Using Bayesian Regression GLM Methods to Analyze Arctic Biomass Density Trends Over Time

ISYE 6420, Spring 2024
Eagle Yuan

GitHub Link: https://github.com/eagleyuan21/bayesian-regression-biomass-climate-change

Introduction

The presence of biomass has shown to be an important attribute in the global ecosystem. In addition to their general influence on an ecosystem [1], their ability to balance energy to the ecosystem as well as their carbon trapping and oxygen releasing capabilities establish their crucial role in a system [2]. Additionally, these biomasses have proven to be extremely sensitive to climate warming and cooling. This has caused an increase in their growth in multiple types of biomasses [3], which in turn causes more changes to the ecosystem in an almost cyclic cause and effect chain. Given the role of biomass in the ecosystem, they are a particular interest of study in understanding my climate.

As a result of the ever-expanding study of climate change, an even greater focus has been placed on the geographical location of the arctic. The primary reason for this focus is the studied effect of climate warming in the arctic is several times faster than anywhere else on the planet [4]. There have been several major efforts to improve the coverage of research and data and analysis in the arctic due to the significance of its nature relative to the rest of the planet, including the dataset in which these Bayesian methods presented in this paper are based upon [5]. This dataset includes years of study and data gathering aggregated over several arctic and subarctic regions across the planet. This dataset includes the measurements of biomass of several specific types of biomasses to have a consistent base to compare between, as well as several important factors measured in conjunction with the biomasses.

With these Bayesian methods, I hope to model a relationship between the biomasses and some of the factors in the dataset to gather more insights into the general relationship between arctic biomass and its trend over time.

Background

Bayesian GLM Regression

There are two major GLM Bayesian regressions that I choose to model my data. My decision process for these two models will be elaborated upon in my methods section. I choose to use the common normal distribution GLM, also known as the linear regression, and I choose to use the exponential distribution GLM regression. The below describes both of my distributions and how I form my linear and distribution fitting parts of my model.

$$y_i|\beta, x_i \sim Normal(\mu, \sigma)$$

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\mu = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\beta_0 \sim N(0, \sigma_0)$$

$$\beta_j \sim N(0, \sigma_j)$$

$$\sigma \sim \frac{1}{\sqrt{\tau}}$$

$$\tau \sim Gamma(0.001, 0.001)$$

The g function is the link function. In the normal case this is derived as the identity function. The inverse of the function is still the identity.

$$y_i|\beta, x_i \sim Exponential(\lambda)$$

$$g(\lambda) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$\lambda = g^{-1}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right) = -\frac{1}{\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}}$$

$$\beta_0 \sim N(0, \sigma_0)$$

$$\beta_j \sim N(0, \sigma_j)$$

The g function in the exponential case is the negative reciprocal. The inverse of the link function is still the negative reciprocal.

With these two models, I can fit my data. One important thing to note is that I use mostly uninformed priors. This is because I was curious to see the effect of how some x variables are modified with the beta coefficient relating to that x variable. I thus have a prior centered around 0, and I fit to see how the coefficient is fitted and what cause and effect the coefficient indicates how the x variable affects my y variable. Additionally, a lack of knowledge in the domain prevents from forming more informed priors.

Python and PyMC and Arviz

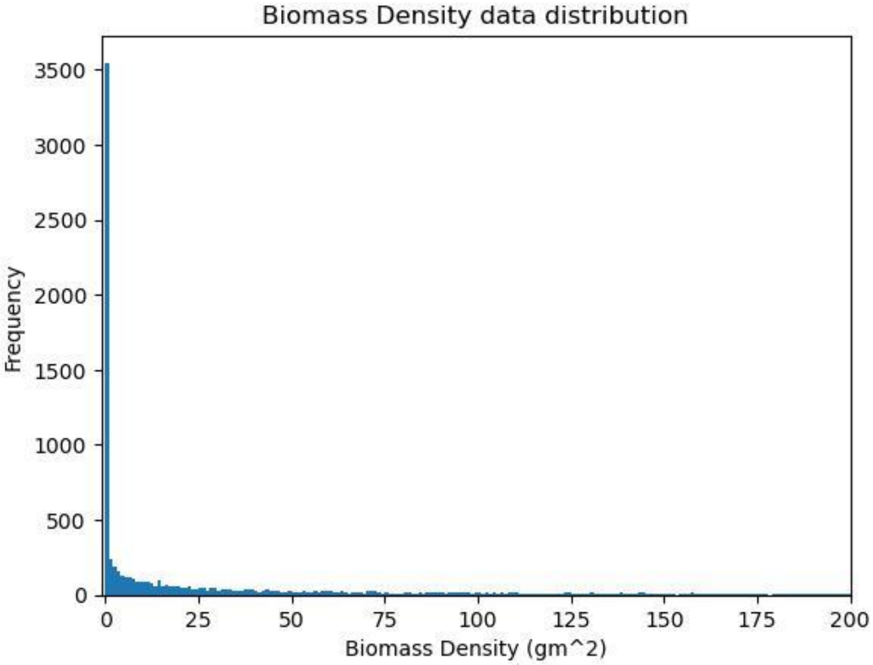
The work of this paper was done in the Python3 programming language. The two main packages used in this investigation were PyMC and Arviz [6 and 7].

PyMC is a probabilistic programming language package written in Python. This package at its core uses Markov chain Monte Carlo simulations to continuously sample and calculate problems to solve multiple probabilistic problems. In my case, I use the package to solve my Bayesian analysis problems using the sampling and computational features that PyMC has implemented and fit my model.

Arviz is a Python package that quickly helps run analysis on Bayesian models. This is particularly convenient in the cases of determining the accuracy and validity of my models. Arviz can provide analysis in several statistical and graphical and visual forms, which will be shown later in this paper.

The Dataset

I use the data gathered from The Arctic Above Ground Biomass Synthesis Dataset. This dataset includes several specific types of plants gathered across multiple years and multiple locations across several areas in the arctic and subarctic. For each entry of the biomass data, there are several corresponding variables that relate to that entry. Although there are a lot of entries, I decide to fit for biomass density, which is units of grams per meter squared. The plot below shows a distribution of all the biomass densities measured.

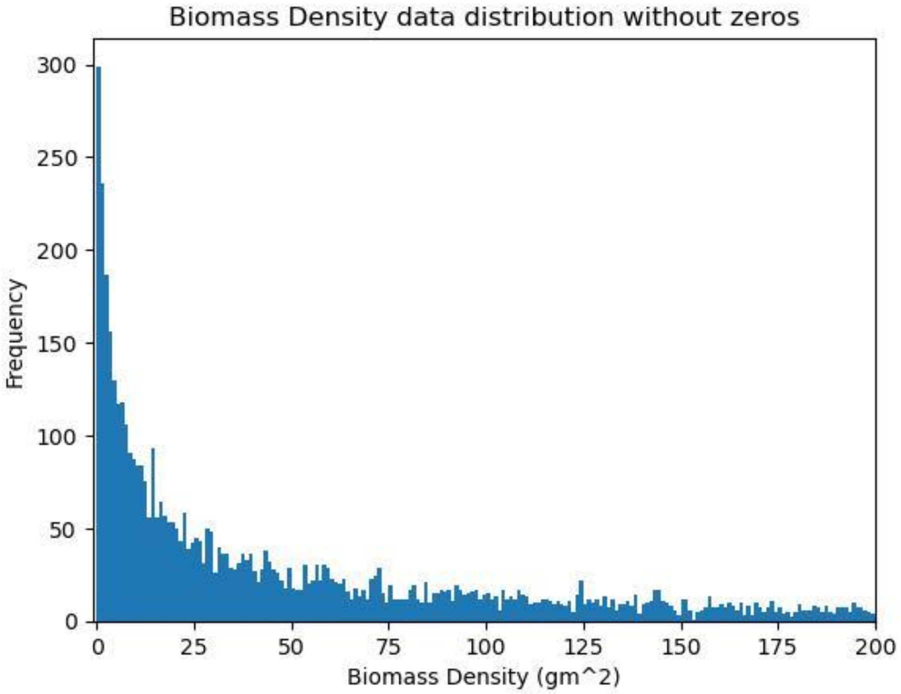


Based off my limited knowledge of biomass density factors, I hand selected a few of the variables that I deemed influential to fit my biomass density. These include the year, plant type, bioclimatic zone, the mean annual temperature, mean annual growing degree days, and mean annual precipitation. For the year, I normalize the year by subtracting all years from my minimum year. For the bioclimatic zone and the plant type, I encode these strings to a set of integers since there’s only a few specific values these variables can be.

Methods

With or without zero densities

As seen in the plots of my biomass densities, there is a strikingly large amount of zero biomass densities. This factor in the data is something that I needed to investigate to determine the effects of all the zero biomass densities in my models, as this can be the effect of multiple factors in my arctic ecosystem. I thus conduct two separate variations of methods for each of my models, one including my zero density data and one excluding my zero density data. The plot below shows the same biomass density data without the measurements of zero.

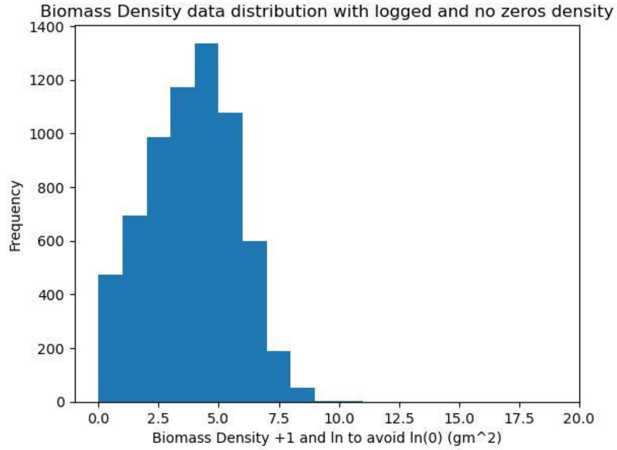


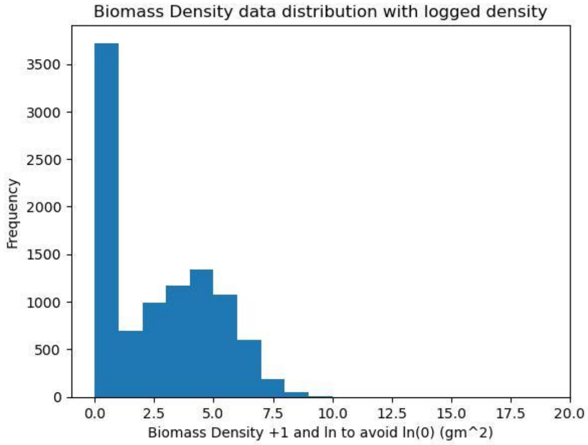
Exponential GLM

As seen from the dataset distributions for both excluded zero densities and all densities, the distribution resembles a distinct exponential graph. This helped justify the first method of my models, which included the exponential GLM. I pull the variables I stated above into my linear model part with a corresponding coefficient multiplier for each variable, and I also include an intercept. As stated earlier in the Bayesian GLM section, I need to use the link function that was derived for the exponential model. I apply this link function to the linear equation generated from my intercept, my covariates, and my variables from the dataset, and I model this against my observed Y, which is my biomass density. Then I use PyMC's sampling function to fit my data with 1000 samples into an exponential distribution, and I report my linear model's coefficients. To compare the results of the model, I use two main metrics. The first metric used is the R2 score from the arviz. Another metric I use is the arviz plot trace and plot posterior functions. I gather 500 samples from the posterior distribution, and I use these two plots to demonstrate both the change of my coefficients over each sample, as well as compare my fitted posterior distribution to the observed data (my actual biomass densities).

Normal GLM with Log Densities

As seen in the dataset, my data strikingly resembles an exponential distribution. I believed that although I can model this data to my exponential distribution, it would be an interesting variation to model this to a normal distribution. Thus, I take the log of all my biomass densities and use a normal distribution to model them. It is important to note that there is a large portion of the dataset that contains zero density, so I therefore add 1 to each value of my density to offset the zero data, and I take the log of that value. The following plots shows a distribution of my biomass densities with this data modification.





As shown above, the distribution now closely resembles my normal distribution. However, the biomass density data that does include the zero densities has a high peak at zero, which was discussed earlier in my methods section.

Now that I have prepared my data, I follow a similar method to my exponential model. I create the same linear model, and I apply my link function to the linear model. As derived in the earlier section, the link function for the normal distribution is just one, so I have no transformation. I then fit to my normal distribution with 1000 samples, and I report the linear model's coefficients. I also use the same strategy of reporting my results as the exponential model.

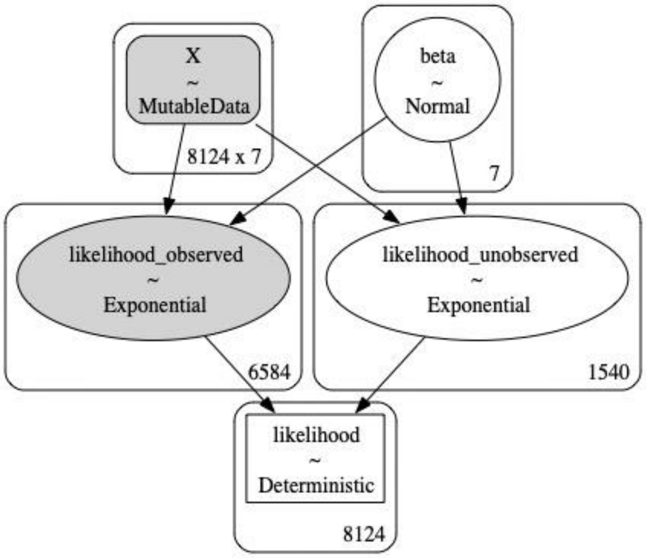
With or without Null/Missing densities (masking)

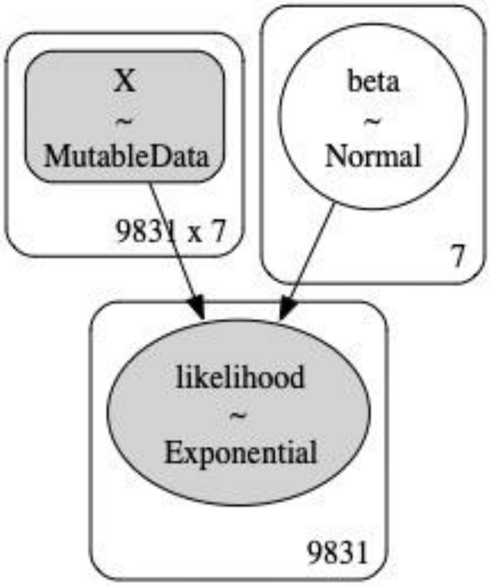
Another interesting characteristic of the dataset is that there are several entries in the data where there is a blank entry for the biomass density. For these entries, the remaining column values still contain data. This can be considered an ignorable missingness, as these values are randomly missing. To address this problem, I apply the Bayesian technique of masking where the biomass density values are missing, and I model the distributions using this masked data. However, for the sake of comparison, I also create a version of each model that do not contain the missing values and just treat like those entries don't exist. This gives us a further indicator of the effect of the missing density values and the impact it has on my models.

Results

Exponential Model

Below is a visual representation of two exponential models. The model on the left includes the censoring of data that have no entries, and the model on the right excludes those data entries.





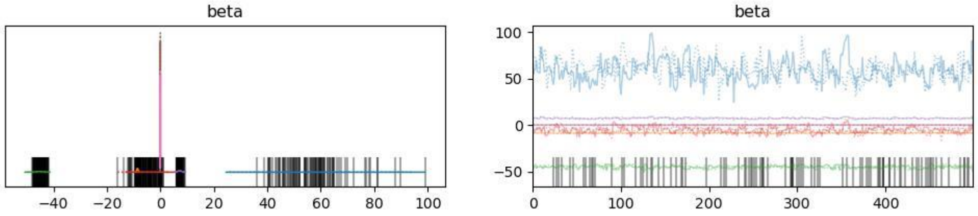
As seen in the visual, the model on the left that considers the missing data also has an additional sampling for the unobserved data. This helps provide an estimation for each entry to counter my ignorable missingness in my dataset.

The table below displays my R2 score and the coefficients for my linear model variables for both with and without data as well as with and without my zero biomass density data. I also report the 95% credible set after the value.

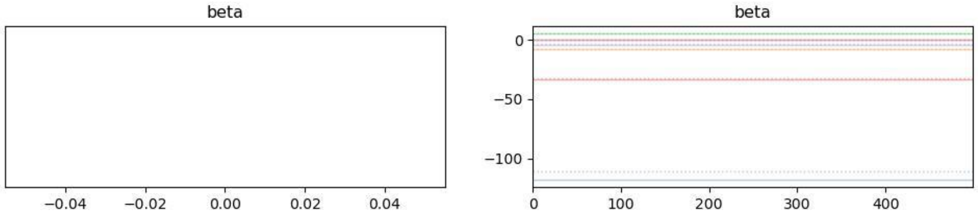
	No Zeros and No Null Data	Zeros and No Null Data	No Zeros and Null Data	Zeros and Null Data
R2	0.158	0.149	0.159	0.146
Intercept	60.077, (36.80,83.95)	-110.539, (-117.41,-103.66)	59.931, (36.02,83.87)	-108.42, (-99.9,-115.9)
C_Year	-8.549, (-9.07,-7.97)	-7.267, (-7.272,-7.263)	-8.553, (-9.044,-8.027)	-7.394, (-7.38,-7.408)
C_PlantType	45.176, (-48.06,42.31)	5.943, (5.593, 6.293)	-45.171, (-47.89,-42.01)	5.943, (5.593, 6.293)
C_BioclimaticZone	-5.211, (-12.09,0.78)	-34.003, (-34.01,-34.0)	-5.139, (-12.08,1.22)	-34.003, (-34.01,-34.0)
C_MeanTemp	7.327, (5.95,8.66)	-3.727, (-4.163, -3.291)	7.32, (5.92,8.65)	-3.84, (-4.24, -3.44)
C_MeanGrowingDays	0.002,	0.014,	0.002,	0.013,

	(-0.013,0.017)	(0.012, 0.016)	(-0.013,0.017)	(0.011, 0.015)
C_MeanPrecipitation	-0.004, (-0.023,0.015)	0.129, (0.124, 0.133)	-0.004, (-0.022,0.019)	0.129, (0.124, 0.133)

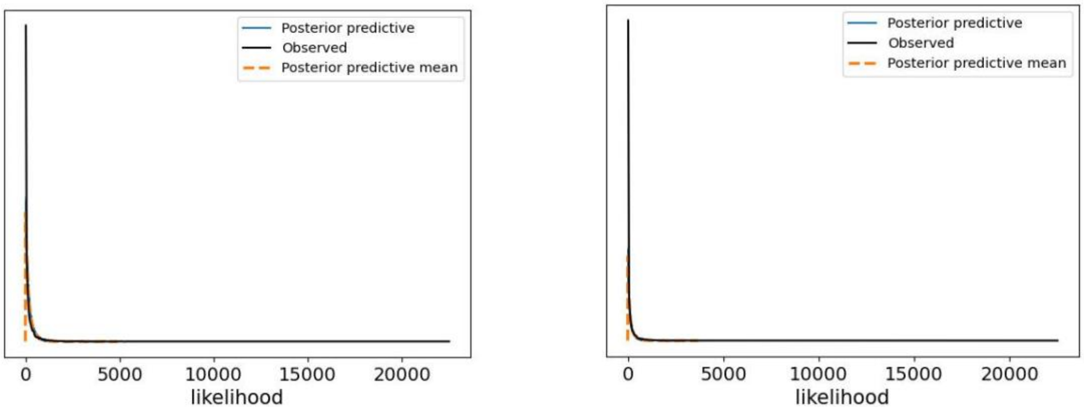
I choose not to report the estimates of the missing data since there are over 1000 values. The following graphs also display a trace of how my coefficients were obtained, and the final distribution compared to the observed distribution. I only report the no null data models because it was compute intensive to visualize the models with the null data since it was sampling not only my coefficients but also the unobserved data points.



Without zeros



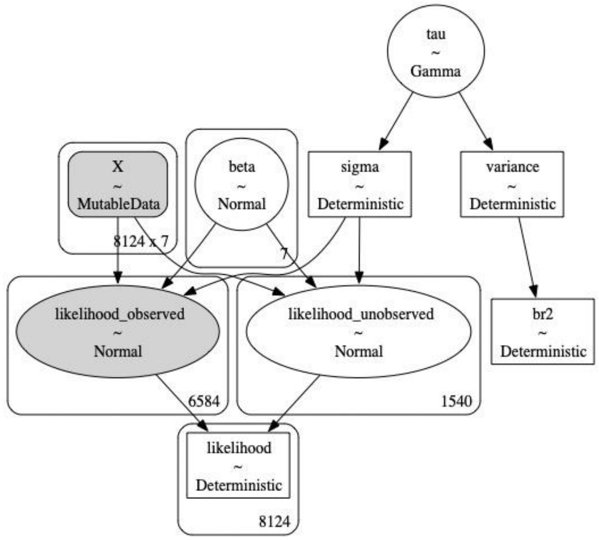
With zeros

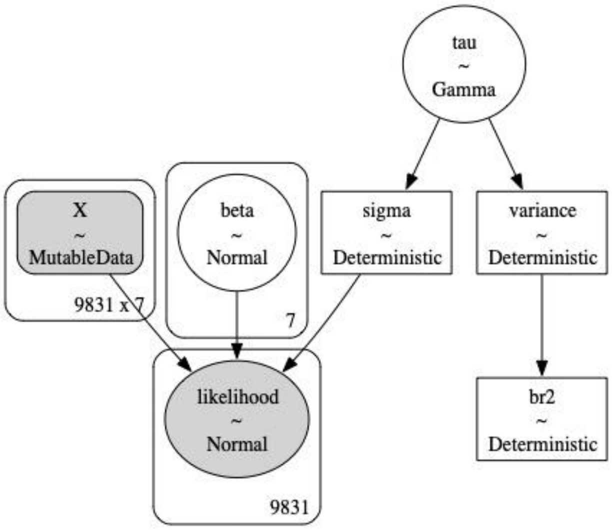


The left graph is for without zeros model. The right graph is for with zeros model.

Normal Model

Like my exponential model, I also show a visual of my normal models. Once again, the model on the left is my model with null data, and the model on the right is the model ignoring null data.

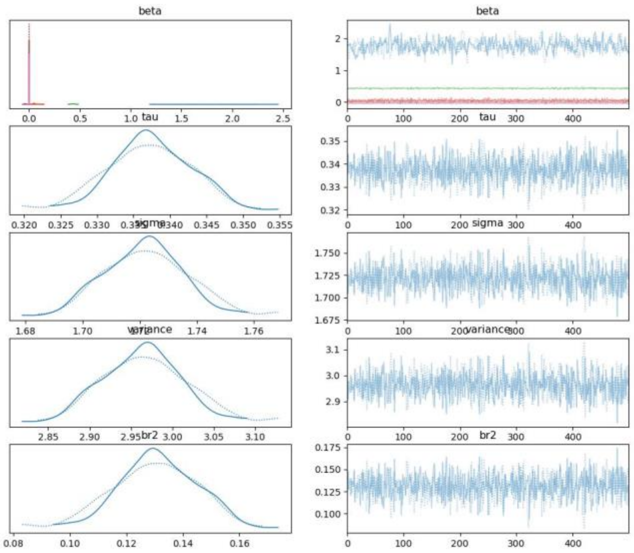




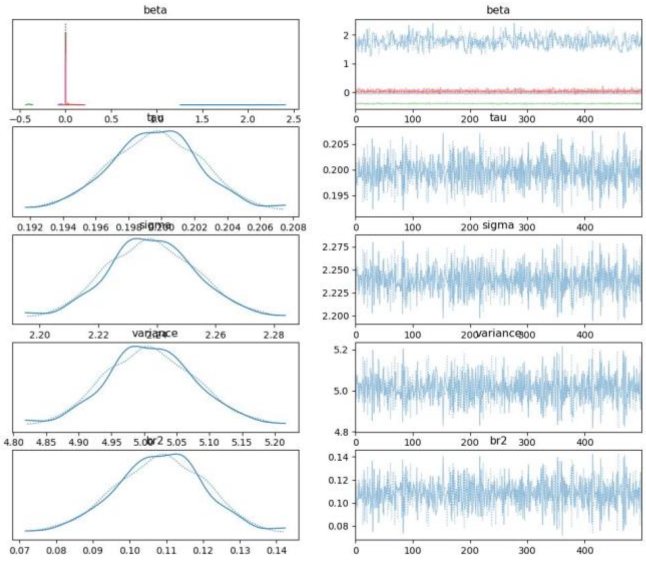
Also like my exponential model, I provide a table of the coefficient values and R2 obtained from these two models with and without zero density data.

	No Zeros and No Null Data	Zeros and No Null Data	No Zeros and Null Data	Zeros and Null Data
R2	0.13	0.108	0.131	0.176
Intercept	1.768, (1.413,2.152)	1.813, (1.39, 2.145)	1.779, (1.406,2.108)	1.793, (1.436,2.195)
C_Year	0.048, (0.039,0.057)	0.028, (0.018,0.038)	0.048, (0.039,0.057)	0.028, (0.018,0.038)
C_PlantType	0.432, (0.397,0.463)	-0.401, (-0.428,-0.376)	0.432, (0.4,0.465)	-0.401, (-0.428,-0.376)
C_BioclimaticZone	0.052, (-0.021,0.129)	-0.058, (-0.025,0.133)	0.054, (-0.029,0.121)	0.057, (-0.019,0.139)
C_MeanTemp	-0.039, (-0.06,-0.018)	-0.05, (-0.07,-0.033)	-0.038, (-0.057,-0.019)	-0.051, (-0.068,-0.031)
C_MeanGrowingDays	0, (0,0)	0, (0,0)	0, (0,0)	0, (0,0)
C_MeanPrecipitation	0, (0,0)	0, (0,0)	0, (0,0)	0, (0,0)
Sigma	1.722, (1.691, 1.75)	2.239, (2.209, 2.271)	1.721, (1.693, 1.754)	2.239, (2.208,2.271)

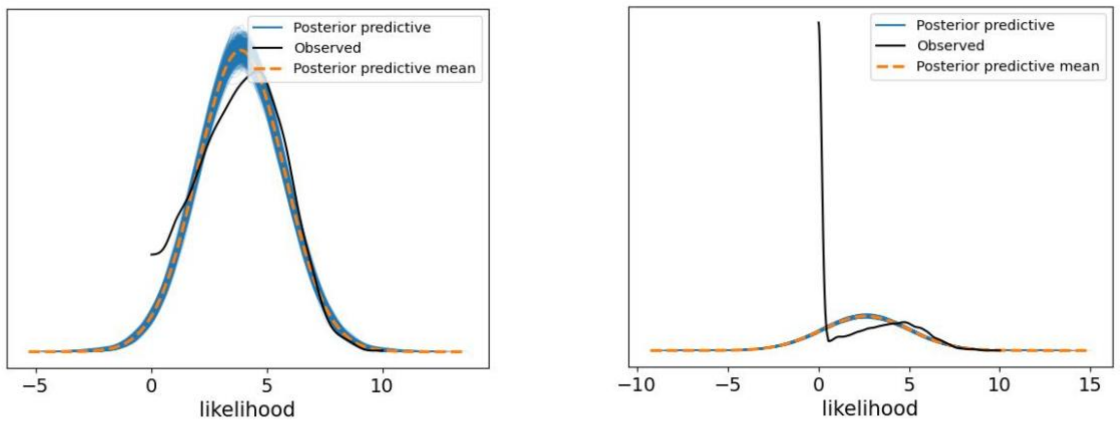
Again, I chose not to report my null data estimations since they are a large quantity. I visualize a trace of models here as well as a posterior computed and observed graph. Again, I do not report the null data models since it was too compute intensive given the number of null data entries.



Without zeros



With zeros



The left graph is for without zeros model. The right graph is for with zeros model.

Discussion

Numerical Results

After obtaining my results, there are a few interesting factors that I have discovered. One of the first things to note is that the normal and exponential models although produce different numerical results, show the same general trends in the variables. However, the exponential models have a higher R2 value, indicating their fit was better. This can be expected.

It is important to analyze my coefficients. In all my different models, the mean growing days and mean annual precipitation have values of around 0 as their coefficients, with a strong 95% credible set around 0 too. This was somewhat surprising. Additionally, it was great to see that for both my models with null data, I was able to obtain estimates of the unknown values.

Analysis of Biomass Density versus Time

On the other hand, the focus of this study was to look at time and its effect on biomass densities. It was shown in my normal models, I have a positive coefficient for my year. This implies that as my year numerically increases, I get a higher average biomass density, indicating that biomass is growing over time. In my exponential models, I also see the similar behavior. As my years increase numerically, I obtain a more negative output for my linear function. However, after applying my link function to get my rate parameter (negative reciprocal), I see that my value decreases and approaches zero. This means my exponential distribution is stretched further out, indicating again a general increase in biomass density over time. I also have strong 95% credible set around my year coefficients, giving us more confidence in this result. As stated in my introduction, this follows studies from literature that over time global warming is causing growth in arctic biomass [3].

Potential Shortcomings

The presence of the vast amount of zero density values was quite an interesting factor in my models. As I can demonstrate, the R2 values for models including zero density was slightly lower than the models without zero density data. This is a good sign of correlation, but there are a few variable coefficients that have a different value because of the change between zero density and no zero density models. In particular, the variable coefficients of plant type, bioclimatic zone, and mean temperature had shifts between zero and no zero models. Perhaps this is an indicator that zero density data that was collected were focused on by particular data collecting groups, since this dataset is an aggregation of a lot of separate groups.

On the topic of R2 values, in all my models these values are low. Perhaps this can be explained by a couple of factors, including the ones listed in the paper describing the dataset. One factor in general is that biomass densities vary a lot in any region, and additionally there are quite a few possible sources of error in the process of data collection that the data collectors have identified and warned about in their paper. Another potential source of issue can be the priors. Although the priors were somewhat informed, the priors could be even more informed in the presence of more knowledge in the domain. For example, perhaps some variables like year can have a positive or negative mean instead of a neutral one had there been specific knowledge on this subject. Another interesting factor that can skew results is that there is better data coverage in recent years compared to previous years. This can present biases to the dataset, and thus skew results in favor of recent years data.

Conclusion

The trends of biomass in a region are an indicator of climate behaviors. Particularly, the region of the arctic are of particular focus since the effects of climate change are several times greater than anywhere else on the planet. I use The Arctic Plant Aboveground Biomass Synthesis Dataset which provides a wide ranging of aggregated biomass density data in several arctic regions and over time. Using a few variations of data configurations of exponential and normal GLM Bayesian models, I was able to see a correlation from the year and the biomass density. In general, I see an increase in biomass density, which correlates with other studies in climate warming [3].

However, the models itself have low R2 scores. This can be due to several factors including in general the variance of biomass density, as well as several factors of errors identified in the authors of the dataset. Perhaps future work can include gathering more informed priors or further data cleaning to rerun these models and see potentially better results.

Citations

1. Downing, A. & Cuerrier, A. A synthesis of the impacts of climate change on the First Nations and Inuit of Canada. Indian Journal of Traditional Knowledge 10, 57–70 (2011).
2. Chapin, F. S. 3rd et al. Role of land-surface changes in arctic summer warming. Science 310, 657–660, https://doi.org/10.1126/science.1117368 (2005).
3. Berner, L. T. et al. Summer warming explains widespread but not uniform greening in the Arctic tundra biome. Nature Communications 11, 4621, https://doi.org/10.1038/s41467-020-18479-5 (2020).
4. Rantanen, M. et al. The Arctic has warmed nearly four times faster than the globe since 1979. Communications Earth & Environment 3, 168, https://doi.org/10.1038/s43247-022-00498-3 (2022).
5. Berner, L.T., Orndahl, K.M., Rose, M. et al. The Arctic Plant Aboveground Biomass Synthesis Dataset. Sci Data 11, 305 (2024). https://doi.org/10.1038/s41597-024-03139-w
6. Abril-Pla O, Andreani V, Carroll C, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. PeerJ Comput Sci. 2023;9:e1516. Published 2023 Sep 1. doi:10.7717/peerj-cs.1516.
7. Kumar et al., (2019). ArviZ a unified library for exploratory analysis of Bayesian models in Python. Journal of Open Source Software, 4(33), 1143, https://doi.org/10.21105/joss.01143