Point Pattern Modelling for Degraded Presence-Only Data Over Large Regions

Avishek Chakraborty, Alan E. Gelfand, Adam M. Wilson, Andrew M. Latimer

Journal of the Royal Statistical Society. Series C: Applied Statistics

H-Index - 78

MTH643 - Spatial Analysis

Instructor: Arnab Hazra

Vrinda Rawal (221466)
Snigdha Taneja (221428)
Rahul Birru (221384)
Kanchan Maan (221327)

Contents

List of Tables

1 Abstract

Point Pattern Modeling involves analyzing spatial arrangements of points, revealing patterns within geographic data. Presence-only data, indicating the occurrence of events without noting absences, is central to ecological studies. The title emphasizes addressing challenges posed by "Degraded Presence-Only Data" over extensive regions, highlighting a focus on compromised data quality. This research aims to advance spatial analysis methods for large-scale datasets, contributing to a deeper understanding of spatial phenomena.

2 Introduction

Ecologists have long been interested in understanding the distribution of species, and this topic has been extensively studied. To gain insights into this complex issue, researchers have developed numerous model-based approaches. While there are many different methods available, we will focus on those that are particularly useful for modeling the presence of species.

Our aim in this paper is to:

- **First collect data from a set of locations which are recorded as presence of the species in question**
- **Examine the distribution of finding a species at a given location.**
- **Finally, understand how this is related to various environmental conditions present there.**

To accomplish these aims, researchers typically use a generalized additive binary regression model or a Gaussian process prior or a logistic regression model. However, the crucial issue in these model-based approaches is that they often ignore spatial dependence.

Spatial dependence refers to the idea that whether a species is present at one location can be influenced by whether it is present at nearby locations. This can be important because of ecological factors like localized dispersal (species spreading in a confined area) and unobserved variables with spatial patterns. This is why aside from covariate information which can be included in the model, accounting for such spatial correlation is also a vital component of modelling the presence of species.

Aside from this, just dealing with presence-only data is gaining more and more traction these days, with it being on par with presence-absence data, if not better. The reason is that presence-only data offers a complete census, whereas presence-absence data has limited information due to being restricted to a specific set of sampling sites.

Therefore, **the study focuses specifically on the presence-only setting, where only the presence of a species is recorded, and the spatial dependence must be modelled accordingly.**

The approach that has been proposed in this paper for species distribution modeling is a game-changer in that it is fully model-based, providing a comprehensive understanding of the uncertainty present in the region. We achieve this by modeling presence-only data as a point pattern, with the intensity specified based on the available environmental covariates across the region. This is done through regression modeling, which allows for a natural interpretation of the coefficients. To introduce spatial structure to the intensity surface, we use a hierarchical model that incorporates spatial random effects. **What sets our approach apart from others is that we do not assume any background or pseudo-absence samples.**

This modeling has been done on species data from the Cape Floristic Region (CFR) in South Africa which offers a unique opportunity to study species diversity.

3 Fundamental Concerns and Biases in Common Modeling Approaches

As mentioned earlier, the simplest approaches to predicting species distributions based on presence-only data are based directly on the environmental envelope that is associated with observed occurrences; there is no spatial component to the prediction. Also, most of these approaches fail to address the bias that may exist in sampling occurrences. Yet such bias in sampling is a common problem.

Some of these biases that may be involved in our data are:

- **Not visiting the locations where species are unlikely to occur.**
- **Certain places being almost inaccessible (interiors of the reserve, mountainous regions, etc)**
- **Chance of missing a presence location, depending on the density of growth, time of the year, etc.**
- **Certain locations being oversampled like those that are close to roads or towns.**
- **Prevalence of false records such as false absence and false presence will affect the effort to construct a model.**

4 Point Process Modelling

We see the observed presence-only data as a degraded point pattern. We give a broad point process specification for this problem in Section 4.1. We formalize the likelihood and posterior in Section 4.2 and suggest a grid cell-level approximation.

4.1 Probability model for presence locations

When analyzing point patterns of species presence, it's common to assume a non-homogeneous Poisson process (NHPP). However, it's crucial to account for degradation in the presence data, which can arise from two main factors:

1. **Sampling bias:** This occurs when data collection introduces a non-random bias, such as oversampling or undersampling certain areas due to accessibility or other factors.

2. **Land transformation:** This refers to changes in the environment caused by human activities, like agriculture or the spread of invasive species, that render certain areas unsuitable for a species, thus impacting its presence.

To address this degradation, two types of intensity are conceptualized:

1. **Potential Intensity:** It represents the intensity of species presence that would exist in an ideal, undisturbed environment.

2. **Realized Intensity:** It is the intensity that operates in the presence of degradation. It takes into account the impact of sampling bias and land transformation on the observed presence data.

These two intensity measures help researchers understand how the actual presence data differs from what would be expected in an ideal scenario. Further, we tile the intensity to reflect our inability to explain it at a spatial resolution finer than our grid cells.

The spatial domain is denoted by D. The potential intensity surface is represented by $\lambda(s)$, and the availability surface over D is denoted by $U(s)$. The sampling effort surface is represented by $T(s)$. Each geographical region corresponding to cell i is denoted by A_i. The probability that a randomly selected location in A_i was available and sampled is given by q_i. The degradation at location s is represented by $\lambda(s)U(s)T(s)$. The conditional probability that a randomly selected location in cell i is sampled given that it is available is denoted by P_i.

Glossary

A_i Geographical region corresponding to cell i. 6

P_i Conditional probability that a randomly selected location in cell i is sampled given that it is available. We set $p_i = 1$ if cell i was sampled for any species in our dataset, otherwise we set $p_i = 0$. 6

$T(s)$ Sampling effort surface over D. It is also a binary surface such that $T(s) = 1$ implies that location s is sampled, $T(s) = 0$ implies that location s is not sampled. 6

$U(s)$ Availability surface over D. It is a binary surface such that $U(s) = 1$ implies location s is untransformed by land use, $U(s) = 0$ implies that location s is transformed by land use, assuming no sampling bias. 6

$\lambda(s)$ Potential intensity surface, i.e., the intensity in the absence of degradation. 6

$\lambda(s)U(s)T(s)$ Degradation at location s. 6

q_i $\int_{A_i} T(s)U(s)\,ds/|A_i|$; The probability that a randomly selected location in A_i was available and sampled. 6

D Spatial domain. 6

4.2 Likelihood and posterior

In this section, we will cover the modeling of the potential intensity surface $\lambda(\cdot)$. **We are not considering the potential intensity surface under transformation, due to lack of information about the locations.** We use a Gaussian Process (GP) prior, which results in a log-Gaussian Cox process model for the observed data. The environmental covariates, denoted as $x(s)$, are going to affect the intensity. For a location $s \in D$, we have $\log \lambda(s) = x^{\top}(s)\beta + w(s)$, where $w(\cdot)$ is a zero-mean stationary, isotropic Gaussian Process over D, aimed at capturing spatial correlation in the λ surface.

Suppose, we have n_i presence locations $(s_{i,1}, s_{i,2}, \ldots, s_{i,n_i})$ within cell i for $i = 1, 2, \ldots, l$. Since these locations are available and sampled, we have $U(s_{i,j})T(s_{i,j}) = 1$ for $0 \leq j \leq n_i$, $1 \leq i \leq l$.

The likelihood function corresponding to the Non-Homogeneous Poisson Process (NHPP) $\lambda(\cdot)$ proposed above becomes

$$L[\lambda(\cdot); \{s_{i,j}\}] \propto \exp\left\{ - \int\limits_{D} \lambda(s)U(s)T(s)\, ds \right\} \prod_{i=1}^{I} \prod_{j=1}^{n_i} \lambda(s_{i,j}) \tag{1}$$

Let D denote our CFR study domain. For each cell, we have information on l covariates as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{il})$. We assume that $\lambda(\cdot)$ is a tiled surface such that for cell i, the height is $\triangle\lambda(s_i)$ where $\triangle$ is the area of the cell and s_i is the centroid. Then $n_i \sim \text{Po}(\Delta\lambda(s_i)q_i)$ where n_i are independent.

From equation 1, $\log \lambda(s)$ follows a Gaussian Process (GP) prior. The posterior distribution can then be taken as:

$$\pi(\lambda_{s_{1:m}}, \beta, \theta | n, x, u, q) \propto \exp\left\{ - \sum_{i=1}^{I} \lambda(s_i)(\triangle_i q_i) \right\} \prod_{i=1}^{m} \lambda(s_i)^{n_i} \times \phi_m[\log \lambda_{s_{1:m}} | \beta, x, \theta] \pi(\beta) \times \pi(\theta) \tag{2}$$

where ϕ_m denotes the m-dimensional Gaussian density and θ the parameters in the covariance function of $w(\cdot)$ in equation (1). We can then sample from this posterior distribution using Markov Chain Monte Carlo (MCMC) methods.

5 Computation

The most critical issue in working with CFR dataset was handling the modeling of over 30000 locations, which is basically analogous to the 'large n' problem in GPs. To deal with this, we employ predictive process approximation for Gaussian random fields.

5.1 Predictive process approximation

In our case, we use this approximation on λ_{s_j}

The predictive process method works in the following way on a high-dimensional GP. If $w(\cdot)$

is the zero-mean GP under consideration, and there are $s_{1:I} = (s_1, s_2, \ldots, s_I)$ locations in our data, where I is large, then the method first selects r locations $s_{1:r}^{\text{knot}}$ from the region. These selected r locations are called knots. Then we can replace $w(s_{1:I})$ in the model equation with $w(1 : r)^{\text{knot}}$ (as mentioned in the paper), where the matrix L is calculated from the spatial dependence structure of $w(\cdot)$.

We introduce bias correction in our model as well. This correction introduces a heteroscedastic error ϵ^* with $V(\epsilon^*) = \text{var}(w(s_j)) - \text{var}(w(s_j^{\text{knot}}))$.

5.2 Markov chain Monte Carlo sampling

We approximate equation 2 using the predictive process method mentioned in the above section. We can rewrite $\Lambda_{0,i} = \lambda(s_i)\triangle$, where $\triangle$ is the expected species count in cell i. Finally, we can now define the hierarchical model as:

$$\begin{aligned} n_i \mid \Lambda_{0,i} &\overset{\text{ind}}{\sim} \text{Poi}(\Lambda_{0,i} q_i), && i = 1, 2, \ldots, I, \\ \log(\lambda(s_i)) &= x_i^T \beta + \widetilde{w}(s_i) + \epsilon_i^*, \\ \widetilde{w}_{s_{1:I}} &= R_{I,r}(\phi) R_r^{-1}(\phi) \widetilde{w}(s^0_{(1:r)}), \\ w(s^0_{1:r}) &\sim \mathcal{N}_r(0_r, R_r(\phi)), \\ \epsilon_i^* &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2(1 - R_{I,r}(\phi) R_r^{-1}(\phi) R_{r,I}(\phi))_{ii}\right), \\ \pi(\beta, \phi, \sigma^2) &= \pi(\beta) \cdot \pi(\phi) \cdot \pi(\sigma^2). \end{aligned}$$

Out of I cells, we first sample only m cells and these contribute to the model fitting then inference for the remaining I-m cells is done from their posterior predictive distributions. This particular approximation yields:

$$\begin{aligned} \log\left\{\lambda(\mathrm{s}_{m+1:I})\right\} &= x_{\mathrm{s}_{m+1:I}} \beta + R_{I-m,r}(\phi) R_r^{-1}(\phi) w(s^0_{1:r}) + \epsilon^*_{m+1:I}, \\ \epsilon_i^* &\overset{\text{ind}}{\sim} \mathcal{N}\left(0, \sigma^2(1 - R_{I,r}(\phi) R_r^{-1}(\phi) R_{r,I}(\phi))_{ii}\right), \quad m < i \leq I. \end{aligned}$$

So, conditioning on posterior samples of β, ϕ, and $w(s_{1:r^{\text{knot}}})$, we can draw samples from the posterior predictive distribution of $\log(\lambda(s_{m+1:I}))$, which is independent of $\log\{\lambda(s_{1:m})\}$. This is computationally very efficient as this saves us from drawing samples from a high-dimensional multivariate Gaussian distribution.

The algorithm that we have used to sample from this model has been described below:

1. We first sample only m cells from the total I cells, and out of these m cells, we select r cells. Here R_{r+I} are exponential correlation terms with a decay parameter ϕ.

$$\begin{aligned}
L_1(\phi) &= R_{m,r}(\phi)R_r^{-1}(\phi),\\
L_2(\phi) &= R_{I-m,r}(\phi)R_r^{-1}(\phi),\\
M_1(\phi) &= I_m - \mathrm{diag}\{L_1(\phi)R_{r,m}(\phi)\},\\
M_2(\phi) &= I_{I-m} - \mathrm{diag}\{L_2(\phi)R_{r,I-m}(\phi)\},\\
X_1 &= x(s_1 : m),\\
X_2 &= x(s_{m+1:I}).
\end{aligned}$$

2. For β parameter the prior distribution is normal distribution and the posterior distribution is t-distribution

$$\begin{aligned}
\pi(\beta) &= \mathcal{N}(\beta_0, \Sigma_0),\\
\beta \mid \ldots &\sim t(\mu_\beta, \Sigma_\beta),\\
\Sigma_\beta^{-1} &= \Sigma_0^{-1} + \sigma^{-2}X_1^\top\{L_1(\phi)R_{r,m}(\phi) + M_1(\phi)\}^{-1}X_1,\\
\mu_\beta &= \Sigma_\beta\left[\Sigma_0^{-1}\beta_0 + \sigma^{-2}X_1^\top\{L_1(\phi)R_{r,m}(\phi) + M_1(\phi)\}^{-1}\log(\lambda_{1:m})\right].
\end{aligned}$$

3. The prior and posterior distribution of σ^2 parameter is the conjugate prior inverse gamma distribution.

$$\begin{aligned}
\pi(\sigma^2) &= \mathrm{IG}(a_0, b_0),\\
\sigma^2 \mid \ldots &\sim \mathrm{IG}(a_{\sigma^2}, b_{\sigma^2}),\\
a_{\sigma^2} &= a_0 + \frac{m}{2},\\
b_{\sigma^2} &= b_0 + e^\top\{L_1(\phi)R_{r,m}(\phi) + M_1(\phi)\}^{-1}e/2,\\
e &= \log(\lambda_{1:m}) - X_1\beta.
\end{aligned}$$

4. The prior for $w(s_1 : r^0)$ is normal distribution

$$\begin{aligned}
w(s_1 : r^0) &\sim \mathcal{N}_r(\mu_w, \Sigma_w),\\
\Sigma_w^{-1} &= R_r(\phi)^{-1} + \sigma^{-2}L_1^\top(\phi)M_1^{-1}(\phi)L_1(\phi),\\
\Sigma_w^{-1}\mu_w &= \sigma^{-2}L_1^\top(\phi)M_1^{-1}(\phi)\{\log(\lambda_{1:m}) - X_1\beta\}.
\end{aligned}$$

5. The distribution for the ϕ parameter is uniform distribution

$$\begin{aligned} &\pi(\phi) \sim U(\phi_0, \phi_1), \\ &S(\phi) = -\frac{\log\left\{\left|L_1(\phi)R_{r,m}(\phi) + M_1(\phi)\right|\right\}}{2} \\ &\qquad - \frac{\{\log(\lambda_{1:m}) - X_1\beta\}^\top \{L_1(\phi)R_{r,m}(\phi) + M_1(\phi)\}^{-1}\{\log(\lambda_{1:m}) - X_1\beta\}}{2\sigma^2}. \end{aligned}$$

6. The distribution for λ parameter is log-normal distribution

$$\begin{aligned} &\pi(\lambda_i \mid \ldots) \sim \mathrm{LN}(\lambda_i; x_i^\top\beta + [L_1(\phi)w(s_1 : r^0)]_i, \sigma^2[M_1(\phi)]_{ii}), \\ &\mathrm{Poisson}(y_i \mid \lambda_i, q_i), \end{aligned}$$

In this algorithm, β, σ^2, and $w(s_{1:r^0})$ were updated using Gibbs steps, whereas, for $\lambda(s_{1:I})$ and ϕ, Metropolis-Hastings steps were used.

6 Data Description

The floral kingdom of CFR has a large diversity of plant species (9000 plant species). We consider **six species within the Proteaceae plant family**. They are **Protea aurea, PRAURE, present at 674 locations, Protea cynaroides, PRCYNA, present at 8412 locations, Leucadendron salignum, LDSG, present at 24294 locations, Protea mundii, PRMUND, present at 821 locations, Protea punctata, PRPUNC, present at 2319 locations, and Protea repens, PRREPE, present at 15271 locations.** The covariates employed for the modeling and study include the following:

1. **July (winter) minimum temperature**
2. **January (summer) maximum temperature**
3. **Wind speed at 2 meters range average**
4. **Evapotranspiration**
5. **Profile soil moisture**
6. **Precipitation corrected sum**

6.1 Data source

Initially, the presence data . i.e. the latitude, longitude and the species data was collected from SANBI, South African National Biodiversity Institute. We collected this data particularly for the year 2001. Using this presence data, the co-variates data for the year 2001 corresponding to the species latitude, and longitude was extracted from NASA Power Data Access Viewer.
The data was collected by utilizing the site's API, employing a loop that iteratively called the API with varying latitude and longitude coordinates. This approach allowed us to obtain the covariates corresponding to the desired presence locations of the species in **Python.**

6.2 Distribution of presence-only data of different species

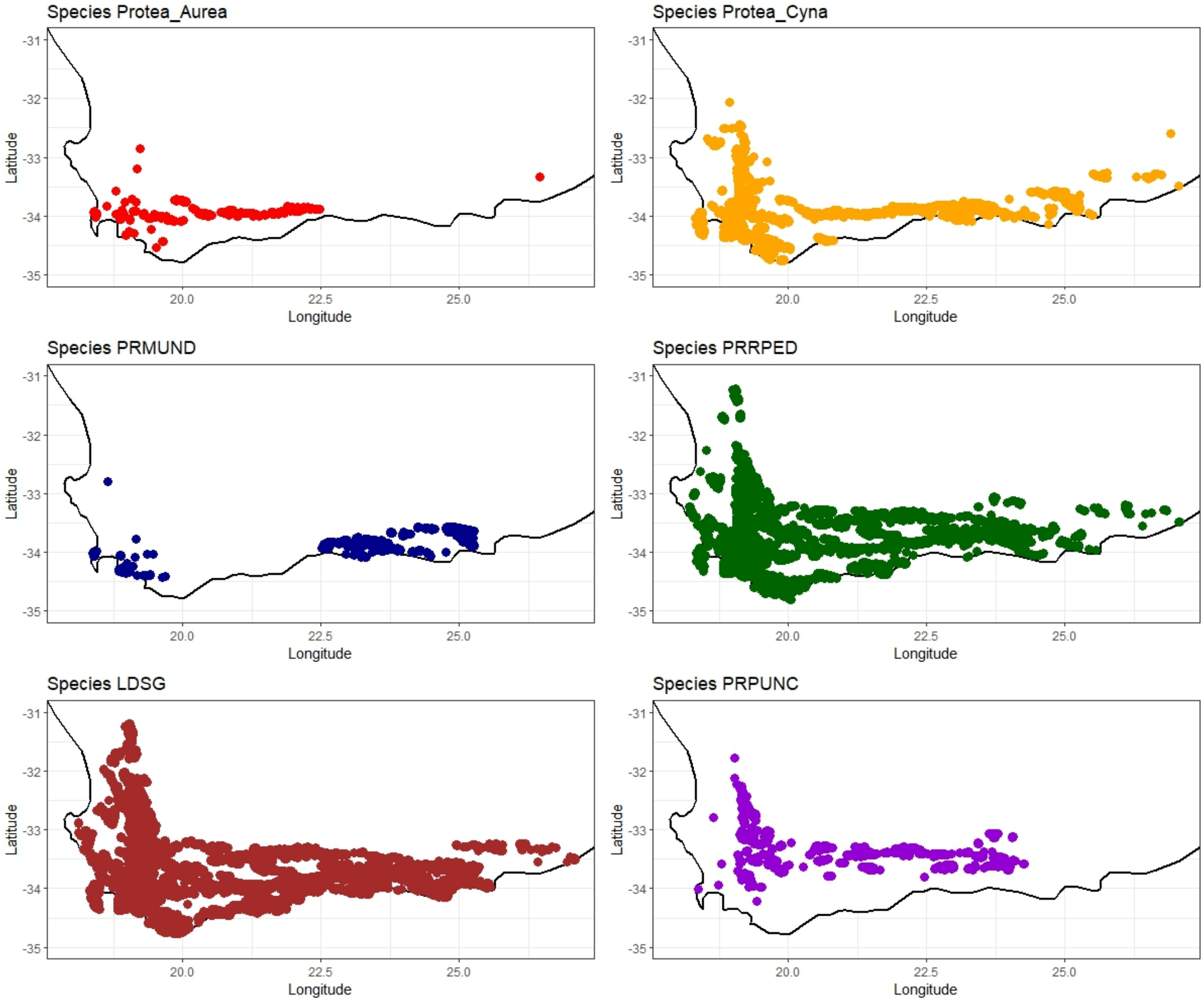


Figure 1:
Maps for the distribution of presence-only data of different species

7 Implementation Strategies, Challenges, and Remedial Approaches

The MCMC algorithm was implemented in R

7.1 Implementation

1. Centering and scaling of all the covariates
2. Intialiazation of all the prior parameters
3. Generation of outputs using Markov Chain Monte Carlo (MCMC) sampling, executing 1000 iterations with a burn-in phase of 250 samples.[1]
4. Parameters were updated using Gibbs Sampling and Metropolis-Hastings Sampling.
5. Examined Outputs and Graphical Representations

7.2 Challenges and Solutions

1. We were unable to sample from the t-distribution since the inverse of Σ_{β}^{-1} was not positive definite. So, to deal with this we used the nearPD function on the matrix.
2. For efficient computing, we used Rcpp package in R for calculating matrix multiplication and its inverse.

[1] * Here we are assuming that there is no inter-species dependence, i.e. why **we are running the MCMC chains separately** on each species.

8 Results and Insights

8.1 Results Summary

8.1.1 Distribution of all the six covariates across the CFR region

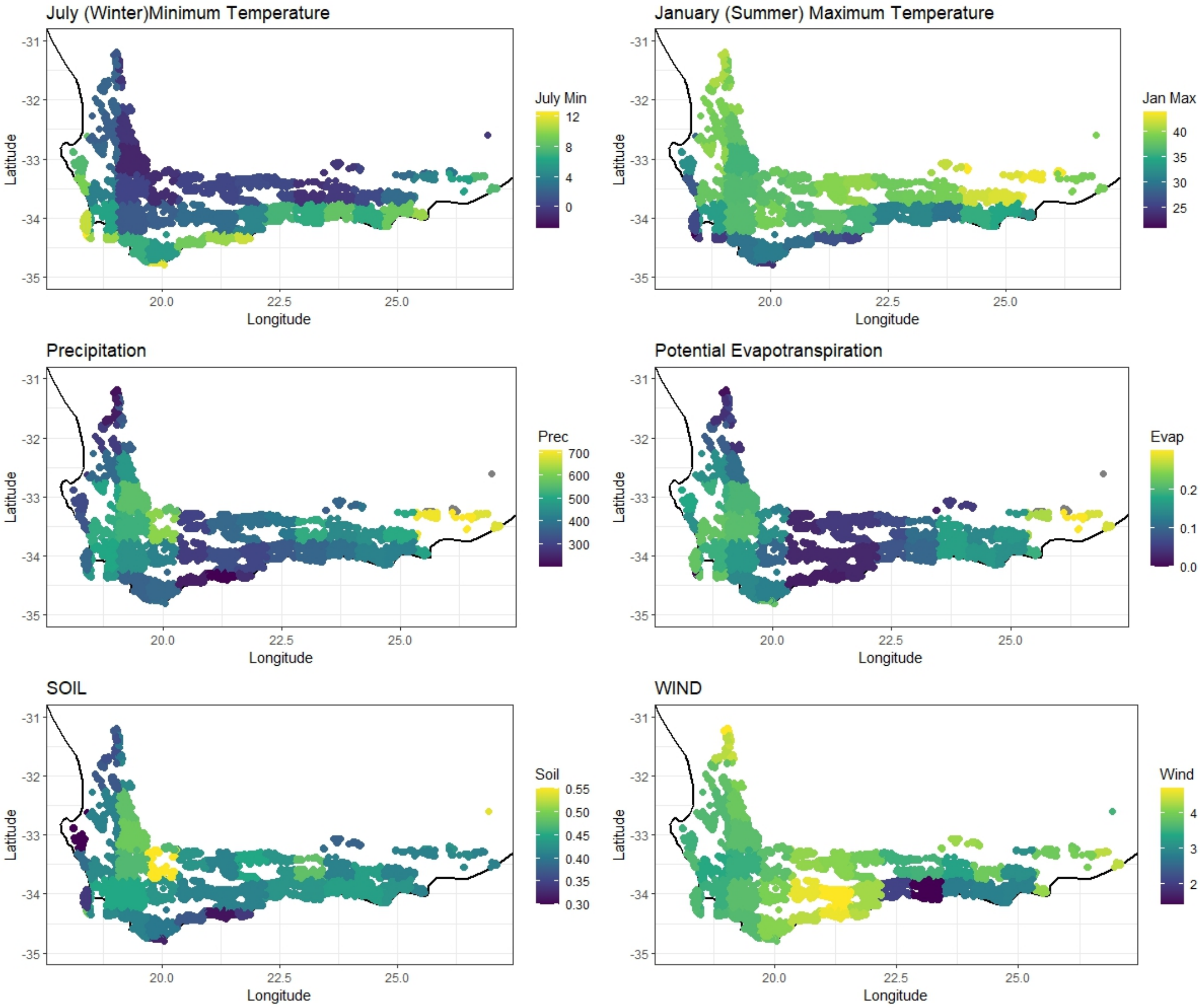


Figure 2: Maps for all six covariate surfaces over the CFR: (a) July (winter) minimum temperature; (b) January (summer) maximum temperature;(c) precipitation corrected sum; (d) evapotranspiration; (e) profile soil moisture days; (f) wind speed at 2m range average

8.1.2 Confidence Intervals of β samples of different covariates for each species

Table 1: **PRAURE, thinning factor of $k = 8$**

	july_min	**jan_max**	**prec**	**evap**	**soil**	**wind**
Mean	-0.1492473	-0.7086169	-0.8594018	1.26415	-0.9400045	-0.9039096
2.50%	-4.197108449	-10.37340532	-45.50533625	-8.518495734	-5.357314949	-6.564557553
97.50%	11.04595699	6.00550603	8.94786305	12.16166976	18.68000318	9.043854999

Table 2: **LDSG, thinning factor of $k = 7$**

	july_min	**jan_max**	**prec**	**evap**	**soil**	**wind**
Mean	1.195057	0.3618308	-0.6519869	1.211469	0.440818	0.7118151
2.50%	-9.10300057	-15.74519807	-9.895043023	-8.567873455	-9.675419566	-6.604494252
97.50%	14.95424781	14.90268721	9.703188152	18.60771161	10.00015327	9.870137981

Table 3: **PRPUNC, thinning factor of $k = 5$**

	july_min	**jan_max**	**prec**	**evap**	**soil**	**wind**
Mean	-1.115739	-1.145595	-1.031668	2.333581	-1.60203	2.097184
2.50%	-10.33510557	-11.21078765	-13.36932432	-8.554846717	-24.76271289	-5.702033227
97.50%	4.166925383	6.177644383	26.52279071	22.96710609	12.98722305	14.45369588

Table 4: **PRREPE, thinning factor of $k = 6$**

	July min	**Jan max**	**Prec**	**Evap**	**Soil**	**Wind**
Mean	-5.069404	-1.39682	-2.922902	8.970779	-0.5127449	-2.433149
2.50%	-33.72290337	-33.71848945	-19.82682536	-10.22328366	-21.33833873	-13.74378858
97.50%	12.48047791	15.59974372	10.60447892	73.30319045	13.16917202	11.28137409

Table 5: **PRMUND, thinning factor of $k = 5$**

	july_min	**jan_max**	**prec**	**evap**	**soil**	**wind**
Mean	-1.806674	-2.214581	3.794876	2.706598	-0.1462244	0.3116929
2.50%	-17.08941656	-20.08839422	-19.57657703	-18.01182115	-9.970441258	-11.83584001
97.50%	15.08535707	22.48554402	23.28997415	21.79371327	4.556744706	13.97009246

Table 6: **PRCYNA, thinning factor of $k = 5$**

	july_min	**jan_max**	**prec**	**evap**	**soil**	**wind**
Mean	13.88582738	8.690377535	3.404857627	-1.119301091	-0.679333401	6.178992329
2.50%	-62.67885427	-35.46306613	-20.7427879	-16.91758291	-24.30314846	-11.41914774
97.50%	67.74685153	49.2761786	27.90916314	27.21928967	11.25716382	48.44382672

Here, to calculate the confidence intervals, we have considered a k^{th} thinning factor, selecting every k^{th} beta sample in the process. The above tables show the posterior mean covariate effects for each species, as well as the respective credible intervals. The direction of significance differs depending on the species. For example, the minimum temperature in July has a negative effect on PRAURE's presence but has a positive effect on LDSG's presence. As a result, we can conclude that LDSG flourishes in cooler environments.

9 Discussion

Until now, using presence-only data, we created a multilayer point pattern model to explain species distribution. This formulation also eliminates the problematic assumptions required to convert the problem to a presence-absence analysis using background samples.

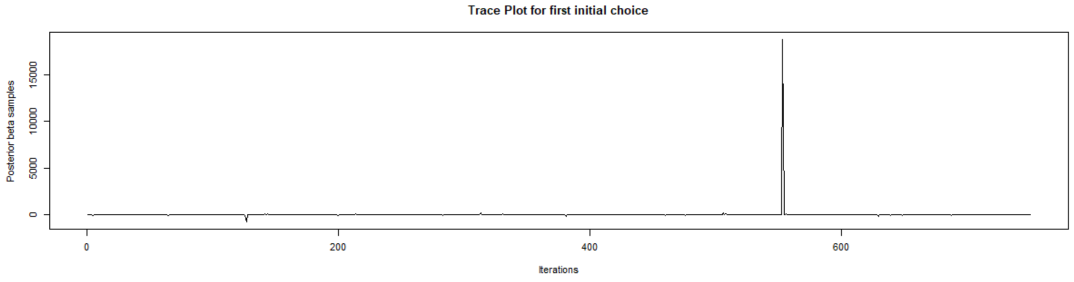
However, our current focus shifts to addressing challenges in our posterior samples. We noticed unusally high peaks in some of our posterior samples for beta and predicted lambda. In response, we initiated hyper parameter tuning and conducted MCMC runs on PRAURE species data with three distinct sets of parameters to enhance the robustness of our analysis.

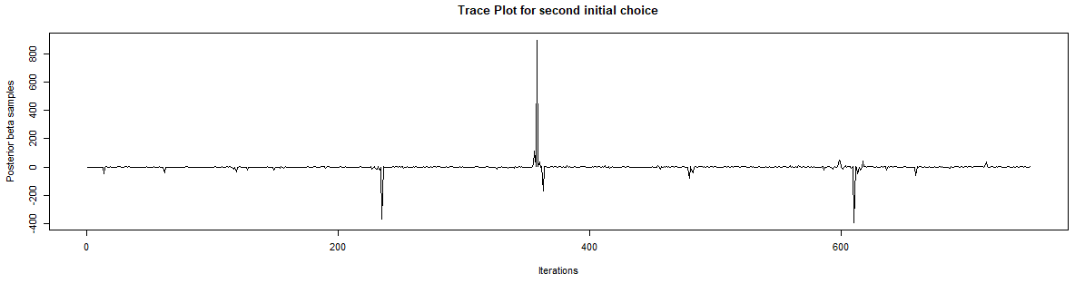
The three different sets are given in the below table along with the corresponding calculated AIC's.

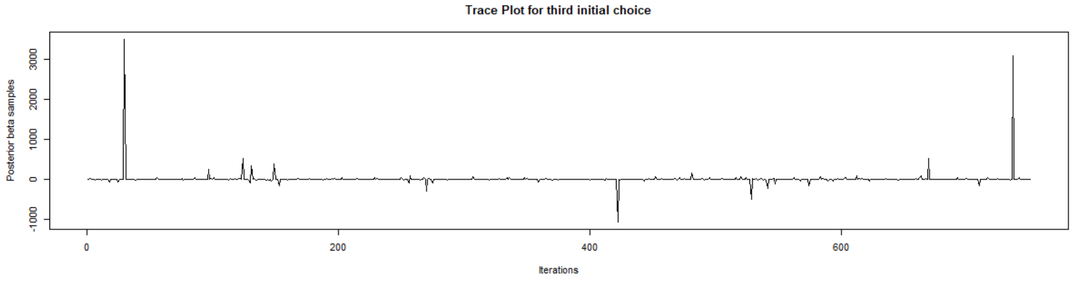
Table 7: **Initial Parameter Choices for different MCMC chains and corresponding AIC Values for PRAURE**

Parameter	**First Initial Choice**	**Second Initial Choice**	**Third Initial Choice**
beta0	`rep(0,7)`	`rep(0,7)`	`rep(1,7)`
Sigma0	`diag(7)`	`diag(7)`	`diag(7)`
a0	0.05	1	1.5
b0	3	1.5	2
phi	0.05	1	1
Sigma2	1.5	0.5	0.05
proposal_mean	0.1	0.5	0.15
proposal_sd	1.5	2	2
beta	`rep(0,7)`	`rep(0,7)`	`rep(1,7)`
lambda	`rep(1,m)`	`rep(1,m)`	`rep(1,m)`
phi0	0	0.5	0.01
phi1	2	1.5	1
phi_current	0.05	1	0.01
phi_proposal_sd	2	1.5	0.05
AIC	1360.778	1457.547	2315.533

Trace plots are commonly used in Bayesian Statistics to assess the convergence of MCMC algorithms, particularly when sampling from the posterior distribution. They help assess whether the MCMC algorithm has converged to the target distribution.
Below are the trace plots for July minimum temperature covariate for corresponding MCMC chains:







Potential Issues

- The MCMC chain might be having difficulty exploring the parameter space effectively. **Poor mixing** might be causing occasional large jumps in parameter values.

- Due to less number of iterations the **burn-in portion is low, which might not be sufficient for the sampler to reach the stationary distribution**. We might be observing spikes due to this.

- The**tuning parameters of the MCMC sampler (e.g., proposal step sizes) might not be well-suited** for the characteristics of the target distribution.

10 Appendix

- The area of the cell i, $\triangle_i$ is the same for all the cells and assumed to be 1, without loss of generality.

- q_i: The probability that a randomly selected location in area A_i was available and sampled is 1 in our case, as we collected presence-only data from the source.

- Our data is structured as follows: The dimension of the species dataset is 15272x9

Table 8: **Presence-data of PRREPE species**

OBS	**Latitude**	**Longitude**	**july_min**	**jan_max**	**prec**	**evap**	**soil**	**wind**
RMC91120502	-33.6375	24.4522	0.54	42.04	453.52	0.14	0.44	3.92
ADA91112702	-33.55	18.4167	9.37	27.08	337.5	0.15	0.39	3.65
TLE91092202	-34.0902	18.4282	11.12	25.9	374.41	0.22	0.34	3.64
PMR91120501	-33.7432	24.2782	0.54	42.04	453.52	0.14	0.44	3.92
PMR91120502	-33.7443	24.2627	0.54	42.04	453.52	0.14	0.44	3.92
PMR91120503	-33.7387	24.2538	0.54	42.04	453.52	0.14	0.44	3.92
PMR91120506	-33.6942	24.2017	0.54	42.04	453.52	0.14	0.44	3.92
AGR91092104	-34.2	19.182	1.57	36.73	458.79	0.16	0.45	3.73
AGR91092107	-34.2	19.1733	1.57	36.73	458.79	0.16	0.45	3.73
AGR91092108	-34.2018	19.17	1.57	36.73	458.79	0.16	0.45	3.73
AGR91092110	-34.2023	19.1637	1.57	36.73	458.79	0.16	0.45	3.73
AGR91092901	-33.7167	18.5458	3.84	37.94	516.8	0.22	0.42	3.4
AGR91093001	-34.331	18.9595	10.17	24.48	474.61	0.22	0.42	3.61
AGR91100601	-34.0595	18.8667	6.06	33.55	506.25	0.19	0.45	3.49

- The predicted λ values are:

Table 9: **Predicted potential intensity surface for PRREPE**

Samples	**V1**	**V2**	**V3**	**V4**	**V5**	**V6**
1	2.245811454	1.169123864	1.130194421	182.030158	0.774834069	0.651254184
2	0.951709393	2.007725494	1.390763587	1.159265907	0.546734255	0.869903865
3	2.091032404	1.516196142	0.46466385	1.690394726	1.696761181	1.400092028
4	0.492872858	0.796613518	0.342565258	0.340053641	0.840293526	0.677347177
5	19427.8976	1.509902958	0.062064392	2.122152334	0.552641377	163368.7065
6	1.480849564	3.421792001	2.999753221	2.534570541	0.715792891	0.95232652
7	0.745458411	2.98348481	0.964274492	2.188436868	5.959888378	0.686150735
8	0.387054351	3.263311286	16.29997107	2.0909646	0.700353934	2.802215669
9	1.318591231	1.075147572	0.482413929	0.90333328	4.345482389	3.233580194
10	1.08518595	1.426704827	2.050022836	1.970185855	4.121311449	0.192470279

The dimension of the table is 1000x4000

11 Acknowledgment

We express our sincere gratitude to our course instructor, Arnab Hazra Sir, for their invaluable guidance and unwavering support throughout this project. Their insightful feedback and encouragement played a pivotal role in shaping the direction of this work. We are truly appreciative of the knowledge and expertise he shared, which significantly enriched our learning experience. His commitment to fostering a positive and challenging academic environment has been instrumental in our growth as a student. His insightful lectures and guidance have been instrumental in shaping the success of this course project.

12 References

- Gaussian predictive process models for large spatial data sets; Banerjee et al 2008
- nearPD R documentation
- Computing the Nearest Correlation Matrix—A problem from Finance

Contributions :

- **Paper Selection :** Vrinda, Snigdha, Rahul
- **Data Collection :** Vrinda, Snigdha, Rahul
- **Code for modelling :** Vrinda, Snigdha
- **Code for plotting :** Rahul
- **Report :** Vrinda, Snigdha