Statistical Consulting Report

Using Advanced Remote Sensing Approaches to characterize and predict juvenile salmonid habitat in small coastal streams

Consultants: Harper Cheng, Ning Shen
Client: Alyssa Nonis

Abstract

Researchers hope to construct a predictive model for estimating fish abundance based on ecological data in order to assess fish habitats in small coastal streams in a more efficient manner. The next step would be to link Airborne Laser Scanning (ALS) data to ecological data if the model performs well in terms of prediction. In this report, general suggestions regarding analyzing different fish species and the use of mixed-effects models are discussed; the model specification, however, should be determined by ecological professionals in order to correctly address the research question of interest. In regard to modelling zero-inflated positive continuous response, depending on the nature of the zero values, we recommend either replacing zeros with an arbitrarily small value and modelling the response with a Gamma or inverse Gaussian generalized linear model(GLM), or it is possible to use the Two-Part Model which uses separate submodels to handle the zero and non-zero responses. One important suggestion is that extrapolation beyond sampled streams should be dealt with care. R code for the application of Gamma GLMs and Two-Part Models can be found in the Appendix.

1 Introduction

The main objectives are to analyze how physical characteristics of streams affect juvenile fish density and biomass at both reach and habitat unit scale, and to explore the possibility of deriving a predictive model that can be applied to juvenile fish habitat assessment. Fish habitat is assessed at the reach scale to evaluate how fish abundance is related to the overall stream characteristics, such as channel width, gradient, and in-stream wood density. Fish abundance is also studied at the habitat unit scale which describes morphological attributes within each stream. To address the research question, the following hypotheses are of particular interest: (1) At the reach scale, juvenile fish density and biomass are associated with gradient, bank full channel width, and in-stream wood. (2) At the habitat unit scale (the main focus is on the "pool" because of its high fish abundance), increased juvenile fish density and biomass with respect to mean density and biomass across all habitat units are associated with pool habitat unit types, increased pool area, and in-stream wood presence.

2 Data Description

Data were collected from 5 streams in the Nahmint Watershed. Approximately 200 meters of each stream were assessed at the habitat unit scale where ecological data were collected. In each study reach, about 50 traps were set at designated locations within habitat units and soaked for 24 hours. There were four sampling events, and the event with the largest number of fish density and biomass was chosen to represent fish abundance for that location. The total sample size across all five streams is 373. For the pool area, the number of observations is 127 which we believe is a decent sample size for regression analysis [2].

3 Statistical Questions

Upon examining the dataset and your current analyses, we address the following statistical questions to help you tackle the research questions.

1. Should different species of fish be analyzed separately?
2. Should the selected five streams be treated as random effects? In other words, is there any noticeable difference between these five streams in terms of hydrological characteristics and fish abundance that prompts the use of a mixed-effects model?
3. What is the most appropriate statistical model for modelling responses that are right-skewed and contain a considerable number of zero values?

4 General Advice

Based on the information we have gathered, the two species of salmonids seem to be similar in terms of their size and weight. If predicting the abundance of a particular species is not of interest, we suggest analyzing different species altogether not only for the ease of execution but for a larger sample size as well. More importantly, the spawning of a fish species might be affected by other factors such as the availability of food and the presence of predators, which are not accounted for in our model. We have learnt that the occurrence of Dolly Varden is lower in streams with Rainbow Trout presence. If the biological interaction between these two species affects their abundance, then it might be sensible to analyze the fish population as a whole to mitigate the influence of possible interaction between species. On the other hand, if a given stream is a known habitat for a particular species and the interest is on predicting fish abundance for this species within their habitat, then it is reasonable to conduct separate analyses on different fish species.

We notice that the mixed-effects model was used to account for the difference between streams. A mixed-effects model is superb at cluster-specific interpolations. In other words, it can provide more accurate predictions for observations drawn from a given cluster. This is because when making predictions on a new observation, the mixed-effects model borrows information from similar observations within the same cluster [2]. However, if the new observation is from a novel cluster (i.e., a cluster that is not used for model fitting), predictions given by the mixed-effects model might not be accurate for out-of-sample observations. That is to say, if you are interested in generalizing the model to other streams, the prediction of fish abundance for a location sampled from a new stream might not be as reliable as they are for a location sampled from the current streams. When extrapolating to a novel cluster, what the mixed-effects model can provide is the population-level prediction where only the fixed effects are considered. However, with random effects being ignored, the population-level prediction is essentially equivalent to the prediction given by the standard linear model, rendering the use of mixed-effects models unnecessary. Thus if the sampled streams are homogenous in terms of hydrological characteristics, we suggest using a standard regression model instead of a mixed-effects model for the simplicity of the analysis. But if the sampled streams are heterogeneous and making stream-specific predictions is of interest, then it might be sensible to use mixed-effects models or conduct standard regression analyses on selected five streams separately.

5 Recommended Statistical Models

The most noticeable feature of the data is the large proportion of zero values. The choice of the statistical model that can be used for modelling such a response depends on the nature of the zero values.

5.1 Modelling Positive and Right-Skewed Responses: Gamma GLM

If the zero responses are merely some extremely small values that went undetected during data collection, it is acceptable to replace zeros with an arbitrarily small value and model the response as a continuous and positive variable. In the context of your research question, this approach would work the best if you have reasons to believe that a small number of fish actually reside in the sampled stream but failed to be captured due to their low population.

Motivations for GLMs

Figure 1 depicts the distribution of the response (biomass per square meter) with zeros replaced with 0.01. When modelling a continuous response that is positive and highly skewed, instead of transforming the response to make it normally distributed, it is more advantageous to use a Gamma or inverse Gaussian generalized linear model (GLM). We prefer GLM over log-transformed linear regression because direct transformations on the response result in the mean response not being modelled on the original scale which complicates the interpretation. In particular, if a log-transformed response $\log(Y)$ is fitted to linear regression, we are modelling the mean of log-transformed response, $\mathrm{E}(\log(Y))$. Thus, it is only interpretable in terms of the (arithmetic) mean change on the log scale. The re-transformation of the result back to its original scale is not straightforward and can be impractical if the error terms are not normally distributed with constant variance [3]. On the other hand, GLM can model the mean response directly by connecting it to the linear predictor $\eta = \mathbf{X}^T\boldsymbol{\beta}$ with a link function $g(.)$, that is, $\eta = g(\mathrm{E}(Y)) = \mathbf{X}^T\boldsymbol{\beta}$. Therefore it is possible to make inference about the arithmetic means $(\mathrm{E}(Y))$ on the original scale. A Gamma or inverse Gaussian distribution are suitable for modelling a positive and continuous response that is highly skewed. In

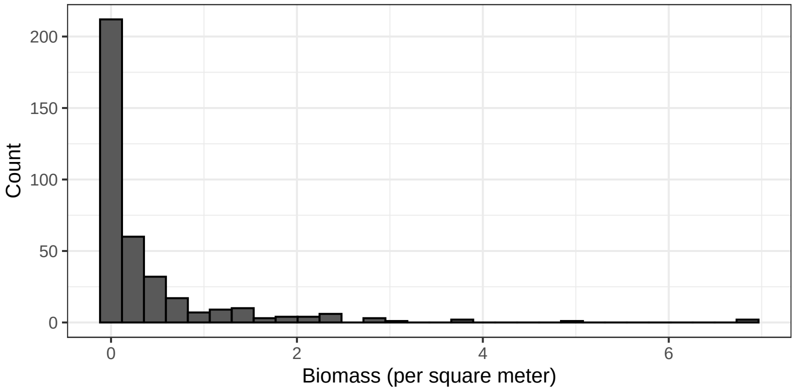


Figure 1: Histogram of biomass per square meter with zeros replaced with 0.01 at the reach scale (all five streams are pooled together for a larger sample size). Distribution is highly right-skewed.

the context of the research question, we believe that a log link best represents the underlying fish breeding process which is usually multiplicative (there exists some other options for specifying the link function which is discussed in the next paragraph). Thus, the mean of the response is linked to the linear predictor through a log link, that is $\eta = \log(\mathrm{E}(Y)) = \mathbf{X}^T\boldsymbol{\beta}$. The mean response can be easily obtained by exponentiating $\log\mathrm{E}(Y)$ which results in $\mathrm{E}(Y) = \exp(\mathbf{X}^T\boldsymbol{\beta})$.

We also provide the discussion for other choices of the link function in the Appendix. Figure 2 shows the relationship between the linearized response (calculation on linearized response is detailed in the Appendix) and linear predictors when a log link is used. There exists a linear relationship indicating that the log link function might be appropriate to use for addressing the relationship between the mean response and the linear predictor.

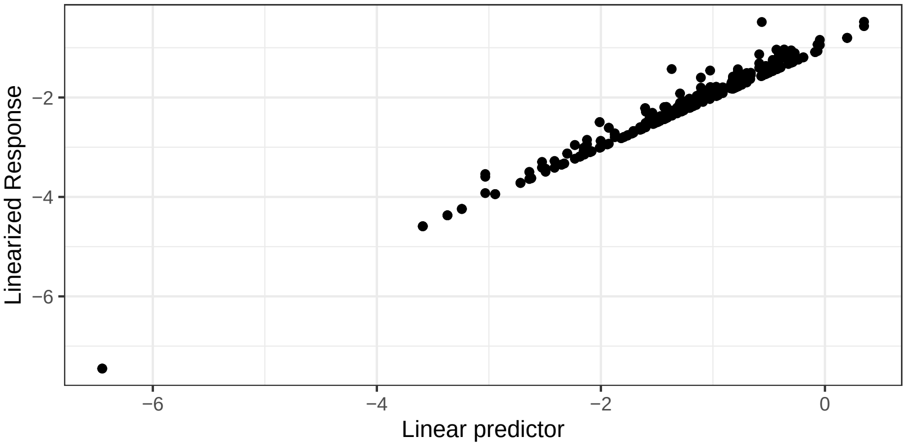


Figure 2: A scatterplot of the linearized response vs linear predictors can be used for examining the appropriateness of the link function. The log link is used. The linear relationship between the linearized response and the linear predictors suggest that the log link is a reasonable choice.

The Choice of Probabilistic Distributions

We have mentioned that both the Gamma and inverse Gaussian GLM can be used for modelling a positive and right-skewed response Y. Unlike in linear regression where the variance of Y is constant, the variance of the Gamma GLM grows with its mean and the variance of the inverse Gaussian GLM increases more rapidly than the Gamma GLM. To determine which distribution would be the most appropriate for modelling the given data, after fitting GLM on the data, we can obtain maximum likelihood estimates of the parameters for the response distribution and generate a fitted density plot which can be overlaid on the histogram of the actual response. Here, we demonstrate the process by fitting the GLM with a Gamma distribution where the response is taken as biomass per square meter and explanatory variables are gradient, bankfull channel width, and wood density. Figure 3 depicts the fitted density curve overlaid on top of the histogram of the response (biomass per square meter). The density curve fits the actual distribution of the response quite adequately suggesting that the use of Gamma distribution for modelling the response is justifiable.

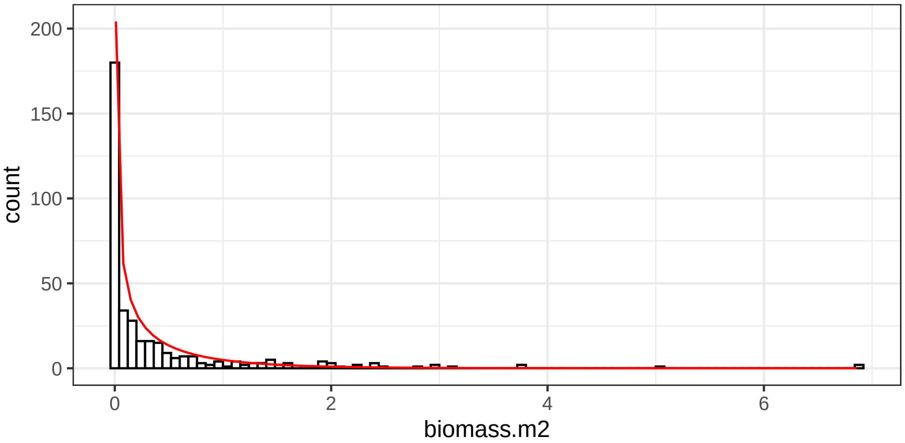


Figure 3: Histogram of the response (biomass per square meter) with the fitted Gamma density curve overlaid. The Gamma density provides a good fit to the actual data meaning that we can use the Gamma distribution for modelling the response.

Diagnostics for GLMs

After fitting a Gamma GLM, it is important to perform diagnostics on the fitted model to check if the assumptions for GLM are met. Similar to the diagnostics for the linear regression, a residual plot can be used to check the variance of the residuals. Since the variance of the response is usually non-constant for GLMs, we adopt the deviance residual which is modified from the regular residual such that it can be used in a similar manner as the regular residual. The deviance residual scales out the non-constant variance function meaning that a good diagnostics plot should show a constant variance pattern when the deviance residuals are plotted against the linear predictor. Another form of residual is called the standardized deviance residual which appears to be more symmetric in shape and close to a normal distribution. The distribution for standardized deviance residuals can be assessed with a QQ plot. In addition, Cook's distance can be used to detect any influential points the same way it is used in the linear regression diagnostics. Figure 4 shows the diagnostics plots for evaluating the fitted Gamma GLM. The deviance residual plot shows an increasing trend of the variance and the standardized deviance residuals do not appear to be normal. Cook's distance plot reveals that the unsatisfactory fit might be due to the presence of influential points, i.e., some sampled habitats are highly populated reservoirs for juvenile salmons.

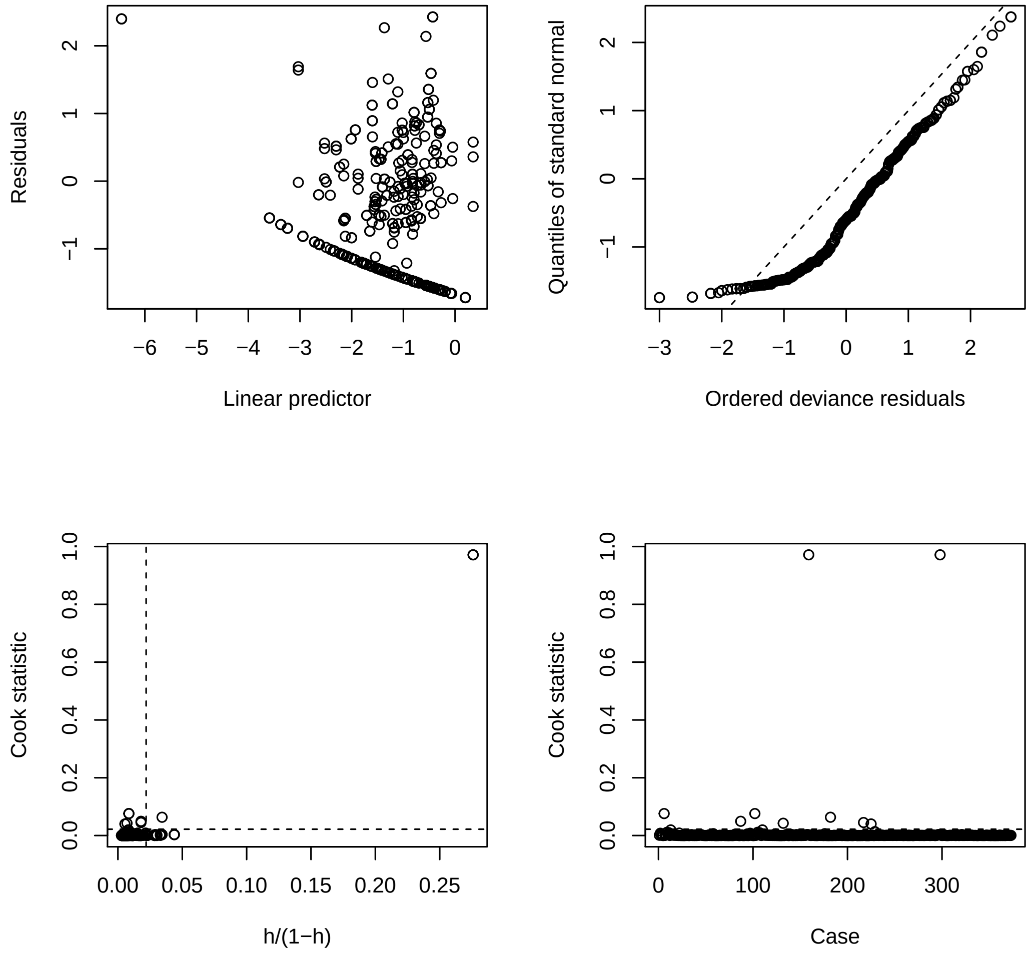


Figure 4: Diagnostics plot for Gamma GLM. Top left: deviance residuals vs linear predictor; top right: QQ plot of standardized deviance residuals; bottom left: Cook's distance vs leverage; bottom right: Cook's distance plot.

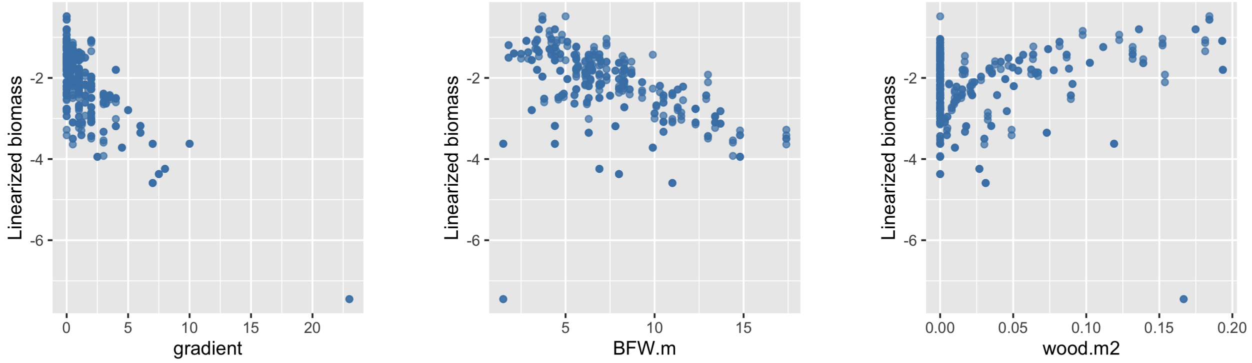
Interpretation of GLM Results

To demonstrate the interpretation, below is an example of `R` output from the Gamma GLM with a log link. The response is biomass per square meter with zeros replaced with 0.01.

```
> Nahmint_pos <- Nahmint %>%   ## tranform response variable biomass
+   mutate(biomass.m2=if_else(biomass.m2==0, 0.01, biomass.m2))
> fit.gamma <- glm(biomass.m2~gradient+BFW.m+wood.m2, data=Nahmint_pos,
+                  family=Gamma(link=log), maxit=10000)  ## Gamma GLM with log link
> a <- gamma.shape(fit.gamma)$alpha  # compute the maximum likelihood estimate of shape parameter
> summary(fit.gamma, dispersion = 1/a)

Call:
glm(formula = biomass.m2 ~ gradient + BFW.m + wood.m2, family = Gamma(link = log),
    data = Nahmint_pos, maxit = 1e+06)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7612  -2.1025  -0.9713   0.2437   3.7976

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.25701    0.21515   1.195   0.2323
gradient    -0.30846    0.03175  -9.714  < 2e-16 ***
BFW.m       -0.16417    0.02359  -6.958 3.45e-12 ***
wood.m2      3.79402    1.58352   2.396   0.0166 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 2.172062)

> round(coef(fit.gamma),3)
 (Intercept)    gradient       BFW.m     wood.m2
       0.257        -0.308        -0.164          3.794

> round(exp(coef(fit.gamma)),2)
 (Intercept)    gradient       BFW.m     wood.m2
         1.29           0.73           0.85           44.43
```

The estimated intercept is 0.257 and the estimated regression coefficient β_i for the three predictors are -0.308, -0.164 and 3.794 respectively. Correspondingly, we observe a negative trend when the linearized response is plotted against *gradient* and *BFW.m* and a slightly positive trend when it is plotted against *wood.m2*

(Figure 5). Detailed calculation on linearized response can be found in the Appendix. We take $\beta_1 = -0.308$ as an instance for interpretation. Each time the gradient increases 1 unit while other variables remain fixed, the log arithmetic mean biomass decreases by 0.308. The exponentiated coefficient $e^{-0.308} = 0.73$ is the factor by which the arithmetic mean biomass on the original scale is multiplied, i.e., if the gradient increases 1 unit, the arithmetic mean of the biomass is expected to reduce $1 - 0.73 = 27\%$ given other predictors remaining the same.



(a) **(b)** **(c)**

Figure 5: Scatter plot of the linearized response versus a) gradient, b) bank full width and c) in-stream wood/m2.

On a side note, the dispersion parameter, which describes the scattering of the data points around the mean, is defined as the inverse of maximum likelihood estimate (MLE) of the shape parameter. The `glm` fit for a Gamma distribution correctly calculates the maximum likelihood estimates of the mean parameters, but it provides only a crude estimate of the dispersion parameter. Instead of using the estimates provided directly by the `glm` fit, we use the function `gamma.shape()` which takes the results of the `glm` fit and solves for the maximum likelihood equation for the shape parameter. The reciprocal of the MLE for the shape parameter provides a better estimate of the dispersion parameter.

All of the above analyses and diagnostics are performed with a Gamma GLM which is suitable for modelling the response (biomass per square meter) with zeros replaced with the value 0.01. The inverse Gaussian is not appropriate in this case. This is because a large proportion of the response resides in the region around 0.01. But the density of an inverse Gaussian near zero is extremely low compared to a Gamma distribution. However, the inverse Gaussian distribution could be of use in other scenarios.

5.2 Modelling Zero-inflated Positive Continuous Responses: Two-Part Models

If the zero values in the response are reflective of the absence of observations, for example, there is absolutely no fish populated in the sampled streams, then the zeros can be kept as they are without replacing them with other values. When modelling a response that has more zeros than expected, a zero-inflated model might be appropriate to allow for excess zero responses. Specifically, for the positive continuous responses in this study, we recommend the use of the Two-Part Model which is composed of two submodels (or parts), namely a Bernoulli distribution for modelling zero and non-zero values as a binary response, and a strictly positive distribution for modelling the positive continuous response.

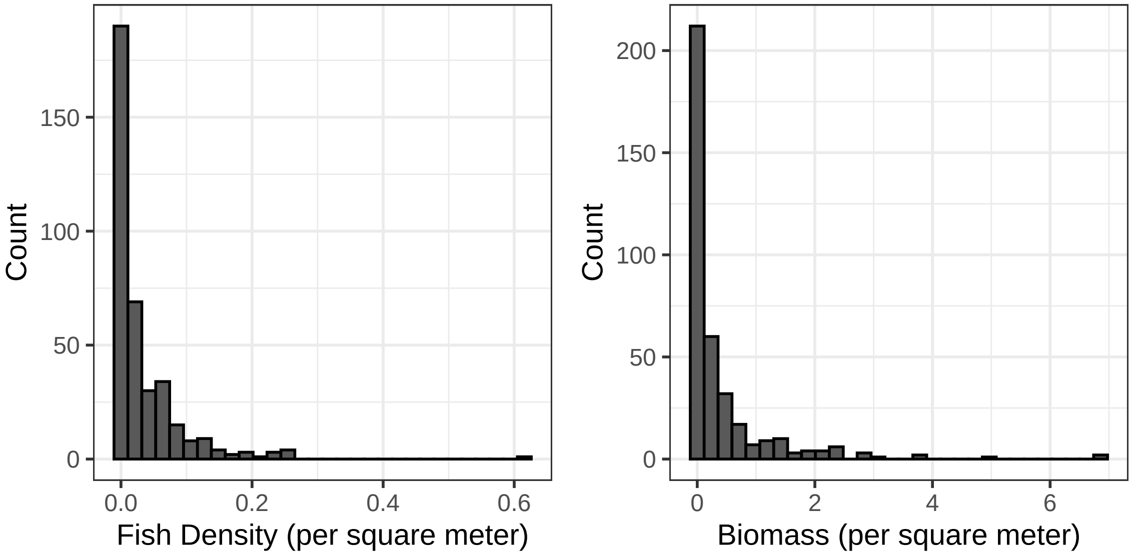


Figure 6: Histograms of the responses (fish density and biomass per square meter) at the reach scale. Responses at the habitat unit scale (i.e., pool area) have similar distributions. Distribution for both responses are highly skewed to the right with a large number of zero values which prompts the use of Two-Part Models.

Motivations

The zero-inflated model is preferred for the following reasons. Sometimes, it is important to distinguish zeros from non-zero responses as the value zero might indicate an interesting characteristic of the study subject. In the context of your research question, if the fish density at a location is zero, it is likely that the sampled location is not a good habitat for juvenile salmonids. It might be of research interest to identify poor fish habitats and differentiate them from high-quality ones.

Another benefit of the Two-Part Model comes from modelling the zero and positive responses in separate submodels. Now that the zeros are out of the way, instead of modelling the transformed response with linear regression, we can use a GLM with distributions, such as the Gamma or inverse Gaussian distribution, which are more suitable for modelling positive but skewed responses.

Theoretical Framework

The Two-Part Model treats the data as if they arise from two distinct data generating processes: one concerning the observance of zeros/non-zeros (i.e., modelling the response as a binary random variable), and the other one determining the observed value if the response is non-zero (i.e., modelling the response as a positive continuous variable). It first models zero and non-zero responses separately and then combines the above-mentioned processes based on conditional probability rules [4].

Formally, we introduce a new binary variable Y_0 to indicate whether the response is zero. You can easily do this in R by dichotomizing the response to 0 and 1. A logistic regression is used to model Y_0 such that the probability of Y_0 greater than 0, i.e., $P(Y_0 > 0|X)$, can be obtained. It describes the probability of observing a positive response given explanatory variables X. Secondly, the positive response is modelled with a Gamma GLM using the log link. The GLM provides estimates on the mean response given non-zero response values, that is $E(Y|Y_0 > 0, X)$. Now, based on the rules for conditional probability, the mean response is estimated as

$$E(Y|X) = P(Y_0 > 0|X) \times E(Y|Y_0 > 0, X). \tag{1}$$

Prediction of Two-Part Models

We demonstrate the prediction process by partitioning the original dataset into a training and a test set. The training set is used for fitting the regression model and the test set is for testing the predictive power of the fitted model. The observed and predicted biomass for the test set are compared using a side-by-side boxplot as shown in Figure 7 and their corresponding density curves are shown in Figure 8. It seems that the fitted model based on the training set gives acceptable predictions when the true observation values are relatively small, though the model sometimes overestimates or underestimates the true values (as suggested by the crossed gray lines). However, the prediction is not very accurate when the true biomass is large.

We can also simulate the distribution of biomass for the test set by using the estimated parameter values obtained from the logistic regression and Gamma GLM. Similar to what is observed in Figure 8, the simulated values predict small responses fairly well. But the density of the fitted model flats out when the response is large making it unsuitable to predict large values. Overall, we think there is a possibility of using the derived model for prediction, but it is not good at capturing outlying observations and the predictive power is not very high.

Apart from predicting the mean fish abundance of a given area, the two modelling processes are also informative in terms of which ecological factors are important for predicting fish habitat/abundance. For example, the logistic regression suggests that gradient is a significant predictor for determining whether a location is a good habitat for salmonids. Results from Gamma GLM indicate that gradient, bank full channel width, and in-stream wood density are significantly associated with fish abundance at the reach scale.

Moreover, correlation within clusters can be accounted for in the Two-Part Model by incorporating random effects to the logistic regression and Gamma GLM [3]. However, as stated earlier, since the out-of-sample prediction given by the mixed-effects model might not be reliable, the use of random effects is not strongly recommended.

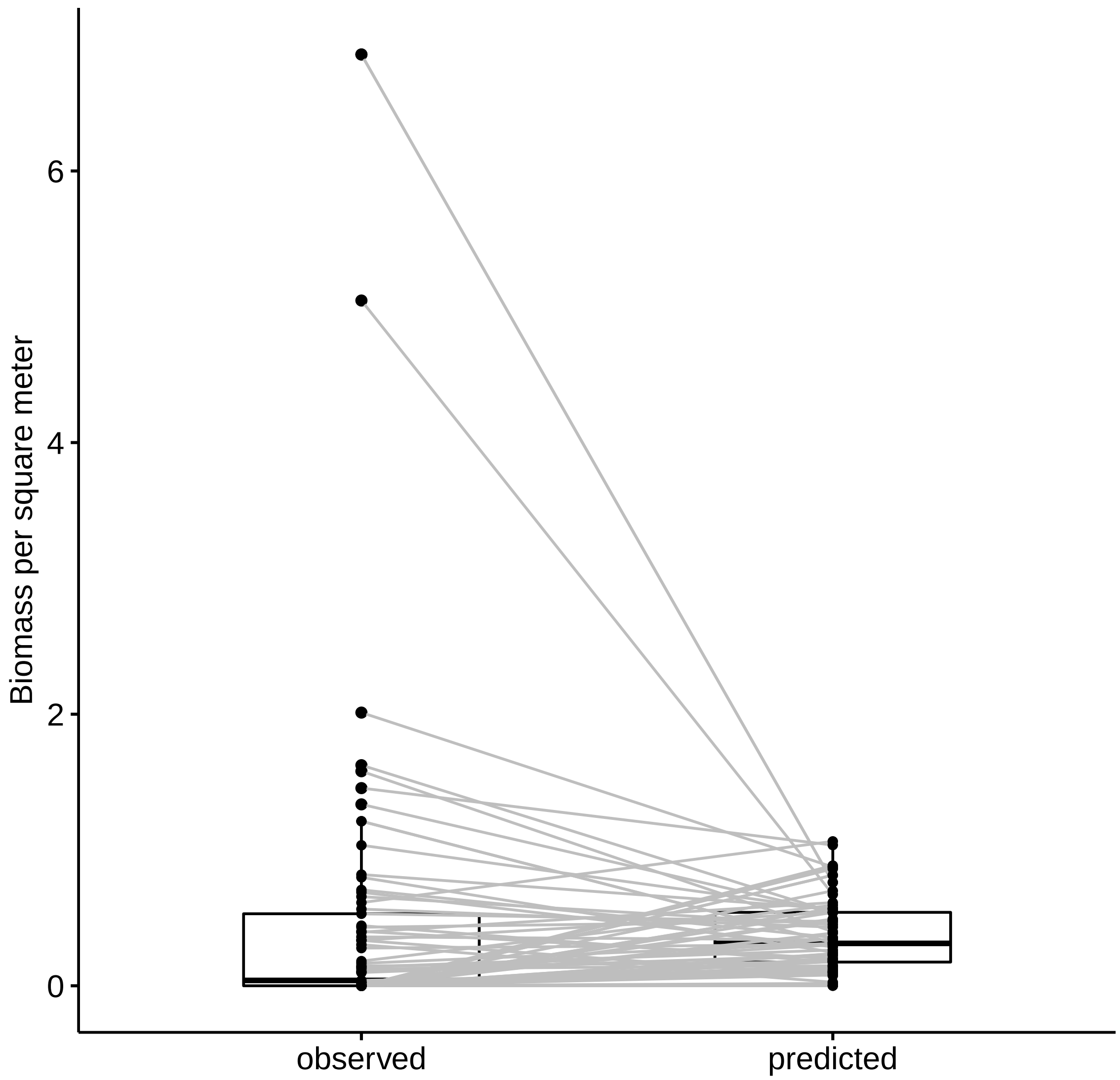


Figure 7: Side-by-side boxplot for observed and predicted biomass for the test set. The gray line connects the observed and predicted value for each data point. The model gives decent predictions when the observations are small in values. But it fails to predict large observations.

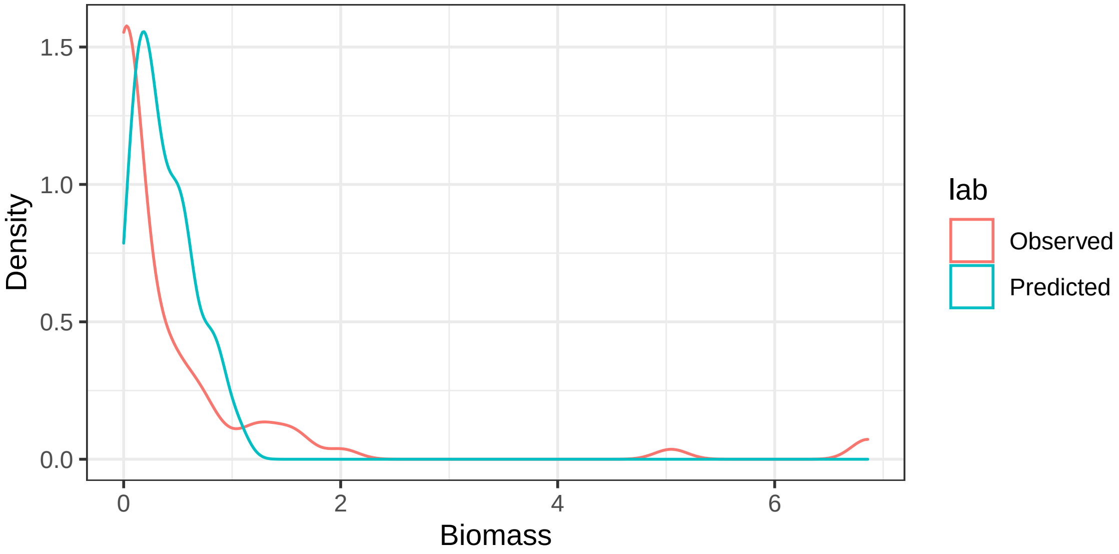


Figure 8: Density plot on observed and predicted biomass for the test set.

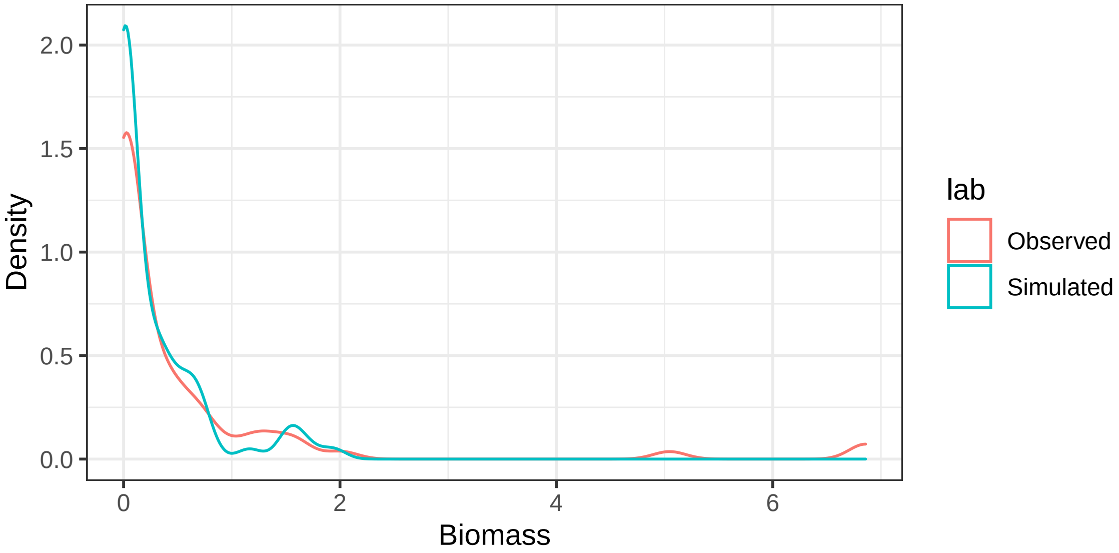


Figure 9: Density plot on observed and simulated biomass. The problem of having low predictive power for large observations persists.

6 Conclusion

To model the fish abundance in selected streams or habitats in the Nahmint watershed, we recommend the use of Gamma GLMs for modelling the positive and skewed response or Two-Part Models to accommodate zero-inflated responses. The choice of the model is dependent on the nature of the observed zero values. For the Two-Part Model, the response is first dichotomized to a binary response which can be modelled by a logistic regression. The second step models the positive continuous response with a Gamma or inverse Gaussian GLM to avoid direct transformation on variables. The mean response can be estimated by combining results from the two steps based on the rule of conditional probability. Furthermore, the Two-Part Model can be applied to correlated data where individual streams are treated as random effects. The cluster-specific prediction given by the mixed-effects model would be more accurate and reliable. However, we should always be cautious when it comes to predicting the value of a sample drawn from a novel cluster since mixed-effects models are not good at out-of-sample prediction. Thus, extrapolation beyond selected streams or beyond the Nahmint watershed might be dangerous.

7 Further Reading

Some tutorials on Two-Part Model implementation using R:

1. `https://devinincerti.com/2015/09/11/twopart.html`
2. `https://seananderson.ca/2014/05/18/gamma-hurdle/`

If you are interested in finding out more detailed mathematical derivation of the Two-Part Model, here are some articles that might help with understanding (clickable link embedded):

1. Statistical Analysis of Zero-Inflated Nonnegative Continuous Data: A Review
2. Modeling zero-modified count and semicontinuous data in health services research Part 1: background and overview

Appendix

Code Availability

R code for data analysis using the Two-Part Model can be accessed from my Github `https://github.com/harpercheng91/StatsConsulting/blob/main/Nahmint.Rmd`.

The Choice of Link Functions

For the Gamma GLM, there are three common choices for the link function [1].

1. The inverse link: $\eta = \frac{1}{\mathrm{E}(Y)}$.
   The inverse link has a nice property that the mean of the response is bounded as the predictor x increases. One problem with the inverse link is that the linear predictor $\mathbf{X}^T\beta$ takes on any real values, that is $-\infty < \mathbf{X}^T\beta < \infty$, which does not guarantee a positive mean response.

2. The log link: $\eta = \log(\mathrm{E}(Y))$.
   It is typically used when the effect of predictors is believed to be multiplicative on the mean. By applying the log link, the positive response is mapped to the real line such that its range matches the domain of the linear predictor.

3. The identity link: $\eta = \mathrm{E}(Y)$.
   If the effect of predictors is believed to be additive as opposed to multiplicative, the identity link can be used.

When the underlying data generating process is unclear, one can choose the link function by the characteristics of the response or by ease of interpretation. It is always prudent to check if the chosen link function is sensible. This can be visually examined by plotting a linearized response $\mathbf{z} = g(\mathbf{y})$ against the linear predictor $\eta = g(\mu) = g(\mathrm{E}(Y)) = \mathbf{X}^T\boldsymbol{\beta}$ and see if there exists a linear relationship. The linear response can be calculated by first-order approximation, that is,

$$\mathbf{z} = \eta + (\mathbf{y} - \mu)\frac{d\eta}{d\mu}. \tag{2}$$

If a log link is used, then $\eta = \log(\mu)$, and the derivative is computed as

$$\frac{d\eta}{d\mu} = \frac{d\log(\mu)}{d\mu} = \frac{1}{\mu} \tag{3}$$

References

[1] Faraway, J. J. (2005). *Extending the Lineal Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC.

[2] Jenkins, D. G. and Quintana-Ascencio, P. F. (2020). A solution to minimum sample size for regressions. *PLOS ONE*, 15(2):1–15.

[3] Liu, L., Shih, Y.-C., Strawderman, R., Zhang, D., Johnson, B., and Chai, H. (2019). Statistical analysis of zero-inflated nonnegative continuous data: A review. *Statistical Science*, 34:253–279.

[4] Neelon, B., O'Malley, A. J., and Smith, V. A. (2016). Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Statistics in Medicine*, 35(27):5070–5093.