# STAT 431 - Bayesian Hierarchical Modeling of Bird Species Richness

Aadya Ranjan[aadyar2], Alarsh Tiwari[alarsht2], Anshika Pradhan[anshika6], Sarthak Morj[smorj2]

May 13, 2025

**Abstract**

Species richness, defined as the count of distinct bird species observed, is an important ecological indicator. Modeling species richness presents challenges due to its non-linear relationship with observational effort, regional variability, and seasonal effects. To address these complexities, this project employs a Bayesian hierarchical NegativeBinomial model to predict species richness across the United States. The primary goal is to estimate the relationship between species richness and multiple factors, including region, day of the year, and observational effort, while accounting for spatial and temporal variability. The model incorporates random effects for U.S. states and employs a Fourier basis to capture seasonal patterns. Gibbs sampling using JAGS is utilized to estimate the posterior distribution of model parameters. This analysis aims to investigate potential relationships between these factors and bird species richness while ensuring model convergence through diagnostic checks.

## 1. Introduction

Biodiversity assessment is crucial for understanding ecological health, and species richness serves as a key indicator in this context. Species richness, defined as the number of distinct species observed in a given area, helps ecologists monitor environmental changes and conservation efforts. However, accurately estimating species richness poses challenges due to the inherent variability in bird observations, influenced by geographical, temporal, and observational factors.

A common approach to modeling species richness involves treating it as a count variable and using a Poisson regression model. However, the simplistic assumption of a linear relationship between species counts and observational effort can lead to inaccuracies, as the relationship is often more complex. In particular, variations in observer effort, regional differences, and temporal patterns significantly impact the observed species richness.

To address these challenges, a hierarchical Bayesian approach is proposed, allowing for the inclusion of random effects to account for regional variability and temporal patterns. The model specifically incorporates random intercepts for U.S. states to capture spatial heterogeneity and employs a Fourier basis to model seasonal variations. Additionally, effort variables such as observation duration and the number of observers are carefully standardized to maintain consistency. Instead of a Poisson model, a Negative Binomial model is used to account for overdispersion.

In this project, we utilize Gibbs sampling through JAGS to estimate the posterior distribution of model parameters. By modeling the underlying relationship between species richness and these factors, we aim to gain a deeper understanding of the ecological dynamics influencing bird biodiversity across the United States.

## 2. Data Description

This project leverages data from the eBird database, which records bird observations across the United States. eBird is one of the world's largest biodiversity-related science projects, managed by the Cornell Lab of Ornithology. It aggregates bird observation data contributed by birdwatchers globally, with more than 100 million bird sightings annually. The dataset contains detailed records of bird species observations, including species identification, location information, date, and observational effort.

### 2.1. Data Acquisition

The data was collected using the eBird API, which provides recent bird observation records for specific states in the U.S. The Python script used for data collection accessed data from the states of Illinois, Indiana, Missouri, Kentucky, and Iowa for the past 30 days. Each record includes the species observed, location (latitude and longitude), date, and effort metrics.

### 2.2. Data Characteristics

The raw dataset initially consisted of multiple columns, including:

- Species Code, Common Name, Scientific Name
- Location ID, Location Name, Observation Date
- Count of Individuals Observed, Latitude, Longitude
- Validation Status, Review Status, Privacy Settings

- Submission ID, Region, Exotic Category

To focus on analyzing species richness effectively, the dataset was preprocessed to retain only the most relevant variables. The primary outcome variable, **species richness**, is quantified as the number of distinct species observed per site visit.

### 2.3. Predictors

The following predictors were incorporated into the analysis:

- **Region:** Encodes the U.S. state where the observation was recorded, allowing for spatial variation.
- **Day:** Represents the day of the year, derived from the observation date to capture seasonal trends.
- **Effort Variables:**
  - Duration: Standardized observation duration to account for variations in effort.
  - Effort: Standardized latitude as a proxy for regional effort variation.
  - Observers: Standardized longitude, reflecting the number of observers involved.

### 2.4. Data Cleaning and Transformation

Data cleaning involved removing rows with missing values in critical variables (e.g., species count, latitude, longitude, date, region). The effort variables were standardized to maintain consistency across observations. To model temporal effects, the day of the year was calculated from the observation date using a Julian day format.

## 3. Model Diagram and Structure: Directed Acyclic Graph

The presented Directed Acyclic Graph represents the hierarchical Bayesian model developed for estimating detection probability and bird observation counts across multiple regions and days of the year, adjusting for observer effort and environmental covariates. We have used `grViz` to output the graph. Due to the package's limitations, Stochastic nodes (e.g., `r`, `Y`, `b_state_raw`) are connected by solid arrows ($\rightarrow$), while deterministic transformations (e.g., from `alpha_raw` to `alpha`) are shown with dashed arrows ($\dashrightarrow$).

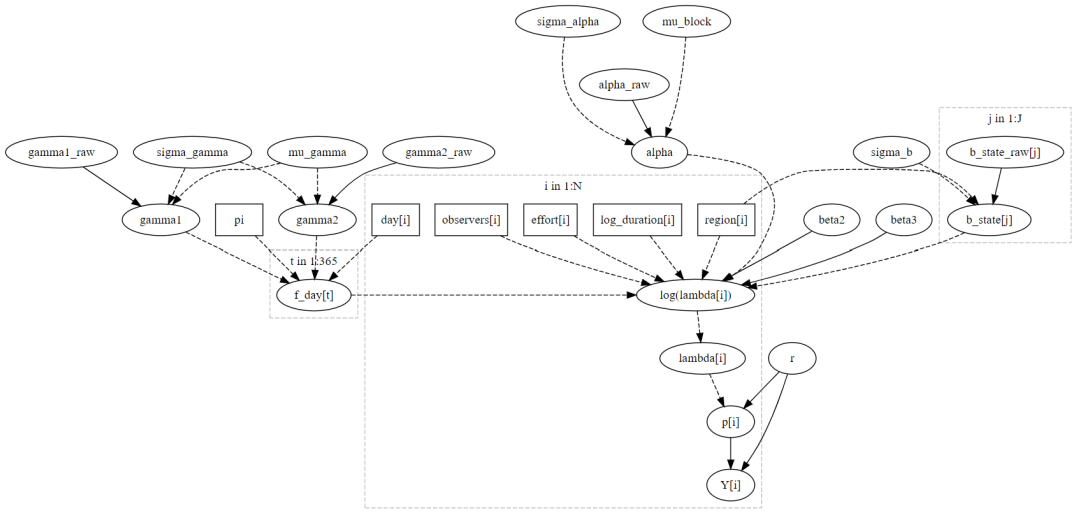


Figure 1: Directed Acyclic Graph (DAG) of the Hierarchical Bayesian Model

**Key Components**

**Observed Data:**

- $Y[i]$: count of birds observed in observation $i$.
- `effort`$[i]$, `observers`$[i]$, `region`$[i]$, `day`$[i]$, `log_duration`$[i]$: covariates associated with each observation.

**Latent Structure:**

- $\lambda[i]$: the expected count rate for observation $i$, modeled on the log scale via log $\_lambda[i]$, which incorporates fixed effects ($\alpha$, $\beta_2$, $\beta_3$), random effects ($b\_state[j]$), and seasonal components ($f\_day[t]$).
- $p[i]$: detection probability derived from $\lambda[i]$ and dispersion $r$, used in the Negative Binomial likelihood for $Y[i]$.

**Random Effects and Priors:**

- $b\_state\_raw[j]$: region-specific raw intercepts modeled as standard normal variables, scaled by $\sigma_b$ to produce $b\_state[j]$.
- $\alpha\_raw$, $\gamma 1\_raw$, $\gamma 2\_raw$: raw intercept and seasonal coefficients, scaled by their respective standard deviations ($\sigma_\alpha$, $\sigma_\gamma$) and centered around priors ($\mu\_block$, $\mu\_gamma$).

- $r$: dispersion parameter for the Negative Binomial distribution, given a Gamma prior.

**Deterministic Transformations:**

- $\alpha$, $\gamma 1$, $\gamma 2$, $b\_state[j]$, $\log\_lambda[i]$, $\lambda[i]$, $p[i]$, and $f\_day[t]$ are all deterministic nodes derived from their respective stochastic inputs and covariates.

**Plates**

- **Observation-level (i):** Models the data across all survey instances.
- **Region-level (j):** Models spatial variability via region-specific intercepts.
- **Day-level (t):** Models temporal variability via seasonal effects across the calendar year.

## 4. Methodology

### 4.1. Bayesian Hierarchical Model with Negative Binomial Distribution

Bird species richness is influenced by a combination of spatial, temporal, and observational factors, including regional differences, seasonal migration patterns, and variations in observational effort. To accurately model these complexities, a Bayesian hierarchical model was chosen, as it effectively incorporates multiple factors and captures the interactions between them.

Bayesian hierarchical modeling is particularly suited for ecological data, as it accounts for variability and uncertainty by providing probability distributions rather than single point estimates. This approach aims to produce robust and accurate predictions, even when dealing with noisy data and inherent variability in bird observations.

Bird species richness data often exhibit overdispersion, where the variance significantly exceeds the mean. The traditional Poisson model assumes equidispersion (variance equal to mean), which is often violated in ecological data. Such violations can result in biased and inefficient estimates. To address **overdispersion**, we employ the Negative Binomial (NB) model, which introduces a dispersion parameter ($r$) to account for additional variability.

### 4.2. Bayesian Hierarchical Model

**Model Specification:**

The response variable $Y_i$ follows a Negative Binomial distribution parameterized by $r$ and $p_i$:

$$Y_i \mid \lambda_i, r \sim \text{NB}(r, p_i)$$

$$p_i = \frac{r}{r + \lambda_i}$$

The mean parameter $\lambda_i$ is modeled as follows:

$$\log(\lambda_i) = \alpha + b_{\text{state}}[\text{region}_i] + \beta_2 \cdot \text{effort}_i + \beta_3 \cdot \text{observers}_i + f(\text{day}_i) + \log(\text{duration}_i)$$

**Spatial Variation (Random Effects):**

To account for variability between different regions, random intercepts are introduced as follows:

$$b_{\text{state}}[j] = \sigma_b \cdot b_{\text{state\_raw}}[j]$$

$$b_{\text{state\_raw}}[j] \sim \mathcal{N}(0, 1)$$

$$\sigma_b \sim \text{Uniform}(0, 10)$$

**Seasonal Variation (Fourier Series):**

The model incorporates seasonal effects through a Fourier series to capture periodicity in bird observations:

$$f(\text{day}_i) = \gamma_1 \sin\left(\frac{2\pi \cdot \text{day}_i}{365}\right) + \gamma_2 \cos\left(\frac{2\pi \cdot \text{day}_i}{365}\right)$$

The Fourier coefficients are modeled as follows:

$$\gamma_1 \sim \mathcal{N}(0, 0.1), \quad \gamma_2 \sim \mathcal{N}(0, 0.1)$$

**Priors for Fixed Effects:**

The fixed effects for the model include the intercept and coefficients related to observational effort:

$$\alpha \sim \mathcal{N}(0, 0.1)$$

$$\beta_2 \sim \mathcal{N}(0, 0.001), \quad \beta_3 \sim \mathcal{N}(0, 0.001)$$

**Dispersion Parameter:**

The dispersion parameter $r$ is modeled as follows:

$$r \sim \text{Gamma}(0.01, 0.01)$$

**Non-Centered Parameterization:**

To improve convergence, non-centered parameterization is applied for the intercept and seasonal effects:

$$\alpha = \mu_{\text{block}} + \sigma_\alpha \cdot \alpha_{\text{raw}}$$

$$\gamma_1 = \mu_\gamma + \sigma_\gamma \cdot \gamma_{1,\text{raw}}$$

$$\gamma_2 = \mu_\gamma + \sigma_\gamma \cdot \gamma_{2,\text{raw}}$$

$$\alpha_{\text{raw}}, \gamma_{1,\text{raw}}, \gamma_{2,\text{raw}} \sim \mathcal{N}(0, 1)$$

**Hyperpriors:**

To introduce flexibility in the model, hyperpriors are used for the intercept and seasonal effects:

$$\mu_{\text{block}} \sim \mathcal{N}(0, 0.1)$$

$$\sigma_\alpha \sim \text{Uniform}(0, 5)$$

$$\mu_\gamma \sim \mathcal{N}(0, 0.1)$$

$$\sigma_\gamma \sim \text{Uniform}(0, 5)$$

**Posterior Inference:**

The posterior distribution is derived from the combination of the likelihood and the prior

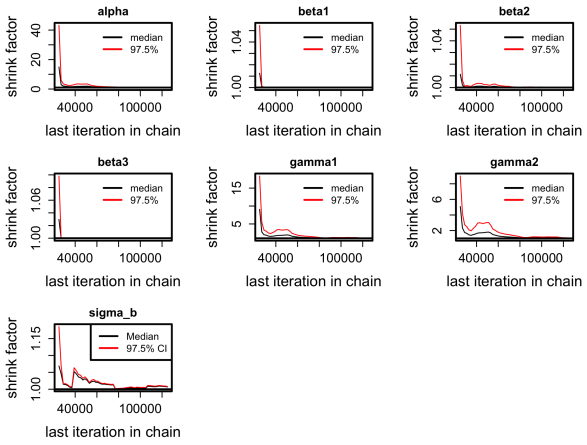


Figure 2: Plots of shrink factors for a subset of parameters

distributions:

$$p(\alpha, \beta_2, \beta_3, \gamma_1, \gamma_2, \sigma_b, r \mid Y) \propto p(Y \mid \lambda, r) \cdot p(\alpha) \cdot p(\beta_2) \cdot p(\beta_3) \cdot p(\gamma_1) \cdot p(\gamma_2) \cdot p(\sigma_b) \cdot p(r)$$

## 5. Implementation

The Bayesian hierarchical model for bird species richness was implemented using JAGS (Just Another Gibbs Sampler) through the R interface. The model was designed to estimate the posterior distributions of the parameters efficiently while ensuring convergence.

### 5.1. MCMC Setup

To optimize computational efficiency, the model was configured to use four parallel Markov Chain Monte Carlo (MCMC) chains. Each chain was run for 100,000 iterations, with an initial adaptation phase of 5,000 iterations to adjust proposal distributions. A burn-in period of 20,000 iterations was used to stabilize the chains, and the remaining 80,000 samples were retained for inference.

The MCMC process was monitored using the Gelman-Rubin diagnostic and trace plots to ensure convergence. The convergence criterion was set as a potential scale reduction factor (PSRF) below 1.05, indicating consistent results across chains.

### 5.2. Convergence Diagnostics and Observations

The convergence of the model was assessed using the following methods:

- **Gelman-Rubin Diagnostic:**

  The Gelman-Rubin plots for all parameters indicate that the shrink factors converge

rapidly towards 1. This indicates that the variability between chains is consistent with the variability within chains, confirming convergence.

| Statistic | Point Est. | Upper C.I. |
|---|---|---|
| alpha | 1.03 | 1.07 |
| beta1 | 1.00 | 1.00 |
| beta2 | 1.00 | 1.00 |
| beta3 | 1.00 | 1.00 |
| gamma1 | 1.03 | 1.07 |
| gamma2 | 1.02 | 1.06 |
| sigma_b | 1.01 | 1.01 |

Table 1: Potential scale reduction factors with point estimates and upper confidence intervals (C.I.).

Multivariate psrf : 1.02 (less than 1.05, indicating convergence)

- **Trace Plots:**
  The trace plots for each parameter ($\alpha$, $\beta_1$, $\beta_2$, $\beta_3$, $\gamma_1$, $\gamma_2$, $\sigma_b$) exhibit good mixing and overlap among the four chains. The lack of apparent trends or patterns in the chains suggests good mixing and indicates that the sampler has adequately explored the posterior distribution. The sigma_b parameter shows slight variability post burn-in but maintains stability, indicating a consistent parameter estimate. These plots are included in Appendix D.

- **Density Plots:**
  The density plots for each parameter show smooth, unimodal distributions that are symmetric around their means, indicating well-defined posterior distributions. There are no heavy tails or skewness, suggesting good statistical properties.

These results confirm that the chains have reached the stationary distribution, demonstrating reliable convergence and consistent posterior estimates.

### 5.3. Efficiency Assessment

To further evaluate MCMC efficiency, the ratio of the standard deviation (SD) to the time-series standard error (SE) was calculated. A ratio greater than 20 for majority of the parameters indicated sufficient effective sample sizes, signifying that the samples within each chain were

largely independent. The exact values of SDs and Monte Carlo errors can be found in Appendix C.

## 6. Conclusion

For the parameters and hyperparameters of the Bayesian hierarchical model, the summary output and each parameter's corresponding confidence interval can be found in Appendix C. After running the JAGS code for 100,000 iterations and discarding the first 20,000 as burn-in, it was observed that all parameters converged to a shrink factor close to one after approximately the 80,000th iteration. The Gelman-Rubin diagnostic indicated that the upper 95% intervals of the estimates converged well with the median, confirming that the MCMC chains had mixed properly.

For the effects of the modeled variables on bird species richness, some notable observations are highlighted here, while the detailed numerical results can be found in Appendix C. The model identified significant associations between species richness and multiple covariates, including observational effort, number of observers, regional variability, and seasonal effects. After convergence, the model identified significant associations between bird species richness and several covariates. Notably, the parameter $\beta_2$ (representing effort) and $\beta_3$ (representing the number of observers) showed slightly negative associations with species richness. This suggests that higher effort and increased number of observers may correspond to more accurate species identification, thereby recording fewer unique species. The confidence intervals for both $\beta_2$ and $\beta_3$ did not cross zero, indicating that the effects are statistically significant.

The Fourier coefficients ($\gamma_1$ and $\gamma_2$) also indicated negative effects, reflecting seasonal variations where species richness decreases during specific times of the year. The random effects parameter ($\sigma_b$) captured moderate regional variability, suggesting that bird species richness differs across U.S. states. The positive intercept ($\alpha$) indicates a baseline richness that can vary depending on spatial and seasonal factors.

Although some parameters, such as $\beta_1$, displayed near-zero effects or had confidence intervals encompassing zero, the overall model provided meaningful insights into the factors influencing bird species richness. The high convergence rate and stable posterior distributions suggest that the model accurately captured the underlying dynamics.

Future research could extend this model by incorporating additional environmental variables such as vegetation cover or climate data, which may offer deeper insights into the ecological factors affecting bird populations. Applying this model framework to longitudinal datasets could also help track changes in biodiversity over time, supporting conservation efforts.

## Appendix A: AI Attribution

This report utilized generative AI tools (such as ChatGPT) for drafting certain sections and formatting in LaTeX. All content generated through AI was critically reviewed, refined, and validated by the project team to ensure accuracy and relevance.

## Appendix B: Group Contributions and Project Modifications

This appendix offers a concise overview of the responsibilities and contributions made by each member of the group, as well as changes made from the original proposal. Although many aspects of the project proposal draft were retained, numerous revisions were made throughout the course of the project.

For this project, Alarsh worked on data acquisition and initiated the first run of the code, which was subsequently enhanced and optimized by Anshika, Sarthak and Aadya. Alarsh also contributed by writing the section on Data Description. Aadya created the Directed Acyclic Graph (DAG) and composed the conclusion, while Anshika wrote the Methodology section and Introduction sections. Sarthak took charge of the Implementation and Abstract section. All four members collaboratively revised the Appendix and References after updates were made to the overall project and the data being utilized.

The initial project proposal aimed to utilize a Poisson model for predicting bird species richness. However, after performing preliminary analysis, it became evident that the data exhibited overdispersion. Consequently, a Bayesian hierarchical model using a Negative Binomial distribution was adopted, as it effectively handles overdispersion while maintaining model accuracy.

Additionally, the originally chosen dataset underwent significant transformations and reduction. Variables that were deemed most relevant were selected, while those that lacked interpretability were excluded. This data refinement process ensured that the final model was both robust and interpretable, thereby enhancing the overall analysis and results.

Through these collective efforts and methodological adjustments, the project was able to meet its objectives and deliver comprehensive insights into bird species richness across the United States.

## Appendix C: Summary Statistics

| Statistic | Mean | SD | Naive SE | Time-series SE | SD/Time series | Confidence Interval (CI) |
|---|---|---|---|---|---|---|
| alpha | 2.40449 | 2.02308 | 3.199e-03 | 0.2723843 | 7.43 | [-0.63377, 7.32536] |
| beta1 | -0.06003 | 31.66404 | 5.007e-02 | 0.0502072 | 630.46 | [-62.08817, 61.99109] |
| beta2 | -0.10152 | 0.03820 | 6.041e-05 | 0.0001610 | 237.37 | [-0.17013, -0.01823] |
| beta3 | -0.05981 | 0.03829 | 6.055e-05 | 0.0001324 | 289.16 | [-0.13793, 0.01439] |
| gamma1 | -2.18796 | 1.86583 | 2.950e-03 | 0.2291487 | 8.14 | [-6.74443, 0.62950] |
| gamma2 | -0.46752 | 0.91746 | 1.451e-03 | 0.0640988 | 14.31 | [-1.89248, 1.71747] |
| sigma_b | 0.07731 | 0.08832 | 1.396e-04 | 0.0009237 | 95.63 | [0.00232, 0.30157] |

Table 2: Summary statistics including mean, standard deviation, naive SE, time series SE, SD/Time series ratio, and confidence intervals (2.5% and 97.5%).

## Appendix D: Trace and Density Plots

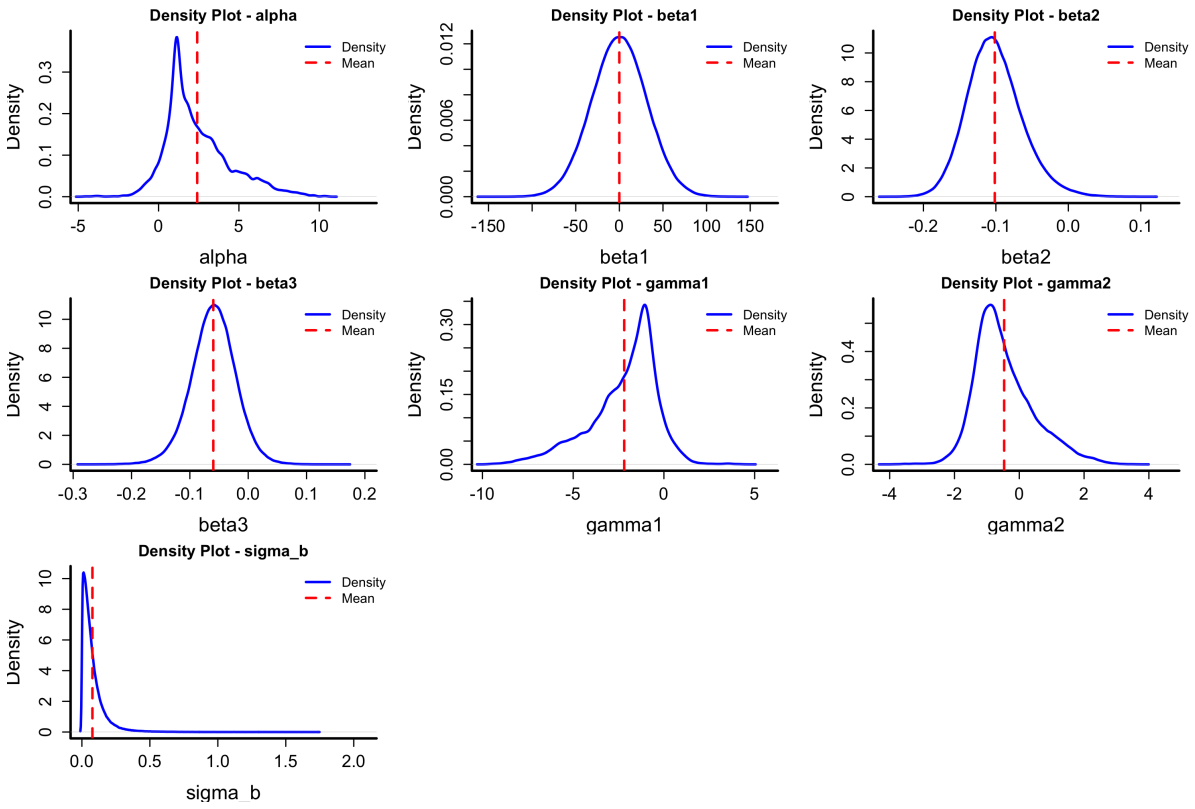


Figure 3: Density plot for parameter distribution.

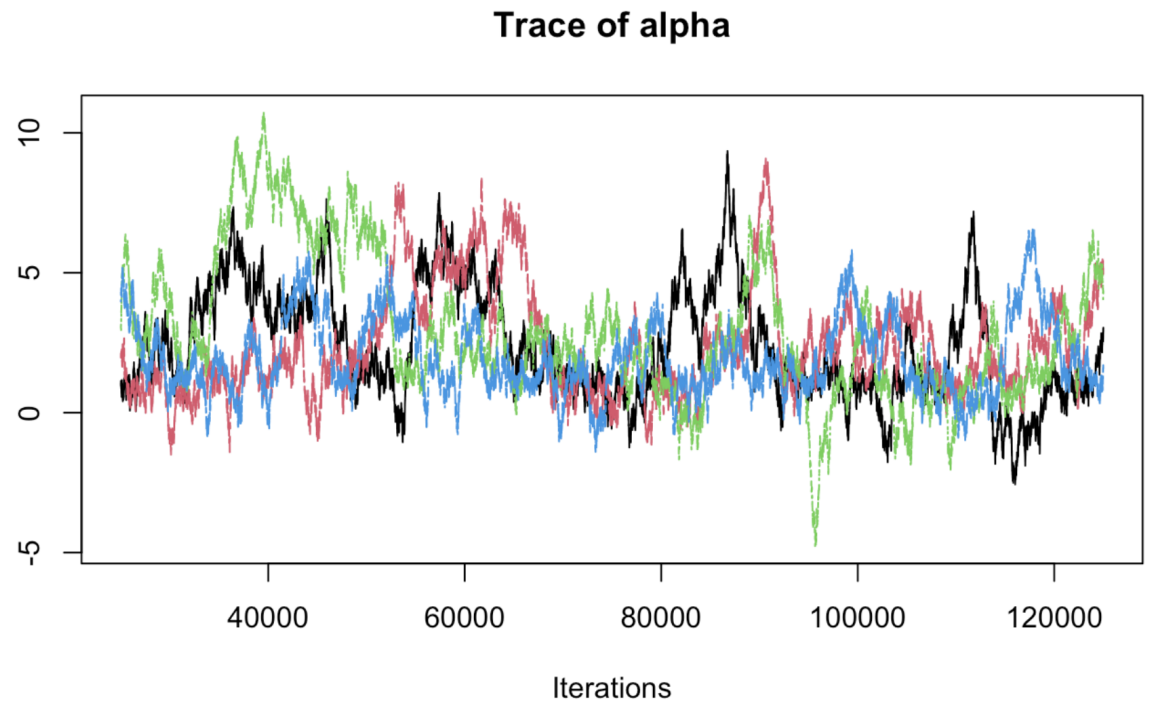


Figure 4: Trace plot for alpha

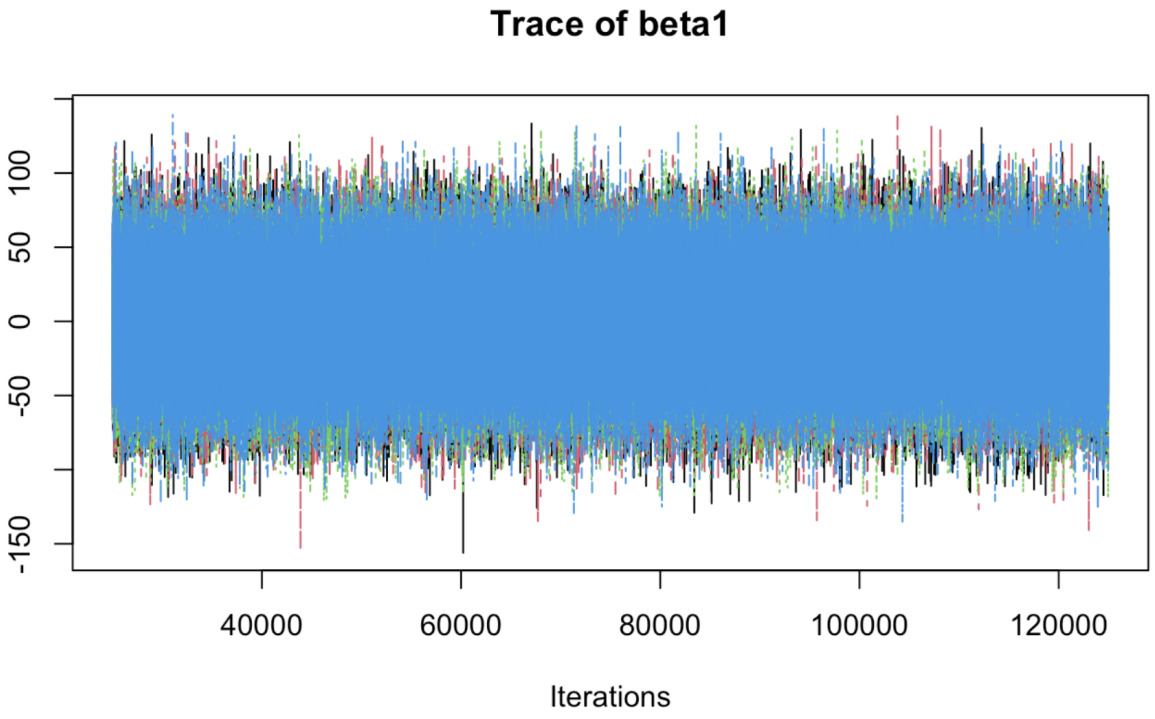


Figure 5: Trace plot for beta1

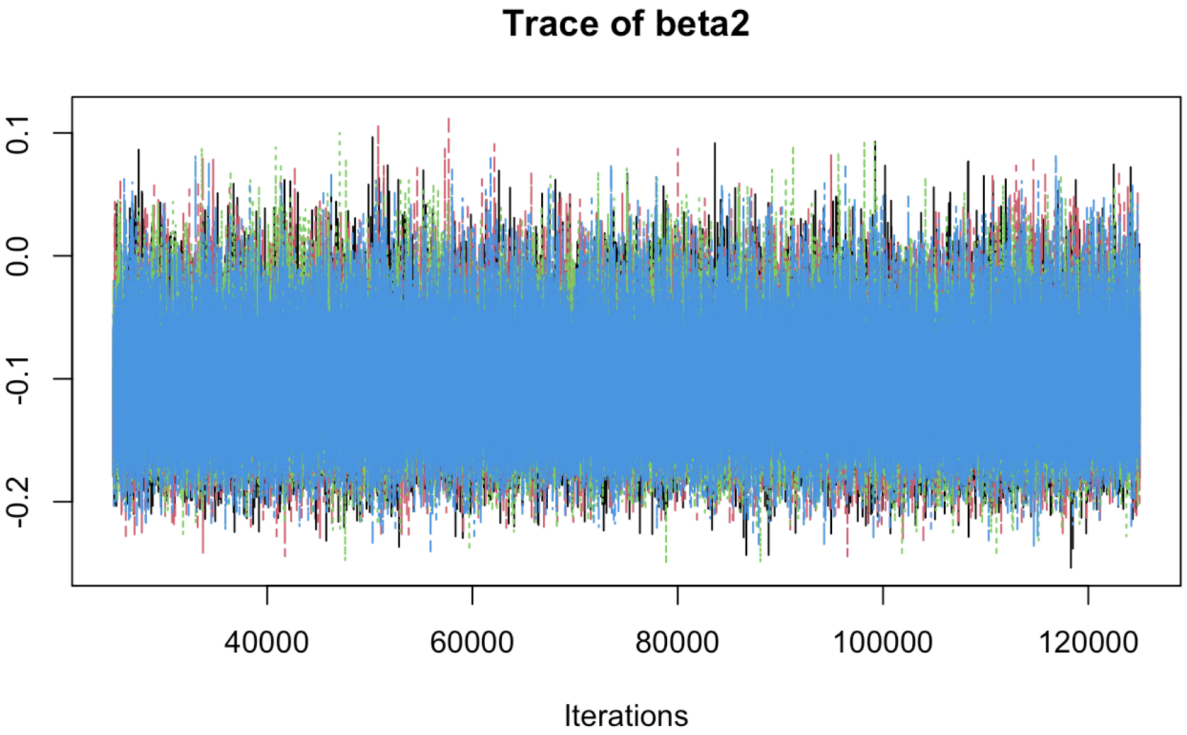


Figure 6: Trace plot for beta2

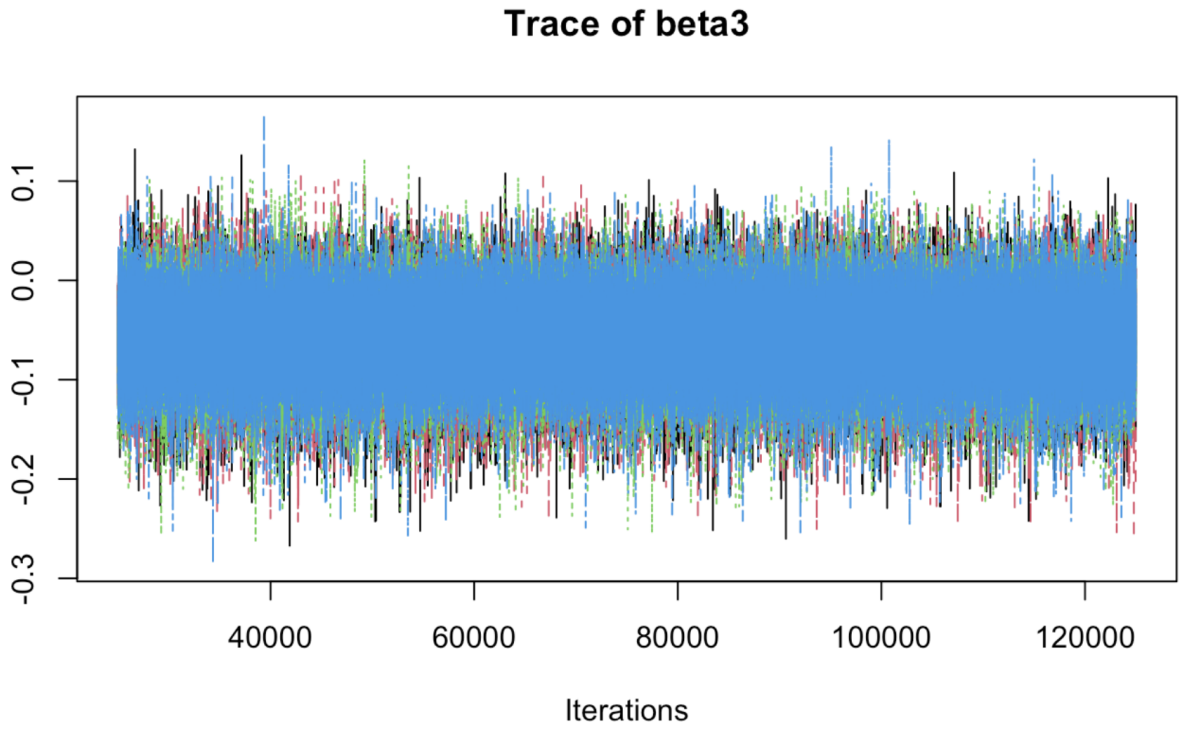


Figure 7: Trace plot for beta3

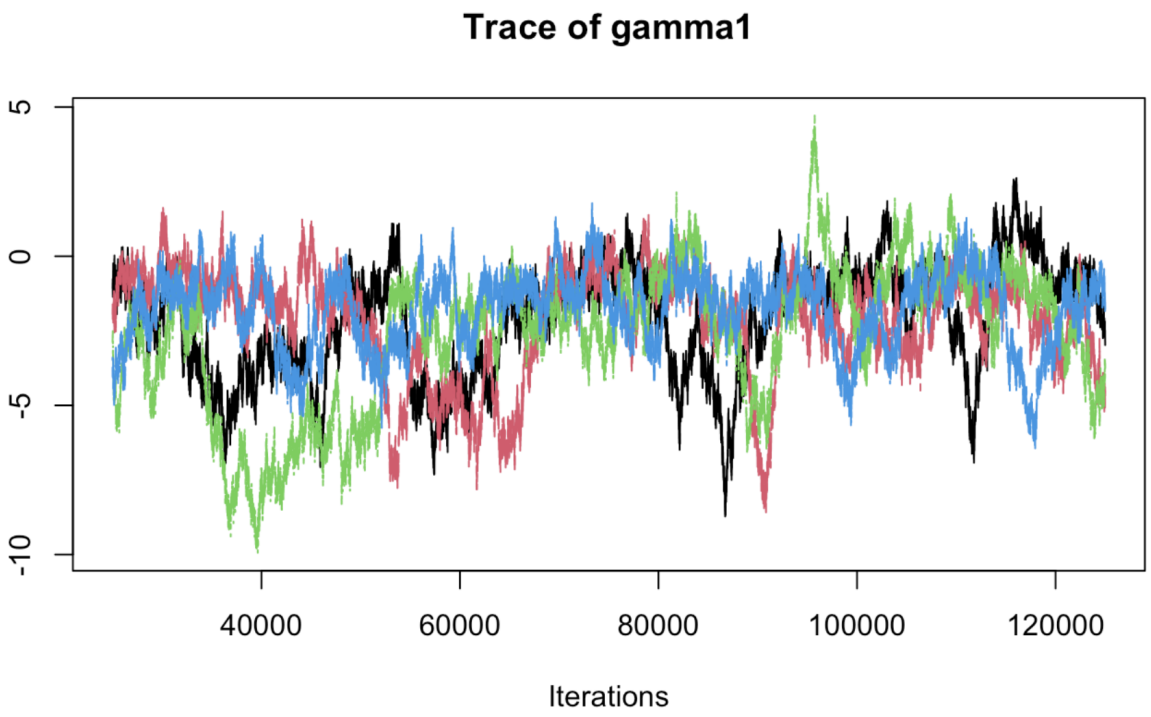


Figure 8: Trace plot for gamma1

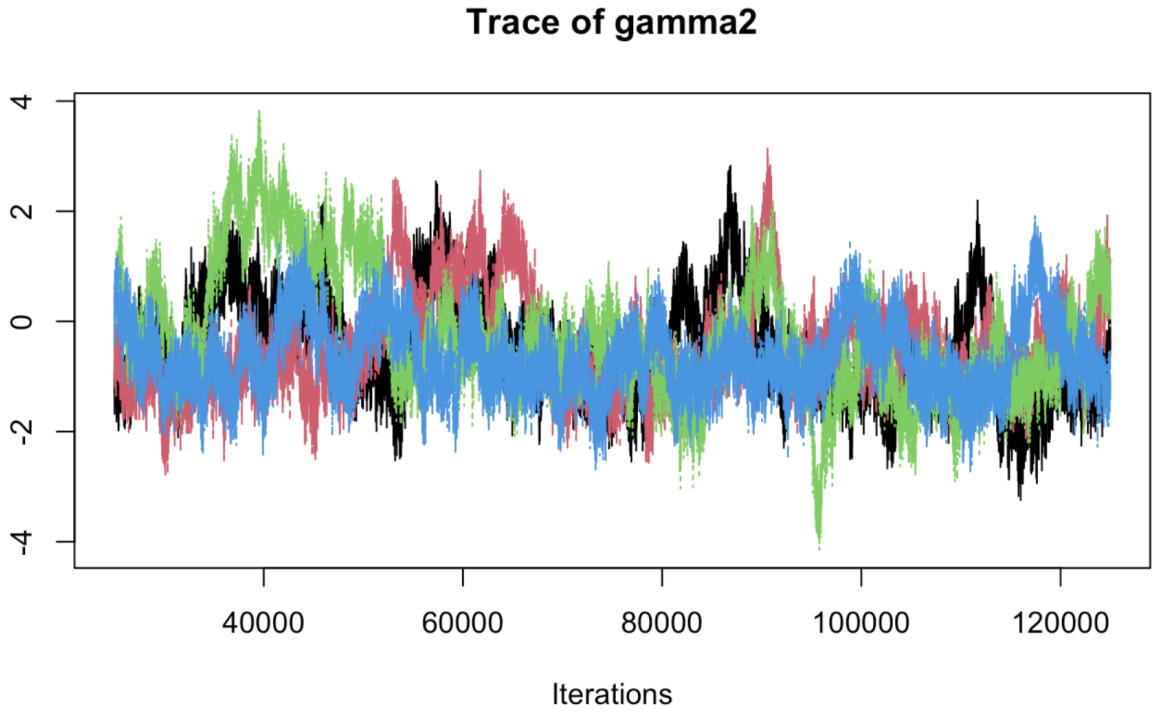


Figure 9: Trace plot for gamma2

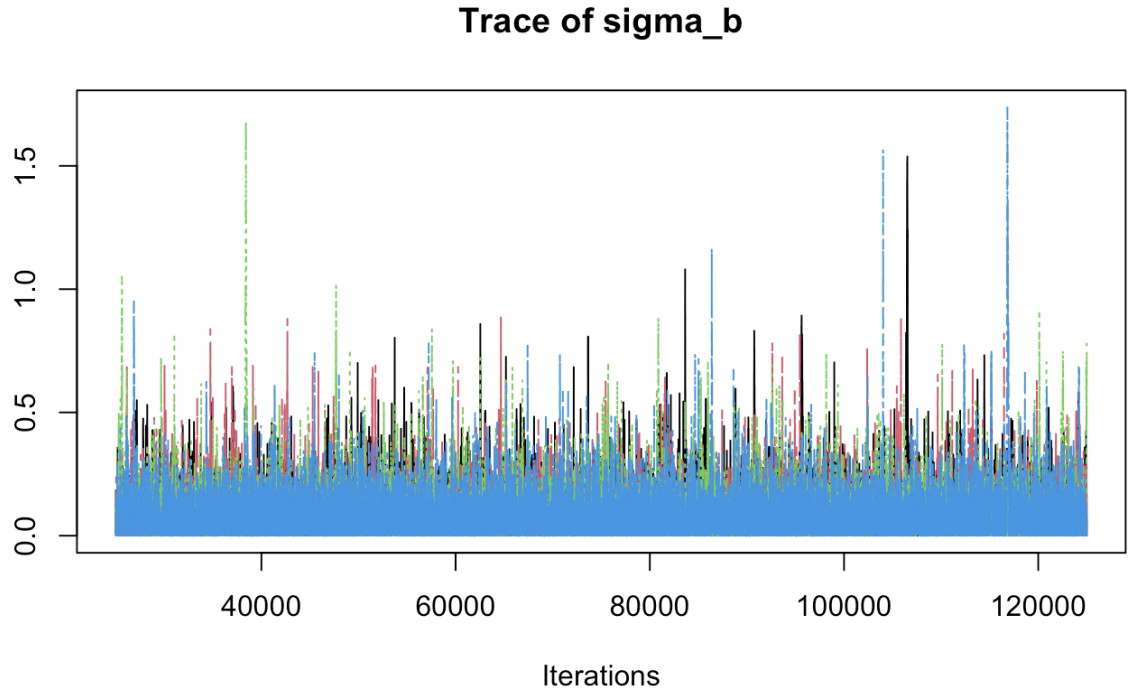


Figure 10: Trace plot for sigmab

## Appendix E: RJAGS Code for Model

```
model_string <- "
model {
  pi <- 3.141592653589793

  # Dispersion parameter for Negative Binomial
  r ~ dgamma(0.01, 0.01)

  for (i in 1:N) {
    # Convert lambda to probability p
    p[i] <- r / (r + lambda[i])

    # Negative Binomial distribution with r and p
    Y[i] ~ dnegbin(p[i], r)

    # Log-linear model for lambda
    log(lambda[i]) <- alpha + b_state[region[i]]
```

```
                    + beta2 * effort[i]
                    + beta3 * observers[i]
                    + f_day[day[i]]
                    + log_duration[i]
}

# Random intercept for each region
for (j in 1:J) {
  b_state_raw[j] ~ dnorm(0, 1)
  b_state[j]     <- sigma_b * b_state_raw[j]
}

# Seasonal effect (Fourier series)
for (t in 1:365) {
  f_day[t] <- gamma1 * sin(2 * pi * t / 365) + gamma2 * cos(2 * pi * t / 365)
}

# Hyperpriors for intercept + seasonals
mu_block      ~ dnorm(0, 0.1)
sigma_alpha   ~ dunif(0, 5)

mu_gamma      ~ dnorm(0, 0.1)
sigma_gamma   ~ dunif(0, 5)

# Non-centered raw parameters
alpha_raw     ~ dnorm(0, 1)
gamma1_raw    ~ dnorm(0, 1)
gamma2_raw    ~ dnorm(0, 1)

# Transform to actual intercept & seasonals
alpha   <- mu_block  + sigma_alpha * alpha_raw
gamma1  <- mu_gamma  + sigma_gamma * gamma1_raw
gamma2  <- mu_gamma  + sigma_gamma * gamma2_raw
```

```
  # Priors for coefficients
  beta1 ~ dnorm(0, 0.001)
  beta2 ~ dnorm(0, 0.001)
  beta3 ~ dnorm(0, 0.001)

  sigma_b ~ dunif(0, 10)
}"
```

## References


[1] Ogawa, R., Wang, G., Burger, L. W., Strickland, B. K., Davis, J. B., & Cunningham, F. L. (2024). Bayesian integrated species distribution models for hierarchical resource selection by a soaring bird. *Ecological Informatics, 82*, 102787. https://doi.org/10.1016/j.ecoinf.2024.102787

[2] Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation, 142*(10), 2282-2292. https://doi.org/10.1016/j.biocon.2009.05.006

[3] Mutiso, F., Pearce, J. L., Benjamin-Neelon, S. E., Mueller, N. T., Li, H., & Neelon, B. (2022). Bayesian negative binomial regression with spatially varying dispersion: Modeling COVID-19 incidence in Georgia. *Spatial Statistics, 52*, 100703. https://doi.org/10.1016/j.spasta.2022.100703

[4] Bloomfield, P. (2004). Fourier Analysis of Time Series: An Introduction. John Wiley & Sons.