# Nonlinear Regression with Outcome Space Bounded in [a,b]

**Pamela M. Chiroque-Solano**, Thomas Jaki

pamela.chiroque-solano@ur.de Faculty für Informatik und Data Science, Universität Regensburg

Universität Regensburg

Women in Data Science Worldwide | Regensburg

## 1. Overview

A problem in Statistics that receives less attention than it deserves is constrained data. Even less so when the observed dataset includes values in the extremes. Say, for example, if $Y_i$ are the observations, then

$$Y_i \in [a, b].$$

Typically, such data is modeled with a (rescaled) hurdle beta regression. However, this approach forces the use of a linear predictor for the (transformed) expected value.

## 2. Framework

The BART approach [Chipman et al., 2010] is a flexible non-parametric regression model. It uses a tree ensemble Machine Learning framework while maintaining desirable Bayesian probabilistic advantages.

While an implementation is available for positive constrained data with a 0 hurdle (Linero, Sinha and Lipsitz, 2019), a solution for an upper bounded support is still missing.

The problem is delicate, since the BART approach requires the trees endpoints to be integrated out analytically when updating the trees themselves. That is to say,

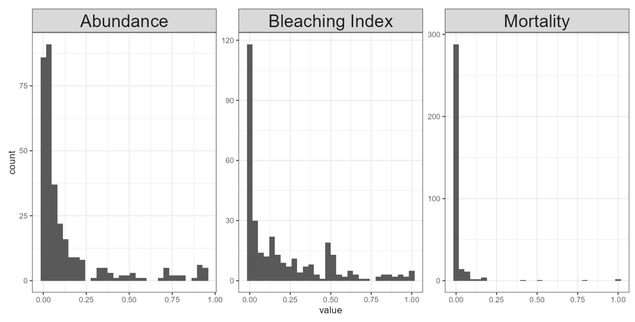
$$\pi(\mathcal{T} \mid \cdot) \propto \int_{\mathcal{M}} \pi(\mathcal{T}, \mu \mid \cdot) d\mu, \tag{1}$$

where $\mu$ are the endpoints, needs to be available in closed form.

## 3. Motivating data

This kind of data is commonly observed in fields like environmental, media, and social research, where many factors contribute to complexity. For instance in marine biology, coral bleaching, mortality, abundance or coverage of coral colonies may include zero and one in their proportions, that is, the data is in the $[0, 1]$ domain.

**Figure: 3 datasets where 0 and 1 are observed in the proportions.**




The middle dataset, Bleaching Index, will be used to exemplify the approach.

## 4. Benchmark

A non-ideal solution is to transform the data so that the typical, already established, continuous and unbounded BART approach can be applied. This is frequently done, however, it is advantageous to avoid transformations, so that the established results will be applied on the data in their original scale. Nonetheless we will use these approaches as benchmark with which to compare our approach.

## 5. Work in progress

The present work emphasizes adapting the BART non-parametric approach to constrained data with hurdle on both extremes. The intended approach is to model $Y_i \in [0, 1]$

$$\begin{aligned} Pr(Y_i = 0) &= \text{probit}(BART_0(x)) && (2)\\ Pr(Y_i = 1 \mid Y_i > 0) &= \text{probit}(BART_1(x)) && (3)\\ Y_i \mid Y_i \in (0,1) &\sim Beta(\eta_i, \cdot);\ g(\eta_i) = BART_{01}(x). && (4) \end{aligned}$$

For the model in Equation (4), we expect that the integral in Equation (1) to be available in closed form. Also, it is parameterized so that

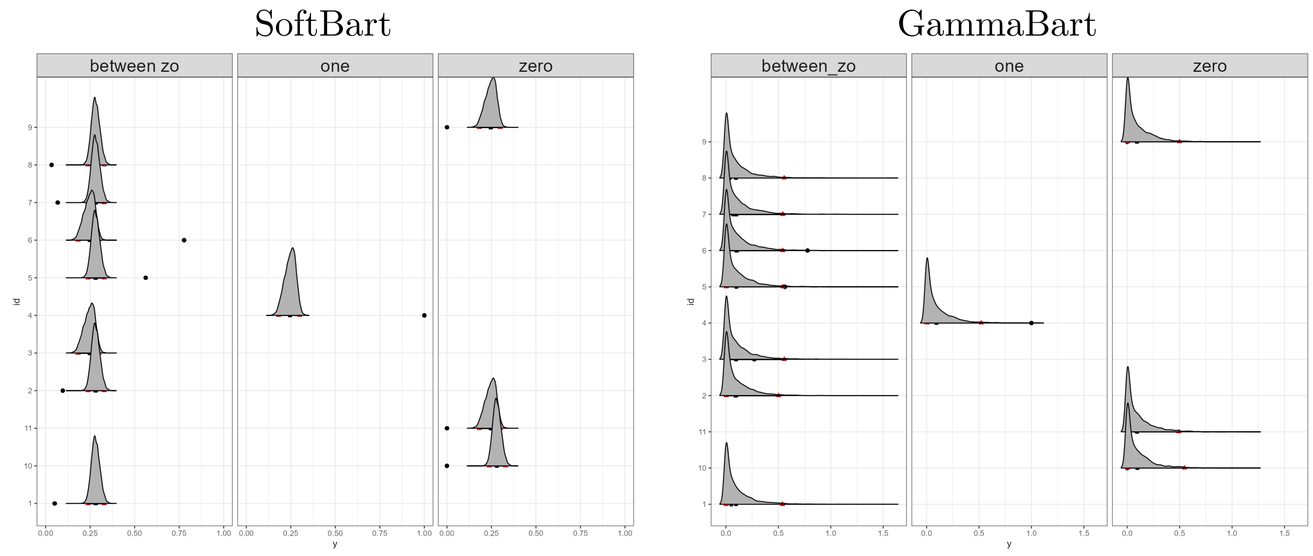
$$E[Y_i \mid Y_i \in (0,1)] = \frac{\exp\{BART(x)\}}{1 + \exp\{BART(x)\}} \tag{5}$$

The SharedBart approach of Linero, Sinha and Lipsitz (2019) will also be used to link the tree structures of the continuous data and the hurdle components.

## 6. Benchmark results

The results using existing packages are less than desirable. One can notice that prediction (marginal) densities fail to encompass the data, which represents a lack of fit.

**Figure: Prediction interval for some of the observed data. Dots represent the true observation.**




This is accentuated by the Root Mean Square Error (RMSE) metric. The GammaBart package [Linero et al., 2019] has a RMSE of 0.28, while for the SoftBart [Li et al., 2022] this is 0.25. The latter is equivalent to the result obtained by Kapelner and Bleich [2016]. Compared to the raw Standard Deviation for the data of 0.26, this is not much of an improvement.

## 7. Discussion

The fit using the transformed data is clearly lackluster. It has already been established that modeling untransformed data yields better results. As such, we are confident that our approach will bring a better fit and prediction, leading to a worthwhile model.

## 8. References


### References

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), March 2010. ISSN 1932-6157. doi: 10.1214/09-aoas285. URL http://dx.doi.org/10.1214/09-AOAS285.

Adam Kapelner and Justin Bleich. bartmachine: Machine learning with bayesian additive regression trees. *Journal of Statistical Software*, 70(4), 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i04. URL http://dx.doi.org/10.18637/jss.v070.i04.

Yinpu Li, Antonio R. Linero, and Jared Murray. Adaptive conditional distribution estimation with bayesian decision tree ensembles. *Journal of the American Statistical Association*, 118 (543):2129–2142, March 2022. ISSN 1537-274X. doi: 10.1080/01621459.2022.2037431. URL http://dx.doi.org/10.1080/01621459.2022.2037431.

Antonio R. Linero, Debajyoti Sinha, and Stuart R. Lipsitz. Semiparametric mixed-scale models using shared bayesian forests. *Biometrics*, 76(1):131–144, November 2019. ISSN 1541-0420. doi: 10.1111/biom.13107. URL http://dx.doi.org/10.1111/biom.13107.