this by assuming that the data were recorded in the first week of each month, and used linear interpolation to find values for the remaining weeks.

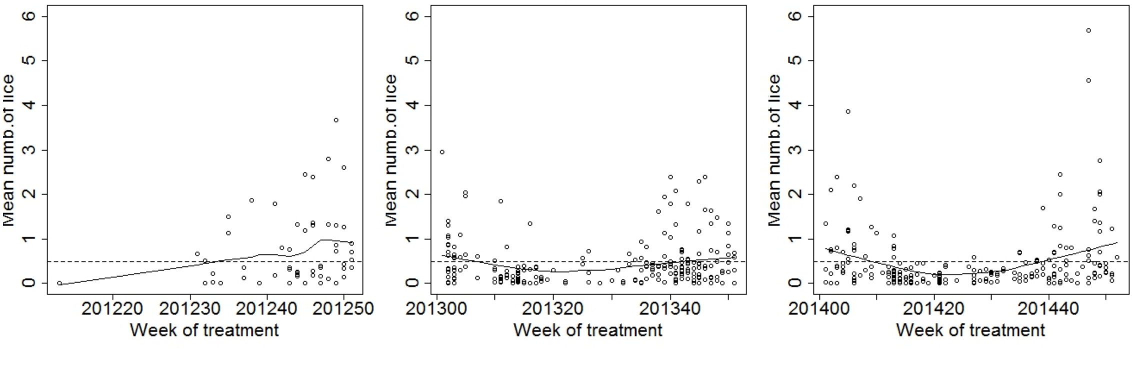
I will review cases with stocking of juvenile smolts at net-pens, excluding cases where the fish were relocated or mixed. Juvenile smolts are defined as being less than 250 grams at the first delivered report. The data are right-censored, which means that for some observations the period of observation, which ends in December 2014, expires before the first treatment occurs, or the fish are removed or harvested without being treated. The data are not left-truncated, meaning that the farms enter the study at the time they stock smolts after January 2012.

I started by constructing the dataset with all possible covariates, and chose to only include fish farms found in all relevant reports, allowing me to have all needed explanatory variables. The analysis was performed on a total of 708 cohorts of fish, where some salmon farms stocked the fish several times. One of the reasons why there are only 708 cohorts from 823 active farms is that I have chosen to look at only the cases where fish cohorts were located for at least six months at the same farm, and where the juvenile smolt weight was lower than 250 grams at the time of stocking. This corresponds to 141 removals due to the above two reasons. Another reason could be that some cohorts had the fish stocked before 2012 and then stopped being active during the observation period. Since I did not wish to have left-truncated data, those cohorts are excluded from the analysis. The total percentage of treated fish (meaning events of interest) during the studytime is 71%, and the percentage of censored events is 29%.

In a preliminary analysis of this dataset, I experienced that the assumptions on which my choice of Cox regression was made, ultimately failed on one point: hazard rates for some of the covariates appeared as non-proportional, and some of the results were difficult to explain. I have chosen to not report on that particular analysis here, and will report on only possible reasons for unsatisfactory results and points of improvement for the dataset. In the preliminary analysis I used all stockings between January 2012 and December 2014 and studied the time until their first treatment. Some of the time-intervals from stocking to first treatment or censoring were up to 100 weeks long. A Cox analysis with time-fixed covariates allows use of only values known at time $t = 0$, and those values need to remain fixed throughout the whole study. Many of the covariates change their values completely over a time-interval which is as long as 100 weeks, and in the worst cases the values at stocking (at time $t = 0$) are radically different from what actually initiates treatment. One possible improvement to this problem, would be to introduce time-dependent covariates, which I will address later. However there are other possible reasons for unexpected and illogical results, which will be explained in the following paragraph.

Sometimes the farmers treat their fish for other reasons than high amounts of lice. To explore this closer, I have plotted treatment date and amounts of lice at the farm one week before treatment. The results are rendered in figure 3.2, which shows that on occasion farmers perform treatments even when amounts of female lice are within FSA limits. Farmers are required to perform delousing in spring, but it is unestablished exactly when spring delousing is supposed to happen at different farms, and this makes it impossible to distinguish those events. Figure 3.2 shows that most of the treatments that occur when the amount of lice is below the FSA limit appear to happen in spring: approximately between week 10 and 20 (March through May). This indicates on spring delousing, which I have decided to keep out of this analysis. For this reason I have chosen to reduce the dataset, and look at stockings between week 14 and week 52 each year (meaning that the first possible treatment will happen in week 18). Censoring time will be week 6 the following year (meaning that treatments which are performed after week 6, will be censored).

Figure 3.2: Week of treatment and number of lice at the farm one week before. Including a line for FSA lice limit equal to 0.5 lice per fish.



This leaves me with 648 cohorts, in which 436 observations (67%) are censored and 212 (33%) are actual treatments. Time-intervals are now smaller because censoring happens sooner. Consequently the amount of censored events is much larger.

3.3 Covariates

Now let us consider each of the possible covariates. A general overview of the possible explanatory variables was presented in section 3.1. As mentioned before the farmers deliver reports monthly and weekly, making the variables vary with time. However, there are also a number of constant variables, such as geographical index and distance covariates. Type of cohort (spring or fall) may also be used as a constant covariate. In the current dataset I distinguish between stockings in 2012, 2013 and 2014, and therefore year can also be used as a categorical covariate in the analysis. This would allow me to examine whether the treatment hazard rate was higher in some years compared to others. Remaining covariates vary with time and need to be chosen carefully when included in the model. As mentioned before, I will perform an analysis in which all of the covariates will be treated as time-fixed. Thereafter I will perform a new analysis with updated values of the covariates that vary with time, expanding the analysis to include time-dependent covariates. When using covariates that vary with time in a Cox PH model, covariate values need to be known at the time of stocking or before stocking, such that they are known at the start of the study. In an extended Cox model the values of those covariates can be updated further, and their values need to be known just before time t.

Fixed variables

Some of the variables have a constant value at every time-point, making them compatible with the analysis without having to consider the time at which they should be obtained. This applies to the following covariates:

- Distance to the nearest slaughterhouse
- Distance to the nearest active neighbor
- Location density
- Type of cohort (spring or fall)
- Geographical index
- Year of stocking

The distance covariates, location density and geographical index are numerical, which means that they require a correct functional form as described in section 1.1.3 when used in both models. The season (spring or fall) and year in which a cohort is stocked are two categorical variables, and therefore they require no linearity check. However, the proportional hazard rates assumption needs to be satisfied for all six covariates in both models.

Should geographical index, due to its non-linearity, appear inappropriate for this analysis, I will group it as done by Jansen et al. 2012 and treat it as categorical variable. When treating the geographical index as a categorical variable, I will distinguish between three groups according to region of Norway: south, mid and north. South-region will be defined as all farms located south of 62 decimal degrees north (WGS 84). Mid-region will be defined as all farms between 62 and 67 decimal degrees north. North-region will be defined as all farms located north of 67 decimal degrees.

For location density and distance covariates I will consider different transformations should the assumption of log-linearity fail, and I will consider grouping only in cases where none of the transformations lead to improvement.

Time-varying variables

The following variables vary with time, and therefore they require closer consideration to determine whether or not they are appropriate for the Cox analyses. Possible time-varying covariates are:

- External infection pressure (EIP)
- Neighbor biomass density
- Seawater temperature
- Lice situation at the nearest active neighbor
- Amount of fish in the cohort
- Weight of fish in the cohort
- Internal infection pressure (IIP)
- In-feed treatment

Other variables that vary with time, and that can be obtained from the Aquaculture Database are several lice-variables, since farmers report several statistics each week, as mentioned in section 1.1.3. Including lice values from the farm of interest would be inappropriate for both a Cox PH model and an extended Cox model. Those covariates have a value of zero at the time of stocking, and are therefore inappropriate as time-fixed covariates. Furthermore, they are directly related to the event of interest when treated as time-dependent. Treatments happen because of high amounts of lice, and including any lice-covariate from the farm of interest will therefore only lead to measuring the same thing twice. The lice situation at the nearest active neighbor, however, can be used in both analyses. I have chosen to use the total amount of female lice at the nearest active farm as a covariate, as the females produce egg-strings and cause the spread of infection. This covariate will be compiled by multiplying the reported mean number of female lice per fish and the number of fish at the nearest active farm. The neighboring farm will be obtained through the information from the distance files obtained from the Aquaculture Database. When the distance to the nearest active neighbor is known, the locality number of the nearest neighbor is known as well, and therefore any information about that farm can be obtained. The lice situation at the neighboring farm may correlate with EIP, and the two covariates may have comparable effects in the model because infection pressure is compiled from lice reproduction.

Another variable that I have access to, but have decided to not use in the analysis is biomass at the farm of interest. Instead I have chosen to use amount and weight of fish as two covariates in the analysis. My goal is that they will provide me with more information than one joint covariate. Whenever I include information about neighbors, I will use their biomass density.

EIP, neighbor biomass density, seawater temperature and neighbors lice situation are external covariates, which means that they are generated by external factors and can be used as both time-fixed and time-dependent explanatory covariates in a Cox analysis. They exist independently of the situation at the farm of interest and can be included at stocking or before stocking in the Cox PH model and at time t or just before time t in the extended Cox model. Since these covariates exist at all time-points, even before studystart, I am able to consider whether their history is more significant than their value at stocking or at time t. This might be the case as appearance and spread of salmon lice to the farm of interest might not necessarily be caused by current seawater temperature and neighbors situation, but rather by the situation in the previous weeks. In cases such as these I will check whether history is more significant in the univariate analysis. I will define the history based on the previous four weeks, which is approximately one month, and will use either mean, max or min value of these four measurements. I will compare the significance of covariates obtained from previous weeks to the significance of covariates obtained at stocking (or at time t in the extended model).

The amount of fish in the cohort can also be used as a time-fixed covariate, and refers to the amount of fish stocked at a farm. The covariate can then be expanded to be treated as a time-dependent variable, because it varies over time. This covariate is not related to the event of interest directly, and is for that reason not an internal covariate in a Cox setting. Ideally I would like to work with a constant number of fish from stocking to first treatment or censoring. In the real world, however, that is not a case, and farmers choose to add, harvest or relocate some of the fish several times during their lifetime.

Mean weight per fish in the cohort increases with time, because the fish grow and gain weight as time passes. The values of this covariate changes at approximately the same rate as studytime, and using weight as a time-dependent covariate will only lead to measuring the time "twice". For this reason, I have decided to define this covariate as time-fixed in both models, measured at the time of stocking.

IIP has a value of zero at the time of stocking, because there is no infection pressure within the farm at the time the fish are stocked. Of this reason IIP can not be used in a Cox PH model. Challenges are also posed when considering IIP in an extended Cox model. An increase in IIP is directly caused by an increased amount of lice at the farm, which in turn leads to treatment. Using IIP as a time-dependent covariate to analyse time to treatment may be somewhat of a self-fulfilling prophecy. For this reason I have chosen to not include IIP as a time-dependent covariate in the analysis, on the basis that it is an internal covariate related directly to the event of interest.

In-feed treatment (whether a farm has been using any in-feed treatment on their cohort) also has a value of zero at the time of stocking. Sometimes smolts receive in-feed treatment before stocking. However, there is not enough information available to track smolt producing farms to the localities in the sea, and therefore the values of this covariate at time 0 is 0. On the other hand in-feed treatment does deserve scrutiny when deciding whether to use it as a time-dependent covariate. Farmers may perform in-feed treatment when they suspect having bad lice situation, making this covariate related to the event of interest. According to manufacturers the effect of in-feed treatment lasts for two months at most, and when the lice situation at a farm has reached a critical state, in-feed treatment will probably be futile. In-feed treatment is used mainly for prevention, not for delousing. By checking the data, I discovered that the mean amount of female lice two weeks before in-feed treatment was larger than 0.2 only in 16 of 497 cases, which is only 3.2%, and larger than 0.4 only in 6 cases. Because the medicaments used for in-feed treatments differ, and some require treatment several weeks in a row to achieve the desired effect, I checked for the amount of lice two weeks before the first in-feed treatment occurred, ignoring cases where next in-feed treatment was performed immediately after the previous. The amount of female lice two weeks before in-feed treatment does not indicate that farmers perform in-feed treatments when they observe large amounts of lice. Farmers do not add medicaments to fish food just before bath treatments, but rather perform the treatment as a prevention based on predictions on the future. By including this covariate as a time-dependent variable, I will be able to determine whether in-feed treatment any time during the study postpones the time to first treatment. I have chosen to include in-feed treatment as a time-dependent categorical covariate in the extended Cox model. Because the medicaments differ, and because treatment is required several times in some cases, I will categorise this covariate in the following way: 0 if there has not been any in-feed treatments just before time t at the farm since stocking, and 1 for minimum one in-feed treatment since stocking.

Chapter 4

Results

4.1 Cox proportional hazards model

4.1.1 Data

As described in section 2.4, careful exploration of the data in the constructed dataset is needed before the analysis is performed. One needs to check whether the covariates correlate with each other. Some of the covariates may be related in ways beyond my knowledge, and including them might lead to explaining the same thing twice. Let us first take a look at the correlation between the potential explanatory variables. I have chosen to report only correlations whose absolute values are higher than 0.6 (see table 4.1). Neighbor biomass density correlates highly with location density. This is understandable since the biomass variable is computed by biomasses as a weighted variable of the neighboring farms within 40 km seaway distances, while location density is a weighted number of active localities within a 40 km distance. Accordingly, the location density somehow contains in the biomass variable, and it is unnecessary to include both of the covariates in an analysis. There is also some correlation between the type of cohort (spring or fall) and temperature at stocking, which is certainly due to the difference in seasonal temperatures. Geographical index and location density correlate, which could be explained by farms being located more closely in the south of Norway, which can also be seen in figure 3.1. When it comes to correlation between EIP and the amount of female lice at the nearest neighbor, the correlation was equal to 0.14, meaning that there is no issue with including both covariates in the model.

Before deciding whether to drop the correlating variables, I will look at their variance inflation factors. In table 4.2 one can see that neighbor biomass density and location density both have VIF values higher than 3. This is caused by their correlation, as seen in table 4.1, and by the correlation between location density and geographical index. Since location

Table 4.1: The highest correlations between some explanatory variables

Biomass neighbor and location density	0.84
Type of cohort and temperature at stocking	0.66
Geo.index and location density	-0.60

Table 4.2: Variance inflation factors for each explanatory variable

	VIF		VIF		VIF
EIP	1.85	number of fish	1.04	weight of fish	1.17
Temperature	2.33	type of cohort	2.21	geo.index	2.04
biomass neighbors	3.66	distance sl	1.31	distance neighbor	1.07
loc. dens	4.98	year	1.16	lice neighbor	1.05

density has the highest VIF, meaning that this covariate is explained by the remaining covariates, I will remove this explanatory variable from the analysis and recalculate the VIF values. After removing it, neighbor biomass density has a VIF value 1.71, geographical index has 1.67 and VIF values for the remaining covariates are still lower than 3. Thus I am ready to perform the analysis with 11 explanatory variables.

Table 4.3 gives a short summary of the explanatory variables that will be used in the further analysis with time-fixed covariates. All values reported in the table are from the time of stocking at the farm of interest. EIP is calculated external infection pressure, which is salmon lice infection from the neighboring farms as described in section 3.1. This covariate gives a relative value computed from the appearance of salmon lice. I have chosen to denote one unit of this covariate as 10^6, since the values of this variable are huge, and this denotation will give a better overview when looking at the hazard ratio further in the analysis. From table 4.3 one can see that the EIP at some farms is 10 times higher than at others, and thus those values need to be explored further and some truncation or transformation might need to be considered. Number of fish in the cohort will be denoted in 10^5, as the smallest cohort is approximately 30 000, and the biggest is approximately 2 350 000. Seawater temperature is measured in degrees Celsius, and vary from 2^oC to 20^oC, such that I believe that there is quite a big difference between summer and winter months, and probably also between regions in the south and north of Norway. Neighbor biomass density

Table 4.3: Description of explanatory time-fixed variables, including units the covariates will be measured in, max and min values, mean, median, range interval for 50% of observations for numerical variables, and distribution of categorical variables.

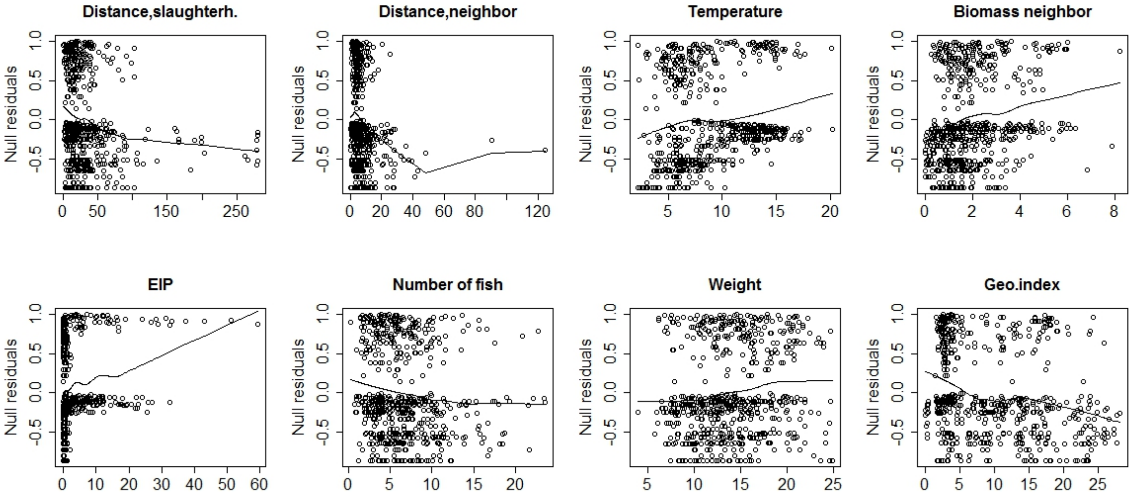
Explanatory variable	Min value	First quantile (25%)	Median	Mean value	Third quantile (75%)	Max value
EIP (10^6)	0.00	0.05	0.35	3.10	2.85	59.30
Number of fish (10^5)	0.30	3.71	5.79	6.52	8.21	23.54
Biomass neighbors (100 t)	0.00	0.96	1.63	1.96	2.60	8.21
Temperature (^{o}C)	2.24	6.00	8.40	9.20	12.71	20.19
Distance: nearest slaughterhouse (km)	1.04	12.01	21.34	33.24	38.76	279.90
Distance: nearest neighbor (km)	0.38	3.06	4.57	7.03	7.93	124.70
Female lice, nearest neighbor (10^4)	0.00	0.00	0.00	3.67	1.90	128.30
Weight of smolts (10 gram)	4.00	10.94	13.59	14.02	16.68	24.95
Geo.index	0.00	2.97	7.21	9.44	14.67	28.21
Type of cohort	Spring:	404 (62.3%)				
	Fall:	244 (37.7%)				
Year	2012:	248 (38.3%)				
	2013:	255 (39.3%)				
	2014:	145 (22.4%)				

is measured in 100 tonnes, and vary between 0 and 800 tonnes. This means that some of the farms might have active neighbors with many big fish close to themselves, while other farms do not have any active farms close by. Seaway distances in kilometres to the nearest active farm and the nearest slaughterhouse will also be included in the analysis. One can see from the table 4.3 that most farms are located within the 40 km seaway distance from a slaughterhouse, but there are also some farms that are located far away, up to 280 km from the nearest slaughterhouse. One can also see that most farms are located within a 8 km seaway distance of each other, but there are also farms that are located within up to 125 km seaway distance of each other. When it comes to the female lice situation at the nearest neighbor one can see that the major part of the covariate is equal to zero. When I studied it more closely, I could see that 65% of the observations of this covariate were equal to 0 at time $t = 0$ in the analysis. However this covariate might be significant in this analysis, even with that many observations equal to 0. Due to the many zeros, I have chosen to categorise this covariate and define group 1 as 0 female lice at the nearest neighbor, and group 2 as more than 0 female lice. In the univariate analysis I will check whether to get the value of this covariate at the study-start, or as history through the past four weeks before study-start in the univariate analysis. Furthermore I will include the mean weight per smolt in the cohort in this analysis to see whether smolt size has any significance on time till first treatment. 50% of the observations of this variable are between 110 and 167 grams, and the heaviest smolts are 250 grams. Thus I have chosen to measure one unit of smolt weight in 10 grams. Geographical index will also be included in the analysis as an explanatory variable, and at first sight it looks like all farms that I am going to work with are evenly spread along the Norwegian coastline. In addition to the explanatory variables mentioned above I will include two categorical variables. One of them is type of cohort: fish that have been stocked during spring (between February and July) or fall (between August and January). Fall season spans over fewer summer months than what spring season does, and it is essential to think that temperature is higher during spring than during fall. However this might not be the case when it comes to seawater temperatures, as the sea has higher heat capacity. After checking the temperatures at stocking, I discovered that max temperature in the spring season was 20.2^{o}C, and 17.6^{o}C in the fall season. However the mean temperature in the spring season was only 7.2^{o}C compared to the mean fall season temperature, which was 12.5^{o}C. Higher temperatures lead to higher lice reproduction, and might lead to a higher amount of lice during upcoming months. There is also a mandatory delousing in spring as mentioned before, which also might lead to lower amounts of lice in that season. Thus I am interested in seeing whether treatments occur more likely when the smolts have been stocked in fall or in spring. The second categorical variable is year of stocking. I use data collected between January 2012 and December 2014, meaning that I can define three factors of this covariate. Since the censoring happens at week 6 the upcoming year, the time-intervals will be of approximately the same length each year. If any years have suffered a bad lice situation, due to for example climate variations, this factor will capture that.

4.1.2 Univariate modelling

To explore each of the covariates more closely, I will start by looking at the null residual plots for each of the numeric covariates with a smoothing spline fitted to the observations. In figure 4.1 one can see that EIP and distance variables have a few outliers. For EIP

Figure 4.1: Residual plot for each numerical covariate, and the smoothing spline.



and distance to the nearest slaughterhouse the outliers do not seem to affect the hazard rate in a "wrong" way. For distance to the nearest active neighbor, one can see that the hazard rate decreases with an increase in distance up to roughly 40 km. After reaching a value of 40 the curve gets pulled up, not necessarily because the hazard starts to increase, but mostly because there are not enough observations after that value (only 5 observations exceed 40 km). For this reason I have chosen to truncate this covariate at 40 km, and define this covariate as distance to the nearest neighbor within 40 km seaway distances. Further one can see that increases in temperature, neighbor biomass density, EIP and smolt weight lead to an increase in the treatment hazard rate, meaning shorter time till first treatment. Increases in distance variables, geographical index and amount of fish at the farm lead to a decrease in the treatment hazard rate, meaning longer time till first treatment. It does not seem reasonable that an increased amount of fish at a farm will lead to longer time till first treatment. One possible explanation of the above result might be as following: when there are a lot of fish at the farm, the amount of lice per fish might be lower than in cases where the number of fish was smaller. This is simply because the lice will have more fish to attach themselves to. However this is a short term effect and should not be practised by farmers because it will lead to a greater spread of lice in the future. One can also see that there is almost no further effect on the treatment hazard rate once the number of fish at a farm reaches approximately one million, which means that I am able to truncate this covariate at 10 if it appears as non-linear. One can also see that the plot of smolt weight shows only a slight increase in the hazard rate, which might mean that an increase in this explanatory variable has only a slight effect on the treatment hazard rate. When it comes to geographical index, the hazard rate seems to be higher for southern locations and lower for northern locations.

I will continue the analysis by checking for log-linearity in the numerical variables in univariate models. In table 4.4 one can find results after log-linearity tests. Some of the explanatory variables appeared as non-linear, and thus I had to try different transforma-

Table 4.4: P-values of linear and non-linear fit of covariates in univariate model using splines, including different transformations. Additionally a decision whether they are appropriate for a univariate Cox regression model.

	Linear	Non-linear		Transformation	Linear	Non-linear	
Smolt weight (10 gram)	< 0.001	0.135	OK				
Distance to slaughterhouse (km)	0.001	0.087	OK				
Distance to neighbor (km)	< 0.001	0.550	OK				
Number of fish (10^5)	0.001	0.655	OK				
Biomass neighbor (100 t)	< 0.001	0.035		$\sqrt{x}$	< 0.001	0.150	OK
Temperature (°C)	< 0.001	0.220	OK				
EIP (10^6)	< 0.001	< 0.001		$\log(x+1)$	< 0.001	0.095	OK
Geo.index	< 0.001	< 0.001					

tions. In the table I only included transformations improving linearity. One can see that smolt weight, distance variables, number of fish and seawater temperature were linear covariates in a univariate model. The remaining covariates had significantly low p-values for the linear part of the smoothing spline fit. However the non-linear part of the fit appeared to be significant too, which means that the variables could not be used in a Cox analysis as they are, and further consideration is needed. Further one can see a result after transformations of neighbor biomass density and EIP, improving the log-linearity. Geographical index was not linear, and was not improved by transformations, meaning that it is likely to behave better if grouped. I have chosen to group this covariate as described in section 4.1.1.

As mentioned in chapter 3 I had to check whether the explanatory variables addressing to EIP, existence of female lice at the nearest farm, neighbor biomass density and seawater temperature should be included at the time of stocking or as history from the previous four weeks. EIP and temperature appeared as most significant in the univariate analysis when included at the time of stocking, according to both the Wald test and AIC. For this reason I will use their value at stocking in the remaining analysis. There was almost no difference between the AIC values addressing to biomass at the neighboring farm, and I have chosen to use its value at time of stocking. The lice-covariate appeared as most significant when included as history from the past four weeks. Thus the distribution of this covariate will be as follows: group 1 (no female lice at the nearest farm in the past four weeks): 355 observations, and group 2 (at least 1 female louse): 293 observations.

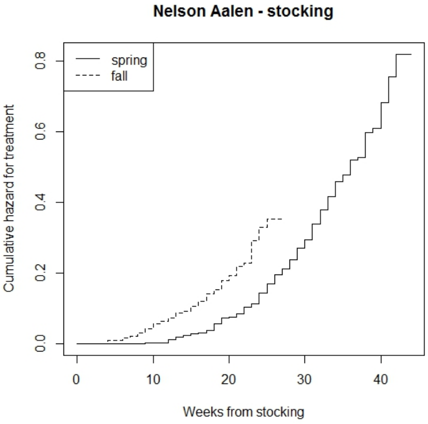
To examine further how all explanatory variables affect the treatment hazard rate and whether the effect is logical, I have studied the univariate models more closely. However, those analyses are similar to each other, and thus I have chosen to report the summary of univariate analyses of one categorical explanatory variable (type of cohort) and one numerical (EIP) in this thesis.

Cohort - spring or fall

Type of cohort is a categorical variable, which takes value 1 if a spring cohort and value 2 if a fall cohort. For the categorical variables it is not necessary to consider linearity and transformations. The only thing that needs to be looked at here is the difference between groups and how the treatment rate changes when the fish are stocked during fall compared to when they are stocked during spring.

From figure 4.2 one can see that the hazard rates in the two groups differ. Graphs do not cross each other, and the hazard rate for first treatment appears to be higher when the smolts were stocked during fall. One can see that there are more observations of treatment for spring cohorts. Only 51 fall cohorts were actually treated, while the treatment amount is 161 for the spring cohorts. However, as one can see in table 4.3, 62% of the observations were spring cohorts, so it makes sense that more of them actually meet the event of interest. Even if there are more spring cohorts, one can see by looking at figure 4.2 that treatments appear more frequently for fall cohorts during the first 25 weeks after stocking. The result after a log-rank test, when testing for difference in those two groups, gave a p-value 2e-05, which indicates that the difference between the two groups is highly significant. One should however keep in mind that for the fall stockings there are fewer actual treatments (21% treatments and 79% censored), and hence the results are more uncertain.

Figure 4.2: Nelson-Aalen plot for the fish cohort covariate.



Despite the uncertainty in this covariate I have chosen to keep it in the analysis to begin with. However, I will have to analyse the results carefully.

If one looks at table 4.5, one can see that when the smolts have been stocked in fall (between August and January) the hazard rate for first treatment is 2.2 times higher than when the stocking was in spring. This result may be explained by spring delousing or higher temperature in spring leading to higher lice reproduction and thus a worse lice situation the following fall. One can see that in a univariate analysis the cohort-type looks highly significant for the model according to both p-value after a Wald test and AIC value compared to AIC in the null model.

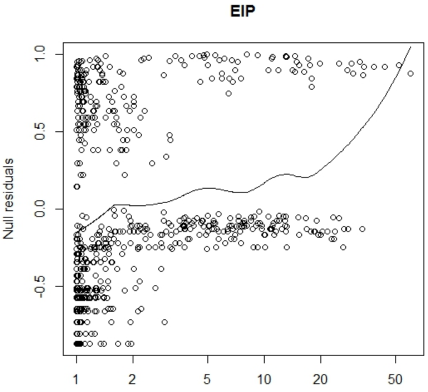
Table 4.5: Summary of a univariate analysis with type of cohort as a categorical covariate. In addition to the values of coefficients and their p-value, the table includes the AIC value of the model, keeping in mind that AIC value in a null model is 2452.82

Explanatory variable	Group	Coefficient	exp(coeff)	exp(-coeff)	p-value	AIC value of the model
Type of cohort *Reference group:* Spring	Fall	0.80	2.22	0.45	< 0.001	2436.37

External infection pressure

EIP is a time-varying covariate with weekly calculated values as described in section 3.1. In the current Cox PH analysis I have chosen to use it as a time-fixed covariate with its value obtained at the time of stocking. By studying table 4.4 one knows that EIP is not a linear variable, and that log-transformation improved the linearity. Since some of the values of this covariate are equal to 0, I needed to add a 1 to all of the observations when transforming, such that the transformation is now expressed by $\log(x+1)$. Remember that one defined unit of this covariate is 10^6, and thus by adding one unit to all of the observations I mean to add 10^6. However since the covariate is modelled as a relative value, this will not lead to any confusions. By looking at the residual plot in figure 4.1 one can see that the hazard rate for first treatment increases as EIP increases, but since I will use log-transformation of this covariate further in the analysis, one can rather look at the residual plot with transformed EIP. In figure 4.3 one can see the residuals in a null model plotted against the log-transformed values of EIP with a smoothing spline fitted to the observations. Main values of the variable are between 0 and 20, and one is able to better see the effect of the small values of this variable in a plot with log-transformation. By looking at the plot one can see that treatment hazard rate at a farm increases as log-transformed EIP increases. One can also see that the treatment hazard increases a lot for the highest values of EIP. The increase is biologically logical since higher infection pressure from neighbors means a greater spread of lice and shorter time till first treatment.

Figure 4.3: Residual plot for log-transformed EIP, and the smoothing spline.



Let us now look at the results after the univariate analysis. From table 4.6 one can see that when EIP increases, the treatment hazard rate increases accordingly. For one unit logarithmic increase in EIP the treatment hazard rate increases 205%. This means that when (x_1+1) increases to $\exp(1)\cdot(x_1+1)$, the hazard rate doubles. To express the relation between this coefficient and treatment hazard rate mathematically, one can calculate the change in the hazard rate caused by a change in EIP, in a univariate model, by:

$$\begin{aligned}\frac{\alpha(x_2)}{\alpha(x_1)} &= \exp\left(0.72\cdot(\log(x_2+1)-\log(x_1+1))\right)\\ &= \frac{\exp(0.72\cdot\log(x_2+1))}{\exp(0.72\cdot\log(x_1+1))} = \left(\frac{x_2+1}{x_1+1}\right)^{0.72}\end{aligned}$$

Table 4.6: Summary of a univariate analysis with time-fixed EIP as a numerical covariate measured in 10^6. In addition to the values of coefficients and their p-value the table includes the AIC value of the model, keeping in mind that AIC value in a null model is 2452.82

Explanatory variable	Transformation	coefficient	exp(coeff)	exp(-coeff)	p-value	AIC value of the model
EIP	$\log(x+1)$	0.72	2.05	0.49	< 0.001	2383.00

To give an example: the treatment hazard rate increases 1.65 times when EIP changes from 0.5 to 2. This was calculated by: $\left(\frac{2+1}{0.5+1}\right)^{0.72} = 1.65$. One can also see that EIP is an important covariate in this model according to both the Wald test and AIC value of the univariate model. AIC decreased with 68 when EIP was included compared to a null model. This is a great improvement.

4.1.3 Multivariate modelling

By comparing the AIC values in the univariate models I could conclude that the explanatory variable addressing to EIP was the most important covariate for the treatment hazard rate with an AIC value of the univariate model equal to 2383. I could also see that AIC values of all univariate analyses were lower than the AIC value of the null model, meaning that all of the covariates were somehow important for the treatment hazard rate univariate. However the covariates addressing to year of stocking and existence of female lice at the nearest neighbor were of less significance, as the AIC values of the univariate models were approximately 2450, while the AIC value of the null model is 2453. This means an improvement of 3 compared to the AIC in a null model.

I will now look at the EIP covariate together with the other covariates in a multivariate model one by one. I could see that the linearity of the variables changed when I went from using the univariate model to using the multivariate model. Once I had adjusted the variables to become linear when put together, I again looked at the AIC values and found that the model with EIP and geographical index was the best one with a value of AIC equal to 2373. By continuing in the same way I ended up with the following model:

$$\begin{aligned}\alpha(\mathbf{x}) = \alpha_0(t)\exp(&\beta_1\log(x_1+1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4\min(x_4, 40) + \beta_5\frac{1}{x_5} + \beta_6 x_6 \\ &+ \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \beta_{14} x_{14},)\end{aligned}$$

where the explanatory variables are as follows, included in order of importance:

- x_1: EIP
- x_2: geographical index, group 2 (mid-region)
- x_3: geographical index, group 3 (north-region)
- x_4: distance to the nearest neighbor
- x_5: distance to the nearest slaughterhouse
- x_6: smolt weight
- x_7: year, group 1 (2012)
- x_8: year, group 3 (2014)
- x_9: type of cohort, group 2 (fall)
- x_{10}: number of fish
- x_{11}: existence of female lice at the nearest neighboring farm, group 2 (at least one female louse at the nearest farm)
- x_{12}: biomass neighbor, group 2 (the largest half)
- x_{13}: temperature at stocking, group 2 (between 7^{o}C and 11^{o}C)
- x_{14}: temperature at stocking, group 3 (warmer than 11^{o}C)

where covariates x_2, x_3, x_7, x_8, x_9, x_{11}, x_{12}, x_{13} and x_{14} are binomial. I could see that there was a clear improvement in the AIC values only for the first covariates, which made it obvious only in the beginning, which covariates were most important. When choosing geographical index as second covariate, I had already seen that distance to the nearest neighbor should be added next, as models with both of the two covariates had much lower AIC values than the rest. Gradually the difference between the AIC values decreased, and it was no longer obvious which covariates improved the model in a best way. When choosing the fourth covariate, the AIC values for smolt weight, distance to the nearest slaughterhouse and year of stocking were almost the same, approximately 2355, and I used results after the Wald test to decide which of those covariates to include. When including remaining covariates I also used results after the Wald test in cases where the difference between AIC values was less than 2.

I have chosen to use year 2013 as reference group for the covariate addressing to year, as it appeared as significantly different from both 2012 and 2014. Years 2012 and 2014 seem to have approximately the same hazard rates for first treatment. I will present the impact of this covariate on the hazard rate later.

When it comes to linearity one can see that it changed for some of the covariates in the multivariate model. Distance to the nearest slaughterhouse (x_5) was improved by transformation $\frac{1}{x}$, while neighbor biomass density (x_{12}) and seawater temperature (x_{13} and x_{14}) did not behave as linear anymore and needed to be grouped. Distance to the nearest neighbor (x_4), smolt weight (x_6) and number of fish in the cohort (x_{10}) are still linear and can be used in the analysis as they are.

Now the model has an AIC value of 2349. Before the last three explanatory variables (x_{11}, x_{12}, x_{13} and x_{14}) were included in the model, the value of AIC was 2343, meaning that those explanatory variables worsen the model according to AIC. However, the difference in AIC values is too small to decide whether any variables are non-significant, and I prefer to look at the p-values after the Wald-test to decide whether to omit any explanatory variables from the model. I will nonetheless include all of them in the model in the first place, as I need to check for interactions first. If it appears that non-significant variables are not included in any significant interactions, they can be omitted from the analysis.

Since some of the explanatory variables are grouped and some are transformed in the multivariate model, I needed to check for collinearity one more time to ensure that the covariates were not explained by each other. The results were surprising: VIF corresponding to log-transformed EIP and to type of cohort were 3.67 and 3.18 respectively. Both values exceed the cut-off value of 3 and thus need to be considered further. I then checked for correlation between the explanatory variables to find out the reason for a high VIF. The transformed EIP appeared to correlate with three other covariates, and type of cohort with two. The result can be seen in table 4.7. One can see that the main reason for high VIFs is the correlation between type of cohort and log-transformed EIP. EIP correlates with more covariates than type of cohort, but in the multivariate analysis without interactions seawater temperature and neighbor biomass density appeared as non-significant, and thus it does not make much of a difference whether they correlate with EIP and type of cohort, as they most likely will be excluded from the analysis. Even if EIP had the highest VIF, I have chosen to keep this covariate in the model, and rather exclude type of cohort. The reason

Table 4.7: The highest correlations between log-transformed EIP and type of cohort with other covariates used in the multivariate model with time-fixed covariates

Type of cohort and log(EIP + 1)	0.72
Type of cohort and temperature at stocking	0.67
Temperature at stocking and log(EIP + 1)	0.61
Biomass neighbor and log(EIP + 1)	0.59

for this is that EIP appeared as more significant covariate both in the univariate and multivariate models, than type of cohort. Furthermore recalls that there is some uncertainty in the covariate addressing to the type of cohort since there are fewer events occurring in fall. Thus there are more problems linked to this covariate compared to EIP. When excluding type of cohort, VIF for EIP becomes 2.47 and less than 2 for remaining covariates. AIC value of the multivariate model increased from 2349 to 2352, which is not much, and indicates that excluding type of cohort does not worsen the model too much. Interactions can now be checked in a multivariate Cox PH model with 10 explanatory variables.

I checked for all possible interactions between all covariates, including non-significant covariates. Only one interaction improved the model according to BIC (and AIC) and is between transformed EIP and existence of female lice at the nearest neighbor. This interaction also improved significance in the lice-covariate, and covariate x_{11} appears to be an important covariate after all.

Summary of the multivariate analysis

When tests for interactions have been performed, the significance of the variables can be examined. In table 4.8 one can see that the signs of significant coefficients are mostly logical and the same as in the univariate analysis. Note that for transformation $\frac{1}{x}$ positive coefficient means negative effect, which leads to the following interpretation: an increase

Table 4.8: Part of a summary of a multivariate Cox PH model. The summary contains coefficients and p-values after Wald-tests. I am only interested in the sign before coefficients in the current table and whether the p-values are higher or lower than 5%.

	coef	p-values
EIP (transformed)	0.90	< 0.01
geo.index (group 2)	-0.34	0.08
geo.index (group 3)	-0.02	0.94
dist.n	-0.06	< 0.01
dist.sl. (transformed)	1.98	0.01
weight	0.05	0.01
year 2012	-0.53	< 0.01
year 2014	-0.47	< 0.01
number of fish	-0.04	0.12
lice n. (group 2)	0.51	0.01
biomass n (group 2)	-0.07	0.68
temp.(group 2)	0.16	0.38
temp.(group 3)	-0.19	0.48
EIP (transformed) : lice n. (group 2)	-0.41	< 0.01

in distance to the nearest slaughterhouse leads to decreased treatment hazard rate, which is expected biologically. Note also that the interaction term between EIP and existence of female lice at the nearest neighbor has a negative coefficient, while coefficients addressing to EIP and the lice-covariate itself are positive, meaning that interaction affects the hazard oppositely to what the covariates alone do. I will present the effects on the hazard rate more detailed later in this chapter.

Let us now look at the significance of the covariates in the model. Seawater temperature and neighbor biomass density did not appear as significant in the model without interaction. The two covariates were not included in any of the significant interactions either, and the significance was not improved when a interaction was added. For this reason they can be excluded from the final model one by one. One reason that they appear as non-significant in the multivariate model could be that most of the effect of these covariates is captured by EIP. Another reason could be that the value of those covariates at the time of stocking is not informative enough, as seawater temperature changes every week and neighbor biomass density changes every month. Therefore when time-intervals increase, the value of those covariates at stocking becomes non-significant. Another covariate that appears to be non-significant is amount of fish in the cohort, and I have chosen to exclude it from the model. Recall that this covariate was not highly significant in the model without interactions, and when one interaction was added, the effect of size of the cohort vanished.

After excluding covariates addressing to temperature, neighbor biomass density and amount of smolts in the cohort from the model, significance in geographical index was improved, and the mid-region of Norway appeared as significant with p-value 0.04 after a Wald test. But when group 1 was chosen as the reference group, only group 2 appeared as significant in comparison. I tried to change the covariate to consist of two groups, but the model was not improved. Thus I will keep this covariate grouped in three groups.

When it comes to year of stocking one can see that 2013 is significantly different from 2012 and 2014. The hazard rate was highest in 2013, and very similar in 2012 and 2014.

The model consists of seven significant covariates and one interaction term. The AIC value of the model is now 2341.

Proportional hazard rates assumption

Before I start analysing the results, the assumption about proportional hazard rates needs to be satisfied. I performed a test as described in section 2.2.4 by using different transformations, and all of them produced quite similar results. In table 4.9 I reported the results with identity transformation $g(t) = t$. A p-value lower than 0.05 indicates a violation of the proportionality assumption.

In table 4.9 one can see that group 3 of geographical index has a p-value lower than 5%, while the rest of the covariates satisfy the proportional hazard rate assumption. This covariate should be considered more closely. In figure 4.4 one can examine the effect on hazard rate caused by group 3 of geographical index. The hazard rate seems to be somehow increasing for group 3 of geographical index from week 10 after stocking until 27 weeks after stocking,

Table 4.9: A summary after test for proportional hazard rates.

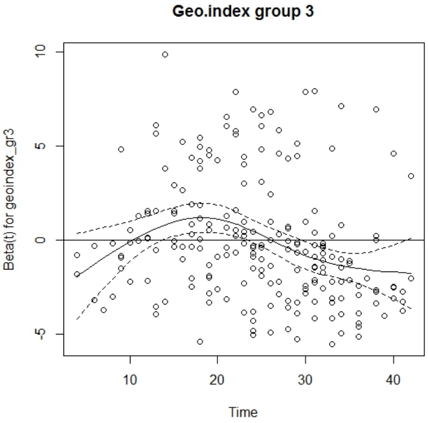
	rho	chisq	p
EIP (transformed)	0.01	0.02	0.88
geo.index (group 2)	-0.02	0.10	0.75
geo.index (group 3)	-0.16	5.40	0.02
dist.n.	0.01	0.01	0.94
dist.sl. (transformed)	-0.02	0.08	0.78
weight	0.04	0.38	0.54
year 2012	0.005	0.005	0.94
year 2014	0.01	0.02	0.90
lice n. (group2)	-0.10	1.84	0.18
EIP (transformed) : lice n. (group2)	-0.003	0.003	0.96
GLOBAL		16.40	0.09

and to be decreasing the rest of the time. The last row in table 4.9 contains a global test of proportional hazard rates for all the covariates tested at once. One can see that p-value after the global test is 9%. This is higher than 5%, which is acceptable. However, I chose to improve the model to eliminate non-proportionality in the geographical index.

After trying both stratification of the covariate and time-interaction, I chose to stratify the covariate. Interaction with time worsened the whole model and proportionality in the hazard rates was not improved. Geographical index is a categorical variable and contains only three levels, which means that it can be stratified without further ado. In fact when I stratified geographical index in the above model, the AIC value decreased to 1915, which constitutes a significant drop. The only disadvantage is that I will not be able to see how the treatment hazard rate is affected by moving the localities between north-, mid- and south-regions of Norway from the R-output, since geographical index will be a part of the baseline now.

After stratifying geographical index, p-values after testing for proportional hazards were much higher than 5% for all covariates, and the p-value after the global test was 75%, meaning that the proportional hazard assumption is satisfied, and I can start interpreting the results.

Figure 4.4: A graph of the scaled Schoenfeld residuals for group 3 of geographical index along with a smooth curve. If the plot shows a random pattern, the PH assumption has been retained, and the fitted line should be approximate horizontal at $\beta(t) = 0$.



Interpretation of the results

Table 4.10 contains a summary of the final covariates used in the Cox PH model. R-output of the model can be found in listing 1 in Appendix B.1. One can see from the output that EIP is the most important explanatory variable in the current analysis according to p-value

Table 4.10: Description of explanatory variables in the constructed Cox PH model, including transformations and truncations, units the covariates will be measured in, max and min values, mean and median of observations for numerical variables, and distribution of categorical variables.

Explanatory variable	Min value	Median	Mean value	Max value
EIP (10^6) ($\log(x+1)$)	0.00	0.30	0.79	4.10
Distance: nearest slaughterhouse (km) ($\frac{1}{x}$)	0.004	0.05	0.07	0.96
Distance: nearest neighbor (km) (truncated at 40 km)	0.38	4.57	6.80	40.00
Weight of fish (10 gram)	4.00	13.59	14.02	24.95
Female lice, nearest neighbor	none	355 (54.8%)		
	at least 1	293 (45.2%)		
Year	2012:	248 (38.3%)		
	2013:	255 (39.3%)		
	2014:	145 (22.4%)		

after Wald-tests. Recall additionally that existence of female lice at the nearest neighbor and year of stocking were the least important covariates in univariate models. However when they are joined with other explanatory variables, their significance increases greatly.

Let us start by looking at how the treatment hazard rate changes when smolt weight increases, keeping all covariates constant. The coefficient addressing to smolt weight is 0.045 as one can see in listing 1. When average smolt weight in a cohort increases by 10 grams, the hazard rate for first treatment increases by 1.05. This does not seem like a huge impact on the hazard rate, but if the increase in smolt weight is greater, the difference becomes sizable. When smolt weight increases by 50 grams, the treatment hazard rate increases by $\exp(0.045 \cdot 5) = 1.25$.

Year of stocking appeared as a significant covariate in the model. Stocking year 2013 seems to have the highest hazard rate. One can see that the hazard rate was 1.70 times higher in year 2013 compared to year 2012, and 1.60 times higher compared to year 2014. This could mean different things: either that 2013 was a warm year (that could actually also be seen from the dataset), or that this was a year with large amounts of salom lice along the Norwegian coastline due to reasons other than contained in the dataset.

A decrease in distance variables leads to lower treatment hazard rate, which means longer time till first treatment. This is biologically reasonable: if a farm has a neighbor or a slaughterhouse nearby, then the probability of getting lice from them is higher than if they were further away. One can see that a decrease in distance to the nearest slaughterhouse seems to be more consequential for the treatment hazard rate than a decrease in distance to the nearest neighbor. However the slaughterhouse-distance is transformed, and because of the transformation, this covariate does not vary as much as neighbor-distance as one can see in table 4.10. When distance to the nearest neighbor decreases by 1 km, the hazard rate for first treatment increases 1.06 times. Since this covariate spans over a range between 0.38 km and 40 km, it is possible to calculate what happens if this distance-covariate decreases by 5 and 10 km. The hazard rate increases 1.36 times if distance to the nearest neigbor decreases by 5 km and 1.84 times if the decrease is 10 km. Distance to the nearest slaughterhouse is a transformed covariate. The change in treatment hazard caused by a

Table 4.11: Change of hazard rate caused by a decrease in distance to the nearest neighbor when the remaining covariates are left constant (a small selection of examples).

From	To	Hazard change	From	To	Hazard change	From	To	Hazard change
3 km	2 km	1.33	8 km	5 km	1.14	30 km	20 km	1.03
4 km	2 km	1.53	9 km	5 km	1.16	35 km	20 km	1.04
5 km	2 km	1.67	10 km	5 km	1.19	40 km	20 km	1.04

change in the covariate can be expressed by

$$\exp\left(1.71 \cdot (\frac{1}{x_2} - \frac{1}{x_1})\right).$$

When distance to the nearest slaughterhouse decreases from 6 to 5 km, the treatment hazard rate increases 6%. Since the covariate is transformed, then a change of one unit in the covariate is expressed by $\frac{1}{x_2} - \frac{1}{x_1} = 1$. Generally when distance to the nearest slaughterhouse changes from x_1 km to $\frac{x_1}{1+x_1}$, the hazard rate for first treatment increases 5.52 times. This means a change from 2 km to 0.67 km, from 5 km to 0.83 km, from 12 km to 0.92 km, from 20 to 0.95 km, and so on. In table 4.11 one can see how the treatment hazard rate changes when the distance to the nearest neighbor decreases. One can see that the hazard change decreases the further away a slaughterhouse is from the farm of interest, meaning that treatment hazard is more vulnerable to increases when this covariate decreases to a low value of distance. After achieving a certain limit, the treatment hazard rate remains almost the same, and if the distance changes to 20 km from 30, 35 or 40 km the hazard rate does not change much.

The remaining covariates in the current model interact with each other, and thus the effect on treatment hazard rate is not only affected by one of the variables, but also by the combinations of variables. This applies to EIP and existence of female lice at the nearest neighbor. EIP is a transformed covariate, while lice-covariate is categorical. Let us look at the general formula that applies to change in the hazard rate as a result of change in both covariates when all other covariates are constant:

$$\begin{aligned}\frac{\alpha(\mathbf{x}_2)}{\alpha(\mathbf{x}_1)} =& \exp\left(0.51 \cdot x_{lice} + 0.96 \cdot (\log(x_{EIP,2}+1) - \log(x_{EIP,1}+1))\right) \cdot \\ & \exp\left(-0.43 \cdot x_{lice} \cdot (\log(x_{EIP,2}+1) - \log(x_{EIP,1}+1))\right),\end{aligned}$$

where x_{lice} takes value 0 if there are no female lice at the nearest neighbor and value 1 if there is at least one female louse. The interaction term has a negative coefficient and affects the hazard rate oppositely to what the two covariates alone do, as their coefficients are positive. Let us look at the parametric part of the expression of hazard rate in 2.2, which is $\exp(\boldsymbol{\beta}^T\mathbf{x})$, to investigate how varying values of EIP and the lice-covariate will affect the parametric part and thus the treatment hazard when the remaining covariates are as follows: year 2013, smolts weighing 100 grams, 5 km to the nearest neighbor and 15 km to the nearest slaughterhouse. The parametric part is constant for the chosen values while the baseline varies with time. However, it is possible to see how the trend in the hazard rate will be when the constant part takes different values. One can see from table 4.12 that the hazard rate will increase mainly when EIP increases. One can also see that when

Table 4.12: Constant to be multiplied with the baseline for different values of EIP and existence of female lice at the nearest neighbor while other covariates are kept constant.

EIP	0	0	2	2	5	5	10	10	20	20
Female lice n.	0	1	0	1	0	1	0	1	0	1
$\exp(\boldsymbol{\beta}^T \mathbf{x})$	1.48	2.44	4.22	4.39	8.17	6.36	14.59	8.76	27.11	12.18

EIP becomes high, the hazard rate will decrease when amount of female lice at the nearest neighbor goes from zero to non-zero. At first sight this is opposite to what is expected, as a larger amount of female lice at the nearest neighbor should not lead to a decreased treatment hazard rate (longer time till first treatment). However since the EIP and lice-covariate roughly explains the same thing, the interaction is negative in order to eliminate a double effect when both are high. The situation is, furthermore, not realistic when EIP is as high as for example 10 or 20 and there are no female lice at the nearest neighbor. Thus the relationship between EIP and lice-variable needs to be taken into account in this analysis.

Let us now look closer at the existence of female lice at the nearest neighbor as a covariate alone. When amount of female lice at the nearest neighbor over the past four weeks before study-start increases from zero to non-zero, the treatment hazard rate is affected in the following way

$$\frac{\alpha(\mathbf{x}_2)}{\alpha(\mathbf{x}_1)} = \exp\left((0.51 - 0.43 \cdot \log(x_{EIP} + 1))\right).$$

This formula applies in an analysis when all other covariates are kept constant, including EIP. The interesting point is where the effect is equal to 1, meaning that the treatment hazard goes from being on the increase, to being on the decrease, as the amount of female lice increases to above zero. Values of transformed EIP vary from 0 to approximately 4, which means that the formula above will achieve a value of 1 for a particular level of EIP. It is possible to calculate that the hazard rate increases as lice-covariate goes from factor 1 to factor 2 before EIP achieves a value of 2.30. Once EIP achieves a value of 2.30, the interaction term starts to eliminate the double effect of EIP and existence of lice, and thus for values of EIP higher than 2.30, one should only look at the cases when amount of female lice at the neighboring farm is higher than zero.

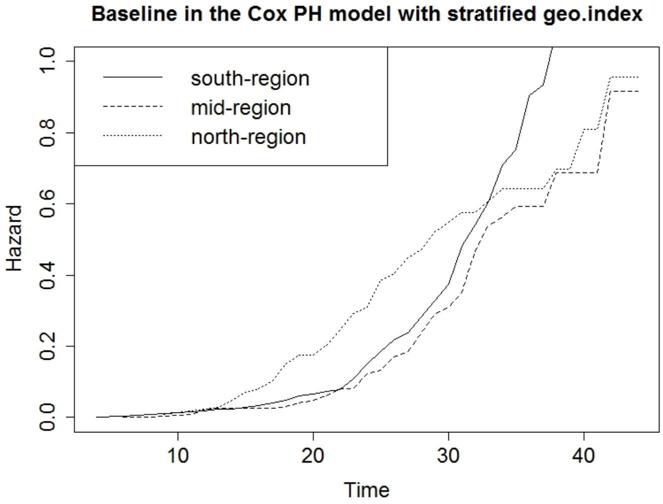
Let us also look at the EIP alone. Hazard rate changes in the following way when EIP changes:

$$\frac{\alpha(\mathbf{x}_2)}{\alpha(\mathbf{x}_1)} = \exp\left((0.96 - 0.43 \cdot x_{lice}) \cdot (\log(x_{EIP,2} + 1) - \log(x_{EIP,1} + 1))\right).$$

One can see that this change will remain higher than one independent of which value the lice-covariate takes, meaning that treatment hazard rate increases with increased EIP. If there are no female lice at the nearest neighboring farm, the treatment hazard rate increases 2.61 times for each logarithmic increase in EIP (the transformation effect is explained in section 4.1.2). And if there are any female lice at the neighboring farm, the hazard rate increases by 1.70 for each logarithmic increase in EIP.

All changes in the treatment hazard rate caused by all covariates have now been considered.

Figure 4.5: Computed baseline survival curve in Cox PH model with stratified geographical index.



The only influence on the hazard rate I was not able to investigate was the one caused by geographical index since it is contained in the baseline. Even if there are no coefficients addressing to this covariate, a plot of the baseline hazard curve can be made to investigate the effect. Figure 4.5 contains the plot of the baseline hazard based on a `coxph` model. One can see that until approximately 32 weeks from stocking the hazard rate is very similar for south- and mid-regions, and is the highest for the north-region. After approximately 32 weeks after stocking hazard is the highest in the south-region, while north- and mid-regions seem to have a similar hazard curve.

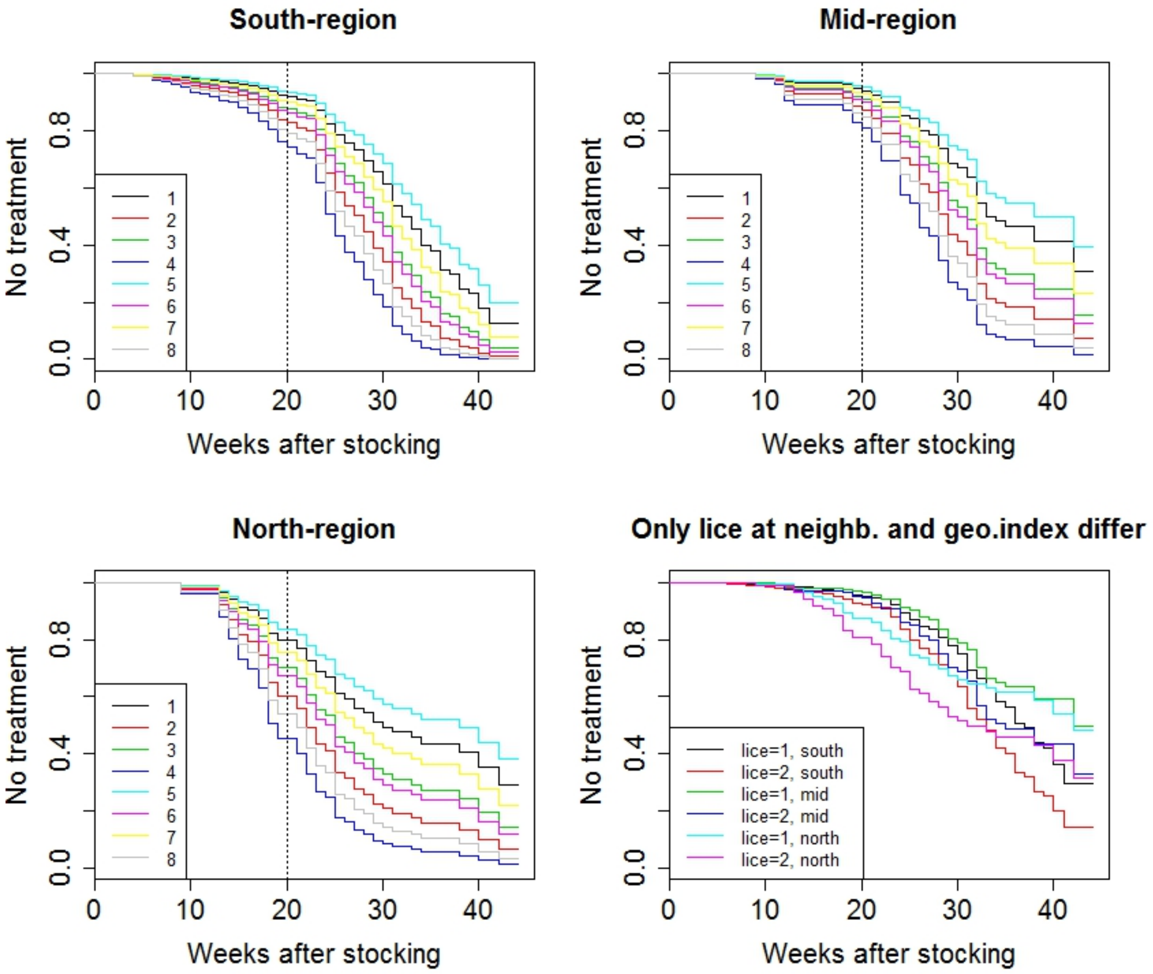
Survival curves

In this analysis survival probability means probability of no treatment. Probability curves for different combinations of covariates will now be presented. Since the model is contained by many explanatory variables, I have chosen to keep two of them constant, such that I will present the plots for the group of cohorts within 15 km distance of the nearest slaughterhouse, and when year of stocking is 2013.

Groups are divided into the following eight groups:

1. EIP = 1, no female lice at the nearest farm, smolt weight = 100 gram, 1 km to nearest neighbor
2. EIP = 5, female lice neighbor > 0, smolt weight = 100 gram, 1 km to nearest neighbor
3. EIP = 1, no female lice at the nearest farm, smolt weight = 200 gram, 1 km to nearest neighbor
4. EIP = 5, female lice neighbor > 0, smolt weight = 200 gram, 1 km to nearest neighbor
5. EIP = 1, no female lice at the nearest farm, smolt weight = 100 gram, 4 km to nearest neighbor
6. EIP = 5, female lice neighbor > 0, smolt weight = 100 gram, 4 km to nearest neighbor
7. EIP = 1, no female lice at the nearest farm, smolt weight = 200 gram, 4 km to nearest

Figure 4.6: Probability of no treatment for different combinations of fixed explanatory variables.



neighbor

8. EIP = 5, female lice neighbor > 0, smolt weight = 200 gram, 4 km to nearest neighbor

And plotted for different values of geographical index: south, mid and north. Even if it was not possible to see the difference between different values of geographical index from the R-output, it is possible to distinguish between them when plotting survival curves. The survival curves of the groups defined above can be seen in the first three plots in figure 4.6. One can see that the trend in survival curves for the 8 groups is the same in all parts of Norway. Group 5 seems to have the highest probability of no treatment, which is expected, as this group contains low EIP, no female lice at the nearest neighbor, small smolts and long distance to the nearest neighbor. The group with the second best result is group 1 with low EIP, no female lice at the nearest neighbor, small smolts and shorter distance to the nearest neighbor. Following that, are groups 7 and 3 respectively. The remaining groups come in the same order, but with high EIP and amount of female lice at the nearest neighbor higher than zero. From this I conclude that a 100 gram change in smolt weight influences treatment hazard more than a change in the distance to the nearest neighboring farm from 4 km to 1 km. One can see that groups with low EIP have higher probability of no treatment than groups with high EIP. This again indicates that EIP is one of the main factors for first bath treatment. However there are not enough observations of high EIP in the north-region to make final conclusions of how probability of no treatment differs there. Thus let us look at the survival probability in the three regions when EIP is low. At first glance one can see that south- and mid-regions of Norway have quite similar survival curves, and that the north-region has a much lower probability of no treatment. Probability of no treatment 20 weeks after stocking, when EIP is low, is approximately 87-95% if a farm is located in the south-region, and 92-97% and 70-83% in mid- and north-regions respectively. Thus a first treatment seems to occur earlier at the farms located in the northern parts of Norway, and later at the farms located in the mid-region of Norway.

In the last plot in figure 4.6 one can see survival curves only when existence of female lice at the nearest neighbor and geographical index differ. The remaining covariates are kept constant at the same level as above and additionally smolt weighing 150 grams, 4 km to the nearest neighbor and EIP equal to 0.5. One can see that probability of no treatment within the first 30 weeks after stocking is the highest when there are no lice at the nearest farm in mid- and south-regions respectively. The lowest probability of no treatment is in the north-region. However 30 weeks after stocking the situation changes, and the farms that have no lice at the nearest neighbor, and that are located in the north, seem to have the second highest probability of no treatment, and the farms located in the south and that have at least one female louse at the nearest neighbor appear to have the lowest probability of no treatment.

4.2 Extended Cox model

The Cox PH model showed some interesting and useful results. However, as mentioned before, many of the variables can be obtained at more updated time points than at the time of stocking. For this reason I will extend the model to contain time-dependent covariates. My hope is that this will lead to an improved understanding, and that I will collect more significant effects than what I did with the Cox PH model. The advantage of introducing time-dependency in covariates is their updated value. Here, the values of each covariate can be obtained not only at the time of stocking, but also just before any defined time t. Obtaning values before time t produces much more precise values.

4.2.1 Data

To extend the dataset to include time-dependent covariates obtained just before time t, I will start by splitting each cohort into multiple sub-cohorts at each cutting time. I have chosen to split the observations at every 4 weeks, such that I will look at approximately one month at a time. Since the longest time from stocking until first treatment or censoring is 44 weeks in this study, I have defined cut times as 4, 8, 12, ..., 40, 44. The new data set is now in ”counting process” format with a start time, a stop time and a treatment status for each record. A small selection of the dataset can be seen in table 4.13. The amount of cohorts is now the same as in the Cox PH model, but the number of time-intervals is 4652, of which 212 contain actual treatments.

Table 4.13: A small selection of the updated dataset (with time-dependent covariates).

id	locality	stocking	observation start	treatment/ censoring	time0	time	status	temp old	temp new
1	10041	2013 39	2013 39	2013 43	0	4	0	15.0	15.0
1	10041	2013 39	2013 44	2013 47	4	8	0	15.0	12.0
1	10041	2013 39	2013 48	2013 51	8	12	0	15.0	10.0
1	10041	2013 39	2013 52	2014 03	12	16	0	15.0	9.0
1	10041	2013 39	2014 04	2014 06	16	19	1	15.0	7.0
2	10078	2012 43	2012 43	2012 47	0	4	0	10.0	10.0
2	10078	2012 43	2012 48	2012 51	4	8	0	10.0	8.0
2	10078	2012 43	2012 52	2013 03	8	12	0	10.0	8.0
2	10078	2012 43	2013 04	2013 06	12	15	0	10.0	7.0
3	10080	2012 19	2012 19	2012 23	0	4	0	9.5	9.5
...									

In section 3.3 it is described which of the covariates vary with time and whether they can be used in an extended Cox model. Fixed variables, as described in section 3.3, will be kept constant. I will not include location density in this analysis, as it appeared to correlate highly with neighbor biomass density. Type of cohort will not be included either, as it appeared to correlate with EIP. In essence I will construct an extended Cox model, based on the already developed PH model, and attempt to keep the same transformations, grouping and truncation of the covariates. I will consider including all time-dependent covariates in the new model, even if some of them appeared as non-significant in the Cox PH model.

Temperature at the time of stocking could be used as a time-fixed covariate alongside temperature as a time-dependent covariate, since the two do not correlate much. However time-fixed temperature at the time of stocking did not appear as significant in the Cox PH model, and thus I have chosen to not include it in the model.

Lice situation at the nearest neighbor was treated as a categorical covariate in the Cox PH model, since 55% of observations were equal to zero. This could be caused by organised stocking, which means that farmers from different localities choose to stock their fish at the same time. In cases of organised stocking there will be no lice at the nearest neighbor at time $t = 0$ in the dataset. However when neighbor's lice-values are obtained just before time t for different time points, the majority of observations obtained at times $t > 0$ may be greater than 0, and the covariate should rather be treated as numerical. However when I checked, I discovered that 68% of observations of the amount of female lice at the nearest active neighbor are equal to 0 when obtained at the beginning of each time-interval (just before time t). Therefore there are no problems related to categorising this covariate in this analysis.

A summary of the values of time-dependent covariates can be seen in table 4.14. One can see that of all covariates only EIP has huge outliers. When treated as time-dependent, the max value of EIP is 230, while when it was treated as time-fixed, the max value was 59. Thus the time-dependent EIP spans over a much wider interval. There is indication that the huge values are contained by outliers, since only 3% of observations show values of

Table 4.14: Description of time-dependent explanatory variables, including units in which covariates will be measured: max and min values, mean, median, range interval for 50% of observations for numerical variables, and distribution of categorical variables. Values of the explanatory variables are taken at the start of each time-interval.

Explanatory variable	Min value	First quantile (25%)	Median	Mean value	Third quantile (75%)	Max value
EIP (10^6)	0.00	0.19	1.21	3.96	4.36	230.10
Number of fish (10^5)	0.30	5.60	8.18	9.08	11.61	32.34
Biomass neighbors (100 t)	0.00	1.04	1.83	2.15	2.94	8.64
Temperature (^{o}C)	1.63	7.20	9.82	9.96	12.60	21.80
Female lice, nearest neighbor	group 1	3157 (67.9%)				
	group 2	1495 (32.1%)				
In-feed treatment	group 1	4155 (89.3%)				
	group 2	497 (10.7%)				

higher than 20. For this reason log-transformation will be very useful in this case. However a more heavy transformation might be more helpful here, for example $\frac{1}{x}$. I will consider the transformations of EIP at a later point in the analysis.

4.2.2 Univariate modelling

The first thing I chose to do, was to study external covariates: EIP, seawater temperature, neighbor biomass density and existence of female lice at the nearest farm to see whether they were significant in a univariate analysis. If they appeared as non-significant, I tried to include them at earlier time points to see whether significance improved. EIP, neighbor biomass density and neighbor's lice situation were all highly significant in the univariate analysis when included at the beginning of each time-interval. Seawater temperature, however, appeared as non-significant. I then performed two univariate analyses: first, with lag-time effect of the temperature (seawater temperature approximately at the beginning of the previous month), and second, with history of temperatures (mean value over the previous four weeks). Lag-time effect appeared as highly significant, and thus I chose to use temperature at time $t - 1$ in further analysis.

Since EIP is contained by a few huge values, I had some concern that there were outliers in this covariate that affected the hazard in a wrong way. I considered a truncation at 59, which is max value of EIP when treated as time-fixed. To understand whether the outliers affected the hazard in an opposite way to what the main effect is, I tried to look at the null residual plot, but with the large number of observations it was not clear. Thus I chose to group the covariate into 5 groups, where group 1 consisted of the smallest values and the last group consisted of values higher than 20. Results from the univariate analysis with EIP as a categorical covariate did not indicate that the outliers affect the hazard rate negatively. In fact the group containing outliers had the highest hazard rate of the 5 groups, and to be more precise the hazard rate in the group with the highest values was approximately nine times higher than in the group with the lowest values. For this reason I chose not to truncate extreme values of EIP, and rather to apply a heavy transformation of the covariate.

In table 4.15 one can see a comparison of AIC values in the univariate models, with explanatory variables treated as time-dependent and time-fixed in the new dataset. When covariates are treated as time-fixed, they have the same value in every time-interval for

Table 4.15: AIC values in univariate, extended Cox models. AIC in a null model is 2448.83.

	time-fix		time-fix	time-dep
Smolt weight	2438.01	Number of fish	2440.28	2417.63
Distance to slaughter ($\frac{1}{x}$)	2433.85	Biomass neighbor ($\sqrt{x}$)	2409.76	2428.25
Distance to neighbor	2414.37	Lice at neighbor	2450.70	2336.97
Geo.index	2410.62	Temperature	2432.52	2433.60
Year of stocking	2447.71	EIP ($\log(x+1)$ and $\frac{1}{x}$)	2380.71	2354.70
		In-feed treatment		2450.73

each cohort, which is the same value as at time of stocking. When covariates are time-dependent, their values vary for each time-interval. I compared the models to see whether time-dependency in covariates led to any improvements in AIC values. Covariates in the left column of the table do not vary with time, and thus I reported only one AIC value for each of them. On the right-hand side one can see how the AIC values change when time-dependency in the time-varying covariates is introduced.

Neighbor biomass density did not appear as log-linear in the univariate analysis when treated as time-fixed and had to be transformed. In order to compare the results more precisely, I performed the same transformation of the time-dependent biomass-covariate. When it comes to EIP, the covariate appeared as non-linear when treated both as time-fixed and as time-dependent. The same transformation did not improve the linearity in both cases, and I had to use different transformations. Furthermore recall that values of EIP span over a much wider interval now, and therefore a heavier transformation appeared to work better on the data.

By studying the table 4.15 one can conclude that introducing time-dependency in the covariates addressing to EIP, amount of fish in the cohort and lice situation at the neighboring farm leads to improved significance. One can see that when the lice-covariate is treated as time-fixed, the AIC value in the model is roughly the same as in a null model, which indicates that existence of lice at the nearest farm at time of stocking is not a significant covariate in a univariate Cox PH model. However when time-dependency is introduced, the AIC value is reduced by 114 and is now the lowest AIC value of all univariate analyses, meaning that the lice-covariate is the most important covariate in the model when it varies over time. For neighbor biomass density it seems that univariate Cox PH model produces better results than the univariate extended model. AIC in the univariate model with in-feed treatment as a covariate is higher than AIC in the null model, which means that in-feed treatment seems to be non-significant. Seawater temperature seems to have a similar impact on the two models univariately. However significance might change when a multivariate model is constructed.

Also I checked the effect on the hazard rate caused by each time-dependent covariate in a univariate analysis by looking at the coefficient addressing to the covariate. The hazard rate at time t increases along with increased neighbor biomass density, existence of female lice at the nearest farm, seawater temperature and EIP. While hazard rate at time t appears to be decreasing for increased amount of fish in the cohort. These results indicate on the same trend as when the covariates were treated as time-fixed. In-feed treatment appears as non-significant in the univariate analysis after a Wald test, and the effect on the hazard is not clear.

4.2.3 Multivariate modelling

I started constructing a multivariate extended Cox model by including all fixed covariates that were significant in the Cox PH model. These are distance variables, geographical index, year of stocking and smolt weight. Further one could see from table 4.15 that existence of female lice at the nearest farm is the most important time-dependent covariate. For this reason I included this covariate as the next one in the model. By adding time-dependent covariates one by one and checking for linearity, I ended up with a model that produced similar results to what the Cox PH model did. I excluded non-significant covariates after performing Wald tests, and chose to report only the final model with all significant covariates, omitting all steps to obtain the model. The final model is then as follows:

$$\alpha(\mathbf{x}(t)) = \alpha_0(t)\exp(\beta_1 \min(x_1, 40) + \beta_2 \frac{1}{x_2} + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 \\ + \delta_1 x_7(t) + \delta_2 \frac{1}{x_8(t)+1} + \delta_3 x_9(t-1) + \delta_4 x_{10}(t))$$

With the following explanatory variables:

- x_1: distance to the nearest neighbor
- x_2: distance to the nearest slaughterhouse
- x_3: geographical index, group 1 (south-region)
- x_4: geographical index, group 2 (north-region)
- x_5: year, group 1 (2012)
- x_6: year, group 3 (2014)
- $x_7(t)$: existence of female lice at the nearest farm, group 2 (at least one female louse)
- $x_8(t)$: EIP
- $x_9(t)$: temperature
- $x_{10}(t)$: number of fish in the cohort

Where covariates x_3, x_4, x_5, x_6 and x_7 are binomial. The AIC value of the model is now 2252. In this model there are three categorical variables and five numerical variables, of which two are transformed. Four of the covariates are time-fixed, and four are time-dependent. There were no interactions that improved the above model according to neither AIC nor BIC.

Amount of fish at the farm and seawater temperature became highly significant covariates when time-dependency was introduced. I did expect seawater temperature to be significant based on results in Jansen et al. 2012.

Smolt weight stopped being significant after time-dependent covariates were introduced. The remaining time-fixed covariates were significant and linear with the same transformations and truncations as in the Cox PH model. Again I chose to use year 2013 as a reference group. Furthermore I chose to use the north-region of Norway as a reference group in this analysis, as it appeared that the hazard rate in the north was significantly different from the other two regions. The hazard rates for the three levels of geographical index appeared as proportional in the final extended Cox model, and therefore I chose to keep this covariate in the parametric part in this analysis, such that it will be possible to see its impact on the hazard from R-output.

In the Cox PH model there was an interaction term between EIP and existence of lice at the neighboring farm that eliminated the double effect of the covariates when both were high. Since this interaction is not significant in the current analysis, the double effect when both are high was eliminated in another way. Transforming EIP as $\frac{1}{x+1}$ prevent the hazard from increasing extremely when EIP changes a lot. Additionally this transformation helped improve log-linearity in this covariate.

In-feed treatment and neighbor biomass density appeared as non-significant covariates in the final model. In-feed treatment was non-significant throughout: both in the univariate model and in the multivariate models. Neighbor biomass density was significant univariately, but stopped being significant when more information about the neighbors was included in the model. This means that the effect of biomass at the neighboring farms within 40 km seaway distance may be explained enough by other covariates in the model.

Also I checked for collinearity and correlation between covariates in the final extended Cox model, and all VIF values were lower than 2, which means that there is no correlation to be concerned about.

I checked for proportional hazard rates in time-fixed covariates, and there was no evidence of non-proportionality. The global test resulted in a p-value equal to 0.93 after testing for non-proportionality in the current model, and the assumption about proportional hazards in time-fixed covariates is satisfied.

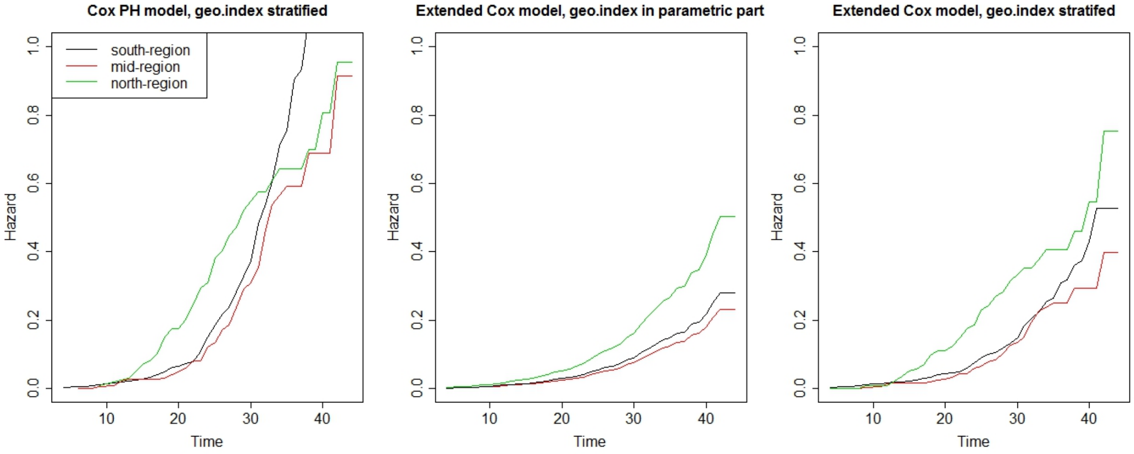
Interpretation of the results

R-output can be found in listing 2 in Appendix B.1. Time-dependent covariates are the most important covariates in the model. This is because their values are more up to date in this analysis. Let us now look at how the explanatory variables affect treatment hazard rate.

The effect of time-fixed covariates on treatment hazard is almost the same in this model as in the Cox PH model, and therefore I choose not to go into the details. The models showed that when distance covariates decrease, treatment hazard increases, and that 2013 was a year with a bad lice situation, while 2012 and 2014 had quite similar treatment hazards.

Geographical index was a part of the baseline in the Cox PH model due to its non-proportionality. In the current model the hazard rates appeared as proportional, and I chose to use this covariate in the parametric part of the model. One can see that the hazard rate is the highest in the north-region of Norway according to the current model. It is actually 1.81 times higher than in the south-region and 2.17 times higher than in the mid-region of Norway. Figure 4.7 shows how the treatment hazard rate changes with time in different regions of Norway, both when it is contained in the parametric part and when it is a part of the baseline. Also I present the baseline plot in Cox PH model, just for comparison. One can see that in the Cox PH model the baselines for the three regions intersected, and therefore geographical index needed to be stratified. In the current, extended Cox

Figure 4.7: Computed treatment hazard rate in developed Cox models. Both when geographical index is a part of the baseline and when contained in the parametric part in the extended Cox model.



model one can see that even if geographical index is stratified, the hazards do not intersect considerably, and the trend in the curves is almost the same when the covariate is included in the parametric part or in the baseline. Since there are only small differences between the impact on the model from geographical index when it is part of the baseline and when it is not, I chose to keep this covariate in the parametric part in the current analysis, such that there will be a coefficient explaining its effect.

Amount of female lice at the nearest farm is now a time-dependent covariate. Interpretation of the results addressing to this covariate is now as follows: at any given time t, the treatment hazard for a cohort whose neighbor does not have any female lice (but may have later) is approximately 5 times lower than the hazard for a cohort whose neighbor already has one or more female lice at that time.

EIP is also time-dependent now. I chose to transform this covariate as $\frac{1}{x+1}$ in this analysis since the logarithmic transformation (as in Cox PH model) did not improve linearity in the covariate as good as the current transformation did. Additionally, the transformation $\frac{1}{x+1}$ is more heavy and does not allow the hazard to increase extremely when EIP increases to the high values, since this covariate varies from 0 to 230. This transformation leads to higher hazard changes for low values of EIP and diminishing hazard changes as EIP grows. One can calculate that if EIP at one farm is 0.5 and 5 at another farm, at any time t, the hazard rate at the farm with the lowest EIP is $\exp(-1.80 \cdot (\frac{1}{6} - \frac{1}{1.5})) = 2.46$ lower at that given time point. And if the values of EIP are 5 and 100 at time t at two farms respectively, the treatment hazard is only 7% higher at the farm with EIP $= 100$ at the same time point. Thus one can see that treatment hazard is more vulnerable for differences in EIP when initial EIP is low. Generally, at any time t, if one farm has an EIP value equal to x, and another farm has an EIP value equal to $\frac{x}{1+x}$ at the same time point, the treatment hazard at the farm with EIP $= x$ is six times higher than at the other farm.

Two covariates appeared as significant only when treated as time-dependent in this analysis. They are amount of fish in the cohort and seawater temperature. Amount of fish in the cohort has a negative coefficient as one can see in the R-output. This means that higher amount of fish essentially leads to longer time till first treatment. As mentioned in section 4.1.2 this is only a short-term effect and should not be practised by farmers. If one farm has 100 000 more fish in the cohort at any given time t than any other farm, the hazard rate for first treatment for that cohort is 6% lower at that time. And if one farm has one million more fish than any other farm at any given time t, the hazard rate at that farm is 80% lower than at the other farm containing fewer fish. This result indicates that the time till first treatments is often postponed at the farms that contain a lot of fish. One reason for this may be that a similar amount of lice in the sea produces a different average number of lice per fish at a large locality compared to at a small locality. Thus a large locality will be affected less than a small locality when one examines time till first treatment

Seawater temperature appeared as significant when lag-time effect was introduced. This means that seawater temperature at the previous time point affects the hazard rate at the current time point. An increase of seawater temperature at time $t-1$ leads to an increase in treatment hazard rate at time t. More precisely if seawater temperature at one farm increased by 1^oC in the previous month, the hazard rate this month will increase by 10% compared to another farm where seawater temperature remains the same.

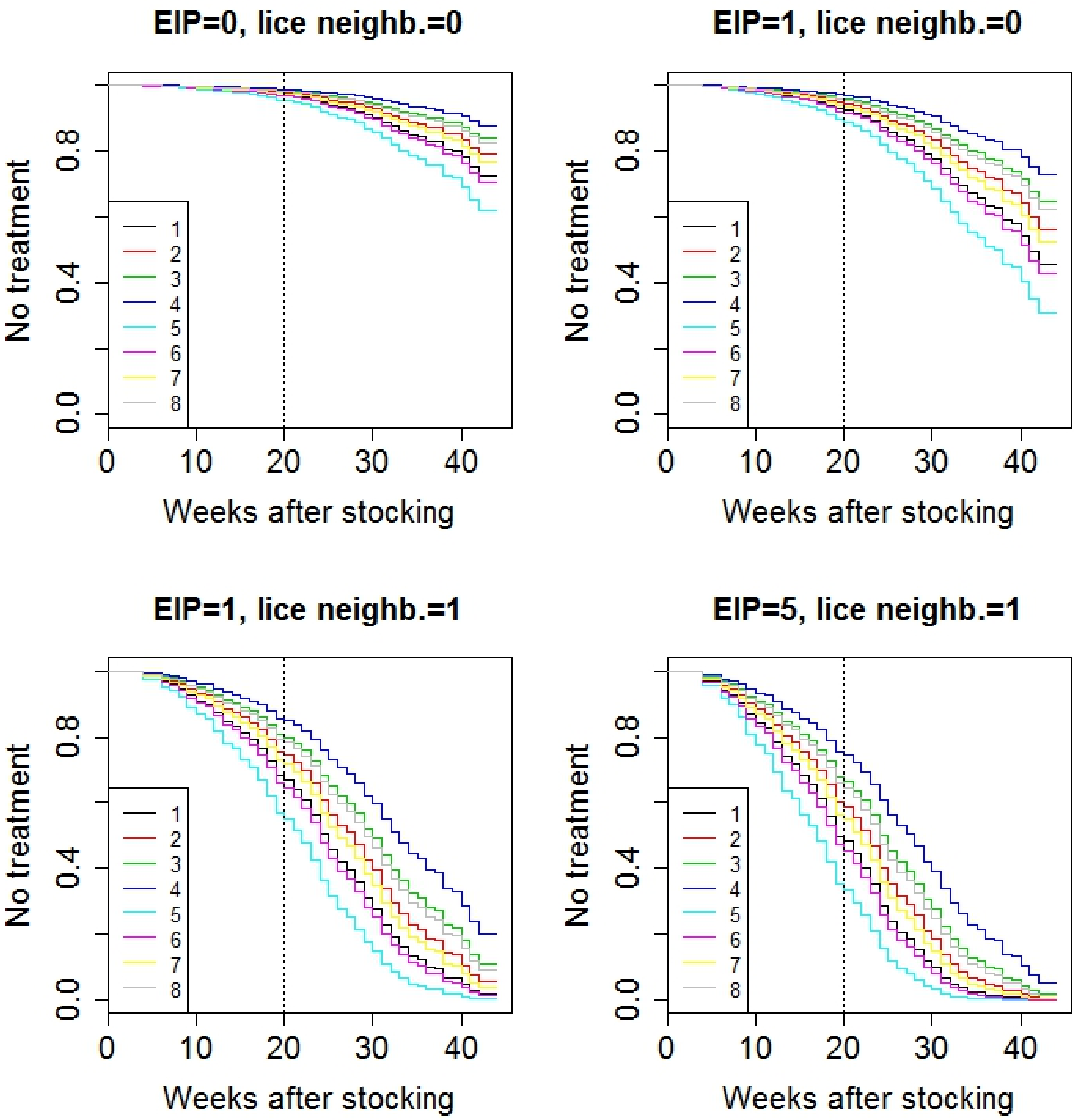
Survival curves

Probability curves can now be presented for different combinations of covariates. I chose to keep year of stocking and amount of fish in the cohort constant, and equal to 2013 and one million respectively. Furthermore, since the distance covariates affect the hazard rate similarly, I chose to use only one of them, keeping the other one constant. Thus distance to the nearest neighbor will be kept constant and equal to 5 km. Since geographical index is not part of the baseline in this analysis, one could see that the north-region has the highest treatment hazard, while the other two regions have a quite similar effect on the hazard. For this reason I chose to distinguish only between the northern part and the southern part of Norway when looking at survival curves.
The combinations of remaining covariates were divided into the following eight groups:

1. 6^oC in the sea, north-region, 5 km to slaughterhouse
2. 6^oC in the sea, north-region, 40 km to slaughterhouse
3. 6^oC in the sea, south-region, 5 km to slaughterhouse
4. 6^oC in the sea, south-region, 40 km to slaughterhouse
5. 10^oC in the sea, north-region, 5 km to slaughterhouse
6. 10^oC in the sea, north-region, 40 km to slaughterhouse
7. 10^oC in the sea, south-region, 5 km to slaughterhouse
8. 10^oC in the sea, south-region, 40 km to slaughterhouse

and plotted for different values of EIP and existence of female lice at the nearest farm. The results can be seen in figure 4.8. As expected when EIP is 0 and there are no female lice at the nearest farm, the probability of no need for treatment is high, and when EIP is 5 and there are female lice at the nearest neighbor, there is a high probability of performing a first treatment. One can see that group 4 (with low seawater temperature, in the southern region and long distance to a slaughterhouse) produces the best result considering the hazard rate for first treatment, while group 5 has the highest probability of early first treatment

Figure 4.8: Probability of no treatment for different combinations of explanatory variables, where some of them are time-dependent.



(high seawater temperature, in northern region and short distance to a slaughterhouse) and produces the poorest result in this analysis. The only difference between the two groups with the best and the second best result is distance to the nearest slaughterhouse, while temperature and region remain the same. Further one can see that groups 1 and 6 have almost the same probability of no treatment, and also groups 2 and 7, and groups 3 and 8. This indicates that probability of first treatment is somewhat similar when it is cold in the sea (6^oC) and the distance to the nearest slaughterhouse is short, and when it is warmer(10^oC) in the sea and the distance to the nearest slaughterhouse is long. Thus an increase in the seawater temperature from 6^oC to 10^oC has almost the same effect on the hazard as a decrease in the distance to the nearest slaughterhouse from 40 km to 5 km.

By looking at the plots one can see that 20 weeks after stocking the probability that there has been no treatment at the farm is approximately 95-100% when EIP is 0 and there are no female lice at the nearest farm, 90-98% when EIP is 1 and no female lice at the nearest neighbor, 58-83% when EIP is 1 and there are female lice at the nearest farm, and 35-78% when EIP is 5 and there are female lice at the nearest neighbor. One can also see that the difference between groups grows bigger as the EIP and lice-covariate increase. When there are no infection pressure and no lice at the nearest neighbor, it does not matter much whether the farm is located in the north or in the south, and whether the seawater temperature is low or high. However as the EIP and lice-covariate increase, the difference grows bigger, and it makes a significant difference whether a farm is contained in the south and seawater temperature is low, compared to a farm that is contained in the north and seawater temperature is high.

4.3 Comparison of results

The two models in this analysis did not differ much. When I compare them, I will look mainly at which covariates are important and which ones are not. For the covariates that appear as significant I will compare their influence on the hazard rate.

In table 4.16 one can see how the coefficients for every covariate differ in the models with time-dependent and time-fixed covariates. The model with time-fixed covariates is contained by one interaction term and seven covariates of which one was stratified. The model with time-dependent covariates is contained by eight covariates, includes no interaction terms and requires no stratification. The proportionality assumption holds true for all time-fixed covariates in the extended Cox model, while in the Cox PH model the assumption was not fully satisfied, and stratification was therefore needed.

Neighbor biomass density appeared as a significant covariate in univariate analyses and as non-significant in multivariate analyses, both when treated as time-fixed, with value obtained at time $t = 0$, and when treated as time-dependent, with values obtained at the start of each time-interval. EIP appeared as a highly significant covariate in both multivariate models. One reason for this result could be that biomass-covariate is somehow contained in EIP, and when EIP is included in the model, neighbor biomass density ”loses” its significance.

Table 4.16: Comparison of the coefficients in analyses with time-fixed covariates and time-dependent covariates.

Time-dependent covariates		**Time-fixed covariates**		
Lice neighbor (group 2):	1.62	Lice neighbor (group 2):	0.51	}+ negative interaction term
EIP ($\frac{1}{x+1}$):	-1.80	EIP ($\log(x+1)$):	0.96	
Amount of fish:	-0.06			
Temperature (lag-time effect):	0.10			
Dist.sl. ($\frac{1}{x}$):	1.81	Dist.sl. ($\frac{1}{x}$):	1.71	
Dist.neighbor:	-0.04	Dist.neighbor:	-0.06	
Year 2012:	-0.44	Year 2012:	-0.53	
Year 2014:	-0.45	Year 2014:	-0.47	
		Smolt weight:	0.05	
Geo.index (south):	-0.59	Geo.index is stratified, but lowest hazard in mid-region and highest hazard in north-region within 30 weeks after stocking.		
Geo.index (mid):	-0.77			

Non-significant covariates	
Biomass neighbors	Biomass neighbors
Smolt weigth	Amount of fish in the cohort
In-feed treatment	Temperature

One can see that when temperature and amount of fish in the cohort are treated as time-dependent, they appear as highly significant covariates. When covariates are used as time-dependent, their values are more up to date, and thus their effect on treatment hazard is captured better.

Smolt weight appears as non-significant in the model with time-dependent covariates. I chose to not introduce time-dependency in this covariate, as it increases almost linearly with time. When the model was extended, time-fixed smolt weight was no longer significant. In the Cox PH model this covariate was not among the highly significant covariates, and thus when other covariates that explained the effect on the hazard rate better were introduced, smolt weight lost its significance.

In-feed treatment could only be included as a time-dependent covariate, but it appeared as non-significant in the current model. Thus it can be concluded that in-feed treatment does not seem to affect time till first treatment of salmon lice at the aquaculture facilities according to this model.

Let us now look at the coefficients addressing to each covariate, by looking at the table 4.16. Covariates, that are significant and time-fixed in both models seem to have approximately equal effect on the hazard rate since their coefficients are very similar. This applies to distance covariates, year and somehow geographical index. Since geographical index is stratified in the analysis with time-fixed covariates, it is contained in the baseline and there are no coefficient values to compare. However one could see in figure 4.5 that treatment hazard rate is the highest mainly at the farms located in the north-region of Norway, and the lowest at farms located in the mid-region of Norway, while treatment hazard in the south-region is quite similar to in the mid-region. The model with time-dependent covariates shows that the hazard rate for first treatment is the highest in the northern part of Norway and the lowest in the mid-region. Covariates addressing to EIP and existence of lice at the neighboring farms have an interaction term when they are treated as time-fixed, and EIP is transformed differently in the two models, and thus it is difficult to compare the coefficient values. However this was discussed in 4.2.3, and one can conclude that EIP and the lice-covariate have a similar impact on the models: as expected their increase leads to a higher treatment hazard rate.

Chapter 5

Validation of Cox PH model

In this study I constructed a Cox PH model and then extended it to contain time-dependent covariates. I built the extended Cox model to increase my understanding since some of the covariates vary with time. The differences between the two models were not big, and a validation of only one of the models can help us get a general indication of the strength of both models. I chose to validate the Cox PH model since there are some technical issues related to validating the extended Cox model using the below methods.

5.1 Methods for validation

Theory material in this chapter is adapted from Royston and Altman 2013.

When validating Cox models, one must distinguish between internal and external validation. Internal validation means to use parts of the same data that was used to build the model to evaluate it, whereas external validation is performed by applying a developed model to an independent dataset. In this study I will use only external validation by using data from year 2015.

An essential part of model validation is computation of prognostic index (PI). Prognostic index is the linear predictor, $\mathrm{PI} = \hat{\boldsymbol{\beta}}^T \mathbf{x}_i$, in the expression 2.2, so the hazard function can be expressed as $\alpha(t) = \alpha_0(t)\exp(\mathrm{PI})$. PI is computed by applying coefficients ($\hat{\boldsymbol{\beta}}$) from the developed model that one wants to validate to the data ($\mathbf{x}$) in derivation and validation datasets. The values of PI in the validation dataset will be compared to the values of PI in the derivation dataset. There are different ways of validating a Cox model, and I will use three methods to validate the model in this study. I chose to validate the constructed Cox PH model by comparing it with the new data, and not with a new model fitted to the new data. The methods are described below.

Regression on PI and check for model misspecification

One method of model validation is to estimate regression coefficient on PI from the validation dataset by modelling the hazard function as $\alpha(t) = \alpha_0(t)\exp(\beta \cdot \mathrm{PI})$. By doing so, one gets insight into whether the relative hazard is specified correctly by the model. If the new data is consistent with the model, the coefficient addressing to PI in the validation dataset should be close to 1. The result after a likelihood-ratio test should then show that PI is not significantly different from 1. It is worth noting that in a perfect case, as in the derivation dataset, the coefficient is exactly 1. If the coefficient differs from 1, the explanatory power of the model, applied to the new dataset, is poorer.

If the coefficient of PI in the validation dataset differs from 1, the data should be analysed more closely. The discrepancy may be due to the difference in regression coefficients for some of the covariates in the two datasets. One can then test this by applying Cox regression to the covariates in the validation dataset and offsetting PI. Offsetting a covariate means to specify its coefficient as 1, instead of estimating it. (I chose to ignore the uncertainty in the estimates in the derivation dataset.) The $\hat{\boldsymbol{\beta}}$'s corresponding to the remaining covariates should now have values close to 0. The reason is that when one offsets PI, the new $\hat{\boldsymbol{\beta}}$'s correspond to the difference between estimated coefficients in the model fitted to the derivation dataset and estimated coefficients in the model fitted to the validation dataset. In a perfect scenario the difference between them should be 0. If lack of fit is found in some of the covariates, they should be examined more closely, as lack of fit could be caused by dissimilarity in the definition, measurements or units of variables between the validation dataset and the derivation dataset.

Measures of discrimination

Another method of validation is to measure discrimination, which is a measure of difference between survival for different risk levels in both validation and derivation datasets. Maintaining a discrimination in the validation data similar to in the derivation data is a substantial part of model validation. There are several ways to measure discrimination, and I will focus on two measurements in this study: Harrell's concordance index and Royston-Sauerbrei D-statistics. Both are based on the PI, and no new models need to be fitted to the data. I am aware that stratified geographical index is not included in the PI in the current model, but I choose to ignore that fact here.

Harrell's concordance index is defined as proportion between all of groups or proportion between all of individuals whose predictions and outcomes are concordant. It can be interpreted as estimated probability that of two randomly chosen cohorts, the cohort with the lower PI will outlive the cohort with the higher PI. If the concordance index is close to 0.5, it implies that the prognostic index is no better than flipping a coin to determine which cohorts are in a better situation. In the current study concordance index will be compiled as proportion of all stockings where PI decreases when time increases. If PI is lower for a cohort that actually has longer time till first treatment than another cohort, the prediction for that pair is concordant with the actual outcome. I will look at all possible pairs of cohorts that were actually treated, excluding censored events. Mathematically concordance index can be expressed as:

$$\frac{1}{N}\sum_{i,j}\left(1\cdot\mathbf{I}(T_i^*>T_j^*)\cdot\mathbf{I}(\mathrm{PI}_i<\mathrm{PI}_j)+0.5\cdot\mathbf{I}(T_i^*>T_j^*)\cdot\mathbf{I}(\mathrm{PI}_i=\mathrm{PI}_j)\right),$$

where T_i^* is time to an actual treatment for cohort i and N is the amount of all possible pairs where $T_i^*>T_j^*$.

Royston-Sauerbrei D-statistics, which will be denoted by R_D^2, is a "measure of explained variation on the log relative hazard scale based on the authors' D statistic. D measures prognostic separation of survival curves, and is closely related to the standard deviation of the PI" [Royston and Altman 2013]. In this thesis I will report only how R_D^2 is calculated, and the results will be interpreted. The motivation behind calculations and formulas can be found in the article by Royston and Sauerbrei 2004. To calculate R_D^2 one needs to start by computing the rankits of ordered PI (rankits are expected values of order statistics of a sample from the standard normal distribution), and dividing them by a factor $\kappa = \sqrt{8/\pi} \simeq 1.60$. From there a Cox regression has to be performed on the scaled rankits, and the estimated coefficient in the model is D. Royston-Sauerbrei D-statistics is then computed by

$$R_D^2 = \frac{D^2\kappa^2}{\sigma^2 + D^2/\kappa^2}$$

where standard deviation is chosen to be $\sigma^2 = \pi^2/6 \simeq 1.64$ for proportional hazard models. R_D^2 is a type of R^2, a coefficient of determination that indicates how well data fits a given statistical model. In this case R_D^2 will measure explained variation. Usually PI in the validation data is less spread out and thus less heterogeneous than in the derivation data, which might be reflected in smaller explained variation.

Survival curves for risk groups and hazard ratios across them

By categorising PI one can define different risk groups in the derivation and validation datasets. PI needs to be grouped at the same cut points because of the further comparison. It can be categorised into several groups, for example at 25th and 75th percentiles in the derivation dataset, producing two smaller groups at relatively low and high risk, and one large, central group. Survival curves can then be constructed for the different risk groups in the derivation and validation datasets, and one can make a visual comparison of the trend in the groups. The exact amount of risk groups, and exactly how to define cut points, is of free choice. At the same time a moderate number of risk groups (fewer than five) is preferable because a large number of groups could cause unstable survival curves, and the difference between groups is likely to be smaller.

By looking at the survival curves one can see the difference between risk groups in each dataset: if the curves are widely separated, the difference between the risk groups is substantial. Furthermore one can compare the risk groups in the two datasets by observing whether the curves (or the trend in the curves) are similar.

A useful complement to survival curves is a table of hazard ratios between the groups. The ratios are easily obtained by fitting Cox regression to the categorical PI and comparing the difference between risk groups and the datasets. However the hazard ratios will not capture whether one of the risk groups is similar between datasets, as the ratios only show the difference between the two groups in the same dataset.

5.2 Validation dataset

As mentioned above I will use external validation in this study. Now I have access to data from 2015 in the Aquaculture Database, and I can see that in year 2015 there were 786 active salmon farms along the Norwegian coast. I will use data from cohorts stocked during the time between week 14 of 2015 and week 52 of 2015. I will then look at the time till first treatment within week 6 of 2016, and censor the events that happen after that date. This produces 144 observations, where 59 are censored and 85 are actual treatments.

Data from 2015 are actually not the best data to use for model validation as a lot of unexpected things happened in 2015 that were not faced earlier. In the south-region of Norway it appeared to rain a lot. This caused salt proportions in the sea to decrease. Salmon lice do not survive in freshwater, and therefore the lice situation in southern parts of Norway was very good in 2015, and farmers did not have any major need for treating their fish and thus could wait longer until first treatment. In the mid-region, however, farmers were not as lucky as farmers in the south since they met some huge treatment resistance problems and lost control over their lice situation, which probably led to unsuccessful treatments and high infection pressure. In northern parts of Norway the Infectious salmon anemia virus (ISA) was spread, leading to a huge harvest of fish at some localities, for example Lofoten. This will lead to early censoring in the dataset, as well as a lower amount of fish in the fjords and thus fewer lice hosts and probably lower infection pressure.

In addition to the above-mentioned problems I discovered some cohorts with some plausible information when the dataset was constructed. At 24 farms treatments occurred within eight weeks after stocking. Delousing happened between weeks 22 and 23 (end of May) even if nothing really indicated on a bad lice situation as EIP was low in all cases. All cohorts were located in the south-region of Norway, and I assume that those treatments were organised spring treatments. The intention is then to lower the level of lice, in regards to wild salmon smolts moving from freshwater into the sea during that period. Such treatments do not take place because the amount of lice is higher than the maximum level defined by FAS, but primarily to control and coordinate treatments when the level of lice is low. There were no similar cases in the original dataset, and thus I chose to exclude those cohorts from the validation dataset.

Another challenge is that in the current model, year is one of the covariates with three factors: years 2012, 2013 and 2014. The validation dataset is constructed by the data from year 2015, meaning that there are no coefficient values addressing to this covariate. There are then two possible choices: I can either refit the model without year as a covariate and validate the new model, or I can calculate the weighted mean of the coefficients addressing to year and use it as coefficient for year 2015. I chose the latter for my validation. Coefficient addressing to year 2015 will then be computed by:

$$\frac{\hat{\beta}_{2012} \cdot c_1 + 0 \cdot c_2 + \hat{\beta}_{2014} \cdot c_3}{c_1 + c_2 + c_3},$$

where c_1 is the amount of observations corresponding to year 2012, c_2 is amount of observations corresponding to year 2013 and c_3 is amount of observations corresponding to year

2014. Since year 2013 is the reference year in the developed model, the corresponding β_{2013} has a value of 0.

One important step after the dataset is constructed is to check whether the values of each covariate in the validation dataset lies in the same interval as in the derivation dataset. If values of some covariates would expand over a larger time-interval, it would lead to extrapolation. However all covariates in the validation dataset lie within interval in the derivation dataset.

5.3 Investigation of the data and validation performance

Since the situation was so different in the three regions of Norway in 2015, and geographical index was a stratified covariate in the Cox PH model, I will perform method 1 model validation (regression on PI) for one region at a time, to investigate whether data from 2015 for some of the regions do not fit to the model due to salt proportion in the sea, resistance or the ISA virus. I will then combine the data from the regions that are the best match for the model.

I started by constructing Nelson-Aalen plots for each covariate separated by geographical index in the validation dataset, and by comparing them with Nelson-Aalen plots for corresponding covariates in the derivation dataset. I chose not to divide the plots from the derivation dataset by regions since I was interested only in comparing validation data to a general trend in the derivation data, and in investigating whether some of the covariates in the validation dataset deviate from the derivation dataset. From the plots I could see that validation data from the mid-region of Norway differed completely from the derivation dataset and from the data from south- and north-regions in the validation dataset. Especially Nelson-Aalen plots addressing to EIP and to distance to the nearest slaughterhouse showed some huge deviance for mid-region data. When plotting hazard curves I divided each covariate into three groups based on the distribution of the data in the derivation dataset. The plots of hazard rates with EIP and with distance to the nearest slaughterhouse as covariates can be seen in figure 5.1. One can see that Nelson-Aalen plots addressing to EIP in the validation dataset for south- and north-regions show a similar trend in the curves as the derivation dataset, while the Nelson-Aalen plot for the mid-region shows an opposite effect on the treatment hazard. When it comes to Nelson-Aalen plots addressing to distance to the nearest slaughterhouse, one can again see that for the mid-region the effect on treatment hazard seems to be opposite to the effect in the derivation dataset. However the plots from south- and north-regions do not have the same clear trend as the derivation dataset either, and the curves are not well separated for the three groups in the validation dataset.

Further I constructed PI by using coefficient values which can be found in listing 1, and by using coefficient addressing to year 2015: -0.31. The distribution of the PI in the two datasets can be seen in figure 5.2. When plotting PI, I centered it on "average risk" according to the derivation dataset by subtracting the mean of 0.71 from PI in both datasets. The mean (and standard deviation) of the values in the histogram is 0.00 (0.97) and -0.26 (0.75) in the derivation dataset and validation dataset respectively. By looking at the histogram

Figure 5.1: Nelson-Aalen plots of time-fixed EIP and distance to the nearest slaughterhouse in the derivation dataset and in the validation dataset divided by the regions.

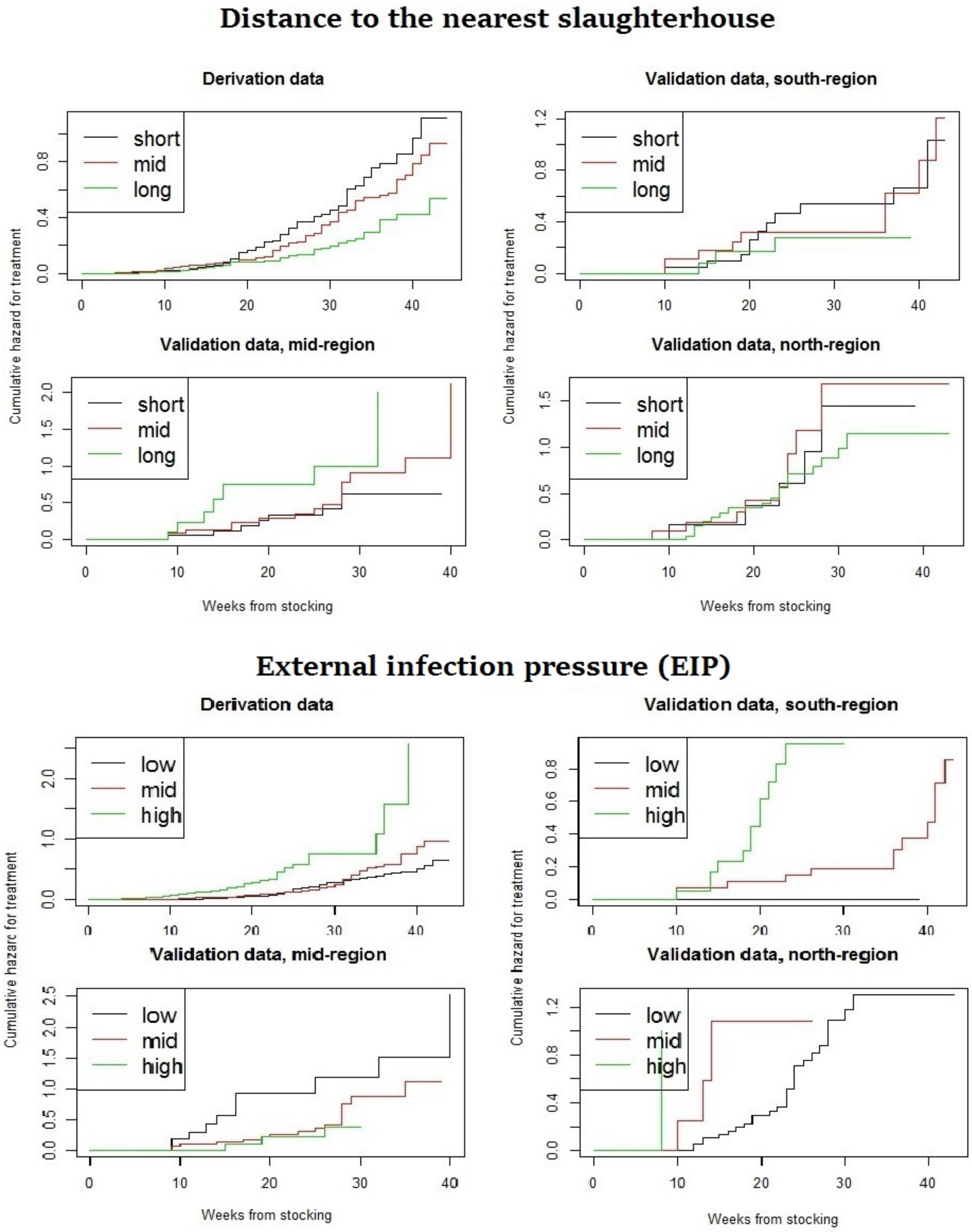
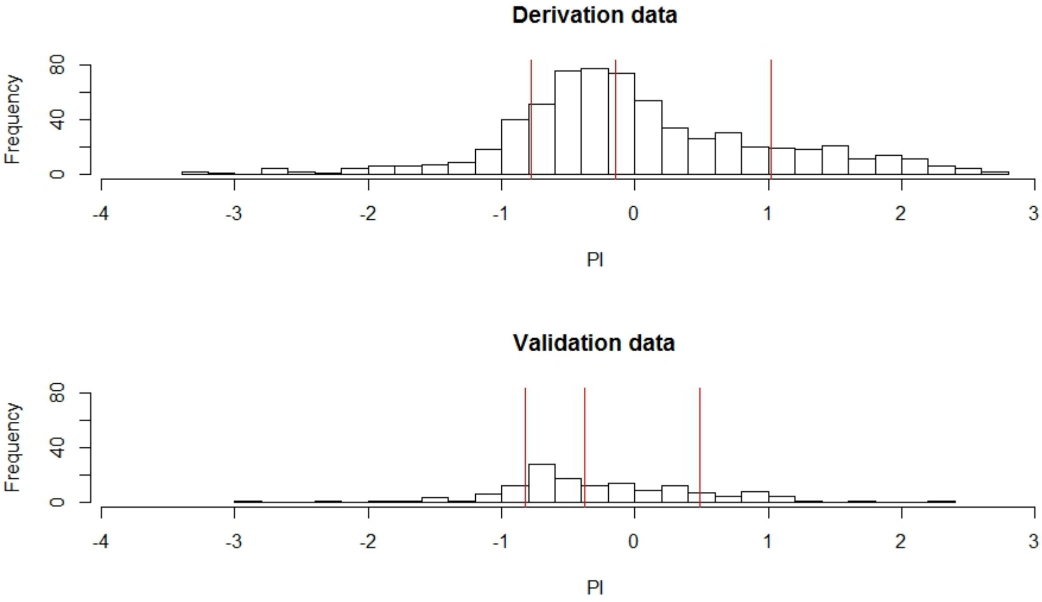


Figure 5.2: Histogram of the PI (time-fixed) in both datasets, centered on the mean from the deviation dataset. The vertical lines show the 16th, 50th and 84th percentiles of PI in each dataset.



one can see how the PI is spread, and one can also see the general level of log relative hazard. One can observe that the validation data is more left-aligned compared to the derivation data. However it seems that PI in both datasets follows the same trend. This type of histogram may also show an indication on outliers in the validation dataset that might lead to extrapolation. In this validation dataset there are no such outliers as one can see in figure 5.2.

When I performed regression on PI in the validation data, I started by using three datasets: data from the south-region, data from the mid-region and data from the north-region. The slopes in Cox models on the PI with the three datasets were then 0.79 (se 0.33), -0.14 (se 0.31) and 0.37 (se 0.44) respectively. Data from the south-region seem to validate the model best, while data from the mid-region of Norway do not validate the model well. As mentioned before the mid-region had huge treatment resistance problems in 2015 that might be the main reason that the data do not fit to the model. Two independent studies, Kaur et al. 2016 and Jansen et al. 2016, show that the mid-region of Norway had the most increasing resistance problems. These studies used data from years 2013 and 2014 which means that there are no particular study addressing to data from 2015. However veterinarians know that the resistance problems worsened in 2015. Especially in the mid-region the lice problem became unmanageable. Kaur et al. 2016 concludes that the worst increasing resistance problems appeared to be in Nord-Trøndelag county. The county is located between 63.3 and 65.2 decimal degrees of latitude. Jansen et al. 2016 concludes that probability of adult female lice dying from a particular treatment in year 2013 was low at latitudes of about 64-66 decimal degrees, and the corresponding predicted probability was even lower in 2014. In year 2015 lice problems increased to high levels in the mid-region of Norway.

Lice levels came out of control due to resistance towards medical treatments [Peder Jansen, personal communication]. Even if treatment resistance problems were increasing from year 2013 to year 2014, farmers could still control them. Since farmers in the mid-region lost control completely in 2015, I chose to exclude some of the cohorts from that region from the validation dataset as there are many indications that the dataset from the mid-region is not compatible with the developed model. In my dataset the mid-region is defined to be between 62 and 67 decimal degrees north. This means that most likely there are some observations from the mid-region that are compatible with the derivation dataset. After a number of attempts – investigation of Nelson-Aalen plots and regression on PI – I decided to keep farms located north of 66 decimal degrees in the mid-region of Norway. Meaning that I chose to exclude farms located at 62-66 decimal degrees north from the validation dataset. The slope in the Cox model on the PI with mid-region data is now 0.93 (se 0.85). A high standard error is caused by too few observations as there are only 11 observations from the mid-region of Norway now.

I tried to explore the data from the north-region, but did not find any cohorts that appeared to be plausible. EIP was quite low in all observations from the north-region in year 2015, independent of whether or not farmers performed treatments. Low EIP values in these cases were most likely caused by early harvesting due to the ISA virus. However there were no obvious indications as to which observations led to a poor validation of the model. Therefore I chose to include all pieces of information from the north-region in the validation dataset.

The datasets from the three regions can now be combined, and a validation on the whole dataset can be performed. There are now 106 observations in the validation dataset of which 42 are censored events and 64 are actual treatments. Since geographical index is part of the baseline in the current Cox PH model, there are no coefficients corresponding to this covariate, and it is not contained in the PI. As mentioned, the lice situation was very different in the three regions in 2015. For these reasons I chose to stratify geographical index when preforming the method 1 validation in the dataset. Additionally separate survival plots for each region will be constructed when a method 3 validation will be performed.

Method 1: regression on PI and checking for model misspecification

The coefficient addressing to PI in the validation dataset, with stratified geographical index, is 0.68 (se 0.26). I performed a likelihood-ratio test where I checked whether the slope is significantly different from 1. P-value after the test was 0.22, which means that it is not significantly different from 1, and the data validates the model with time-fixed covariates well according to this method.

Even if the likelihood-ratio test indicated that the coefficient addressing to PI in the validation dataset is acceptable, the coefficient is not equal to 1. Thus I would like to check which covariates lead to that small lack of fit in the validation dataset. In table 5.1 one can see the result after a Cox regression on the covariates in the validation dataset when PI is offset. The covariates are transformed and grouped in the same way as in the original model. In an optimal case all coefficients should take zero-values. By looking at the p-values in table 5.1, one can see that the model spesification is maintained in this case since all

Table 5.1: Cox regression on the time-fixed covariates in the validation dataset with the PI offset.

	coef	se(coef)	p-value
EIP (transformed)	-0.13	0.33	0.68
dist.sl. (transformed)	-2.60	1.99	0.19
dist.neighb.	0.03	0.03	0.29
weight	-0.02	0.04	0.60
lice neighb. (group 2)	-0.31	0.33	0.36
EIP (transformed) : lice neighb. (group 2)	0.43	0.34	0.20

p-values addressing to the covariates are high enough, meaning that none of the covariates are significantly different from 0. However one can see that the main lack of fit is caused by the covariate addressing to distance to the nearest slaughterhouse (since the coefficient value is the highest and p-value is the lowest for this covariate). This was expected, as the Nelson-Aalen plots constructed earlier indicated that this covariate was partially different in the validation and derivation datasets. Most important: both regression on PI and the model check indicated that the new data are consistent with the constructed model.

Method 2: measures of discrimination

Harrell's concordance index and Royston-Sauerbrei D-statistic are shown in table 5.2 for both datasets. Concordance index is high enough in the derivation dataset and indicates that time to actual treatment and PI are concordant in approximately 65% of the cases. When it comes to the validation dataset, one can see that the concordance index is only 50%. This implies that the discrimination in the validation dataset is not well maintained. D-statistics too differs between the datasets, and shows that there is more explained variation of the treatment hazard in the derivation dataset. An indication of this could also be seen in figure 5.3, as one could see that PI was less spread in the validation dataset than in the derivation dataset, and thus the explained variation is smaller. These results simply indicate that the model works better on the data of which the model is made.

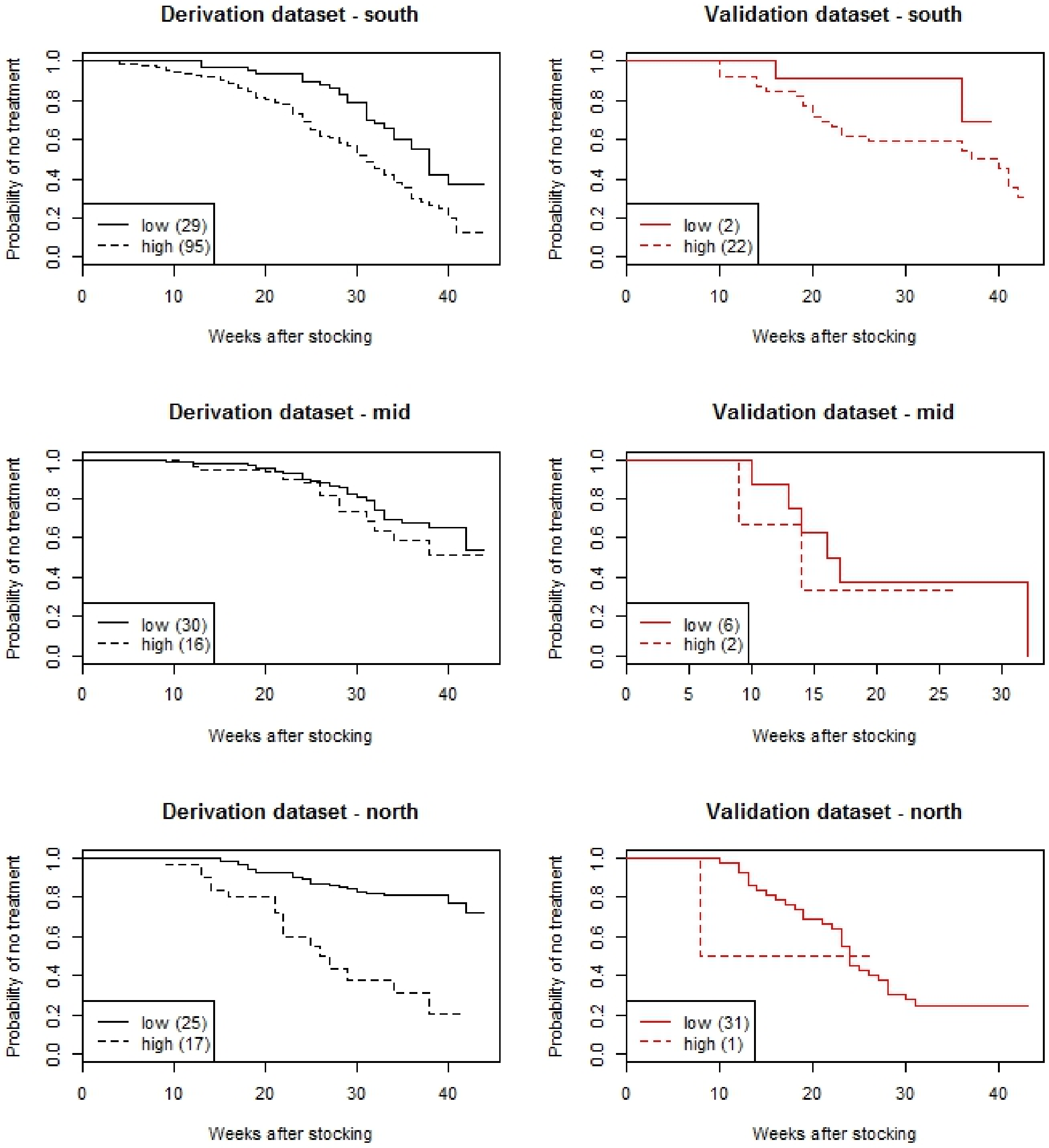
Table 5.2: Discrimination measures evaluated in the derivation and valuation datasets with time-fixed covariates.

	Derivation data	Validation data
Measure	Estimate	
Concordance index	0.64	0.51
R_D^2	0.34	0.13

Method 3: survival curves and hazard ratios

Since there are only 106 observations in the validation dataset (where only 64 are actual treatments), I chose to divide PI into only two risk groups: low risk and high risk of first treatment. As mentioned before I will distinguish between the three regions, as the lice situation was different in different parts of Norway in 2015 and survival curves of PI in the current model do not take into account the difference between regions. I chose to divide PI at the 50th percentile in the derivation dataset. The groups are then consequentially defined as PI lower than 0.56 and PI higher than 0.56. Amount of actual treatments in each group is shown in table 5.3 since only actual treatments are used when survival curves

Figure 5.3: Survival curves of time-fixed PI in the three regions in both datasets. Groups are divided at 50th percentiles in the derivation dataset, such that there are low and high risk groups. Amount of actual treatments in each group can be seen in the parentheses.



are constructed, and when hazard ratios are computed. The amount of observations in the validation dataset is small, and therefore dividing PI into not more than two groups is the best choice in regards to survival curves. One can see that in the validation dataset the north-region has only one actual treatment at high risk, and the south-region has only two treatments at low risk. Also in the derivation dataset the highest amount of observations at high risk are in the south-region, and at low risk in the north-region. The mid-region is contained of few observations in the validation dataset as a consequence of excluding localities with resistance problems.

Figure 5.3 shows survival curves for the two risk groups in different regions. The curves divided by regions are separated clearly in the two datasets, and high risk groups seem to have higher probability of first performed bath treatment than low risk groups. For mid-region data the difference between the groups is not so clear, but one can still observe that the high risk group appears to have a higher treatment hazard rate than the low risk group. One can also see that the groups of PI in the validation dataset seem to have lower probabilities of no treatment compared to the groups of PI in the derivation dataset, and thus the survival curves in the validation dataset do not agree perfectly with those in the derivation dataset. The curves indicate that the model leads to overestimation of risk when applied to the validation dataset. However there are not enough observations in the validation dataset to make any conclusions, apart from that the hazard rate is higher in high risk groups in both datasets.

Hazard ratios between risk groups are presented in table 5.3, and one can see that they differ between the datasets. For mid-region data there is almost no difference between hazard ratios in the two datasets. The biggest difference seems to be in the north-region, as the treatment hazard seems to be approximately five times higher in the high risk group compared to the low risk group in the derivation dataset, and only 12% higher in the validation dataset. However there is only one observation of actual treatment in the high risk group in the north-region in the validation dataset, which means that this result is highly uncertain. The same applies to the south-region data, as there are only two observations of actual treatments in the validation dataset. However the hazard ratios indicate that high risk groups have higher hazard rates than what low risk groups do, and this is a satisfactory result when the validation dataset is so small.

Table 5.3: Distribution of actual treatments in low risk and high risk groups in three regions in the derivation and validation datasets, together with hazard ratios of PI's evaluated in both datasets

	Derivation data			Validation data		
	Distribution		Hazard ratio	Distribution		Hazard ratio
Regions	Low risk	High risk	high vs low risk	Low risk	High risk	high vs low risk
South-region	29	95	2.14	2	22	3.15
Mid-region	30	16	1.44	6	2	1.44
North-region	25	17	5.36	31	1	1.12

5.4 Conclusion of validation

I conclude that the Cox PH model is well validated according to some of the chosen methods. The discrimination is not well maintained in the validation dataset, but this is most likely caused by the different lice situation in year 2015. Not all of the methods led to the same conclusions about validation, but it is useful to look at different methods to understand what exactly lead to different validation results.

The discrimination measures indicated that the validation dataset is not perfect for the valuation of this model. However the regression on PI indicated that the 2015 data validated the model well.

As mentioned before the validation dataset is contained by too few data. Especially when the data need to be divided in three region-groups, the amount in every group is small. In a perfect scenario I would like to have more observations in a validation dataset, and to have data that contains no plausible information. In real life this is unfortunately not the case, and situations change from year to year due to factors beyond one's control. If there was some exact information about lice resistance at the farms and salt proportion in the sea, some reasons for poor validation would probably be eliminated.

A possible improvement to the model would be to construct a new model by using 80% of the data from years 2012, 2013, 2014 and 2015. This would then give an exact coefficient addressing to year 2015 that could have captured some of the problems in 2015. The remaining 20% of the data could be used to validate the model. This would definitely improve the discrimination results. However when I started this study, the 2015 data was not ready, and for practical reasons I constructed the model by using data from 2012, 2013 and 2014, while 2015 data was used for model validation.

Problems linked to validation of the extended Cox model

As already mentioned there are some technical problems linked to validation of the extended Cox model. For this reason I chose to perform the validation of the Cox PH model and assume that since the results after the two models were so similar, the extended Cox model is also validated well by the data. In this section I will discuss briefly the main problems that prevent one from performing the above validation on the extended Cox model.

In the dataset with time-dependent covariates there are several observations linked to each time-interval from stocking till first treatment or censoring. There are several measurements of each covariate at each interval divided by four weeks, and since time-intervals are split, there are many more censored events in the dataset with time-dependent effects. This leads to elimination of methods 2 and 3 since method 3 and parts of method 2 are based on observations that contain actual treatments. The values of covariates linked to the actual treatments in the extended Cox model are only contained in the last four-week interval, eliminating the previous observations linked to the same event.

Harrels concordance index is based on the cases where PI decreases as time till first treatment increases. However PI linked to time till actual treatment is only the last value of PI in the whole time-interval. This will not necessarily reflect the proportion of concordant events, as the values of PI might be quite equal at the time of treatment, and not necessarily much lower when time till treatment is high, since PI develops over time for each cohort. Royston-Sauerbrei D-statistic is based not only on the observations linked to every actual treatment, but on all observations. However there are no methods in the calculations that link the observations that are related to the same observation of treatment or censoring, and all PI's will be treated independently. Therefore this validation method is not possible either, in an extended Cox model with ancillary time-dependent covariates.

When survival curves, as seen in figure 5.3, are plotted, jumps in the curves are based on observations that contain actual events – even in an extended Cox model. Therefore, again, since there are more than one observation linked to each actual treatment, it is inappropriate to perform such a validation.

Method 1 validation, which is regression on PI, ought to be possible, at least technically, since one now has access to new data, which can be constructed to contain similar time-intervals to the ones in the derivation dataset with time-dependent covariates. This method should take into account that there are several measurements linked to each time-interval and thus should produce a correct result. However it is unclear whether this is a proper approach, and I will not pursue it further.

Chapter 6

Discussion

6.1 Challenges with the dataset

Modelling the time till first lice treatment based on various sources of information about Norwegian salmon farms has been an enormous cross-disciplinary and rewarding challenge. One of the main challenges in this thesis has been to construct the dataset. I had access to a lot of data from the Aquaculture Database, and as a statistician I could not see the things that were obvious to veterinarians: what values of the covariates should and should not be used, and that some covariates had to be modified before they were used (for example smolts that weigh less than 250 grams, and that I needed to check whether the nearest neighbor was active in the period of interest). This resulted in several attempts to perform the analysis before new possible and necessary modifications to the dataset were discovered and made. A continuous dialogue with veterinarians has been essential in this work.

Additional challenging decisions addressing to covariates was whether updated values of the covariates could be used. Since I had access to weekly and monthly reports of the data from years between 2012 and 2014, it seemed logical to be able to use mean, max or min values for each covariate from the time of stocking until first treatment. However the baseline in Cox regression models is computed for all possible time points, and therefore at the first time point the future values of the covariate are not known. If one uses future values in the regression, the results after Cox regression will be useless. Thus one always needs to be careful that the values used in regression are known at the beginning of each defined time-interval.

Another problem I encountered when I constructed the datasets (for both analyses and the validation) was spring delousing. There is not enough information to identify whether or not a treatment happens due to spring delousing. Farmers are known to treat their fish in March/April to prevent a high level of lice in May due to requirements of a lower level of lice in that time period. Sometimes several localities unite to treat their fish simultaneously to ensure that the lice level will be kept low in the whole fjord. Those treatments happen independently of the lice level and often when the level of lice is low, and are not of the main interest in this analysis. However there are also cases where farmers actually have a high level of lice between March and May, and those treatments are of interest in this analysis. If there was any information available indicating whether a treatment was spring delousing, I could have defined it as a censoring event and thus censor an observation when spring delousing occurred. I would then have an improved and more precise dataset. Since I did not have that information available, I have chosen to exclude all of the cohorts that are stocked before week 14 every year and censor the events that occur after week 6 the following year as described in section 3.2.

More challenges appeared when I performed the analysis. Only a few covariates appeared as linear throughout the analysis, and the majority of the covariates changed their linearity when new covariates were included. Thus I had to perform new linearity tests every time a new covariate was added to the model, which delayed the construction of the final model.

6.2 Covariates and estimated effects

The effects identified as important for time till first lice treatment are as expected (in direction) by the salmon lice experts at the Veterinary Institute. Firstly let us consider EIP as a risk factor. This covariate appeared to be highly significant both when treated as time-fixed and as time-dependent. This means that infection from neighbors plays a major role in whether time till first treatment will be short or long. When lice levels at the neighbors within a 100 km distance increase, EIP increases, which in turn leads to an increased treatment hazard rate at the farm of interest. Thus the amount of lice in the area affects the time till first treatment. Furthermore both distance variables appeared to be significant in both models, meaning that when a farm is located further away from another farm containing fish, the need to treat decreases. In practice farmers cannot control the distance to neighbors, nor to the slaughterhouses. They can, however, keep that fact in mind. If they are located close to a neighbor or a slaughterhouse, they are highly exposed to infection. Geographical location is also a significant covariate that addresses to localities, but that unfortunately cannot be controlled fully. One could see that the hazard rates for first treatment were different in different parts of Norway, and that the situations changed from one year to the next, making it difficult to determine what regions are best or worst considering the lice-levels.

One interesting point of view is to look at the significant covariates that can be controlled by farmers. Even if farmers cannot control their neighbors, they can control their own lice levels and the amount and weight of fish at their own farms. The results of the analyses in this thesis showed that smolt weight was a significant covariate in the Cox PH model, and that amount of fish at the farm was a significant covariate in the extended Cox model. Farmers would then benefit from stocking smolts when they are as small as possible to prolong time till first treatment, according to the constructed Cox PH model. However if the smolts are small at stocking, they need to stay longer in the sea before they reach proper harvest weight. This in turn results in higher costs of keeping the fish in net-pens, and thus farmers lose part of the win. Farmers would do well to decide what risk are they willing to take: early first treatment or the cost of keeping fish longer in the sea. Since this covariate was not one of the highly important ones, I would suggest that farmers avoid stocking smolts when they are too small, for financial reasons. When it comes to amount of fish in the cohort, the results from the analysis in this thesis indicates that the treatment hazard is lower when there are many fish at a farm. Since only time till first treatment is studied in the current analysis, this result indicate that the infecting lice will have more fish to distribute onto, when many lice-hosts are present at a locality. This leads to longer time until there are approximately 0.5 female lice per fish. Farmers, therefore, could benefit from keeping many fish on their farms to postpone time till first treatment. However, as mentioned before, this is a short term effect, since a farm that contains many lice hosts will get a higher amount of reproductive female lice in the long run. This will lead to a higher infection pressure within the farm by the end of the production period, and probably to the need of treating their fish several times after achieving higher lice levels. Furthermore since the neighbor situation too is highly significant, farms that stock many fish will not only cause problems for themselves in the long run, but also for their neighbors.

Type of cohort is another covariate that deserves closer consideration. Type of cohort refers to the season (spring or fall) during which the cohort has been stocked. In the current analysis this covariate appeared to correlate highly with EIP and was for that reason omitted from the analysis. Additionally there is some uncertainty related to fall cohorts, as described in section 4.1.2. Furthermore spring fish are stocked when the lice level is at its lowest (between April and May), while in fall fish are stocked when the lice level is at its highest (between September and October). One possible improvement or extension of the analysis is to make two independent models, where the first one includes only the data with spring cohorts, and the second one includes only data with fall cohorts. I could then analyse whether there would be any differences between the results.

Another covariate that should be discussed here is in-feed treatment. In the current analysis it appeared as non-significant, both in univariate and multivariate analyses. This cannot be explained by in-feed treatment being contained in other covariates since it is an independent, not modelled covariate that did not correlate with other covariates. The question for discussion is then: is it really true that in-feed treatment has no impact on time till first bath treatment? If the answer is yes, why do farmers insist on in-feed treatments? One possible reason for the covariate being non-significant is that the value of the covariate used in the analysis may not be optimal. I used it as a history variable: whether farmers performed in-feed treatments during the time-interval from stocking until time t. If I had enough information to distinguish between the different in-feed treatments, I could have used more precise values of the covariate, but unfortunately that information was not available. Another improvement could be to fit a new model with in-feed treatment as a part of the response variable. I could for example then look at the time from in-feed treatment until bath treatment or a given amount of lice at the farm, and analyse whether they were related in any way.

About the results related to the two different models: both of them produce satisfactory results. However there are benefits and drawbacks to both. The model with time-dependent covariates is trickier to compile, analyse and validate. However by extending the Cox PH model to include time-dependent covariates, I found more information about what affects time till first treatment since two new covariates appeared to be significant: seawater temperature and number of fish in the cohort. I expected that seawater temperature would be significant in this analysis, as several studies (as Stien et al. 2005 and Jansen et al. 2012) showed that temperature is highly significant when it comes to lice reproduction, and thus it ought to be significant when it comes to lice treatments. The extended Cox model showed that a lag-effect of seawater temperature that varies with time is highly significant for the treatment hazard, meaning that it is not that important what the seawater temperature is during the current month, but rather what it was during the previous month. The covariate addressing to amount of fish in the cohort indicated that having more fish at a farm leads to a lower hazard for first treatment, as discussed above. This is an interesting observation, however, this is not a result that I would like to present to farmers, asking them to construct large fish farms to prevent high lice levels. Getting this result increased understanding, but it should not be used to make prognoses. If I had constructed a model with time-fixed covariates only, I would not have seen that lag-effect of temperature is significant, and that farmers at the biggest farms wait longer before treating their fish. Thus extending the model was helpful.

An improvement of the model with time-fixed covariates would be to include seawater temperature. However this covariate appeared to be significant only when it could vary with time, which is meaningful since the difference in temperature varies a lot from one season to the next. I could have chosen to add an interaction term in the model that would capture both temperature and season of stocking. However this would again lead to a model with a time-dependent covariate. I could also have chosen to stratify temperature and include it in the baseline, but then it would not be possible to see how this covariate affects the hazard rate directly, and one would only see its effect through the baseline. Furthermore I would have been unable to include any of its values at all of the possible time points, except for at time $t = 0$. Grouping this covariate would also be a challenge.

If I constructed the model using all of the data (between years 2012 and 2015), I would have preferred to stratify geographical index since stratification of a covariate allows it to have different effects on the hazard at different time points. Since this covariate is categorical in the models already, there is no need to choose a grouping. It is known that the situation was completely different in the regions in year 2015 compared to previous years. For that reason including geographical index in the baseline would be a better choice than including it in the parametric part of the model because it would capture the difference in the hazard rate in the three regions at different time points.

6.3 Modelling choices

6.3.1 Time-fixed vs time-dependent covariates

In this study two models were developed: one with time-fixed covariates, and one with time-dependent covariates. The results from the two analyses were not that different, and a majority of the covariates appeared to be significant in both models. But what are the advantages and disadvantages to introducing time-dependency in covariates?

The main advantage with time-dependent covariates is that values are more updated and thus more precise. When the data consists of weekly reports, the values of variables change all the time for different reasons. If one then looks at treatments that happen 40 weeks after stocking, many of the covariate values may change a lot throughout the time from stocking till first treatment. Time-dependent covariates, where the values of the covariates are used at the beginning of each four-week interval, capture an updated value, and the model is then built on a more precise basis.

However there are also some disadvantages with time-dependent covariates in Cox-regression. One could imagine that introducing shorter time-intervals to capture the updated values of covariates might lead to a higher uncertainty in the estimates. However the uncertainty in the estimates did not increase much in this analysis when I introduced time-dependency in the covariates. The increase in standard error addressing to β's was inconsiderable as one can see in listings 1 and 2, and thus this disadvantage is eliminated in this case. Another disadvantage is that a model with time-dependent covariates cannot be used for prediction, again because of the updated values. When one has time-fixed covariates, all values are known at time $t = 0$, and thus the model can be used when one is in a situation where the values of all significant covariates are known today and one wants to predict the risk of first salmon lice treatment in the future. The prediction method can then be applied straight to the covariates. This cannot be done when one has a model with time-dependent covariates because the future values of the covariates are not known. However this analysis was not constructed to be used for prediction, but rather to understand what factors affect the hazard rate for first treatment, and thus this disadvantage is also irrelevant here. The last disadvantage that is worth mentioning here is that there are some technical issues with validation of a model with time-dependent covariates. However the results in the extended model were quite similar to the ones in the Cox PH model, which was well validated, and therefore I assume that the extended model is not that bad either.

6.3.2 Frailty

A possible improvement of the constructed models would be to introduce shared frailty in the covariate addressing to year. This would be a good idea especially if a new model would be constructed by using data from between years 2012 and 2015, since year 2015 was so different.

Frailty is helpful if there is more variation in the model than what can be explained. Normally in Cox models it is assumed that survival times for individuals are independent, given the covariates. By introducing frailty one can allow for dependencies between survival times in groups of individuals. "A frailty corresponds to random block effect that acts multiplicatively on the hazard rates of all subjects in a group" [Fan and Li 2002]. The Cox model can then be expressed as:

$$\alpha_f(t|\mathbf{x}_i) = Z \cdot \alpha(t|\mathbf{x}_i),$$

where Z is a random variable that cannot be observed. Individuals with higher Z will have higher probability of experiencing the event of interest earlier than what individuals with lower Z [Fan and Li 2002] will. One can distinguish between unshared and shared frailty. In shared frailty models the random effect accounts for dependencies within the groups sharing the same frailty.

In the models developed in this study, the cohorts can be grouped by years, and thus shared frailty can be used to catch an unobserved effect that is equal for all cohorts in each year-group. If one additionally divides those groups of cohorts by geographical index, and introduces shared frailty, then the problem with different situations in the regions from one year to the next might be eliminated.

6.3.3 Other options

The aim of this study was primarily to analyse time till first bath treatment, calling for a Cox analysis. The response variable was the possibly censored time from stocking till first lice bath treatment, and I investigated what factors could lead to shorter or longer time till first treatment at the farms. However there are models other than Cox that could be used to preform a similar analysis using the same or similar data from the Aquaculture Database. As one remembers from sections 2.1 and 2.2, Cox regression is one special case of the hazard function on the form $\alpha(t|\mathbf{x}_i(t)) = \alpha_0(t) r(\boldsymbol{\beta}, \mathbf{x}_i(t))$, with relative risk function set to $r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \exp(\boldsymbol{\beta}^T \mathbf{x}_i(t))$. In Aalen, Borgan, and Gjessing 2008 some other possible choices of the relative risk function are specified:

- Additive relative risk function: $r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = 1 + \boldsymbol{\beta}^T \mathbf{x}_i(t)$
- Excess relative risk function: $r(\boldsymbol{\beta}, \mathbf{x}_i(t)) = \prod_{j=1}^{p} (1 + \beta_j x_{ij}(t))$

However Cox regression is the only relative risk regression model implemented in `R`. For relative risk regression models (which can also be called proportional hazard models, when all covariates are fixed) the covariates act multiplicatively on the baseline hazard, and the effect of the covariates is assumed to be constant over time. An alternative choice of the hazard rate in a survival analysis is to use an additive non-parametric model named Aalen's additive regression model, which is on the form

$$\alpha(t|\mathbf{x}_i) = \beta_0(t) + \beta_1(t) x_{i1}(t) + ... + \beta_p(t) x_{ip}(t).$$

In a Cox model the hazard rates are always non-negative, which is a natural boundary for the hazard rates. Aalen's model does not have this restriction because the model can stray into negative values for the hazard rate, and therefore an additive model is a bit unconventional for hazard rates. However there are some advantages with Aalen's model because this model allows for a very simple estimation of the change in effects of covariates over time, which is not that simple in a Cox model. For more on this see Aalen, Borgan, and Gjessing 2008.

A completely different way of analysing these data could be to model directly the number of female lice with the same explanatory variables as in this study. A very simple approach could be to construct a generalized linear model (GLM) by using the data from one specific time point, where one assumes that the distribution for the response variable stems from the exponential class. For theory on generalized linear models see De Jong and Heller 2008. In this case one could have primarily assumed Poisson distribution. Poisson distribution is used for counting data, and the number of female lice at a farm can be viewed as such. In reality the amount of lice is reported as a continuous variable, but could in theory be modelled as a discrete variable. Farmers report a mean amount of salmon lice at the farm from a sample of fish, without reporting the sample amount. By using Poisson, one can for example assume that lice are counted from a sample of 20 fish and then model the amount of lice on the assumed sample. The Poisson distribution has one parameter, that is a rate parameter and hence has to be positive. Using a log-link one gets $\lambda = \exp(\boldsymbol{\beta}^T\mathbf{x})$ as the rate parameter for suitable covariates $\mathbf{x}$. A drawback with the Poisson distribution is that in some cases it will underestimate the number of zeros in the response variable. In this study I have experienced that the amount of female lice at the farms is often equal to zero, and therefore underestimation would be a problem when using Poisson GLM. Instead, to fix this, one could use a zero-inflated Poisson distribution assuming a separate parameter as the probability for zero. However the GLM analysis would be too simple and not informative enough since the Aquaculture Database contains data from many time points, and thus modelling amount of lice at only one of them would not explain enough. One could have used this approach to understand what affects the appearance of salmon lice at any given time point. However as one could see from the developed analysis, the situation was not the same from one year to the next, and thus the results would probably differ depending on which time point one would choose to analyse. Furthermore one would probably have faced some problems addressing to dependencies.

Another approach is to construct a temporal mixed effects model with a time-dependent component where one could model the number of female lice in cohort i at time t. Just to define a starting point for the analysis, one could try to fit one of the simpler available models for such an analysis. Let

$$\begin{aligned} Y_t &= \alpha Y_{t-1} + W_t, && \text{where } W_t \overset{iid}{\sim} N(\mu_w, \sigma_w^2) \\ Z_{t,i} &= \exp\left(\boldsymbol{\beta}^T \mathbf{x_i}(t) + \delta Y_t\right), \end{aligned}$$

where one assumes the latter to be Poisson distributed. The component Y_t is a time-series that represents the development of salmon lice over time. This part of the model represents the temporal variation and dependency of the model. For the sake of simplicity it is common to assume a normal distribution of the first observation in the time-series: $Y_1 \sim N(\mu_1, \sigma_1^2)$ [Cressie and Wikle 2011]. The variable $Z_{t,i}$ is the number of salmon lice in cohort i at time t. However a similar analysis was done in Aldrin et al. 2013, where the expected salmon lice abundance at farm i for month t was modelled. Thus it is not necessary to repeat that analysis here. The expectation of lice was modelled as a function of observed lice abundances in the previous month at the current farm and at the neighbouring farms, including other explanatory factors such as distances, seawater temperatures, etc. The response variable was constructed from the reported salmon lice abundances and was first assumed to be zero-inflated Poisson distributed since there was an excess frequency of zeroes. However over-dispersion was found, and therefore a zero-inflated negative binomial distribution was chosen. This is because the negative binomial distribution is considered to be a better fit for modelling over-dispersed Poisson counts. In Aldrin et al. 2013 many of the same covariates where chosen to be significant as in the Cox analysis in this thesis, and to some extent this validates the results found in this analysis.

One could also have performed a spatial analysis on the data since the Aquaculture Register contains geographical coordinates for every salmon farm in Norway. One could furthermore have analysed whether treatments appear in clusters since an indication of organised treatments could be seen in the data. Piecing this together requires a lot of work and is viewed to be too comprehensive for this particular thesis.

6.4 Further work

An extension of this thesis would be to analyse time between first and second treatment by using the same survival analysis approach as in this thesis. This would improve the understanding of the situations at the farms even more, and show an insight in further development of lice levels at the farms. This approach would probably show that the covariate addressing to amount of fish in the cohort affects the hazard rate for second treatment oppositely to how it affects the hazard rate for first treatment. Furthermore one could have used amount of lice at the farm and IIP, obtained at time $t = 0$ when analysing time between first and second treatment. This is because amount of lice and IIP after a treatment are not necessarily reduced to a number of zeros. Analysing the above would also indicate on any resistance problems.

Another possible extension would be to analyse the amount of total treatments of the cohorts during the period when they are located at the farms. This could be done by using a GLM approach, and by starting the analysis by assuming Poisson distribution of the explanatory variable. However with this approach one would encounter some complications addressing to the explanatory variables. If the covariate values would be obtained at the start of the study, and if the study would last between 18 and 24 months, the covariates would change their values a lot during that time. This would call for more advanced modelling.

6.5 Conclusion

Using a survival analysis approach, and more precisely Cox analysis, allowed me to study the time-changing hazard rates related to salmon lice treatment. I was able to determine what factors are associated with shorter or longer time till the first bath treatment applied on cohorts of farmed salmon. There are many advantages of performing multiple Cox regression, which has been mentioned throughout the thesis. Here is a summary of them:

- No assumptions about the shape of the distribution of the response variable (time related to treatment) is needed
- It would work for both discrete-time and continuous-time data
- Right-censoring is easily handled
- The explanatory variables can be stratified
- Both time-fixed and time-dependent covariates can be used
- Can be validated in different ways
- Can be extended to non-proportional hazards

The results from the analysis indicated that lice levels at neighboring farms play a major role in the timing of the first treatment of a farmed salmon cohort after stocking. The variables accounting for the amount of female lice at the nearest neighbor, infection pressure from neighbors within a 100 km seaway distance, and distances to both the nearest neighbor and the nearest slaughterhouse are of primary importance in the model. This applies both when covariates are time-fixed and when they are allowed to vary with time. The significant variables for time till first treatment unrelated to neighbors are seawater temperature and amount of fish at the farm of interest. However both variables appeared as significant only when they were allowed to vary each month. In the model without time-varying covariates, smolt weight in the cohort appeared as a significant variable, but was not significant in the extended model. Furthermore the results showed that the time till first treatment is different in different parts of Norway, and vary from one year to the next.

Both of the constructed models gave similar results, and a validation indicated that the Cox PH model is proper for the dataset. However one could see that the situation was different in year 2015, which resulted in poor discrimination in the validation dataset.

Ideally I would conclude and suggest that salmon farmers place their farms as far away from other farms as possible if they want to postpone time till first treatment. If one discovers a fjord where no other farms are located, one should place a salmon farm there. However in real life this is easier said than done: farmers are not able to decide such things themselves because fish farm locations are decided by the government. The fish farming industry is now developing fish tanks in the open sea lying far away from the coast of Norway to prevent farms from being placed close to each other. Based on the analyses in this thesis, this is a smart thing to do.

Bibliography

Aalen, O.O., Borgan, Ø., and Gjessing, H.K. (2008). *Survival and event history analysis: a process point of view*. Springer Science, Business Media.

Aldrin, M., Storvik, B., Kristoffersen, A.B., and Jansen, P.A. (2013). "Space-time modelling of the spread of salmon lice between and within Norwegian marine salmon farms". In: *PloS one* 8.5, e64039.

Bellera, C.A., MacGrogan, G., Debled, M., Lara, C.T. de, Brouste, V., and Mathoulin-Pélissier, S. (2010). "Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer". In: *BMC medical research methodology* 10.20, pp. 1–12.

Bernhoft, A.C. and Fardal, A. (2007). "IFRS og fiskeoppdrett". In: *Magma – Tidsskrift for økonomi og ledelse*, pp. 49–58.

Cressie, N.A. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

De Jong, P. and Heller, G.Z. (2008). *Generalized linear models for insurance data*. Cambridge University Press.

Fan, J. and Li, R. (2002). "Variable selection for Cox's proportional hazards model and frailty model". In: *The annals of Statistics* 30.1, pp. 74–99.

Fisher, L.D. and Lin, D.Y. (1999). "Time-dependent covariates in the Cox proportional-hazards regression model". In: *Annual review of public health* 20.1, pp. 145–157.

Fisheries.no (2014). *Aquaculture*. URL: `http://www.fisheries.no/aquaculture/Aquaculture/#.VrBv-PkrLIW` (visited on 03/19/2015).

Forskning.no (2005). *Hva er lakselus?* URL: `http://forskning.no/fisk-fiskehelse-fiskesykdommer-oppdrett-miljovern/2008/02/hva-er-lakselus` (visited on 03/19/2015).

Forskrift om bekjempelse av lakselus i akvakulturanlegg (2013). Forskrift 5. desember 2012, nr. 1140. §7, §8, §10.

Fox, J. and Weisberg, S. (2011). “Cox proportional-hazards regression for survival data in R”. In: *An R Companion to Applied Regression, Second Edition*, pp. 1–20.

Helland, I.P., Finstad, B., Uglem, I., Diserud, O.H., Foldvik, A., Hanssen, F., Bjørn, P.A., Nilsen, R., and Jansen, P.A. (2012). “Hva avgjør lakselusinfeksjon hos vill laksefisk?” In: *Statistisk bearbeiding av data fra nasjonal lakselusovervåking, 2004-2010* 891, pp. 1–51.

Heuch, P.A., Nordhagen, J.R., and Schram, T.A. (2000). “Egg production in the salmon louse [Lepeophtheirus salmonis (Krøyer)] in relation to origin and water temperature”. In: *Aquaculture Research* 31.11, pp. 805–814.

Jansen, P.A., Kristoffersen, A.B., Viljugrein, H., Jimenez, D., Aldrin, M., and Stien, A. (2012). “Sea lice as a density-dependent constraint to salmonid farming”. In: *Proceedings of the Royal Society of London B: Biological Sciences* 279.1737, pp. 2330–2338.

Jansen, P.A., Grøntvedt, R.N., Tarpai, A., Helgesen, K.O., and Horsberg, T.E. (2016). “Surveillance of the Sensitivity towards Antiparasitic Bath-Treatments in the Salmon Louse (Lepeophtheirus salmonis)”. In: *PloS one* 11.2, e0149006.

Kalbfleisch, John D and Prentice, Ross L (2002). *The statistical analysis of failure time data*. Vol. 360. John Wiley & Sons.

Kaur, Kiranpreet, Jansen, Peder Andreas, Aspehaug, Vidar Teis, and Horsberg, Tor Einar (2016). “Phe362Tyr in AChE: A Major Factor Responsible for Azamethiphos Resistance in Lepeophtheirus salmonis in Norway”. In: *PloS one*, e0149264.

Kristoffersen, A.B., Jimenez, D., Viljugrein, H., Grøntvedt, R., Stien, A., and Jansen, P.A. (2014). “Large scale modelling of salmon lice (Lepeophtheirus salmonis) infection pressure based on lice monitoring data from Norwegian salmonid farms”. In: *Epidemics* 9, pp. 31–39.

Lusedata (2012). *Forklaring statistikk behandling*. URL: `http://lusedata.no/statistikk/forklaring/behandling/` (visited on 04/09/2015).

Royston, P. and Altman, D.G. (2013). “External validation of a Cox prognostic model: principles and methods”. In: *BMC medical research methodology* 13.33, pp. 1–15.

Royston, P. and Sauerbrei, W. (2004). “A new measure of prognostic separation in survival data”. In: *Statistics in medicine* 23.5, pp. 723–748.

Stien, A., Bjørn, P.A., Heuch, P.A., and Elston, D.A. (2005). “Population dynamics of salmon lice Lepeophtheirus salmonis on Atlantic salmon and sea trout”. In: *Marine Ecology Progress Series* 290, pp. 263–275.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009). *Mixed effects models and extensions in ecology with R*. Springer Science & Business Media.

Appendix A

Useful terms

Cohort The salmon population in a given farm, usually stocked in the same time-period and of the same age

Farm A given geographical locality, with an amount of net-pens containing a number of fish

Harvesting Removing fish from the farms after completed production

Salmon In this study Atlantic salmon (*Salmo salar*) and rainbow trout (*(Onchorhynchus mykiss*) are termed as salmon. Atlantic salmon represent more than 90% of the fish

Smolt Juvenile fish being stocked to seawater from their nursery farms in fresh water

Stocking Releasing the smolts into sea water

Treatment Bath treatment, as described in 1.1.4 if nothing else is specified

Lice stages:

Larval stages - the first two nauplii stages, which are non-parasitic and planktonic (freely drifting in the water current)

Copepodid stage - the infectious stage

Chalimus stages - two sessile stages attached to the fish, both being parasitic

Pre-adult - Two stages, both being mobile (able to move about on the fish) and parasitic (also able to harm the fish when in large numbers)

Adult (female and male) - fully reproductive stages, where the females produce eggstrings

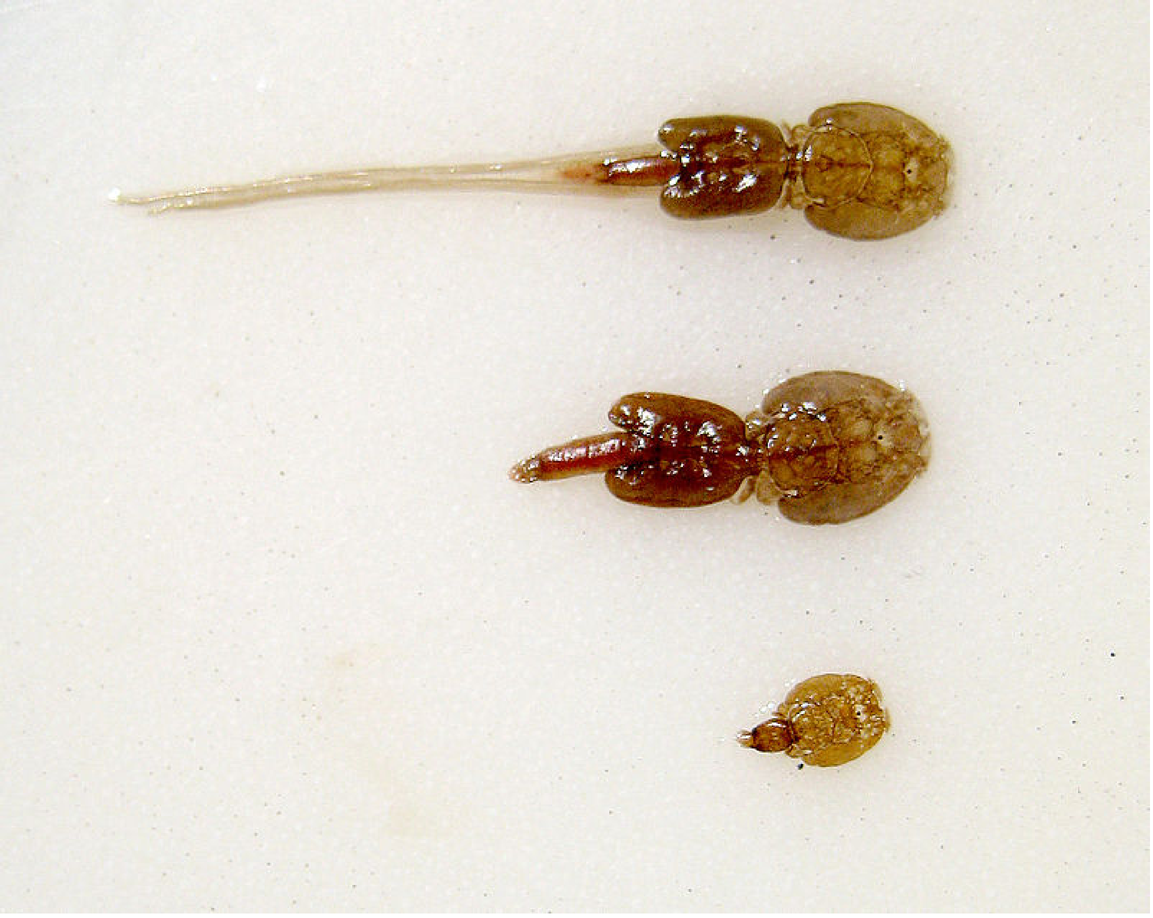


Figure A.1: From top: 1. Mature female with egg strings. 2. Mature female without eggstings. 3. Immature louse. Picture taken by Thomas Bjørkan at Norwegian Aquaculture Center, Brønnøy, Norway, 02.06.09.

Appendix B

Computer code and output

B.1 R-outputs

Listing 1: R-output of Cox regression with time-fixed covariates (and stratified geographical index)

```
###########----------- TIME-FIXED COVARIATES -----------###########
Call:
coxph(formula = Surv(time, status) ~ 1 + EIP_tr + strata(geoindex_gr) +
    dist_slakt_tr + dist_nabo2 + vekt + year + lusnabo + EIP_tr:lusnabo,
    data = tidsint2)

  n= 648, number of events= 212

                    coef exp(coef) se(coef)      z Pr(>|z|)
EIP_tr           0.95975   2.61105  0.12378  7.753 8.99e-15 ***
dist_slakt_tr    1.70867   5.52159  0.81394  2.099  0.03579 *
dist_nabo2      -0.06100   0.94082  0.01880 -3.245  0.00117 **
vekt             0.04517   1.04621  0.01765  2.559  0.01050 *
year2012        -0.52824   0.58964  0.16310 -3.239  0.00120 **
year2014        -0.46970   0.62519  0.19524 -2.406  0.01614 *
lusnabo2         0.51026   1.66573  0.18806  2.713  0.00666 **
EIP_tr:lusnabo2 -0.42779   0.65195  0.13856 -3.087  0.00202 **
---
                exp(coef) exp(-coef) lower .95 upper .95
EIP_tr             2.6110     0.3830    2.0486    3.3280
dist_slakt_tr      5.5216     0.1811    1.1201   27.2201
dist_nabo2         0.9408     1.0629    0.9068    0.9761
vekt               1.0462     0.9558    1.0106    1.0830
year2012           0.5896     1.6959    0.4283    0.8117
year2014           0.6252     1.5995    0.4264    0.9166
lusnabo2           1.6657     0.6003    1.1522    2.4081
EIP_tr:lusnabo2    0.6519     1.5339    0.4969    0.8554

Concordance= 0.698  (se = 0.037 )
Rsquare= 0.138   (max possible= 0.954 )
Likelihood ratio test= 96.27  on 8 df,   p=0
Wald test            = 91.1  on 8 df,   p=3.331e-16
Score (logrank) test = 100.8  on 8 df,   p=0   p=0
```

Listing 2: R-output of Cox regression with time-dependent covariates

```
######################################################################
###------------------ TIME-DEPENDENT COVARIATES -------------------###
######################################################################
Call:
coxph(formula = Surv(time0, time, status) ~ 1 + t_lusnabo + t_EIP_tr +
    t_antall + t_temp_lag + dist_slakt_tr + year + geoindex_gr +
    dist_nabo, data = tidsavh_ny)

  n= 4652, number of events= 212

                  coef exp(coef) se(coef)      z Pr(>|z|)
t_lusnabo2     1.61820   5.04401  0.19514  8.292  < 2e-16 ***
t_EIP_tr      -1.79808   0.16562  0.32925 -5.461 4.73e-08 ***
t_antall      -0.05905   0.94266  0.02013 -2.933  0.00335 **
t_temp_lag     0.09915   1.10423  0.03111  3.187  0.00144 **
dist_slakt_tr  1.80955   6.10772  0.85751  2.110  0.03484 *
year2012      -0.43703   0.64595  0.16490 -2.650  0.00804 **
year2014      -0.45465   0.63467  0.20469 -2.221  0.02634 *
geoindex_gr2  -0.77421   0.46107  0.23487 -3.296  0.00098 ***
geoindex_gr1  -0.59179   0.55334  0.27002 -2.192  0.02841 *
dist_nabo     -0.04355   0.95739  0.01856 -2.347  0.01893 *
---
              exp(coef) exp(-coef) lower .95 upper .95
t_lusnabo2       5.0440     0.1983   3.44087    7.3941
t_EIP_tr         0.1656     6.0380   0.08687    0.3158
t_antall         0.9427     1.0608   0.90619    0.9806
t_temp_lag       1.1042     0.9056   1.03892    1.1737
dist_slakt_tr    6.1077     0.1637   1.13753   32.7942
year2012         0.6460     1.5481   0.46756    0.8924
year2014         0.6347     1.5756   0.42493    0.9479
geoindex_gr2     0.4611     2.1689   0.29097    0.7306
geoindex_gr1     0.5533     1.8072   0.32595    0.9394
dist_nabo        0.9574     1.0445   0.92320    0.9928

Concordance= 0.787  (se = 0.022 )
Rsquare= 0.045   (max possible= 0.409 )
Likelihood ratio test= 216  on 10 df,   p=0
Wald test            = 173  on 10 df,   p=0
Score (logrank) test = 205.7  on 10 df,   p=0
```

.

B.2 R-code

I choose to present only a part of the computer code used to construct the models in this thesis. This is because of similarities between the R-procedures that are used, and reuse of code.

R-code related to the data:

```
#---> Correlatioan and collinearity check
source("HighstatLibV6.R")

Z = cbind(tidsint2$time, tidsint2$EIP, tidsint2$geoindex_gr, tidsint2$dist_slakt,
    tidsint2$dist_nabo, tidsint2$vekt, tidsint2$year, tidsint2$utsett,
    tidsint2$antall, tidsint2$lusnabo, tidsint2$temp, tidsint2$biomass_nabo)

colnames(Z) = c("time", "EIP", "geoindex_gr", "dist_slakt", "dist_nabo",
    "vekt", "year", "utsett", "antall", "lusnabo", "temp", "biomass_nabo")

cor(Z,Z)
corvif(Z[,-1])


#---> Null residual plots

library(survival)
fit.pr = coxph(Surv(time0, time,status)~1, data=tidsavh_ny)
rr = resid(fit.pr)

a = cbind(tidsint2$dist_slakt, tidsint2$dist_nabo, tidsint2$temp,
    tidsint2$biomass_nabo, tidsint2$EIP, tidsint2$antall,
    tidsint2$vekt, tidsint2$geoindex)

navn = c("Distance,slaughterh.","Distance,neighbor","Temperature",
    "Biomass neighbor","EIP","Number of fish","Weight","Geo.index")

par(mfrow=c(2,4))
for (i in 1:ncol(a)){
  plot(a[,i], rr,main=navn[i],xlab="",
    ylab="Null residuals",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
  lines(lowess(a[,i], rr, iter=0))
}

# Null residual plot of log-transformed EIP:
E = tidsint2$EIP +1
plot(E,rr,log="x",main="EIP",xlab="",ylab="Null residuals")
lines(lowess(E, rr, iter=0))
```

R-code related to the Cox PH model:

```
#---> Call of a univariate model and log-linearity check:

cox.psEIP=coxph(Surv(time,status)~1+pspline(EIP),data=tidsint2)
print(cox.psEIP);termplot(cox.psEIP,se=T)

cox.EIP=coxph(Surv(time,status)~1+EIP,data=tidsint2)
summary(cox.EIP); AIC(cox.EIP)
```

```
#---> Grouping and Nelson-Aalen plot construction:

tidsint2$geoindex_gr=cut(tidsint2$geoindex, breaks=c(-1,4.8,14.4,30),labels =1:3)
survdiff(Surv(time,status)~geoindex_gr,data=tidsint2)

na.geoindex=survfit(coxph(Surv(time,status)~strata(geoindex_gr),data=tidsint2))
plot(na.geoindex,fun="cumhaz",mark.time=F,main="Nelson Aalen geoindex",
    xlab="Weeks from stocking",ylab="Cumulative hazard for treatment",col=1:3)
legend("topleft",c("south","mid","north"),col=1:3,lty=1,cex=1.5)


#---> Checking for interactions:

cox.alt_s = coxph(Surv(time,status)~1+EIP_tr+geoindex_gr+dist_slakt_tr+
    dist_nabo2+vekt+year+antall+lusnabo+biomass_nabo_gr+
    temp_gr,data=tidsint2)
summary(cox.alt_s)

a = c("EIP_tr", "geoindex_gr", "dist_slakt_tr", "dist_nabo2", "vekt",
    "year", "antall", "lusnabo", "biomass_nabo_gr", "temp_gr")

AIC.v = c(); cov1 = c(); cov2 = c()

for (i in 1:10){
  for (j in 1:10){
    if (j > i){
      cox.inter=tryCatch(coxph(Surv(time,status)~1+EIP_tr+geoindex_gr+
    dist_slakt_tr+dist_nabo2+vekt+year+antall+lusnabo+biomass_nabo_gr+
    temp_gr+
      eval(parse(text=a[i])):eval(parse(text=a[j])),data=tidsint2))
    AIC.v = c(AIC.v, tryCatch(AIC(cox.inter)))
    cov1 = c(cov1, a[i])
    cov2 = c(cov2, a[j])
}}}

inter = data.frame(cov1, cov2, AIC.v)
inter2 = inter[which(AIC.v<AIC(cox.alt_s)),]
inter2[order(inter2[,3]),]


#---> Checking for proportional hazard assumption:

cox.endelig2 = coxph(Surv(time,status)~1+EIP_tr+geoindex_gr+dist_slakt_tr+
    dist_nabo2+vekt+year+lusnabo+EIP_tr:lusnabo,data=tidsint2)

cox.zph(cox.endelig2, transform="identity")
plot(cox.zph(cox.endelig2, transform="identity")[2], main="Geo.index group 3")
abline(h=0)

#---> Call of the final Cox PH model:

cox.endelig3 = coxph(Surv(time,status)~1+EIP_tr+strata(geoindex_gr)+
    dist_slakt_tr+dist_nabo2+vekt+year+lusnabo+
    EIP_tr:lusnabo,data=tidsint2)
summary(cox.endelig3); AIC(cox.endelig3)


#---> Survival curves:

# For illustration I plotted the estimated survival function for the following
# combinations of the covariates:
# (keeping year constant at 2013 and distance to slaughterhouse at 15km)

#    1) EIP=1, lice_neighb=0, weight smolt=100, dist_neighb=1
#    2) EIP=5, lice_neighb=1, weight smolt=100, dist_neighb=1
#    3) EIP=1, lice_neighb=0, weight smolt=200, dist_neighb=1
#    4) EIP=5, lice_neighb=1, weight smolt=200, dist_neighb=1
#    5) EIP=1, lice_neighb=0, weight smolt=100, dist_neighb=4
#    6) EIP=5, lice_neighb=1, weight smolt=100, dist_neighb=4
#    7) EIP=1, lice_neighb=0, weight smolt=200, dist_neighb=4
#    8) EIP=5, lice_neighb=1, weight smolt=200, dist_neighb=4

par(mfrow=c(2,2))

new.covariates=data.frame(EIP_tr=rep(c(0.6931472,1.791759),4),year=rep("2013",8),
    dist_slakt_tr=rep(0.06667,8),
    lusnabo=rep(c("1","2"),4),
    dist_nabo2=c(rep(2,4),rep(6,4)),vekt=rep(c(10,10,20,20),2),
    geoindex_gr=rep("1",8))
surv.final=survfit(cox.endelig3,newdata=new.covariates)
plot(surv.final,mark.time=F, xlab="Weeks after stocking", ylab="No treatment",col
    =1:8,lty=1,
    main="South-region",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
legend("bottomleft",c("1","2","3","4","5","6","7","8"),col=1:8,lty=1)
abline(v=20, lty=3)

new.covariates=data.frame(EIP_tr=rep(c(0.6931472,1.791759),4),year=rep("2013",8),
    dist_slakt_tr=rep(0.06667,8),
    lusnabo=rep(c("1","2"),4),
    dist_nabo2=c(rep(2,4),rep(6,4)),vekt=rep(c(10,10,20,20),2),
    geoindex_gr=rep("2",8))
surv.final=survfit(cox.endelig3,newdata=new.covariates)
plot(surv.final,mark.time=F, xlab="Weeks after stocking", ylab="No treatment",col
    =1:8,lty=1,
    main="Mid-region",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
legend("bottomleft",c("1","2","3","4","5","6","7","8"),col=1:8,lty=1)
abline(v=20, lty=3)

new.covariates=data.frame(EIP_tr=rep(c(0.6931472,1.791759),4),year=rep("2013",8),
    dist_slakt_tr=rep(0.06667,8),
    lusnabo=rep(c("1","2"),4),
    dist_nabo2=c(rep(2,4),rep(6,4)),vekt=rep(c(10,10,20,20),2),
    geoindex_gr=rep("3",8))
surv.final=survfit(cox.endelig3,newdata=new.covariates)
plot(surv.final,mark.time=F, xlab="Weeks after stocking", ylab="No treatment",col
    =1:8,lty=1,
    main="North-region",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
legend("bottomleft",c("1","2","3","4","5","6","7","8"),col=1:8,lty=1)
abline(v=20, lty=3)

# And an additional survival plot, when only geographical index and amount of lice
    at the nearest farm differ:
# (EIP = 0.5, dist_neighb=5km, weight_smolt=150gr)

new.covariates=data.frame(EIP_tr=rep(0.09531018,6),year=rep("2013",6),vekt=rep(15,6)
    ,
    dist_nabo2=rep(5,6),dist_slakt_tr=rep(0.06666667,6),
    lusnabo=rep(c("1","2"),3),geoindex_gr=c("1","1","2","2","3","3"))
surv.final=survfit(cox.endelig3,newdata=new.covariates)
plot(surv.final,mark.time=F, xlab="Weeks after stocking", ylab="No treatment",col
    =1:6,lty=1,
    main="Only lice at neighb. and geo.index differ",cex.axis=1.5,cex.main=1.5,cex.
        lab=1.5)
legend("bottomleft",c("lice=1, south","lice=2, south","lice=1, mid",
    "lice=2, mid","lice=1, north","lice=2, north"),col=1:6,lty=1)
```

R-code related to the extended Cox model:

```
#---> Extending the dataset with time-fixed covariates, by splitting the intervals:

#Starting by only collecting the significant time-fixed covariates from the previous
     dataset:
tidsavh = tidsint2[,c(1:20,22,24,28)]

#Making an indicator for each cohort:
for (i in 1:nrow(tidsavh)){
  tidsavh$id[i] = i
}

#Constructing the new dataset:
cut.points = seq(4, 44, by=4)
tidsavh_ny = survSplit(data = tidsavh, cut = cut.points, end = "time",start = "time0
    ", event = "status")
tidsavh_ny = tidsavh_ny[order(tidsavh_ny$id), ]
tidsavh_ny = tidsavh_ny[,c("id","lokalitet","mnd_utsatt","uke_utsatt","mnd_slaktet",
    "uke_slaktet", "mnd_1.beh", "uke_1.beh", "dist_slakt", "dist_nabo", "nabo_
    lokalitet","temp_utsett", "biomass_nabo", "EIP", "lusnabo", "antall", "vekt", "
    utsett", "year","geoindex", "geoindex_gr", "geoindex_gr_ny", "time0", "time", "
    status")]


tidsavh_ny$obs_start = NA
tidsavh_ny$obs_slutt = NA
for (i in 1:nrow(tidsavh_ny)){
  t = which(colnames(Behandling)==tidsavh_ny$uke_utsatt[i])
  tidsavh_ny$obs_start[i] = colnames(Behandling)[t+tidsavh_ny$time0[i]]
  if (tidsavh_ny$status[i]==1){
    tidsavh_ny$obs_slutt[i] = colnames(Behandling)[t+tidsavh_ny$time[i]] }
  else {
    tidsavh_ny$obs_slutt[i] = colnames(Behandling)[t+tidsavh_ny$time[i]-1] }
}
tidsavh_ny[,5:8] = list(NULL)


#----> Lag-effect of the sea water temperature:

tidsavh_ny$t_temp_lag = NA
for (i in 1:nrow(tidsavh_ny)){
  loci = as.character(tidsavh_ny$lokalitet[i])
  t = which(colnames(temp)==tidsavh_ny$obs_start[i])
  tidsavh_ny$t_temp_lag[i] = as.numeric(temp[loci,t-4])
}


#---> Call of the final model:
cox.endelig_T = coxph(Surv(time0,time,status)~1+t_lusnabo+t_EIP_tr+t_antall+t_temp_
    lag+dist_slakt_tr+year+geoindex_gr+dist_nabo,data=tidsavh_ny)
summary(cox.endelig_T); AIC(cox.endelig_T)


#---> Baseline curves:
par(mfrow=c(1,3))

baseline=basehaz(cox.endelig3)
plot(baseline$time[which(baseline$strata==1)],baseline$hazard[which(baseline$strata
    ==1)],type="l",main="Cox PH model, geo.index stratified",xlab="Time",ylab="
    Hazard",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
lines(baseline$time[which(baseline$strata==2)],baseline$hazard[which(baseline$strata
    ==2)],col=2)
lines(baseline$time[which(baseline$strata==3)],baseline$hazard[which(baseline$strata
    ==3)],col=3)
legend("topleft",c("south-region", "mid-region","north-region"),col=1:3,lty=1,cex
    =1.5)
```

```
baseline2=basehaz(cox.endelig_T)
plot(baseline2$time, baseline2$hazard,col=3,cex.axis=1.5,cex.main=1.5,cex.lab=1.5,
    main="Extended Cox model, geo.index in parametric part",xlab="Time",ylab="Hazard
    ")
points(baseline2$time, baseline2$hazard*exp(-0.7743),col=2)
points(baseline2$time, baseline2$hazard*exp(-0.5923),col=1)

cox.endelig_T2 = coxph(Surv(time0,time,status)~1+t_lusnabo+t_EIP_tr+t_antall+t_temp_
    tr+dist_slakt_tr+year+strata(geoindex_gr)+dist_nabo,data=tidsavh_ny)
baseline3=basehaz(cox.endelig_T2)
plot(baseline3$time[which(baseline3$strata==1)], baseline3$hazard[which(baseline3$
    strata==1)],type="l",main="Extended Cox model, geo.index stratifed",xlab="Time",
    ylab="Hazard",lty=2,cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
lines(baseline3$time[which(baseline3$strata==2)], baseline3$hazard[which(baseline3$
    strata==2)],col=2,lty=2)
lines(baseline3$time[which(baseline3$strata==3)], baseline3$hazard[which(baseline3$
    strata==3)],col=3,lty=2)
```

R-code related to the model validation:

```
#----> Calculating prognostic index:
### Validation dataset
tid$PI = NA
for (i in 1:nrow(tid)){
  tid$PI[i] =  -0.3072684 + 0.86562*tid$EIP_tr[i] + 1.85330*tid$dist_slakt_tr[i] -
      0.06009*tid$dist_nabo[i] + 0.03987*tid$vekt[i] + 0.45085*(as.numeric(tid$
      lusnabo[i])-1) - 0.39263*tid$EIP_tr[i]*(as.numeric(tid$lusnabo[i])-1)
}
### Original dataset
tidsint2$PI = NA
for (i in 1:nrow(tidsint2)){
  coeff = 0.86562*tidsint2$EIP_tr[i] + 1.85330*tidsint2$dist_slakt_tr[i] - 0.06009*
      tidsint2$dist_nabo2[i] + 0.03987*tidsint2$vekt[i] + 0.45085*(as.numeric(
      tidsint2$lusnabo[i])-1) - 0.39263*tidsint2$EIP_tr[i]*(as.numeric(tidsint2$
      lusnabo[i])-1)
  if (tidsint2$year[i]==2013){tidsint2$PI[i] = coeff}
  if (tidsint2$year[i]==2012){tidsint2$PI[i] = coeff -0.52823970}
  if (tidsint2$year[i]==2014){tidsint2$PI[i] = coeff -0.46970021}
}


#---> Histogram of PI's, centered on the mean in the deviation dataset:

par(mfrow=c(2,1))
hist(tidsint2$PI-0.7080577, ylim=c(0,80),xlim=c(-3.8,3.2),main="Derivation data",
    xlab="PI",ylab="Frequency",breaks =40)
abline(v=(quantile(tidsint2$PI,probs = c(0.16, 0.5, 0.84))-0.7080577),col=2)
hist(tid$PI-0.7080577, ylim=c(0,80),xlim=c(-3.8,3.2), main="Validation data",xlab="
    PI",ylab="Frequency",breaks =20)
abline(v=(quantile(tid$PI,probs = c(0.16, 0.5, 0.84))-0.7080577),col=2)

mean(tidsint2$PI-0.7080577);sd(tidsint2$PI-0.7080577)
mean(tid$PI-0.7080577);sd(tid$PI-0.7080577)


#----> Regression on the PI separated for regions:
cox.val_fix1 = coxph(Surv(time,status)~PI, data=tid[which(tid$geoindex==1),]);
    summary(cox.val_fix1)
cox.val_fix2 = coxph(Surv(time,status)~PI, data=tid[which(tid$geoindex==2),]);
    summary(cox.val_fix2)
```

```
cox.val_fix3 = coxph(Surv(time,status)~PI, data=tid[which(tid$geoindex==3),]);
    summary(cox.val_fix3)


#----> Removing cohorts betwee 62 and 66 degrees north:
tid_ny = tid

for (i in nrow(tid_ny):1){
  if ((tid_ny$north[i] > 62) & (tid_ny$north[i]<66)){
    tid_ny = tid_ny[-i,]
}}


#----> Regression on the PI:
cox.val_fix = coxph(Surv(time,status)~PI+strata(geoindex_gr), data=tid_ny)
summary(cox.val_fix)

# Testing whether PI is significantly different from 1:
z = (0.6777-1)/0.2635
1 - (pnorm(-z) - pnorm(z))


#----> Checking model fit:
cox.val_fix_all = coxph(Surv(time,status)~EIP_tr+strata(geoindex_gr)+dist_slakt_tr+
    dist_nabo+vekt+lusnabo+EIP_tr:lusnabo+offset(PI), data=tid_ny)
summary(cox.val_fix_all)$coef[,c(1,3,5)]


#----> Measures of discrimination

# Concordance index:

d = data.frame("time"=tidsint2$time,"PI"=tidsint2$PI,"status"=tidsint2$status)
d = d[order(tidsint2$time), ]; d = d[which(d$status==1),]
v = data.frame("time"=tid_ny$time,"PI"=tid_ny$PI,"status"=tid_ny$status)
v = v[order(tid_ny$time), ]; v = v[which(v$status==1),]

s_d = 0
n_d = 0
for(i in 1:nrow(d)){
  for (j in 1:nrow(d)){
    if (i>j){
      if (d$time[i] > d$time[j]){
        n_d = n_d + 1
        s_d = s_d + (d$PI[i] < d$PI[j]) + 0.5 * (d$PI[i] == d$PI[j])
      }
}}}
c.index_d = s_d/(n_d)

s_v = 0
n_v = 0
for(i in 1:nrow(v)){
  for (j in 1:nrow(v)){
    if (i>j){
      if (v$time[i] > v$time[j]){
        n_v = n_v + 1
        s_v = s_v + (v$PI[i] < v$PI[j]) + 0.5 * (v$PI[i] == v$PI[j])
      }
}}}
c.index_v = s_v/(n_v)


c.index_d;c.index_v


# T-test:
```

```
d = data.frame("time"=tidsint2$time,"PI"=tidsint2$PI,"status"=tidsint2$status)
d0 = d[which(d$status==0),];d1 = d[which(d$status==1),];

v = data.frame("time"=tid_ny$time,"PI"=tid_ny$PI,"status"=tid_ny$status)
v0 = v[which(v$status==0),];v1 = v[which(v$status==1),];

t.test(d0$PI,d1$PI)
t.test(v0$PI,v1$PI)


# D-statistics
t_v = data.frame(time = tid_ny$time,status = tid_ny$status,PI=tid_ny$PI,GI=tid_ny$
    geoindex_gr)
t_v = t_v[order(t_v$PI), ]

rankit = qqnorm(t_v$PI)$x

# Can also use the following formula instead of the command "rankit":
# z=c(); for (i in 1:nrow(tid_ny)){z = c(z, qnorm((i-0.5)/(nrow(tid_ny)+1-2*0.5)))}


kappa = sqrt(8/pi)
t_v$rankit = rankit/kappa
D_v = coxph(Surv(time,status)˜rankit+strata(GI), data=t_v);summary(D_v)
D = summary(D_v)$coef[,1]
sigma2 = piˆ2/6
R_v = (Dˆ2/kappaˆ2)/(sigma2 + Dˆ2/kappaˆ2)

t_d = data.frame(time=tidsint2$time,status=tidsint2$status,PI=tidsint2$PI,GI=
    tidsint2$geoindex_gr)
t_d = t_d[order(tidsint2$PI), ]
rankit = qqnorm(t_d$PI)$x
t_d$rankit = rankit/kappa
D_d = coxph(Surv(time,status)˜rankit+strata(GI), data=t_d);summary(D_d)
D = summary(D_d)$coef[,1]
R_d = (Dˆ2/kappaˆ2)/(sigma2 + Dˆ2/kappaˆ2)

R_d; R_v


#----> Survival curves (reporting only for south-region):

quantile(tidsint2$PI,probs = 0.5)
tidsint2$PI_gr=cut(tidsint2$PI, breaks=c(-3,0.56,4),labels =1:2)
tid_ny$PI_gr=cut(tid_ny$PI, breaks=c(-3,0.56,4),labels =1:2)

km.PI3=survfit(Surv(time,status)˜PI_gr, data=tidsint2[which(tidsint2$geoindex_gr==1)
    ,])
plot(km.PI3, mark.time=F, xlab="Weeks after stocking",ylab="Probability of no
    treatment",main="Derivation dataset - south", col=1,lty=1:2)
legend("bottomleft", c("low (29)","high (95)"), col=1,lty=1:2)
km.PI4=survfit(Surv(time,status)˜PI_gr, data=tid_ny[which(tid_ny$geoindex_gr==1),])
plot(km.PI4, mark.time=F, xlab="Weeks after stocking",ylab="Probability of no
    treatment",main="Validation dataset - south", col=2,lty=1:2)
legend("bottomleft", c("low (2)","high (22)"), col=2,lty=1:2)


#----> Hazard ratios (reporting only for south-region):
cox.hr2 = coxph(Surv(time,status)˜PI_gr+strata(geoindex_gr), data=tidsint2[which(
    tidsint2$geoindex_gr==1),])
summary(cox.hr2)$coeff[,1:2]
cox.hr1 = coxph(Surv(time,status)˜PI_gr+strata(geoindex_gr), data=tid_ny[which(tid_
    ny$geoindex_gr==1),])
summary(cox.hr1)$coeff[,1:2]
```