Master's thesis

NTNU
Norwegian University of Science and Technology
Department of Mathematical Sciences

Lyder Bøe Iversen

Data Integration for Species Distribution Models

Master's thesis in Physics and Mathematics
Supervisor: Robert B. O'Hara
June 2020

NTNU
Norwegian University of
Science and Technology

Lyder Bøe Iversen

Data Integration for Species Distribution Models

Master's thesis in Physics and Mathematics
Supervisor: Robert B. O'Hara
June 2020

Norwegian University of Science and Technology
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

In this thesis the distribution of freshwater fish in Norwegian lakes is estimated by using data from two standardized data sets and one opportunistic data set in addition to environmental covariates to create a combined model. The model used is a Bayesian hierarchical model and model fitting is done using Integrated nested Laplace approximation (INLA), a tool for fast Bayesian inference. The use of INLA allows for the use of Gaussian random fields in the model, and parameterization of this random field is analysed. The combined model is shown to be better than all individual models for three of the four species examined.

Sammendrag

I denne oppgaven blir fordelingen av ferskvannsfisk i norske innsjøer estimert ved å bruke data fra to standardiserte datasett og ett opportunistisk datasett sammen med miljøbaserte kovariater til å lage en kombinert modell. Modellen er en Bayesiansk hierarkisk modell, og modelltilpasning blir utført med integrert nøstet Laplace approksimasjon, et verktøy for å utføre rask Bayesiansk inferens. Bruken av INLA gjør åpner opp for å bruke romlige Gaussiske stokastiske felter i modellen, og parametrisering av disse feltene blir analysert. Forskjellige typer data blir utforsket, og resultatene fra den kombinerte modellen blir sammenlignet med resultater fra modeller basert på de individuelle datasettene. Den kombinerte modellen får bedre resultater enn de individuelle modellene for tre av de fire artene som undersøkes.

Preface

I would like to say thank you so much to my supervisor Bob O'Hara for introducing me to the world of biostatistics, and for ideas and comments that helped me write this thesis. He has been immensely helpful throughout the whole process, and I could not have done this without him.
Thanks to Emma Skarstein for helping me with writing code and comparing models. Having someone to discuss results with who could provide a different perspective has been very helpful.

Thank you!

Lyder Bøe Iversen, June 2020

Table of Contents

Chapter 1

Introduction

Species distribution models (SDMs) are widely used in statistical ecology in order to model the distribution of species over a given area, and to predict how these species will be distributed in the future. SDMs are built using data sets containing observations of the species of interest in addition to environmental covariate data for the geographical areas being examined. In recent years there has been a lot of focus on how the observation data is being collected and in what way the different types of data should be used to create the best models (Miller et al., 2019). Observations gathered from sampling where a set of guidelines for sample location, effort and method is followed is usually called standardized data. This type of data is usually preferred for modelling as it makes comparison between areas, times and different data sets easier. With standardized data it is also more feasible to account for uncertainty in observations and various types of sampling bias.

For most species however, the majority of available data is of the non-standardized type as there are a lot less requirements for recording such data. Museum records and Citizen science data are examples of this type of data. Non-standardized data could be missing information about exact location, time, sampling effort and more, which makes estimating uncertainty in observations more difficult. As this type of data is often more opportunistic than standardized data, it is also more often heavily affected by sampling bias; citizen science data for example is biased towards more densely populated areas.

While non-standardized data is usually considered weaker and less informative than standardized data, it is much easier to collect as it does not require large scale complex surveys. The vast amount of non-standardized data available means there is still a lot of knowledge to be gained from it. The development of methods to combine different types of data in recent years (Miller et al., 2019) is therefore important, as this allows SDMs to be built in ways that utilizes the strengths and diminishes the weaknesses of the different data types.

In this thesis the topic of data integration will be explored by implementing a model that can estimate species distributions by combining data from multiple

data sets. The model will be used on observation data of freshwater fish species in Norwegian lakes from three different data sets. The modelling will follow a Bayesian approach, where all model parameters are given prior distributions based on prior knowledge. This type of modelling makes it possible to get uncertainty estimates for all parameters included in the model. The Bayesian approach also opens up the use of the Integrated nested Laplace approximations (INLA) method for model fitting (Rue et al., 2009), which allows for fast inference with complex models allowing for the use of spatial random fields.

Chapter 2 presents the data of freshwater fish observations used in this thesis. The environmental covariates used to model the distributions are also examined. In Chapter 3 the theoretical background needed to explain the modelling is presented, and Chapter 4 shows the model details and the model variations that are tested. Chapter 5 presents the results from the model analysis, while Chapter 6 is used to discuss the results and possible improvements to the model.

Chapter 2

Data

2.1 Data preparation

This thesis uses data on observations of freshwater fish in Norwegian lakes from three different data sets. For all the data sets, the data is filtered based on the following steps:

1. Species not in list of Norwegian freshwater fish, as described by SNL (Pethon and Vøllestad, 2019), are removed from the data.
2. The coordinates of all observations are matched to the closest lake, and all observations further than 10 meters from the closest lake are removed.
3. All observations with no time variable are removed.

To ensure there is enough data of each species to do proper inference, only observations of the four most prevalent species will be modelled and analysed, and observation numbers mentioned here will only account for these species. The four species of interest here will be *Trout*, *Arctic Char*, *Perch* and *Minnow*. The number of observations of these species is shown in Figure 2.1. Observations of other freshwater species will still be used to create pseudo-absences, which will be explained later.

2.2 Nordic fish status survey

The Nordic fish status survey data set consists of 54 species surveyed over lakes in Fennoscandia, including the 734 Norwegian lakes that are used here. 98% of the data is from 1996. The data were downloaded from GBIF[1]. The data set is based on presence/absence data, meaning both presences and absences of a

[1] https://gbif.vm.ntnu.no/ipt/resource?r=fish_status_survey_of_nordic_lakes

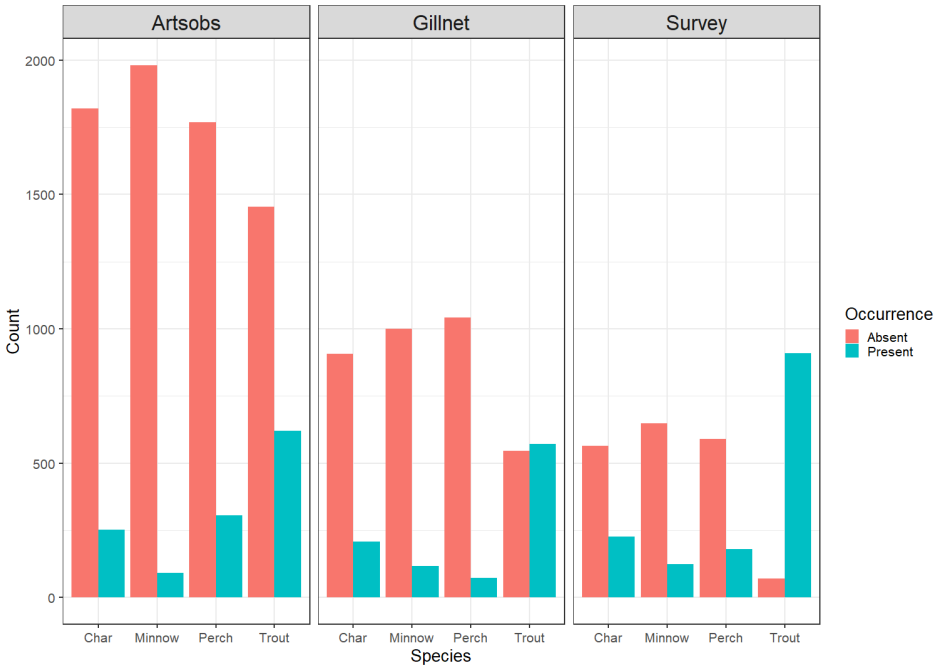


Figure 2.1

species are recorded for all lakes. After the filtering described in Section 2.1 the data set consists of 1439 presences and 1870 absences. The data were collected by questioning fishermen/locals/etc., which means it is considered a standardized data set. Because of this and the fact that this is the only data set where absences are recorded, the Nordic fish status survey is considered the main data set in this thesis.

2.3 Transcribed gillnet test fishing

The gillnet test-fishing data set is based on transcriptions of gillnet test-fishing results from Norwegian grey literature (technical reports) in the time period between 1970 and 1998. These data were downloaded from GBIF[2]. For each test-fishing event the number of caught fish of each species is recorded, meaning the data set consists of count data. There are however no absences explicitly recorded in the data. Pseudo-absences have therefore been added to the data set by assuming that a species is absent in a given lake if the species has not been recorded there, while one or more other species have been recorded in the lake. After filtering, the data set consists of 966 presences, and 3494 absences have been added with the described method.

[2]https://gbif.vm.ntnu.no/ipt/resource?r=transcribed_gillnet_test_fishing_data_norway

2.4 Citizen science observations from Artsobservasjoner

The data set from Artsobservasjoner (Artsobs) includes observations from as early as the 1960s all the way to 2020, but the majority of the observations are from the time period 1995-2019. The data were downloaded from GBIF[3]. The data set consists of Citizen Science data in the form of presence-only observations, meaning it is non-structured data recorded by various observers. Because of this it is assumed that the data from Artsobs is likely to have a stronger effect of sampling bias than the other data sets. Following the same reasoning as with the Gillnet data, pseudo-absences have been added to the Artsobs data set. This attempt at "upgrading" the data from presence-only to presence/absence will be examined later. After data preparation the data set consists of 1267 presences, and 7025 absences have been added.

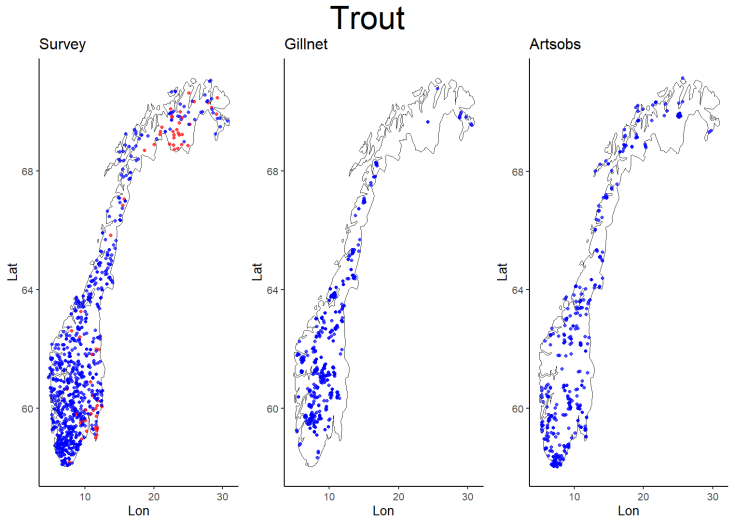


Figure 2.2: Lakes with observations of trout in the three data sets Survey, Gillnet and Artsobs. Blue points indicate an observed presence of the species, red points indicate an observed absence.

[3]GBIF.org (18 March 2020) GBIF Occurrence Download https://doi.org/10.15468/dl.nrjqsx

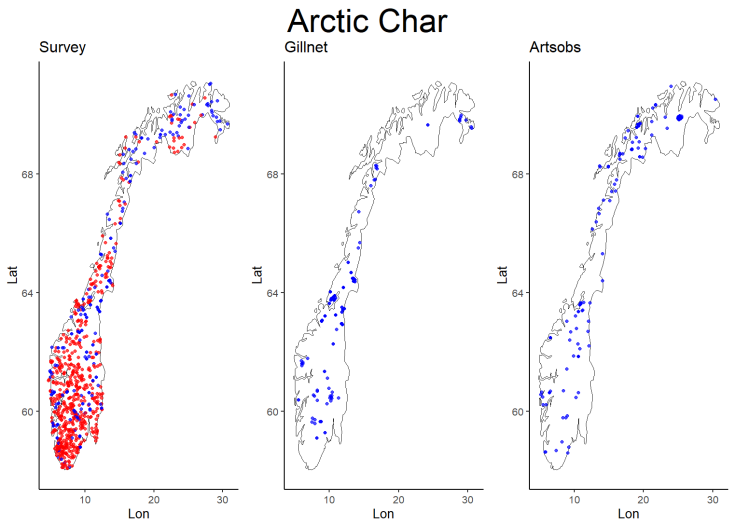


Figure 2.3: Lakes with observations of arctic char in the three data sets Survey, Gillnet and Artsobs. Blue points indicate an observed presence of the species, red points indicate an observed absence.

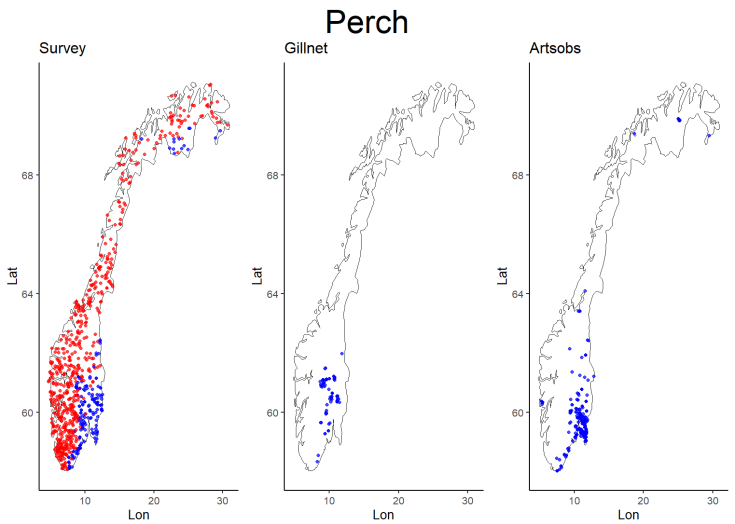


Figure 2.4: Lakes with observations of perch in the three data sets Survey, Gillnet and Artsobs. Blue points indicate an observed presence of the species, red points indicate an observed absence.

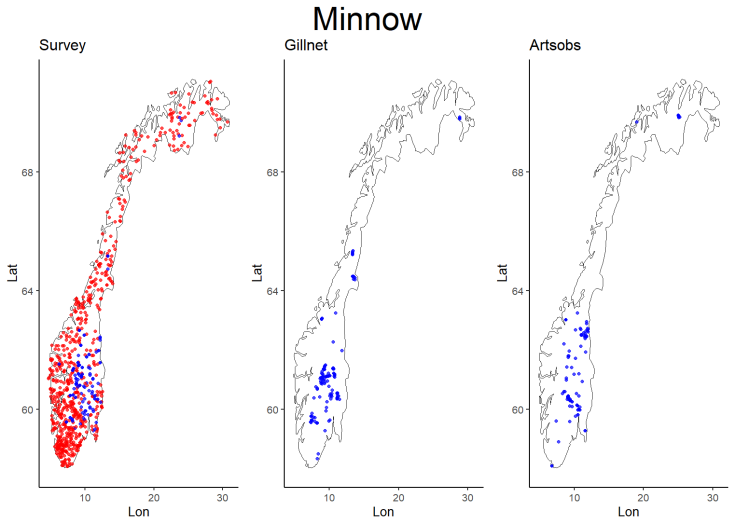


Figure 2.5: Lakes with observations of minnow in the three data sets Survey, Gillnet and Artsobs. Blue points indicate an observed presence of the species, red points indicate an observed absence.

2.5 Covariates

The covariates used in modelling are:

- Longitude, in degrees
- Latitude, in degrees
- Land surface temperature during summer, in degrees Celsius times ten
- Human Footprint Index, score between 0 and 50
- Distance to nearest road, in meters
- Lake area, in square meters (on the log-scale)

While the longitude and latitude of a lake are not environmental covariates, they are known to often be correlated with the distribution of freshwater fish. These covariates are also often correlated with the temperature, especially the latitude, and as such are often considered proxies for this environmental covariate. The land surface temperature during summer is taken from the article by Metz (2014) and is based on data from 2010. This covariate can be seen as a proxy for the temperature in the lakes, which is often an important factor for whether a species is found in an area or not.

The Human Footprint Index is a score based on the combination of eight different human impact variables which approximate the impact from humans on nature. These variables include human population density, nearby roads and farmlands, and

HFI is measured for grid cells of one square kilometer (Venter et al., 2016). The HFI data is from 2009 and is downloaded from `https://wcshumanfootprint.org`. HFI and distance to nearest road are included as covariates in an attempt to model the effect of human population distribution on the data. A lake being more accessible to humans is often thought to lead to more observations, especially in non-standardized data sets.

The lake area is included as the intensity of a species distribution is believed to increase approximately linear as a function of the lake area. The lake area is included on the log-scale since the intensity is modelled as a log-linear function of the covariates, as described in Section 3.2.1.

Figure 2.6 shows the correlation between the covariates. It is clear that there is significant correlation between all covariates excluding the lake area. This is not great, as correlated covariates can lead to higher variability, and the correlation can make it difficult to distinguish which covariates are significant to the model. The effects of the correlated covariates will be examined more in Chapter 5.

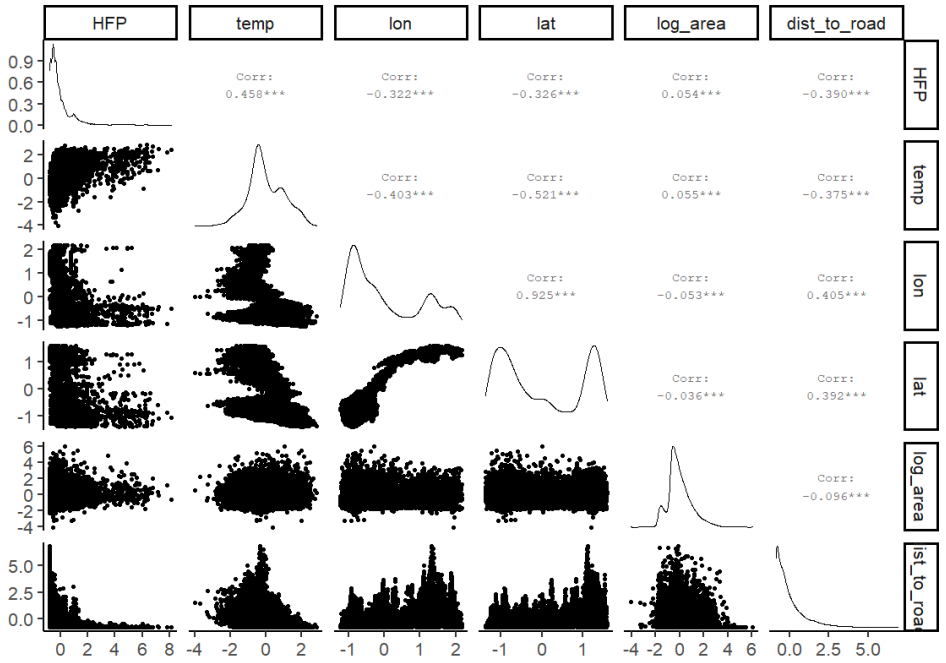


Figure 2.6: Pairs-plot of the covariates used in modelling. The lower triangle of the matrix shows the values of the covariates plotted against each other. The upper triangle shows the correlation values of the covariates from -1 to 1.

Chapter 3

Theory

Most species distribution models attempting to integrate different data types do so using spatial point processes (Miller et al., 2019), (Isaac et al., 2019). A spatial point process can be used to model the expected distribution of a species at a location s in space by having points be generated independently from a random process. $\lambda(s)$ describes the expected density of points at the location s, which can be modelled as the intensity of a Poisson process. This Poisson point process can then be described as a function $f(\lambda(s), X, \phi)$ of environmental covariates X and other parameters ϕ that would influence the species distribution. A major advantage of using spatial point processes over other modelling approaches is that there is no need to discretize the data, as the parameters of the model do not depend on the spatial scale (Dorazio, 2014). This is convenient when using data from data sets with differing spatial resolution, which is often the case when integrating standardized and non-standardized data.

3.1 Bayesian hierarchical models

The distribution of a species can be modelled by a Poisson point process, but this distribution cannot be directly observed, as that would require observing every single individual inside the area of interest. Because of this, there is need for a modelling structure that can be used to estimate a hidden process using this type of imperfect data. Bayesian hierarchical models are used to create a layered dependence structure consisting of observation models, process models and parameter models, in descending order. One or more observation models will be used to describe how the observational data were produced, including information about possible sampling bias and detection probability. A process model will model the hidden (or latent) process of interest and its uncertainty. Finally, a parameter model will give all the unknown parameters of the model probability distributions that can be controlled by hyperparameters. If $\pi(b)$ is the distribution of a random variable B, the Bayesian hierarchical model system is based on

Observation model: $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$
Process model: $\pi(\boldsymbol{x}|\boldsymbol{\theta})$
Parameter model: $\pi(\boldsymbol{\theta})$
where $\boldsymbol{y}$ is the observation data, $\boldsymbol{x}$ is the latent process and $\boldsymbol{\theta}$ is the list of unknown parameters. REF!!!!!!! In Bayesian statistics the parameter model $\pi(\theta)$ is called the prior distribution and reflects any knowledge of the distribution prior to the use of any observation data. This prior knowledge is used together with the data to construct the posterior distribution, which is the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ given $\boldsymbol{y}$, using Bayes' theorem REF!!!!!!!!!

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \tag{3.1}$$

The goal of Bayesian inference in the context of species distribution models is thus to use any prior knowledge of the species and covariates and combine it with the observations from available data sets to estimate the posterior distribution of the species of interest.

When integrating different data sets there will be one observation model for each data set, and all these models will be conditional on the same latent process. With Y_d being the data from data set d, the likelihood for the data is $Pr(Y_d|\lambda(s), \theta_d)$ where θ_d is a set of parameters effecting the observation model. This leads to a joint likelihood approach, where each data type is used to fit a likelihood for a set of shared parameters describing the latent process. These parameters are set to be equal across the different likelihoods, and the parameter estimates are the ones that best fit the likelihoods all together (Miller et al., 2019). This approach makes it possible to create a set of estimators for the species distribution which is based on all the different data sources. Based on these definitions of the latent process and the observation models, the full likelihood for the state-space model is

$$L(Y_d|X, \phi, \theta_d) \propto f(\lambda(s), X, \phi) \prod_{d=1}^{M} Pr(Y_d|\lambda(s), \theta_d) \tag{3.2}$$

3.2 Distribution and observation models

3.2.1 Process model

Modeling the actual distribution of the species can be done in different ways, one of which is by using point process models. Here the species density is modeled with a continuous surface, where higher density at a location corresponds to a species being more likely to exist at said location. The specific case of point process models for freshwater fish is unique in the sense that the density of the continuous surface will be forced to zero at all points outside of lakes, as freshwater fish naturally can not exist at these points. Modelling the process as a continuous point process opens up the possibility of using some methods developed for continuous models, even though the restriction on observations to discrete lakes might advocate for the use of discrete methods.

In this thesis the continuous surface is modeled as a log-Gaussian Cox process with intensity $\lambda(s)$ at location s (Møller and Waagepetersen, 2004). The intensity is formulated as a log-linear function of $i = 1, .., P$ covariates with corresponding fields $X_i(s)$ and unknown parameters β_i

$$\log \lambda(s) = \eta(s) = \sum_{i=1}^{P} \beta_i X_i(s) + u(s) \tag{3.3}$$

$u(s)$ is a spatial field that is included to model the effects that are not explained by the covariates. $u(s)$ creates spatial autocorrelation between observations and will be modelled as a Gaussian Markov Random Field (Rue and Held, 2005) in such a way that $u(s)$ will have Matérn covariance defined as $Cov(u(s_i), u(s_j)) = \sigma_u Corr(u(s_i), u(s_j))$ where σ_u^2 is the marginal variance and

$$Corr(u(s_i), u(s_j)) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa ||s_i - s_j||)^\nu K_\nu(\kappa ||s_i - s_j||) \tag{3.4}$$

Fixing $\nu = 1$ and parameterising the covariance as $\theta = \{log(\tau), log(\kappa)\}$ with τ being a parameter for the local variance, yields the marginal variance as

$$\sigma_u^2 = \frac{1}{4\pi\tau^2\kappa^2} \tag{3.5}$$

which gives

$$\log(\tau) = \frac{-\log(4\pi\sigma_u^2\kappa^2)}{2} \tag{3.6}$$

The model can be expanded to include the time dimension by adding an AR(1) term to the right hand side of (3.3)

$$\eta(s,t) = \sum_{i=1}^{P} \beta_i X_i(s,t) + \rho\eta(s,t-1) + u(s,t) \tag{3.7}$$

where ρ is the parameter of the AR(1) series and $u(s,t)$ is separable. The inclusion of the time dimension to the model will due to time constraints not be examined further in this thesis.

3.2.2 Presence-only observation model

The observation process for the presence-only data is based on the assumption that the observations come from a thinned point process, modelled as a random sample of where the species actually appear. If the probability of observing an individual is $q(s)$ then the observation intensity is $\phi(s) = q(s)\lambda(s)$. The parameter $q(s)$ is usually not known and is therefore estimated. In most data sets however, this thinning is not done evenly across the geographic area but heavily affected by sampling bias. For example, observations can be weighted towards more densely populated areas or areas that are more attractive for research. These biases can heavily devalue the information gained from presence-only data, and it is important to account for

this in some way, or at least be aware of the effect. Symmonds (2020) shows that it's possible to attach a second spatial field $\zeta(s)$ to only the presence-only data to help account for the sampling bias. This spatial field is intended to describe the spatial variation which is not explained by the either the covariates or the shared spatial field $u(s)$.

3.2.3 Presence/absence observation model

Integrating one or more data set consisting of presence-absence data would be a natural extension of the model based on presence-only data. As described by Isaac et al. (2019) presence-absence data is usually modeled as a Bernoulli random variable giving the probability of observing at least one individual at a given lake as

$$Pr(N(s,t) > 0) = p(s,t) = 1 - e^{-\eta(s,t)} \tag{3.8}$$

which on log scale for η gives the inverse of the complimentary loglog link function

$$\log(-\log(1 - Pr(N(s,t) > 0))) = \eta(s,t) \tag{3.9}$$

where $\eta(s,t) = pt \int_s \lambda(s)ds$. Because the area of the lake is very small compared to the whole region, the intensity surface is unlikely to change significantly. The intensities can therefore be assumed constant within each lake, and $\eta(s,t) \approx ptA(s)\bar{\lambda}(s)$ with $A(s)$ being the area of the lake.

3.2.4 Point count observation model

An extension of the presence-absence data would be count data. While presence-absence data only provides information about whether a species has been observed or not in a given lake, count data also provides the number of individuals observed for each observation period. The number of individuals observed then follows a Poisson distribution given by

$$Pr(N(s,t) = r) = \frac{\eta(s,t)^r e^{-\eta(s,t)}}{r!} \tag{3.10}$$

with $\eta(s,t)$ defined in the same way as with the presence-absence data.

3.2.5 The intercept and effort

The total abundance N of a species in the region of interest is given as the integral of the intensity λ of the process model over said region. The model is designed on the assumption that not all individuals in the region are being observed, and as such it can be difficult to estimate this integral. The probability p of observing each individual of the species, which is often also too difficult to estimate, can help find the total abundance as

$$N = \int \lambda(s)ds = \int e^{log(p)+\eta(s)}ds \tag{3.11}$$

Equation (3.11) shows that imperfect species detection ($p < 1$) affects the model by changing the intercept of η. While not being able to estimate p means we cannot estimate the total abundance, this also means that we don't need to model the intercept precisely as it cannot be estimated. This also means that any covariates that are known to be constant across a data set do not need to be estimated, as they cannot be separated from the intercept. The observation time t, the probability of observation p and the area of the observation site $A(s)$ are parameters that can be combined into a parameter called effort, $E(s) = ptA(s)$, and these parameters can sometimes be assumed constant dependent on the data set. If any parameters of the effort varies in an unknown way, said parameters should be included as additional terms to be estimated in the model. If the parameters vary in a known way however, for example the area of observation $A(s)$ being the area of a lake, this can be included as a known term by adding $\log(A(s))$ as an offset in the model.

3.3 Model fitting with INLA

In order to perform any form of statistical inference, the model above needs to be fitted to the data. The hierarchical structure and the possible large number of parameters in the model makes Bayesian modelling a natural choice. Due to the possibility of high model complexity, as well as the importance of including spatial autocorrelation in the model, computational efficiency is important when it comes to model fitting. Markov Chain Monte Carlo (MCMC) methods have traditionally been the most used way to fit Bayesian Hierarchical models, but recently developed methods using Integrated Nested Laplace Approximation (INLA) have been proven able to give similar results with significantly lower computation time (Rue et al., 2009). While MCMC is sampling based, INLA closely approximates the posterior distribution and solves the necessary integrals numerically.

To be able to use INLA to do inference on a model, it needs to be a latent Gaussian model where the latent field is a Gaussian Markov Random Field (REF?). The class of latent Gaussian models consist of models where all elements of the predictor η, and therefor η itself, are assumed to be Gaussian. A Gaussian Markov random field is a random field where all finite-dimensional distributions are Gaussian (REF, Stein?), and all points in the field have the Markov property, i.e. points that are not in the same neighborhood are considered independent of one another.

To approximate the Gaussian random field in a computationally efficient way, the INLA-package can use stochastic partial differential equations (SPDE), where the stochastic terms are Gaussian white noise. The SPDE is solved numerically and discretized on a triangulated mesh grid. This is implemented in the INLA-package, where a particular SPDE called the linear fractional SPDE is used to give a Gaussian random field with Matérn covariance function when solved (Lindgren et al., 2011), (Hem, 2017).

Chapter 4

Method

As described in Section 3.3 INLA will be used to do inference in this thesis, and this will be carried out using the R-package `R-INLA`. The models used for this inference is shown in Section 4.1. To find the model best suited to examine the effect of data integration, comparing different model variations will be done in three steps.

First modelling of the non-standardized data from Artsobservasjoner is tested to see whether adding absences to the data improves the model. After that the INLA default prior distributions for the Gaussian random field are compared to using Penalized Complexity prior distributions, and different hyperparameters are examined. More details on prior distributions being tested is given in Section 4.2. Finally the full model based on all three data sets with the preferred configurations from the previous two comparisons is compared to individual models based on only one data set. This is done to examine the effects of integrating multiple data sets into one full model and whether this leads to improved inference. Model validation is carried out by cross-validation as described in Section 4.3.

4.1 Model

The models used for inference is based on the theory in Chapter 3, and the models for the different data types used is shown here.

As described in Section 3.2.3 for presence/absence data, for a location s, the linear predictor in the model is given as

$$\begin{aligned}\eta(s) = \beta_0 + \beta_{lon}X_{lon} + \beta_{lat}X_{lat} + \beta_{temp}X_{temp} + \beta_{HFI}X_{HFI} \\ +\beta_{road}X_{road} + \beta_{area}\log(X_{area}) + u(s)\end{aligned} \tag{4.1}$$

with $y(s) \sim \text{Bernoulli}(\mu(s))$, $\eta(s) = \log(-\log(1-\mu(s)))$ and where β_0 is a data set specific intercept. $u(s)$ is the Gaussian random field described in Section 3.2.1.

For presence-only data:

$$\begin{aligned}\eta(s) = \beta_0 + \beta_{lon}X_{lon} + \beta_{lat}X_{lat} + \beta_{temp}X_{temp} + \beta_{HFI}X_{HFI} \\ +\beta_{road}X_{road} + \beta_{area}\log(X_{area}) + u(s) + \zeta(s)\end{aligned} \tag{4.2}$$

with $y(s) \sim \text{Poisson}(\mu(s))$, $\eta(s) = \log(\mu(s))$ and where β_0 is a data set specific intercept. $u(s)$ and ζ are Gaussian random fields, where ζ is included while modelling presence-only data to help account for the sampling bias (Symmonds, 2020).

4.1.1 Mesh

As described in Section 3.3 the SPDEs approach in INLA requires a mesh grid to discretize the SPDEs on. The mesh used in all models in this thesis can be seen in Figure 4.1. This mesh consisting of 5852 nodes is divided into an inner and outer part, where the observations are found inside the finer inner grid, and the rougher outer grid is used to avoid numerical issues with the SPDEs due to boundary effects.

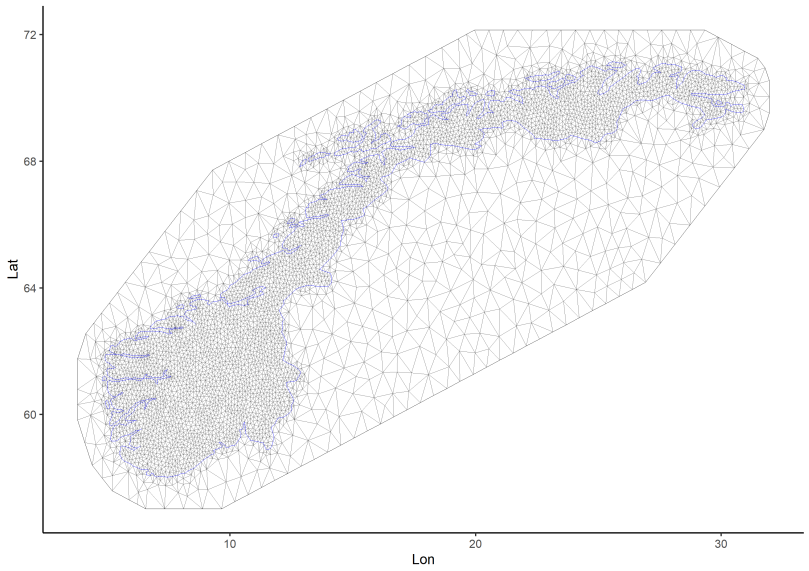


Figure 4.1: The mesh used for solving the stochastic partial differential equations in all models in this thesis. The total number of nodes in the mesh is 5852.

4.1.2 Artsobs - Poisson versus Bernoulli

The data from Artsobservasjoner includes presences only, which is usually considered worse than a data set including absences as well. As described in Chapter 2 however, because the observations are in discrete areas (lakes), it is possible to add absences to this data set. This is done by assuming a species is absent in a lake

if there is no recording of the species, and some other species is observed in that lake. A model where the Artsobs data is modelled as a thinned Poisson process based on presence-only data will be compared to a model where it is modelled by a Bernoulli process based on the data including absences. It is worth noting that while absences can be added to the data in this way, the data is still unstructured and likely suffers from sampling bias. Because of this, the spatial field ζ is kept in the model also when the data is modelled as presence/absence data.

4.1.3 Gillnet data

As described in Section 2.3, the Gillnet data set consists of count data, which in theory by Section 3.2.4 can be modelled using a poisson process. However due to numerical and programming issues, modelling the count data unfortunatly did not work as intended. Because of this, absences is added to the Gillnet data set, and the data is modelled as presence-absence data for the remainder of this thesis.

4.2 Priors

4.2.1 Covariate fixed effects

The coefficients of the longitude and latitude fixed effects β_{lon} and β_{lat} are given Gaussian priors with mean 0 and default variance 1000. The coefficients of the other fixed effects, β_{temp}, β_{HFI}, β_{road}, β_{area} are given Gaussian priors with mean 0 and variance 1. This is an attempt to emphasize the value of these covariates over the longitude and latitude. These covariates are actual environmental covariates expected to affect the underlying process, while longitude and latitude are often considered proxies for other effects like temperature.

4.2.2 Spatial random field

It is of interest to see how the priors of the Gaussian random field u affect the results. In all model variations both fields are assumed to have mean 0, but the parameters of the covariance function of u are given different prior distributions and hyperparameters, and the results of these variations are investigated. Four different configurations of parameters are tested, and in all these model variations the field ζ uses the INLA default priors, as analysing the effects of ζ is outside the scope of this thesis. The four model variations and the parameters differing between them are shown in Table 4.1.

Param.	Prior-model 1	Prior-model 2	Prior-model 3	Prior-model 4
ρ_u	INLA default	PC(2, 0.5)	PC(10, 0.5)	PC(2, 0.5)
σ_u	INLA default	PC(1, 0.1)	PC(1, 0.1)	PC(0.01, 0.1)

Table 4.1: Table describing prior distributions and hyperparameters used by the covariance function of the Gaussian random field u in the four model variations. PC$(\cdot, \cdot)$ has different interpretation for the range and standard deviation.

The INLA default priors are set based on the mesh, and so the model here have a unique set of default priors. In addition to this the priors are based on various parameters which make the priors complicated to state. There will therefor be no attempt at stating the default priors in this thesis, but they will still be used to make comparisons. More details on INLA default priors can be found in Lindgren (2012).

The penalized complexity (PC) prior has been established as a suitable option for GRFs with Matérn covariance, and it's often used instead of the more complicated default priors. The PC prior attempts to reduce overfitting by penalizing complexity in the prior distribution (Hem, 2017). From Fuglstad et al. (2019) the PC prior for a GRF with Matérn covariance in two dimensions is given by

$$\pi(\theta, \rho) = \tilde{\lambda_1}\tilde{\lambda_2}\rho^{-2}\exp(-\tilde{\lambda_1}\rho^{-1} - \tilde{\lambda_2}\sigma), \quad \sigma > 0, \rho > 0, \tag{4.3}$$

where definitions $P(\rho < \rho_0) = \alpha_1$ based on the range ρ and $P(\sigma > \sigma_0) = \alpha_2$ based on the standard deviation σ can be found by setting

$$\tilde{\lambda_1} = -\log(\alpha_1)\rho \quad \text{and} \quad \tilde{\lambda_2} = -\frac{\log(\alpha_2)}{\sigma_0} \tag{4.4}$$

With little actual prior knowledge of the spatial autocorrelation effects of freshwater fish, it is hard to know what prior distributions to set on the range and standard deviation of the random field u. A prior median range of 2 degrees as chosen in Prior-model 2 is expected to capture short to medium-scale spatial variation relative to the domain that is Norway. Prior-model 3 is included to see how the model is affected if the random field is given a much larger prior range. Prior-model 4 is included in the same way to see the effects of reducing the prior standard deviation of u.

4.3 Model validation

Ideally one would have an additional data set of standardized data available to use as validation data, but a suitable data set is not available. As a substitute, the different model variations will be validated by cross-validation. This means that a fraction of the data is left out when fitting the model, and the model attempts to predict the left out data. To reduce the amount of spatial bias in this validation process the domain is divided in rectangular blocks, and the blocks are randomly assigned to five-folds. One fold at the time is chosen as the test set while the other four folds become the training set. Test set data from all data sets are withheld from the model, and the model is evaluated by its ability to predict the test data *from the Survey data.* Only the Survey data set is being predicted, as it is considered the most reliable data source. This is important since the model should be evaluated by its ability to reconstruct the species intensity without the sampling bias that is likely present in the presence-only data (Fithian et al., 2015). The validation statistic used is the deviance, defined as twice the log-likelihood of the model.

Chapter 5

Results

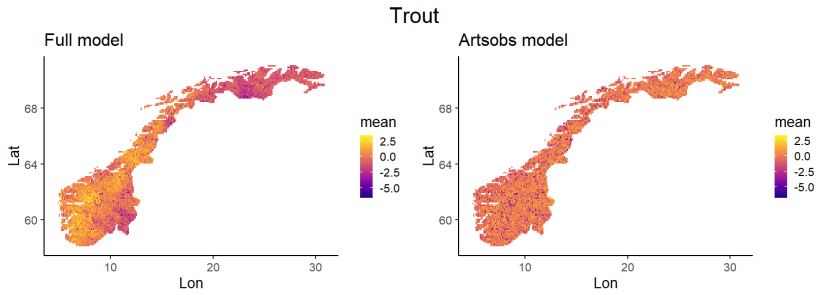
5.1 Artsobs as presence absence

Before any further modelling can be done, a choice needs to be made whether or not absences should be added to the Artsobs data set. As described in Chapter 2, these absences are added by assuming that a species is absent in a lake if it has not been observed there, while some other species has. A full model of all data sets where Artsobs has absences is compared to a similar full model where Artsobs has only presences. This comparison is done by cross-validation as described in Section 4.3, and the marginal predicted deviance is the validation statistic.

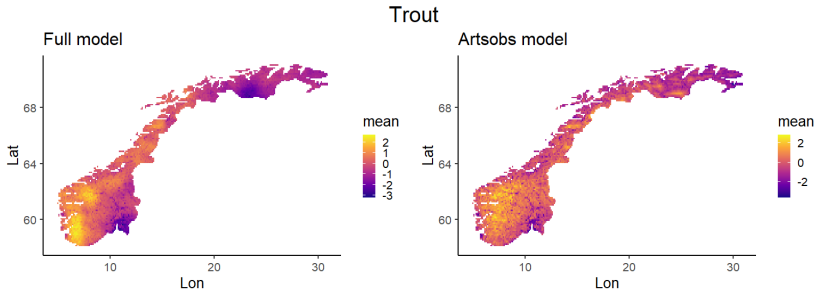
The results from the cross-validation is that the full model has a marginal predicted deviance of 466.7 when the Artsobs data is presence-only and 317.0 when the Artsobs data includes absences. The difference of 149.7 is significant and indicates that the model is improved by the addition of absences. Figure 5.1 shows a significant difference of modelling with and without absences. The model for Artsobs without absences shows little to no spatial pattern and looks more like noise, while the model with absences shows a clearer spatial effect. This is also reflected in the posteriors for the full model, where the model with absences estimates high intensity in the west and lower in the east, while the model without absences estimates a more flat intensity. Due to these results, the Artsobs data set will include the generated absences in all further analysis in this thesis.

5.2 Priors

Some Bayesian models can be very sensitive to the prior distribution and hyperparameters used. Because of this, examining the effects the prior distributions have on the posteriors is important. Four models were fit, from here on referred to as models 1-4, with prior distribution and hyperparameters as described in Table 4.1. All other parameters were kept equal for all four models, including modelling the



(a) Posterior means with Artsobs having presences only.



(b) Posterior means with Artsobs including pseudo-absences.

Figure 5.1: Posterior means of the estimated intensity $\lambda = \exp(\eta)$ of trout for the full model and the model based only on the Artsobs data set. Note that as the probability of detection is not known, the plots show relative intensity. The spatial fields are given the default prior distributions.

data from Artsobs as presence/absence, as concluded in the previous section. Due to time and computational constraints, this prior analysis was done using only data on the species trout. The marginal predicted deviances from doing cross-validation with the four models as described in Section 4.3 is shown in Table 5.1.

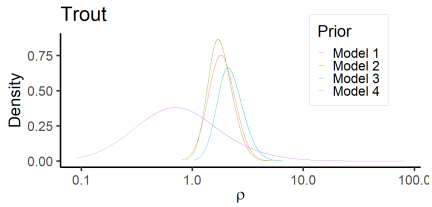
Parameter	Prior-model 1	Prior-model 2	Prior-model 3	Prior-model 4
Deviance	317.0	285.5	316.4	61.0

Table 5.1: The marginal predicted deviance values for the prior-models defined in Table 4.1. All values are obtained by cross-validation and predicting on the Survey data set.

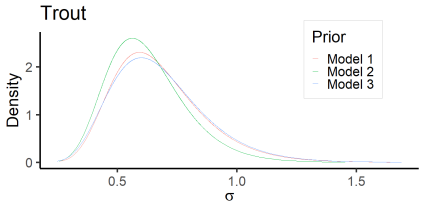
As the deviance depends on the data being predicted, the deviance values of the models relative to one another should be of interest, not the absolute value. As models with lower deviance is preferred, it is clear from the table that model 4 is the best model according to the cross-validation, with model 2 being the second best. The difference between the two models is that model 4 puts the PC(0.01, 0.1) prior on the standard deviation σ_u, forcing it to be much smaller than in model 2. It's clear from Figure 5.2c that the posterior standard deviation of model 4 has far less than 10% of the density at values larger than 0.01, suggesting that the data

pulls the posterior towards lower values.

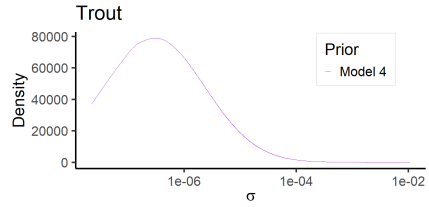
Figure 5.2a shows that the posterior ranges of models 1-3 are all quite similar with little difference in modes, even though the range in model 3 is given a much larger prior than model 2. This could indicate that the data prefers a lower range, which could also explain the higher deviance in model 3. The posterior range of model 4 has both a lower mode and a flatter curve than the other models. The flatter curve is likely due to the prior forcing a low variance on the spatial field, making the posterior of the range more vague to compensate. The density being non-zero at values closer to 100 is somewhat worrying, since at these values the random field is capturing effects from outside the domain of the data, which is not desired. However as the part of the density at these higher values is small, whether this is significant or not could be argued.



(a) Posterior distributions of the range ρ_u. The x-axis is shown on the log-scale.



(b) Posterior distributions of the standard deviation σ_u for prior-models 1-3.



(c) Posterior distributions of the standard deviation σ_u for prior-model 4. The x-axis is shown on the log-scale.

Figure 5.2: Posterior marginal distributions of the range ρ and the standard deviation σ of the random field u for the prior-models defined in Table 4.1 for the species trout.

Figure 5.3 shows the estimated regression coefficients for all the covariate fixed effects included in the models for trout. Similar plots for the other three species are included in Figure A.1, A.2 and A.3. Once again models 1-3 give very similar results, with only small differences in both point estimates and confidence intervals for the coefficients. Also here model 4 deviates from the rest, estimating stronger effects of longitude, latitude and distance to road. Combined with the results from Figure 5.2 this could indicate that model 4 is more heavily influenced by the fixed effects and relies less on the random field u, as compared to models 1-3. It is worth noting that model 4 gives notably smaller confidence intervals for the estimates of multiple covariate effects. In general it is also encouraging to see that all models are somewhat in agreement on the significance and magnitude of the fixed effects.

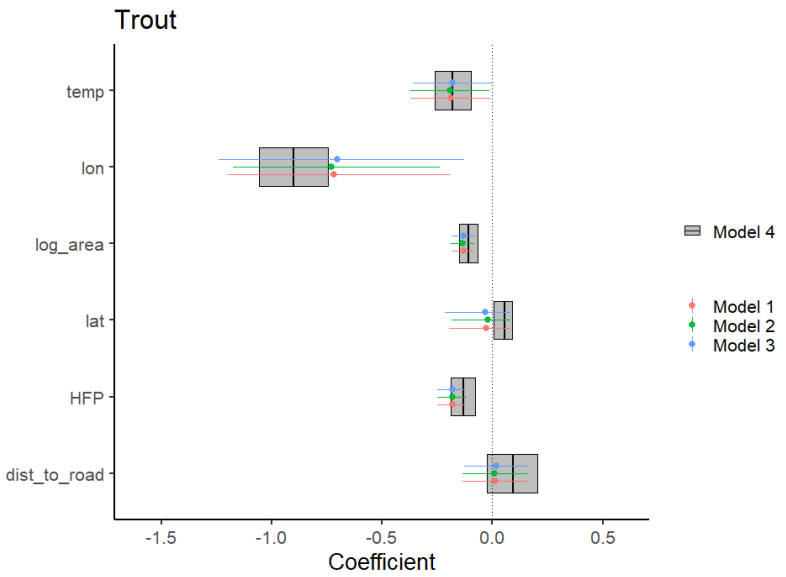


Figure 5.3: Posterior distributions of regression coefficients for fixed effects of the prior-models defined in Table 4.1 for the species trout, with posterior means and 95% confidence intervals.

Taking all the previous results into consideration, model 2 is similar to model 1 and 3, while also offering a notably lower deviance value. Thus the choice of the best prior-model for further analysis is between model 2 and 4. Model 4 having large variation in the posterior range and very low posterior standard deviation makes it somewhat hard to interpret. However, the smaller confidence intervals for the fixed effects and in particular the significantly lower deviance in comparison to the other models is desirable. Because of this, the priors of model 4 will be used in all the models of the next section.

5.3 Data integration

The main goal of this thesis is to examine the effect of combining multiple data sets to create one model, which hopefully models the species distribution better than models based on the individual data sets. In addition to models for the three individual data sets, full models were fit combining the data from all data sets. The prior distribution from Prior-model 4, being $\rho_u \sim PC(2, 0.5)$ and $\sigma_u \sim PC(0.01, 0.1)$, was used on the Gaussian random field u in all models. The models were validated by cross-validation as described in Section 4.3. It is important to note that due to time and computational constraints there has been no attempt at fitting models combining only two data sets.The results of the validation can be seen in Table 5.2.

Notably, the full model outperformed the individual models in all cases except for trout, where the Survey model had better results. Survey outperformed Gillnet for trout and perch, with it being the other way around for char and minnow. The Survey model was notably worse at predicting the minnow species, which will be looked at more later. In general it is difficult to conclude which of the individual models did best based on this validation.

It is however evident that the model based only on Artsobs data is by far the worst, having the highest deviance value for all species. This result is not surprising, as this is the only data set consisting of non-standardized data. This shows that the non-standardized data is still effected by sampling bias, even after adding pseudo-absences and the second spatial field ζ.

Parameter	Survey	Gillnet	Artsobs
Trout	-11.1	14.2	287.0
Char	18.2	2.6	663.0
Perch	18.4	31.0	83.1
Minnow	127.2	15.8	947.5

Table 5.2: Difference in marginal predicted deviance from the full model for each of the three individual data set models, for each of the four species of interest. All deviance values are obtained by cross-validation and predicting on the Survey data set.

The posterior mean and standard deviation of the Gaussian random fields u and ζ used in the full model for trout is shown in Figure 5.4. Similar plots for the other three species can be found in Figure A.5, A.6 and A.7 in the Appendix. The posterior mean of u is low in the south-east compared to the west, which is somewhat reflected in the posterior mean of the intensity in Figure 5.5. The standard deviation of u is however very high compared to the mean, indicating that the model might be influenced more by the fixed effects than the random field.

Figure 5.5 shows the the posterior means of the estimated intensity of trout. All models show trout in most of Norway, with a lower mean in Finnmark in the north-east and in the south-east around Oslo. All four posterior means look reasonably similar, although there is some difference, mainly the mean of the Survey model being a bit more even than that of the other models.

Figure 5.6 shows that outside of a few differences, the four models mostly agree on the distributions of regression coefficients for fixed effects when modelling trout. Once again the Artsobs model looks the weakest with large confidence intervals, while the full model usually has the smallest intervals. All models agree on negative effects of longitude, likely due to the lower density of trout in the north-eastern parts of Norway.

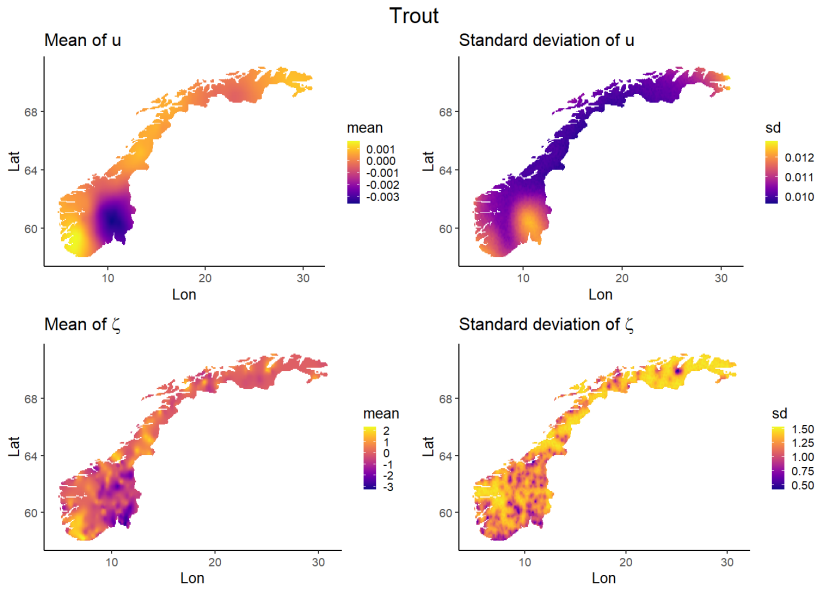


Figure 5.4: Posterior mean and standard deviation of the random fields u and ζ for the trout species.

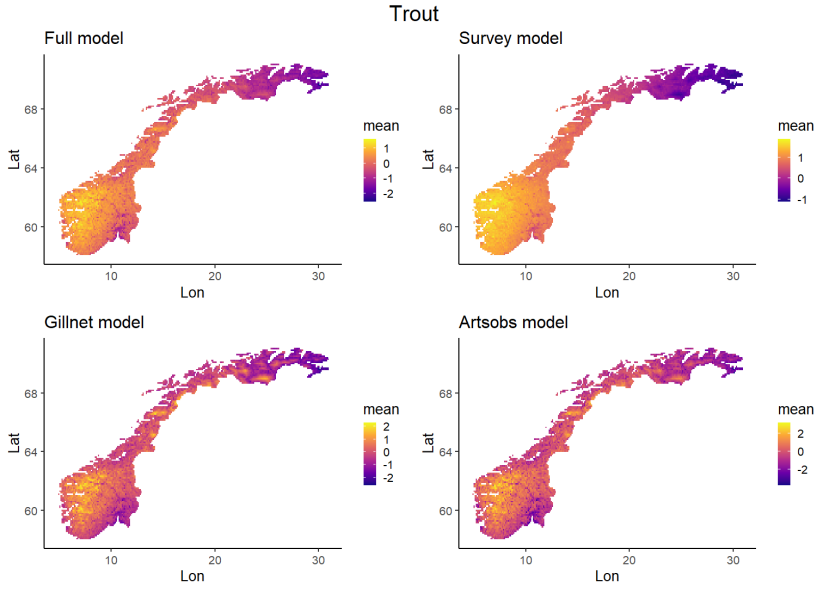


Figure 5.5: Posterior means of the estimated intensity $\lambda = \exp(\eta)$ of trout for the full model and the three individual data set models. Note that as the probability of detection is not known, the plots show relative intensity.

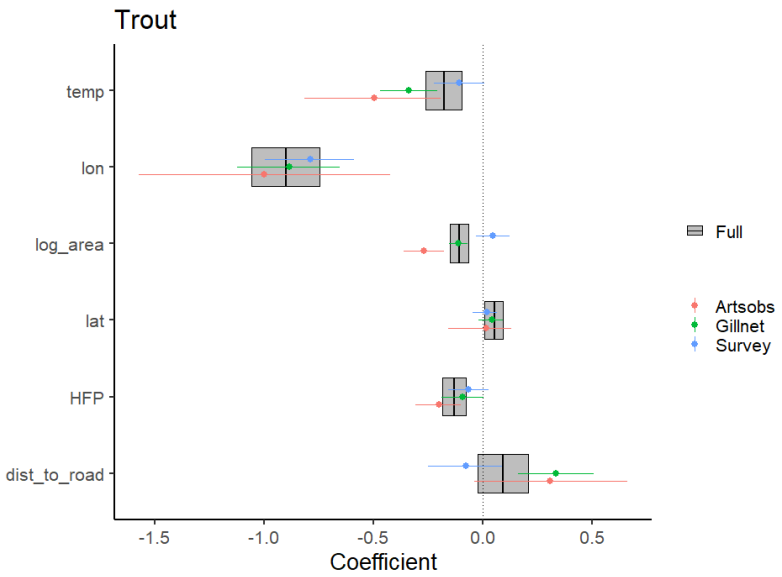


Figure 5.6: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species trout.

Figure 5.7 shows that the Artsobs model estimates the intensity of arctic char to be moderately high in most of the country, although higher in the north. The other three models show a much lower intensity in the southern half of Norway, more clearly showing that although the species exists in the south, it is way more prevalent in the north. The Artsobs model not being able to catch this effect is likely because the non-standardized data is biased towards the larger cities and more populated areas in the south. This can also explain the reason for the weak validation result of the Artsobs model when predicting char, as shown in Table 5.2.

From Figure 5.8 it can be seen that the full model indicates a strong positive effect of longitude on arctic char, corresponding to the higher intensity in the north-eastern part of Norway. The negative effect of latitude and positive effect of temperature is surprising however, as arctic char is known to usually appear in colder areas in the north. These unexpected effects are likely because temperature and latitude are negatively correlated, as shown in Section 2.5.

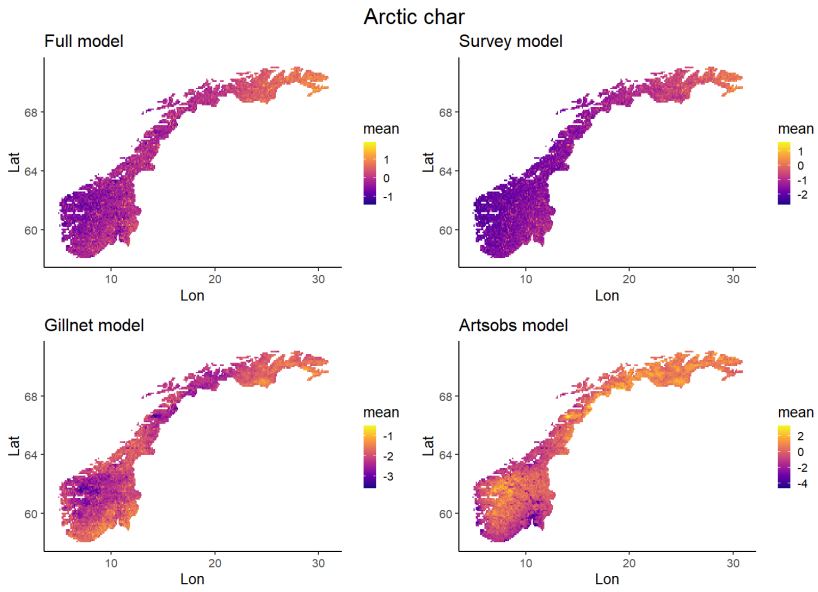


Figure 5.7: Posterior means of the estimated intensity $\lambda = \exp(\eta)$ of arctic char for the full model and the three individual data set models. Note that as the probability of detection is not known, the plots show relative intensity.

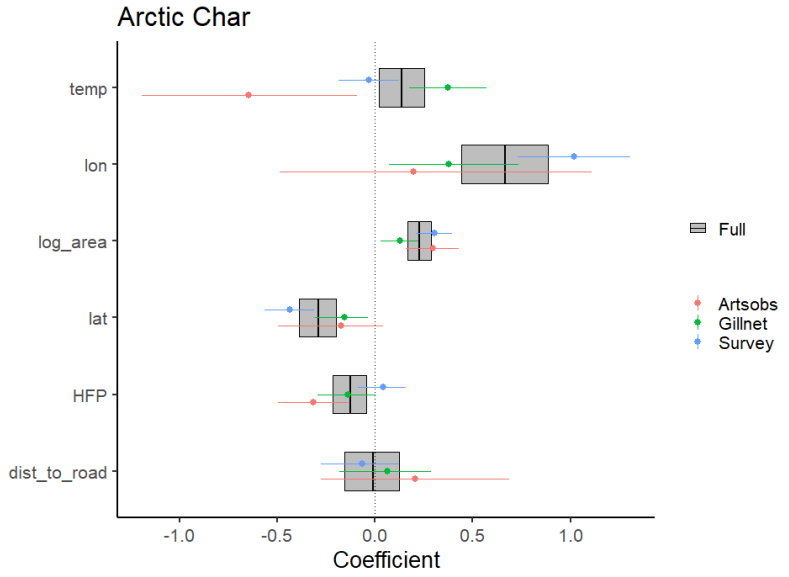


Figure 5.8: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species arctic char.

The data in Figure 2.4 show that Perch is mainly found in the south-east of Norway, and this is clearly reflected in the plot of the full model in Figure 5.9. While the Artsobs model estimates a higher density in the south-east as expected, the scale of the mean is very large compared to the other models. This is explained by Artsobs estimating way larger coefficients for the fixed effects than the other models, shown in Figures 5.10 and A.13. The Survey and Gillnet models seem to be heavily affected by lake area, creating the darker dots in the figure. The plot of the fixed effects reflect this, where especially Gillnet models a very large negative effect of lake area. This effect is difficult to explain, and it can be seen that the full model estimates the effect of area to be zero.

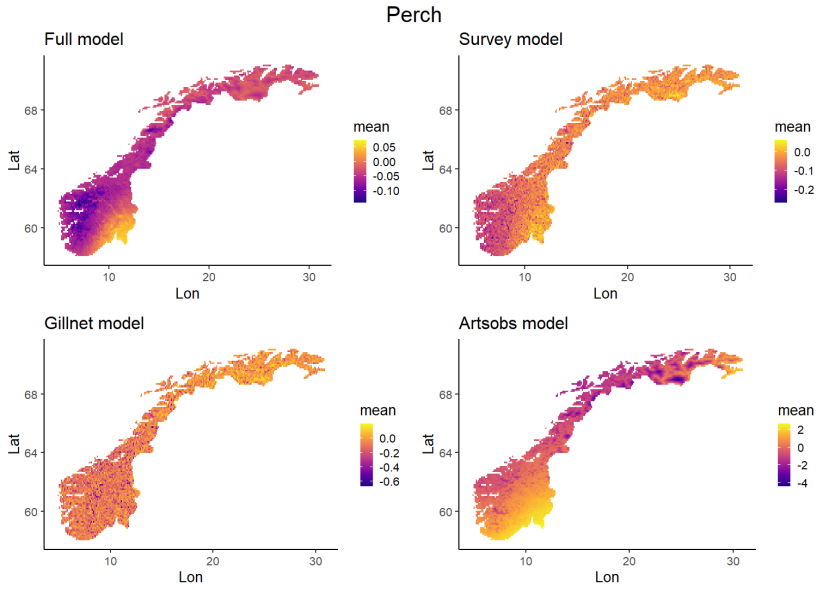


Figure 5.9: Posterior means of the estimated intensity $\lambda = \exp(\eta)$ of perch for the full model and the three individual data set models. Note that as the probability of detection is not known, the plots show relative intensity.

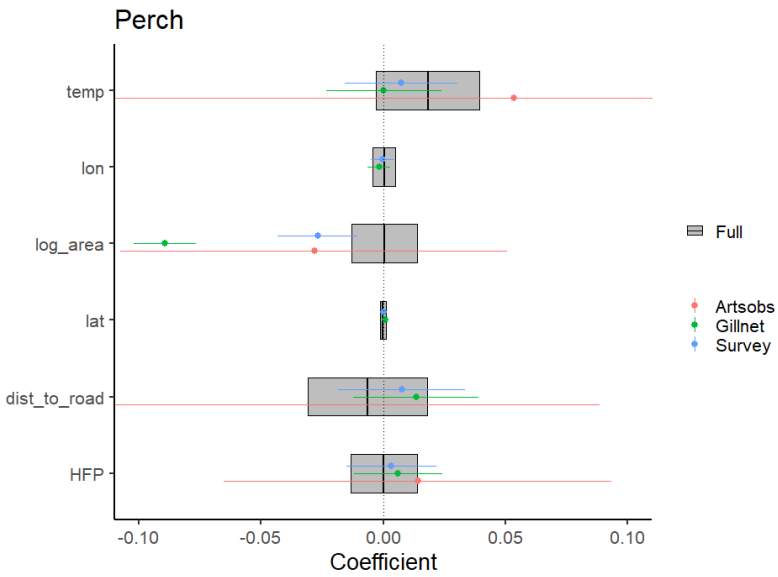


Figure 5.10: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species minnow. The plot is zoomed in around zero to better see the values close to zero.

There are clearly some issues with modelling minnow, especially for the Survey model. The scale of the intensity is extremely large compared to the other models, and Figure 5.12 shows that the model estimates large coefficients for the fixed effects compared to the other three models. The problems with modelling minnow is likely a combination of the correlation between the covariates and a lack of data, as minnow is the least reported species of the four examined here.

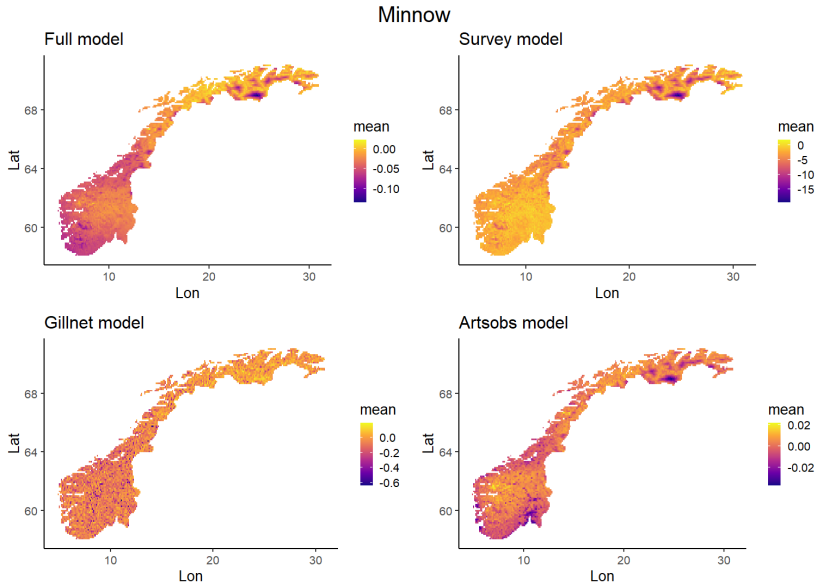


Figure 5.11: Posterior means of the estimated intensity $\lambda = \exp(\eta)$ of minnow for the full model and the three individual data set models. Note that as the probability of detection is not known, the plots show relative intensity.

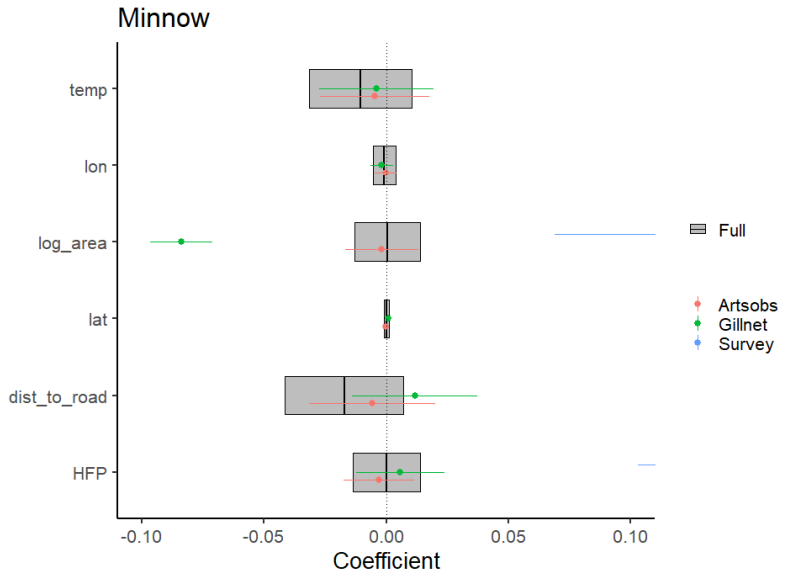


Figure 5.12: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species minnow. The plot is zoomed in around zero to better see the values close to zero.

Chapter 6

Discussion and conclusion

The goal of this thesis was to implement a model that combines data from multiple data sets in order to estimate the distribution of freshwater fish in Norway. The goal was also to compare the results from this combined model to models based on individual data sets, to examine the value of data integration for freshwater fish data and species distribution modelling in general. For most species the full model presented in this thesis estimates the species distributions better than any of the individual models, and the data integration is seen as a success.

It is shown that most of the covariate effects used in the model are significantly correlated, and this affects the results. Correlated covariates provide less relevant information to the model, and the regression coefficients of the correlated covariates tend to have higher variance. Because of this it is more difficult to know which covariates affect the species distribution, which can lead to issues when applying the model to other data because of potential overfitting. The model could possibly be improved by removing some of the correlated covariates or replacing them with other less correlated covariates.

For some species the fixed effects estimates are very unstable for models based on individual data sets, while the estimates of the combined model are more stable with shorter confidence intervals. This might indicate that the data integration helps create clearer and more stable estimates, although more research is likely needed to confirm that this effect is due to the data integration.

The prior analysis done indicates that the model is sensitive to the prior distributions chosen on the parameters of the Gaussian random field. It is shown that the model performs significantly better when the standard deviation of the field is given a smaller prior distribution, but this also drastically changes the posterior range of the field. It is interesting to see that the model performs best when the spatial field is very vague. This could indicate that the field and the fixed effects give similar explanations, and that the model performs better when the spatial field is used to explain small scale autocorrelation only. These results show that it is important to examine the effects of different hyperparameter values for both the

range and the standard deviation, and perhaps also other prior distributions like the Log-Gaussian prior used by Hem (2017).

It is very unfortunate that the modelling the Gillnet data using the Poisson distribution did not succeed, as there is likely valuable information to be gained from the point count data. While the Gillnet data contributed to the full integrated model as a presence/absence data set, more future work should be aimed at modelling the point counts correctly to improve the model further.

The value of adding absences to the non-standardized data set from Artsobservasjoner was measured, and the addition of absences clearly improved the model. While this means that all three data sets were modelled as presence-absence data, the weakness of non-standardized data is still visible in the results. The strength of non-standardized data is usually the larger amount of available data, but as the data sets used in this thesis are of similar size that strength is not a part of this model. Adding more of this type of data would therefore be useful to improve the model and utilise the non-standardized data better.

Bibliography

Dorazio, R.M., 2014. Accounting for imperfect detection and survey bias in statistical analysis of presence-only data. Global Ecology and Biogeography 23, 1472–1484. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/geb.12216, doi:10.1111/geb.12216, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/geb.12216.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6, 424–438. doi:10.1111/2041-210X.12242.

Fuglstad, G.A., Simpson, D., Lindgren, F., Rue, H., 2019. Constructing priors that penalize the complexity of gaussian random fields. Journal of the American Statistical Association 114, 445–452. doi:10.1080/01621459.2017.1415907.

Hem, I.G., 2017. A statistical approach to spatial mapping of temperature change URL: http://hdl.handle.net/11250/2450462.

Isaac, N.J., Jarzyna, M.A., Keil, P., Dambly, L.I., Boersch-Supan, P.H., Browning, E., Freeman, S.N., Golding, N., Guillera-Arroita, G., Henrys, P.A., Jarvis, S., Lahoz-Monfort, J., Pagel, J., Pescott, O.L., Schmucki, R., Simmonds, E.G., O'Hara, R.B., 2019. Data integration for large-scale models of species distributions. Trends in Ecology Evolution doi:https://doi.org/10.1016/j.tree.2019.08.006.

Lindgren, F., 2012. Continuous domain spatial models in r-inla. The ISBA Bulletin 19. URL: http://www.r-inla.org/examples/tutorials/spde-from-the-isba-bulletin.

Lindgren, F., Rue, H., Lindström, J., 2011. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 423–498. doi:10.1111/j.1467-9868.2011.00777.x.

Metz, M.; Rocchini, D.N.M., 2014. Surface temperatures at the continental scale: Tracking changes with remote sensing at unprecedented detail. Remote Sens. 6, 3822–3840. doi:https://doi.org/10.3390/rs6053822.

Miller, D.A.W., Pacifici, K., Sanderlin, J.S., Reich, B.J., 2019. The recent past and promising future for data integration methods to estimate species' distributions. Methods in Ecology and Evolution 10, 22–37. doi:10.1111/2041-210X.13110.

Møller, J., Waagepetersen, R., 2004. Statistical Inference and Simulation for Spatial Point Processes. New York: Chapman and Hall/CRC. doi:https://doi.org/10.1201/9780203496930.

Pethon, P., Vøllestad, A., 2019. Store norske leksikon: Ferskvannsfisk i norge. URL: https://snl.no/ferskvannsfisk_i_Norge.

Rue, H., Held, L., 2005. Gaussian Markov Random Fields. Theory and Applications. Chapman and Hall.

Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71, 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Symmonds, Emily G.; Jarvis, S.G.H.P.A.I.N.J.B.O.R.B., 2020. Is more data always better? a simulation study of benefits and limitations of integrated distribution models .

Venter, O., Sanderson, E., Magrach, A.e.a., 2016. Sixteen years of change in the global terrestrial human footprint and implications for biodiversity conservation. Nature Communications 7. doi:https://doi.org/10.1038/ncomms12558.

Appendix

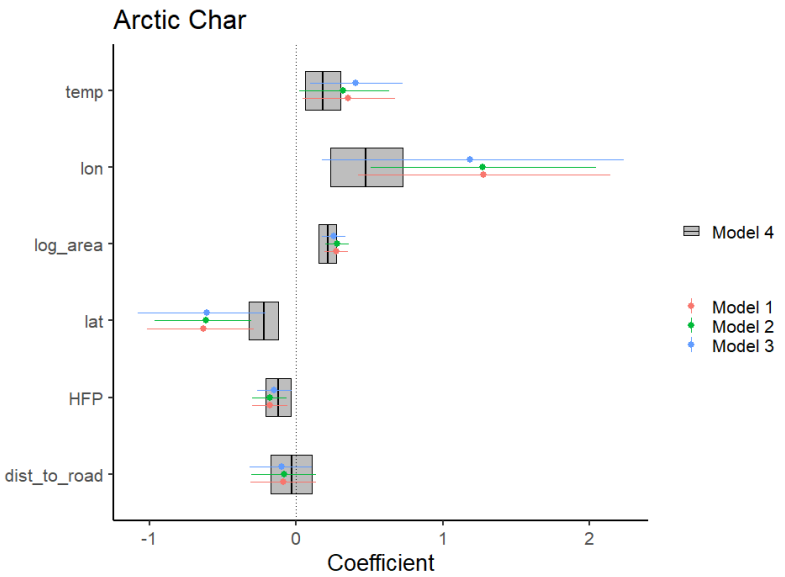


Figure A.1: Posterior distributions of regression coefficients for fixed effects of the prior-models defined in Table 4.1 for the species arctic char, with posterior means and 95% confidence intervals.

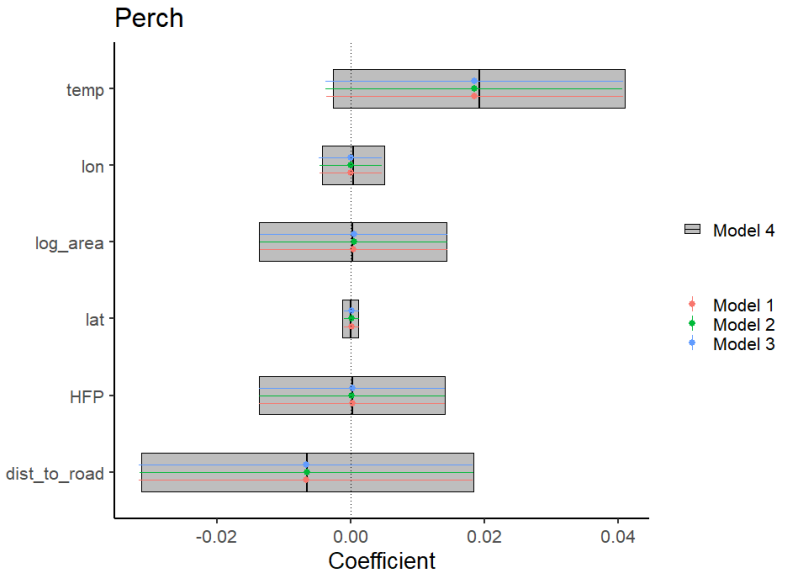


Figure A.2: Posterior distributions of regression coefficients for fixed effects of the prior-models defined in Table 4.1 for the species perch, with posterior means and 95% confidence intervals.

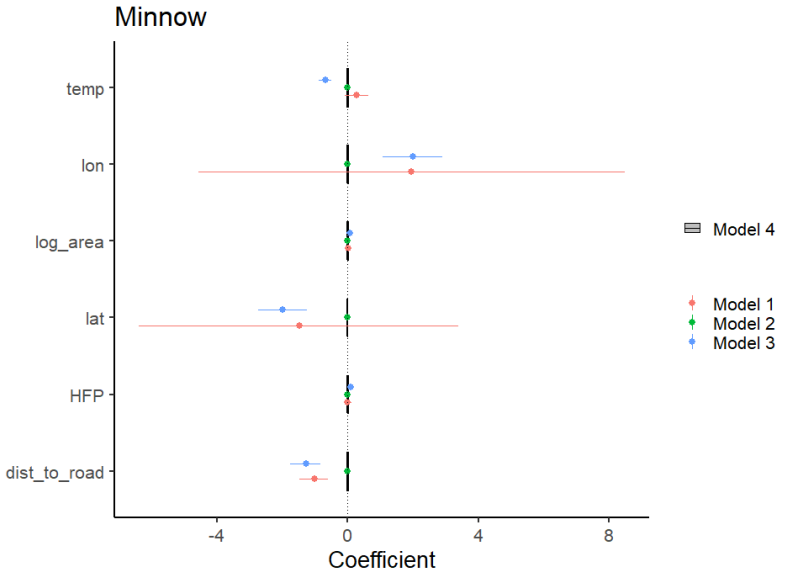


Figure A.3: Posterior distributions of regression coefficients for fixed effects of the prior-models defined in Table 4.1 for the species minnow, with posterior means and 95% confidence intervals.

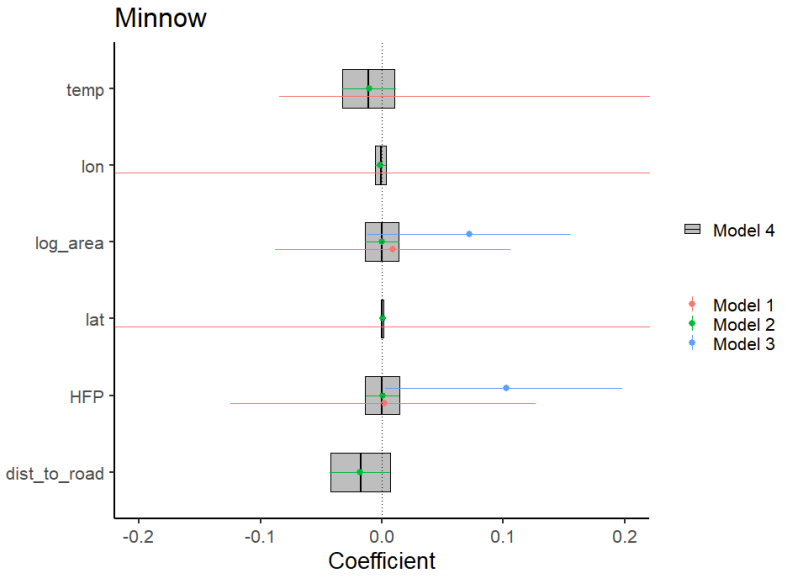


Figure A.4: Posterior distributions of regression coefficients for fixed effects of the prior-models defined in Table 4.1 for the species minnow, with posterior means and 95% confidence intervals. The plot is zoomed in around zero to better see the values close to zero.

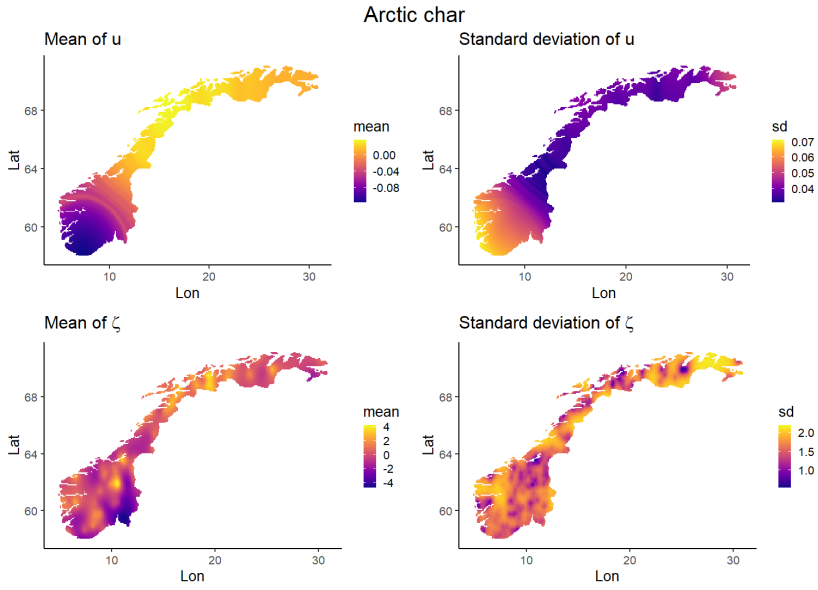


Figure A.5: Posterior mean and standard deviation of the random fields u and ζ for the arctic char species.

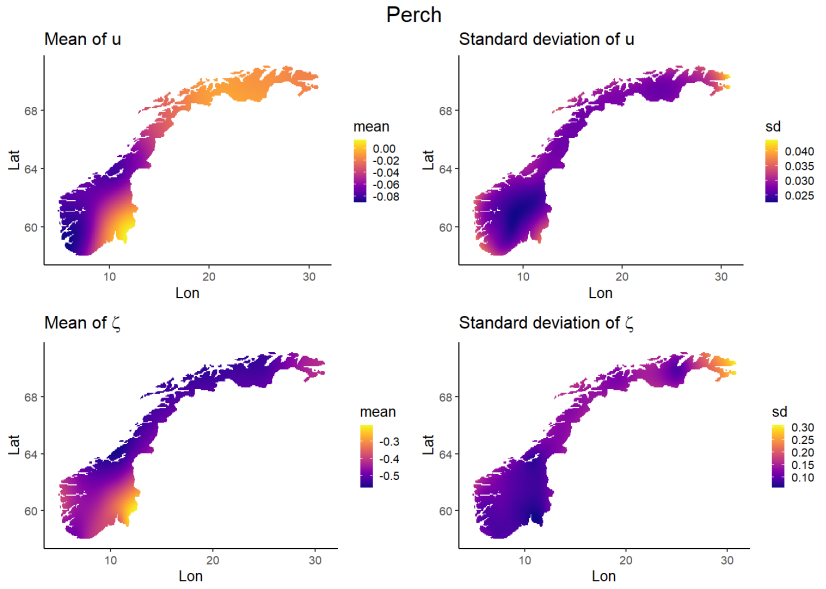


Figure A.6: Posterior mean and standard deviation of the random fields u and ζ for the perch species.

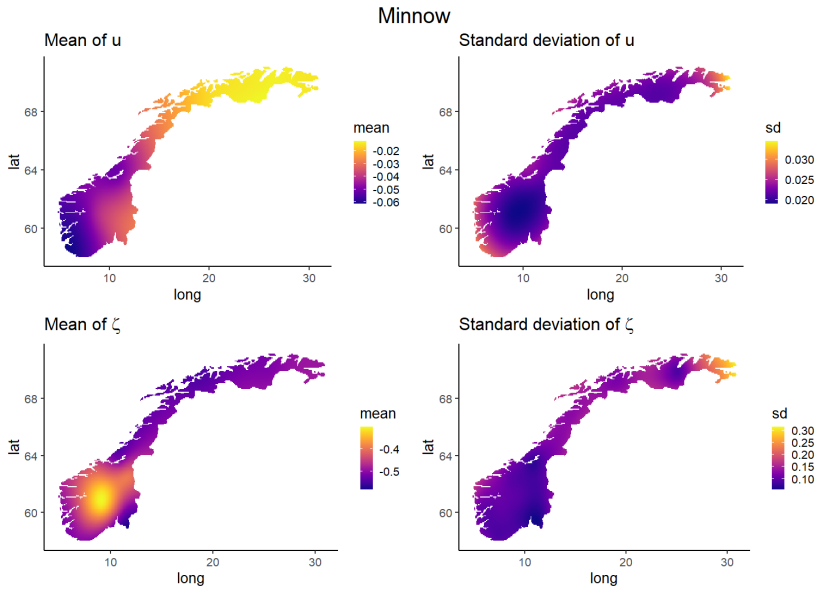


Figure A.7: Posterior mean and standard deviation of the random fields u and ζ for the minnow species.

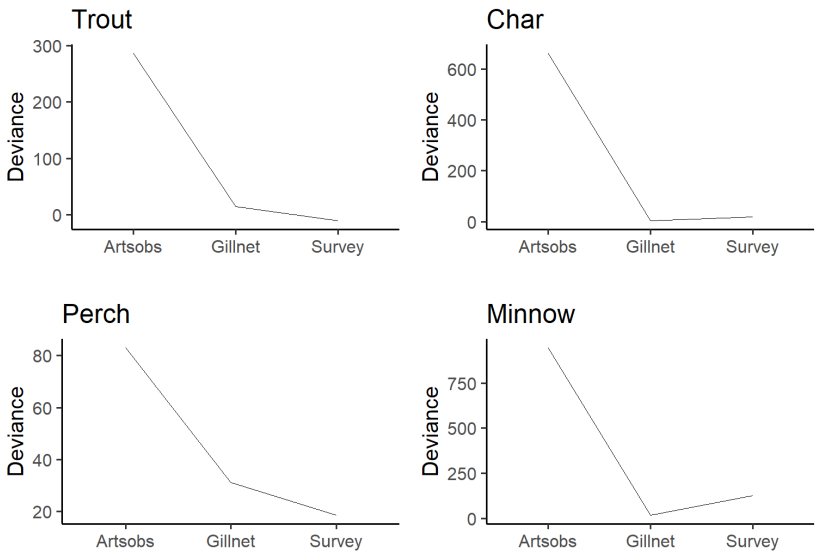
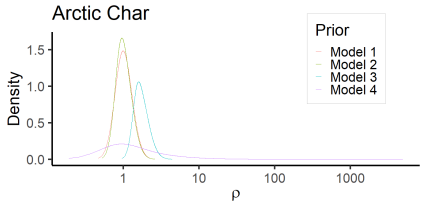
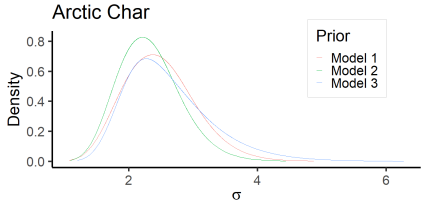


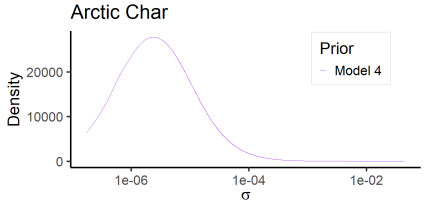
Figure A.8: Difference in marginal predicted deviance from the full model for each of the three individual data set models, for each of the four species of interest. All deviance values are obtained by cross-validation and predicting on the Survey data set.



(a) Posterior distributions of the range ρ_u.

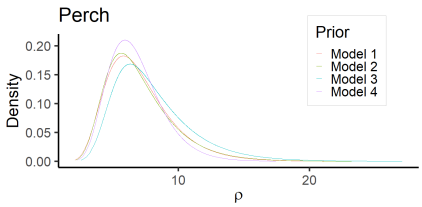


(b) Posterior distributions of the standard deviation σ_u for prior-models 1-3.

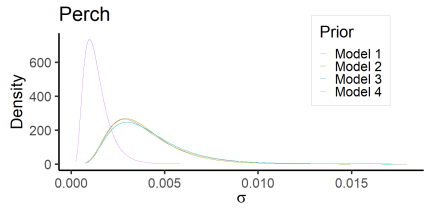


(c) Posterior distributions of the standard deviation σ_u for prior-model 4.

Figure A.9: Posterior marginal distributions of the range ρ and the standard deviation σ of the random field u for the prior-models defined in Table 4.1 for the species arctic char.

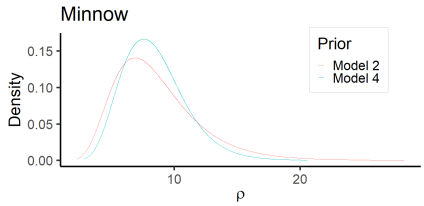


(a) Posterior distributions of the range ρ_u.

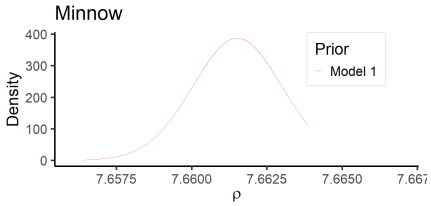


(b) Posterior distributions of the standard deviation σ_u.

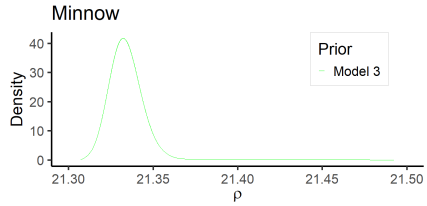
Figure A.10: Posterior marginal distributions of the range ρ and the standard deviation σ of the random field u for the prior-models defined in Table 4.1 for the species perch.



(a) Posterior distributions of the range ρ_u for prior-models 2 and 4.

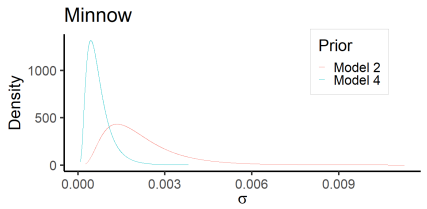


(b) Posterior distributions of the range ρ_u for prior-model 1.

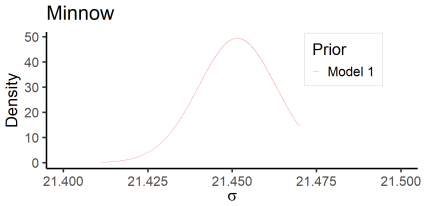


(c) Posterior distributions of the range ρ_u for prior-model 3.

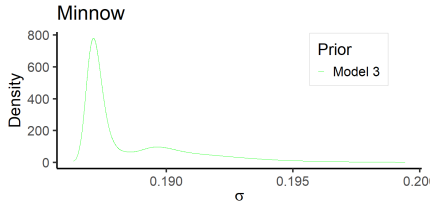
Figure A.11: Posterior marginal distributions of the range ρ of the random field u for the prior-models defined in Table 4.1 for the species minnow.



(a) Posterior distributions of the standard deviation σ_u for prior-models 2 and 4.



(b) Posterior distributions of the standard deviation σ_u for prior-model 1.



(c) Posterior distributions of the standard deviation σ_u for prior-model 3.

Figure A.12: Posterior marginal distributions of the standard deviation σ of the random field u for the prior-models defined in Table 4.1 for the species minnow.

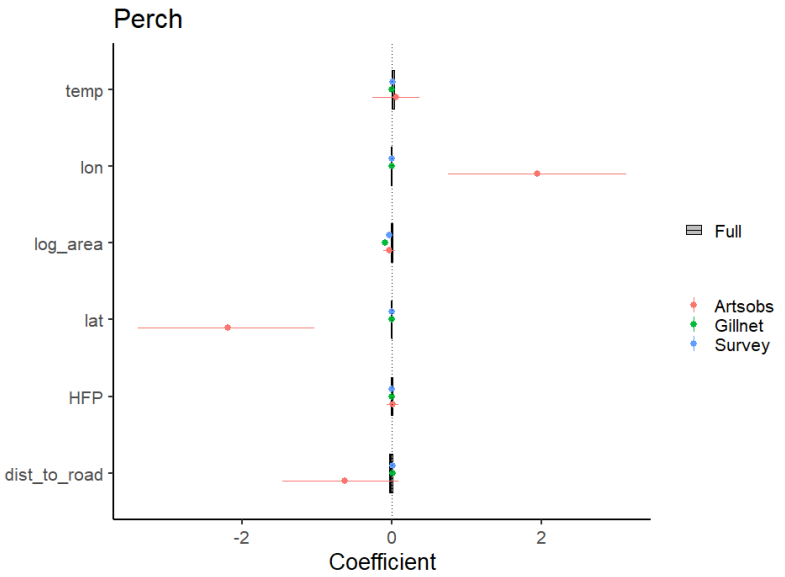


Figure A.13: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species minnow.

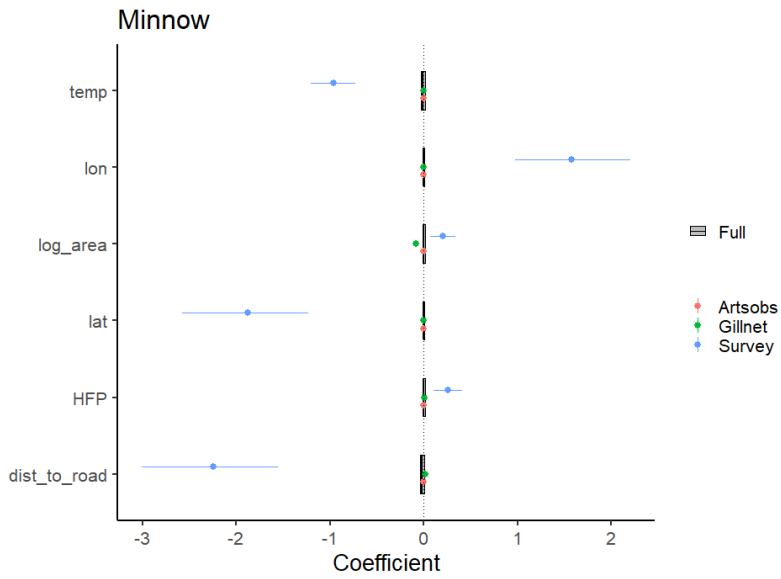
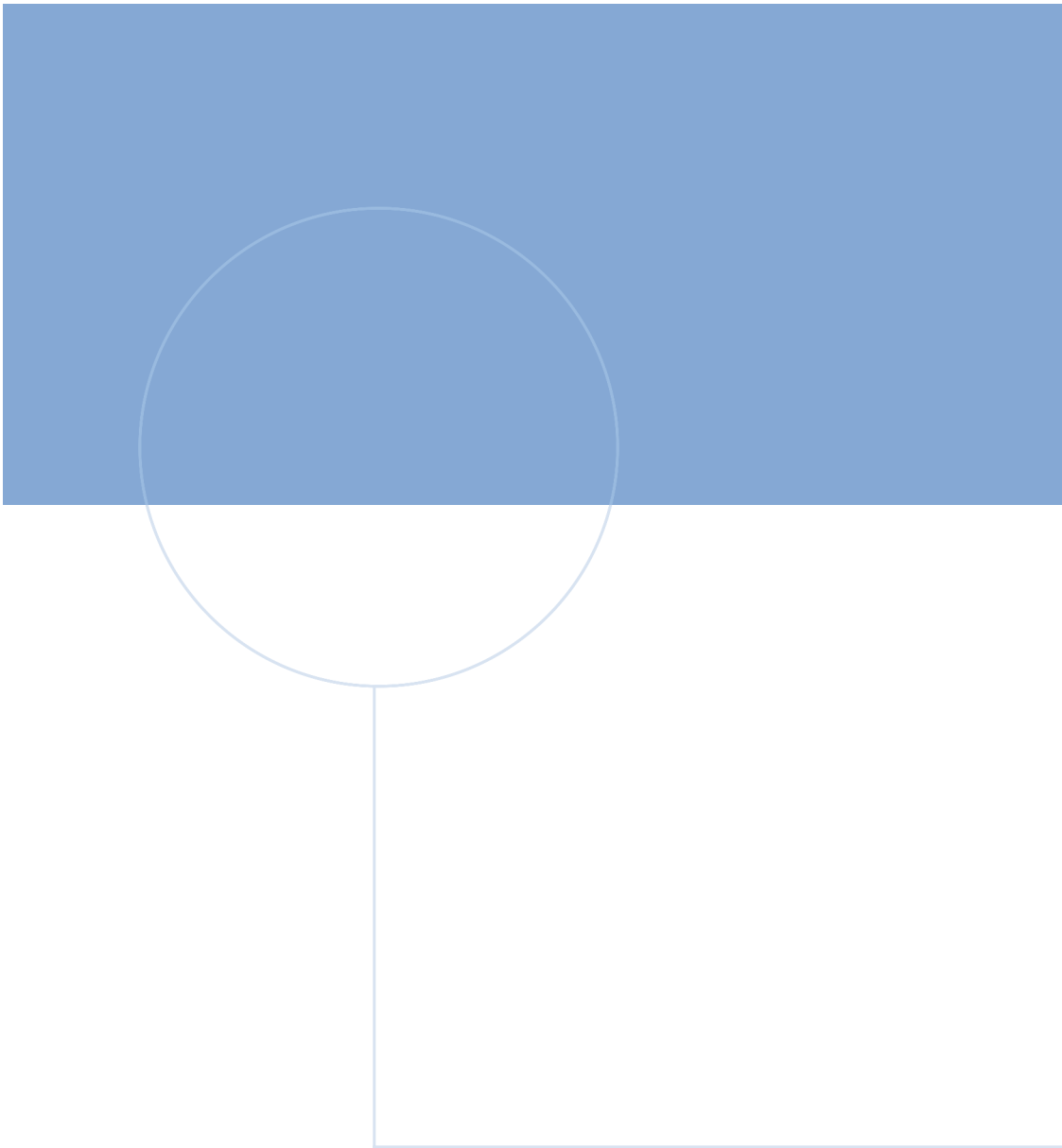


Figure A.14: Posterior distributions of regression coefficients for fixed effects of the models using individual data sets, as well as the full model in grey shading, with posterior means and 95% confidence intervals. All models are for the species minnow.



NTNU

Norwegian University of
Science and Technology