SCC Client Report
Methods for the Analysis of Nitrogen Fixation Time Series Data

November 16, 2023

1 Background

Nitrogen is an essential building block of organic molecules and thus the biosphere must maintain healthy nitrogen levels. Inorganic nitrogen is abundant in the atmosphere as N_2 and can combine with other elements to form more reactive compounds (i.e. ammonia, nitrates and nitrites). These compounds can be converted into nitrogen's organic form (i.e. bounded with carbon, oxygen and hydrogen, typically) through industrial (i.e. Haber Process) or biological means. Biological nitrogen fixation ("BNF") occurs typically in soil or aquatic systems with the help of diazotrophs, which are microorganisms that fix gaseous inorganic nitrogen into organic nitrogen. The mechanics behind their BNF and the conditions best suited for such procedures vary widely in both diazotrophs in soil and marine diazotrophs; however, this discussion only focuses on the latter. Within marine ecosystems, open ocean tropical and subtropical regions (oligotrophic areas and tropical regions) are believed to have higher BNF rates; however, the coastal and polar regions remain understudied and undersampled. Thus, you have collected BNF rate data in understudied regions on four scientific expeditions, where you first used nitrogen gas incubation and then the FARACAS measurement method on subsequent expeditions. In addition to the BNF rates measured, each exploration vessel collected an abundance of auxiliary data. While observing charts of BNF rates across various expeditions in relation to time-of-day and light intensity, certain patterns became apparent, suggesting diel cycle periodicity where higher BNF rates align with increased light intensity, or on certain expeditions, with decreased light intensity. Thus, you want to identify the best method to quantify the significance of periodicity in BNF rates over diel cycles.

2 Recommendations

The recommendation entails utilizing generalized additive models in order to capture variability and account for heteroskedasticity. This paper will first provide an overview on generalized additive models and the applicability to the client data and objectives. Following this overview, there will be an explanation of spline bases and recommendation on spline selection. It is recommended to use the `mgcv` software package. An overview of key functions, model fitting and model comparison will be provided. Next, extension of generalized least squares models will be described including detection and categorization of residual heteroskedasticity and temporal patterns of correlation in residuals. In addition, adjustments to mixed models and generalized additive models to account

for heteroskedasticity and residual autocorrelation will be reviewed, as well as the process for evaluating competing formulations. Next, a step-by-step guide to construct a model for the nitrogen fixation data will be detailed. Lastly, alternative approaches to analysis including periodograms and wavelet-based time-scale decompositions will be evaluated.

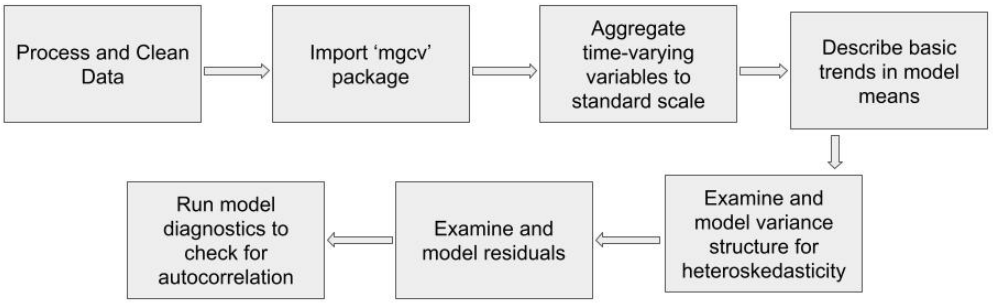

Figure 1: Graphical depiction of the proposed analysis pipeline.

2.1 GAMs, smoothing splines and the R Package `mgcv`

Generalized additive models (GAMs) are an extension of linear regression which allow for much greater flexibility and modelling of non-linear relationships. They provide interpretable results and are able to model seasonal effects and trends, making GAMs a powerful tool for analysis of complex environmental data.

2.1.1 High-level description of GAMs and Spline Bases

GAMs allow us to move past the assumption that the relationship between our predictors (in your case, most notably time) and our response (BNF) is strictly linear. In a normal linear regression model, you can illustrate the predictor as having a constant slope; its relationship with the response does not change as the value of the predictor changes. With GAMs, the slope for each predictor need not be constant. As such, the impact of predictors is expressed through a series of (not necessarily linear) smooth functions. A more detailed outline of GAMs can be found in Hastie (2017).

To build the components of GAMs, use smoothing functions known as splines. Think of splines as piecewise functions; they allow for the modeling of nonlinear relationships by connecting a series of curves through several knots (connector points). There are a variety of spline types you can choose from to model the relationship between time and BNF. Polynomial splines offer different degrees of flexibility, with common choices being cubic splines and natural cubic splines. Cubic splines are versatile and simple but can exhibit instability near data boundaries. Natural cubic splines address this issue by providing stability but may over-smooth the data. See "Polynomial Interpolation" in De Boor and De Boor (1978). B-splines use basis functions associated with control points to create the final spline. Control points are specific points that influence the shape and behavior of the spline, allowing for adaptability and versatility in modeling various scenarios. B-splines are known for their adaptability, with control point adjustments having a limited impact on the overall curve. They are not constrained to specific data points, making them versatile for various modeling scenarios. See "B-Splines" in De Boor and De Boor (1978). Smoothing splines are ideal for data with inherent smoothness, reducing noise through a smoothing parameter. This choice of spline type allows users to control the trade-off between model complexity and goodness of fit. P-splines combine the flexibility of B-splines with a penalty parameter from smoothing splines. They capture non-linear relationships while the penalty term controls curve smoothness and prevents overfitting. See "Introduction" in Eilers and Marx (2021). In the subsequent sections, you will explore the practical application of these spline types in our analysis, considering their suitability for modeling the relationship between time and BNF.

2.1.2 Using the mgcv package

Within the `mgcv` package, there are 3 functions that can create GAMs (Wood, 2023). The function `gam()` creates a Generalized Additive Model. The function `gamm()` creates a Generalized Additive Mixed Model, used for clustered data. The function `bam()` creates a Generalized Additive Model that is specialized for big datasets. See gam, gamm, and bam in Wood (2023)

To create splines for individual predictors, use `s(pred, fx = FALSE, k = -1, bs = "cr")` within the predictor section of the relevant GAM function. See s in Wood (2023)

Note that `bs` specifies the type of spline, for example use `bs = "cr"` for cubic regression splines. See smooth.terms in Wood (2023) `k` is the smoothing parameter, the upper limit on the degrees of freedom associated with an s smooth. See choose.k in Wood (2023). `fx` controls whether it is a fixed regression spline (TRUE) or a penalized regression spline (FALSE). See s in Wood (2023)

Some types of splines to consider include cyclic cubic splines, `bs = "cr"`, for cyclical data such as our nitrogen data. You could also consider thin plate splines, `bs = "tp"` , that allow for interactions between predictors. See smooth.terms in Wood (2023)

An example of an implementation is:

`gam(y  s(x) + ns(z), family = gaussian(), data = list(), weights = NULL, subset = NULL, na.action, method = "GCV.Cp", knots = NULL` See gam in Wood (2023)

Some other things to consider are parameters such as:

`family`: Default is gaussian, but can changed for different responses. See family.mgcv in Wood (2023) for technical guidance on family choice see pages 62 and 63 in Wood (2017).

`knots`: Specify where you want the knots to be. Default means knots are selected by R. See the knots parameter under gam in Wood (2023).

`method`: There's a penalty method employed when running the function and there are a wide variety of different penalties. Discussion regarding the methods can be seen in Section 1.8.5 of page 50 of Wood (2017).

To further discuss the 3 methods, the default is Mallow's Cp, but the function includes REML or NCV. Thorougher explanation of these penalty method options are found in Wood's package document Mallow's Cp is a penalization to the Residual Sum of Square Error by adding a penalty based on the number of predictors. REML is a likelihood-based method used to estimate model parameters while accounting for the variance-covariance structure of the data. On the other hand, NCV refers to "Neighbourhood Cross Validation", a technique for estimating smoothing parameters by optimizing the average ability of a model to predict "neighboring" subsets of data that have been omitted from fitting. See the method parameter under gam in Wood (2023)

For model comparison, you can run a Analysis of Variance, or ANOVA, within the mcgv package. See anova.gam in Wood (2023) To model select for the smoothness parameter, select for the `method` parameter, as discussed above. For trade-offs between REML, NCV, and GCV.Cp, see gam.selection in Wood (2023).

To model select for automatic term selection, there are 2 options. First, you could add an additional shrinkage term. Alternatively, you could add an additional penalty for each smooth. For more details, , see gam.selection in Wood (2023).

To model select for interactive terms, candidates for removal can also be identified by reference to the approximate p-values. For more details, see summary.gam and gam.selection in Wood (2023).

2.2 Generalized Least Squares (GLS) Model Extensions

2.2.1 How is residual heteroskedasticity detected and characterized?

Residual heteroskedasticity refers to the inconstant variance of residuals. The easiest way to detect heteroskedasicity is to create a plot of the covariate and the same summaries, eg. day, and look for the distinctive non-constant shape in the data. One can also create a Scale-Location plot, where the square root of the standardized residuals are plotted with the predicted values. If there is a pattern in this plot, then residual heteroskedasticity is detected. There are also a few statistical tests that can be used to detect heteroskedasicity, including, but not limited to, the Breusch-Pagan test. The Breusch-Pagan test is a hypothesis test that uses a variance function and an χ^2-test with the null hypothesis of constant variance. Overall, each of these methods are used to detect the variation in the variance for all observations in a dataset (Frost, 2021).

2.2.2 Temporal Patterns of Correlation in Residuals

Temporal patterns of correlation in residuals can have a significant effect on the accuracy and reliability of the model, and is an important factor to consider when modeling time-series data. An effective tool to detect and characterize these temporal patterns is the variogram. The variogram is a statistical tool to describe the behavior of spatial, dynamic, and random processes. The variogram serves as a metric to quantify dissimilarity by examining pairs of data points separated by specific lag distances. You have the flexibility to choose how to measure these lags based on the characteristics inherent to the type of data you are analyzing, ranging from the distance between observations, the time difference between two data points, or even the number of observations between two points for irregularly-spaced data. Regarding temporal variograms, they explore how the variance of the time series changes as the time lag increases. If there is a strong autocorrelation, you can expect the variance to increase slowly with larger lags, and vice versa. The variogram will generally illustrate this relationship, often resembling a curve that levels off as time lags become larger, which in turn indicates decreasing autocorrelation (Zuur, 2009).

You can calculate the variogram using the standard equation or the weighted least squares (WLS) method, along with other types of spatial interpolation - proving especially helpful when the data is irregularly spaced. Computationally, the `variogram()` function in R allows users to build, compute, and plot variograms, which can be extremely useful for temporal datasets. Before using the variogram function for temporal data in R, you need to organize your time series data and define the temporal lags / distance metrics that you are using to compute the variogram. The function takes the time series data, the time lag values, and other optional parameters for customization as inputs. After calculation is complete, you can inspect the variogram's shape / form for insights about the temporal or spatial correlation structure.

Overall, the variogram serves as a powerful statistical tool in understanding the temporal pattern of correlation of residuals, especially providing insights into time-series data and the noise present within it. In the context of biological research and sampling, you can apply these concepts to infer trends within data, such as environmental conditions or traits that exhibit continuity over time, or even external factors that lack a discernible trend. The application of variograms to biological nitrogen fixation can offer valuable insights into the temporal dynamics and spatial trends associated with these broader processes and the other factors present in the data.

2.2.3 Accounting for heteroskedasticity and residual autocorrelation

In terms of implementing Generalized Least Squares Models (which extends the linear regression by modelling the heterogeneity with covariates, in the case that the homogeneity assumption is violated), we recommend R packages `nlme` and `mgcv` (Zuur, 2009). The `nlme::lme()` function is used for fitting linear mixed-effects models, which accommodates for both fixed effects (typically called "predictor variables" in simple linear regression) and random effects, such as random deviations of groups. The `lme` function specifically allows for complex variance structures, including different levels of variability at different levels of grouping. In essence, fixed effects are analogous to the predictor variables in linear regression, assumed to be constant across all individuals or groups in the study, whereas random effects account for variations that are not captured by the fixed effect since random effects are assumed to be different for individual or group differences in the response.

The `mgcv::gamm()` function in R is used for fitting Generalized Additive Mixed Models (GAMMs), allowing for non-linear relationships between predictors and responses using smooth terms. It effectively handles heteroskedasticity using the `weights` argument for different variance levels, and autocorrelation by incorporating correlation structures within random effects via the the `correlation` argument. This is particularly beneficial for data with temporal or spatial dependencies.

Similarly, `nlme::lme()` is employed for fitting linear mixed-effects models. It also uses the `weights` argument to manage heteroskedasticity by specifying variance functions via the `varFunc` object. Additionally, `lme()` employs a `corStruct` object to address within-group autocorrelation, ensuring that data dependencies and variability in residuals are adequately considered in the analysis. Both functions share syntax elements for weights and correlation, providing flexible frameworks to address data-specific needs in modeling complex error structures.

2.2.4 Evaluating Competing Formulations

Evaluating competing formulations involves several key steps:
1. Understanding The Base Model And Its Residuals. Gain a true understanding of the base model and its structure and be sure to examine its residuals.
2. Identifying Temporal Correlation in Residuals. Plot the residuals over time to help formulate an understanding of temporal covariation and reveal trends and seasonality.
3. Formulating Competing Models. Based on the results of step 2, formulate alternative models by examining the temporal structure, correlation structures, and variance models.
4. Use Of GOF Tests And Comparing Non-Nested Models. Using tools like ANOVA in mgcv, check goodness-of-fit in nested models. If the models are not nested, using AIC will help.
5. Use Of Variance And Covariance Models. Consider if variance or covariance structures can improve the performance. Errors may be correlated over time in time series data and have non-constant variance.
Overall, the process involves a careful analysis of the base model, identification of potential improvements, formulation of competing models, and comparison using appropriate statistical criteria to select of final choice of model.

2.3 Overview of Recommended Data Analysis Approach

2.3.1 Data

In constructing a model for the nitrogen fixation data using the mgcv::gamm() function, first and foremost, you would need to focus on data preparation. You can start by temporally aggregating the time-varying variables, including nitrogen fixation rate (the response variable) and other relevant covariates, such as salinity, oxygen level, temperature, depth, coastline proximity, species distribution, and light intensity, to a common frequency, perhaps on a minute scale. Following this, you would then create derived variables, including time-of-day on a 0-24 scale and day-of-year represented both nominally as an integer and continuously. Assemble these variables into a cohesive data matrix for further analysis.

2.3.2 Model Mean Structure

Moving on to the mean structure of the model, you can initiate the process with a daily cyclic spline (`bs = cc`). This initial model would help capture inherent daily patterns. Additionally, you would then introduce an additive, unconstrained day-of-year random effects spline (`bs = re`) to account for potential random variations in the data explained by individual days. As you progress, consider adding relevant covariates to the model and assessing their impact on the overall fit.

As an alternative approach, you can explore incorporating a smooth day-of-year spline (`bs = tp`) into the model. Compare the model fit with the daily cyclic spline to determine if the smooth spline offers improvements, particularly in capturing nuanced variations.

In the subsequent stage, you can experiment with a joint model that includes an interaction term between time-of-day and day-of-year using the `te()` function. This approach aims to capture both daily and long-range variations in the nitrogen fixation data and allow patterns in daily cycling to vary slowly over time and space.

2.3.3 Model Variance Structure

In the third step of your GAM analysis using `mgcv::gamm()`, the primary focus is on examining the residuals for signs of heteroskedasticity. After fitting the initial model, it's crucial to analyze the residuals to check if their variance remains constant across different levels of fitted values. This involves creating and inspecting residual plots against the fitted values and each predictor variable. In these plots, patterns such as funnel shapes or a widening spread in residuals at higher fitted values are indicative of heteroskedasticity (R Core Team, 2020). Alternatively, boxplots are a concise and efficient method for comparing distributions across categorized variables with less systematic load. While visual inspection of these plots is a key method for identifying non-constant variance in residuals, supplementing this analysis with statistical tests like the ANNOVA LRT test can provide a more formal assessment for nested alternatives. If these evaluations reveal that the variance of the residuals is not constant, it signals the need for explicitly modeling this variance to improve the accuracy and reliability of your model.

Upon identifying heteroskedasticity in your GAM, the next critical step is to model the variance structure. This involves selecting an appropriate variance function that can be integrated into your model using the `mgcv::gamm()` function. The choice of this variance function, which is typically incorporated through the 'weights' argument in `gamm()`, depends on how the variance of the residuals relates to the predictors or the fitted values. Common approaches include using functions like

`varExp`, `varIdent`, or `varPower`, which allow the variance to change in relation to a predictor or the fitted values in different ways, such as exponentially or as a power function (Pinheiro, 2000). For example, if you suspect that variance is a function of a predictor, you might use `weights = varExp(form = predictor)`. This tells the model to allow the variance to change exponentially based on the values of the predictor. Implementing this variance structure often requires an iterative approach, where the model is repeatedly fitted with different variance functions, each time checking the residuals to ensure that the heteroskedasticity has been adequately addressed. Once a satisfactory variance model is established, the GAM is refitted, hopefully improving the reliability of parameter estimates and standard errors. This iterative process is essential to refine the model, ensuring that both the mean and variance structures accurately represent the underlying ecological processes in your nitrogen fixation data study.

2.3.4 Model Residual Correlation

In the final stage of GAM analysis using the mgcv::gamm() function, the process pivots to a thorough evaluation of the residuals, focusing on uncovering any temporal autocorrelation. After the initial Generalized Additive Mixed Model (GAMM) is fitted, a crucial step is to meticulously examine the residuals for patterns or trends indicative of temporal dependencies. This examination is achieved not only through the creation of scatterplots of residuals against the fitted values and key predictor variables but also by employing comprehensive diagnostics, such as Durbin-Watson test or Breusch-Godfrey Test.

Supplementing visual inspection with statistical tests enhances your ability to quantify and validate temporal dependencies in the residuals. If evidence of temporal autocorrelation is found, it signals the need for adjustments to the model structure to account for this correlation and improve its accuracy.

Upon identifying temporal autocorrelation in your GAMM, the subsequent critical step is to address and model the residual correlations. This involves modifying the model to incorporate components that explicitly capture the temporal structure in the residuals. Potential adjustments include introducing additional smooth terms, altering correlation structures, or incorporating temporal components such as lagged terms.

Similar to modeling variance structure, addressing temporal autocorrelation often requires an iterative approach. The model is repeatedly fitted with different adjustments to the structure, and the residuals are examined after each iteration to ensure that temporal autocorrelation has been appropriately addressed. This iterative refinement process is crucial to enhance the model's ability to capture and represent the temporal dependencies in the data accurately.

2.4 Alternate Approaches: Promise and Shortcomings

2.4.1 Brief introduction to periodograms

A periodogram is a graphical data analysis technique used to identify dominant periods (or frequencies) in a time series, particularly when the cycles are hidden within seemingly random data.

The Lomb-Scargle periodogram is a well-known variant of the traditional Fourier periodogram, used specifically for the frequency analysis of unequally sampled data, including regular time-series with missing data (VanderPlas, 2018). It fits a sinusoidal model (i.e., sine waves) to the data at each frequency and then evaluates the power (or goodness-of-fit) at each frequency to generate a power spectrum. A larger power reflects a better fit, so peaks in the periodogram correspond to the frequencies that are most likely to represent real periodic phenomena in the data.

Although the Lomb-Scargle periodogram has the notable advantages of being able to handle unevenly spaced data and being computationally efficient, it is important to note that it requires stationarity (Ruf, 1999). A stationary time series is one whose properties do not depend on the time at which the series is observed (i.e., the time series cannot have trends or seasonality). Additionally, the Lomb-Scargle periodogram relies on the relatively strong assumption that the underlying pattern is sinusoidal, so it may not accurately capture more complex forms and can potentially yield misleading estimates (VanderPlas, 2018). The reliance on sinusoidal models also means outliers and irregularities can significantly impact the results to yield false positives or inaccurate estimates. Data should be preprocessed to remove noise and outliers. Lastly, determining the statistical significance of peaks in a periodogram may require a more thorough understanding of the noise characteristics and often involves the use of Monte Carlo simulations or other resampling techniques.

2.4.2 Wavelet-based time-scale decompositions

Wavelet analysis is a powerful tool for time-series analysis used throughout science and engineering. It has been particularly utilized for analysis of ecological time patterns, in areas such as climate, epidemics, and population dynamics.

In most ecological data, the data is non-stationary - they have varying dynamics and scales with time, and the data is transient, meaning that it has short-term, abrupt fluctuations. This phenomena could look like the impact from a wildfire or disease outbreak. Wavelet analysis accounts for these properties by utilizing a local time-scale decomposition. It estimates the spectral characteristics as a function of time, allowing one to identify the different scales of the spectral characteristics over time, which is useful because ecological time series data often have complex variations that occur at different time scales. For example, there may be a short-term variation from daily weather changes, or there may be long-term variation from seasonal trends. Wavelet transformations allow the separation of these different scales of variability so that you can study both ecological processes.

"Wavelet analysis of ecological time series" published by Cazelles details wavelet transformations, specifically in ecological time series, which informed this discussion. Wavelet transformation is a mathematical function that allows you to analyze signals in both time and frequency domains. It decomposes a time series into its constituent wavelet components at different scales. The transform works by sliding the wavelet function along the time series and measuring the similarity between the wavelet and the local features of the data at various scales.

The Continuous Wavelet Transform (CWT) takes translation parameters a and b, where 'a' is the scaling factor (determines the width of the wavelet) and 'b' is the translation (determines th exposition along the time axis). The Discrete Wavelet Transform (DWT) is a more computationally efficient alternative to CWT. In the DWT, the time series is repeatedly divided into segments and analyzed at discrete scales. Mother wavelet function is the determining property to choose because the shape impacts the sensitivity to different frequency components. The scalogram is the output of the CWT. It is a plot that shows intensity and magnitity of the wavelet coefficients at each scale and time position, which indicates the strength of specific frequency components.

Obstacles for this nitrogen fixed data include data preparation, methodological choices, and complexity of statistical analysis. The main shortcoming of this method for our data is that it is unable to analyze data with missing values. The lack of continuity will cause edge effects. The data quality is important - noisy data, data with missing values, or data with measurement errors will impact the accuracy of the analysis. In this case, synchronizing the data will also be a challenge, since there are multiple different time series - comparing the intervals will be difficult. This spatial variability and sparsity drive obstacles in the data. Next, methodological choices have a significant impact on results. Selecting different mother wavelet function and scales can lead to a different analysis. Nitrogen fixation can have strong seasonal and temporal variation, so choosing the relevant scale will be a challenge in order to capture a meaningful pattern. Lastly, the complex statistical analysis is required and inability to reveal underlying ecological mechanisms is a shortcoming. Wavelet analysis is computationally intensive and may require significant resources. Analysis can also be complex and require expertise in both wavelet mathematics and specific ecological context in order to determine relevance and significance of the findings. Wavelet analysis also does not provide insights to underlying ecological mechanisms, and different mechanisms can generate similar patterns of associations between series. In this case, we would not be able to use wavelet analysis to prove relationship between light and fixation - the analysis could show correlation but not causation.

2.4.3 Any other alternatives?

There are several other alternatives that should be looked into in this analysis of ecological time series data. One alternative approach that can specifically target ecological-based data is a Hidden Markov Model, which has the ability to catch obscured or hidden unobservable modes and their transformations in time series data. This is useful for distinguishing specific ecological processes that contain concealed dynamics. Another alternative approach is using a change point detection algorithm, which is beneficial for recognizing sudden structural transitions and movements within time series data. Using this method would lead to more specific results that contain distinct points when the measured ecological processes are going through crucial structural changes. The last alternative approach you could try is using a Bayesian Time Series Model. You should use this approach if the data contains a substantial amount of uncertainty when estimating parameters, as the Bayesian model will allow you to include and account for prior information and continuous updating when new data is received. When deciding which alternative approaches to try, you should account for the specific characteristics and qualities of your ecological time series data rather than just assuming all of these approaches will work for your distinct set of data.

3 Technical Writing Instructions – Remove at Final Revision

NOTE 1: Here is an example of how to cite a reference in–line and have it appear in the references section:

We propose that you use the R programming language (R Core Team, 2020) for your analysis.

Note that there is a BibTex citation entry call 'Rcitation' in the bibliography file SCC.bib (see menu on left of OverLeaf Screen). Most online links to journal articles will have a link that provides citations in a variety of formats; usually BibTex is an option. There are different entry formats for books, journal articles, software manuals and online links, so be sure to use the appropriate format.

NOTE 2: Please use typewriter font (as shown `here`) for references to `R`, `R` functions and variable names.

References

De Boor, C. and C. De Boor (1978). *A practical guide to splines*, Volume 27. springer-verlag New York.

Eilers, P. H. and B. D. Marx (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge.

Pinheiro, J. C., . B. D. M. (2000). *Mixed effects models in S and S-plus*. Springer.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ruf, T. (1999). The lomb-scargle periodogram in biological rhythm research: Analysis of incomplete and unequally spaced time-series. *Biological Rhythm Researc 30*.

VanderPlas, J. T. (2018). Understanding the lomb–scargle periodogram. *The Astrophysical Journal Supplemental Series 236*(1).

Wood, S. (2023, Jul). Package mgcv.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*. CRC Press/Taylor amp; Francis Group.