University of Helsinki
Faculty of Science
Master's programme in Mathematics and Statistics

Master's thesis

Bayesian approach for modeling zero-inflated plant percent covers using spatial left-censored beta regression

Juho Pietilä

Supervisor:
Jarno Vanhatalo

May 2025

Title: Bayesian approach for modeling zero-inflated plant percent covers using spatial left-censored beta regression
Author: Juho Pietilä
Month and year: May 2025
Page count: 63 pp.

Abstract:
Species distribution models are tools to combine species observations to environmental data. These models learn about the species' ecological preferences and can be used to predict their distributions in unsampled locations or under climate change scenarios. This information is essential to learn about species' responses to climate change and further to aid conservation decisions.

Typical challenge with ecological data is the large number of zeros: most of the time species are not observed at the sampling location. Zeros can be thought to arise from ecological reasons or due to randomness. This problem of zero inflation is well studied for discrete positive data, such as counts and occurrences. When sampling plant species, it is logical to record their percent cover, leading to observations between zero and one. With improving technologies, this kind of data is becoming increasingly popular while the need to model vegetation communities is high. However, zero-inflated models for continuous positive data have not been widely studied.

One way to build models for percent cover data is to use beta distribution, utilizing left-censoring to create probability mass for zero observations. In this study, alternative versions of left-censored beta regression are implemented and their behavior studied when applied to real data from the Baltic Sea. Specifically, the model is extended by introducing spatial random effects and a separate process for zeros arising from a sites unsuitability for a species.

This study finds that including spatial random effects improves the predictive performance of the model, whereas accounting for two sources of zeros does not make a big difference. It is also found that modeling the precision of the beta distribution with covariates is highly beneficial for predictive ability and essential to make the model catch the patterns in the observed data. This result is considered important, since the existing studies typically assume a constant precision parameter across sampling locations.

On top of the predictive performance, this study shows that the model specification can have a large effect on the identified hotspot areas. This could further affect the conservation actions informed by these models. The result serves as a reminder for proper model comparison and model assessment when doing Bayesian analysis.

Keywords: species distribution modeling, percent cover data, beta regression, Bayesian hierarchical models, spatial random effects

Contents

Chapter 1

Introduction

Species distribution models (SDM) are a set of tools to link species observations with environmental covariates [11]. These models aim to learn about the species' ecological preferences and further predict the distributions in unsampled locations to end up with thematic maps of, for example, species' probability of occurrence or expected abundance. Further, climate change scenarios can be used to study the effects of climate change and the results used to identify and plan conservation areas [20, 28].

Typical challenge for species observation data is the large number of zeros. Dataset is called zero-inflated if it includes more zeros than assumed by standard statistical distributions (e.g. Gaussian, Poisson, negative-binomial, beta) [30]. Zeros can be thought to arise from different sources. Species might not be observed for ecological reasons: the habitat is not suitable for the species, and thus it does not appear. As well, species might not be observed even if the environment was suitable. This can be related to pure randomness: the species does not saturate its entire suitable habitat [30]. In the literature these zeros have been called structural and random (or stochastic) zeros, respectively [2].

Ignoring zero inflation in the dataset may lead to biased estimation of the parameters and overestimation of standard errors [27, 29]. A way to tackle zero inflation is to use models that account for extra zeros, typically by introducing a latent binomial process. In zero-inflated models, this binomial process predicts the suitability of the location. If the location is unsuitable, the process outputs structural zeros. If the location is suitable, a separate count process (e.g. Poisson) produces non-negative observations. Zeros coming from the count process represent the stochastic zeros. This way the model can produce zeros from two distinct processes, representing the two different types of zeros (structural and stochastic).

Another typical challenge for modeling task is spatial autocorrelation: observations nearby tend to be more similar to each other than observations far apart. This spatial autocorrelation caused by different ecological factors

tends to remain in the residuals even after multiple environmental covariates are included in the model [28]. This can be thought to arise from missing covariates or unobserved environmental processes. As with zero inflation, ignoring spatial autocorrelation may harm the statistical inference through inaccurate parameter estimation and uncertainty quantification [45]. One way to approach spatial dependence in the data is to introduce spatially correlated random effects into the model [28]. This can be achieved using Gaussian processes, potentially improving the predictive ability of the model [41].

Species distribution models and the zero-inflated versions of them are well studied for discrete (presence/absence, count) data, but not that well for continuous positive data [39]. Counting individuals does not always make sense. For example vegetation observations are typically measured as percent covers, each value representing the relative area of the species when projected onto the surface [7]. This sort of data is also expected to become increasingly popular with improving technologies [26]. Modeling of vegetation patterns might be of interest, for example, when studying carbon intake from vegetation communities.

A natural approach for modeling percent covers between in $[0, 1]$ interval is to use beta regression. An obvious problem with continuous positive distributions is that they do not have separate mass at 0. One way to correct this is to use left-censored beta regression, where a beta-distributed random variable V is rescaled to have its support on $(-a, 1)$ for some $a > 0$ and setting $\mathbb{P}(Y = 0) = \mathbb{P}(V \leq 0)$. This approach is used by Tang et al. [39] to introduce spatial and zero-inflated models for modeling percent cover data.

In this thesis, left-censored beta regression is implemented to study the distribution of an algae species, using zero-inflated percent cover data from the Baltic Sea. On top of the basic left-censored beta model, also zero-inflated, spatial and their combination are implemented. The goal is to study the behavior of these four different models while keeping in mind the typical objectives of species distribution models: predictive ability, ecological inference and hotspot identification. Three main questions asked are:

1. Which of the alternative models is best in terms of predictive ability? This is answered by calculating leave-one-out cross-validation log-scores for each model.

2. How does the model specification affect the ecological inference? This is answered by qualitatively examining the response curves (how each environmental covariate relates to the expected coverage of the species).

3. How does the model specification affect the hotspot/conservation area identification for the species? This is done by creating maps of species

hotspots for each model and examining their differences both qualitatively and quantitatively.

To conclude, the idea of this thesis is to take recently introduced model for zero-inflated percent cover data and study behavior of its different versions (zero-inflated, spatial and both) for an algae species from the Baltic Sea. Rather than interpreting the results from ecological perspective, the interest is more in the statistical model and it's behavior.

Chapter 2

Theoretical Background

2.1 Bayesian inference

Statistical inference is mainly about drawing conclusions about model parameters $\boldsymbol{\theta}$ given observations $\mathbf{D}$, which typically consists of observations y and covariates $\mathbf{x}$, such that $\mathbf{D} = \{y_i, \mathbf{x_i}\}_{i=1}^n$. In Bayesian inference, it all comes down to Bayes theorem. The fundamentals of Bayesian inference, Bayesian data analysis and Bayesian computation are well covered by Gelman et al. [17]. The theory for Sections 2.1-2.3 comes mainly from that book, which serves as a great introduction to Bayesian statistics for a curious reader.

Let $p(\cdot)$ denote a probability distribution. Bayes theorem states that

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})}{p(\mathbf{D})} \ . \tag{2.1}$$

Note that here $\boldsymbol{\theta}$ is bolded for vector notation, but it could be a scalar as well. Let's see what is inside the equation. $p(\boldsymbol{\theta}|\mathbf{D})$ is called posterior distribution, and this is what we ultimately are after. It gives the probability distribution for the model parameters given the observed data. To achieve the posterior distribution, the two following components must be specified.

Prior distribution $p(\boldsymbol{\theta})$ includes the modelers prior beliefs about the values of model parameters. It can be non-informative, such as entirely flat over the possible parameter values, or informative, leading the posterior distribution to direction where we believe the values should be. Informative priors can be formed e.g. using previous data or information gathered from experts of the subject. A priori model parameters are typically assumed independent, such that the joint prior factorizes. For example in linear regression $p(\boldsymbol{\theta}) = p(\mu, \sigma^2) = p(\mu)p(\sigma^2)$, meaning that mean and variance can be given their own univariate prior distributions.

The sampling distribution (or likelihood function) $p(\mathbf{D}|\boldsymbol{\theta})$ tells the probability distribution for observed data given the model parameters. Since

typically the covariates $\mathbf{x}$ are considered fixed, in our notation the sampling distribution can be written $p(\mathbf{D}|\boldsymbol{\theta}) = p(y|\mathbf{x}, \boldsymbol{\theta})$. Again using linear regression as an example, $y|\mathbf{x}, \boldsymbol{\theta} \sim \mathrm{N}(\mathbf{x}^\mathbf{T}\boldsymbol{\beta}, \sigma^2)$ and $p(y|\mathbf{x}, \boldsymbol{\theta})$ would be the corresponding probability density function.

$p(\mathbf{D}) = \int \mathbf{p}(\mathbf{D}, \boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta} = \int p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ in the denominator is called marginal likelihood (or normalizing constant). It does not depend on the model parameters $\boldsymbol{\theta}$ and considering the inference might be uninteresting. It is still needed to make a proper posterior distribution that integrates to 1. Another name for the denominator, prior predictive distribution [17], is quite informative name as well. Indeed, it gives the distribution for observations, weighting over the prior distribution $p(\boldsymbol{\theta})$. This distribution can be e.g. used when selecting prior distribution to see what kind of data they produce (so-called prior predictive check [17]).

Once the posterior distribution is obtained, we are usually interested in predicting new data $\tilde{y}$ given the covariates $\tilde{\mathbf{x}}$. Especially we are interested in the posterior predictive distribution $p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D})$, which can be obtained by

$$p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{D}) = \int p(\tilde{y}, \boldsymbol{\theta}|\tilde{\mathbf{x}}, \mathbf{D})d\boldsymbol{\theta} \tag{2.2}$$

$$= \int p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta}, \mathbf{D})p(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\theta} \tag{2.3}$$

$$= \int p(\tilde{y}|\tilde{\mathbf{x}}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{D})d\boldsymbol{\theta} \ . \tag{2.4}$$

Once again, posterior predictive distribution can be in certain cases analytically solved, but usually not. However, if we are able to sample from posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$, we are also able to sample from posterior predictive distribution for $\tilde{y}|\tilde{\mathbf{x}}$. In addition to making predictions, posterior predictive distributions can be utilized in model checking (posterior predictive checks [17]) and model comparison, both discussed in Section 2.3.

2.2 Bayesian inference in practice

The major interest in Bayesian inference is the posterior distribution in (2.1). In simple cases (e.g. Normal prior, Normal likelihood) this distribution can be analytically solved, and the inference can be done with pen and paper. However, typically solving by hand is not possible. Especially the integral in the denominator causes problems in analytical solutions. However, there are multiple ways of sampling from the posterior distribution using the unnormalized posterior density $p(\boldsymbol{\theta})p(\mathbf{D}|\boldsymbol{\theta})$, which is always well defined. Great overview of these sampling methods can be found from Chapters 11 and 12 of the book by Gelman et al. [17].

Perhaps the most utilized method is so-called Markov chain Monte Carlo (MCMC) simulation [18]. The goal is to draw samples from Markov chain

(sequence of random variables $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \dots$) such that the chain convergences to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$, and eventually we are sampling from the posterior distribution.

One way to construct such a chain is called Metropolis-Hastings algorithm [21]. It works such that we start from initial value $\boldsymbol{\theta}_0$ and at each step propose a new parameter value $\boldsymbol{\theta}_*$ from proposal distribution $J_t(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})$, which depends on the previous parameter value. We accept the new proposed value $\boldsymbol{\theta}_*$ with probability

$$p = \min\{1, r\} \ , \tag{2.5}$$

where

$$r = \frac{p(\boldsymbol{\theta}_*)p(\mathbf{D}|\boldsymbol{\theta}_*)}{p(\boldsymbol{\theta}_{t-1})p(\mathbf{D}|\boldsymbol{\theta}_{t-1})} \times \frac{J(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_*)}{J(\boldsymbol{\theta}_*|\boldsymbol{\theta}_{t-1})} \ . \tag{2.6}$$

This algorithm works, but may be inefficient in converging to the target distribution because of the random walk behavior [17]. We might easily take too long steps to accept the new proposed values with high enough probability, or take too small steps such that the chain convergences but slowly. Hamiltonian Monte Carlo [10] borrows ideas from physics and uses gradient information of the log-posterior density to make better proposals and explore the target distribution more efficiently [17]. This allows for faster convergence. Widely used software Stan [4] is an user-friendly computer program to automatically apply Hamiltonian Monte Carlo (HMC) given a Bayesian model [17]. Stan is used for posterior sampling in this thesis.

After sampling from the posterior, what remains is to verify that our chains are actually converging to a target distribution. The main idea is to start multiple chains using different initial values and compare them. Especially we are interested in the variance within the chains versus the variance between the chains [17]. Intuitively, if the between-chains variance is much larger than within-chains variance, there is indication that our chains have not converged to same target distribution. A measure introduced by Gelman and Rubin [16], that relies on these two variances is called potential scale reduction factor (or $\hat{R}$-statistics)

$$\hat{R} = \sqrt{\frac{V}{W}} \ , \tag{2.7}$$

where

$$V = \frac{n-1}{n}W + \frac{1}{n}B \ , \tag{2.8}$$

and B, W are the between and within chain variances respectively. The minimum $\hat{R}$ value is 1, and any values above 1.05 are typically considered as indication of the chains not converging. Before proceeding to analyse the results, it is necessary to check that the posterior chains have converged.

It is noticeable that since Markov chain is used for sampling from posterior, the samples include autocorrelation and are not fully independent.

Therefore, they do not carry as much information as independent samples. So-called effective number of simulation draws (n_{eff}) can be calculated as explained in Section 11.5 by Gelman et al. [17], and it, broadly speaking, tells the number of independent draws that the autocorrelated sample corresponds to.

2.3 Bayesian model checking and comparison

There are two fundamental parts of Bayesian data analysis that come after fitting your Bayesian model. First, we want to confirm the fit of the model. Second, if we have multiple alternative models, we want to be able to compare them to pick the best for our purposes. Again, great introductions to the topics can be found from the book by Gelman et al. [17] from Chapters 6 and 7. In this section, the basic ideas are discussed.

2.3.1 Model assessment

After fitting a model, it is relevant to examine whether it makes sense or not. One approach is the following: if the model fits well, it should generate data similar to the observed dataset. This can be done by posterior predictive checking, fully explained in Gelman et al. [17]. See also Conn et al. [5] for a review for ecologists. Samples from $p(\mathbf{y}_{\text{rep}}|\mathbf{D})$ can be generated and the replications $\mathbf{y}_{\text{rep}}$ compared to observations $\mathbf{y}$ in different ways.

One quite intuitive way is to plot a histogram of $\mathbf{y}_{\text{rep}}$ and compare it to histogram of $\mathbf{y}$. To see an example, Figure 5.3a shows histogram of observed percent cover with 15 replicated histograms that were produced by fitted model. However, if multiple response variables are modeled with many alternative models, this could become quite a laborious way to check the model fits.

Thus, more compact way to examine discrepancies is to use test quantities $T(\mathbf{y})$. A simple test quantity could be the sample mean

$$T(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \ . \tag{2.9}$$

Again, histograms of $T(\mathbf{y}_{\text{rep}})$ can be drawn and a vertical line included for the observed test quantity $T(\mathbf{y})$. Further tail-area probabilities (or Bayesian p-values) can be calculated by $\mathbb{P}(T(\mathbf{y}_{\text{rep}}) \geq T(\mathbf{y}))$ [17]. Since this thesis deals with zero-inflated data, one can think of other test quantities to measure the disrepancy between predicted and observed datasets. The most obvious is the proportion of zeros

$$T(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i = 0) \ . \tag{2.10}$$

As well, instead of using the mean of the whole data, we could instead use the mean of the positive observations to get a better understanding of how the model handles the non-zero part of the data

$$T(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{for all } y_i > 0. \tag{2.11}$$

Yet another test quantity used in this thesis is the maximum observed coverage

$$T(\mathbf{y}) = \max(\mathbf{y}) \ . \tag{2.12}$$

To see a practical example, Figure 5.3b shows histograms of these quantities for 200 replicated datasets, while the vertical line indicates the test quantity for the observed dataset.

This approach of using test quantities makes it easier to examine the results, as one model with one test quantity produces only one histogram. Further, in case of multiple species the Bayesian p-values could be represented as a table where each row corresponds to species and each column to a model.

2.3.2 Model comparison

Model comparison is a central part of the analysis to find out the model best suitable for the task, and is well revised for ecologists by Hooten and Hobbs [22]. Vehtari et al. have published multiple ([33, 42, 43]) comprehensive papers about Bayesian model selection as well.

A typical approach when considering alternative models is to measure their predictive ability. Model is considered good if it predicts future data well. For point predictions typical measure of accuracy is the mean square error

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ , \tag{2.13}$$

where $\hat{y}_i$ is the predicted value, such as posterior mean $\mathbb{E}[y_i|\mathbf{x}_i, \mathbf{D}]$. For probabilistic predictions, common approach is to use log-predictive density

$$\log p(y_i|\mathbf{x}_i, \mathbf{D}) \ , \tag{2.14}$$

which is known to have desirable properties, for example being the unique local and proper scoring rule [17]. A great model would on average predict well the future observations. If we knew the true data generating process $p(\tilde{y}_i|\tilde{\mathbf{x}}_i)$ for future data, we could calculate the expected log-predictive density

$$\text{ELPD} = \int \log p(\tilde{y}_i|\tilde{\mathbf{x}}_i, \mathbf{D}) p(\tilde{y}_i|\tilde{\mathbf{x}}_i) d\tilde{y}_i \ . \tag{2.15}$$

Similarly for n data points one could sum up expected pointwise log-predictive density

$$\text{ELPPD} = \sum_{i=1}^{n} \int \log p(\tilde{y}_i|\tilde{\mathbf{x}}_i, \mathbf{D}) p(\tilde{y}_i|\tilde{\mathbf{x}}_i) d\tilde{y}_i \ . \tag{2.16}$$

Of course, the true data generating process for future data is not known (otherwise there would be no need for statistical modeling), and we need tools to approximate the equation (2.16). One way we could approximate this is to use the observed data to calculate the pointwise log-predictive density

$$\text{LPPD} = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \mathbf{D}) \ . \tag{2.17}$$

This can easily be calculated from posterior sample, but intuitively overestimates the predictive accuracy of the model in general. Many widely used information criteria (AIC, BIC, WAIC) indeed are based on approximating ELPPD. The main idea is to start from (2.17) and correct for the bias coming from using the same data as used for fitting the model by subtracting a value e.g. proportional to the number of parameters used in the data [17].

Another way of approximating ELPPD is to use leave-one-out (LOO) cross-validation. There the model is fit n times, each time leaving one data point out of the training data, and then evaluating it's posterior log-predictive density. Eventually we get to estimate the ELPPD with

$$\text{LOO-LPPD} = \sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \mathbf{D}_{-i}) \ , \tag{2.18}$$

where $\mathbf{D}_{-i}$ denotes a dataset where the i:th observation has been left out. The problem with this approach is that it becomes computationally expensive to run the model separately n times. Typical alternative is to use K-fold cross-validation, dividing the data into K distinct parts, having each at a time as a validation set. This still requires K runs of the model, but can be much more efficient if $n >> K$.

Fortunately there are methods to get the LOO values using the posterior sample from the single model using all the data, one introduced by Vehtari et al. [42]. Using S posterior samples, one way is to use importance sampling approach where we evaluate

$$p(\tilde{y}_i|\tilde{\mathbf{x}}_i, \mathbf{D}_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^s p(\tilde{y}_i|\tilde{\mathbf{x}}_i, \boldsymbol{\theta}^s)}{\sum_{s=1}^{S} r_i^s} \tag{2.19}$$

with importance ratios

$$r_i^s = \frac{1}{p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^s)} \propto \frac{p(\boldsymbol{\theta}^s|\mathbf{D}_{-i})}{p(\boldsymbol{\theta}^s|\mathbf{D})} \ . \tag{2.20}$$

Plugging in $\tilde{y}_i = y_i$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i$ for the LOO-LPD at the heldout point, we get

$$p(y_i|\mathbf{x}_i, \mathbf{D}_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^s)}} \ . \tag{2.21}$$

This can be already used to estimate the predictive density for heldout point, but can cause instability due to high variance of importance ratios [42]. In Pareto smoothed importance sampling (PSIS) the importance ratios are smoothed by fitting a generalized Pareto distribution to the possibly heavy right tail of the ratios. By using fitted values from this distribution instead of raw importance ratios, we get more stable ratios. PSIS-LOO approach is fully explained by Vehtari et al. [42], and is implemented in R-package 'loo' [44]. PSIS-LOO is used in this thesis to evaluate the predictive performance of the model.

Predictive accuracy in extrapolation tasks

Recalling the nature of typical ecological data, we could write the expected log-predictive density as

$$\int \log p(\tilde{y}|\mathbf{x}, \mathbf{s}, \mathbf{D}) p(\tilde{\mathbf{D}}) d\tilde{\mathbf{D}} \tag{2.22}$$

$$= \int \log p(\tilde{y}|\mathbf{x}, \mathbf{s}, \mathbf{D}) p(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}) p(\tilde{\mathbf{x}}|\tilde{\mathbf{s}}) p(\tilde{\mathbf{s}}) d\tilde{y} d\tilde{\mathbf{x}} d\tilde{\mathbf{s}} \ . \tag{2.23}$$

It is notable that in typical tasks in species distribution modeling we might be either extrapolating in space (predicting in new locations) or environment (predicting in new conditions), and thus the distributions of locations and covariates in our training data do not match the corresponding distributions for predictions. More specifically, $p(\mathbf{x}|\mathbf{s}) \not\approx p(\tilde{\mathbf{x}}|\tilde{\mathbf{s}})$ or/and $p(\mathbf{s}) \not\approx p(\tilde{\mathbf{s}})$. Thus the methods introduced do not estimate the correct measure, since they are inherently assuming we are interpolating. To account for covariate shift, for example importance weighted cross-validation have been suggested by Sugiyama et al. [38]. More generally for extrapolation tasks in ecology, so-called blocking methods have been introduced by Roberts et al. [35]. There the idea is to divide the data into e.g. spatially or environmentally distinct blocks and use K-fold cross-validation to estimate the predictive accuracy in the extrapolation task. However, due to simplicity of implementation and avoidance of extra computational burden, standard leave-one-out cross-validation is used in this thesis, recognizing it's possible problems in extrapolation tasks.

2.4 Species distribution modeling

Ecology studies interactions between species as well as interactions between species and environment. A basic tool in an ecologist's toolbox are species

distribution models (SDM), that are models to link species observations with environmental information, letting us understand the species niche preferences as well as predicting its abundance across a landscape [11]. Approaches for modeling species distributions had a significant rise in the late 1990s and are well reviewed by Guisan and Zimmermann [19]. Bayesian hierarchical modeling [15] is used in this thesis to construct spatially explicit species distribution models.

Typical tasks performed with species distribution models involve inferring the ecological responses or predicting the species abundance over region of interest, having sampled some locations within that region (interpolation). As well we could see how species abundances are predicted to change under different climate change scenarios (extrapolation). The latter usually requires extra care, since the model might be used with covariate values not seen by the model while fitting. These models have a great potential in aiding conservation decisions by providing information about critical areas for species preferences [36], although there have been a gap between theory coming from modelers and conservation acts made by practitioners [20, 46]. Still, many successful cases involving the use of SDMs in conservation planning exist, but have been poorly reported in scientific literature [20].

Since species distribution data are spatial in nature, typically the residuals are spatially autocorrelated even when environmental covariates have been used for modeling [28]. Points nearby tend to be more similar to each other than points far apart. This can be due to missing important covariates or due to ecological processes difficult to measure, such as competition, reproduction or dispersal [14]. Using models that ignore spatial dependence may lead to incorrect parameter estimation and quantification of uncertainty around them [14, 28]. Hierarchical modeling approach allows us to introduce latent spatial random effects to tackle this problem [28]. One way for producing spatially correlated random effects is to use Gaussian processes as done by Vanhatalo et al. [41].

This thesis works with single species distribution models, meaning that the outcome variable is one-dimensional. When interested in modeling a whole species community, these single SDMs could be stacked to generate community-level predictons. However, recently there has been a shift towards joint species distribution models (JSDM) that have the ability to take into account the interactions and correlations between species. For example a successful application in jointly modeling plant communities with competition for space has been recently published [25].

2.5 Zero inflation in ecological data

Species observation data have a tendency of containing many zeros, more than expected by standard probability distributions such as Poisson for

counts. If number of zeros is so large that the data do not fit the standard distributions, the dataset is called zero-inflated [30]. The zeros can be "true zeros" (as called by Martin et al. [30]), if they are caused by real ecological effect. A rare species might not be observed, because it has small population size and therefore low occurrence rate. Alternatively, species might not occur just because the sampling locations were not suitable for it. Therefore, true zeros can also be called "unsuitability zeros". Zeros can as well be "false zeros". Two main sources of false zeros are detection errors (species was there but not observed) and stochastic zeros (as called in by Tang et al. [39]). Stochastic zeros are such that the sampling location was indeed suitable for the species, but it was not there at the sampling time.

Not accounting for extra zeros in the data may lead to biased inference through, for example, biased estimates and uncertainties around parameters [2, 30]. Modeling approaches for zero-inflated data have been widely studied in the case of discrete (e.g. count, binary) data but not that much attention has been given to zero-inflated continuous data [39].

Data with many true zeros are usually modeled with either hurdle models or mixture models [30]. The idea in the former is to model the data in two stages. The presence of the species is modeled with binomial process, and positive observations are modeled with count process (e.g. truncated Poisson). Therefore, in this approach all the zeros arise from the same process. Eventually, the modeler is able to predict the probability of presence, and further the distribution of abundance given that the species is present.

In mixture models the zeros can arise from two processes. The suitability (instead of presence as in a hurdle model) is modeled with binomial process. If the location is suitable, the counts are modeled with count process (e.g. Poisson), that can also produce zeros. Therefore, the binomial process produces unsuitability zeros, and the count process produces either positive counts or stochastic zeros. Both the binomial process and the count process are typically modeled with environmental covariates, giving tools to infer how each process links to environmental conditions.

Another way of tackling the zero inflation due to true zeros is to use an alternative distribution, such as by replacing Poisson with negative binomial, that can account for overdispersion.

If data include false zeros, the mixture approach is required since we want to be able to separate between true and false zeros. Failing to account for false zeros may lead to false inferences about the relationship between species abundances and environmental covariates, and further harm management actions [30].

Chapter 3

Data

Typical ecological dataset can denoted by $\mathbf{D} = \{\mathbf{y}_i, \mathbf{x}_i, \mathbf{s}_i\}_{i=1}^n$ where $\mathbf{y}_i$ are the species observations at site i, $\mathbf{x}_i$ are the corresponding environmental covariates and $\mathbf{s}_i$ are the spatial coordinates. Let J denote the number of species and p the number of environmental covariates. Then $\mathbf{y}_i \in \mathbb{R}^J$, $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{s}_i \in \mathbb{R}^2$.

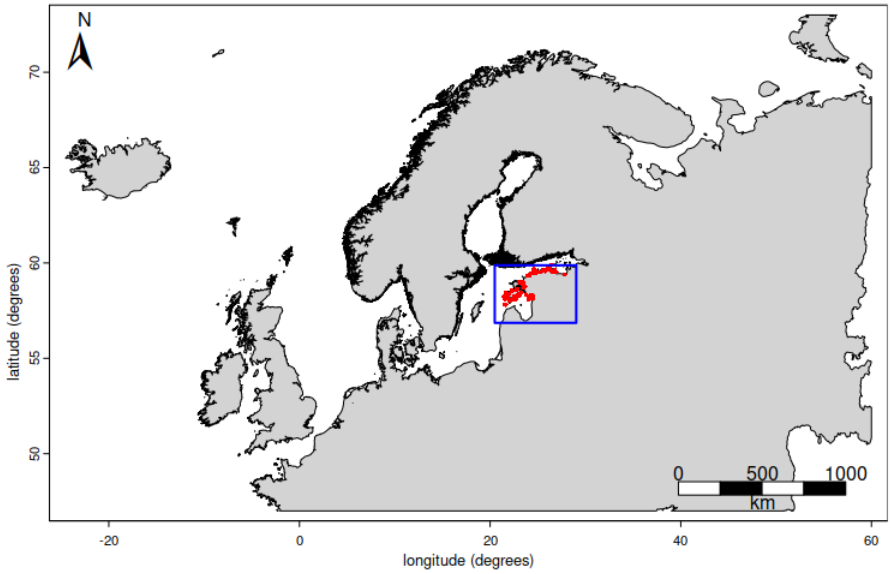
The dataset analysed in this thesis includes $n = 500$ species observations gathered in 2020-2021 over the coastal area of northern Estonia, visualized in Figure 3.1. All the samples are from summer months (May to September). Observations are coverages of different marine species and each data point comes with TM35FIN–coordinates $\mathbf{s}_i \in \mathbb{R}^2$ representing the easting and northing in meters. For the modeling task, several environmental covariates were also available. In the following two sections, the data are described in more detail.

3.1 Plant coverage data

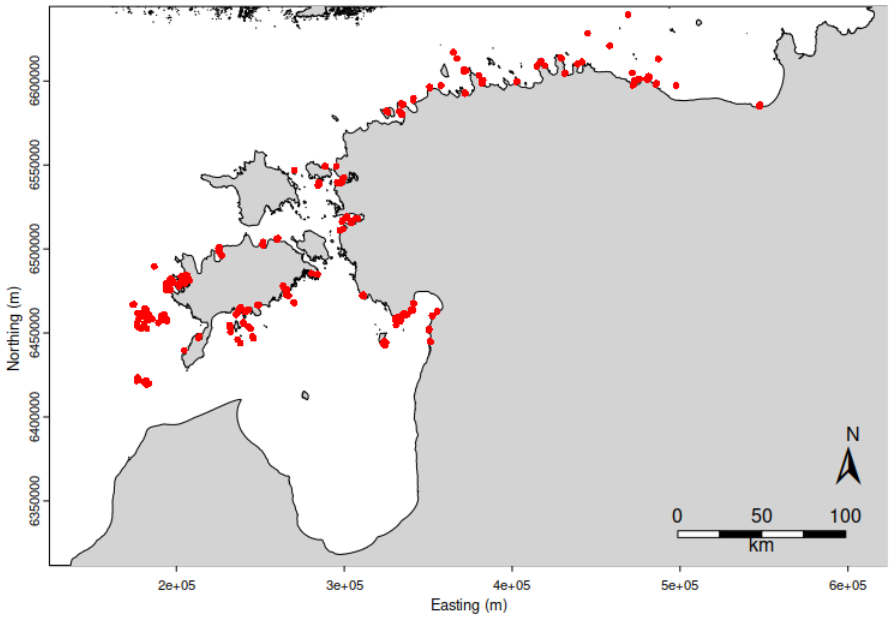
The species observations $\mathbf{y}_i \in \mathbb{R}^J$ tell the coverage for J water plant species. In practice, area from the bottom of the sea was examined, and the percentage of the area occupied by each J species was recorded. Let y_{ij} denote the coverage of species j at location i. Now each $y_{ij} \in [0, 100]$, but the sum over species $\sum_{j=1}^J y_{ij}$ can exceed 100, since vegetation can grow on top of each other.

From total of 52 species in the dataset, 16 were selected for further examination. All the relevant species for the Baltic Sea were tried to be included. The comprehensive list of the species along with their functional groups can be read from table C.1. However, this thesis eventually performs the modeling task for only one species.

As typical for ecological observations, there are many zeroes in the data. Figure 3.2 shows the distribution of coverages for each species. Every species have a clear peak at zero and the prevalences (proportion of non-zero obser-



(a) Study area is on the Northern coast of Estonia.



(b) Observation locations as red points.

Figure 3.1: Visual description of the study area.

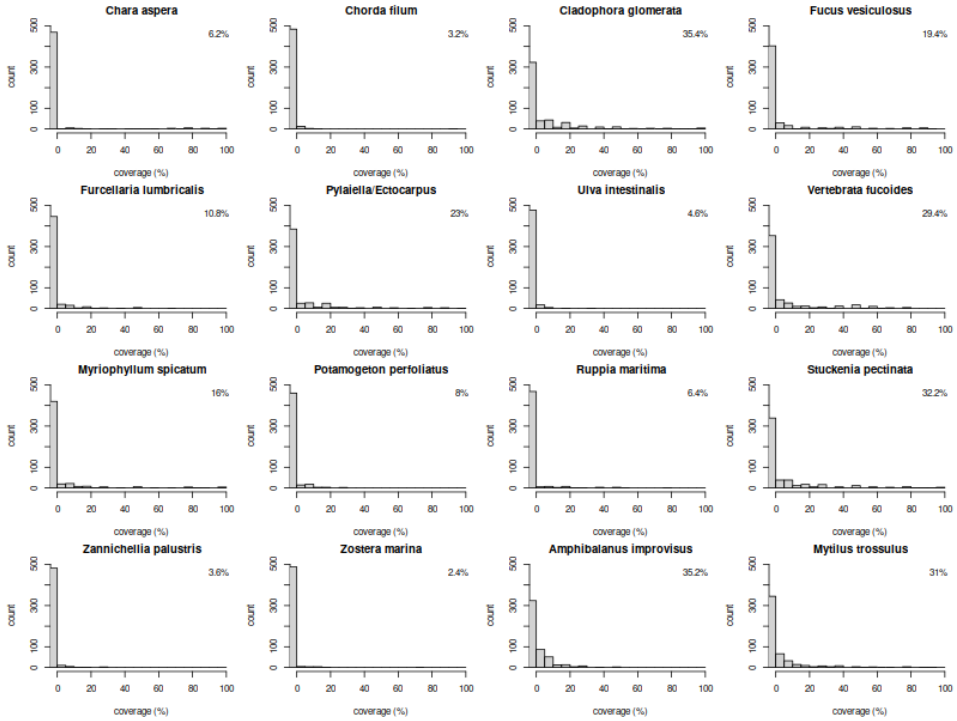


Figure 3.2: Distributions of percent covers in the data. Every species have a peak at zero. Number in the top corner tells the proportion of non-zero observations.

vations) vary between 2.4% and 35.4% among the species.

As mentioned, only one species is fully examined in this thesis, focus being in the differences between increasingly complex models. However, gathering the results from multiple species could give more information on the model behavior in general, and is discussed in Chapter 7.

The species selected for the full analysis was *Cladophora glomerata*. Looking at the Figure 3.2, there are qualitatively two different types of histograms. Roughly half of the species have prevalence around 5% and thus the histogram is very concentrated at zero. The other half has larger prevalence around 30% and there is considerable mass on small positive values. The chosen species *Cladophora glomerata* belongs to the latter group, which was in the analysis showing interesting problems in the posterior predictive checks, as seen later in Section 5.1. If yet another species was to be chosen for further analysis, it would be logical to take it from the group having almost all the mass at 0.

Further, the species consisted mostly of macro algaes (first two rows in Figure 3.2) and vascular plants (third row and first two of the fourth row), only exceptions being *Amphibalanus improvisus* (crustacean) and *Mytilus*

Table 3.1: List of covariates used for modeling.

Covariate	Description	Resolution
depth	Depth of the water column	1 km
nitrate	Nitrate concentration in the near-bottom layer	$\approx$ 1.6km
oxygen	Oxygen concentration in the near-bottom layer	$\approx$ 1.6km
phosphate	Phosphate concentration in the near-bottom layer	$\approx$ 1.6km
Secchi depth	Maximum depth a Secchi disk is seen underwater	$\approx$ 1.6km
temperature	Seawater temperature near the sefloor	$\approx$ 1.6km
salinity	Salinity in the bottom layer of seawater	$\approx$ 1.6km
current	Current velocity magnitude	$\approx$ 1.6km
chlorophyll	Chlorophyll-a concentration in the near-bottom layer	$\approx$ 1.6km
light level	$\exp(-1.7\times\frac{\text{depth}}{\text{Secchi depth}})$, light availability at the bottom layer	$\approx$ 1.6km

trossulus (mollusc). Since the analysed species *Cladomora glomerata* belongs to macro algaes, the logical second species to analyse would be for example *Zannichellia palustris*, vascular plant with very small prevalence. This way we would have examples from two qualitatively different species, giving more understanding on the performance of the model introduced.

3.2 Environmental covariates

In addition to the species coverages, multiple environmental covariates are presented. These include covariates such as depth of water, salinity, temperature and concentrations for multiple nutrients at the bottom layer. The full list of covariates used for modeling can be read from Table 3.1.

The availability of light directly affects the occurrence of aquatic vegetation [37]. Two covariates relevant to the bottom layer light level are depth and Secchi depth. The latter informs the depth in which a white disk is still visible [37]. Thus, it serves as a proxy for water clarity: the deeper the water and the smaller the Secchi depth, the light reaches the bottom layer. On the contrary, the lower the water and the higher the Secchi depth, the more light reaches the bottom layer. Secchi depth itself may not be that interesting as a covariate, but these two can be used to approximate the light level at the bottom layer, being ecologically more meaningful covariate.

One approach (used by Sahla et al. [37]) is to divide the water depth by Secchi depth, the ratio serving as a measure for light availability. However, another approach was taken to calculate the proportion of light reaching the bottom layer, having clear interpretation. It is known that if the light intensity in the surface is I_0, the light intensity at depth d is $I_0 e^{-kd}$ where a widely used estimate for light attenuation coefficient $k = 1.7/\text{secchi}$ [23, 34]. Thus, using $e^{-1.7\times\text{depth/secchi}}$ serves as an estimate for proportion of light intensity compared to intensity on the surface. This measure was

used instead of Secchi depth in the modeling task to represent the light availability. Depth was still used as one of the covariates, since it may have ecological meaning itself. For example, species might have more shelter in the deep, or depth might correlate with some unobserved but ecologically meaningful covariate.

For predictions, the depth was available as 1km $\times$ 1km raster. All the other variables were available as an even grid in terms of latitudes and longitudes (0.017×0.028 degrees). The grid does not remain even after projecting this data into TM35FIN–coordinates, distance to the nearest measure location being approximately 1.6km. The prediction was conducted at 1km $\times$ 1km resolution, such that covariate values from the nearest available location was used. The environmental covariates have been collected monthly over multiple years (1993-2021), and the dataset used for modeling is from summer months (May to Septemeber) in 2021 and 2022. Thus, the covariate values for prediction task were decided to be taken from July 2021.

The environmental covariates had very different ranges. Thus, for the modeling task all the covariates were zero-centered and scaled to have variance of 1. For covariate x_i, zero-centered and scaled version $\dot{x}_i$ is achieved by

$$\dot{x}_i = \frac{x_i - \bar{x}_i}{\mathrm{sd}(x_i)} \tag{3.1}$$

where $\bar{x}_i = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the mean of the covariate and $\mathrm{sd}(x_i)$ it's standard deviation.

Chapter 4

Methods

4.1 Left-censored beta regression

Since the observations are proportions and hence restricted between 0 and 100, using beta distribution as an observation model is tempting. Indeed, beta distribution is routinely used to model continuous plant cover data [7, 9]. However, beta distribution has support on open interval $(0, 1)$, which causes problems with data including large number of exact zeros. Separate mechanism is needed to create mass at zero. For this purpose, the left-censored Beta regression was introduced by Tang et al. [39].

Let $V \sim \text{Beta}(\alpha, \beta)$. More convenient way for modeling purposes is to reparameterize the beta distribution using mean and precision [12]. Denote $\mu = \frac{\alpha}{\alpha+\beta}$ and $\rho = \alpha + \beta$. Now

$$\mathbb{E}[V] = \frac{\alpha}{\alpha + \beta} = \mu \tag{4.1}$$

and

$$\text{Var}[V] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = \frac{\mu(1 - \mu)}{1 + \rho}. \tag{4.2}$$

Now the original parameters could be recovered by $V \sim \text{Beta}(\mu\rho, (1 - \mu)\rho)$. Instead, from now on, for clarity, $\text{Beta}(\mu, \rho)$ is used.

So let $V \sim \text{Beta}(\mu, \rho)$ and let $a > 0$ be a real number. Now a transformation $W = (a+1)V - a$ has a support on $(-a, 1)$. Moreover, $\mathbb{P}(W \leq 0) = \mathbb{P}\left(V \leq \frac{a}{a+1}\right)$.

Set $Y = \max\{0,\text{W}\}$. The mass at zero can be calculated as

$$\mathbb{P}(Y = 0) = \mathbb{P}\left(V \leq \frac{a}{a + 1}\right). \tag{4.3}$$

For positive entries $y > 0$ the density can be calculated in terms of density for beta-distributed V

$$f_Y(y) = f_V(v)\frac{dv}{dy} = f_V\left(\frac{y + a}{a + 1}\right)\frac{1}{a + 1}\ . \tag{4.4}$$

From this on, we denote by $y \sim$ LC-Beta(μ, ρ) the random variable constructed in the way described above.

When analyzing the results, it is useful to be able to compute the expected values. If $y \sim$ LC-Beta(μ, ρ), what is the expectation $\mathbb{E}[y]$? Recalling that $y = \max\{0, W\}$ where $W = (a+1)V - a$ and $V \sim$ Beta$, (\mu, \rho)$

$$\mathbb{E}[y] = 0 \times \mathbb{P}(y = 0) + \int_0^1 y f_y(y) dy \tag{4.5}$$

$$= \int_0^1 y f_V \left(\frac{y+a}{a+1} \right) \frac{1}{a+1} dy \ , \tag{4.6}$$

where (4.4) was used to get the density for $y > 0$ and f_V denotes the probability density function of beta distribution. This integral can be approximated with numerical integration.

This formulation serves as a basis for modeling proportions including exact zeros. Note that the scalar a has to be set in the model formulation, but the choice is quite arbitrary. This parameter has no interpretation and is just a mechanism for generating mass at zero. Further, the recovery of probability of zero has been found to be robust to the choice of a [39]. This parameter could be used as a model parameter, but it would be competing with μ, since they both determine the probability of zero. Thus, throughout this thesis the typical choice of $a = 1$ is used, if not otherwise mentioned.

4.2 Including covariates

Now we can introduce covariates in the model as is usually the case with generalized linear models. Mean parameter μ is restricted between 0 and 1 and therefore logit-link is a logical choice to use. For the positive precision parameter ρ log-link can be used. A basic left-censored beta regression can be thus formulated as

$$\begin{aligned} y_i &= \max\{0, W_i\} \\ W_i &= (a+1)V_i - a \\ V_i &\sim \text{Beta}(\mu_i, \rho_i) \\ \text{logit}(\mu_i) &= \alpha + \mathbf{x}_i^T \boldsymbol{\beta} \\ \log(\rho_i) &= \delta + \mathbf{x}_i^T \boldsymbol{\gamma} \\ p(\alpha, \delta, \boldsymbol{\beta}, \boldsymbol{\gamma}) &\sim p() \ . \end{aligned}$$

Let's first forget the modeling of precision ρ but instead assume it to be constant over sampling locations. Thus, at first we focus on modeling mean μ_i with covariates. Modeling of ρ will be discussed in Section 4.6. Mean μ_i is modeled through second-order polynomial, where the second order coefficients are restricted to be strictly negative. The justification for

this restriction comes from the ecological niche theory, where the species' responses to environmental covariates are typically unimodal and bell-shaped [1]. This means that species tend to have ecological preferences such that they have their niche, and deviating to any direction is suboptimal. Let $\mathbf{x}_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]^T$ be a vector of p covariates at location i and further $\mathbf{z}_i = [x_{i1}^2, x_{i2}^2, \ldots, x_{ip}^2]^T$ be a vector of second-order terms of these covariates. Now, the simplest model used in this thesis takes the form

$$\begin{aligned} y_i &= \max\{0, W_i\} \\ W_i &= (a+1)V_i - a \\ V_i &\sim \text{Beta}(\mu_i, \rho) \\ \text{logit}(\mu_i) &= \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{\gamma} \\ p(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho) &\sim p() \ , \end{aligned}$$

where the probability distribution for y_i can be written

$$p(y_i) = \begin{cases} \mathbb{P}\left(V_i \leq \frac{a}{a+1}\right) & y_i = 0 \\ f_{V_i}\left(\frac{y_i+a}{a+1}\right) \times \frac{1}{a+1} & y_i > 0 \ . \end{cases}$$

4.3 Including spatial random effects

Basic generalized linear model does not account for the spatial autocorrelation typically observed in the ecological data. The occurrence of species at certain location is likely associated with the occurrence nearby. It can be expected that the covariates in our model do not fully explain this spatial dependency, and spatial patterns due to unobserved covariates remain. This problem can be approached by introducing spatial random effects into the model [28, 41]. This can be achieved using e.g. Gaussian processes, as in this thesis.

Let $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n]^T \in \mathbb{R}^{n\times 2}$ be a matrix of sampling locations and $k(\mathbf{s}, \mathbf{s}'|\boldsymbol{\theta})$ be a covariance function that calculates the covariance between two locations $\mathbf{s}$ and $\mathbf{s}'$. An example of covariance function, used throughout this thesis is the exponential covariance function

$$k(\mathbf{s}, \mathbf{s}'|l, \sigma^2) = \sigma^2 e^{-||\mathbf{s}-\mathbf{s}'||/l} \ . \tag{4.7}$$

Here σ^2 is the magnitude of the process and l is the length-scale parameter determining how fast the correlation decays as a function of distance. With the exponential covariance function in (4.7), it can for example be computed that the covariance decreases to 5% of it's maximum (σ^2) when $||\mathbf{s} - \mathbf{s}'|| = -\log(0.05)l \approx 3l$, namely, when the distance between locations $\mathbf{s}$ and $\mathbf{s}'$ equals three times the length-scale l.

Now consider $\mathbf{K}(\mathbf{S}|l,\sigma^2) \in \mathbb{R}^{n\times n}$ is a covariance matrix calculated such that $\mathbf{K}_{ij} = k(\mathbf{s}_i, \mathbf{s}_j|l,\sigma^2)$. Zero-centered, spatially correlated random effects can be introduced with

$$\boldsymbol{\phi}(\mathbf{S}) \sim \mathrm{N}(\mathbf{0}, \mathbf{K}(\mathbf{S}|l,\sigma^2)) \ . \tag{4.8}$$

where $\boldsymbol{\phi}(\mathbf{S}) = [\phi(\mathbf{s}_1), \phi(\mathbf{s}_2), \ldots, \phi(\mathbf{s}_n)]^T \in \mathbb{R}^{n\times 1}$ is a vector of spatial random effects for each sampling location. Another way of expressing this so called Gaussian Process (GP) is to write

$$\phi(\mathbf{s}) \sim \mathrm{GP}(0, k(\mathbf{s}, \mathbf{s}'|l,\sigma^2) \ . \tag{4.9}$$

Now the base model introduced in Section 4.2 can be extended by adding the spatial random effect to the linear predictor modeling the mean of the distribution μ_i

$$\begin{aligned}
y_i &= \max\{0, W_i\} \\
W_i &= (a+1)V_i - a \\
V_i &\sim \mathrm{Beta}(\mu_i, \rho) \\
\mathrm{logit}(\mu_i) &= \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{\gamma} + \phi(\mathbf{s}_i) \\
\phi(\mathbf{s}) &\sim \mathrm{GP}(0, k(\mathbf{s}, \mathbf{s}'|l,\sigma^2) \\
p(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, l, \sigma^2) &\sim p() \ .
\end{aligned}$$

Fitting a spatial model and predicting with it comes with extra computational burden, especially if the spatial resolution is high. For this reason, the spatial random effect was introduced in the model in coarse 20km$\times$20km resolution (see Figure 4.1). Consider observations are gathered from m spatial grid cells and the coordinates of grid cell centers are gathered in a matrix $\mathbf{C} \in \mathbb{R}^{m\times 2}$. The Gaussian process is now set on the coarse spatial grid by letting $\phi(\mathbf{c}) \sim \mathrm{GP}(0, k(\mathbf{c}, \mathbf{c}'|l,\sigma^2))$. Now let $P \in \mathbb{R}^{n\times m}$ be a matrix indicating in which of the m grid cells each of the n observations belong to, such that

$$P_{ij} = \begin{cases} 1 \ , \text{if } i\text{:th observation is in } j\text{:th grid cell} \\ 0 \ , \text{otherwise} \ . \end{cases}$$

Note that each row sums up to 1. Now we can select a spatial random effect for each observation using $\mathbf{P}\boldsymbol{\phi}$. The final form of the spatial model used in this thesis becomes

$$\begin{aligned}
y_i &= \max\{0, W_i\} \\
W_i &= (a+1)V_i - a \\
V_i &\sim \mathrm{Beta}(\mu_i, \rho) \\
\mathrm{logit}(\mu_i) &= \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + \mathbf{z}_i^T\boldsymbol{\gamma} + \mathbf{p}_i^T\boldsymbol{\phi} \\
\phi(\mathbf{c}) &\sim \mathrm{GP}(0, k(\mathbf{c}, \mathbf{c}'|l,\sigma^2) \\
p(\alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \rho, l, \sigma^2) &\sim p() \ ,
\end{aligned}$$

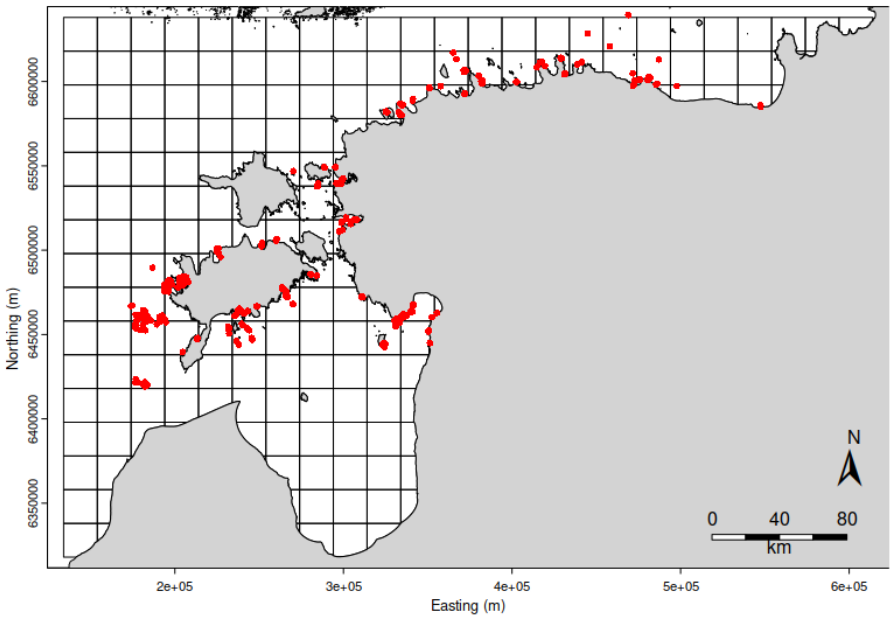


Figure 4.1: Spatial random effects are introduced in 20km×20km resolution. A grid cell can include multiple observations.

where $\mathbf{p}_i^T \in \mathbb{R}^{1\times m}$ denotes the i-th row of matrix $\mathbf{P}$ and $\boldsymbol{\phi} \in \mathbb{R}^{m\times 1}$ are the spatial random effects over coarse spatial grid.

4.4 Including zero-inflation

As seen from Figure 3.2, the observations for each species are heavily concentrated on exact zeros. All these zeros are not probably due to unsuitability of the sampling location, but there might be observations where the species did not happen to be on the site, even though the habitat was suitable. These zeros due to randomness can be tried to account for in modeling. Indeed, ignoring the zero inflation can lead to biased estimates and poor inference [2]. Intuitively, model can for example account for extra zeros by shrinking the coefficients $\boldsymbol{\beta}$ falsely towards zero.

Hurdle models and zero-inflated models are basic tools for tackling zero inflation, well covered, for example, in Chapter 11 of the book by Zuur et al. [48]. The basic idea is to introduce two processes, binomial process determining the suitability of the location and count process producing observations if the location is suitable. The key difference between hurdle models and zero-inflated models is that in hurdle models, the count process cannot produce zeros. In both models, unsuitability directly leads to zero observation.

With zero-inflated models, zero produced by count process corresponds to stochastic zeros, where the species is not around, even though the habitat is suitable. The latter approach is used in this thesis, since two sources of zeros are thought to be a reasonable assumption for plant species: there are probably locations suitable for the plant, but it did not disperse there yet and is thus not observed.

Let's formalize the model, with similar approach to [39]. Let Z_i be a binary latent variable such that

$$\begin{cases} Z_i = 0 & \text{location i is unsuitable} \\ Z_i = 1 & \text{location i is suitable .} \end{cases}$$

and denote $\mathbb{P}(Z_i = 1) = \pi_i$. The probability of suitability can be modeled with covariates using $\text{logit}(\pi_i) = \alpha + \mathbf{x}_i^T \boldsymbol{\beta}$. If $Z_i = 0$, it follows that $y_i = 0$ automatically. If $Z_i = 1$, then y_i comes from a left-censored beta distribution as in Section 4.2. The zero-inflated version of the model formalizes as

$$\begin{aligned} y_i &= \begin{cases} 0, \text{ if } Z_i = 0 \\ \max\{0, W_i\}, \text{ if } Z_i = 1 \end{cases} \\ Z_i &\sim \text{Bernoulli}(\pi_i) \\ W_i &= (a+1)V_i - a \\ V_i &\sim \text{Beta}(\mu_i, \rho) \\ \text{logit}(\pi_i) &= \alpha_\pi + \mathbf{x}_i^T \boldsymbol{\beta}_\pi + \mathbf{z}_i^T \boldsymbol{\gamma}_\pi \\ \text{logit}(\mu_i) &= \alpha_\mu + \mathbf{x}_i^T \boldsymbol{\beta}_\mu + \mathbf{z}_i^T \boldsymbol{\gamma}_\mu \\ p(\alpha_\pi, \alpha_\mu, \boldsymbol{\beta}_\pi, \boldsymbol{\beta}_\mu, \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\mu, \rho) &\sim p() \end{aligned}$$

and the probability distribution for y_i can be written

$$p(y_i) = \begin{cases} 1 - \pi_i + \pi_i \mathbb{P}\left(V_i \leq \frac{a}{a+1}\right) & y_i = 0 \\ \pi_i f_{V_i}\left(\frac{y_i + a}{a+1}\right) \times \frac{1}{a+1} & y_i > 0 \ . \end{cases}$$

Note that calculating the expectation for a zero-inflated random variable is exactly as without zero inflation part in (4.6), but the density is just multiplied by π_i.

In general, the covariates for modeling abundance and zero-inflation part do not have to be the same, and indeed they typically differ. For example, abiotic covariates could be used to model suitability, and biotic covariates to model abundance [40]. However, in this thesis all the covariates were used for both components.

4.5 Spatial zero-inflated model

For the final and the most complex model tested in this thesis, spatial random effects are included both for the part describing the mean of the beta

distribution and the binomial part of the zero-inflated model. The structure of this spatial zero-inflated model is

$$
\begin{aligned}
y_i &= \begin{cases} 0, \text{ if } Z_i = 0 \\ \max\{0, W_i\}, \text{ if } Z_i = 1 \end{cases} \\
Z_i &\sim \text{Bernoulli}(\pi_i) \\
W_i &= (a+1)V_i - a \\
V_i &\sim \text{Beta}(\mu_i, \rho) \\
\text{logit}(\pi_i) &= \alpha_\pi + \mathbf{x}_i^T \boldsymbol{\beta}_\pi + \mathbf{z}_i^T \boldsymbol{\gamma}_\pi + \mathbf{p}_i^T \boldsymbol{\phi}_\pi \\
\text{logit}(\mu_i) &= \alpha_\mu + \mathbf{x}_i^T \boldsymbol{\beta}_\mu + \mathbf{z}_i^T \boldsymbol{\gamma}_\mu + \mathbf{p}_i^T \boldsymbol{\phi}_\mu \\
\phi_\pi(\mathbf{c}) &\sim \text{GP}(0, k(\mathbf{c}, \mathbf{c}'|l_\pi, \sigma_\pi^2) \\
\phi_\mu(\mathbf{c}) &\sim \text{GP}(0, k(\mathbf{c}, \mathbf{c}'|l_\mu, \sigma_\mu^2) \\
p(\alpha_\pi, \alpha_\mu, \boldsymbol{\beta}_\pi, \boldsymbol{\beta}_\mu, \boldsymbol{\gamma}_\pi, \boldsymbol{\gamma}_\mu, \rho) &\sim p() \ .
\end{aligned}
$$

4.6 Modeling precision with covariates

As seen later in Section 5.1, the models introduced so far showed systematic problems in posterior predictive checks. Fitted models had e.g. hard time predicting small positive values, leading to overly large mean of the replicated datasets, as well as extra proportion of zeros. This seemed to be related to the fact that the precision parameter ρ of the beta distribution was common across sampling locations, not allowing for smaller variances for example near zero. This behavior is further examined in Section 5.1. We could expect to see, for example, a pattern where predicted values near zero have smaller variance than predicted values near 50% coverage.

Multiple papers ([9, 39]) mention that the precision parameter could be modeled with covariates using log-link

$$
\log(\rho) = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} \ , \tag{4.10}
$$

but none was found to actually implement it for the modeling task, obviously the main interest being in the mean μ. For this thesis this approach was tried but it introduced computational problems for MCMC sampling, most likely due to extreme values of ρ. Thus, another approach was used to model the precision with covariates. Namely, a scaled sigmoid function seemed to be a logical choice, giving natural tools to control the maximum value of ρ

$$
\rho = C \times \frac{1}{1 + \exp(-(\alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \mathbf{z}_i^T \boldsymbol{\gamma}))} \ , \tag{4.11}
$$

where again second order terms are included in the linear predictor to allow for U- or bell-shaped curves. This leaves a question, how to choose the

constant C controlling the maximum value of ρ. In our dataset, there is already error in measurements that arises from the fact that each coverage is reported with a number ending with 0 or 5. For example, 23% would round up to 25%, 82% to 80% etc. Thus, one should not expect precision larger than that in general. The beta distribution has the largest variance with $\mu = 0.5$, and thus one can, for example, examine Beta($\mu = 0.5, \rho$) with different choices of ρ. Our choice was $C = 1000$ to avoid sharper distributions than expected by pure measurement error, still providing the possibility of having sharp enough distribution with $\mu = 0.5$.

4.7 Prior selection

The summary of the priors used for modeling is in Table 4.1, and the following section includes the ideas behind the choice of prior distributions.

Since the covariates were scaled and zero-centered for the modeling, Gaussian distribution with variance of two was used as prior for the regression coefficients $\boldsymbol{\beta}$ in every model component, including the coefficients for the mean of the beta distribution, coefficients for binomial process in zero-inflated models and coefficients for modeling the precision ρ. This is not highly informative prior, but still sets some regularization on the coefficients. The coefficients $\boldsymbol{\gamma}$ for second order terms were restricted to be negative, half-Normal distribution with the same variance of two being a logical choice. Thus, for $j = 1, \ldots, p$

$$\begin{aligned} \beta_j &\sim \mathrm{N}(0, 2) \\ \gamma_j &\sim \mathrm{N}^{-}(0, 2) \ . \end{aligned}$$

For intercept terms α, same $\mathrm{N}(0, 2)$ prior was used for modeling the mean μ. This prior leads to quite uniform prior for $\mu = \frac{1}{1+e^{(-\alpha)}}$, which can with zero-centered covarites be thought as a mean in average environmental conditions. A bit more disperse $\mathrm{N}(0, 4)$ was given for the intercept α_ρ related to ρ. This gives a bit more prior probability also for small values of ρ.

The difficulty in setting priors turned out to be in α_π related to probability of suitability and α related to mean of the beta distribution. The identification problem of α and α_π is well discussed by Tang et al. [39, 40]. Since

$$\mathbb{P}(y = 0) = 1 - \pi + \pi \left(V \leq \frac{a}{a+1} | \mu, \rho \right) \ ,$$

where $\mathrm{logit}(\pi) = \alpha_\pi + \mathbf{x}^T \boldsymbol{\beta}_\pi$ and $\mathrm{logit}(\mu) = \alpha + \mathbf{x}^T \boldsymbol{\beta}$, intercepts α, α_π are competing to explain the zeros in the data. Intuitively, it is difficult to distinguish between unsuitability zeros and stochastic zeros in the data. The use of informative prior for either α or α_π is suggested by Tang et al.

[39], more natural choice being α_π since it can be easier to think of prior probabilities for suitability than abundance. The choice in this study was to give an informative $\mathrm{N}(-1, 0.25)$ for α_π to encourage unsuitability zeros instead of stochastic zeros. Since the data are zero-centered, one can think of the prior through probability of suitability in average conditions. In our data set, this means for example 7m water depth, 14.7°C water temperature and 43% light level at the bottom. The prior for probability of suitability in these conditions with the choice of $\alpha_\pi \sim \mathrm{N}(-1, 0.25)$ is centered around $1/(1 + e^1) \approx 0.27$.

What about the prior for ρ when it is not modeled with covariates? The precision parameter ρ for beta distribution is inversely linked to the variance of the distribution, as seen from equation (4.2). Thus, a prior with largest mass towards zero, but heavy tails was used. This ensures that the parameter can get large values if the likelihood is strong towards that direction. The choice for the prior distribution was half-Cauchy with mean $\mu = 0$ and scale $\sigma^2 = 10$.

The same logic as with precision parameter ρ was used with the length-scale and magnitude for the spatial covariance function. Half-Cauchy distribution was used to prefer smaller values, but heavy tails allowing large deviations if the data is informing towards that direction.

Since the coarse grid for the spatial random effects, the minimum distance between sampling points is 20km. 5% of the maximum covariance is retained at 20 km distance if the lengh-scale l equals 6.7. Thus, values of length-scale below that tells that there is basically no spatial correlation. Therefore, the location parameter for Half-Cauchy was set to eight instead of zero for the other parameters.

In opposite, one wants to avoid letting length-scale l too large: $l \to \infty$ corresponds to perfect correlation at any distance, spatial random effects being constant along the study area. The largest distance between spatial grid centers is around 560km. For example 50% of the maximum covariance is retained at 560km distance if $l \approx 808$. Much larger length-scales than that will lead to unneccessarily correlated random effects and can be controlled by the scale parameter of the prior distribution. For l_μ, half-Cauchy$(8, 50)$ was used. Length-scale l_π was showing unrealistically large values and student-t with 2 degrees of freedom, location parameter of 8 and scale of 50 prior was used instead. This prior distribution has less heavy tails than the corresponding half-Cauchy prior.

For the magnitude parameter σ^2, small scale parameter 0.1 was used to shrink it more heavily towards zero. The logic behind this is trying to prevent it from dominating the fixed effects if possible. Again, heavy tails still allowing for larger values. However, using Half-Cauchy resulted in problems in convergence. Therefore, a distribution with slightly thinner tails, student-t was used with zero mean, scale 0.1 as well as two and four degrees of freedom for magnitudes of spatial random effects ϕ_μ, ϕ_π respectively.

Table 4.1: Table of priors used in modeling.

Parameter	Prior distribution
α	N(0, 2)
α_ρ	N(0, 4)
α_π	N(−1, 0.25)
β_j	N(0, 2)
γ_j	N^-(0, 2)
ρ	Half-Cauchy(0, 10)
l_μ	Half-Cauchy(8, 50)
l_π	Student-t(2, 8, 50)
σ^2_μ	Student-t(2, 0, 0.1)
σ^2_π	Student-t(4, 0, 0.1)

Problems in the convergence of covariance function parameters were not a surprise. It is well known that the length-scale l and magnitude σ^2 are not identifiable, their ratio σ^2/l being more important than individual values for l and σ^2 [47]. This can be tackled by informative priors, and ways of constructing complexity penalizing priors for these parameters have recently been introduced, for example in [13]. However, since the problems with convergence were achieved with lighter tail priors, these more sophisticated approaches were omitted.

The idea of using heavy-tailed (half-Cauchy, student-t) priors for strictly positive parameters was taken from discussion on weakly informative priors for variance parameters in Gelman et al. [17].

4.8 Predictions

Fitting a model produces samples from $p(\boldsymbol{\theta}|\mathbf{D})$. Eventually, given set of covariates $\tilde{\mathbf{x}}$ and locations $\tilde{\mathbf{x}}$, one is interested in the posterior predictive distribution

$$p(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \mathbf{D}) = \int p(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \ . \tag{4.12}$$

Sampling from this distribution can be conducted by sampling $\tilde{y}^{(h)} \sim p(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \boldsymbol{\theta}^{(h)})$ for posterior draws $\boldsymbol{\theta}^{(h)} \sim p(\boldsymbol{\theta}|\mathbf{D})$, where $h = 1, \ldots, H$ and H is number of posterior samples.

In this thesis, rather than predicting $\tilde{y}$, the interest is in quantities such as probability of zero observation $\mathbb{P}(\tilde{y} = 0|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \mathbf{D})$ or expected covarage $\mathbb{E}(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \mathbf{D})$. In those cases, drawing $\tilde{y}^{(h)}$ can be skipped, and quantities $\mathbb{P}(\tilde{y} = 0|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \boldsymbol{\theta}^{(h)})$ or $\mathbb{E}(\tilde{y}|\tilde{\mathbf{x}}, \tilde{\mathbf{s}}, \boldsymbol{\theta}^{(h)})$ stored. In such cases, we ignore the aleatory uncertainty [31] in the observations, and the focus remains in the uncertainty over model parameters.

4.8.1 Non-spatial models

For models without spatial components, predicting is straightforward. For each $h = 1, \ldots, H$ draw

$$\tilde{y}^{(h)} \sim \text{LC-Beta}(\mu^{(h)}, \rho^{(h)}) \tag{4.13}$$

$$\text{logit}(\mu^{(h)}) = \alpha^{(h)} + \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}^{(h)} + \mathbf{z}_i^T \boldsymbol{\gamma}^{(h)} \ . \tag{4.14}$$

If the zero-inflation component is included, an extra step is taken by first sampling the suitability from the Bernoulli process

$$Z^{(h)} \sim \text{Bernoulli}(\pi^{(h)}) \tag{4.15}$$

$$\text{logit}(\pi^{(h)}) = \alpha_\pi^{(h)} + \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_\pi^{(h)} + \mathbf{z}_i^T \boldsymbol{\gamma}_\pi^{(h)} \tag{4.16}$$

If $Z^{(h)} = 0$, then $y^{(h)} = 0$ and $\tilde{y}^{(h)} \sim \text{LC-Beta}(\mu^{(h)}, \rho^{(h)})$ otherwise.

4.8.2 Spatial models

For the spatial models, procedure is exactly as for non-spatial models in Section 4.8.1, only difference being the spatial random effects $\phi_\mu^{(h)}$, $\phi_\pi^{(h)}$that are added to the linear terms. However, posterior samples for these spatial random effects have to be obtained first.

The spatial model produces posterior samples of length-scale l, magnitude σ^2 and spatial random effects $\boldsymbol{\phi}$ at observed locations. The objective is to predict the spatial random effects $\tilde{\boldsymbol{\phi}}$ at prediction locations. This is achieved by utilizing the properties of multivariate Gaussian distribution. Jointly one has

$$\begin{pmatrix} \boldsymbol{\phi}^{(h)} \\ \tilde{\boldsymbol{\phi}}^{(h)} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\phi,\phi}^{(h)} & \boldsymbol{\Sigma}_{\phi,\tilde{\phi}}^{(h)} \\ \boldsymbol{\Sigma}_{\tilde{\phi},\phi}^{(h)} & \boldsymbol{\Sigma}_{\tilde{\phi},\tilde{\phi}}^{(h)} \end{pmatrix} \right) \ , \tag{4.17}$$

where each block of covariance matrix is calculated using $l^{(h)}, \sigma^{2(h)}$. For multivariate Gaussian the conditional distribution is well defined, and hence the random effects for prediction locations come from

$$\tilde{\boldsymbol{\phi}}^{(h)} | \boldsymbol{\phi}^{(h)}, \sigma^{2(h)}, l^{(h)} \sim \text{N}\left(\boldsymbol{\Sigma}_{\tilde{\phi},\phi}^{(h)} \boldsymbol{\Sigma}_{\phi,\phi}^{-1(h)} \boldsymbol{\phi}^{(s)}, \boldsymbol{\Sigma}_{\tilde{\phi},\tilde{\phi}}^{(h)} - \boldsymbol{\Sigma}_{\tilde{\phi},\phi}^{(h)} \boldsymbol{\Sigma}_{\phi,\phi}^{-1(h)} \boldsymbol{\Sigma}_{\phi,\tilde{\phi}}^{(h)} \right) \tag{4.18}$$

From this onwards the procedure is exactly same than described in previous section for non-spatial case, but now

$$\text{logit}(\mu^{(h)}) = \alpha^{(h)} + \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}^{(h)} + \mathbf{z}_i^T \boldsymbol{\gamma}^{(h)} + \mathbf{p}_i^T \tilde{\boldsymbol{\phi}}_\mu^{(h)} \tag{4.19}$$

and

$$\text{logit}(\pi^{(h)}) = \alpha_\pi^{(h)} + \tilde{\mathbf{x}}_i^T \boldsymbol{\beta}_\pi^{(h)} + \mathbf{z}_i^T \boldsymbol{\gamma}_\pi^{(h)} + \mathbf{p}_i^T \tilde{\boldsymbol{\phi}}_\pi^{(h)} \tag{4.20}$$

where $\tilde{\boldsymbol{\phi}}_\mu^{(h)}, \tilde{\boldsymbol{\phi}}_\pi^{(h)} \in \mathbb{R}^{M \times 2}$ are the spatial random effects at M coarse grid cells over the study area and $\mathbf{p}_i \in \mathbb{R}^{M \times 1}$ indicates the grid cell that the i:th prediction location belongs to.

4.9 Implementation of the models

All the eight models introduced in this chapter were implemented in Stan. Four chains with 1000 iterations for each were run and the first half of the chains were discarded as burn-in period (to let the chains converge after initialization). This resulted in 2000 posterior samples of the model parameters.

The convergence of the chains was examined by looking at the $\hat{R}$-statistics. All the $\hat{R}$-statistics values were below the typical threshold 1.05, maximum value being around 1.02. Thus, no indications of problems in convergence were present.

Chapter 5

Results

The alternative models were fitted for only one species, *Gladophora glomerata*, selected from the set of 16 alternative species. For this species, the following sections examine the model assessment and model comparison as well as the effect of model specification on ecological inference and hotspot identification.

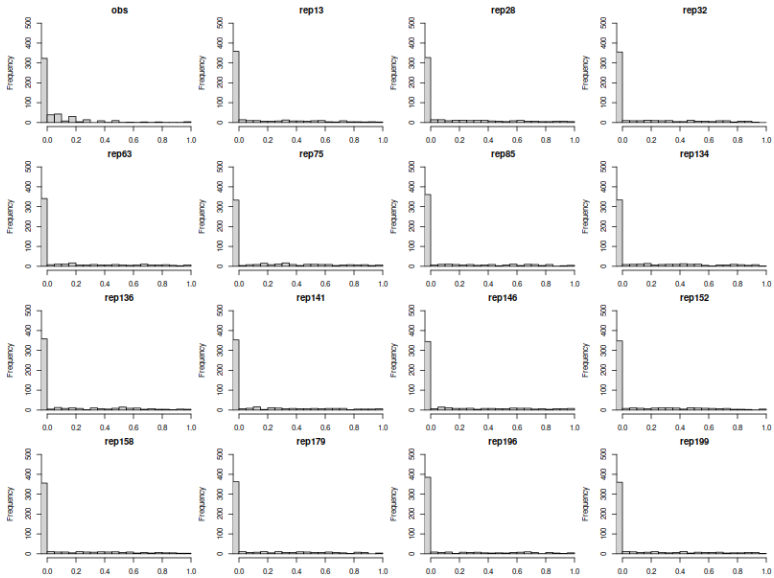
5.1 Model assessment

Model assessment was performed as described in Section 2.3. Fitted models were used to create replicated datasets $\mathbf{y}_{\text{rep}}$ by using 200 randomly taken posterior samples from the total of 2000. 15 replicated datasets were compared to the observed data $\mathbf{y}$ purely by plotting the histogram of the observed percent covers. Further, the replicated datasets were compared to the observed one through test quantities $T(\mathbf{y})$. Four test quantities used were proportion of zeros, maximum value, sample mean and mean of non-negative observations.

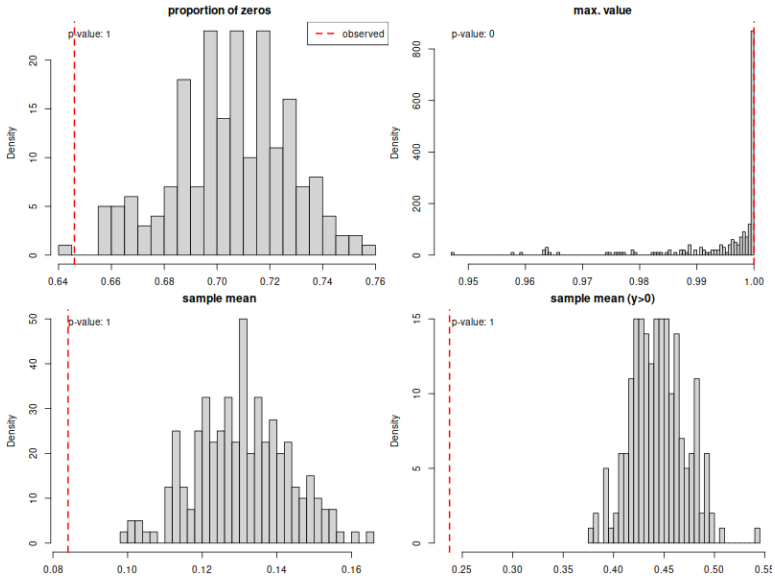
As an example, let's look at the zero-inflated model with common ρ parameter and without spatial random effects. Figure 5.1 shows the posterior predictive checks for the model. In Figure 5.1a histogram of the observed coverages is plotted next to 15 replicated datasets. It is very visible that the replicated datasets do not match the observed one very well: there are too many zeros predicted, and the small coverages in interval $(0, 0.2)$ are underrepresented.

In Figure 5.1b histograms of four test quantities are plotted from 200 replicated datasets. Vertical red line shows the observed test quantity. Now problems in the model fit become clear. Model overpredicts zeros and has too large sample mean, still not always capturing the maximum value of one. Replicated histograms seem to have too heavy tails but the small observations tend to be predicted as exact zeros.

To get an idea why this may occur, a simple approach was taken. By

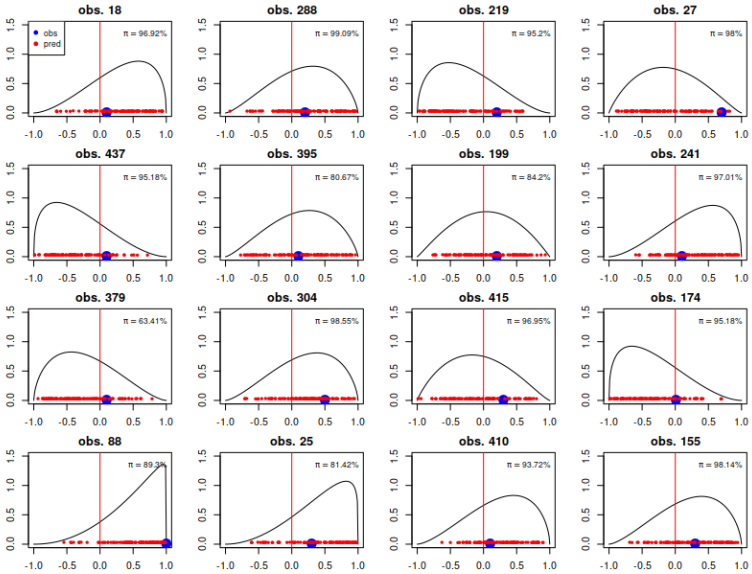


(a) Histogram of observed percent covers along with 15 replicated datasets.

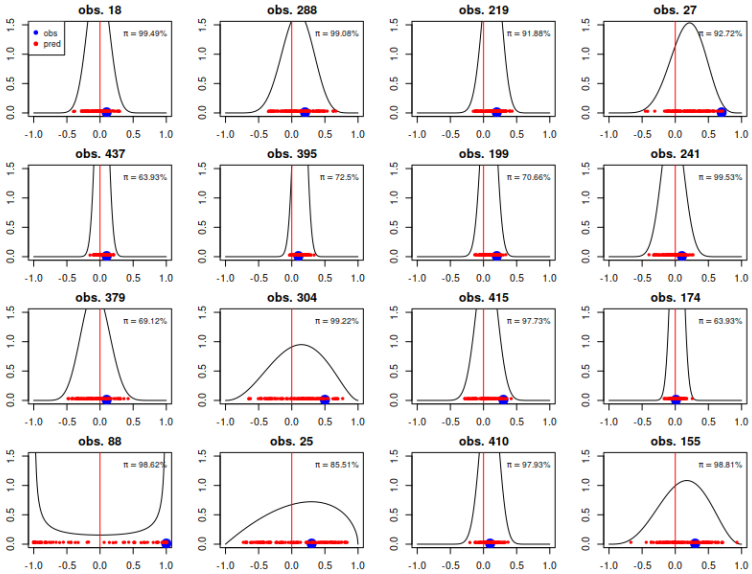


(b) Histograms of four different test quantities from 200 replicated datasets. Vertical red line shows the observed test quantity.

Figure 5.1: Posterior predictive checks for zero-inflated model with common ρ parameter.



(a) Latent beta distributions W_i for zero-inflated model with common ρ.



(b) Latent beta distributions W_i for zero-inflated model with ρ modeled using covariates.

Figure 5.2: Latent beta distributions W_i for 16 observations. Upper right corner shows the probability of suitability for that cite. Curves and probability of suitability were calculated by fixing model parameter values at their posterior means. Modeling the precision parameter ρ has significant effects on the shapes of predictive distributions.

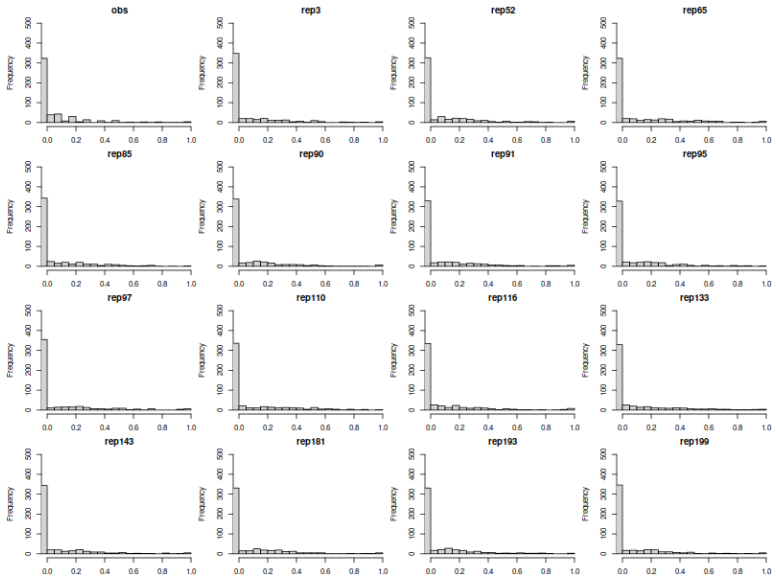
fixing the model parameters at their posterior means (i.e. $\hat{\boldsymbol{\beta}} = \mathbb{E}[\boldsymbol{\beta}|\mathbf{D}]$), one can examine the underlying beta distributions in the data generating process $\text{Beta}(\text{logit}^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\mu), \hat{\rho})$ for any given $\mathbf{x}_i$. Or, as was done, the scaled version that has its support on $(-a, 1)$. These distributions were visually examined for 16 randomly taken nonnegative observations, and the results are visible in Figure 5.2a. The upper right corner shows the probability of suitability for that site, calculated in a similar fashion $\hat{\pi}(x_i) = \text{logit}^{-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\pi)$.

The problem seems to be quite clear. The distributions are too wide, and especially the problem is that from the assumption of common scale parameter ρ, the distributions are equally wide, no matter what the mean is. For example, it would be more reasonable to expect less variation with smaller mean (observations with small expected percent cover) than mean around 0.5. For this reason the problem of overpredicting zeros and missing the small positive values becomes obvious: the wideness of the distributions gives a low probability for producing small positive values.

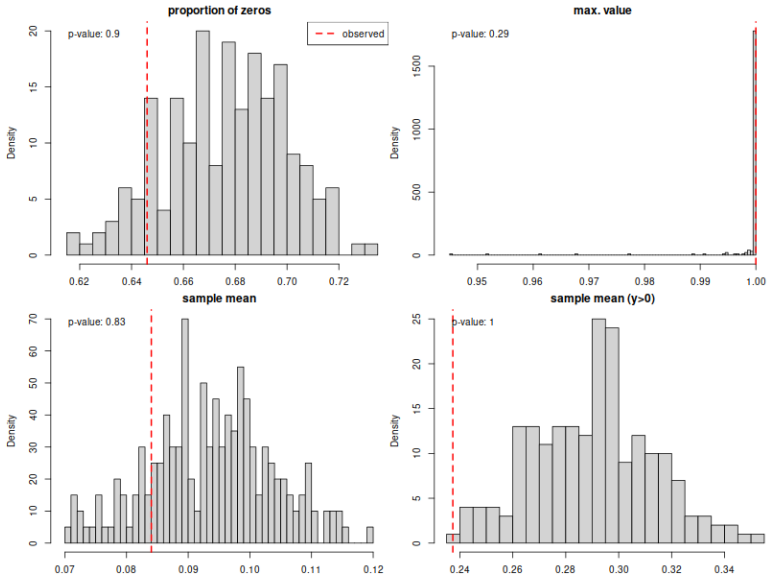
Since the problem seems to originate from the assumption of common precision ρ, an alternative model as described in Section 4.6 was introduced. This version models the precision ρ with covariates, using scaled sigmoid function such that ρ stays within the range of $(0, 1000)$. In Figure 5.2b are the corresponding latent beta distributions than previously in this section described for common ρ. The result is exactly as wished, and now the distribution changes it's width with mean. Now the distributions with mean close to 0 are having smaller variance, making it more probable to produce small non-negative observations.

Figure 5.3 shows the results from posterior predictive checks for the corrected model with varying ρ. From replicated datasets (Fig 5.3a) it is visible that the model starts to catch the small positive values without leading to extra zeros. Looking at the histograms of test quantities (Fig 5.3b), it indeed seems like now the model is able to capture the main properties of the observed datasets, only showing some difficulties with sample mean for positive coverages.

An interesting note to take from Figure 5.2 is the bottom left corner, which corresponds to large (≈ 1) observed percent cover. The model with common ρ correctly puts a large mass on the right end of the distribution. However, model with varying ρ actually ends up in a U-shaped distribution resulting from small value $\rho_i < 1$. This gives a very large mass on values near one and is probably the reason why this version of the model well finds the maximum value of the dataset in the replicated datasets. The U-shape also has an intuitive interpretation. In the environment with highly suitable conditions for the species, one either expects to see very high percent covers, or to see the species absent since it has not found the place yet (stochastic zero). Another possible reason for the model to prefer the U-shape for observations near one is the fact that for beta distribution, even with large mean (and not U-shape), the density approaches zero when we approach



(a) Histogram of observed percent covers along with 15 replicated datasets.



(b) Histograms of four different test quantities from 200 replicated datasets. Vertical red line shows the observed test quantity.

Figure 5.3: Posterior predictive checks for zero-inflated model with ρ modeled using covariates.

Table 5.1: LOO-CV log-predictive densities for all 8 fitted models. First column (base) corresponds to left-censored beta regression introduced in Section 4.2. Extensions with zero-inflation (ZI) and spatial random effects (RE) were introduced in Sections 4.4 and 4.3 respectively. The most complicated model (ZI+RE) includes both zero-inflation and spatial random effects and was constructed in Section 4.5.

	base	ZI	RE	ZI+RE
ρ common	-192.26	-198.86	-162.95	-155.07
ρ modeled	-98.36	-102.63	**-74.77**	-81.39

one. This is for the reason that, similar to the case with exact zeros, beta distribution does not have mass for exact ones.

The last notice can be found from the top right corner of Figure 5.2. With moderately large observations, the distributions do not seem to match the observations that well. It can be reasoned to be resulting from the fact that the beta distribution still has to account for stochastic zeros by having mass below zero. Having a peak at moderately large value while at the same time leaving mass below zero seems like a difficult task.

Approaching with hurdle model would remove the need of forcing mass on the negative side to produce stochastic zeros. With hurdle model, beta distribution would be responsible for only observations larger than zero. This would possibly give better ability to catch the moderately large percent covers as in the scenario at hand. However, using hurdle model would remove the intuitive approach for two types of zeros. Also, Tang et al. [39] found the hurdle model performing worse than the zero-inflated, or the basic left-censored beta regression, suggesting the need of modeling two sources of zeros.

Posterior predictive checks for all the models are visible in Appendix A. They show very similar behavior to the zero-inflated model went through above, indicating that modeling ρ is necessary for capturing important properties of the data.

5.2 Model comparison

The models were compared using their leave-one-out log-predictive densities (Section 2.3). The results can be read from Table 5.1 and there are three main observations.

First observation is that including covariates for modeling the precision parameter ρ noticeably increases the predictive ability of every model, much more than introducing random effects or zero inflation component into the model. This result, as well as the better ability of the model to succeed in posterior predictive checks, is very interesting in light of the following note:

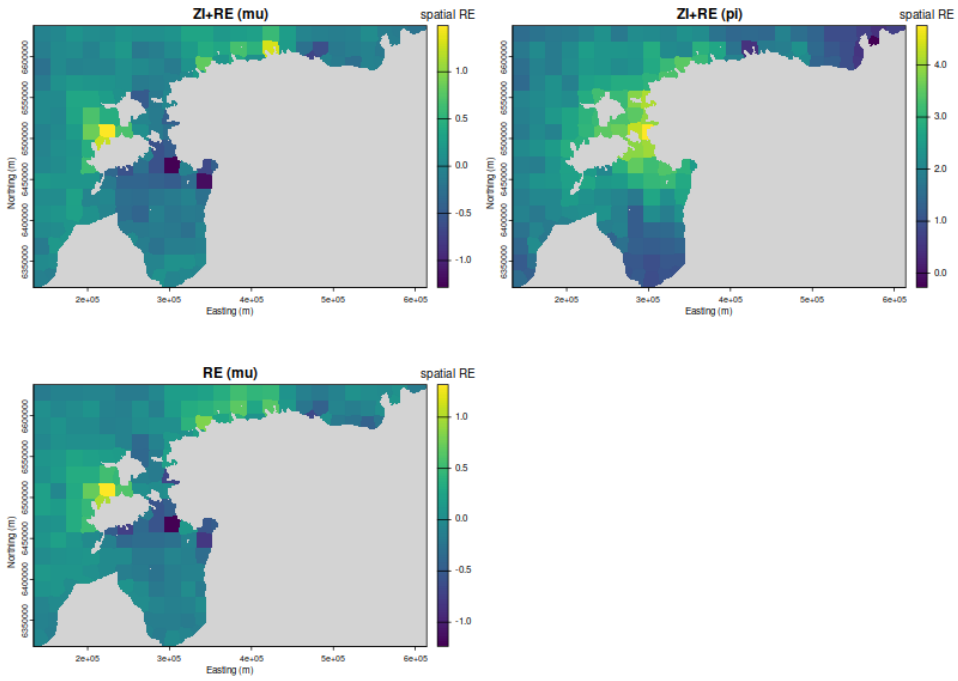


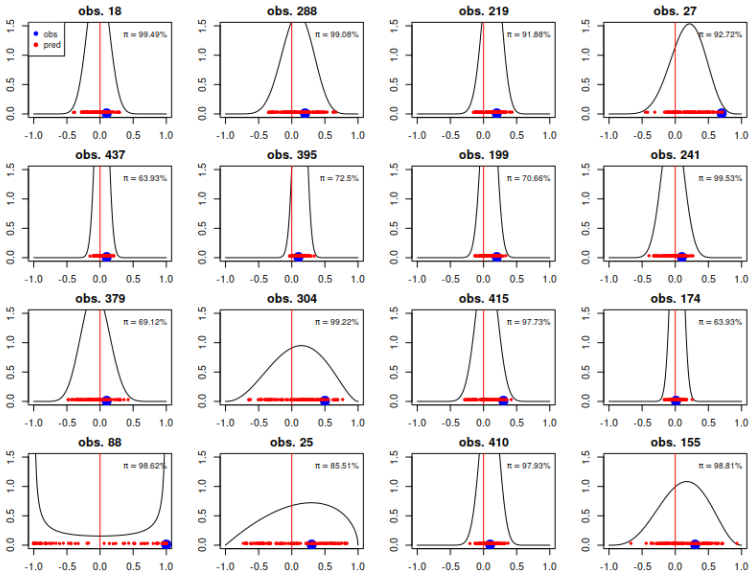
Figure 5.4: Spatial random effects for spatial (lower row) and spatial zero-inflated (upper row) models with common ρ parameter.

even though modeling ρ with covariates was often mentioned as an option (e.g. in [9, 39]), not a single study was found where in the modeling task it was actually done. The results so far indicate that modeling ρ indeed significantly improves the model.

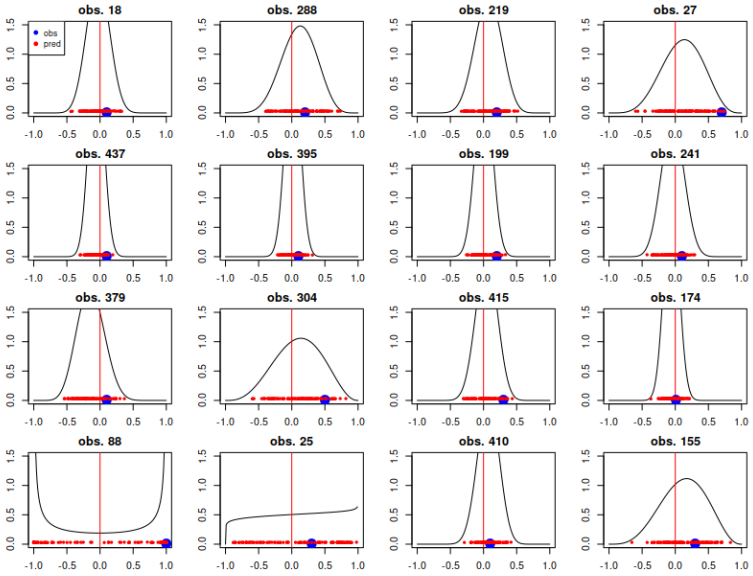
Second observation is that including spatial random effects improves the predictive accuracy in all of the cases. This was not an unexpected result, since including spatial random effects can typically be expected to improve the models for spatial data.

In Figure 5.4 the spatial random effects are shown for spatial and spatial zero-inflated model with common ρ parameter. From those plots it can be read, where the spatial random effects are affecting the model prediction. For modeling the mean μ of the beta distribution (left panel of the figure), there are at least two clear spots (in yellow) where the random effects considerably increase the expected percent cover and two clear spots (dark blue) where the random effects lower the expected percent cover. For modeling the probability of suitability π (right panel), there is one spot with highly increased probability (in yellow). Spatial process for π shows less variation due to the larger estimated length-scale parameter $l_\pi > l_\mu$.

Finally, modeling the two sources of zeros with zero-inflated models



(a) Latent beta distributions W_i for zero-inflated model with ρ modeled using covariates. Value in the top-right corner tells the probability of suitability for the location.



(b) Latent beta distributions W_i for base model with ρ modeled using covariates.

Figure 5.5: Latent beta distributions W_i for 16 observations using posterior mean values for the model parameters. Including zero-inflation seems not to make much of a difference in the distributions.

seemed not to perform any better than using one source of zeros only. Actually, the only increase was when adding zero inflation component to spatial model in the case of common ρ parameter. However, this result is in line with the model comparison results of Tang et al. [39]. In their results, the basic left-censored beta regression performed better than it's zero-inflated version for both two species they examined. Their best model included spatial random effects in the mean μ, and is again in line with the results of this study.

Let's try to demonstrate the similarity between the models with and without component for zero inflation. In Figure 5.5 models with and without component for zero inflation are compared. It is visible that the differences are very small, shapes of the distributions being quite the same. Top right corner of Figure 5.5a shows the probabilities of suitability from zero-inflated models. Overall, those probabilities seem quite large, even for small observed percent covers, meaning that most zeros are expected to be stochastic zeros. This seems to happen even though an informative prior was used to prefer the unsuitability zeros over stochastic zeros. This raises question, whether still more heavily informative priors could be set, and further, how can one justify the selected prior?

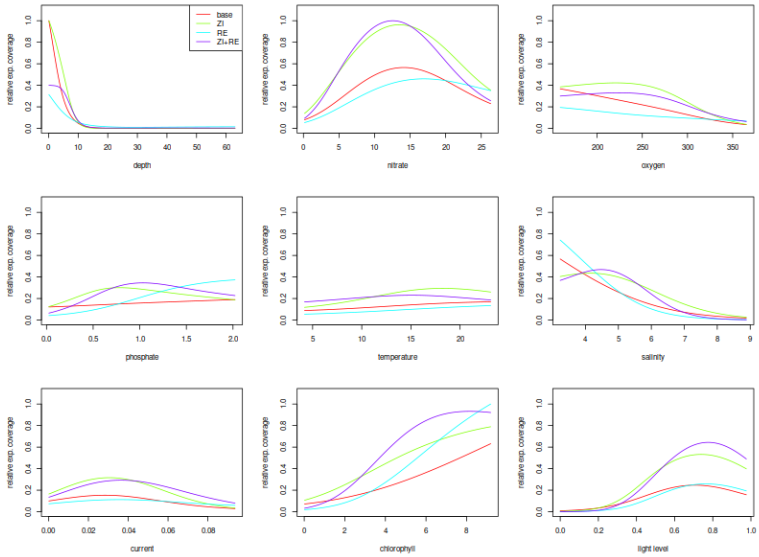
5.3 Ecological inference

The differences in ecological inference between different models are examined by producing the response curves for each nine variables in the model. This is done by varying each variable x at a time from $\min(x)$ to $\max(x)$ while keeping the rest of the variables at their mean $\bar{x}$. For fixed parameters $\boldsymbol{\beta}$ and ρ, expected value of $y \sim \text{LC-Beta}(\mu(\mathbf{x}_i, \boldsymbol{\beta}), \rho)$ can be calculated using equation (4.6). Further, using posterior sample to produce multiple curves, posterior mean is taken to end up with single response curve indicating the posterior mean of expected percent covers.

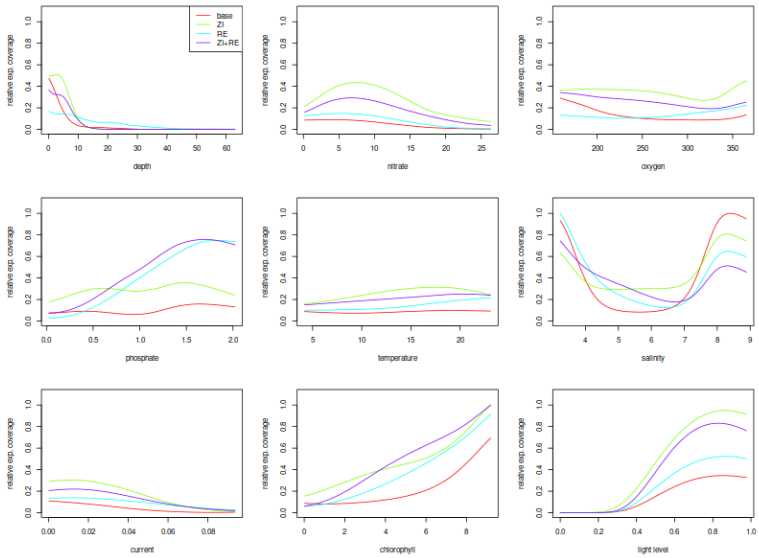
Further, dividing by the maximum posterior mean rescales the curves between 0 and 1, where value 1 indicates where the expected percent cover is at its maximum. These relative expected percent covers can be useful when examining multiple species having very different average percent covers.

In Figure 5.6 the relative expected coverages are shown for all 8 models. Figure 5.6a shows the models with common precision parameter ρ. Qualitatively the response curves are quite similar for most of the variables.

Curves for zero-inflated models (purple and green) seem to go hand-in-hand as well as curves for models without zero inflation component (red and blue). For two covariates, phosphate concentration and salinity, zero-inflated models have response curves with clear peak, whereas the curves of the other two models are monotonically increasing and decreasing, respectively. Also, nitrate concentration and light level at the bottom seem to be



(a) Relative expected coverages for models with common ρ parameter.



(b) Relative expected coverages for models with ρ modeled using covariates.

Figure 5.6: Relative expected coverages for all eight models considered. Value of 1 means that posterior mean of expected coverage reaches it's maximum.

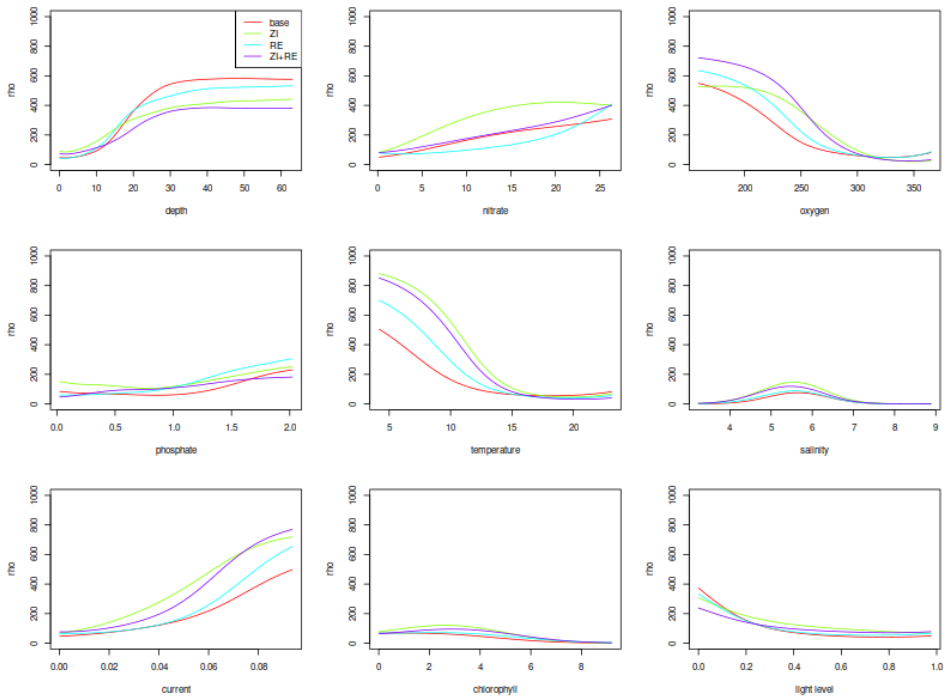


Figure 5.7: Changes of ρ along environmental covariates.

more important for zero-inflated models. However, this is due to responses for water depth, which are not as strong for zero-inflated models near zero. Strong positive effect with small depths makes water depth the most important variable for models without zero inflation. Overall, the response curves are qualitatively similar except for minor differences (especially depth, phosphate concentration and salinity).

Figure 5.6b shows the same response curves for models with varying ρ. The results are quite similar compared to those of common ρ parameter except for one clear difference. Response curves for salinity take the U-shape, having now also a peak at high salinity values (around 8.2). The reason of this behavior can be explained looking at how the ρ varies along the covariates, shown in Figure 5.7. Looking at the curve for salinity, it has a peak around 5.5 but gets really small values at the both ends of the range. These small values of ρ lead to U-shaped beta distribution and thus have an effect for the expected percent cover. Now instead of gathering the mass near zero, the beta distribution also has additional mass near one. As a consequence from this, instead of reaching zero the expectation starts to increase.

This, however, creates somewhat weird behavior in the ecological interpretation, even though it can be explained by the behavior of the under-

lying beta distribution. Large values of salinity first interpreted as being very unpreferred, suddenly seem to be beneficial for the species, even the most beneficial for the base model reaching relative expected coverage of 1 with very high salinity, as well as with very low salinity. This contradicts the ecological niche theory for unimodal responses, which in the first place was the reasoning for restricting second-order terms to produce bell-shaped response curves.

Thus, the result raises question of whether examining the response curves of this sort make sense, or would it be beneficial to only look for example how the covariates effect the mean μ_i of the beta distribution, giving less care for the values of ρ_i. However, since the expected cover is what we are interested in, and is affected by ρ, I find this approach to be the most appropriate one.

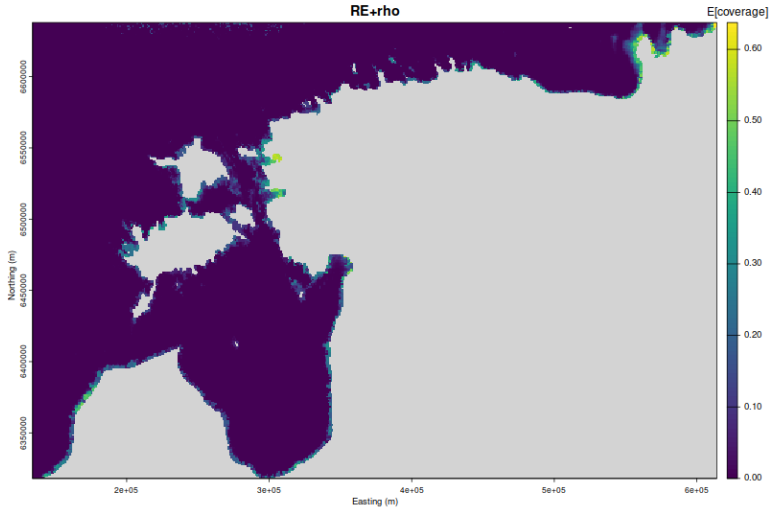
5.4 Hotspot identification

Fitted models can be used to produce maps of posterior means of expected percent covers at unsampled locations $\tilde{\mathbf{s}}_i$ with environmental covariates $\tilde{\mathbf{x}}_i$. An example map produced by the best performing model (spatial random effects with varying ρ) can be seen in Figure 5.8a.

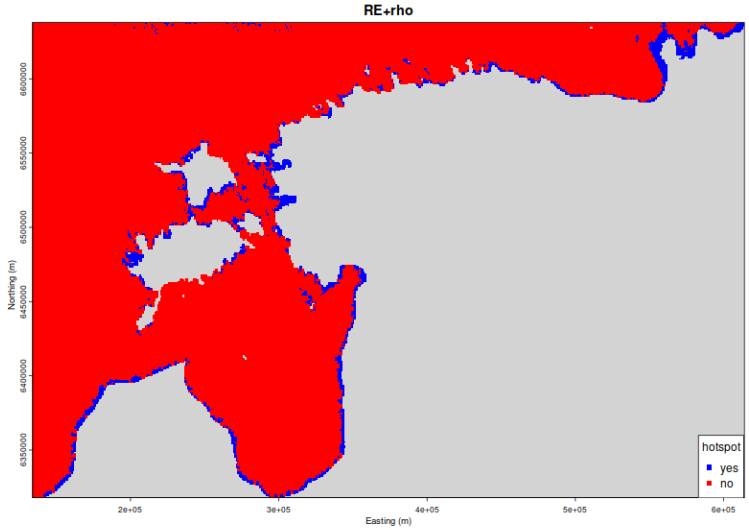
Further, these maps can be turned into binary maps indicating the hotspot areas for the species. This is done by ordering the locations with respect to their expected coverages, starting from the highest value. A location is added to hotspot area until 80% of the total expected coverage over the study area is reached. This approach is very similar to the one used by Kallasvuo et al. [24] to identify the most important fish reproduction areas in the northern Baltic Sea. In Figure 5.8b is the hotspot map produced by the best performing model.

These hotspot areas can be compared between models. In Figure 5.9 are the expected coverage and hotspot maps for the base model. It can be already seen that they differ somewhat from the ones produced by the best model: it covers smaller area in the far east of the study area and it emphasizes the eastern coasts of the two small islands. This can be at least partly explained by the patterns in spatial random effects seen in Figure 5.4. Indeed, the spatial random effects had a positive effect on the western coasts and negative effect on the eastern coasts of these small islands. Figure 5.10 shows a map of agreement between these two hotspot maps, showing on red the areas of disagreement.

Examining these maps for every combination of two models becomes laborious, and thus a measure for similarity between two hotspot maps was considered. A natural approach for measuring the overlap of two binary maps was taken by calculating Jaccard index [6]. Denoting A and B the areas identified as hotspots for two different models, Jaccard index is defined

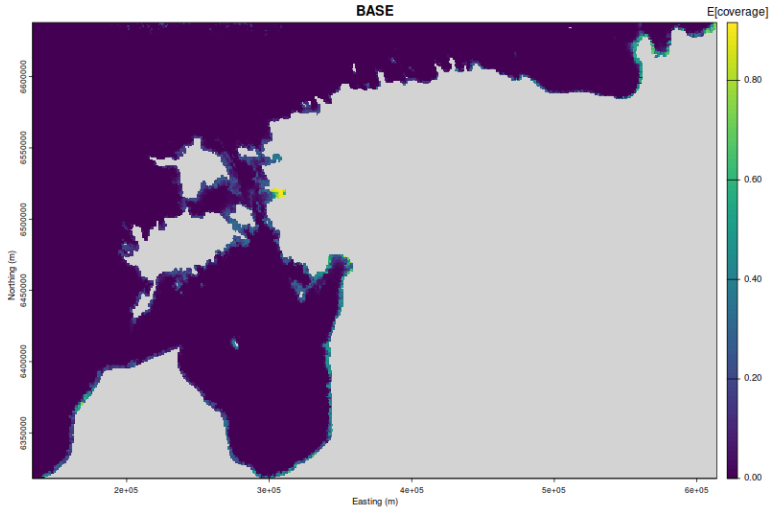


(a) Posterior mean of expected coverages over the study area.

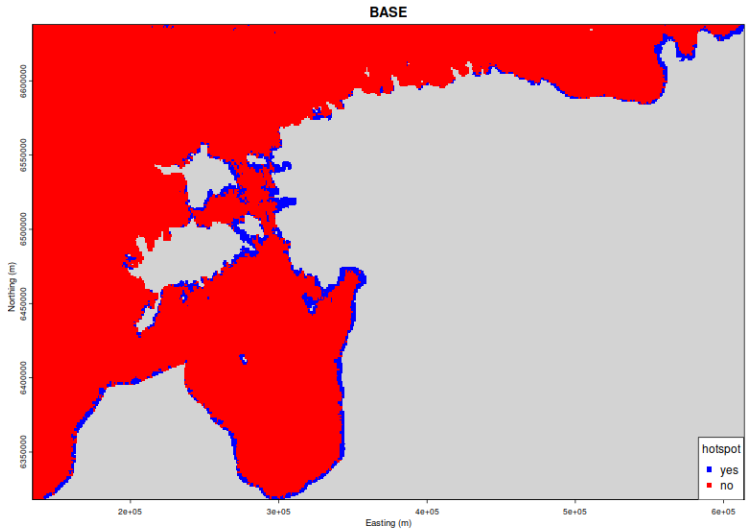


(b) Hotspot areas are identified by accumulating expected coverage until 80% of the total expected coverage is reached.

Figure 5.8: Maps of expected coverages can be turned into binary maps identifying hotspot areas. Results demonstrated using the best model with spatial random effects and ρ modeled by covariates.



(a) Posterior mean of expected coverages over the study area.



(b) Hotspot areas are identified by accumulating expected coverage until 80% of the total expected coverage is reached.

Figure 5.9: Corresponding maps to Figure 5.8 demonstrated using the base model with common ρ.

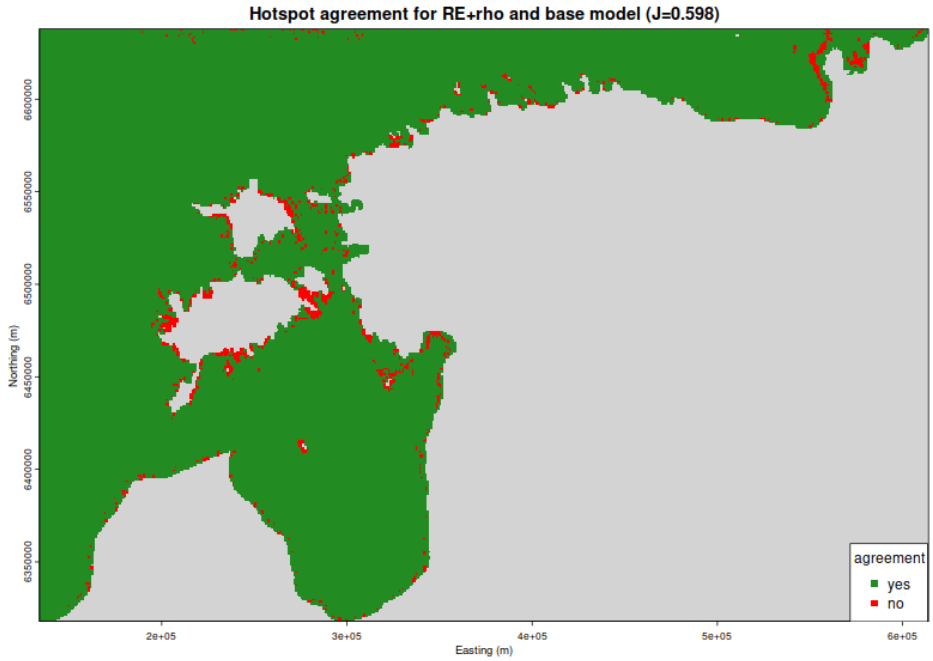


Figure 5.10: Agreement of hotspots for the best performing model (in Figure 5.8b) and the base model (in Figure 5.9b). Overlapping of two binary maps can be quantified with Jaccard Index (J), indicating the proportion of shared hotspot area from total area identified as hotspot.

as

$$J = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \ , \tag{5.1}$$

where $A \cap B$ is the intersection of these areas and $|A|$ denotes the size of an area. Index gives values between 0 and 1, intuitively giving the proportion of shared area of the total area. Value 0 means that the models produced hotspot areas that did not overlap at all, value 1 means that two models produced exactly the same hotspot areas. In the Figure 5.10, Jaccard index 0.598 tells that the shared hotspot area was 59.8% of the total area identified as hotspots by these two models.

These values can be calculated for any combination of two models, and the results are shown in Table 5.2. The results are somewhat similar to the ones obtained in the model comparison part.

Firstly, letting ρ vary in the model seems to have the largest effect on the hotspot area identification. Comparing any model with common ρ and varying ρ (but otherwise similar model components) gives Jaccard index between 0.65 and 0.69. So, in addition to significantly increasing the predictive accuracy of the model, including covariates for modeling ρ also heavily

Table 5.2: Jaccard indexes between eight models considered. It measures the similarity between hotspot areas produced by two models.

	base	base+rho	ZI	ZI+rho	RE	RE+rho	ZI+RE	ZI+RE+rho
base	1.00	0.65	0.88	0.67	0.75	0.60	0.72	0.63
base+rho	-	1.00	0.63	0.87	0.64	0.76	0.61	0.74
ZI	-	-	1.00	0.68	0.68	0.57	0.74	0.62
ZI+rho	-	-	-	1.00	0.62	0.69	0.65	0.79
RE	-	-	-	-	1.00	0.69	0.82	0.69
RE+rho	-	-	-	-	-	1.00	0.65	0.83
ZI+RE	-	-	-	-	-	-	1.00	0.70
ZI+RE+rho	-	-	-	-	-	-	-	1.00

affects the hotspot identification.

Secondly, including spatial random effects into the model has an influence on the identified hotspot areas, but not as strong as including the varying ρ. Comparing any otherwise identical models, including spatial random effects results to Jaccard index between 0.74 and 0.79.

Thirdly, including zero inflation has the least effect on hotspot identification. Taking any model with zero-inflated component and comparing to the one without, Jaccard indexes change between 0.82 and 0.88. The similaries are larger than comparing the models with and without the varying ρ.

Overall, the values changing between 0.57 and 0.88 indicate that model specification has a clear effect on the hotspot identification, possibly affecting the following conservation area identification. In a similar fashion, Domisch et al. [8] found the spatially explicit SDMs superior to non-spatial version in predictive accuracy, drastically affecting prediction-based conservation plans. These observations emphasize the role of proper model assessment as well as model comparison before proceeding to conservation acts.

Chapter 6

Conclusion

In this thesis, zero-inflated percent cover data from the Baltic Sea were used to model the spatial distribution of an algae species. Models for zero-inflated, continuous and positive data are not as well addressed in the literature than e.g. binary or count data. Various versions of left-censored beta regression introduced recently by Tang et al. [39] were implemented to examine their behavior. Specifically, the basic model was extended by introducing spatial random effects, zero inflation mechanism or their combination. To compare the alternative models, three typical properties regarding the use of species distribution models were considered: predictive ability, ecological inference, and hotspot identification. Instead of interpreting the results from ecological perspective (e.g. where the example species prefers to live), the focus of the thesis was in the behavior of the statistical model and it's alternatives.

Posterior predictive checks showed structural problems in the model fits for all the alternative models. The underlying reason was found to be the precision parameter ρ that was assumed to be constant across sampling locations. This produced unnecessarily wide distributions especially for small positive observations, making the model unable to catch these values well enough.

Modeling the precision parameter with environmental covariates fixed the problem. In addition to fixing problems of the model fit, including precision was the major model component to improve predictive accuracy. Also, the largest difference between identified hotspot areas were found to be between identical models with and without common ρ. These results were interesting, since in practical studies found from the literature, the common precision parameter was always assumed.

Even though modeling the precision improved the model performance, it showed ecologically awkward results in the response curves. Especially an U-shaped response curve observed for salinity contradicted the unimodal, bell-shaped response expected by the ecological niche theory. Reasons for

this behavior could be explained from the model components, but it remains open, whether this result is something we would like to avoid in the future.

Spatial random effects were found to be important in terms of predictive ability as well. This was something to be expected, spatial random effects capturing the spatial autocorrelation in the observations, that could not be captured by the environmental covariates used for modeling.

Finally, introducing mechanism to produce two types of zeros (unsuitability and stochastic zeros) seemed to not make much difference compared to the models with only one source for zeros. Possible reasons for this were speculated, but no clear answers were achieved. However, the result was in line with the study by Tang et al. [39] who introduced the model.

It was shown that the model specification overall had clear effects on the identified hotspot areas. Since this would further impact the possible conservation actions followed by modeling, it serves as a reminder for proper model assessment and comparison before proceeding to acts that might be influential.

Chapter 7

Future Directions

This thesis eventually examined the behavior of a statistical model for one species only. Logical way to continue from here would be to perform similar analysis for species with different prevalences to see how the model works more generally. Is some model superior to others consistently? Further, this sort of analysis could give ideas of how the alternative models perform with respect to prevalence: is for example including zero-inflation more important for species with very low prevalence? Since the sample size ($n = 500$) was also kept quite small for this study, yet another way to proceed from here would be to examine the model behavior when increasing the sample size.

Considering the modeling of multiple species, there is another interesting way to proceed from this study. One might model multiple species by first modeling each species separately and then stacking the results to end up with joint predictions. However, this does not take into account the interactions between species. Species interactions can be included in joint species distribution models (JSDM), where multiple species are modeled simultaneously. Plant community modeling is moving from non-spatial SDMs to spatial JSDMs and the growing literature on the topic is well reviewed by Gelfand et al. [14]. Extending the left-censored beta regression to multivariate model (Dirichlet regression) would give tools for modeling this percent cover data jointly for multiple species. The natural question of interest would be to compare the results of stacked single-species models and the extended joint model. Tang et al. [40] have recently introduced a joint model for zero-inflated multivariate data for observations on $[0, \infty)$. In the realm of multivariate percent cover data, Kettunen et al. [25] have introduced a joint model for plant cover data with competition for space and Korhonen et al. [26] have compared joint species distribution models for percent cover data, also accounting for zero inflation.

Also something that could be looked at more carefully are the prior distributions of the model. With the most complex versions of the models introduced in the thesis, there are quite a number of regression coefficients

to estimate. Some regularization was tried to achieve by using quite narrow Gaussian priors, but more could be done by using shrinkage priors such as Laplace (Bayesian Lasso [32]) or group inverse-gamma gamma [3] priors. Similarly, for the parameters of the Gaussian process, complexity penalizing priors [13] could be considered.

Bibliography

[1] Mike P Austin. "Spatial prediction of species distribution: an interface between ecological theory and statistical modelling". In: *Ecological modelling* 157.2-3 (2002), pp. 101–118.

[2] Anabel Blasco-Moreno et al. "What does a zero mean? Understanding false, random and structural zeros in ecology". In: *Methods in Ecology and Evolution* 10.7 (2019), pp. 949–959.

[3] Jonathan Boss et al. "Group inverse-gamma gamma shrinkage for sparse linear models with block-correlated regressors". In: *Bayesian Analysis* 19.3 (2024), pp. 785–814.

[4] Bob Carpenter et al. "Stan: A probabilistic programming language". In: *Journal of statistical software* 76 (2017), pp. 1–32.

[5] Paul B Conn et al. "A guide to Bayesian model checking for ecologists". In: *Ecological Monographs* 88.4 (2018), pp. 526–542.

[6] Luciano da F Costa. "Further generalizations of the Jaccard index". In: *arXiv preprint arXiv:2110.09619* (2021).

[7] Christian F Damgaard and Kathryn M Irvine. "Using the beta distribution to analyse plant cover data". In: *Journal of Ecology* 107.6 (2019), pp. 2747–2759.

[8] Sami Domisch et al. "Spatially explicit species distribution models: A missed opportunity in conservation planning?" In: *Diversity and Distributions* 25.5 (2019), pp. 758–769.

[9] Jacob C Douma and James T Weedon. "Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression". In: *Methods in Ecology and Evolution* 10.9 (2019), pp. 1412–1430.

[10] Simon Duane et al. "Hybrid monte carlo". In: *Physics letters B* 195.2 (1987), pp. 216–222.

[11] Jane Elith and John R Leathwick. "Species distribution models: ecological explanation and prediction across space and time". In: *Annual review of ecology, evolution, and systematics* 40.1 (2009), pp. 677–697.

[12] Silvia Ferrari and Francisco Cribari-Neto. “Beta regression for modelling rates and proportions”. In: *Journal of applied statistics* 31.7 (2004), pp. 799–815.

[13] Geir-Arne Fuglstad et al. “Constructing priors that penalize the complexity of Gaussian random fields”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 445–452.

[14] Alan E Gelfand. “Spatial modeling for the distribution of species in plant communities”. In: *Spatial Statistics* 50 (2022), p. 100582.

[15] Alan E Gelfand et al. “Explaining species distribution patterns through hierarchical modeling”. In: *Bayesian Analysis* 1.1 (2006), pp. 41–92.

[16] Andrew Gelman and Donald B Rubin. “Inference from iterative simulation using multiple sequences”. In: *Statistical science* 7.4 (1992), pp. 457–472.

[17] Andrew Gelman et al. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

[18] Walter R Gilks, Sylvia Richardson, and David Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

[19] Antoine Guisan and Niklaus E Zimmermann. “Predictive habitat distribution models in ecology”. In: *Ecological modelling* 135.2-3 (2000), pp. 147–186.

[20] Antoine Guisan et al. “Predicting species distributions for conservation decisions”. In: *Ecology letters* 16.12 (2013), pp. 1424–1435.

[21] W Keith Hastings. “Monte Carlo sampling methods using Markov chains and their applications”. In: *Biometrika* 57.1 (1970), pp. 97–109.

[22] Mevin B Hooten and N Thompson Hobbs. “A guide to Bayesian model selection for ecologists”. In: *Ecological monographs* 85.1 (2015), pp. 3–28.

[23] Sherwood B Idso and R Gene Gilbert. “On the universality of the Poole and Atkins Secchi disk-light extinction equation”. In: *Journal of Applied Ecology* 11.1 (1974), pp. 399–401.

[24] Meri Kallasvuo, Jarno Vanhatalo, and Lari Veneranta. “Modeling the spatial distribution of larval fish abundance provides essential information for management”. In: *Canadian Journal of Fisheries and Aquatic Sciences* 74.5 (2017), pp. 636–649.

[25] Juho Kettunen et al. “Joint species distribution modeling with competition for space”. In: *Environmetrics* 35.2 (2024), e2830.

[26] Pekka Korhonen et al. “A comparison of joint species distribution models for percent cover data”. In: *Methods in Ecology and Evolution* 15.12 (2024), pp. 2359–2372.

[27] Diane Lambert. “Zero-inflated Poisson regression, with an application to defects in manufacturing”. In: *Technometrics* 34.1 (1992), pp. 1–14.

[28] Andrew M Latimer et al. “Building statistical models to analyze species distributions”. In: *Ecological applications* 16.1 (2006), pp. 33–50.

[29] Darryl I MacKenzie et al. “Estimating site occupancy rates when detection probabilities are less than one”. In: *Ecology* 83.8 (2002), pp. 2248–2255.

[30] Tara G Martin et al. “Zero tolerance ecology: improving ecological inference by modelling the source of zero observations”. In: *Ecology letters* 8.11 (2005), pp. 1235–1246.

[31] Tony O’Hagan. “Dicing with the unknown”. In: *Significance* 1.3 (2004), pp. 132–133.

[32] Trevor Park and George Casella. “The bayesian lasso”. In: *Journal of the american statistical association* 103.482 (2008), pp. 681–686.

[33] Juho Piironen and Aki Vehtari. “Comparison of Bayesian predictive methods for model selection”. In: *Statistics and Computing* 27 (2017), pp. 711–735.

[34] Hl H Poole and WRG Atkins. “Photo-electric measurements of submarine illumination throughout the year”. In: *Journal of the Marine biological Association of the United Kingdom* 16.1 (1929), pp. 297–324.

[35] David R Roberts et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure”. In: *Ecography* 40.8 (2017), pp. 913–929.

[36] Jon Paul Rodríguez et al. “The application of predictive modelling of species distribution to biodiversity conservation”. In: *Diversity and Distributions* (2007), pp. 243–251.

[37] Matti Sahla et al. “Assessing long term change of Fucus spp. communities in the northern Baltic Sea using monitoring data and spatial modeling”. In: *Estuarine, Coastal and Shelf Science* 245 (2020), p. 107023.

[38] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. “Covariate shift adaptation by importance weighted cross validation.” In: *Journal of Machine Learning Research* 8.5 (2007).

[39] Becky Tang et al. “Zero-inflated Beta distribution regression modeling”. In: *Journal of Agricultural, Biological and Environmental Statistics* 28.1 (2023), pp. 117–137.

[40] Becky Tang et al. “Zero-inflated multivariate tobit regression modeling”. In: *Journal of Statistical Planning and Inference* 236 (2025), p. 106229.

[41] Jarno Vanhatalo, Lari Veneranta, and Richard Hudd. “Species distribution modeling with Gaussian processes: A case study with the youngest stages of sea spawning whitefish (Coregonus lavaretus L. sl) larvae”. In: *Ecological Modelling* 228 (2012), pp. 49–58.

[42] Aki Vehtari, Andrew Gelman, and Jonah Gabry. “Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC”. In: *Statistics and computing* 27 (2017), pp. 1413–1432.

[43] Aki Vehtari and Janne Ojanen. “A survey of Bayesian predictive methods for model assessment, selection and comparison”. In: *Statistics Surveys* 6 (2012), pp. 142–228.

[44] Aki Vehtari et al. *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. R package version 2.8.0. 2024. URL: `https://mc-stan.org/loo/`.

[45] Jay M Ver Hoef et al. “Uncertainty and spatial linear models for ecological data”. In: *Spatial uncertainty in ecology: implications for remote sensing and GIS applications*. Springer, 2001, pp. 214–237.

[46] Dani Villero et al. “Integrating species distribution modelling into decision-making to inform conservation actions”. In: *Biodiversity and Conservation* 26 (2017), pp. 251–271.

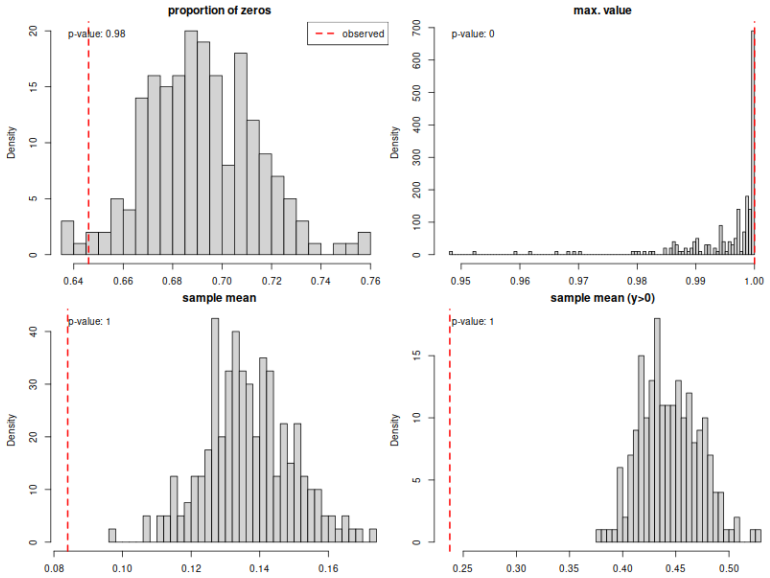
[47] Zhiliang Ying. “Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process”. In: *Journal of Multivariate analysis* 36.2 (1991), pp. 280–296.

[48] Alain F Zuur et al. *Mixed effects models and extensions in ecology with R*. Vol. 574. Springer, 2009.

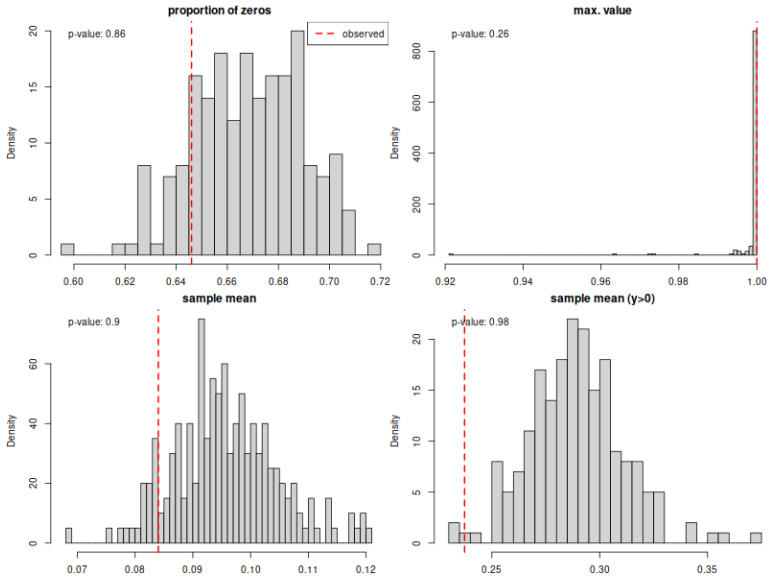
Appendix A

Posterior predictive checks

In this chapter posterior predictive checks from all eight models are visible. Only the histograms of the test quantities were decided to be included, as they more compactly gather the information about the model fit than histograms of replicated datasets.

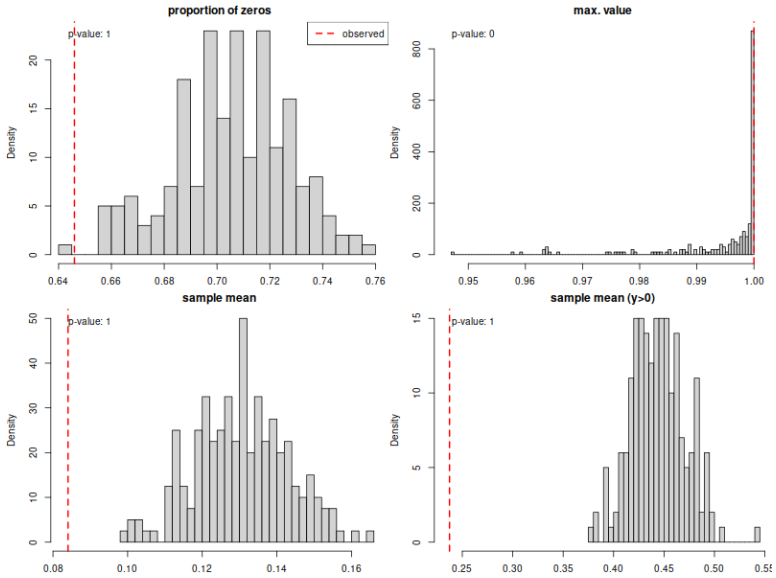


(a) Histograms of test quantities for left-censored beta regression with common ρ parameter.

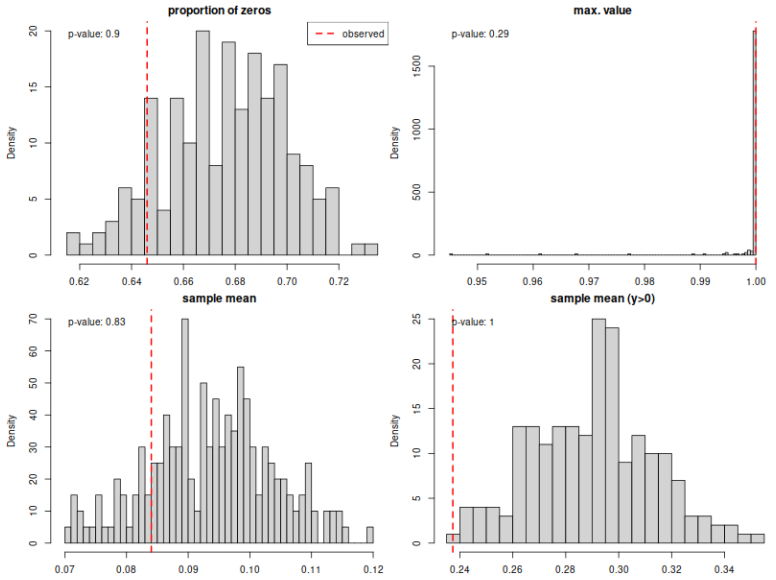


(b) Histograms of test quantities for left-censored beta regression with ρ modeled by covariates.

Figure A.1: Posterior predictive checks for left-censored beta regression (BASE).

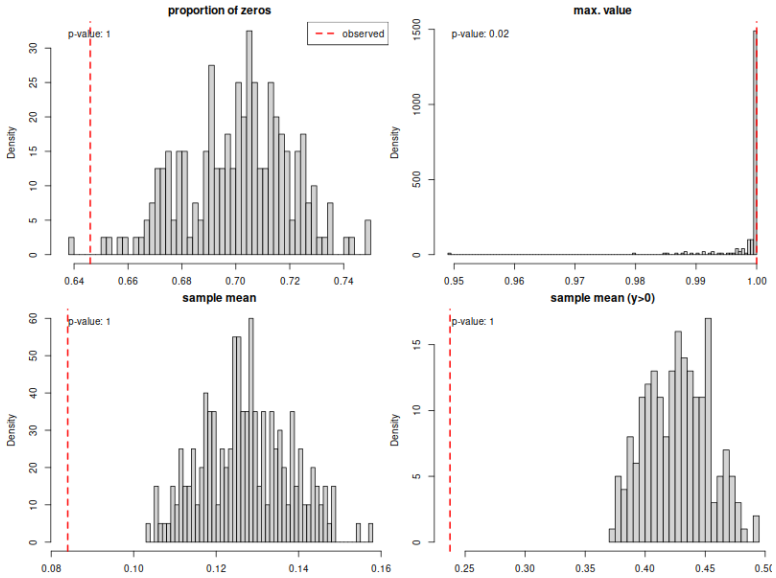


(a) Histograms of test quantities for zero-inflated left-censored beta regression with common ρ parameter.

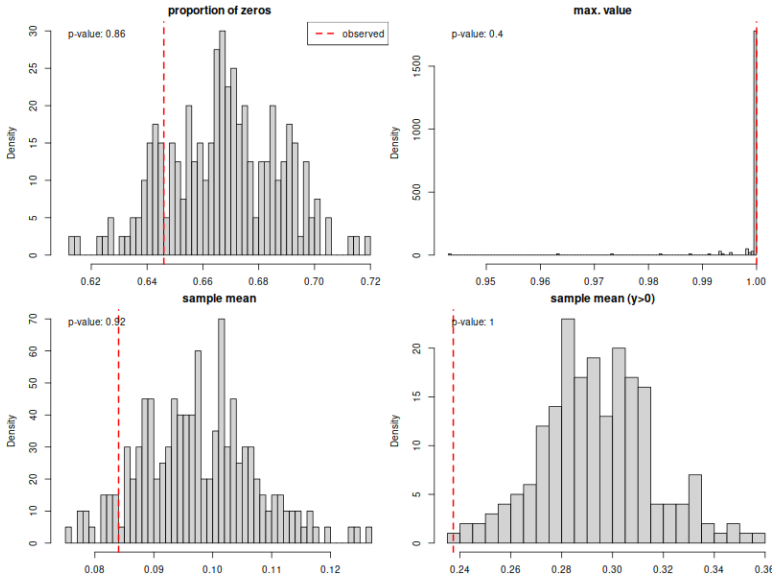


(b) Histograms of test quantities for zero-inflated left-censored beta regression with ρ modeled by covariates.

Figure A.2: Posterior predictive checks for zero-inflated left-censored beta regression (ZI).

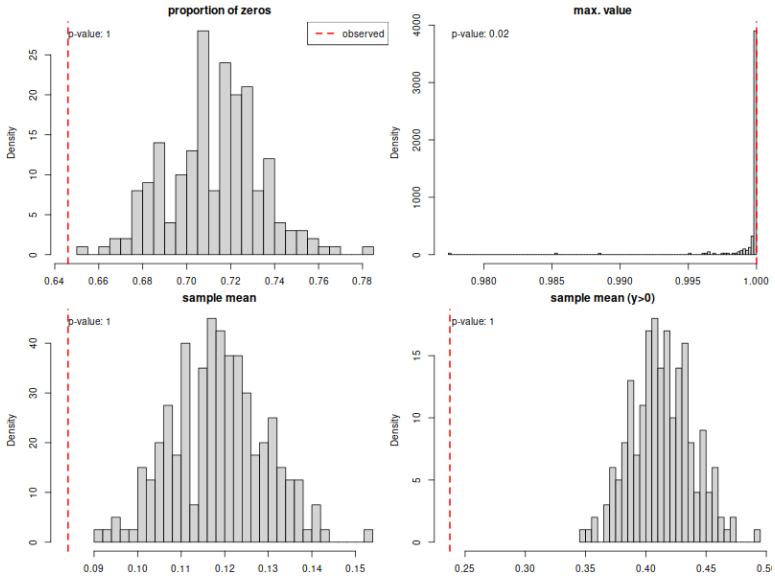


(a) Histograms of test quantities for spatial left-censored beta regression with common ρ parameter.

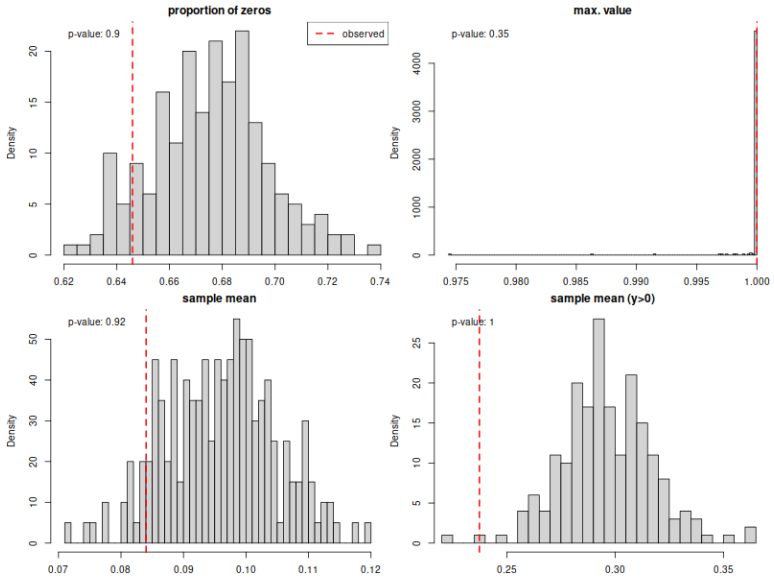


(b) Histograms of test quantities for spatial left-censored beta regression with ρ modeled by covariates.

Figure A.3: Posterior predictive checks for spatial left-censored beta regression (RE).



(a) Histograms of test quantities for spatial and zero-inflated left-censored beta regression with common ρ parameter.



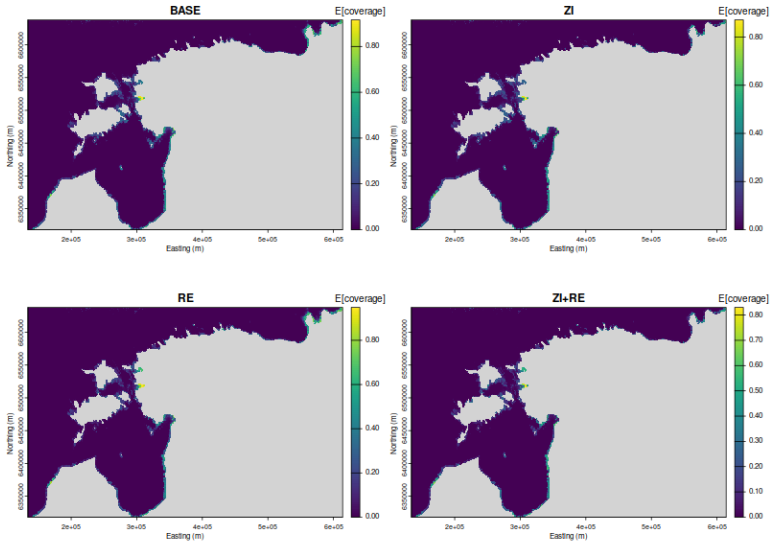
(b) Histograms of test quantities for spatial and zero-inflated left-censored beta regression with ρ modeled by covariates.

Figure A.4: Posterior predictive checks for spatial and zero-inflated left-censored beta regression (ZI+RE).

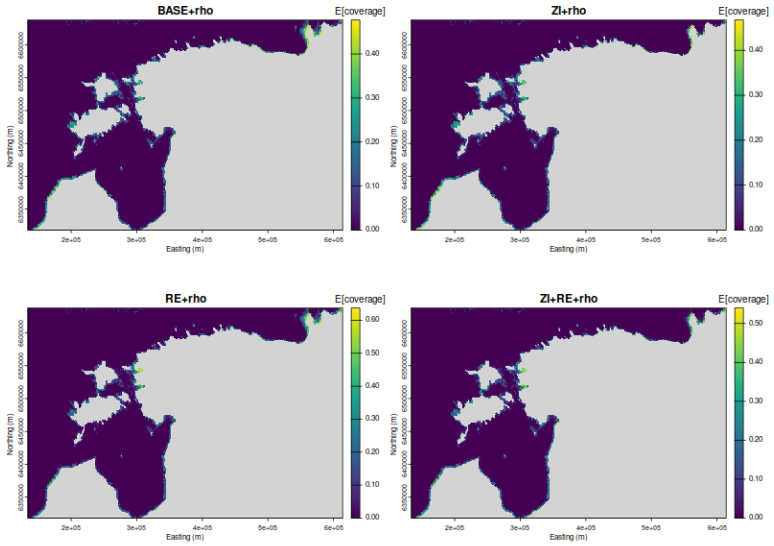
Appendix B

Maps

In Figure B.1 is visible the posterior predictive means of expected coverages for all the eight models considered in the thesis. In Figure B.2 are the corresponding hotspot areas.

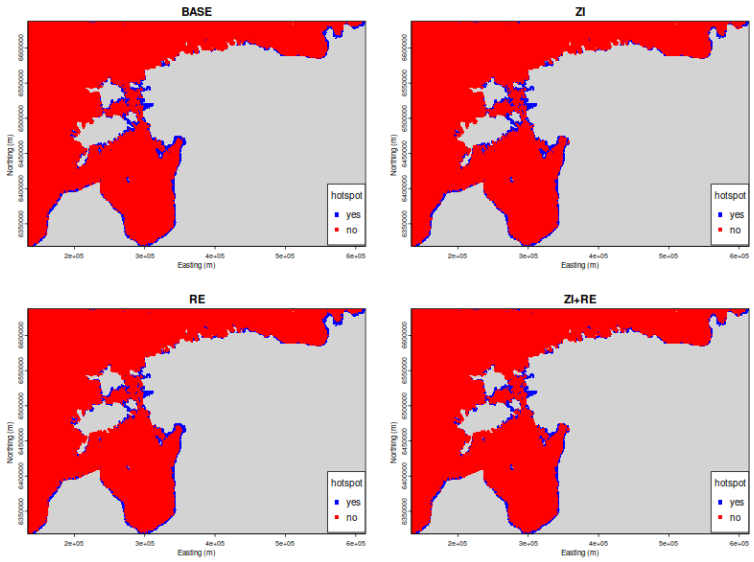


(a) Posterior predictive mean of expected coverages for models with common ρ parameter.

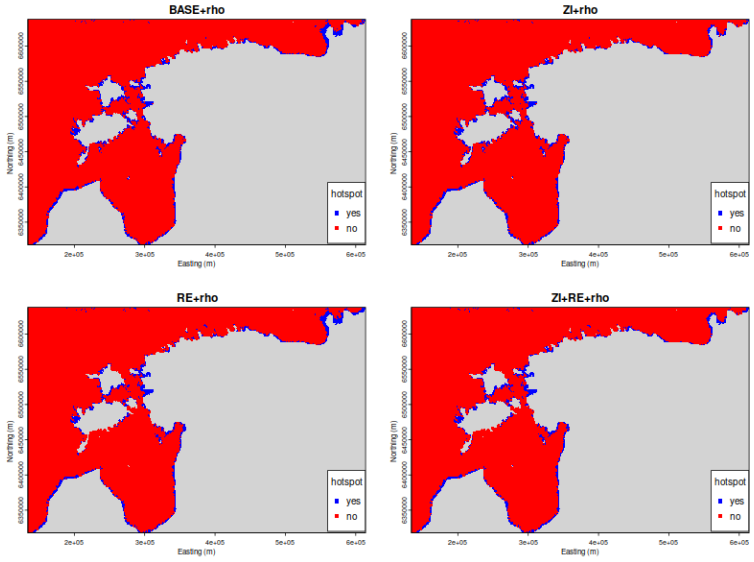


(b) Posterior predictive mean of expected coverages for models with ρ modeled by covariates.

Figure B.1: Posterior predictive mean of expected coverages for all eight models.



(a) Hotspot areas for models with common ρ parameter.



(b) Hotspot areas for models with ρ modeled by covariates.

Figure B.2: Hotspot areas (cumulative 80% of the expected coverage) for all eight models.

Appendix C

List of species

In Table C.1 are listed all the species consideded for modeling task, although only one (*Cladophora glomerata*) was selected as an example species during this thesis.

Table C.1: List of species considered for modeling task.

Species	Functional group
Amphibalanus improvisus	Crustaceans
Chara aspera	Macro algae
Chorda filum	Macro algae
Cladophora glomerata	Macro algae
Fucus vesiculosus	Macro algae
Furcellaria lumbricalis	Macro algae
Myriophyllum spicatum	Vascular plants
Mytilus trossulus	Molluscs
Potamogeton perfoliatus	Vascular plants
Pylaiella/Ectocarpus	Macro algae
Ruppia maritima	Vascular plants
Stuckenia pectinata	Vascular plants
Ulva intestinalis	Macro algae
Vertebrata fucoides	Macro algae
Zannichellia palustris	Vascular plants
Zostera marina	Vascular plants