BAYESIAN ADDITIVE REGRESSION TREES FOR BIOLUMINESCENCE

Narmadha Mohankumar*

Department of Statistics, Kansas State University

meenu@ksu.edu

Bioluminescence is the light emitted by a bioluminescent organism, produced by energy released from chemical reactions occurring inside the organism. The main focus in this study is to predict the expected number of bioluminescence sources in each depth in deep sea, using Bayesian regression trees (BART). Data for this analysis consists of number of bioluminescence sources and the depth of the sea it was recorded. BART is mostly used for higher prediction purposes and to determine the distribution of a species where the distribution over a spatial area does not seem continuous. BART is an efficient method to use when we have complex Bayesian hierarchical structures with uncertainty. The BART is a summation of nonparametric regression trees with priors to regularize the parameters in the tree model. Metropolis algorithm in MCMC is used to get the posterior distribution of the parameters where the distribution is unknown and the full conditionals are used for other known parameters. Tree prior was defined using the probability of a split in each terminal node and the probability of assigning splitting rules for each node. BART package in R-Studio was mainly used for this analysis and the results for the expected number of bioluminescence sources were obtained for each spitted groups of the explanatory variable. Here in this analysis, splitting rules are assigned as equally spaced across the explanatory variable. The resulting credible intervals showed an over fit in the model and this may be because the number of nodes in each tree is not controlled.

Keywords: *Bioluminescence , Bayesian Additive Regression Tree, MCMC algorithm, Metropolis algorithm*

BAYESIAN ADDITIVE REGRESSION TREES FOR BIOLUMINESCENCE

Introduction

Bioluminescence is a chemical processes that certain living organisms are able to synthesize and emit light. They are more commonly found in marine environments than in non-marine environments. It is found in many organisms: bacteria, algae, jellyfish, worms, crustaceans, sea stars, fish, and sharks. It is the predominant source of light in the largest fraction of the habitable volume of the earth, the deep ocean. The benefits of marine species co-existing with the bioluminescence are attraction of prey, diversion of predators, and communication. This study was conducted to discover and predict the distribution of bioluminescence sources in deep sea with respect to the depth of the sea using Bayesian additive regression trees. Data for the study was collected from an online journal (Mixed Effects Models and Extensions in Ecology with R; Zuur, Ieno, Walker, Saveliev and Smith. Springer 2009.) and number of sources and the depth of the sea were considered for the analysis. “Bart" package in R-studio is used to conduct the analysis.

Methods

Bayesian additive regression trees (BART) are accepted in many fields of research because they are easy to interpret, more flexible than conventional parametric regression models and have a good predictive power. These models can be used to identify and estimate complex hierarchical relationships in ecology. It is important to find a method to improve predictions and to handle uncertainty in inference. BART model is an efficient way for this and it will lead to accurate and more reliable estimations.

The BART is a summation of nonparametric regression trees and sampling of the posterior distribution is achieved using a Markov chain Monte Carlo (MCMC) algorithm. Unknown full conditionals of parameters are sampled using the Metropolis algorithm, which is a MCMC method. This allows us to obtain a sequence of samples to approximate the unknown posterior distribution.

- **BART Structure**

Single tree model

$Y = g(X; T, \mu) + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$

Sum of tree model

$Y = \sum g(X; T_i, \mu_i) + \epsilon \qquad \epsilon \sim N(0, \sigma^2), i=1,2,\ldots\ldots.,n$

Where,

$Yi | \mu_{i,} \sigma^2 \sim N(\mu_{i,} \sigma^2)$ and $\mu_i = X\beta_i$

n: Number of trees

T_i: Sequence of decision rules

μ_i: Set of parameter values associated with each terminal node ($\mu_{1i}, \mu_{2i}, \ldots. \mu_{bi}$) of T_i

b: Number of terminal nodes in a particular tree T

- **Tree Prior**

The process of growing a tree depends on two main parts, the probability of splitting a terminal node and the probability of assigning a splitting rule to that node.

Probability that a terminal node m is split

$Psplit(m,T) = \alpha(1+d_m)^{-h}$: d_m is the depth of node m

If it splits, then the probability of assigning splitting rule ρ to m

$Prule(\rho|m,T)$

A simple choice is choosing the splitting value uniformly from the observed values X)

- **Prior distributions**

Prior distribution for μ_i

$\mu_i|T_i, \sigma^2 \sim Normal\ (\bar{\mu}, \sigma^2/\sigma^2_\beta)$

Prior distribution for β_i

$\beta_i \sim Normal(\ 0\ , \sigma^2_\beta)$

Prior distribution for σ^2

$\sigma^2 \sim Inverse\ Gamma\ (q, r)$

- **Full Conditionals**

Full Conditional for β_i

$$[\boldsymbol{\beta}| \quad \bullet \quad] \sim \mathrm{N}\left(\left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}\right)^{-1}\mathbf{X}'\mathbf{y}, \sigma^2\left(\mathbf{X}'\mathbf{X} + \frac{1}{\sigma_\beta^2}\mathbf{I}\right)^{-1}\right)$$

Full Conditional for σ^2

$$[\sigma_\varepsilon^2| \quad \bullet \quad] \sim \text{inverse gamma}\left(q + \frac{n}{2}, r + \frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)$$

Metropolis Algorithm is used to get full conditional for Ti. This allows us to obtain a sequence of samples to approximate the unknown posterior distribution of Ti.

"BART" Package is used in R-studio to conduct Bayesian additive trees. Here the splits are determined using values equally spaced across the range of explanatory variable.

Results

Following figures show the obtained predicted values of expected values of the response variable and the credible intervals using BART. Even though the predicted values show a good match with the original distribution, the credible interval shows an over fit.

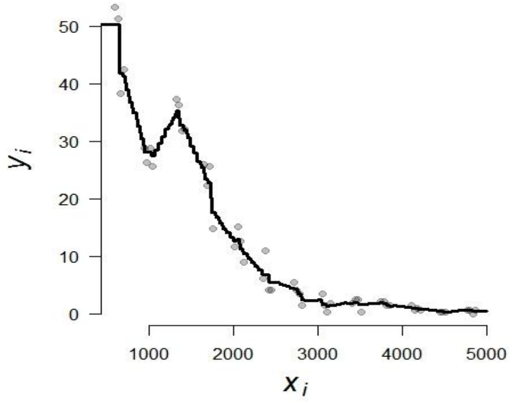

Figure 1: Predicted expected values of Y

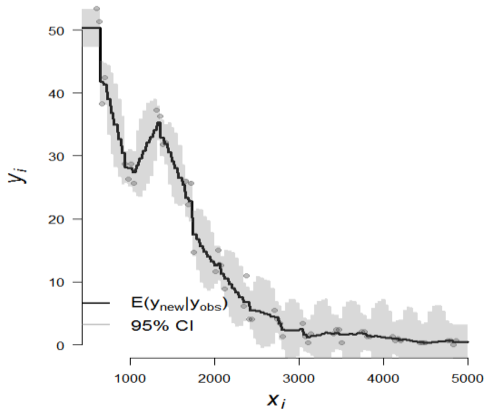

Figure 2: Credible interval

The following figures show the output for multiple numbers of splits in the Bayesian regression tree.

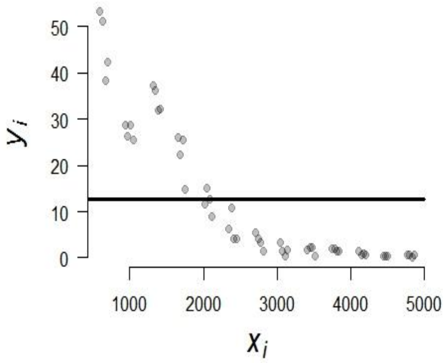

Figure 3: One split in the BART

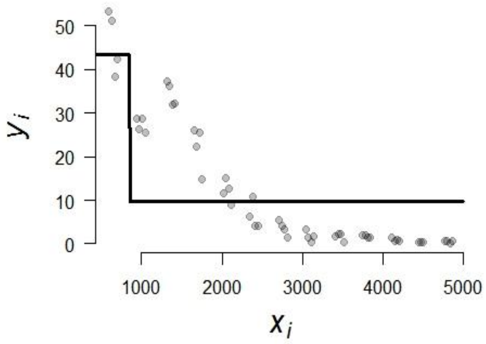

Figure 4: Two split in the BART

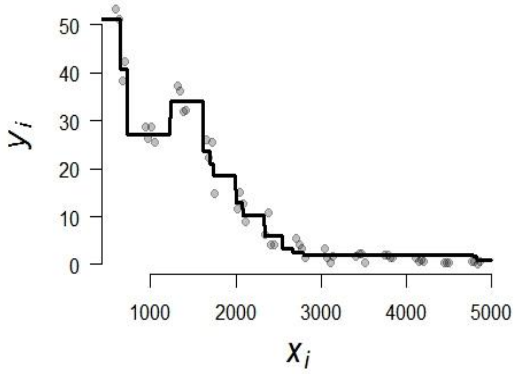

Figure 5: Multiple splits in the BART

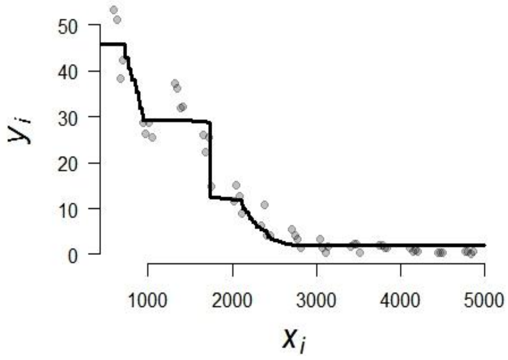

Figure 6: Multiple splits in the BART

Discussion

The results obtained showed an over fit and this can be solved by restricting the number of terminal nodes. This may allow us to obtain the number of terminal nodes that gives the maximum predictive score, so that the problem of over fitting can be solved. Moreover, this method can be improved by using multiple splitting rules for nodes rather than using equally spaced criteria over the explanatory variable. Embedding Bayesian additive regression trees in Bayesian hierarchical model (such as the occupancy model), may also lead to more efficient results.

References

- Mixed Effects Models and Extensions in Ecology with R (2009). Zuur, Ieno, Walker, Saveliev and Smith. Springer. (n.d.).
- Bayesian Additive Regression Trees Hugh A. Chipman, Edward I. George, Robert E. McCulloch. (n.d.).
- Bayesian CART Model Search Hugh A. Chipman , Edward I. George & Robert E. McCulloch. (n.d.).