Great Lakes Coastal Wetlands Data Analysis using Generalized Additive Models for Location, Scale and Shape (GAMLSS) to Measure Percent Coverage

Juan M. Cánovas, Jimmy G. Moore and Jasvinder Singh

December 13, 2018

Abstract

Generalized Additive Models for Location Shape and Scale (GAMLSS) have been growing in popularity over the past decade. This work showcases a methodology for modeling proportional vegetation coverage data in the context of GAMLSS models. More specifically, a unique approach to modeling overdispersed data that follows a zero-inflated beta distribution (BEZI) is explained. Data of this structure is very common in the field of ecology. Exploratory analysis, model selection, diagnostic procedures, and inference are all discussed and related back to the larger context of ecology and environmental treatment.

Keywords: GAMLSS, vegetation, zero-inflation, overdispersion, beta distribution

Introduction

Statistical modeling is used to obtain a deeper understanding of a particular phenomena of interest. To begin the statistical modeling process we define a response variable which measures the phenomena of interest and set this response to be a function of explanatory variables. In order to maximize the power of our model we can make distributional assumptions about our response variable. Distributional assumptions rely heavily on the empirical nature of the data as well as the motivation of the experimenter. It is common in ecological situations for researchers to model data using the Poisson, negative binomial, Bernoulli, and many other classical parametric distributions. One such classical distribution is the beta distribution. Models using the beta distribution are very versatile and allow for a wide variety of uncertainties to be usefully modelled. Their flexibility encourages

its empirical use in a wide range of applications [1]. More specifically, beta distributions are useful when modeling proportional data. In the context of ecology this could mean measuring forest canopy cover estimation [2], or used as an evaluation metric of the proportion of land containing invasive species in coastal wetlands (Team Typha).

The beta distribution is a 2 parameter distribution used to define a random variable over the range $0 < y < 1$. This distribution has 2 parameters for shape and has the following probability density function (pdf):

$$f_Y(y|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1} \quad ; \quad 0 < y < 1$$

The parameters α and β are the shape parameters that characterize this distribution and are both greater than 0. Because these parameters are only constrained by a lower bound greater than 0, beta models exhibit a great deal of flexibility. Additionally, because it is used for random variables between 0 and 1, beta models are appropriate when our random variable of interest is a proportion. However, it should be noted that the beta distribution only handles variables between 0 and 1 and cannot model random variables that take on the value of exactly 0 or 1. In practice, this limitation has many drawbacks.

In ecological research it is common to observe zeros in data. As a result there is increasing interest in methods to properly model data that contains more zeros than expected. Datasets that contain more zeros than expected are said to be zero-inflated. In this case the observed probability density/mass functions at zero exceed its theoretical value [3]. There are many methods proposed to handling zero inflation. Martin et al.[4] provide a framework for understanding how zero-inflated datasets originate and bring attention to methods for handling zero inflation to help ecologist navigate data in these situations. The paper focuses primarily on the use of zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), and zero-inflated Bernoulli (ZIB) models and demonstrates that failing to properly account for zero inflation can lead to substantially different parameter and precision estimates.

The underlying issue of zero inflation is that it introduces a phenomenon known as overdispersion. Overdispersion occurs when the observed variance is higher than what is theoretically expected under a given set of distributional assumptions. Because beta distributions cannot handle

zero observations, a common procedure is to transform zero observations to be non-zero [5]. However, this is a large problem because when such transformations occur the high frequency of zeros is simply reformatted to be an equally high frequency of the value to which 0 is transformed [6]. Therefore the underlying issue of overdispersion is not properly accounted for. Not accounting for the overdispersion can lead to underestimating variance and thus overconfidence in a given model [4]. Such overconfidence can lead to increased Type I and Type II error rates. Additionally, transforming zero observations creates a loss of information. Thus, a beta model that is capable of handling zero inflation is desired in order to preserve the integrity of the data as well as minimize the risk of Type I and Type II error.

Ospina and Ferrari [7] discuss in detail the intricacies of inflated beta models and derive the Likelihood and Moment Generating function for zero-inflated beta (BEZI) Distributions. BEZI distributions are very similar to standard beta distributions. The only difference is the addition of a third parameter, p_0, which is interpreted as the proportion of zeros in our sample. As a result, the modified pdf for BEZI distributions is as follows:

$$f_Y(y|\alpha,\beta,p_0) = \begin{cases} p_0 & y = 0 \\ (1-p_0)\frac{1}{B(\alpha,\beta)}y^{\alpha-1}(1-y)^{\beta-1} & 0 < y < 1 \end{cases}$$

To showcase the application of BEZI models in practice Ospina and Ferrari [7] consider three examples. These examples include modeling the percentage of qualified nurses in a given region, proportion of infant deaths from unknown causes, and proportion of inhabitants who lived within a given threshold of coastal areas. In each example, the BEZI distribution was fit with a Generalized Additive Model for Location Scale and Shape (GAMLSS). Recently, GAMLSS models have been increasing in popularity as a semi-parametric way of modeling all parameters of the underlying distribution as a function of the explanatory variables in a given dataset. Because of their growing popularity, we next discuss their origin and briefly explain their statistical nature.

Originally introduced by Rigby and Stasinopoulos [8] [9], and Akantziliotou, Rigby, and Stasinopoulos [10], GAMLSS models serve to overcome some of the shortcomings of Generalized Linear Models (GLM) [11] by relaxing the exponential family assumption for the response variable, and replacing this assumption to be a general distribution family that encompasses more diverse non-exponential family forms such as leptokurtic, platykurtic, skewed, or overdispersed

distributions [12]. The intuitive idea of GAMLSS models is that they are distributional based semi-parametric regression models. They are regression type models in that they use explanatory variables to model a single response, distributional in that a full parametric distribution assumption is made for the response variables, and semi-parametric in that all parameters of the distribution are modeled as a function of explanatory variables using both parametric and non-parametric smoothing functions. The term "all parameters" is in reference to the four statistical moments of mean, variance, skewness, and kurtosis (μ, σ, ν, τ). As such, the basic form of a GAMLSS model is as follows:

$$g_1(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} \mathbf{Z}_{j1}\gamma_{j1} \tag{1}$$

$$g_2(\boldsymbol{\sigma}) = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=2}^{J_2} \mathbf{Z}_{j2}\gamma_{j2} \tag{2}$$

$$g_3(\boldsymbol{\nu}) = \mathbf{X}_3\boldsymbol{\beta}_3 + \sum_{j=3}^{J_3} \mathbf{Z}_{j3}\gamma_{j3} \tag{3}$$

$$g_4(\boldsymbol{\tau}) = \mathbf{X}_4\boldsymbol{\beta}_4 + \sum_{j=4}^{J_4} \mathbf{Z}_{j4}\gamma_{j4} \tag{4}$$

Where:

$\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$ are vectors of lenght n

$\mathbf{X}_k$ is a fixed known design matrix of order $n \times J_k^{'}$

$\beta_k^T = (\beta_{1k}, \beta_{2k}, \ldots, \beta_{J_k^{'}k})$ is a parameter vector of length $J_k^{'}$

$\mathbf{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix

γ_{jk} is a q_{jk} dimensional random variable which is assumed to be distributed as $\gamma_{jk} \sim N_{qjk}(\mathbf{0}, \mathbf{G}_{jk}^{-1})$ where $\mathbf{G}_{jk}^{-1}$ is the generalized inverse of a $q_{jk} \times q_{jk}$ symmetric matrix $\mathbf{G}_{jk} = \mathbf{G}_{jk}(\boldsymbol{\lambda}_{jk})$ which may depend on a vector of hyperparameters $\boldsymbol{\lambda}_{jk}$

It should be noted that GAMLSS models are a generalization of GLM and as such, each model parameter utilizes a specific link function to map the explanatory variables to the response. Using link functions guarantees that parameter estimates remain within the appropriate range of the response variable. The specific link functions used are conditional on the distributional assumptions of the response. Rigby and Stasinopolous [13] list all of the distributions compatible to GAMLSS

models and the respective link functions for each parameter of these distributions.

GAMLSS models are an attractive way method to model BEZI models for a number of reasons. The first being their ability to handle non-exponential family forms. When we analyze the Fisher Information matrix of beta (and BEZI) distributions it can be shown that the α and β parameters have issues with orthogonality, which contradicts the requirements of a GLM [14]. Therefore, such methods should not be used and alternative methods should be considered. Because GAMLSS models are merely a generalization of GLM's and GAM's [13], we can relax the requirement of orthogonality and adequately model response variables following a BEZI distribution.

A motivating example of BEZI based GAMLSS models being used in practice can be seen in a study conducted by Korhonen, Ali-Sisto, and Tokola [2] in which optical satellite images were used to calculate the proportional measures of tropical forest canopy coverage for a plot in the country of Laos.

In this present study our goal is to showcase the previously discussed concepts of zero-inflated beta distributions and GAMLSS models as a way to model the proportion of vegetation coverage in coastal wetlands as a function of multiple covariates, treatment methods, and random effects. Such model was developed using R-Statistical Software and the external library `gamlss` developed by Stasinopoulos and Rigby [12].

Research Objectives

Team Typha, a wetland research group from Loyola University Chicago, have a rich history of studying coastal wetlands using approaches that integrate components biogeochemistry, plant ecology and community ecology. One of the objectives of their research is to study the ecology and management of the invasive species *Typha x Glauca* in coastal wetlands of North America. This species has invaded the wetlands and is disrupting the flora and fauna within the ecosystem as seen in Figure 1. Ultimately the goal of this research is to see what kinds of land management techniques have the most effect on reducing *Typha x Glauca* and improving natural vegetation in these regions. This project was assigned to us by Dr. Brian Ohsowski, a member of Team Typha, with the goal of using statistical models to determine the effects of certain factors on the percentage of *Typha x Glauca* in plots in coastal wetlands.

Figure 1: Invasive Typha

Data

The data for this project was collected in a three-year span from 2011-2013 across two wetland sites in Michigan: Cedarville and Munuscong. The data collected consisted of 72 observations with 36 for each site respectively. There were 12 variables, 7 of which were numerical and 5 that were categorical. 6 of the variables are proportions that represented the percentage of the plot covered or uncovered by vegetation. To be concise, the scope of this paper is limited to model percent Typha coverage as our only response variable. The predictor variables consisted of various land management techniques (treatment effects), water depth, year, and site.

The first predictor, treatment effects, was categorical with 4 levels; control, above, below, and mow. For the control treatment nothing was done to the plot containing *Typha x Glauca*. For the above treatment Typha was cut above the sediment surface and biomass was removed. In the below treatment, Typha was cut below the sediment surface and biomass was removed. The mow treatment was when Typha was cut above the water line and biomass was left inside the plot.

Additionally, water depth was a numerical variable that measured the depth of wetland water for a given plot in centimeters. Year was categorical and represented 2011, 2012 or 2013 which correspond to the three increments in which data was collected. Site was categorical as well and

was either Cedarville or Munuscong. The goal of our analysis was to evaluate the efficacy our treatment methods on the proportion of Typha coverage.

Upon receiving the data, we conducted exploratory data analysis to improve our understanding of the data. Figure 2 shows the distribution of Typha in the form of a histogram. The x-axis ranges from 0-1 because our response variable is a percentage. As a result, modeling our response variable under the assumptions of a beta distribution seem reasonable. However, one can visually observe a high amount of zeros in this plot. Because of this we decided it appropriate to extend the distributional assumptions of our response to follow a BEZI distribution.

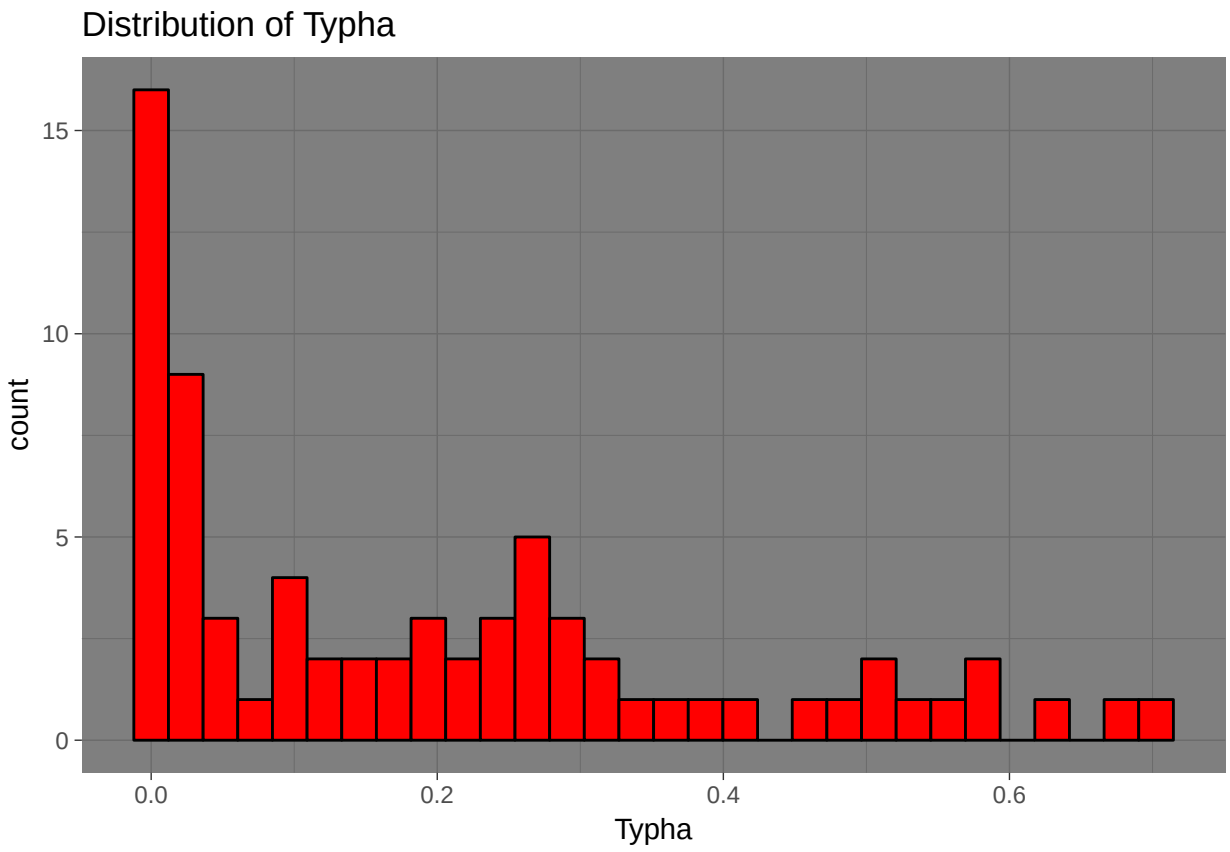

Figure 2: Distribution of Typha

Figures 3a and 3b are boxplots showing the proportion of Typha coverage as functions of each treatment method at the two sites, Cedarville and Munuscong. Figure 3a only considers data from Cedarville. By observing this plot, one can see that the control group had the highest mean Typha vegetation percentage when compared to the other treatments, as indicated by the white dot on the

plot. This plot also give slight insight into the variance of each treatment. It can be observed that for Cedarville, the below treatment had the greatest variation in Typha percentage.

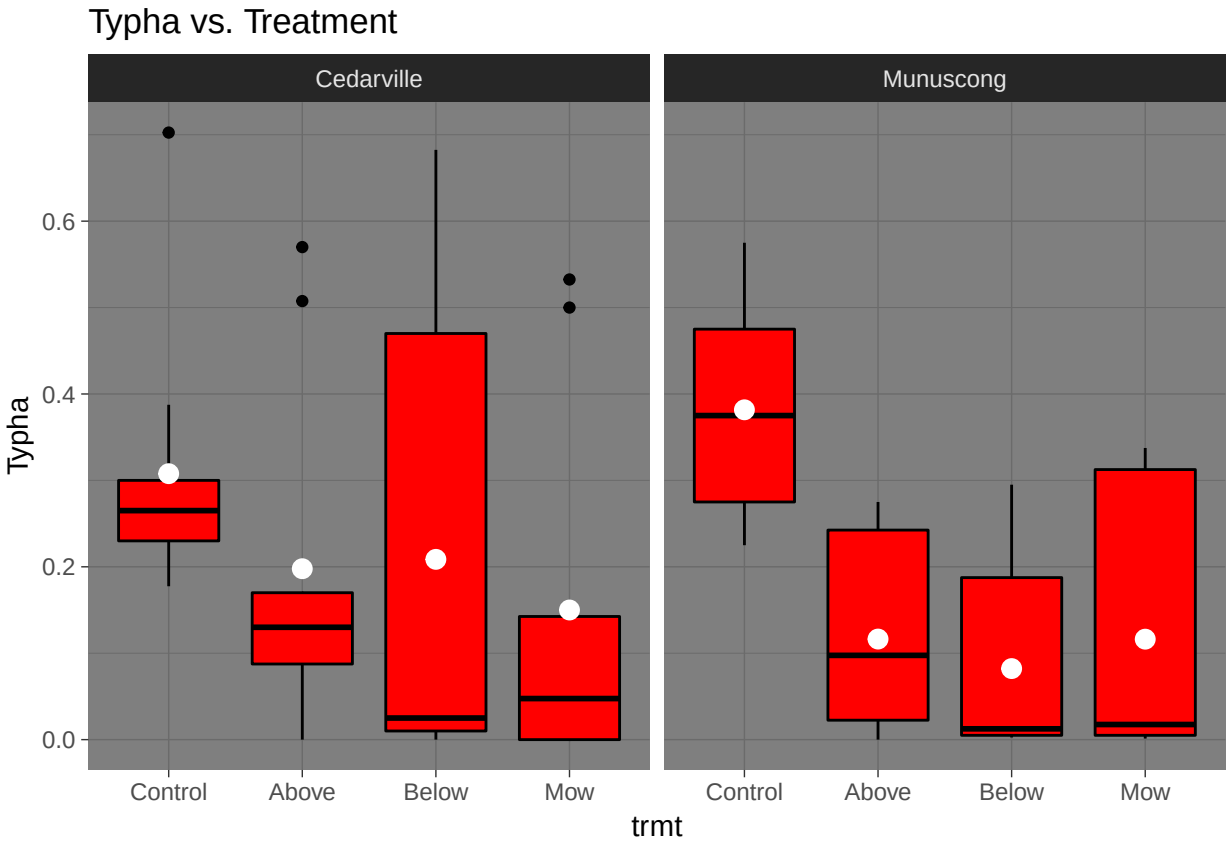

Figure 3: Boxplot representation of Typha coverage by treatment and site

Figure 3b only considers data from Munuscong. By observing the plot it can again be seen that the control treatment had a smaller effect on Typha reduction when compared to the other treatments. Again, one can see that the mean Typha vegetation percentages for control differs from the other 3, however the means for the other 3 treatments do not differ as much.

Although there does not appear to be much significant difference in mean between the treatments of above, below, and mow we were also interested in dispersion effects. Figure 3 is an illustration of how the different types of treatment result in different variation in the Typha coverage percentage. This can be seen by the varying inter quartile ranges for each treatment effect. Using this insight we consider the potential dispersion effects when building our final model and conducting inference on our data.

Methods

After our exploratory data analysis, we shifted our focus to modeling. Initially, it was suggested that a Generalized Linear Mixed Model (GLMM) can be used. Because GLMM's are an extension of linear mixed models with a relaxed normality assumption, and our BEZI response variable is not normal, this method seemed like a good approach. We first attempted to fit a GLMM using the `glmmTMB` package in R. This process required us to transform our data to remove zero observations. Once the zeros were transformed we were able to fit a GLMM and analyze the residuals. However, as previously explained in the introduction, our transformation did not solve the underlying issue of overdispersion and thus we were decided to examine alternative approaches.

The alternative approach we settled on was to fit our data with a GAMLSS model. As previously described, each parameter representing the four statistical moments for the response variable is modeled as a function of the predictors. Therefore, in our final model we assumed our response variable Typha to be from a BEZI distribution and executed manual variable selection using AIC criterion to determine which predictors would be used to model μ, σ, and ν. τ was not considered in our model. As a starting point we considered predictors treatment, site, water depth, year and used the R function `drop1()` to determine which variables to include. The respective link functions for each parameter in our final model were logit for both μ and σ, and log for ν. Upon building our final model we conducted inference on our data.

Results

As initially suggested by our client, we fit a GLMM to model the behavior of Typha given a set of covariates. One of the problems with this approach is that GLMMs cannot handle responses that come from a zero-inflated beta distribution. In order to avoid this issue we decided a transformation of our response is the best approach to try to convert the zero-inflated values to values more appropriate and suitable to fit the assumptions of this technique. So we transformed the response using the formula $y' = \frac{y*(N-1)+0.5}{N}$ as used by Smithson and Verkuilen [5]. In this formula y is the original response, N is the sample size, and y' is the transformed response. After the transformation of the response, we fit the model and to determine how well the model fits the data, we proceeded

with some visual residual diagnostics. Figure 4 is a plot of the residuals vs. index for our fitted GLMM model.

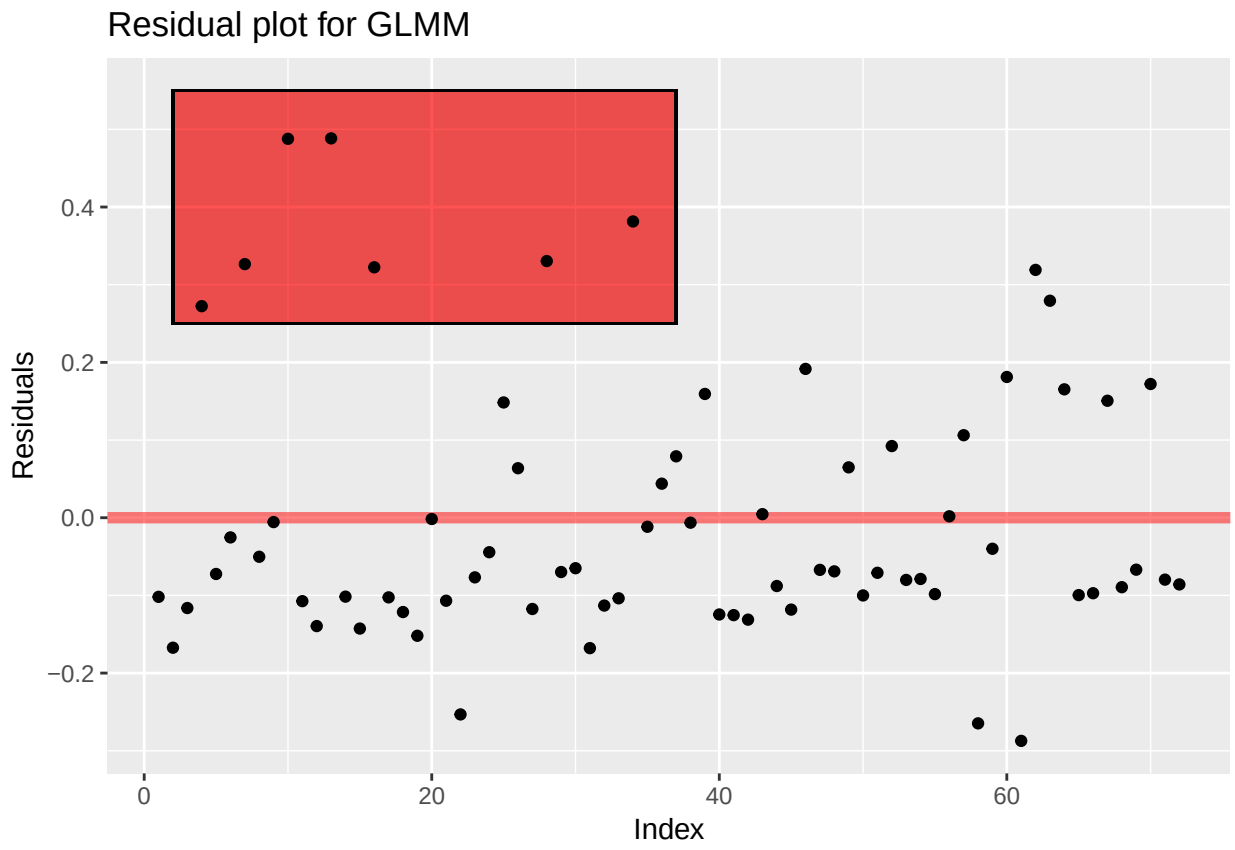

Figure 4: Residual plot for GLMM

The residual plot indicates that our assumptions of homoscedasticity and independence are upheld. However, there appears to be outliers. These outliers are highlighted in the transparent red box in Figure 4. Because of the high number of outliers and the loss of information we obtained from transforming the zero observations, we conclude that the GLMM model was not an adequate approach to modeling our data. Thus, we continued on with GAMLSS as none of these issues arised with that technique.

Using the R package `gamlss` we fit a GAMLSS model. Our variable selection method yielded the following results. Location (μ) was modeled as a function of treatment, water depth and year. Scale (σ) was modeled as a function of treatment. Shape (ν) was modeled as intercept only. The diagnostic residual plots are contained in Figure 5. The plot relating residuals on fitted values,

shows that the residuals uphold independence. Additionally, the residuals on index plot shows that our homoscedasticity assumption is upheld. Finally, the last two plots in Figure 5 show that our assumption of normality of the residuals is preserved. Therefore, we conclude that the model does not violate our residual assumptions and is a good fit for the data.

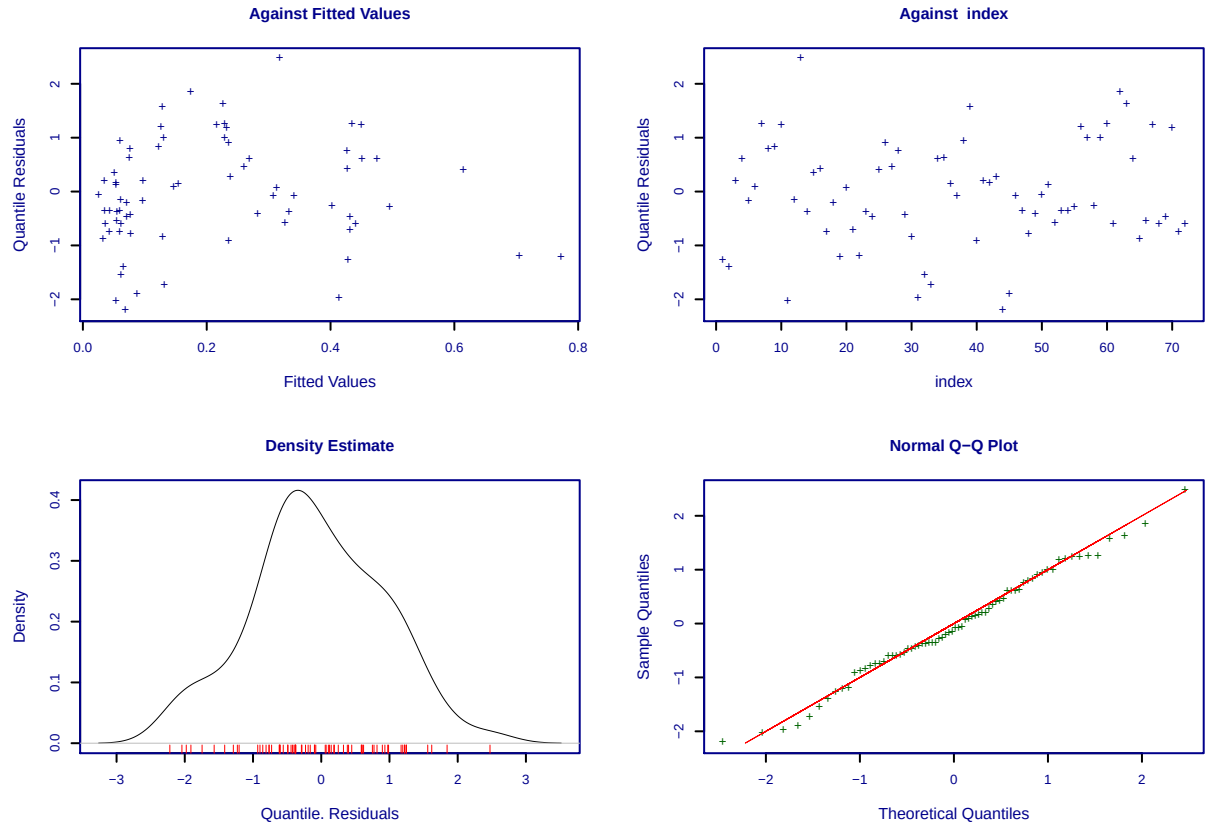

Figure 5: Residual plot for GAMLSS

Post-hoc analysis compared different levels of treatments. Figure 6 shows Tukey Honest Significant Difference (HSD) comparisons for the four treatment levels for each year based on our final model. The blue bar indicates the confidence interval for the mean of that specific treatment. The red arrow is for the comparison range of the means via Tukey HSD. If the red arrows between treatments overlap, then there is no significant difference between the treatments. For each year there is a significant difference in Typha vegetation coverage between the control vs. below, above and mow. However, there is not a significant difference in treatment for above, below and mow.

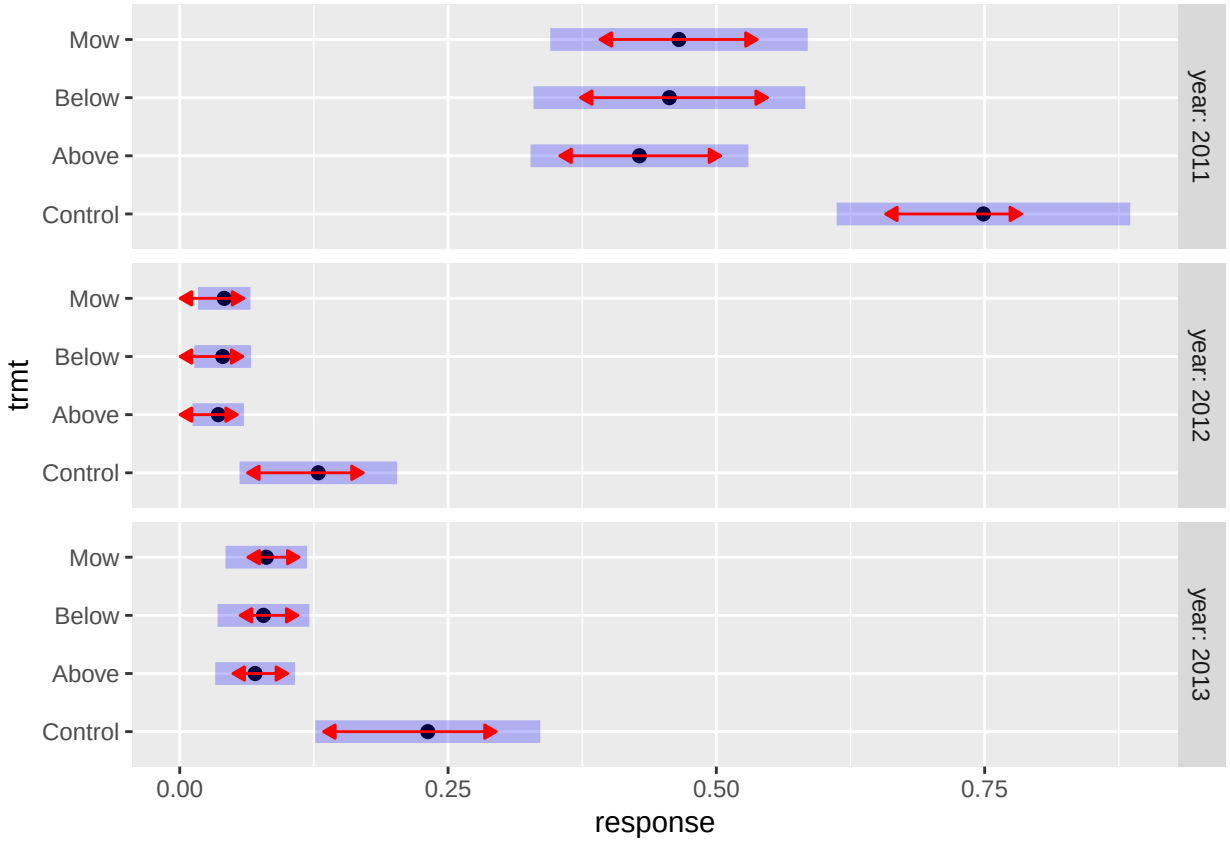

Figure 6: Post-hoc comparisons

Finally, the model coefficients for each parameter are analyzed in detail in Figure 7. Recall, that in order to fully understand the coefficients as they relate to their respective parameters, they must be transformed via the link function.

```
## ******************************************************************
## Family:  c("BEINF0", "Beta Inflated zero")
##
## Call:  gamlss(formula = Typha ~ factor(trmt) + waterDepth +
##     factor(year), sigma.formula = ~factor(trmt), family = BEINF0(mu.link = "logit",
##     sigma.link = "logit", nu.link = "log"), data = dat,
##     trace = FALSE)
##
## Fitting method: RS()
```

```
##
## ------------------------------------------------------------------
## Mu link function:  logit
## Mu Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.4111     0.5611   4.297 6.45e-05 ***
## factor(trmt)Above  -1.3810     0.3319  -4.161 0.000103 ***
## factor(trmt)Below  -1.2679     0.3466  -3.658 0.000537 ***
## factor(trmt)Mow    -1.2319     0.3182  -3.872 0.000269 ***
## waterDepth         -5.2217     1.1170  -4.675 1.71e-05 ***
## factor(year)2012   -3.0010     0.4366  -6.873 4.11e-09 ***
## factor(year)2013   -2.2942     0.3745  -6.126 7.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Sigma link function:  logit
## Sigma Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -0.2401     0.2367  -1.014  0.31462
## factor(trmt)Above  -1.0064     0.3162  -3.183  0.00231 **
## factor(trmt)Below  -0.6299     0.3992  -1.578  0.11985
## factor(trmt)Mow    -1.0339     0.4470  -2.313  0.02418 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ------------------------------------------------------------------
## Nu link function:  log
## Nu Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -2.2285     0.3978  -5.602 5.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## ----------------------------------------------------------------
## No. of observations in the fit:  72
## Degrees of Freedom for the fit:  12
##       Residual Deg. of Freedom:  60
##                       at cycle:  15
##
## Global Deviance:     -107.3404
##             AIC:     -83.34042
##             SBC:     -56.02042
## ****************************************************************
```

For example, the estimate for ν parameter is -2.2285. When we exponentiate this value (because of our log link function) we observe a value of 0.1077 which is approximately the proportion of zero observations in our data set (0.0972). By reversing the logit transformation for the sigma coefficients we see that the variance of each treatment are 0.44 for control, 0.26 for above, 0.34 for below, and 0.26 for mow. Indicating the mowing yields the smallest variability in our response variable.

Discussion

The scope of this project was to model the proportion of vegetation coverage in coastal wetlands dependent on a set of covariates, treatments and random effects. Since our target variable was a proportion of Typha vegetation coverage in a plot, we assumed it to follow a beta distribution. After exploring the data further, we noticed that the response was zero-inflated, so we extended our distributional assumption of the response to follow a BEZI distribution. Initially, our approach was to build a GLMM to try to explain the variation in the proportion of Typha plot coverage. This approach seemed ideal in the sense that GLMMs are an extension of linear mixed models with a relaxed normality assumption, and our BEZI response variable is not normal.

However, after fitting the GLMM we encountered issues related to zero-inflation and overdispersion. Thus, we deemed our GLMM model inadequate. Next, we transitioned into a more competent model type known as GAMLSS. This model was an improvement from the GLMM in the sense that they are able to handle non-exponential family forms. One intriguing element of the GAMLSS model is that four models are created, each relating to the respective statistical moments: one model for the location (μ), one model for the scale (σ), and two models for the shape (ν and τ).

Upon fitting the GAMLSS model, residual diagnostics indicated that the GAMLSS model was an improvement from the GLMM. This claim is further supported by AIC criterion. Post-hoc test comparisons determined that there was no difference in Typha vegetation coverage for the above, below, and mow treatment. However, each of these three treatments had a significant difference from the control. It should also be noted that GAMLSS provide information about the variation characteristics of each treatment. The results determined the mow treatment yielded the lowest variability of all treatment methods. This insight can be paired with domain expertise to determine which treat may be most eco-friendly or economical.

The intention of this paper is to bring attention to the richness of GAMLSS models as they apply to proportional metrics of vegetation coverage. GAMLSS models are very powerful and the limited scope of this project did not include all capabilities. We did not include any non-parametric smoothing, nor did we include random effects. Both of these are tasks are possible in the context of GAMLSS models. For example, in future work it may be beneficial to include plot as a random effect and include interaction terms.

Acknowledgements

We would like to acknowledge Dr. Swarnali Banerjee of the Mathematics and Statistics Department at Loyola University Chicago and Dr. Brian Ohsowski of the Institute of Environmental Sustainability at Loyola University Chicago

Bibliography

[1] Johnson, N L, Kemp, A W and Kotz, S (2005). *Univariate Discrete Distributions*. John Wiley & Sons

[2] Korhonen, L, Ali-Sisto, D, Tokola, T and others (2015). Tropical forest canopy cover estimation using satellite imagery and airborne lidar reference data. *Silva Fennica*. **49** 1–18

[3] Tu, W (2002). Zero inflated data. *Encycl environ 4:2387–2391*

[4] Martin, T G, Wintle, B A, Rhodes, J R, Kuhnert, P M, Field, S A, Low-Choy, S J, Tyre, A J and Possingham, H P (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*. Wiley Online Library. **8** 1235–46

[5] Smithson, M and Verkuilen, J (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*. American Psychological Association. **11** 54

[6] Hall, D B (2000). Zero-inflated poisson and binomial regression with random effects: A case study. *Biometrics*. Wiley Online Library. **56** 1030–9

[7] Ospina, R and Ferrari, S L (2010). Inflated beta distributions. *Statistical Papers*. Springer. **51** 111

[8] Rigby, R and Stasinopoulos, D (2001). The gamlss project: A flexible approach to statistical modelling. *New trends in statistical modelling: Proceedings of the 16th international workshop on statistical modelling*. University of Southern Denmark. **337** 345

[9] Rigby, R and Stasinopoulos, D (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*. Wiley Online Library. **54** 507–54

[10] Akantziliotou, K, Rigby, R and Stasinopoulos, D (2002). The r implementation of generalized additive models for location, scale and shape. *Statistical modelling in society: Proceedings of the*

17th international workshop on statistical modelling. Statistical Modelling Society. 75–83

[11] Nelder, J A and Wedderburn, R W (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*. Wiley Online Library. **135** 370–84

[12] Stasinopoulos, D, Rigby, R and others (2007). Generalized additive models for location scale and shape gamlss in r. *Journal of Statistical Software*. **23** 1–46

[13] Stasinopoulos, D, Rigby, R, Heller, G, Voudouris, V and De Bastiani, F (2017). *Flexible Regression and Smoothing: Using Gamlss in R*

[14] McCullagh, P and Nelder, J A (1989). *Generalized Linear Models*. CRC press