Methods

Table of contents

# 1. Preamble

## 1.1 Why Bayesian

The models presented below are Bayesian predictive models. We have opted for a Bayesian approach because of ability to deal with the small number of observations for some species and the presence of outliers in these small samples.[[1]](#footnote-20)

Bayesian models are particularly well-suited for handling hierarchical data structures like the one we faced in this analysis.[[2]](#footnote-22) Finally, carefully chosen priors help regularize estimates and thus prevent overfitting, which is essential in prediction modeling.

## 1.2 Goals of analysis

We used a two-stage model to predict the number of individuals breeding on 217 unsurveyed atolls for 37 seabird species.

In the first step, we fit a Bayesian multilevel logistic regression model to presence / absence data from 199 surveyed atolls. Next, posterior samples were used to predict the presence of 37 species on 81 unsurveyed atolls. We combined knowledge from the literature and model predictions to compile a list of atolls for which count predictions should be made. Including a given species *s* from a given unsurveyed atoll *a* in the list of atoll-species pairs for which count predictions were made, required that species’ *s* mean predicted probability of being present on atoll *a* exceeded 80%. In total, we made count predictions for 2997 atoll-species pairs, as described below, of which 114 were predictions of our presence model. Conversely, the model confidently predicted the absence of 2067 atoll-species pairs, i.e., their mean predicted probability was below 20%.

In the second step, we fit a Bayesian multilevel regression model to log-transformed bird counts from surveyed atolls. Lastly, posterior samples were used to predict bird counts for those atolls-species pairs for which either the literature or the predictions of our logistic model indicated that a given species was present on a given atoll.

# 2. Explain the model

Both the presence model and count model use the same independent variables to model presence or counts, respectively, of species of nesting type on atoll , with environmental parameters , in region (for full lists of these variables, see supplements?). Species, nesting type, atoll, and region are categorical variables. The matrix of environmental parameters contains the results of a principal component analysis conducted on a selection of 16 environmental variables (for further detail, see supplements?).

## 2.1 Presence model

Let , where , denote the presence of species of nesting type on atoll with environmental parameters in region .

### Likelihood function

The parameter encodes the probability that species occurs in the region in which atoll is located.

The parameter encodes how environmental parameters of atoll affect the probability that species within a nesting type occurs on atoll . This effect is allowed to differ between nesting types, such that the probability that a ground nesting species occurs on atoll increases with some environmental parameter, while the probability of a tree nesting species decreases. By partially pooling species within nesting types, we express our assumption that species from the same nesting type are similarly affected by certain environmental parameters.

### Priors

The priors of the presence model are weakly informative. Their small variance serves to regularize the estimates and prevent overfitting where few samples are available in our hierarchical data structure.

We assume , the parameter encoding regional variation in species occurence, to have the largest variation. It is often the case that species do not occur in one region while they are common in another.

In contrast, we assume the variation resulting from , i.e., nesting type, environmental factors, and their interaction, to be much smaller than that caused by the regional variation encoded in .

## 2.2 Count model

Let , where , denote the log-transformed count of individuals of species of nesting type on atoll in region .

### Likelihood function

The parameter encodes the abundance of species across regions by nesting regional estimates within species. As such, is the average abundance of species across regions. The hyperparameter on encodes the variation between regions across species. By partially pooling regions within species, we express our assumption that species forms colonies of similar size across regions. This is unlike the presence model, where the fixed effect prevents information flow between species ’s estimates for different regions.

The parameter encodes how environmental parameters affect the abundance of species within a nesting type. This effect is allowed to differ between nesting types, such that the abundance of a species of ground nesters may be positively related to some environmental parameters, while the abundance of tree nesters may be negatively related. By partially pooling species within nesting types, we express our assumption that species from the same nesting type are similarly affected by certain environmental parameters.

### Priors

The priors of the count model are in part informed and in part weakly informed, with the intention of regularizing estimates.

The informed prior on is set to the median log-transformed count of each species. Our choice of a species-specific prior informed by the data is due to the large differences in colony sizes across species, and the small sample sizes for some species. While reasonable prior settings were consistently overwhelmed in species with medium to large sample size, species with small sample sizes () require more careful specification of priors. Consider how a species with small sample size and very small colony size is impacted by setting the prior to the average colony size across species: due to the small sample size, the prior skews our predictions towards the global average, which leads to poor predictions.

Our choice of the median over the mean is motivated by the presence of very few, very large colonies in some species. In summary, the goal of the prior on is to enable our model to make reasonable and conservative predictions even for species with very small samples. For further detail on the influence of different prior settings, see the results of the sensitivity analysis in the supplements.

The priors on is weakly informed in that it places most probability mass on values near 0.

The priors on the variance parameters , , and place more probability mass on higher variances than in the presence model to allow uncertain predictions where there are few data.

## 2.3 Prior predictive check

We have conducted and visually inspected prior predictive checks for both models. The prior predictive checks generated data in plausible value ranges, consistent with our prior knowledge. For further details, see supplements.

# 3. Computational details

## 3.1 Software

We used the Julia Programming Language[[3]](#footnote-35) and most notably the Turing package ecosystem[[4]](#footnote-37) to fit, diagnose, visualize and summarize the Bayesian models presented here. For an exhaustive list of all packages used and the versions, see the Manifest.toml file in the GitHub repository associated with this project.

## 3.2 MCMC chain convergence

We made sure that all chains have converged by examining the statistic for all models. For detailed convergence diagnostics, see supplements.

## 3.3 MCMC chain resolution

We made sure that all chains have high resolution by examining the effective sample size (ESS) in the bulk and tail regions of each parameter distribution. For detailed convergence diagnostics, see supplements.

# 4. Describe the posterior distribution

## 4.1 Posterior predictive check

We have conducted posterior predictive checks to make sure that our model usefully mimics the observed data. For further details on the posterior predictive checks, see supplements.

# 5. Report sensitivity analysis

Sensitivity analyses were carried out for a wide range of different prior settings. In summary, the results show that the results of the presence model are robust across different prior settings. The same can be said for the count model, with the exception of very few species for which the small sample sizes are extremely low. In situations with extremely small sample sizes, sensitivity to the prior is inevitable. Thoughtful prior specification is critical in these situations, as described above.

1. [Gelman et al., 2020](https://www.cambridge.org/highereducation/books/regression-and-other-stories/DD20DD6C9057118581076E54E40C372C#overview) [↑](#footnote-ref-20)
2. [McGlothlin & Viele, 2018](doi:10.1001/jama.2018.17977) [↑](#footnote-ref-22)
3. [Bezanson et al. 2017](https://doi.org/10.1137/141000671) [↑](#footnote-ref-35)
4. [Ge et al. 2018](https://github.com/TuringLang/Turing.jl/blob/master/CITATION.bib) [↑](#footnote-ref-37)