Master's thesis

NTNU
Norwegian University of Science and Technology

Inge Hoven Brakestad

Analysis of salmon lice count data for production zones 6 and 7 in Norway from 2017 to 2019

Master's thesis in Lektorutdanning i realfag
Supervisor: John Sølve Tyssedal
June 2020

NTNU
Norwegian University of Science and Technology

Inge Hoven Brakestad

Analysis of salmon lice count data for production zones 6 and 7 in Norway from 2017 to 2019

Master's thesis in Lektorutdanning i realfag
Supervisor: John Sølve Tyssedal
June 2020

Norwegian University of Science and Technology

NTNU
Norwegian University of
Science and Technology

Abstract

The salmon louse (*Lepeophtheirus salmonis*), is a parasite that represents direct threats to wild salmonids and indirectly threats to salmon farmers and farmed salmon. This thesis is written in cooperation with ”Taskforce Salmon lice”, a R & D project, where the objective is to ”establish fundamental knowledge on how sea lice infest farmed salmon and the mechanisms of how the parasites spread within and between farmed and wild populations of salmonids”.

Salmon lice count data were downloaded from the sources Barentswatch, the Norwegian Directorate of Fisheries and eKlima. Barentswatch gives information on activity in sea and coastal areas, and is based on cooperation between different Norwegian state agencies and research institutes. The Norwegian Directorate of Fisheries is responsible for fisheries and aquaculture management. Eklima is a web site which gives access to the database of climate data of the Norwegian Meteorological Institute.

To analyse the data, generalized linear models (GLMs) such as a Poisson regression model, a quasi-Poisson regression model, a negative binomial regression model and a zero inflated negative binomial regression model were fitted. A regression tree and a random forest analysis were applied. Finally, a generalized additive model (GAM) was fitted. The study suggested that the fitted GLMs did not give a good fit to the data.

Samandrag

Lakselus (*Lepeophtheirus salmonis*), er ein parasitt som er ei stor utfordring for vill laksefisk og for oppdrettsnæringa. Denne oppgåva er skriven i samarbeid med ”Taskforce Lakselus”, som er eit FoU-prosjekt, der målet er å ”etablere grunnleggjande kunnskapar om korleis lakselus angrip oppdrettslaks og mekanismane om korleis parasitten spreier seg innad og mellom oppdrettslaks og ville bestandar av laks”.

Teljedata om lakselus vart lasta ned frå Barentswatch, Fiskeridirektoratet og eKlima. Barentswatch har informasjon om aktivitet i kyst- og sjøområde, og baserer seg på samarbeid mellom norske, statlege etatar og forskningsinstitusjonar. Fiskeridirektoratet er ansvarlege for fiskeri- og havbruksforvaltninga i Noreg. eKlima er ei nettside som gir tilgang til databasen for klimadata frå Meterologisk Institutt.

For å analysere dataane har generaliserte lineære modellar (GLMar) som ein Poisson regresjonsmodell, ein quasi-Poisson regresjonsmodell, ein negativ-binomisk regresjonsmodell og ein null-inflatert negativ binomisk regresjonsmodell blitt tilpassa. Eit regresjonstre og ein random forest-analyse vart deretter laga. Til slutt vart det tilpassa ein generalisert additiv modell (GAM). Studien antyda at GLM-ane som vart brukt, ikkje gav ei god tilpasning til dataane.

Preface

This thesis was completed during the spring 2020, and marks the end of my studies at NTNU in Trondheim. It has been five fantastic years and I am grateful to everything I have learned during this period. I want to thank all my friends in Trondheim for making this such a great time. I also want to thank my family for supporting me during my time at the university.

To learn and apply statistics can be used in many situations. As a teacher, it is important to have a good understanding of the background of what the students learn, but also how mathematics and statistics can be applied in various fields. During the time I have worked with this thesis, I have hopefully developed a better understanding of both these aspects, which can make me able to teach in an interesting and meaningful way.

A special thank goes to my supervisors John Sølve Tyssedal at the Department of Mathematical science at NTNU, Lone Sunniva Jevne at NTNU SeaLab and
Arnfinn Aunsmo at the Faculty of Veterinary Medicine at NMBU. Thank you all for sharing your knowledge and your patience with me.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The challenges with salmon lice

The salmon louse (*Lepeopheheirus salmonis* Krøyer, 1837) represents one of the biggest challenges in The Norwegian aquaculture industry. The parasites cause physical damage to the fish by their attachment and feeding activities (Woo and Buchmann, 2012). When salmon lice are not feeding, they cling to the host by digging into their skin with claw-like antennae (Lester and Hayward, 2006). The presence of salmon lice on the skin is enough to cause stress to the fish (Ho, 2000). The skin damage caused by salmon lice also makes the fish more exposed to secondary bacterial infections (Thorstad and Finstad, 2018). Salmon lice increase risk of mortality of wild salmon smolt when migrating from the river into the sea (Thorstad and Finstad, 2018).

At the salmon farms, there are more hosts available which increases the abundance and thus the risk of spreading salmon lice in marine habitats (Thorstad and Finstad, 2018). Regulations in Norway state that there, on average, have to be fewer than 0.5 adult female salmon lice per salmon in a salmon farm. Fish farmers must count the number of lice per salmon, along with eventual treatment used and sea temperature. When the sea temperature is below $4°$C, the number of lice per salmon must be counted once every fourteenth days, and when the sea temperature exceeds $4°$C, the number of lice per salmon must be counted once a week. The limit of 0.5 adult female salmon lice per salmon is reduced to 0.2 in Trøndelag and southern regions in week 16 to Sunday in week 21. In Nordland and Troms og Finnmark the limit is reduced to 0.2 adult female salmon lice per salmon in week 21 to Sunday in week 26 (Regulations on salmon lice control, 2012).

1.2 Lifecycle and counting regulations of the salmon louse

The salmon louse has eight stages of development. After hatching from a string of eggs in the water, the salmon louse develops through two naupliar stages. In the naupliar stages, it is a free-living ocean drifter and unable to feed. It then develops into a copepodid. If the copepodid attaches to a host, it will progress through two sessile chalimus stages and thereafter through two mobile pre-adult stages. Finally, it develops into a mobile adult female or an adult male. In the two pre-adult stages and in the adult stage, the louse is mobile, and is therefore able to move across the surface of the fish and swim in the water column (Hayward et al., 2011; Maran et al., 2013; Hamre et al., 2013).

Fish farmers count and report the number of salmon lice, per salmon, for all stages on the fish. They count the number of sessile lice, which corresponds to the copepodid stage and the two sessile chalimus stages, and the number of mobile salmon lice which corresponds to the two mobile pre-adult stages and adult male. They count the number of adult female lice as a separate category, and these are not included in counts of mobile lice (Regulations on salmon lice control, 2012). The salmon lice count data can thus be stored in three different categories.

1.3 Salmon lice treatments

In this thesis, the salmon lice treatment methods were divided into three: Medicinal treatments, deployment of cleaner fish and mechanical delousing.

Medicinal treatments are administered either as a feed supplement or in a bath. Medicinal treatments have traditionally been used in order to control salmon lice in the fish farms (Helgesen et al., 2019). There are several challenges to the use of medicinal treatments. Salmon lice can become resistant (Poley et al., 2018). The treatments may harm the farms' surrounding environments, because uneaten medicinal feed can accumulate under cages that can damage non-target organisms (Olsvik et al., 2015).

Cleaner fish eat the salmon lice on the salmon without stressing the salmon. The majority of cleaner fish used in Norway until 2016 were species of wrasse, but between 2015 and 2018 the production of farmed lumpfish increased strongly, and surpassed the number of wrasses combined in 2017 (Rueness et al., 2019). There are different challenges in order to succeed with the use of cleaner fish. Cleaner fish efficiency is affected by sea temperature. Lumpfish are more effective than wrasse at low sea temperatures (Hjeltnes et al., 2019). Cleaner fish themselves can also be affected by diseases and can therefore have high mortality rates (Hjeltnes et al., 2019).

Salmon lice can be removed by mechanical delousing. These treatment methods include the use of lasers, flushing, brushing and fresh water baths. The salmon farming industry in Norway has largely retreated from medicinal treatments, and the use of mechanical delousing has become more common in recent years (Overton et al., 2018). Mechanical treatments are reportedly most effective against mobile lice stages, and less so against sessile stages (Torrissen et al., 2013).

1.4 Outline and objective of the thesis

The main goal of this thesis was to investigate possible factors determining the number of mobile salmon lice, reported from salmon farms in production zones 6 and 7, in Norway from 2017 to 2019.

In Chapter 2 the necessary statistical theory is presented. Chapter 3 contains information about datasets and variables. Chapter 4 gives a visualization of the data, and Chapter 5 is about analysis and validation of the fitted models. Finally, Chapter 6 contains a discussion and a conclusion with challenges and recommendations for further work.

Chapter 2

Theory

2.1 Poisson regression

Counts are non-negative integers that represent count of events per unit. A normal approximation can be useful, especially when the number of occurrences happen with high frequencies. On the other hand, discrete distributions are often more suitable for count data. One of the most common discrete distributions for count data is the Poisson distribution (Casella and Berger, 2002).

For a random Poisson distributed variable, Y_i, the probability mass function is

$$P(Y = y) = f(y|\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}, \quad y = 0, 1, 2, ... \tag{2.1}$$

For example, Y can correspond to the salmon lice counts at a salmon farm.

In a Poisson regression model each observation $Y_i, \quad i = 1, 2, ..., n$ are assumed to be independent Poisson distributed variables. For each Y_i, we have the covariates $\boldsymbol{x}_i^T = (x_{0i}, x_{1i}, ..., x_{ni})$ and the linear predictor $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$, where p is the number of parameters (intercept not included). A natural choice of link function is to choose the log because it allows the linear predictor to span the entire real line, and it ensures that the expected value is a positive number. Hence, we will have that $\mu_i = E[Y_i] = \lambda_i = \exp(\eta_i)$ (Fahrmeir et al., 2013).

2.1.1 Asymptotic estimation of the parameters

To derive the asymptotic estimation of the parameters in the Poisson distribution we have to derive the log-likelihood, the score function and the expected Fisher information. We recall the probability mass function from 2.1. Then, we take the logarithm of the distribution of a random variable Y_i.

$$l_i(\boldsymbol{\beta}) = \log f(y_i|\boldsymbol{\beta}) = \log\left(\frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}\right) = y_i \log \lambda_i - \lambda_i - \log(y_i!) \tag{2.2}$$

Looking at n equally distributed variables, we can find an expression for the log-likelihood by utilizing that $\lambda_i = \exp(\eta_i)$ along with equation (2.2).

$$\sum_{i=1}^{n} l_i(\boldsymbol{\beta}) = \sum_{i=1}^{n} \Big(y_i \eta_i - \exp(\eta_i) - \log(y_i!) \Big) \tag{2.3}$$

By differentiating l_i with respect to $\boldsymbol{\beta}$ and use the chain rule, we get

$$\frac{\partial l_i}{\partial \boldsymbol{\beta}} = \boldsymbol{x}_i (y_i - \exp(\eta_i)) \tag{2.4}$$

When we sum this expression from i to n we get the score function as

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i (y_i - \exp(\eta_i)) \tag{2.5}$$

The Fisher information is found by the expression

$$\boldsymbol{F}(\boldsymbol{\beta}) = E(\boldsymbol{s}(\boldsymbol{\beta}) \boldsymbol{s}(\boldsymbol{\beta})^T) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T E(y_i - \lambda_i)^2 = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \lambda_i \tag{2.6}$$

In order to solve $\boldsymbol{s}(\boldsymbol{\beta}) = 0$, we use the Fisher scoring algorithm

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{F}^{-1}\big(\hat{\boldsymbol{\beta}}^{(t)}\big) \boldsymbol{s}\big(\hat{\boldsymbol{\beta}}^{(t)}\big) \tag{2.7}$$

When $n \to \infty$ we get the asymptotic result that $\hat{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta}, \boldsymbol{F}^{(-1)}(\hat{\boldsymbol{\beta}})\big)$. Then, the estimated variances for each parameter are on the diagonal in the inverse Fisher information matrix.

The parameters in the model are estimated from the Fisher scoring algorithm. In order to construct confidence intervals, we use that $\hat{\boldsymbol{\beta}} \sim N\big(\boldsymbol{\beta}, \boldsymbol{F}^{(-1)}(\hat{\boldsymbol{\beta}})\big)$. We denote the square root of each diagonal element as $\sqrt{F^{-1}(\hat{\beta})_{jj}}$, for $j = 1, 2, ..., p+1$. We have, for each element of the parameter vector, that

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{F^{-1}(\hat{\beta})_{jj}}}, \quad \approx N(0,1) \tag{2.8}$$

A 95% confidence interval for a parameter β_j is then given by

$$\hat{\beta}_j \pm 1.96 \cdot \sqrt{F^{-1}(\hat{\beta})_{jj}} \tag{2.9}$$

(Fahrmeir et al., 2013).

2.1.2 Pearson and deviance residuals and goodness of fit

In the Poisson regression model the deviance statistic is defined as

$$D = 2 \sum_{i=1}^{n} \Big[y_i \log\Big(\frac{y_i}{\hat{\mu}_i}\Big) - (y_i - \hat{\mu}_i) \Big], \tag{2.10}$$

where $\hat{\mu}_i = \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. Then, the deviance residuals are given as

$$d_i = \text{sign}(y_i - \hat{\mu}_i) \cdot 2\sqrt{\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right)}, \tag{2.11}$$

where $\text{sign}(y_i - \hat{\mu}_i) = -1$ if $y_i - \hat{\mu}_i < 0$ and $\text{sign}(y_i - \hat{\mu}_i) = 1$ if $y_i - \hat{\mu}_i \geq 0$.
The Pearson statistic is given by

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{2.12}$$

The Pearson residuals are given by

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \tag{2.13}$$

The deviance and Pearson statistics are approximately χ^2- distributed with $n - p$ degrees of freedom, and can be used to evaluate the goodness of fit. If $D < \chi^2_{0.05,n-p}$, there is no evidence to believe that the model is a bad fit to the data. The Pearson statistic is often used as a test for overdispersion.

2.1.3 Poisson random intercept model

In these models we add a random effect to the linear predictor from section 2.1. The random effects are assumed to be independent and identically normally distributed variables, i.e. $\gamma_{0i} \sim N(0, \tau_0^2)$. Then we have the expression $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_{0i}$. If we condition on the random effects $\gamma_{01}, ..., \gamma_{0n}$, where $i = 1, 2, ..., n$, it is assumed that $y_i | \gamma_{0i} \sim \text{Po}(\lambda_i)$ are independent, where $\lambda_i = \exp(\eta_i)$, with $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \gamma_{0i}$ (Fahrmeir et al., 2013).

2.1.4 Dispersion parameter

For a Poisson distributed random variable Y_i, $E(Y_i) = \text{Var}(Y_i)$. Overdispersion is the case where the variance in a dataset is higher than what is expected by the Poisson regression model. For the Poisson distribution, this means that $\text{Var}(Y_i) > E(Y_i)$. The dispersion parameter can be estimated by

$$\hat{\phi} = \frac{D}{n - p}, \quad \text{or} \quad \hat{\phi} = \frac{P}{n - p}, \tag{2.14}$$

where D is the deviance statistic given from equation (2.10) and P is the Pearson statistic given from equation (2.12), n is the number of observations and p is the number of parameters. We can take this into account by multiplying the covariance matrix, $\boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}})$, by the dispersion parameter. This leads us to the quasi-Poisson model (Fahrmeir et al., 2013).

2.1.5 quasi-Poisson model

In the quasi-Poisson model, $\mu_i = E[Y_i] = \lambda_i = \exp(\eta_i)$ and $\text{Var}(Y_i) = \phi \mu_i$, where ϕ is the dispersion parameter estimated in equation (2.14). Thus, this changes the estimated

variance in the Fisher information matrix. Hence, the standard errors of the maximum likelihood estimate are the square root of the diagonal of $\boldsymbol{F}^{-1}$ multiplied with $\hat{\phi}$. Then, for an element j in the parameter vector, the distribution is

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\phi} F^{-1}(\beta)_{jj}}}, \quad \approx t_{n-p} \tag{2.15}$$

(Fahrmeir et al., 2013).

2.2 Negative binomial regression

2.2.1 Assumptions in the Negative binomial regression model

In cases where we have overdispersion, i.e. the dispersion parameter $\phi > 1$, it can be useful to use a negative binomial response function because, since a negative binomial distributed random variable Y, allows that $\text{Var}(Y) > E(Y)$.
One way to write the probability mass function of the negative binomial distribution is

$$P(Y = y) = \binom{r + y - 1}{y} p^r (1 - p)^y, \quad y = 0, 1, ... \tag{2.16}$$

where $Y =$ "number of failures before the r-th success", where r is a fixed number. We can alternatively look at $X = Y + r$, where it follows that $X =$ "trial at which the r-th success occurs". The negative binomial distribution can, like the Poisson, be used to model phenomena in which we are waiting for an occurrence. In the negative binomial case we are waiting for a specified number of successes (Casella and Berger, 2002).

We can derive the expected value, $E(Y)$. By use of equation (2.16) and by the definition of the expected value, this can be written as

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} y \binom{r + y - 1}{y} p^r (1 - p)^y \\ &= \sum_{y=1}^{\infty} \frac{(r + y - 1)!}{(y - 1)!(r - 1)!} p^r (1 - p)^y \\ &= \sum_{y=1}^{\infty} r \binom{r + y - 1}{y - 1} p^r (1 - p)^y \end{aligned} \tag{2.17}$$

If we set $z = y - 1$, we can write the expression for $E(Y)$ as

$$E(Y) = r \frac{(1 - p)}{p} \sum_{z=0}^{\infty} \binom{(r + 1) + z - 1}{z} p^{r+1} (1 - p)^z \tag{2.18}$$

(Casella and Berger, 2002). The last term is the sum over all point probabilities in a negative binomial distributed variable, $Z \sim NB(r + 1, p)$, which means that this term is equal to 1. Hence, the expected value is equal to $\text{E}(Y) = r\frac{(1-p)}{p}$. In order to find the

variance, the same idea can be used, or it can be found via the moment generating function. We find that $\text{Var}(Y) = \frac{r(1-p)}{p^2}$. If we write $\mu = \text{E}(Y) = r\frac{1-p}{p}$, we get that $\mu = \frac{r}{p} - r$. Then, $\text{Var}(Y) = \mu + \frac{1}{r}\mu^2$. If we solve this for p, we get that $p = \frac{r}{\mu+r}$. Further, if we replace p in equation (2.16) by the previous expression, the probability mass function of Y can be written in another form as

$$\begin{aligned} P(Y = y) &= \binom{r+y-1}{y}\Big(\frac{r}{\mu+r}\Big)^r\Big(1-\frac{r}{\mu+r}\Big)^y, \quad y = 0, 1, ... \\ &= \frac{(r+y-1)!}{y!(r-1)!}\Big(\frac{r}{\mu+r}\Big)^r\Big(1-\frac{r}{\mu+r}\Big)^y, \quad y = 0, 1, ... \\ &= \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)}\Big(\frac{r}{\mu+r}\Big)^r\Big(1-\frac{r}{\mu+r}\Big)^y, \quad y = 0, 1, ... \end{aligned} \tag{2.19}$$

Another way to derive the distribution from (2.19) is to consider a gamma distributed random variable T. The probability density function of this variable can be written

$$P(T = t) = \frac{1}{\Gamma(\alpha)\beta^\alpha}t^{\alpha-1}\exp(-t/\beta), \quad t > 0, \quad \alpha, \beta > 0, \tag{2.20}$$

where α is the shape parameter and β is the scale parameter (Casella and Berger, 2002). Let us consider a gamma distributed random variable, T, with shape parameter $\alpha = r$ and rate parameter $\beta = 1/r$. If the distribution of Y given T is Po(μT), hence if

$$f(y|t) = \frac{\exp(-\mu t)(\mu t)^y}{y!}, \tag{2.21}$$

we can derive the marginal distribution of $f(y)$ as

$$f(y) = \int_0^\infty f(y|t)g(t)dt, \tag{2.22}$$

where $g(t) = \frac{r^r}{\Gamma(r)}t^{r-1}\exp(-rt)$. Multiplying out, we get that

$$\begin{aligned} f(y) &= \int_0^\infty f(y|t)g(t)dt \\ &= \int_0^\infty \frac{\exp(-\mu t)(\mu t)^y}{y!}\frac{r^r}{\Gamma(r)}t^{r-1}\exp(-rt)dt \\ &= \frac{r^r\mu^y}{\Gamma(r)\Gamma(y+1)}\int_0^\infty t^{r-1+y}\exp(t(-\mu-r))dt \\ &= \frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)}\Big(\frac{r}{\mu+r}\Big)^r\Big(\frac{\mu}{\mu+r}\Big)^y \end{aligned} \tag{2.23}$$

which is equal to the probability mass function in equation (2.19) (Casella and Berger, 2002).

We derive the asymptotic estimation of the parameters in the negative binomial distribution. We take the logarithm of the distribution of Y:

$$\begin{aligned}\log f(y) &= \log\Big(\frac{\Gamma(r+y)}{\Gamma(y+1)\Gamma(r)}\Big(\frac{r}{\mu+r}\Big)^r\Big(\frac{\mu}{\mu+r}\Big)^y\Big)\\ &= \log\Gamma(r+y) - \log\Gamma(y+1) - \log\Gamma(r) + r\log(r) - r\log(\mu+r)\\ &+ y\log(\mu) - y\log(\mu+r)\end{aligned} \tag{2.24}$$

(Dominique and Park, 2010). Let now $Y_i = NB(\mu_i, r), \quad i = 1, 2, ..., n$ be independent distributed variables. An expression for the log-likelihood function using that $\mu_i = E(Y) = \exp(\eta_i) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})$ is

$$\begin{aligned}\sum_{i=1}^{n} l_i(\boldsymbol{\beta}, r) = \sum_{i=1}^{n}\Bigg(&\log\Big(\frac{\Gamma(r+y_i)}{\Gamma(r)}\Big) - \log\Gamma(y_i+1) +\\ &r\log(r) - r\log(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}+r) + y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i\log(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r)\Bigg)\end{aligned} \tag{2.25}$$

By utilizing the relationship that $\log\Big(\frac{\Gamma(y_i+r)}{\Gamma(r)}\Big) = \sum_{k=0}^{y_i-1}\log(k+r)$, we can write that

$$\begin{aligned}\sum_{i=1}^{n} l_i(\boldsymbol{\beta}, r) = \sum_{i=1}^{n}\Bigg(&\sum_{k=0}^{y_i-1}\log(k+r) - \log\Gamma(y_i+1) + r\log(r)\\ &- r\log(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r) + y_i\boldsymbol{x}_i^T\boldsymbol{\beta} - y_i\log(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r)\Bigg)\end{aligned} \tag{2.26}$$

By looking at the derivative of l_i with respect to $\boldsymbol{\beta}$, the first three terms from (2.26) disappears and we end up with

$$\frac{\partial l_i}{\partial\boldsymbol{\beta}} = y_i\boldsymbol{x}_i - \frac{r\boldsymbol{x}_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + y_i\boldsymbol{x}_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r} = \frac{\boldsymbol{x}_i r(y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r} \tag{2.27}$$

Here, we have used that $\frac{\partial}{\partial\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta}) = \boldsymbol{x}_i$ and then $\frac{\partial}{\partial\boldsymbol{\beta}}(\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\boldsymbol{x}_i$. Similarly, we find the derivative of l_i with respect to r

$$\frac{\partial l_i}{\partial r} = \sum_{k=0}^{y_i-1}\Big(\frac{1}{k+r}\Big) + \log\Big(\frac{r}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r}\Big) + \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+y_i}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r} \tag{2.28}$$

When we sum this expression from 1 to n we get the score function as

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\Big(y_i\boldsymbol{x}_i - \frac{r\boldsymbol{x}_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) + y_i\boldsymbol{x}_i\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r}\Big) = \sum_{i=1}^{n}\frac{\boldsymbol{x}_i r(y_i - \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}))}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r} \tag{2.29}$$

and

$$\boldsymbol{s}(r) = \sum_{i=1}^{n}\Bigg(\sum_{k=0}^{y_i-1}\Big(\frac{1}{k+r}\Big) + \log\Big(\frac{r}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r}\Big) + \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+y_i}{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta})+r}\Bigg) \tag{2.30}$$

Then, we find the matrix, $\boldsymbol{F}_{11}(\boldsymbol{\beta})$, by the expression

$$\boldsymbol{F}_{11}(\boldsymbol{\beta}) = E(\boldsymbol{s}(\boldsymbol{\beta})(\boldsymbol{s}(\boldsymbol{\beta})^T) = \sum_{i=1}^{n} \frac{\boldsymbol{x}_i \boldsymbol{x}_i^T (r^2\mu_i + r\mu_i^2)}{(\mu_i + r)^2}, \tag{2.31}$$

where $\mu_i = \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})$ and the matrix, $\boldsymbol{F}_{22}(r)$, as

$$\boldsymbol{F}_{22}(r) = E(\boldsymbol{s}(r)\boldsymbol{s}(r)^T) = \sum_{i=1}^{n} \left(\Big(\sum_{k=0}^{y_{i-1}} \Big(\frac{1}{k+r}\Big) + \log\Big(\frac{r}{\mu_i + r}\Big)\Big)^2 + \frac{\mu_i}{(\mu_i + r)^2} \right) \tag{2.32}$$

In order to solve $\boldsymbol{s}(\boldsymbol{\beta}) = 0$, we can use the Fisher scoring algorithm

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \hat{\boldsymbol{\beta}}^{(t)} + \boldsymbol{F}_{11}^{-1}\big(\hat{\boldsymbol{\beta}}^{(t)}\big)\boldsymbol{s}\big(\hat{\boldsymbol{\beta}}^{(t)}\big) \tag{2.33}$$

and we can solve $\boldsymbol{s}(r) = 0$ similarly by

$$\hat{r}^{(t+1)} = \hat{r}^{(t)} + \boldsymbol{F}_{22}^{-1}\big(\hat{r}^{(t)}\big)\boldsymbol{s}\big(\hat{r}^{(t)}\big) \tag{2.34}$$

When $n \to \infty$ we get the asymptotic result that

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{r} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\beta} \\ r \end{bmatrix}, \begin{bmatrix} \boldsymbol{F_{11}(\hat{\beta})^{-1}} & 0 \\ 0 & \boldsymbol{F}_{22}(\hat{r})^{-1} \end{bmatrix} \right) \tag{2.35}$$

We have that $\text{Cov}(\hat{r}, \hat{\boldsymbol{\beta}}) = \text{Cov}(\hat{\boldsymbol{\beta}}, \hat{r}) = \boldsymbol{F}_{12} = \boldsymbol{F}_{21} = 0$, because that $E\Big(\frac{-\partial \boldsymbol{s}(\boldsymbol{\beta})}{\partial r}\Big) = E\Big(\frac{-\partial \boldsymbol{s}(r)}{\partial \beta}\Big) = E\Big(\frac{x_i(y_i - \mu_i)(\mu + r) - x_i r(y - \mu)}{(\mu + r)^2}\Big) = 0$, (Nakashima, 1997).
Then we have that, for each element of the parameter vector,

$$Z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{F_{11}^{-1}(\hat{\beta})_{jj}}}. \tag{2.36}$$

We can find a 95% confidence interval for a parameter β_j by

$$\hat{\beta}_j \pm 1.96 \cdot \sqrt{F_{11}^{-1}(\hat{\beta})_{jj}} \tag{2.37}$$

Further, we have that,

$$Z = \frac{\hat{r} - r}{\sqrt{F_{22}^{-1}(\hat{r})}}. \tag{2.38}$$

We can find a 95% confidence interval for a parameter r by

$$\hat{r} \pm 1.96 \cdot \sqrt{F_{22}^{-1}(\hat{r})} \tag{2.39}$$

In order to construct the confidence intervals, we have used the asymptotic result from equation (2.35).

2.2.2 Pearson and deviance residuals and goodness of fit

In the negative binomial regression model, the deviance statistic is defined as

$$D = 2\sum_{i=1}^{n}\Big[y_i \log\Big(\frac{y_i}{\hat{\mu}_i}\Big) - (y_i + r)\log\Big(\frac{r + y_i}{r + \hat{\mu}_i}\Big)\Big], \tag{2.40}$$

where $\hat{\mu}_i = \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})$. The deviance residuals are given as

$$d_i = \text{sign}(y_i - \hat{\mu}_i) \cdot \sqrt{2\Big(y_i \log\Big(\frac{y_i}{\hat{\mu}_i}\Big) - (y_i + r)\log\Big(\frac{r + y_i}{r + \hat{\mu}_i}\Big)\Big)} \tag{2.41}$$

The Pearson statistic in the negative binomial regression model is given as

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i + r^{-1}\hat{\mu}_i^2} \tag{2.42}$$

The formula for the Pearson residuals is

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + r^{-1}\hat{\mu}_i^2}} \tag{2.43}$$

2.3 Zero-inflated Poisson and negative binomial regression model

In data where there are more zeros than expected by a Poisson regression model or negative binomial regression model, a ZIP or ZINB model can be fitted to adjust for this. Such models mix two generating processes. The first have a probability of being zero given by π and follows a binomial distribution. The other may follow a Poisson or negative binomial distribution, which also include zero. Hence, the probability mass function can be partitioned into two parts, the probability of a zero count, and a probability of a count bigger than zero

$$P(Y = y) = \begin{cases} \pi + (1 - \pi)g(y = 0), & \text{if} \quad y = 0. \\ (1 - \pi)g(y), & \text{if} \quad y > 0. \end{cases}, \tag{2.44}$$

where $g(y)$ can, for example, be either the Poisson probability mass function or the negative binomial probability mass function given from equation (2.1) and (2.19), respectively. If we choose the probability mass function from equation (2.1), we have that Y is zero-inflated Poisson distributed or if we choose the probability mass function from equation (2.19) we have that Y is zero-inflated negative binomial distributed. It can be shown that, if $g(y)$ is the probability mass function of a Poisson distributed variable from equation (2.1), the mean and the variance of the ZIP distributed variable is equal to

$$\begin{aligned} \text{E}(Y) &= \mu(1 - \pi) \\ \text{Var}(Y) &= (1 - \pi)(\mu + \pi\mu^2) \end{aligned} \tag{2.45}$$

Otherwise, if $g(y)$ is is the probability mass function of a negative binomial distributed variable from equation (2.19) it can be shown that the mean and the variance of the ZINB distributed variable is equal to

$$\begin{aligned} \mathrm{E}(Y) &= \mu(1-\pi) \\ \mathrm{Var}(Y) &= (1-\pi)(\mu + \frac{\mu^2}{r}) + \mu^2(\pi^2 + \pi) \end{aligned} \tag{2.46}$$

Let us consider a regression model, with each observation $Y_i, \quad i = 1, 2, ..., n$ being independent variables with probability mass function given by equation (2.44). To model the positive count, we can use the log link function. We then have, $\mu_i = \exp \eta_i$, where $\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$. In order to model the probability of being zero given by π, we can use the logit function, which gives

$$\pi_i = \frac{\exp \gamma_0 + \gamma_1 z_{i1} + ... + \gamma_q z_{iq}}{1 + \exp \gamma_0 + \gamma_1 z_{i1} + ... + \gamma_p z_{iq}}, \tag{2.47}$$

where we use $\boldsymbol{z}_i$ because these points might be different from $\boldsymbol{x}_i$ (Zuur et al., 2009).

2.3.1 Asymptotic distribution of parameters

We derive the asymptotic distribution of the parameters. We assume that Y_i is a zero inflated negative binomial distributed random variable. We start by finding the likelihood function as

$$\begin{aligned} L[\boldsymbol{\beta}, r, \boldsymbol{\alpha}] = \prod_{i=1}^{n} &\Big[\pi_i + (1-\pi_i)\Big(\frac{r}{\mu_i + r}\Big)^r\Big]^{t(y_i)} \\ &\cdot \Big[(1-\pi_i)\frac{\Gamma(r+y_i)}{\Gamma(y_i+1)\Gamma(r)}\Big(\frac{r}{\mu_i+r}\Big)^r\Big(1 - \frac{r}{\mu_i+r}\Big)^{y_i}\Big]^{1-t(y_i)} \end{aligned} \tag{2.48}$$

$$\text{where } t(y_i) = \begin{cases} 1, & \text{if} \quad y_i = 0 \\ 0, & \text{else} \end{cases}$$

Then, we can find the log-likelihood function as

$$\begin{aligned} \sum_{i=1}^{n} l_i(\boldsymbol{\beta}, r, \boldsymbol{\gamma}) = &\sum_{i=1, y_i=0}^{n} \log\Big(\pi_i + (1-\pi_i)\Big(\frac{r}{\mu_i+r}\Big)^r\Big) \\ &+ \sum_{i=1, y_i>0}^{n} \Bigg(\log(1-\pi_i) + \sum_{k=0}^{y_i-1} \log(k+r) - \log\Gamma(y_i+1) + r\log(r) \\ &- r\log(\mu_i + r) + y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - y_i \log(\mu_i + r) \Bigg) \end{aligned}$$

The log-likelihood function can be divided into L_1 and L_2, where L_1 corresponds to the case if $t(y_i) = 1$, that is, for $y_i = 0$. L_2 corresponds to when $t(y_i) = 0$ which means

$y_i > 0$. Hence, the log-likelihood function can be written as

$$L = L_1 + L_2$$

where

$$\begin{aligned} L_1 &= \sum_{i=1,y_i=0}^{n} \Big(\log\Big(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \Big(\frac{r}{\mu_i + r}\Big)^r \Big) - \log(1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})) \Big) \\ L_2 &= \sum_{i=1,y_i>0}^{n} \Big(\Big(\sum_{k=0}^{y_i-1} \log(k+r) \Big) - \log\Gamma(y_i+1) + r\log(r) \\ &\quad - r\log(\mu_i + r) + y_i \boldsymbol{x}_i^T\boldsymbol{\beta} - y_i \log(\mu_i + r) - \log(1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})) \Big) \end{aligned} \tag{2.49}$$

We have to study the derivatives of first and second order

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{y_i=0,i=1}^{n} -\frac{r^{r+1}\boldsymbol{x}_i\mu_i}{(\mu_i + r)((\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}))(\mu_i + r)^r + r^r)} \\ &\quad + \sum_{y_i>0,i=1}^{n} \frac{\boldsymbol{x}_i r(y_i - \mu_i)}{\mu_i + r} \end{aligned} \tag{2.50}$$

$$\begin{aligned} \frac{\partial l}{\partial \boldsymbol{\gamma}} &= \sum_{y_i=0,i=1}^{n} \Big(-\frac{\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + (\frac{r}{\mu+r})^r)} - \frac{\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})} \Big) \\ &\quad - \sum_{y_i>0,i=1}^{n} \frac{\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})} \end{aligned} \tag{2.51}$$

$$\begin{aligned} \frac{\partial l}{\partial r} &= \sum_{y_i=0,i=1}^{n} \Big(-\frac{r^r\left((r+\mu_i)\log(r+\mu_i) - (r+\mu_i)\log(r) - \mu_i\right)}{(r+\mu_i)\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)} \Big) \\ &\quad + \sum_{y_i>0,i=1}^{n} \Big(\sum_{k=0}^{y_i-1} \big(\frac{1}{k+r}\big) + \log\big(\frac{r}{\mu_i + r}\big)\frac{\mu_i + y_i}{\mu_i + r} \Big) \end{aligned} \tag{2.52}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} &= \sum_{y_i=0,i=1}^{n} \frac{r^{r+2}\boldsymbol{x}_i^2\mu_i((\mu_i+r)^r(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i - \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) - r^r))}{(\mu_i+r)^2(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + r)^r + r^r)^2} \\ &\quad + \sum_{y_i>0,i=1}^{n} \frac{-\boldsymbol{x}_i^2 r\mu_i(\mu_i + r) - (\boldsymbol{x}_i r y_i - \boldsymbol{x}_i r\mu_i)(\boldsymbol{x}_i\mu_i + r)}{(\mu_i + r)^2} \end{aligned} \tag{2.53}$$

$$\begin{aligned} \frac{\partial^2 l}{\partial \boldsymbol{\gamma}^2} &= \sum_{y_i=0,i=1}^{n} \Bigg(\frac{\boldsymbol{z}_i^2 \exp(2\boldsymbol{z}_i^T\boldsymbol{\gamma})}{\Big(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \frac{r^r}{(r+\mu)^r}\Big)^2} + \frac{\boldsymbol{z}_i^2 \exp(2\boldsymbol{z}_i^T\boldsymbol{\gamma})}{\big(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + 1\big)^2} - \frac{\boldsymbol{z}_i^2 \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + \frac{r^r}{(r+\mu)^r}} \\ &\quad - \frac{\boldsymbol{z}_i^2 \exp(\boldsymbol{z}_i\boldsymbol{\gamma})}{\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma} + 1)} \Bigg) \\ &\quad - \sum_{y_i>0,i=1}^{n} \frac{\boldsymbol{z}_i^2 \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})}{(1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}))^2} \end{aligned} \tag{2.54}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial r^2} = \sum_{y_i=0,i=1}^{n} \Bigg(& \frac{r^r \left((r+\mu_i)\log(r+\mu_i) - (r+\mu_i)\log(r) - \mu_i\right)\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r\right)}{(r+\mu_i)\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^2} \\ & \cdot \left(\log(r+\mu_i) + \frac{r}{r+\mu_i}\right) r^r \left(\log(r)+1\right)\Bigg) \\ & - \frac{r^r\left(\ln(r)+1\right)\left((r+\mu_i)\log(r+\mu_i) - (r+\mu_i)\log(r) - \mu_i\right)}{(r+\mu_i)\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)} \\ & + \frac{r^r\left((r+\mu_i)\log(r+\mu_i) - (r+\mu_i)\log(r) - \mu_i\right)}{(r+\mu_i)^2\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)} - \frac{r^r\left(\log(r+\mu_i) - \log(r) - \frac{r+\mu_i}{r} + 1\right)}{(r+\mu_i)\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)}\Bigg) \\ & + \sum_{y_i>0,i=1}^{n}\left(\left(\sum_{k=0}^{y_i-1} -\frac{1}{(r+k)^2}\right) - \frac{yr-\mu^2}{r(r+\mu)^2}\right)\end{aligned} \tag{2.55}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\gamma}\partial r} = \sum_{y_i=0,i=1}^{n} \frac{\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) r^r (r+\mu_i)^{r-1}\left((r+\mu_i)\log(r+\mu_i) + (-r-\mu_i)\log(r) - \mu_i\right)}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^2} \tag{2.56}$$

$$\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial \boldsymbol{\gamma}} = \sum_{y_i=0,i=1}^{n} \frac{r^{r+1}\boldsymbol{x}_i(\mu_i+r)^{r-1}\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i}{\left((\mu_i+r)^r \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma}) + r^r\right)^2} \tag{2.57}$$

$$\begin{aligned}\frac{\partial^2 l}{\partial \boldsymbol{\beta}\partial r} = \sum_{y_i=0,i=1}^{n} -\Bigg(& \frac{2\boldsymbol{x}_i\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i r^{r+1}(r+\mu_i)^{r-1}\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r\right)}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^3} \\ & \cdot \left(\log(r+\mu_i) + \frac{r}{r+\mu_i}\right) - \frac{2\boldsymbol{x}_i\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i r^{r+1}(r+\mu_i)^{r-1}\left(r^r\log(r)+1\right)}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^3} \\ & + \frac{\boldsymbol{x}_i\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i r^{r+1}(r+\mu_i)^{r-1}\left(\log(r+\mu_i) + \frac{r-1}{r+\mu_i}\right)}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^2} \\ & + \frac{\boldsymbol{x}_i\boldsymbol{z}_i \exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})\mu_i r^{r+1}(r+\mu_i)^{r-1}\left(\log(r) + \frac{r+1}{r}\right)}{\left(\exp(\boldsymbol{z}_i^T\boldsymbol{\gamma})(r+\mu_i)^r + r^r\right)^2}\Bigg) \\ & + \sum_{y_i>0,i=1}^{n} \frac{\boldsymbol{x}_i\mu_i(y_i-\mu_i)}{(r+\mu_i)^2}\end{aligned} \tag{2.58}$$

When $n \to \infty$ we get the asymptotic result that

$$\begin{bmatrix}\hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\gamma}} \\ \hat{r}\end{bmatrix} \sim N\left(\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\gamma} \\ r\end{bmatrix}, \begin{bmatrix} -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}} & -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}} & -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial r} \\ -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}} & -\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}} & -\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial r} \\ -\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial r} & -\frac{\partial^2 l}{\partial\boldsymbol{\gamma}\partial r} & -\frac{\partial^2 l}{\partial r\partial r}\end{bmatrix}^{-1}\right) \tag{2.59}$$

2.3.2 Pearson residuals and goodness of fit

The Pearson statistic is given as

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i(1-\hat{\pi}_i))^2}{(1-\hat{\pi}_i)(\hat{\mu}_i + \frac{\hat{\mu}_i^2}{\hat{r}}) + \hat{\mu}_i^2(\hat{\pi}_i^2 + \hat{\pi}_i)} \tag{2.60}$$

The formula for the Pearson residuals for the zero inflated negative binomial regression is given as

$$r_i = \frac{y_i - \hat{\mu}_i(1-\hat{\pi}_i)}{\sqrt{(1-\hat{\pi}_i)(\hat{\mu}_i + \frac{\hat{\mu}_i^2}{\hat{r}}) + \hat{\mu}_i^2(\hat{\pi}_i^2 + \hat{\pi}_i)}} \tag{2.61}$$

2.4 AIC and BIC

For the models that we have discussed above, the Akaike information criterion (AIC) is given as

$$\text{AIC} = -2l + 2p - 2 \tag{2.62}$$

where l is the maximum log-likelihood and p is the number of regression parameters where the intercept is not included. The Bayesian information criterion (BIC) which also takes the number of data points, n, into account, is defined as

$$\text{BIC} = -2l + \log(n)(p+1) = -2l + p\log(n) + \log(n) \tag{2.63}$$

Goodness of fit is rewarded by both AIC and BIC, but these criteria also include a penalty term for the number of parameters. BIC also include a penalty for the number of data-points, n. The penalty terms discourage overfitting, because adding parameters that does not improve the goodness of fit, results in higher AIC and BIC. Hence, both AIC and BIC are used to assess the quality of statistical models. The model with the lowest AIC or BIC is preferred (Fahrmeir et al., 2013).

When fitting models, backward elimination and forward selection can be applied. In backward elimination we start with a full model and look at the AIC and BIC value and then remove one term in turn and study the change in AIC and BIC. On the other hand, in forward selection we start with an empty model and add terms and then extract the corresponding AIC and BIC value from each model, and finally compare these. In both approaches, the model that gives the lowest AIC or BIC values is considered the best.

2.5 Hypothesis testing

2.5.1 Likelihood ratio test

We consider a parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_p)^T \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is the parameter space. $\boldsymbol{\theta}$ can, for example, correspond to $\boldsymbol{\beta}$ of regression coefficients in a Poisson regression model. We want to test the nullhypothesis, $H_0 : \boldsymbol{C\theta} = \boldsymbol{d}$ against the alternative hypothesis, H_1 that $\boldsymbol{C\theta} \neq \boldsymbol{d}$. Here, $\boldsymbol{C}$ is a $r \times (p+1)$ matrix and $\boldsymbol{d}$ is a column vector of dimension r, where r corresponds to the number of hypothesis that is being tested (Fahrmeir et al., 2013). Then, we can consider the likelihood ratio statistic

$$\lambda(\boldsymbol{x}) = \frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})} \tag{2.64}$$

Here, $L(\hat{\boldsymbol{\theta}})$ corresponds to the maximum of the likelihood under H_0, and $L(\tilde{\boldsymbol{\theta}})$ under H_1. $\lambda(\boldsymbol{x})$ will have values ≥ 1 and we reject H_0 for large values of $\lambda(\boldsymbol{x})$. If we take the logarithm of equation (2.64) and multiply with -2, we obtain the log-likelihood ratio test statistic as

$$-2\log(\lambda(\boldsymbol{x})) = -2\log\left(\frac{L(\hat{\boldsymbol{\theta}})}{L(\tilde{\boldsymbol{\theta}})}\right) = -2(\log(l(\hat{\boldsymbol{\theta}})) - \log(l(\tilde{\boldsymbol{\theta}}))) \tag{2.65}$$

$-2\log(\lambda(\boldsymbol{x}))$ is asymptotically χ^2 distributed under the null hypothesis where the degrees of freedom is equal to r, that is the difference in dimensionality from $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ (Fahrmeir et al., 2013), (Casella and Berger, 2002). The likelihood ratio test can be used to test nested models. Two models are nested when one model contains all the other terms of the other, and at least one additional term. We can for example think about the regression coefficients in a Poisson regression with the parameter vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$. We want to test the null hypothesis that $\beta_p = 0$, *versus* the alternative hypothesis that $\beta_p \neq 0$. Let model A correspond to the model where $\beta_p = 0$ and model B correspond to the model where $\beta_p \neq 0$. Model A is then nested within model B because model B contains all the terms of model A and also β_p. The log-likelihood ratio test statistic can be calculated as

$$-2\log(\lambda(\boldsymbol{x})) = -2(\log(l(\boldsymbol{\beta}_A) - \log(l(\boldsymbol{\beta}_B)) \tag{2.66}$$

This statistic is asymptotically χ^2-distributed with degrees of freedom equal to $p+1-p = 1$.

2.5.2 Wald test

For the Wald test we can calculate the Wald test statistic defined as

$$\boldsymbol{w} = (\boldsymbol{C\hat{\theta}} - \boldsymbol{d})^T(\boldsymbol{C\hat{V}C}^T)^{-1}(\boldsymbol{C\hat{\theta}} - \boldsymbol{d}) \tag{2.67}$$

We have that $\text{Cov}(\boldsymbol{C\hat{\theta}} - \boldsymbol{d}) = \text{Cov}(\boldsymbol{C\hat{\theta}}) = \boldsymbol{C\hat{V}C}^T$, and $\hat{\boldsymbol{V}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$. The Wald test statistic measures the distance from the estimate of $\boldsymbol{C\theta}$ that is $\boldsymbol{C\hat{\theta}}$ and the value of $\boldsymbol{d}$ under H_0. A large value of $\boldsymbol{w}$ indicates that the distance from

the estimate and $\boldsymbol{d}$ is large, and then H_0 should be rejected. If we consider a test for one single element of the parameter vector in a Poisson regression model, where H_0 is that $\beta_j = 0$ *versus* the alternative hypothesis, H_1 that $\beta_j \neq 0$ the Wald test statistic is equal to $\boldsymbol{w} = t_j^2 = \frac{\hat{\beta}_j^2}{F^{-1}(\hat{\beta})_{jj}}$. When $n \to \infty$ we get the asymptotic result that $t_j \approx N(0,1)$, and we reject the null hypothesis if the test statistic $|z_j| > z_{1-\alpha/2}$ (Fahrmeir et al., 2013).

2.6 Methods from statistical learning

In the rest of this chapter, important concepts that will be introduced, will use the re-sampling methods cross validation and bootstrapping. Therefore, the ideas behind these two methods will be explained.

2.6.1 Cross-validation

In the validation set approach, the observations n are randomly divided into two groups, that is, a training set and a validation set. A model is fitted to the training set, and the validation set is used to evaluate the fit of the model. The mean squared error (MSE) is used to provide an estimate for the test error rate (James et al., 2013).

In the previous method, the MSE can be highly variable, depending on which observations that are in the training set and which that are in the validation. set. One way to try to handle this is to use leave-one-out cross-validation. The idea here is to include every observation point but one in the training set, and use these points to fit a model. If point i is excluded from the training set, we can use this point to evaluate the model fit by calculating the mean squared error as $\text{MSE}_i = (y_i - \hat{y}_i)^2$. This is repeated for all n points such that each point in the dataset has been excluded once from the training set. The total estimate for the MSE can then be calculated as

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

In polynomial regression, the leave-one-out cross-validation (LOOCV) estimate for the test MSE is calculated as

$$\text{CV} = \sum_{i=1}^{n}\Big(\frac{y_i - \hat{y}_i}{1 - h_i}\Big),$$

where h_i is the i-th diagonal element of the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}$ and $\hat{y}_i$ is the i-th fitted value from the least squares fit.

K-fold cross validation can also be used. The idea here is to divide the data into k folds, and include all k folds but one in the training set in order to fit a model. The left-out fold is used as validation set, and this is done in turn, such that every fold is left out of the training set once. For every fold that is left out, the associated MSE is calculated. The average of the MSE is calculated as

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\text{MSE}_i$$

(James et al., 2013).

2.6.2 Bootstrapping

A resampling method that can be applied in order to make inference about an estimate for a parameter, θ, based on sample data, is bootstrapping. The idea behind bootstrapping is to sample, with replacement, from a dataset with n observations. With replacement means that we have the possibility to extract the same value the next time we sample. We sample B times, with replacement, and hence create B bootstrap datasets containing n observations. We evaluate the statistics of θ for each sample. These B bootstrap statistics can be used to create a sampling distribution, which can be used to do inference (James et al., 2013).

2.7 Tree models

2.7.1 Regression tree

The idea behind building regression trees, is to divide the possible values for the predictors $X_1, X_2, ..., X_p$ into J regions, that have rectangular shapes, $R_1, R_2, ..., R_J$. When the dataset is partitioned into a training set and test set, we find the mean of the response values for the training observations that fall into the region R_J. In order to construct these regions we minimize the residual sum of squares (RSS), where RSS = $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$, $\hat{y}_{R_j}$ is the mean response within the J-th box from the training dataset. To do this, we begin at the top of the tree and split the predictor space into two branches further down the tree. The way this is done, is that we find values for j and s such that $R_1(j, s) = X|X_j < s$ and $R_2(j, s) = X|X_j \geq s$ will minimize the RSS from above. This process is continued by splitting each of the two previous branches, in a way that minimizes the RSS. This process is known as *recursive binary splitting*. This process continues until we reach a stopping criterion, which we can define as the minimum number of observations within one region. For the test observations, we can predict the response by using the mean of the region estimated from the training set, to which the given test observation belong (James et al., 2013).

2.7.2 Pruning

A typical problem with regression trees is overfitting, where a big and complex tree is chosen, that does not perform well for the test data and might lead to high variance. A smaller tree, with fewer splits, is easier to interpret and might lead to lower variance. The way we choose the size of the tree is to start with a large tree T_0, and then find a subtree $T \subset T_0$ in order to minimize the equation

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 + \alpha |T|, \tag{2.68}$$

where $|T|$ is the number of terminal nodes and α is a penalty term for having a large tree with many terminal nodes. The way we can choose this subtree is to use *recursive binary*

splitting as explained above and apply pruning that create subtrees, as a function of α. In order to choose α we can use K-fold cross-validation. The idea is then to split the data randomly into K parts of equal size. We leave out part k, and *recursive binary splitting* and pruning is used to create subtrees as a function of α to the $K-1$ parts , and predict for part k. This is done for each part $k = 1, 2, ..., K$. Then the results are averaged and the α that minimizes the average error is picked, and the corresponding tree to this α is chosen (James et al., 2013).

2.7.3 Bagging and out-of-bag error

Bagging is a way to reduce the variance produced by a statistical learning method. Bagging is used for regression trees by considering B bootstrapped training sets. From these B bootstrapped training datasets, we create B regression trees, from which we average the predictions in order to reduce the variance.

It can be difficult to interpret results from a regression tree where bagging has been performed. A way to measure the prediction error is to use what is called out-of-bag error (OOB). In fact, two thirds of the observations are used to create each of the bagged trees. Consider one observation i, that is not used to create the bagged tree. For this observation, we can predict a response value based on all the $1/3$ bagged trees where this observation was excluded from. Finally, we can average over these. We can repeat this process for the n observations we have, and then calculate the overall out-of-bag MSE. (James et al., 2013).

2.7.4 Random forests

Assume we have m number of predictors. Random forests build regression trees similar to the bagging method, but for each split in the tree, a random sample is chosen from the m predictors as split candidates. In classification problems, $m \approx \sqrt{p}$ is chosen, which means that the square root of the number of predictors are considered at each split in the tree. In regression problems, $m = \frac{p}{3}$ is chosen. If $m = p$ this method is similar to bagging (James et al., 2013).

2.8 Non-linear estimation

2.8.1 Polynomial regression with least square estimation

It is possible to model the relationship between the response variable and the predictors as an n-th degree polynomial in order to fit a nonlinear model to the data. The polynomial regression model is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i, \tag{2.69}$$

where d is the degree of the polynomial and ϵ is the random error term. Polynomial regression is a special case of multiple linear regression, because y is linear in the unknown parameters $\beta_1, \beta_2, ..., \beta_d$ that are estimated from the data with predictors $x_i, x_i^2, ..., x_i^d$.

Thus, we can estimate the unknown parameters in equation (2.69) by the least squares method. We can translate the model from equation (2.69) to matrix notation, by writing

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^d \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \tag{2.70}$$

which can be written in the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.71}$$

In order to derive the least square estimator, we want to minimize the residual sum of squares, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. We set $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Then $\boldsymbol{\epsilon} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ are orthogonal to the columns of $\boldsymbol{X}$, and hence

$$\begin{aligned} \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &= 0 \\ \Rightarrow \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} &= 0 \\ \Rightarrow \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{y} \end{aligned} \tag{2.72}$$

If we solve the previous equation (2.72) for $\hat{\boldsymbol{\beta}}$, we find the least square estimator to be

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2.73}$$

We can arrive at the same expression for $\hat{\boldsymbol{\beta}}$ by taking the first derivatives of the sum of squares, $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$, with respect to $\boldsymbol{\beta}$ and setting this equal to zero. This gives

$$\frac{\partial((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}} = 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}^T\boldsymbol{y} \tag{2.74}$$

Setting equation (2.74) equal to zero gives $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, as in equation (2.73) (James et al., 2013).

2.8.2 Piecewise, continuous polynomials

In the previous section, we explained polynomial regression and how this is used to fit the data when the relationship between the predictors and the responses is non-linear. It is also possible to fit different polynomials for different regions for the predictor. The points where the polynomials change shape are called knots. For example, we can consider a piecewise cubic polynomial with two knots, one at the point c_1 and one at the point c_2. Then, the polynomial take the following form,

$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if} \quad x_i < c_1 \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if} \quad c_1 < x_i < c_2 \\ \beta_{03} + \beta_{13}x_i + \beta_{23}x_i^2 + \beta_{33}x_i^3 + \epsilon_i & \text{if} \quad x_i > c_2 \end{cases}$$

For C number of knots, we get $C + 1$ different polynomials. We can use least square

estimation for each of the polynomials to estimate the coefficients. In order to make the piece-wise cubic polynomial above smoother, it is possible to add constraints that both the first and second derivatives are continuous at the knots. In general we define a degree-d spline as a piece-wise degree-d polynomial where, for every knot, there are continuity in derivatives up to degree $d-1$. In order to choose the number of knots, it is possible to use leave-one-out cross-validation and use the value of C resulting in the smallest RSS (James et al., 2013).

2.8.3 Smoothing splines

Smoothing splines is another approach to find a smooth function that fits the data well, hence reduce RSS=$\sum_{i=1}^{n}(y_i - g(x_i))^2$, while not overfitting the data. Overfitting means that the function fits too well to the data, and if we were given a new set of data, the same function would possibly give a bad fit. To manage to create a smooth function g that is not overfitted, one approach is to find the function g in order to minimize the penalized residual sum of squares

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx, \tag{2.75}$$

where integration is over the range of x and $\lambda \geq 0$ is a smoothing parameter. The larger λ is, the smoother g will be. The first term in equation (2.75) is the residual sum of squares, and we want this term to be as small as possible in order to give a good fit to the data, but we must also have this second term in (2.75) that ensures that we are not overfitting the data, and this can be interpreted as a *penalty term*. We have that $\int g''(x)^2 dx$ gets big if the variation in g is big. If g is smooth, $\int g''(x)^2 dx$ is smaller. Thus, the term $\lambda \int g''(x)^2 dx$ wants g to be smooth and not to overfit. In order to choose λ it is possible to use cross validation (James et al., 2013).

2.8.4 Generalized additive models

A generalized additive model (GAM) can include non-linear functions for multiple predictors. We can write the model as

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned} \tag{2.76}$$

(James et al., 2013). To fit the functions $f_1(x_{i1}), f_2(x_{i2}), \cdots, f_p(x_{ip})$ we can use smoothing splines. We then use an extension of the method explained in the previous section, where we want to minimize the expression for penalized residual sum of squares (PRSS) given by

$$\text{PRSS}(\beta_0, f_1, f_2, \cdots, f_p) = \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} f_j(x_{ij})\Big)^2 + \sum_{i=1}^{p} \lambda_j \int f_j''(x_j)^2 dx_j, \tag{2.77}$$

where integration is over the region of x_j and λ_j are smoothing parameters. It can be shown that if each of the functions $f_j, \quad j = 1, ..., p$ are cubic splines, then this additive cubic spline will minimize equation (2.77) (Hastie et al., 2017).

Chapter 3

Datasets and variables

3.1 Data from Barentswatch

From Barentswatch, the dataset ”lice per salmon” was downloaded from the period 2017-2019, which is denoted as Dataset 1. Dataset 1 includes the number of sessile lice, mobile lice and adult female lice, all given per salmon, at all the salmon farms across the 13 production zones in Norway. At each farm, a number of salmon per cage are counted, averaged for all cages per week. Corresponding week number and year are included in Dataset 1. The measured sea temperatures, as well as the latitude and longitude coordinates from all the farms could also be found in Dataset 1. In addition, all the salmon farms are registered with an individual farm number in the dataset. In this study, data from production zones six and seven, which include parts of Møre og Romsdal, Trøndelag and parts of Nordland, were studied. Thus, Dataset 1 only includes observations from these two production zones from the period 2017–2019.

The dataset ”lice treatments” was downloaded from Barentswatch. It contains information about deployment of cleaner fish, as well as the use of mechanical delousing and medicinal treatment at every farm, with corresponding week number and year, and is denoted as Dataset 2. The use of cleaner fish, medicinal treatment and mechanical delousing were coded as factor variables in the analysis, in the following way and merged with Dataset 1:

$$\text{Cleaner fish} = \begin{cases} 0, & \text{if cleaner fish were not deployed in the given quarter} \\ 1, & \text{if cleaner fish were deployed in the given quarter} \end{cases}$$

$$\text{Medicinal} = \begin{cases} 0, & \text{if medicinal treatments were not used in the given quarter} \\ 1, & \text{if medicinal treatments were used in the given quarter} \end{cases}$$

$$\text{Mechanical} = \begin{cases} 0, & \text{if mechanical delousing was not used in the given quarter} \\ 1, & \text{if mechanical delousing was used in the given quarter} \end{cases}$$

It is the deployment of new cleaner fish which is used in the model, and a 0 value can

therefore mean that cleaner fish were deployed at an earlier time. These variables were coded in the same way for the 12 quarters during the three years 2017, 2018 and 2019.

3.2 Data from Norwegian Directorate of Fisheries

Each salmon farm has their own maximum biomass capacity, which is a limit of the allowed biomass of living salmon at any given time at the farm, and cleaner fish are excluded from this. The biomass capacity is measured in kilograms. The maximum biomass capacity per farm was downloaded from The Norwegian Directorate of Fisheries, denoted as Dataset 3. To estimate the number of lice, the assumption that 90% of the maximum biomass capacity is utilized at any given time was used. In addition, the mean weight per salmon was assumed to be 5 kg. The number of salmon was estimated by multiplying the maximum biomass capacity by 0.9 and divide by 5. To get an estimate of the number of lice for the three different categories, the number of salmon was multiplied by the number of lice per salmon from Dataset 1. Dataset 1 and Dataset 3 were merged in order to include the estimation of the number of lice for each farm in Dataset 1, per category.

3.3 Distance from the salmon farms to the coastline

The smallest distance from each of the localities to the coastline was parameterized by using the function getbb() in the R package *osmdata* as the search engine for OpenStreetMap coastline data for Møre og Romsdal, Trøndelag and Nordland (Padgham et al., 2017) that includes fjords and islands. OpenStreetMap is a collaborative project in order to create free wiki world map. From the *Geosphere* R package, the function dist2Line() was used to measure the distance from each salmon farm to the coastline (Hijmans, 2019), and the distances were stored in a dataset, which is denoted as Dataset 4. Dataset 4 was merged with Dataset 1.

3.4 Shortest distance from one salmon farm to another

All the unique rows in terms of location number, latitude and longitude coordinates were extracted and stored in a dataset, denoted as Dataset 5. With use of the functions distm() and distHaversine() from the *Geosphere* package, the shortest distance between each farm and all the others was estimated (Hijmans, 2019). When calculating the distances, all ellipsoidal effects were ignored, and a spherical earth was assumed. These values were stored in a matrix and the shortest distance from one farm to another was extracted and stored in the variable ”Shortest distance” in Dataset 5. Finally, Dataset 5 was merged with Dataset 1.

3.5 Wind data from eKlima

Wind data from eKlima were downloaded. A daily measurement of FFM, FFN and FFX, was included in the dataset, denoted as Dataset 6. FFM is the mean of the wind speed given in m/s, FFN is the lowest measured wind speed given in m/s and FFX is the highest measured wind speed given in m/s (see Table 3.1).
The weather stations that had incomplete data within the period from 2017–2019 in terms of measurements from the variables above, were filtered out. The latitude and longitude coordinates from the different weather stations were given Dataset 6. It was possible to find out which weather station was closest to each of the localities by using the distHaversine() function for each farm coordinate in Dataset 1. The daily data was averaged to give weekly measurements by using the timeAverage() function from the R package *openair* (Carslaw and Ropkins, 2012). Finally, Dataset 6 was merged with Dataset 1 giving wind measurements at each farm equal to the measurements at the closest weather station.

3.6 Full dataset

The dataset used for analysis in this thesis were the now combined, Dataset 1, which consists of salmon lice count data from production zones 6 and 7, during 2017–2019. The variables in the dataset, are shown in Table 3.1. Each data point is the total estimated salmon louse numbers at a given salmon farm each week each year. The dataset ”lice per salmon” contained rows where localities had not reported salmon lice numbers, sample points from Dataset 1 corresponding to these missing data were removed from Dataset 1.

Table 3.1: Variables used in the analysis with explanation of the variables.

Variables used in the analysis	
Variable Name	**Explanation of variable**
Adult female lice	Estimation of average weekly number of adult female lice at the farm
Sessile lice	Estimation of average weekly number of sessile lice at the farm
Mobile lice	Estimation of average weekly number of mobile lice at the farm
Year	Year of the observation, 2017,2018 or 2019
Week number	Week number from 1 to 52 during the three years
Sea temperature	The weekly sea temperature reported in the different localities, measured in °C
FFM	Weekly mean of the wind speed, measured in m/s.
FFN	The lowest measured weekly wind speed, measured in m/s
FFX	The highest measured weekly wind speed, measured in m/s
Cleaner fish	Use of cleaner fish, coded 0 or 1
Medicinal	Use of medicinal treatments, coded 0 or 1
Mechanical	Use of mechanical treatments, coded 0 or 1
Distance from coastline	Estimated smallest distance from coastline to each farm measured in m
Shortest distance	Shortest distance from one salmon farm to another salmon farm measured in m

Chapter 4

Visualization of the data

This chapter gives a visualization of the data, where Mobile lice, Adult female lice and Sessile lice are plotted against the explanatory variables.

Figure 4.1 shows Mobile lice, Adult female lice and Sessile lice plotted against Sea temperature, respectively. In the interval from $10°$C to $15°$C, the number of lice seems to increase with increasing sea temperatures.

Figure 4.2 shows Mobile lice, Adult female lice and Sessile lice plotted against FFM, FFN and FFX, respectively. The number of mobile lice seems to increase with increasing wind speed measurements, up to FFM values of about 5 m/s, FFN values around 2 m/s and FFX around 8 m/s, respectively. For wind measurements above these values, the number of mobile lice seems to decrease with increasing wind values.

Figure 4.3 shows Mobile lice, Adult female lice and Sessile lice plotted against the Shortest distance form one farm to another. It seems to be more lice when the farms are closer together.

Figure 4.4 shows Mobile lice, Adult female lice and Sessile lice plotted against Week number. From the plots it seems that louse numbers increase for week number 25–40.

Figure 4.5 shows Mobile lice, Adult female lice and Sessile lice plotted against Distance from coastline. It seems to be more lice on localities closer to the coastline.

Figure 4.6 shows Mobile lice, Adult female lice and Sessile lice plotted against Cleaner fish, Medicinal and Mechanical. It seems to be more lice when cleaner fish and medicinal treatments were not used.

The correlation plot in 4.7 shows that there is high correlation between some variables. Between Mobile lice and Adult female lice the Pearson correlation coefficient was 0.671. The wind variables were highly correlated with $\text{Corr}(\text{FFM}, \text{FFX}) = 0.968$ and $\text{Corr}(\text{FFN}, \text{FFM}) = 0.936$ and $\text{Corr}(\text{FFX}, \text{FFN}) = 0.85$. The Pearson correlation coefficient was calculated to be 0.524 between Week number and Sea temperature. All correlations had p-values $< 2.2 \cdot 10^{-16}$.

Figure 4.8 shows histograms for Mobile lice and a log-transformation of Mobile lice. The natural logarithm was utilized in the transformation.

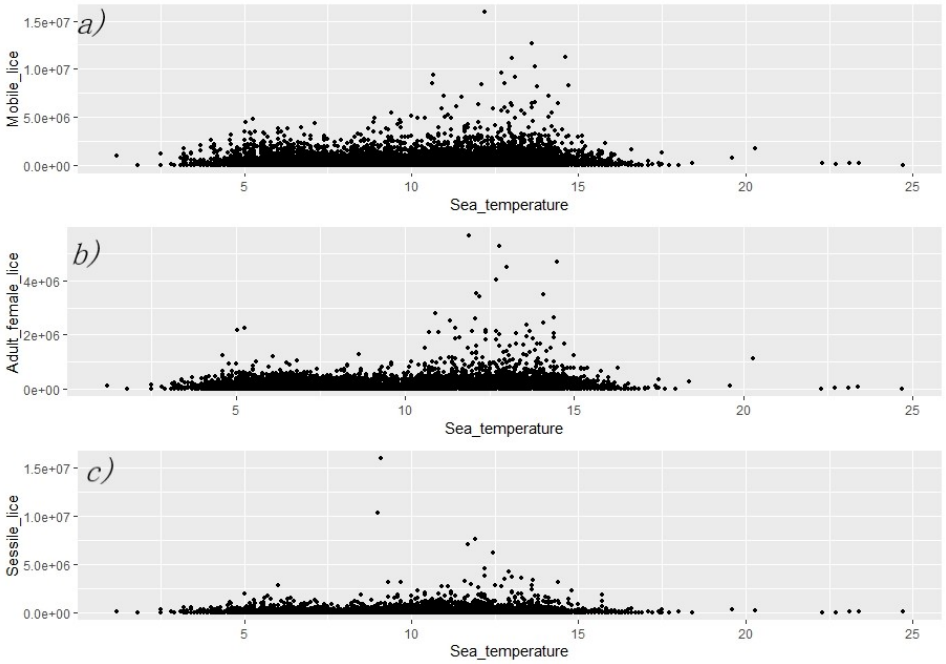


Figure 4.1: Lice *versus* Sea temperature. a) Mobile lice *versus* Sea temperature, b) Adult female lice *versus* Sea temperature and c) Sessile lice *versus* Sea temperature

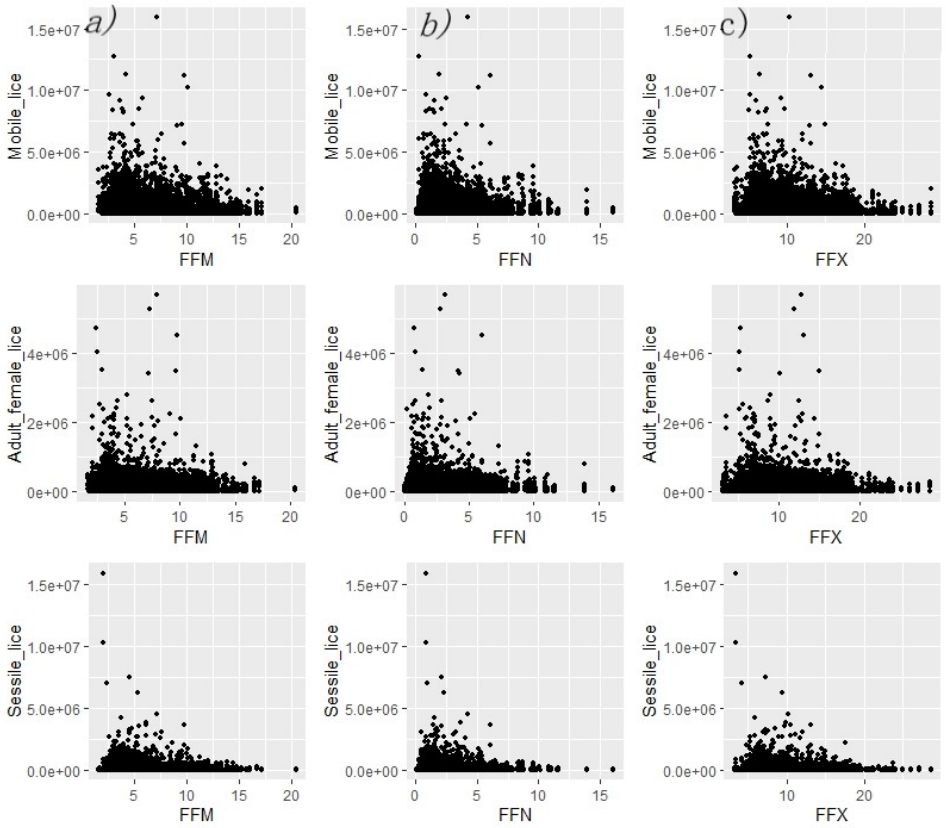


Figure 4.2: Lice *versus* wind. a) (first column) Mobile lice, Adult female lice and Sessile lice plotted against FFM, respectively, b) (second column) Mobile lice, Adult female lice and Sessile lice plotted against FFN, respectively, c) (third column), Mobile lice, Adult female lice and Sessile lice plotted against FFX, respectively.

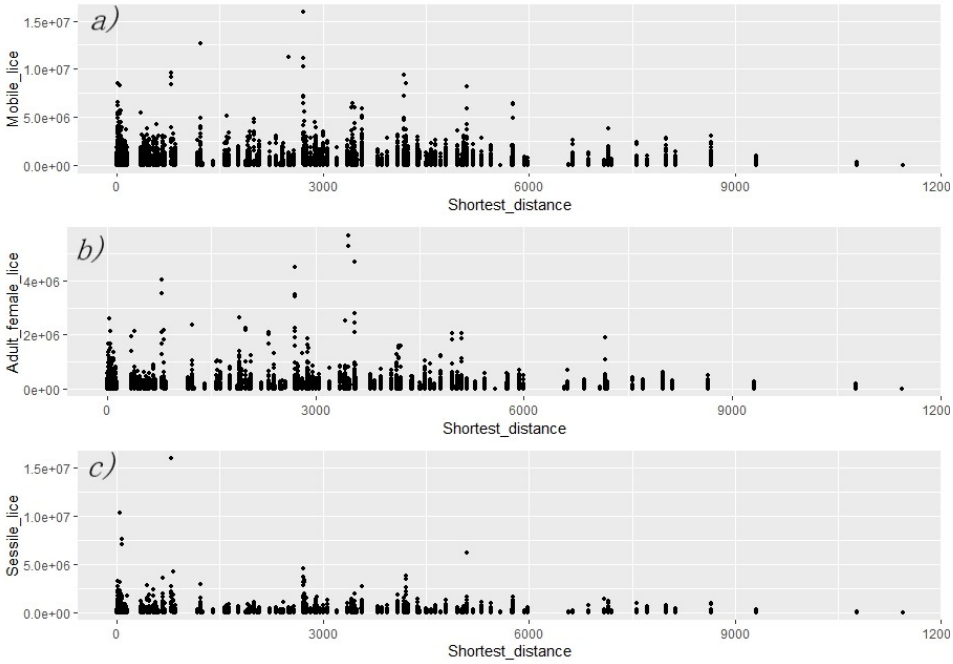


Figure 4.3: Lice *versus* Shortest distance. a) Mobile lice *versus* Shortest distance, b) Adult female lice *versus* Shortest distance, c) Sessile lice *versus* Shortest distance.

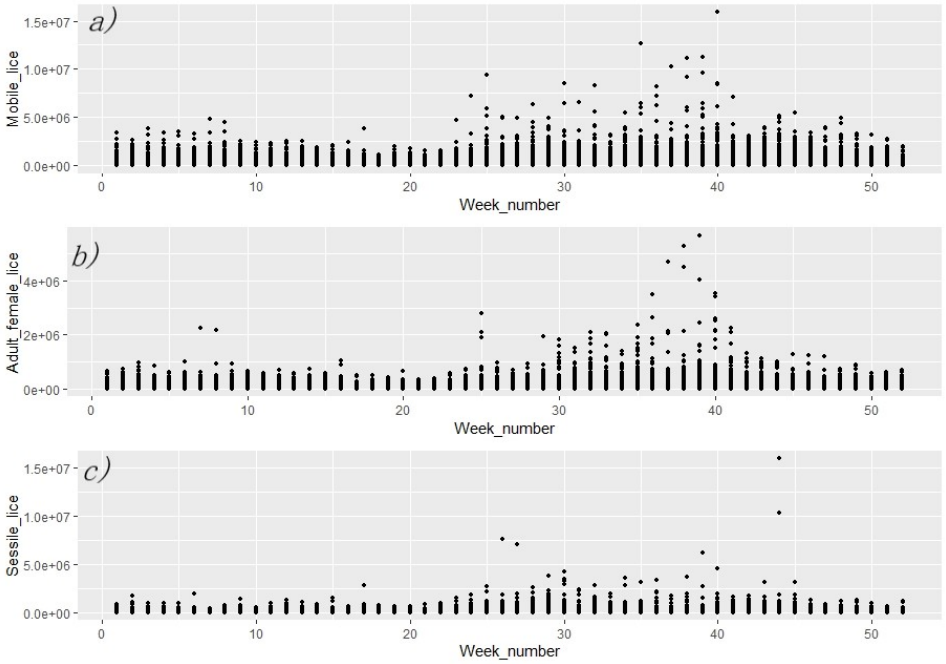


Figure 4.4: Lice *versus* Week number, that is, week number from 1 to 52. a) Mobile lice *versus* Week number, b) Adult female lice *versus* Week number, c) Sessile lice *versus* Week number.

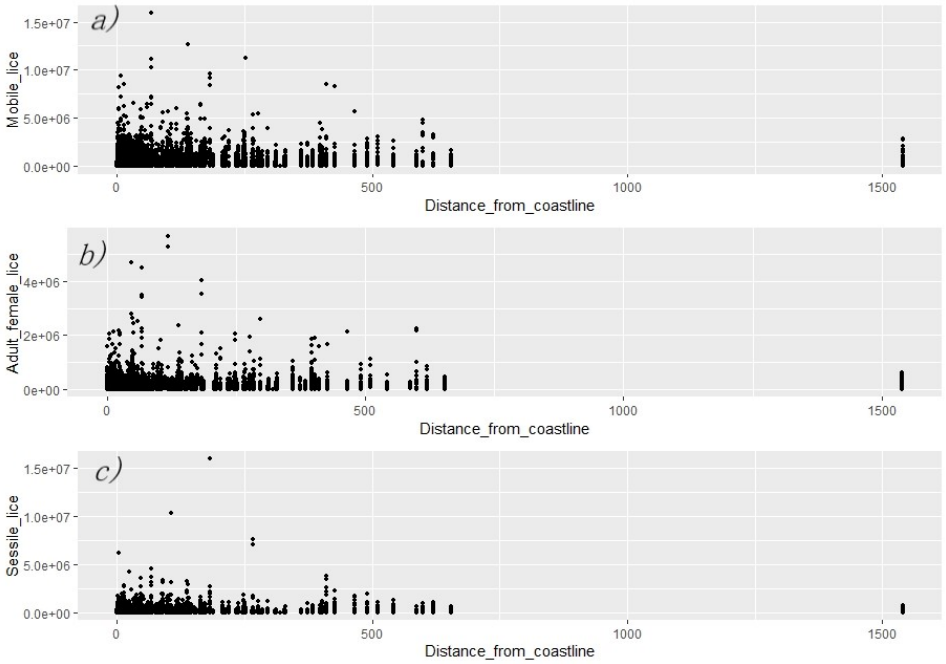


Figure 4.5: Lice *versus* the estimated distance form each salmon farm to the coastline. a) Mobile lice *versus* Distance from coastline, b) Adult female lice *versus* Distance from coastline and c) Sessile lice *versus* Distance from coastline.

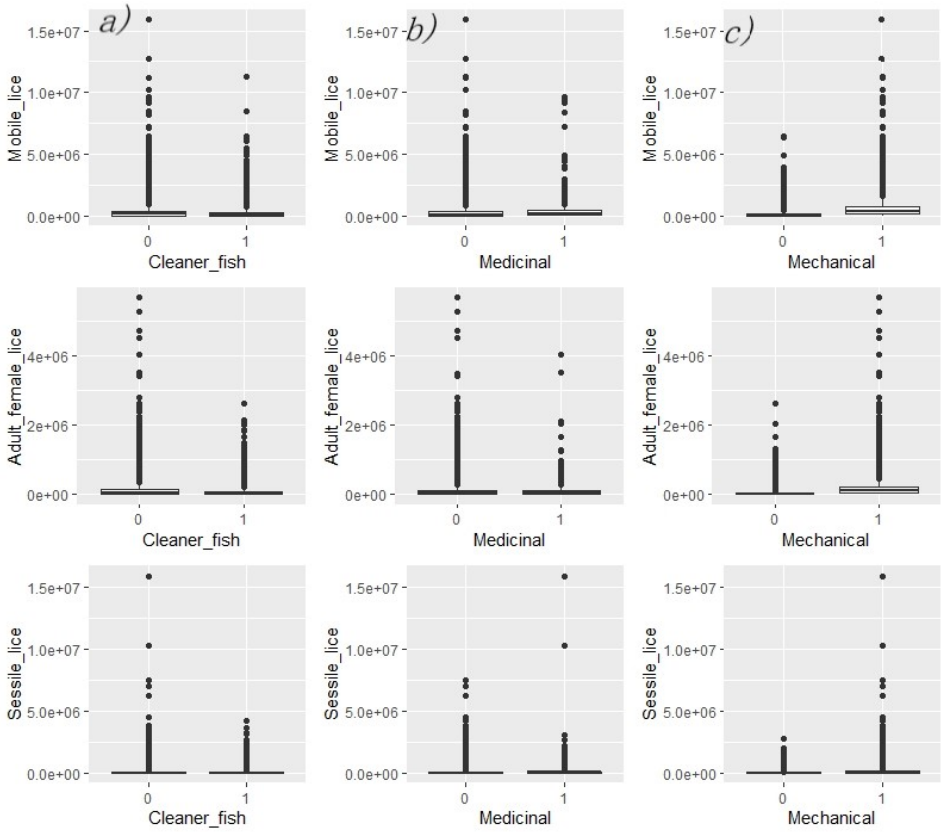


Figure 4.6: Lice *versus* the use of cleaner fish, medicinal treatment and mechanical delousing. a) (first column) Mobile lice, Adult female lice and Sessile lice plotted against Cleaner fish, respectively, b) (second column) Mobile lice, Adult female lice and Sessile lice plotted against Medicinal, respectively, c) (third column) Mobile lice, Adult female lice and Sessile lice *versus* Mechanical, respectively.

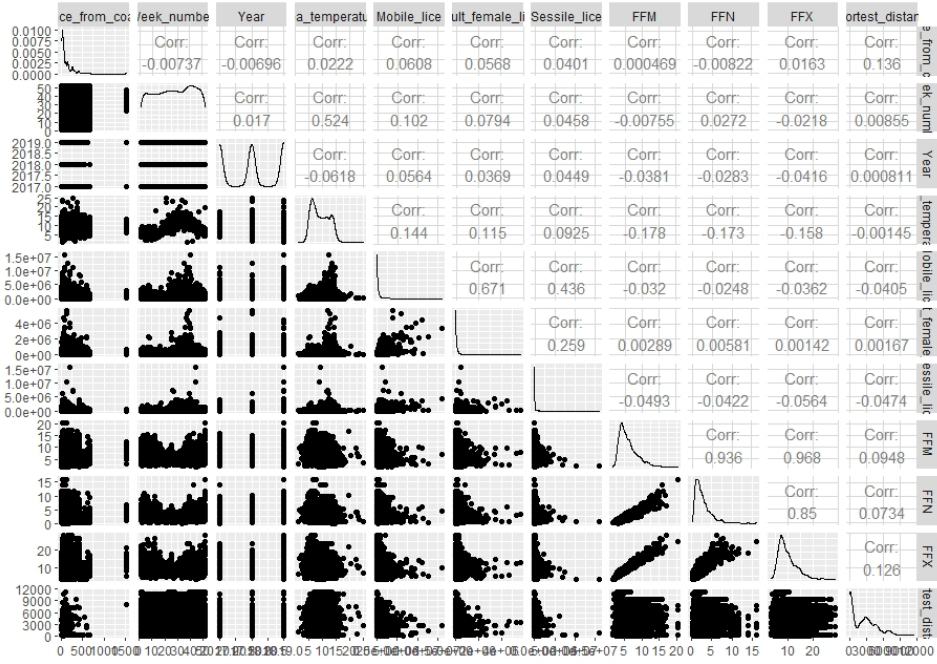


Figure 4.7: Pearson correlation coefficient between variables and distribution plot of variables.

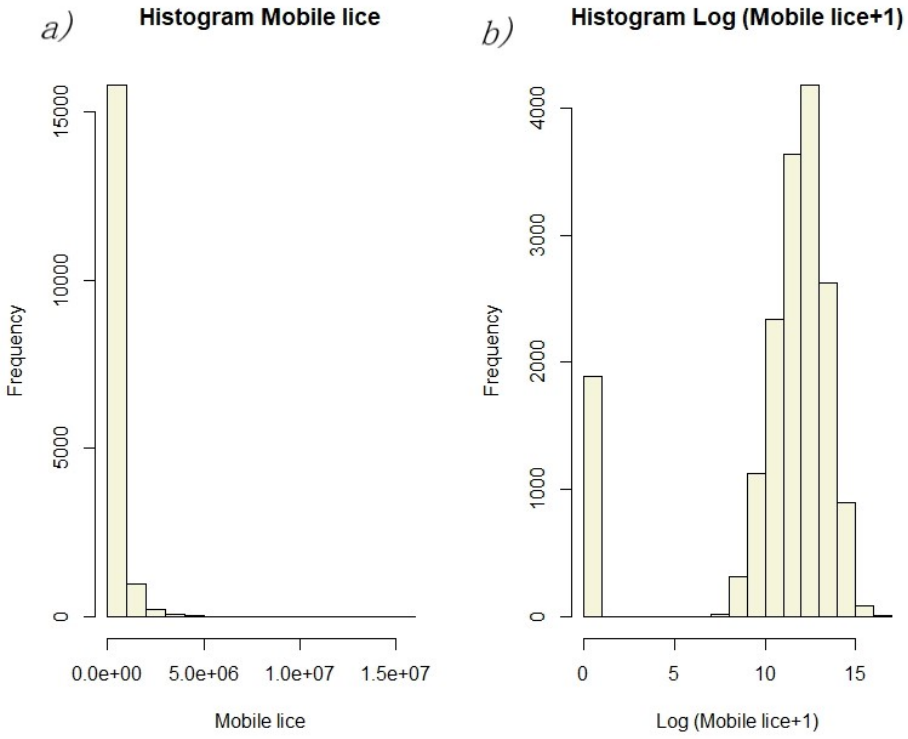


Figure 4.8: a) Histogram of Mobile lice, with a corresponding skewness of 6.60 b) Histogram of the log-transform of Mobile lice+1 with a skewness of -1.97

Chapter 5

Analysis and validation

5.1 Specifying the full model

Mobile lice was used as the response variable, in the fitting of the different regression models. The full model had the explanatory variables Sea temperature, FFX, Distance from coastline, Shortest distance, Mechanical, Medicinal and Cleaner fish, with pairwise interaction terms among the last three variables (three in total). The full model was used to compare different regression models. Nested models were checked if they performed better by backward elimination and forward selection. In the selection, we focused on AIC and BIC-values, and used the stepAIC-function from the *MASS* package in R (Venables and Ripley, 2002). Hypothesis testing for nested models using likelihood ratio test were conducted by the anova-function from the *stats* package (R Core Team, 2019). The *summary*-function in R was used to perform a Wald test. In the Wald test, one term at a time is dropped from the model, and tested against the full model. The lrtest-function from the *lmtest* package was used to conduct the likelihood ratio test. Likelihood ratio test was used to test the Poisson regression model *versus* the negative binomial regression model, and the ZIP regression model *versus* the ZINB regression model (Zeileis and Hothorn, 2002).

5.2 Poisson regression

A Poisson regression model was fitted in R, and the output is given in Table 5.1. In this model, it was assumed that the number of mobile lice at each salmon farm followed a Poisson distribution and independent observations, $y_i \sim \text{Po}(\lambda_i)$, where

$$\begin{aligned}\log \lambda_i = \hat{\beta}_0 &+ \hat{\beta}_1 \cdot \text{Sea temperature} + \hat{\beta}_2 \cdot \text{FFX} + \hat{\beta}_3 \cdot \text{Distance from coastline} \\ &+ \hat{\beta}_4 \cdot \text{Shortest distance} + \hat{\beta}_5 \cdot \text{Mechanical1} \\ &+ \hat{\beta}_6 \cdot \text{Medicinal1} + \hat{\beta}_7 \cdot \text{Cleaner fish1} \\ &+ \hat{\beta}_8 \cdot \text{Mechanical1:Medicinal1} + \hat{\beta}_9 \cdot \text{Mechanical1:Cleaner fish1} \\ &+ \hat{\beta}_{10} \cdot \text{Medicinal1:Cleaner fish1}\end{aligned} \tag{5.1}$$

From Table 5.1, the estimates for both Medicinal1 and Mechanical1 are positive, whereas the estimate for Cleaner fish1 is negative. The estimate for the interaction term Mechanical1:Medicinal1 is negative. For example, if only mechanical delousing was used during the quarter, the expected number of lice increased by a factor of $\exp(1.232) = 3.428$, whereas if both mechanical delousing and medicinal treatments were used, the expected number of lice increased by a smaller factor, namely $\exp(1.232 + 0.5874 - 0.5950) = \exp(1.2244) = 3.4021$ (when other terms were kept constant). From the Wald test, all the terms in Table 5.1 were significant. Backward elimination and forward selection were performed to compare AIC and BIC values for nested models, and the full model, specified in Section 5.1 gave the lowest AIC and BIC values. R^2 was calculated as $R^2 = 1 - \frac{\text{Deviance}}{\text{Null deviance}} = 1 - \frac{7241289472}{9992057431} = 0.275$, which is a quite low number and indicates that a rather low portion of the variation was explained by the model.

In Figures 5.1 and 5.2, the residuals are plotted against the fitted values from the Poisson regression model. Neither the deviance residuals nor the Pearson residuals show a random spread. High variance is observed for low fitted values. The variance then decreases before it increases, which forms a U-shape. This indicates heteroscedasticity in the model. Hence, the Poisson regression model was not a good fit to the data.

A goodness-of-fit test for the model was done by using the residual deviance. The critical value in the χ^2_{17114}-distribution is $z_{0.05,17114} = 17419$. Since $7241289472 > 17419$, the conclusion was that this model was not a good fit.

A hypothesis test for overdispersion was done by the Pearson statistic. The null hypothesis, H_0, of no overdispersion, is written $H_0 : \phi \leq 1$. H_0 was tested against the alternative hypothesis, $H_1 : \phi > 1$, that there is overdispersion. Under the null hypothesis the Pearson statistic, P, is χ^2-distributed with $n - p = 17114$ degrees of freedom. Since $P > \chi^2_{0.05,17114} = 17419$, the null hypothesis was rejected, and a conclusion of overdispersion was drawn.

Table 5.1: Regression coefficients with associated estimate, std. error, z-value and p-value in Poisson regression.

Coefficients	Estimate	Standard error	z-value	p-value
Intercept	$1.173 \cdot 10^{1}$	$7.020 \cdot 10^{-5}$	167031.6	$< 2 \cdot 10^{-16}$
Sea temperature	$5.278 \cdot 10^{-2}$	$4.481 \cdot 10^{-6}$	11780.3	$< 2 \cdot 10^{-16}$
FFX	$-2.976 \cdot 10^{-3}$	$3.934 \cdot 10^{-6}$	-756.4	$< 2 \cdot 10^{-16}$
Distance from coastline	$5.128 \cdot 10^{-4}$	$7.580 \cdot 10^{-6}$	6765.9	$< 2 \cdot 10^{-16}$
Shortest distance	$-4.832 \cdot 10^{-5}$	$6.189 \cdot 10^{-9}$	-7807.0	$< 2 \cdot 10^{-16}$
Mechanical1	1.232	$4.385 \cdot 10^{-5}$	28092.7	$< 2 \cdot 10^{-16}$
Medicinal1	$5.874 \cdot 10^{-1}$	$7.686 \cdot 10^{-5}$	7642.9	$< 2 \cdot 10^{-16}$
Cleaner fish1	$-3.279 \cdot 10^{-1}$	$5.033 \cdot 10^{-5}$	-6515.2	$< 2 \cdot 10^{-16}$
Mechanical1:Medicinal1	$-5.950 \cdot 10^{-1}$	$7.812 \cdot 10^{-5}$	-7617.2	$< 2 \cdot 10^{-16}$
Mechanical1:Cleaner fish1	$1.887 \cdot 10^{-1}$	$5.774 \cdot 10^{-5}$	3268.0	$< 2 \cdot 10^{-16}$
Medicinal1:Cleaner fish1	$-1.321 \cdot 10^{-1}$	$7.572 \cdot 10^{-5}$	-1744.5	$< 2 \cdot 10^{-16}$

The null deviance was 9992057431 on 17124 degrees of freedom. Residual deviance was 7241289472 on 17114 degrees of freedom. Pearson residuals was 12335072703 on 17114 degrees of freedom. The AIC value was 7241499526 and BIC was 7241499611.

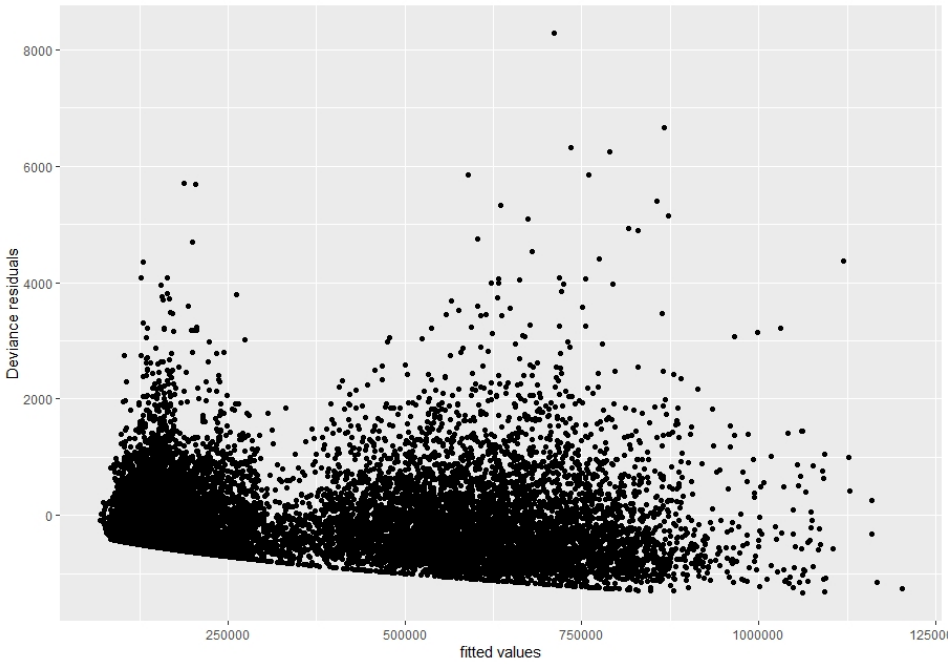


Figure 5.1: Plot of deviance residuals against fitted values for the Poisson regression model.

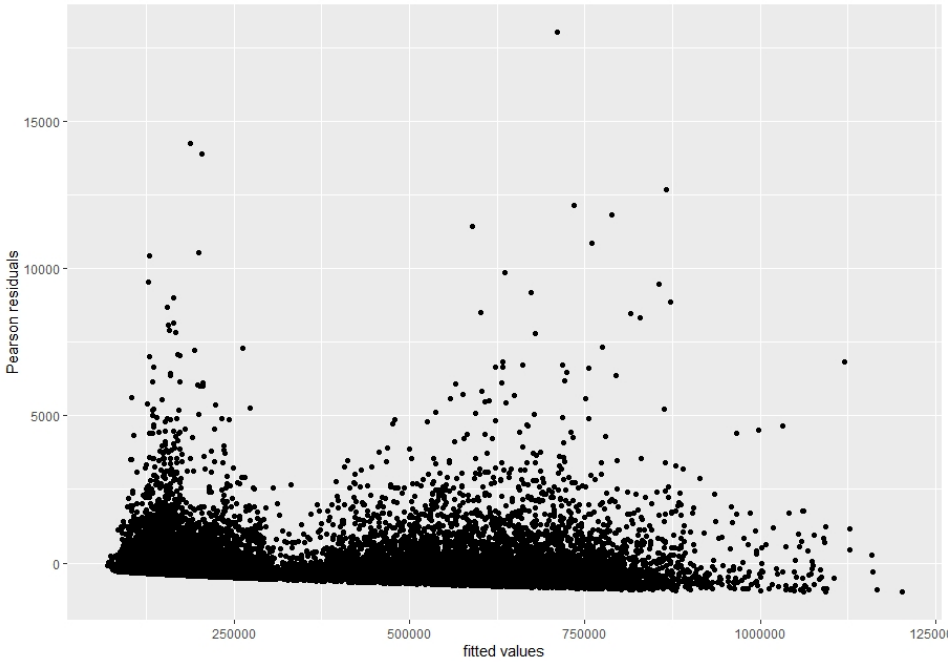


Figure 5.2: Plot of Pearson residuals against fitted values for the Poisson regression model.

5.3 Quasi-Poisson model

The dispersion parameter for the Poisson regression model in Section 5.2 was estimated to be $\hat{\phi} = \frac{P}{n-p} = \frac{12335072703}{17114} = 720759$. The standard error for the estimates in Table 5.2 was obtained by multiplying the standard error in Table 5.1 by a factor of $\sqrt{\hat{\phi}} = \sqrt{720759}$. FFX was not significant in the quasi-Poisson model, while significant in the Poisson regression model. The residual plots for the quasi-Poisson regression model, seen in Figures 5.3 and 5.4, have a similar shape as the residuals in the Poisson regression model (see Figures 5.1 and 5.2).

Table 5.2: Regression coefficients with associated estimate, std. error, t-value and p-value in quasi-Poisson regression.

Coefficients	Estimate	Standard error	t-value	p-value
Intercept	$1.173 \cdot 10^{1}$	$5.960 \cdot 10^{-2}$	196.745	$< 2 \cdot 10^{-16}$
Sea temperature	$5.278 \cdot 10^{-2}$	$3.804 \cdot 10^{-3}$	13.876	$< 2 \cdot 10^{-16}$
FFX	$-2.976 \cdot 10^{-3}$	$3.340 \cdot 10^{-3}$	-0.891	0.372934
Distance from coastline	$5.128 \cdot 10^{-4}$	$6.435 \cdot 10^{-5}$	7.969	$1.69 \cdot 10^{-15}$
Shortest distance	$-4.832 \cdot 10^{-5}$	$5.254 \cdot 10^{-6}$	-9.196	$< 2 \cdot 10^{-16}$
Mechanical1	1.232	$3.723 \cdot 10^{-2}$	33.090	$< 2 \cdot 10^{-16}$
Medicinal1	$5.874 \cdot 10^{-1}$	$6.525 \cdot 10^{-2}$	9.002	$< 2 \cdot 10^{-16}$
Cleaner fish1	$-3.279 \cdot 10^{-1}$	$4.272 \cdot 10^{-2}$	-7.674	$1.75 \cdot 10^{-14}$
Mechanical1:Medicinal1	$-5.950 \cdot 10^{-1}$	$6.632 \cdot 10^{-2}$	-8.972	$< 2 \cdot 10^{-16}$
Mechanical1:Cleaner fish1	$1.887 \cdot 10^{-1}$	$4.902 \cdot 10^{-2}$	3.841	0.000119
Medicinal1:Cleaner fish1	$-1.321 \cdot 10^{-1}$	$6.428 \cdot 10^{-2}$	-2.055	0.039907

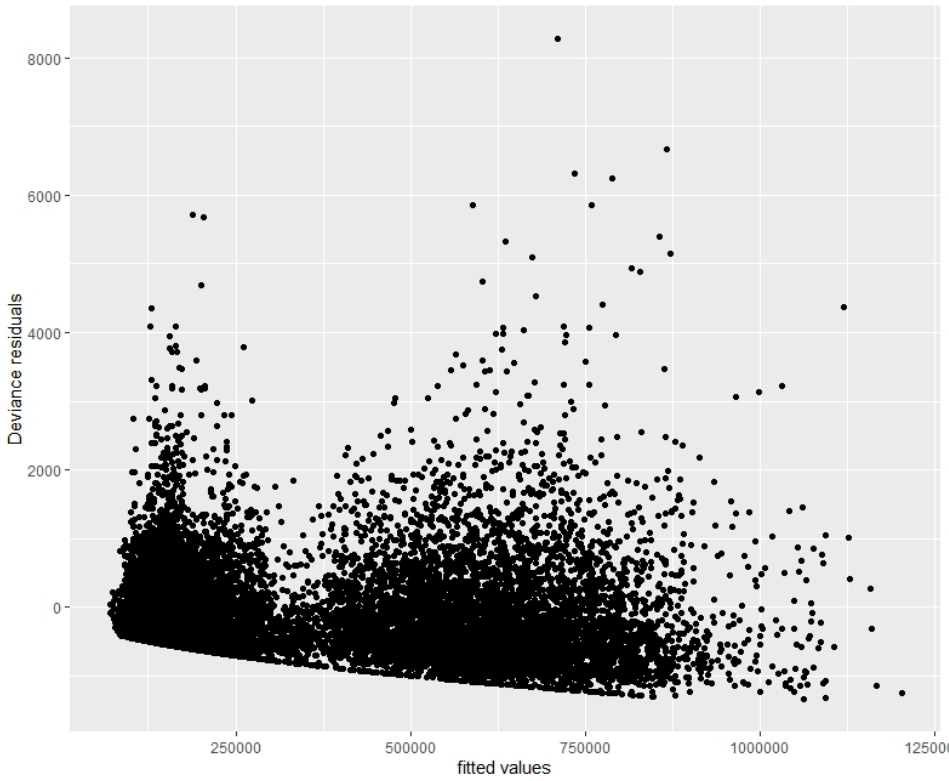


Figure 5.3: Plot of deviance residuals against fitted values for the quasi-Poisson regression model.

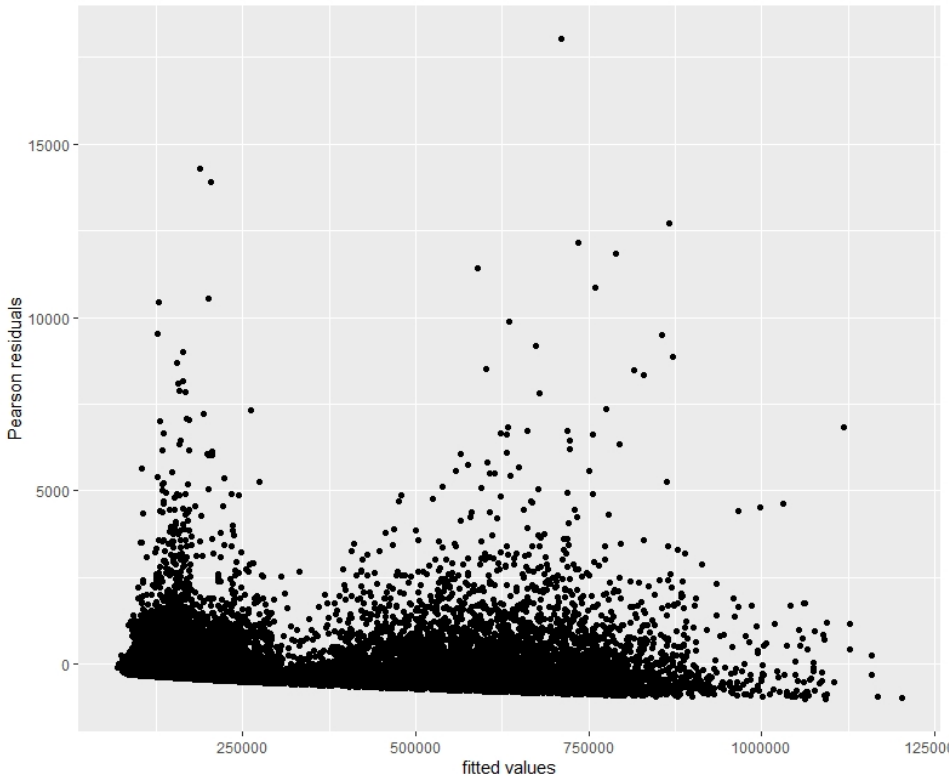


Figure 5.4: Plot of Pearson residuals against fitted values for the quasi-Poisson regression model.

5.4 Negative binomial regression

A negative binomial regression model was fitted to the data, and results are shown in Table 5.3. From the Wald test, the estimated coefficient for FFX and the interaction term Medicinal1:Cleaner fish1 were not significant. After performing backward elimination, excluding FFX and Medicinal1:Cleaner fish1 yielded a slight improvement in terms of AIC and BIC with values 438821 and 438800, respectively. The residual plots with these two terms were excluded, looks similar to the plots in Figures 5.5 and 5.6. From the summary in R, an estimate for r from equation (2.19) was $\hat{r} = 0.31242$ with a standard deviation equal to $SD(\hat{r}) = 0.00291$. The regression equation was

$$\begin{aligned}\log \lambda_i = {} & 1.183 \cdot 10^{1} + 4.118 \cdot 10^{-2} \cdot \text{Sea temperature} - 2.133 \cdot 10^{-3} \cdot \text{FFX} + \\ & 9.080 \cdot 10^{-4} \cdot \text{Distance from coastline} \\ & - 7.082 \cdot 10^{-5} \cdot \text{Shortest distance} + 1.253 \cdot \text{Mechanical1} \\ & + 5.996 \cdot 10^{-1} \cdot \text{Medicinal1} - 3.659 \cdot 10^{-1} \cdot \text{Cleaner fish1} \\ & - 6.589 \cdot 10^{-1} \cdot \text{Mechanical1:Medicinal1} \\ & + 2.286 \cdot 10^{-1} \cdot \text{Mechanical1:Cleaner fish1} \\ & - 2.653 \cdot 10^{-2} \cdot \text{Medicinal1:Cleaner fish1}\end{aligned} \tag{5.2}$$

Even though a lot of the points seem to be randomly spread out in Figure 5.5, there are outliers for deviance residuals below -2. Figure 5.6 shows a similar shape for the Pearson residuals against fitted values, as for the Poisson regression model (see Figure 5.2). Both residual plots indicate that the model was not a good fit to the data. R^2 was calculated to be $R^2 = 1 - \frac{\text{Deviance}}{\text{Null deviance}} = 1 - \frac{22052}{24840} = 0.112$, which is a low number and very little of the variation in the data was explained by the model.

For the Poisson regression model, the parameter vector is $\boldsymbol{\theta_0} = \boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{10})^T$ and the parameter vector for the negative binomial regression model is $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_{10}, r)^T$. Then $\boldsymbol{\theta_0} \in \boldsymbol{\theta}$, and the Poisson regression model is nested within the negative binomial regression model. A likelihood ratio test was used to compare the two models. The null hypothesis H_0: $r = 0$ was tested against the alternative hypothesis, H_1, that $r > 0$. As the parameter space for r is $r \in (0, \infty)$ the null hypothesis is on the boundary of the parameter space. It can be shown that the likelihood ratio statistic can be written as a mixture of half a probability mass at zero and half of χ^2_1 (Lawless, 1987). The test results are given in Table 5.4. Although the negative binomial regression model explained less of the variation in the data than the Poisson regression model, the p-value $< 2.2 \cdot 10^{-16}$ from the likelihood ratio test strongly suggested that the negative binomial regression model, estimating the dispersion parameter, was more appropriate than the Poisson regression model.

Table 5.3: Regression coefficients with associated estimate, std. error, z-value and p-value in negative binomial regression.

Coefficients	Estimate	Standard error	z-value	p-value
Intercept	$1.183 \cdot 10^{1}$	$6.633 \cdot 10^{-2}$	178.342	$< 2 \cdot 10^{-16}$
Sea temperature	$4.118 \cdot 10^{-2}$	$4.651 \cdot 10^{-3}$	8.855	$< 2 \cdot 10^{-16}$
FFX	$-2.133 \cdot 10^{-3}$	$3.950 \cdot 10^{-3}$	-0.540	0.589
Distance from coastline	$9.080 \cdot 10^{-4}$	$9.416 \cdot 10^{-5}$	9.644	$< 2 \cdot 10^{-16}$
Shortest distance	$-7.082 \cdot 10^{-5}$	$6.175 \cdot 10^{-6}$	-11.468	$< 2 \cdot 10^{-16}$
Mechanical1	1.253	$4.324 \cdot 10^{-2}$	28.981	$< 2 \cdot 10^{-16}$
Medicinal1	$5.996 \cdot 10^{-1}$	$8.226 \cdot 10^{-2}$	7.289	$3.12 \cdot 10^{-13}$
Cleaner fish1	$-3.659 \cdot 10^{-1}$	$3.729 \cdot 10^{-2}$	-9.813	$< 2 \cdot 10^{-16}$
Mechanical1:Medicinal1	$-6.589 \cdot 10^{-1}$	$8.632 \cdot 10^{-2}$	-7.634	$2.28 \cdot 10^{-14}$
Mechanical1:Cleaner fish1	$2.286 \cdot 10^{-1}$	$5.755 \cdot 10^{-2}$	3.972	$7.12 \cdot 10^{-5}$
Medicinal1:Cleaner fish1	$-2.653 \cdot 10^{-2}$	$8.737 \cdot 10^{-2}$	-0.304	0.761

The null deviance was 24840 on 17124 degrees of freedom. Residual deviance was 22052 on 17114 degrees of freedom. Pearson residuals was 14386 on 17114 degrees of freedom. AIC and BIC values corresponding to this model were 438826 and 438919, respectively. The coefficients have the same sign as for the Poisson regression model, although some coefficients differ quite a bit. Distance from coastline is almost twice the coefficient from Table 5.1, and Medicinal1:Cleaner fish1 differs a lot.

Table 5.4: Result from the likelihood ratio test between the Poisson regression model and the negative binomial regression model.

number of Df	LogLik	Df	Chisq	p-value
11	3620749752			
12	-219401	1	7241060701	$< 2.2 \cdot 10^{-16}$

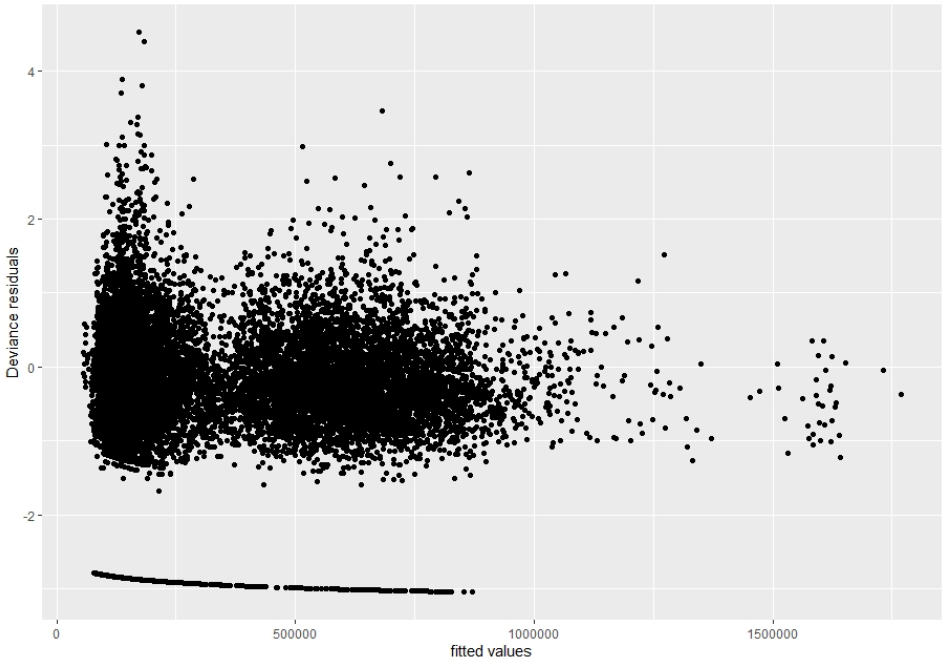


Figure 5.5: Plot of deviance residuals against fitted values for the negative binomial regression model.

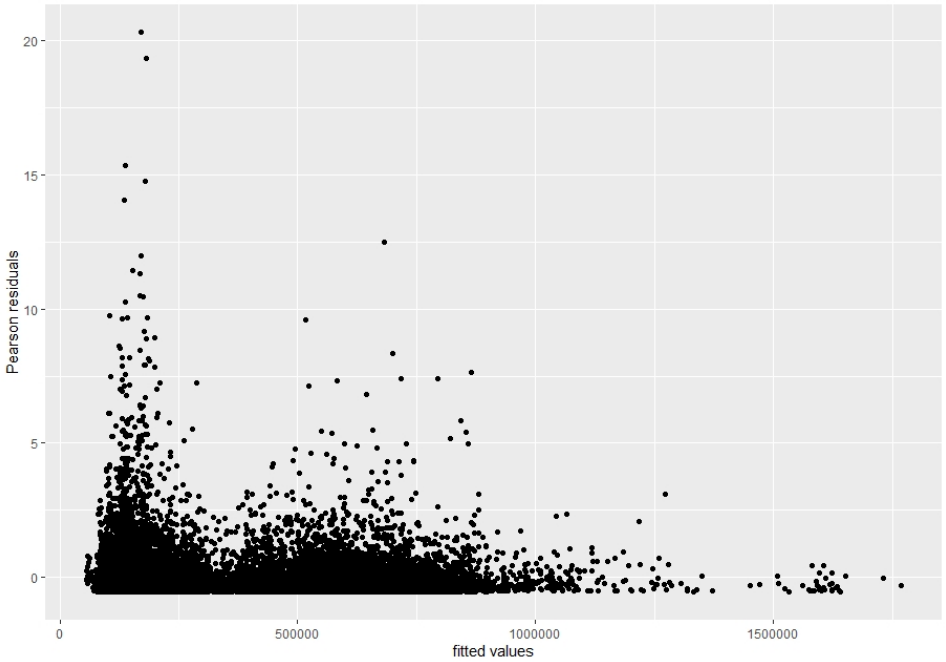


Figure 5.6: Plot of Pearson residuals against fitted values for the negative binomial regression model.

5.5 Zero-inflated models

In Dataset 1, 11% of the datapoints were zeros, which is more than predicted by the Poisson regression model and the negative binomial model. To check if a zero inflated model (ZI) fit the data better, both the Poisson regression model and the negative binomial regression model were compared with their zero inflated counterparts (ZIP and ZINB). The parameter vector in the ZIP regression model is $\boldsymbol{\theta}_0 = (\beta_0, \beta_1, ..., \beta_{10}, \alpha_0, \alpha_1, ..., \alpha_{10})^T$. In the ZINB regression model, the parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1, ..., \beta_{10}, r, \alpha_0, \alpha_1, ..., \alpha_{10})^T$. Hence, $\boldsymbol{\theta_0} \in \boldsymbol{\theta}$, and the ZIP model is nested within the ZINB model. These two models were compared with a likelihood ratio test. The following null hypothesis $\mathrm{H}_0 : r = 0$ was tested against the alternative hypothesis H_1 that $r > 0$, which is a test on the boundary of the parameter space. Then, the likelihood ratio statistic is a 50% mixture of a probability of zero and a χ^2- distribution with one degree of freedom (Stram and Lee, 1994). The lrtest-function in R was used to run the test, and test results are given in Table 5.6. The null hypothesis was rejected in favour of the alternative hypothesis (p-value $< 2.2 \cdot 10^{-16}$) and the ZINB regression model was preferred.

A ZINB regression model was fitted and the output is shown in Table 5.5. The same covariates were used both for the zero part and the count part, as *a priori* we did not know which component that might be affected by this predictor, and an effect on both parts was assumed. Hence, $\boldsymbol{x}_i = \boldsymbol{z}_i$ in this case. From Table 5.5, the odds for a zero inflation was written as

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) = {} & -1.327 + 1.183 \cdot 10^{-2} \cdot \text{Sea temperature} - 3.788 \cdot 10^{-2} \cdot \text{FFX} - \\ & 5.994 \cdot 10^{-5} \cdot \text{Distance from coastline} \\ & + 7.903 \cdot 10^{-5} \cdot \text{Shortest distance} - 2.105 \cdot \text{Mechanical1} \\ & - 6.195 \cdot 10^{-1} \cdot \text{Medicinal1} + 2.320 \cdot 10^{-2} \cdot \text{Cleaner fish1} \\ & + 6.950 \cdot 10^{-1} \cdot \text{Mechanical1:Medicinal1} \\ & - 8.201 \cdot 10^{-2} \cdot \text{Mechanical1:Cleaner fish1} \\ & - 7.970 \cdot 10^{-1} \cdot \text{Medicinal1:Cleaner fish1}\end{aligned} \tag{5.3}$$

From the model, Sea temperature, Shortest distance, Cleaner fish1 and the interaction term between Mechanical1 and Medicinal1 increased the odds of a zero inflation, and the other terms reduced the odds of a zero inflation. Sea temperature, Distance from coastline, Shortest distance and Cleaner fish1 were not significant, hence did not seem to be important for the zero inflation. Therefore the model was refitted without the non-significant terms, and a similar residual plot to that of the full model was observed.

From the residual plot in Figure 5.7, as for previous fitted models, the points are not randomly spread out, which indicates that the model was not a good fit to the data. The full model gave the lowest AIC when backward elimination was performed.

Table 5.5: Regression coefficients with associated estimate, std. error, z-value and p-value in zero inflated negative binomial regression.

Count model coefficients	Estimate	Standard error	z-value	p-value
Intercept	$1.202 \cdot 10^{1}$	$4.288 \cdot 10^{-2}$	280.428	$< 2 \cdot 10^{-16}$
Sea temperature	$4.347 \cdot 10^{-2}$	$2.957 \cdot 10^{-3}$	14.699	$< 2 \cdot 10^{-16}$
FFX	$-5.506 \cdot 10^{-3}$	$2.456 \cdot 10^{-3}$	-2.242	0.025
Distance from coastline	$8.476 \cdot 10^{-4}$	$6.010 \cdot 10^{-5}$	14.104	$< 2 \cdot 10^{-16}$
Shortest distance	$-6.702 \cdot 10^{-5}$	$3.944 \cdot 10^{-6}$	-16.996	$< 2 \cdot 10^{-16}$
Mechanical1	1.086	$2.740 \cdot 10^{-2}$	39.635	$< 2 \cdot 10^{-16}$
Medicinal1	$5.133 \cdot 10^{-1}$	$5.449 \cdot 10^{-2}$	9.420	$< 2 \cdot 10^{-16}$
Cleaner fish1	$-3.555 \cdot 10^{-1}$	$2.411 \cdot 10^{-2}$	-14.744	$< 2 \cdot 10^{-16}$
Mechanical1:Medicinal1	$-5.532 \cdot 10^{-1}$	$5.459 \cdot 10^{-2}$	-10.134	$< 2 \cdot 10^{-16}$
Mechanical1:Cleaner fish1	$2.222 \cdot 10^{-1}$	$3.597 \cdot 10^{-2}$	6.178	$6.49 \cdot 10^{-10}$
Medicinal1:Cleaner fish1	$-8.487 \cdot 10^{-2}$	$5.580 \cdot 10^{-2}$	-1.521	0.128
Log(r)	$-1.318 \cdot 10^{-1}$	$9.986 \cdot 10^{-3}$	-13.202	$< 2 \cdot 10^{-16}$
Zero-inflation coefficients	Estimate	Standard error	z-value	p-value
Intercept	-1.327	$1.184 \cdot 10^{-1}$	-11.203	$< 2 \cdot 10^{-16}$
Sea temperature	$1.183 \cdot 10^{-2}$	$8.539 \cdot 10^{-3}$	1.386	0.165770
FFX	$-3.788 \cdot 10^{-2}$	$7.609 \cdot 10^{-3}$	-4.979	$6.41 \cdot 10^{-7}$
Distance from coastline	$-5.994 \cdot 10^{-5}$	$1.817 \cdot 10^{-4}$	-0.330	0.741565
Shortest distance	$7.903 \cdot 10^{-5}$	$1.114 \cdot 10^{-4}$	0.709	0.478198
Mechanical1	-2.105	$1.256 \cdot 10^{-1}$	-16.769	$< 2 \cdot 10^{-16}$
Medicinal1	$-6.195 \cdot 10^{-1}$	$1.718 \cdot 10^{-1}$	-3.605	0.000312
Cleaner fish1	$2.320 \cdot 10^{-2}$	$5.727 \cdot 10^{-2}$	0.405	0.685420
Mechanical1:Medicinal1	$6.950 \cdot 10^{-1}$	$2.910 \cdot 10^{-1}$	2.389	0.016906
Mechanical1:Cleaner fish1	$-8.201 \cdot 10^{-2}$	$1.760 \cdot 10^{-1}$	-0.466	0.641328
Medicinal1:Cleaner fish1	$-7.970 \cdot 10^{-1}$	$2.168 \cdot 10^{-1}$	-3.676	0.000237

The AIC value was 425604

Table 5.6: Result from the likelihood ratio test between the ZIP and ZINB regression model

number of Df	LogLik	Df	Chisq	p-value
22	-3226381595			
23	-212779	1	6452337633	$< 2.2 \cdot 10^{-16}$

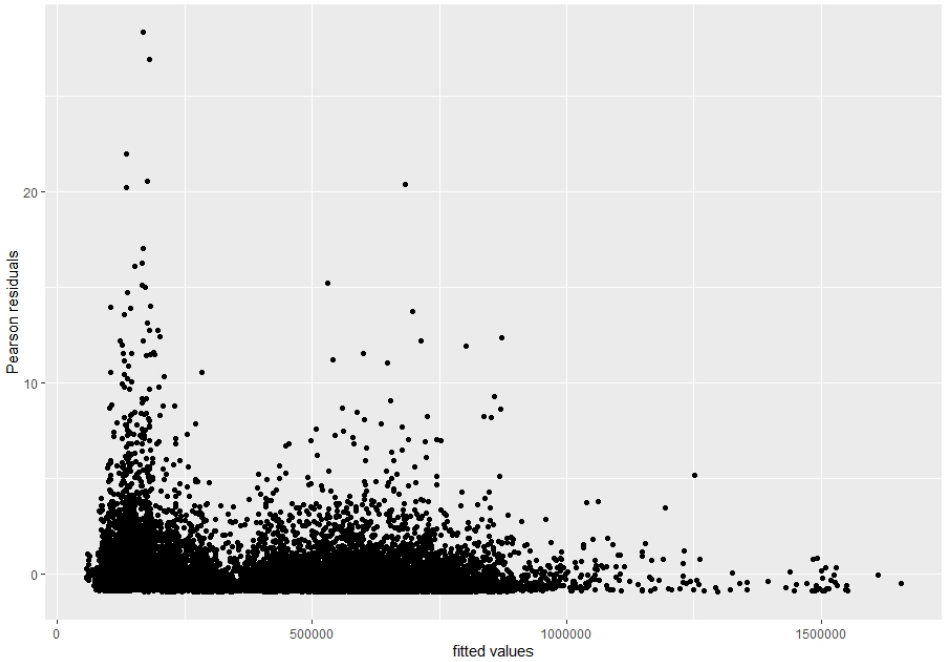


Figure 5.7: Plot of Pearson residuals against fitted values for the zero- inflated negative binomial regression model.

5.6 Regression tree and random forest

In the regression tree, independent observations were assumed. The response variable, Mobile lice, was log-transformed because its distribution was very right-skewed (log+1 - transformation was used because of right skewness and zero-inflation). The transformed variable was left-skewed, mainly due to the zeros (see Figure 4.8). Then, the dataset was divided into a training set and a test set, by dividing the set in two, randomly half by half, before a regression tree was fitted to the training set. The fitted regression tree was pruned in order to minimize equation (2.68). For $\alpha = 0$, no pruning was carried out. For a large enough α, no partition was done and only the root node was predicted by the tree. For the latter case, the value for α was denoted as α_0, and the penalty term was re-scaled as

$$cp = \frac{\alpha}{\alpha_0}, \tag{5.4}$$

where cp is the complexity parameter. Figure 5.8 shows the MSE relative to when $cp = 0.1$, which is the default value for the complexity parameter, and gave an error $= 1$. The choice to decrease cp, by decreasing α, gave more nodes in the tree. In order to still keep a tree that is pruned, and simultaneously minimize the relative error $cp = 0.01$ was chosen. This gave six nodes in the corresponding tree shown in Figure 5.9. The rpart-function from the $rpart$ package was used to create the tree (Therneau and Atkinson, 2019).

The number of mobile lice was predicted based on the regression tree in Figure 5.9. If mechanical delousing was used, a number of $\exp(12.48) = 263024$ mobile lice at the farm were predicted. If mechanical delousing was not used, we proceeded to the left in the tree. Medicinal treatments can be interpreted in the same way. If medicinal treatments were not used, Distance from coastline and Shortest distance come into account. The regression tree suggested some important distances, these were Distance from coastline $= 11.32$ m, Shortest distance $= 1987$ m and Shortest distance $= 2153$ m. If the distance from the coastline to the farm was < 11.32 m the number of mobile lice at that farm was predicted to be $\exp(9.597) = 14721$. If Distance from coastline was ≥ 11.32 m, and Shortest distance was < 1987 m, according to the tree, $\exp(9.93) = 20537$ mobile lice were predicted. If all the previous conditions hold, and $1987\text{m} \leq$ Shortest distance < 2153 m the tree predicted a number of $\exp(2.465) = 12$ mobile lice. Finally, if Shortest distance was ≥ 2153 m a number of $\exp(8.163) = 3509$ mobile lice were predicted at the farm. All predictions were made regardless of time of year. From a single regression tree, the numbers should be carefully interpreted.

The method of random forest with bootstrapping was used to create regression trees. From the $randomForest$ package the randomForest-function in R was utilized to create the default value of 500 trees (Liaw and Wiener, 2002). For the $p = 7$ predictors in the dataset, $m = 3$ predictors were used to be considered in each split in the tree. Figure 5.11 consists of the two columns %IncMSE and IncNodePurity. %IncMSE measures the mean decrease of accuracy in the out-of-bag predictions when variable i was excluded from the model.

For each split in a tree, it was calculated how much this split reduced the node impurity, which is the difference between the RSS, before and after the split. In the second column in Figure 5.11 the reduction in node impurity is seen, which is summed over all splits, in all trees, for that variable. From Figure 5.11 Mechanical, Distance from coastline, Shortest

distance and Sea temperature were the most important variables in the prediction. The test set MSE from the random forest model was 8.39, an improvement from 13.39 from the single regression tree without using bagging. The variance explained by the random forest model is a measure of how well out-of-bag predictions explain the variance of the training set.

The variance explained by the random forest model was 47.67%.

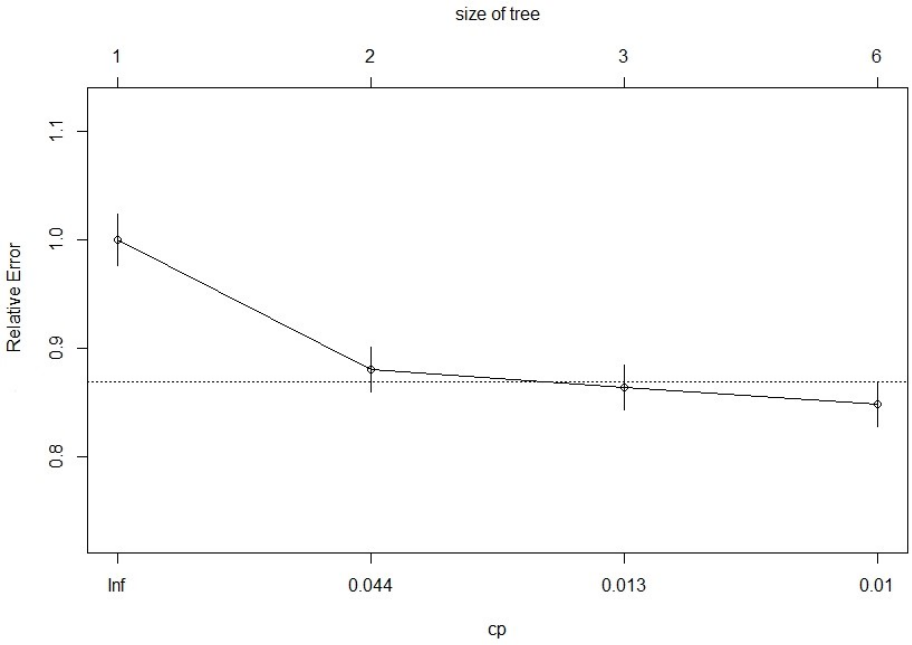


Figure 5.8: The relative error in MSE (error = 1 corresponds to the choice of cp = 0.1, the default) plotted against the complexity parameter, cp. The formula for cp was given in equation (5.4) and the lowest relative error was to choose 6 nodes in the tree with $cp = 0.01$. The corresponding regression tree is shown in figure 5.9.

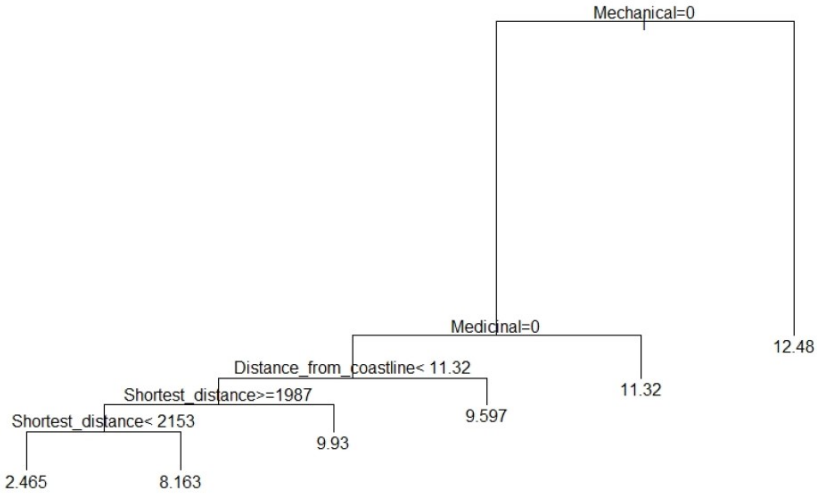


Figure 5.9: Plot of a single regression tree where the complexity parameter, cp, was chosen to be 0.01. The tree had six nodes and five splits.

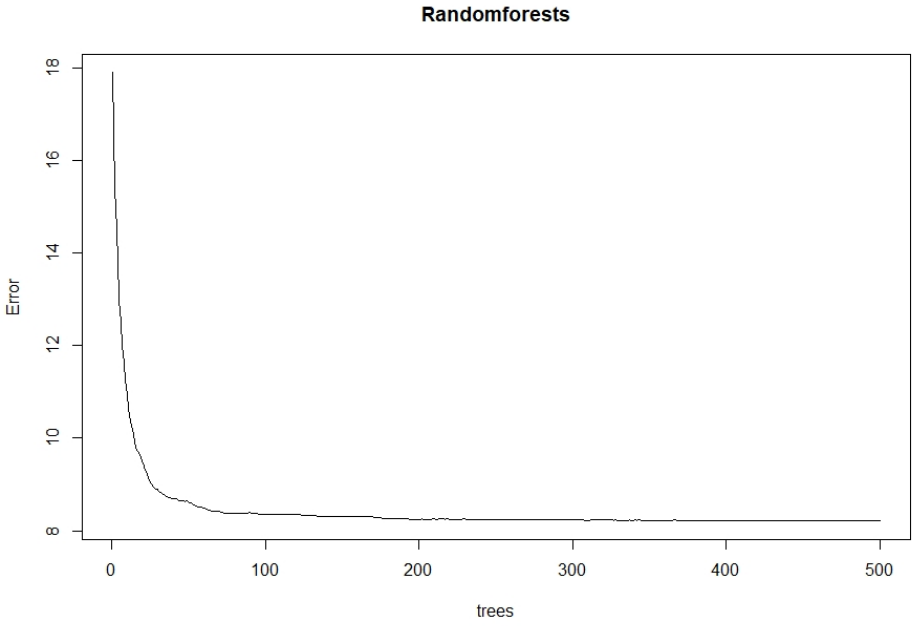


Figure 5.10: Plot of MSE *versus* number of trees in the random forest model.

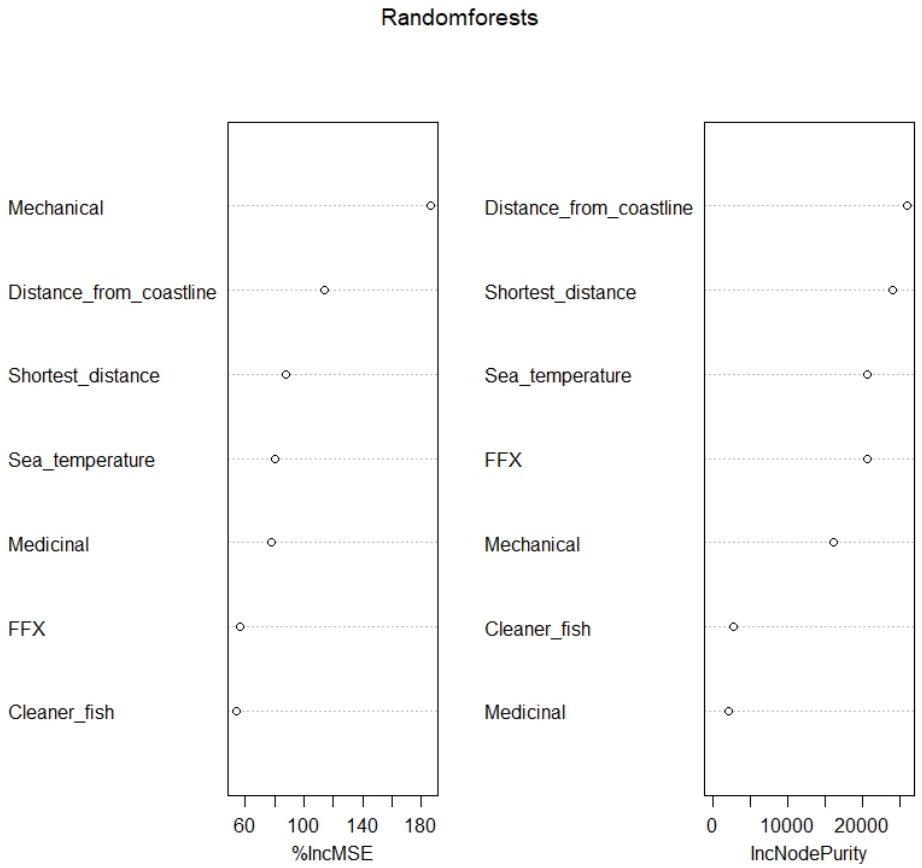


Figure 5.11: Variance important plots for the random forest model.

5.7 Generalized additive models

A generalized additive model was fitted to the data by the gam-function from the *gam* package in R (Hastie, 2019). Mobile lice was fitted against smoothing splines of degree three for Sea temperature, FFX, Distance from coastline and Shortest distance. The three last variables, the treatment variables, were fitted with a separate constant for each level. The resulting equation for this model was given from (2.76), and is

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + f_5(x_{i5}) + f_6(x_{i6}) + f_7(x_{i7}) + \epsilon_i,$$

where $\beta_0 = 53513$. Or, written in a more interpretative way as

$$\begin{aligned}\text{Mobile lice} = 53513 &+ f_1(\text{Sea temperature}) + f_2(\text{FFX}) + f_3(\text{Distance from coastline}) \\ &+ f_4(\text{Shortest distance}) + f_5(\text{Mechanical}) \\ &+ f_6(\text{Medicinal}) + f_7(\text{Cleaner fish}) + \epsilon\end{aligned} \tag{5.5}$$

The null hypothesis, H_0, of a linear relationship between the quantitative variables and the response was tested against the alternative hypothesis, H_1, of a non-linear relationship, i.e. smoothing spline of degree three, between the quantitative variables and the response. From the ANOVA-test from Table 5.7, there is reason to believe that the smoothing splines were sufficient for the variables Sea temperature, FFX, Distance from coastline and Shortest distance(all p-values < 0.05). In Figure 5.13, as in Figure 5.14, the variance increases with higher fitted values. This indicates that the model was not a good fit to the data. R^2 was calculated to be $R^2 = 1 - \frac{\text{Deviance}}{\text{Null deviance}} = 1 - \frac{5.43 \cdot 10^{15}}{6.418 \cdot 10^{15}} = 0.154$, which is a low number and the model explained little of the variation in the data.

Figure 5.12 shows the results from the fitted generalized additive model in equation (5.5). Fixing the other variables, a rapid increase in lice levels around 7–14°C is seen, before it flattens out and decreases around 14–18°C. There were few data points for temperatures > 18°C, hence this region should be carefully interpreted. For FFX, a decrease in the interval around 2–7m/s is seen, and a slight increase for wind measurements > 7m/s. From Figure 5.12, the number of mobile lice increases when the distance from the coastline increases, and the number of lice decreases when the farms are further away from each other. The linear trend in lice levels for Distance from coastline > 650m should be carefully interpreted. There were few data points in this interval, which increases the risk of overfitting. The single point at around 1500m decides the shape of the graph for values > 700m. For the same reason, the decrease in the number of mobile lice for Shortest distance > 6000m should be carefully interpreted. A rapid decrease in Mobile lice is seen for Shortest distance in the interval from 4000–6000m.

Table 5.7: DF for terms and F-values for Nonparametric Effects

	Df	Npar F	Pr(F)
(Intercept)			
s(Sea temperature, 3)	2	37.236	$< 2 \cdot 10^{-16}$
s(FFX, 3)	2	3.344	0.03532
s(Distance from coastline, 3)	2	4.608	0.00998
s(Shortest distance, 3)	2	22.353	$2.019 \cdot 10^{-10}$

The AIC value was 502156

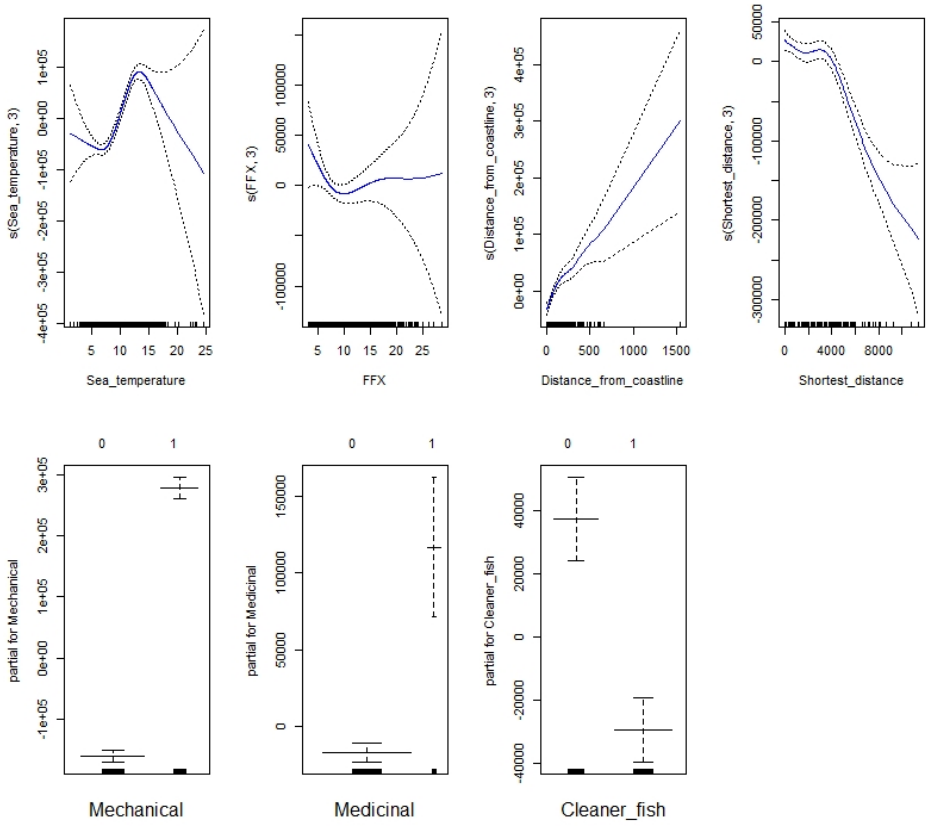


Figure 5.12: The first four plots in the first row show fitted natural splines in Sea temperature, FFX, Distance from coastline and Shortest distance with point-wise standard errors, respectively. The three last plots are step functions fitted to the factor variables Mechanical, Medicinal and Cleaner fish.

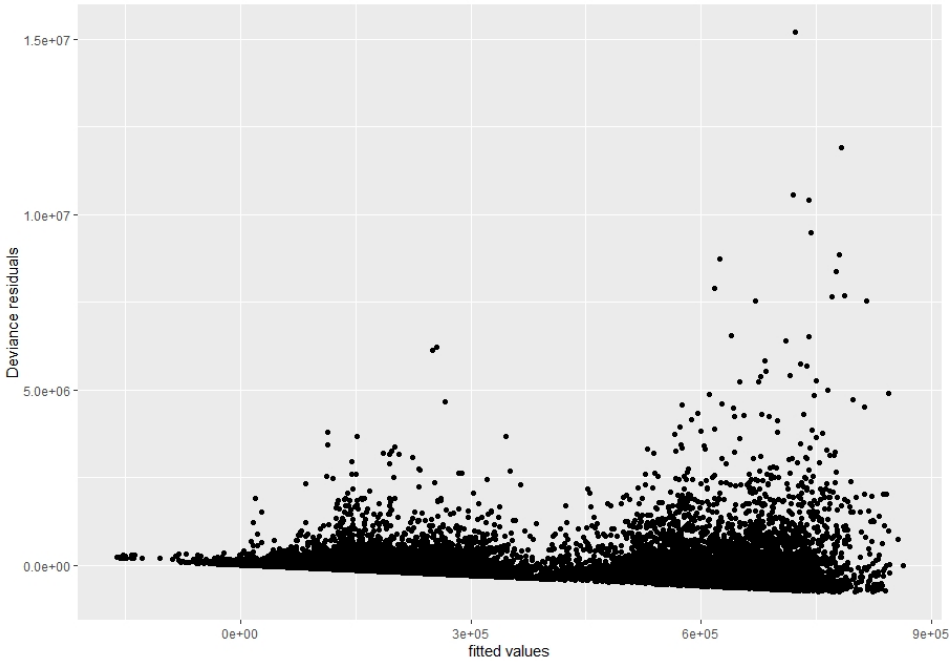


Figure 5.13: Plot of deviance residuals *versus* fitted values for the generalized additive model.

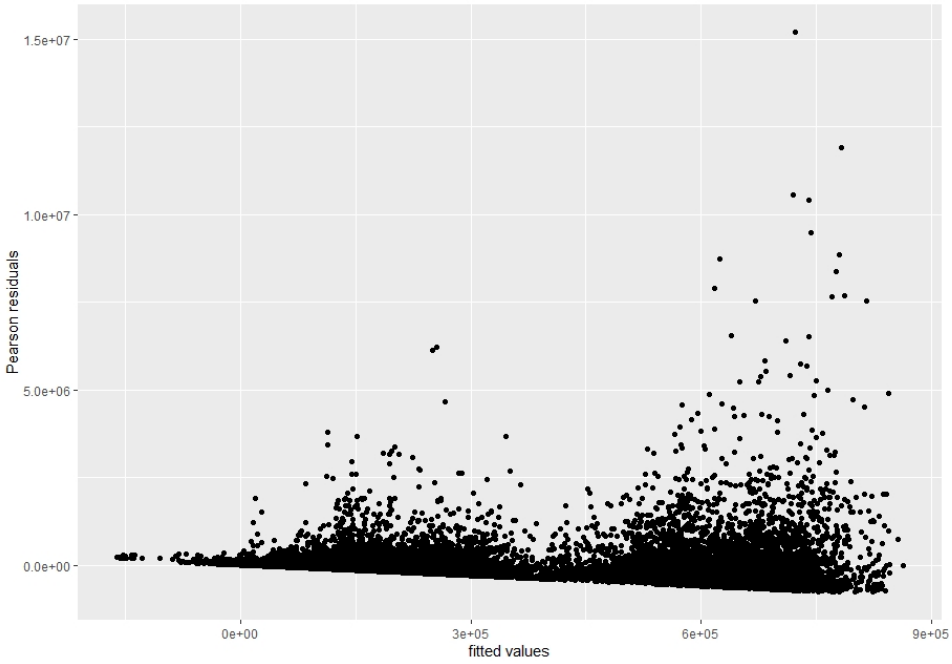


Figure 5.14: Plot of Pearson residuals *versus* fitted values for the generalized additive model.

Chapter 6

Discussion

The estimated count coefficients from the ZINB regression model, seen in Table 5.5, can be compared with the plots in Chapter 4. Some coefficients did not correspond particularly well with the plots in Chapter 4. The positive coefficients for Sea temperature and Distance from coastline indicate that increased sea temperatures and distances from the coastline correspond to more mobile lice. For Sea temperature seen in Figure 4.1, a positive coefficient seems appropriate for temperatures up to $15°$C, but the number of mobile lice drops with higher temperatures. From Figure 4.5, larger distances from the coastline to the farms seems to coincide with less mobile lice.

The plots in Figure 5.12 can be compared to the figures from Chapter 4. From Figure 4.2 the number of mobile lice decreases for higher FFX measurements, which does not correspond well to FFX $>$ 7m/s from Figure 5.12. From Figure 4.5 it could seem that increased distance from the coastline reduces the number of mobile lice, which does not correspond well with the monotonous increase from Figure 5.12. From Figure 4.3 it seems to be a decrease in the number of mobile lice when Shortest distance increases, although not as rapid a decrease as seen in Figure 5.12 for Shortest distance in the interval from 4000–6000 m.

The fitted models estimated an increase in Mobile lice when mechanical delousing and medicinal treatments were used, and a reduction when cleaner fish were deployed (with other terms kept constant). The treatments are used when the number of lice exceeds the limits given in Section 1.1. It is therefore not surprising that treatment variables were associated with higher lice numbers. Hence, these variables are not a reflection on the effect of the treatment, nor a cause to salmon lice. The 0/1 -coding in this study made it impossible to distinguish between cause and effect of the treatments, and the design in this study made it impossible to say if the treatments could be considered successful or not. The estimated interaction coefficients for Mechanical1:Medicinal1 and Medicinal1:Cleaner fish1 were negative for the GLMs. Several treatments methods are used simultaneously (in the same quarter) as an attempt to reduce the salmon lice abundance.

Distance from coastline and Shortest distance are variables that in this study, to a greater extent, can explain the cause of salmon lice. Although conservative evidence is

lacking, general agreement now exists that the risk of salmon louse infection increase in areas with intensive fish farming (Bjørn et al., 2001, 2008; Gargan et al., 2007). A study in Norway found that the distance to the closest fish farms played a key role in the success of protected salmon fjords, where salmon lice levels were consistently low over time when the farms were further away from each other (Serra-Llinares et al., 2014). Distances looked at were $>$ 30 km, i.e. larger distances than analysed in this thesis. In a study from Ireland, it was demonstrated that the greatest infestation of *L. salmonis* on sea trout were seen close to salmon farms (Gargan et al., 2007). A time series analysis from (Aldrin et al., 2019) on salmon lice count data also found that the infection pressure on neighbouring farms decreases by increasing seaway distances to the neighbours. In this study, Shortest distance was estimated by a straight line. Using seaway distances, rather than straight lines, would have been more accurate.

Mobile lice predictions for Distance from coastline, Shortest distance and Sea temperature from the different models can be compared. In Figure 6.1 predictions of Mobile lice *versus* Distance from coastline can be seen, where Sea temperature =12°C, FFX = 10 m/s, Shortest distance = 2000 m and all treatment variables equal to zero were kept fixed. The GLMs and the GAM model predict an increase in Mobile lice with increased distances. On the other hand, the random forest model predicts a decrease in the number of mobile lice, before it levels out at ≈ 0 for Distance from coastline > 600 m. The prediction from the single regression tree can be found directly from Figure 5.9. For Distance from coastline ≤ 11 m a number of $\exp(2.465) = 12$ lice were predicted. For Distance from coastline ≥ 12 m the regression tree predicted a number of $\exp(9.597) = 14721$ lice. As there is rather few data points for Distance from coastline > 650m, the prediction of ≈ 0 mobile lice by the random forests model seems most reasonable according to Figure 4.5.

Figure 6.2 shows Mobile lice predicted by Shortest distance where Sea temperature =12°C, FFX = 10 m/s, Distance from coastline = 50 m and all treatment variables equal to zero were kept constant. The single regression tree predicts 14721 lice over the whole interval, seen directly from Figure 5.9, whereas the GLMs and the GAM predict a decrease in Mobile lice for greater distances between farms. The random forest model predicts the highest abundance of lice in the region 3000m $<$ Shortest distance $<$ 4500m. Mobile lice decreases for Shortest distance > 6000 m and approaches 0, which seems reasonable according to Figure 4.3, as there are relatively few observations for Shortest distance $>$ 6000 m. It should be mentioned that in Figure 6.2 the prediction from the ZINB regression model approaches 0 for Shortest distance > 6000 m which seems reasonable.

In Figure 6.3 Mobile lice is predicted by Sea temperature when FFX = 10 m/s, Shortest distance = 2000 m, Distance from coastline = 50 m and all treatments variables = 0. The GLMs predicted an monotonous increase in Mobile lice on the interval. Although the numbers differ a lot, the random forest model and the GAM model both predict an increase in the mobile lice numbers for Sea temperature from around 7°C to around 14°C and it drops with higher sea temperatures, which seems reasonable according to Figure 4.1. A study in western Canada, where a multiple regression model was used, found that sea temperature had little effect on the number of mobile lice (Revie et al., 2004). On the other hand, a space-time analysis from Norway found that the number of lice increased with increasing sea temperatures (Aldrin et al., 2013). A study also found that increasing sea temperature had an increased effect on the maturation rate in both the naupliar stages

and in the chalimus stage (Stien et al., 2005).

The predictions considered (seen in Figures 6.1, 6.2 and 6.3) clarified the problem by using GLMs that use the same function throughout the whole interval.

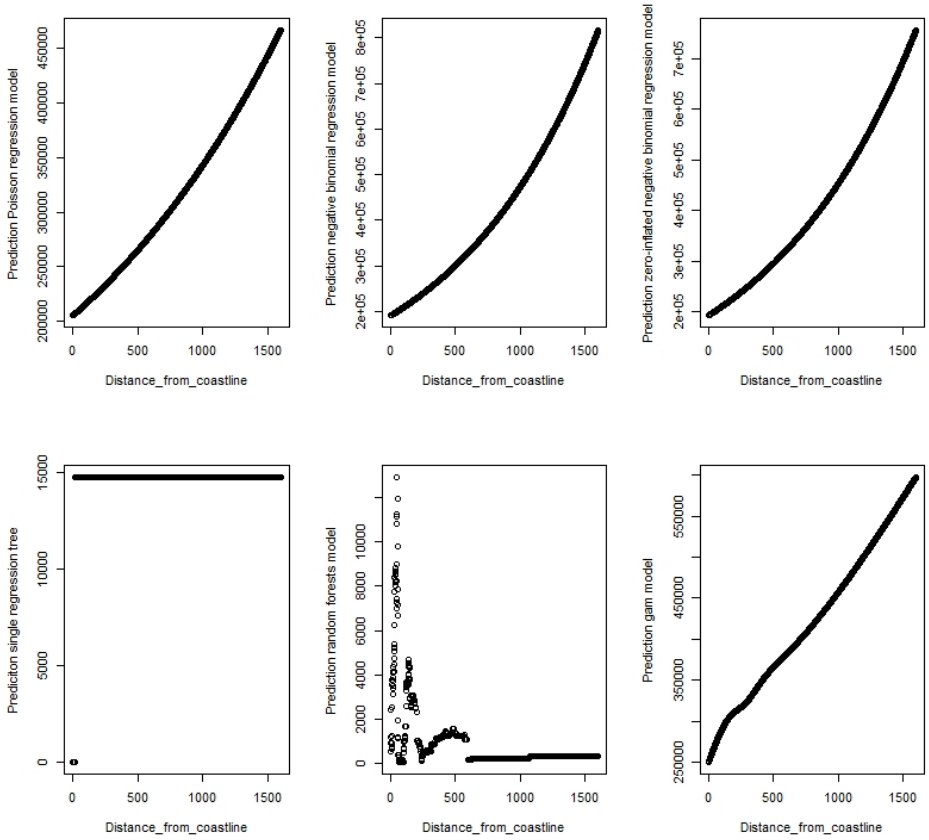


Figure 6.1: Prediction of Mobile lice *versus* Distance from coastline for observations where Sea temperature =12°C, FFX = 10 m/s, Shortest distance = 2000 m and all treatment variables 0

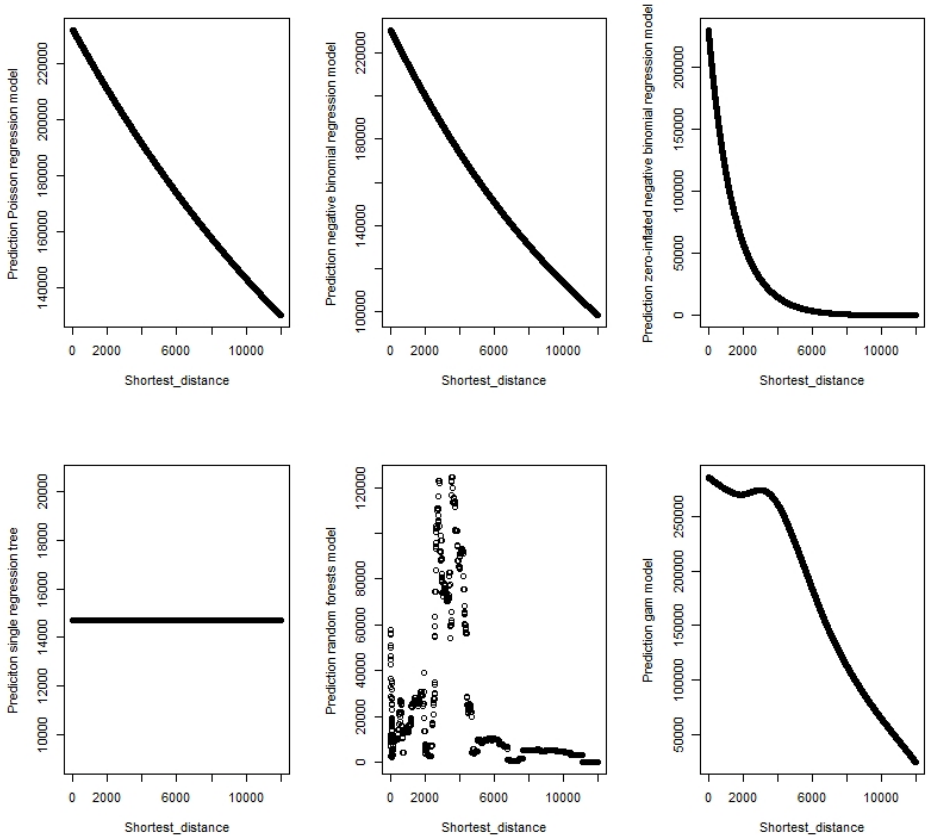


Figure 6.2: Prediction of Mobile lice *versus* Shortest distance for observations where Sea temperature =12°C, FFX = 10 m/s, Distance from coastline = 50 m and all treatment variables 0

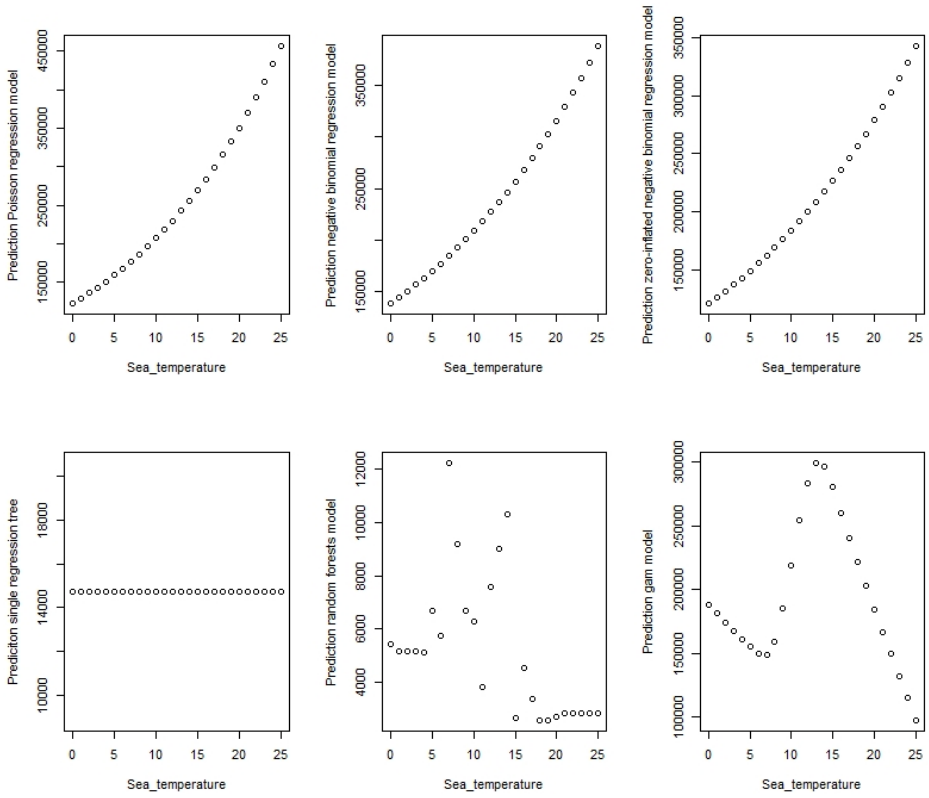


Figure 6.3: Prediction of Mobile lice *versus* Sea temperature for observations where Shortest distance = 2000 m, FFX = 10 m/s, Distance from coastline = 50 m and all treatment variables 0

6.1 Concluding remarks on models used

The starting point in this analysis was a Poisson regression model. That model was over-dispersed and a quasi-Poisson regression model and a negative binomial regression model were fitted. A zero-inflated negative binomial regression model was fitted as an attempt to account for the zero-inflation in the data. The great spread in the variance seen from the residual plots from the fitted Generalized linear models indicated lack of fit. From the AIC criterion, the zero-inflated negative binomial regression model was considered the best among the fitted GLMs. In search for models that could fit the data better, the GAM model and the tree based models were fitted. Such models have an advantage of being more flexible and dynamic, but at the cost of being harder to interpret. This concerns especially the random forest model. The GAM model did not give a particularly good fit either. Log-transforming Mobile lice and fitting a random forest model gave the best result in terms of variance explained by the models. The random forest model stated that Mechanical, Distance from coastline, Shortest distance and Sea temperature were the most important variables for the number of mobile salmon lice. The model predicted less lice with higher sea temperatures, for farms further away from other farms(> 6000 m), and further away from the coastline (> 650 m), in correspondence with Figures 4.1, 4.3 and 4.5. The predictions from Figures 6.1, 6.2 and 6.3 emphasized that GLMs using the same analytic functions on the interval, did not give a good fit to the data.

6.2 Challenges and recommendations for further work

A more dynamic way to estimate the number of lice should be investigated. This could be done by finding a proportionality term between the weight of the salmon and the time after the salmon smolt sea transfer, which may give a more accurate estimation of the number of salmon, and therefore also lice. A more accurate number could have resulted in improved fit by the models. Exact data on the percentage of total biomass capacity utilized at any time, could also be very useful in this estimation. The number of salmon at the farm could have been collected directly, which would have given a more accurate estimation of the lice number.

It is reasonable to assume that salmon farms far away from the coastline, also are further away from other farms, since the majority of the farms are relatively close to the coast. Thus, the variable Distance from coastline can be influenced by Shortest distance. Distance from coastline should hence be investigated further.

For the wind estimates, the directions were not taken into account. To include the wind direction in a reasonable way in a model could be demanding, but worthwhile, as wind direction affect the current in the sea that might influence the spread of salmon lice. There were also cases where the weather stations were far away from the farms. Measuring wind speed at locations closer to the farms, could provide more accurate measurements.

To see any effect of the use of cleaner fish, mechanical delousing or medicinal treatment methods it could have been interesting to work with a smaller area with fewer farms and preferably for larger periods. It could be interesting to include the exact number of cleaner fish deployed at the different farms, to see if this had a significant impact on the model. Furthermore, it could be interesting with a further analysis to compare weeks

where farms used the treatments method wholly, or only in parts. One could also have focused more specifically at which kind of medicinal or mechanical treatment that was used at the farm.

One can use a time series analysis approach to model dependency between data points and study how the number of lice at one farm one week will affect the number of lice for future weeks at the farm. This could be done by using a similar strategy as (Aldrin et al., 2019) where the count of mobile lice at each farm were treated as time series, that also included potential dependency within the same week between time series of lice counts at other farms. Other time-dependent approaches should also be considered.

It could have been interesting to add more variables in the analysis, such as the time since the salmon smolts were released.

Bibliography

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

Aldrin, M., Storvik, B., Kristoffersen, A.B., Jansen, P.A., 2013. Space-time modelling of the spread of salmon lice between and within norwegian marine salmon farms. *PLOS ONE* 8, 1–10. URL: https://doi.org/10.1371/journal.pone.0064039, doi:10.1371/journal.pone.0064039.

Bjørn, P.A., Finstad, B., Kristoffersen, R., 2001. Salmon lice infection of wild sea trout and Arctic char in marine and freshwaters: the effects of salmon farms. *Aquaculture Research*, 32(12), 947–962.

Bjørn, P.A., Finstad, B., Nilsen, R., Asplin, L., 2008. Nasjonal overvåkning av lakselusinfeksjon på ville bestander av laks, sjøørret og sjørøye i forbindelse med nasjonale laksevassdrag og laksefjorder. *NINA Rapport*, 377, 1–33.

Carslaw, D.C., Ropkins, K., 2012. openair — an r package for air quality data analysis. Environmental Modelling Software 27–28, 52–61. doi:10.1016/j.envsoft.2011.09.008.

Casella, G., Berger, R.L., 2002. *Statistical Inference*. 2nd Edition. Belmont: Brooks/cole, Cengage Learning.

Dominique, L., Park, B., 2010. Negative binomial regression models and estimation methods. URL: https://www.icpsr.umich.edu/CrimeStat/files/CrimeStatAppendix.D.pdf.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. *Regression. Models, Methods and Application*. Springer-Verlag Berlin Heidelberg.

Gargan, P., Tully, O., Poole, R., 2007. Relationship Between Sea Lice Infestation, Sea Lice Production and Sea Trout Survival in Ireland, 1992-2001. pp. 119 – 135. doi:10.1002/9780470995495.ch10.

Hamre, L., Eichner, C., Caipang, C., Dalvin, S., Bron, J., 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): e73539. URL: https://doi.org/10.1371/journal.pone.0073539.

Hastie, T., 2019. gam: Generalized Additive Models. URL: https://CRAN.R-project.org/package=gam. r package version 1.16.1.

Hastie, T., Tibshirani, R., Friedman, J., 2017. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd Edition. Springer Series in Statistics.

Hayward, C., Andrews, M., Nowak, B., 2011. Introduction: Lepeophtheirus salmonis—a remarkable success story. *Salmon Lice: An Integrated Approach to Understanding Parasite Abundance and Distribution* , 1 – 28doi:10.1002/9780470961568.ch.

Helgesen, K.O., Horsberg, T.E., Tarpai, A., 2019. The surveillance programme for resistance to chemotherapeutants in salmon lice (*Lepeophtheirus salmonis*) in Norway 2018. Norwegian Veterinary Institute.

Hijmans, R.J., 2019. geosphere: Spherical Trigonometry. URL: https://CRAN.R-project.org/package=geosphere. r package version 1.5-10.

Hjeltnes, B., Bang Jensen, B., Bornø, G., Haukaas, A., Walde, C.S., 2019. Fiskehelserapporten 2018. Norwegian Veterinary Institute.

Ho, J.S., 2000. The major problem of cage aquaculture in Asia relating to sea lice. In: Cage Aquaculture in Asia, Proceedings of the First International Symposium on Cage Aquaculture in Asia (eds I. Liao and C. Lin), pp.13-19. Asian Fisheries Society, Manila and World Aquaculture Society, Southeast Asian chapter, Bangkok.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. *An Introduction to Statistical Learning with Applications in R*. New York, Heidelberg, Dordrecht, London: Springer.

Lawless, J., 1987. Negative binomial and mixed Poisson regression. *The Canadian Journal of Statistics* 15(3) , 209–225URL: https://doi.org/10.2307/3314912.

Lester, R., Hayward, C., 2006. Phylum Arthropoda. In: *Fish Diseases and Disorders*, Volume 1: Protozoan and Metazoan Infections 2nd Edition (ed P.T.K. Woo), 463-562. CABI Publishing, Oxon, UK.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. *R News* 2, 18–22. URL: https://CRAN.R-project.org/doc/Rnews/.

Maran, B., Moon, S., Ohtsuka, S., Oh, S., Soh, H., Myoung, J., Iglikowska, A., Boxshall, G., 2013. The caligid life cycle: new evidence from Lepeophtheirus elegans reconciles the cycles of Caligus and Lepeophtheirus (Copepoda: Caligidae), 20 (15). *Parasite* (France, Paris). doi:10.1051/parasite/2013015.

Nakashima, E., 1997. Some Methods for Estimation in a Negative-Binomial model, 49 (1), 101-115. URL: https://www.ism.ac.jp/editsec/aism/pdf/049_1_0101.pdf.

Olsvik, P., Samuelsen, O., Agnalt, A.L., Lunestad, B., 2015. Transcriptional responses to teflubenzuron exposure in european lobster (homarus gammarus). *Aquatic toxicology* (Amsterdam, Netherlands) 167, 143–156. doi:10.1016/j.aquatox.2015.07.008.

Overton, K., Dempster, T., Oppedal, F., Kristiansen, T.S., Gismervik, K., Stien, L.H., 2018. Salmon lice treatments and salmon mortality in norwegian aquaculture: a review. *Reviews in Aquaculture*, 1-20. doi:10.1111/raq.12299.

Padgham, M., Rudis, B., Lovelace, R., Salmon, M., 2017. osmdata. The Journal of Open Source Software 2. URL: https://doi.org/10.21105/joss.00305, doi:10.21105/joss.00305.

Poley, J.D., Braden, L.M., Messmer, A.M., Igboeli, O.O., Whyte, S.K., Macdonald, A., Rodriguez, J., Gameiro, M., Rufenerd, L., Bouvierd, J., Dorota, W.W., Koop, B.F., Hosking, B. C. andFast, M.D., 2018. High level efficacy of lufenuron against sea lice (*Lepeophtheirus salmonis*) linked to rapid impact on moulting processes. *International Journal for Parasitology: Drugs and Drug Resistance*, 8 (2), 174-188.

R Core Team, 2019. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Regulations on salmon lice control, 2012. Norwegian directorate of fisheries. forskrift om lakselusbekjempelse. *Forskrift om bekjempelse av lakselus i akvakulturanlegg* nr 6 måling av sjøtemperatur og telling av lakselus og nr 8 grenser for lakselus og tiltak [regulations on salmon lice control. regulations on combating salmon lice in aquaculture plants. no 6 measurement of sea temperature and counting of sea lice and no 8 limits for salmon lice and treatments]. in: Fiskeridepartementet n-o (ed). https://lovdata.no/dokument/SF/forskrift/2012-12-05-1140.

Revie, C., Gettinby, G., Treasurer, J., Wallace, C., 2004. Identifying epidemiological factors affecting sea lice lepeoptherius salmonis abundance on scottish salmon farms using general linear models. *Diseases of aquatic organisms* 57, 85–95. doi:10.3354/dao057085.

Rueness, E., Berg, P.R., Gulla, S., Halvorsen, K., Järnegren, J., Malmstrøm, M., Mo, T., Rimstad, E., de Boer, H., Eldegard, K., Hindar, K., Hole, L.R., Kausrud, K., Kirkendall, L., Inger Måren, I., Erlend B. Nilsen, E.B., Thorstad, E.B., Nielsen, A., Velle, G., 2019. Assessment of the risk to norwegian biodiversity from import of wrasses and other cleaner fish for use in aquaculture, 1-111. Norwegian Scientific Committee for Food and Environment (VKM), Oslo, Norway.

Serra-Llinares, R.M., Bjørn, P.A., Finstad, B., Nilsen, R., Harbitz, A., Berg, M., Asplin, L., 2014. Salmon lice infection on wild salmonids in marine protected areas: an evaluation

of the Norwegian 'National Salmon Fjords'. *Aquaculture Environment Interactions*, 5, 1-16. doi:10.3354/aei00090.

Stien, A., Bjørn, P., Heuch, P., Elston, D., 2005. Population dynamics of salmon lice lepeophtheirus salmonis on atlantic salmon and sea trout. *Marine Ecology-progress Series - MAR ECOL-PROGR SER* 290, 263–275. doi:10.3354/meps290263.

Stram, D., Lee, J., 1994. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50 (4) , 1171–1177doi:10.2307/2533455.

Therneau, T., Atkinson, B., 2019. rpart: Recursive Partitioning and Regression Trees. URL: https://CRAN.R-project.org/package=rpart. r package version 4.1-15.

Thorstad, E., B., Finstad, B., 2018. Impacts of salmon lice emanating from salmon farms on wild atlantic salmon and sea trout. *NINA Report 1449*, 1-22. URL: https://www.salmon-trout.org/wp-content/uploads/2018/01/Thorstad-Finstad-2018-Impacts-of-salmon-lice-NINA-Report-1449-2.pdf.

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

Venables, W.N., Ripley, B.D., 2002. *Modern Applied Statistics with S*. 4th ed., Springer, New York. URL: http://www.stats.ox.ac.uk/pub/MASS4. iSBN 0-387-95457-0.

Woo, P.T.K., Buchmann, K., 2012. Fish Parasites. Pathobiology and Protection. *Fish and Fisheries*, 13 (4), 477. CAB International.

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

Zuur, A., Ieno, E., Walker, N., Saveliev, A., Smith, G., 2009. *Mixed Effects Models and Extensions in Ecology with R*. Springer science+Business media.

Appendix

R-code used to create the dataset

```
library(readxl)
library(dplyr)
library(psych)
library(tidyverse)
library(sf)
library(geosphere)
library(osmdata)
library(rnaturalearthdata)
library(progress)
library(openair)
library(lubridate)
#The code to create the initial dataset 1 in 3.1
data = read_excel("lakselus_per_fisk17-19.xlsx")
licedata= subset(data,
Produksjonsomr deId <8 & Produksjonsomr deId>5)
licedata2 = licedata[complete.cases(licedata), ]
#####################################################################
#The code to create dataset 3 explained in 3.2
biomassedata<- read_excel("biomasse_lokalitet.xlsx")
#Correponds to dataset 3
biomassekg=select(biomassedata, Lokalitetsnummer, kapasistetkg)
licedatabio=left_join(licedata2, biomassekg,
by="Lokalitetsnummer")
licedatabio$antallfisk=licedatabio$kapasistetkg*0.9/5
licedatabio$lusibevegelse=licedatabio$`Lus i bevegelige stadier`
*licedatabio$antallfisk
licedatabio$voksnehunnlus=licedatabio$`Voksne hunnlus`
*licedatabio$antallfisk
licedatabio$fastsittendelus=licedatabio$`Fastsittende lus`
*licedatabio$antallfisk
#####################################################################
#The code to create dataset 2 explained in 3.1
#Look at cleaner fish, mechanical delousing and
medicinal treatment
rens=read_excel("tiltak_mot_lakselus17-19.xlsx")
rens=subset(rens, Produksjonsomr deId <8
& Produksjonsomr deId>5)
medi=subset(rens, rens$Tiltak=="medikamentell")
meka=subset(rens, rens$Tiltak=="mekanisk_fjerning")
rens=subset(rens, rens$Tiltak=="rensefisk")
```

```
for (i in 1:nrow(licedatabio)){
  if (licedatabio$Uke[i]<14){
    licedatabio$Kvartal[i]=1
  } else if (licedatabio$Uke[i]>13 &licedatabio$Uke[i]<27){
    licedatabio$Kvartal[i]=2
  } else if (licedatabio$Uke[i]>26 &licedatabio$Uke[i]<40){
    licedatabio$Kvartal[i]=3

  } else {
    licedatabio$Kvartal[i]=4
  }
}
for (i in 1:nrow(rens)){
  if (rens$Uke[i]<14){
    rens$Kvartal[i]=1
  } else if (rens$Uke[i]>13 &rens$Uke[i]<27){
    rens$Kvartal[i]=2
  } else if (rens$Uke[i]>26 &rens$Uke[i]<40){
    rens$Kvartal[i]=3

  } else {
    rens$Kvartal[i]=4
  }
}

for (i in 1:nrow(medi)){
  if (medi$Uke[i]<14){
    medi$Kvartal[i]=1
  } else if (medi$Uke[i]>13 &medi$Uke[i]<27){
    medi$Kvartal[i]=2
  } else if (medi$Uke[i]>26 &medi$Uke[i]<40){
    medi$Kvartal[i]=3

  } else {
    medi$Kvartal[i]=4
  }
}
for (i in 1:nrow(meka)){
  if (meka$Uke[i]<14){
    meka$Kvartal[i]=1
  } else if (meka$Uke[i]>13 &meka$Uke[i]<27){
    meka$Kvartal[i]=2
  } else if (meka$Uke[i]>26 &meka$Uke[i]<40){
    meka$Kvartal[i]=3

  } else {
    meka$Kvartal[i]=4
  }
}
rens=rens[c(2,3,5,18,20)]
```

```
medi=medi[c(2,3,5,18,20)]
meka=meka[c(2,3,5,18,20)]
sammensattmedrens =
left_join(licedatabio, rens,
by=c(" r ", "Lokalitetsnummer", "Kvartal"))
sammensattmedmedi =
left_join(licedatabio, medi,
by=c(" r ", "Lokalitetsnummer", "Kvartal"))
sammensattmedmeka=
left_join(licedatabio, meka,
by=c(" r ","Lokalitetsnummer", "Kvartal" ))
#Remove duplicated rows
sammensattmedrens1=
subset(sammensattmedrens,
!duplicated(subset(sammensattmedrens,
select=c(Uke, Kvartal,  r , Lokalitetsnummer,Fylke))))
sammensattmedmedi1=
subset(sammensattmedmedi,
!duplicated(subset(sammensattmedmedi,
select=c(Uke, Kvartal,  r , Lokalitetsnummer,Fylke))))
sammensattmedmeka1=
subset(sammensattmedmeka,
!duplicated(subset(sammensattmedmeka,
select=c(Uke, Kvartal,  r , Lokalitetsnummer,Fylke))))

#Order the data points in ascending order
sammensattmedrens1 =
sammensattmedrens1
[order(sammensattmedrens1$ r , sammensattmedrens1$Uke),]
sammensattmedmedi1=
sammensattmedmedi1
[order(sammensattmedmedi1$ r , sammensattmedmedi1$Uke), ]
sammensattmedmeka1=
sammensattmedmeka1
[order(sammensattmedmeka1$ r , sammensattmedmeka1$Uke), ]

#Create the variables for lice treatments
sammensattmedmeka1$mekaniskfjerning=0
sammensattmedmedi1$medikamentell=0
sammensattmedrens1$rensefiskniv =0
sammensattmedrens1[is.na(sammensattmedrens1)]=F
sammensattmedmedi1[is.na(sammensattmedmedi1)]=F
sammensattmedmeka1[is.na(sammensattmedmeka1)]=F
#Look at 2017
for (i in 1369:2727){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv [i]=1
  }

}
```

```
for (i in 2728:4270){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}
for (i in 4271:5850){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}
Look at 2018
for (i in 5851:7221){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}

#Secon quarter
for (i in 7222:8623){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}
#Third quarter
for (i in 8624:10187){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}
#Forth quarter
for (i in 10188:11780){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}

#Look at 2019
for (i in 11781:13149){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}
```

```
#Second quarter
for (i in 13150:14586){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}

#Third quarter
for (i in 14587:16163){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}

#Forth quarter
for (i in 16164:17746){
  if(sammensattmedrens1$Tiltak[i]=="rensefisk"){
    sammensattmedrens1$rensefiskniv[i]=1
  }

}




#Medicinal treatment
#Look at 2017
#first quarter
for (i in 1:1368){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Second quarter
for (i in 1369:2727){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Third quarter
for (i in 2728:4270){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }
```

```
}

#Forth quarter
for (i in 4271:5850){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}




#Look at 2018
for (i in 5851:7221){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Second quarter
for (i in 7222:8623){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Third quarter
for (i in 8624:10187){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Fourth quarter
for (i in 10188:11780){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Look at 2019
for (i in 11781:13149){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
```

```
  }

}

#Second quarter
for (i in 13150:14586){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Third quarter
for (i in 14587:16163){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

#Fourth quarter
for (i in 16164:17746){
  if(sammensattmedmedi1$Tiltak[i]=="medikamentell"){
    sammensattmedmedi1$medikamentell[i]=1
  }

}

Mechanical treatment
#Look at 2017
#First quarter
for (i in 1:1368){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Second quarter
for (i in 1369:2727){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}
#Third quarter
for (i in 2728:4270){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }
```

```
}

#Fourth quarter
for (i in 4271:5850){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Look at 2018
for (i in 5851:7221){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Second quarter
for (i in 7222:8623){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Third quarter
for (i in 8624:10187){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Fourth quarter
for (i in 10188:11780){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Look at 2019
for (i in 11781:13149){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}
```

```
#Second quarter
for (i in 13150:14586){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Third quarter
for (i in 14587:16163){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

#Fourth quarter
for (i in 16164:17746){
  if(sammensattmedmeka1$Tiltak[i]=="mekanisk␣fjerning"){
    sammensattmedmeka1$mekaniskfjerning[i]=1
  }

}

sammensattmedmedi2=select(sammensattmedmedi1,  r , Uke,
Lokalitetsnummer, medikamentell)

sammensattmedmeka2=select(sammensattmedmeka1,  r , Uke,
Lokalitetsnummer, mekaniskfjerning)


sammensattmedrens2=
left_join(sammensattmedrens1, sammensattmedmedi2)

sammensattmedrens2=
left_join(sammensattmedrens2, sammensattmedmeka2)


sammensattmedrens2=select(sammensattmedrens2, -c(28,30,32))
############################################################
#Create dataset 4 that estimated the distance
to the coastline in 3.3

sammensattmedrens3=select(sammensattmedrens2,
Lokalitetsnummer,Lon,Lat,
Produksjonsomr deId.x, Fylke)
sammensattmedrens3=subset(sammensattmedrens3,
!duplicated(subset(sammensattmedrens3,
select=c(Lokalitetsnummer,Lon,Lat,Produksjonsomr deId.x))))
```

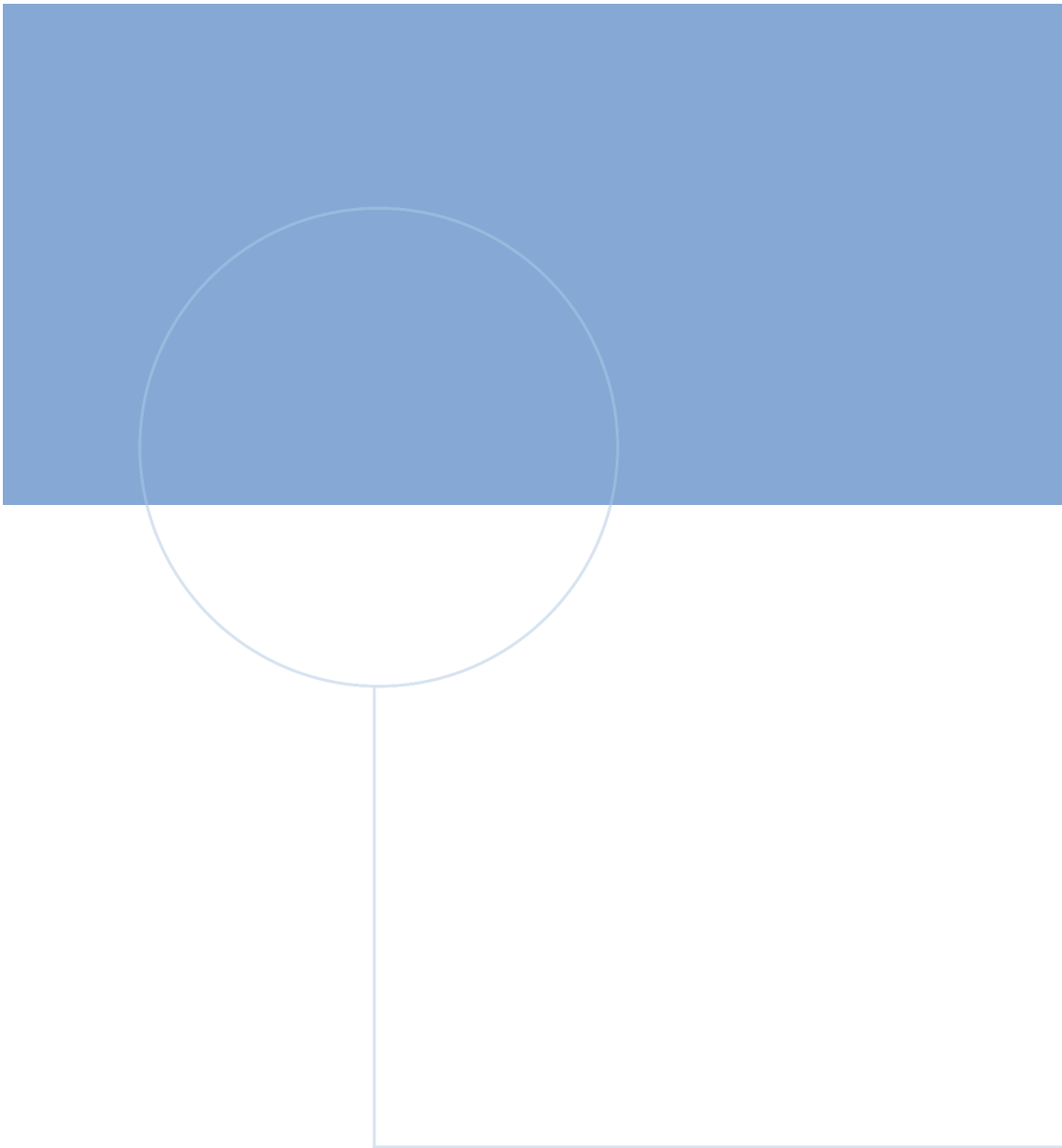
```
rownames(sammensattmedrens3) <- 1:nrow(sammensattmedrens3)
d1_sf = sammensattmedrens3 %>%
st_as_sf(coords = c('Lon','Lat')) %>%
  st_set_crs(4326)
d1_sf = select(d1_sf, Lokalitetsnummer,geometry)
osm_box = getbb (place_name = "T r  n d e l a g") %>%
  opq () %>%
  add_osm_feature("natural", "coastline") %>%
  osmdata_sf()
osm_boxmore = getbb (place_name = "M   r e␣og␣Romsdal") %>%
  opq () %>%
  add_osm_feature("natural", "coastline") %>%
  osmdata_sf()
osm_boxnordland = getbb (place_name = "Nordland") %>%
  opq () %>%
  add_osm_feature("natural", "coastline") %>%
  osmdata_sf()

library(geosphere)
# use dist2Line from geosphere
dist = dist2Line(p = st_coordinates(d1_sf),
                 line =
                st_coordinates(osm_box$osm_lines)[,1:2])
dist2= dist2Line(p = st_coordinates(d1_sf),
                line =
                st_coordinates(osm_boxmore$osm_lines)[,1:2])
dist3= dist2Line(p = st_coordinates(d1_sf),
                line =
                st_coordinates(osm_boxnordland$osm_lines)[,1:2])

#combine initial data with distance to coastline


distt=as.data.frame(dist)
distt=select(distt,distance)
colnames(distt)[colnames(distt)=="distance"] <- "distanceT"

distm=as.data.frame(dist2)
distm=select(distm, distance)
colnames(distm)[colnames(distm)=="distance"] <- "distanceM"

distn=as.data.frame(dist3)
distn=select(distn, distance)
colnames(distn)[colnames(distn)=="distance"] <- "distanceN"

disttot=cbind(distt, distm, distn)
disttot$Distance=apply(disttot, 1, FUN=min)
disttot1=select(disttot, Distance)

#Merge this with the location numbers
```

```
locdist=
select(sammensattmedrens3, Lokalitetsnummer, Lon, Lat)
locdist$Distance=disttot1$Distance
#We must merge this with the rest
sammesattmedrensogdist=
left_join(sammensattmedrens2, locdist,
by=c("Lokalitetsnummer", "Lon", "Lat"))
##########################################################
#Code to create dataset 6 from section 3.5
vinddatasett = read_excel("Stasjonermedvindogretning.xlsx")
stasjonsnavnkoordinater=read_excel("Stasjonsnavnkoordinater.xlsx")
stasjonkoordogvind=
left_join(vinddatasett, stasjonsnavnkoordinater, by=c("Stnr"))
stasjonskoordogvind1
=select(stasjonkoordogvind, Stnr, Dato, FFM, FFN, FFX, Lat, Lon)

vinddatasett$date=as.Date(vinddatasett$Dato)
split= split.data.frame(vinddatasett, vinddatasett$Stnr)
vinddataveke10380=timeAverage(split$`10380`, avg.time = "week")
vinddataveke59610=timeAverage(split$`59610`, avg.time = "week")
vinddataveke59680=timeAverage(split$`59680`,avg.time = "week")
vinddataveke59695=timeAverage(split$`59695`, avg.time = "week")
vinddataveke59800=timeAverage(split$`59800`, avg.time = "week")
vinddataveke60190=timeAverage(split$`60190`, avg.time = "week")
vinddataveke60240=timeAverage(split$`60240`, avg.time = "week")
vinddataveke60500=timeAverage(split$`60500`, avg.time = "week")
vinddataveke60810=timeAverage(split$`60810`, avg.time = "week")
vinddataveke60930=timeAverage(split$`60930`, avg.time = "week")
vinddataveke60990=timeAverage(split$`60990`, avg.time = "week")
vinddataveke61060=timeAverage(split$`61060`, avg.time = "week")
vinddataveke61410=timeAverage(split$`61410`, avg.time = "week")
vinddataveke61420=timeAverage(split$`61420`, avg.time = "week")
vinddataveke62270=timeAverage(split$`62270`, avg.time = "week")
vinddataveke62480=timeAverage(split$`62480`, avg.time = "week")
vinddataveke62980=timeAverage(split$`62980`, avg.time = "week")
vinddataveke63420=timeAverage(split$`63420`, avg.time = "week")
vinddataveke63630=timeAverage(split$`63630`, avg.time = "week")
vinddataveke63705=timeAverage(split$`63705`, avg.time = "week")
vinddataveke63820=timeAverage(split$`63820`, avg.time = "week")
vinddataveke64330=timeAverage(split$`64330`, avg.time = "week")
vinddataveke65310=timeAverage(split$`65310`, avg.time = "week")
vinddataveke65451=timeAverage(split$`65451`, avg.time = "week")
vinddataveke65940=timeAverage(split$`65940`, avg.time = "week")
vinddataveke66150=timeAverage(split$`66150`, avg.time = "week")
vinddataveke67280=timeAverage(split$`67280`, avg.time = "week")
vinddataveke67560=timeAverage(split$`67560`, avg.time = "week")
vinddataveke68010=timeAverage(split$`68010`, avg.time = "week")
vinddataveke68290=timeAverage(split$`68290`, avg.time = "week")
vinddataveke68860=timeAverage(split$`68860`, avg.time = "week")
vinddataveke69100=timeAverage(split$`69100`, avg.time = "week")
```

```
vinddataveke69380=timeAverage(split$`69380`, avg.time = "week")
vinddataveke70150=timeAverage(split$`70150`, avg.time = "week")
vinddataveke71000=timeAverage(split$`71000`, avg.time = "week")
vinddataveke71550=timeAverage(split$`71550`, avg.time = "week")
vinddataveke71850=timeAverage(split$`71850`, avg.time = "week")
vinddataveke71990=timeAverage(split$`71990`, avg.time = "week")
vinddataveke72580=timeAverage(split$`72580`, avg.time = "week")
vinddataveke73466=timeAverage(split$`73466`, avg.time = "week")
vinddataveke73500=timeAverage(split$`73500`, avg.time = "week")
vinddataveke73550=timeAverage(split$`73550`, avg.time = "week")
vinddataveke74350=timeAverage(split$`74350`, avg.time = "week")
vinddataveke75220=timeAverage(split$`75220`, avg.time = "week")
vinddataveke75410=timeAverage(split$`75410`, avg.time = "week")
vinddataveke75550=timeAverage(split$`75550`, avg.time = "week")
vinddataveke76240=timeAverage(split$`76240`, avg.time = "week")
vinddataveke76330=timeAverage(split$`76330`, avg.time = "week")
vinddataveke76450=timeAverage(split$`76450`, avg.time = "week")
vinddataveke76530=timeAverage(split$`76530`, avg.time = "week")
vinddataveke76750=timeAverage(split$`76750`, avg.time = "week")
vinddataveke77230=timeAverage(split$`77230`, avg.time = "week")
vinddataveke77280=timeAverage(split$`77280`, avg.time = "week")
vinddataveke77425=timeAverage(split$`77425`, avg.time = "week")
#We find which of the weather stations that are complete,
#and which that are not and store these in a dataframe.

vinddatany=rbind(vinddataveke10380, vinddataveke59680,
vinddataveke59695,
vinddataveke59800, vinddataveke60190, vinddataveke60240,
vinddataveke60500,
vinddataveke60810, vinddataveke61060, vinddataveke61410,
vinddataveke61420,
vinddataveke62270, vinddataveke62480, vinddataveke62980,
vinddataveke63420,
vinddataveke63630, vinddataveke63705, vinddataveke63820,
vinddataveke64330,
vinddataveke65310, vinddataveke65451, vinddataveke65940,
vinddataveke66150,
vinddataveke67280, vinddataveke67560, vinddataveke68290,
vinddataveke68860,
vinddataveke69100, vinddataveke69380, vinddataveke71000,
vinddataveke71550,
vinddataveke71850, vinddataveke72580, vinddataveke73550,
vinddataveke74350,
vinddataveke75220, vinddataveke75410, vinddataveke76240,
vinddataveke76330,
vinddataveke76450, vinddataveke76530, vinddataveke76750,
vinddataveke77230,
vinddataveke77280, vinddataveke77425)

vinddatany$date=as.Date(vinddatany$date)
```

```
vinddatany$Uke = week(vinddatany$date)
vinddatany$ r  = year(vinddatany$date)
vinddatany=subset(vinddatany, Uke!=53)

stasjonsnavnkoordinater$location_id=1:45
DB2=select(stasjonsnavnkoordinater,location_id, Lat, Lon )
sammesattmedrensogdist$location_id=1:17746
DB1=select(sammesattmedrensogdist, location_id, Lat, Lon)
DistFun = function(ID){
  TMP = DB1[DB1$location_id==ID,]
  TMP1 = distHaversine(TMP[,3:2],DB2[,3:2])
  TMP2 = data.frame(DB1ID=ID,DB2ID=DB2[which.min(TMP1),1],
  DistanceBetween=min(TMP1)      )
  print(ID)
  return(TMP2)
}

DistanceMatrix = bind_rows(lapply(DB1$location_id, DistFun))
DistanceMatrix$Stnr=0
DistanceMatrix=select
(DistanceMatrix, location_id, DistanceBetween)
DistancewithStnr=left_join
(DistanceMatrix, stasjonsnavnkoordinater)

for (i in 1:17746){
  DistancewithStnr$Lokalitetsnummer[i]=
  sammesattmedrensogdist$Lokalitetsnummer[i]
}

DistancewithStnr=select(DistancewithStnr, Stnr, Lokalitetsnummer)
sammesattmedrensogdist$Stnr=DistancewithStnr$Stnr

vinddatany=select(vinddatany,  r , Uke, Stnr, FFM, FFX, FFN)

Dataset1=left_join(sammesattmedrensogdist, vinddatany)
write.table(Dataset1, file = "Datasetttotalrens1.csv",
row.names=FALSE, na="",col.names=TRUE, sep=",")
##################################################################
#We create dataset 5 from section 3.4
closestloc=subset(data, !duplicated(subset(Dataset1,
select=c(Lokalitetsnummer,Lon,Lat))))
closestloc=select(closestloc, Lokalitetsnummer, Lon, Lat)
distanse=data.frame()

for (i in 1:nrow(closestloc)){
  for (j in 1:nrow(closestloc)){
    distanse=
    append(distanse, distm(c(closestloc$Lon[i],closestloc$Lat[i]),
    c(closestloc$Lon[j],closestloc$Lat[j]),fun=distHaversine))
  }
```

```
}
output=matrix(unlist(distanse), ncol = 240, byrow = F)
outputframe=as.data.frame(output)
for ( i in 1:240){
  closestloc$Mindist[i]=min
  (outputframe[,i][which(outputframe[,i]>0)])
}

closestloc=left_join(Dataset1, closestloc)
#Create final dataset, dataset 1 in 3.6
write.table(closestloc, file = "Datasettrensclosdist.csv",
row.names=FALSE,na="",col.names=TRUE, sep=",")
```



NTNU

Norwegian University of
Science and Technology