NTNU

Norwegian University of
Science and Technology

Analysis of The Norwegian Veterinary Institute's Model for Salmon Lice Abundance Prediction

Johannes Bogen

Master of Science in Physics and Mathematics
Submission date: February 2018
Supervisor: Jon Andreas Støvneng, IFY
Co-supervisor: Kerstin Bach, IDI
Lars Martin Sandvik Aas, Ecotone AS
Paul Anton Letnes, Ecotone AS

Norwegian University of Science and Technology
Department of Physics

Abstract

Empirical studies have found that lice from salmon farms are a main source of infection of wild salmonids. Due to Norway's responsibility to conserve wild stocks of salmon, Veterinærinstituttet has created a statistical model to predict the lice infections of wild salmon. In this work, this model has been analyzed with respect to how well is predictions correlate with reported weekly manual lice counts from Norwegian salmon farms between 2012 and 2017, and how different factors affects its predictive abilities. The model was found to perform best before 2016, in the summer, when the water temperature is 8 °C to 11 °C and rising, and when a farm has few neighboring farms in its vicinity. The model was also compared to the hydrodynamical model created by Havforskningsinstituttet. While Veterinærinstituttet's model was found to be generally better, it was outperformed by the former when predicting lice abundances on farms with low current velocities. A set of new models were also created using linear- and support vector regression. These were on par with Veterinærinstituttet's model during its best months, but superior the rest of the year, illustrating the possibilities for using machine learning in this field .

Preface

I would first like to thank my advisors for their assistance and support throughout the writing of this thesis. Paul Anton Letnes and Lars Martin Sandvik Aas at Ecotone AS have given valuable input and helped me shape the focus of the thesis, and Kerstin Bach at NTNU has been indispensable for the creation of the regression models, among other things. I would also like to thank Jon Andreas Støvneng at NTNU for supervising my thesis.

This thesis would not have been possible without the help and support from key people in the industry. Anja Kristoffersen and Peder Jansen at Veterinærinstituttet, Ingrid Johnsen at Havforskningsinstituttet and Ingrid Ellingsen at SINTEF Ocean have provided me with their data, answered my questions and given me valuable information necessary for a proper analysis and comparison of their models.

Table of Contents

List of Tables

List of Figures

Abbreviations

AF	=	Adult Female (lice)
CH	=	Chalimus (lice)
HI	=	Havforskningsinstituttet
IQR	=	Inter-Quartile Range
PAAM	=	Pre-Adult and Adult Male (lice)
ROC	=	Rate Of Change
SVM	=	Support Vector Machine
SVR	=	Support Vector machine for Regression
VI	=	Veterinærinstituttet
ZINB	=	Zero-Inflated Negative Binomial

Chapter 1

Introduction

Salmon farming is major industry in Norway, and accounts for more than 80 percent of the country's total aquaculture production [1]. In 2016, more than 1.2 billion kilograms of salmon was produced, at a total value of almost 60 billion NOK [2].

As the number of farmed salmon are increasing rapidly, the high density of fish gives great conditions for salmon lice to thrive, causing high levels of infection on salmon. This is not only a major cost to the industry, which suffered an estimated loss of 3 billion NOK in 2014 [3], but empirical studies have found that lice from salmon farms are a main source of infection of wild salmonids [4].

Due to Norway's responsibility to conserve wild stocks of salmon [5], measures to control the amount of lice have been put in place. Manual counts of salmon lice on farmed salmon must be performed on regular intervals, and treatments must be applied if the counted amount exceeds the government regulations [6]. Due to the inherent difficulty in monitoring the lice infections on wild salmonids, a cost-effective and accurate method of monitoring lice levels is desired [7, 8].

Two of the institutions that are invested in the modeling of sea lice abundances is Veterinærinstituttet (VI) and Havforskningsinstituttet (HI). The former has developed a statistical model for the spreading of sea lice from farms, while the latter has combined hydrodynamic and biological models to simulate the movement of lice in the current. However, there have been a lack of analyses and comparisons of the models, although they have been frequently used to predict lice abundances [9]. This work aims to provide an analysis of the models' abilities to predict lice abundances in salmon farms, and how this ability varies with various conditions. This work is mainly focused on the model created by VI, as this had the highest amount of data available.

The model is first analyzed on its own, and then compared with HI's model to see whether their predictive abilities differ. The prediction ability has been measured by the correlation

between time series of counted and predicted lice, as well as the mean absolute difference between the time series.

Machine learning is in the progress of taking a big step into the area of salmon lice prevention through AquaCloud [10], a pilot project launched in April 2017 which aims to use big data to predict lice development and recommend treatment methods to keep the lice population under control. However, while AquaCloud will likely be very useful for farmers and the Norwegian aquaculture in general, it appears to not be usable for predicting lice abundances on wild salmon. In this work, new models have been created using linear regression and support vector regression, and compared to the VI model as an alternative approach.

Chapter 2

Theory

The first section gives a short introduction to the biology of salmon lice and the factors that are known to affect infection of salmon. Section 2.2 follows with an overview of the salmon farming cycle and the lice-related regulations imposed on the farms. The existing models created by VI and HI are then presented in Section 2.3, and Section 2.4 contains an overview of the tools used for the evaluation of these models. Lastly, Section 2.5 gives the theory behind the regression techniques utilized in this work, and Section 2.6 outlines the machine learning techniques employed when training and evaluating the regression models.

2.1 Salmon lice

2.1.1 About the louse

The salmon louse, or Lepeophtheirus salmonis (Krøyer, 1837), is a parasitic copepod and a part of the larger family Caligidae, which is collectively known as sea lice. The life cycle of the salmon louse consists of eight distinct stages: After hatching from a string of eggs in the water, the louse enters two naupliar stages in which it is living free in the water and is unable to feed. It then moves onto a copepodid stage, in which it has to find and infect a fish. Once it has, it moves onto two chalimus stages, in which it uses a frontal filament to attach itself onto a fish. The final stages are two pre-adult and one adult stage, in which the louse is mobile and able to move across the surface of the fish as well as swim in the water column [11].

The duration of each of the stages in the life cycle depends directly on the water temperature [12]. In aquaculture, the duration of a stage is therefore given in degree days (°d),

defined as the sum of the daily temperature degrees. For instance, a louse can gain 100 °d by being in water holding 10 °C for 10 days, or 2 °C for 50 days.

The combined duration of the naupliar and copepodid stages has been found to vary between 22 days at 5 °C and about 12 days at 15 °C, in which the lice is in its infectious copepodid stage for about 50 °d and 140 °d, respectively [13]. The duration of these stages is the most vital to know when modeling the movement of lice, as it directly limits how long the lice can drift with the currents before it has to find a host. As the louse cannot feed until finding a host, it will die if unable to find one before the duration of its copepodid stage is complete.

The water temperature also affects the reproduction of lice, with a reported hatching time from about 45 days at 2 °C to 9 days at 10 °C [14]. It has also been found that at lower temperatures, the number of eggs in the egg strings increase, but a higher percentage are nonviable [15].

The louse is able to swim up and down in the water column, but can only move horizontally by drifting with the currents [16, 17]. It also rises to the the upper part of the water column during the day and sinks to deeper layers at night, in a pattern known as diurnal migration [18, 19]. In waters where there are salinity gradients in the water column, the louse has been found to swim up or down in the water column to stay at salinities above 27 parts per thousand (ppt). At salinities below 29 ppt, the survival of the louse is severely lowered [18, 20].

2.1.2 Factors affecting lice infections

Several studies have been done to investigate how different factors affect the abundance of lice on fish. These factors include the temperature, salinity and current velocity of the water in which the salmon reside, the amount of salmon in a farm and how large they are, and the presence of neighboring salmon farms.

From farms located in western Canada, at a latitude of about 50°, Saksida et al. [21] constructed a general linearized model to find that water temperature had very little effect on the number of mobile lice. The temperatures ranged from 6 °C to 13 °C. The same results were found from Scottish salmon farms [22, 23] with a similar temperature range, located at a latitude of about 55° to 58°. On the other hand, Aldrin et al. investigated lice counts from Norwegian salmon farms and found that the water temperature is a strong predictor of lice abundances, with the number of lice increasing with increasing temperatures, and much stronger at a latitude of 68° than 60° [24].

As previously mentioned, the salinity of the water is also known to affect the lice, and in addition to the increased mortality of the lice, a lower salinity has also been found to reduce the ability of copepodids to remain attached to fish [20]. Statistical modeling of lice counts on a small number of salmon farms in Hardangerfjorden in Norway showed that lice abundance was significantly lower where the salinity levels were lower [25]. On the other hand, Saksida et al. [21] found that salinity, like water temperature, had a very limited effect.

Current velocities at farms have been found to correlate negatively with the lice abundance [23]. This correlation has also been shown from laboratory tests [26], where lice subjected to stronger currents were less likely to succeed in attaching itself to a host.

The size of the fish have been found to have a strong positive correlation with lice abundances in Norwegian farms [24, 25], while analyses on Scottish farms has found it to be both unimportant [23] and important [27]. Similarly, Aldrin et al. [24] found from Norwegian farms that increasing the number of fish in a farm increases the lice abundance, while Revie et al. [23] found the density of fish in a farm to have no effect. Lastly, Aldrin et. al. also found that a higher amount of nearby salmon farms increases the abundance of lice. In addition, several models have been created showing how lice can float with the currents from neighboring farms (see for example [5, 17, 28]).

2.2 Salmon farming

2.2.1 Salmon farming cycle

The production cycle of a Norwegian salmon farm broadly consists of two phases: Production and fallowing. At the start of the production phase, a group of young fish of the same year class (termed a cohort) are introduced to the farm. This will be referred to as a farm 'becoming active' for the remainder of this thesis. The cohort is then grown for roughly 18 months before they are slaughtered. After slaughtering, the farm must be fallowed for at least two months before the production cycle can be repeated [29].

2.2.2 Lice counting regulations

All Norwegian salmon farms are required to count and report the number of lice that are found on the salmon. Starting from 2012, the regulations require that the lice are counted at least every seven days if water temperatures exceed 4 °C, and at least every 14 days if the temperature is below that. The counting does not have to take place if all the fish in the farm are to be slaughtered within 14 days after the counting was to be completed. The weekly or biweekly lice count has to be performed on least 10 randomly selected fish (20 from week 14 to 21 in southern Norway and week 19 to 26 in northern Norway) from half of the cages at the farm, with the other half of the cages being counted the following week[1]. The numbers must be reported no later than Tuesday the following week [6].

The salmon lice are counted in in three different categories according to their gender and stage in the developmental cycle: Chalimus (CH); mobiles, or pre-adult and adult males and pre-adult females (PAAM); and adult females (AF). The number of adult females is the most important, as lice in this stage are the only ones capable of reproduction. Salmon farms are required to keep the average number of adult female lice below the threshold set by the government, which is 0.5 lice per fish (0.2 per fish from week 14 to 21 in southern

[1] From March 2017, new regulations requires all cages to be counted every week

Norway and week 19 to 26 in northern Norway) [6]. The counting of lice in the chalimus stage is the most difficult, as the lice can be very small (down to a length of about 1 mm) and therefore hard to spot [30]. It has been estimated that only 9 % to 19 % of chalimus are actually counted [31].

In addition to the counted lice, the water temperature at 3 m depth and information on any treatments that have been performed are also reported weekly [6]. Production numbers are also reported, but only at the end of every month. These numbers include the number and the mean mass of the fish in the farm [29]. As new fish can be introduced and existing fish can die, escape, be relocated or slaughtered, the number of fish in the farm can fluctuate on a sub-month level. A monthly report of these numbers can therefore be inaccurate when used together with weekly lice counts.

2.3 Existing Models

In this section the existing models are introduced, the statistical model created by Veterinærinstituttet in Section 2.3.1, and then Havforskningsinstituttet's model in Section 2.3.2. The models are briefly explained and results from previous validations are recounted. Lastly, Section 2.3.3 contains a summary of previous comparisons of the two models.

2.3.1 Veterinærinstituttet's model

The model created by Veterinærinstituttet (the VI model) is described by Kristoffersen et.al. [32]. The model combines counted values of adult female salmon lice with simple biological models for the fecundity and the growth and mortality rates of the lice. A model for the infection pressure from neighboring farms, based on the seaways distance between the farms, is added to account for the spread of lice between farms. The output from the model is the infection pressure of chalimus, which can be used in regression to estimate the abundance of lice on salmon. The basis for the model is the reports of adult female lice counts from a total of 1.645 salmon farms in Norway, from week 1 of 2012 until week 40 of 2017.

Model description

First, the model calculates the number of eggs released from one farm at a given week. This is done by multiplying the average counts of adult female lice n_{AF} by the number of salmon n_{salmon}. Each of these female lice are assumed to produce 300 eggs, and the total number of larvae and when they are released are calculated using temperature-dependent models for the fecundity F given by Stien et.al [33]. Biological parameters for the further development of the lice, i.e. the growth rate from one stage to the next and the mortality rate, are given in the same article.

The development time from hatching until reaching the copepodid stage is set to 126 °d, during which the lice has a mortality of 17 % every day. This gives a total mortality of $S_{CO} = (1-0.17)^{\Delta d_{CO}}$, where Δd_{CO} is the number of days it takes before a total of 126 °d are accumulated, and the louse has reached the copepodid stage.

After reaching the copepodid stage, a delay of four days is included to account for the time it takes for a louse to find and infect a host. To calculate the infection pressure of PAAM lice, a further development time of 155 °d with a mortality of 0.05 each day is added, $S_{CH} = (1-0.05)^{\Delta d_{CH}}$.

The infection pressure on a farm is divided into the internal infection pressure (IIP) and the external infection pressure (EIP). To calculate the internal pressure, the total amount of adult female lice are used with the aforementioned parameters and summed over all days that contribute, to calculate the total amount of chalimus lice at day d:

$$IIP(d) = \sum_{\forall \Delta d_*} n_{AF}(d') n_{\text{salmon}}(d') F(d') S_{CO}(\Delta d_{CO}) S_{CH}(\Delta d_{CH}), \tag{2.1}$$

where

$$d' = d - (\Delta d_{CO} + \Delta d_{CH} + 4)$$

and $\Delta d_* = \Delta d_{CO} + \Delta d_{CH} + 4$ represents all previous days that contributes with chalimus to day d. The temperature dependence of the biological parameters is implied.

The external infection pressure on farm i from other farms is then calculated by weighting the internal infection pressure of all farms within a seaways distance of 100 km with a function of the seaways distance between them, and summing the values,

$$EIP_i(t) = \sum_{j=i} IP_j(t) R_{ij}, \tag{2.2}$$

where R_{ij} is a measure of the relative risk of infection between farms i and j,

$$R_{ij} = \frac{\exp(-1.444 - 0.351 \cdot (d_{ij}^{0.57} - 1)/0.57)}{\exp(-1.444 + 1/0.57))} \tag{2.3}$$

with d_{ij} the distance between the farms in kilometers. The distance function R_{ij} was modeled by Aldrin et.al. [24] using zero-inflated negative binomial regression. The total infection pressure IP is then given by adding Eq. (2.1) and Eq. (2.2), giving

$$IP = IIP + EIP. \tag{2.4}$$

Previous evaluations

This thesis employs a similar method to the one employed by Kristoffersen et.al. [32]. There, the counts of mobile lice from 370 cohorts, during the first 16 weeks without treatment, was compared to the 16 week average EIP. That is, the authors investigated the fit of the modeled amount of lice coming from other farms, to the number of counted lice. The cohorts were grouped into tertiles based on the their average EIP during the 16 weeks, and

the average weekly abundance of lice was calculated within each group. The relationship between predictors and counted lice was then modeled with zero-inflated negative binomial regression. In addition to the *EIP*, other predictors included the lice counts at the previous week, the water temperature and the week of the year. The results showed that the *EIP* had a significant contribution to the model performance.

In 2017, scientists from VI again evaluated this model by predicting lice counts on salmon cages placed in different locations along the Norwegian fjord [34]. These sentinel cages were deployed yearly from 2012 to 2017, 87 % in May or June, the remaining in July or August, and left for a period of 12 to 30 days. After each trial period, the fish were removed from the cage and the lice were counted and reported for each fish [5]. The model was applied to this data and the relationship between infection pressure and lice counts was modeled using mixed model regression with a negative binomial variance structure. The results showed a linear dependence between the predicted and counted numbers of lice, albeit with a significant variance. The results also appeared to differ strongly between the different regions the cages were placed in.

2.3.2 Havforskningsinstituttet's model

The second model [16] is developed by Havforskningsinstituttet. It is a numerical model that simulates the movement and development of lice using information on water currents, temperatures and salinity. HI has published several papers in which a hydrodynamic model have been used to simulate the dispersion of salmon lice [5, 16, 17, 35, 36]. One result of this research is a salmon lice map[2] which shows the estimated density of copepodids along the Norwegian coastline, as well as the reported lice counts from salmon lice farms.

Model description

The hydrodynamic model consists of a coastal ocean and a fjord model, and is based on the Regional Ocean Model System (ROMS) [37, 38]. The NorKyst800 model gives a 800×800m grid resolution for the coastal area. In the fjords, a 200×200m resolution is obtained using boundary values from NorKyst800, fresh water input from rivers connected to the fjord and atmospheric models [39].

This model is used as input to the salmon lice growth model, which is a modified version of a Lagrangian Advection and Diffusion Model [40]. This model uses hourly values of currents, salinity and temperature from the hydrodynamic model to simulate the movement and growth of salmon lice. It uses linear interpolation to achieve sub-grid spatial resolution, and sets a temporal resolution of 180 s [16].

The model works by first releasing simulated salmon lice particles in a planktonic stage at the farms. The number of larvae to releasewere calculated in the same way as for VI's model in Section 2.3.2. These numbers were linearly interpolated to daily values, and released hourly. Each of the particles were given a random movement in vertical and

[2] `http://hi.no/lakseluskart/html/lakseluskart.html`

horizontal direction at each timestep to simulate sub-grid turbulence, and were otherwise only carried by the currents [41]. The model also accounts for diurnal migration by letting the simulated lice swim upwards during the day and downwards at night. The infective copepodid stage is set to be between 40 °d and 170 °d [5], after which the lice are assumed to be dead. The resulting map gives the density (number of lice per square meter) of copepodids with a grid resolution of 160 m.

Previous evaluations

The results from the model were evaluated by comparing predicted values with counted lice from cages with salmon smolts in Hardangerfjorden in May and/or June between 2012 and 2015 [5]. In the research, 18 cages each containing 30 salmon, were left out for a period of 2-3 weeks. After each trial the fish were extracted and the number of lice were counted. The total data consisted of counts on 122 such groups of salmon. The time-integrated mean, median and maximum modeled lice densities in a 3×3 and 5×5 grid around each cage were compared to the lice counts with the Spearman's rank correlation [42] and the accuracy as metrics. When calculating the accuracy, the lice infestation level was divided into four groups by the mean number of lice on each fish - low (0-1 lice), moderate (1-5), medium (5-10) and high(>10) - giving an accuracy of 0.78.

2.3.3 Previous comparisons of the models

The abilities of the two models to predict lice abundances on 99 salmon cages in 2014 were compared in Aldrin et.al. [43]. The analysis found that the VI model had a higher R^2-value [44] than the HI model, and was thus concluded to be a better model. Another comparison by Qviller et. al. [45] was done using reported lice from 2019 salmon farms in 2014. Once again, the results found that VI's model was better than HI's due to having a lower BIC (Bayesian information criterion) [46] score.

2.4 Statistical tools

2.4.1 Zero-inflated negative binomial regression

The number of lice counted on one fish is assumed to be Poisson distributed. However, the variability in counted lice between the fish is expected to be larger than the variance of the Poisson distribution, as some fish carry few and others many lice. This situation where the variance is greater than the mean is called over-dispersion, and is better modeled by a negative binomial distribution [24].

The negative binomial distribution is a generalization of the Poisson distribution, and describes the number of successes k in a series of Bernoulli trials (an experiment with a

True/False outcome) before r failures occur. The probability mass function of the distribution is

$$f(k \mid r,p) = \binom{k+r-1}{r-1} p^r (1-p)^k$$

where p denotes the probability of success.

In negative binomial regression, one assumes that the response variable Y has a negative binomial distribution, and that the logarithm of its expected value can be modeled by a linear combination of parameters,

$$log(E(Y \mid \vec{x})) = \vec{\theta}^{\intercal}\vec{x},$$

where $\theta \in \mathbb{R}^n$ are the regression parameters and $\vec{x}$ the input vector. This gives a predicted mean of the regression model of

$$E(Y) = \exp^{\vec{\theta}^{\intercal}\vec{x}}. \tag{2.5}$$

The zero-inflated negative binomial (ZINB) distribution is useful for when there is an excessive amounts of zeros in the data. It separates the probability outcome into two distinct mechanisms, the first being the normal distribution and the second a separate probability of being zero,

$$E(Y) = \begin{cases} 0 & \text{with probability } p = \pi \\ \exp^{\vec{\theta}^{\intercal}\vec{x}} & \text{with probability } p = 1-\pi \end{cases}$$

giving an expectation value of

$$E(Y) = (1-\pi)\exp^{\vec{\theta}^{\intercal}\vec{x}} \tag{2.6}$$

where π is the probability that the zero-inflation mechanism will produce a zero [47].

If during the regression the probability of being zero is set to depend on the value of the modeled parameters $\vec{x}$, the regression returns a probability vector $\vec{\Pi} \in \mathbb{R}^n$ from which the probabilities $\vec{\pi}$ for each of the count values can be calculated as [48]

$$\vec{\pi} = \frac{exp^{\vec{\Pi}^{\intercal}\vec{x}}}{1+exp^{\vec{\Pi}^{\intercal}\vec{x}}}. \tag{2.7}$$

2.4.2 Pearson correlation coefficient

The Pearson correlation coefficient is a measure of the strength of the linear relationship between two variables. In other words, the coefficient measures the deviance of the data points from a fitted straight line. The correlation coefficient for a sample of n points $\{(x_1,y_1),...,(x_n,y_n)\}$ is given as

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{2.8}$$

where $\bar{x}$ and $\bar{y}$ are the respective average values of $\{x\}$ and $\{y\}$. The coefficient r can take values between -1 and $+1$, where a value of $r = +1$ corresponds to a perfect positive linear correlation, $r = -1$ to a perfect negative linear correlation, and $r = 0$ to no correlation. Fig. 2.1 shows examples of time series and how they correlate to different correlation values.

2.4.3 Mean positive and negative error

As a complementary measure to the correlation, the positive and negative error of the models has been used. The error here refers to the difference between modeled and counted lice numbers for each week. The error is denoted positive if the modeled value is higher than the counted, and negative if lower. With the error $e(i) = modeled(i) - counted(i)$, where i denotes a week between 1 and n, where n is the length of the time series, the mean positive error e_+ and negative error e_- are calculated as

$$\begin{aligned} e_+ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} e(i), \quad \forall e(i) > 0 \\ e_- &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} e(i), \quad \forall e(i) < 0 \end{aligned} \tag{2.9}$$

2.5 Regression techniques

2.5.1 Linear regression

Linear regression [49] is a statistical method for analyzing the relationship between two or more variables. In its simplest form, the method aims to fit a line through a set of predictor values x_i and a set of target values y_i. This line will be on the form

$$\hat{y}_i = a_0 + a_1 x_i,$$

where the constants a_0 and a_1 are chosen so that the least squares error

$$LSE = \sum_i (y_i - \hat{y}_i)^2$$

is minimized. Once this line has been found, it can be used to predict the value of the target value when given a predictor variable.

If there are $p > 1$ sets of predictor values x_i, the method is called multiple linear regression [50] and the fitted line will be on the form

$$\hat{y}_i = a_0 + \sum_{j=1}^{p} a_j x_{i,j}. \tag{2.10}$$

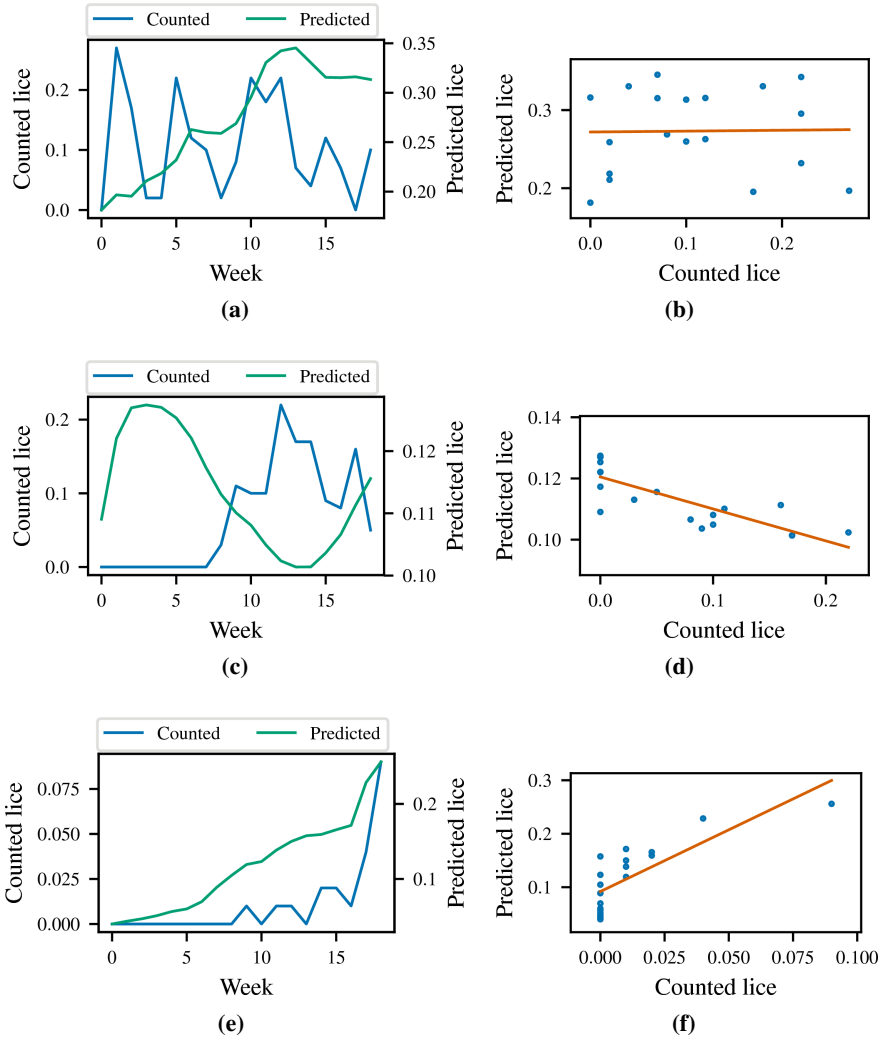


Figure 2.1: Examples of Pearson correlation coefficients for time series, and how they appear as time series plots (left column) and scatter plots (right column). The drawn line in the scatter plot is the best linear fit. The series have correlation coefficients of a-b) $r = 0$, c-d) $r = -0.8$ and e-f) $r = 0.8$.

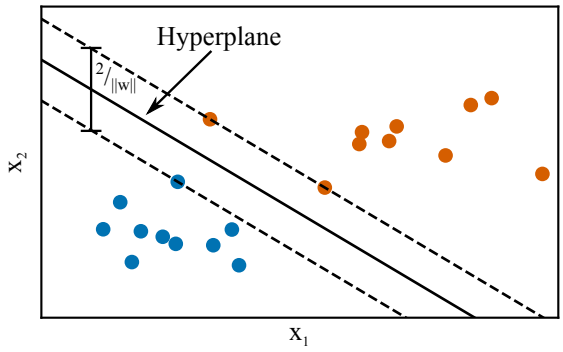


Figure 2.2: An example of support vector machine classification, and how the optimal hyperplane is chosen.

2.5.2 Support vector regression

Support vector machine for regression (SVR) [51] is a machine learning technique that, like linear regression, uses a set of predictor and target values to create a prediction model. In order to understand how SVR works, it is useful to begin with the more well-known support vector machine (SVM) [52] for classification.

An SVM classifier attempts to find the hyperplanes that maximizes the margins between two classes. Let the predictor values be given by $\vec{x}_i$ and the target values by y_i, where $i \in \{1,...,n\}$. Each y_i is either -1 or 1, corresponding to the two classes which the classifier tries to separate. Any hyperplane in the space spanned by $\vec{x}$ can be written as

$$\vec{w} \cdot \vec{x} - b = 0, \tag{2.11}$$

where $\vec{w}$ is the vector normal to the hyperplane.

For simplicity, let $\vec{x} = [x_1, x_2]$. If the data is linearly separable by the two classes, the situation can look like in Fig. 2.2. The classifier will find two parallel hyperplanes that separate the two classes so that the distance between the hyperplanes, the margin, is maximized. These two hyperplanes are given by

$$\vec{w} \cdot \vec{x}_i - b = \pm 1,$$

and the distance between them is $\frac{2}{\|w\|}$, where $\|w\|$ is the euclidean norm of $\vec{w}$. To prevent the hyperplanes from letting any data points be located between them, they are constrained by

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \quad \forall i. \tag{2.12}$$

With this constraint, the hyperplanes are found by maximizing $\frac{1}{\|w\|}$, or minimizing $\|w\|$.

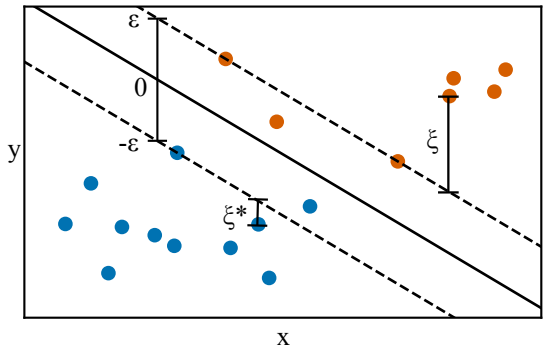


Figure 2.3: An example of support vector regression, the tolerance margin ε and the errors ξ and ξ^*.

This equation must be generalized for the classifier to be used when the classes are not linearly separable. This is done by adding the slack variable $\xi_i \geq 0$ to penalize data points that are located within the margin. The constraint in Eq. (2.12) then becomes

$$y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1 - \xi_i, \quad \forall i,$$

and the separating hyperplanes are found by minimizing

$$\frac{1}{2}\|w\| + C\sum_i \xi_i. \tag{2.13}$$

The value of C determines the level of trade-off between having a wider margin and keeping the vectors $\vec{x}_i$ outside of the margins.

In the case of SVR, the target values are continuous, and a different formulation is therefore required. The goal is now to find a linear function

$$\hat{y}(\vec{x}) = \vec{w} \cdot \vec{x} - b \tag{2.14}$$

that fits the set of predictor and target values well. The most common form of SVR uses is called ε-SVR, and uses the epsilon-insensitive loss function. This introduces a margin of tolerance ε which defines a width around the regression line within which a data point is considered to be correct, and have no error. A data point located outside of this margin is given an error $\xi_i \geq 0$ equal to the distance from the margin if its target value is above it, and $\xi_i^* \geq 0$ if it is below. Fig. 2.3 shows how these parameters relate to each other.

The minimization problem in Eq. (2.13) then becomes

$$\frac{1}{2}\|\vec{w}\| + C\sum_i (\xi_i + \xi_i^*), \tag{2.15}$$

subject to the conditions

$$\xi_i \geq y(\vec{x}_i) - (\hat{y}(\vec{x}_i) - \varepsilon),$$

$$\xi_i^* \geq -y(\vec{x}_i) + (\hat{y}(\vec{x}_i) - \varepsilon).$$

In Eq. (2.15), $\|\vec{w}\|$ is a measure of the flatness of $\hat{y}(\vec{x})$, and the value of C now determines the trade-off between having a flatter $\hat{y}$ and keeping the data points closer to the regression line (a flatter $\hat{y}$ is better as it makes the model less sensitive to outliers).

By using kernels to map the data points into a new feature space, the model can turn non-linear. In this paper, however, only the linear SVR has been used, and so non-linear kernels will not be discussed further.

2.6 Machine learning formalism

Feature scaling

In many classification and regression algorithms, including linear regression and SVR, the euclidean distance between two points is used as an error metric. If there are several feature vectors with different scales, for instance one with values between 0 and 1 and another with values up to a million, the calculated distance will be dominated by the latter. By scaling these features by normalization, each feature can be given equal importance by the algorithm.

There are several different ways in which the features can be normalized. They can be rescaled to within a set range, scaled to unit length or standardized by giving each feature a zero mean value and unit variance,

$$x' = \frac{x - \text{mean}(x)}{\text{std}(x)}. \tag{2.16}$$

Preventing overfitting

When training a machine learning model, one must be cautious not to overfit. Overfitting occurs when the model becomes so specialized at predicting the values for one set of data that it is no longer applicable to other sets. To prevent the results from being affected by this, the available data can be split into a training and test set. The model is trained and optimized using the training set and used to predict the values of the test set.

A sign of the model having been overfitted on the training data is that it has a much higher accuracy on the training set than on the test set, and will typically occur when the hyper-parameters of the model have been optimized on the training set. This can be prevented by splitting the training set into even smaller parts in a process called cross-validation.

Table 2.1: Confusion matrix.

		Predicted class	
		False	True
True class	False	True negative	False positive
	True	False negative	True positive

The model is then trained on one subset and tested on another (the validation set), and its ability to predict the validation set is used to optimize the hyper-parameters.

There are also here several different methods that can be used, the simplest of which would be to split the training set into one training and one validation set. Another is called k-fold cross-validation, in which the training set is split into k equal parts and each are in turn used as the validation set. The average score for all k validation sets can then be used to optimize the model.

Evaluation

The most intuitive metric for evaluating a classification method is its accuracy, i.e. the proportion of the data it manages to label correctly. While this can intuitively seem like a good metric, it has some major flaws. Say we want to predict whether a random salmon will have lice or not based on a set of features. There are then four prediction outcomes, as the salmon either does or does not have lice, and we predict that the salmon has or does not have lice. This can be summarized in a confusion matrix like Table 2.1.

The accuracy of a prediction only measures the proportion of the data points being classified as true negative or true positive. The issue with the accuracy arises if the amount of each class is disproportionate, for instance if we attempt to predict the lice on $n = 100$ salmon of which only five have lice. A model that predicts that only one of the five salmon have lice, while all others don't will have an accuracy of $^{96}/_{100} = 96\%$, although it does not appear to be very good for salmon with lice. A supplementary metric which highlights this inaccuracy is the precision and recall of the model. Let T_p be the number of true positives, F_p the number of false positives and F_n the number of false negatives. The precision is then defined as

$$\text{precision} = \frac{T_p}{T_p + F_p},$$

and the recall as

$$\text{recall} = \frac{T_p}{T_p + F_n}.$$

The precision and recall values of a classifier are unified in the F_1 score, which is the harmonic mean of the precision and recall,

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{2.17}$$

In this example, the precision answers the question "of all the salmon predicted to have lice, how many actually did have lice?", which in this case would be 100% as the one salmon predicted to have lice did have lice. The recall, however, answers "of all the the salmon that did have lice, how many did we predict to have lice?", which is only 20%. The precision and recall values shows how skewed the classifier is, and the F1 score of $0.2/1.2 = 0.17$ reflects this.

Chapter 3

Method

3.1 Gathering and pre-processing of data

This section outlines the methodology employed prior to the data analysis. The first two sections, Sections 3.1.1 and 3.1.2, describe the process of collecting data on the lice counts and biomass reported by the farms, respectively, Section 3.1.3 the gathering of environmental data and Section 3.1.4 how the seaways distances were calculated. First, however, an overview of the available resources for the model analysis is presented.

The available resources can be split into four categories. The first and most important category is the information on lice counts reported by salmon farms. The second category is data from the farms on the salmon it keeps, which include the amount and mean mass of the fish, and the treatments that are performed to decrease the numbers of lice. The third is the environmental conditions at the farm. The environmental conditions included in this work is the water temperature, salinity and current velocity at the farm, which are all known to affect the lice. In addition, the location of the farm and the time during which the cohort was active are included. The final category is information on the seaways distances between the farms.

3.1.1 Reported lice counts

All reported salmon lice data from all Norwegian salmon farms, continuously from week 1 of 2012, are made publicly available through BarentsWatch Fishhealth[1] in a cooperation between BarentsWatch and Mattilsynet (The Norwegian Food Safety Authority), who supplies the data. BarentsWatch provides an intuitive graphical user interface to easily see

[1] https://www.barentswatch.no/fiskehelse/

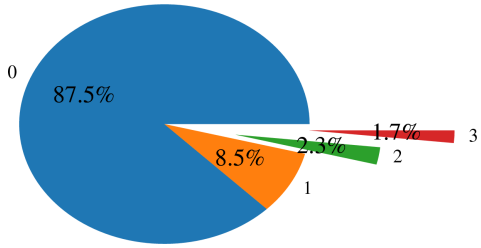


Figure 3.1: Distribution of weeks with consecutive missing lice counts.

statistics, and also a publicly available API through which much of the data used in this thesis was gathered.

The BarentsWatch API is a REST-ish API and uses JSON via HTTP. To gain access to the data, one first has to register as a user for the API. Then one can receive a token that enables one to send GET requests to download the data. The API provides several different endpoints, of which only the one providing the most detailed information was used for the data gathering. This endpoint gives detailed information about a particular salmon farm for one week. This information was downloaded for all 1.670 Norwegian salmon farms for each week between week 1 of 2012 and week 35 of 2017, totaling 296 weeks. After being downloaded to JSON-files, one file for each location, the Python library pandas [53] was used to concatenate the data into one two-dimensional tabular data structure (a pandas DataFrame).

The raw data downloaded from BarentsWatch required pre-processing before being analyzed. First, the locations that had never reported lice counts were excluded, leaving 1.021 farms. To make the data format more convenient, and to account for the difference in what date each week starts at for the different years, the reported information from each week was assigned to the Wednesday of that week, as was also done by Kristoffersen et.al. [32].

Due to regulations as well as possible holes in reporting, lice counts are missing for some of the weeks in which a farm is active. The amount and lengths of these holes in the subset of the data used for the analysis of the VI model are shown in Fig. 3.1. These missing lice counts were filled in using linear interpolation. The same data subset also had 12.7% farm weeks for which no temperature was reported, or the temperature was zero and assumed to be erroneous (the latter was the case for less than 0.1% of the farm weeks). These values were also filled in with linear interpolation. In the cases where linear interpolation was impossible due to no temperature reports for an extended period of time, the values were replaced by the modeled temperatures at 3 m depth (see Section 3.1.3).

3.1.2 Reported biomass data

To supplement the data from BarentsWatch, information on the average mass and the number of salmon at each farm was acquired from Fiskeridirektoratet (Norwegian Directorate of Fisheries). This data is not openly accessible, but can be released for research purposes.

According to the regulations specified in Section 2.2.2, the data consists of reports made by each farm at the end of every month. The data spanned from December 2011 until December 2016, as later reports will not be made available until the summer of 2018. The data contained reports from a total of 987 farms. Out of the 1.021 farms with lice count reports, 946 had available biomass data.

It is notable that there were farms with multiple reports in the same month (in many of these cases, the first report was likely erroneous as it had zero-values for both numbers and mass, even though the location was not fallow that month). In these cases, all but the last reported value for the given month were discarded. The reports were all assumed to be given for the last day of every month. In order to get the monthly biomass data on the same frequency as the weekly lice counts, it was grouped by farms and re-sampled from a monthly to a daily frequency using nearest-neighbor interpolation, so that the values assigned to each day were set equal to those of the preceding or following end-of-month values, depending on which is closest in time. Then, all but the values corresponding to a day with lice counts, i.e. every Wednesday, were kept while the rest were discarded. Values for when a location was reported to be fallow were set to zero.

This technique causes some errors in the transitions between being fallow and being active, resulting in a farm having no reported fish numbers for some weeks after going active or before going fallow. In the case of going active, this was remedied by setting the biomass values to the first following values, while for going fallow the values were set to the last occurring values.

3.1.3 Environmental data

The only environmental data included in the BarentsWatch data is the weekly measured sea temperature, but other environmental effects are also known to affect the lice infection pressure, as mentioned in Section 2.1.2. To investigate the impact of these factors, additional environmental data was gathered from Meteorologisk institutt (the Norwegian Meteorological Institute), who has weather and ocean forecasts openly available on the their THREDDS Data Server[2]. From the ROMS NorKyst800m coastal ocean forecasting system, daily average values with a grid resolution of 800 m for modeled water temperatures, horizontal currents and salinities from a 3 m depth were downloaded on the NetCDF format. The data is only available from the end of June 2012, so values corresponding to the first six months of the BarentsWatch data was not included.

The values for each of the farm locations were extracted from the NetCDF files using the kd-tree algorithm [54] from the Python SciPy library [55]. A kd-tree was constructed from the latitudes and longitudes of all the grid points, and the grid point closest to each of the farm locations was found using the *query*-function. If the closest point was masked, i.e. not located in the sea, the closest sea-point was chosen instead. The data from this one point was then extracted. To match the temporal resolution of the BarentsWatch data, these values were re-sampled to weekly frequencies. For every Wednesday, the average value

[2] `http://thredds.met.no/`

of the preceding week for each of the variables was calculated (for current velocities, the maximum value was calculated as well).

3.1.4 Seaways distances

As mentioned in Section 2.1.2, one of the factors known to affect the abundance of lice is the presence of neighboring salmon farms. To investigate this relation further, the distance to the farms have to be known. A straight-line distance between two farms is easily calculable, but is obviously not an accurate measure of the movement of lice. Therefore, the seaways distances between all farms were calculated.

To find the seaways distances between locations, accurate maps are necessary. For this purpose, depth maps for all regions that contain salmon farms were downloaded from Geonorge [56]. These maps cover the entire Norwegian coast on a 50 m grid, in Cartesian coordinates using the WGS84 coordinate system. The depth information was discarded, and the maps were stitched together and converted to a boolean matrix, distinguishing between land (boolean 1) and sea (boolean 0). In order to speed up computations, the map was re-binned to a 800 m grid, and an array element was denoted land if at least half of its 16 sub-elements were land tiles.

To find the distance between two farms, their geodetic (latitude and longitude) coordinates were converted to Cartesian coordinates, and their equivalent points were located in the boolean matrix. Then, the A* algorithm [57] was used to find the shortest path between the two points that only goes through the nodes denoting sea. A Python implementation of the algorithm [58] was used as the basis. To increase efficiency, the movement was limited to eight directions, and the square of the octile distance was used as a heuristic. Squaring the distance was done to reduce the computation time to a manageable timescale, but also causes the algorithm to no longer be guaranteed to find the shortest path. To limit this error, the algorithm was run twice for each combination, while alternating which points were used for start-points and which were goals. The shortest of the two distances were then chosen for each of the paths.

The maximum distance to calculate between two points was set to 100 km: The calculation was canceled if the path exceeded this limit. This is the limit used for calculating distances in VI's model [24, 59], and is also a sensible upper limit for how far lice can travel before needing to find a host. Using HI's hydrodynamical model, lice were released continuously in a fjord on the west coast and drifted with the currents. From the position of the lice after the model had run for 39 days, it was very unlikely that a louse had traveled more than 100 km in that time.

3.2 Data analysis

Due to the use of treatments and techniques to limit the abundance of lice in salmon farms, a direct time series comparison of modeled and counted lice counts will give dubious

results. Although treatments can be accounted for [31], the lack of data from the farms about the treatments and the limited knowledge of the effectiveness of treatments based on different parameters, will cause a high degree of uncertainty. It is therefore deemed beneficial to exclude weeks in which treatments have been performed. Due to uncertainties in when the treatment starts having effects on the lice abundance, and the duration of the effect, all weeks in a cohort after which a treatment has taken place are also excluded.

This leaves a subset of the original data, consisting of a number of active periods with various lengths for each farm. This data is further standardized by fixing the number of weeks for each of the periods: Periods that are below this limit are excluded, while periods longer than the limit are cut off at the limit. In addition, periods in which modeled values for any of the weeks are missing are discarded.

Two measures have been used to investigate how well the models predict the counted amounts of lice: The Pearson correlation given in Eq. (2.8) and the average positive and negative error in Eq. (2.9). Together, they can give a good picture of how the models performs in various conditions. The correlation gives information on how well the predicted time series follows the lice counts, while the errors show whether there is a systematic over- or under-prediction of the number of lice.

Due to the significant difference in available data for the VI and HI models, the analysis of the models has been performed twice. The pre-processing steps taken before the analyses are outlined in Section 3.2.1. In Section 3.2.2 and Section 3.2.3, the data used in each of the two analyses, hereby termed Analysis 1 and Analysis 2, are presented. Section 3.2.4 gives an overview of the investigated factors, and lastly, Section 3.2.5 contains an overview of the visualization techniques used when presenting the results.

3.2.1 Model pre-processing

Due to the different natures of the existing models, the pre-processing steps required for each of them were different. These steps are now outlined, and the model fitting process explained.

VI model

The statistical model constructed by VI gives a relative infection pressure of mobile lice per location for each week, from week 1 of 2012 until week 35 of 2017, and therefore required little pre-processing. Of the 1.021 farms with lice counts, the VI model had values for 988 of them. As done by Kristoffersen et. al. [32, 34], the natural logarithm of these values were taken before further calculations were performed. Any weeks with an infection pressure of zero were set to the minimum non-zero value in the dataset.

HI model

As previously mentioned, the HI model gives a density of chalimus lice per square meter which covers all Norwegian salmon farms with lice counts. Only model results from 2016 and 2017 were available for this thesis: For 2016, the model was run for 176 days between April 1 and September 24, and in 2017 for 237 days between March 5 and October 28. The model results were given in a set of NetCDF-files in which the copepodid density was saved with a temporal frequency ranging from 2 to 8 days.

In order to compare these values to the reported lice counts, a series of steps were taken: The density values were interpolated in time, the chalimus density for each farm was extracted, and the densities were integrated over a temperature-dependent number of days to calculate the number of infective copepodids in the water. After these steps, which will now be presented in more detail, the values were on the same format as the VI data and could be easily compared.

First, the density values were linearly interpolated to a daily frequency. To extract the chalimus density for each farm, a kd-tree were used in the same way as explained in Section 3.1.3. However, in this case certain precautions had to be taken. As discussed in Sandvik et.al. [5], the lice density is a patchy field and a point-to-point comparison of modeled and counted numbers of lice is not trivial. To remedy this, the mean and maximum densities of the 3×3 and 5×5 grid neighborhood of each farm was calculated, using the same kd-tree method. Any cell that had zero density (i.e. a land cell) was not included in the mean calculation.

Next, the copepodid densities were converted to densities of chalimus, in order to be comparable to the counted values. This conversion was performed by estimating what percentage copepodids will evolve into chalimus, and how many days it would take to reach that stage. This was done using the same parameters as was used in VI's model calculation, and is described in Section 2.3.1: The time delay was estimated to four days plus the number of days required to accumulate 155 °d. During this period, the lice were given a daily mortality of 5 %, as in the VI model [32]. The resulting densities were then summed over each week to give a weekly chalimus density for each farm.

Model fitting

Both of the models now give a weekly relative value for the number of new lice that will be present at a given farm. To simulate the accumulation of lice over time, these values are integrated over time, starting from the first week of each cohort. In this calculation, the daily mortality was set to 5 % (as in [43]) giving a weekly mortality of $1-(1-0.05)^7 = 30\,\%$. This is an estimation based on research on the mortality of lice of different stages [33].

To convert the modeled values to a measure of the amount of lice for each location week, the modeled values were fitted to the counts of PAAM and adult female lice. This was done independently for each analysis, using only the cohorts included in the analysis. This value was converted to integer values by multiplying by 30 and rounding to the nearest

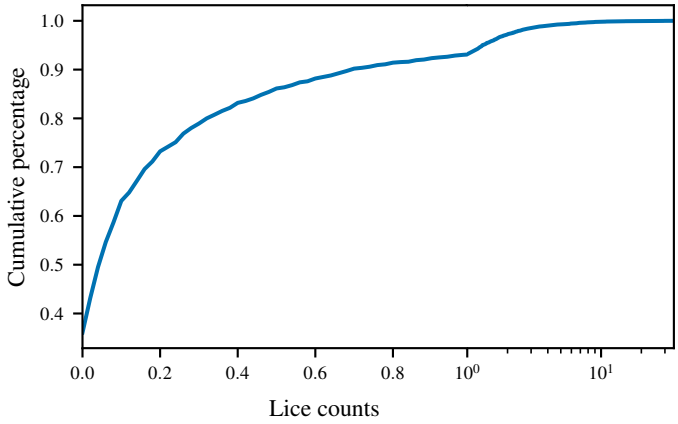


Figure 3.2: Cumulative percentage of lice counts per week for Analysis 1.

integer, as integer targets are required for the regression function. The regression was performed using the *zeroinfl*-function with logit-link from the *R*-package *pscl* [60], in which the zero-inflation was set to be dependent on the modeled values. This gave a vector of zero-probabilities $\vec{\Pi}$ and the regression parameters $\vec{\theta}$, which were used with Eq. (2.6) and Eq. (2.7) to compute the modeled lice counts. The correlation coefficients as well as positive and negative errors could then be calculated for each of the cohorts using Eq. (2.8) and Eq. (2.9), respectively.

3.2.2 Analysis 1: VI model

In the analysis of the VI model, the minimum number of weeks a cohort is required to have been active without any use of treatments was set to 19 weeks, giving a total of 639 cohorts. 31 of these cohorts reported no lice for the whole 19 weeks, and are excluded from the dataset as the Pearson correlation is not defined in this case. This gives a total of $608 \cdot 19 = 11,552$ data points from 409 different farms. With 608 cohorts, the results are expected to be statistically significant while still having sufficiently long time series for the Pearson correlation to be robust.

Fig. 3.2 shows how the counts of lice are distributed. 36.1 % of the counts are zero, and in 90 % of the cases there were less than 0.7 lice per salmon.

3.2.3 Analysis 2: Comparison of the VI and HI models

In Analysis 2, the HI and VI models are compared with each other. Due to data from the HI model only being available from 2016 and 2017, the number of cohorts are sharply reduced in this analysis. The number of weeks was therefore reduced to 17, which gave 101 cohorts, 91 of which had reported more than zero lice, from 90 different farms. Fig. 3.3

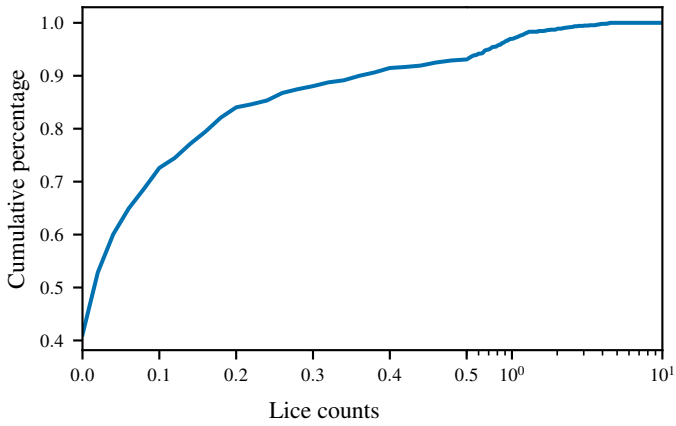


Figure 3.3: Cumulative percentage of lice counts per week for Analysis 2.

shows the distribution of lice for this data. There are now much fewer lice, with 41.1 % of the counts being zero and 90 % having less than 0.38 lice per salmon.

3.2.4 Investigated factors

A number of factors were investigated to see whether they affected the performance of the models. These factors are divided into four main groups: Time, location, internal and external factors. Internal factors include all conditions at the farm in question, while external factors are the conditions at neighboring farms.

Time

Many of the analyzed factors are expected to vary within a year (for example the temperature). There are also interannual variations, not only in the environmental parameters but also in e.g. the use of treatments and other farming techniques. Three measures of time were therefore looked into: The year and the month a cohort became active, and all the months during which the cohort was active. The last of these was included to smooth the data and perhaps get a better idea of the full impact of the time of year on the model. The correlation and errors of each cohort was in this case assigned to all the months during which it was active.

Location

The latitude and longitude of each farm was used to analyze the variation in model performance depending on the farm location. However, coordinates themselves is not enough to adequately describe differences in the geography of two locations. The correlations of each cohort were therefore plotted on a map using the python package *folium*, with each

Table 3.1: External factors.

• The number of active (non-fallow) farms
• The amount of CH lice
• The amount of PAAM lice
• The amount of AF lice
• The number of salmon
• The mass of the salmon
• The number of treatments

circle corresponding to a farm location and its color the correlation value. For farms with more than one active period and therefore more than one correlation value, the correlation value has been replaced with the average. A map for Analysis 2 was also created, in which the colors designate the difference between the correlation values of the VI and the HI models.

Internal factors

The internal factors can again be divided into two parts: Environmental and farming factors. The values have been aggregated over all the 17 or 19 weeks that are studied. Environmental factors include the mean water temperature of the cohort and the temperature's rate of change (ROC), the mean salinity and the mean and maximum current velocities. The ROC is calculated as the slope of a linear regression line on the temperature as a function of time. The farming factors consists of the median amounts and mass of the fish, as well as the mean amounts of counted lice of the different stages.

External factors

Using the seaways distances described in Section 3.1.4, each farm was linked to all others within a distance of 100 km. These linked farms have not been filtered in the way described in Section 3.2, so all available data is included. For each farm, the set of parameters given in Table 3.1 were then weighted based on the distance to them using (2.3), as this function is used in the VI model as a measure of the connectedness of two farms. The densities of lice have been calculated by multiplying the average counts by the number of salmon in the farm.

3.2.5 Visualization

The results are presented using a few different visualization techniques. The parameter values used for these plots are as follows:

Boxplot

A boxplot is a simple way to show the distribution of numerical data. The boxplots used in this thesis are called Tukey boxplots [61], in which the rectangle indicates the upper and lower quartiles of the data and the whiskers extend to the lowest data point within $1.5 \cdot \text{IQR}$ of the lower quartile, and the highest within $1.5 \cdot \text{IQR}$ of the upper quartile. The IQR is the interquartile range, and is equal to the difference between the upper and lower quartiles. Values outside these ranges are plotted as single points.

Violin plot

Violin plots are similar to box plots, except that they also show the probability density of the data. The probability density has been calculated with a kernel bandwidth of 0.2 [62], and the interior of the violins contain a boxplot computed with the same parameters as above.

Data binning

Many of the results presented in the next chapter have been binned along the x-axis. In order to exclude outlier values, a minimum amount of data points has been set for a bin to be included in the plot. This minimum amount is 10 for Analysis 1 in Section 4.1, and 5 for Analysis 2 in Section 4.2. The figures shows the median value for each of the bins, as well as the IQR (shaded region) and the range between the 10th and 90th percentiles (lighter shaded region).

3.3 Creating a new model

The VI model uses known statistical models and is built upon the foundations of earlier research in statistics and lice biology. To get an indication as to how the VI model performs against machine learning models, two models have been created using the gathered data. The first model is a multiple linear regression model, and the second uses support vector regression (SVR) with a linear kernel. Both models are trained to predict the total number of lice every week for each cohort.

For both models, the data has been pre-processed in a similar manner as the VI model, and cohorts were limited to 19 weeks as before. As all available data was necessary as inputs to the models, all cohorts active in 2017 and before June 2012 were excluded as no biomass or no environmental data was available. This left a total of 400 cohorts from 307 farms, a total of 7.600 data points.

As is standard procedure when training machine learning models, the samples have been split into two sets: A training set and a test set. The training set was set to all cohorts

Table 3.2: Parameters used for the regression models. Values in parentheses are how many values the parameter consists of. *Sine and cosine transform, **mean and maximum, †the current and past three weeks, ††the past three weeks.

Parameter	Model 1	Model 2
Month (2)*	X	X
Start year	X	X
Weeks active	X	X
Sea temperature	X	X
Latitude	X	X
Longitude	X	X
Current speed (2)**	X	X
Mean salinity	X	X
Temperature change	X	X
Density of active farms (4)†	X	X
Density of PAAM (4)†	X	X
Density of CH (4)†	X	X
Density of AF (4)†	X	X
Counted PAAM (3)††	X	
Counted CH (3)††	X	
Counted AF (3)††	X	

becoming active in 2012-2014, while the test set contains those that became active in 2015 and 2016. The test set consisted of 32.3% of all available data.

Before training the models, the predictor values of both the training and the test sets were normalized by the mean and standard deviation of the training set. The test set was not used for calculating these values to prevent any information about the test set having an impact on the training of the models.

In addition to using the parameters for a cohort from one week to predict the amount of lice the same week, certain of the parameters, namely the lice counts and the densities, have been included from each of the last three weeks. In order to have values available from the past three weeks, the time series of each of the cohorts were reduced to only include the last 16 weeks, leaving the first three to be used for training the models, reducing the total number of data points to 6.400. Table 3.2 shows which parameters were included in the two models. The amount and average mass of the fish were not included because the large uncertainty in these values are believed to have had a large effect on the regression.

Time data was included in the model through two predictors: The time of year and the month. While including the year is straight-forward, the month is more challenging due to its cyclical nature. This parameter was therefore split into two separate parameters prior to the normalization by taking the sine and the cosine transform of the month number.

Each of the models were trained twice based on two groups of input parameters. The first group contained all available information, while the second did not contain any information about the amounts of lice in the farm during the previous weeks. The second group is

thus expected to give poorer results than the first, but will be more comparable to the VI model and also more applicable to predicting the lice impact on wild salmonids, for which a manual counting of lice in the preceding weeks is not possible.

When including the lice counts (in case of group 1) and densities for the last three weeks as parameters, the number of parameters grows quickly, which is likely to cause the models to over-fit. In addition, not all the parameters are expected to be good (linear) predictors of the lice counts. Therefore, to generalize the models and increase their accuracies, both models have been optimized with respect to the number of input parameters. The input parameter optimization was done by maximizing the prediction score as a function of the input parameters. The prediction score was defined as the average score of a 5-fold cross validation on the training data, with the score calculated as the mean absolute value of the differences between predictions and target values.

The parameter optimization was done by calculating the mean score from the cross-validation when including only one parameter. The parameter giving the highest score was then added to the model. This process was then repeated, and a new parameter was added to the model if it improved the score by at least a threshold amount, here set to 0.1%. If no single parameter being added would increase the score, the procedure was repeated but now while trying to add combinations of two parameters. If the score was increased above the threshold, the process continued with trying to add a single parameter again. If it did not, the process was ended.

In the case of SVR, the score for each set of input parameters was set to the maximum score after doing a grid search on the values of the hyper-parameters C and ε in Eq. (2.15). The optimal value of C was searched for among the seven numbers evenly spaced on the log scale between 10^{-5} and 10^{2}, and ε among the five numbers on the log-scale between 10^{-6} and 10^{-1}. After the optimum amount of parameters with an optimal value for the hyper-parameters were found, the hyper-parameters were fine-tuned. In the end, each of the models were used to predict lice values for the test set.

3.3.1 Evaluating the model

In addition to the correlation and errors, the accuracy and F_1-score in Eq. (2.17) was used to evaluate and compare the regression models. To do so, the actual and predicted lice counts were binned in order to turn the regression results into a classification. The bins and the number of data points falling into each of them are shown in Table 3.3. The calculation of the F_1-score in Section 2.6 is strictly speaking only applicable to a binary classification problem, but was straightforwardly generalized by calculating the score for each class separately, i.e. by treating the classification of each class as a binary classification.

3.3.2 Comparison with the VI model

To ensure a fair comparison between the two models, the zero-inflated negative binomial fitting of the VI data was done again, and including only the data points that were used in

Table 3.3: The bins used for the calculation of the F_1-scores, and the number of data points within each bin.

Number of lice	Number of data points
0-0.1	1319
0.1-0.2	280
0.2-0.5	224
0.5-1	133
1-2	67
2-5	41

the regression models (both training and test sets). All 19 weeks in each cohort were used to calculate the growth and mortality of lice in the VI model, but only the last 16 weeks were included in the correlation and error calculations.

Chapter 4

Results

The results will be presented in the same order the methods were described. The results from the analysis of the VI model are presented in Section 4.1, the results from the comparison with the HI model in Section 4.2 and lastly the results from the new regression models in Section 4.3. Most of the results presented in this chapter are of the same format, in which the median values of the positive and negative errors, the correlation coefficient, the lice counts, and the number of cohorts are all displayed in a single figure (for Analysis 2, the lice counts are not displayed). The data has been binned in the way explained in Section 3.2.5. Table 4.1 shows the zero-probabilities and regression parameters for Analysis 1 and 2, as well as for the comparison of the VI model with the regression models.

4.1 Analysis 1: VI model

The results from the analysis of the VI model will be presented in the order of the factors outlined in Section 3.2.4: Time, location, internal factors and external factors. As not all results are deemed equally important, only some are included in this section, while the remainder can be seen in Appendix A.

Table 4.1: ZINB regression parameters for the analyses. θ are the regression parameters and π the probability of a zero, with the subscripts 0 and 1 denoting intercepts and coefficients, respectively. Numbers in parentheses denote which analysis they correspond to.

Model	θ_0	θ_1	π_0	π_1
VI (1)	-0.4917	0.1891	1.306	-0.2337
VI (2)	-1.725	0.2135	6.807	-0.6566
HI (2)	0.2426	0.1050	-14.83	0.1413
VI (3)	0.9488	0.06049	3.883	-0.4062

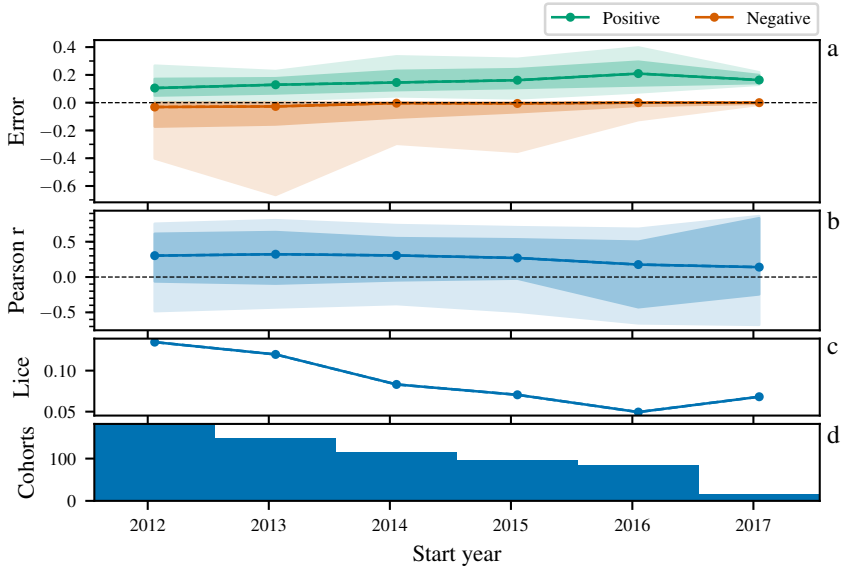


Figure 4.1: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the year in which a cohort became active. The shaded areas show the middle 50th and 80th percentiles.

4.1.1 Time

Fig. 4.1 shows the yearly variability for the model. It is interesting to see that the error in Fig. 4.1a has shifted from being more or less equally negative and positive in 2012, to more and more positive errors as time passes. More light can be shed on this relationship by looking at the steady decline in lice over the years. The model appears to fail to incorporate this declining trend, as it underestimates the lice in 2012 and 2013, when the amounts are higher, and overestimates in 2016 and 2017. The correlation in Fig. 4.1b is high and quite stable for cohorts starting from 2012 to 2015, but then starts declining and is considerably lower in 2017. As Fig. 4.1d shows, there is a much smaller number of cohorts in 2017 so the values for this year are less statistically significant. However, it does looks to fit well in with the downward trend since 2015.

The inter-annual trend in model performance is even higher, as Fig. 4.2 clearly shows. The model performs substantially better on cohorts becoming active from April to July, with both low error and high correlations. After July, the correlation proceeds to decrease and even drops to negative values between September and December.

A different view of the impact of the time of year on the model can be seen in Fig. 4.3, in which the correlations and errors of each cohort have been assigned to all months during which the cohort was active. A clear cyclical trend is now visible for both the error and the correlation. Cohorts that were active between May and September have substantially fewer

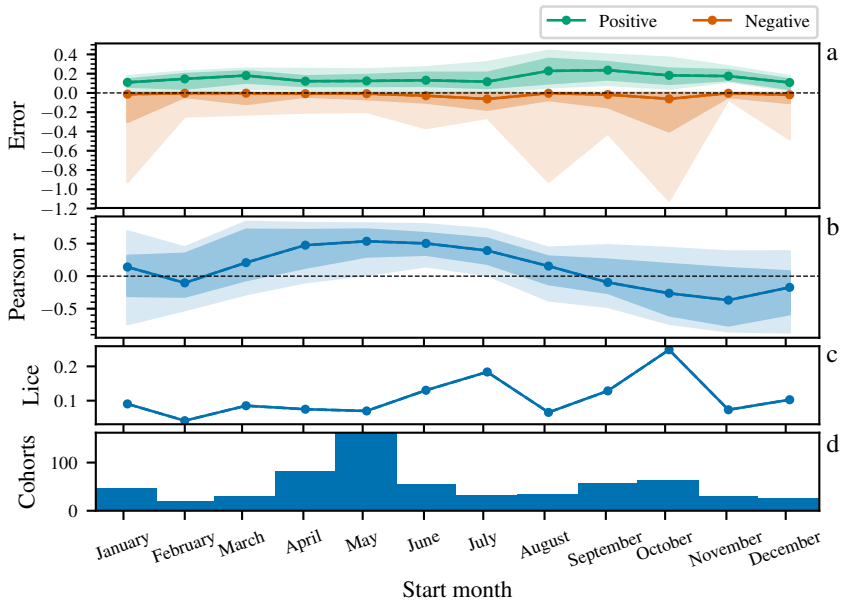


Figure 4.2: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the month in which a cohort became active. The shaded areas show the middle 50th and 80th percentiles.

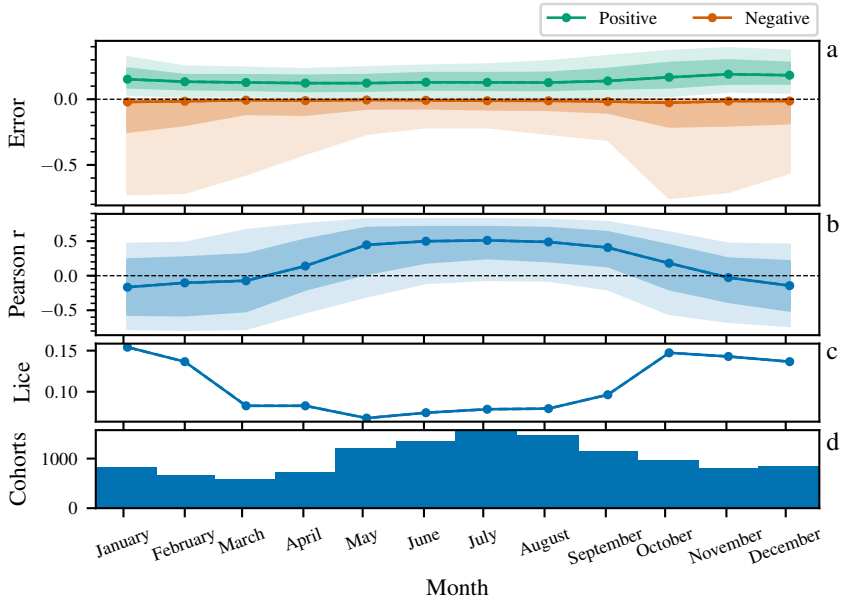


Figure 4.3: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of all the months during which a cohort was active. The shaded areas show the middle 50th and 80th percentiles.

lice, and the model has a higher correlation and lower error than the remaining months. The winter months give the worst results, with the model giving high errors and negative correlation with counted values.

4.1.2 Location

Fig. 4.4 shows the results as a function of the latitude of each farm. From Fig. 4.4a it is obvious that the error is much smaller at latitudes above about 65° than otherwise. Fig. 4.4c explains this as the amount of lice at higher latitudes is significantly lower than the rest. In the lower latitudes, however, there is no clear linear trend in whether the model predicts too many or too few lice, with the 3rd quartiles extending to about 0.2-0.3 for both negative and positive errors.

By comparing Fig. 4.4b to Fig. 4.4a, it is seen that the model performs well at latitudes of 63° to 65° and 67° to 71°, albeit with a significant error in the former range. In the lower latitudes, the correlation drops even to negative values. Fig. 4.4d shows that there are many cohorts both at the lowest and highest latitudes, so this difference is clearly supported by the data. At 62° to 63° there seems to be a pocket in which the model quite consistently overestimates the number of lice.

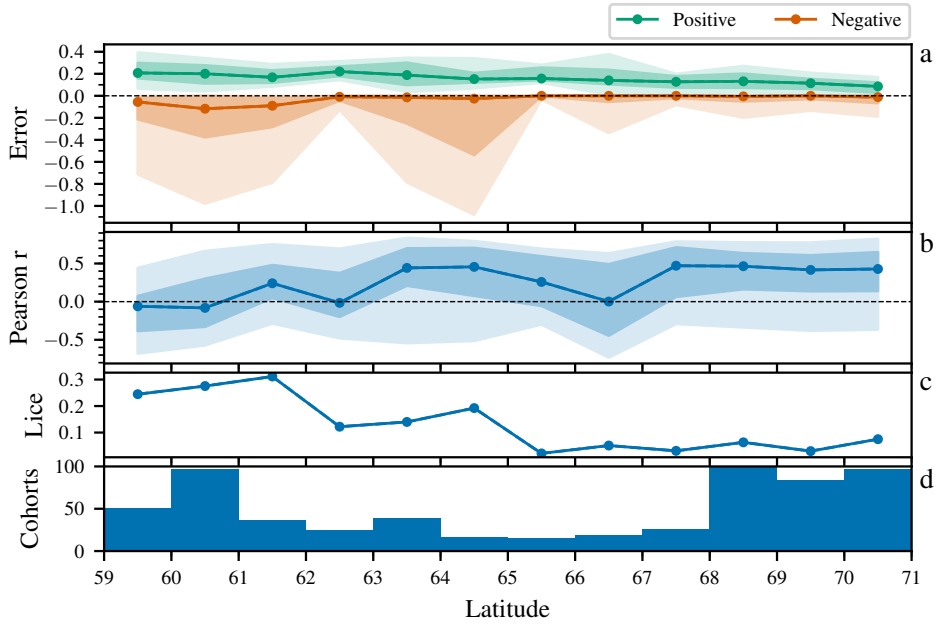
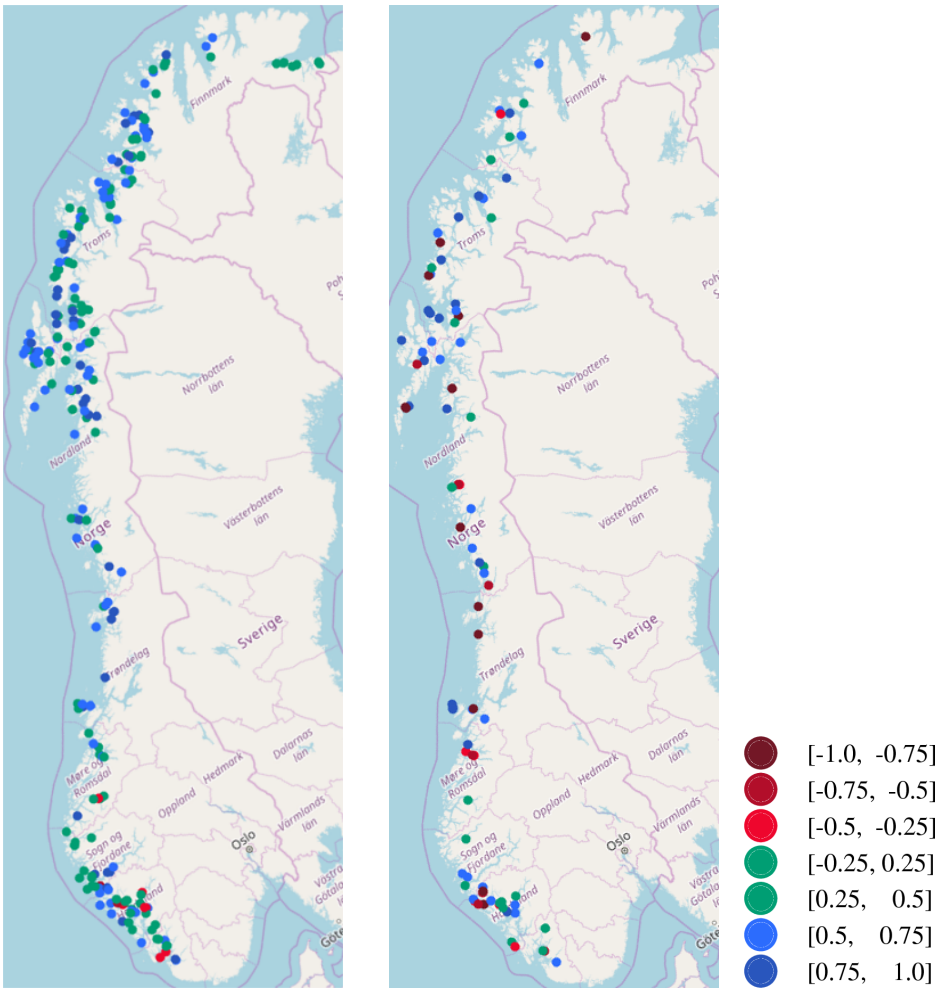


Figure 4.4: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the latitude of the farms. The shaded areas show the middle 50th and 80th percentiles.



(a) **(b)** **(c)**

Figure 4.5: Maps showing the correlation values for the VI model for cohorts becoming active in (a) April-July and (b) August-March. (c) shows the legend.

The maps in Fig. 4.5 shows the correlation of each farm, as explained in Section 3.2.4. As the results in the map are not binned like those in Fig. 4.4, the large variation in time will have a large impact on the result. To rectify this, the cohorts were split into those becoming active between April and July in Fig. 4.5a, and those becoming active at all other times in Fig. 4.5b, an approximately 50/50 split. Both maps contain several clusters of farm with similar correlation values. Fig. 4.5a in particular has many clusters with high correlation values in the northern half of the country, and many of these are located along the coastline and not in the fjords. In the southern parts, the correlations are generally worse and there seems to be less cohesion among the farms located close to one another. In Fig. 4.5b the correlation values are overall lower and there are more clusters of farms with low correlations. The farms located between Møre og Romsdal and Trøndelag are the exception, with decent correlation in the summer cohorts and very high correlation the rest of the year.

4.1.3 Internal factors

The impact of internal factors on the correlation will now be presented: Environmental ones, the statistics of the biomass of the salmon, and the amount of counted lice.

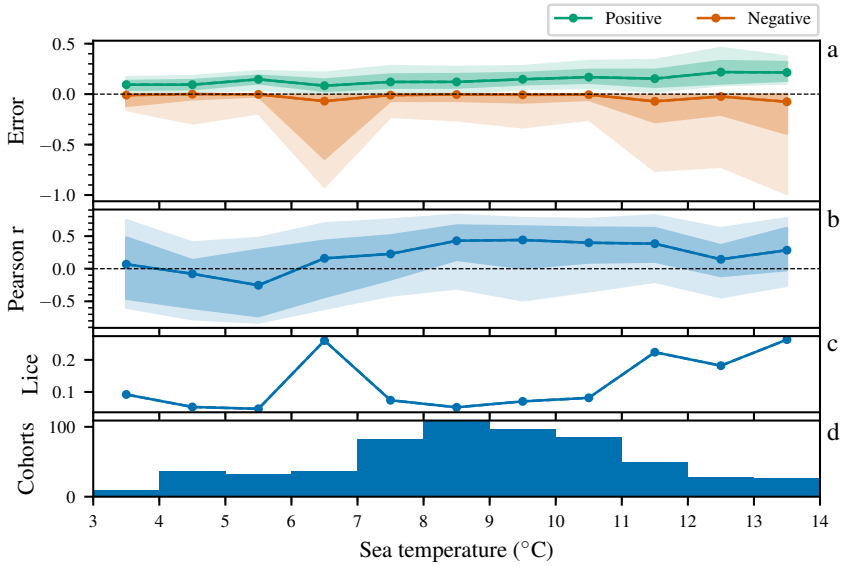


Figure 4.6: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the median sea temperature at each farm. The shaded areas show the middle 50th and 80th percentiles.

Among the environmental factors, the sea temperature is perhaps the most important as it plays a vital role in the calculation of the predicted amounts of lice. Figs. 4.6 and 4.7 shows the impact of the mean sea temperature and the rate of which the temperature changes for

each cohort. From Fig. 4.6a,b, the optimal temperature appears to be in the range 8 °C to 11 °C, where the correlation is high and the error low. Although the higher error at 6 °C to 7 °C can be explained by the up-tick in lice at this temperature, the correlation decreases as the temperature gets lower, implying a decrease in model performance. Conversely, at higher temperatures, the error increases to higher levels than at 7 °C although the lice counts at higher temperatures is similar. The correlation stays quite high, however, indicating that the model performs quite well at these temperatures despite the increase in lice.

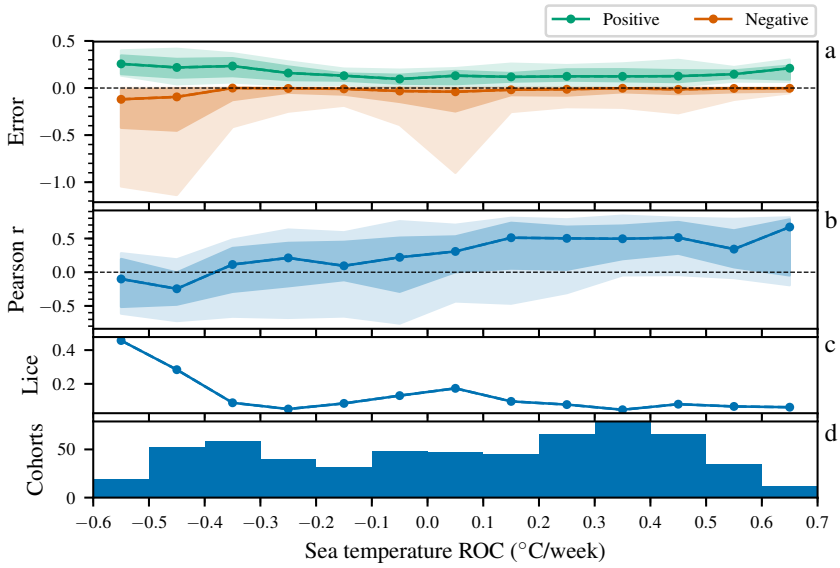


Figure 4.7: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the rate of change (ROC) of the median sea temperature at each farm. The shaded areas show the middle 50th and 80th percentiles.

The rate of change of the temperature has a perhaps even greater impact on the model performance. Fig. 4.7b clearly shows that the model performs best when the sea temperature increases by 0.1 °C/week to 0.7 °C/week, and very bad when it decreases rapidly. At a weekly temperature change of −0.5 °C to −0.4 °C, the median correlation is even significantly negative. It is also very interesting to note the sharp increase in lice counts at negative temperature slopes. Again, however, Fig. 4.7a is ambiguous as to whether the model over- or underestimates the amount of lice here.

The remaining two environmental factors are the water current and the salinity. Fig. 4.8 shows how the maximum daily current velocity affects the model performance. The error increases slightly with an increase in current velocity, but the correlation shows that the model is significantly better for lower current velocities. Due to the ambiguity in Fig. 4.8a, however, it is not clear whether an increase in current velocity causes the model to predict

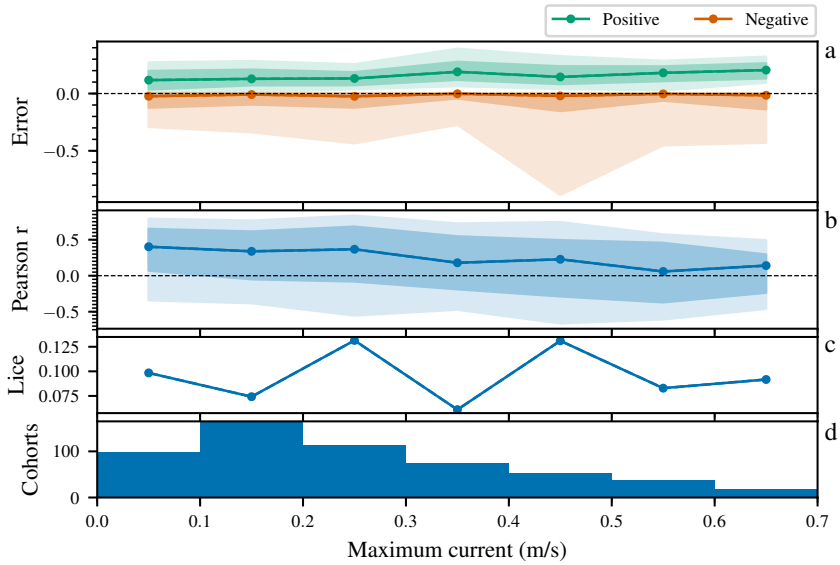


Figure 4.8: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the maximum current velocity at each farm. The shaded areas show the middle 50th and 80th percentiles.

fewer or more lice. Both the average current velocity and the salinity, however, show weaker trends in both correlation and error, and are therefore not discussed here.

The last category of internal factors is the farm-related factors. In Fig. 4.9, the impact of the average number of counted PAAM lice on the model is shown. One might expect that as the number of counted lice increases, so would the prediction error. Fig. 4.9a shows an overall increasing negative error, while the positive error remains quite stable. However, the correlation in Fig. 4.9b is quite stable. It therefore appears that the number of counted lice does not have a significant effect on the model's ability to predict the trends in counted lice, but for cohorts with an average amount higher than 0.3, the negative error is higher than the positive and the model consistently underestimates the lice abundances. There are two more metrics for lice counts, namely the counts of CH and AF lice. These show the same trends as Fig. 4.9, and are not discussed here.

The number of salmon in a farm shows a surprisingly large impact on the error in the model, as Fig. 4.10a shows. The cohorts with a median amount of salmon less than 100.000 has a 75th percentile negative error of 1.5, and a 90th percentile of 4.0, although the median error is very low. The error decreases quickly as the number of fish increases. The correlation in Fig. 4.10b also follows the same pattern, with a lower correlation for the first few bins, after which it is quite stable. The high error could indicate that cohorts with few salmon have a lower average amount of lice than others, but this is disproved by Fig. 4.10c. The model must be concluded to drastically underestimate the amount of lice in cohorts with few salmon.

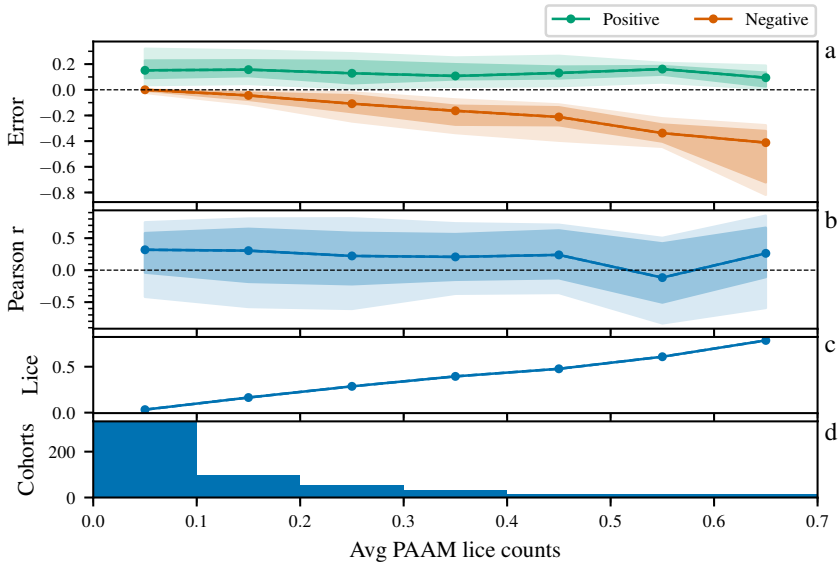


Figure 4.9: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the average PAAM lice counts at each farm. The shaded areas show the middle 50th and 80th percentiles.

The weight, and therefore also the size of the fish, appears to contain little valuable information. As most cohorts only contains fish with low weights during the first 19 weeks after fallowing, the distribution is highly skewed.

4.1.4 External factors

The external factors and how the density around each farm is calculated is explained in Section 3.2.4. Fig. 4.11 shows that the error increases and the correlation mostly decreases as the density of active farms surrounding a farm increases. Also, Fig. 4.11c shows that there is an almost linear relationship between farm density and lice counts. The density of the total number of fish and of the total mass of fish show very similar results to Fig. 4.11, and are therefore not discussed her.

The density of lice around a farm is expected to show similar results as the density of active farms. Overall, the results for all the stages of lice show similar trends, so only the results for PAAM lice in Fig. 4.12 are included here. Similarly to Fig. 4.11, the error increases as the lice density increases, and the correlation slightly decreases. However, while the positive error in Fig. 4.11a is constant and very low, in Fig. 4.12a, it increases slightly with an increasing density of PAAM lice, while the negative error does not . There is also here an almost linear relationship to the lice counts in the farm.

The last external factor is the density of treatments, which did not give much information,

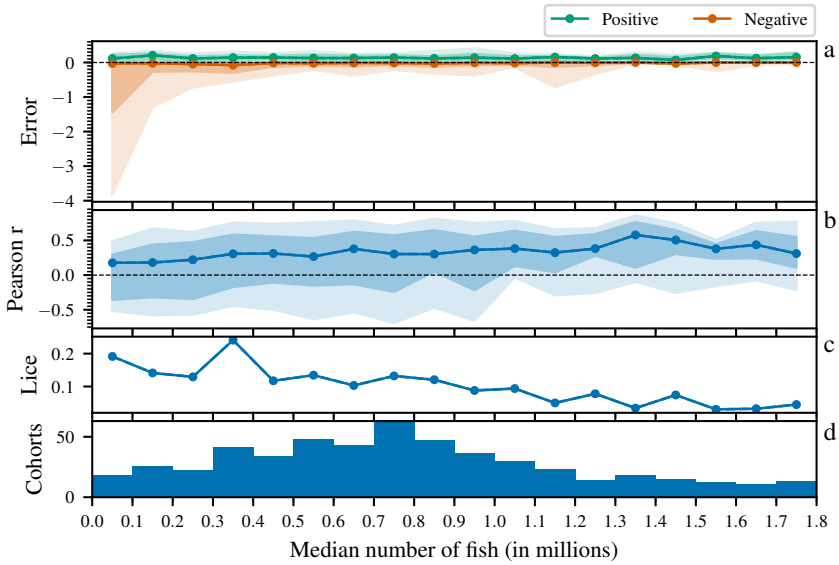


Figure 4.10: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the median amount of fish at each farm. The shaded areas show the middle 50th and 80th percentiles.

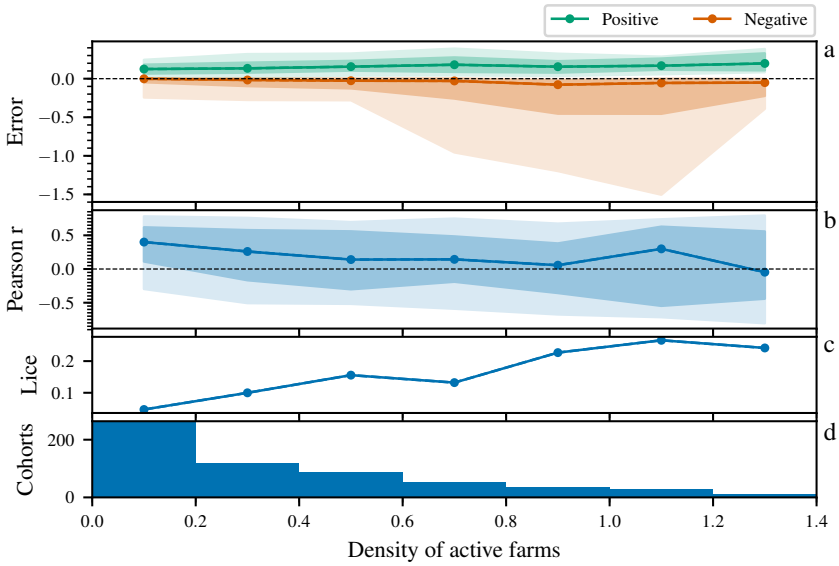


Figure 4.11: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the density of active farms around each farm. The shaded areas show the middle 50th and 80th percentiles.

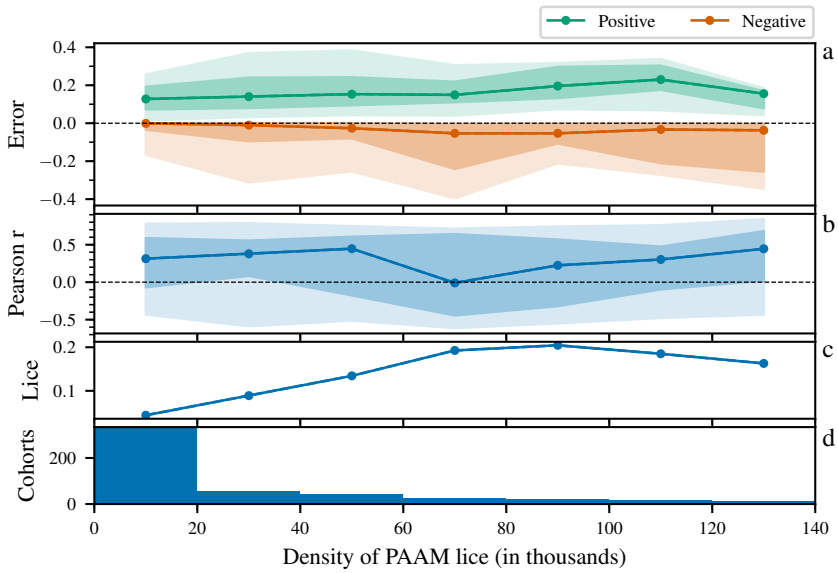


Figure 4.12: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the density of PAAM lice around each farm. The shaded areas show the middle 50th and 80th percentiles.

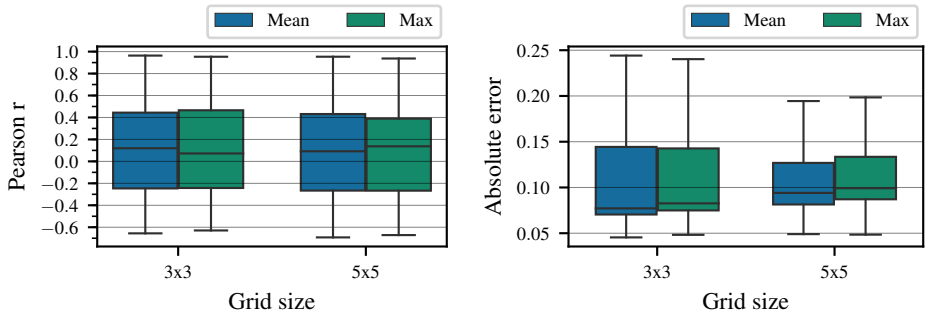


Figure 4.13: The (a) correlation with counted values and (b) absolute error for the four HI models.

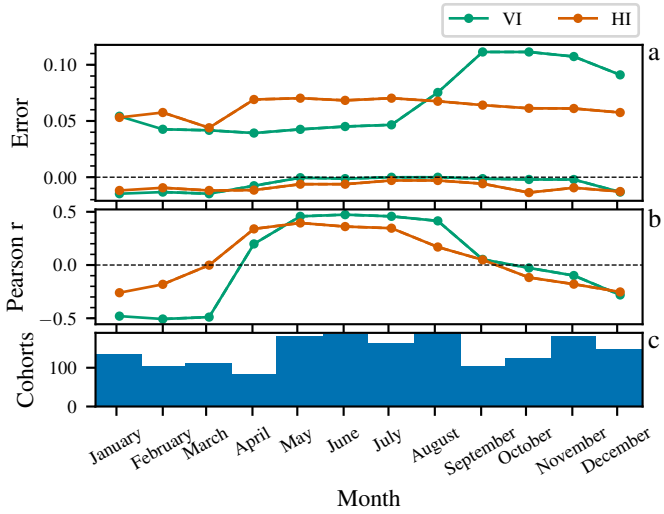


Figure 4.14: The median a) positive and negative errors and b) correlation values for the VI model, and the total amount of cohorts, as a function of all the months during which a cohort was active.

mostly due to the distribution being highly skewed towards no treatments.

4.2 Analysis 2: Comparison of the VI and HI models

After following the pre-processing steps outlined in Section 3.2.1, the HI model gave four different sets of lice predictions. Fig. 4.13a shows a comparison of the four sets with respect to their correlations with the lice counts, while Fig. 4.13b shows their absolute errors. In the latter figure, outlier values are not displayed because they are many orders of magnitudes higher than the median values. The maximum absolute errors range from $1.7 \cdot 10^2$ for the mean value in a 3×3 grid to $3.6 \cdot 10^3$ for the maximum value in a 5×5 grid. The model using the maximum amount of predicted lice counts in a 5×5 grid around the farm gave the highest median correlation value but also the highest median absolute error. The second highest correlation value is given by the mean number of lice counts in a $3x3$ grid, which also has the lowest median absolute error. Only the latter model has been compared against the VI model.

The rest of this section will follow the same structure as Section 4.1, but due to the limited amount of data available for the HI model, less of the factors show significant enough results to be presented here. All results not shown here can be seen in Appendix B.

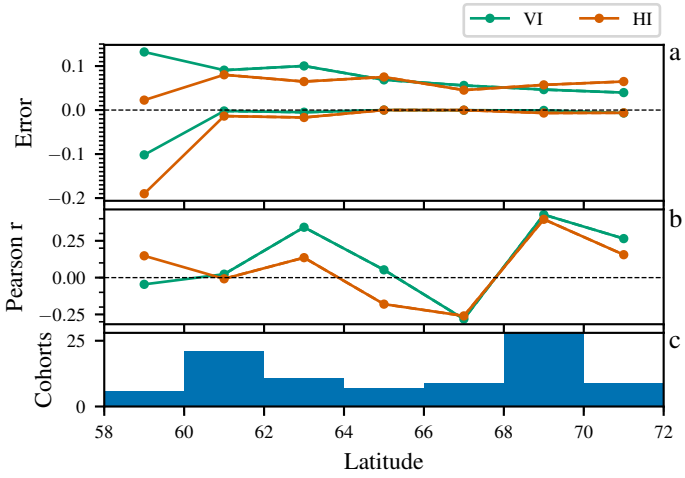


Figure 4.15: The median a) positive and negative errors and b) correlation values for the VI model, and the c) total amount of cohorts, as a function of the latitude of each farm.

4.2.1 Time

The year in which the cohort began makes little difference for the models. They both perform better in 2017 than 2016, and the VI model has the higher correlation in 2016, while the HI model is better in 2017. However, the differences are very small. The results for the start month of the cohorts are also ambiguous due to the lack of data, but looking at all the months during which a cohort was active gives very telling results. Fig. 4.14 shows that the models follow a similar trajectory through the year. They are both best between April and August, and have around zero or negative correlation values from October through March. From the error plot in Fig. 4.14a, the HI model has a more consistent error, while the VI model highly overestimates the amount of lice between September and January. From the available data, it thus seems that the HI model is superior from January to April, the VI model is better from May to August, and that they are quite equal between September and December.

4.2.2 Location

A comparison of the models with respect to the latitude of the farms show results that will be shown to be indicative of many of the factors: In terms of the correlation in Fig. 4.15b, the HI model is generally worse or as good as the VI model, expect for in some subset. In this case, the models are approximately equally good (or bad) at latitudes above 66°, while the VI model is better between 60° to 66° and the HI model better between 58° to 60°. The error is also quite similar across the range expect for in the lowest latitudes, where the VI model has a rather large but symmetric error, while the HI model quite consistently

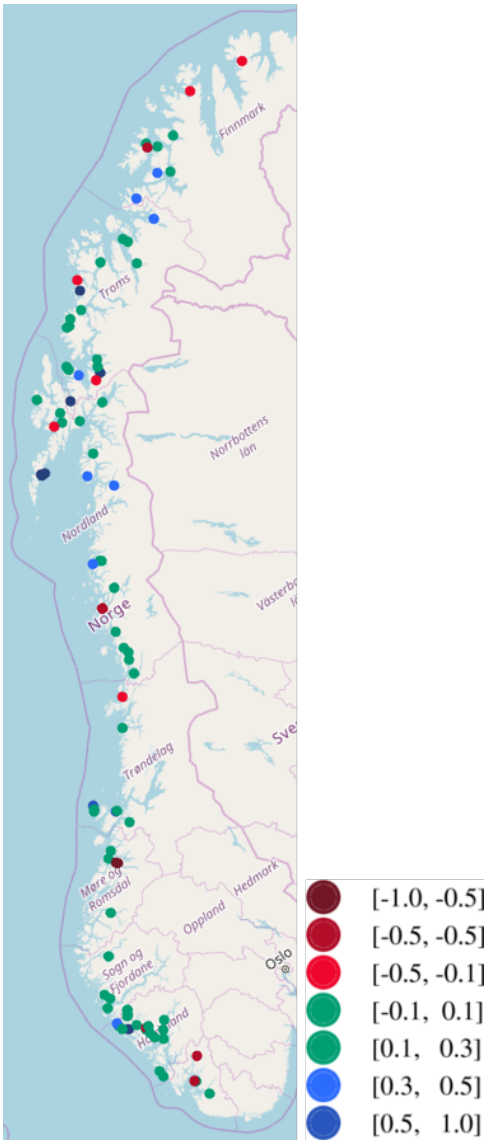


Figure 4.16: Map showing the difference in correlation values between the VI and HI models, $r = r_{\mathrm{VI}} - r_{\mathrm{HI}}$.

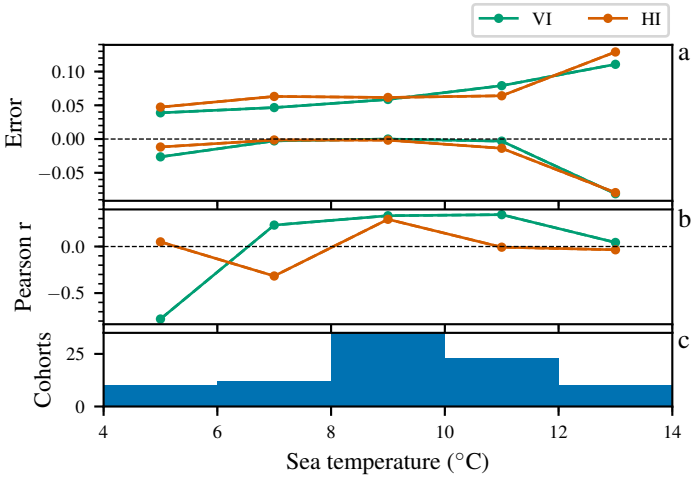


Figure 4.17: The median a) positive and negative errors and b) correlation values for the VI model, and the d) total amount of cohorts, as a function of the median sea temperature of each farm.

predicts too few lice.

Fig. 4.16 shows a map of the differences in correlation values between the VI and HI models. It can appear as though the VI model tends to be better in the fjords, while the locations for which the HI model is better are more often located along the coast. However, the map clearly shows the limited amount of data available for this comparison, and it is difficult to draw conclusions from this.

4.2.3 Internal factors

The results for the sea temperature in Fig. 4.17b shows that the VI model is consistently better than the HI model at temperatures above 6 °C. The errors in Fig. 4.17a shows that the errors of the HI model is much more consistent than the VI model, for which the positive error increases rapidly above 9 °C. Another observation is that the HI model appears to be vastly better than the VI model at the temperatures below 6 °C.

The rate of change of the temperature in Fig. 4.18 shows similar results. The VI model is significantly better than the HI model when the temperature increases, while at decreasing temperatures the results are less clear, albeit with both models having zero or negative correlations. However, when the ROC is sufficiently negative, the correlation of the HI model shoots up. The errors in Fig. 4.18a show that the errors follow each other quite closely, but with the VI model having a twice as high error as the HI model when the ROC is below -0.25 °C/week.

These results show that the models are approximately equally good at temperatures around 8 °C to 10 °C, and the VI model better at other temperatures and when the temperature is

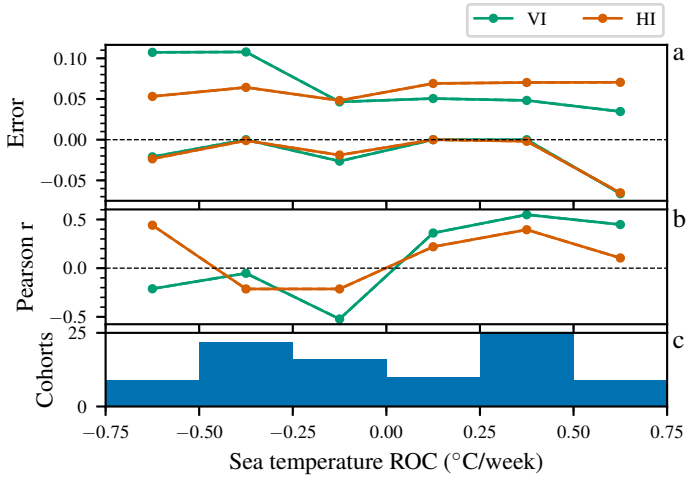


Figure 4.18: The median a) positive and negative errors and b) correlation values for the VI model, and the c) total amount of cohorts, as a function of the rate of change (ROC) of the sea temperature at each farm.

rising, except at temperatures below 6 °C and when the temperature decreases quickly.

The maximum current in Fig. 4.19 show a rather large disparity between the two models' correlation. First of all, the HI model is much better at locations with a slow current. The current is an integral part of the HI model, and is also indirectly important when the density of lice is estimated from a grid around the farm location, as the spread of lice around it is likely very dependent on the current velocity. At velocities above 0.2 m/s, however, the VI model is mostly much better.

As for the salinity in Fig. 4.20, which is also a part of the HI model, we also see a large difference between the models at the lower values. For cohorts experiencing salinity levels of 30 ppt to 31 ppt, the HI model has a very high median correlation while it is negative for the VI model. At higher salinities, the HI model has a mostly decreasing correlation, while the VI model does not show any significant pattern.

4.2.4 External factors

It is challenging to interpret the results from the external factors due to the limited amount of cohorts, giving a very limited variation in the parameter values. The densities of the weights and number of salmon show that both models are best when there are small densities of mass and numbers, and mostly get worse as the densities increase. This is also the case for the densities of PAAM, AF and CH lice, all for which the models have similar trends, but with the HI model having a worse correlation. The density of active farms, though, does gives some interpretable results. The perhaps most interesting thing to note is

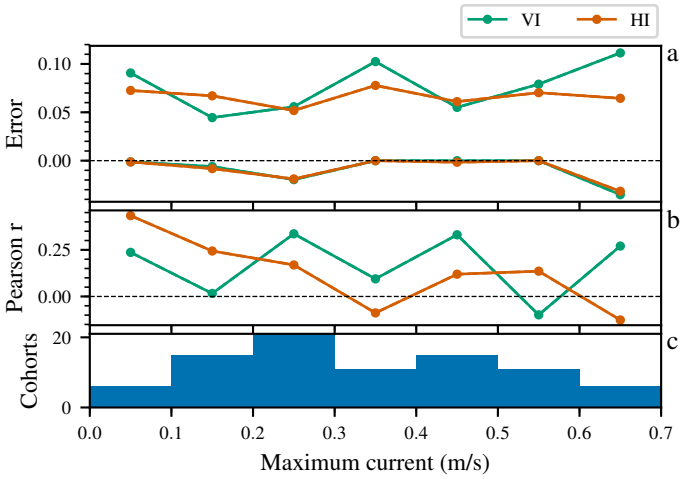


Figure 4.19: The median a) positive and negative errors and b) correlation values for the VI model, and the c) total amount of cohorts, as a function of the maximum current velocity at each farm.

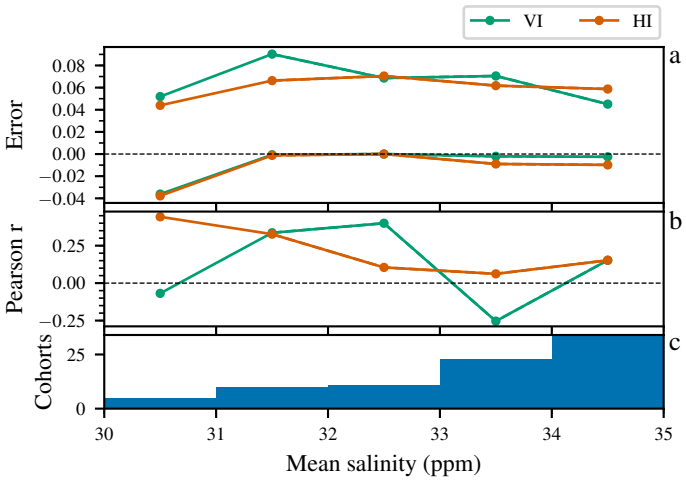


Figure 4.20: The median a) positive and negative errors and b) correlation values for the VI model, and the c) total amount of cohorts, as a function of the mean salinity at each farm.

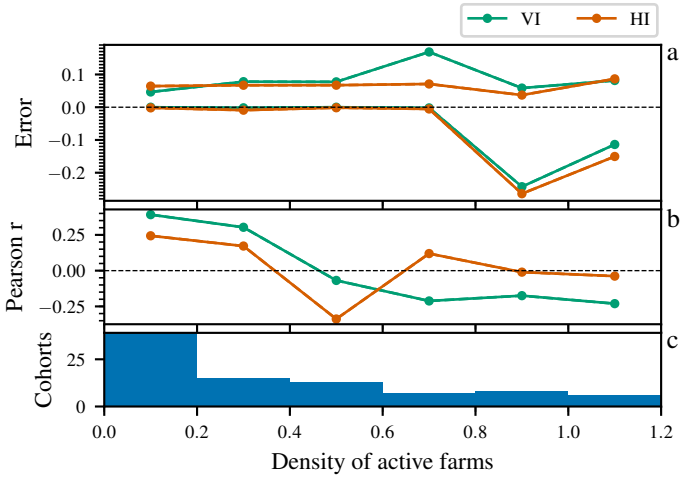


Figure 4.21: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the density of active farms around each farm.

that in Fig. 4.21b, the models' correlations are not very correlated to each other. Whereas the VI model becomes gradually worse as the density of active farms increases, the HI model does not have the same linear tendency.

4.3 Regression models

Table 4.2 and Section 4.3 shows which parameters were used in each of the models, how much the addition of each parameter improved the model score, and the parameter coefficients for the models using the first and the second sets of parameters, respectively. Table 4.2 shows that if the lice counts in previous weeks are available, the models primarily uses the amount of counted PAAM lice three weeks earlier and adult female lice two weeks earlier to predict the lice abundances. The SVR model complements these values with densities of lice from the same week, two and three weeks earlier. If the counts of lice from previous weeks are not included as parameters, Section 4.3 shows that the density of PAAM lice the same week is the most important.

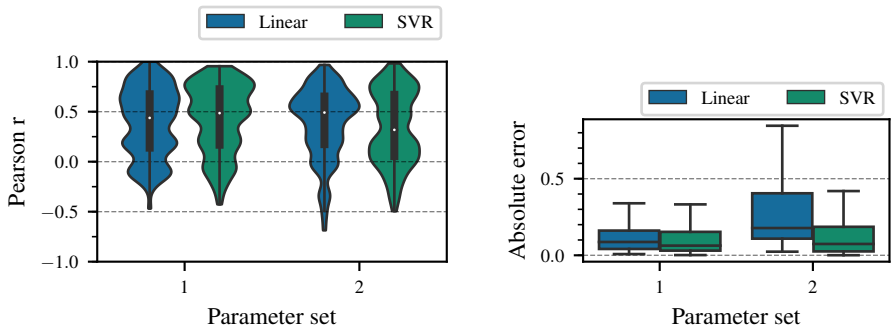
Fig. 4.22 shows how the correlation values and absolute errors compare for the four models. As expected, the models using the first set of parameters give overall better correlations than those using the second, among which the SVR model is quite superior. For the second set of parameters, linear regression gives a significantly higher median correlation compared to SVR, albeit also with a much higher error. In Fig. 4.23 the SVR model using the first parameter set (hereby named the SVR 1-model) and the linear regression for the second (Linear 2-model) are compared against the VI model.

Table 4.2: The order in which parameters were added to the regression models using the first set of parameters, how much they increased the score by and the coefficients for the final model. A ρ indicates density, and a superscript denotes that the value is from an earlier week.

Linear			SVR		
Parameter	Increase	coef	Parameter	Increase	Coefficient
Intercept		0.384	Intercept		0.328
$PAAM^3$		0.724	$PAAM^3$		0.773
CH^3	4.9%	0.226	AF^2	2.8%	0.178
AF^2	3.7%	0.374	CH^3	0.62%	0.0786
Latitude	0.85%	-0.032	$PAAM_\rho$	0.29%	0.0883
AF^3	0.34%	-0.219	$PAAM^2_\rho$	0.32%	0.0294
			CH_ρ	0.63%	$4.23 \cdot 10^{-3}$
			AF^3_ρ		0.0788

Table 4.3: The order in which parameters were added to the regression models using the second set of parameters, how much they increased the score by and the coefficients for the final model. A ρ indicates density, and a superscript denotes that the value is from an earlier week.

Linear			SVR		
Parameter	Increase	coef	Parameter	Increase	Coefficient
Intercept		0.384	Intercept		0.143
$PAAM_\rho$		0.516	$PAAM_\rho$		0.207
Latitude	2.0%	-0.119	Weeks active	0.73%	0.0143
Weeks active	0.74%	0.0555	CH^3_ρ	0.42%	0.0568
Current velocity (max)	0.47%	0.0368	$PAAM^2_\rho$	0.26%	0.0709
Salinity (mean)	0.12%	0.0111	Latitude	0.11%	$-1.73 \cdot 10^{-3}$
			Year		-0.0146
			Month (sin)	0.18%	$4.96 \cdot 10^{-3}$
			$Farms^3_\rho$		$-9.74 \cdot 10^{-3}$



(a) **(b)**

Figure 4.22: The (a) correlation values and (b) absolute errors for the four regression models.

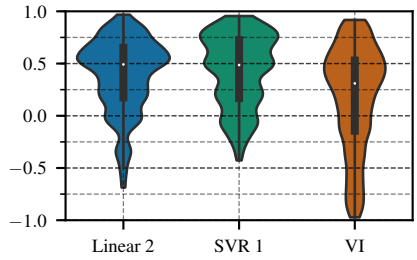


Figure 4.23: Comparison of the distribution of correlation values between the SVR 1-, Linear 2- and VI models.

Table 4.4: Classification scores for the four regression models and the VI model on the test dataset.

		F_1 score				
		Parameters 1		Parameters 2		
Lice range	n	Linear	SVR	Linear	SVR	VI
0-0.1	1319	84.3	87.6	44.0	82.0	25.7
0.1-0.2	280	33.6	33.7	13.5	19.7	13.2
0.2-0.5	224	39.5	38.6	17.8	23.0	23.8
0.5-1	133	38.4	35.8	22.7	11.9	18.9
1-2	67	31.8	39.7	9.4	9.5	-
2-5	41	43.0	52.0	3.1	0.0	-
Mean		45.1	47.9	18.4	24.4	13.6
Weighted mean		67.1	69.3	33.7	58.7	22.0
Accuracy		65.3	68.4	28.2	63.9	21.7

The F_1-scores for each of the models as well as the VI model are given in Table 4.4. Although all the models, except for the Linear 1, are quite adept at predicting whether or not there will be more than 0.1 lice, classification into other bins are not equally good. In particular, Linear 2 and SVR 2 have very low scores for the two highest bins. The VI model has lower mean and weighted mean scores than the other models. Furthermore, it did not predict more than one lice for any of the cohorts at any week, and the scores for these bins are therefore undefined.

It seems obvious that the regression models in general outperform the VI model on this data. To gain some more intuition into why that is, look at Fig. 4.24 which shows how the models vary with the months during which each cohort was active, equivalently to Fig. 4.3 on page 36. In Fig. 4.24 we see that the regression models does not suffer the fall in correlation values after September as is the case with the VI model, but has quite consistently high correlations until December. The SVR 1 model even performs best during the months in which the VI model is worst. In the VI model's best months, the three models give very similar results.

Fig. 4.22b showed that the Linear 2-model had the highest median absolute error, but Fig. 4.24a shows that it is still overall lower than that of the VI model. The SVR 1-model,

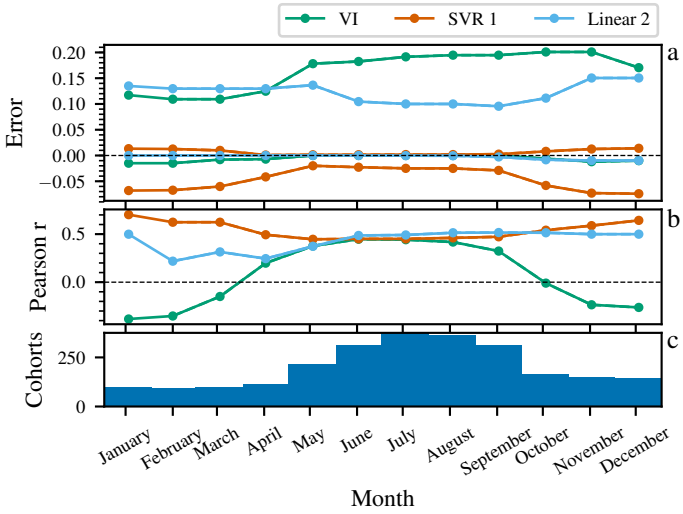


Figure 4.24: The median a) positive and negative errors and b) correlation values for the SVR 1-, Linear 2-, and VI models, and the c) total amount of cohorts, as a function of all the months during which each farm was active.

on the other hand, has a higher negative error but also a much lower positive error, than the VI model.

As the Linear 2-model is the most comparable to the VI model based on the parameters available for the modeling, it is shown in Fig. 4.25 with percentiles included. Compared to the equivalent figure for the VI model, Fig. 4.3 on page 36, the spread in errors are quite similar, except that the negative error of the regression model does not increase sharply until January, while this happens already in October for the VI model. Fig. 4.25b also shows that there is a large variability in the correlation values from October to December, and even though the median value is high, the 10th percentile extends down to around zero.

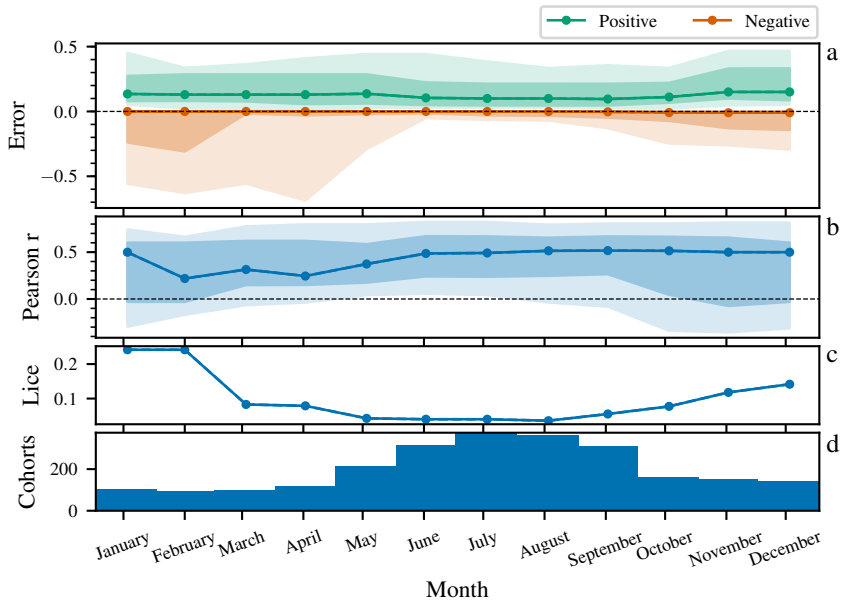


Figure 4.25: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of all the months during which each cohort was active. The shaded areas show the middle 50th and 80th percentiles.

Chapter 5

Discussion

The largest difficulty with analyzing these models is that there is a wide range of uncertainties and assumptions made at all stages in both the creation, processing and analysis of the models. In Section 5.1, the known sources of inaccuracies will be presented, and their impact on the model discussed. First, the reported values of lice counts, biomass statistics and sea temperatures from the farm will be discussed. The density calculations and the environmental factors will follow. Lastly treatments and use of other techniques that might have impacted the analysis will be substantiated. In the next section, Section 5.2, the uncertainties in the three models, the VI, HI and the regression model, are discussed. The results from the analysis are discussed in Section 5.3.

5.1 Uncertainties

5.1.1 Reported numbers

The most important numbers for the models discussed in this thesis are the amount of counted lice at each farm. It is therefore very unfortunate that there is a big uncertainty in these numbers. As previously mentioned, it is very difficult to get an accurate count of chalimus. Due to their larger sizes, PAAM and adult female lice do not suffer from the same difficulties when it comes to counting. However, the lice on only 10 or 20 salmon are taken to be representative for a whole cohort of up to several million fish. This is one of the biggest challenges for using counts of salmon lice to create or validate models.

Furthermore, the lice are only required to be counted once every week or two weeks, which combined with the small sample size gives patchy time series. In addition, the day which the lice were counted is not reported. Assigning all counts to one day, as was done in this thesis, will cause an error which is impossible to quantify without more data. In

periods with high amounts of lice, the lice count is sometimes conducted more frequently than what the government requires, as the farmers want better control of the lice impact on their cohorts. These counts are also not reported, and are thus unavailable for research such as this. As Fig. 3.1 (p. 20) shows, 12.5% of the lice values used in Analysis 1 were interpolated, and as many as 1.7% interpolated over three weeks. This can be a large source of inaccuracy, but due to the inherent inaccuracy in the lice counts it is difficult to estimate just how large impact this interpolation has.

The salmon farms are also required to report the amount and weight of the salmon on a monthly basis, which were interpolated to weekly values in this work. The amount of salmon is very important for the models. The infrequent reports can therefore be a large source of uncertainty, especially when salmon are added to the cohort which can cause big spikes in the numbers, without the reported data reflecting when the salmon were added. For the analysis of the VI model, this uncertainty is less important as only the median values from the 19 or 17 weeks were used. Whether the median values is a good measure for the mean mass and number of salmon in a cohort is however very uncertain. One salmon farm will typically have several separate salmon cages, and the number and mean mass of the salmon in each of them can differ. However, the reports of biomass as well as of lice counts from the farms only contain the average values from all the cages combined. This might be a reason why the model performance appeared to not be influenced by these statistics.

Finally, the reported sea temperature, which is used for calculating the growth of the lice at the farms, is only measured at 3 m depth, and like the lice counts, at an unknown day in the week. The linear interpolation between the gaps in reports is not expected to cause a large error as there are not large fluctuations in the water temperature, as the case was for lice counts, but the replacement with modeled temperatures might have. This point will be touched upon again in Section 5.1.3.

5.1.2 Density calculation

The calculation of the seaways distances between farms is flawed for several reason. First of all, the grid resolution of 800 m is expected to have had a larger impact on farms that are located close to one another. As the distance between two farms is only calculated as the distance between the grids each of them are located within, the calculated distances will have errors in the order of several hundred meters, which can have a large impact if the distance between two farms is within a few kilometers. For the longer distances, this error is no longer as important, but the effect of using the cartesian distance between the points as well as the possibility that the path selected is not the shortest path due to the inaccuracy in the A^* algorithm used will have a larger impact. As all distances are weighted with the exponentially decreasing distance function in Eq. (2.3), however, the resulting error at longer distances is less important. For the same reason, the cut-off at 100 km is not considered a cause of error as the distance function is so close to zero at this value as to make no difference.

The binning of the values in the presentation of the results will also have made the errors

less important. Nevertheless, a more accurate distance metric would have been preferable. If enough time and computational power is available, the original A* algorithm can be used to loop through all the farm combinations on the original 50 m grid maps. If not, algorithm optimizations such as jump-point search [63] can be used to speed up the algorithm while still obtaining an optimal solution.

5.1.3 Environmental factors

The accuracy of the investigated modeled environmental factors (temperatures, salinities and currents) are not expected to have had a large impact on this analysis, as only the mean or maximum values over one week have been used. The modeled current flow is however the backbone of the HI model. A test of the model was conducted in Hardangerfjorden in 2009 [17], where a buoy was used to measure the currents at 12 m depth. A time series comparison reportedly showed a good correlation between modeled and actual currents.

Inaccuracies in the modeled temperature could have an impact on the model in the cases where measured temperatures were missing and could not be filled through linear interpolation. The median absolute difference between measured and modeled temperatures is 1.2 °C. This is likely to a large degree caused by the temperature measurements being done on one day, while the modeled temperatures are the mean values for the whole week. Additionally, inaccuracies in the model, its grid size of 800×800 m and potentially also errors in the measurements will also have contributed.

For all the environmental factors, only the values in the grid in which the farm is located have been used for the analysis. Using for instance the mean value of a larger grid around the farm might in some cases have given more accurate results, especially for the current as it can be highly variable even within smaller areas due to geographical factors. Additionally, all the modeled values were taken from a 3 m depth. However, as mentioned in Section 2.1.1, the louse has been shown to move in the water column to avoid low salinity levels. At the locations with a relatively lower salinity at 3 m, it might therefore have been more prudent to use environmental values from a depth in which the lice are more likely to reside.

As mentioned in Section 5.1.3, the current velocity is believed to have two distinct effects on the lice abundance at a farm: First, a stronger current makes it more difficult for lice to attach itself to a host, and second, it can be indicative of a higher amount of lice being transported to the farm from neighboring farms. The current at the location itself is considered sufficient to show whether the models are affected by the first of these effects. For the second, though, a much more informative metric would be the current flow from all other farms to this one. An attempt was made to include this metric in the analysis by using results from SINTEF's 3D hydrodynamical model SINMOD [64] for the flow of currents between salmon farms in the central parts of Norway [65]. However, this approach was not pursued further due to the limited amount of farms included in the model, and the even smaller subset which were also among the ones used in the analysis.

5.1.4 Treatments and other techniques

A range of different treatments are used to manage the lice population at a farm, and while they are necessary they make the model evaluation process more difficult. Farms are required to report what week and which kind of treatment they use, but it is not known whether the treatment that week was performed before or after the lice counting. In addition, the effect and duration of the different kinds of treatments are unknown.

If a farm has deployed cleaner fish (fish that feed on the lice), the total number of fish introduced to the cages are known, but no further information about them are reported. Their population in the proceeding weeks is therefore not known. In addition, farms might use different types of cleaner fish and feed them differently, both of which will affect how much lice the fish can eat. The combination of these factors makes it difficult to model the impact of using cleaner fish, although attempts have been made [31].

In addition to the treatments that are reported every week, a farm might utilize other methods to get ride of lice or to keep them from accessing the cages where the salmon resides. Over the last few years, several new techniques have been introduced, for example lice skirts [66] and snorkel cages [67]. The farms are not required to report their use of these kinds of techniques, and no information about it has been available for this study. A farm which has success with using such a technique would likely have to use less treatments, and therefore have a higher chance of being part of this study, which could cause outliers and skew the results.

5.2 Models

5.2.1 Model data

As both data on the usage of treatments and the effects of different kinds of treatment to reduce the numbers of salmon lice is very sparse, only cohorts in which such treatments had not taken place were included in the analysis. In addition to severely reducing the available data, it also only chooses a subset of the cohorts in which the amount of lice is relatively low: The median value of mobile lice in the first 19 weeks of cohorts is 3 times higher than if only cohorts with no treatments are selected. This is likely due to farms only use treatments when there is a considerable amount of lice.

Both the VI and HI models use the counts of adult female lice as a basis for their models. This number is highly volatile as the number of adult female lice in a cohort is generally quite low, and the report is often zero. Of all the data reported since 2012, about 25% of the reports with a non-zero number of PAAM lice had zero adult female lice. As the number of adult female lice is multiplied with the amount of salmon to calculate the number of eggs they hatch, a relatively small error in the count can have a large impact on the model predictions.

5.2.2 VI model

The VI model is quite simple compared to hydrodynamic models such as the HI one, which has both positive and negative consequences. The simplicity leads to the model having relatively few inputs: From the salmon farms, it requires the counts of salmon and adult female lice and the water temperature, and certain biological parameters such as the fertility and mortality of lice at different stages. This makes the model very easy to run, which in turn can be used to tweak the parameters to improve the model performance. A disadvantage is that the model does not take other parameters into account, most importantly the current between farms. For example, there is a coastal current going up the Norwegian coast. It is therefore very difficult if not impossible for lice to be transported against this current, but this is neglected by the model, which only looks at the distance between farms.

The fecundity, growth rate, mortality and stage durations were all estimated from research on the biology of salmon lice and synthesized and modeled by Stien et.al. [33]. The parameter values are sometimes quite different depending on the source, and all have non-negligibly wide confidence intervals. Moreover, the values do not take into account mortality due to e.g. predators or higher mortality rates at lower salinities, and only depends on the temperature.

5.2.3 HI model

The most important part of the HI model is the simulated currents. As mentioned in Section 5.1.3, the modeled current is believed to be quite accurate, but due to the patchy distribution of particles transported in the ocean [68], the model has a high variation in both time and space [5]. In addition, due to the uncertainty in the number of lice in each farm, when they are counted and when eggs are hatched, as well as the modeled currents, the currents which the lice encounter might be quite different than what is modeled. For this reason, the high spatial and temporal accuracy of the simulations might have been unnecessary.

5.2.4 Pre-processing of model data

Zero-inflated negative binomial model was then fitted to the data to match the infection pressure (for the VI model) and the lice density (for the HI model) to the lice counts. For the data used in the first analysis, 39% of the lice counts were zero, so a zero-inflation was likely necessary to account for these values. The lice counts entering into the model were not pre-processed in any way, so outlier counts can have skewed the model. Removing outliers or smoothing the time series before the model fitting might have given more realistic data, and thus also improved the models. Of course, what values one determine as being outliers, or the amount of smoothing deemed necessary, is highly subjective, and such practices can give less reliable results. This has therefore not been conducted in this research.

The fitting and subsequent analysis of the models was done with the average number of all PAAM and adult female lice. The mixing of the different stages and of male and female lice makes it difficult to estimate the mortality of the lice. As mentioned earlier, the value of the mortality rate is not easily measured even for a single stage or sex, making it even more difficult when they are combined. However, neither of the models incorporate the growth of the lice beyond the pre-adult stage (the HI model only up to the copepodid stage).

The mortality is highly dependent on the stage and sex of a louse: For instance, one experiment found a mortality of 2% per day for adult female lice and 6% per day for adult male lice [69]. As model data on the ratio of the stages and sex of the lice in a cohort was unavailable, the mortality rate was set to a flat 5% per day, equal to the mortality rate of chalimus which both models use. An optimization to find the best fitting mortality rate was not performed.

The usage of counts of both PAAM and AF as true values must also be commented. It would have been much better to use only the number of chalimus, as this would require a lesser dependence on accurate biological parameters for the growth and mortality until reaching the PAAM stage. As mentioned previously, however, the high difficulty of accurately counting these lice, leading to highly uncertain numbers, makes this difficult. The second best option would be to have counts of pre-adult males and females, which would still reduce the uncertain parameters somewhat. However, these are grouped together with adult male lice in the reports from the salmon farms, and counting them separately would require an extra effort from the farmers.

5.2.5 Regression models

In an attempt to prevent overfitting in the regression models, the feature selection process was designed to return a minimal amount of features required to get a good fit. There are likely several other combinations of features whose score was not tested that would have given better results. Testing all combinations would indeed be beyond the scope of this research, as there are about 10^{12} possible combinations for the first set of parameters. One way to circumvent this problem would be to do a more thorough analysis on the effects of the different parameters on the lice numbers and select only a portion to include as features in the regression. Furthermore, during the hyper-parameter optimization in the SVR, a more thorough search for the best parameter values during the features selection process might have changed which features were selected in the end.

The split of the training and test sets by the year in which the cohorts became active likely had a negative effect on the resulting accuracy of the prediction. Fig. 4.1 (p. 34) shows that the average weekly counts of lice in each cohort steadily decreased from 2012 to 2016. This difference between training and test data could have been avoided by randomly dividing the data into the two sets, but was not done here to ensure comprehensibility of the results. However, this also indicates a future research direction to create a more generalizable prediction.

In this work, only regression with linear kernels was used, in order to make the results

easier to interpret and to provide a basis for future work in this field. If the goal is to provide the best possible prediction for lice abundances, a next step could be to utilize non-linear regression SVR with for instance a polynomial or radial basis function kernel.

Another way to potentially improve the accuracies of the models is to include every cohort for as long as they are active or until treatments were performed, instead of setting the fixed limit to 19 weeks. This would drastically increase the available training data and likely improved the model. Furthermore, instead of excluding cohorts after treatments are performed, the usage of treatments could be included in the model and further increased the available data.

5.3 Discussion of results

5.3.1 Time

The apparent decrease in correlation since 2015 might well be caused by an influx of new technologies and treatments from 2016. The observation that the error becomes less negative and more positive over the years lends credence to this hypothesis, as use of technologies would decrease the amounts of lice. The VI model does not take this into account, which leads to a poor prediction and an overestimation of lice abundances.

The high inter-annual variability in the model is undoubtedly connected to the water temperature and its rate of change, as these change with the seasons too. The higher correlation for cohorts becoming active in the spring and summer must be seen together with the higher correlation when the median temperature is 8 °C to 12 °C and the water is heating. An investigation into the respective effects of the time of year and water temperature on the model performance has not been performed.

5.3.2 Location

While the location of a farm does appear to have a large impact on the VI model's prediction strength, the results do not clearly show the reasons why, or how the model can be changed to reflect it. The observation that the model performs well at latitudes above 67° does however provide a starting point for a future undertaking regarding the spatial dependency of the model.

5.3.3 Internal factors

Temperature

The observation that the VI model appears to give the best results at temperatures around 8 °C to 11 °C is not all surprising if one looks at the sources for the biological parameter values the model uses. Stien et.al. [33] uses values found from several studies to estimate

these parameters, and studies have determined the development time at different louse stages with a range of different water temperatures. For the egg and planktonic stages, it ranges from 2 °C to 19 °C, which covers the water temperatures that are included in Fig. 4.6 (p. 39). However, the equivalent range for the infective stages is limited to 6.5 °C to 15 °C, and Stien et.al. notes that there have been no studies investigating the development times of parasitic stages at temperature below 7 °C over a sufficient time period. This lack of studies may explain why the model performs so much worse at lower water temperatures, as the parameters it uses might not be valid at these temperatures.

Another difficulty for achieving correct predictions at low temperatures is that the development time of the lice stages becomes very long, which will cause errors in the mortality or growth rates to have a much larger effect than at higher temperatures. This hypothesis is supported by the error plot in Fig. 4.6a, which shows a dominant negative error at the lowest temperatures, which indicates that at least for these temperatures, the mortality might be set too low or the growth too slow. It must also be noted that in Fig. 4.17 (p. 48), the HI model appears to give much better results at low water temperatures. The sample size at these temperatures is however so low that one should be careful to draw any conclusions from this. The reason for the high dependence on the slope of the temperature change in Fig. 4.7 (p. 40), and why the correlation is higher when the temperature increases, is unknown.

Currents

The almost linear slope of the correlation in Fig. 4.8 (p. 41) is perhaps not surprising, as the current directly affects the louse's ability to attach itself to a host, as mentioned in Section 2.1.2. For this reason, one might expect to see the positive error increasing with increasing current speeds, as the model overestimates how many lice will be present at farms with higher current speeds. This trend is to a small degree present in Fig. 4.8, but does not give conclusive results of this hypothesis.

Another interesting result is the comparison with the HI model in Fig. 4.19 (p. 50), in which the HI model gives a significantly higher correlation value in cohorts with a maximum current below $0.2\,\mathrm{m/s}$, and especially good if below $0.1\,\mathrm{m/s}$. However, at higher current speeds, the HI model is generally inferior. One possible explanation for this is that the HI model quite accurately predicts the lice density at any location. Still, if the current at the location is too high, the spatial and temporal inaccuracies in the model as well as the lack of information on when lice counts are performed causes the model to become highly inaccurate. However, the amount of cohorts at these lower currents are few, and the results are therefore not statistically strong enough to warrant firm conclusions.

Lice counts

The increasing negative error as the average number of counted PAAM lice increases indicates a systematic error in the model. Furthermore, the confidence intervals in Fig. 4.9a (p. 42) are tight and shows that this error is representative for the cohorts within the bins.

However, the amounts of cohorts with high average lice counts are low, as Fig. 4.9d shows. This might also be the reason for the negative error for these cohorts, as there are so many more data points with smaller values for lice counts than larger, making the negative binomial fitting too skewed towards fewer counts of lice.

Biomass

The large negative error when a cohort consists of less than 100.000 salmon appears to point to a large flaw in the model. As Eq. (2.1) shows, the infection pressure in the VI model is directly dependent on the number of salmon in the farm. The results indicate that instead of a linear dependence, it should rather be modeled as for instance $y \approx \sqrt{x}$, or with a steeper curve at lower values, in order to predict more lice when there are few salmon.

5.3.4 External factors

The results from the external factors show that the amount of lice at a farm is highly dependent on the density of active farms surrounding it, as was seen in Fig. 4.11 (p. 43), and that the correlation also decreases as the density increases. At density values above 1.0, however, the results are less clear, which can likely be attributed to the smaller number of cohorts in these bins. The increasing negative error at higher densities is quite clear though, showing that the model clearly underestimates the amount of lice at farms with many close neighbors. The other external factors do not appear to have an equally large impact on the model performance.

The comparison to the HI model in Fig. 4.21 (p. 51) appears to show that the HI model is better, with both a lower error and higher correlation, at densities above 0.6 active farms. However, neither model is good at these higher densities. Furthermore, the large dip in the correlation value at densities between 0.4 and 0.6 indicates that the farm density does not have the same effect on the HI model as it does on the VI model.

5.3.5 Regression models

An important result from the regression analysis is which of the features that were most important for the prediction. For the first parameter group, the amount of PAAM lice three weeks earlier has the highest coefficient in both models, followed by the amount of adult female lice two weeks prior. The results suggest that the number of counted lice in the previous weeks can be good estimators of the lice in the current week.

For the second parameter group, the density of PAAM lice the same week is the most important parameter. As discussed in Section 5.1.2, however, there are uncertainties in the density calculation that can have a large impact on the regression. And even if these uncertainties are minimized or depleted, there is still the question of whether the distance function in Eq. (2.3) has the right shape. A way to counter this problem could be to include a parameter of the distance function as a hyper-parameter in the regression. This would

not only give the potential for a better model but also give more insight into what exactly the shape of this function should be.

In Fig. 4.24 (p. 54), which shows how the SVR 1- and Linear 2-models compare to the VI model throughout the year, the regression models have a much higher correlation than the VI model from October to March, with a peak in January. This shows that although the VI model fails to accurately predict lice counts at these months, possibly due to the decreasing water temperature, the regression models are actually best at these times.

Based on the decline of the prediction ability of the VI model since 2015, it is likely that the regression models would have achieved much better correlation values and prediction scores if they had been applied to cohorts from between 2012 and 2014. There is however little use in predicting the past, and a model that cannot predict the future would be useless.

Chapter 6

Conclusion

In this work, the prediction ability of the salmon lice dispersion model created by Veterinærinstituttet has been thoroughly analyzed. The results show that the model is highly dependent on time, the water temperature and the current velocity. It performs best before 2016, between May and September, when the median water temperature is approximately 8 °C to 11 °C and increasing, and the current is slow. Additionally, it tends to underestimates the lice abundances at farms with a median amount of salmon less than 100.000 or with many neighboring active farms.

The HI model has a similar inter-annual time dependency as the VI model, although its predictive abilities are generally worse. However, it appears to give better predictions at low water temperatures, when the temperature is decreasing quickly, at low water salinities and when the currents are slow. However, the much smaller amount of data available from the HI model results makes the comparison results less significant than they could have been. However, they do indicate that a merging of the two models can prove to give better and more generalizable results than either of the models can give separately.

The regression models that were used to predict lice abundances have several advantages over the more traditional VI- and HI models, the biggest being that it does not require biological parameters such as mortality rates or development times to model lice abundances. The results show that it performs about equally well with the VI model from May to August but are superior the rest of the year. While the models using lice counts from previous weeks perform better, as expected, even the ones without them perform better than the VI model. The result that these machine learning models perform better than the existing models shows that the future of salmon lice predictions might not lie in improving the already existing models, but rather in using the increasing amounts of available data in machine learning.

It must be noted that the F_1-scores of the models are not overly convincing, but with more data and more complex models, this is expected to improve quickly. Moreover,

these models can easily be extended to also account for treatments and other factors which have an impact on the lice abundances. Such models will not only be useful for salmon farmers hoping for better forecasts to aid their decision-making processes, but also as tools to estimate the infection pressure from salmon farms on wild salmon and in effect assist lawmakers when new laws or routines are being developed.

6.1 Suggestions for future work

The analysis of the VI model has shown that some parameters has a large impact on the model performance, in particular the water temperature and how quickly it changes, which is reflected in the temporal results; the current speed at the location; the amount of lice; and the density of neighboring active farms. The bad correlation at lower water temperatures is likely at least partially caused by the biological parameters not being suitable for the full range of water temperatures. In lieu of existing or new research investigating the growth and mortality rates of salmon lice at low temperatures, the existing parameters can be tweaked through trial and error. The framework for model testing introduced in this work can be used to re-evaluate the tweaked models.

As mentioned previously, a new evaluation of the distance function in Eq. (2.3) would be beneficial, and letting it be merged or weighted with a measure of the current flow between farms should be looked into. This would in effect constitute a merging of the VI and HI models and utilize the best of each model: The simplicity and lack of dependencies of the former, and the use of currents in the latter. By avoiding accurate spatial and temporal modeling of the transportation of lice but rather relying on trends and seasonal effects on the currents, the problem of the patchy field might be partially resolved. The observation that the HI model performs much better than the VI model when the current speed is low supports this idea: Modeling the dispersion of lice in the currents is good if the current velocity at the destination is low, but at higher velocities the patchiness in time and space necessitates a less sensitive model, such as the one created by VI.

Another factor that future research can probe further into is time delay. By offsetting the predicted values with respect to the counted values by different amounts of time, and plotting the time delay giving the best correlation as a function of e.g. the median temperatures of each cohort. This would likely show that at temperatures of 8 °C to 11 °C, for which the correlation is already quite high, the correlation would not increase much by an offset. However, at lower temperatures the results might show that an offset of several days or even weeks gives a much better correlation. If this is the case, it strongly indicates that the biological parameters at these low temperatures are wrong, and the suggested time delay can be used to see what the parameters ought to be. Due to the uncertainties in when the lice are counted at each farm though, the data might however be too noisy for this approach to give much insight.

There is likely much more that can be learned from comparing the VI and HI models than what was found through this work. If results from the HI model had been available for more than just 2016 and 2017, and if the model had been run for the whole year and not

only in the spring, a thorough comparison could have been undertaken.

As the amount of different treatments and use of technologies to combat salmon lice increases, these models are likely to become less and less accurate at predicting lice abundances at salmon farms. However, in addition to the lice counts from salmon farms, another resource for model validation exists, namely the cages of salmon that HI places in Norwegian fjords to estimate the impact of lice on wild salmon smolts, as mentioned in Section 2.3.2. The lice on these salmon are counted with high accuracy, but time series information is unavailable and the amount of data is small compared to the information from the farms. Regardless, this data can be used to validate (or invalidate) the findings from this research.

One of the main advantages of machine learning techniques compared to the more traditional techniques used by VI and HI is the lack of dependence on biological and other sub-models to build the main model. Instead, they only require data. The regression techniques used in this work are quite simple and work well even with the limited amount of data used to train the models. However, more complex techniques such as ones utilizing deep learning often requires a much larger dataset, but can also attain much higher accuracies. New technologies are enabling automatic lice counts with a much higher precision than the manual counts today, for instance Ecotone's SpectraLice [70], which uses hyperspectral imaging to identify lice on the salmon. These counts can the be used not only to evaluate existing models in a much more precise way, but can also be fed into a deep learning model and combined with parameters discussed in this research to predict future lice abundances in the same farm, or the impact of lice on wild salmon.

An example of a type of deep learning algorithm that could be used for this is a recurrent neural network (RNN). RNNs, and particularly ones using long short-term memory networks [71], are widely used for e,g, speech recognition [72, 73] and time series prediction [74, 75]). What makes RNNs especially useful for predicting time series (such as the abundance of salmon lice over time) is that they remember input from all previous time steps and use that to predict the future. For comparison, SVR does not inherently see the input as time series but rather as a set of data in which the temporal information is only included as features to the model.

6.2 Contributions

The main contribution of this work to the field of salmon lice abundance prediction is the design of a framework for analyzing, evaluating and comparing salmon lice dispersion models. The framework can easily be expanded when new data or models become available. Furthermore, the results from the analysis of the existing models can be used as a guideline for improving them, for which the framework can function as a metric to evaluate the improved version. In addition, the results from the regression models implemented in this work shows that machine learning models can provide significantly better predictions than the existing models.

Bibliography

[1] T. Venvik. National aquaculture sector overview. norway. national aquaculture sector overview fact sheets., 2005. http://www.fao.org/fishery/countrysector/naso_norway/en.

[2] Statistisk sentralbyrå. Akvakultur, 2017. https://www.ssb.no/jord-skog-jakt-og-fiskeri/statistikker/fiskeoppdrett/aar.

[3] A. Iversen, Ø. Hermansen, and O. Andreassen. The cost impact of lice in norwegian salmon farming. http://www.seafoodinnovation.no/download/317, 2016.

[4] P. A. Jansen, A. B. Kristoffersen, H. Viljugrein, D. Jimenez, M. Aldrin, and A. Stien. Sea lice as a density-dependent constraint to salmonid farming. *Proceedings of the Royal Society B: Biological Sciences*, 279(1737):2330–2338, 2012.

[5] A. D. Sandvik, P. A. Bjørn, B. Ådlandsvik, L. Asplin, J. Skarðhamar, I. A. Johnsen, M. Myksvoll, and M. D. Skogen. Toward a model-based prediction system for salmon lice infestation pressure. *Aquaculture Environment Interactions*, 8:527–542, 2016.

[6] Nærings- og fiskeridepartementet. Forskrift om bekjempelse av lakselus i akvakulturanlegg. I 2012 hefte 13, 2012. https://lovdata.no/dokument/SF/forskrift/2012-12-05-1140.

[7] I. P. Helland, I. Uglem, P. A. Jansen, O. H. Diserud, P. A. Bjørn, and B. Finstad. Statistical and ecological challenges of monitoring parasitic salmon lice infestations in wild salmonid fish stocks. *Aquaculture Environment Interactions*, 7(3):267–280, 2015.

[8] Anonymous. Meld. st. 16 forutsigbar og miljømessig bærekraftig vekst i norsk lakse- og ørretoppdrett, 2015.

[9] O. Karlsen, B. Finstad, O. Ugedal, and T. Svåsand. Kunnskapsstatus som grunnlag for kapasitetsjustering innen produksjonsområder basert på lakselus som indikator. Technical Report 14, Havforskningsinstituttet, 2016.

[10] The Seafood Innovation Cluster. Aquacloud- the use of artificial intelligence in sea lice management. http://www.seafoodinnovation.no/article/213/AquaCloud_The_use_of_artificial_intelligence_in_sea_lice_management.

[11] C. J. Hayward, M. Andrews, and B. F. Nowak. Introduction:lepeophtheirus salmonis-a remarkable success story. In *Salmon Lice*, pages 1–28. Wiley-Blackwell, aug 2011.

[12] R. J. G Lester and C. J. Hayward. Phylum arthropoda. In *Fish Diseases and Disorders, Volume 1: Protozoan and Metazoan Infections*. Cabi Publishing, Oxon, UK, 2001.

[13] F. Samsing, F. Oppedal, S. Dalvin, I. Johnsen, T. Vågseth, and T. Dempster. Salmon lice (lepeophtheirus salmonis) development times, body size, and reproductive outputs follow universal models of temperature dependence. *Canadian Journal of Fisheries and Aquatic Sciences*, 73(12):1841–1851, 2016.

[14] K. Boxaspen and T. Næss. Development of eggs and the planktonic stages of salmon lice (lepeophtheirus salmonis) at low temperatures. *Contributions to Zoology*, 69, 2000.

[15] P. A. Heuch, J. R. Nordhagen, and T. A. Schram. Egg production in the salmon louse [lepeophtheirus salmonis (kroyer)] in relation to origin and water temperature. *Aquaculture Research*, 31(11):805–814, 2000.

[16] L. Asplin, I. A. Johnsen, A. D. Sandvik, J. Albretsen, V. Sundfjord, J. Aure, and K. Boxaspen. Dispersion of salmon lice in the hardangerfjord. *Marine Biology Research*, 10(3):216–225, 2013.

[17] I. A. Johnsen, Ø. Fiksen, A. D. Sandvik, and L. Asplin. Vertical salmon lice behaviour as a response to environmental conditions and its influence on regional dispersion in a fjord system. *Aquaculture Environment Interactions*, 5(2):127–141, 2014.

[18] P. A. Heuch, A. Parsons, and K. Boxaspen. Diel vertical migration: A possible host-finding mechanism in salmon louse (lepeophtheirus salmonis) copepodids? *Canadian Journal of Fisheries and Aquatic Sciences*, 52(4):681–689, 1995.

[19] E.M Hevrøy, K. Boxaspen, F. Oppedal, G. L Taranger, and J. C Holm. The effect of artificial light treatment and depth on the infestation of the sea louse lepeophtheirus salmonis on atlantic salmon (salmo salar l.) culture. *Aquaculture*, 220(1-4):1–14, 2003.

[20] I. R. Bricknell, S. J. Dalesman, B. O'Shea, C. C. Pert, and A. J. Mordue Luntz. Effect of environmental salinity on sea lice lepeophtheirus salmonis settlement success. *Diseases of Aquatic Organisms*, 71:201–212, 2006.

[21] S. Saksida, G. A. Karreman, J. Constantine, and A. Donald. Differences in lepeophtheirus salmonis abundance levels on atlantic salmon farms in the broughton archipelago, british columbia, canada. *Journal of Fish Diseases*, 30(6):357–366, 2007.

[22] C. W. Revie, G. Gettinby, J. W. Treasurer, G. H. Rae, and N. Clark. Temporal, environmental and management factors influencing the epidemiological patterns of sea lice (Lepeophtheirus salmonis) infestations on farmed Atlantic salmon (Salmo salar) in Scotland. *Pest Management Science*, 58(6):576–584, 2002.

[23] C. W. Revie, G. Gettinby, J. W. Treasurer, and C. Wallace. Identifying epidemiological factors affecting sea lice lepeophtheirus salmonis abundance on scottish salmon farms using general linear models. *Diseases of Aquatic Organisms*, 57:85–95, 2003.

[24] M. Aldrin, B. Storvik, A. B. Kristoffersen, and P. A. Jansen. Space-time modelling of the spread of salmon lice between and within norwegian marine salmon farms. *PLoS ONE*, 8(5):e64039, 2013.

[25] P. A. Heuch, R. S. Olsen, R. Malkenes, C. W. Revie, G. Gettinby, M. Baillie, F. Lees, and B. Finstad. Temporal and spatial variations in lice numbers on salmon farms in the hardanger fjord 2004-06. *Journal of Fish Diseases*, 32(1):89–100, 2009.

[26] A. Mustafa, W. D. Peters, G. A. Conboy, and J. F. Burka. Do water temperature and flow affect sea lice development and settlement? *Aquaculture Association of Canada Special Publication*, 5:53–55, 2001.

[27] F. Lees, G. Gettinby, and C. W. Revie. Changes in epidemiological patterns of sea lice infestation on farmed Atlantic salmon, Salmo salar L., in Scotland between 1996 and 2006. *Journal of Fish Diseases*, 31(4):259–268, 2008.

[28] P. A. Gillibrand and T. L. Amundrud. A numerical study of the tidal circulation and buoyancy effects in a scottish fjord: Loch torridon. *Journal of Geophysical Research*, 112(C5), 2007.

[29] Nærings- og fiskeridepartementet. Forskrift om drift av akvakulturanlegg (akvakulturdriftsforskriften). I 2008 hefte 8, 2008. `https://lovdata.no/dokument/SF/forskrift/2008-06-17-822`.

[30] L. M. S. Aas, P. A. Letnes, H. L. Braa, R. Pettersen, and K. Sæther. Klassifisering og telling av lakselus. resreport, Akvaplan-niva AS, 2017.

[31] M. Aldrin, R. B. Huseby, A. Stien, R. N. Grøntvedt, H. Viljugrein, and P. A. Jansen. A stage-structured bayesian hierarchical model for salmon lice populations at individual salmon farms – estimated from multiple farm data sets. *Ecological Modelling*, 359:333–348, 2017.

[32] A. B. Kristoffersen, D. Jimenez, H. Viljugrein, R. Grøntvedt, A. Stien, and P. A. Jansen. Large scale modelling of salmon lice (lepeophtheirus salmonis) infection pressure based on lice monitoring data from norwegian salmonid farms. *Epidemics*, 9:31–39, 2014.

[33] A. Stien, P. A. Bjørn, P. A. Heuch, and D. A. Elston. Population dynamics of salmon lice lepeophtheirus salmonis on atlantic salmon and sea trout. *Marine Ecology Progress Series*, 290:263–275, 2005.

[34] A. B. Kristoffersen, L. Qviller, K. O. Helgesen, K. W. Vollset, H. Viljugrein, and

P. A. Jansen. Quantitative risk assessment of salmon louse-induced mortality of seaward-migrating post-smolt atlantic salmon. Unpublished article.

[35] L. Asplin, K. Boxaspen, and A. Sandvik. Modelled distribution of salmon lice in a norwegian fjord. *ICES CM 2004/P:11*, 2004.

[36] I. A. Johnsen, L. C. Asplin, A. D. Sandvik, and R. M. Serra-Llinares. Salmon lice dispersion in a northern norwegian fjord system and the impact of vertical movements. *Aquaculture Environment Interactions*, 8:99–116, 2016.

[37] A. F. Shchepetkin and J. C. McWilliams. The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, 9(4):347–404, 2005.

[38] IMCS, Ocean Modeling Group. Regional ocean modeling system (roms), 2018. `http://myroms.org/`.

[39] J. Albretsen, A. K. Sperrevik, A. Staalstrøm, A. Sandvik, F. Vikebø, and L. Asplin. Norkyst-800 report no. 1. user manual and technical descriptions, 2011.

[40] B. Ådlandsvik and S. Sundby. Modelling the transport of cod larvae from the lofoten area. *ICES Marine Science Symposia*, 198:379–392, 1994.

[41] F. Samsing, I. Johnsen, L. H. Stien, F. Oppedal, J. Albretsen, L. Asplin, and T. Dempster. Predicting the effectiveness of depth-based technologies to prevent salmon lice infection using a dispersal model. *Preventive Veterinary Medicine*, 129:48–57, 2016.

[42] J. L. Myers and A. D. Well. *Research Design & Statistical Analysis*. Routledge, 2002.

[43] M. Aldrin. Havforskningsinstituttets spredningmodell for kopepoditter validert mot burdata fra 2014, 2016.

[44] D. N Gujarati and D. C. Porter. *Basic Econometrics*. McGraw-Hill Education, 2008.

[45] L. Qviller, A. Kristoffersen, and P. Jansen. Validering av havforskningsinstituttets luselarvespredningsmodell, 2016.

[46] E. Wit, E. van den Heuvel, and J. Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

[47] W. H. Greene. Accounting for excess zeros and sample selection in poisson and negative binomial regression models. *NYU Working Paper*, (No. EC-94-10), 1994.

[48] A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, and G. M. Smith. Zero-truncated and zero-inflated models for count data. In *Statistics for Biology and Health*, pages 261–293. Springer New York, 2009.

[49] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246, 1886.

[50] T. L. Lai, H. Robbins, and C. Z. Wei. Strong consistency of least squares estimates in

multiple regression. *Proceedings of the National Academy of Sciences*, 75(7):3034–3036, 1978.

[51] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. Vapnik. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161. MIT Press, 1997.

[52] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[53] W. McKinney. pandas: a foundational python library for data analysis and statistics. In *PyHPC 2011 : Python for High Performance and Scientific Computing*, 2011.

[54] S. Maneewongvatana and D. M. Mount. On the efficiency of nearest neighbor searching with data clustered in lower dimensions, 1999. `https://www.cs.umd.edu/~mount/Papers/iccs01-kflat.pdf`.

[55] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python. `http://www.scipy.org/`, 2001-.

[56] Kartverket. Dybdedata 50m grid, 2016. `https://kartkatalog.geonorge.no/metadata/kartverket/`.

[57] R. Sedgewick and K. Wayne. *Algorithms*. Addison Wesley, 2011.

[58] C. Careaga. Python A* pathfinding (with binary heap), 2014. `http://code.activestate.com/recipes/578919-python-a-pathfinding-with-binary-heap/`.

[59] A.B. Kristoffersen, H. Viljugrein, R.T. Kongtorp, E. Brun, and P.A. Jansen. Risk factors for pancreas disease (PD) outbreaks in farmed atlantic salmon and rainbow trout in norway during 2003–2007. *Preventive Veterinary Medicine*, 90(1-2):127–136, 2009.

[60] A. Zeileis, C. Kleiber, and S. Jackman. Regression models for count data in R. *Journal of Statistical Software*, 27(8), 2008.

[61] M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some implementations of the boxplot. *The American Statistician*, 43(1):50, 1989.

[62] J. L. Hintze and R. D. Nelson. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998.

[63] D. Harabor and A. Grastien. Online graph pruning for pathfinding on grid maps. In *25th National Conference on Artifical Intelligence. AAAI.*, 2011.

[64] D. Slagstad and Tm A. McClimans. Modeling the ecosystem dynamics of the barents sea including the marginal ice zone: I. physical and chemical oceanography. *Journal of Marine Systems*, 58(1-2):1–18, 2005.

[65] SINTEF Fiskeri og havbruk. Modellering av strøm, hydrografi og smittespredning i midt-norge. `http://midtnorge.sinmod.com/`.

[66] M. Næs, P. A. Heuch, and R. Mathisen. Bruk av «luseskjørt» for å redusere påslag av lakselus Lepeophtheirus salmonis (Krøyer) på oppdrettslaks, 2012.

[67] F. Oppedal, T. Dempster, and L. H. Stien. Snorkelmerd: Produksjonseffektivitet, adferd og velferd, 2016.

[68] A.P. Martin. Phytoplankton patchiness: the role of lateral stirring and mixing. *Progress in Oceanography*, 57(2):125–174, 2003.

[69] C. S. Tucker, R. Norman, A. P. Shinn, J. E. Bron, C. Sommerville, and R. Wootten. A single cohort time delay model of the life-cycle of the salmon louse lepeophtheirus salmonis on atlantic salmon salmo salar. *Fish Pathology*, 37(3):107–118, 2002.

[70] Ecotone AS. Automatic sea lice counting, 2018. `https://ecotone.com/automatisk-luseteller-fra-ecotone/?lang=en`.

[71] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[72] X. Li and X. Wu. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4520–4524, 2015.

[73] Y. Miao M. Gowayyed and F. Metze. Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding. 2015.

[74] J.T. Connor, R.D. Martin, and L.E. Atlas. Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2):240–254, 1994.

[75] P. Coulibaly and C. K. Baldwin. Nonstationary hydrological time series forecasting using nonlinear dynamic methods. *Journal of Hydrology*, 307(1-4):164–174, 2005.

Appendices

A Analysis 1

Internal factors

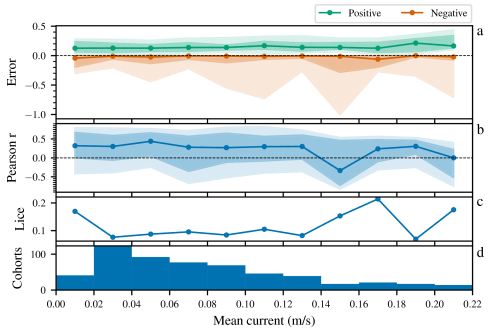


Figure 1: The median a) positive and negative errors and b) correlation values for the VI model, and the c) median counts of lice and d) total amount of cohorts, as a function of the mean current velocity at each farm. The shaded areas show the middle 50th and 80th percentiles.

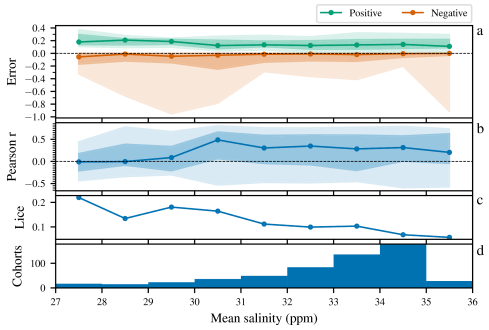


Figure 2: The error and correlation of the VI model as a function of the mean salinity. a-d) labeled as in Fig. 1.

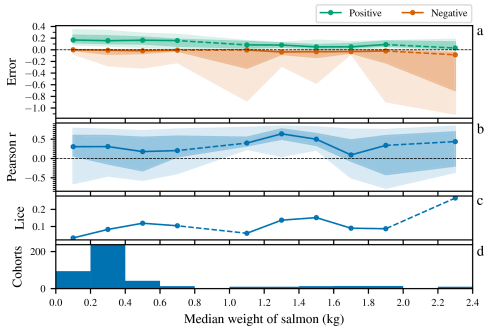


Figure 3: The error and correlation of the VI model as a function of the median salmon mass. a-d) labeled as in Fig. 1.

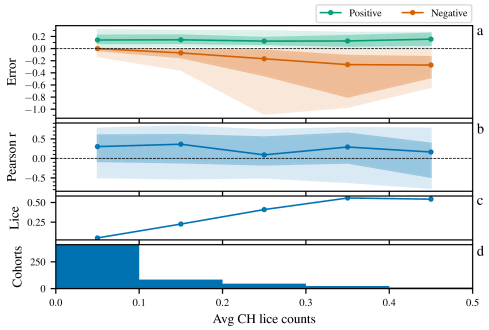


Figure 4: The error and correlation of the VI model as a function of the average CH lice counts. a-d) labeled as in Fig. 1.

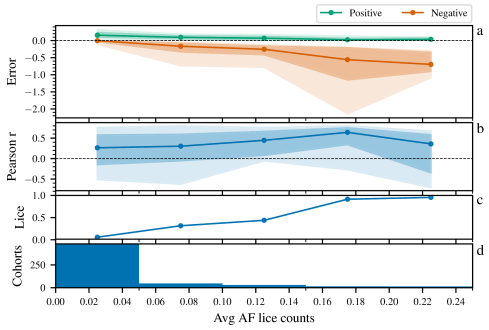


Figure 5: The error and correlation of the VI model as a function of the average AF lice counts. a-d) labeled as in Fig. 1.

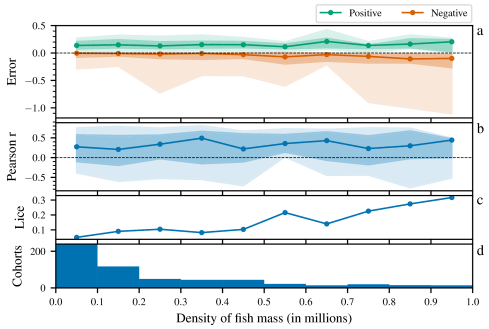


Figure 8: The error and correlation of the VI model as a function of the density of farmed salmon mass around each farm. a-d) labeled as in Fig. 1.

External factors

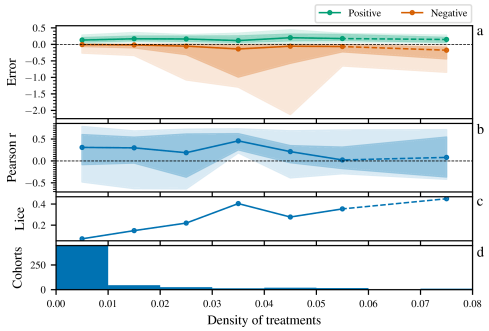


Figure 6: The error and correlation of the VI model as a function of the density of treatments around each farm. a-d) labeled as in Fig. 1.

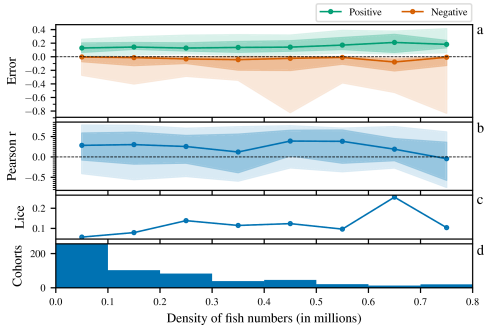


Figure 7: The error and correlation of the VI model as a function of the density of farmed salmon around each farm. a-d) labeled as in Fig. 1.

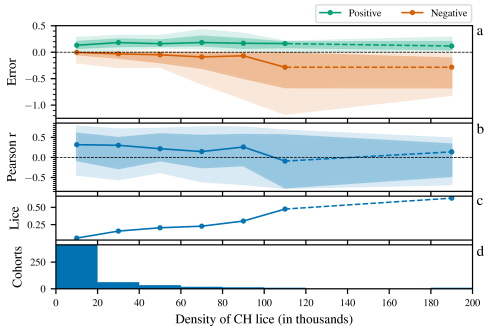


Figure 9: The error and correlation of the VI model as a function of the density of CH lice counts around each farm. a-d) labeled as in Fig. 1.

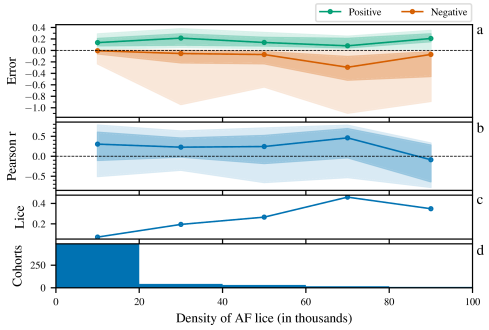


Figure 10: The error and correlation of the VI model as a function of the density of AF lice counts around each farm. a-d) labeled as in Fig. 1.

B Analysis 2

Internal factors

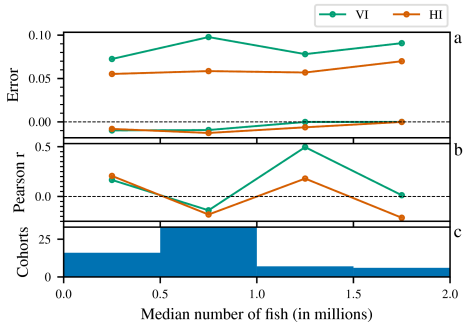


Figure 11: The median a) positive and negative errors and b) correlation values for the VI and HI models, and the c) total amount of cohorts, as a function of the median amount of salmon.

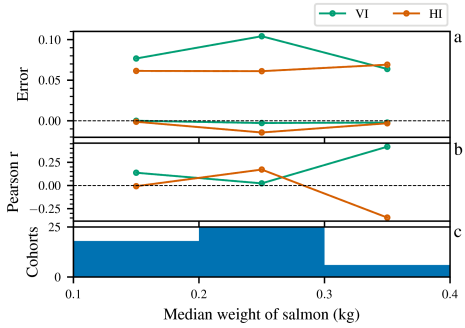


Figure 12: The error and correlation of the VI and HI models as a function of the median mass of salmon. a-c) labeled as in Fig. 11.

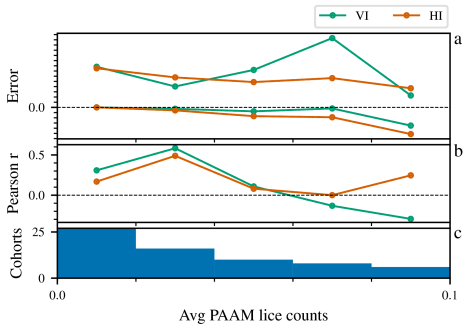


Figure 13: The error and correlation of the VI and HI models as a function of the average PAAM lice counts. a-c) labeled as in Fig. 11.

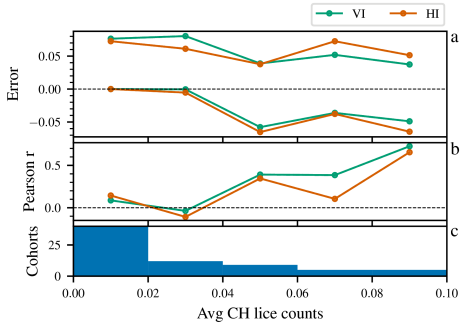


Figure 14: The error and correlation of the VI and HI models as a function of the average CH lice counts. a-c) labeled as in Fig. 11.

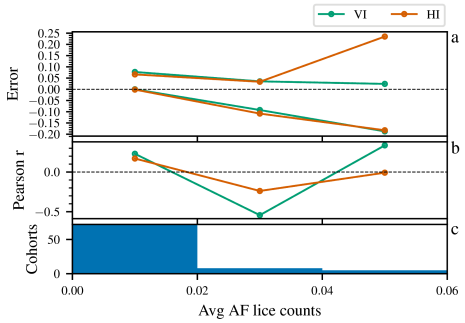


Figure 15: The error and correlation of the VI and HI models as a function of the average AF lice counts. a-c) labeled as in Fig. 11.

External factors

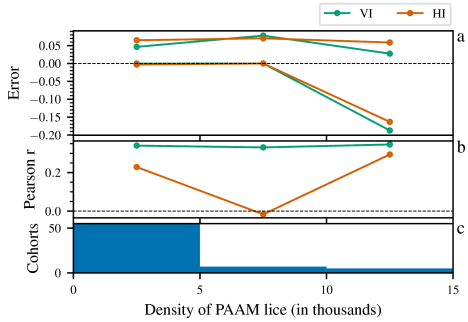


Figure 16: The error and correlation of the VI and HI models as a function of the average density of PAAM lice counts around each farm. a-c) labeled as in Fig. 11.

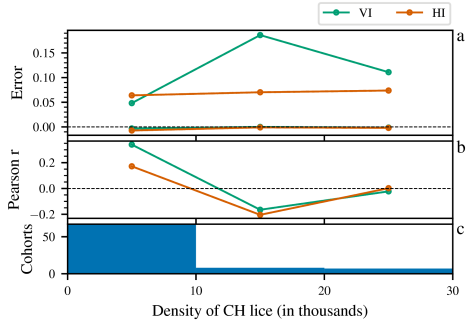


Figure 17: The error and correlation of the VI and HI models as a function of the average density of CH lice counts around each farm. a-c) labeled as in Fig. 11.

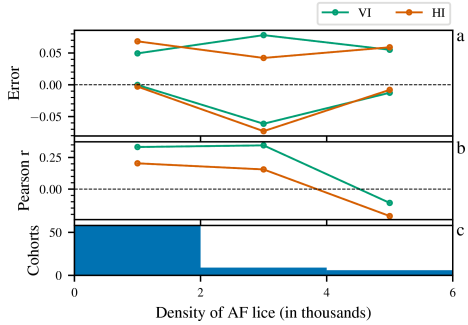


Figure 18: The error and correlation of the VI and HI models as a function of the average density of AF lice counts around each farm. a-c) labeled as in Fig. 11.

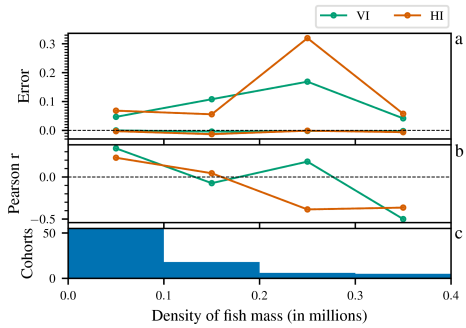


Figure 19: The error and correlation of the VI and HI models as a function of the average density of farmed salmon mass around each farm. a-c) labeled as in Fig. 11.

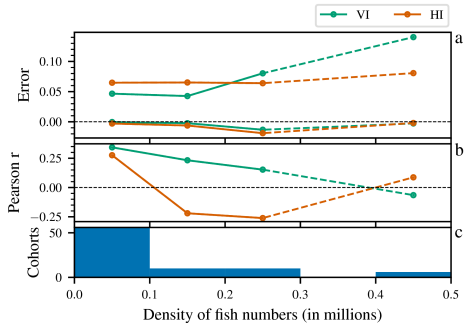


Figure 20: The error and correlation of the VI and HI models as a function of the average density of farmed salmon around each farm. a-c) labeled as in Fig. 11.

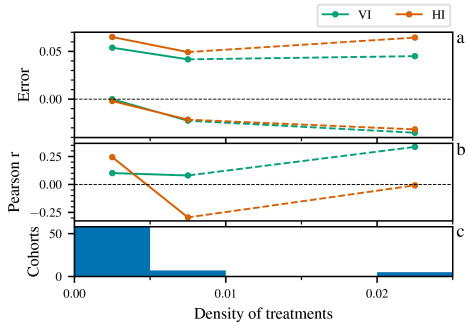


Figure 21: The error and correlation of the VI and HI models as a function of the average density of treatments around each farm. a-c) labeled as in Fig. 11.