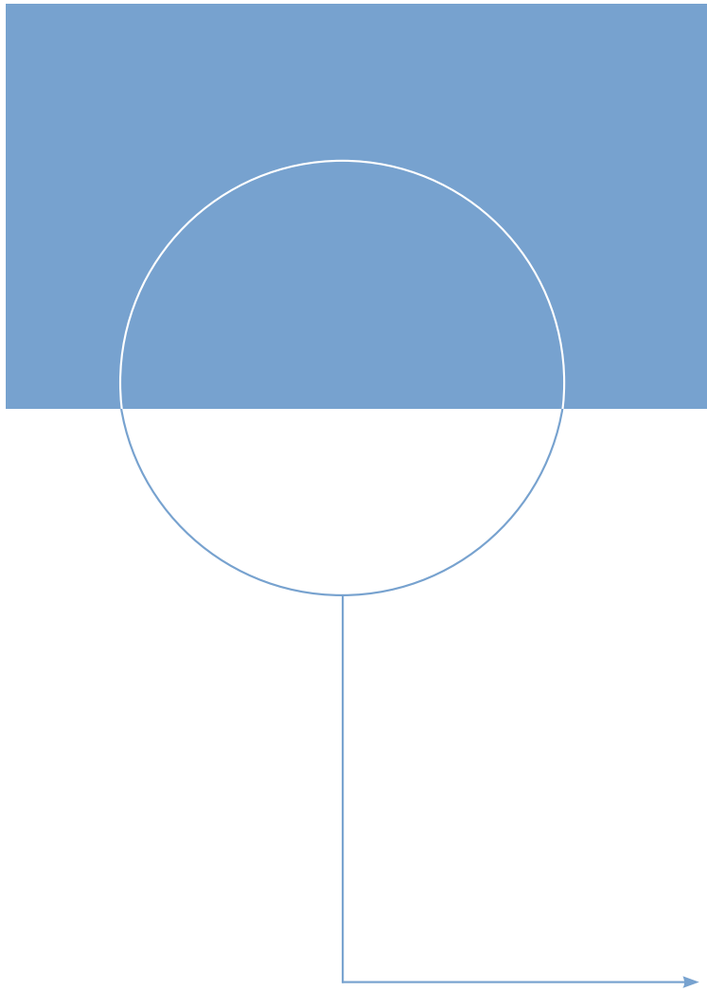

Master's thesis

2021

Camilla Karlsen

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

Camilla Karlsen

Investigation of shortened sea phase effect on salmon lice

June 2021

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of
Science and Technology

Investigation of shortened sea phase effect on salmon lice

Camilla Karlsen

Applied Physics and Mathematics
Submission date: June 2021
Supervisor: John Sølve Tyssedal
Co-supervisor: Anna Solvang Båtnes

Norwegian University of Science and Technology
Department of Mathematical Sciences

Preface

This master thesis was completed during the spring 2021, and marks the end of my studies at Norwegian University of Science and Technology (NTNU) in Trondheim. The five years as a student have been wonderful, and I am grateful for everything I have learned and all the friends I have got. The thesis is written in cooperation with Taskforce Salmon Lice, which is a R&D project that focuses on salmon lice.

To be able to finish my studies by writing a thesis aimed at the salmon farming industry is absolutely fantastic. Throughout my studies, I have wanted to combine my statistics knowledge with the aquaculture industry, so many thanks to Taskforce Salmon Lice and the partner companies for giving me the opportunity to do so. By working on this thesis, I have developed a better understanding of both the theoretical statistics and how it can be used in different fields. My skills in `R` have also been greatly improved. In addition, I have learned a lot about salmon farming, and look forward to continuing with statistical analyzes in the aquaculture industry, in my new job at Aqua Kompetanse.

First, I would like to thank my supervisor John Sølve Tyssedal at the Department of Mathematical science at NTNU and my co-supervisor Anna Solvang Båtnes at NTNU SeaLab. Thanks for the guidance, quick feedback and great advice. The master thesis was a continuation of my project assignment (Karlsen, 2020), where also Lone Sunniva Jevne was my co-supervisor. I would therefore also like to thank Lone for sharing great tips and experiences from her work at Taskforce Salmon Lice. Finally, I would also like to thank my partner and family for support through my studies, and for relevant knowledge and input about salmon lice and the aquaculture industry.

Abstract

The salmon louse (*Lepeophtheirus salmonis*) is a parasite that inflicts major economic and ecological consequences for the Atlantic salmon aquaculture. In an attempt to gain control of the salmon lice, the aquaculture industry is testing various methods and operating models. The aim of this thesis is to see if a shortened sea phase has a positive effect on salmon lice by analyzing weekly lice numbers. In this thesis, a shortened sea phase is achieved by placing salmon in the inner fjord system during the first part of the production. After 7-9 months the salmon are moved to more exposed sites along the coast, where it remains until it is ready for slaughter.

The development of salmon lice have been studied in production area 7 and 8 in the period 2012-2021, by examining count data from the salmon farmers. To compare different operating strategies, both weekly data from production cycles operated with a shortened sea phase and data from year-round productions, both in the fjords and along the coast, have been analyzed. The reported lice numbers have been plotted against different explanatory variables to assess the variables impact on lice numbers.

To investigate the effect of the shortened sea phase, generalised linear models, zero-inflated models and zero-altered models have been fitted to the lice count data. The results in this thesis suggest that a zero-altered negative binomial model seems to fit the lice count data, but the regression model could have been improved by including numerical variables for salinity and other environmental measurements. According to the fitted regression model, lower lice numbers are associated with the cages inside the fjords compared with cages along the coast. The model also indicates that compared with year-round operations in the fjord and along the coast, there are expected a lower count of adult female lice for the investigated operating model, where the sea phase is shortened by keeping the salmon in the inner fjord systems the first seven to nine months after deployment.

Sammendrag

Lakselus (*Leopeohtheirus salmonis*) er en parasitt som påfører store økonomiske og økologiske konsekvenser for havbruksnæringen. I forsøk på å få kontroll over lakselusen tester oppdrettsnæringen ut ulike metoder og driftsmodeller. Målet med masteroppgaven er å undersøke om en forkortet sjøfase har en positive effekt på lakselusen ved å analysere ukentlige lusetall. I dette studiet er forkortet sjøfase oppnådd ved å plassere laksen i indre fjordsystem den første delen av produksjonen. Etter 7-9 måneder blir laksen flyttet ut til mer eksponerte lokaliteter langs kysten, hvor den blir værende til den er slakteklar.

Utviklingen av lakselus ble studert i produksjonsområde 7 og 8 i perioden 2012-2020, ved å undersøke lusetall fra lakseoppdretterne. For å sammenligne ulike driftsstrategier, har både ukentlige data fra produksjonssykluser driftet med en forkortet sjøfase og data fra helårsproduksjoner, både i fjordene og langs kysten, blitt analysert. De rapporterte lusetallene har blitt plottet mot ulike forklaringsvariabler for å undersøke deres påvirkning på lusetallene.

For å studere effekten av forkortet sjøfase, har både generaliserte lineære modeller, zero-inflated modeller og zero-altered modeller blitt tilpasset lusetallene. Resultatene i denne masteroppgaven indikerer at en zero-altered negativ binomisk modell ser ut til å passe til dataene, men regresjonsmodellen kunne ha blitt forbedret ved å inkludere numeriske variabler for saltinnhold og andre miljømålinger. I følge den tilpassede regresjonsmodellen er det lavere lusetall på oppdrettslaks inne i fjordene, sammenlignet med oppdrettslaks langs kysten. Regresjonsmodellen indikerer også at det forventede antallet voksne hunnlus for den undersøkte driftsstrategien, hvor forkortet sjøfasen oppnås ved å holde laksen inne i fjorden de første syv til ni månedene etter utsett, er mindre enn for helårsdrift i fjorden og langs kysten.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Sea Lice

The salmon louse, *Leopeohtheirus salmonis* (Krøyer, 1837), has been a serious problem for the Norwegian aquaculture industry since the 1970s (Torrissen et al., 2013). In recent years, *Caligus elongatus* (Normann, 1832) has also caused problems for the salmon farmers (Gaasø, 2019; Hemmingsen et al., 2020). These two parasites are referred to as sea lice in this thesis, and the term salmon lice will be reserved for only *Leopeohtheirus salmonis*. The sea louse infiltrate the salmon by attaching itself to the skin of salmon with grip-hooks, and feeding on mucus, blood and skin (Overton et al., 2019, Gaasø, 2019). These infestations can lead to physical damage, chronic stress and skin damage, which makes the salmon more exposed to secondary bacterial infections (Overton et al., 2019). The high production of salmon the last years has led to high density of hosts year-round, and thus created good conditions for the sea louse growth and transmission (Torrissen et al., 2013).

1.1.1 Developmental Stages

Following Hamre et al. (2013), the life cycle of the salmon louse consists of eight stages (Figure 1.1). In the initial stages, the salmon louse flow freely in the water and may spread over large areas. When the louse attach itself to the salmon, it starts growing and develops through several stages until it can move and gradually become a full-grown adult louse. In the first stages after the salmon louse has attached itself to the salmon, Chalimus I and Chalimus II, the louse is sessile. From the Chalimus stages, the louse develops into pre-adult I and then pre-adult II. In these stages, the louse can move around on the surface of the salmon. Finally, it becomes an adult male or an adult female. The adult female louse lays eggs which becomes free-living parasites, and the life cycle is started again.

The salmon lice count data can be sorted in three different categories: sessile lice, mobile lice and adult female lice. Sessile lice corresponds to the Chalimus stages and mobile lice to the preadult and adult male stages. Adult female lice is a separate category, and is not included in the count of mobile lice (Hamre et al., 2013; Jevne, 2020). In this thesis, all the preadult and adult stages, including the adult female lice, are referred to as all mobile lice. *Caligus elongatus* (C. *elongatus*) develops trough four Chalimus stages before it becomes an adult (Hemmingsen et al., 2020), but are not divided into different stages in this thesis.

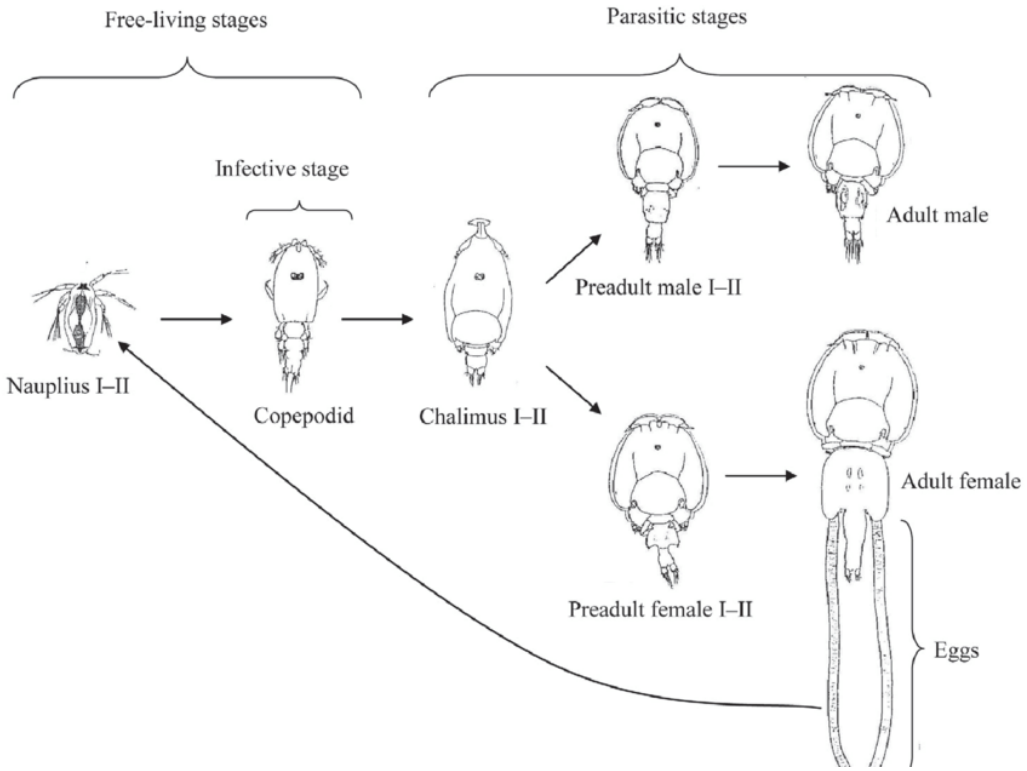

Figure 1.1: Developmental stages of the salmon lice, Lepeophtheirus salmonis (diagram, not to scale).

Source: Igboeli and Burka, 2013

1.1.2 Regulations

The Ministry of Trade, Industry and Fisheries have laid down regulations for salmon farmers to reduce the occurrence of salmon lice, such that the harmful effects on the salmon are minimized (Forskrift om lakselusbekjempelse, 2012). According to these regulations, the salmon farmers must report the average number of salmon lice per salmon at the site each week. The sea temperature at 3 meters depth and eventually delousing treatments used must also be reported. The reported average for the site is based on the calculated sample mean of salmon lice for each of the cages at the salmon farm. Once per week, salmon lice must be counted on at least 10 random salmons in each cage. If the temperature is below 4 °C, the number of lice per salmon can be counted every fourteen days instead of once a week. The average number of salmon lice per salmon calculated from a sample from the cage, is referred to as the lice number, and is divided into the three stages defined in Section 1.1.1: adult female lice, mobile lice and sessile lice. The regulations state that the allowed limit of average adult female lice per salmon at a site is 0.5, but the limit is reduced to 0.2 during 6 weeks in the spring. For Trøndelag and counties further south, this applies to the weeks 16-21, while for counties north of Trøndelag the lice limit is reduced in the weeks 21-26. During these 6 weeks, salmon lice must be counted on at least 20 random salmons in each cage (Forskrift om lakselusbekjempelse, 2012). C. *elongatus* is currently not regulated in Norway.

1.1.3 Treatments

The regulation on the control of salmon lice state that measures must be implemented to ensure that the amount of salmon lice does not exceed the limit for adult female lice (Forskrift om lakselusbekjempelse, 2012). There are several different delousing treatments that are used to reduce the lice pressure in cages with high lice numbers, and they can be divided into five categories: bath treatment, oral treatment, lice flusher, freshwater treatment and thermic treatment. Medicinal bath treatments were the first delousing treatments used in Norway, but its use has declined in recent years (Overton et al., 2019; Poppe et al., 1999). Medicinal oral treatment are use of specialized feed to get rid of the lice. The salmon louse has developed resistance to most of the chemotherapeutants which are approved for delousing in Norway today. Therefore, new delousing methods that do not involve the use of chemotherapeutants have been developed. Lice flusher, freshwater and thermic treatment are mechanical methods developed in the last ten years, and are the most utilized treatments today (Myhre Jensen et al., 2020; Norwegian Seafood Research Fund, n.d.; Overton et al., 2019).

1.2 Start Sites and Growth Sites

The sea lice survival and development are optimal in high-salinity sea water (Heuch et al., 2009; Torrissen et al., 2013). Bricknell et al. (2006) shows that both the salmon louse survival and infectivity are impaired by a reduction in salinity levels. In this context, an operating model with sites in inner fjord systems, where the salinity is low, is considered to give the farmers better control over the sea lice. According to Torrissen et al. (2013), a high density of salmon in the fjord throughout the year creates good conditions for the sea lice. To avoid this, the salmon is therefore transported further out after a period in the inner fjord system. In this way, the salmon farms will be fallowed (emptied and not restocked for a period) more frequently. It is thus possible that the environmental impact will be reduced through less emissions, and that the lice pressure on wild fish and farmed salmon will be lower. However, moving salmon between salmon farms increases the risk of spreading infections, and must be taken into account when such a model is considered (Veiledning: Flytting av laksefisk mellom oppdrettsanlegg, 2019).

A salmon production operated after this model can be divided into two stages, start and growth. This operating method is referred to as the stage model in this thesis. With the stage model, the smolts are deployed in inner fjord systems, called start sites, and after seven to nine months, the salmon are moved to more exposed sites further out, called growth sites. The traditional production of salmon in Norwegian salmon farms does not include the movement of salmon between the salmon farms. The salmon are then kept in the same site from deployment to slaughter, which takes around 1.5 years. The salmon that are deployed together in a cage, is in this thesis referred to as a generation, and the sites where the salmon are located throughout production are called whole-generation sites. This operating method is divided into whether the salmon is kept along the coast or in inner fjord systems throughout production, and is referred to as the coast model and the fjord model, respectively.

1.3 Earlier Studies

Data from the first part of the production of the stage model, the start stage, were analysed in Karlsen (2020). A Poisson model and a zero-inflated Poisson model were fitted for the estimated total number of adult female in the cage, and possible factors that affected the number of lice were investigated. This master thesis is a continuation of Karlsen (2020), where the entire production cycle, from deployment to slaughter, have been analysed. The regression analysis have been improved by adding more explanatory variables to the regression model, and other response variables have been considered.

Following Brakestad (2020), a zero-inflated negative binomial model is a better choice for modelling the lice count than the zero-inflated Poisson model. In Brakestad (2020), the analysis was based on open data material on the web, and not collected from the salmon farmers. This led to other explanatory variables, data at site level instead of at cage level, and that some more assumptions were made. Thus, both a Poisson model and a negative binomial model were fitted in this thesis, in addition to associated zero-inflated models and hurdle models.

1.4 Aims and Outline of the Thesis

The main aim of this master thesis was to investigate the effect shortened sea phase, obtained with the stage model, has on salmon lice. Data from several salmon farms in production area 7 and 8, both in inner fjord systems and along the coast, have been analysed to achieve this. To evaluate the stage model, some sub-aims were formulated:

- Compare the stage model against whole-generation sites along the coast (coast model)
- Compare the stage model against whole-generation sites in inner fjord systems (fjord model)
- Compare the two fjords, are there any similarities in the development of lice?

The necessary statistical theory is presented in Section 2. Information about the study area and the data set are given in Section 3, followed by a visualization of the data in Section 4. The data analysis and the validation of the fitted models are presented in Section 5. Finally, a discussion with recommendations for further work is presented in Section 6.

2 Theory

To investigate the effect of the stage model, a regression model for the weekly count of salmon lice in each cage is fitted to the data. The lice count is endeavoured explained by the covariates describing the sea temperature, the location of the salmon farm, operating model, season of the year, cage biomass, production week, delousing treatment, distance to the coastline, number of neighbours, last weeks lice number and an intervention variable for weeks with high lice pressure (defined in Section 3.2). This section provides theory used in the analysis of the lice count data, where both generalized linear models, zero-inflated models and zero-altered models are fitted.

2.1 Generalized Linear Models

Generalized linear models consist of three parts: the distribution of the response variable, a function of the explanatory variables (the systematic part) and the link between the mean of the response variable and the systematic part. The Poisson distribution and the negative binomial distribution are two of the most common distributions used for count response variables (Zuur et al., 2009).

2.1.1 Poisson GLM

For a Poisson GLM, each observation Y_i, for $i = 1, \ldots, n$ is assumed to be an independent Poisson distributed variable with mean and variance equal to λ_i. The probability mass function for each variable Y_i is given by

$$f(y_i|\lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}, \qquad y_i = 0, 1, 2, \ldots, \quad \lambda_i > 0. \tag{2.1}$$

Given p covariates $x_{i1}, x_{i2}, \ldots, x_{ip}$, where p is the number of parameters without intercept, the systematic part can be specified by the linear predictor, $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. To ensure that the fitted values always are non-negative, the link function $\ln(\lambda_i) = \mathbf{x}_i^T\boldsymbol{\beta} = \eta_i$ is used to link the covariates with the mean λ_i. The expected response function, which is the inverse of the link function, is thus $\mathrm{E}[Y_i] = \lambda_i = \exp(\eta_i) = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$ (Fahrmeir et al., 2013; Zuur et al., 2009).

Following Fahrmeir et al. (2013), the parameter vector of interest, $\boldsymbol{\beta}$, which includes the intercept and the slopes of the covariates, is estimated by maximizing the log-likelihood function. By assuming that the response variables y_i are conditionally independent Poisson distributed variables, the log-likelihood is given by

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ln f(y_i|\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln\left(\frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}\right) = \sum_{i=1}^{n} \left[y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\right] \\ &= \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T\boldsymbol{\beta} - \exp(\mathbf{x}_i^T\boldsymbol{\beta}) - \ln(y_i!)\right]. \end{aligned} \tag{2.2}$$

The maximum likelihood estimate is found by solving $\mathbf{s}(\hat{\boldsymbol{\beta}}) = 0$, where the score function, $\mathbf{s}(\boldsymbol{\beta})$, is the first-order derivatives of the log-likelihood. The score function can be derived as

$$\begin{aligned}\mathbf{s}(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \frac{\partial l_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{s}_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \mathbf{x}_i \right] \\ &= \sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \right). \end{aligned} \tag{2.3}$$

By using that $\mathrm{E}[\mathbf{s}_i(\boldsymbol{\beta})] = 0$ for all i, the covariance matrix of the score function can be expressed as

$$\begin{aligned}\mathbf{F}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} Cov[\mathbf{s}_i(\boldsymbol{\beta})] = \sum_{i=1}^{n} E\left[\mathbf{s}_i(\boldsymbol{\beta}) \mathbf{s}_i^T(\boldsymbol{\beta}) \right] \\ &= \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T E\left[(y_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}))^2 \right] = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \lambda_i, \end{aligned} \tag{2.4}$$

where $E\left[(y_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}))^2\right] = E\left[(y_i - \lambda_i)^2\right] = Var[y_i] = \lambda_i$ is used in the last transition. The covariance matrix $\mathbf{F}(\boldsymbol{\beta})$ is the expected Fisher information matrix, and its inverse is used in solving $\mathbf{s}(\hat{\boldsymbol{\beta}}) = 0$ by the Fisher Scoring Algorithm

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \mathbf{F}^{-1}\left(\hat{\boldsymbol{\beta}}^{(t)} \right) \mathbf{s}\left(\hat{\boldsymbol{\beta}}^{(t)} \right). \tag{2.5}$$

The maximum likelihood estimate is asymptotically distributed as $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \mathbf{F}^{-1}(\hat{\boldsymbol{\beta}}))$. The diagonal elements in the inverted expected Fisher information matrix evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ are thus the asymptotically variance for the parameters (Fahrmeir et al., 2013).

If the data shows greater variability in the response counts than presumed by the Poisson model ($\mathrm{Var}[Y_i] > E[Y_i] = \lambda_i$), the model is overdispersed. A dispersion parameter ϕ can be introduced by assuming $\mathrm{Var}[Y_i] = \phi \lambda_i$, and it can be estimated as the average deviance or Pearson statistic

$$\hat{\phi}_D = \frac{D}{n-p}, \quad \text{or} \quad \hat{\phi}_P = \frac{P}{n-p}, \tag{2.6}$$

where n is the number of observations, p is the number of parameters, D is the residual deviance and P is the Pearson statistic. If the dispersion parameter ϕ is larger than 1 it provides evidence for overdispersion. The residual deviance, D, and the Pearson statistics, P, for the Poisson GLM are in Zuur et al. (2009) defined as

$$D = 2 \sum_{i=1}^{n} \left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right] \quad \text{and} \quad P = \sum_{i=1}^{n} \frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i}, \tag{2.7}$$

where $\hat{\lambda}_i = \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$.

The deviance residuals, d_i, and the Pearson residuals, r_i, are important tools for model validation. For the Poisson model, they are in Zuur et al. (2009) defined as

$$d_{i,P} = \mathrm{sign}(y_i - \hat{\lambda}_i) \sqrt{2 \left[y_i \ln\left(\frac{y_i}{\hat{\lambda}_i} \right) - (y_i - \hat{\lambda}_i) \right]} \quad \text{and} \quad r_{i,P} = \frac{y_i - \hat{\lambda}_i}{\sqrt{\hat{\lambda}_i}}, \tag{2.8}$$

where $\text{sign}(y_i - \hat{\lambda}_i) = 1$ if $y_i - \hat{\lambda}_i > 0$ and $\text{sign}(y_i - \hat{\lambda}_i) = -1$ if $y_i - \hat{\lambda}_i < 0$. For a good model there should not be any patterns in the residuals (Zuur et al., 2009).

2.1.2 Negative Binomial GLM

In cases with overdispersion, a negative binomial response function can be useful. The negative binomial distribution allows that the variance of the response variable is larger than the mean, and has in addition to the mean, a dispersion parameter to capture the amount of over-dispersion. Assume that each reported count of lice are independent Bernoulli(p) trials and let the random variable Y denote the number of failures before the rth success, where r is a fixed integer. The probability mass function for Y is in Casella and Berger (2002) given by

$$f(y|p,r) = \binom{r+y-1}{y} p^r (1-p)^y, \quad y = 0, 1, 2, \ldots. \tag{2.9}$$

The expected value for the negative binomial distributed variable Y can following Casella and Berger (2002) be derived as

$$\begin{aligned} \text{E}[Y] &= \sum_{y=0}^{\infty} y \binom{r+y-1}{y} p^r (1-p)^y = \sum_{y=1}^{\infty} \frac{(r+y+1)!}{(y-1)!(r-1)!} p^r (1-p)^y \\ &= \sum_{y=1}^{\infty} \binom{r+y-1}{y-1} p^r (1-p)^y. \end{aligned} \tag{2.10}$$

By using $z = y - 1$, the expression can be simplified to

$$\begin{aligned} \text{E}[Y] &= \sum_{z=0}^{\infty} \binom{r+z}{z} p^r (1-p)^{z+1} = r\frac{1-p}{p} \sum_{z=0}^{\infty} \binom{(r+1)+z-1}{z} p^{r+1} (1-p)^z \\ &= r\frac{1-p}{p}, \end{aligned} \tag{2.11}$$

where it is in the last transition used that the sum-term is equal to 1, since it is the sum over all values of a negative distributed variable Z with $r = r + 1$. The variance, $\text{Var}[Y]$, can be calculated as $\text{Var}[Y] = \text{E}[Y^2] - \text{E}[Y]^2$. By deriving $\text{E}[Y^2]$ in the same way as for the expected value, the variance can be expressed as $\text{Var}[Y] = \frac{r(1-p)}{p^2}$ (Casella & Berger, 2002).

By defining the parameter λ for the mean, $\lambda = \text{E}[Y] = r\frac{1-p}{p}$, the variance for the negative binomial distribution can be expressed as $\text{Var}[Y] = \lambda + \frac{\lambda^2}{r}$. By replacing p with $\frac{r}{\lambda+r}$ in Equation (2.9), the probability function for the negative binomial model in terms of the mean, λ, follows as

$$\begin{aligned} f(y|\lambda,r) &= \binom{r+y-1}{y} \left(\frac{r}{\lambda+r}\right)^r \left(1 - \frac{r}{\lambda+r}\right)^y \\ &= \frac{(r+y-1)!}{y!(r-1)!} \left(\frac{r}{\lambda+r}\right)^r \left(1 - \frac{r}{\lambda+r}\right)^y \\ &= \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)} \left(\frac{r}{\lambda+r}\right)^r \left(\frac{\lambda}{\lambda+r}\right)^y, \end{aligned} \tag{2.12}$$

for $y = 0, 1, 2 \ldots$ (Casella & Berger, 2002). The covariates for observation i can be linked to the mean of Y_i with the link function $\ln(\lambda_i) = \eta_i$, where $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, p is the number of parameters without intercept and the vector $\boldsymbol{\beta}$ includes the intercept and the p regression coefficients. The expected mean is thus determined by $\lambda_i = \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$.

The maximum likelihood estimate for the negative binomial model is derived in Cameron and Trivedi (1998) with the parameterization $\alpha = r^{-1}$, which is another parameter than used in this thesis. The parameters of interest in this thesis are $\boldsymbol{\beta}$ and r. The log-likelihood with this parameterization is found in Zuur et al. (2009), but no further calculation of the maximum likelihood estimates has been found in the literature. These have therefore been derived in this thesis. By assuming that the response variables y_i are conditionally independent, the log-likelihood for the negative binomial distribution is derived as

$$\begin{aligned} l(\boldsymbol{\beta}, r) &= \sum_{i=1}^{n} \ln f(y_i|\lambda_i, r) = \sum_{i=1}^{n} \ln\left(\frac{\Gamma(r+y_i)}{\Gamma(y_i+1)\Gamma(r)}\left(\frac{r}{\lambda_i+r}\right)^r\left(\frac{\lambda_i}{\lambda_i+r}\right)^{y_i}\right) \\ &= \sum_{i=1}^{n}\left[\frac{\ln\Gamma(r+y_i)}{\ln\Gamma(r)} - \ln\Gamma(y_i+1) + r\ln\left(\frac{r}{\lambda_i+r}\right) + y_i\ln\left(\frac{\lambda_i}{\lambda_i+r}\right)\right] \\ &= \sum_{i=1}^{n}\left(\sum_{j=0}^{y_i-1}\ln(j+r)\right) - \sum_{i=1}^{n}[\ln\Gamma(y_i+1) + r\ln r - r\ln(\lambda_i+r) \\ &\quad + y_i\ln(\lambda_i) - y_i\ln(\lambda_i+r)], \end{aligned} \tag{2.13}$$

where it is in the last transition used that $\frac{\Gamma(r+y)}{\Gamma(r)} = \prod_{j=0}^{y-1}(j+r)$ for integers y (Cameron & Trivedi, 1998). By substituting $\lambda_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$ in the log-likelihood and taking derivatives with respect to $\boldsymbol{\beta}$ and r, respectively, the following score functions are obtained

$$\begin{aligned} \mathbf{s}(\boldsymbol{\beta}) &= \frac{\partial l(\boldsymbol{\beta}, r)}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\left[y_i\frac{\partial\ln\left(\mathbf{x}_i^T\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}} - (r+y_i)\frac{\partial\ln\left(\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r\right)}{\partial\boldsymbol{\beta}}\right] \\ &= \sum_{i=1}^{n}\left[y_i\mathbf{x}_i^T - (r+y_i)\frac{\mathbf{x}_i\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r}\right] = \sum_{i=1}^{n}\left[r\mathbf{x}_i\frac{y_i-\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r}\right], \end{aligned} \tag{2.14}$$

$$\begin{aligned} \mathbf{s}(r) &= \frac{\partial l(\boldsymbol{\beta}, r)}{\partial r} = \sum_{i=1}^{n}\left[\frac{\partial\left(\sum_{j=0}^{y_i-1}\ln(j+r)\right)}{\partial r} + \frac{\partial(r\ln r)}{\partial r} - \frac{\partial\left((r+y_i)\ln\left(\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r\right)\right)}{\partial r}\right] \\ &= \sum_{i=1}^{n}\left[\sum_{j=0}^{y_i-1}\frac{1}{j+r} + \ln r + 1 - \ln(\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r) - \frac{r+y_i}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r}\right] \\ &= \sum_{i=1}^{n}\left[\sum_{j=0}^{y_i-1}\frac{1}{j+r} + \ln r - \ln(\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r) + \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})-y_i}{\exp(\mathbf{x}_i^T\boldsymbol{\beta})+r}\right]. \end{aligned} \tag{2.15}$$

The maximum likelihood estimates (MLE) of $\boldsymbol{\beta}$ and r are found by solving $\mathbf{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ and $\mathbf{s}(r) = 0$. The asymptotically distribution of the negative binomial MLE $\boldsymbol{\beta}$ and $\alpha = r^{-1}$,

is in Cameron and Trivedi (1998) given as

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\alpha}\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\boldsymbol{\beta}\\ \alpha\end{pmatrix}, \begin{pmatrix}\mathbf{F}_{11}(\hat{\boldsymbol{\beta}}) & \mathbf{F}_{12}(\hat{\boldsymbol{\beta}},\hat{\alpha})\\ \mathbf{F}_{21}(\hat{\boldsymbol{\beta}},\hat{\alpha}) & \mathbf{F}_{22}(\hat{\alpha})\end{pmatrix}^{-1}\right], \tag{2.16}$$

where $\mathbf{F}(\hat{\boldsymbol{\beta}},\hat{\alpha})$ is the expected Fisher information matrix for the maximum likelihood estimates. The expected Fisher information matrix can be calculated as the expected value of the observed Fisher information matrix $\mathbf{H}(\hat{\boldsymbol{\beta}},\hat{\alpha})$, which contains the negative Hessian matrix of the log-likelihood (Fahrmeir et al., 2013). By using the score functions obtained above, the asymptotically distribution of $\boldsymbol{\beta}$ and r can be derived. The elements of the expected Fisher information matrix $\mathbf{F}(\hat{\boldsymbol{\beta}},r) = \mathrm{E}[\mathbf{H}(\hat{\boldsymbol{\beta}},r)]$, are obtained as

$$\mathbf{F}_{11}(\boldsymbol{\beta}) = E\left[-\frac{\partial \mathbf{s}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] = E\left[-\frac{\partial}{\partial \boldsymbol{\beta}}\sum_{i=1}^{n}\left(r\mathbf{x}_i\frac{y_i-\lambda_i}{\lambda_i+r}\right)\right] = \sum_{i=1}^{n}\frac{r\lambda_i\mathbf{x}_i\mathbf{x}_i^T}{\lambda_i+r} \tag{2.17}$$

$$\begin{aligned}\mathbf{F}_{22}(r) &= E\left[-\frac{\partial \mathbf{s}(r)}{\partial r}\right] = E\left[\sum_{i=1}^{n}\left(\frac{1}{\lambda_i+r}-\frac{1}{r}+\frac{\lambda_i-y_i}{(\lambda_i+r)^2}+\sum_{j=0}^{y_i-1}\frac{1}{(j+r)^2}\right)\right]\\ &= \sum_{i=1}^{n}\left(E\left[\sum_{j=0}^{y_i-1}\frac{1}{(j+r)^2}\right]-\frac{\lambda_i}{r(\lambda_i+r)}\right)\end{aligned} \tag{2.18}$$

$$\mathbf{F}_{12}(\boldsymbol{\beta},r) = \mathbf{F}_{12}(\boldsymbol{\beta},r) = \mathbf{0}, \tag{2.19}$$

where $\mathbf{F}_{12}(\boldsymbol{\beta},r)$ and $\mathbf{F}_{21}(\boldsymbol{\beta},r)$ are equal to $\mathbf{0}$, since

$$E\left[\frac{\partial^2 l(\boldsymbol{\beta},r)}{\partial\boldsymbol{\beta}\partial r}\right] = E\left[\frac{\partial \mathbf{s}(\boldsymbol{\beta})}{\partial r}\right] = E\left[\frac{\partial \mathbf{s}(r)}{\partial \boldsymbol{\beta}}\right] = E\left[\sum_{i=1}^{n}\frac{\lambda_i\mathbf{x}_i(y_i-\lambda_i)}{(\lambda_i+r)^2}\right] = \mathbf{0}. \tag{2.20}$$

This simplifies the covariance matrix, and the maximum likelihood estimates $\hat{\boldsymbol{\beta}}$ and $\hat{r}$ follows the asymptotically distribution

$$\begin{pmatrix}\hat{\boldsymbol{\beta}}\\ \hat{r}\end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix}\boldsymbol{\beta}\\ r\end{pmatrix}, \begin{pmatrix}\mathbf{F}_{11}^{-1}(\hat{\boldsymbol{\beta}}) & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_{22}^{-1}(\hat{r})\end{pmatrix}\right], \tag{2.21}$$

where $\mathbf{F}_{11}(\boldsymbol{\beta})$ and $\mathbf{F}_{22}(r)$ are defined in Equation (2.17) and (2.18), respectively.

The Pearson residuals are calculated as the difference between the observed values y_i and the estimated values $\hat{\lambda}_i$, divided by the square root of the variance of Y_i (Fahrmeir et al., 2013). For the negative binomial model the Pearson residuals can then be calculated as

$$r_{i,NB} = \frac{y_i-\hat{\lambda}_i}{\sqrt{\hat{\lambda}_i+\frac{\hat{\lambda}_i^2}{r}}}. \tag{2.22}$$

The deviance residuals for the negative binomial model is in Hilbe (2011) expressed as

$$d_{i,NB} = \mathrm{sign}(y_i-\hat{\lambda}_i)\sqrt{2\left[y_i\ln\left(\frac{y_i}{\hat{\lambda}_i}\right)-(y_i+r)\ln\left(\frac{r+y_i}{r+\hat{\lambda}_i}\right)\right]} \tag{2.23}$$

2.1.3 Hypothesis Testing

The likelihood ratio test can be used to compare two nested models. Let $\boldsymbol{\beta}$ be a parameter vector of regression coefficient and $\boldsymbol{\beta}_r$ be a r-dimensional sub-vector of $\boldsymbol{\beta}$. For testing the significance of the r covariates the following hypothesis can be used

$$H_0 : \boldsymbol{\beta}_r = \mathbf{0} \quad \text{vs.} \quad H_1 : \boldsymbol{\beta}_r \neq \mathbf{0}.$$

Under H_1 the full model, denoted as A, is considered, while the smaller model, B, without the r regression coefficients is considered under H_0. Model B is nested within model A, and the likelihood ratio test statistic can be calculated as

$$-2 \ln \boldsymbol{\lambda} = -2 \left(\ln L(\hat{\boldsymbol{\beta}}_B) - \ln L(\hat{\boldsymbol{\beta}}_A) \right). \tag{2.24}$$

Under the null hypothesis the test statistic $-2 \ln \boldsymbol{\lambda}$ are asymptotically χ^2_r distributed with r degrees of freedom, which is the difference in degrees of freedom from $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_B$ (Fahrmeir et al., 2013).

If the same covariates are used in fitting both a Poisson GLM and a negative binomial GLM, the negative binomial GLM contains all the terms in the Poisson GLM, and the models are thus nested. As stated earlier, the variance for the Poisson GLM and the negative binomial GLM are λ_i and $\lambda_i + \lambda_i^2/r$, respectively. By defining $\alpha = 1/r$, the variance for the negative binomial model can be expressed as $\text{Var}[Y_i]_{NB} = \lambda_i + \alpha\lambda_i^2$. For $\alpha = 0$, the variance is the same for both the models, and a likelihood ratio test with the null hypothesis $H_0 : \alpha = 0$ can be used to compare the two regression models. The alternative hypothesis is $H_1 : \alpha > 0$, and under H_0 the test statistic, $-2 \ln \boldsymbol{\lambda}$, follows a $0.5(\chi^2_0 + \chi^2_1) = 0.5\chi^2_1$ distribution. This can be taken into account by dividing the p-value by 2 before evaluating the result (Zuur et al., 2009).

2.1.4 Model Validation

The residual deviance is twice the difference between the log-likelihood of the observed values y_i and the log-likelihood of the fitted model. The expected mean from the model fit is defined as $\hat{\lambda}_i = \exp(\mathbf{x}_i^T \hat{\boldsymbol{\beta}})$, which gives the residual deviance $2[L(y) - L(\hat{\lambda}_i)]$. The explained deviance can be used as a measure of goodness of fit, and is calculated as

$$\frac{\text{null deviance - residual deviance}}{\text{null deviance}} \cdot 100\%, \tag{2.25}$$

where the null deviance is the residual deviance in the model that only contains an intercept (Zuur et al., 2009).

Another goodness of fit criteria is the Pearson statistic, which is Chi-squared distributed with $n-p$ degrees of freedom. The Pearson statistic, P, is found by squaring and summing all the Pearson residuals, which are in general given as

$$r_i = \frac{y_i - \widehat{\mathrm{E}[Y_i]}}{\sqrt{\widehat{\mathrm{Var}[Y_i]}}}, \tag{2.26}$$

If the Pearson statistic P is larger than $\chi^2_{\alpha,n-p}$ for a significance level α, the null hypothesis is rejected and the test indicates that the model does not fit with the observed distribution. (Fahrmeir et al., 2013; Zuur et al., 2009).

The Akaike information criterion (AIC) is in general defined as

$$AIC = -2l(\hat{\boldsymbol{\beta}}) + 2p, \tag{2.27}$$

where l is the log-likelihood and p is the number of regression parameters (without intercept). If the model contains a dispersion parameter ϕ, the total number of parameter must be increased with one, since the maximum likelihood estimator of ϕ should be substituted into the model. For model selection the model with the lowest AIC is preferred. The penalty term for the number of parameters are included in the AIC to prevent overfitting (Fahrmeir et al., 2013).

2.2 Models for Zero-Inflated Count Data

A large set of the reported lice numbers are zero, and alternative methods which can deal with excessive number of zeroes are therefore fitted to the data. In this thesis, both the zero-inflated and the zero-altered models for the Poisson and the negative binomial model are used. These models performs two processes, one binomial logit model causing zeroes and one Poisson or negative binomial model generating counts. The models will be presented in more detail later in this subsection.

The distinctions between the Poisson and the negative binomial model are, as presented in the section for the generalised linear model, the ability to handle overdispersion in the count part. The zero-inflated Poisson (ZIP) is nested in a zero-inflated negative binomial (ZINB) model, and can, as for the Poisson GLM and the negative binomial GLM, be compared by the likelihood ratio test presented in Section 2.1.3. This also applies to the zero-altered Poisson (ZAP) and the zero-altered negative binomial (ZANB) models (Zuur et al., 2009).

The difference between the zero-inflated models and the zero-altered models is related to how the zeroes are modelled. The zero-inflated model is a mixture model and can predict zeros in both processes, while the zero-altered model is a two-part model, where one part predicts zero and the other predicts non-zero counts with a truncated Poisson or negative binomial model. The zero-inflated models distinguish between the type of zero; false zero or true zero. According to Zuur et al. (2009), false zeroes can be due to observed errors, a suitable habitat which is not used or poor experimental design and sampling practises (design errors). The zero-altered models, also called hurdle models, does not distinguish between types of zeros, only absence and presence.

Due to a lack of literature on the maximum likelihood estimates (MLE) for several of these models, the MLE for the ZINB model and the hurdle models have been derived in this thesis. The maximum likelihood estimate for the ZIP model is presented in NCSS Statistical Software (2021). There also exists literature on the MLE for the ZINB model, but with an other parameterization than used here. For the hurdle model, however, no literature has been found on the derivation of the maximum likelihood estimates.

2.2.1 ZIP and ZINB Models

In the zero-inflated models, the zero process models the probability of observing a false zero with a binomial logistic model and the count process generates a true zero or a positive count. Let the probability of a false zero be π and $g(y)$ be the probability mass function for the Poisson distribution or the negative binomial distribution. Following Zuur et al. (2009), the probability mass function for the zero-inflated models can then be written as

$$f(y_i) = \begin{cases} \pi_i + (1-\pi_i)g(y_i = 0), & y_i = 0 \\ (1-\pi_i)g(y_i), & y_i > 0. \end{cases} \tag{2.28}$$

By substituting $g(y)$ with the probability mass function given in Equation (2.1), the pmf for the zero-inflated Poisson model is obtained. By using the negative binomial probability

mass function from Equation (2.12) instead, y_i becomes a zero-inflated negative binomial distributed variable.

The expected mean and the variance for the zero-inflated models can be derived by using the basic rules $\mathrm{E}(Y) = \sum_{y=0} yf(y)$, $\mathrm{Var}(Y) = \mathrm{E}(Y^2) - \mathrm{E}(Y)^2$ and $\Gamma(y+1) = y\Gamma(y)$, where $f(y)$ is the probability mass function for the ZIP or the ZINB model. The expected mean for both the ZIP and ZINB models can be expressed by $\mathrm{E}(Y_i) = \lambda_i(1-\pi_i)$, where λ_i is the expected count from the Poisson model and the negative binomial model, respectively. The variance for the ZIP model is in Zuur et al. (2009) given as $\mathrm{Var}(Y_i) = (1-\pi_i)(\lambda_i + \pi_i\lambda_i^2)$ and the variance for the ZINB model is given as $\mathrm{Var}(Y_i) = (1-\pi_i)(\lambda_i + \frac{\lambda_i^2}{r}) + \lambda_i^2 + (\pi_i^2 + \pi_i)$. By substituting the obtained mean and variance of ZIP and ZINB, respectively, in Equation (2.26), the Pearson residuals for the two zero-inflated models can be calculated as

$$r_{i,ZIP} = \frac{y_i - \hat{\lambda}_i(1-\hat{\pi}_i)}{\sqrt{(1-\hat{\pi}_i)(\hat{\lambda}_i + \hat{\pi}_i\hat{\lambda}_i^2)}} \tag{2.29}$$

$$r_{i,ZINB} = \frac{y_i - \hat{\lambda}_i(1-\hat{\pi}_i)}{\sqrt{(1-\hat{\pi}_i)(\hat{\lambda}_i + \frac{\hat{\lambda}_i^2}{r}) + \hat{\lambda}_i^2 + (\hat{\pi}_i^2 + \hat{\pi}_i)}}. \tag{2.30}$$

For both ZIP and ZINB, the linear predictor $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ is used to specify the systematic part in the count model and the log-link function links the covariates with the mean. This gives the response function $\lambda_i = \exp(\eta_i) = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$for the count part. The probability of a false zero from the logistic model may not include the same covariates as for the count model. Therefore, the probability of a false zero given by π_i, is set to be a function of the intercept and q covariates $z_{i1}, z_{i2}, \ldots, z_{iq}$. Following Zuur et al. (2009), and using the logit link function and the linear predictor $\phi_i = \mathbf{z}_i^T\boldsymbol{\gamma}$, the following relationship for the zero-part is obtained

$$\pi_i = \frac{\exp(\phi_i)}{1+\exp(\phi_i)} = \frac{\exp(\gamma_0)\exp(\gamma_1 z_{i1})\cdots\exp(\gamma_q z_{iq})}{1+\exp(\gamma_0)\exp(\gamma_1 z_{i1})\cdots\exp(\gamma_q z_{iq})}. \tag{2.31}$$

Assume that Y_i is a zero-inflated Poisson distributed variable, then the likelihood function for the ith observation can be given as

$$L_{i,ZIP}(\boldsymbol{\beta},\boldsymbol{\gamma}) = f_{ZIP}(y_i|\boldsymbol{\beta},\boldsymbol{\gamma}) = \begin{cases} \pi_i + (1-\pi_i)\exp(-\lambda_i), & y_i = 0 \\ (1-\pi_i)\frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}, & y_i > 0. \end{cases} \tag{2.32}$$

The response variables y_i are assumed to be conditionally independent, and the likelihood function for the ZIP model is given as $L_{ZIP}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{i=1}^n L_{i,ZIP}(\boldsymbol{\beta},\boldsymbol{\gamma})$. By taking the logarithm of the likelihood, and replacing π_i with $\frac{\mu_i}{1+\mu_i}$, where $\mu_i = \exp(\mathbf{z}_i^T\boldsymbol{\gamma})$, the log-likelihood function can be expressed as

$$\begin{aligned} l_{ZIP}(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{\substack{i=1 \\ y_i=0}}^n \ln[\pi_i + (1-\pi_i)\exp(-\lambda_i)] + \sum_{\substack{i=1 \\ y_i>0}}^n \ln\left[(1-\pi_i)\frac{\lambda_i^{y_i}\exp(-\lambda_i)}{y_i!}\right] \\ &= \sum_{\substack{i=1 \\ y_i=0}}^n \ln[\mu_i + \exp(-\lambda_i)] + \sum_{\substack{i=1 \\ y_i>0}}^n [y_i\ln(\lambda_i) - \lambda_i - \ln(y_i!)] - \sum_{i=1}^n \ln(1+\mu_i). \end{aligned} \tag{2.33}$$

The second order derivatives of the log-likelihood is needed to derive the asymptotic distribution of the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The first order derivatives of the log-likelihood are given by

$$\frac{\partial l_{ZIP}}{\partial \boldsymbol{\beta}} = -\sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{\mathbf{x}_i \lambda_i}{\mu_i \exp(\lambda_i) + 1} + \sum_{\substack{i=1 \\ y_i>0}}^{n} (y_i - \lambda_i)\mathbf{x}_i \tag{2.34}$$

$$\frac{\partial l_{ZIP}}{\partial \boldsymbol{\gamma}} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{\mathbf{z}_i \mu_i \exp(\lambda_i)}{\mu_i \exp(\lambda_i) + 1} - \sum_{i=1}^{n} \frac{\mu_i}{1+\mu_i}\mathbf{z}_i. \tag{2.35}$$

Thereby, the second order derivatives are

$$\frac{\partial^2 l_{ZIP}}{\partial \beta_m \partial \beta_n} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{x_{im} x_{in} \lambda_i [(\lambda_i - 1)\mu_i \exp(\lambda_i) - 1]}{[\mu_i \exp(\lambda_i) + 1]^2} - \sum_{\substack{i=1 \\ y_i>0}}^{n} \lambda_i x_{im} x_{in}, \quad m, n = 1, 2, \ldots, p \tag{2.36}$$

$$\frac{\partial^2 l_{ZIP}}{\partial \gamma_r \partial \gamma_n} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{z_{im} z_{in} \mu_i \exp(\lambda_i)}{[\mu_i \exp(\lambda_i) + 1]^2} - \sum_{i=1}^{n} \frac{z_{im} z_{in} \mu_i}{[1+\mu_i]^2}, \qquad m, n = 1, 2, \ldots, q \tag{2.37}$$

$$\frac{\partial^2 l_{ZIP}}{\partial \beta_m \partial \gamma_n} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{x_{im} z_{in} \mu_i \lambda_i \exp(\lambda_i)}{[\mu_i \exp(\lambda_i) + 1]^2}, \qquad m = 1, 2, \ldots, p \quad n = 1, 2, \ldots, q. \tag{2.38}$$

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ be the maximum likelihood estimates for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively, in the zero-inflated Poisson regression model. The asymptotically distribution of the parameters $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are in NCSS Statistical Software (2021) given as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \begin{pmatrix} -\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_n} & -\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\gamma}_n} \\ -\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\gamma}_n \partial \boldsymbol{\beta}_m} & -\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\gamma}_m \partial \boldsymbol{\gamma}_n} \end{pmatrix}^{-1} \right], \tag{2.39}$$

where $\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_n}$ is defined to be a $(p \times p)$ matrix including the second order derivatives in Equation (2.36), where the (m, n)th entry is $\frac{\partial^2 l_{ZIP}}{\partial \beta_m \partial \beta_n}$. Similarly, the $(q \times q)$ matrix $\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\gamma}_m \partial \boldsymbol{\gamma}_n}$ includes the second order derivatives in Equation (2.37). The matrices $\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\gamma}_n \partial \boldsymbol{\beta}_m}$ and $\frac{\partial^2 l_{ZIP}}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\gamma}_n}$ are respectively, $(q \times p)$ and $(p \times q)$, and contains the second order derivatives in Equation (2.38).

Similarly as for zero-inflated Poisson, the parameter estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{r}$ for the zero-inflated negative binomial model can be derived. The maximum likelihood estimates are thus calculated by maximizing the likelihood function $L_{ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r) = \prod_{i=1}^{n} L_{i,ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r)$, where

$$L_{i,ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r) = \begin{cases} \pi_i + (1 - \pi_i)\left(\frac{r}{\lambda_i + r}\right)^r, & y_i = 0 \\ (1 - \pi_i)\frac{\Gamma(r + y_i)}{\Gamma(y_i + 1)\Gamma(r)}\left(\frac{r}{\lambda_i + r}\right)^r \left(\frac{\lambda_i}{\lambda_i + r}\right)^{y_i}, & y_i > 0. \end{cases} \tag{2.40}$$

The log-likelihood $l_{ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r) = \sum_{i=1}^{n} \ln L_{i,ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r)$ is derived as

$$\begin{aligned} l_{ZINB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r) &= \sum_{\substack{i=1 \\ y_i=0}}^{n} \ln \left[\pi_i + (1-\pi_i) \left(\frac{r}{\lambda_i + r} \right)^r \right] + \sum_{\substack{i=1 \\ y_i>0}}^{n} [\ln(1-\pi_i) + l_{i,NB}(\boldsymbol{\beta}, r)] \\ &= \sum_{\substack{i=1 \\ y_i=0}}^{n} \ln \left[\mu_i + \left(\frac{r}{\lambda_i + r} \right)^r \right] + \sum_{\substack{i=1 \\ y_i>0}}^{n} l_{i,NB}(\boldsymbol{\beta}, r) - \sum_{i=1}^{n} \ln(1+\mu_i), \end{aligned} \tag{2.41}$$

where $l_{i,NB}(\boldsymbol{\beta}, r)$ is the negative binomial log-likelihood for the ith observation expressed in Equation (2.13), and the substitution $\pi_i = \frac{\mu_i}{1+\mu_i}$, where $\mu_i = \exp(\mathbf{z}_i^T \boldsymbol{\gamma})$, is used in the last transition. The ZINB log-likelihood has also similarities to the ZIP likelihood in Equation (2.33), and the last part $\sum_{i=1}^{n} \ln(1+\mu_i)$ is same in both likelihoods. By using calculations from the negative binomial model and the ZIP model, in addition to a little algebra, the score functions are derived as

$$\mathbf{s}_{ZINB}(\boldsymbol{\beta}) = \frac{\partial l_{ZINB}}{\partial \boldsymbol{\beta}} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{r^{r+1} \mathbf{x}_i \lambda_i}{(\lambda_i + r)(\mu_i(\lambda_i + r)^r + r^r)} + \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[r \mathbf{x}_i \frac{y_i - \lambda_i}{\lambda_i + r} \right] \tag{2.42}$$

$$\mathbf{s}_{ZINB}(\boldsymbol{\gamma}) = \frac{\partial l_{ZINB}}{\partial \boldsymbol{\gamma}} = \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{\mu_i \mathbf{z}_i}{\mu_i + \left(\frac{r}{\lambda_i + r} \right)^r} - \sum_{i=1}^{n} \frac{\mu_i \mathbf{z}_i}{1+\mu_i}. \tag{2.43}$$

$$\begin{aligned} \mathbf{s}_{ZINB}(r) = \frac{\partial l_{ZINB}}{\partial r} &= \sum_{\substack{i=1 \\ y_i=0}}^{n} \frac{r^r \left(\ln \left(\frac{r}{\lambda_i + r} \right) + \frac{\lambda_i}{\lambda_i + r} \right)}{\mu_i(\lambda_i + r)^r + r^r} \\ &\quad + \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[\sum_{j=0}^{y_i - 1} \frac{1}{j+r} + \ln r - \ln(\lambda_i + r) + \frac{\lambda_i - y_i}{\lambda_i + r} \right]. \end{aligned} \tag{2.44}$$

The maximum likelihood estimates $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and $\hat{r}$ for the ZINB model follows the multivariate normal distribution

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \\ \hat{r} \end{pmatrix} \sim \mathrm{N} \left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ r \end{pmatrix}, \begin{pmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} & \mathbf{F}_{13} \\ \mathbf{F}_{12} & \mathbf{F}_{22} & \mathbf{F}_{23} \\ \mathbf{F}_{13} & \mathbf{F}_{23} & \mathbf{F}_{33} \end{pmatrix}^{-1} \right], \tag{2.45}$$

where $\mathbf{F}$ is the expected Fisher information matrix of $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\gamma}}$ and r. The covariance matrix in Equation (2.45) is equal to the inverse of $E[\mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, r)]$, where the observed Fisher information matrix $\mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, r)$ contains the negative Hessian matrix of the log-likelihood in Equation (2.42) (Fahrmeir et al., 2013). The (m, n)th entry for each of the elements in

the covariance matrix in Equation (2.45) can thus be calculated as

$$\mathbf{F}_{11,mn} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial\beta_m\partial\beta_n}\right] = -\sum_{\substack{i=1\\y_i=0}}^{n}\frac{r^{r+2}x_{im}x_{in}\lambda_i\left[\mu_i(1-\lambda_i)(\lambda_i+r)^r+r^r\right]}{(\lambda_i+r)^2[(\mu_i(\lambda_i+r)^r+r^r)]^2},$$
$$+\sum_{\substack{i=1\\y_i>0}}^{n}\frac{r\lambda_i x_{im}x_{in}}{\lambda_i+r}, \quad m,n=1,2,\ldots,p, \tag{2.46}$$

$$\mathbf{F}_{12,mn} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial\beta_m\partial\gamma_n}\right] = -\sum_{\substack{i=1\\y_i=0}}^{n}\frac{r^{r+1}x_{im}z_{in}\lambda_i\mu_i(\lambda_i+r)^{r-1}}{(\mu_i(\lambda_i+r)^r+r^r)^2}, \tag{2.47}$$
$$m=1,2,\ldots,p, \quad n=1,2,\ldots,q,$$

$$\mathbf{F}_{13,m1} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial r\partial\beta_m}\right] = \sum_{\substack{i=1\\y_i=0}}^{n}\frac{x_{im}\lambda_i r^{r+1}\left[(\lambda_i+r)\left(C_1\left(\ln r+\frac{r+1}{r}\right)-C_2\right)-C_1\right]}{(\lambda_i+r)^2(\mu_i(\lambda_i+r)^r+r^r)^2}, \tag{2.48}$$
$$m=1,2,\ldots,p,$$

$$\mathbf{F}_{22,mn} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial\gamma_m\partial\gamma_n}\right] = -\sum_{\substack{i=1\\y_i=0}}^{n}\frac{z_{im}z_{in}\mu_i\left(\frac{r}{\lambda_i+r}\right)^r}{\left[\mu_i+\left(\frac{r}{\lambda_i+r}\right)^r\right]^2}+\sum_{i=1}^{n}\frac{z_{im}z_{in}\mu_i}{[1+\mu_i]^2}, \tag{2.49}$$
$$m,n=1,2,\ldots,q,$$

$$\mathbf{F}_{23,m1} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial r\partial\gamma_m}\right] = \sum_{\substack{i=1\\y_i=0}}^{n}\frac{z_{im}\mu_i r^r C_3}{(\lambda_i+r)^r[\mu_i(\lambda_i+r)^r+r^r]^2}, \quad m=1,2,\ldots,q \tag{2.50}$$

$$\mathbf{F}_{33} = \mathrm{E}\left[-\frac{\partial^2 l_{ZINB}}{\partial r^2}\right] = r^r\sum_{\substack{i=1\\y_i=0}}^{n}\frac{C_3C_2-\left[(\ln r+1)C_3+\frac{\lambda_i^2}{r(\lambda_i+r)^2}\right]C_1}{C_1^2} \tag{2.51}$$
$$+\sum_{\substack{i=1\\y_i>0}}^{n}\left(E\left[\sum_{j=0}^{y_i-1}\frac{1}{(j+r)^2}\right]-\frac{\lambda_i}{r(\lambda_i+r)}\right),$$

where

$$C_1 = \mu_i(\lambda_i+r)^r+r^r,$$
$$C_2 = \frac{\partial C_1}{\partial r}, = \mu_i(\lambda_i+r)^r\left(\ln(\lambda_i+r)+\frac{r}{\lambda_i+r}\right)+r^r(\ln r+1)$$
$$C_3 = \ln\left(\frac{r}{\lambda_i+r}\right)+\frac{\lambda_i}{\lambda_i+r}.$$

2.2.2 ZAP and ZANB Models

In the hurdle models all the zeros are treated equally. The zero-process models the probability of presence versus absence with a logistic model, while the count-process generates a non-zero count with a truncated Poisson (ZAP) or a truncated negative binomial (ZANB) model. Let the probability of absence ($y_i = 0$) be defined by π_i, and $g(y)$ be the probability mass function for either the Poisson distribution given in Equation (2.1) or the negative binomial distribution given in Equation (2.12). Following Zuur et al. (2009), the probability mass function for the hurdle models can be written as

$$f(y_i) = \begin{cases} \pi_i & y_i = 0 \\ (1-\pi_i)\frac{g(y_i)}{1-g(y_i=0)}, & y_i > 0. \end{cases} \tag{2.52}$$

Let the covariates $\mathbf{x}_i$ and $\mathbf{z}_i$, with the regression parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, be used in the hurdle models as well. The linear predictors $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ and $\phi_i = \mathbf{z}_i^T\boldsymbol{\gamma}$ links the covariates and the regression parameters for the count model and the zero model, respectively. The log-link function is used to link the count covariates with the expected mean and the logistic link is used for the zero-part. The expected mean for observation i in the count-part is thus calculated as $\lambda_i = \exp(\eta_i)$, and the probability of lice absence is modelled by $\pi_i = \frac{1}{1+\exp(\phi_i)}$ (Cameron & Trivedi, 1998). This gives the following odds for a zero count versus a positive count

$$\frac{\pi_i}{1-\pi_i} = \frac{\frac{1}{1+\exp(\phi_i)}}{\frac{\exp(\phi_i)}{1+\exp(\phi_i)}} = \frac{1}{\exp(\phi_i)} = (\exp(\mathbf{z}_i^T\boldsymbol{\gamma}))^{-1}. \tag{2.53}$$

By taking the logarithm of both sides, the logarithm of the odds for a zero count is $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = -\mathbf{z}_i^T\boldsymbol{\gamma}$. The logarithm of the odds for a positive count $\ln\left(\frac{1-\pi_i}{\pi_i}\right)$ is thus given by $\mathbf{z}_i^T\boldsymbol{\gamma}$.

The expected mean for ZAP and ZANB are in Zuur et al. (2009) formulated as

$$\mathrm{E}_{ZAP}(Y_i) = \frac{1-\pi_i}{1-\exp(-\lambda_i)}\lambda_i, \tag{2.54}$$

$$\mathrm{E}_{ZANB}(Y_i) = \frac{1-\pi_i}{1-P_0}\lambda_i, \quad P_0 = \left(\frac{r}{\lambda_i + r}\right)^r, \tag{2.55}$$

where λ_i are the expected Poisson count and the expected negative binomial count, respectively. Zuur et al. (2009) defines the variance for the two hurdle models as

$$\mathrm{Var}_{ZAP}(Y_i) = \frac{1-\pi_i}{1-\exp(-\lambda_i)}(\lambda_i + \lambda_i^2) - \left(\frac{1-\pi_i}{1-\exp(-\lambda_i)}\lambda_i\right)^2, \tag{2.56}$$

$$\mathrm{Var}_{ZANB}(Y_i) = \frac{1-\pi_i}{1-P_0}\left(\lambda_i^2 + \lambda_i + \frac{\lambda_i^2}{r}\right) - \left(\frac{1-\pi_i}{1-P_0}\lambda_i\right)^2. \tag{2.57}$$

By substituting the mean and the variance in Equation (2.26), the following Pearson residuals for the two hurdle models are obtained

$$r_{i,ZAP} = \frac{y_i - \frac{1-\hat{\pi}_i}{1-\exp(-\hat{\lambda}_i)}\hat{\lambda}_i}{\sqrt{\frac{1-\hat{\pi}_i}{1-\exp(-\hat{\lambda}_i)}(\hat{\lambda}_i + \hat{\lambda}_i^2) - \left(\frac{1-\hat{\pi}_i}{1-\exp(-\hat{\lambda}_i)}\hat{\lambda}_i\right)^2}}, \tag{2.58}$$

$$r_{i,ZANB} = \frac{y_i - \hat{\lambda}_i(1-\hat{\pi}_i)}{\sqrt{(1-\hat{\pi}_i)(\hat{\lambda}_i + \frac{\hat{\lambda}_i^2}{r}) + \hat{\lambda}_i^2 + (\hat{\pi}_i^2 + \hat{\pi}_i)}}. \tag{2.59}$$

Assuming that the response variables y_i are conditionally independent zero-altered Poisson distributed (or zero-altered negative binomial distributed) variables, the maximum likelihood estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and r for ZANB) can be derived by maximizing the log-likelihood function. The log-likelihood function for the ith observation from a hurdle Poisson model is calculated as $l_{i,ZAP}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \ln\left(f_{ZAP}(y_i)\right)$, which gives

$$\begin{aligned} l_{i,ZAP}(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \begin{cases} \ln(\pi_i), & y_i = 0 \\ \ln(1-\pi_i) + l_{i,P}(\boldsymbol{\beta}) - \ln(1-\exp(-\lambda_i)), & y_i > 0, \end{cases} \\ &= \begin{cases} -\ln(1+\mu_i), & y_i = 0 \\ \ln(\mu_i) - \ln(1+\mu_i) + l_{i,P}(\boldsymbol{\beta}) - \ln(1-\exp(-\lambda_i)), & y_i > 0, \end{cases} \end{aligned} \tag{2.60}$$

where the substitution $\pi_i = \frac{1}{1+\mu_i}$ with $\mu_i = \exp(\mathbf{z}_i^T\boldsymbol{\gamma})$ is used in the last transition, and $l_{i,P}(\boldsymbol{\beta})$ is the Poisson log-likelihood function for the ith observation in Equation (2.2). Similarly, let $l_{i,NB}(\boldsymbol{\beta}, r)$ be the ith observation for the negative binomial log-likelihood function in Equation (2.13). The log-likelihood function for the ith observation for the hurdle negative binomial model can then be expressed as

$$l_{i,ZANB}(\boldsymbol{\beta}, \boldsymbol{\gamma}, r) = \begin{cases} -\ln(1+\mu_i), & y_i = 0 \\ \ln(\mu_i) - \ln(1+\mu_i) + l_{i,NB}(\boldsymbol{\beta}, r) - \ln\left(1 - \left(\frac{r}{\lambda_i + r}\right)^r\right), & y_i > 0. \end{cases} \tag{2.61}$$

The score functions are obtained by taking the partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and r, by using the link functions $\lambda_i = \exp(\mathbf{x}_i^T\boldsymbol{\beta})$ and $\mu_i = \exp(\mathbf{z}_i^T\boldsymbol{\gamma})$. The score functions for the ZAP model are derived as

$$\mathbf{s}_{ZAP}(\boldsymbol{\beta}) = \frac{\partial l}{\partial \boldsymbol{\beta}} = \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[\mathbf{s}_{i,P}(\boldsymbol{\beta}) - \frac{\lambda_i \exp(-\lambda_i)\mathbf{x}_i}{1-\exp(-\lambda_i)}\right] = \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[\mathbf{s}_{i,P}(\boldsymbol{\beta}) - \frac{\lambda_i \mathbf{x}_i}{\exp(\lambda_i) - 1}\right], \tag{2.62}$$

$$\mathbf{s}_{ZAP}(\boldsymbol{\gamma}) = \frac{\partial l}{\partial \boldsymbol{\gamma}} = \sum_{\substack{i=1 \\ y_i>0}}^{n} \mathbf{z}_i - \sum_{i=1}^{n} \frac{\mu_i}{1+\mu_i}\mathbf{z}_i, \tag{2.63}$$

where $\mathbf{s}_{i,P}(\boldsymbol{\beta})$ is the ith element in the score function derived from the Poisson distribution with respect to $\boldsymbol{\beta}$, given in Equation (2.3). Similar, let $\mathbf{s}_{i,NB}(\boldsymbol{\beta})$ and $\mathbf{s}_{i,NB}(r)$ be the score functions from the negative binomial distribution, given in Equation (2.14) and Equation (2.15), respectively. The score function with respect to $\boldsymbol{\gamma}$ is the same for both the ZAP

model and the ZANB model. The following score functions are derived for the ZANB model

$$\mathbf{s}_{ZANB}(\boldsymbol{\beta}) = \frac{\partial l_{ZANB}}{\partial \boldsymbol{\beta}} = \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[\mathbf{s}_{i,NB}(\boldsymbol{\beta}) + \frac{\lambda_i r^{r+1} \mathbf{x}_i}{(\lambda_i + r)[(\lambda_i + r)^r - r^r]} \right], \tag{2.64}$$

$$\mathbf{s}_{ZANB}(\boldsymbol{\gamma}) = \frac{\partial l_{ZANB}}{\partial \boldsymbol{\gamma}} = \sum_{\substack{i=1 \\ y_i>0}}^{n} \mathbf{z}_i - \sum_{i=1}^{n} \frac{\mu_i}{1+\mu_i} \mathbf{z}_i, \tag{2.65}$$

$$\mathbf{s}_{ZANB}(r) = \frac{\partial l_{ZANB}}{\partial r} = \sum_{\substack{i=1 \\ y_i>0}}^{n} \left[\mathbf{s}_{i,NB}(r) + \frac{r^r \left(\ln\left(\frac{r}{\lambda_i + r}\right) + \frac{\lambda_i}{\lambda_i + r} \right)}{(\lambda_i + r)^r - r^r} \right]. \tag{2.66}$$

For the ZAP model, the second order derivatives does not depend on y_i, so the expected Fisher information matrix is equal to the observed Fisher information matrix, $\mathbf{F}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \mathrm{E}[\mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})] = \mathbf{H}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ (Fahrmeir et al., 2013). The negative Hessian matrix of the log-likelihood can thus be used to obtain the asymptotically distribution of the ZAP model. The zero-part and the count-part in the hurdle models are independent, such that $\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}} = \frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\beta}} = \mathbf{0}$. The asymptotically distribution of the ZAP models then simplifies to

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \end{pmatrix} \sim \mathrm{N} \left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \begin{pmatrix} -\frac{\partial^2 l_{ZAP}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} & \mathbf{0} \\ \mathbf{0} & -\frac{\partial^2 l_{ZAP}}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}} \end{pmatrix}^{-1} \right], \tag{2.67}$$

where the second order derivatives with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are derived as

$$\frac{\partial^2 l_{ZAP}}{\partial \beta_m \partial \beta_n} = \sum_{\substack{i=1 \\ y_i>0}}^{n} x_{im} x_{in} \lambda_i \left(1 - \frac{(1-\lambda_i)\exp(\lambda_i) - 1}{(\exp(\lambda_i) - 1)^2} \right), \quad m, n = 1, 2, \ldots, p \tag{2.68}$$

$$\frac{\partial^2 l_{ZAP}}{\partial \gamma_m \partial \gamma_n} = -\sum_{i=1}^{n} \frac{z_{im} z_{in} \mu_i}{(1+\mu_i)^2}, \qquad m, n = 1, 2, \ldots, q. \tag{2.69}$$

For the ZANB model, the observed and expected Fisher information matrix is not equal, and the expected value of the negative Hessian matrix of the log-likelihood needs to be calculated. As for the ZAP model, the second derivatives across the zero-part and the count-part are zero. The asymptotically result for the maximum likelihood estimated for the negative binomial hurdle model is then given as

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \\ \hat{r} \end{pmatrix} \sim \mathrm{N} \left[\begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ r \end{pmatrix}, \begin{pmatrix} E\left[-\frac{\partial^2 l_{ZANB}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}\right] & \mathbf{0} & E\left[-\frac{\partial^2 l_{ZANB}}{\partial \boldsymbol{\beta} \partial r}\right] \\ \mathbf{0} & E\left[-\frac{\partial^2 l_{ZANB}}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}}\right] & \mathbf{0} \\ E\left[-\frac{\partial^2 l_{ZANB}}{\partial r \partial \boldsymbol{\beta}}\right] & \mathbf{0} & E\left[-\frac{\partial^2 l_{ZANB}}{\partial r \partial r}\right] \end{pmatrix}^{-1} \right], \tag{2.70}$$

where

$$E\left[-\frac{\partial^2 l_{ZANB}}{\partial\beta_m\partial\beta_n}\right] = -\sum_{\substack{i=1\\y_i>0}}^{n}\frac{r^{r+2}x_{im}x_{in}\lambda_i\left[(1-\lambda_i)(\lambda_i+r)^r-r^r\right]}{(\lambda_i+r)^2[(\lambda_i+r)^r-r^r]^2},$$
$$+\sum_{\substack{i=1\\y_i>0}}^{n}\frac{r\lambda_i x_{im}x_{in}}{\lambda_i+r}, \quad m,n=1,2,\ldots,p \tag{2.71}$$

$$E\left[-\frac{\partial^2 l_{ZANB}}{\partial\gamma_m\partial\gamma_n}\right] = E\left[-\frac{\partial^2 l_{ZAP}}{\partial\gamma_m\partial\gamma_n}\right] = \sum_{i=1}^{n}\frac{z_{im}z_{in}\mu_i}{(1+\mu_i)^2}, \qquad m,n=1,2,\ldots,q. \tag{2.72}$$

$$E\left[-\frac{\partial^2 l_{ZANB}}{\partial\beta_m\partial r}\right] = -\sum_{\substack{i=1\\y_i>0}}^{n}\frac{\lambda_i r^{r+1}x_{im}(\lambda_i+r)^{r-1}\left(\ln\left(\frac{r}{\lambda_i+r}\right)+\frac{\lambda_i}{\lambda_i+r}\right)}{[(\lambda_i+r)^r-r^r]^2}$$
$$+\sum_{\substack{i=1\\y_i>0}}^{n}\frac{\lambda_i^2 r^r x_{im}}{(\lambda_i+r)^2[(\lambda_i+r)^r-r^r]}, \quad m=1,2,\ldots,p \tag{2.73}$$

$$\mathrm{E}\left[-\frac{\partial^2 l_{ZANB}}{\partial r^2}\right] = r^r\sum_{\substack{i=1\\y_i=0}}^{n}\frac{C_3C_2-\left[(\ln r+1)C_3+\frac{\lambda^2}{r(\lambda_i+r)^2}\right]C_1}{C_1^2} \tag{2.74}$$
$$+\sum_{\substack{i=1\\y_i>0}}^{n}\left(E\left[\sum_{j=0}^{y_i-1}\frac{1}{(j+r)^2}\right]-\frac{\lambda_i}{r(\lambda_i+r)}\right),$$

where

$$C_1 = (\lambda_i+r)^r - r^r,$$
$$C_2 = \frac{\partial C_1}{\partial r}, = (\lambda_i+r)^r\left(\ln(\lambda_i+r)+\frac{r}{\lambda_i+r}\right)-r^r(\ln r+1)$$
$$C_3 = \ln\left(\frac{r}{\lambda_i+r}\right)+\frac{\lambda_i}{\lambda_i+r}.$$

2.3 Indicator Variables

Qualitative variables, such as delousing treatment and operating model, have no natural scale of measurement. To account for the effect that qualitative variables may have on the response, indicator variables which includes a set of levels can be assigned. Suppose a response variable y is represented by two regression variables, x_1 and x_2. Let x_1 be quantitative and x_2 be qualitative with two levels, A and B. The indicator variable takes the values 0 and 1 to identify the classes of the regression variable x_2, i.e.

$$x_2 = \begin{cases} 0, & \text{if the observation is from level A} \\ 1, & \text{if the observation is from level B.} \end{cases}$$

The first-order model can then be written as $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, where ϵ is the error term. In this model, the two response functions for $x_2 = 0$ and $x_2 = 1$ have the same slope, β_1, but different intercepts, β_0 and $\beta_0 + \beta_2$, respectively. For both levels, the variance of the errors, ϵ, is assumed to be the same. A change from level A to level B, leads to a change of β_2 in the expected mean $\mathrm{E}[x_2]$. Thus, the expected response functions represents two parallel lines, where the parameter β_2 expresses the difference in heights between the two regression lines. If the regression lines are expected to differ in both intercept and slope, it is possible to add a cross-product between x_1 and the indicator variable x_2, i.e.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon.$$

The parameter β_2 still expresses the change in the intercept resulting from a change from level A to level B, while β_3 reflects the change in the slope associated with a change from level A to level B (Montgomery & Peck, 1983).

For a qualitative variable with a levels, $a - 1$ indicator variables, which takes the values 0 and 1, are needed to represent the variable (Montgomery & Peck, 1983). Suppose a qualitative variable with three levels, A, B and C, should be incorporated in the model. Two indicator variables x_1 and x_2 are then required,

$$\begin{aligned} &\text{if the observation is from level A: } x_1 = 0,\ x_2 = 0, \\ &\text{if the observation is from level B: } x_1 = 1,\ x_2 = 0, \\ &\text{if the observation is from level C: } x_1 = 0,\ x_2 = 1. \end{aligned}$$

When a qualitative factor is included in the generalized linear model, the first level is used as a reference and the remaining levels are interpreted relative to this level. Let x_1 and x_2 be as defined above and x_3 be a quantitative variable. Assume that the linear predictor is expected to only differ in intercept. Thus, the linear predictor can be expressed as $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$, where level A is included in the intercept, $x_{i1} = 1$ if the ith observation is from level B, $x_{i2} = 1$ if the ith observation is from level C and 0 otherwise. For the Poisson regression with the log-link function, the expected mean are linked to the covariates with $\mathrm{E}[Y_i] = \lambda_i = \exp(\eta_i) = \exp(\beta_0)\exp(\beta_1 x_{i1})\exp(\beta_2 x_{i2})\exp(\beta_3 x_{i3})$. If a observation y_i changes from level A to level B and x_3 is kept constant, the estimated expected mean λ_i would increase with a factor $\exp(\beta_1)$.

2.4 Multicollinearity

When there is close to a linear dependence between some of the explanatory variables in a regression model, the problem of multicollinearity occurs. This can lead to misleading or incorrect conclusions based on the regression model, and one must therefore be aware of possible multicollinearity during model building and variables selecting.

Consider the multiple regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{y}$ is a vector of n responses, $\mathbf{X}$ is a $(n \times p)$ matrix of the explanatory variables, $\boldsymbol{\beta}$ is a vector of p unknown constants and $\boldsymbol{\epsilon}$ is a vector of n random errors, with $\epsilon_i \sim \text{NID}(0, \sigma^2)$. The problem of multicollinearity is presented following Montgomery and Peck (1983). Assume that the explanatory variabels has been centered and scaled to unit length, and that the model thus not contain a intercept. $\mathbf{X}^T\mathbf{X}$ is then a $(p \times p)$ correlation matrix between the explanatory variables, and the correlations between the variables and the response is the p dimensional vector $\mathbf{X}^T\mathbf{y}$. Let $\mathbf{X}_j$ denote the jth column of the $\mathbf{X}$ matrix. If there is a set of constants c_1, c_2, $\ldots$, c_p, where at least one constant is not zero, and

$$\sum_{j=1}^{p} c_j \mathbf{X}_j = \mathbf{0}, \tag{2.75}$$

the vectors $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$, $\mathbf{X}_p$ are linearly dependent. If Equation (2.75) is approximately true for some subset of the columns of $\mathbf{X}$, it will be a near linear dependency in $\mathbf{X}^T\mathbf{X}$ and the problem of multicollinearity will occurs.

2.4.1 Variance Inflation Factor

There are several techniques for detecting multicollinearity, but the variance inflation factor is used in this thesis. The variance inflation factors are defined as the diagonal elements of the $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$ matrix, and are useful in detecting multicollinearity. Let C_{jj} be the jth diagonal element of $\mathbf{C}$. For a model, with only two explanatory variable x_1 and x_2, the inverse of $\mathbf{X}^T\mathbf{X}$ is

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2}, \end{bmatrix} \tag{2.76}$$

where r_{12} is the correlation between x_1 and x_2. The jth diagonal element of $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$ for models with several explanatory variables, is in Montgomery and Peck (1983) given as

$$C_{jj} = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \ldots, p, \tag{2.77}$$

where R_j^2 is the coefficient of determination from the regression of x_j on the remaining $p-1$ explanatory variables. If there is strong multicollinearity between x_j and some subset of the remaining explanatory variables, the value of R_j^2 is close to unity and C_{jj} will thus be large. The variance of the jth coefficient is $C_{jj}\sigma^2$, where the variance inflation factor, C_{jj}, measures the combined effect of the dependencies among the explanatory variables on the variance of $\hat{\beta}_j$. If a variance inflation factor (VIF) exceeds 5, it indicates multicollinearity

and the current regression coefficient is poorly estimated because of this (Montgomery & Peck, 1983).

For categorical variable, a generalised variance inflation factor (GVIF) is calculated. This gives a common GVIF value for each separate category type, such that the effect of the different explanatory variables can be identified. Let the matrix of explanatory variables be partitioned in two, $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_1$ consists of r columns of $\mathbf{X}$ belonging to the category of interest and $\mathbf{X}_2$ is formed by the remaining $s = n - r$ columns. Following O'Driscoll and Ramirez (2015), the generalised variance inflation factor can then be defined as

$$\mathrm{GVIF}([\mathbf{X}_1|\mathbf{X}_2]) = \frac{\det(\mathbf{X}_1^T\mathbf{X}_1)\det(\mathbf{X}_2^T\mathbf{X}_2)}{\det(\mathbf{X}^T\mathbf{X})}. \tag{2.78}$$

The generalised variance inflation factor can be interpreted in the same way as the variance inflation factor by using the square of $\mathrm{GVIF}^{1/2df}$, where df are the number of coefficients in the subset. Thus, if $(\mathrm{GVIF}^{1/2df})^2$ is less than 5, the factor indicates that the categorical variable is not collinear with the remaining explanatory variables (Montgomery & Peck, 1983).

2.5 Hypergeometric Distribution

To evaluate the sample size used to calculate lice numbers, a hypergeometric distribution is used. The hypergeometric distribution describes the probability of k successes in n draws, without replacement, from a finite population of size N, wherein each draw is either a success or a failure. The probability mass for a hypergeometric random variable X is in Casella and Berger (2002) given as

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \qquad k = 0, 1, \ldots, n \tag{2.79}$$

where k, N, n are as defined above and m is the total number of successes in the population of size N.

3 Dataset

3.1 Study Area

The stage model was studied in two different fjords in production area 7 (Figure 3.1), which due to confidentiality in this thesis are named Fjord 1 and Fjord 2. In Fjord 1, one site had been operated according to the stage model in three production cycles in the period 2012 to 2018. Three different sites in Fjord 2 operated in the period 2015-2016, one production cycle each with the stage model strategy. After seven to nine months at the start sites, the salmon were split into different sites along the coast in production area 7 and 8. With the exception of one production cycle in Fjord 1, which was moved to another site inside the fjord. This and later production cycles, which have been operated in one of the two fjords throughout the production cycle were used as basis of comparison. Data were also collected from some whole-generation productions at coast sites, that previously have been used as growth sites. In total, data from 9 salmon farms in inner fjord systems and 9 salmon farms along the coast from the period 2012-2021 were collected. Each of these consisted of 5 to 16 cages, with a mean of 9.28. The data were studied at cage level, and contained a total of 235 production cycles. The average duration time of the production cycles was 68 weeks. The number of generations that were operated with start

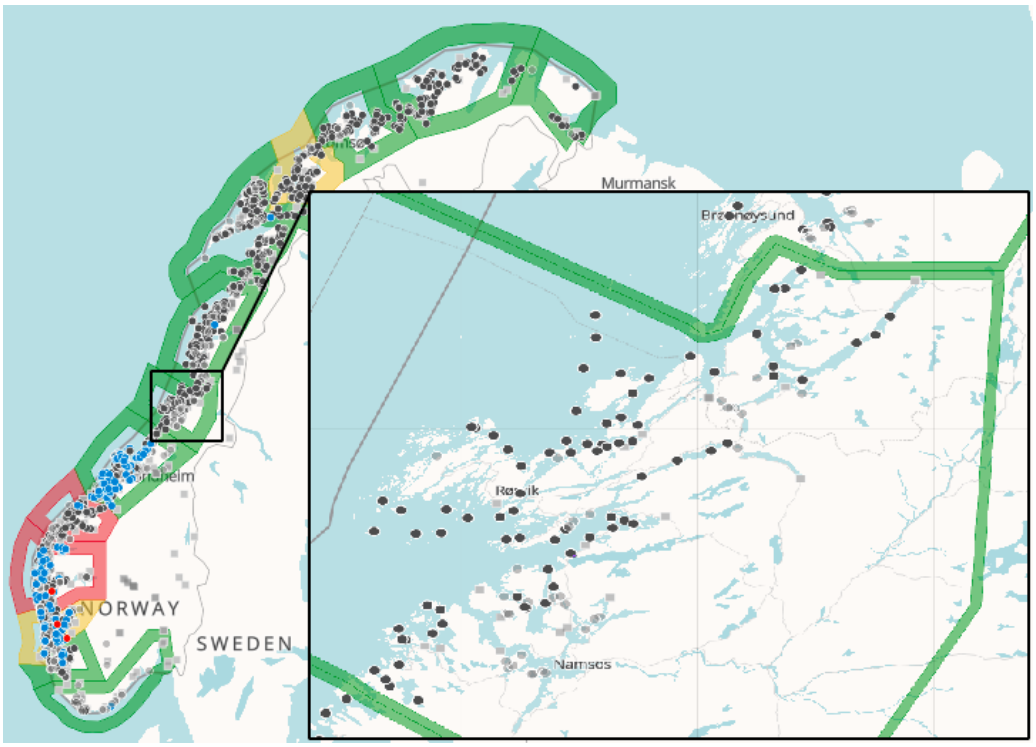

Figure 3.1: Map of the study area. The fjords are located in production area 7 (inside black square), which covers Bindal and the northern part of Trøndelag. The growth sites are located in both production area 7 and 8. All circles, independent of color, illustrates salmon farms. The green, yellow and red sections represent production areas, where production area 1 is the section furthest south and production area 13 is the northernmost section. (Modified from BarentsWatch, 2021b)

sites in the fjord and growth sites along the coast was 99.

3.2 Preparation of the Dataset

Weekly data on lice numbers, sea temperatures, number of salmon, mean weight, biomass and delousing treatments for each cage in operation in the period 2012 to 2021 were gathered from the companies producing salmon in the study area. The lice numbers were the average number of lice per salmon calculated on a sample of the salmon in the cage. Every week at least 10 salmon in each cage were inspected, but the received lice numbers were mainly based on a sample of 20 salmon. Under the weekly inspection, the number of adult female lice, mobile lice, sessile lice and C. *elongatus* were counted. The sample mean for each of these groups were reported as the lice numbers for the current week and cage.

Data for each site were sent as separate documents and the formats varied for the different sites. For sites in Fjord 1 and associated growth sites, there were separate datasets for lice numbers, temperature, biomass and delousing treatments, while lice numbers, temperature and biomass were in the same dataset for sites belonging to Fjord 2. Before these could be merged into one full dataset, a number of adjustments were therefore made. As an example of how the data has been reshaped, the R-code for the preparation of the data from the growth sites belonging to Fjord 1 is presented in Appendix C. The division of production cycles and the preparation of several explanatory variables are also included in the Appendix.

3.2.1 Weekly Reported Data

To begin with, data for sites from the same fjord and stage (start, growth or whole-generation) were merged. Let data from the start site in Fjord 1 be named dataset 1, and data from the growth sites belonging to Fjord 1 be referred to as dataset 2. Similarly for Fjord 2, dataset 3 contains data from the three start sites in Fjord 2, and dataset 4 contains data for the associated growth sites. In addition, data from the sites that were operated according to the coast model were merged and stored in dataset 5. The data from the whole-generation sites in the fjords were merged with the start cages from the same fjord (dataset 1 and dataset 3), but marked with whole-generation. Before the merge, the datasets were checked and updated so that lice numbers, sea temperature, biomass, cage numbers and delousing data existed for all rows. The data from the start sites were also analysed in Karlsen (2020), thus parts of the R-code from this project were reused when dataset 1 and 3 were updated.

In cases where lice numbers had not been registered, the latest number of lice in the same category (adult female/mobile/sessile/scottish) in that cage was used until next registration. It was mainly only periods of one week that lice numbers were missing before they were registered again. In 18 cases, the lice numbers were reported for the last week in the production cycle, but not the sea temperature. In these cases, the sea temperature was set to be the same as for the previous week in the same cage.

The delousing treatments were divided into the five methods presented in Section 1.1.3: bath treatment, oral treatment, freshwater treatment, lice flusher and thermic treatment. Repeated treatments of the same method on the same cage within 3 days were counted as one treatment. Based on this, the treatments were grouped into periods and the start date and end date were updated. The duration of the treatment period was also calculated. In cases where the treatment period extended over several weeks, as for oral treatment, all weeks were registered with ongoing delousing. In this way, the explanatory variable for treatment in the regression model covers all weeks where there has been a treatment, and not only the first week in the treatment. The total number of delousing treatments completed in the study period was 809, while the total number of weeks with ongoing treatment was 1053. Since the number of treatments applied to a small proportion of the data (6.8%), all the treatments were merged into one explanatory variable, *Treatment*, in the regression analysis.

During operations, such as delousing and slaughter, the salmon were often moved to new cages within the farm. During data recording, this was not taken into account, as the data registered the salmon that were in the cage at the time of registration and not on the basis of which cage the salmon originated. In order to follow the salmon from start to slaughter at cage level, the cage numbers had to be manually changed such that they corresponded to the movements that had taken place. For this, the companies sent timelines for each production cycle, which gave an overview of all the relocations, and the cage numbers were thus updated after these movements. These were not available for all production cycles, so biomass and the number of salmon were also used to update the cage numbers after the relocations. In cases where the salmon were split into two cages, a new cage index was made, and the week and duration of the production cycle were updated such that they matched the original cage. The cage number for the start cages, which was used in the analysis in Karlsen (2020), were already updated, but the cages at the whole-generation sites and the growth sites had to be updated.

3.2.2 Merging Start and Growth Stages

The cages at the start site and the growth site had to be linked together with a cage index before the data from the start cage and the growth cage could be merged. Two variables *Start_locality* and *Start_cage* were created, which referred to the site and cage the salmon originated. For the start sites (dataset 1 and dataset 3), these two variables were the same as the current site and cage. The start cages for the growth cages in dataset 4 were manually added in a new column in the excel documents based on received timelines over the production cycles. In dataset 2, biomass and number of salmon were used to link the growth cages with the original cages at the start site. As the data for the growth cages from Fjord 1 were spread over several sheets in excel, an overview of the corresponding start cage and site for each of the growth cages was made in an excel document. This was then added to the original dataset 2. Dataset 1 and 2, which contains the entire production cycle for all the generation deployed in Fjord 1, were merged together into dataset 6. Dataset 3 and dataset 4 was merged to dataset 7, which contains the entire production cycle for all the generations deployed in Fjord 2.

The number of weeks the salmon have been in the sea was also necessary to include as

a variable in the analysis. In order to find this, the data had to be sorted by *start_cage* and date. To comply with the ISO 8601 standard (International Organization for Standardization., 2019), the date on Thursday in the current week was stored in the variable *date*. The data were grouped by operating model and start cage, and arranged by the date. If the difference in the date were larger than seven days, a new production cycle was started. For each production cycle the start and end dates were set, in addition to the duration of the production cycle. The variable *ProductionWeek* was found for each week by subtracting the start date of the production cycle from *date*. In several cases, the growth cage contained salmon from two different start cages, and thus from two different productions. Therefore, each of the starting cages of the growth cage was handled separately during the division of the production cycles.

3.2.3 Environmental Data

By merging dataset 5, 6 and 7, all the data gathered from the companies producing salmon in the study area were collected in one common dataset, referred to as dataset 8. According to Torrissen et al. (2013), the density of salmon farms has a clear effect on the levels of sea lice at the individual sites within an area. Therefore, the distance to the nearest site and the number of sites within a radius of 10km were calculated. To get the coordinates of the salmon farms in the study area, the dataset "lice per salmon" from BarentsWatch (2021a) was downloaded for the period 2012-2021. The dataset was filtered such that only the 18 sites which are studied, were included. All the unique rows in terms of site, latitude and longitude coordinates were extracted and stored as a separate dataset, denoted dataset 9. There were several salmon farms in the study area that were not included in the analysis, so an overview of all salmon farms in production areas 7 and 8 was downloaded from Fiskehelsedirektoratet (2021), and stored in dataset 10.

The distance from the sites in dataset 9 to each of the 158 salmon farms in dataset 10, were estimated by using the functions `distm()` and `distVincentyEllipsiod()` from the `Geosphere` package (Hijmans, 2019). The function `distVincentyEllipsiod()` estimates the shortest distance between two points, and assumes that the earth is an ellipsoid (Hijmans, 2019). The salmon farms that were registered with a distance of less than 10 km from one of the sites in dataset 9, were investigated if they had been in operation during the study period. The reported lice numbers for the last 10 years for all salmon farms in Norway are stored in Barentswatch's database (BarentsWatch, 2021a). Thus, the dataset "lice per salmon" was examined for each of the salmon farms which were closer than 10km from one of the studied sites. For 3 of the sites in dataset 10, no data had been registered in the period 2012-2021. These salmon farms were thus removed from dataset 10. The smallest distance to the remaining salmon farms and the number of salmon farms within a radius of 10km were added to dataset 9 with the variables *MinDist* and *Neighbours*, respectively. The distance from the site to the coastline was also found, as this could be a factor that affected the number of lice. By using the function `getbb()` in the `R`-package `osmdata`, the coastline data for Trøndelag and Nordland were constructed (Padgham et al., 2017). To measure the distance from each site to the coastline, the `R`-function `dist2Line()` in the `Geosphere` package was used (Hijmans, 2019). The result was stored in the variable *Distance* and added to dataset 9. Finally, the environmental data in dataset 9 was added to dataset 8 by the site variable.

Salinity data were not available for all sites in the dataset, and thus could not be included as a variable in the regression analysis. Instead, an indicator variable *Location* was created, which took into account where the salmon farm was located. Salinity data were available for three sites in each of the two fjords and for three sites along the coast. These were thus studied, and it was found that there were large variations in the salinity data for the different salmon farms. Due to different salinity inside the fjords, the location indicator was divided into part A and B, based on how far into the fjord the site was located. The sites in the innermost part of the fjord were included in part A, while the sites further out in the fjord were included in part B. The average salinity at each of the locations are presented in Table 3.1. From the table, Fjord 1A and Fjord 2A stand out with lowest salinity, while Fjord 1B and Fjord 2B have relatively similar salinity as the sites along the coast.

Table 3.1: Average salinity at the different locations in the dataset

Location	Fjord 1A	Fjord 1B	Fjord 2A	Fjord 2B	Coast
Salinity mean	8.41‰	28.67‰	13.10‰	34.12‰	32.20‰

3.2.4 Response Variable

The reported lice numbers from the salmon farms were the average of lice per salmon calculated on a sample of usually 20 salmon in the cage. The total number of lice counted on twenty salmon or the estimated total number of lice in the cage were used to get the response variable as an integer. The estimated total number of lice in the cage was calculated by multiplying the number of salmon in the cage by the reported lice number. It is the adult female lice that are most important to avoid, as it is these that multiply further. Due to many registrations with zero adult female lice, a response variable that included both adult lice and the developmental stage before the lice becomes fully grown were tested. In this way, one could achieve a smaller proportion of zeros in the response variable, higher lice numbers and possibly more changes from week to week. Thus, both adult female lice and all mobile lice, which then include both preadult and adult lice regardless of sex, were tested as response variables.

The reported lice numbers of adult female lice (*AdultFemaleLice*) during the study period are presented together with the two others response variables for adult female lice in Figure 3.2. The counted number of adult female lice on 20 salmon (*CountAdultFemale*) are in the middle and the estimated total number of adult female lice in the cage (*AdultFemaleCage*) are at the bottom. The lice numbers have been studied at cage level, and there were therefore several registered lice numbers from each site for each week. The figure does not show all the registered data points, as there is a large amount of data points in a small time interval and points with the same value overlap. The data points are therefore made partially transparent to give an indication of whether there are overlapping points. The lice numbers of all mobile lice, *AllMobile*, were calculated as the sum of the reported lice numbers of adult female lice and mobile lice. The count of all mobile lice on 20 salmon (*CountAllMobile*) and the estimated total number of all mobile lice in the cage (*AllMobileCage*) are plotted together with *AllMobile* in Figure A.1 in the Appendix.

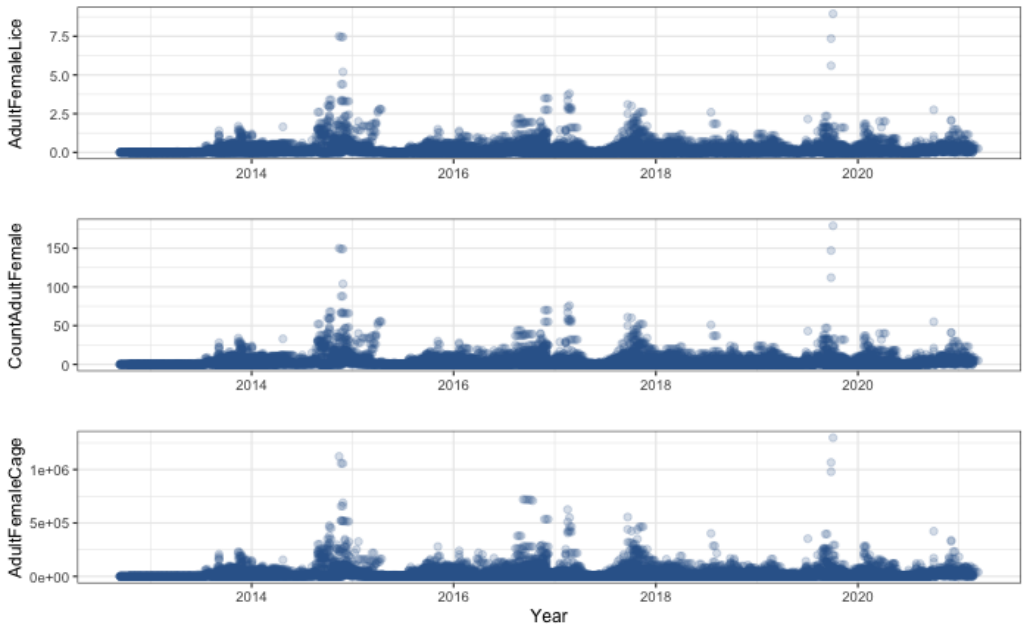

Figure 3.2: The reported lice number of adult female lice (top), the counted number of adult female lice on twenty salmon (middle) and the estimated total number of adult female lice (bottom) for each cage plotted against time. To visualize overplotting, the data points are partially transparent. The data points with the lightest blue color indicates a single data point, while a darker blue color indicates overlapping points.

3.2.5 Periods of High Lice Pressure

To check for outliers and other unexplained records before the analysis, the different variables in the dataset were plotted and compared with other variables. The term lice pressure is used to describe the lice situation, and indicates whether there are high incidences of lice or not. In some periods the lice numbers were much higher than in the rest of the data, and are in this thesis referred to as periods of high lice pressure. To avoid that this periods would affect all the explanatory variables in the regression analysis, an indicator variable for the high period was added to the regression model. It was set to 1 for periods with high lice pressure and 0 otherwise. One variable was created for the model with adult female lice as response variable and one for the model with all mobile lice as response variable, as the periods of high lice pressure were not exactly the same.

The periods were determined from the plot of the weekly reported lice numbers of adult female lice and all mobile lice, respectively, during the study period (Figure 3.3). Weeks with a reported sample mean of adult female lice larger than 3 or a sample mean of all mobile lice larger than 10 were considered to have high lice pressure, and were marked as high periods in Figure 3.3. The high periods for adult female lice and all mobile lice corresponded well, with the exception of a small period at the end of 2017, when there was only high lice pressure of adult female lice and not all mobile lice. The periods of high lice pressure were also a few weeks longer for all mobile lice than for only adult female lice.

The time plot of the reported lice numbers of adult female lice in Figure 3.2, indicated

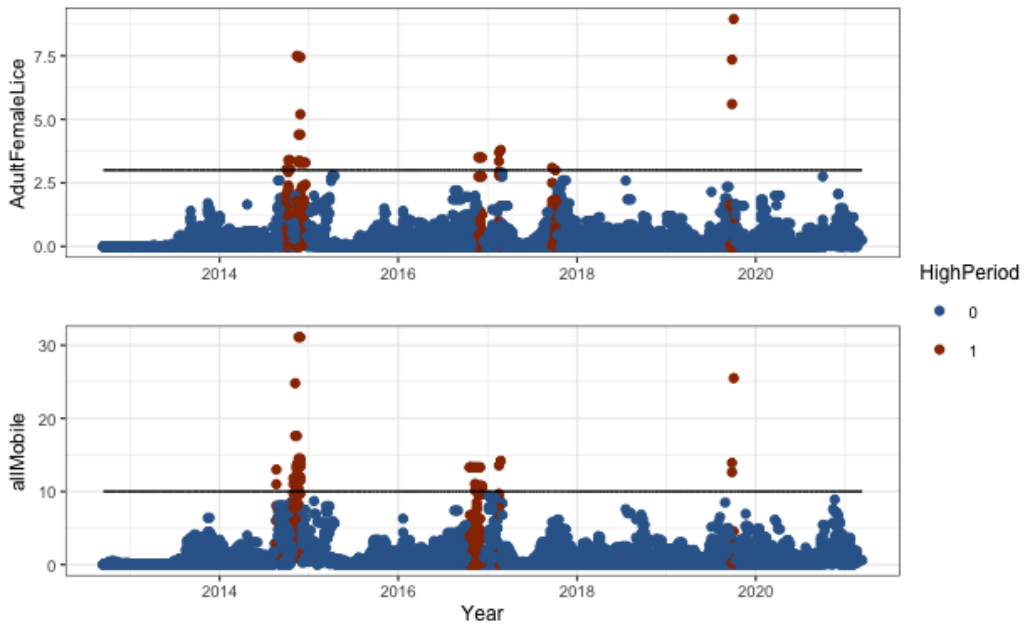

Figure 3.3: Weekly reported lice number of adult female lice (top) and all mobile lice (bottom) in each cage during the study period. The red data points represent data from the periods with high lice pressure and the blue data points the rest of the data.

that there was a small proportion of the data that had lice number of adult female lice above 3. By marking the weeks which included lice numbers above this limit as a high period, cages with normal lice numbers these weeks were also marked. To avoid this, one could have marked only those cages with abnormally high lice numbers as *HighPeriod*. The reason for these high lice numbers was unknown, so it could be that all the cages were affected, but that the composition of other factors had limited the number of lice. There were fewer registrations with zero mobile lice at the turn of the year 2014/2015, than for the rest of the study period. This supported the assumption that all cages were affected, and all lice numbers in these high periods were thus marked with high lice pressure.

3.2.6 Missing Data

The dataset lacked 5 weeks for 16 cages operated after the fjord model. This applied to production week $55 - 61$ in year 2020. These were not included in the analysis, but were included with empty values when the duration of the production cycles was determined. For the cages operated after the coast model, no lice numbers were received for C. *elongatus*. This affected the visualization of C. *elongatus* in Section 4, but the analysis was not affected by this, as it was a regression model for the salmon lice that was fitted.

3.3 The Full Dataset

The full dataset contained weekly data for each cage in operation at the studied sites since 2012 (2015 for Fjord 2). The variables used in the analysis with explanation and units are presented in Table 3.2.

Table 3.2: Variables used in the analysis with explanation and units

Variable name	**Explanation of variable**
AdultFemaleLice	Lice number of adult female lice - Reported average of adult female lice calculated from a sample of at least 10 salmon
MobileLice	Lice number of mobile lice - Reported average of mobile lice calculated from a sample of at least 10 salmon
AllMobile	Lice number of all mobile lice - The sum of AdultFemaleLice and MobileLice
SessileLice	Lice number of sessile lice - Reported average of sessile lice calculated from a sample of at least 10 salmon
C.elongatus	Lice number of *Caligus elongatus* - Reported average of *Caligus elongatus* calculated from a sample of at least 10 salmon
CountAdultFemale	The count of adult female lice on a sample of 20 salmon (Calculated as AdultFemaleLice $\cdot$ 20)
CountAllMobile	The count of all mobile lice on a sample of 20 salmon (Calculated as AllMobile $\cdot$ 20)
AdultFemaleCage	The estimated total number of adult female lice in the cage (Calculated as AdultFemaleLice $\cdot$ NumberOfFish)
AllMobileCage	The estimated total number of all mobile lice in the cage (Calculated as AllMobile $\cdot$ NumberOfFish)
Year	Year of the observation, 2012-2021
Week	Week number, 1 to 52 (53 in 2015 & 2020)
SeaTemperature	Sea temperature, °C
Salinity	Salinity at the site, ‰
Weight	Average weight of the salmon, g
Biomass	Biomass in the cage, metric ton
NumberOfFish	Number of salmon in thousand
OperatingModel	Operating method, coded as an indicator variable: StageModel, FjordModel, CoastModel
Stage	Stage in production, coded as an indicator variable: start, growth, normal (whole-generation),
Location	The location of the salmon site, coded as an indicator variable: Fjord1A, Fjord1B, Fjord2A, Fjord2B, Coast
LastWeek	Last week's reported lice number of all mobile lice (both adult female lice and mobile lice)
Method	Delousing method used, divided into bath, oral, freshwater, thermic and lice flusher
Treatment	Delousing indicator variable, coded as: 0 - no treatment, 1 - ongoing treatment
ProductionWeek	Number of weeks since the smolts were deployed, 0-89
Distance	Distance from the salmon farm to the coastline
MinDist	Distance to the closest salmon farm
Neighbours	Number of salmon farms within a radius of 10km
HighPeriod	Periods with high lice pressure (defined in Sec 3.2.5), coded as an indicator variable: 0 - normal, 1 - high lice pressure

4 Visualization of the Data

In this section, a presentation of the data is given in the form of summary statistics and visualizations. To get most information from the summary statistics of the lice numbers, the zeroes have been disregarded when the summary has been retrieved. The summary statistics for lice abundance and several other quantities are presented in Table 4.1. The percent of zeros in each lice group are given in parentheses.

Table 4.1: Summary statistics for various quantities in the study period.

Variable	Mean	1st Qu.	Median	3rd Qu.	Min.	Max.
AdultFemaleCage* (50.70%)	43805	8488	23365	52156	169	1298430
AllMobileCage* (30.64%)	128918	14716	48966	145644	874	4418843
AdultFemaleLice* (50.70%)	0.32	0.07	0.20	0.40	0.007	8.95
MobileLice* (39.9%)	0.76	0.05	0.3	0.88	0.007	23.7
SessileLice* (68.8%)	0.39	0.05	0.15	0.40	0.005	23.3
C.elongatus* (80.9%)	0.11	0.05	0.10	0.15	0.009	0.95
NumberOfFish (in 1000)	138.3	113.9	142.7	161.5	0.564	372.5
Weight (gram)	1727	523	1153	2723	68.8	7291
Biomass (metric ton)	231.9	68.7	146.7	368.0	0.349	1075
SeaTemperature (°C)	8.2	6.0	7.8	10.4	3.0	16.1

*: Without zero counts, the percent of zeros are given in parentheses

The total number of salmon and the estimated total number of salmon lice in the study area, both in millions, are plotted against time in Figure 4.1. Both the estimated number of adult female lice and all mobile lice (pre-adult and adult lice) in the cage are plotted. The band above the x-axis shows the number of cages treated against salmon lice, where darker colors indicates that a higher number of cages were treated and black equals 15-19 cages treated. In Figure A.2 in the Appendix, the total number of salmon lice in the study area grouped into adult female lice, mobile lice and sessile lice are presented. The total number of salmon treated against salmon lice during the study period is presented in Figure A.3 in the Appendix.

The amount of salmon varied throughout the study period and reached a small peak each year. The time of year the smolts were deployed varied for the various productions in the dataset, but most productions were started in the spring. During the study period, there were always some cages in the data set that were in operation and contained salmon. From Figure 4.1, the salmon production was highest for the period 2016-2018, and it was for this period that most data had been received. The periods with high lice pressure presented in Figure 3.3 corresponds to the high peaks with all mobile lice in year 2015 and 2016/2017 in Figure 4.1. The fact that both adult female lice and all mobile lice fluctuated equally throughout the study period confirmed that both a regression model for adult female lice and for all mobile lice could be used to model the lice pressure in the study area.

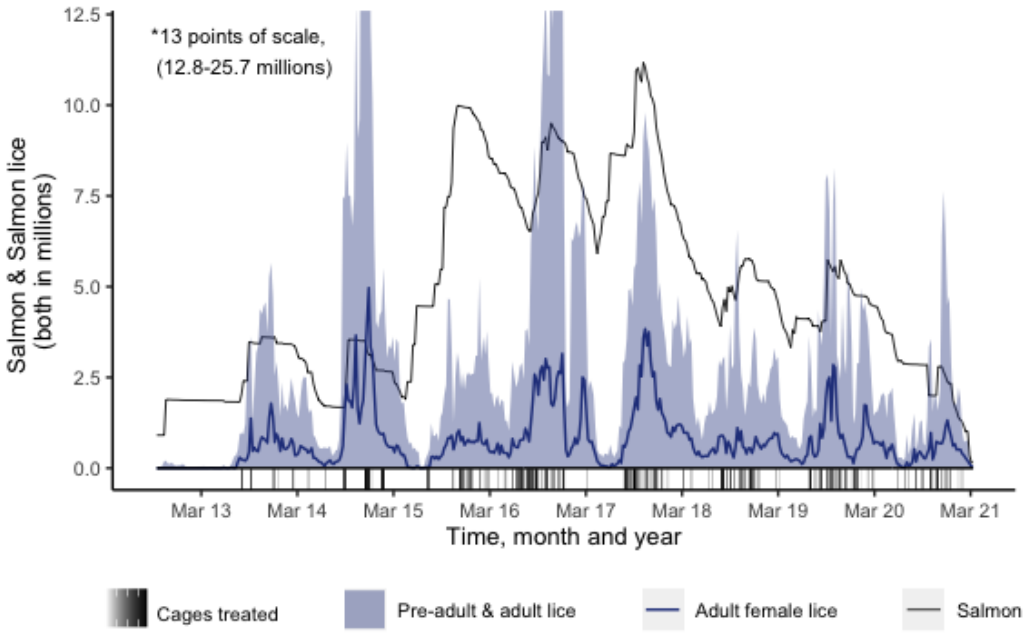

*Figure 4.1: Total number of salmon and salmon lice in the study area during the period 2012-2021. The total numbers are found by summarising over all cages in the study area. The number of salmon lice in each cage is estimated as the reported lice number multiplied with the number of salmon in each cage. The band above the x-axis shows the number of cages treated against salmon lice, where white indicates that zero cages have been treated and black indicates that 15-19 cages have been treated. *The 13 points above the scale are; 13, 17, 22, 26, 24, 24, 13, 15, 14, 18, 16, 14 and 16 millions pre-adult and adult lice in year/week 2014/41, 2014/44 - 2014/48, 2016/39, 2016/42 and 2016/44 - 2016/48, respectively.*

The weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation have each been plotted against the explanatory variables used in the analysis. The reported lice numbers of mobile lice and sessile lice were not included in the visualization, as they were not used in the analysis and it was expected that they had a relatively similar trend as adult female lice and all mobile lice. All the different development stages of salmon lice were visualized in Karlsen (2020), but there were no remarkable differences in the trend for the different stages of the salmon lice. The visualization of C. *elongatus* was included as this behaved differently from the salmon lice. The data were grouped after the indicator variable *HighPeriod*, so that the effect of the various variables both with and without the period of high lice pressure became visible.

The variables *AdultFemaleLice* and *allMobile* included 15596 observations each, which led to an overlap of the data points when plotted against the various explanatory variables. In order to be able to distinguish between overlapping points and individual points, the data points were made partially transparent. This only differed between the degree of overlap where there were fewer than 4 overlapping points, and for most points there was a greater degree of overlap. To get a clearer indication of how the lice numbers were related to the various explanatory variables, the explanatory variables were divided into intervals and an average for each of these interval was calculated. The average lice number, for

each of the explanatory variables, is plotted as a line together with the lice numbers below 1 adult female lice and 2 mobile lice (including adult female lice) in Appendix A.

In Figure 4.2, the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* are plotted against the explanatory variable *SeaTemperature*. Neither the salmon lice nor the C. *elongatus* seemed to be affected of the sea temperature largely, but the reported lice numbers were a bit smaller for sea temperatures below $4°C$ and above $14°C$. The average of *AdultFemaleLice* and *allMobile* for each degree Celsius is plotted together with the lice numbers in Figure A.4 in the Appendix. There were no major differences in the average for the different sea temperatures.

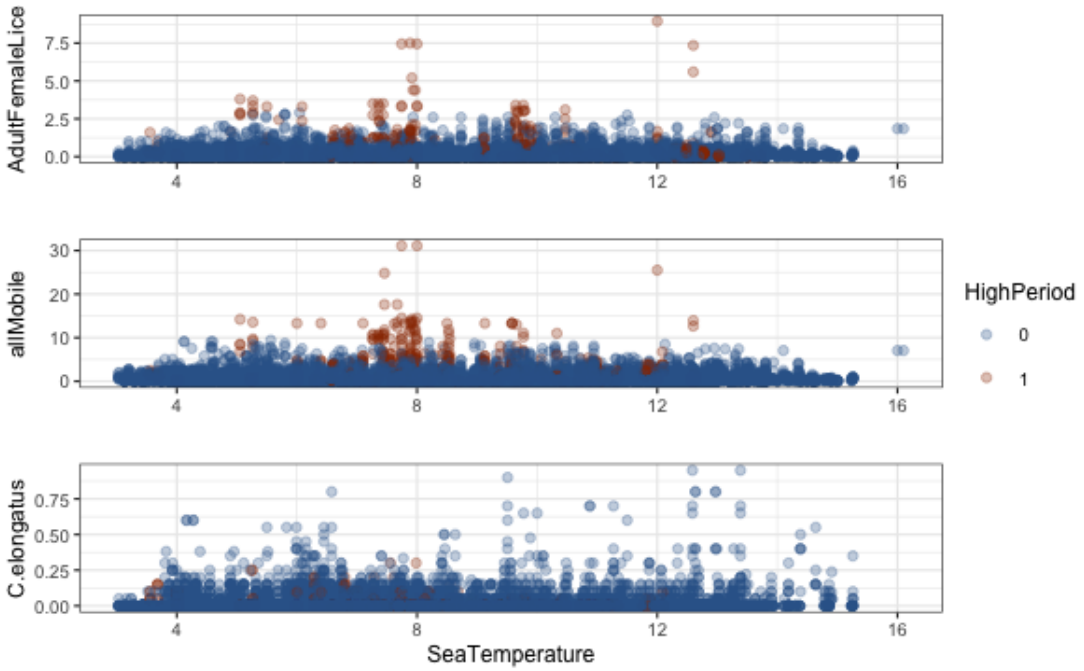

Figure 4.2: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the sea temperature (°C). The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

In Figure 4.3, box-plots of the explanatory variable *Location* and the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* are presented. The salmon lice numbers seemed to be least for Fjord 1A and Fjord 2A, which is the inner part of the two fjords. For the period with normal lice pressure (HighPeriod=0), there were no clear differences in the lice numbers between coast sites and the sites in the outer parts of the fjords (Fjord 1B and Fjord 2B). For the period with high lice pressure, the highest lice numbers were observed at the coast and in Fjord 1B. There were no salmon production in Fjord 1A during the periods with high lice pressure, so Fjord 1A is therefore not represented for HighPeriod=1. The average lice number of adult female lice and all mobile lice for coast sites and Fjord 1A were, according to Figure A.5 in the Appendix, highest. Most of the data from Fjord 1A were not zero, so the average became high even though no particularly high lice numbers were registered. The median of the lice numbers in Fjord 2A was zero, and the median for Fjord 1B and 2B were close to zero.

The main part of the lice numbers from these locations were low, but both Fjord 1B and 2B had cases where the lice number deviated strongly from the average. From Figure 4.3, it seemed to be most C. *elongatus* in Fjord 1B and along the coast. Note that the lice numbers of C. *elongatus* were below 1 lice per salmon in the entire data set, while the maximum lice number of adult female lice was 9 lice per salmon. Thus, the C. *elongatus* did not seem to be as big a problem as the salmon lice.

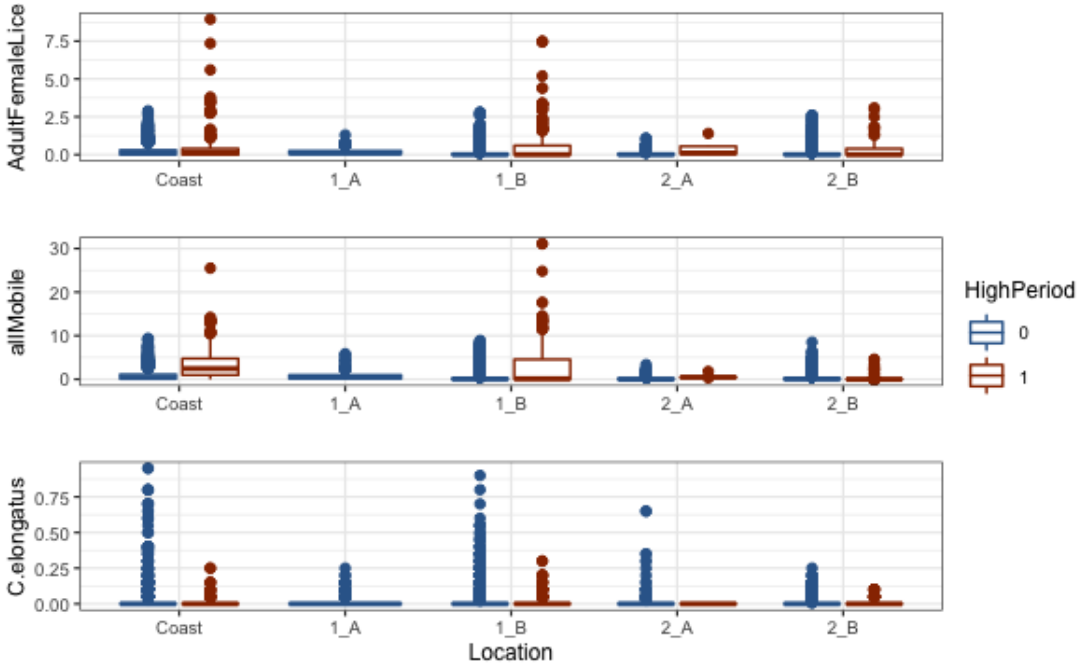

Figure 4.3: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the explanatory variable Location in a box-plot. The explanatory variable differs between whether the site was located along the coast or where in Fjord 1 or Fjord 2 it was located. Part A is the innermost part of the fjords where the salinity is lowest, and part B is the outer part of the fjords. The red data points represent data from the period with high lice pressure and the blue points the rest of the data.

In Figure 4.4, box-plots of the explanatory variable *OperatingModel* and the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* are presented. When focusing on the period with normal lice pressure, there were no clear differences in the lice numbers between the various operating models. The inner quartile range for all appeared to be approximately 0 and all operating methods had cases of lice up to 3 adult female lice per salmon and 10 mobile lice per salmon, respectively. In the high period, there were a little more differences between the various operating methods. The lice numbers for the coast model were most often low, but some lice numbers were very high. The stage model and the fjord model had several cases with lice, but the stage model did not have as high lice numbers as the highest registered numbers for the coast model. Lice numbers for C. *elongatus* had not been received for cages operated after the coast model, and is therefore not represented in the figure. The average lice number of adult female lice and all mobile lice for each of the operating models is presented in Figure A.6 in the Appendix. The average was highest for the coast model and lowest for the fjord model.

The medians of *AdultFemaleLice* for the stage model and the fjord model were zero. This indicated that there were normally least lice on the salmon that were operated in inner fjord systems throughout production.

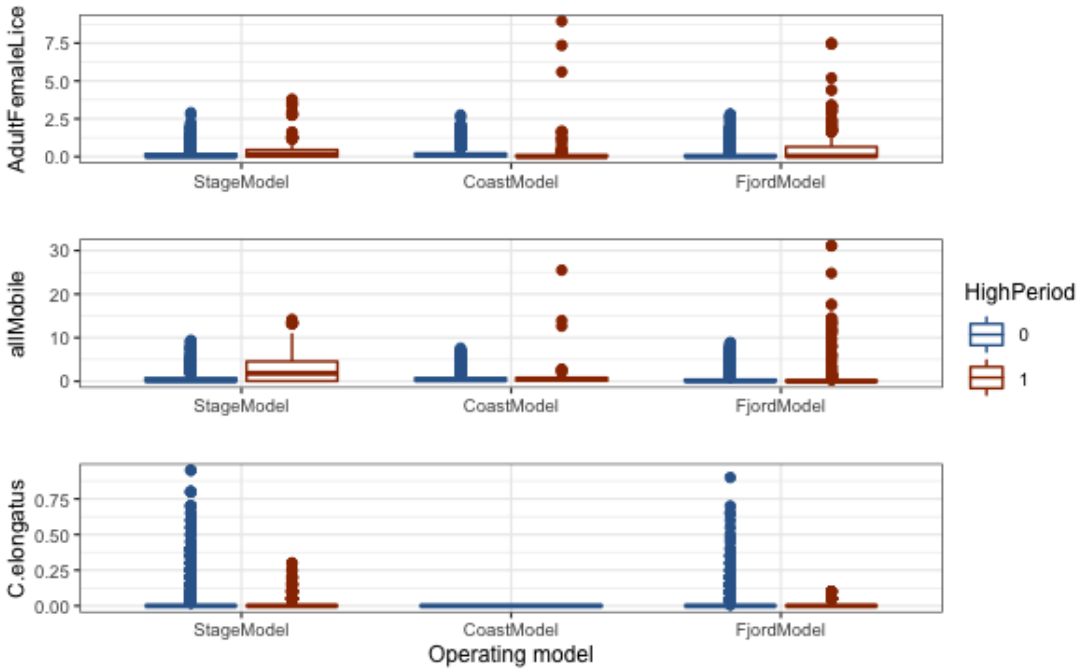

Figure 4.4: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the explanatory variable Operating-Model in a box-plot. The red data points represent data from the period with high lice pressure and the blue points the rest of the data.

In Figure 4.5, the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *Week*. There appeared to be less lice during a period in the summer and greatest lice pressure in the fall. The average lice number of adult female lice and all mobile lice for each week are presented in Figure A.7 in the Appendix. The average shows the same trend as Figure 4.5, and the year could be divided into three parts according to the lice numbers. The lice pressure was lowest in the middle of the year, from around week 14 to week 32. In late summer and autumn, the lice pressure was highest, before it decreased when winter came. According to Figure 4.5, there were no major changes in lice number for the C. *elongatus* during the study period, but it seemed to fluctuate slightly throughout the year.

In Figure 4.6, the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *NumberOfFish*. The number of salmon in a cage decreased slightly each week, and there were usually no major changes in the number of salmon during a production. From Table 4.1, the average number of salmon in a cage was 138.3 thousand, and the 1st and 3rd quartile were 113.9 and 161.5 thousand, respectively. The number of salmon lice tended to increase up to around 150 thousand salmon, before it decreased with increasing number of salmon. In some cases, there were cages with a much higher number of salmon

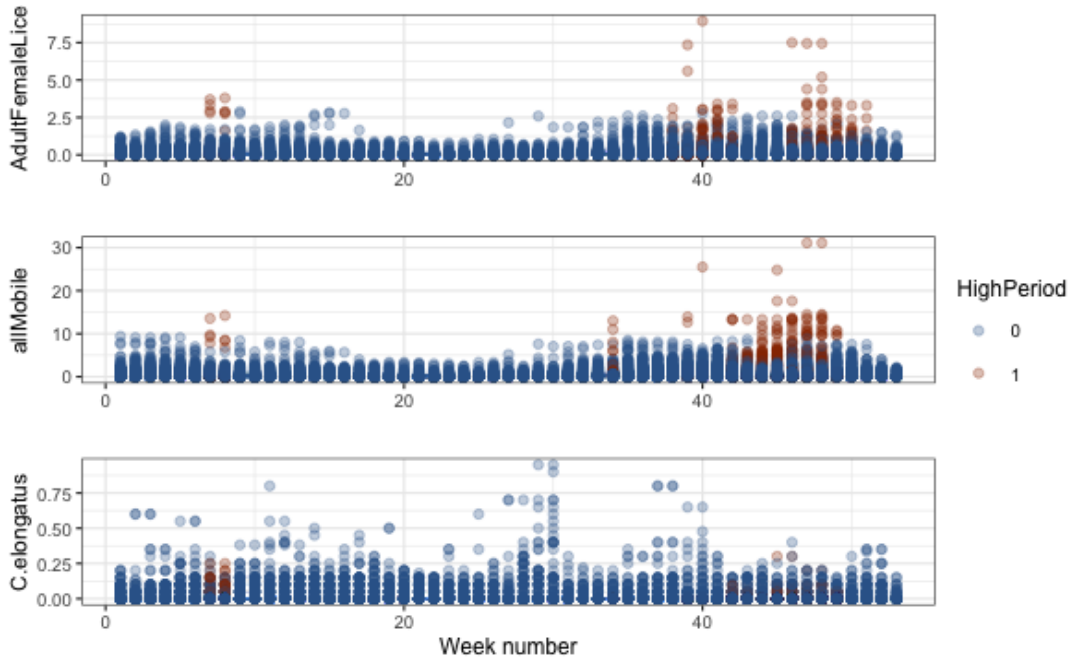

Figure 4.5: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the week number. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

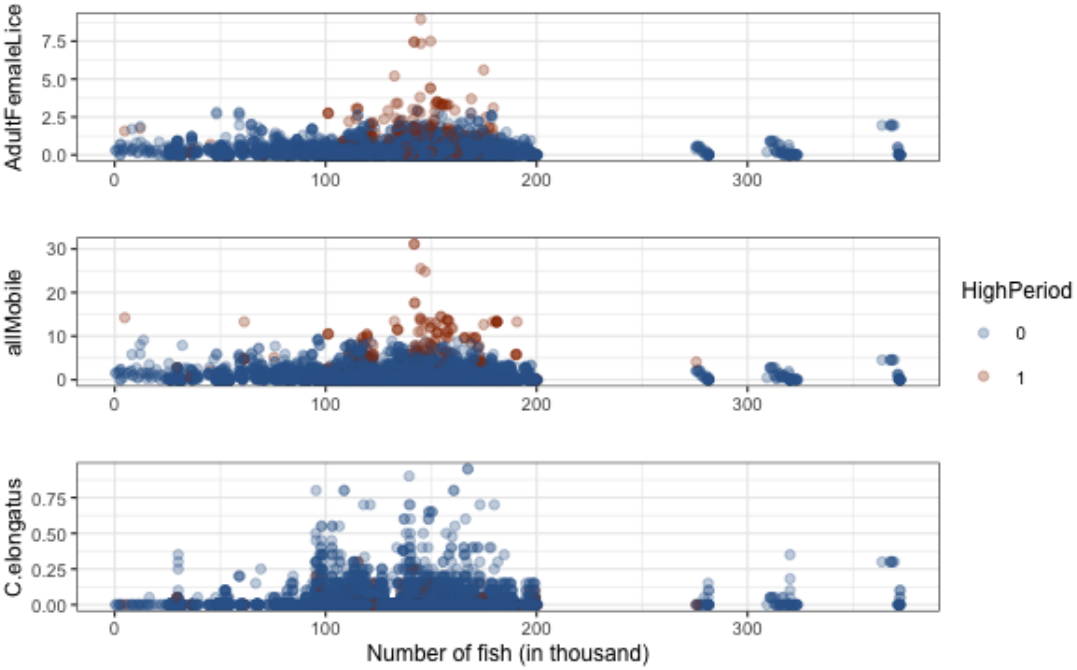

Figure 4.6: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the number of salmon in the cage, given in thousand. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

than usual, and for these the lice numbers seemed to increase with increasing numbers of salmon. Quite similar to the salmon lice, the C. *elongatus* numbers were highest for 100-170 thousand salmon, which covers the deployed amount of salmon in most of the cages in the dataset.

The number of salmon was divided into intervals of 21 thousand, and an average lice number for each of the intervals were calculated and presented in Figure A.8. The average curve indicated a different relationship than assumed from Figure 4.6, as it appeared that the lice number decreased with the number of salmon. The cages studied usually contained between 100 and 200 thousand salmon, but under deploying, transport and slaughtering there could be large changes in the number of salmon in a cage. These were only temporary changes that last until all the salmon had been moved or slaughtered, and only applied to a small part of the data. There were therefore not enough data to make a representative average for the cages with less than 80 thousand salmon or for cages with over 200 thousand salmon. Due to this and the low change in number of salmon in the cage throughout a production, *NumberOfFish* was not a preferred explanatory variable.

In Figure 4.7, the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *Weight*. The salmon lice numbers tended to increase with increasing weight up to 3kg and then decreases with increasing weight. It was a peak in the beginning, which seemed to be from one of the periods of high lice pressure, and was thus not necessarily common for the majority. C. *elongatus* behaves differently than salmon lice, and for this the highest

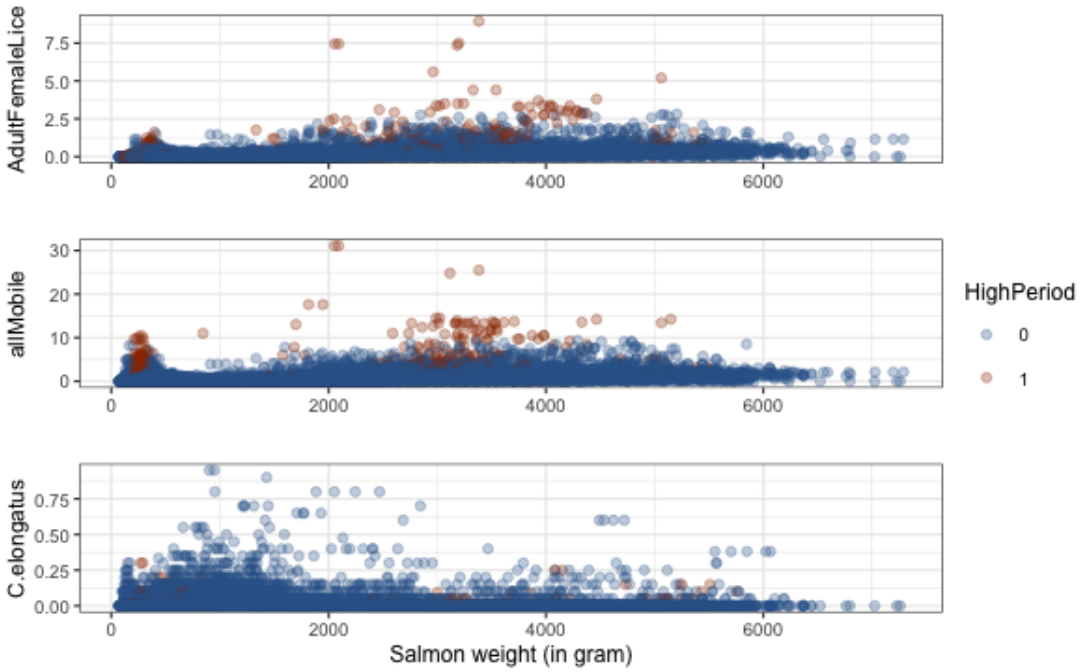

Figure 4.7: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the average weight of the salmon in the cage, given in gram. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

lice numbers had been registered for salmon below 2 kg. The average lice number of adult female lice and all mobile lice calculated for every 200 gram are presented in Figure A.9 in the Appendix. The average seemed to increase with increasing weight, as indicated by Figure 4.7, but after the highest average value were reached, around 3500g, the average did not decrease but fluctuate. It also seemed to be fewer cases with absence of salmon lice on salmon larger than 3kg.

In Figure 4.8, the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *Biomass*. The biomass are the total weight of salmon in the cage, and was calculated as *NumberOfFish* multiplied with *Weight*. The salmon lice number seemed to increase with increasing biomass up to 500 metric ton, and for biomass above 500 metric ton there were fewer cases of zero adult female lice. Thus, the lower limit of the salmon lice seemed to increase with increasing biomass. The biomass were divided into intervals of 20 metric ton, and the average of adult female lice and all mobile lice for each of the interval were calculated and plotted in Figure A.10 in the Appendix. The average increased with increasing biomass up to 500 metric ton and then fluctuated before it increased for the largest biomasses.

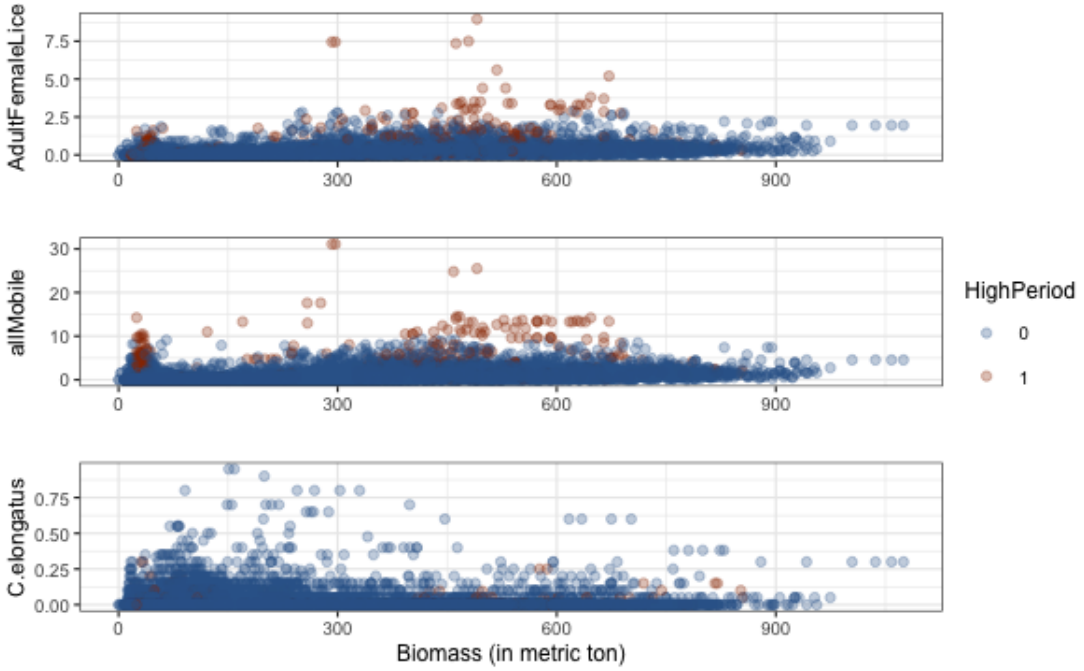

Figure 4.8: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the biomass of the cage, given in metric ton. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

In Figure 4.9, the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *ProductionWeek*. The figure indicated the same as mentioned for Figure 4.7, the salmon lice numbers had a peak in the beginning, and the lice numbers increased with increasing

weeks up to around 60 weeks. It was expected that the salmon weight and the number of weeks since deployment correlate positively, since the salmon weight increases with time. For the C. *elongatus*, the lice numbers seemed to peak after around 25 weeks in production and then, with some exceptions, decrease again. The average lice number of adult female lice and all mobile lice for each production week are plotted with the lice numbers in Figure A.11 in the Appendix. The average values of the lice numbers were low until they increased rapidly in the middle of the production, from around production week 40 to production week 55. After this, the average of adult female lice fluctuated around 0.4 adult female lice per salmon.

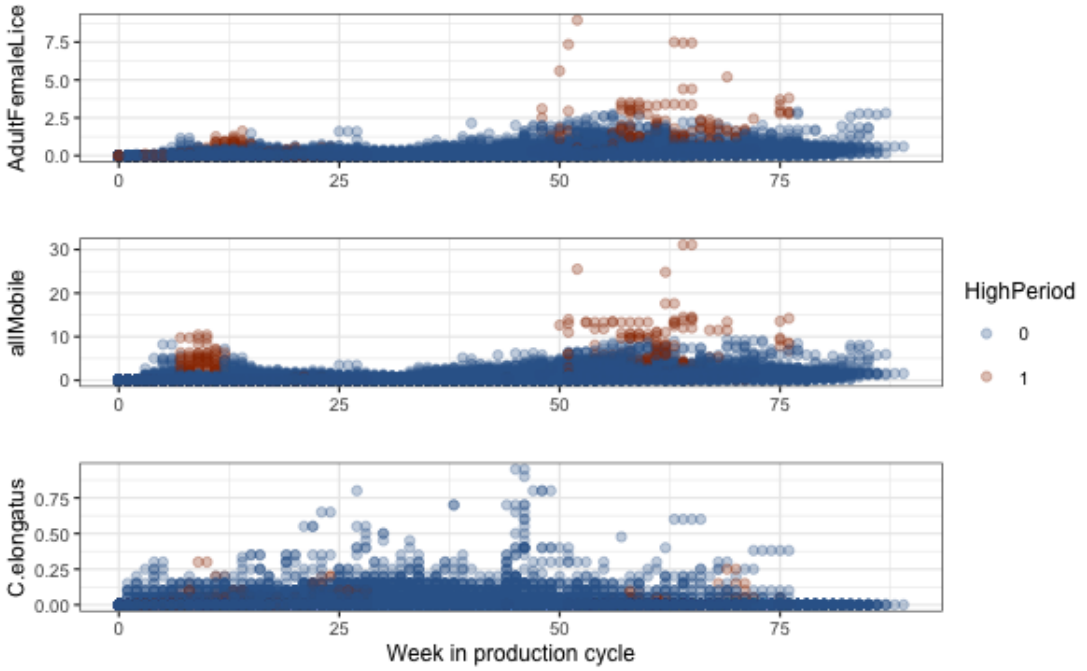

Figure 4.9: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against number of weeks since deployment. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

In Figure 4.10, box plots of the explanatory variable *Treatment* and the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are presented. The indicator variable *Treatment* is 1 if a delousing method has been used and 0 otherwise. One interesting result was that *AdultFemaleLice* and *allMobile* were highest for no treatment (Treatment0), especially for the high period. The lice numbers that were registered the week the treatment started could have been counted after the end of the treatment, and the lice numbers could thus have been reduced as a result of the delousing. In such situations, it is natural that the lice numbers were higher in the week before the delousing treatment. There were no clear differences in ongoing treatment or none treatment for C. *elongatus*. The average lice number of adult female lice and all mobile lice for no treatment and ongoing treatment are presented in Figure A.12 in the Appendix. The average was highest for ongoing delousing treatment (Treatment1), which is natural since there are cages with high lice numbers that tend to be treated

against salmon lice. It is not common to delouse a cage with zero lice, so most of the lice numbers for ongoing delousing treatment were greater than zero. The average for each of the delousing methods are also presented in Figure A.12 in the Appendix. The bath treatment appeared to be the delousing methods that had been used for cages with highest lice pressure, while oral treatment had been used for cages with lowest lice pressure. This is also related to when the cages has been deloused, as before 2016 only bath and oral treatment were used to delouse the studied cages.

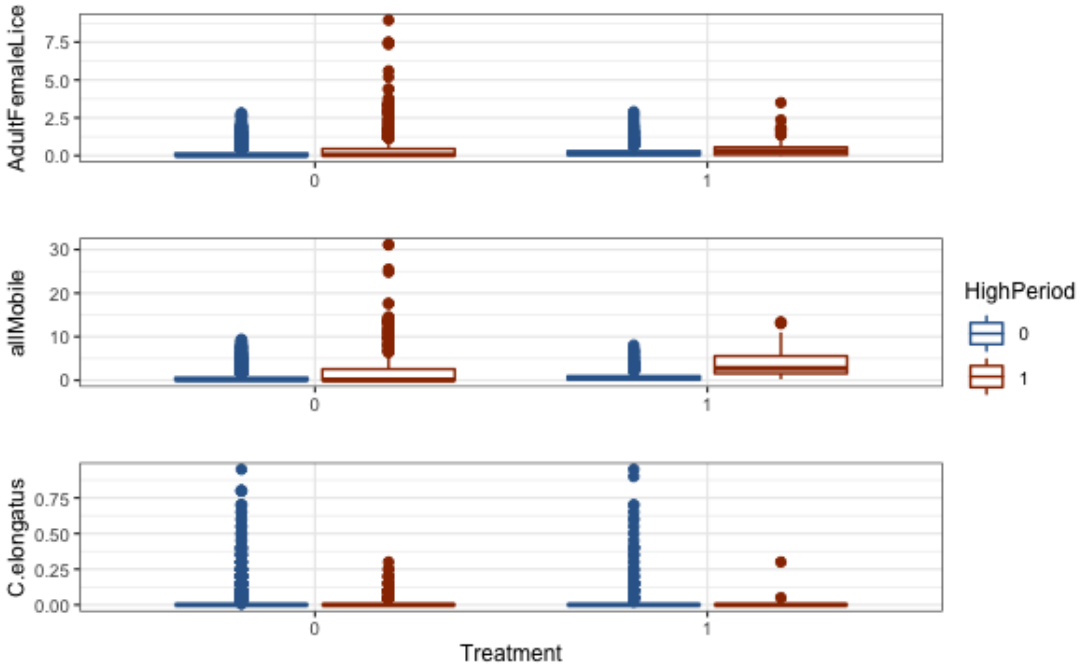

Figure 4.10: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the delousing indicator variable in a box-plot. Treatment0 indicates that the cage not had been deloused the current week, while Treatment1 indicates that there was an ongoing treatment. The red data points represent data from the period with high lice pressure and the blue point the rest of the data.

In Figure 4.11, the weekly reported lice numbers of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *Distance*. With the exception of one site, the distance from the sites to the coastline was less than 300m. The salmon lice number tended to decrease with increasing distance for the sites below 300m. For the site 1650m from the coastline, the lice number increased a bit compared to the lice numbers from the sites 300m from the coast. There were no linear trend between the lice numbers of C. *elongatus* and the distance to the coastline. The average lice number of adult female lice and all mobile lice for every 150 meter from the coastline, plotted in Figure A.13 in the Appendix, indicated the same as mentioned above. The lice number decreased with increasing distance from the coastline, but increased somewhat again for the salmon farm far from the coastline.

In Figure 4.12, the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *MinDist*, which contains the distance to the nearest salmon farm in meters. The smallest

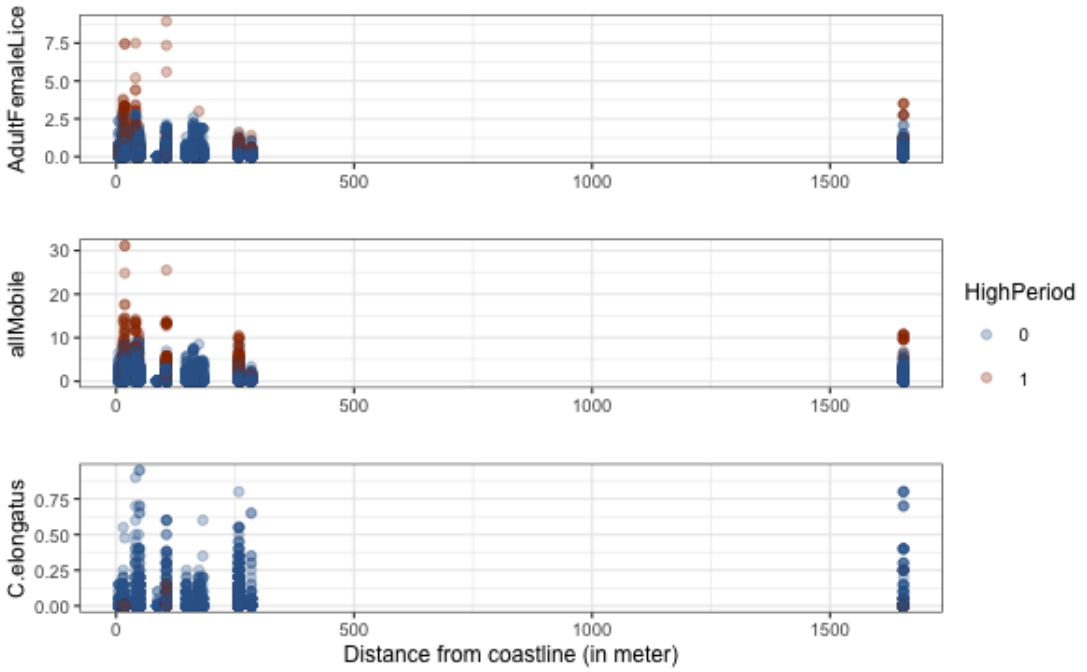

Figure 4.11: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the distance from the salmon farm to the coastline, given in meter. The red data points represent data from the period with high lice pressure and the blue points the rest of the data.

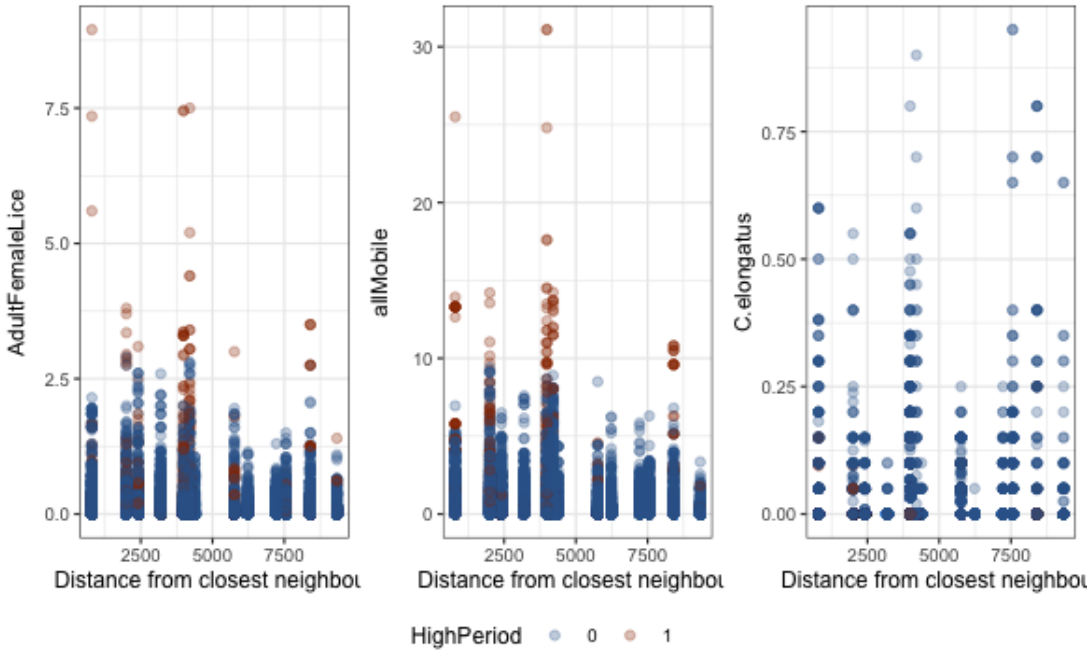

Figure 4.12: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the distance to the nearest salmon farm, given in meter. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

distance to the nearest salmon farm for the studied sites was 794 meter and the longest distance was 9334 meter. There were no clear trend between the distance and the salmon lice, but it seemed like the lice pressure was highest for sites with neighbours within 5km. For the C. *elongatus*, it did not appear that the lice number was affected by the distance to the nearest salmon farm. The average lice number of adult female lice and all mobile lice calculated for every half kilometer are presented in Figure A.14. The average was highest for the salmon farms with a close neighbour, but there were no clear trend between the average of salmon lice and the distance to the nearest neighbour.

In Figure 4.13, the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are each plotted against the explanatory variable *Neighbours*. This variable contained the number of neighbours the site had within a radius of 10km, and the number of neighbours to the sites varied from 1 to 4 neighbours. The figure indicates that the number of lice, both salmon lice and C. *elongatus*, increased with the number of neighbours. The average lice number of adult female lice and all mobile lice for 1,2,3 and 4 neighbours, respectively, are presented in Figure A.15 in the Appendix. There were no major differences in the average for sites with one, two and three neighbours, but for sites with four neighbours the average was twice as high as for the others.

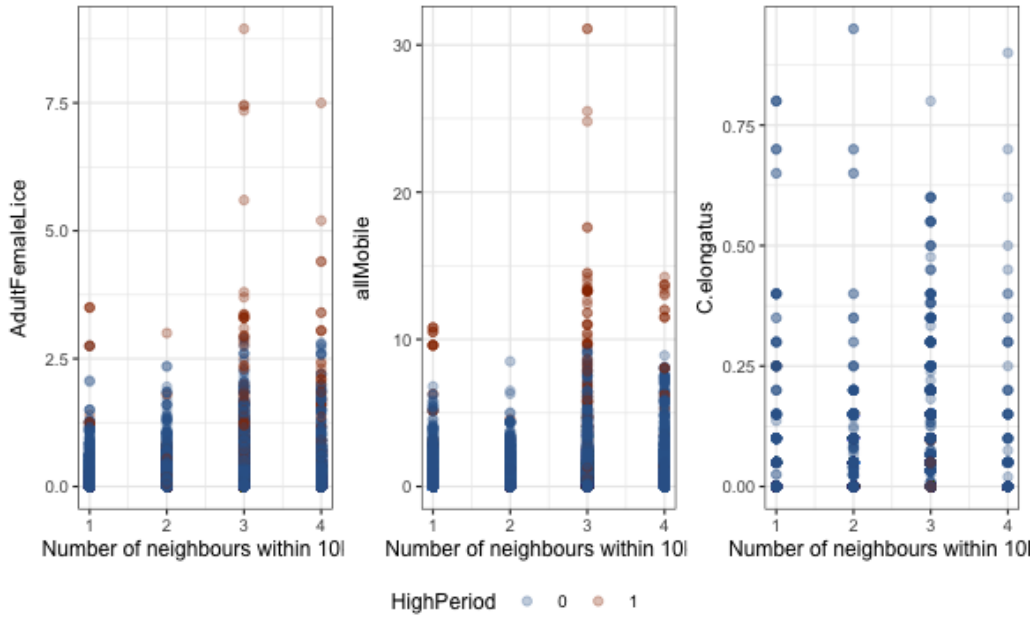

Figure 4.13: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the number of salmon farms within a radius of 10km. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

In Figure 4.14, the weekly reported lice number of adult female lice, all mobile lice and C. *elongatus* for each cage in operation are plotted against the explanatory variable *LastWeek*. The last weeks reported lice number of all mobile lice, which included both adult female lice and mobile lice, was used since the number of adult female lice depends on the survival

of adult female lice and the number of mobile lice which has become adult female from last week. There were a positive correlation between the salmon lice numbers and the last weeks reported lice numbers, while the lice numbers of *C.elongatus* seemed to be highest when none mobile lice had been observed last week. The lice numbers for the C. *elongatus* decreased for increasing lice numbers last week. As presented in Figure 1.1, only half of the mobile lice are assumed to become adult female lice, while the other half becomes adult male lice. The correlation between the adult female lice and the last weeks reported number of all mobile lice was therefore a bit lower than for all mobile lice. The average lice number of adult female lice and all mobile lice for each of last weeks reported lice number with one significant digit is plotted in Figure A.16 in the Appendix. There were few observations with last weeks lice numbers above 10, so no average was calculated for these. The average increased linearly with increasing number of mobile lice last week up to 5 all mobile lice per salmon.

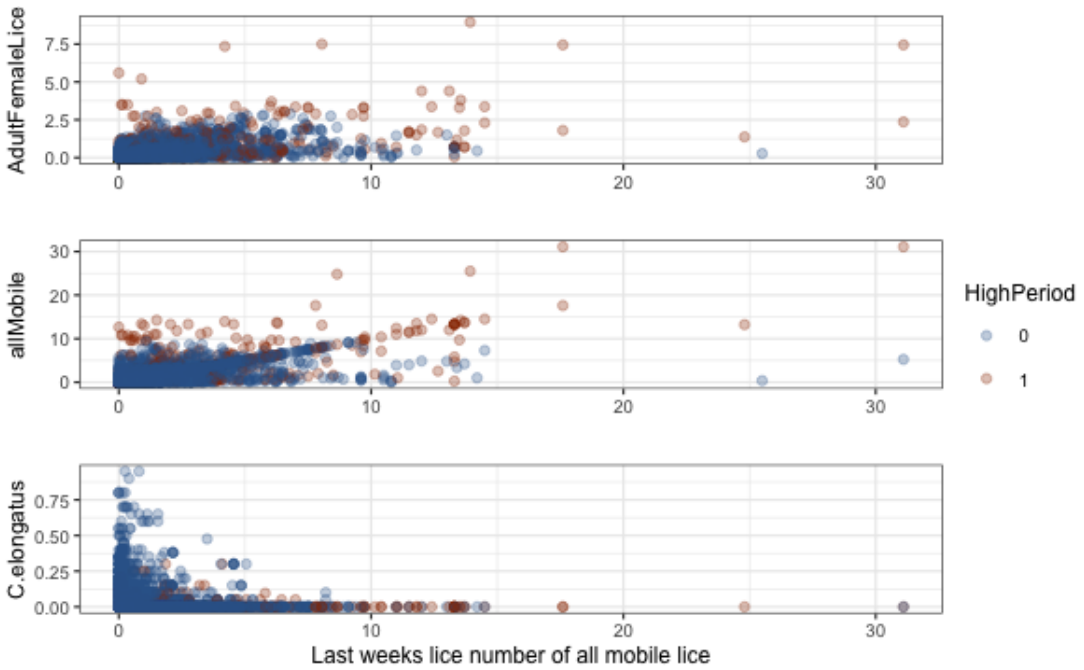

Figure 4.14: The weekly reported lice numbers of adult female lice, all mobile lice and C. elongatus for each cage in operation plotted against the last weeks reported lice numbers of all mobile lice. The red data points represent data from the period with high lice pressure and the blue points the rest of the data. To visualize overplotting, the data points are partially transparent.

In Figure 4.15, the correlation between each pair of the numerical variables are presented. In the upper triangle, the Pearson correlation coefficients are presented, and in the lower triangle, scatter plots of each pair are drawn. The variable distribution are plotted on the diagonal. The Pearson correlation indicated high correlation between some of the variables. *AdultFemaleLice* and *allMobile* were highly correlated with correlation coefficient 0.784, and the scatter plot of these two indicated that there had not been observed a low lice number of adult female lice while the lice number of all mobile lice was high. This supported that an alternative model with *allMobile* as a response variable could be used to model the lice pressure in the study area.

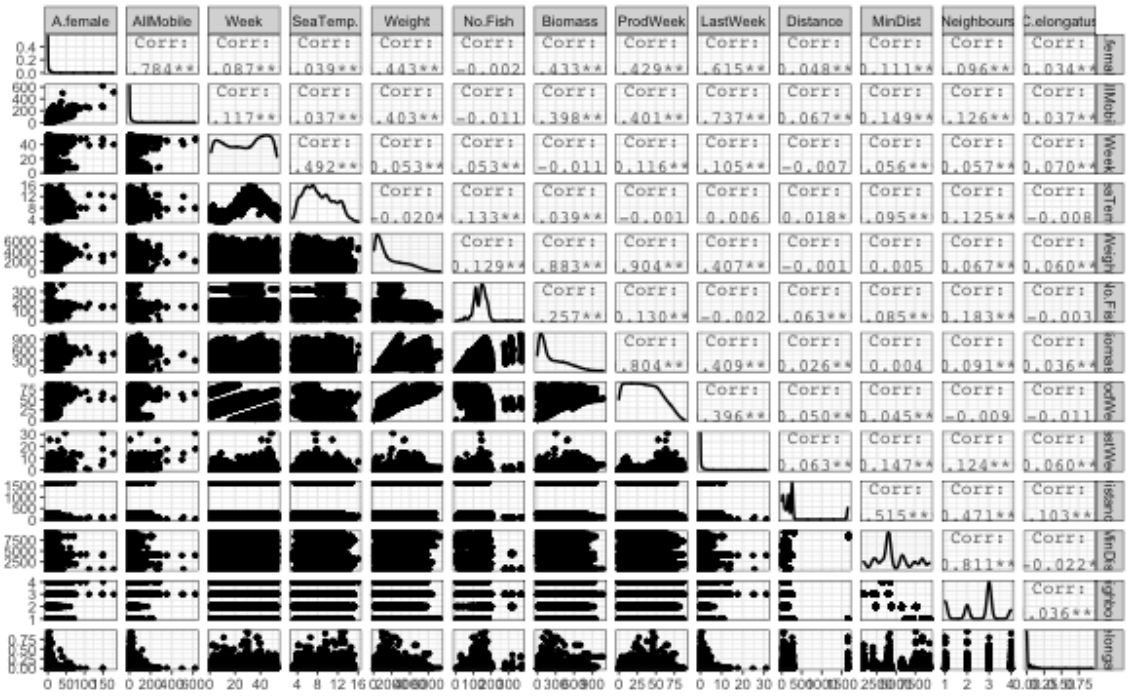

Figure 4.15: Pearson correlation coefficients between each pair of the numeric variables, scatter plots of each pair and distribution plot of each variable.

As indicated in Figure 4.14, the last weeks reported lice number of all mobile lice, *LastWeek*, was positive correlated with *AdultFemaleLice* and *allMobile*. The Pearson correlation coefficient between these pairs were calculated to be 0.617 and 0.737, respectively. The variable for production week was highly correlated with both the salmon weight and the biomass, with the correlation coefficient 0.904 and 0.804, respectively. The Pearson correlation coefficient between *Weight* and *Biomass* was calculated to be 0.883.

The distance to the closest neighbour, *MinDist*, and the number of neighbours within 10km, *neighbours*, were negative correlated with correlation coefficient -0.811. This make sense, since a lower distance to the nearest neighbour could lead to more salmon farms within a radius of 10 km. There was also a correlation between these two variables and the distance to the coastline, *Distance*. The Pearson correlation coefficient for *Distance* and *MinDist* was 0.515, and -0.471 for *Distance* and *neighbours*. Scatter plots of each of these pairs are plotted in Figure A.17 in the Appendix. There were no clear connection between the distance to the coastline and the sites nearby in these plots, but the site with much longer distances to the coastline probably affected the associated correlation coefficients.

According to the calculated Pearson coefficients in Figure 4.15 there were no major differences in the correlation between the different explanatory variables for *AdultFemaleLice* and *allMobile*. This indicated that a model for all mobile lice could be a good alternative if it became difficult to model only adult female lice due to a high number of zeros. From the distribution plot of *Week* the data seemed to be well spread over the whole year. This was also supported by the Pearson correlation between *ProductionWeek* and *SeaTemperature*, which was not significant. *C.elongatus* did not appear to correlate strongly with either the lice numbers for salmon lice or with any of the other variables.

5 Analysis and Validation

The lice numbers from the salmon farms were the reported average of lice per salmon calculated from a sample of at least 10 random salmon in a cage. The lice numbers gathered in this thesis were usually calculated from a sample of 20 salmon. To get a counting variable as response variable, the lice numbers needed to be converted to integers. Both the counted number of lice on a sample of 20 salmon in the cage and the estimated total number of lice in the entire cage were tested as response variable. The response variable was intended to apply to adult female lice, but to reduce the number of zeros, a model that included all mobile salmon lice was also tested. Thus, 4 different response variables were used in the fitting of a regression model for salmon lice: *CountAdultFemale*, *CountAllMobile*, *AdultFemaleCage* and *AllMobileCage*. As presented in Table 3.2, *CountAdultFemale* and *CountAllMobile* were calculated as the reported sample mean of adult female lice and all mobile lice, respectively, multiplied by 20 salmon. *AdultFemaleCage* and *AllMobileCage* were estimated as the reported sample mean of adult female lice and all mobile lice, respectively, multiplied by the reported number of salmon in the cage. Regression models based on the count of lice on a sample of 20 salmon are further referred to as sample models, while models for the estimated count of lice in the entire cage are referred to as cage models.

5.1 Model Selection

The explanatory variables for the regression model were selected based on experiences of the data set and the visualization in Section 4. In addition, the variance inflation factor, calculated with the `R`-function `vif` from the `car`-package (Fox & Weisberg, 2019), was used to assess multicollinearity in the regression model. The number of salmon in a cage varied little throughout the production cycles, so instead of using *Weight* and *NumberOfFish* as explanatory variables, the product of these two, *Biomass*, was used. The variance inflation factor for *MinDist* and *Neighbours* indicated multicollinearity (VIF$\geq$ 5), so the distance to the nearest salmon farm was not included in the model. The number of neighbors within 10km was assumed to be a better explanatory variable, since *MinDist* only took into account the nearest salmon farm and did not say anything about whether there were several salmon farms nearby. Nor was there a clear trend between the distance to the nearest salmon farm and the salmon lice numbers, while the lice pressure seemed to increase with the number of neighbors within 10km.

As presented in the correlation plot (Figure 4.15), the variable *LastWeek* was positive correlated with the response variables. An explanatory variable with last week's reported lice numbers could be interpreted as a reproduction number, where the exponential of the estimated coefficient indicates the expected factor increase in the lice count if the previous lice number increases by one and other factors are constant. The highest reported lice number of all mobile lice in the study period was 31 lice per salmon, while 81.4% of the data had lice number below 1. By using the last week's lice numbers as an explanatory variable, too high values were predicted for the observations with highly registered lice numbers previous week and too low otherwise.

Farming companies are required to introduce measures to prevent the lice number from rising above 0.5 adult female lice per salmon. It was therefore most interesting to adapt a suitable model for lice counts below this limit, as this is where other factors have the greatest impact on the number of lice. Once the lice numbers in the cage becomes too high, the lice multiply quickly and a delousing treatment are needed to reduce the lice pressure. To avoid that the highest lice numbers in *LastWeek* affected the coefficient estimate, all reported lice numbers from the previous week that were greater than 1 were registered as 1 in the explanatory variable *LastWeek1*. In this way, the variable ranged from 0 all mobile lice to 1 all mobile lice per salmon, which indicated that 20 mobile lice (including adult female lice) had been counted on a sample of 20 salmon. By using this variable as explanatory variable instead, the correlation with the response variables became less and the regression model fitted the low lice numbers better. The correlation coefficient for *LastWeek1* and the lice numbers *AdultFemaleLice* and *allMobile* were calculated as 0.531 and 0.576, respectively. The average lice number of adult female lice and all mobile lice calculated for every tenth in *LastWeek1* are presented in Figure A.18. High lice numbers have been observed regardless of last weeks lice numbers, but the average lice number increased for higher reported lice numbers last week.

The visualization of the explanatory variables in Section 4, gave an indication that the lice numbers not had a clearly increasing or decreasing trend for all the variables. Therefore, some polynomial terms and indicator variables were added to the model. Quadratic terms of the sea temperature and the biomass were added to the model. This lead to multicollinearity, and instead of using the quadratic term of *Biomass*, an indicator variable, *BiomassIndicator*, was used. According to Figure 4.8, the lice number increased with the biomass up to 490 metric ton and there were fewer cases of zero lice in cages with larger biomasses. Therefore, *BiomassIndicator* were used to indicate whether the biomass in the cage was smaller or larger than 490 metric ton.

To take into account how the lice numbers changed with the time of the year an indicator variable, *Season*, was created. It distinguished between the three periods with different lice pressure, as commented in Figure 4.5. The period with lowest lice pressure, week 14-32 was denoted spring, the period with highest lice pressure, week 33-48, was denoted autumn and the last period was named winter. One site was located 1650m from the coastline, while all other sites were closer than 300m to the coastline. An indicator variable, *DistToCoast*, was thus preferred instead of the continuous variable *Distance*. The distance to the coastline were divided into three levels; 0: less than 150m, 1: 150-300m and 2: $>$ 300m. In Figure 4.9 and Figure A.11 it was a clear trend that the lice numbers became higher after the salmon had been in the sea for around 50 weeks. Instead of using the continuous variable for production week, which was highly correlated with the biomass, an indicator variable, *ProdWeek50*, which indicated whether it had been more than 50 weeks since deployment, was used in the regression model. The final explanatory variables used in the full model with definitions and units are presented in Table 5.1.

Table 5.1: The explanatory variables used in the analysis with explanation and units

Variable name	**Explanation of variable**
SeaTemperature	Sea temperature, °C
Location	The location of the salmon site (Fjord 1A, 1B, 2A, 2B or Coast), coded as an indicator variable with Coast as reference variable
OperatingModel	The operating method used (Stage, Fjord or Coast), coded as an indicator variable with the stage model as reference
Season	Period of the year (Spring: week 14-32, Autumn: week 33-48, Winter: week 49-13), coded as an indicator variable with Autumn as reference
Biomass	The reported biomass in the cage, in metric ton
BiomassIndicator	An indicator variable indicating whether the biomass was larger than 490 metric ton. Biomass under 490 metric ton was used as a reference.
ProdWeek50	An indicator variable indicating whether there was over 50 weeks since the smolts were deployed. Production weeks below 50 were used as the reference.
Treatment	An indicator variable indicating whether there was an ongoing delousing treatment in the cage, where no treatment was used as the reference
DistToCoast	Distance from the salmon farm to the coastline, coded as an indicator variable: 0 - less than 150m (used as reference), 1 - 150-300m, 2 - over 300m
Neighbours	Number of salmon farms within a radius of 10km (1-4 neighbours)
LastWeek1	Last week's reported lice number of all mobile lice (adult female lice and mobile lice), where all reported lice numbers larger than 1 were equal to 1.
HighPeriod	An indicator variable indicating whether the week was defined with high lice pressure, where periods with normal lice pressure were used as reference. The variable was defined in Sec 3.2.5, and there were some small differences compared to whether it was a model for adult female lice or all mobile lice that was adapted.

5.2 Poisson Regression

A Poisson regression model for the data was fitted by assuming that the counts of salmon lice from each cage and week followed an independent Poisson distribution, $y_i \sim \text{Po}(\lambda_i)$, with

$$\begin{aligned}\ln \hat{\lambda}_i = \hat{\beta}_0 &+ \hat{\beta}_1 \cdot \text{SeaTemperature} + \hat{\beta}_2 \cdot \text{SeaTemperature}^2 \\ &+ \hat{\beta}_3 \cdot \text{Location1A} + \hat{\beta}_4 \cdot \text{Location1B} + \hat{\beta}_5 \cdot \text{Location2A} \\ &+ \hat{\beta}_6 \cdot \text{Location2B} + \hat{\beta}_7 \cdot \text{OperatingModelCoast} \\ &+ \hat{\beta}_8 \cdot \text{OperatingModelFjord} + \hat{\beta}_9 \cdot \text{SeasonSpring} + \hat{\beta}_{10} \cdot \text{SeasonWinter} \\ &+ \hat{\beta}_{11} \cdot \text{Biomass} + \hat{\beta}_{12} \cdot \text{BiomassIndicator} + \hat{\beta}_{13} \cdot \text{ProdWeek50} \\ &+ \hat{\beta}_{14} \cdot \text{Treatment} + \hat{\beta}_{15} \cdot \text{DistToCoast1} + \hat{\beta}_{16} \cdot \text{DistToCoast2} \\ &+ \hat{\beta}_{17} \cdot \text{Neighbours} + \hat{\beta}_{18} \cdot \text{LastWeek1} + \hat{\beta}_{19} \cdot \text{HighPeriod}.\end{aligned} \tag{5.1}$$

For modelling the count of adult female lice, regression models for both the sample model and the cage model were fitted in `R`. The summary output from the model fit of the sample model with *CountAdultFemale* as response variable is given in Table 5.2. The estimated regression coefficients with standard error, z-value and p-value for the cage model with *AdultFemaleCage* as response variable are presented in Table B.1 in the Appendix.

Table 5.2: Regression coefficients with associated estimate, standard error, z-value and p-value from the Poisson regression for the sample model of adult female lice.

Coefficients	Estimate	Std. Error	z value	p-value
Intercept	-0.6864	0.0670	-10.24	$< 2 \cdot 10^{-16}$
SeaTemperature	-0.1023	0.0127	-8.04	$8.69 \cdot 10^{-16}$
SeaTemperature^2	0.0069	0.0007	10.28	$< 2 \cdot 10^{-16}$
Location1_A	-0.5788	0.0423	-13.68	$< 2 \cdot 10^{-16}$
Location1_B	-0.4112	0.0282	-14.58	$< 2 \cdot 10^{-16}$
Location2_A	-0.4658	0.0490	-9.50	$< 2 \cdot 10^{-16}$
Location2_B	-0.5487	0.0328	-16.75	$< 2 \cdot 10^{-16}$
OperatingModelCoast	0.2858	0.0146	19.60	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.6447	0.0275	23.46	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.3154	0.0151	-20.94	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.1277	0.0166	-7.71	$1.24 \cdot 10^{-14}$
Biomass	0.0013	$4.6 \cdot 10^{-5}$	29.49	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.3289	0.0162	-20.36	$< 2 \cdot 10^{-16}$
ProdWeek50	0.5182	0.0165	31.46	$< 2 \cdot 10^{-16}$
Treatment	-0.2288	0.0151	-15.17	$< 2 \cdot 10^{-16}$
DistToCoast1	-0.1125	0.0138	-8.14	$3.89 \cdot 10^{-16}$
DistToCoast2	0.2037	0.0239	8.53	$< 2 \cdot 10^{-16}$
Neighbours	0.1361	0.0082	16.50	$< 2 \cdot 10^{-16}$
LastWeek1	2.0244	0.0181	112.07	$< 2 \cdot 10^{-16}$
HighPeriod	0.7181	0.0132	54.35	$< 2 \cdot 10^{-16}$

AIC: 73406, Null deviance: 125348 on 15595 degrees of freedom
Residual deviance: 49321 on 15576 degrees of freedom

The p-values from the Wald test indicated that all the terms were significant for both the sample model and the cage model. There were no major changes in the coefficient estimates for the two models, but the intercept and the linear term of *SeaTemperature* changed sign and became positive for the cage model. By substituting the estimated mean $\hat{\lambda}_i$ in Equation (2.7), the Pearson statistic for the sample model and the cage model were calculated as $P = 67362$ and $P = 463842913$, respectively. The corresponding quantile of the $\chi^2_{\alpha,n-p}$ distribution was $\chi^2_{0.05,15576} = 15867$. Thus, both models were rejected at significance level $\alpha = 0.05$, since the test statistic P was larger than $\chi^2_{0.05,15576}$ for both models.

Residual plots of the Pearson residuals plotted against the fitted values for both models are presented in Figure 5.1. The shape of the residual plots were quite similar for both the sample model and the cage model, but there were large differences in the magnitude of the fitted values. A count of 50 adult female lice on 20 salmon will give a lice number of 2.5, and for a cage with 140 thousand salmon this corresponds to 350 000 estimated adult female lice in the cage. There were thus large differences in the size of the two response variables, and naturally then also in the fitted models.

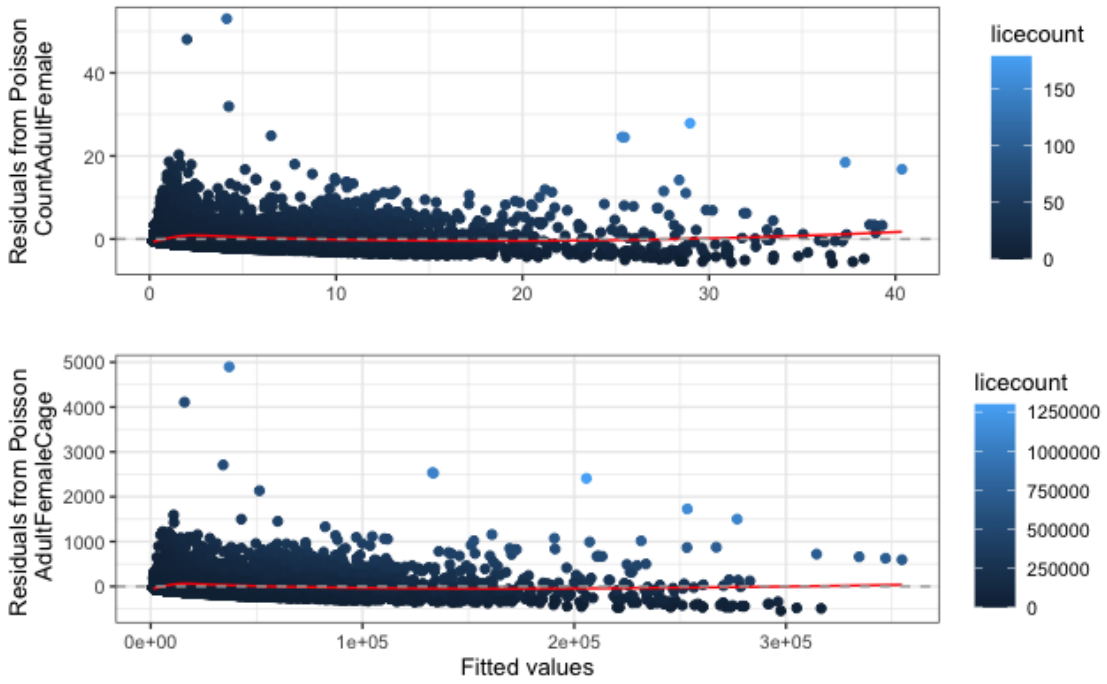

Figure 5.1: Plot of Pearson residuals against fitted values from the Poisson regression for the sample model of adult female lice (top) and the cage model of adult female lice (bottom). The points are colored after the corresponding lice count of adult female lice, which is the observed count of adult female lice on a sample of 20 salmon (top) and the estimated total count of adult female lice in the cage (bottom).

According to the goodness of fit test and the residual plots, the Poisson regression model was not a good fit to the data. By using the observed Pearson statistic, the overdispersion parameter was estimated as $\hat{\phi}_P = \frac{P}{n-p} = 4.32$ for the sample model and 29779.33 for the cage model. Both indicated overdispersion, and since the Pearson statistic was chi-squared distributed under $H_0 : \phi_P \leq 1$ and $P > \chi^2_{0.05,15576}$, this supported the impression

of overdispersion.

A frequency plot of fitted and observed values of the count of adult female lice on a sample of 20 salmon are presented to the left in Figure 5.2. The plot shows that there was an excessive number of zeros in the observed data and that the Poisson model not was able to fit enough zeroes. Instead, it predicted more ones than what had been observed. It had been reported absence of adult female lice on the sample in 8125 observations. The expected mean from the Poisson model was used to calculated the probability of observing zero salmon lice for each row in the data set. By summarizing the probability for all rows, the expected number of observations with zero salmon lice from the model was estimated. The estimated expected number of zeroes predicted from the sample model was 5748 zeros, which gave a ratio of 0.71.

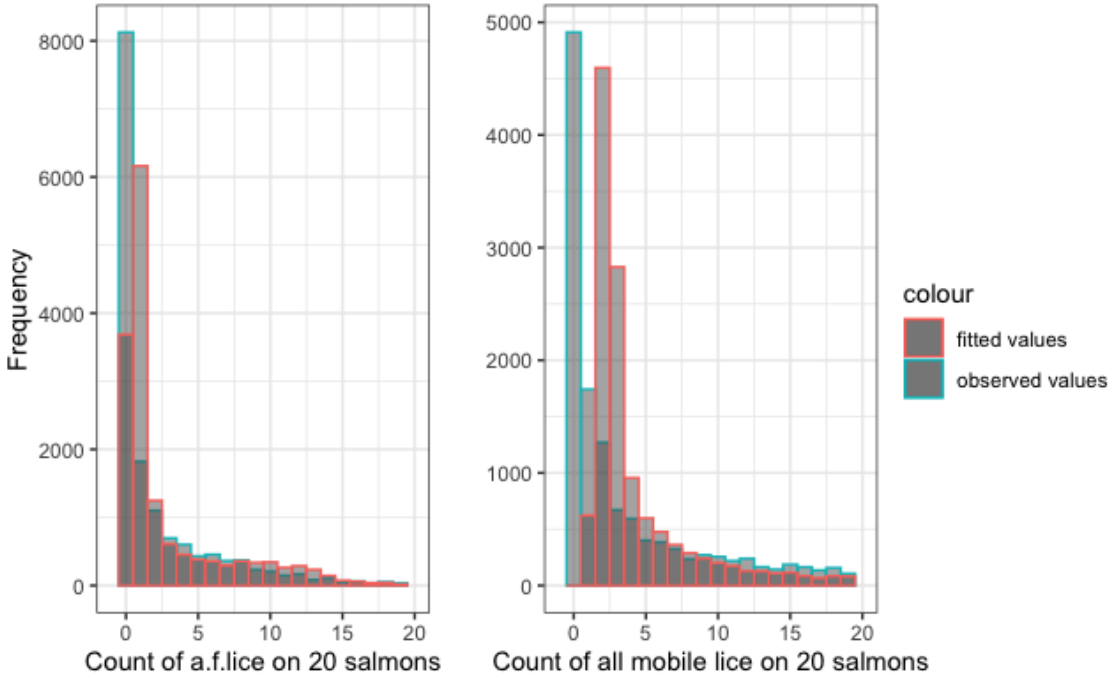

Figure 5.2: The frequency of observed and fitted values from the sample model of adult female lice and all mobile lice, respectively. Only counts up to 20 lice are presented here (lice numbers ≤ 1), but the highest observed counts were 179 adult female lice and 622 all mobile lice. The maximum fitted values from the two models were 40 adult female lice and 195 all mobile lice, respectively.

To reduce the number of observed zeroes, a sample model for all mobile lice on 20 salmon was fitted. *CountAllMobile* was thus used as the response variable, and the summary output from the Poisson regression is presented in Table B.2 in the Appendix. The frequency plots for the observed values and the fitted values are presented to the right in Figure 5.2. The frequency of observed values equal to zero was reduced compared to the model for adult female lice, but none of the fitted values for the count of all mobile lice were zero. By using the expected mean from the regression model, the model was expected to predict 993 zeroes, while it was observed 4911 zeroes. This gave a ratio of 0.20, which was lower than for the sample model of adult female lice. The overdispersion parameter was estimated as 14.7, and a model with all mobile lice thus did not appear to

improve the dispersion problem.

The expected means from the Poisson regression for the cage model of adult female lice were high, and the model was thus not expected to fit any zeroes. Frequency plots of both observed values and fitted values from the Poisson regression for the cage model of adult female lice are presented in Figure 5.3. The lowest fitted values was 1118 adult female lice, while absence of salmon lice had been observed in 7907 cases. The observed number of adult female in the cage, *AdultFemaleCage*, was estimated as the number of salmon in the cage multiplied with the average number of adult female lice calculated from a small sample of the cage. If none adult female lice had been observed on the sample of salmon, it was for the cage model assumed that there were zero lice in the cage. There was thus great uncertainty in the estimated number of lice in the cage. The sample model, which was based on the sample of 20 salmon, also comes better out on the goodness-of-fit tests, so the sample model was used further in the analysis.

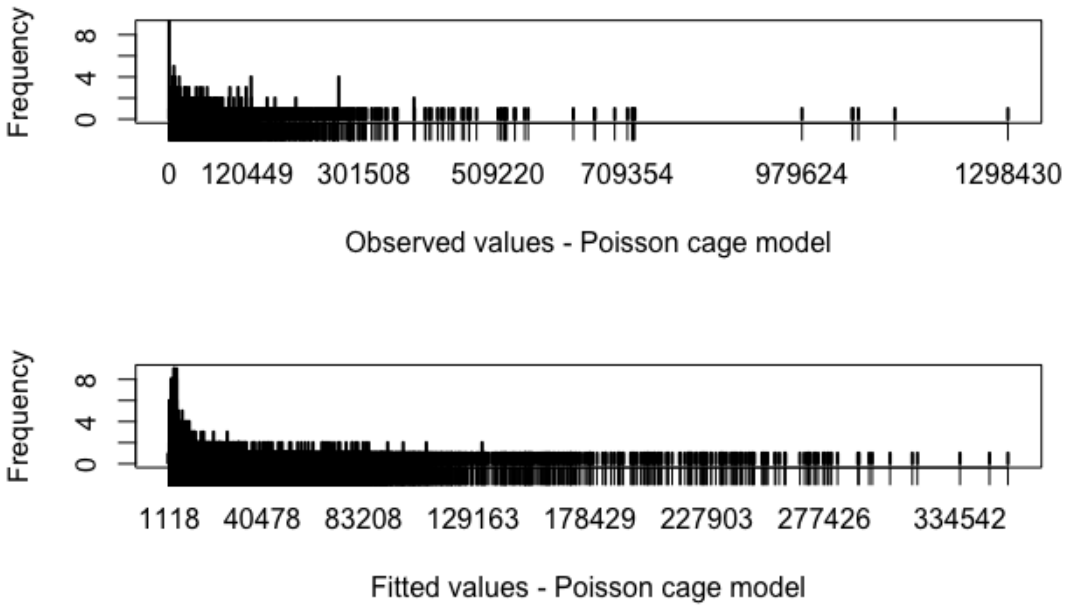

Figure 5.3: The frequency of observed (top) and fitted (bottom) values from the Poisson model for the cage model of adult female lice. The frequency of observed zeroes are cut at 9, but there was estimated that there were 7907 observed values of zero adult female lice in the cage.

Censored frequency plots for observed values of both adult female lice and all mobile lice on a sample of 20 salmon, with frequency below 100 and an unlimited x-axis are presented in Figure A.19 in the Appendix. This indicated that on a sample of 20 salmon, it was most often counted below 20 adult female lice and 80 all mobile lice, but there were cases which differs greatly from this. The highest fitted values from the Poisson model was 40 adult female lice and 195 all mobile lice, while there had been observed up to 179 adult female lice and 622 all mobile lice. It therefore appeared that there was greater variance in the positive data than expected from a Poisson model. Thus, the overdispersion was probably due to both zero-inflation and extra variance in the positive count data.

5.3 Negative Binomial Regression

The negative binomial regression model allows larger variance in the count data than the Poisson model, and was thus fitted for the sample model of adult female lice. The summary output from the model fit in R is given in Table 5.3. The p-values from the Wald test indicated that all the terms without *SeasonWinter* and *Treatment* was significant up to a significance level 0.05. There were no changes in sign from the estimated Poisson regression coefficient in the sample model in Table 5.2. The Pearson statistic for the negative binomial model was calculated as $P = 19077$, which was lower than the Pearson statistic from the Poisson model. The corresponding quantile $\chi_{0.05,15576} = 15867$ was still lower than the Pearson statistic, which indicated that the fitted model did not fit with the observed distribution. From the null deviance and the residual deviance in Table 5.3, the explained deviance was calculated as 61.3% by using Equation (2.25).

Table 5.3: Regression coefficients with associated estimate, standard error, z-value and p-value from the negative binomial regression for the sample model of adult female lice.

Coefficients	Estimate	Std. Error	z value	p-value
Intercept	-1.0065	0.1527	-6.59	$4.37 \cdot 10^{-11}$
SeaTemperature	-0.1684	0.0295	-5.70	$1.18 \cdot 10^{-08}$
SeaTemperature^2	0.0113	0.0016	7.22	$5.39 \cdot 10^{-13}$
Location1_A	-0.5572	0.0839	-6.64	$3.06 \cdot 10^{-11}$
Location1_B	-0.3072	0.0542	-5.66	$1.49 \cdot 10^{-8}$
Location2_A	-0.4220	0.0853	-4.95	$7.58 \cdot 10^{-7}$
Location2_B	-0.7044	0.0611	-11.53	$< 2 \cdot 10^{-16}$
OperatingModelCoast	0.4982	0.0354	14.07	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.5905	0.0453	13.05	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.1826	0.0320	-5.70	$1.18 \cdot 10^{-8}$
SeasonWinter	-0.0273	0.0406	-0.67	0.5015
Biomass	0.0026	0.0001	23.10	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.6989	0.0444	-15.74	$< 2 \cdot 10^{-16}$
ProdWeek50	0.6473	0.0390	16.59	$< 2 \cdot 10^{-16}$
Treatment	-0.0272	0.0387	-0.70	0.4821
DistToCoast1	-0.0673	0.0310	-2.17	0.0299
DistToCoast2	0.3052	0.0521	5.85	$4.82 \cdot 10^{-9}$
Neighbours	0.0900	0.0184	4.88	$1.06 \cdot 10^{-6}$
LastWeek1	2.2157	0.0378	58.59	$< 2 \cdot 10^{-16}$
HighPeriod	0.6590	0.0438	15.05	$< 2 \cdot 10^{-16}$

AIC: 50335, Null deviance: 34814 on 15595 degrees of freedom
Residual deviance: 13468 on 15576 degrees of freedom

Both the Pearson residuals and the deviance residuals for the negative binomial sample model are plotted against the fitted values in Figure 5.4. It seemed to be a dominance of positive residuals for fitted values close to zero. The residuals then decreased with increasing fitted values and it became an dominance of negative residuals. A positive sign of the residuals indicates that the observed value was higher than the predicted value from the model. The variance of the Pearson residuals was higher for fitted values close to zero, and decreased for higher fitted values. There seemed to be heteroscedasticity in

the residuals, which violated the assumption of constant variance for the residuals.

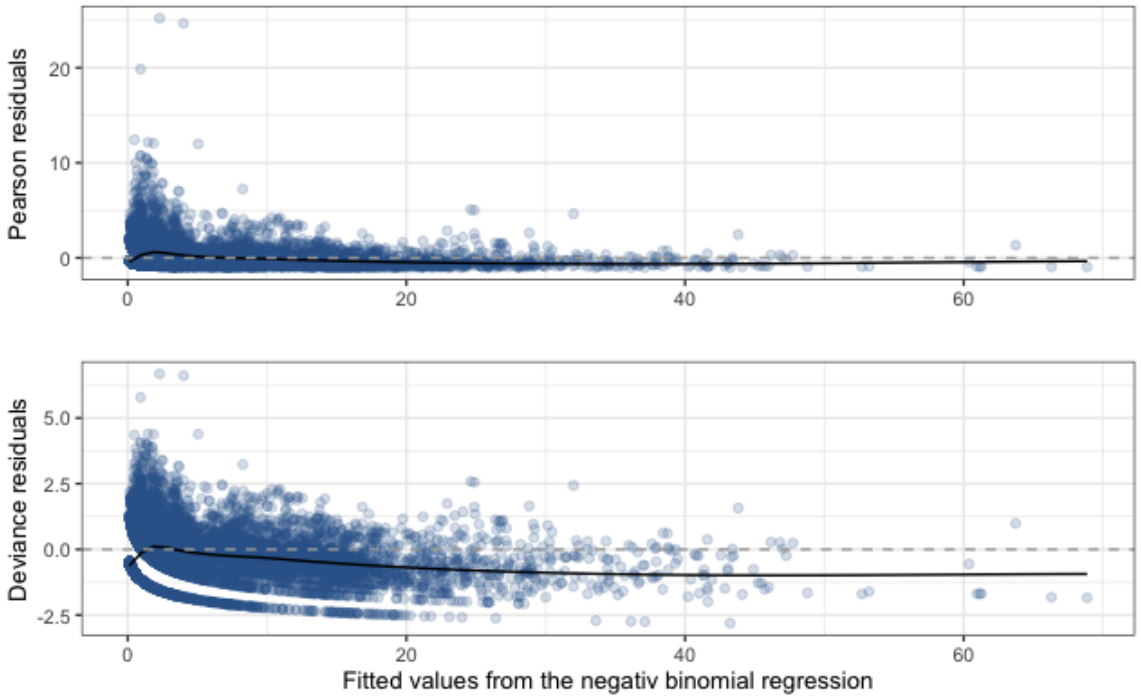

Figure 5.4: Pearson residuals (top) and deviance residuals (bottom) plotted against the fitted values from the negative binomial regression model for the sample model of adult female lice. Because of the large data set, the data points are partly transparent and the black line is a loess smoothing curve.

As presented in Section 2.1.3, a likelihood ratio test could be used to compare the sample models obtained from the Poisson distribution and the negative binomial distribution, respectively. The Poisson model was nested within the negative binomial model, since the same covariates were used in both models. If $r \to \infty$, the variance function $\text{Var}_{NB}[Y_i] = \lambda_i + \frac{\lambda_i}{r} \to \lambda_i$, which is the variance of a Poisson model. Thus, by defining $\alpha = r^{-1}$, the null hypothesis $H_0 : \alpha = 0$ could be tested against the alternative hypothesis $H_1 : \alpha > 0$. The likelihood ratio test was conducted by using the `R`-function `lrtest` from the `lmtest` package (Zeileis & Hothorn, 2002). The results from the test is presented in Table 5.4. The p-value was lower than $2.2 \cdot 10^{-16}$, which suggested that the negative binomial model was a more appropriate model than the Poisson model.

Table 5.4: Results from the likelihood ratio test between the Poisson sample regression model and the negative binomial sample model for adult female lice.

Number of df	log likelihood	df	Chisq	p-value
20	-36683			
21	-25147	1	23073	$< 2.2 \cdot 10^{-16}$

The parameter r in the variance function $\text{Var}_{NB}[Y_i] = \lambda_i + \frac{\lambda_i}{r}$ was estimated as 1.049 with standard error 0.022 in the model fit. The negative binomial sample model was expected to predict 7810 zeros, which gave a ratio of observed and predicted zero as 0.96. The

ratio was within the tolerance range, and the model did not seem to be underfitting zeros as the Poisson model did. By using the ratio of the Pearson statistic and the degrees of freedom the dispersion parameter was estimated as 1.22. This was an improvement from the Poisson model, but the dispersion parameter still indicated slightly overdispersion.

5.4 Zero-Inflated Regression

The reported lice numbers of adult female lice included a lot of zeros (50.7%). The Poisson model was underfitting zeros, while the negative binomial model seemed to fit almost enough zeros. An alternative was to fit both a zero-inflated Poisson regression model and a zero-inflated negative binomial model for the sample model of adult female lice. Which covariates that affected the zero part and the count part were not known, so for the full model, the same covariates were used for both the zero part and the count part, $\mathbf{x}_i = \mathbf{z}_i$. The expected mean for observation i in the count part was given by $\lambda_i = \exp(x_i^T \boldsymbol{\beta})$, and the probability of observing a false zero was given by $\pi_i = \frac{\exp(z_i^T \boldsymbol{\gamma})}{1+\exp(z_i^T \boldsymbol{\gamma})}$. The parameters in the zero-inflated Poisson (ZIP) model were thus $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_{19}, \gamma_0, \gamma_1, \ldots, \gamma_{19})^T$, and for the zero-inflated negative binomial (ZINB) model, the parameter r in Equation (2.12) also had to be included. The ZIP model was thus nested within the ZINB model, and the sample models could be compared with a likelihood ratio test using `lrtest` from the `R`-package `lmtest` (Zeileis & Hothorn, 2002).

The zero-inflated models were fitted in `R` by the function `zeroinfl` from the `countreg` package (Zeileis et al., 2008). The summary output from the model fit of the sample models ZIP and ZINB are presented in Table B.3 and Table B.4 in the Appendix, respectively. The results from the likelihood ratio test are presented in Table 5.5. The null hypothesis was rejected (p-value $< 2.2 \cdot 10^{-16}$), and there was a strong support for the ZINB model. The ratio of observed and predicted zeros for the ZINB sample model was 0.96, which was within the tolerance range 0.05. The ratio for the ZIP sample model was 0.81, which was an improvement from the Poisson model, but the ZIP model did still not predict enough zeroes compared to the observed numbers of zeros.

Table 5.5: Results from the likelihood ratio test between the full regression models ZIP and ZINB, both sample models for adult female lice.

Number of df	log likelihood	df	Chisq	p-value
40	-32384			
41	-23928	1	16913	$< 2.2 \cdot 10^{-16}$

Several of the regression parameters for the ZINB model in Table B.4 were not significant at a 5%-level, so a model selection was preferred. The standard error for the zero-inflated coefficient *ProdWeek50* was 567, which was very large compared to the estimated regression coefficient. The p-value 0.9769, from the Wald test, indicated that the indicator variable was not significant and thus not important for the zero-inflation. From a biological perspective, it was not expected that the probability of observing a false zero changed after production week 50, and since the variables for biomass correlated with the production week, it seemed that *ProdWeek50* lead to multicollinearity in the zero-inflation part. The model was thus refitted without this term, and the estimated zero-inflated coefficient

for *BiomassIndicator* then increased from 1.72 to 3.26, and became significant with a p-value $4.08 \cdot 10^{-7}$.

From a new Wald test, the count coefficient *SeaTemperature* was not significant. In addition, the significance of each of the categorical variables with more than 2 levels were calculated by dropping all the levels from the model, and comparing the refitted model with the full model using likelihood ratio test. The likelihood ratio test suggested that the zero-inflation coefficient *Season* was not significant. Sequentially, models where either the count-coefficient *SeaTemperature* or the zero-inflation coefficient *Season* had been dropped was fitted. The model where *Season* was dropped in the zero-part gave the lowest AIC, and the likelihood ratio test with p-value 0.41 supported that the season term was not significant for the zero-inflation. The ZINB model was refitted without the zero-inflation coefficient *Season*.

The significant of the remaining categorical variables where tested with the likelihood ratio test, and the result from the test are given in Table 5.6. The summary output for the sample model without the zero-inflation coefficients *ProdWeek50* and *Season* is presented in Table 5.7. From the likelihood ratio test in Table 5.6 and the Wald test in Table 5.7, no further terms could be dropped from the model, and the preferred ZINB sample model was obtained. The signs of the count coefficient estimates for the ZINB model were the same as for the standard models Poisson and negative binomial.

Table 5.6: Results from the likelihood ratio tests where the significance of each of the categorical variables in the optimal ZINB sample model for adult female lice (Table 5.7) are tested. λ_i is the expected mean in the count part and π_i is the probability of observing a false zero.

Dropped term	#Df	LogLik	Df	Chisq	p-value
None	38	-23974			
Location from λ_i	34	-23995	4	41.16	$2.48 \cdot 10^{-8}$
OperatingModel from λ_i	36	-24055	2	161.8	$< 2.2 \cdot 10^{-16}$
Season from λ_i	36	-24043	2	137.4	$< 2.2 \cdot 10^{-16}$
DistToCoast from λ_i	36	-23992	2	35.74	$1.73 \cdot 10^{-8}$
Location from π_i	34	-24057	4	165.04	$< 2.2 \cdot 10^{-16}$
OperatingModel from π_i	36	-23995	2	40.75	$1.42 \cdot 10^{-9}$
DistToCoast from π_i	36	-23982	2	15.01	$5.49 \cdot 10^{-4}$

The Pearson residual for the ZIP model (Table B.3), the negative binomial model (Table 5.3) and the ZINB model (Table 5.7) are plotted against their fitted values in Figure 5.5. The residual plot for the ZIP model seemed to be randomly spread around the horizontal axis, while the residuals for the negative binomial model and the ZINB model were highest for the low fitted values. It appeared that there was an improvement in the residual plot for the zero-inflated negative binomial model, compared to the residuals from the negative binomial model. The residuals for the low fitted values were smaller for the ZINB model, and the residuals seemed thus to be more randomly spread and the variance in the residuals were more constant compared to the residuals from the negative binomial model. The negative binomial model fitted values up to 70 adult female lice counted on 20 salmon, while the zero-inflated models did not fit values above 40 adult

Table 5.7: Regression coefficients with associated estimate, standard error, z-value and p-value from the zero-inflated negative binomial regression for the preferred sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z-value	p-value
Intercept	-0.3983	0.1535	-2.5941	0.0095
SeaTemperature	-0.0594	0.0296	-2.0062	0.0448
SeaTemperature^2	0.0049	0.0016	3.1054	0.0019
Location1_A	-0.4938	0.0921	-5.3646	$8.11 \cdot 10^{-8}$
Location1_B	-0.2749	0.0630	-4.3635	$1.28 \cdot 10^{-5}$
Location2_A	-0.1948	0.1198	-1.6258	0.1040
Location2_B	-0.3115	0.0780	-3.9912	$6.57 \cdot 10^{-5}$
OperatingModelCoast	0.2995	0.0328	9.1304	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.4407	0.0630	6.9902	$2.75 \cdot 10^{-12}$
SeasonSpring	-0.3502	0.0310	-11.3098	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.0748	0.0378	-1.9821	0.0475
Biomass	0.0012	0.0001	11.3202	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.2723	0.0408	-6.6679	$2.59 \cdot 10^{-11}$
ProdWeek50	0.5479	0.0344	15.9290	$< 2 \cdot 10^{-16}$
Treatment	-0.1480	0.0348	-4.2554	$2.09 \cdot 10^{-5}$
DistToCoast1	-0.0870	0.0329	-2.6435	0.0082
DistToCoast2	0.2547	0.0477	5.3445	$9.06 \cdot 10^{-5}$
Neighbours	0.1395	0.0183	7.6397	$2.18 \cdot 10^{-14}$
LastWeek1	1.5114	0.0374	40.3599	$< 2 \cdot 10^{-16}$
HighPeriod	0.7270	0.0428	16.9778	$< 2 \cdot 10^{-16}$
Log(r)	0.3609	0.0226	15.9654	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z-value**	**p-value**
Intercept	1.4059	0.6731	2.0888	0.0367
SeaTemperature	0.3884	0.1254	3.0974	0.0020
SeaTemperature^2	-0.0202	0.0073	-2.7662	0.0057
Location1_A	-0.9492	0.5588	-1.6987	0.0894
Location1_B	-1.0449	0.3256	-3.2095	0.0013
Location2_A	0.2212	0.3910	0.5659	0.5715
Location2_B	1.2056	0.3140	3.8397	0.0001
OperatingModelCoast	-1.4175	0.2709	-5.2335	$< 2 \cdot 10^{-16}$
OperatingModelFjord	-0.5163	0.1378	-3.7475	0.0002
Biomass	-0.0139	0.0011	-13.2144	$< 2 \cdot 10^{-16}$
BiomassIndicator	3.2210	0.6311	5.1041	$3.32 \cdot 10^{-7}$
Treatment	-1.2014	0.5634	-2.1324	0.0330
DistToCoast1	0.0993	0.1704	0.5831	0.5598
DistToCoast2	1.1604	0.3017	3.8464	0.0001
Neighbours	0.2321	0.1077	2.1551	0.0312
LastWeek1	-20.9839	1.4012	-14.9753	$< 2 \cdot 10^{-16}$
HighPeriod	0.5285	0.2366	2.2334	0.0255
AIC=48025.01				

female lice. The highest observed lice count of adult female lice on 20 salmon was 160, but the lice count on 20 salmon was normally below 10 adult female lice (lice number 0.5).

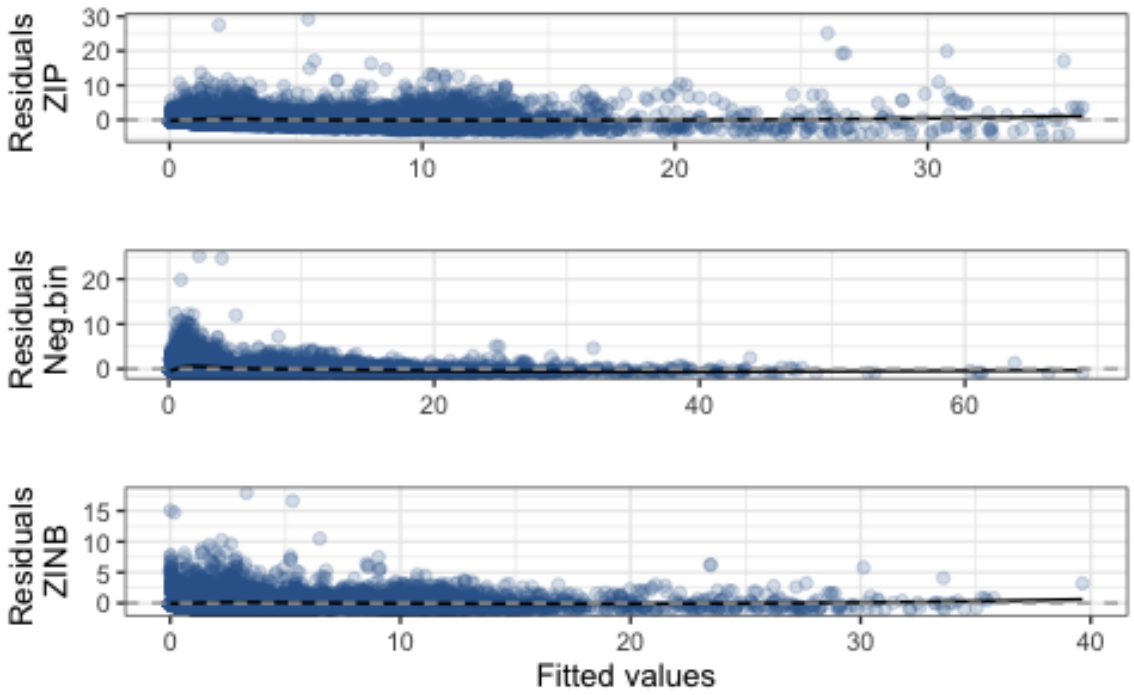

Figure 5.5: Pearson residuals plotted against the fitted values for the regression models ZIP (top), negative binomial (middle) and ZINB (bottom) for the sample model of adult female lice. Because of the large data set, the data points are partly transparent and the black line is a loess smoothing curve.

The fitted values from the model fit of the negative binomial model (Table 5.3) and the zero-inflated negative binomial model (Table 5.7) are plotted against the time in Figure 5.6. To compare fitted values against observed values, a plot with the observed count of adult female lice on 20 salmon is also included. Due to some high observed values, a censored plot, where $CountAdultFemale > 70 = 70$, is used for the comparison. The observed values are presented without censoring in Figure 3.2. The average fitted count of adult female lice on a sample of 20 salmons was calculated for each week in the study period for each of the regression models, and plotted as a black line in the plots. The average reported lice count was calculated from *CountAdultFemale* and not the censored variable. The fitted values from both the negative binomial model and the zero-inflated negative binomial model seemed to fit the lice count well, but the high lice numbers fitted from the negative binomial model in the middle of 2018, did not match with the observed values from this period. The ZINB model did not fit as high values as the negative binomial model, but seemed to hit the periods of high and low lice numbers better than the negative binomial model.

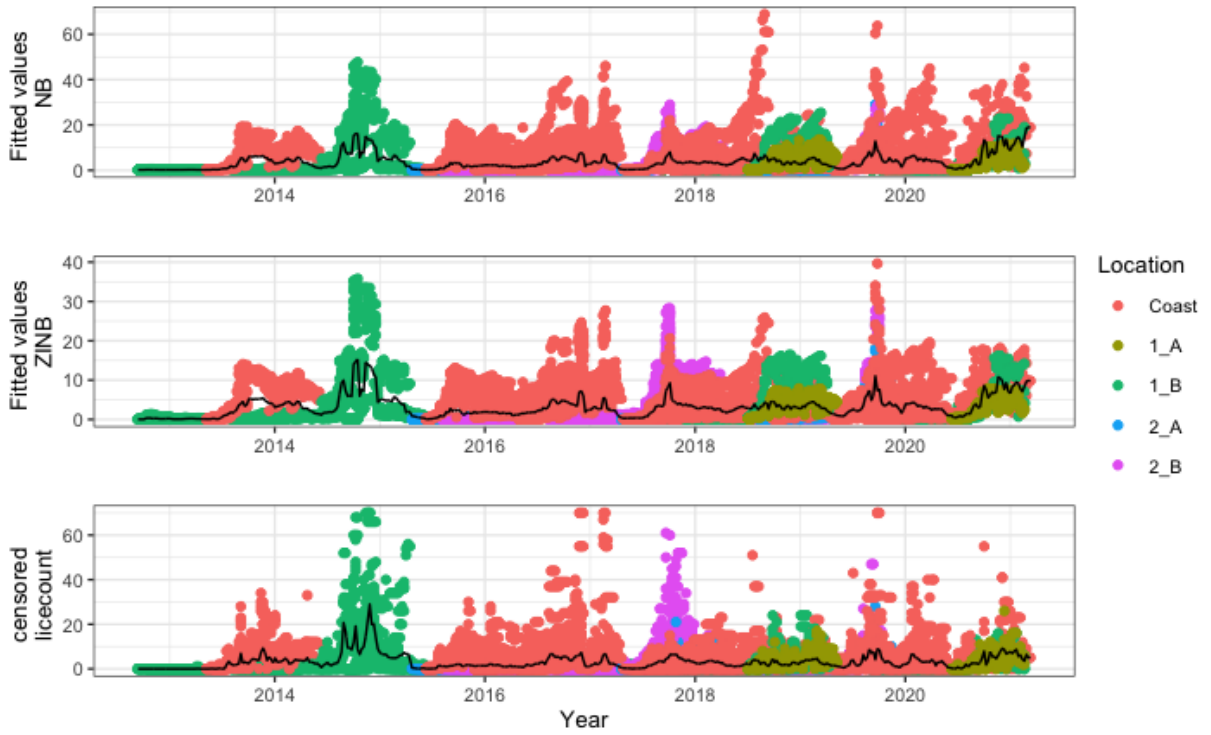

Figure 5.6: The fitted values for the negative binomial sample model (top) and the ZINB sample model for adult female lice (middle) during the study period. The observed number of adult female lice on a sample of 20 salmon (bottom) are plotted for comparison of fitted and observed values. The reported lice count are censored, such that lice count above 70 adult female lice are plotted as 70 adult female lice. The points are colored after the location of the site. The black line is the average value calculated for each week in the study period. The average reported lice count are calculated from CountAdultFemale and not the censored variable.

5.5 Hurdle Regression

Another model for data with zero-inflation is the zero-altered regression model, also called the hurdle regression model. Both a Poisson hurdle (ZAP) model and a negative binomial hurdle (ZANB) model were fitted for the count of adult female lice on 20 salmon (sample model). All the explanatory variables presented in Table 5.1 were used for both the count-part and the zero-part. The hurdle models were fitted by using the `R`-function `hurdle` from the `countreg` package (Zeileis et al., 2008), and the summary outputs from the model fit for the ZAP model and the ZANB model are presented in Table B.5 and Table B.6 in the Appendix, respectively. The ZAP model was nested within the ZANB model, and the models were thus compared with the likelihood ratio test. The AIC for the ZAP model was 64881.76, while it was 48604.38 for the ZANB model. From the likelihood ratio test presented in Table 5.8 and the calculated AICs, the ZANB model was preferred.

Table 5.8: Results from the likelihood ratio test between the full regression models ZAP and ZANB, both sample models for adult female lice.

Number of df	log likelihood	df	Chisq	p-value
40	-32401			
41	-24261	1	16279	$< 2.2 \cdot 10^{-16}$

From the Wald test for the ZANB model, the count variable *SeaTemperature* and the zero-inflation variables *Neighbours* and *HighPeriod* were not significant for a significance level of 0.05 (Table B.6). The model was thus simplified by using backward selection, and the results from the likelihood ratio tests are presented in Table 5.9, together with the refitted model's AIC. The expected mean for observation i in the count part was given by $\lambda_i = \exp(x_i^T \boldsymbol{\beta})$, and the probability of observing zero lice was given by $\pi_i = (1 + \exp(z_i^T \boldsymbol{\gamma}))^{-1}$. The model where *Neighbours* was dropped from π_i gave the lowest AIC, and the likelihood ratio test supported that *Neighbours* could be dropped from π_i. The model selection was continued, and the AIC was slightly improved by also dropping *HighPeriod* from π_i, with a p-value 0.258 from the likelihood ratio test. The count coefficient *SeaTemperature* was still not significant according to the Wald test, but the AIC became higher by dropping it from the model. Thus, no further terms were dropped, and the optimal model based on the AIC was the model without *Neighbours* and *HighPeriod* in π_i. The summary output for the preferred model is presented in Table 5.10.

Table 5.9: Results of the first step of model selection in the ZANB sample model for adult female lice presented in Table B.6

Dropped term	df	AIC	Likelihood ratio test	
None	41	48604.38		
SeaTemperature from λ_i	40	48605.12	$X^2 = 2.74$	$(df = 1, p = 0.098)$
Neighbours from π_i	40	48602.84	$X^2 = 0.46$	$(df = 1, p = 0.496)$
HighPeriod from π_i	40	48603.67	$X^2 = 1.29$	$(df = 1, p = 0.257)$

Table 5.10: Regression coefficients with associated estimate, standard error, z-value and p-value from the hurdle negative binomial regression for the preferred sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z value	p-value
Intercept	-0.3524	0.1744	-2.0210	0.0433
SeaTemperature	-0.0546	0.0330	-1.6548	0.0980
SeaTemperature^2	0.0041	0.0017	2.3291	0.0199
Location1_A	-0.5534	0.1029	-5.3751	$7.66 \cdot 10^{-8}$
Location1_B	-0.2973	0.0713	-4.1683	$3.07 \cdot 10^{-5}$
Location2_A	-0.3419	0.1302	-2.6247	0.0087
Location2_B	-0.3350	0.0868	-3.8604	0.0001
OperatingModelCoast	0.2284	0.0361	6.3256	$2.52 \cdot 10^{-10}$
OperatingModelFjord	0.4227	0.0708	5.9692	$2.38 \cdot 10^{-9}$
SeasonSpring	-0.3778	0.0355	-10.6330	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.1290	0.0425	-3.0310	0.0024
Biomass	0.0012	0.0001	10.9251	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.2934	0.0438	-6.6931	$2.18 \cdot 10^{-11}$
ProdWeek50	0.4412	0.0382	11.5612	$< 2 \cdot 10^{-16}$
Treatment	-0.0986	0.0390	-2.5297	0.0114
DistToCoast1	-0.0535	0.0355	-1.5092	0.1312
DistToCoast2	0.3227	0.0540	5.9718	$2.35 \cdot 10^{-9}$
Neighbours	0.1483	0.0205	7.2206	$5.18 \cdot 10^{-13}$
LastWeek1	1.5610	0.0404	38.6266	$< 2 \cdot 10^{-16}$
HighPeriod	0.7896	0.0459	17.2063	$< 2 \cdot 10^{-16}$
Log(r)	0.3325	0.0333	9.9967	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z value**	**p-value**
Intercept	-1.7274	0.2909	-5.9380	$2.89 \cdot 10^{-9}$
SeaTemperature	-0.3051	0.0588	-5.1874	$2.13 \cdot 10^{-7}$
SeaTemperature^2	0.0197	0.0031	6.2688	$3.64 \cdot 10^{-10}$
Location1_A	0.0626	0.1737	0.3606	0.7184
Location1_B	0.2691	0.1135	2.3705	0.0178
Location2_A	-0.0491	0.1287	-0.3819	0.7025
Location2_B	-0.6734	0.1123	-5.9966	$2.01 \cdot 10^{-9}$
OperatingModelCoast	1.0877	0.0846	12.8647	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.6384	0.0701	9.1074	$< 2 \cdot 10^{-16}$
SeasonSpring	0.1510	0.0631	2.3915	0.0168
SeasonWinter	0.2173	0.0853	2.5467	0.0109
Biomass	0.0056	0.0003	21.2712	$< 2 \cdot 10^{-16}$
BiomassIndicator	-1.4586	0.1324	-11.0152	$< 2 \cdot 10^{-16}$
ProdWeek50	1.2844	0.0968	13.2691	$< 2 \cdot 10^{-16}$
Treatment	-0.2105	0.0991	-2.1246	0.0336
DistToCoast1	-0.1686	0.0647	-2.6079	0.0091
DistToCoast2	-0.1205	0.0857	-1.4065	0.1596
LastWeek1	3.3572	0.0933	35.9970	$< 2 \cdot 10^{-16}$

AIC=48602.13

The significance of the categorical variables with more than 2 levels in Table 5.10 were tested with the likelihood ratio test, where each of the categorical variables in turn was dropped from the preferred model. The obtained p-values for the categorical variables *Location*, *OperatingModel*, *Season* and *DistToCoast* obtained from the likelihood ratio test are presented in Table 5.11. All the terms in Table 5.11 were significant up to a significance level 0.05, which indicated that they affected the salmon lice count and should not be dropped from the regression model.

Table 5.11: Results from the likelihood ratio test where the significance of the categorical variables in the optimal ZANB sample model for adult female lice (Table 5.10) are tested.

Dropped term	#Df	LogLik	Df	Chisq	p-value
None	39	-24262			
Location from μ_i	35	-24278	4	32.05	$1.87 \cdot 10^{-6}$
OperatingModel from μ_i	37	-24308	2	92.12	$< 2.2 \cdot 10^{-16}$
Season from μ_i	37	-24319	2	113.38	$< 2.2 \cdot 10^{-16}$
DistToCoast from μ_i	37	-24281	2	38.33	$4.75 \cdot 10^{-9}$
Location from π_i	35	-24334	4	144.22	$< 2.2 \cdot 10^{-16}$
OperatingModel from π_i	37	-24401	2	278.71	$< 2.2 \cdot 10^{-16}$
Season from π_i	37	-24266	2	7.93	0.0190
DistToCoast from π_i	37	-24266	2	7.24	0.0268

Residual plots for the zero-inflated negative binomial model and the hurdle negative binomial model are presented in Figure 5.7. The residual plots for the two models were

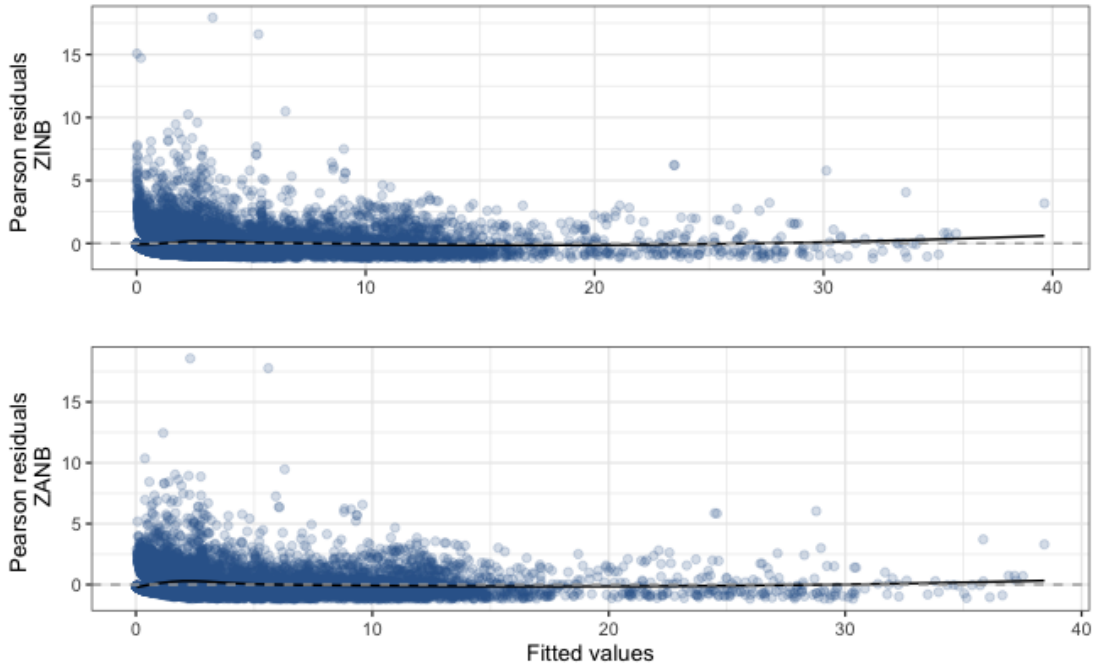

Figure 5.7: Pearson residuals plotted against the fitted values for the regression models ZINB (top) and ZANB (bottom) for the sample model of adult female lice. Because of the large data set, the data points are partly transparent and the black line is a loess smoothing curve.

relatively similar, but there was an improvement in the residuals for the hurdle model. The residuals for fitted values close to zero were smaller for the hurdle model than for the zero-inflated model. The residuals seemed to be quite randomly spread around the horizontal axis for the ZANB model, but there were some high positive residuals for fitted values below 10.

Rootograms for the zero-inflated negative binomial model and the negative binomial hurdle model (Figure 5.8) were obtained by the function `rootogram` from the `R`-package `countreg` (Kleiber & Zeileis, 2016). The smooth red line is the expected counts given from the regression model and the bars, which hangs from the curve, represent the observed counts. By the construction of the hurdle models, the expected number of zeros

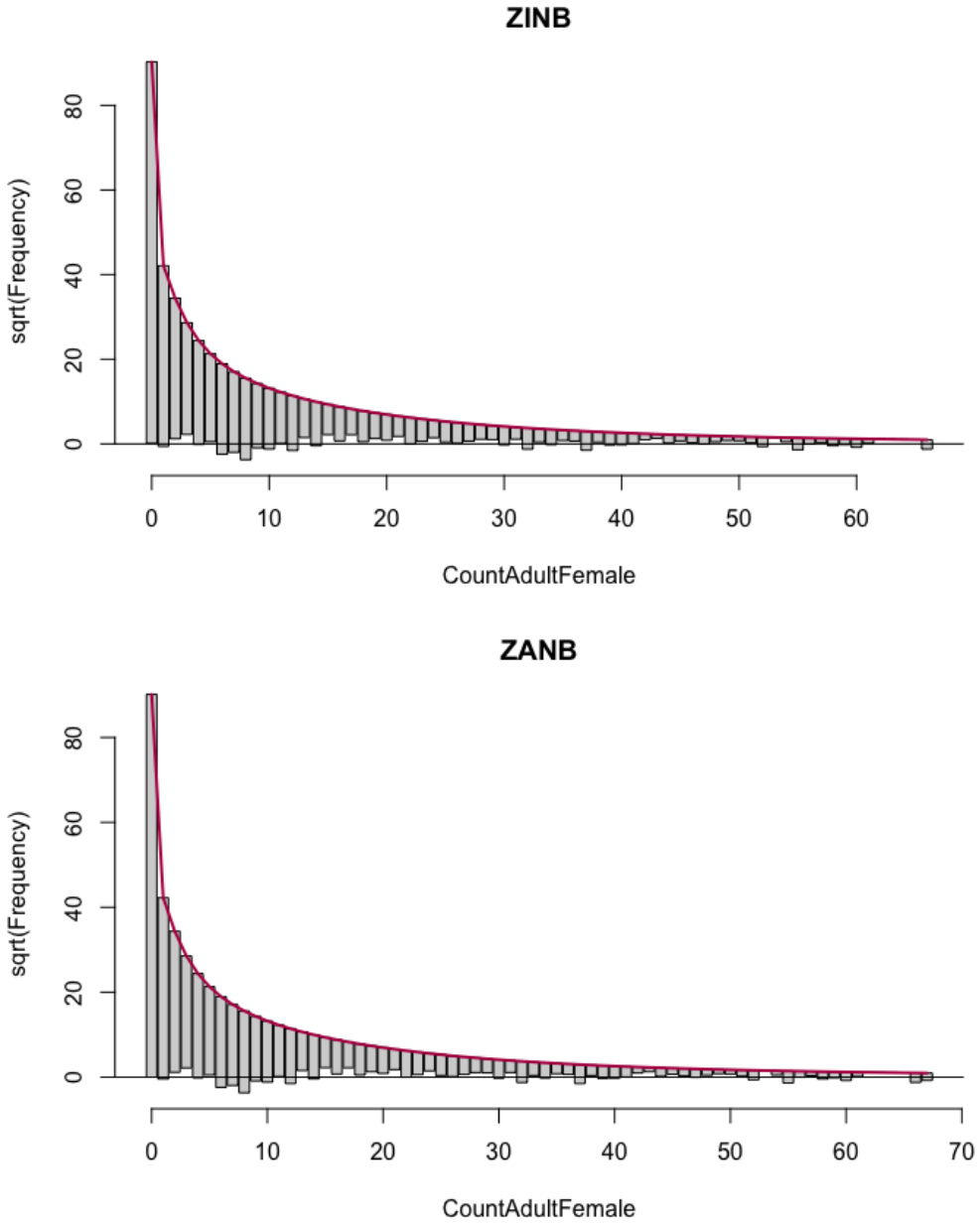

Figure 5.8: Rootograms for the sample regression models ZINB and ZANB.

corresponded to the observed number of zeroes. There were no big differences between the two models, and for both models the bar exceeded the zero line for counts 6-8, which indicated that the models were underfitting at this counts. There were no major deviations from observed and predicted values, but there was a slight tendency of overfitting for counts between 12 and 30, and underfitting for higher counts. As visualized in Figure A.19, the highest number of adult female lice counted on 20 salmon was 160. The models was thus also underfitting for counts above 70, but this only applied to 11 observations.

The estimated regression coefficients of the zero-inflated negative binomial model in Table 5.7 were compared with the estimated regression coefficients of the negative binomial hurdle model in Table 5.10. For the significant parameters in the count part, the sign and the magnitude were very similar. The biological conclusion for the count part was thus the same for both the models. For the logit part, there were a few more differences for the two models. For the ZINB model *ProdWeek50* and *Season* were dropped, while *Location1A*, *Location2A*, *DistToCoast1* were included with a p-value < 0.05. The zero-inflated coefficients of *Neighbour* and *HighPeriod* were significant in the ZINB model, but were dropped from the ZANB model as they were non-significant. The dropped terms in the ZINB model, *ProdWeek50* and *Season*, were significant in the hurdle model. Same for the both models, *Location1A* and *Location2A* were not significant up to a significance level 0.05, while the common variable *Location* were significant according to the likelihood ratio test. The magnitude of the significant parameters were in most cases similar, but the sign changed for all the parameters without *Treatment*. The definition of π_i for the ZINB model and the ZANB model are, as presented in Section 2.2.2, different. It was therefore expected that the zero-inflation regression coefficients changed sign for the two models. The odds for a false zero in ZINB was given by $\frac{\pi_i}{1-\pi_i} = \exp(\eta_i)$, while the odds for a zero count in ZANB was given by $\frac{\pi_i}{1-\pi_i} = \exp(-\eta_i)$.

From the estimated regression coefficients from the model fit of the negative binomial hurdle model in Table 5.10, the odds for a zero count could be estimated as

$$\begin{aligned}\ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) &= 1.727 + 0.305 \cdot \text{SeaTemperature} - 0.020 \cdot \text{SeaTemperature}^2 \\ &\quad - 0.063 \cdot \text{Location1A} - 0.269 \cdot \text{Location1B} + 0.049 \cdot \text{Location2A} \\ &\quad + 0.673 \cdot \text{Location2B} - 1.088 \cdot \text{OperatingModelCoast} \\ &\quad - 0.638 \cdot \text{OperatingModelFjord} - 0.151 \cdot \text{SeasonSpring} \\ &\quad - 0.217 \cdot \text{SeasonWinter} - 0.006 \cdot \text{Biomass} + 1.459 \cdot \text{BiomassIndicator} \\ &\quad - 1.284 \cdot \text{ProdWeek50} + 0.211 \cdot \text{Treatment} + 0.169 \cdot \text{DistToCoast1} \\ &\quad + 0.121 \cdot \text{DistToCoast2} - 3.357 \cdot \text{LastWeek1}.\end{aligned} \tag{5.2}$$

The odds for a zero count for a biomass above 490 metric ton would thus increase with a factor $\exp(1.459) = 4.300$ over the odds for a zero count for biomass below 490 metric ton. From the estimated count coefficients in Table 5.10, the expected count of adult female lice would decrease with a factor $\exp(-0.293) = 0.746$ if the biomass in the cage exceeded 490 metric ton and all other terms were kept constant. In addition, the odds of a zero count would decrease with a factor $\exp(-0.006) = 0.999$ and the expected count of adult female lice would increase with a factor $\exp(0.001) = 1.001$ if the biomass in the cage increased with one unit (1 metric ton) and all other coefficients were kept constant.

5.6 Time Series Analysis

The number of adult female lice counted on 20 salmon each week during the study period in a selected cage forms a time series. The reported counts for one of the cages, where it had been started 7 production cycles during the study period, are presented in Figure 5.9. Several of the production cycles had been operated according to the stage model, so the salmon had been moved from the starting cage after a while. The lice numbers were then obtained from the growth cage that contained the salmon from the given starting cage. After the salmon had been moved to the growth area and the salmon farm had been empty for a while, a new production cycle could be started and smolts could be deployed in the cage. Therefore, an overlap in the data could be seen, and separate time series for each production were studied.

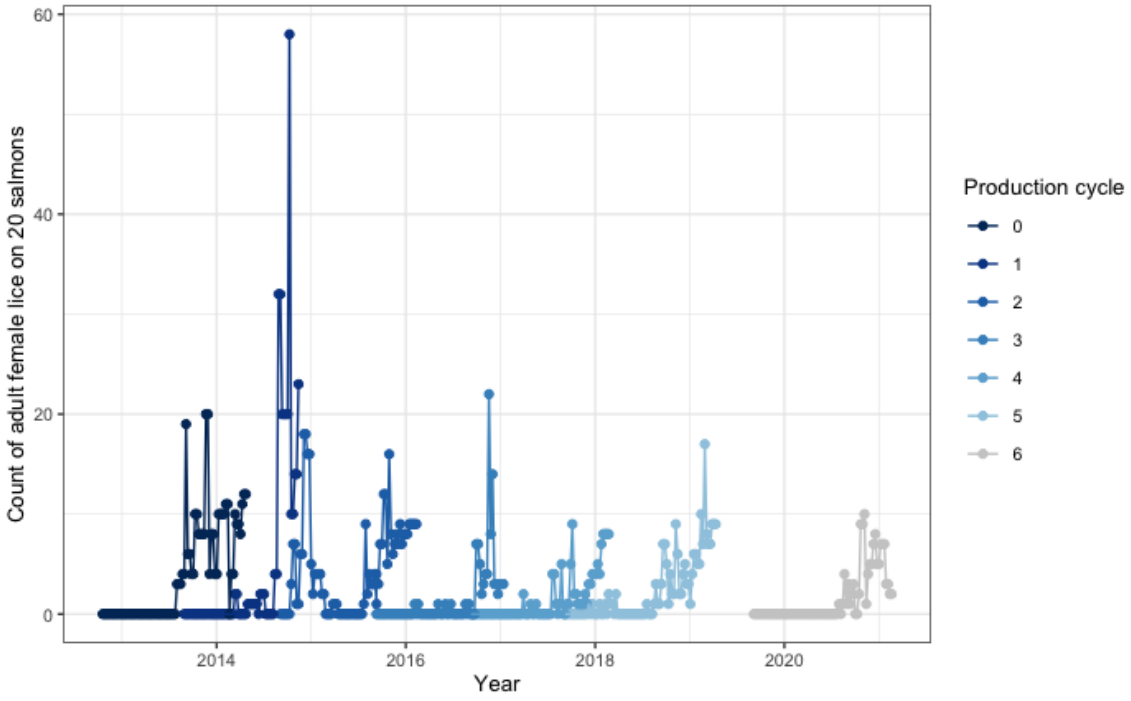

Figure 5.9: The weekly count of adult female lice on a sample of 20 salmon in a given cage throughout the study period. The points are grouped after production cycles.

The residuals from the negative binomial hurdle model belonging to the selected cage were studied to see if there was any correlation in the residuals. The auto correlation function (ACF) of the residuals from production cycle 3,4 and 5 are presented to the left in Figure 5.10. From these, it was clear that the time correlation varied for the different production cycles. By dropping the explanatory variable *LastWeek1* in both the count and the zero-inflation part, the residuals were expected to be more correlated. Following Aldrin et al. (2019), the count of adult females for a given week and cage depends on the survival of adult females from previous week and recruitment of females from mobile lice. For high lice numbers, it was thus expected that if no delousing had been used, the data would be more correlated than for low lice numbers. The auto correlation function of the residuals from the ZANB model without last weeks reported lice number is presented to the right in Figure 5.10. The residuals were collected from the same production cycles as for the

full ZANB model, and for production cycle 4 and 5 an increase in correlation between the residuals could be seen. This indicated that *LastWeek1* reduced the correlation in the residuals, but there was still some correlation in the residuals.

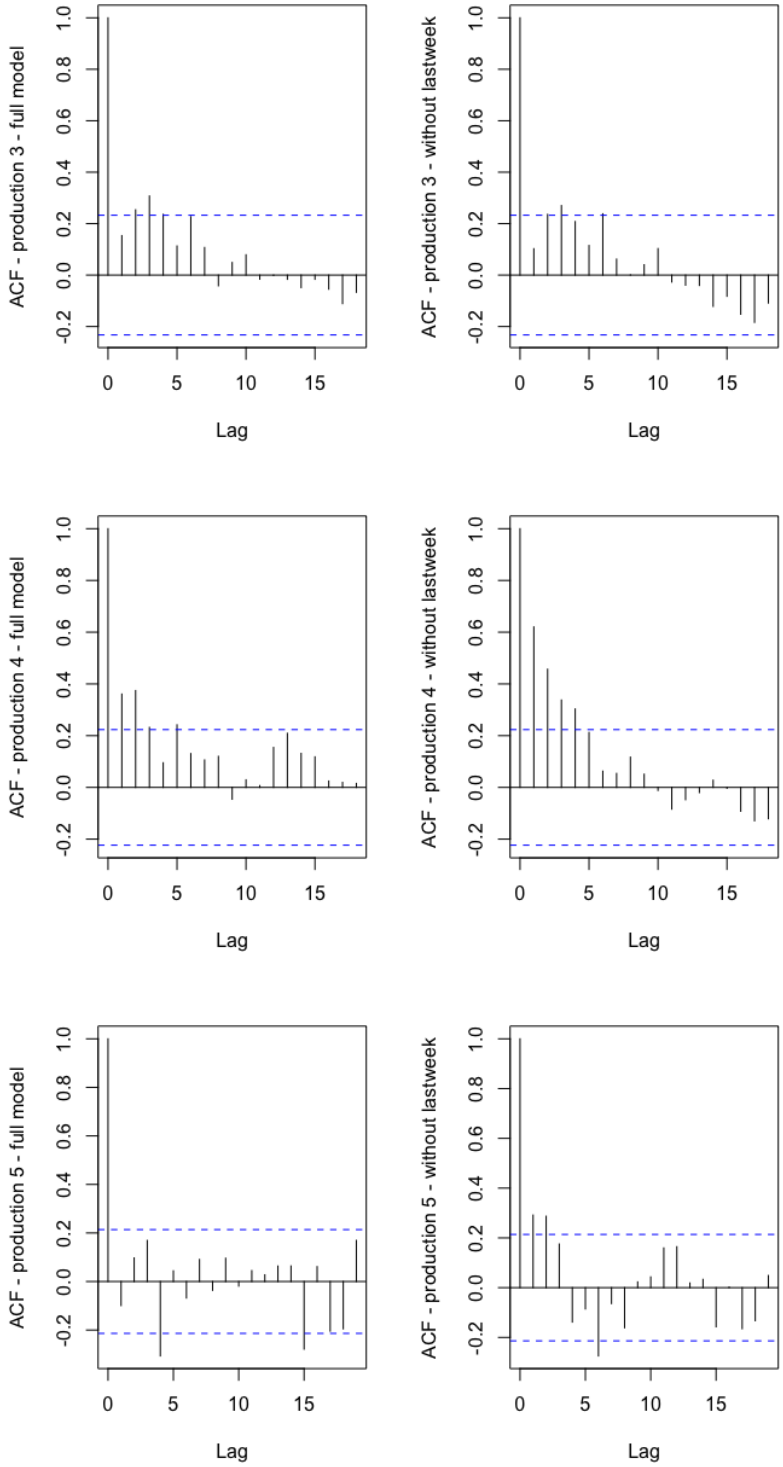

Figure 5.10: Auto correlation function plot of the residuals from the negative binomial hurdle regression model with (left) and without (right) the explanatory variable LastWeek1. The plots corresponds to production cycle 3,4 and 5 in Figure 5.9

5.7 Evaluation of the Sample Size

As presented in Section 1.1.2, *The Ministry of Trade, Industry and Fisheries* states that salmon lice should be counted on at least 10 random salmon in each cage each week in operation. There are some exceptions from these regulations, and during 6 weeks in the spring the salmon lice should be counted on at least 20 random salmon in each cage (Forskrift om lakselusbekjempelse, 2012). To assess whether a sample of 10 salmon were enough to obtain a representative average for the cage, hypergeometric distribution was used. The lice number gathered in this thesis were mainly based on a sample of 20 random salmon in each cage, so the difference between a sample size of 10 and 20 salmon was also investigated.

It was assume that there was no distinction between the number of lice on the salmon, but only whether salmon lice was observed or not. Thus, for the hypergeometric distribution, a salmon with lice was seen as a success, while a salmon without lice was seen as a failure. The number of salmon in the cage was set to be $N = 140000$, which was close to the average number of salmon in a cage (Table 4.1). The probability of drawing at least one salmon with salmon lice from a cage with $N = 140000$ salmon, where one tenth of the salmon has salmon lice, $m = 14000$, is plotted against the number of draws in Figure 5.11. The probability of observing salmon lice during 10 draws was calculated as 65.1%, while the probability increased to 87.8% by increasing the number of draws to 20. This indicated that the count was more representative for the lice pressure in the cage for higher number of draws, but the effect of increasing the number of draws with one unit was highest in the beginning and declined for values above 15.

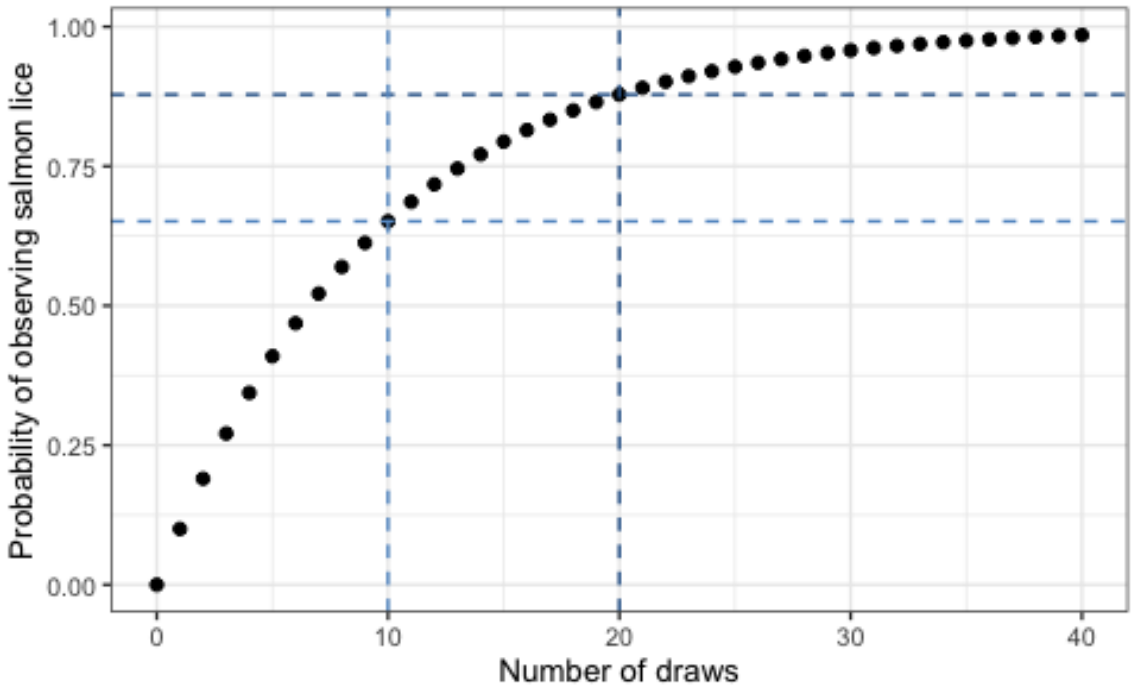

Figure 5.11: The probability of observing salmon lice in at least one draw ($P \neq 0$), plotted against the total number of draws, without replacement, from a sample size of 140000 salmon where 1/10 has salmon lice.

The probability of not observing salmon lice in 20 draws without replacement from a cage with 140000 salmon for different ratios of salmon lice are presented in Table 5.12. How many salmon with salmon lice this corresponded to for a sample size of 20 salmon is represented in column x. The probability of getting the same sample ratio as the given ratio for the whole cage was estimated as $P(X = x)$. The probability that zero lice was observed in 20 draws if one quarter of the salmon had lice was 0.32%, while if there were salmon lice on 1/20 of the salmon the probability was 35.8%. Both the probability of observing zero and the probability of getting the same sample ratio as the true ratio increased with a decreasing amount of lice in the cage. For ratios smaller than 1/20, the number of salmon with salmon lice needed to get a representative sample ratio were smaller than 1. For the last three rows, the number was closer to zero than one, and the probability $P(X = 0)$ was thus most appropriate to use. One salmon less without lice in the sample would for small ratios have a great effect on the sample ratio.

Table 5.12: The probability of not observing salmon with salmon lice in $n = 20$ draws without replacement calculated with the hypergeometric distribution for different cases. The finite population size are $N = 140000$ and the ratio are defined as the number of salmon lice divided by the number of salmon in the cage. The probability of getting the same sample ratio as the ratio for the whole cage, are estimated as $P(X = x)$, where x are nearest integer of $20 \cdot ratio$.

Ratio	m	N-m	P(X=0)	x	P(X=x)
1/2	70000	70000	$< 0.001\%$	10	17.6%
1/3	46667	93333	0.003%	7	18.2%
1/4	35000	105000	0.32%	5	20.2%
1/8	17500	122500	6.92%	3	23.0%
1/10	14000	126000	12.2%	2	28.5%
1/16	8750	131250	27.5%	1	36.7%
1/20	7000	133000	35.8%	1	37.7%
1/40	3500	136500	60.3%	1	30.9%
1/80	1750	138250	77.8%	1	19.7%
1/200	700	139300	90.5%	1	9.09%

5.8 Number of Delousing Treatments Performed

The number of delousing treatments that have been performed for each of the operating methods were used in the comparison of the different operating strategies, as this indicates how the lice pressure has been. The number of delousing treatments performed on the cages operated according to the stage models was 364, while 251 and 194 treatments had been performed for the coast model and the fjord model, respectively. Only the first week of a treatment was included in the count, such that treatments that ran over two weeks were not counted twice. The amount of data from the different operating models varied, so it was most appropriate to use the proportion presented in Figure 5.12. The proportion was related to how many weeks each of the operating models had been in operation, and thus the maximum number of delousing treatments that could have been performed for each of the models. The operating model with the highest number of cages treated against

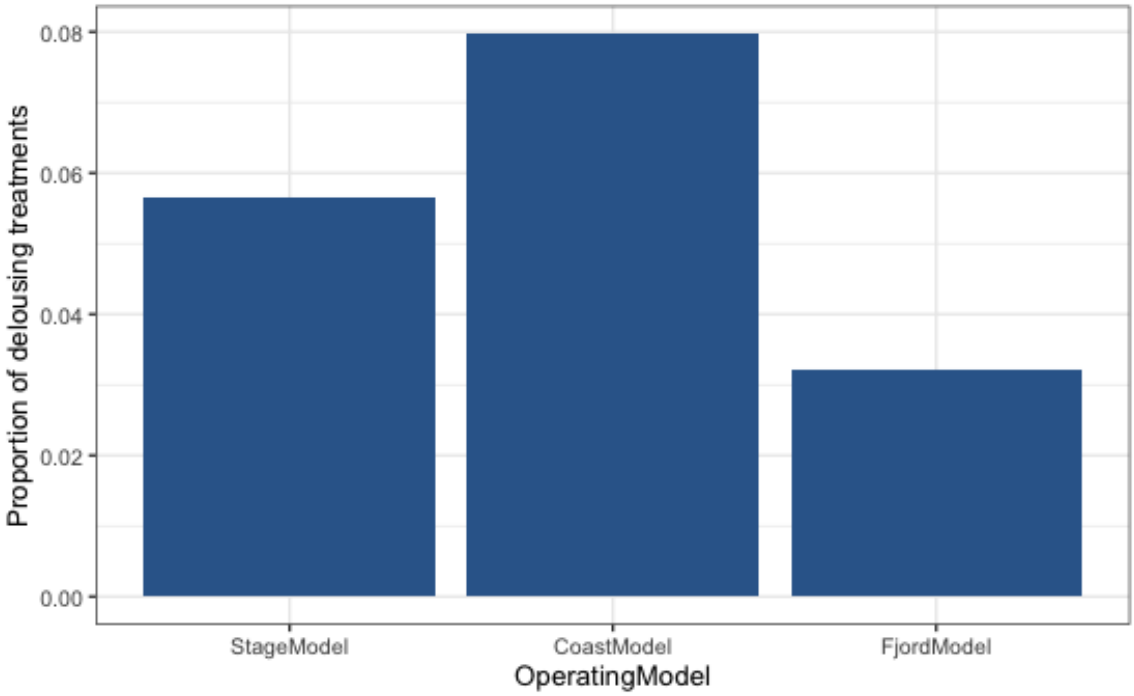

Figure 5.12: The proportion of delousing treatments relative to the number of weeks in operation for each of the operating models.

salmon lice was the stage model, but from the figure there were more frequent delousing treatments on the cages operated according the coast model. Delousing treatments were most rarely performed at salmon that had been in the fjord throughout the production.

5.9 Violations of the Lice Limit

The allowed limit of adult female lice per salmon is mainly 0.5 (0.2 in the six weeks presented in Section 1.1.2), and has to be calculated for a sample of at least 10 salmon (Forskrift om lakselusbekjempelse, 2012). Violations of this limit were used to compare the lice pressure for the different operating models. Number of violations are represented in Table 5.13, together with the number of counts performed at cage level. Only the weeks when the lice numbers went from legal value to above 0.5 (or 0.2 during the 6 weeks specified in Section 1.1.2) were registered as violations. Subsequent weeks where the lice number was above the lice limit were not included, as these depended on the effect of the delousing and not the operating models. There were different amounts of data for the different operating models, so the ratio between violations and the number of counts was used in the comparison of the operating models. The number of counts reflected the total number of weeks the cages had been operated according to the given model.

Table 5.13: Number of violations of the lice limit for each of the operating models and the locations of the salmon farms. Counts represent the total number of registrations performed at cage level for each of the operating models/locations. The percentage represents the ratio between violation of the lice limit and the number of registrations.

OperatingModel	Location	Stage	Violation	Counts	Violation %
StageModel	1_B	start	15	2303	0.65
	2_A		0	417	0.00
	2_B		0	735	0.00
	Fjord	start	16	3455	0.46
	Coast	growth	190	2972	6.39
	Fjord + Coast		**206**	**6427**	**3.21**
CoastModel	Coast	normal	**142**	**3149**	**4.51**
FjordModel	1_A	normal	25	538	4.65
	1_B		59	3013	1.96
	2_A		15	446	3.36
	2_B		59	2023	2.92
	Fjord	normal	**158**	**6020**	**2.62**

From the table, the coast model had the highest number of violations of the lice limit relative to the number of counts (4.51%). The stage model violated the limit of adult female lice in 3.21% of the counts, while 2.62% of the counts in the fjord model violated the limit. Most of the violations of the lice limit for the stage model occurred after the salmon had been moved from the fjord to the coast. For the stage model, violations of the lice limit for cages in Fjord 1 had been observed in 16 cases, while the lice numbers for Fjord 2 remained below the limit throughout the start stage. The ratio for the growth stage in the stage model was 6.39%, while the ratio for the coast model was 4.51%. From Figure 4.9, the lice numbers increased after around 50 weeks in production, which was around the same time the salmon were moved to the growth stages. It therefore made sense that the growth stage, which was the last part of the production for the stage model, had a higher ratio of violations of the lice limit, than the coast model, which was based on the entire production cycle.

6 Discussion

6.1 Remarks on Fitted Models

Alternative response variables with lice counts of adult female lice and all mobile lice for a sample of 20 salmon and the estimated total number of salmon lice in the cage have been tested. There were a lot of uncertainty linked to the observed values for the cage model, as these were estimated as the number of lice calculated from a small sample multiplied by the number of salmon in the cage. The probability of not observing salmon lice during 20 draws from a cage with 140 000 salmon, where one out of twenty has salmon lice, was 35.8% (Table 5.12). This indicated that even if no lice had been counted on 20 salmon, there may was lice in the cage, and the estimated number of lice in the cage thus became too inaccurate. There were also great variation between the observed values in the cage model (min 0 - max 1298430), and thus it became difficult to adapt a suitable model. The cage model for adult female lice was not expected to predict a single zero. By basing the regression model on the sample count of adult female lice instead, the model fit were largely improved, and the sample model was thus preferred for modelling the salmon lice. Due to the large number of zeros in the observed counts, it was tested whether a model for all mobile lice would fit the data better. The number of observed zeros was thus reduced, but the ratio between observed and predicted zeros was not improved. Since the count of adult female lice was most interesting to model, the sample count of adult female lice stood out as the preferred response variable in the analysis.

The sample model for adult female lice were fitted by the generalised linear models Poisson and negative binomial, the zero-inflated Poisson and negative binomial models and the Poisson and negative binomial hurdle models. The sign of the estimated regression coefficients were the same for all the models, while there were some differences in the magnitude and the significance of the coefficients, especially between the GLMs and the models for zero-inflation. The model statistics from the model fit of the full sample model for the different regression models are summarized in Table 6.1. The Poisson model was clearly inferior to all other fits, and the GLM negative binomial improved the fit dramatically. The model was further improved by the zero-inflated and the hurdle models with the negative binomial distribution for the counts. This indicated that a model which handled zero-inflation and extra variance in the count data seemed to be appropriate for the lice count data.

Table 6.1: Model comparison for the fitted regression models for the full sample model for adult female lice.

	Poisson	Neg.bin	ZIP	ZINB	ZAP	ZANB
Degrees of freedom	20	21	40	41	40	41
Log-likelihood	-36683	-25147	-32384	-23923	-32401	-24261
Expected number of 0	5748	7810	6580	7788	8125	8125
AIC	73406	50335	64848	47938	64882	48604

The ZINB model and the ZANB models gave almost identical fits and lead to the same qualitative results. There was a slight improvement in the residual plot for the ZANB model compared with the residuals from the ZINB model. The interpretation of the hurdle model was also nicer, since the hurdle model is a two-part model, where one part generates whether lice have been observed and the other determines how many lice that have been observed. The cause of the excessive number of zeros should also be considered when choosing between the zero-inflated model and the hurdle model. The hurdle models do not discriminate between the different types of zeroes, while the zero-inflation models discriminate between true and false zeroes. There may were some zeros in the observed dataset that were false zeros due to observed errors, but in most cases the recorded zeroes were correct and there were no lice on the salmon in the sample. It was therefore most natural to choose a hurdle model, which in the zero part looked at the probability of the absence or presence of lice, and not the probability of false zeroes as in the zero-inflated model. The negative binomial hurdle model was thus the preferred regression model for the sample count of adult female lice.

From the rootograms in Figure 5.8, the negative binomial hurdle model did not predict as high values as the highest observed lice counts. This also appeared in the residual plot (Figure 5.7), where there were some high positive residuals that stood out. Otherwise, the residuals seemed to be randomly spread around the horizontal axis. By including an uncensored variable of last week's lice numbers, the model would predict higher lice numbers, but then far too high. For the purpose of evaluating the effect of the stage model and other factors, it was the lowest lice numbers that were most interesting, since their impact on lice numbers was greatest under low lice pressure. The estimated count coefficient for *LastWeek1* was 1.56 for the ZANB model. The data for this variable was the last weeks lice number, which was the average number of lice per salmon calculated on a sample of the cage. An increase of one unit in *LastWeek1* was therefore the same as that 20 extra salmon lice had been counted on a sample of 20 salmon. Thus, if all other coefficients were kept constant, and the last weeks observed number of salmon lice on a sample of 20 salmon increased with one, the expected number of lice on 20 salmon was in the ZANB model estimated to increase with a factor $\exp(1.56/20) = 1.081$ while the odds for zero lice would decrease with a factor $\exp(-3.357/20) = 0.845$.

The estimated regression coefficients from the ZANB regression model, seen in Table 5.10, were compared with the plots in Section 4. From the plot of adult female lice and sea temperature (Figure 4.2), it was expected a quadratic trend of the sea temperature, where the lice number increased with the temperature up to around $10°C$, and then decreased again. The linear term for the sea temperature in the count-part was not significant for a significance level 0.05. A quadratic term of the sea temperature was thus probably not the best way to model the sea temperature, but since most of the explanatory variables were categorical, a numerical variable with a linear and a quadratic term were chosen in this analysis. Otherwise, all the estimated regression coefficients for the count corresponded well to the visualization in Section 4.

There were several potential problems with the analysis of the salmon lice data. The data were reported from different salmon farms and there may have been some dependency between the data from same generation or site. From Figure 5.10, most of the temporal correlation within each of the residuals time series were removed by adding last weeks

reported lice number of all mobile lice to the regression model. The study period extended over a period of nine years, so there may have been changes in the operation and routines throughout the study period that have affected the lice numbers. There were no available data on this, and it was therefore assumed that there had been no major differences in operations outside the factors that were included in the model. In order to achieve a more accurate model, environmental factors such as salinity, wind and current should also have been included in the model. The regression model used in this thesis included most categorical variables, so by including continuous environmental factors the lice numbers could have been analysed more dynamically.

6.2 Comparison of the Site Locations

Figure 4.3 indicated that the lice numbers were highest along the coast and lowest in the inner part of the two fjords. The coast sites were used as a reference in the regression model, and from the estimated ZANB regression coefficients in Table 5.10, there was expected a significant decrease in the lice count if the cage was moved from the coast to one of the two fjords. The lice count was expected to decrease most if the cage was moved from the coast to Fjord 1A, which from Table 3.1 had the lowest average salinity. The inner part of Fjord 1 was also, according to Figure 4.3, the location with lowest lice numbers. This corresponds with the results in Torrissen et al. (2013), which shows that salmon lice prefer high-salinity sea water. If the cage was moved from the coast to the inner part of Fjord 1 and all other terms were kept constant, the expected count of lice on a sample of 20 salmon would decrease with a factor $\exp(-0.553) = 0.575$. The odds for a zero-count for the inner part of the two fjords were not significantly different from the odds for a coast site, while the odds for absence increased for Fjord 2B and decreased for Fjord 1B. For Fjord 1A, most of the lice numbers in Figure A.5 were larger than zero. This corresponded to the estimated regression coefficients, but there were no clear differences in the amount of zeros for Fjord 2A and the outer part of the two fjords in the figure. The results from the model fit of the ZANB model was thus a bit surprising, and would have been interesting to study further.

The smallest factor change in the expected count of lice due to changed location was $\exp(-0.297) = 0.743$, and occurred if all other terms were fixed and the cage was moved from the coast to Fjord 1B (Table 5.10). The outer part of Fjord 1 was also the location in Figure 4.3 that had the highest lice numbers of the fjord locations. According to the gathered salinity data (Table 3.1), the average salinity was highest for Fjord 2B. There were some shortcomings in the salinity data, so the salinity for each of the locations should have been studied in more detail with an improved dataset. By including environmental measurements as salinity and currents in the regression model, the indicator variable for the location of the cage could give an indication of whether there were other differences between the two fjords.

6.3 Evaluation of the Stage Model

The estimated regression coefficients for the ZANB sample model in Table 5.10 indicated that the count of adult female lice was expected to increase if a coast model or a fjord model were used instead of the stage model. The lice count on a sample of 20 salmon were from the ZANB model expected to increase by a factor $\exp(0.228) = 1.26$ if a coast model was used instead of a stage model, and all other terms were fixed. If a fjord model was used instead of the stage model, the lice count was expected to increase by a factor $\exp(0.423) = 1.53$. A location along the coast was used as reference, and for all four fjord locations, the count of lice was expected to be lower than at the coast. A cage operated after the coast model was always located by the coast, and the other location levels in *Location* were thus irrelevant for the coast model. For the fjord model, it was expected that the count of lice, in addition to being increased by a factor of 1.53 compared to the stage model, would be reduced by a factor between 0.58 and 0.74 depending on where in the two fjords the cage was located. It was therefore the coast model that had the highest expected lice count if both the location and the operating model were considered.

The probability of absence of lice were studied from the estimated zero-inflated regression coefficients in the ZANB model. The odds of a positive count on a sample of 20 salmon was for the coast model and the fjord model increased by a factor $\exp(1.088) = 2.97$ and $\exp(0.638) = 1.89$, respectively, if the stage model were used as reference and all other terms were fixed. Thus, there were less likely to observe lice on a sample of 20 salmon from a cage operated according to the stage model than from a cage with whole-generation operation along the coast or in the fjord.

The number of delousing treatments and violations of the lice limit (0.5 adult female lice per salmon) were used as an indicator for the lice pressure for the different operating methods. From Table 5.13, 3.21% of the lice numbers from the stage model violated the limit, while 4.51% of the coast model and 2.62% of the fjord model violated the lice limit. From Figure 5.12, the proportion of delousing treatments relative to the number of weeks in production for each of the operating models stage, coast and fjord were 5.66%, 7.97%, 3.22%, respectively. The number of treatments was higher than the violations, which is natural, since the salmon should be deloused before the limit is exceeded. If the limit is exceeded, several different treatments may be needed before the lice number is at an acceptable level. Both the ratio of the violations of the lice limit and the proportion of delousing treatments indicates that the lice pressure is highest for the coast model and lowest for the fjord model.

6.4 Conclusion and Further Work

The negative binomial hurdle model for the sample count of adult female lice seemed to fit the lice count data best, and were used to evaluate the effect of a shortened sea phase. From the estimated regression coefficients, the stage model seems to be significantly better than the coast model and the fjord model, as a lower lice count is expected for the stage model. The odds of absence of lice is higher for the stage model than for the coast model and the fjord model. Both the estimated regression coefficient, the number of violations of the lice limit and the proportion of delousing treatments indicates that there is a lower lice pressure for cages operated in the fjord versus along the coast.

There were fewer delousing treatments and violations of the lice limit associated with the fjord model compared with the stage model. There have been no consecutive whole-generation productions in the two fjords, so a possible increasing lice situation for future productions in the fjords due to that there is a lot of salmon in the fjord throughout the year, was not visible in this analysis. Thus, the full effect of the shortened sea phase, that is achieved by moving the salmon to the coast after 7 to 9 months of production in the fjord, could not be assessed. The regression model indicates that keeping the salmon in inner fjords systems has good effect on the lice pressure, but the effect of moving the salmon out instead of keeping it in the fjord should have been investigated further.

For further work, it would therefore be interesting to set up a research project across several production cycles for the fjord model and the stage model, where environmental variables and other factors that may affect the lice pressure are reported. This would improve the database, and the differences between the fjord model and the stage model could have been studied more accurate. Then the infection dynamics between sites and lice development in the two fjords by the various operating methods could also be investigated. Several of the salmon farms that were considered as neighbour sites in this analysis may not have been in operation for several years, and should therefore not have been considered as a neighbour site in the regression model. The number of neighbors should therefore also be reported during a possible research project, so that the neighbor variable becomes more accurate. By including several variables, a more general regression model would be obtained, and the count of lice would be described more dynamically using numerical variables such as salinity and wind.

In this study, only lice numbers were used as a basis in the evaluation of the stage model, but there are also other factors that should be taken into account when such a operating strategy is considered. Such as the salmon growth and the risk of spreading infections associated with transporting the salmon from inner fjord systems to the coast (Veiledning: Flytting av laksefisk mellom oppdrettsanlegg, 2019). The environmental impacts and the lice pressure on wild fish are also factors that should be consider when comparing the operating models.

Bibliography

Aldrin, M., Jansen, P. & Stryhn, H. (2019). A partly stage-structured model for the abundance of salmon lice in salmonid farms. *Epidemics*, *26*, 9–22. https://doi.org/https://doi.org/10.1016/j.epidem.2018.08.001

BarentsWatch. (2021a). *Download fish health data*. https://www.barentswatch.no/nedlasting/fishhealth/lice?lang=en

BarentsWatch. (2021b). *Fishhealth*. https://www.barentswatch.no/fiskehelse/

Brakestad, I. H. (2020). *Analysis of salmon lice count data for production zones 6 and 7 in norway from 2017 to 2019* (Master's thesis). Norwegian University of Science and Technology.

Bricknell, I. R., Dalesman, S. J., O'Shea, B., Pert, C. C. & Luntz, A. J. (2006). Effect of environmental salinity on sea lice *Lepeophtheirus salmonis* settlement success. *Diseases of aquatic organisms*, *71*(3), 201–212. https://doi.org/https://doi.org/10.3354/dao071201

Cameron, A. C. & Trivedi, P. K. (1998). *Regression analysis of count data.* Cambridge University Press. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=120799&site=ehost-live

Casella, G. & Berger, R. (2002). *Statistical inference* (2nd ed.). Brooks/cole, Cengage Learning.

Fahrmeir, L., Kneib., T., Lang, S. & Marx, B. (2013). *Regression. models, methods and application*. Springer-Verlag Berling Heidelberg.

Fiskehelsedirektoratet. (2021). *Akvakulturregisteret* [The Norwegian Directorate of Fisheries]. https://www.fiskeridir.no/Akvakultur/Registre-og-skjema/akvakulturregisteret

Forskrift om lakselusbekjempelse. (2012). *Forskrift om bekjempelse av lakselus i akvakulturanlegg* [Regulations on salmon lice control. Regulations on combating salmon lice in aquaculture plants]. https://lovdata.no/dokument/SF/forskrift/2012-12-05-1140

Fox, J. & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gaasø, M. (2019). *Sea lice (Leopeophtheirus salmonis and Caligus elongatus) during freshwater treatment* (Master's thesis). Norwegian University of Science and Technology. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2656633

Hamre, L. A., Eichner, C., Caipang, C. M. A., Dalvin, S. T., Bron, J. E., Nilsen, F., Boxshall, G. & Skern-Mauritzen, R. (2013). The salmon louse *Lepeophtheirus salmonis* (copepoda: Caligidae) life cycle has only two chalimus stages. *PLOS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0073539

Hemmingsen, W., MacKenzie, K., Sagerup, K., Remen, M., Bloch-Hansen, K. & Dagbjartarson Imsland, A. K. (2020). *Caligus elongatus* and other sea lice of the genus *Caligus* as parasites of farmed salmonids: A review. *Aquaculture*, *522*, 735160. https://doi.org/https://doi.org/10.1016/j.aquaculture.2020.735160

Heuch, P. A., Olsen, R. S., Malkenes, R., Revie, C. W., Gettinby, G., Baillie, M., Lees, F. & Finstad, B. (2009). Temporal and spatial variations in lice numbers on salmon farms in the hardanger fjord 2004–06. *Journal of Fish Diseases*, *32*(1), 89–100. https://doi.org/https://doi.org/10.1111/j.1365-2761.2008.01002.x

Hijmans, R. J. (2019). *Geosphere: Spherical trigonometry* [R package version 1.5-10]. https://CRAN.R-project.org/package=geosphere

Hilbe, J. (2011). *Negative binomial regression.* (Vol. 2nd ed). Cambridge University Press. http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=357443&site=ehost-live

Igboeli, O. & Burka, J. (2013). *Lepeophtheirus salmonis*: A persisting challenge for salmon aquaculture. *Animal Frontiers*, *4*, 22–32. https://doi.org/10.2527/af.2014-0004

International Organization for Standardization. (2019). *Iso 8601-2:2019*. https://www.iso.org/standard/70908.html

Jevne, L. S. (2020). *Development and dispersal of salmon lice (Lepeophtheirus salmonis krøyer 1837) in commercial salmon farming localities* (Doctoral thesis). Norwegian University of Science and Technology. https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2656444/Lone%20Sunniva%20Jevne.pdf?sequence=3&isAllowed=y

Karlsen, C. (2020). *Analysis of salmon louse count data from salmon farms in inner fjord systems* (Project assignment), Norwegian University of Science and Technology.

Kleiber, C. & Zeileis, A. (2016). Visualizing count data regressions using rootograms. *The American Statistician*, *70*(3), 296–303. https://doi.org/10.1080/00031305.2016.1173590

Montgomery, D. C. & Peck, E. A. (1983). *Introduction to linear regression analysis*. John Wiley Sons.

Myhre Jensen, E., Horsberg, T. E., Sevatdal, S. & Helgesen, K. O. (2020). Trends in delousing of norwegian farmed salmon from 2000–2019—consumption of medicines, salmon louse resistance and non-medicinal control methods. *PLOS ONE*, *15*(10), 1–17. https://doi.org/10.1371/journal.pone.0240894

NCSS Statistical Software. (2021). Zero-inflated poisson regression [LLC. Kaysville, Utah, USA, Chapter 321]. https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Poisson_Regression.pdf

Norwegian Seafood Research Fund. (n.d.). *Kampen mot lusa – en samlet oversikt*. https://www.fhf.no/resultater/utvalgte-tema/lakselus/

O'Driscoll, D. & Ramirez, D. E. (2015). Response surface designs using the generalized variance inflation factors (G. Zou, Ed.). *Cogent Mathematics*, *2*(1), 1053728. https://doi.org/10.1080/23311835.2015.1053728

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T. S., Gismervik, K. & Stien, L. H. (2019). Salmon lice treatments and salmon mortality in norwegian aquaculture: A review. *Reviews in Aquaculture*, *11*(4), 1398–1417. https://doi.org/https://doi.org/10.1111/raq.12299

Padgham, M., Rudis, B., Lovelace, R. & Salmon, M. (2017). Osmdata. *The Journal of Open Source Software*, *2*(14). https://doi.org/10.21105/joss.00305

Poppe, T., Bergh, Ø. & Keeping, D. (1999). Del 12. behandling. *Fiskehelse og fiskesykdommer* (pp. 334–337). Universitetsforl.

Torrissen, O., Jones, S., Asche, F., Guttormsen, A., Skilbrei, O. T., Nilsen, F., Horsberg, T. E. & Jackson, D. (2013). Salmon lice – impact on wild salmonids and salmon aquaculture. *Journal of Fish Diseases*, *36*(3), 171–194. https://doi.org/https://doi.org/10.1111/jfd.12061

Veiledning: Flytting av laksefisk mellom oppdrettsanlegg. (2019). *Veiledning om rammene for flytting av laksefisk mellom oppdrettsanlegg* [Guidance: Moving salmonids between salmon farms]. https://www.mattilsynet.no/fisk_og_akvakultur/akvakultur/drift_av_akvakulturanlegg/mattilsynet_brev_om_veiledning_om_rammene_for_flytting_av_

laksefisk_mellom_oppdrettsanlegg.34412/binary/Mattilsynet%20Brev%20om%20veiledning%20om%20rammene%20for%20flytting%20av%20laksefisk%20mellom%20oppdrettsanlegg

Yang, S., Harlow, L., Puggioni, G. & Redding, C. (2017). A comparison of different methods of zero-inflated data analysis and an application in health surveys. *Journal of Modern Applied Statistical Methods*, *16*, 518–543. https://doi.org/10.22237/jmasm/1493598600

Zeileis, A. & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. https://CRAN.R-project.org/doc/Rnews/

Zeileis, A., Kleiber, C. & Jackman, S. (2008). Regression models for count data in r. *Journal of Statistical Software, Articles*, *27*(8), 1–25. https://doi.org/10.18637/jss.v027.i08

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A. & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. Springer, New York. https://doi.org/10.1007/978-0-387-87458-6

Appendix

A Additional figures

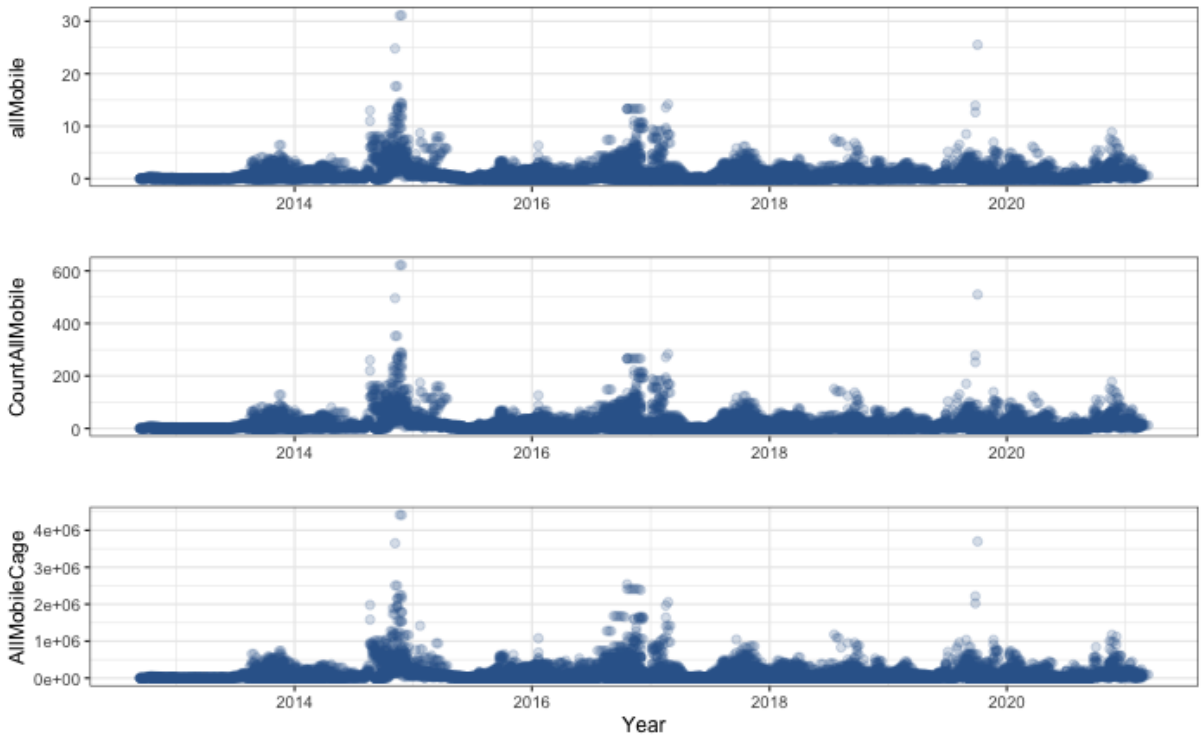

Figure A.1: Presentation of the response variables for all mobile lice. The reported mean of all mobile lice (top), the counted number of all mobile lice on twenty salmon (middle) and the estimated total number of all mobile lice (bottom) in each cage plotted against time.

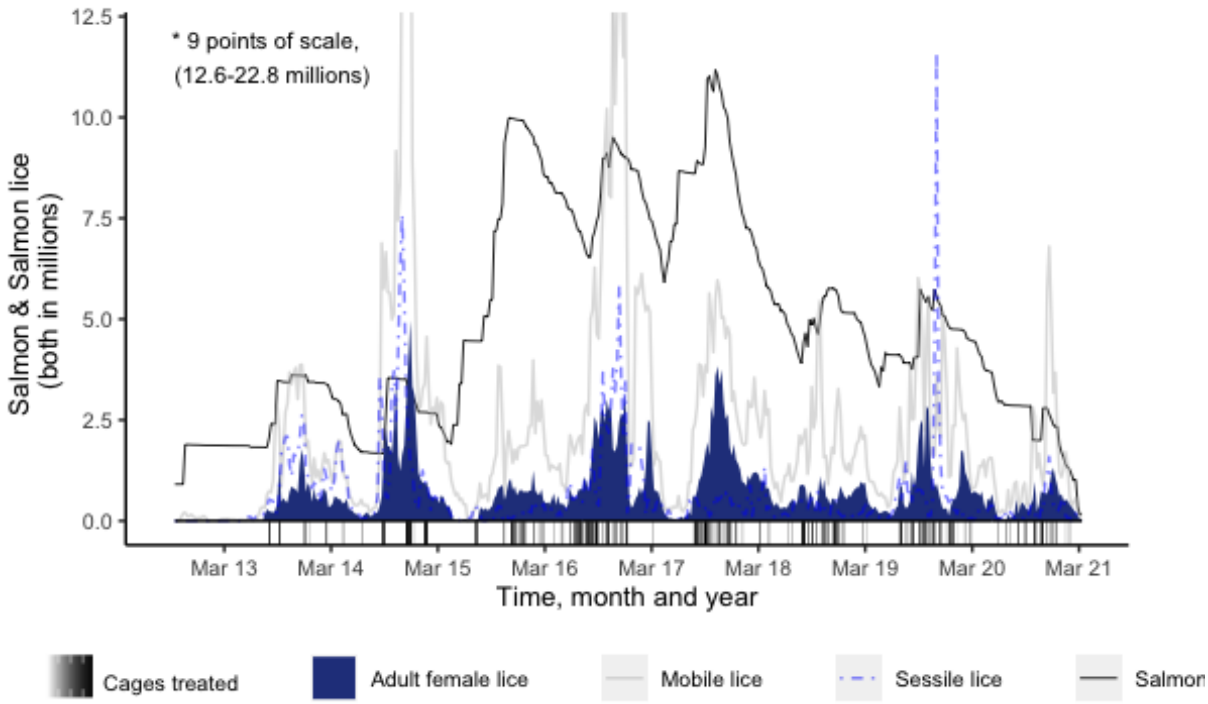

Figure A.2: Total number of salmon and salmon lice in the study area during the period 2012-2021. The total number is found by summarising over all cages in the study area. The number of salmon lice in each cage is estimated as the reported lice number for the cage multiplied with the number of salmon in the cage.

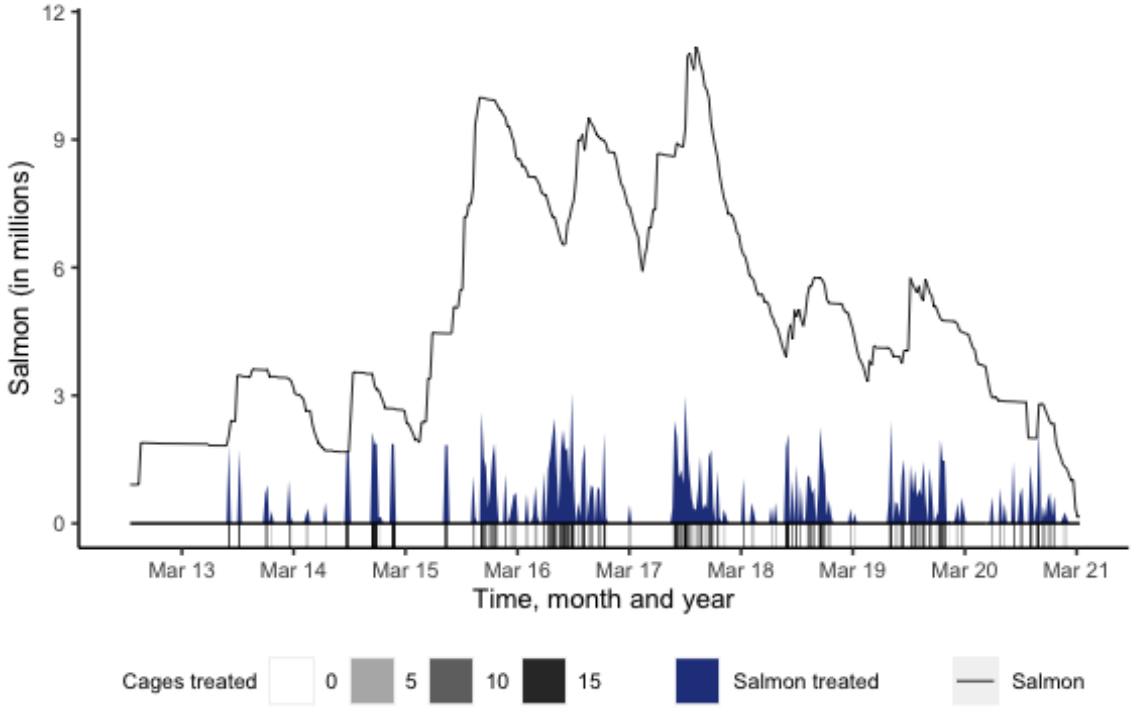

Figure A.3: Total number of salmon and salmon treated against salmon lice in the study area during the period 2012-2021. The total number is found by summarising over all cages in the study area.

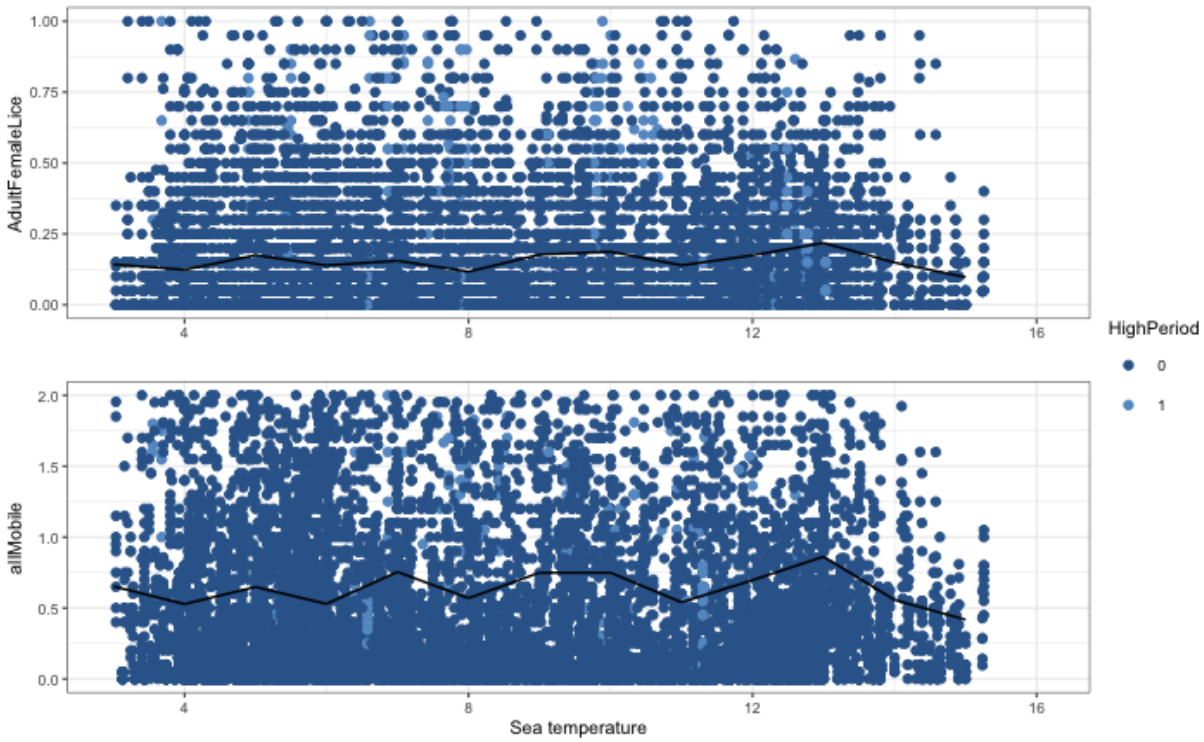

Figure A.4: The average of AdultFemaleLice and allMobile for every degree Celsius in the sea temperature plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.2).

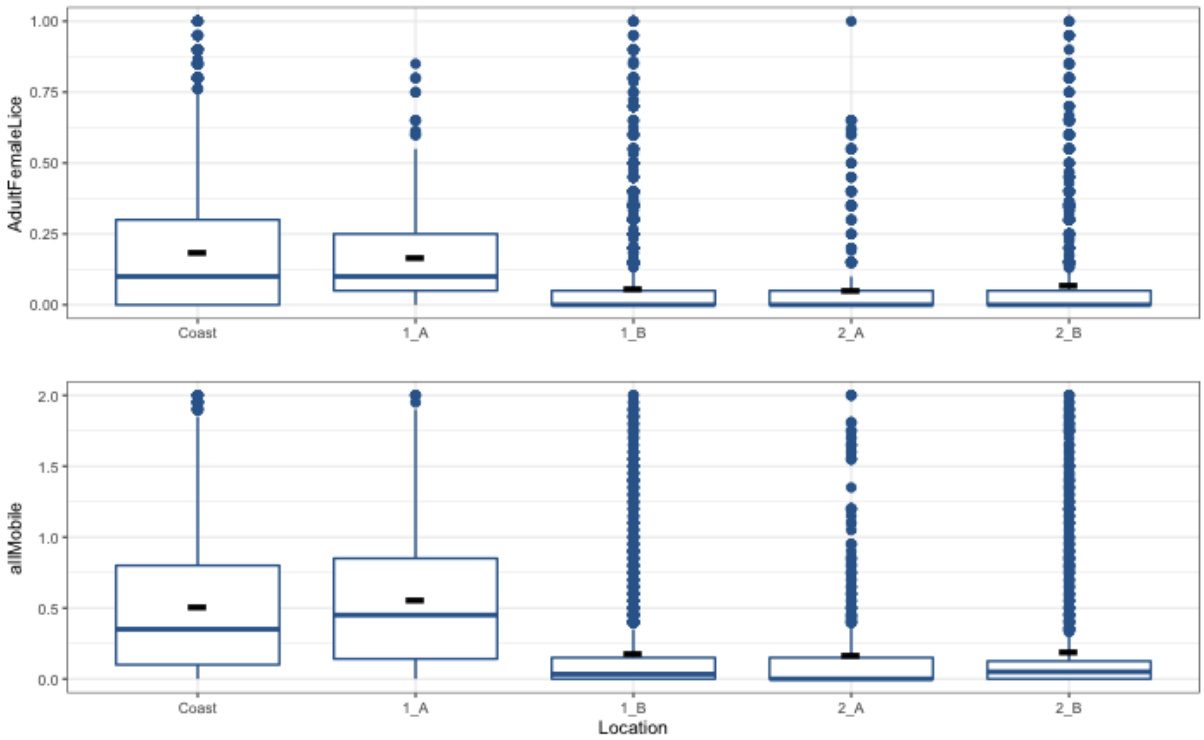

Figure A.5: The average of AdultFemaleLice and allMobile for the locations Coast, Fjord 1A, Fjord1B, Fjord2A and Fjord2B, plotted as a black bar. Box-plots of the reported lice numbers for adult female lice and all mobile lice, respectively, and the explanatory variable Location are also presented. Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.3).

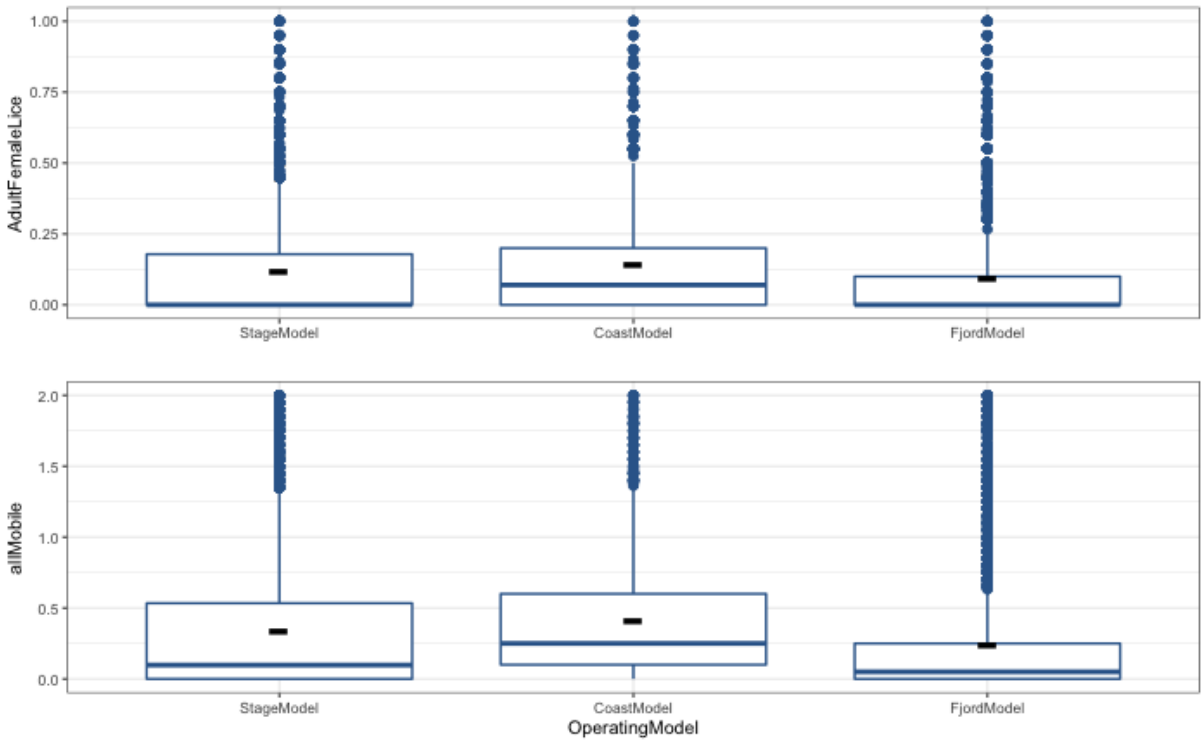

Figure A.6: The average of AdultFemaleLice and allMobile for the stage model, coast model and fjord model, plotted as a black bar. Box-plots of the reported lice numbers for adult female lice and all mobile lice, respectively, and the explanatory variable Operating-Model are also presented. Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.4).

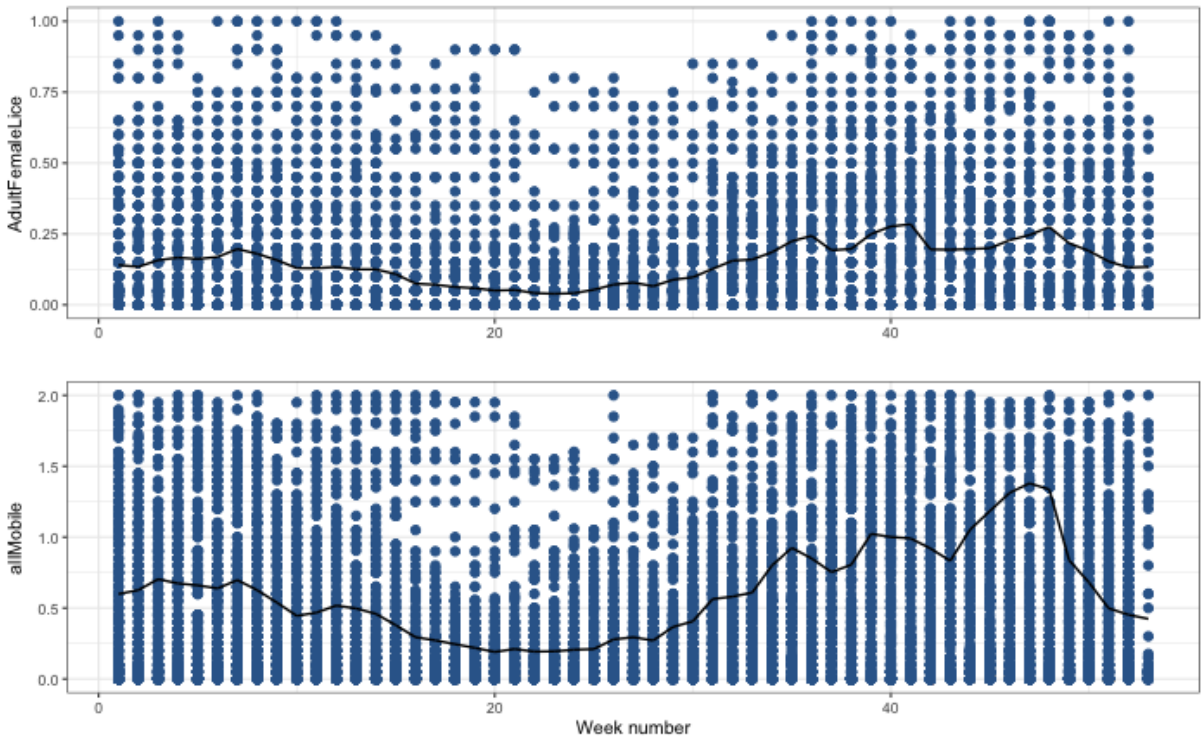

Figure A.7: The average of AdultFemaleLice and allMobile per week plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.5).

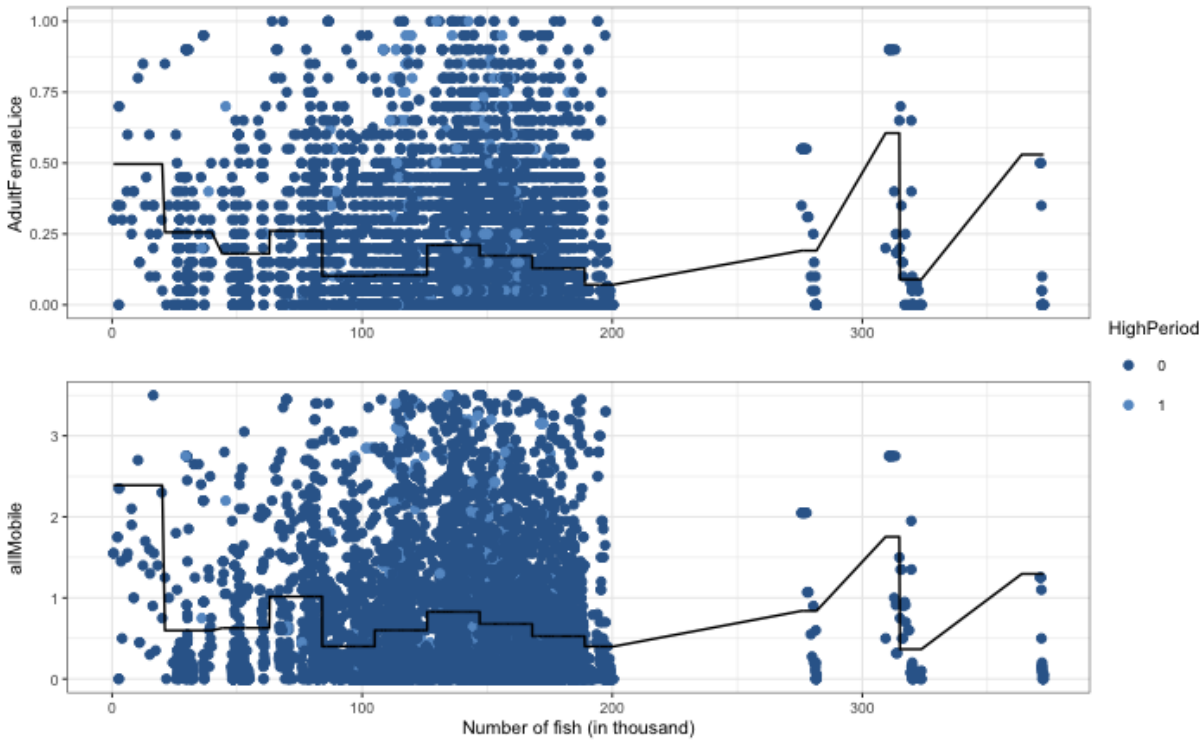

Figure A.8: The average of AdultFemaleLice and allMobile for every 21000 salmon in the cage plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.6).

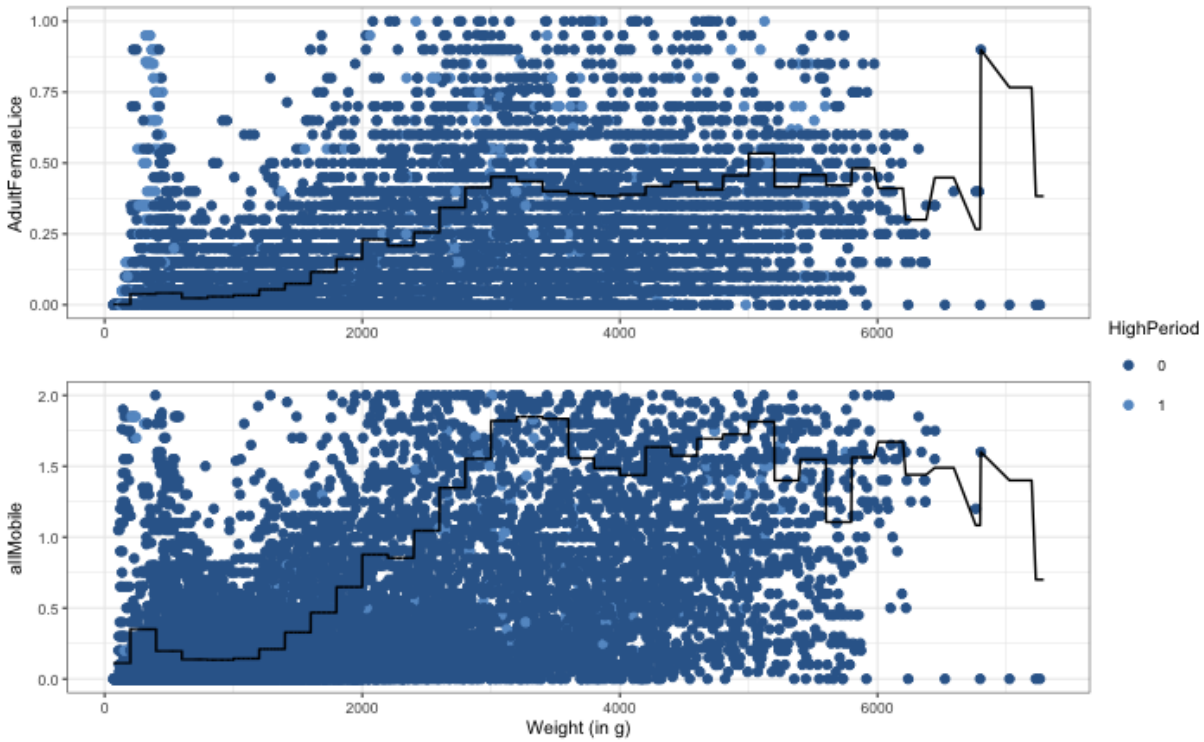

Figure A.9: The average of AdultFemaleLice and allMobile for every hundred gram plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.7).

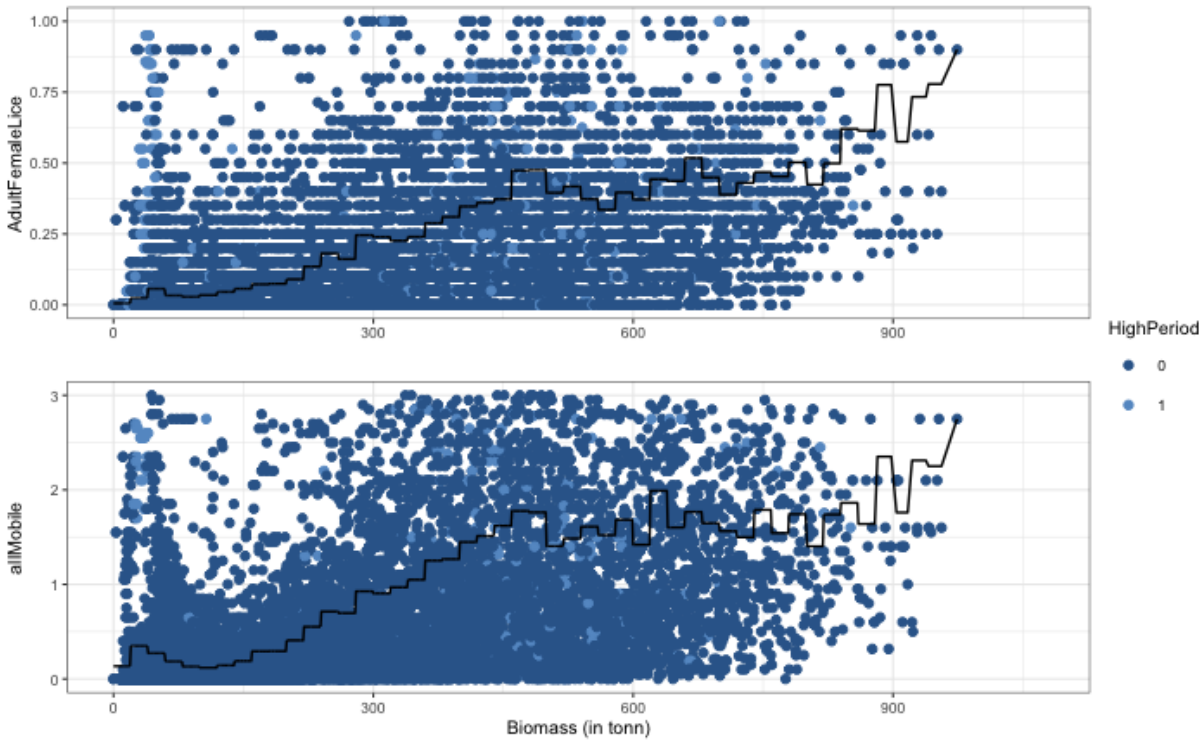

Figure A.10: The average of AdultFemaleLice and allMobile for every twenty metric tons plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.8).

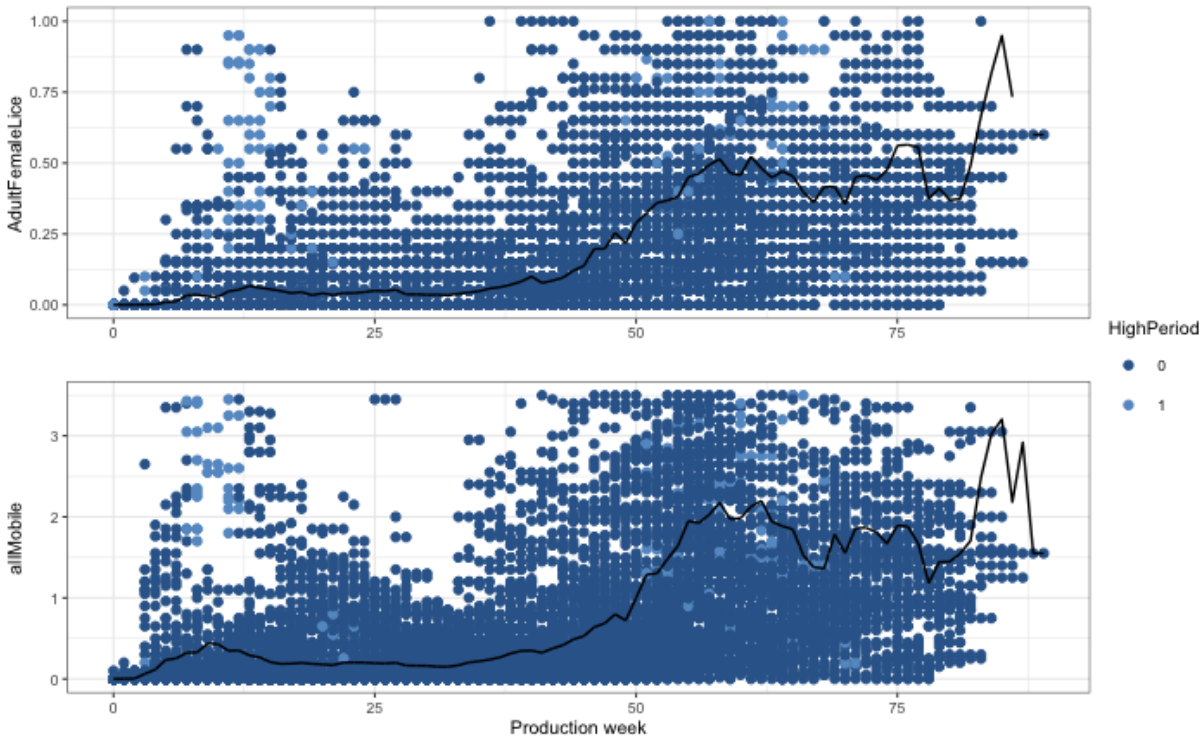

Figure A.11: The average of AdultFemaleLice and allMobile for each production week plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.9).

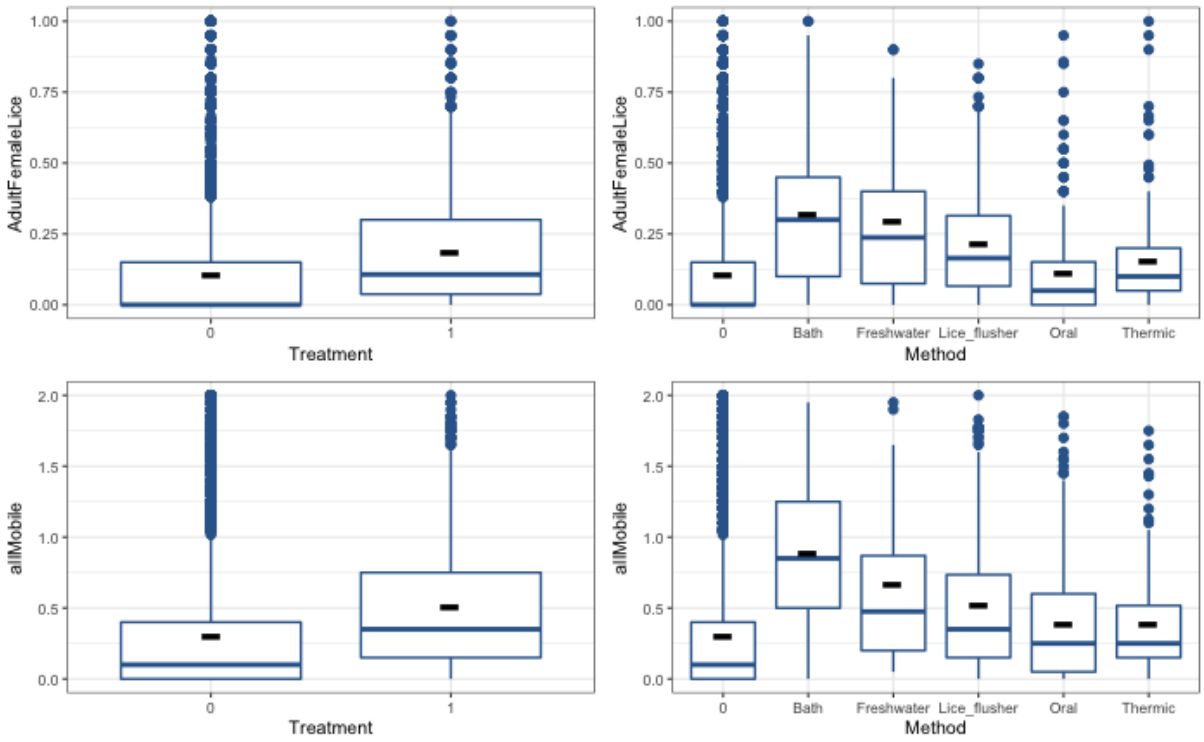

Figure A.12: The average of AdultFemaleLice and allMobile for no treatment and ongoing treatment, as well as the average for each of the different delousing methods, plotted as a black bar. Box-plots of the reported lice numbers for adult female lice and all mobile lice, respectively, for both Treatment and Method are also presented. Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.10).

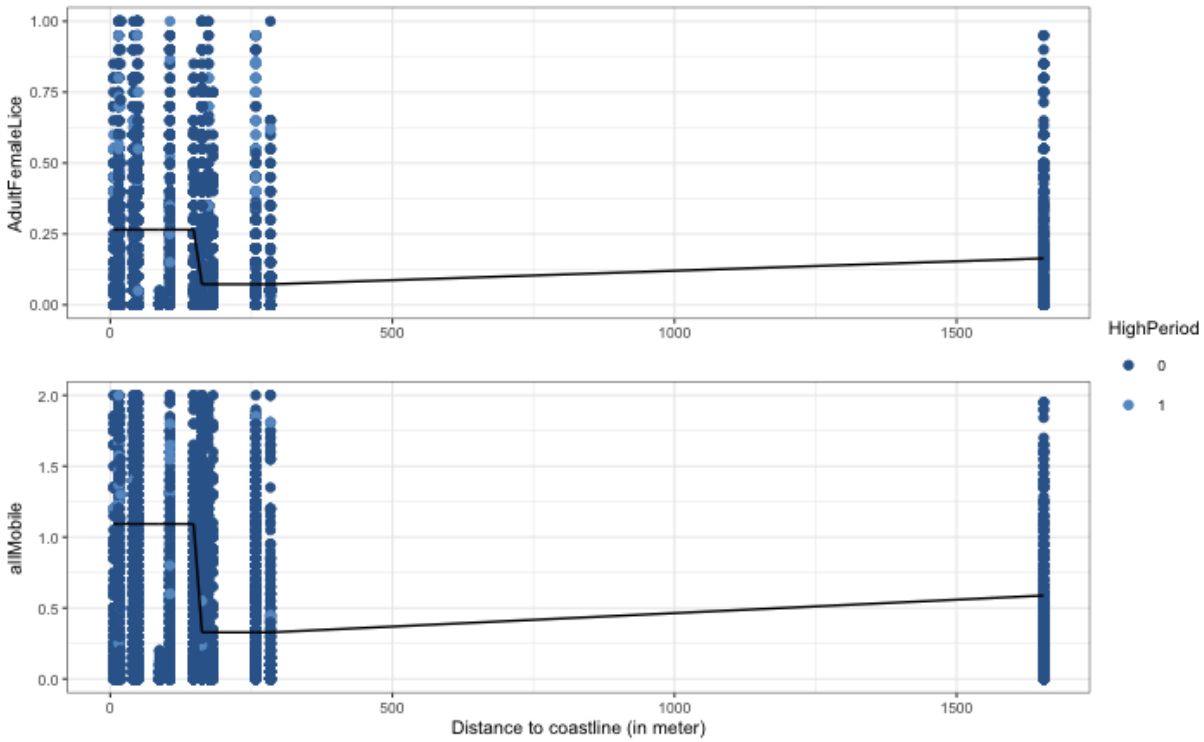

Figure A.13: The average of AdultFemaleLice and allMobile for every 150 meter from the coastline plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.11).

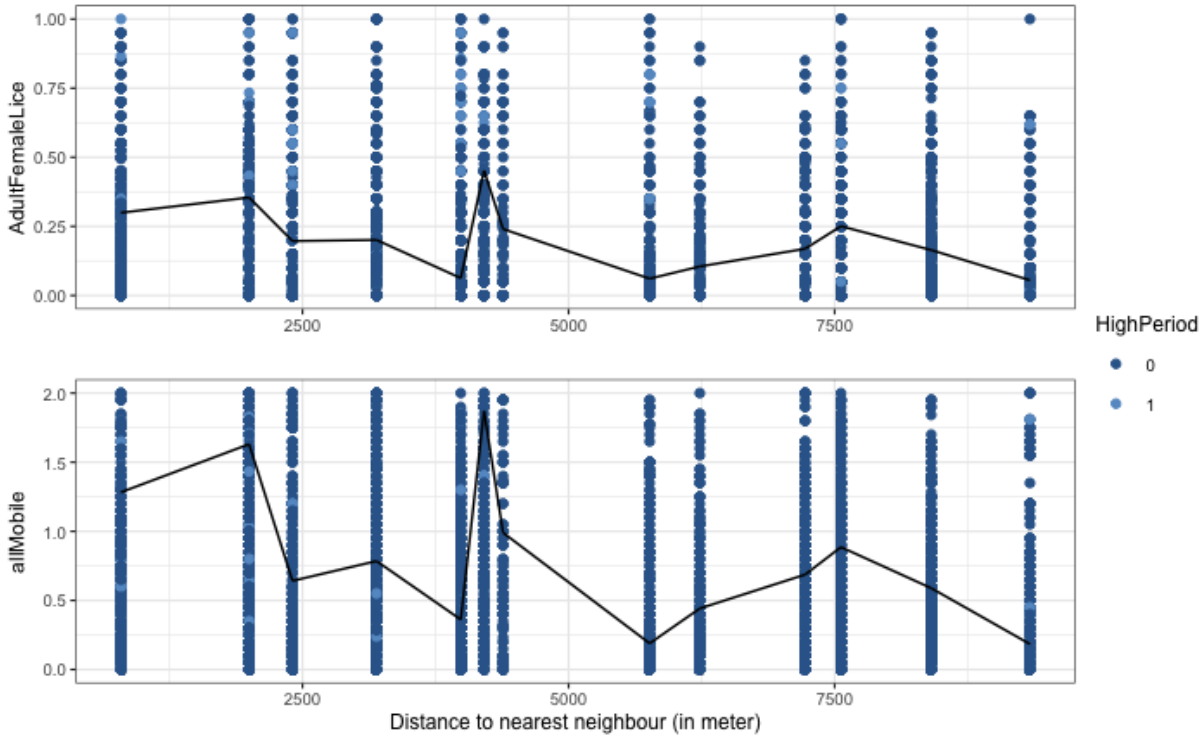

Figure A.14: The average of AdultFemaleLice and allMobile for every half kilometer from the nearest neighbour plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2 are plotted in this figure (all points are plotted in Figure 4.12).

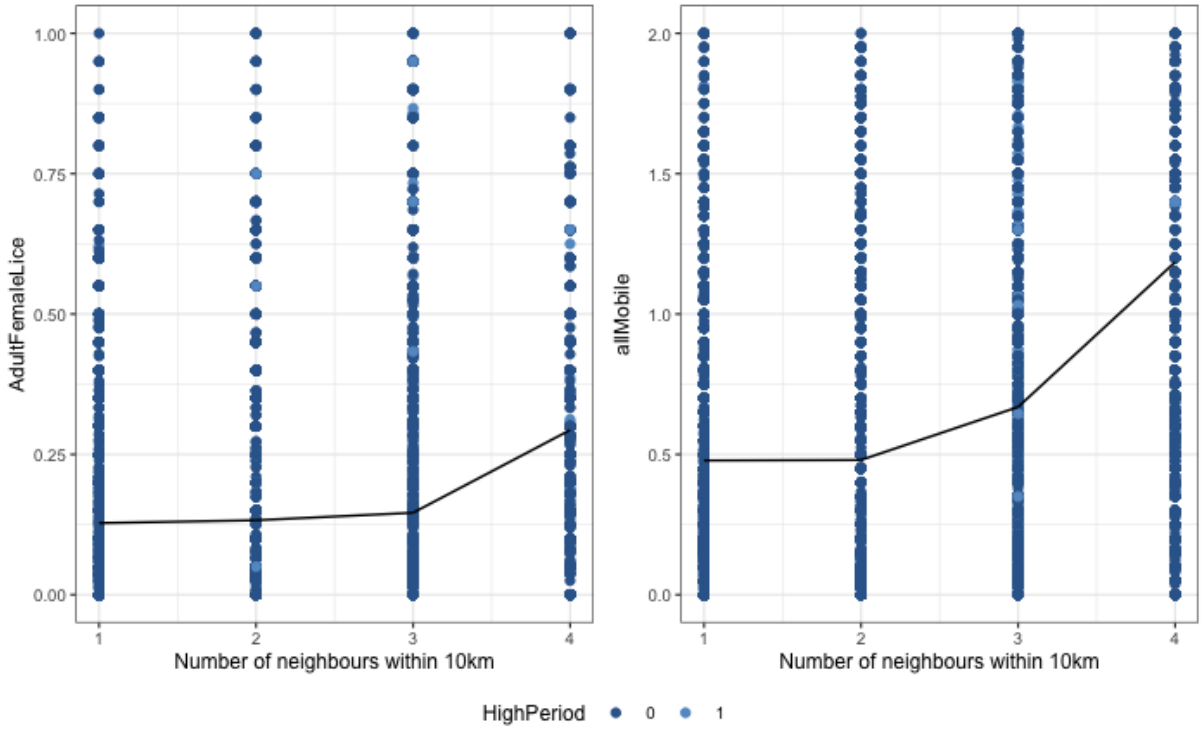

Figure A.15: The average of AdultFemaleLice and allMobile for each number of neighbours within 10km plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 2, respectively, are plotted in this figure (all points are plotted in Figure 4.13).

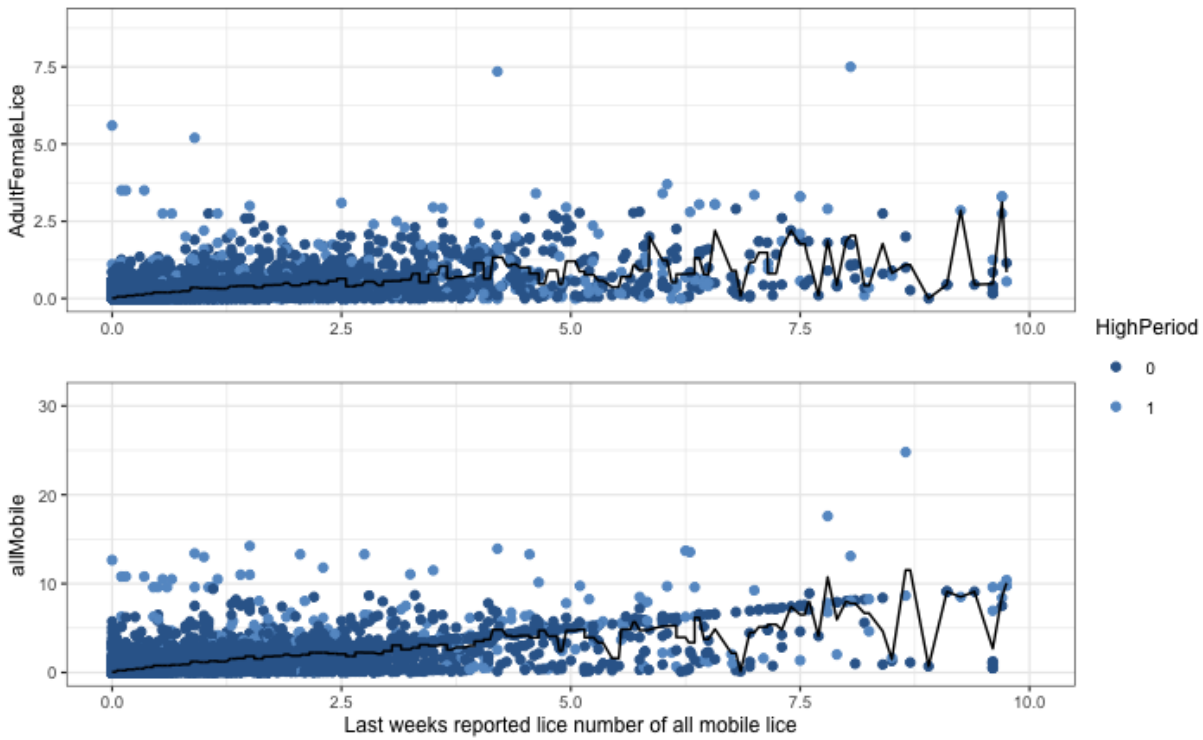

Figure A.16: The average of AdultFemaleLice and allMobile for the last weeks reported lice number with one significant digit are plotted as a black line. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers with last weeks reported lice number below 10 are plotted in this figure (all points are plotted in Figure 4.14).

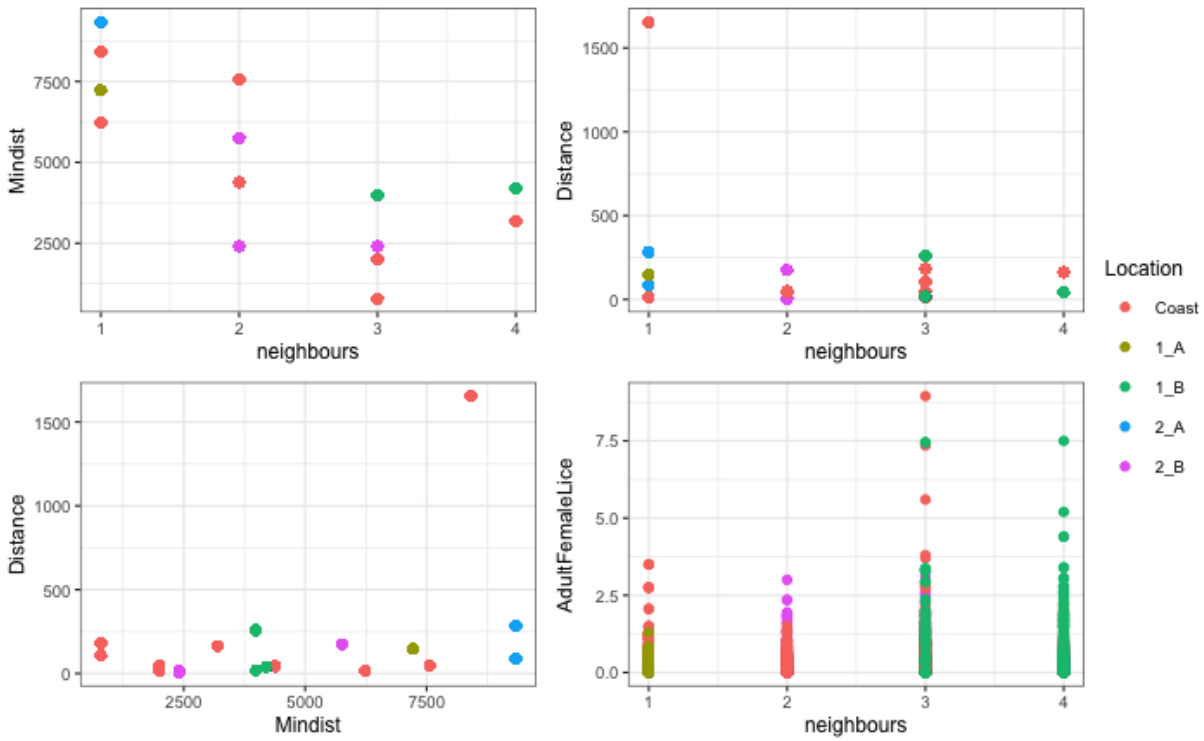

Figure A.17: Top left: The distance to the closest salmon farm in meter (Mindist) plotted against the number of neighbours within 10km (neighbours). Top right: The distance to the coastline in meter (Distance) plotted against the number of neighbours within 10km (neighbours). respectively. Bottom left: The distance to the coastline in meter (Distance) plotted against the distance to the closest salmon farm in meter (Mindist). Bottom right: The reported lice number of adult female lice plotted against the number of neighbours within 10km. The data points are grouped into the location of the salmon farm.

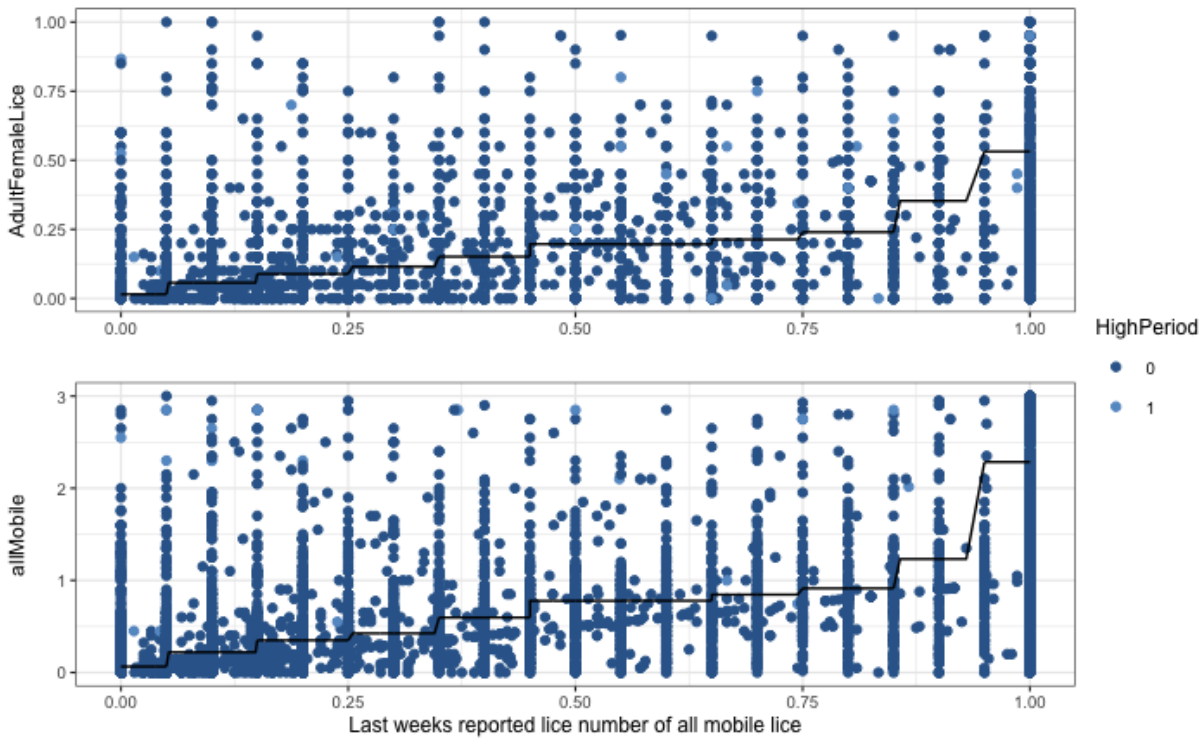

Figure A.18: The average of AdultFemaleLice and allMobile for the censored last weeks reported lice number with one significant digit are plotted as a black line. All the reported lice numbers from last weeks which are larger than 1, are registered as 1 in this figure. The blue points are the reported lice numbers for adult female lice and all mobile lice, respectively, and are grouped into the periods with high lice pressure (light blue) and normal lice pressure (dark blue). Only lice numbers below 1 and 3, respectively, are plotted in this figure (all points are plotted in Figure 4.14).

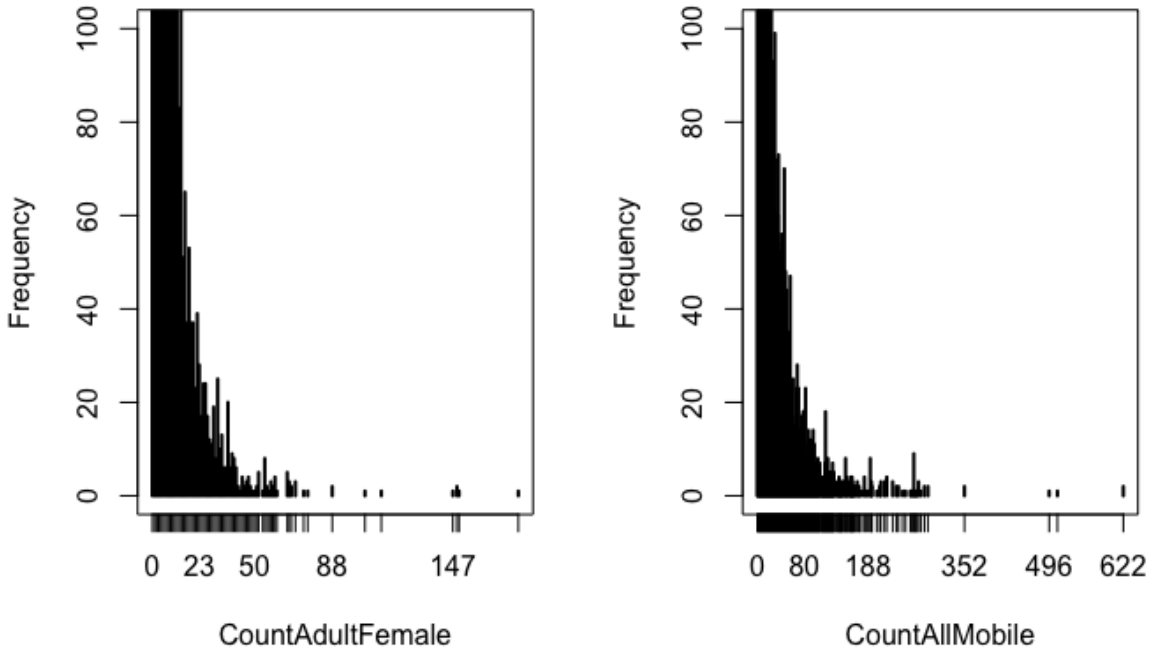

Figure A.19: The frequency of the CountAdultFemale and CountAllMobile with a censoring at frequency above 100.

B Additional results

Table B.1: Regression coefficients with associated estimate, standard error, z-value and p-value from the Poisson regression for the cage model of adult female lice.

Coefficients	Estimate	Std. Error	z value	p-value
Intercept	7.6732	0.0008	9339.33	$< 2 \cdot 10^{-16}$
SeaTemperature	0.0054	0.0002	34.46	$< 2 \cdot 10^{-16}$
SeaTemperature^2	0.0021	$8.1 \cdot 10^{-6}$	261.55	$< 2 \cdot 10^{-16}$
Location1_A	-0.7952	0.0006	-1407.63	$< 2 \cdot 10^{-16}$
Location1_B	-0.3735	0.0004	-1062.35	$< 2 \cdot 10^{-16}$
Location2_A	-0.5490	0.0006	-929.82	$< 2 \cdot 10^{-16}$
Location2_B	-0.5105	0.0004	-1275.69	$< 2 \cdot 10^{-16}$
OperatingModelCoast	0.1899	0.0002	1131.68	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.5173	0.0003	1504.54	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.4651	0.0002	-2611.70	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.2412	0.0002	-1189.81	$< 2 \cdot 10^{-16}$
Biomass	0.0027	$5.3 \cdot 10^{-7}$	5122.60	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.3406	0.0002	-1761.47	$< 2 \cdot 10^{-16}$
ProdWeek50	0.0199	0.0002	107.47	$< 2 \cdot 10^{-16}$
Treatment	-0.1577	0.0002	-897.24	$< 2 \cdot 10^{-16}$
DistToCoast1	-0.2517	0.0002	-1528.92	$< 2 \cdot 10^{-16}$
DistToCoast2	0.1592	0.0003	576.04	$< 2 \cdot 10^{-16}$
Neighbours	0.1417	0.0001	1449.25	$< 2 \cdot 10^{-16}$
LastWeek1	1.8225	0.0002	8468.28	$< 2 \cdot 10^{-16}$
HighPeriod	0.7441	0.0002	4724.04	$< 2 \cdot 10^{-16}$

AIC: 339990937, Null deviance: 903657462 on 15595 degrees of freedom
Residual deviance: 339899458 on 15576 degrees of freedom

Table B.2: Regression coefficients with associated estimate, standard error, z-value and p-value from the Poisson regression for the sample model of all mobile lice.

Coefficients	Estimate	Std. Error	z value	p-value
Intercept	1.1667	0.0335	34.84	$< 2 \cdot 10^{-16}$
SeaTemperature	-0.1777	0.0063	-28.21	$< 2 \cdot 10^{-16}$
SeaTemperature^2	0.0108	0.0003	32.24	$< 2 \cdot 10^{-16}$
Location1_A	0.2518	0.0196	12.81	$< 2 \cdot 10^{-16}$
Location1_B	0.4719	0.0114	41.37	$< 2 \cdot 10^{-16}$
Location2_A	0.0885	0.0236	3.75	0.0002
Location2_B	0.1103	0.0146	7.57	$3.69 \cdot 10^{-14}$
OperatingModelCoast	0.2768	0.0075	37.09	$< 2 \cdot 10^{-16}$
OperatingModelFjord	-0.1797	0.0112	-16.05	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.3701	0.0076	-48.80	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.3223	0.0085	-37.84	$< 2 \cdot 10^{-16}$
Biomass	0.0009	$2.29 \cdot 10^{-5}$	38.93	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.2266	0.0080	-28.19	$< 2 \cdot 10^{-16}$
ProdWeek50	0.4560	0.0083	54.69	$< 2 \cdot 10^{-16}$
Treatment	-0.2793	0.0072	-38.81	$< 2 \cdot 10^{-16}$
DistToCoast1	-0.1682	0.0071	-23.85	$< 2 \cdot 10^{-16}$
DistToCoast2	0.0606	0.0122	4.96	$6.89 \cdot 10^{-7}$
Neighbours	0.1084	0.0042	26.06	$< 2 \cdot 10^{-16}$
LastWeek1	2.3073	0.0086	267.27	$< 2 \cdot 10^{-16}$
HighPeriod	0.8592	0.0067	128.13	$< 2 \cdot 10^{-16}$

AIC: 197655, Null deviance: 477241 on 15595 degrees of freedom
Residual deviance: 156004 on 15576 degrees of freedom

Table B.3: Regression coefficients with associated estimate, standard error, z-value and p-value from the zero-inflated Poisson regression for the full sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z value	p-value
Intercept.	0.2540	0.0702	3.6164	0.0003
SeaTemperature	-0.0546	0.0132	-4.1387	$3.49 \cdot 10^{-5}$
SeaTemperature^2	0.0034	0.0007	4.9110	$9.06 \cdot 10^{-7}$
Location1_A	-0.5362	0.0491	-10.9152	$< 2 \cdot 10^{-16}$
Location1_B	-0.2234	0.0330	-6.7642	$1.34 \cdot 10^{-11}$
Location2_A	-0.4358	0.0629	-6.9272	$4.29 \cdot 10^{-12}$
Location2_B	-0.3523	0.0403	-8.7322	$< 2 \cdot 10^{-16}$
OperatingModelCoast	0.1719	0.0154	11.1715	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.4768	0.0344	13.8767	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.3310	0.0160	-20.7151	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.1793	0.0169	-10.5887	$< 2 \cdot 10^{-16}$
Biomass	0.0008	$4.42 \cdot 10^{-5}$	18.3968	$< 2 \cdot 10^{-16}$
BiomassIndicator1	-0.1598	0.0166	-9.6394	$< 2 \cdot 10^{-16}$
ProdWeek50	0.3122	0.0165	18.9312	$< 2 \cdot 10^{-16}$
Treatment	-0.1782	0.0153	-11.6076	$< 2 \cdot 10^{-16}$
DistToCoast1	-0.0452	0.0150	-3.0159	0.0026
DistToCoast2	0.1973	0.0248	7.9561	$1.78 \cdot 10^{-15}$
Neighbours	0.1213	0.0088	13.7118	$< 2 \cdot 10^{-16}$
LastWeek1	1.3833	0.0192	72.1942	$< 2 \cdot 10^{-16}$
HighPeriod	0.7141	0.0134	53.2057	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z value**	**p-value**
Intercept	1.2440	0.3498	3.5564	0.0004
SeaTemperature	0.3000	0.0667	4.4985	$6.84 \cdot 10^{-6}$
SeaTemperature^2	-0.0196	0.0036	-5.4691	$4.52 \cdot 10^{-8}$
Location1_A	-0.3350	0.2318	-1.4450	0.1485
Location1_B	-0.5282	0.1303	-4.0528	$5.06 \cdot 10^{-5}$
Location2_A	-0.1984	0.1801	-1.1013	0.2708
Location2_B	0.4592	0.1306	3.5146	0.0004
OperatingModelCoast	-1.1619	0.0986	-11.7791	$< 2 \cdot 10^{-16}$
OperatingModelFjord	-0.3760	0.0796	-4.7206	$2.35 \cdot 10^{-6}$
SeasonSpring	-0.3361	0.0731	-4.6001	$4.22 \cdot 10^{-6}$
SeasonWinter	-0.2930	0.0943	-3.1083	0.0019
Biomass	-0.0058	0.0003	-18.9606	$< 2 \cdot 10^{-16}$
BiomassIndicator	1.5964	0.1528	10.4498	$< 2 \cdot 10^{-16}$
ProdWeek50	-1.3151	0.1104	-11.9158	$< 2 \cdot 10^{-16}$
Treatment	0.1514	0.1116	1.3559	0.1751
DistToCoast1	0.1757	0.0743	2.3636	0.0181
DistToCoast2	0.3589	0.1291	2.7806	0.0054
Neighbours	0.1270	0.0471	2.6985	0.0070
LastWeek1	-3.0410	0.1045	-29.0914	$< 2 \cdot 10^{-16}$
HighPeriod	0.2604	0.1133	2.2986	0.0215

AIC: 64848, 47, Loglik: -32384.23, Expected zeroes: 6580

Table B.4: Regression coefficients with associated estimate, standard error, z-value and p-value from the zero-inflated negative binomial regression for the full sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z value	p-value
Intercept	-0.4338	0.1545	-2.8084	0.0050
SeaTemperature	-0.0436	0.0297	-1.4656	0.1428
SeaTemperature^2	0.0041	0.0016	2.5999	0.0093
Location1_A	-0.5380	0.0923	-5.8286	$5.59 \cdot 10^{-9}$
Location1_B	-0.3090	0.0632	-4.8896	$1.01 \cdot 10^{-6}$
Location2_A	-0.2966	0.1206	-2.4585	0.0140
Location2_B	-0.3628	0.0781	-4.6434	$3.43 \cdot 10^{-6}$
OperatingModelCoast	0.2801	0.0327	8.5560	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.4920	0.0634	7.7573	$8.68 \cdot 10^{-15}$
SeasonSpring	-0.3522	0.0316	-11.1482	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.0607	0.0386	-1.5729	0.1158
Biomass	0.0012	0.0001	11.3619	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.2658	0.0405	-6.5571	$5.49 \cdot 10^{-11}$
ProdWeek50	0.5040	0.0346	14.5814	$< 2 \cdot 10^{-16}$
Treatment	-0.1503	0.0348	-4.3144	$1.60 \cdot 10^{-5}$
DistToCoast1	-0.0685	0.0330	-2.0774	0.0378
DistToCoast2	0.2464	0.0475	5.1817	$2.20 \cdot 10^{-7}$
Neighbours	0.1313	0.0182	7.2041	$5.84 \cdot 10^{-13}$
LastWeek1	1.5229	0.0368	41.3490	$< 2 \cdot 10^{-16}$
HighPeriod	0.7253	0.0427	16.9855	$< 2 \cdot 10^{-16}$
Log(r)	0.3625	0.0225	16.1445	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z value**	**p-value**
Intercept	0.4472	0.7831	0.5711	0.5679
SeaTemperature	0.5257	0.1379	3.8127	0.0001
SeaTemperature^2	-0.0229	0.0076	-3.0330	0.0024
Location1_A	-0.5299	0.5262	-1.0071	0.3139
Location1_B	-0.8434	0.3222	-2.6176	0.0089
Location2_A	-0.1878	0.3855	-0.4873	0.6260
Location2_B	1.0521	0.3145	3.3450	0.0008
OperatingModelCoast	-1.4959	0.2663	-5.6177	$1.94 \cdot 10^{-8}$
OperatingModelFjord	-0.2560	0.1382	-1.8522	0.0640
SeasonSpring	-0.1438	0.1593	-0.9024	0.3669
SeasonWinter	0.2502	0.1933	1.2939	0.1957
Biomass	-0.0126	0.0010	-12.7064	$< 2 \cdot 10^{-16}$
BiomassIndicator	1.7293	0.9901	1.7465	0.0807
ProdWeek50	-16.4463	566.9700	-0.0290	0.9769
Treatment	-1.0825	0.5302	-2.0416	0.0412
DistToCoast1	0.1697	0.1629	1.0416	0.2976
DistToCoast2	0.9646	0.2975	3.2422	0.0012
Neighbours	0.1256	0.1049	1.1970	0.2313
LastWeek1	-18.6639	1.3311	-14.0214	$< 2 \cdot 10^{-16}$
HighPeriod	0.3403	0.2388	1.4253	0.1541

AIC: 47973, 94, Loglik: -23927.97, Expected zeroes: 7788

Table B.5: Regression coefficients with associated estimate, standard error, z-value and p-value from the zero-altered Poisson regression for the full sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z value	p-value
Intercept	0.2567	0.0701	3.6645	0.0002
SeaTemperature	-0.0546	0.0132	-4.1382	$3.50 \cdot 10^{-5}$
SeaTemperature^2	0.0034	0.0007	4.9226	$8.54 \cdot 10^{-7}$
Location1_A	-0.5236	0.0486	-10.7630	$< 2 \cdot 10^{-16}$
Location1_B	-0.2056	0.0327	-6.2885	$3.21 \cdot 10^{-10}$
Location2_A	-0.4088	0.0615	-6.6438	$3.06 \cdot 10^{-11}$
Location2_B	-0.3288	0.0398	-8.2563	$< 2 \cdot 10^{-16}$
OperatingModelCoast	0.1751	0.0154	11.3795	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.4567	0.0339	13.4685	$< 2 \cdot 10^{-16}$
SeasonSpring	-0.3304	0.0159	-20.7285	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.1807	0.0170	-10.6462	$< 2 \cdot 10^{-16}$
Biomass	0.0008	$4.43 \cdot 10^{-5}$	18.2293	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.1602	0.0166	-9.6610	$< 2 \cdot 10^{-16}$
ProdWeek50	0.3161	0.0164	19.2206	$< 2 \cdot 10^{-16}$
Treatment	-0.1811	0.0153	-11.8036	$< 2 \cdot 10^{-16}$
DistToCoast1	-0.0529	0.0149	-3.5446	0.0004
DistToCoast2	0.1962	0.0248	7.9250	$2.28 \cdot 10^{-15}$
Neighbours	0.1211	0.0088	13.8271	$< 2 \cdot 10^{-16}$
LastWeek1	1.3806	0.0191	72.1755	$< 2 \cdot 10^{-16}$
HighPeriod	0.7132	0.0134	53.1529	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z value**	**p-value**
Intercept	-1.6625	0.3060	-5.4326	$5.55 \cdot 10^{-8}$
SeaTemperature	-0.3050	0.0589	-5.1772	$2.25 \cdot 10^{-7}$
SeaTemperature^2	0.0198	0.0032	6.2750	$3.50 \cdot 10^{-10}$
Location1_A	0.0216	0.1844	0.1169	0.9069
Location1_B	0.2814	0.1141	2.4666	0.0136
Location2_A	-0.0946	0.1453	-0.6510	0.5150
Location2_B	-0.6774	0.1142	-5.9320	$2.99 \cdot 10^{-9}$
OperatingModelCoast	1.0948	0.0850	12.8857	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.6378	0.0702	9.0899	$< 2 \cdot 10^{-16}$
SeasonSpring	0.1393	0.0645	2.1604	0.0307
SeasonWinter	0.2079	0.0857	2.4257	0.0153
Biomass	0.0055	0.0003	21.1693	$< 2 \cdot 10^{-16}$
BiomassIndicator	-1.4562	0.1326	-10.9858	$< 2 \cdot 10^{-16}$
ProdWeek50	1.2837	0.0969	13.2416	$< 2 \cdot 10^{-16}$
Treatment	-0.2058	0.0992	-2.0742	0.0381
DistToCoast1	-0.1542	0.0669	-2.3058	0.0211
DistToCoast2	-0.1608	0.1063	-1.5128	0.1303
Neighbours	-0.0268	0.0393	-0.6804	0.4963
LastWeek1	3.3828	0.0954	35.4698	$< 2 \cdot 10^{-16}$
HighPeriod	-0.1281	0.1133	-1.1315	0.2579

AIC: 64881.76, Loglik: -32400.88, Expected zeroes: 8125

Table B.6: Regression coefficients with associated estimate, standard error, z-value and p-value from the zero-altered negative binomial regression for the full sample model of adult female lice.

Count-model coefficients	Estimate	Std. Error	z value	p-value
Intercept	-0.3524	0.1744	-2.0210	0.0433
SeaTemperature	-0.0546	0.0330	-1.6548	0.0980
SeaTemperature^2	0.0041	0.0017	2.3291	0.0199
Location1_A	-0.5534	0.1029	-5.3751	$7.66 \cdot 10^{-8}$
Location1_B	-0.2973	0.0713	-4.1683	$3.07 \cdot 10^{-5}$
Location2_A	-0.3419	0.1302	-2.6247	0.0087
Location2_B	-0.3350	0.0868	-3.8604	0.0001
OperatingModelCoast	0.2284	0.0361	6.3256	$2.52 \cdot 10^{-10}$
OperatingModelFjord	0.4227	0.0708	5.9692	$2.38 \cdot 10^{-9}$
SeasonSpring	-0.3778	0.0355	-10.6330	$< 2 \cdot 10^{-16}$
SeasonWinter	-0.1290	0.0425	-3.0310	0.0024
Biomass	0.0012	0.0001	10.9251	$< 2 \cdot 10^{-16}$
BiomassIndicator	-0.2934	0.0438	-6.6931	$2.18 \cdot 10^{-11}$
ProdWeek50	0.4412	0.0382	11.5612	$< 2 \cdot 10^{-16}$
Treatment	-0.0986	0.0390	-2.5297	0.0114
DistToCoast1	-0.0535	0.0355	-1.5092	0.1312
DistToCoast2	0.3227	0.0540	5.9718	$2.35 \cdot 10^{-9}$
Neighbours	0.1483	0.0205	7.2206	$5.18 \cdot 10^{-13}$
LastWeek1	1.5610	0.0404	38.6266	$< 2 \cdot 10^{-16}$
HighPeriod	0.7896	0.0459	17.2063	$< 2 \cdot 10^{-16}$
Log(r)	0.3325	0.0333	9.9967	$< 2 \cdot 10^{-16}$
Zero-inflated coefficients	**Estimate**	**Std. Error**	**z value**	**p-value**
Intercept	-1.6625	0.3060	-5.4326	$5.55 \cdot 10^{-8}$
SeaTemperature	-0.3050	0.0589	-5.1772	$2.25 \cdot 10^{-7}$
SeaTemperature^2	0.0198	0.0032	6.2750	$3.50 \cdot 10^{-10}$
Location1_A	0.0216	0.1844	0.1169	0.9069
Location1_B	0.2814	0.1141	2.4666	0.0136
Location2_A	-0.0946	0.1453	-0.6510	0.5150
Location2_B	-0.6774	0.1142	-5.9320	$2.99 \cdot 10^{-9}$
OperatingModelCoast	1.0948	0.0850	12.8857	$< 2 \cdot 10^{-16}$
OperatingModelFjord	0.6378	0.0702	9.0899	$< 2 \cdot 10^{-16}$
SeasonSpring	0.1393	0.0645	2.1604	0.0307
SeasonWinter	0.2079	0.0857	2.4257	0.0153
Biomass	0.0055	0.0003	21.1693	$< 2 \cdot 10^{-16}$
BiomassIndicator	-1.4562	0.1326	-10.9858	$< 2 \cdot 10^{-16}$
ProdWeek50	1.2837	0.0969	13.2416	$< 2 \cdot 10^{-16}$
Treatment	-0.2058	0.0992	-2.0742	0.0381
DistToCoast1	-0.1542	0.0669	-2.3058	0.0211
DistToCoast2	-0.1608	0.1063	-1.5128	0.1303
Neighbours	-0.0268	0.0393	-0.6804	0.4963
LastWeek1	3.3828	0.0954	35.4698	$< 2 \cdot 10^{-16}$
HighPeriod	-0.1281	0.1133	-1.1315	0.2579

AIC: 48604.38, Loglik: -24261.19, Expected zeroes: 8125

C R-code examples

Packages and Functions

```
\usemintedstyle{tango}
#Install packages
library(easypackages)

#install.packages("countreg", repos="http://R-Forge.R-project.org")
#library(devtools)
#install_github("vqv/ggbiplot")

packages("tidyverse", "readxl", "conflicted", "gt", "gridExtra", "pscl",
    "performance", "plotly", "ISOweek", "GGally", "AER", "MASS",
    "wesanderson", "GGally", "ggpubr", "osmdata", "geosphere",
    "sf", "reshape2", "ggcorrplot", "countreg", "ggbiplot", "xtable")

conflict_prefer("filter", "dplyr")
conflict_prefer("lag", "dplyr")
conflict_prefer("select", "dplyr")
conflict_prefer("mutate", "dplyr")
conflict_prefer("arrange", "dplyr")
conflict_prefer("first", "dplyr")
conflict_prefer("summarise", "dplyr")
conflict_prefer("summarize", "dplyr")
conflict_prefer("group_by", "dplyr")
conflict_prefer("startsWith", "gdata")
conflict_prefer("zeroinfl", "pscl")
conflict_prefer("zeroinfl.control", "pscl")
conflict_prefer("hurdle", "pscl")

#Functions

###Preparation of the dataset - growth sites fjord1 ###
fjord_1 <- function(path){
  biomasse_opprinnelig <- read_excel(path, sheet = "Biomasse", skip = 6)
  temperatur_opprinnelig <- read_excel(path, sheet = "Temperatur ", skip = 6)
  lusetall_opprinnelig <- read_excel(path, sheet = "Lusetall", skip = 8)
  avlusing_opprinnelig <- read_excel(path, sheet = "Avlusninger")
  biomasse <- biomasse_opprinnelig %>%
    mutate(YearWeek = ifelse(startsWith(`Row Labels`, "20"), `Row Labels`, NA))
  while(anyNA(biomasse$YearWeek)==TRUE){
    biomasse <- biomasse %>%
      mutate(YearWeek = ifelse(is.na(YearWeek), lag(YearWeek), YearWeek))
  }
  biomasse <- biomasse %>%
    filter(startsWith(`Row Labels`,"20")==FALSE & `Row Labels`!="Grand Total")%>%
    separate("YearWeek", c("Year", "Week"), sep = "w", remove=TRUE) %>%
    mutate(Cage=`Row Labels`, Year=as.numeric(Year), Week = as.numeric(Week))%>%
    select(-c(`Row Labels`))%>%
    filter(`IB Biomasse`!=0)

  temperatur <- temperatur_opprinnelig %>%
    mutate(YearWeek = ifelse(startsWith(`Row Labels`, "20"), `Row Labels`, NA))
  while(anyNA(temperatur$YearWeek)==TRUE){
```

```
  temperatur <- temperatur %>%
    mutate(YearWeek = ifelse(is.na(YearWeek), lag(YearWeek), YearWeek)) }
temperatur <- temperatur %>%
  filter(startsWith(`Row Labels`,"20")==FALSE & `Row Labels`!="Grand Total")%>%
  separate("YearWeek", c("Year", "Week"), sep = "w", remove=TRUE) %>%
  mutate(Cage=`Row Labels`, Year=as.numeric(Year),
      Week = as.numeric(Week))%>%
  select(-c(`Row Labels`))

lusetall <- lusetall_opprinnelig %>%
  mutate(YearWeek = ifelse(startsWith(`Row Labels`, "20"),
      `Row Labels`, NA))
while(anyNA(lusetall$YearWeek)==TRUE){
  lusetall <- lusetall %>%
    mutate(YearWeek = ifelse(is.na(YearWeek), lag(YearWeek), YearWeek))}
if(path=="data_R.xlsx"){
  lusetall <- lusetall %>%
      filter(startsWith(`Row Labels`,"20")==FALSE &
          `Row Labels`!="Grand Total")%>%
      separate("YearWeek", c("Year", "Week"), sep = "w", remove=TRUE) %>%
      mutate(Mobile = Mobile (large) + Mobile (small)) %>%
      mutate(Cage=`Row Labels`, Year=as.numeric(Year),
          Week = as.numeric(Week))%>%
      select(-c(`Row Labels`))}
else{
  lusetall <- lusetall %>%
    filter(startsWith(`Row Labels`,"20")==FALSE & `Row Labels`!="Grand Total")%>%
    separate("YearWeek", c("Year", "Week"), sep = "w", remove=TRUE) %>%
    mutate(Cage = `Row Labels`, Year=as.numeric(Year),
      Week = as.numeric(Week))%>%
    mutate(Mobile = ifelse(is.na(Mobile),
      Mobile (large) + Mobile (small),Mobile)) %>%
    select(-c(`Row Labels`))}

avlusing <- avlusing_opprinnelig %>%
  mutate(Date = as.Date(TreatmentDate, "%d.%m.%Y"))%>%
  mutate(Method = ifelse(Method=="Mekanisk",
      str_replace_all(`Active Substance`, c("Hot water" = "Thermic",
         "Lice flusher" = "Lice_flusher",
          "Seawater" = "Lice_flusher",
          "fresh water" = "Freshwater")), Method)) %>%
  mutate(Method = str_replace_all(Method, c("Feed" = "Oral"))) %>%
  mutate(Cage = as.factor(UnitName), AC = `Active Substance`,
         Method = as.factor(Method)) %>%
  group_by(Cage, Method) %>%
  arrange(Date) %>%
  mutate(date.diff = c(1,diff(Date)))%>%
  mutate(period_treatment = cumsum(date.diff>3))%>%
  ungroup()%>%
  group_by(Cage, Method, period_treatment)%>%
  summarize(start_treatment = min(Date),
            end_treatment = max(Date))
treatment <- avlusing %>%
  mutate(duration_treatment=end_treatment-start_treatment,
         Week = as.numeric(strftime(start_treatment, format = "%V")),
         Year = as.numeric(strftime(start_treatment, format = "%Y")))
treatments = treatment
```

```
  if (any(as.numeric(strftime(treatment$end_treatment[],
    format = "%V"))-treatment$Week[]==1)) {
    week2 <- treatment %>%
      filter(as.numeric(strftime(end_treatment, format = "%V"))-Week==1)%>%
      mutate(Week = as.numeric(strftime(end_treatment, format = "%V")))
    treatments <- rbind(treatments, week2)}
  if (any(as.numeric(strftime(treatment$end_treatment[],
    format = "%V"))-treatment$Week[]==2)) {
    week3 <- treatment %>%
      filter(as.numeric(strftime(end_treatment, format = "%V"))-Week==2) %>%
      mutate(Week = as.numeric(strftime(start_treatment, format = "%V"))+1)
    treatments <- rbind(treatments, week3)}
  total <- biomasse %>%
    left_join(temperatur, by=c("Year", "Week", "Cage"))
  total <- total %>%
    left_join(lusetall, by=c("Year", "Week", "Cage"))
  total <- total %>%
    left_join(treatments, by=c("Year", "Week", "Cage")) %>%
    mutate(Locality=substr(Cage, 1, 1))%>%
    mutate(SeaTemperature = `Temperatur gj.snitt.`,
           Biomass = `IB Biomasse`,
           NumberOfFish = `IB Antall`, Weight = `IB Vekt`,
           AdultFemaleLice = `Adult female ovig.`,
           MobileLice = `Mobile`,
           SessileLice = `Chalimus`,
           ScottishLice = `Scottish lice`) %>%
    select(Locality, Cage, Year, Week, SeaTemperature, AdultFemaleLice,
        MobileLice, SessileLice, ScottishLice, NumberOfFish, Weight,
        Biomass, start_treatment, end_treatment, Method, period_treatment,
        duration_treatment) %>%
    filter(Year>2011)
  return(total)
}

###Sort in production cycles  ###
Production_cycle <- function(dataset){
  dataset %>%
    group_by(base) %>%
    arrange(base, date) %>%
    mutate(date.diff = c(1,diff(date))) %>%
    mutate(production_cycle = cumsum(abs(date.diff) > 7))  %>%
    ungroup() %>%
    group_by(base, production_cycle) %>%
    mutate(start = date[1], end = date[length(date)]) %>%
    mutate(duration_production_cycle = as.numeric((end-start)/7)) %>%
    mutate(week_in_production_cycle = as.numeric((date-start)/7))
}

###Temperature for each week###
temperatur <- function(dataset){
  notemp <- subset(dataset, is.na(dataset$SeaTemperature))
  for (i in 1:nrow(notemp)){
    index <- which(dataset$Locality==notemp$Locality[i] &
                     dataset$Week==notemp$Week[i] &
                     dataset$Year==notemp$Year[i] &
                     dataset$Cage==notemp$Cage[i])
```

```
    if((notemp$Week[i]==1) & (notemp$Year[i]==2016)){ #Check last year (2015 has 53
    ↪  weeks)
      last_number <- which(dataset$Locality==notemp$Locality[i] &
                           dataset$Week==53 &
                           dataset$Year==(notemp$Year[i]-1) &
                           dataset$Cage==notemp$Cage[i])}
    else if ((notemp$Week[i]==1) & (notemp$Year[i]!=2016)){
      last_number <- which(dataset$Locality==notemp$Locality[i] &
                           dataset$Week==52 &
                           dataset$Year==(notemp$Year[i]-1) &
                           dataset$Cage==notemp$Cage[i]) }
    else{ #forrige uke
      last_number <- which(dataset$Locality==notemp$Locality[i] &
                           dataset$Week==(notemp$Week[i]-1) &
                           dataset$Year==notemp$Year[i] &
                           dataset$Cage==notemp$Cage[i])}
    dataset$SeaTemperature[index] <- dataset$SeaTemperature[last_number]}
  return(dataset)
}

###Merge start and growth site after base1 (start site nr 1) ###
Prod1 <- function(data){
  data <- data %>%
    group_by(base)%>%
    arrange(date) %>%
    mutate(date.diff = c(1,diff(date))) %>%
    mutate(cycle = cumsum(abs(date.diff) > 7))  %>%
    ungroup() %>%
    group_by(base, cycle) %>%
    mutate(start = date[1], end = date[length(date)]) %>%
    mutate(duration_production_cycle = as.numeric((end-start)/7)) %>%
    mutate(week_in_production_cycle = as.numeric((date-start)/7)) %>%
    mutate(start = as.Date(start), end=as.Date(end)) %>%
    ungroup()%>%
    select(-cycle)
  return(data)
}

###Group into production cycles ###
Prod11 <- function(data){
  data <- data %>%
    group_by(base)%>%
    arrange(date) %>%
    mutate(date.diff = c(1,diff(date))) %>%
    mutate(cycle = cumsum(abs(date.diff) > 7))  %>%
    ungroup() %>%
    group_by(base) %>%
    mutate(start = date[1], end = date[length(date)]) %>%
    mutate(duration_production_cycle = as.numeric((end-start)/7)) %>%
    mutate(week_in_production_cycle = as.numeric((date-start)/7)) %>%
    mutate(start = as.Date(start), end=as.Date(end)) %>%
    ungroup()%>%
    select(-cycle)
  return(data)
}

###Merge start and growth site after base2 (start site nr 2) ###
```

```
Prod2 <- function(data){
  b <- data%>%
    filter(!is.na(base2))
  cage <- as.character(b$base2[!duplicated(b$base2)])
  b2 <- as.character(b$base[!duplicated(b$base2)])
  for (i in 1:length(cage)){
    data <- data %>%
      mutate(end = as.Date(ifelse(Cage==cage[i],
        data$end[which(data$base==b2[i] &
        data$duration_production_cycle==data$week_in_production_cycle)],
        end), format= "%Y-%m-%d"))}
  data <- data %>%
    mutate(duration_production_cycle = as.numeric(end-start)/7)
  return(data)
}

###Group into production cycles if two start sites ###
Prod22 <- function(data){
  b <- data%>%
    filter(!is.na(base2))
  cage <- as.character(b$base2[!duplicated(b$base2)])
  b2 <- as.character(b$base[!duplicated(b$base2)])
  for (i in 1:length(cage)){
    data <- data %>%
      mutate(end = as.Date(ifelse(base==cage[i],
                  data$end[which(data$base==b2[i] &
                  data$duration_production_cycle==data$week_in_production_cycle)],
                  end), format= "%Y-%m-%d"))}
  data <- data %>%
    mutate(duration_production_cycle = as.numeric(end-start)/7)
  return(data)
}

###Load salinity data - preparation ###
salinity <- function(path){
  miljo_opprinnelig <- read_excel(path, sheet = "Miljødata", skip = 7, guess_max =
  ↪  5000)
  miljo <- miljo_opprinnelig %>%
    select(-c(1,2, 4, 5, 6, 8, 9, 10))%>%
    subset(!is.na(Date)) %>%
    mutate(YearWeek = strftime(Date, format = "%Y-W%V")) %>%
    separate(YearWeek, c("Year", "Week"), sep = "-W", remove=TRUE) %>%
    mutate(Locality = path) %>%
    subset(!is.na(`Salinity(5m) (‰)`)) %>%
    mutate(Salinity = as.numeric(sub(",", ".", `Salinity(5m) (‰)`, fixed = TRUE)))%>%
    filter(Salinity<50) %>%
    group_by(Locality, Year, Week) %>%
    summarize(Salinity = mean(Salinity, na.rm=T))%>%
    subset(!is.na(Salinity))
  return(miljo)
}
```

Preparation of the Dataset: Growth Sites - Fjord 1

```
M <- fjord_1("data_M.xlsx")
N <- fjord_1("data_N.xlsx")
O <- fjord_1("data_O.xlsx")
P <- fjord_1("data_P.xlsx")
Q <- fjord_1("data_Q.xlsx")
R <- fjord_1("data_R.xlsx")

mowing <- read_xlsx("Flytting_A.xlsx")

#Rename the cages after relocations
##Location M is used as an example

M_oppd <- M %>%
  filter(Cage!="M01" | (Year!=2014 | Week<15))%>%
  filter(Cage!="M02" | (Year!=2014 | Week<15))%>%
  filter(Cage!="M03" | (Year!=2014 | Week<15)) %>%
  mutate(Cage =
    ifelse((Cage=="M02" & Year==2013 & Week>45 & Week<49), "M01",
    ifelse((Cage=="M05" & Year==2013 & Week>45 & Week<49), "M02",
    ifelse((Cage=="M07" & Year==2013 & Week>45 & Week<49), "M03",
    ifelse((Cage=="M03" & Year==2013 & Week>45 & Week<49), "M04",
    ifelse((Cage=="M04" & Year==2013 & Week>45 & Week<49), "M05",
    ifelse((Cage=="M01" & Year==2013 & Week>45 & Week<49), "M06",
    ifelse((Cage=="M06" & Year==2013 & Week>48 & Week<52), "M02",
    ifelse((Cage=="M02" & Year==2013 & Week>48 & Week<52), "M04",
    ifelse((Cage=="M04" & Year==2013 & Week>48 & Week<52), "M06",
    ifelse((Cage=="M05" & Year==2013 & Week>51 & Week<53), "M02",
    ifelse((Cage=="M02" & Year==2013 & Week>51 & Week<53), "M04",
    ifelse((Cage=="M04" & Year==2013 & Week>51 & Week<53), "M05",
    ifelse((Cage=="M07" & Year==2013 & Week>51 & Week<53), "M06",
    ifelse((Cage=="M05" & Year==2014 & Week<9), "M02",
    ifelse((Cage=="M02" & Year==2014 & Week<9), "M04",
    ifelse((Cage=="M04" & Year==2014 & Week<9), "M05",
    ifelse((Cage=="M07" & Year==2014 & Week<9), "M06",
    ifelse((Cage=="M05" & Year==2014 & Week>8 & Week<17), "M02",
    ifelse((Cage=="M06" & Year==2014 & Week>8 & Week<30), "M03",
    ifelse((Cage=="M02" & Year==2014 & Week>8 & Week<17), "M04",
    ifelse((Cage=="M04" & Year==2014 & Week>8 & Week<17), "M05",
    ifelse((Cage=="M07" & Year==2014 & Week>8 & Week<17), "M06",
    ifelse((Cage=="M04" & Year==2014 & Week>16 & Week<30), "M06",
       Cage))))))))))))))))))))))))))) %>%
  mutate(Cage =
    ifelse((Cage=="M02" & Year==2016 & Week>39 & Week<44), "M01",
    ifelse((Cage=="M03" & Year==2016 & Week>39 & Week<44), "M02",
    ifelse((Cage=="M04" & Year==2016 & Week>39 & Week<44), "M03",
    ifelse((Cage=="M01" & Year==2016 & Week>39 & Week<44), "M05",
    ifelse((Cage=="M05" & Year==2016 & Week>39 & Week<43), "M06",
    ifelse((Cage=="M06" & Year==2016 & Week>39 & Week<43), "M07",
    ifelse((Cage=="M02" & Year==2016 & Week>47), "M01",
    ifelse((Cage=="M03" & Year==2016 & Week>47), "M02",
    ifelse((Cage=="M04" & Year==2016 & Week>47), "M03",
    ifelse((Cage=="M01" & Year==2016 & Week>47), "M05",
    ifelse((Cage=="M05" & Year==2016 & Week>47), "M06",
    ifelse((Cage=="M06" & Year==2016 & Week>47), "M07",
```

```
    ifelse((Cage=="M02" & Year==2017 & Week<10), "M01",
    ifelse((Cage=="M03" & Year==2017 & Week<10), "M02",
    ifelse((Cage=="M04" & Year==2017 & Week<10), "M03",
    ifelse((Cage=="M01" & Year==2017 & Week<10), "M05",
    ifelse((Cage=="M05" & Year==2017 & Week<10), "M06",
    ifelse((Cage=="M06" & Year==2017 & Week<10), "M07",
    ifelse((Cage=="M07" & Year==2017 & Week>9 & Week<43), "M01",
    ifelse((Cage=="M05" & Year==2017 & Week>9 & Week<43), "M06",
           Cage))))))))))))))))))))))


#Plotting the biomass to control the changes

M_oppd%>%
  filter(Year>2015)%>%
  group_by(Cage) %>%
  ggplot(aes(x=Week, y = `Biomass`))+
  geom_point(aes(color=Cage))+
  #geom_line(aes(color=Cage))+
  labs(x="Year")+
  theme_bw()


#Lice number for each week

liceM <- lice(M_oppd)
liceN <- lice(N_oppd)
liceQ <- lice(Q_oppd)
growth <- rbind(liceM,liceN,liceQ) %>%
  mutate(OperatingModel = "StageModel", Stage="growth", Fjord="Coast")
liceO <- lice(O_oppd)
liceP <- lice(P_oppd)
liceR <- lice(R_oppd)
normal <- rbind(liceO, liceP, liceR)%>%
  mutate(OperatingModel = "NormalModel", Stage="normal",
         Fjord = ifelse(Locality=="R","1_A", "1_B"))

total <- rbind(growth, normal) %>%
  mutate(Start_Locality = "A",
         base = substr(Cage, 1,3),
         base2 = NA)


#Sea temperature for each week

total[which(is.na(total$SeaTemperature)),]
total <- temperatur(total)

#Dato-column

data <- total %>%
  mutate(Year = as.character(Year), Week = as.character(Week)) %>%
  group_by(Year, Week, Cage) %>%
  mutate(Week = ifelse(nchar(Week)==1, paste("0",Week, sep=""), Week)) %>%
  mutate(Week = paste("W",Week, sep="")) %>%
  mutate(date = ISOweek2date(paste(Year, Week,4, sep = "-" ))) %>%
```

```
  mutate(Year = as.numeric(Year),
         Week = as.numeric(str_replace_all(Week, "W", "")))


#Which cage did the salmon originate from

mow <- mowing %>%
  select(-c(`Til lok`, `Fra lok`))

data2 <- Production_cycle(data)

data <- left_join(data2, mow,
    by=c("Year"="År", "Week"="Uke", "base"="Cage", "Stage")) %>%
  group_by(base, production_cycle) %>%
  arrange(base, date) %>%
  mutate(Start_cage = Start_cage[1],
         A_production_cycle = A_production_cycle[1])


#Save dataframe

saveRDS(data, file="A_growth")
```

Merging Start and Growth Sites - Fjord 1

```
#Merge together data from start and growth

start <- readRDS("FjordA") %>%
  mutate(A_production_cycle = production_cycle) %>%
  mutate(Fjord = "1_B")

growth <- readRDS("A_growth")%>%
  select(Locality, Cage, Year, Week, OperatingModel, Fjord, Stage,
    Start_cage, Start_Locality, SeaTemperature, AdultFemaleLice,
    MobileLice, SessileLice, ScottishLice, NumberOfFish, everything())

data <- rbind(start, growth) %>%
  mutate(Locality = as.factor(Locality), Cage=as.factor(Cage),
         OperatingModel = as.factor(OperatingModel),
         Fjord=as.factor(Fjord), Stage=as.factor(Stage)) %>%
  mutate(Method = as.factor(ifelse(is.na(Method), 0 ,Method)))


summary(data)


#Egen datokolonne

data <- data %>%
  mutate(Year = as.character(Year), Week = as.character(Week)) %>%
  group_by(Year, Week, Cage) %>%
  mutate(Week = ifelse(nchar(Week)==1, paste("0",Week, sep=""), Week)) %>%
  mutate(Week = paste("W",Week, sep="")) %>%
  mutate(date = ISOweek2date(paste(Year, Week,4, sep = "-" ))) %>%
  mutate(Year = as.numeric(Year),
```

```
        Week = as.numeric(str_replace_all(Week, "W", "")))


#Produksjonssyklus

#If the growth stage is the sum of two start cages
#Functions defined in the top of this section
staged <- data %>%
  mutate(base = ifelse(nchar(Start_cage)==3, Start_cage,
    substr(Start_cage, 1,3))) %>%
  mutate(base2 = ifelse(nchar(Start_cage)==3, NA,
    substr(Start_cage, 5,7))) %>%
  ungroup()

prod0 <- staged %>%
  filter(A_production_cycle==0)
prod1 <- staged %>%
  filter(A_production_cycle==1)
prod2 <- staged %>%
  filter(A_production_cycle==2)
prod3 <- staged %>%
  filter(A_production_cycle==3)
prod4 <- staged %>%
  filter(A_production_cycle==4)
prod5 <- staged %>%
  filter(A_production_cycle==5)
prod6 <- staged %>%
  filter(A_production_cycle==6)

P0 <- Prod1(prod0)
P0 <- Prod2(P0)

P1 <- Prod1(prod1)
P1 <- Prod2(P1)

P2 <- Prod1(prod2)
P2 <- Prod2(P2)

P3 <- Prod1(prod3)
P3 <- Prod2(P3)

P4 <- Prod1(prod4)
P4 <- Prod2(P4)

P5 <- Prod1(prod5)
P5 <- Prod2(P5)

P6 <- Prod11(prod6)
P6 <- Prod22(P6)

P <- rbind(P0,P1,P2,P3, P4, P5, P6)

summary(P)

#Plot to control that everything is okay
P6%>%
```

```
  group_by(base) %>%
  ggplot(aes(x=date, y = `Biomass`))+
  geom_point(aes(color=Cage))+
  #geom_line(aes(color=Cage))+
  labs(x="Year")+
  theme_bw()

#Save dataframe

saveRDS(P, file="x")
```

Merging to a Common Dataset and Making Several Variables

```
#Loading data
y <- readRDS("y") %>%
  select(-LiceCount_date) %>%
  select(Locality, Cage, Year, Week, date, everything())%>%
  mutate(Stage=as.factor(Stage),
    OperatingModel = as.character(OperatingModel))%>%
  mutate(OperatingModel = ifelse(OperatingModel != "NormalModel",
    OperatingModel, ifelse(Fjord=="Coast", "CoastModel","FjordModel")),
         A_production_cycle = production_cycle)

x <- readRDS("x")%>%
  select(Locality, Cage, Year, Week, date,everything()) %>%
  mutate(start_treatment = strftime(start_treatment, format = "%Y/%V"),
         end_treatment = strftime(end_treatment, format = "%Y/%V"),
         duration_treatment = as.numeric(duration_treatment),
         OperatingModel = as.character(OperatingModel)) %>%
  mutate(OperatingModel = ifelse(OperatingModel != "NormalModel",
        OperatingModel, ifelse(Fjord=="Coast", "CoastModel", "FjordModel")),
         A_production_cycle = A_production_cycle)

Coast <- readRDS("CoastData") %>%
  mutate(A_production_cycle = production_cycle)

Latlon <- read_xlsx("barentswatch.xlsx", sheet="LatLon")

sites <- read_xlsx("Akvakulturregisteret.xlsx", sheet="Prod78")

salinity <- readRDS("Salinity")

start_data <- bind_rows(x, y, Coast)

summary(start_data)


##Make different variables to the visualization and the analysis

data <- start_data%>%
  mutate(NumberOfFish1000 = NumberOfFish/1000,
         Biomass = Biomass/1000) %>%
  mutate(censoredAdult = ifelse(AdultFemaleLice>3, 3, AdultFemaleLice)) %>%
  mutate(allMobile = AdultFemaleLice+MobileLice) %>%
  mutate(censored = ifelse(allMobile>10, 10, allMobile)) %>%
```

```
  ungroup()%>%
  mutate(Treatment = as.factor(ifelse(Method!=0, 1, 0)),
         Stage = factor(Stage, levels=c("normal", "start", "growth")),
         Fjord = factor(Fjord,
            levels=c("Coast", "1_A", "1_B", "2_A", "2_B")),
         Locality = factor(Locality,
            levels=c("A", "B", "C", "D", "E", "F",
                     "G", "H", "I", "J", "K", "L",
                     "M", "N", "O", "P", "Q", "R")),
         Location=as.factor(Fjord),
         OperatingModel = as.factor(OperatingModel)) %>%
  mutate(OperatingModel = factor(OperatingModel,
            levels=c("StageModel", "CoastModel", "FjordModel"))) %>%
  group_by(Cage, production_cycle) %>%
  arrange(Cage, production_cycle, week_in_production_cycle) %>%
  mutate(LastWeek = dplyr::lag(allMobile, default=0)) %>%
  ungroup() %>%
  mutate(AllMobileCage = as.integer(allMobile*NumberOfFish),
         AdultFemaleCage= as.integer(AdultFemaleLice*NumberOfFish),
         CountAllMobile = as.integer(allMobile*20),
         CountAdultFemale = as.integer(AdultFemaleLice*20),
         A_production_cycle = factor(A_production_cycle, ordered = TRUE))

summary(data)

#Distance from coastline
{r echo=True, include=FALSE}
osm_box <- getbb(place_name = "Trøndelag")%>%
  opq() %>%
  add_osm_feature("natural","coastline") %>%
  osmdata_sf()

osm_box_nordland <- getbb(place_name = "Nordland")%>%
  opq() %>%
  add_osm_feature("natural","coastline") %>%
  osmdata_sf()

saveRDS(osm_box, file="osm_box")
saveRDS(osm_box_nordland, file="osm_box_nordland")

{r echo=True, include=FALSE}
osm_box <- readRDS("osm_box")
osm_box_nordland <- readRDS("osm_box_nordland")

d1_sf = Latlon %>%
  st_as_sf(coords = c("Lon", "Lat")) %>%
  st_set_crs(4326) %>%
  select(Locality, geometry)

coastline <- ggplot() +
  geom_sf(data=osm_box$osm_lines) +
  geom_sf(data=osm_box_nordland$osm_lines)
  geom_sf(data=d1_sf)

dist <- dist2Line(p=st_coordinates(d1_sf),
              line=st_coordinates(osm_box$osm_lines)[,1:2])
dist2 <- dist2Line(p=st_coordinates(d1_sf),
```

```
              line=st_coordinates(osm_box_nordland$osm_lines)[,1:2])

distt = as.data.frame(dist)
colnames(distt)[1]<-"distanceT"
distn = as.data.frame(dist2)
colnames(distn)[1]<-"distanceN"
distancetot = cbind(distt,distn)%>%
  select(!starts_with("L"))

distancetot$Distance=apply(distancetot,1,FUN=min)

saveRDS(distancetot, file="distancetot")

distancetot <- readRDS("distancetot")
locdist <- Latlon %>%
  select(Locality,Lat,Lon,ProduksjonsområdeId)%>%
  mutate(Distance = distancetot$Distance)

#Distance from other sites
site_overwiev <- sites %>%
  select(LOK_NAVN,N_GEOWGS84,Ø_GEOWGS84,PROD_OMR)
site_overwiev <- site_overwiev[which(!duplicated(site_overwiev)),] %>%
  mutate(Lat = as.numeric(N_GEOWGS84),
    Lon = as.numeric(Ø_GEOWGS84),.keep="unused")

distance=data.frame()
for (i in 1:nrow(site_overwiev)){
  for (j in 1:nrow(locdist)){
    distance=append(distance,distm(c(locdist$Lon[j],locdist$Lat[j]),
        c(site_overwiev$Lon[i],site_overwiev$Lat[i]),
        fun=distVincentyEllipsoid))}
}

output=matrix(unlist(distance),ncol=nrow(locdist),byrow=T,
              dimnames=list(site_overwiev$LOK_NAVN,locdist$Locality))

outputframe=as.data.frame(output)

#Remove sites that not has been in operation the last 10 years
neighbours <- outputframe

for (i in 1:ncol(outputframe)){
  for (j in 1:nrow(outputframe)){
    if(outputframe[j,i]>10000){
      neighbours[j,i]=NA}
  }
}

remove = c("xxx", "xxx", "xxx")
outputframe <- outputframe[!rownames(outputframe) %in% remove, ]

#Mindist og neighbours
locdist$Mindist=NA
locdist$neighbours=NA
for (i in 1:nrow(locdist)){
  locdist$Mindist[i] = min(outputframe[,i][which(outputframe[,i]>1)])
  locdist$neighbours[i] = sum(outputframe[,i]<10000 & outputframe[,i]>1)
```

```
}

#Merge with the rest
total_data <- left_join(data, locdist)
total_data2 <- left_join(total_data, salinity) %>%
  mutate(Locality = as.factor(Locality),
         Fjord = as.factor(Fjord))

#High period
total_data2%>%
  ggplot(aes(x=date, y = allMobile))+
  geom_point(aes(color=Locality))+
  labs(x="Year")+
  theme_bw()

total_data2%>%
  filter(allMobile>=10) %>%
  arrange(date)

total_data2%>%
  ggplot(aes(x=date, y = AdultFemaleLice))+
  geom_point(aes(color=Locality))+
  labs(x="Year")+
  theme_bw()

total_data2%>%
  filter(AdultFemaleLice>=3) %>%
  arrange(date)

total <- total_data2%>%
  mutate(HighPeriod = as.factor(ifelse((Year==2014 & Week==34) |
         (Year==2014 & Week>43 & Week<49) |
         (Year==2016 & Week>41 & Week<50) |
         (Year==2017 & Week>6 & Week<9) |
         (Year==2019 & Week>38 & Week<41),1,0))) %>%
  mutate(HighPeriod_adult = as.factor(ifelse(((Year==2014 & Week>39 & Week<43) |
         (Year==2014 & Week>45 & Week<52) |
         (Year==2016 & Week>46 & Week<50) |
         (Year==2017 & Week>6 & Week<9) |
         (Year==2017 & Week>37 & Week<41) |
         (Year==2019 & Week>37 & Week<41)),1,0)))

#Plot lice numbers vs date - grouped into highperiod
g1 <- total%>%
  ggplot(aes(x=date, y = AdultFemaleLice))+
  geom_point(aes(color=HighPeriod_adult))+
  geom_line(y=3)+
  labs(x=NULL, color="HighPeriod")+
  scale_color_manual(values=c("#336699","#993300"))+
  theme_bw(base_size = 9)

g2 <- total%>%
  ggplot(aes(x=date, y = allMobile))+
  geom_point(aes(color=HighPeriod))+
  geom_line(y=10)+
  labs(x="Year", color="HighPeriod")+
  scale_color_manual(values=c("#336699","#993300"))+
```

```
  theme_bw(base_size = 9)

ggarrange(g1,g2, common.legend = TRUE, ncol=1, nrow=2,
    align="hv", legend = "right")

#Save dataframe
saveRDS(total, file="AnalyseData")
```