Master's thesis

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

Emma Sofie Skarstein

Accounting for spatial bias in citizen science observations of Norwegian freshwater fish by using an effort spatial field

Master's thesis in Mathematical Sciences
Supervisor: Robert Brian O'Hara
July 2020

NTNU
Norwegian University of Science and Technology

Emma Sofie Skarstein

Accounting for spatial bias in citizen science observations of Norwegian freshwater fish by using an effort spatial field

Master's thesis in Mathematical Sciences
Supervisor: Robert Brian O'Hara
July 2020

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

NTNU
Norwegian University of
Science and Technology

Preface

This thesis is the assignment for the course MA3911 *Master Thesis in Mathematical Sciences*.

I first want to thank Professor Robert O'Hara for his patience and very thoughtful feedback and support throughout the whole thesis. In particular, I want to thank him for including me in his community, through weekly group meetings with the other people he also supervises. This really became a highlight of the week, especially due to everything closing down as a result of the Covid-19. The regularity of those meetings, and the feedback and support of everyone in this group has been greatly appreciated, and I think my thesis would be a lot worse off without it.

I also want to thank Lyder Bøe Iversen, who has been completing his masters thesis in parallel with me, and looking at the same data. Being able to exchange thoughts and experiences, especially as we were both new to a lot of the topics here, was really, really helpful.

Lastly, thanks to my (extended) family for criticizing my language and commenting on my ecological incompetence, and to Christoffer for letting me discuss all my problems and being so supportive and sending me angry looks whenever I was procrastinating.

Emma Skarstein

Trondheim, July 2020

Abstract

Model-based data integration provides a promising framework for fitting species distribution models using citizen science data together with structured survey data, but a common challenge is how to properly include biased citizen science data in an integrated model.

I implement an integrated species distribution model using two data sets of freshwater fish in Norway: one which is a structured survey data set and one which is a citizen science data set. For the underlying distribution, I use a log-Gaussian Cox-process. Together with this, I assume separate observation processes for each data set, but with shared environmental covariates and a shared spatial field. In addition, the observation process for the citizen science data is given a separate spatial field which is estimated only from the citizen science data, referred to as the effort spatial field. This allows us to estimate the spatial bias of these observations.

By comparing the estimated separate spatial field across four different species of freshwater fish, we see that even in fish with very different distributions, the effort spatial field is very similar. When comparing variations of integrated models to a survey-only model, the integrated models perform consistently better than the single-dataset model.

The integrated nested Laplace approximation (INLA) methodology is used to fit all models, and gives great flexibility as well as very efficient computation.

Sammendrag

Modellbasert dataintegrasjon gir et lovende rammeverk for konstruksjon av artsfordelingsmodeller ved bruk av folkeforsknings-data sammen med strukturerte data fra undersøkelser, men en vanlig utfordring er hvordan man skal forholde seg til romlige skjevheter i folkeforsknings-dataene i slike modeller.

Jeg implementerer en integrert artsfordelingsmodell ved å bruke to datasett med observasjoner av ferskvannsfisk i Norge: et strukturert datasett, og et folkeforskningsdatasett. Det antas en log-Gaussisk Cox-prosess for den underliggende fordelingen til dataene. I tillegg antas det individuelle observasjonsprosesser for hvert datasett, men med felles miljømessige kovariater og et felles romlig felt. Observasjonsprosessen for folkeforsknings-dataene blir også gitt et eget romlig felt som estimeres fra folkeforsknings-data alene. Dette lar oss estimere den romlige skjevheten til disse observasjonene.

Ved å sammenligne dette estimerte separate romlige feltet på tvers av fire forskjellige arter av ferskvannsfisk, ser vi at selv i fiskearter med svært forskjellige fordelinger, er det romlige feltet veldig likt. Når vi sammenligner varianter av integrerte modeller med en modell basert kun på undersøkelses-datasettet, yter de integrerte modellene konsekvent bedre enn modellen med bare ett datasett.

All inferens er utført med metodikken "Integrated nested Laplace approximation" (INLA), som gir god fleksibilitet og effektiv utregning.

Contents

Chapter 1

Introduction

Citizen science species observations are becoming more and more interesting as a source of data in ecology, as there are both more efforts to gather citizen science data, as well as more ways of accessing it, for example through the open database GBIF.[1] The term *citizen science data* is broad, even within ecology. In some cases, data may be reported through a website, such as the Norwegian Species Observation Service, Artsobservasjoner.[2] There are also several mobile applications that seek to make it easier to report various species, notably iNaturalist for species in general and eBird for birds. The more specific Bumble Bee Watch seeks to track and conserve North America's bumble bees. For a more local example, the app Målerjakt[3] seeks to track the northward spread of the scarce umber moth (*Agriopis aurantiaria*, gul frostmåler). This is of great interest since similar species tend to cause great damage to many tree species in Northern Norway, and this particular species has never been observed further north than Troms. There are also more organized types of citizen science projects, such as the North American Breeding Bird Survey,[4] where participants are asked to walk a specific route and record all the birds they observe along it.

As is clear from these examples, the purpose of citizen science data collection can be anything from engaging people in the nature around them, to collecting data for very specific tracking and conservation efforts. For some purposes citizen science data works very well, for instance in the moth example from above. This particular moth species is easy to distinguish from other moths, it is important to track its

[1] `https://www.gbif.org/`
[2] `https://www.artsobservasjoner.no/`
[3] `https://www.malerjakt.no/`
[4] `https://www.pwrc.usgs.gov/bbs/`

spread north, but sending individual researchers out would be expensive. Provided one is able to engage the locals, this seems like a perfect solution, because we are only interested in knowing about single presences. But for more general ecological purposes, there are several challenges with this type of data.

First of all, citizen science data is often opportunistic. The citizen scientist will commonly only report the species they happen to come across, and realistically not even all of them. Often they will report a species that was out of the ordinary, since reporting all the common species would be quite uninteresting, and so rarer species may be over-represented when looking at citizen science species observations as a whole. People may also easily misclassify species they are less familiar with. They will also report only species that happen to be at the location they have gone to, instead of following a systematic observation procedure, such as observing at a set of random locations. This results in sites near cities, landmarks and roads being over-represented, and so the observations will be spatially biased. It also means that we only have information about when a species is *present*, and the fact that a species is not reported in a given location does not mean that the species is absent, it *could* mean that; or it could mean that it was just not observed and recorded there. This is an important distinction in many ecological models, and has been (and still is) the topic of many studies, see for instance Hastie and Fithian (2013).

The main advantages of citizen science data is that it is usually not very costly to collect, and that we often have large quantities of it. But there is a large variety of citizen science data types and collection procedures, and both the advantages and disadvantages will vary widely based on this. Although attempts have been made to assess the quality of citizen science data in a larger sense (see e.g. Kosmala et al. (2016)), it is hard to say anything general due to the wide variety of collection methods as well as project aims. Kosmala et al. (2016) also point out that biases that scientists are aware of in citizen science data might often be present in survey or more organized data as well.

Overall, the aim of this thesis is to develop a model that combines structured species observations (in an organized survey with both presences and absences reported) with citizen science data (which gives only the presences of a species), in order to produce better predictions than we would have if we used only structured data. Ideally, this approach will enable us to take advantage of the best aspects of both data sets: the survey data may have higher quality and more information in the sense that it is presence/absence data, but the citizen science data is cheaper to collect and will in many cases be more plentiful. This motivates the idea of creating a model that combines the two types of data. There are several aspects and challenges to this.

Firstly, we need to know how to model single data sets, which will differ depending

on the type of data we have at hand. Topics like spatial models and species distribution models tie into this. Secondly, we need to have some framework for combining models of different data types. This motivates using what is referred to as *integrated distribution models* (or model-based data integration).

Species distribution models are a large class of ecological models that seek to describe or predict the distribution of some species across a geographic space, using environmental explanatory variables. Originally the interest was often in understanding ecological connections, but lately the interest has to some degree shifted more towards prediction into the future or onto some other new space (Elith and Leathwick, 2009). This especially relates to how we can expect species' distributions to change as temperatures rise in connection to global warming, which is of great interest for many parties.

An extension from the basic species distribution model is to construct a model from two or more data sets reporting on the same species. There are many ways to do this. In this thesis, I will be using an integrated model. This approach allows for sharing parameters across the different sub-models representing the individual data sets, and thereby better capturing the underlying distribution as well as taking into account potential biases in each of the individual data sets (Isaac et al., 2020; Miller et al., 2019). This becomes particularly useful when dealing with citizen science data.

In a recent simulation study, Simmonds et al. (2020) compare different integrated models fit on simulated data sets representing a structured data set and an unstructured citizen science data set. They compare a variety of models, some based on individual data sets and others based on both data sets. They also introduce the concept of including a separate spatial field informed only by the citizen science data, in an attempt to capture the bias of the data. This gave significant model improvements, and this thesis will be among the first attempts to recreate this using real data. Thus one of the most important points of discussion will be how this has contributed to the model performance, and how this can be explored in future applications.

Specifically, I will be looking at observations made of three freshwater fish species both from a 1996 survey of approximately 800 lakes in Norway, in addition to citizen science data from Artsobservasjoner of the same fish species. Following the approach of Isaac et al. (2020), I will use a model formulation that explicitly separates the biological and data generation processes, emphasizing the different observation processes that generate the different data types. I will be modeling them in an integrated model using binomial regression with a cloglog link for the presence/absence survey data, and a Poisson regression with a log link for the presence-only citizen science data. In addition to various environmental variables

and a shared spatial field that explains spatial autocorrelation of the observations, I will also examine models with variables that in some way explain the human impact at the location, and a second spatial field fit only from the citizen science data, that will attempt to explain the spatial variance unique to the citizen science data.

With multiple data types, the flexibility of a Bayesian model makes it an attractive choice, but due to the complexity of Bayesian models this may be computationally expensive. However, integrated nested Laplace approximation (INLA) (Rue et al., 2009) provides a computationally feasible framework for approximating a continuous surface, and modeling complex point process models can be done in a relatively short period of time (Isaac et al., 2020).

I begin by introducing the data and initial analysis in chapter 2, to make clear what we are dealing with and how I chose to clean the data. In chapter 3 I move on to necessary background theory. Here I will give a brief background in some concepts from spatial statistics and from ecology, covering how to model one or more data sets in an integrated model, as well as giving some insight into INLA.

In chapter 4, I will introduce the specific observation models as well as the model components I use in this study. Some of these will already have been introduced in chapter 3, but here I will present them in a more specific context. I will describe the model fitting, evaluation and validation. The results of this will be presented in chapter 5, where I compare five different models, and also examine the results of the chosen model in more detail.

Finally, I will discuss my results in further detail in chapter 6, before concluding in chapter 7.

Chapter 2

Data presentation and initial exploration

2.1 Observation data sets

The data used in this thesis consists of two data sets of freshwater fish observations: one that has been collected in a systematic way, and one opportunistic citizen science data set.

As citizen science data, I have used observations from the Artsobservasjoner dataset, available through GBIF (see appendix A for download links). I have chosen to look at the four most prevalent freshwater fish in this data set, which are the brown trout

Table 2.1: Number of observations of each of the fish species examined in the citizen science data set, before and after matching them to the closest lake (observations further than 30 m from a lake are removed)

Species	Number of observations		Fraction removed
	Originally	After data cleaning	
Brown trout	1220	661	0.45
European perch	417	321	0.23
Arctic char	280	254	0.09
Northern pike	312	237	0.24

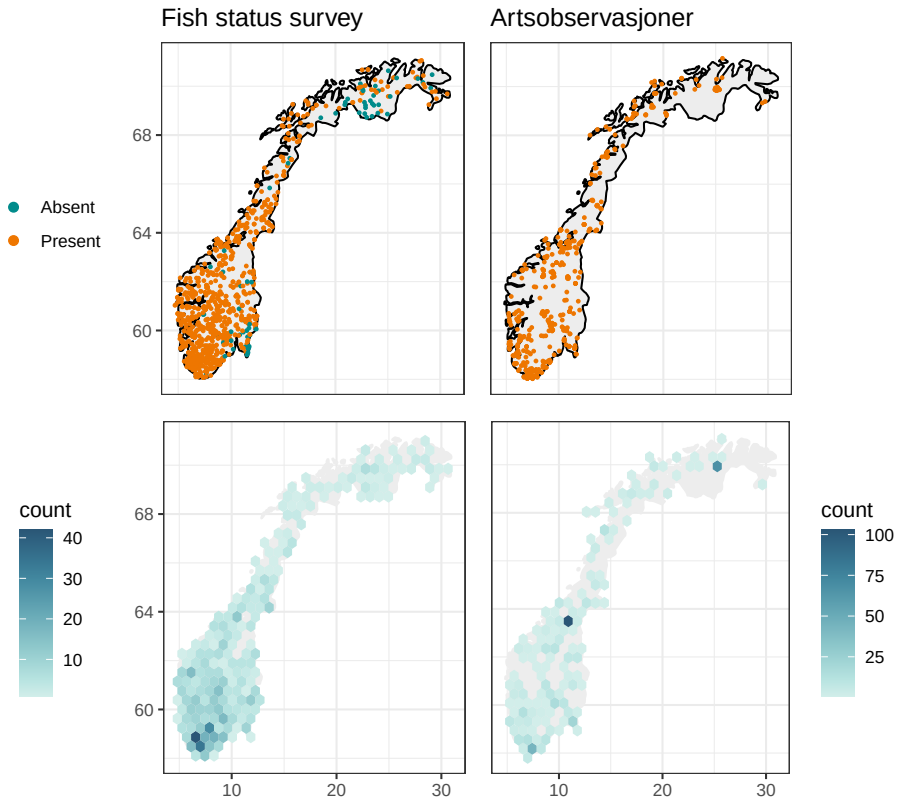


Figure 2.1: Top row: point maps of all cleaned observations of trout in both data sets. Bottom row: Hexagon maps of the observations of trout in both data sets, to illustrate the density of observations in areas where the points land on top of each other.

(*Salmo trutta*), the European perch (*Perca fluviatilis*), the Arctic char (*Salvelinus alpinus*) and the northern pike (*Esox lucius*).

For presence-absence data, I have used the Fish Status Survey of Nordic Lakes (Tammi and Finstad, 2019). This was a survey conducted over several lakes in Norway, Sweden and Finland, where the presence or absence of a number of freshwater fish was recorded in 1996. I selected the Norwegian observations of the four species of interest from this data set, which left me with the occurrence status of these species in around 800 lakes in Norway.

One of the challenges in using citizen science data is that the spatial location for the observation may not be completely accurate. In this project, I have the advantage of working with individuals within lakes, and I am not interested in the exact location of the fish within the lake.

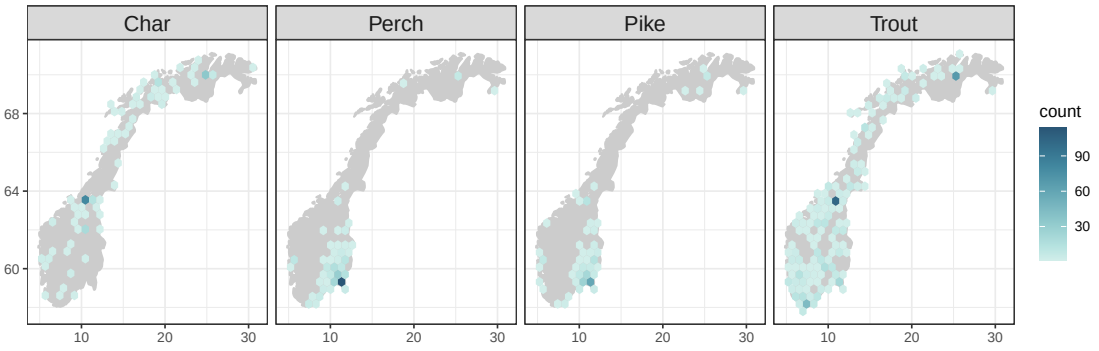


Figure 2.2: Hexagon maps of the citizen science observations (from Artsobservasjoner) of all four fish species (note that for the survey data, locations are the same for all species, so the hexagon maps will all look like the one in figure 2.1).

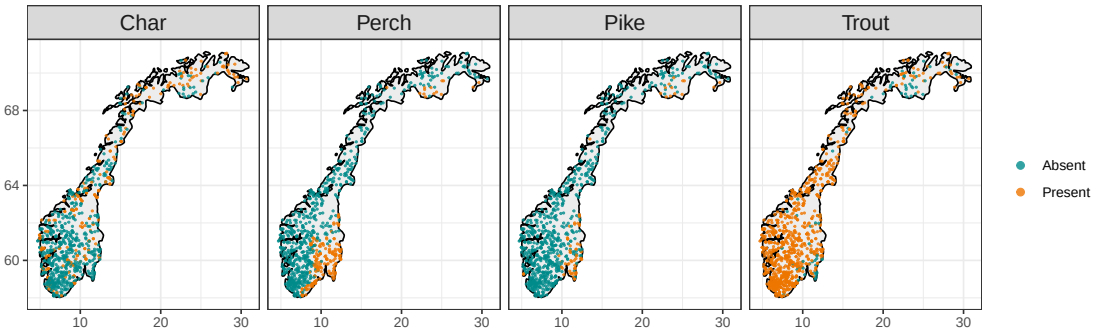


Figure 2.3: Presences and absences from the survey data for all four species.

Often the citizen scientist will specify a location outside of the lake itself, but it is fairly easy to check if an observation is within some reasonable distance of a lake and then match it to the closest lake. This strategy does potentially allow for some error, in the cases where there are several lakes close to each other.

One might try to remedy this somehow, for instance one could favour large lakes in cases where there are multiple lakes within a given threshold of the observation, or one could try to use the name of the lake to a larger degree if this is given, although this gives rise to a whole new set of problems since lake names are not at all unique, see for example Storvatnet and Langvatnet (or just look at the occurrences of Lomtjønna in the Trondheim area alone!), and if the user is asked to spell the name themselves this could give rise to more problems, for example,

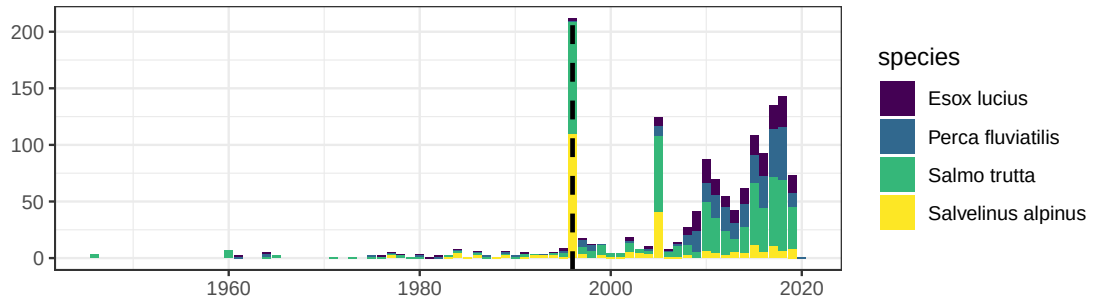


Figure 2.4: The number of observations in the citizen science data set per year, for all four fish species. The black dashed line at 1996 is included to mark that this is the year the observations in the survey data set were made.

should they write 'vatn' or 'vann', 'tjern' or 'tjønn'?[1]

In this case, I excluded observations that were further than 30 meters from the lake shore. This was done by using a list of all Norwegian lakes, and checking if the observation was within 30 meters of some lake. If it was, the observation was marked as coming from that lake, if not it was removed from the considered data set. A fairly large proportion of the data was marked as further than 30 meters from a lake, see table 2.1 for exact numbers. This was done since it was crucial for further analysis to be able to connect the individual fish with the lake they came from. Note that the removed observations may just be fish observed in rivers, which seems plausible when looking at how the removed fractions differ between species: the brown trout is the only species that is commonly found in rivers, and it also has the highest removal percentage. After this, any observations missing the observation time were also removed.

2.2 Explanatory variables

As potential covariates I had access to the area of each lake, in square kilometers; the average air temperature by each lake of the warmest annual quarter, measured in degrees Celsius multiplied by 10 (estimated in Metz et al. (2014)); the perimeter of the lake in meters; the shoreline complexity index (SCI); and the size of the catchment area (this is the area that drains into the lake in question) in square kilometers. I also have the longitude and latitude location of each lake.

[1] 'Storvatnet' means 'the big lake', 'Langvatnet' is 'the long lake'. 'Lom' is a bird, the black-throated loon (*Gavia arctica*), and 'tjønn' is a term for a small lake. 'vatn' and 'vann' both mean lake, and 'tjern' and 'tjønn' are local variations on the same term.

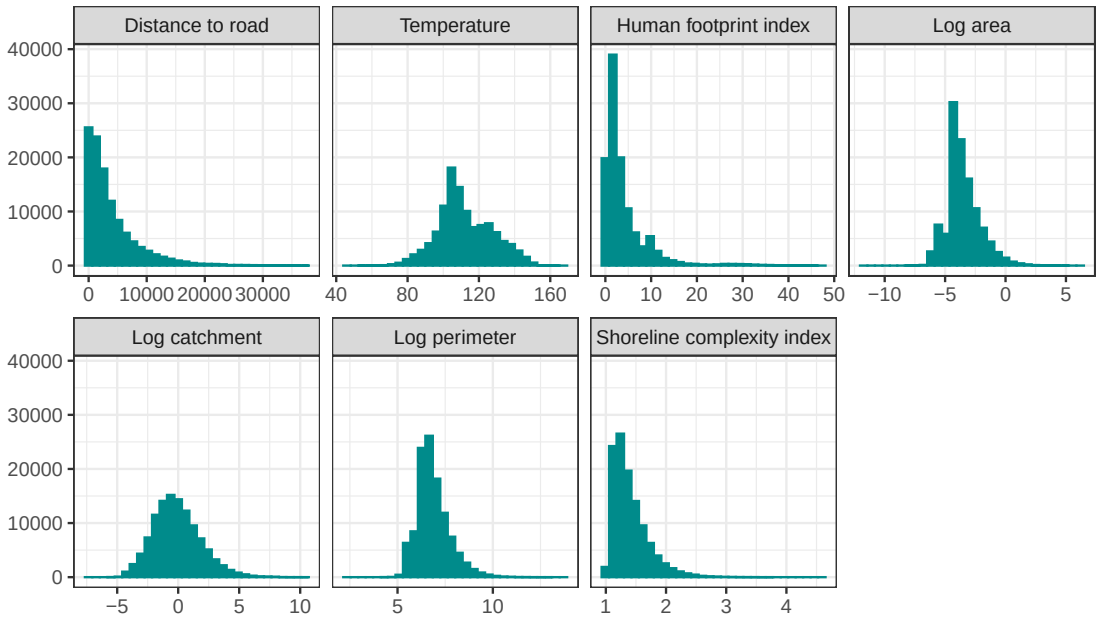


Figure 2.5: Histograms showing the distributions of the explanatory variables.

These variables all have reasonable interpretations in an ecological context when explaining occurrences of different freshwater fish. I will be referring to this set of variables as the *environmental* variables.

In addition to these, I have two variables that can be used as a measure of effort needed to make the observations: the distance to the closest road for each lake; and the human footprint index (HFP) at each lake. The latter is a score made up of eight human impact variables (the variables are: built-up environments, population density, electric power infrastructure, crop lands, pasture lands, roads, railways, and navigable waterways) that approximate the level of human pressure in this area (from Venter et al. (2016)). I will refer to the set of these two variables as the *effort* variables.

Due to strongly skewed distributions, three of the environmental variables, the lake area; catchment area; and lake perimeter, were log-transformed in all subsequent analysis.

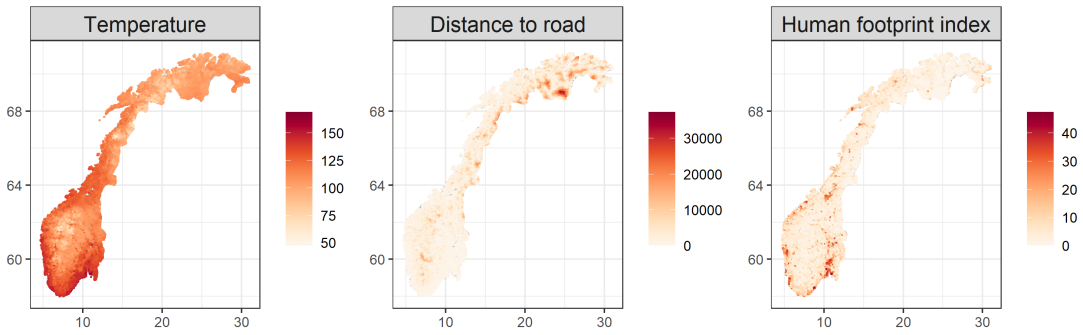


Figure 2.6: Spatial plots of the explanatory variables that appear to display some spatial structure.

Chapter 3

Background

The motivation of this thesis is the challenge of using citizen science data to model species distributions. This has in itself been addressed previously by many different studies. I will first give some ecological background, to point out what is unique for this context. I will then cover some general statistical background, and introduce some point process models that will be used in this project. Following that, I will cover a few different topics that are instrumental in understanding my further analysis.

3.1 The ecological context

Models for describing the distributions of species exist in an intersection between statistics and ecology. The statistical framework for considering point observations will be presented in subsequent sections, but first I will cover some considerations that are specific for the ecological setting.

3.1.1 Types of observation data

When examining the distribution of species, we may encounter many different types of data, and knowing how to best model the specific data on hand is important (Guillera-Arroita et al., 2015). A common type of observation type is *presence-only data*, which, not surprisingly, consists of only the locations of presences of a given species. This is common in cases such as citizen science data

collection, or records from museum samples. In surveys and more structured data collections, we often have *presence/absence data*. As the name suggests, this gives us double the information in the sense that we get information about where the species is recorded as absent, as opposed to presence-only data, where an absence of observations in an area may mean that the species is not there, or it may mean that it was simply not observed there despite the fact that it was present. However, note that detection may not be perfect for presence/absence data either. Another common data type is *abundance data* (or *count data*), where counts of individuals or some index based on the counts is reported.

We also distinguish between *structured* and *unstructured* data, where structured data is derived from a well-defined sampling protocol, often repeated over time, and unstructured data is often more opportunistic in nature (and more often tends to be presence-only data). Commonly we assume that structured data is less spatially biased than unstructured data (as this is often one of the aims of having structured data in the first place), but this does not need to be the case (Isaac et al., 2020).

3.1.2 Species distribution models

The broad class of models that connect environmental covariates to species' records over geographic region are referred to as *species distribution models*. The aim is either to get further understanding of a species or to predict the species' distribution across a landscape or into the future, for example as a result of changing temperatures (Elith and Leathwick, 2009). The term encompasses multiple different technical approaches, though the most common models today are maximum entropy models (MaxEnt) (Phillips et al., 2006) and generalized linear models (GLM). The covariates might include longitude and latitude, some measure of temperature, some measure of human involvement in the area or altitude, just to mention a few. Many also incorporate some term to capture spatial autocorrelation between the observations, such as a spatial field (Elith and Leathwick, 2009).

3.1.3 Integrated distribution models

There are several different ways to combine different data sets in order for them to inform one species distribution model (Fletcher Jr. et al., 2019; Pacifici et al., 2017). For instance, if one has access to one data set with both presences and absences reported, and another data set with presence-only data, one could convert the presence/absence data into presence-only and then combine the data sets in order to inform one model. This approach is referred to as data pooling (Fletcher Jr.

et al., 2019), and it is an approach that combines the actual data sets before feeding the data to the model. The disadvantage with this approach is that it essentially degrades the presence/absence data, and we loose information that we had access to in the original data set.

A different approach is to explicitly describe the differences in how data were collected by using individual observation models for each data set. In this approach, we are combining the *models*, rather than the data directly, which is why it is sometimes referred to as *model-based data integration* (Isaac et al., 2020), though the term integrated distribution modelling is more widely used. When we have access to multiple data sets that are observed from the same underlying population, we can combine these data sets in an integrated distribution model. The idea is that the data all arise as separate realizations of the true distribution model, and that by taking into account separate observation processes for each data set, we will be able to capture more of the underlying distribution than when using only one data set. Given M data sets with individual observation processes, the total model likelihood can then be found by

$$L(Y_1, \ldots, Y_M \mid X, \phi, \theta_1, \ldots, \theta_M) \propto \underbrace{p(\lambda(s), X, \phi)}_{\text{model for unobserved state}} \prod_{i=1}^{M} \underbrace{\Pr(Y_i \mid \lambda(s), \theta_i)}_{\text{likelihood for dataset i}}, \tag{3.1.1}$$

where Y_i is observation data set number i, X is the environmental covariates, the parameters ϕ for the underlying model, and the parameters θ_i for the likelihood of data set i.

Integrated distribution models become particularly interesting in connection to citizen science data, since we may have access to, for instance, a survey data set and a citizen science data set. Here, the citizen science data is most often presence-only, while the survey could be presence-absence, or maybe it even provides counts of some type. A range map could also be included, such as in Merow et al. (2017), where they integrate occurrence data and expert range maps. In these cases it is a great advantage to be able to integrate several data sets to inform the same model.

However, in practice it is not always guaranteed that an integrated model will perform better than one based on a single data set (Isaac et al., 2020), and benefits observed in some cases may not be universal. Especially interesting for this thesis are the results of Simmonds et al. (2020), where they explore integrated distribution models on structured and unstructured data in a simulation study. They, among other things, find that the integrated distribution model does not improve over the model using only structured data, if the bias in the unstructured data is not accounted for in any way. This issue of accounting for the bias in some

meaningful way is at the core of this thesis, and will be addressed in detail in later sections.

3.2 Bayesian inference and hierarchical models

In the classical, frequentist approach to statistics, the parameter(s) $\boldsymbol{\theta}$ is considered to be some unknown, but fixed, value. Based on a sample observed from a population generated in some way from $\boldsymbol{\theta}$, we can obtain some knowledge about $\boldsymbol{\theta}$. In the Bayesian approach however, $\boldsymbol{\theta}$ is considered to itself have some probability distribution, which we denote the *prior distribution*. This is subjective, usually provided by the statistician constructing the model, and reflects any prior knowledge from before the data has been observed. The fact that the parameters are given a distribution may represent either the fact that the parameters are truly varying, or it could reflect the fact that our knowledge of the parameters is imperfect. Either way, it provides an additional layer of flexibility.

After data has been observed, we update our prior distribution with information from the data we have observed, to obtain a *posterior distribution*. Specifically, we denote the prior distribution by $p(\boldsymbol{\theta})$, and the likelihood by $p(\text{data}|\boldsymbol{\theta})$. Then the posterior distribution $p(\boldsymbol{\theta}|\text{data})$ is proportional to $p(\text{data}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})$, that is, the likelihood times the prior.

In *hierarchical modeling*, we allow for models with more levels to estimate the parameters. If, as above, $\boldsymbol{\theta}$ is our parameter determining the data generating process, and ϕ is some *hyperparameter* governing $\boldsymbol{\theta}$, then we can define the three levels or stages of the model as

1. the likelihood or data model, $\pi(\boldsymbol{y}|\boldsymbol{\theta}, \phi)$,
2. the prior distribution (or process model) $\pi(\boldsymbol{\theta}|\phi)$,
3. the hyperprior distribution (or parameter model) $\pi(\phi)$,

see e.g. Cressie and Wikle (2011) or Gelman et al. (2003). These give the joint distribution

$$\pi(\boldsymbol{y}, \boldsymbol{\theta}, \phi) = \pi(\boldsymbol{y}|\boldsymbol{\theta}, \phi)\pi(\boldsymbol{\theta}|\phi)\pi(\phi),$$

and finally the joint posterior distribution

$$\pi(\boldsymbol{\theta}, \phi|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{\theta}, \phi)\pi(\boldsymbol{\theta}|\phi)\pi(\phi)}{\pi(\boldsymbol{y})}.$$

This is the posterior distribution I will be interested in estimating in this thesis.

3.3 Spatial point processes

A spatial point process is a stochastic process that determines the locations of some events $\{s_i\}$ in a set $S \subseteq \mathbb{R}^d$, often on the plane, so $S \subseteq \mathbb{R}^2$ (see e.g. Cressie and Wikle (2011), section 4.3). There exist a great number of spatial point processes and variations on these. I will be using a log Gaussian Cox point process, a version of a Cox point process, which is a generalization of a Poisson point process.

3.3.1 Poisson point processes

Poisson point processes can be used to model occurrences of random events in time by defining the process on the real line, or occurrences of some event in space by looking at the process defined on a plane, or even higher dimensions where that is appropriate.

An inhomogeneous Poisson point process is characterized by the conditions that the number of points in disjoint sets are distributed independently of each other; and that the intensity $\lambda(s)$ varies spatially according to covariates indexed by location s (Cressie, 2015). Then the number of points m in a region A is a realization of a Poisson random variable M with mean $\int_A \lambda(s)ds$. From this, the likelihood can be derived, and is

$$L(\beta; \boldsymbol{s}) = \exp(-\int_A \lambda(s)ds) \prod_{i=1}^{m} \lambda(s_i),$$

which then leads to the log-likelihood

$$l(\beta; \boldsymbol{s}) = \sum_{i=1}^{m} \ln \lambda(s_i) - \int_A \lambda(s)ds.$$

However, an assumption here is that data are conditionally independent given the covariates. This is often not the case, and does not account for spatial dependence (Renner et al., 2015). Therefore, I move on to consider a practical generalization: the Cox process.

3.3.2 Cox processes

A Cox process is a generalization of the class of Poisson processes, designed to allow for more flexible models. They were originally known as doubly stochastic Poisson processes, which stems from the fact that the intensity is itself a stochastic process, in contrast to the standard Poisson process (Møller and Waagepetersen, 2003). Like Poisson processes, Cox processes can model events both in time or events in space. Here, I am interested in the latter. Using the definition given by Møller and Waagepetersen (2003), we can define a Cox process as follows.

Definition 1. Suppose that $Z = \{Z(\xi) : \xi \in S\}$ is a nonnegative random field so that with probability one, $\xi \to Z(\xi)$ is a locally integrable function. If the conditional distribution of X given Z is a Poisson process on S with intensity function Z, then X is said to be a *Cox process driven by Z*.

The statement that $\xi \to Z(\xi)$ is a locally integrable function just means that its integral is finite, which is required so we can actually find the intensity measure, where the intensity measure of the Poisson process $X \mid Z$ is

$$M(B) = \int_B Z(\xi) d\xi, \quad B \subseteq S.$$

In other words, this measure will be the mean of the Poisson distribution that describes the number of observations in the area B.

3.3.3 Log Gaussian Cox processes

For our purposes however, we consider the specific case when $Y = \log(X)$ is a Gaussian field, Again, using the definition from Møller and Waagepetersen (2003), we define this process.

Definition 2. Let X be a Cox process on $\mathbb{R}^d$ driven by $Z = \exp(Y)$ where Y is a Gaussian field. Then X is said to be a *log Gaussian Cox process (LGCP)*.

Now, the distribution of (X, Y) is completely determined by the mean and covariance function

$$m(\xi) = \mathrm{E}[Y(\xi)] \qquad \text{and} \qquad c(\xi, \eta) = \mathrm{Cov}(Y(\xi), Y(\eta)).$$

We can understand this as the point process equivalent of a generalized linear mixed model with a random intercept that is normally distributed (Renner et al., 2015).

The wide range of possibilities for these make log Gaussian Cox processes very flexible. In general, the intensity $\lambda(s)$ of a log Gaussian Cox process at a point s is

$$\log(\lambda(s)) = \boldsymbol{x}(s)^T\beta + \xi(s),$$

where $\boldsymbol{x}(s)^T\beta$ is the standard linear regression predictor while $\xi(s)$ is a spatial Gaussian random field with mean zero and a covariance function that varies such that observations that are closer together have higher correlation than observations that are further apart.

3.4 Computational tools for fast inference

All model fitting in this thesis has been carried out by using integrated nested Laplace approximations (INLAs) (Rue et al., 2009), and in this section I hope to de-mystify this to anyone unfamiliar with the INLA methodology. I will not go further into detail than what is necessary to understand my model implementation, as the inner workings of INLA are not directly relevant to the questions addressed in this thesis. I will also focus part of this section on stochastic partial differential equations, which are essential to model spatial fields through INLA. For a more in-depth introduction to the spatial methods used in this thesis, the book by Krainski et al. (2019) is an excellent guide.

3.4.1 Integrated nested Laplace approximations

When fitting Bayesian models, the most common approach has been to use Markov chain Monte Carlo (MCMC). However, as models have become more complex, faster methods have been necessary. INLA is an alternative that is growing more and more accessible and popular.

Latent Gaussian models

In order to be able to use INLA for a model, the model must belong to the class of *latent Gaussian models* (LGMs). This is a very wide class of models that encompasses generalized linear models, generalized additive models, time series and spatial models, and several more (Rue et al., 2009). The latent Gaussian model is a hierarchical model with three levels. First, there is the likelihood

function

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta} \sim \prod_i \pi(y_i|\eta_i(\boldsymbol{x}),\boldsymbol{\theta}).$$

Next, there is the latent field

$$\boldsymbol{x}|\boldsymbol{\theta} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(\mu,\boldsymbol{\Sigma}),$$

and finally, there are the hyperpriors,

$$\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}).$$

Here, $\boldsymbol{y}$ is an observed data set, $\boldsymbol{x}$ is the latent field (where the latent field can be understood as the joint distribution of the parameters in the linear predictor), $\boldsymbol{\theta}$ is the hyperparameter vector and $\eta_i(\boldsymbol{x})$ is the i^{th} linear predictor connecting the data to the latent field.

The general form of the linear predictor is

$$\eta_i = \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f_k(v_{ki}) + \varepsilon_i, \tag{3.4.1}$$

where β_0 is the intercept, the remaining β are the regression coefficients for the fixed effects $\boldsymbol{z}$, and the functions $\boldsymbol{f}$ are a set of functions on some covariates $\boldsymbol{v}$. These functions could be non-linear, for example they might give a spatially correlated random effect.

Latent Gaussian models are a special case of Bayesian hierarchical models with a structured additive predictor, where the elements of the predictors are assumed to follow a Gaussian distribution (Rue et al., 2009). This can be achieved by assigning Gaussian priors.

Now, in our case, using a log Gaussian Cox process, we have that the intensity is $\exp(\eta)$, where η is the linear predictor, and just as in equation 3.4.1, it is a sum of some fixed effects, a random spatial field (specifically a Gaussian Markov random field with a Matérn covariance function, so this term is also Gaussian), and an unstructured error term. So we see that the log Gaussian Cox process is an example of a LGM, which is again a special case of a hierarchical Bayesian model, and this all fits into the framework of INLA.

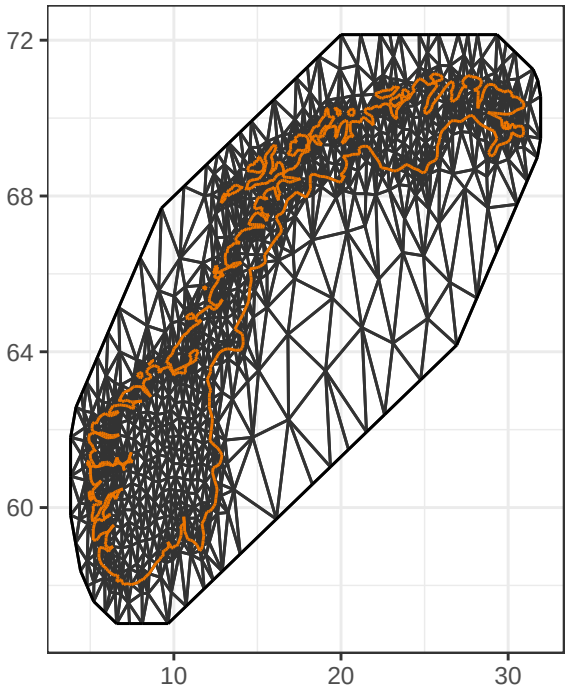


Figure 3.1: The triangulated mesh used to calculate the SPDE representation of the Gaussian field.

3.4.2 Stochastic partial differential equations

In order to fit the spatial model in practice with INLA, I use a stochastic partial differential equation (SPDE) representation of the Gaussian fields in the models I fit. This approach was originally proposed by Lindgren et al. (2011), where they show how Gaussian fields with Matérn covariance, which result in dense matrices and are therefore computationally expensive to calculate, can in some cases be represented by a Gaussian Markov random field, which has a sparse precision matrix and is therefore a lot quicker to carry out. This is done by using a solution to stochastic partial differential equations, and is implemented in the `R-INLA` package. I will not go into detail on the general theory of this, which can be found in Lindgren et al. (2011), as well as Simpson et al. (2016) for log Gaussian Cox processes in particular. For details on how to use the SPDE results in INLA, Krainski et al. (2019) gives great insight into this.

When implementing the SPDE representation of the Gaussian field in practice, we must first define a mesh over the spatial region we wish to model. This mesh will consist of a triangulation of the region, which is what is used to define the basis functions that approximate the spatial process. See figure 3.1 for the mesh used in this thesis. The reason for using a triangulation as opposed to a regular

grid, is that it allows for irregular boundaries and finer resolution where that is needed (Lindgren et al., 2011). Usually a Delaunay triangulation is used, which maximizes the smallest angle of each triangle.

Each basis function will be equal to one at a given vertex and zero outside the triangles that meet at that vertex. The functions then decrease linearly from the vertex. The result is that for any one triangle in the mesh, there are only three functions in the basis that are non-zero, so that the estimate of the random spatial effect at any point is just a linear combination of these three basis functions. Then, using a projector matrix, the projection of the SPDE is mapped to the observed points (Krainski et al., 2019; Gomez-Rubio, 2020).

3.4.3 Prior distributions in INLA: Penalized complexity priors

When fitting any Bayesian model, and thus also any model we are to fit in INLA, an important question is which priors to use. The idea of the priors is that they will convey any information the modellers knows about the behavior of the hyperparameters *a priori*, based on previous experience on similar data or common knowledge. Priors can be very subjective to the problem at hand, and so general guidelines are hard to come by.

I will be using *penalized complexity priors* (PC priors) (Simpson et al., 2017), which have been showed to be especially robust, and make user-defined scaling easier. In the case when the prior is for a Gaussian random field (GRF), we can follow the derivation from Fuglstad et al. (2018). They show that for dimension $d = 2$ and smoothness parameter $\nu = 1$, the Matérn covariance function giving the covariance between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is

$$C(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sigma^2 \left(\frac{\sqrt{8}}{\rho} |\boldsymbol{s}_i - \boldsymbol{s}_j| \right) K_1 \left(\frac{\sqrt{8}}{\rho} |\boldsymbol{s}_i - \boldsymbol{s}_j| \right), \qquad \rho > 0, \sigma > 0. \tag{3.4.2}$$

The joint PC prior for the range and the standard deviation of the field is

$$\pi(\rho, \sigma) = \lambda_\sigma \lambda_\rho \rho^{-2} \exp(-\lambda_\sigma \sigma - \lambda_\rho \rho^{-1}), \tag{3.4.3}$$

which is specified by the user through the two probabilities $\Pr(\sigma > \sigma_0) = \alpha_\sigma$ and $\Pr(\rho < \rho_0) = \alpha_\rho$ through the relationship

$$\lambda_\sigma = -\frac{\log(\alpha_\sigma)}{\sigma_0} \quad \text{and} \quad \lambda_\rho = -\log(\alpha_\rho)\rho_0.$$

From the joint distribution in 3.4.3 we can easily find the marginal distributions for ρ and σ,

$$\begin{aligned}\pi(\rho) &= \lambda_\rho \rho^{-2} \exp(-\lambda_\rho \rho^{-1}) \int_0^\infty \lambda_\sigma \exp(-\lambda_\sigma \sigma) d\sigma \\ &= \lambda_\rho \rho^{-2} \exp(-\lambda_\rho \rho^{-1}),\end{aligned}$$

and

$$\begin{aligned}\pi(\sigma) &= \lambda_\sigma \exp(-\lambda_\sigma \sigma) \int_0^\infty \lambda_\rho \rho^{-2} \exp(-\lambda_\rho \rho^{-1}) d\rho \\ &= \lambda_\sigma \exp(-\lambda_\sigma \sigma).\end{aligned}$$

So we have that the range ρ has an inverse exponential distribution with parameter λ_ρ, and the standard deviation σ has an exponential distribution with parameter λ_σ.

Chapter 4

Method

In this section, I will derive and describe the models used in a theoretical sense, stating the underlying point process model as well as the different observation models for each of the two data sets, and then describe the step-by-step implementation, model fitting and selection in detail.

4.1 Models

As described in equation 3.1.1, the total model likelihood will consist of both an underlying model for the unobserved state, as well as an observation process for each of the two data sets used. I will first describe how the intensity $\lambda(s)$ is modelled in the underlying model, and then explain how the two data sets are modelled individually.

4.1.1 Underlying process model

I model the underlying process as a log-Gaussian Cox process (Møller et al., 1998). That means that the density of the points is described by the intensity $\lambda(s) = \exp\{\eta(s)\}$, with

$$\eta(s) = \log(\lambda(s)) = \alpha_0 + \boldsymbol{x}(s)^T\boldsymbol{\beta} + \xi(s) + \varepsilon(s),$$

where $\boldsymbol{x}(s)^T\boldsymbol{\beta}$ are the fixed effects, $\xi(s)$ is a Gaussian Markov random field that ensures spatial autocovariance, and $\varepsilon(s)$ is some random error. The random field $\xi(s)$ is determined by a Matérn covariance function (see equation 3.4.2) (Lindgren et al., 2011), which in our specific case is determined by two parameters controlling the standard deviation and the range. Then the total number of presences in a given region A is Poisson distributed with mean given by integrating $\lambda(s)$ over A,

$$\mu(A) = \int_A \lambda(s)ds = \int_A \exp(\eta(s))ds.$$

That means that the probability that A has at least one individual is

$$Pr(N_A > 0) = 1 - Pr(N_A = 0) = 1 - \exp(-\mu(A)).$$

The integral in equation 4.1.1 may be tough to solve, and is therefore done numerically. In this case, we solve it using the approach of Simpson et al. (2016). As described in section 3.4.2 on the SPDE method, the region A is discretized into a tesselation of triangles, and each center point of a triangle is referred to as an *integration point*. We then calculate the value of the intensity at each integration point. The mean for the Poisson distribution at area A will then be approximated by the sum

$$\mu(A) \approx \sum_{s \in \mathbb{P}_A} |T_s| \exp\{\eta(A(s))\},$$

where $\mathbb{P}_A$ is the set of integration points in A and $|T_s|$ is the area of the triangle around s. This means that we only need to calculate the intensity at the integration points, and to estimate the intensity of any other point we just interpolate between the three vertex points of that integration point's triangle.

This gives the specifications for the actual abundance in any area. We will now move on to look at models describing the observation processes of the two data sets in question.

4.1.2 Observation processes

Depending on which type of data we are dealing with, we have different observation processes. I work with two different types of data: presence/absence data and presence only data.

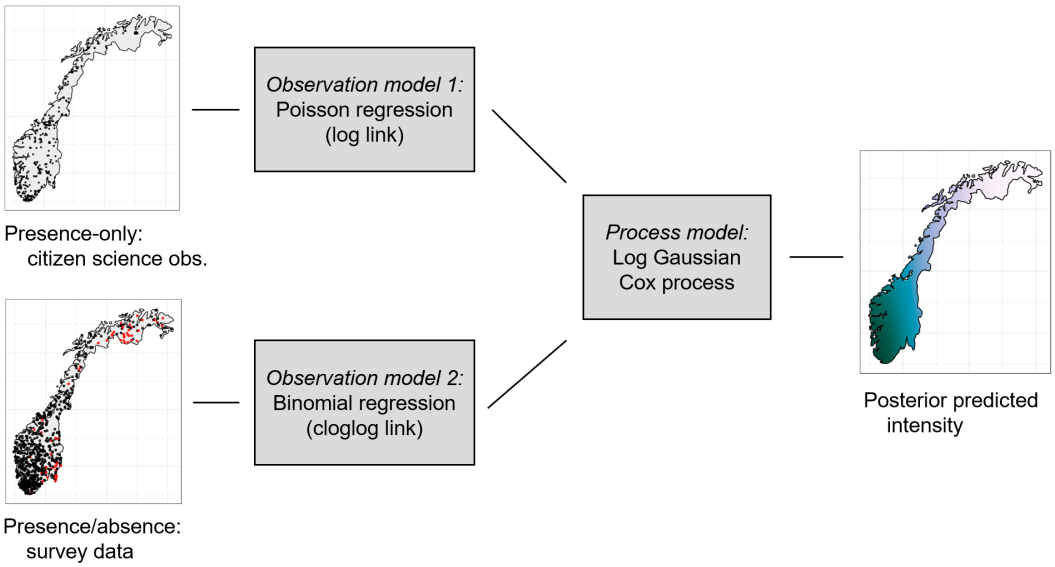


Figure 4.1: Schematic representation of observation models.

Count data: The more general case

In order model the presence/absence data we have in our survey data set, I first take a look at how to model *count data*, that is, data where for each location we have a number of observed individuals. We observe a region S, and the probability of observing each individual in the site is p. Then the number of individuals counted is

$$Pr(N(S) = r) = \frac{\mu(S)^r e^{-\mu(S)}}{r!},$$

where

$$\mu(S) = pt \int_S \lambda(s) ds.$$

If S can be assumed to be small, then we can treat the intensity and covariates as constant on the site, so

$$\mu(S) \approx p|S|\lambda(S).$$

This is reasonable in my application, as the surface is discretized into lakes where the intensity and covariates are constant over the lake surface anyway.

We are unlikely to know p and $|S|$, but since both can be seen as the observation effort we combine them to the parameter $E(S) = p|S|$. In practice, this is modeled on the log-scale as

$$\log(\mu(S)) \approx \log(E(S)) + \log(\lambda(S)) = \log(E(S)) + \eta(S).$$

Presence/absence data: Binomial regression with cloglog link

Now, if instead of count data we are only given the information of whether the species is present or not, we have *presence/absence data*. From equation 4.1.2, if we have only a presence or absence, this simplifies to

$$Pr(N(S) > 0) = 1 - Pr(N(S) = 0) = 1 - \exp\{\mu(S)\}.$$

When operating on the log scale for $\mu(S)$ and inverting this expression, we see how this results in the complementary log-log (cloglog) link function:

$$\log(\mu(S)) = \log(-\log(1 - Pr(N(S) > 0)))$$

If we have multiple visits to the site, we can set $p = Pr(N(S) > 0)$ and extend this from a Bernoulli to a binomial model. Then the likelihood that the species is observed n times in N visits is

$$Pr(n = r \mid N, p(S)) \propto p(S)^n (1 - p(S))^{N-n}$$

So for modeling presence/absence data I will be using a binomial regression model with a cloglog link function in order to link the probability of presence to the log intensity.

Presence-only data: Poisson regression with log link

The other data type I will be dealing with is *presence-only data*, which is typical for citizen science data. For the presence-only data a thinned point process model is used. Had all the individuals been observed, this would have simply been the process model, but this is probably not the case in the citizen science observations of lake fish. If we assume that each individual is observed with probability $q(s)$, then the intensity of observation is $\phi(s) = q(s)\lambda(s)$. Then if we observe M points, at locations $s_1, \ldots, s_M$, the log-likelihood is

$$l(\beta \mid s_1, \ldots, s_M) = \sum_{i=1}^{M} \phi(s) - \int_A \phi(a) da$$

(see Renner et al. (2015) for the details of this). This turns out to have the form of a Poisson likelihood, and so we may use a standard GLM formulation of a Poisson regression with a log link for the model (Renner et al., 2015). The log intensity is now

$$\log(\phi(s)) = \log(\lambda(s)q(s)) = \eta(s) + \log(q(s)),$$

and so any other additive terms may be included here, for example terms explaining observation bias.

So to sum up the process and observation models: we assume the data all come from the same underlying process, but that the data sets have different observation processes. For estimating the intensity, we may construct a linear predictor consisting of any environmental covariates we choose, as well as a spatial field to account for any spatial autocorrelation in the data. The specifications of the linear predictors will be detailed in section 4.2.

4.1.3 An effort field for describing the spatial bias of citizen science data

As mentioned, the linear predictors of the observation models will include a spatial field. This will be shared between the models, that is, it will be estimated from both of the data sets. In addition, in some of the models I will include a *second* spatial field, estimated by only using the citizen science data. This field will explain variation in the data not explained by the shared spatial field or the environmental covariates, and the intention is for it to capture the observation bias of the citizen science data. I will be referring to this spatial field as the *effort field*.

This approach is inspired by that of Illian (2019) (although that second field had a slightly different purpose), and more directly from Simmonds et al. (2020), where the authors use the second spatial field to capture the spatial bias of simulated "citizen science" data, which is exactly my aim as well. The advantage with this is that when dealing with citizen science data, we may not always know exactly what the source of bias is. This method does not depend on that, as opposed to other methods of accounting for spatial bias, such as including covariates that one might hope capture the bias of the citizen science data.

Fithian et al. (2015) has done something similar, although they looked at capturing the bias using predictors instead of spatial fields. They model the distribution of 36 eucalyptus species in a region of Australia, using biased presence-only data and systematic presence/absence data to build a multispecies model. They model sampling bias as a function of observation predictors, and assume this to be common across species, so that it can be shared across species.

By comparing the effort fields across different fish species, we may be able to identify such a bias field, that can then be used on other species.

4.2 Implementation in R

All analysis for this thesis was done in R (R Core Team, 2019). For the models themselves, these were implemented using the integrated nested Laplace approximation (INLA) methodology (Rue et al., 2009), specifically the `R-INLA` package, with the stochastic partial differential equation (SPDE) approach (Lindgren et al., 2011). I also used the package `PointedSDMs` (O'Hara, 2017), which is designed to make the INLA methodology more accessible when making species distribution models.

4.2.1 Model options

For all models, I included a base set of environmental covariates consisting of latitude, longitude, log lake area, log catchment area, shoreline complexity index and mean summer temperature. The log perimeter was excluded as it was strongly correlated (r = 0.96) with log area.

I was interested in comparing the four models that result from varying the spatial fields used and the covariate sets used. In addition, it is interesting to include a model fit only to the survey data, as I would be interested in examining if the models including citizen science data in some way can outperform this. Specifically, the five following models are of interest (see table 4.1 as well):

0. One spatial field fit only on survey data, and using only environmental covariates.
1. One spatial field fit on both survey data and citizen science data, and using only environmental covariates.
2. One spatial field fit on both survey data and citizen science data, one spatial field fit on citizen science data only, and and using only environmental covariates.
3. One spatial field fit on both survey data and citizen science data, and using both environmental and effort covariates.
4. One spatial field fit on both survey data and citizen science data, one spatial field fit on citizen science data only, and and using both environmental and effort covariates.

Apart from the varied components, all other settings were kept the same throughout all models. Notably, I used penalized complexity (PC) priors (Simpson et al.,

Table 4.1: The different models that were fit. The environmental covariates are $\boldsymbol{x}(s)$, while the effort covariates are $\boldsymbol{z}(s)$. $\hat{\xi}_1(s)$ is the spatial field informed by both data sets, while $\hat{\xi}_2(s)$ is the effort field estimated only from the citizen science data.

#	Data sets	Second sp. field?	Effort covs.?	Linear predictor
0	Survey only	No	No	$\hat{\alpha}_{PA} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$
1	Both	No	No	$\hat{\alpha}_{PA} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$ $\hat{\alpha}_{PO} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$
2	Both	Yes	No	$\hat{\alpha}_{PA} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$ $\hat{\alpha}_{PO} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s) + \hat{\xi}_2(s)$
3	Both	No	Yes	$\hat{\alpha}_{PA} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$ $\hat{\alpha}_{PO} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \boldsymbol{z}(s)^T\hat{\boldsymbol{\beta}}_z + \hat{\xi}_1(s)$
4	Both	Yes	Yes	$\hat{\alpha}_{PA} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$ $\hat{\alpha}_{PO} + \boldsymbol{x}(s)^T\hat{\boldsymbol{\beta}}_x + \boldsymbol{z}(s)^T\hat{\boldsymbol{\beta}}_z + \hat{\xi}_1(s) + \hat{\xi}_2(s)$

2017), through the function `inla.spde.pcmatern` (Fuglstad et al., 2018), for the parameters of the spatial fields. This function requires the user to specify two parameters, `prior.range` and `prior.sd`, which control the joint prior on range and standard deviation of the spatial field, as described in equation 3.4.3. I specify the the range ρ and the standard deviation σ through

$$\Pr(\rho < 10) = 0.1, \qquad \Pr(\sigma > 0.1) = 0.1.$$

I use the same prior specifications for both spatial fields.

For the fixed effects of the model, I use the INLA default priors, which for the intercepts means a normal distribution with mean 0 and precision 0, and for the remaining coefficients a normal distribution with mean 0 and precision 0.001.

4.2.2 Model validation

In order to compare these models, I used block cross-validation. This simply means dividing the data into blocks spatially, and then iteratively using all but

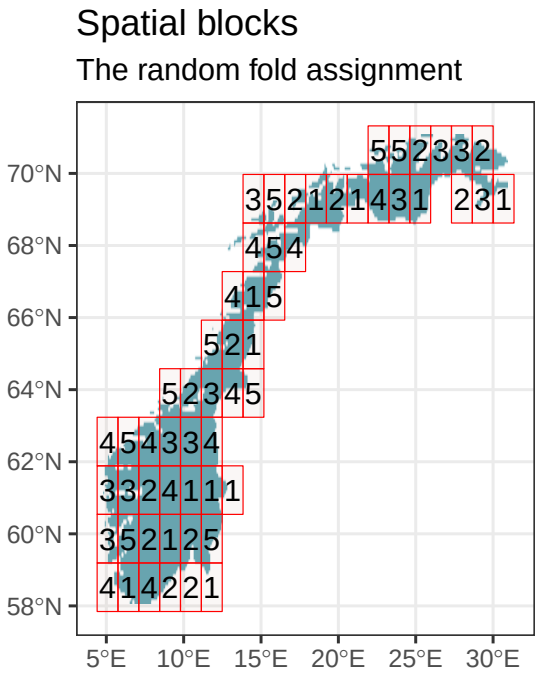


Figure 4.2: The spatial folds used for cross validation.

one of these blocks for fitting the model, and the last one for testing the model. This means that each block has been used for model fitting and testing separately. Block cross validation differs from regular cross validation in that the observations are not assigned to folds completely at random, they are assigned in spatial chunks, as shown in figure 4.2, and each chunk of observations is then assigned to a fold. The reason for choosing block CV, as opposed to a completely random fold assignment, is that our data has an underlying spatial structure, and a completely random cross validation procedure will quite probably give underestimated predictive error (Roberts et al., 2017). To implement the block cross validation, I used the package `blockCV` (Valavi et al., 2019). I used five folds ($k = 5$) with a random fold assignment, see figure 4.2 for the fold assignment used.

When the folds are established based on the survey data, the models are fit based on all the data in the training folds, both survey and citizen science (except in model 0). Only the survey data is used as validation data, the corresponding citizen science data in the same spatial block is not used for validation or training, since we do not wish to reconstruct the sampling bias of the citizen science data. Although there may be biases present in the survey data as well, it is probably the most reliable source of data.

For validating the models, I calculated the linear predictor for the validation sites

and then fitted an intercept-only model with these as an offset. Then I use the marginal deviance (marginalized over the prediction and the parameter uncertainty) from this as the validation statistic. Note that with the cross validation procedure all the models are compared on exactly the same data (since we use only survey data for validation), and thus the marginal deviances should be comparable.

When the model with the lowest marginal deviance has been established, I fit this model on the entire data sets to examine the final prediction as well as the posterior coefficients, spatial fields and hyperparameters of this model.

Running the model with two spatial fields takes no more than 10 minutes, and the whole cross-validation procedure with five models takes approximately two hours.

Up until this point, I have only been looking at observations of brown trout. As a final step, I fit the selected model on three additional fish species: Arctic char, European perch and Northern pike. It should be noted that this may not be the optimal procedure: the model is selected based on the trout observations, and then this selected model is used on all fish species. A more accurate procedure might have been to do the cross-validation procedure for each species separately, but since the three additional fish species had so few observations, the choice was made to use the model selected by the trout data. More importantly, I want to be able to compare the same model across species to study how that affects the first and second spatial field.

Chapter 5

Results

5.1 Choosing between models with different predictors

The first requirement for the models is that the integrated models need to be able to outperform the model based only on the survey data. The averaged marginal deviance of the single data set model on the validation data is 65.27, and all the other models have deviances below 14.00, so I move on to compare the integrated models to each other. The marginal deviances for each model on each cross-validation fold can be seen in table 5.1, and the averaged marginal deviances, which are used to select the final model, are compared in table 5.2.

Table 5.1: The marginal deviance resulting from predicting onto the validation fold in the cross validation procedure.

	Validation fold					Averaged
	1	2	3	4	5	
Model 0	58.30	76.18	57.46	58.78	75.63	65.27
Model 1	16.32	15.67	13.37	10.62	10.17	13.23
Model 2	16.28	15.63	13.00	10.24	9.60	12.95
Model 3	16.34	15.67	13.38	10.62	10.19	13.24
Model 4	16.29	15.64	13.01	10.24	9.62	12.96

We see that the models with an effort field do slightly better than the models without, but the difference is not large. Including effort covariates or not does

Table 5.2: Averaged marginal deviance values for each model, averaged across results of five-fold block cross validation.

		Spatial fields	
		One	Two
Covariates	Environmental	13.23	12.95
	Env. and effort	13.24	12.96

not seem to improve or worsen the model in any significant way. But either way, including the effort covariates does not seem to be helping the predictive power of the model, and so as my final model I will be using model 2, that is, a model with an effort field but no effort covariates.

5.2 Examining the selected model

I will now examine model 2 (see table 4.1), that is, the model with an effort field and no effort covariates. Again, the linear predictors are specified by

$$\eta_1(s) = \hat{\alpha}_{PA} + \boldsymbol{x}(s)^T \hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s)$$
$$\eta_2(s) = \hat{\alpha}_{PO} + \boldsymbol{x}(s)^T \hat{\boldsymbol{\beta}}_x + \hat{\xi}_1(s) + \hat{\xi}_2(s)$$

Note that the model has now been fit on the entire data sets, not sub-folds as in the cross-validation previously. That means that measures of error are probably over-optimistic and should not be relied upon without caution (Elith and Leathwick, 2009). The aim now is to study the behavior of the model and the predictions as they are in the case with the most available data.

First, looking at the environmental coefficients of the model, displayed in figure 5.1, they are all found to be insignificant. However, when looking at the predicted posterior log intensity of the model, plotted spatially on the map of Norway, as seen in figure 5.2, we see that there definitely seems to be a clear tendency of variation across the longitude and latitude, in that the log intensity is higher to the south-west and lower in the north-east of the map (although the effect is not very strong). In other words, there does seem to be that kind of environmental variation, at least in the distribution of the brown trout, but it has not been picked up in the latitude/longitude covariates. I suspect this variation has been completely captured in the spatial fields.

The two spatial fields of the model, $\xi_1(s)$ and $\xi_2(s)$, have been projected to the

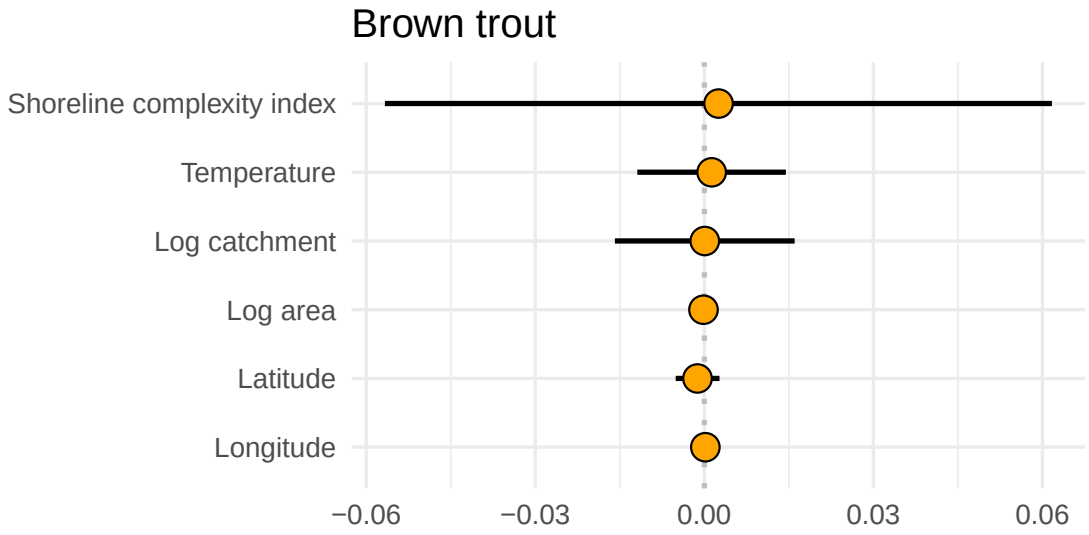


Figure 5.1: Estimated coefficients and 95% credible intervals for the environmental covariates of the final model, on observations of brown trout.

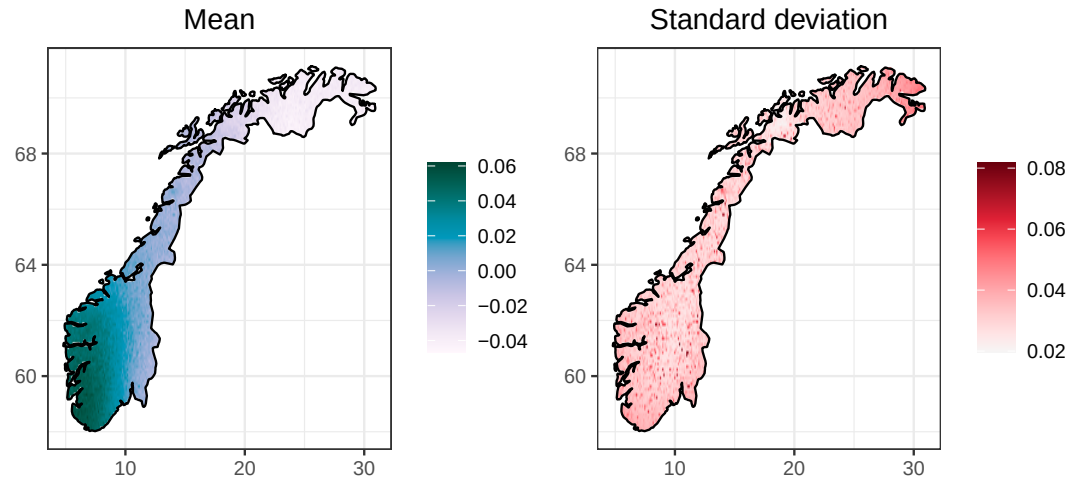


Figure 5.2: Predicted posterior log intensity $\log(\lambda(s))$ for the selected model, as well as standard deviation, for brown trout data.

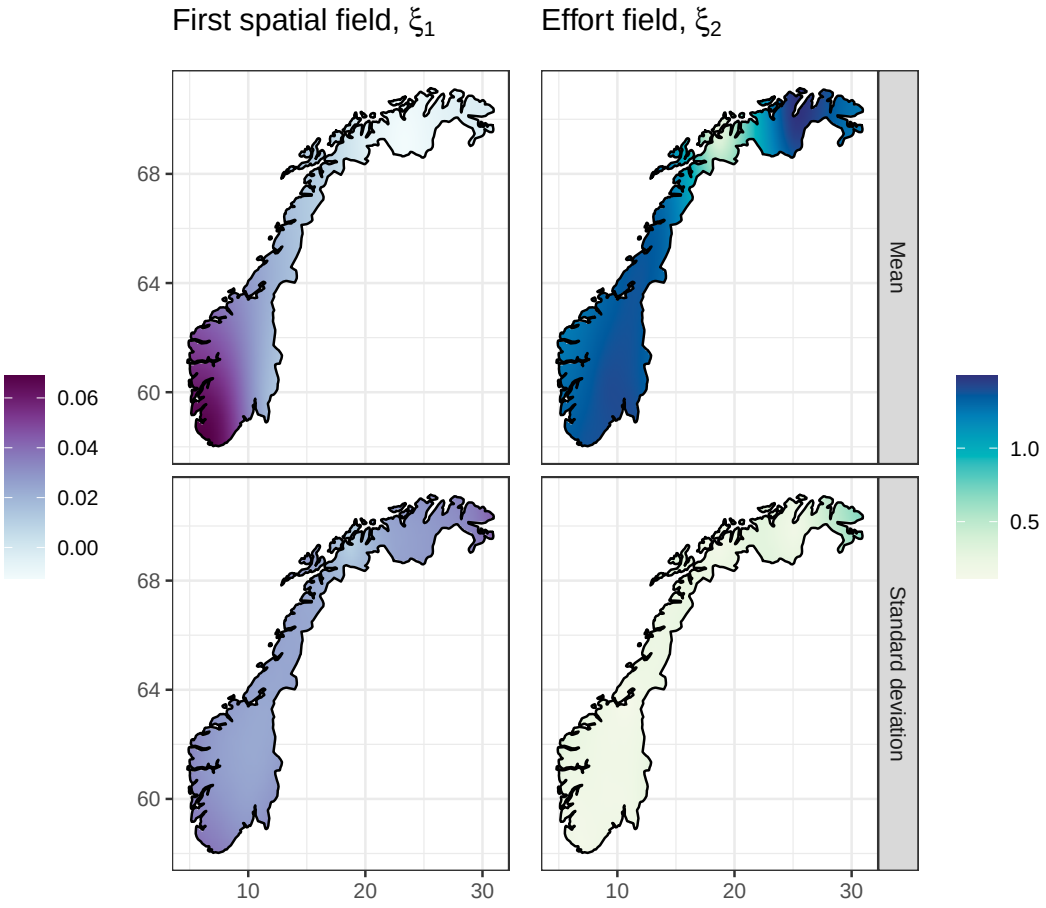


Figure 5.3: The random spatial fields of the final model. The first random field reflects the variation that was not explained by the covariates and that is shared between the two data sets, and the second spatial field shows the variation that is specific to the citizen science data set.

entire surface of Norway, in order to allow for better interpretation, since originally the values for the spatial field are only estimated at the integration points (the centroids between the vertices of the triangles in the mesh) of the model. The maps can be seen in figure 5.3. We may note that the standard deviation of the first spatial field is fairly high, compared to the estimated mean of the field. This is not the case for the second spatial field, to the same degree.

I was interested in seeing if there are any tendencies in the locations of the observation points in the citizen science data that are recognizable in the effort field, but this does not seem to be the case; in the citizen science data there are hotspots near the southern tip of Norway, as well as in Trondheim (in the center of Norway) and at the very north of the country (see figure 2.1), and this does not correspond

to any visible tendency in the effort field (see figure 5.3). This is not very unexpected though, since the effort field may only be picking up on residual variation left after the first spatial field has captured the variation that is shared between the two spatial fields, and thus not dependant on the citizen science observations alone.

5.3 Beyond trout: Comparing the selected model on four fish species

The estimated coefficients are not found to be significant for any of the other species either, see appendix C for the coefficients of these.

For each of the four fish species, the posterior mean and standard deviation is displayed in figure 5.4. There are clearly differing spatial trends for each of the species, I will discuss further how these correspond to the actual distributions of the fish in the discussion.

In addition to the posterior mean and standard deviation, the mean of the two spatial fields for each fish species has been plotted in figure 5.5. Here, it is interesting to see that although the first spatial field varies widely between species, the second spatial field seems to show the same tendency across species. I will discuss this further in the next chapter as well.

I also examine the posterior marginal distributions of the hyperparameters of both spatial fields for all fish species, see figure 5.6. As mentioned when specifying priors, each spatial field has two hyperparameters; the range ρ and the standard deviation σ. The posterior estimates for these are given in figure 5.6.

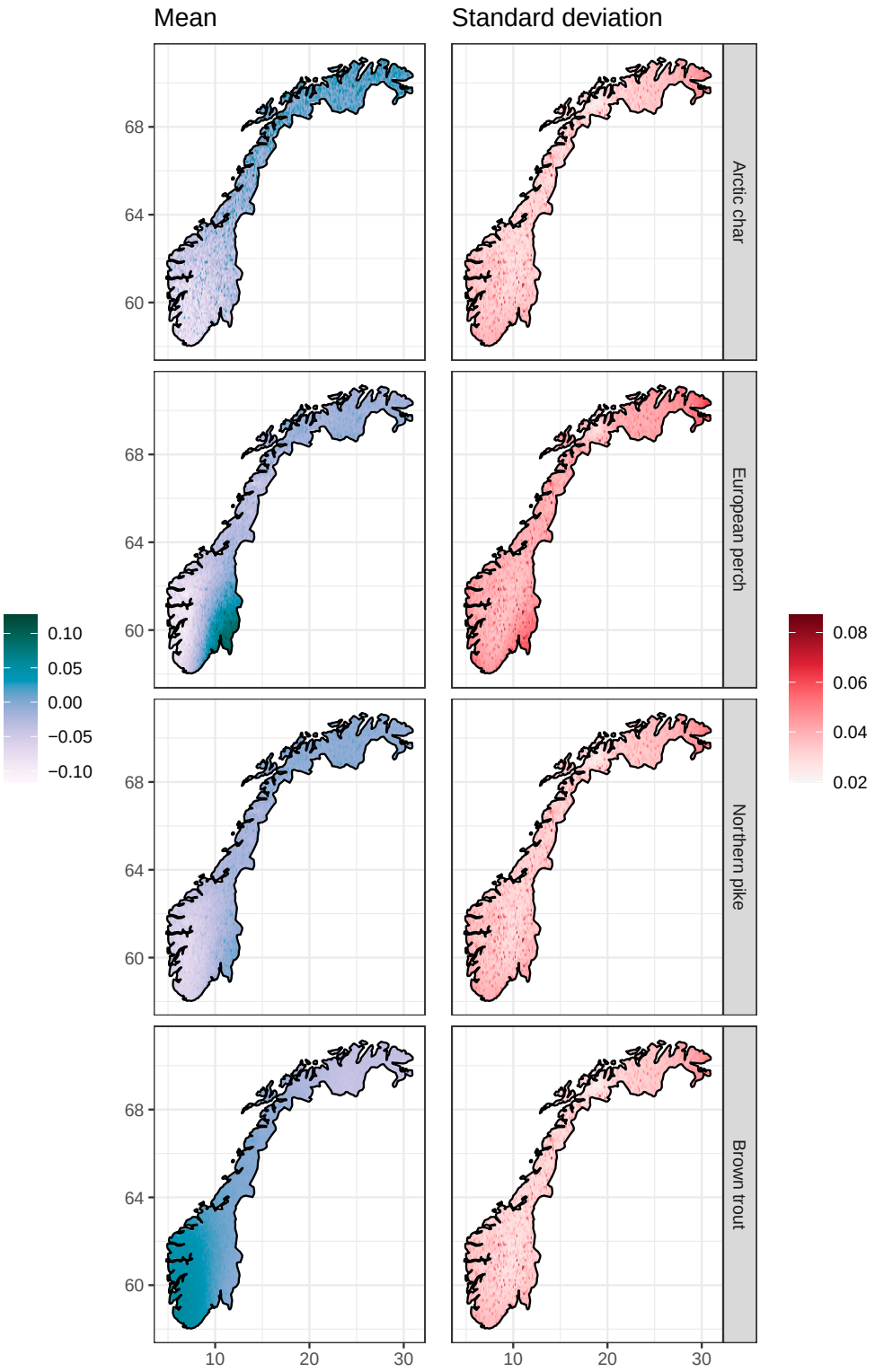


Figure 5.4: Posterior predicted log intensities $\log(\lambda(s))$ for all four fish species

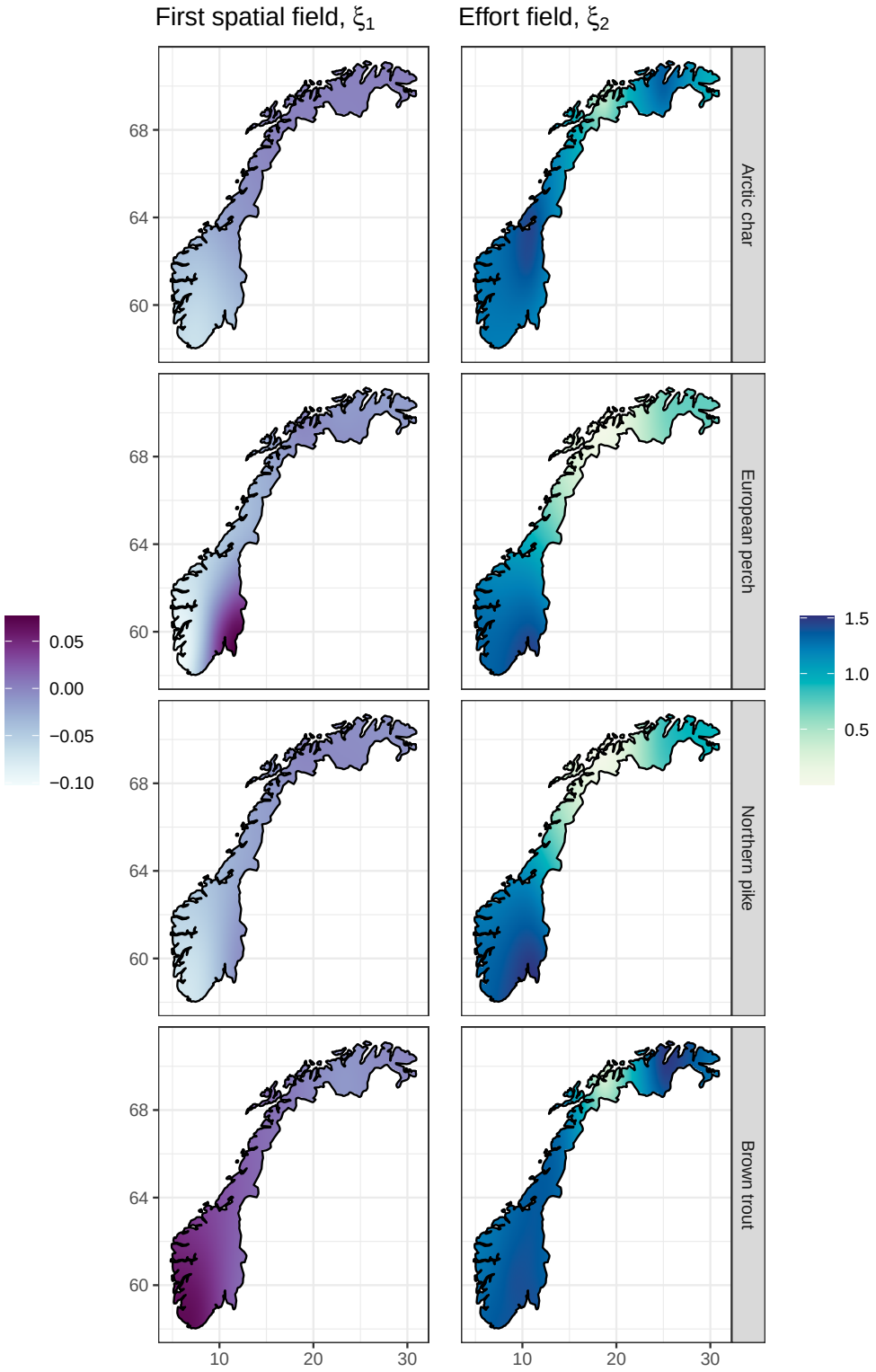


Figure 5.5: Comparing the two spatial fields across all four species. The first random field reflects the variation that was not explained by the covariates and that is shared between the two data sets, and the effort field shows the variation that is specific to the citizen science data set.

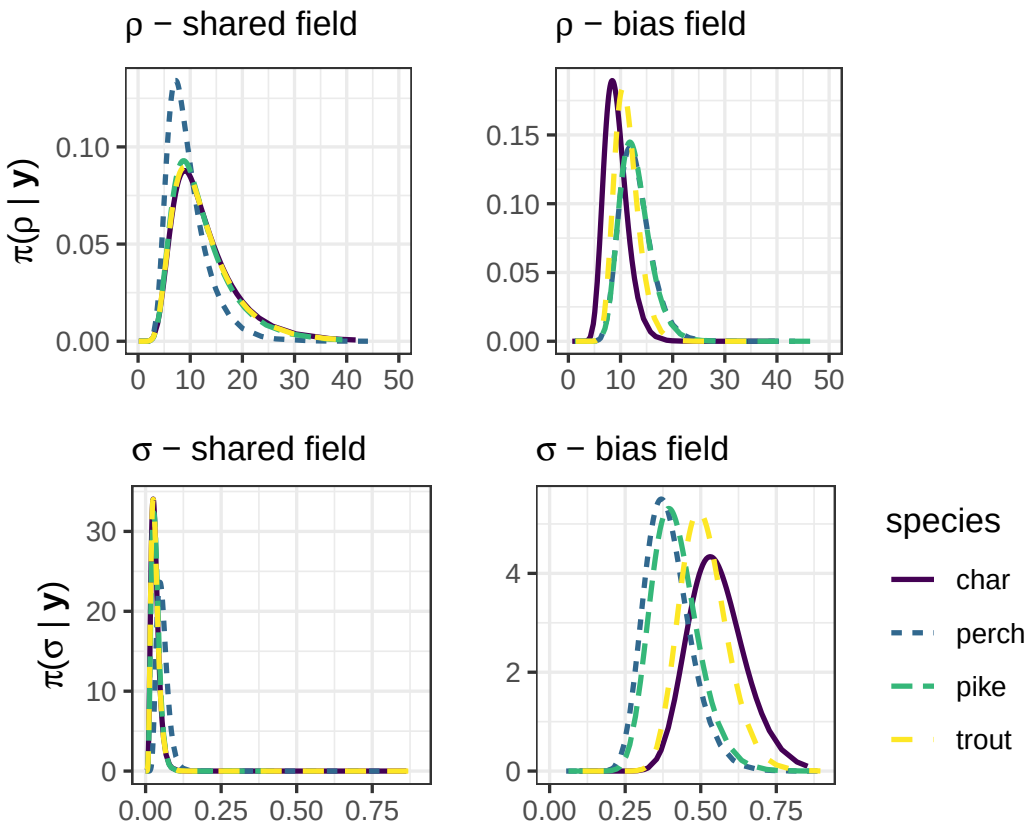


Figure 5.6: The posterior marginal distributions of the hyperparameters for the spatial fields of all fish species.

Chapter 6

Discussion

By comparing four different integrated models to the model using only survey data, we can establish that the integrated models outperform the survey-only model. Including a second spatial field (effort field) informed only by the citizen science data, in addition to a first spatial field fit to both data sets, did slightly improve predictions on new data, which was established through cross-validation. But it should be emphasized that the improvement, as measured using the averaged marginal deviance when predicted onto new data through cross validation, is not large.

Another finding from comparing the four candidate models was that the two covariates meant to measure citizen scientists' access to locations - distance to road and human footprint - did not seem to improve the model in any way. This agrees with the findings of Simmonds et al. (2020), and it does make sense that including both an effort field and effort covariates is redundant, since these may be explaining a lot of the same variation.

Another point is that our effort variables may just be poor choices for describing the spatial bias of the citizen science observations, maybe other variables than the human footprint index and distance to road would be better at describing this? Assuming that we can include these variables as a way to describe effort for the observation process of the citizen science data may also be problematic, since these variables might just as well be influencing the actual distribution of the species. For instance, areas with a higher human footprint are likely to be more polluted.

It is worth considering if a different model would have been better. For instance, given that none of the covariates were found to be significant, and the spatial

fields seem to be able to capture the spatial variation fully, a model without any covariates could be considered. The priors for the spatial fields should also be studied further, for instance, maybe a smaller range value should be used in order to capture more local processes.

Looking at the spatial fields for all four fish species in figure 5.5, there is clearly a greater difference across species in the first spatial field than in the effort field. This is definitely an interesting result, as it shows that the effort field *is* seemingly picking up on something underlying beyond the distribution of the individual species itself. Once the effort field is estimated based on some more common species, a possible next step might then be to transfer this field to models of less observed species as well.

At the same time, one must be careful in applying spatial fields as a mean of capturing the bias in citizen science data, as the bias might not only be spatial. As an example, Courter et al. (2012) looked at how observations of first arrivals of migratory birds tend to be reported on weekends rather than weekdays. Other unknown non-spatial biases may also be affecting the data and should be considered, especially if one wishes to extend this to a spatio-temporal model.

In all examined models, the specifications for the penalized complexity priors were kept the same, and they were also the same for both spatial fields, but there is no reason why this needs to be the case. The reasoning was simply that there was not good enough prior information to set the priors to any specific differing values, so they were instead rather arbitrarily set to the same values. Finding good values for the prior specification was not a major focus of this project, and the chosen priors gave reasonable results. They were not found to influence the model dramatically, so this approach seemed reasonable. But setting the prior specifications to separate values for the two spatial fields, if there was found to be some motivation for that, might be of interest.

One of the main benefits of including citizen science data is that this often gives us access to a lot of observations, compared to structured data sets that require more work to gather. However, fish are not the most popular animals to make citizen science observations of, and in this case I actually had more survey observation sites than I had citizen science, for any one of the most prevalent freshwater fish species. This means that the structured data likely played a central role in the final models.

Situations where one has access to more citizen science data may give more accurate models, since more data gives more information, but it may also strengthen the effect of the sampling bias of this data on the model. At the same time, simply having access to larger amounts of data may help to cancel out some of the

bias. If the citizen science data set becomes very large, additional measures may be necessary to ensure that the likelihood is not swamped by the citizen science data (Simmonds et al., 2020), such as a weighted likelihood approach (Fletcher Jr. et al., 2019).

The advantage of using freshwater fish in lakes is that it gives a nice variety of comparable species and good grounds for comparison across fish species. In order to have access to more citizen science data, one could easily look at the same data for Sweden or Finland. The survey data set covers Sweden and Finland as well as Norway, and from looking briefly at citizen science observations in Sweden, they seem to be significantly more plentiful than in Norway. Regardless, the methods of this thesis should be easily transferable to other species, even beyond freshwater fish, which should be done to examine further the behavior and challenges of the model with an effort field.

The posterior predictions of the distributions of the species seem pretty much in accordance with common knowledge about these species. The the Arctic char is predicted to be more common in the north of Norway than the south, which is opposite of the European perch and brown trout, and this is indeed the case. We also know that the pike is commonly found in northern Norway as well as in the east of Norway, and that the perch is also mostly found in eastern Norway (Pethon and Vøllestad, 2019). Looking more into the details of the predictions, maybe also on a smaller spatial scale so we can look at individual lakes to a larger degree, would be interesting as a second type of validation for the model.

Chapter 7

Conclusion

I propose an integrated species distribution model with two spatial fields in order to capture two different underlying spatial trends: the distribution of the species itself, and the spatial bias of the citizen science data. The effort field captures whatever variance is unique for the citizen science data after the first spatial field has captured the variance that is shared in both data sets. Interestingly, the effort field is similar across very different fish species, showing that the effort field does indeed capture something beyond the distribution of the species itself.

The proposed model is intuitive in its components, it shows predicted distributions for each species that seem reasonable, and picks up on a significant trend in the effort field. The model is relatively fast to run, taking around 10 minutes.

An interesting extension to this would be to estimate the effort field from one or more species that are rich in observations, and then use this to account for spatial bias in citizen science observations of less common fish species. In addition to this, it would be interesting to extend the model to a spatio-temporal model, especially as the amount of citizen science observations seems to be consistently increasing for every year.

A point that needs to be addressed further is the priors for both spatial fields. It does seem reasonable to assign different prior hyperparameter values for each of the spatial fields, but it takes time to do this in a systematic manner, and further insight is needed in order to decide how these should differ.

Bibliographic notes

A lot of the topics of this thesis were completely new to me at the time I started working with it. I have relied on a large number of sources throughout the period, and although these have all been cited throughout my thesis, I would like to give an extra mention here of my main sources for each topic, both for my future self as well as for others looking for helpful resources in similar projects.

First of all, my approach with the second spatial field was fully inspired from Simmonds et al. (2020). I really think the concept of applying a separate spatial field to the citizen science data is a promising idea that deserves further exploration beyond what I have done in my thesis, and this is a great read for anyone looking to try it.

My challenge with learning about species distribution models is that it is such a huge and general topic, that it is hard to know where to begin. Elith and Leathwick (2009) approaches the topic with language that is not confusing, and doesn't take the basics for granted, but covers them efficiently and then moves on to the more interesting topics surrounding SDMs. This article helped me feel more confident about the ecological aspects of my thesis.

There are already a few review articles on integrated species distribution models, but I have especially appreciated Isaac et al. (2020). They provide useful explanations of terminology, which is particularly helpful for a non-ecologist like me, and they also give a very nice explanation of why one might want to use integrated models, as well as intuitive explanations of the modeling framework, and considerations to make when applying integrated models.

For the model validation through block cross validation, Roberts et al. (2017) gives a thorough explanation of the reasoning behind it, not only in the spatial case but also for other cases where one might expect some dependence structure in the data, such as temporal or grouped data. The article is well reasoned and a very

informative read, and at the end gives a very specific step-by-step procedure to model validation.

A large part of my learning curve throughout this project consists of learning how to use INLA for spatial models. I used multiple sources for this, but the one that resulted in the most aha-moments and was the most directly applicable to my thesis was Krainski et al. (2019) (in particular chapter 4 on point processes). This became my go-to source for spatial models in INLA. The book is direct and to the point, and very user-oriented, with specific examples of code, but it also provides the theoretical framework necessary to get a complete understanding.

Bibliography

Courter, J., Johnson, R., Stuyck, C., Lang, B., and Kaiser, E. (2012). Weekend bias in citizen science data reporting: Implications for phenology studies. *International journal of biometeorology*, 57.

Cressie, N. (2015). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics. Wiley.

Cressie, N. A. C. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley Series in Probability and Statistics. Wiley.

Elith, J. and Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40(1):677–697.

Fithian, W., Elith, J., Hastie, T., and Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6(4):424–438.

Fletcher Jr., R. J., Hefley, T. J., Robertson, E. P., Zuckerberg, B., McCleery, R. A., and Dorazio, R. M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100(6):e02710.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gomez-Rubio, V. (2020). *Bayesian Inference with INLA*. Taylor & Francis Limited.

Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R., and Wintle, B. A. (2015). Is my species distribution model fit for purpose? Matching data and models to applications. *Global Ecology and Biogeography*, 24(3):276–292.

Hastie, T. J. and Fithian, W. (2013). Inference from presence-only data; the ongoing controversy. *Ecography*, 36 8:864–867.

Illian, J. B. (2019). Spatial and spatio-temporal point processes in ecological applications. In Gelfand, A. E., Fuentes, M., Hoeting, J. A., and Smith, R. L., editors, *Handbook of Environmental and Ecological Statistics*, Chapman & Hall/CRC handbooks of modern statistical methods. CRC Press.

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., and O'Hara, R. B. (2020). Data integration for large-scale models of species distributions. *Trends in Ecology & Evolution*, 35(1):56 – 67.

Kosmala, M., Wiggins, A., Swanson, A., and Simmons, B. (2016). Assessing data quality in citizen science. *Frontiers in Ecology and the Environment*, 14(10):551–560.

Krainski, E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Simpson, D., Lindgren, F., and Rue, H. (2019). *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. CRC Press/Taylor and Francis Group.

Lindgren, F., Rue, H., and 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(4):423–498.

Merow, C., Wilson, A. M., and Jetz, W. (2017). Integrating occurrence data and expert maps for improved species range predictions. *Global Ecology and Biogeography*, 26(2):243–258.

Metz, M., Rocchini, D., and Neteler, M. (2014). Surface temperatures at the continental scale: Tracking changes with remote sensing at unprecedented detail. *Remote Sensing*, 6(5):3822–3840.

Miller, D. A. W., Pacifici, K., Sanderlin, J. S., and Reich, B. J. (2019). The recent past and promising future for data integration methods to estimate species' distributions. *Methods in Ecology and Evolution*, 10(1):22–37.

Møller, J. and Waagepetersen, R. (2003). *Statistical inference and simulation for spatial point processes*. Number 100 in Monographs on Statistics and Applied Probability. Chapman & Hall.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. (1998). Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

O'Hara, B. (2017). *PointedSDMs: Fit Models derived from point processes to ppecies distributions using INLA*. R package version 0.2.1.9004.

Pacifici, K., Reich, B. J., Miller, D. A. W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A., and Collazo, J. A. (2017). Integrating multiple data sources in species distribution modeling: a framework for data fusion*. *Ecology*, 98(3):840–850.

Pethon, P. and Vøllestad, A. (2019). Ferskvannsfisk i Norge. `https://snl.no/ferskvannsfisk_i_Norge`. Accessed: 2020-07-11.

Phillips, S. J., Anderson, R. P., and Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3):231 – 259.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renner, I. W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S. J., Popovic, G., and Warton, D. I. (2015). Point process models for presence-only analysis. *Methods in Ecology and Evolution*, 6(4):366–379.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rue, H., Martino, S., and 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(2):319–392.

Simmonds, E. G., Jarvis, S., Henrys, P. A., Isaac, N. J. B., and O'Hara, R. B. (2020). Is more data always better? A simulation study of benefits and limitations of integrated distribution models. *Ecography*. In press.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H., and Rue, H. (2016). Going off grid: computationally efficient inference for log-Gaussian Cox processes. *Biometrika*, 103(1):49–70.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.

Tammi, J. and Finstad, A. G. (2019). Fish status survey of nordic lakes. `https://gbif.vm.ntnu.no/ipt/resource?r=fish_status_survey_of_nordic_lakes`. Accessed: 2020-02-06.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. (2019). blockcv: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 10(2):225–232.

Venter, O., Sanderson, E. W., Magrach, A., Allan, J. R., Beher, J., Jones, K. R., Possingham, H. P., Laurance, W. F., Wood, P., Fekete, B. M., Levy, M. A., and Watson, J. E. (2016). Global terrestrial human footprint maps for 1993 and 2009. *Scientific Data*, 3.

Appendix A

Data accessibility

All observation data used in this thesis are publicly available. The exact citizen science observations used in this thesis can be found at https://www.gbif.org/occurrence/download/0006251-200127171203522 and more recent citizen science observations for Norway can be found through the Norwegian Species Observation Service (Artsobservasjoner) through GBIF at https://www.gbif.org/dataset/b124e1e0-4755-430f-9eab-894f25a9b59c. The survey data is found at https://gbif.vm.ntnu.no/ipt/resource?r=fish_status_survey_of_nordic_lakes.

All R-scripts used for analysis are also publicly available. The static version at the point of completing this thesis is available at https://doi.org/10.5281/zenodo.3941072, and the latest version (which may change after completion of this thesis) can be found at https://github.com/emmaSkarstein/Citizen_Science_Skarstein_master.

Appendix B

Further data exploration

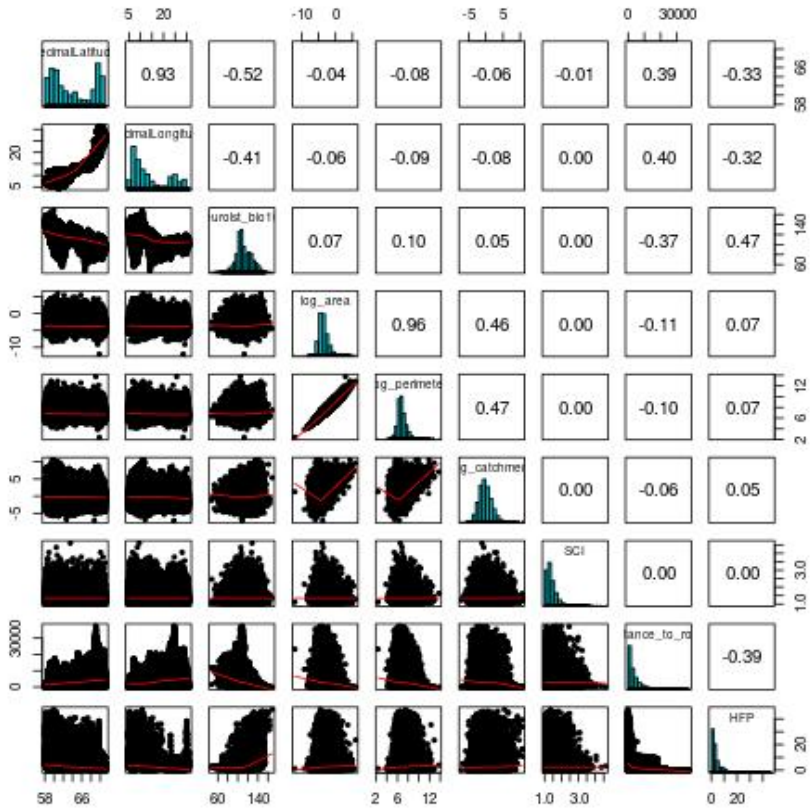


Figure B.1: Pairs plot to examine relationships between variables.

Appendix C

Additional results from comparing different species

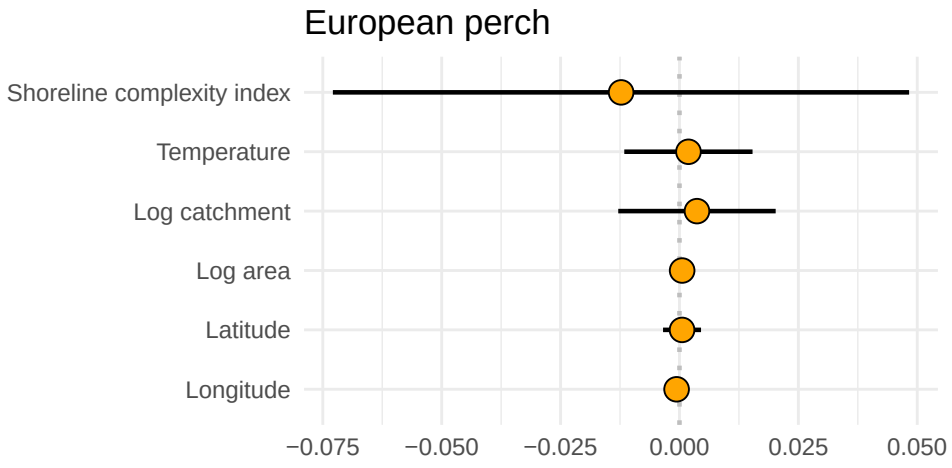


Figure C.1: Estimated coefficients and 95% credible intervals for the environmental covariates of the final model, on observations of European perch.

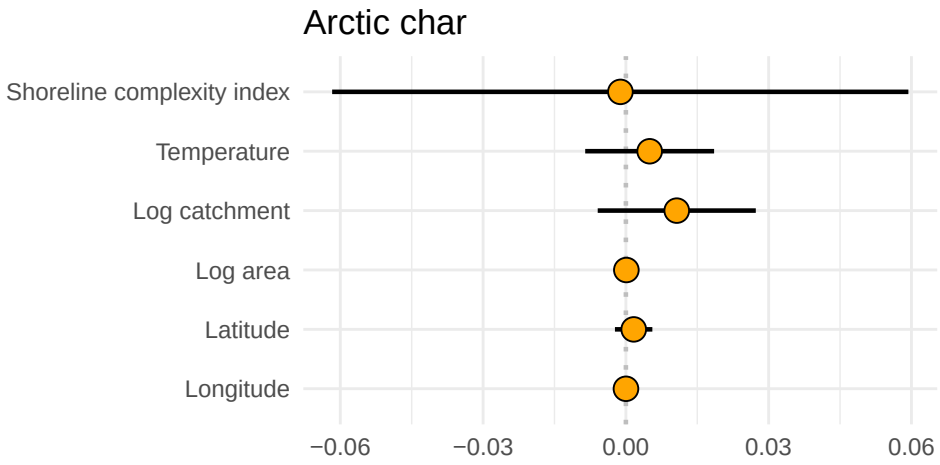


Figure C.2: Estimated coefficients and 95% credible intervals for the environmental covariates of the final model, on observations of Arctic char.

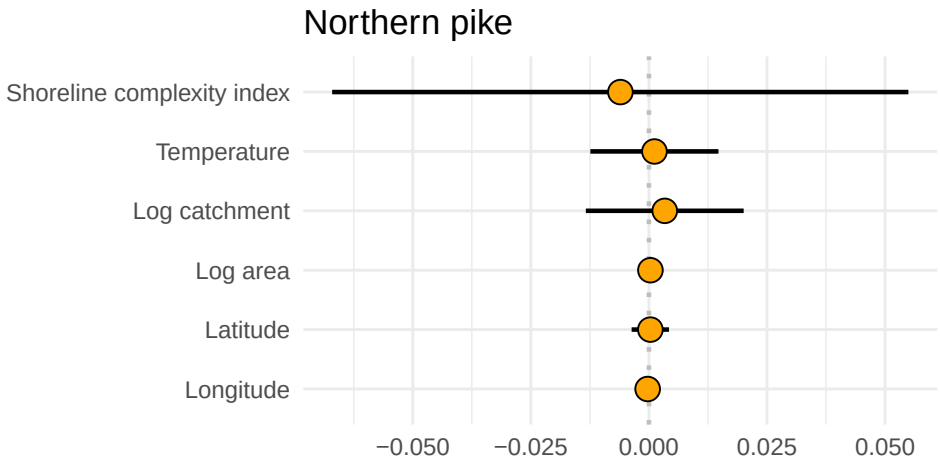
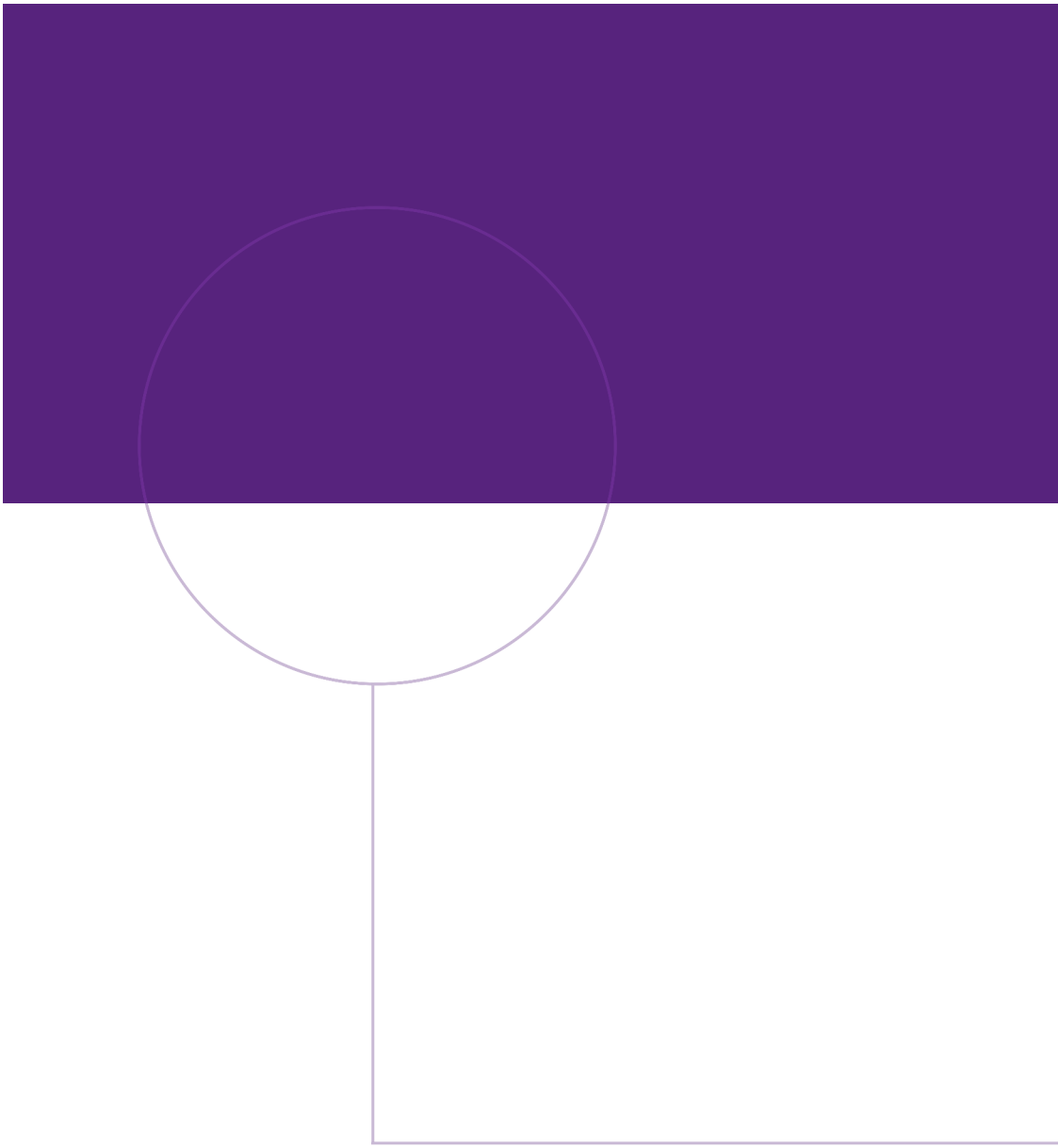


Figure C.3: Estimated coefficients and 95% credible intervals for the environmental covariates of the final model, on observations of northern pike.



NTNU

Norwegian University of
Science and Technology