USING PLANKTON EDNA TO ESTIMATE WHALE ABUNDANCES OFF THE CALIFORNIA COAST: DATA INTEGRATION AND STATISTICAL MODELING

A Thesis

presented to

the Faculty of California Polytechnic State University,

San Luis Obispo

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Statistics

by

Katherine Chan

June 2024

COMMITTEE MEMBERSHIP

TITLE: Using Plankton eDNA to Estimate Whale Abundances off the California Coast: Data Integration and Statistical Modeling

AUTHOR: Katherine Chan

DATE SUBMITTED: June 2024

COMMITTEE CHAIR: Trevor D. Ruiz, Ph.D.
Assistant Professor of Statistics

COMMITTEE MEMBER: Ulric Lund, Ph.D.
Professor of Statistics

COMMITTEE MEMBER: Lauren Chan, Ph.D.
Associate Professor of Biological Sciences

ABSTRACT

Using Plankton eDNA to Estimate Whale Abundances off the California Coast: Data Integration and Statistical Modeling

Katherine Chan

Understanding marine mammal populations and how they are affected by human activity and ocean conditions is vital, especially in tracking population declines and monitoring endangered species. However, tracking marine mammal populations and their distribution is challenging due to difficulties in observation and costs. Using surrounding plankton environmental DNA (eDNA) has the potential to provide an indirect measure of monitoring cetacean abundances based on ecological associations. This project aims to apply statistical methods to assess the relationship of visual abundances of common species of baleen whales with amplicon sequence variants (ASV) of plankton eDNA samples from the NOAA-CalCOFI Ocean Genomics (NCOG) project. Modeling this relationship of eDNA with marine mammal sightings may greatly aid the ability to predict the abundance of whales in the ocean.

There are several key challenges associated with the analysis of this NCOG data. Plankton eDNA samples are an example of compositional data, where the proportions of each ASV must sum to one; this provides a challenging constraint for statistical analysis and interpretation. High dimensionality (the number of parameters exceeds the observations) and sparsity (many observed zeros) of the genetic sequencing data also pose challenges in estimating parameters. Finally, the model associations should be adjusted for related factors, including seasonality and oceanographic factors, the latter of which goes beyond this project's scope.

This thesis develops and fits models to estimate cetacean abundance from plankton eDNA by leveraging methods of compositional data analysis and high-dimensional

regression. This project applies log-ratio data transformations and corresponding log-contrast models to address the compositional aspect of eDNA reads. Regression methods involving high-dimensional data typically rely on dimensionality reduction or regularization. This project implements both reduction and regularization through sparse partial least squares (sPLS) regression. In addition to the data modeling objective of using plankton eDNA to predict baleen whale abundances, this project also identifies ecological correlations between whale abundance and plankton eDNA.

Keywords: marine mammal, cetacean, abundance estimation, sPLS, eDNA

ACKNOWLEDGMENTS

Thank you to my thesis advisor, Dr. Trevor Ruiz, for their patient guidance, support, and encouragement throughout this journey.

Thank you also to my committee members, Dr. Lauren Chan and Dr. Ulric Lund for their expressive willingness, positivity, and excitement to join me in this thesis process. Thanks for your valuable and thorough revisions, input and feedback.

Thank you to those at Scripps Institution of Oceanography for providing initial guidance for this project. Thank you to Erin Sattethwaite, Michaela Alksne, Nastassia Patin, Brice Semmens, Julie Dinasquet, and Simone Baumann-Pickering.

Thank you to Nick Patrick, Gabrielle Low, and Samantha Ward for their research efforts alongside my thesis. I've appreciated our conversations and your bright personalities.

Thank you to my family: Mom, Dad, Christopher, Zachary, and Chantelle, for your love and care, always.

Thank you to my housemates who have believed in me, cheered me on, and celebrated this journey with me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Over the past 10-15 years, high-throughput sequencing of environmental DNA (eDNA) extracted from soil, water, sediment, and other media has become widely adopted to study communities. It presents both opportunities for discovery alongside novel challenges for methodology and data analysis (Rees et al., 2014; Shokralla et al., 2012; Bohmann et al., 2014; Ruppert et al., 2019). Environmental DNA refers to DNA extracted from an environmental sample without first isolating any target organisms, characterized by a combination of shed cellular material from many different organisms and possibly degradation of DNA molecules (Taberlet et al., 2012). Single species detection using eDNA has previously been applied to aquatic ecosystems (e.g., Ficetola et al. (2008)) and continues to be a common alternative to traditional surveys, especially for rare and endangered species. Increasingly, studies have demonstrated positive correlations between eDNA abundance, including metabarcoding data, and abundance estimation using traditional tools, such as quantifying abundance associations in fish and amphibian mesocosms (Evans et al., 2016).

Environmental DNA can be analyzed through different methodologies to achieve various research purposes. Several studies have used eDNA to monitor marine mammals, including single species detection via barcoding, biodiversity assessment via metabarcoding, and genetic characterization within a species (Foote et al., 2012; Riaz et al., 2011; Suarez-Bregua et al., 2022). eDNA studies typically use either barcoding or metabarcoding analysis methods. Quantitative PCR (qPCR) or digital droplet PCR (ddPCR) for barcoding uses species-specific primers that can be used to detect eDNA from a single target species and quantify target-species DNA abundance. Metabar-

coding uses primers that target genome regions for taxon specificity but sufficient variation for species distinction, allowing for simultaneous identification of multiple taxa in a single environmental sample (Székely et al., 2021). Metabarcoding of eDNA from seawater samples has been used to assess associations with marine species as early as Thomsen et al. (2012) for assessing marine fish biodiversity using eDNA metabarcodes. Morey et al. (2020) investigated a multiple marker approach with metabarcoding of eDNA for biodiversity and multiple species detection, recovering 50% of target species and 80% of target taxa in a closed marine system.

Environmental DNA is primarily used for species detection (presence/absence) and biodiversity monitoring. Abundance estimation approaches based on eDNA present non-invasive, cost-effective, and potentially sensitive alternatives to survey techniques, especially for elusive species such as the harbor porpoise (Parsons et al., 2018), manatee (Hunter et al., 2018), and killer whale (Baker et al., 2018). A limited but growing number of eDNA studies for species abundance estimation has produced mixed results in the reliability of eDNA to assess abundance (Székely et al., 2021). With single-species abundance using qPCR or ddPCR, Yates et al. (2019) found that eDNA concentration could only explain 50% to 57% of the observed variation, on average, of aquatic species abundances in natural environments. This apparent limitation may be partly due to low concentrations of genetic material shed from the species of interest into the surrounding environment.

Despite its rapid increase as a tool for species monitoring, the sampling, processing, and sequencing of eDNA is not without its challenges and limitations. These biases can stem from all steps in the process of eDNA analysis, including degradation of eDNA in environmental conditions, high variability in PCR amplification, which may not match specific target organism's DNA, and incomplete reference databases and taxonomic assignment biases, among others (Adams et al., 2019; Beng and Corlett,

2020). To investigate eDNA degradation, Thomsen et al. (2012) performed an experiment, finding that very small (100-bp) eDNA fragments degraded beyond the detection threshold within a few (0.9–6.7) days for two fish species. Other experiments suggest eDNA in freshwater can persist up to several (2-4) weeks after an organism has been removed from the controlled water environment (Dejean et al., 2011). Notably, oceanographic and local weather conditions significantly impact the possible distance that eDNA is dispersed in aquatic environments (Barnes et al., 2014). While sequencing technologies and sampling methodology raise concerns about biases in eDNA data, collection and analysis of eDNA remains a rapidly developing area with a record of adoption and success across many fields in ecology (Ruppert et al., 2019).

This project seeks to explore the potential of eDNA metabarcoding from species at lower trophic levels to predict cetacean abundance, hypothesizing that ecological relationships may give rise to correlations that can be leveraged for this purpose. This novel approach has the additional potential to identify an "ecological habitat" or set of species whose abundances collectively correlate with marine mammal abundances. This project does not propose specific mechanisms (e.g., feeding, species interaction, habitat use) for ecological associations; instead, it seeks to identify potential correlations as solid groundwork for marine ecologists to investigate underlying ecological mechanisms. The approach presented here integrates two datasets in a correlative/associative analysis: visual estimates of marine mammal abundances and plankton eDNA metabarcodes. Both datasets comprise measurements recorded on monitoring cruises conducted by the California Cooperative Oceanic Fisheries Investigations (CalCOFI); Campbell et al. (2015) describe the methodology for collecting visual estimates and James et al. (2022) describe the methodology for collecting and sequencing eDNA. This current project develops and implements a framework for identifying and estimating associations between plankton eDNA and marine mam-

mal abundances using existing statistical methodology; this approach could be applied to other ecosystems or species of interest.

Statistical analysis of eDNA data — which are typically high-dimensional, extremely sparse, and compositional (*i.e.*, capture relative rather than absolute information) — largely parallels approaches developed for microbiome studies (see, *e.g.*, Li (2015); Zhou et al. (2023); Combettes and Müller (2021); Calle (2019); Tsilimigras and Fodor (2016)). This project, in particular, applies methods for high-dimensional regression with a continuous response and sparse compositional covariates to select and estimate associations between relative abundances of individual plankton species and visual estimates of baleen whale abundances. That is, consider the linear model framework:

$$Y = X\beta + \epsilon \tag{1.1}$$

Here Y is an $n \times 1$ vector representing the response of interest (*e.g.*, whale abundances for a particular species), X is an $n \times p$ matrix representing covariate information (*e.g.*, relative abundances of a large number of gene variants), the model coefficients β represent associations between the covariate set and the response, and ϵ represent random errors. The problem of interest here is fitting such a model when $p \gg n$ and X are compositional, *i.e.*, subject to a row-sum constraint, and doing so while performing subset selection on the columns of X. Each row of X is typically represented as a set of proportions under unit closure, meaning that the rows are normalized to sum to one. This sum constraint arises because the selected explanatory variables must be re-weighted to account for ones dropped during subset selection such that the proportions for each row still sum to one.

For data that is high-dimensional where the number of explanatory variables exceeds the number of observations ($p \gg n$), ordinary least squares regression (OLS) and

similar traditional approaches cannot be applied to estimate (1.1) directly. While there are many approaches to this problem, the partial least squares (PLS) method allows the linear model to be fit with latent variables consisting of linear combinations of the original variables (Rosipal and Krämer, 2005). In other words, instead of fitting the model (1.1) directly, PLS estimates a $p \times q$ linear transformation A with low-dimensional $q \ll n$ and then fits:

$$Y = T\gamma + \epsilon \qquad \text{where} \qquad T = XA \tag{1.2}$$

The original model (1.1) is recovered via the relationship:

$$\beta = A\gamma$$

PLS thus allows for simultaneous dimension reduction and estimation of variable coefficients (Liu et al., 2013), and several computationally efficient algorithms and implementations are available. For such reasons, PLS is recognized as a valuable and versatile tool for modeling genomic data (Boulesteix and Strimmer, 2007).

However, compositional data exhibit a unique geometry due to the row-sum constraint on X (Aitchison, 1982). Applying PLS (or any other statistical or multivariate analysis technique, for that matter) directly to compositional data can produce potentially misleading results due to biases arising purely from the native geometry in which compositional variables are negatively correlated (relative increases in one variable must coincide with relative decreases in another). The class of log-ratio transformations provides a means of mapping compositional data into Euclidean space to facilitate statistical analysis. It forms the basis for a wide range of compositional data analysis methods (Pawlowsky-Glahn and Buccianti, 2011). In the regression

context specifically, Aitchison and Bacon-Shone (1984) introduced log-contrast models that fit model (1.1) after applying a log-ratio transformation to X, and Hinkle and Rayens (1995) proposed log-contrast PLS (LCPLS) as an extension of this approach to fitting the PLS model (1.2).

While LCPLS addresses the issues of high dimensional and correlated compositions, this approach does not automatically result in a selection of only essential variables. Chun and Keleş (2010) introduced sparse partial least squares (sPLS), which uses a regularization framework to filter irrelevant variables during model fitting. This approach constrains the L_1 norm of the rows of the projection A during estimation (*i.e.*, adding the condition $\|a_i^T\|_1 < t$ for some $t > 0$) and results in sparse linear combinations of the original predictors. The sparsity of the projections translates directly to a sparse estimate for the coefficients β, which has the dual advantage of improving model interpretability and predictions.

This thesis applies the estimation method of Chun and Keleş (2010) to the model framework of Hinkle and Rayens (1995) to estimate an LCPLS model relating baleen whale abundances to sparse subsets of plankton eDNA. Separate models are fit for three species: fin whales, blue whales, and humpback whales. The fitted models identify sets of 27, 52, and 39 plankton species, respectively, from 3,248 candidate species, that explain 79-83% of the variation in visually observed whale abundance after adjusting for seasonality in the data. The thesis is organized as follows: Chapter 2 provides a detailed overview of the datasets, integration approach, and statistical methodology; Chapter 3 presents the analysis results; and Chapter 4 summarizes the main findings of the analysis and discusses challenges and potential future steps.

Chapter 2

MATERIALS AND METHODS

This project integrates metabarcoding data from 18SV9 rRNA amplicon sequencing of environmental DNA extracted from water samples with visual estimates of baleen whale abundances and develops a log-contrast modeling framework for predicting whale abundances using a sparse subset of 18S rRNA amplicons for three common species. The measurements in each dataset are co-located in space and time and collected over six years (2014-2020) but recorded at differing spatial resolutions. Thus, in addition to common data processing steps involved in the analysis of genomic data — specifically, zero imputation, conversion to relative abundances, and log-ratio transformations — nontrivial spatial aggregation is required to reconcile differences in resolution and obtain observations on a common spatial scale suitable for fitting statistical models. Furthermore, as observations in both datasets exhibit seasonality over time, removing seasonal patterns prior to modeling is necessary to avoid spurious correlations due to seasonality alone. Finally, an estimation framework amenable to both high predictor dimensionality (many amplicons) and variable selection through sparsity (for interpretability) is required. The analysis presented in this thesis applies sparse partial least squares (sPLS) for this purpose.

This chapter (1) describes the datasets in detail and provides elaboration on how the materials were gathered; (2) reviews methods used to process, aggregate, and integrate the datasets; (3) presents the log-contrast modeling framework for analyzing the integrated data; and (4) describes the sPLS estimation procedure used to fit models along with considerations surrounding hyperparameter tuning.

2.1 Datasets

Each dataset used in the analyses was obtained from the California Cooperative Oceanic Fisheries Investigations (CalCOFI) and the NOAA-CalCOFI Ocean Genomics (NCOG) project [1]. CalCOFI manages a long-term monitoring program that collects large volumes of oceanographic and biological data every quarter from cruises conducted in the California current. Two distinct datasets were utilized:

1. On-transect and on-effort marine mammal sightings
2. Environmental DNA (eDNA) data from seawater samples collected at CalCOFI stations

While each dataset captures distinct information, sampling for both datasets was conducted on the same quarterly CalCOFI cruises from winter 2014 through winter 2020, for a total of 25 cruises. As a result, measurements are co-located in space and time.

2.1.1 Sampling

CalCOFI is a long-term monitoring program established in 1949 that operates off the coast of California. In the summer and fall, CalCOFI has seventy-five stations — specified geographical coordinates — located along the Pacific coast from Point Conception down to San Diego. Time and weather permitting, thirty-eight additional stations off central and northern California are typically scheduled during winter and

[1]Data were obtained from the California Cooperative Oceanic Fisheries Investigations and the NOAA-CalCOFI Ocean Genomics project and are available at https://calcofi.org/, and from the National Center for Biotechnology Information Sequence Read Archive BioProject accessions PRJNA555783, PRJNA665326, and PRJNA804265.

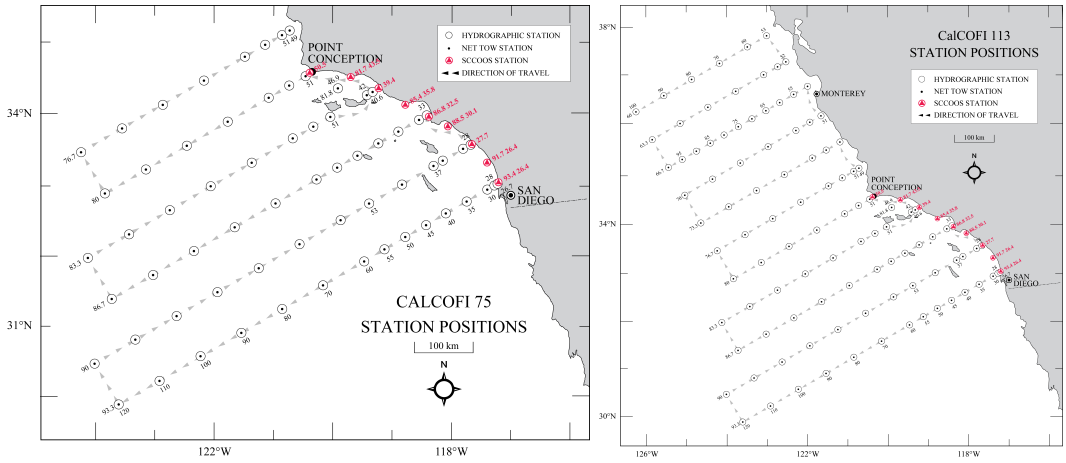

Figure 2.1: CalCOFI sampling grids. A 75-station pattern (left) is utilized for summer and fall cruises; a 113-station pattern (right) is used for winter and spring cruises.

spring. Cruises occur quarterly, during which ships travel along transects connecting CalCOFI stations. The sampling grids showing station locations and indicating direction of travel are shown in Figure 2.1. There is some variability in which stations are visited from cruise to cruise.

While a large volume of oceanographic and marine ecosystem data is collected on CalCOFI cruises, this project focuses on eDNA and marine mammal sighting data. The eDNA data used in this project was obtained from sequencing genetic material extracted from seawater samples collected at CalCOFI stations on each cruise from winter 2014 to winter 2020. Since 1990, CalCOFI has used CTD-Rosette sampling as the primary method of seawater collection; this device is depicted in Figure 2.2. Seawater is collected from various depths at each station in the bottles attached to CTD-Rosette frames; eDNA is extracted, amplified, and sequenced from select samples according to methods described in detail by James et al. (2022). Sightings of marine mammals are recorded at and between stations ("on-effort" and "on-transect") as well as opportunistically. Visual estimates of group sizes are recorded for each

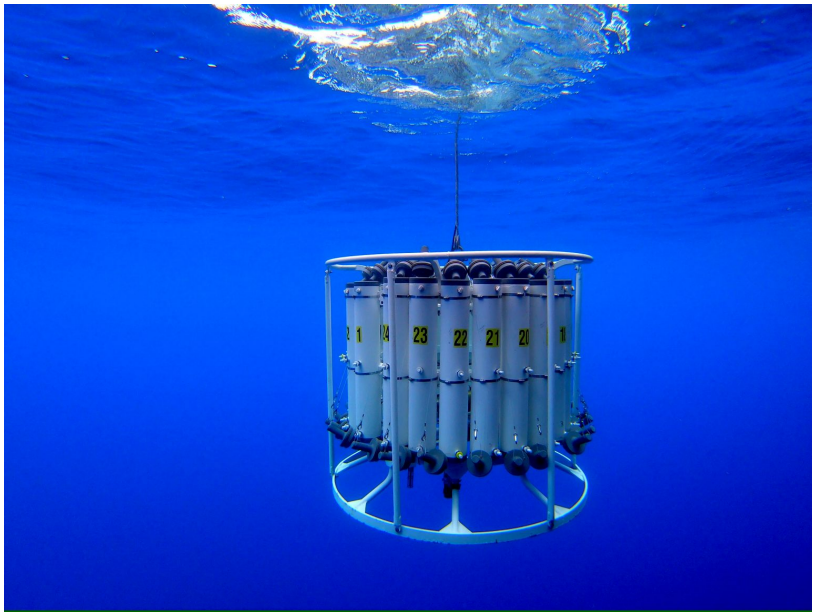

Figure 2.2: CalCOFI CTD-Rosette used to obtain samples for eDNA and oceanographic data, pictured underwater.

sighting according to methods described briefly below and in detail in Campbell et al. (2015).

2.1.2 Baleen whale sightings

On each quarterly cruise, two trained observers monitor the ocean surface for cetaceans and other animals and record sightings following line-transect protocol (Burnham et al., 1980). Weather permitting, observations occur during local daylight hours as the cruise travels between CalCOFI stations. "On-effort" sightings occur when two observers actively search in sufficiently calm sea conditions (Beaufort sea state 0-5) with a visibility of at least 1 km and a minimum ship speed of 11 km/h. "On-transect" sightings are observations made while the ship is traveling along one of the predefined transects within the CalCOFI study area.

On-effort and on-transect sightings are scaled to obtain measurements of the number of individuals per 1000 km of transect length for each species sighted — effectively,

density estimates assuming uniform probability of detection and effective strip width (visible distance in either direction from the transect). These scaled sightings are available for ten whale species and constitute the response variables of interest. This project focuses on three common species: blue whales, fin whales, and humpback whales.

2.1.3 Plankton eDNA

The NOAA-CalCOFI Ocean Genomics (NCOG) project is a partnership between the National Oceanic and Atmospheric Administration (NOAA), the J. Craig Venter Institute (JCVI), and the Scripps Institution of Oceanography (SIO) formed to collect seawater samples for RNA and DNA sequencing. Since 2014, NCOG samples have been collected on all CalCOFI cruises from two depths on primary productivity stations: 10 m and the chlorophyll-A maximum depth (the highest chlorophyll concentration). The latter varies considerably from station to station, between as little as approximately 30 m and as much as 130 m in depth.

Available NCOG data include 16S ribosomal RNA (rRNA) and 18S rRNA metabarcoding reads from amplicon sequencing; 16S and 18S refer to specific rRNA genes. Variation in the 16S gene differentiates bacterial species, and this gene is typically sequenced to study bacterial composition. By contrast, variation in the 18S region is well-resolved to eukaryotes, and, specifically in this context, the gene is sequenced to study plankton. In short, the NCOG data targets microbial and plankton assemblage. Amplicon sequencing involves isolating strands of genetic material corresponding to a specific region of the target gene and amplifying those strands through polymerase chain reactions (PCR) (Callahan et al., 2019). The amplicons are sequenced, and the numbers of instances of unique sequences after amplification — "amplicon sequence

variants" (ASVs) — in a sample are counted and identified taxonomically using a reference library.

This work focuses on read counts obtained from amplicon sequencing of the V9 region of the 18S rRNA gene. Associations between plankton assemblage and baleen whale abundances are more likely to be ecologically interpretable than the associations involving bacteria that could be identified by analyzing 16S data; the latter may be too far removed at the tropic level to produce meaningful statistical results.

2.2 Data processing and integration

The datasets used for analysis, although collected contemporaneously in time and space, reflect different spatial sampling resolutions. Reconciling these differences involves aggregating observations to the cruise level; additional steps are taken to account for the compositionality of eDNA data and seasonal trends present in both datasets. These steps are described in detail below.

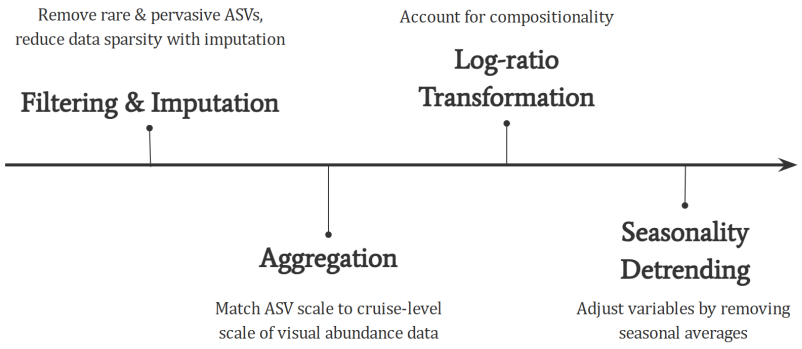

Figure 2.3: Flowchart outlining preprocessing steps of the integrated visual abundance/eDNA dataset before modeling

2.2.1 Scaled marine mammal sightings

For each on-effort and on-transect sighting $j = 1, \ldots, n_i$ on cruise $i = 1, \ldots, n$, an estimated group size s_{ij} is recorded along with location, time, observation conditions, and other contextual information. For this analysis, scaled sightings are defined as the number of individuals sighted per 1000km of transect length:

$$y_i \propto \frac{1}{L} \sum_{j=1}^{n_i} s_{ij}$$

It is noted that density per area in the sampling region could also be estimated following the methodology of Marques et al. (2007) by adjusting for the probability of detection, visible distance from the transect line, and other covariates.[2]

Individuals are likely present in at least small numbers year-round for the common species considered in this project. Therefore, the absence of sightings is not indicative of truly zero abundance in the sampling region. To account for this imperfect detection and mathematically allow for modeling scaled sightings on the log scale, a uniform random number between zero and the corresponding season's minimum scaled sighting is imputed for cruises where no whale observations were sighted.

[2]For example, one might form transect-specific estimates of the detection probability and strip width (visible distance) and compute a weighted average:

$$d_i = \sum_v \frac{n_{iv}\bar{s}_{iv}}{L_v \hat{p_{iv}} \hat{w_{iv}}}$$

Here, $n_{iv}\bar{s}_{iv}$ is the total number of individuals sighted of a species along transect v, L_v is the total on-effort transect length, $\hat{p}_{iv}$ is the average estimated probability of detection, and $\hat{w}_{iv}$ is the estimated effective strip width. This density measure captures the number of species observations, weighted inversely by the probability of detection and search area along the transect. Various additional strategies exist for performing density estimation using distance sampling methodology (Marques and Buckland, 2003; Marques et al., 2007; Buckland et al., 2015).

2.2.2 Filtering, imputation, and aggregation of eDNA data

The initial processing of 18S data consists of removing rare and common ASVs, imputing zeroes, and converting them to relative abundances.

Raw data are read counts for a large number of ASVs in each sample. Let r_{ijklm} denote the read count for ASV j from the sample taken at station l on transect k and depth m on cruise i.[3] For the purposes of this project, rare ASVs are defined as those that appear in at most 5% of samples, and common ASVs are defined as those that appear in at least 90% of samples. Both rare and common ASVs are likely to be statistically insignificant; rare ASVs may have too little representation to be statistically significant, while common ASVs may not have enough differentiation to detect statistically significant differences between samples. Both are removed to reduce the candidate set of ASVs to ones more likely to have significant associations with whale abundance. In detail, with $N = \sum_i \sum_k \sum_l \sum_m 1$ denoting the total number of samples, and $p_j = \frac{1}{N} \sum_i \sum_k \sum_l \sum_m \mathbb{1}\{r_{ijklm} > 0\}$ denoting the proportion of samples in which ASV j has a nonzero read count, the ASVs are filtered to the index set:

$$\mathcal{J} = \{j \in 1, \ldots, J : 0.05 \leq p_j \leq 0.9\}$$

From this point forward notationally, assume the filtered ASVs are reindexed consecutively so that, *e.g.*, $j = 1$ refers to the relative abundance of the first ASV in the index set $\mathcal{J}$, $j = 2$ to that of the second ASV in the index set $\mathcal{J}$, and so forth, and that $\mathcal{J} = |\mathcal{J}|$.

[3] So the indices are: cruise i, ASV j, transect k, station l, and depth m.

The filtered data retain many zeroes due to the high variability of eDNA detection across samples. In this context, zeroes are generally thought to simply indicate that the corresponding ASV is below the detection threshold after amplification in the physical sample (rather than being "true zeroes") (Li, 2015). A widely used practice is to replace zeros with a small number, called a pseudo-count (Kaul et al., 2017). While a variety of imputation methods are available, this project utilized Bayesian multiplicative count replacement (Martín-Fernández et al., 2015) for its computational efficiency. The resulting counts after zero imputation $\tilde{r}_{ijklm}$ are then converted to relative abundances by normalizing on a per-sample basis by the total number of reads (both counts and pseudo-counts) for the corresponding sample:

$$x_{ijklm} = \frac{\tilde{r}_{ijklm}}{\sum_j \tilde{r}_{ijklm}}$$

The `zCompositions` package in R (Palarea-Albaladejo and Martín-Fernández, 2015) implements both count replacement and relative abundance conversion simultaneously. Limitations of this method are that it introduces bias and tends to be conservative.

Finally, relative abundances are aggregated to the cruise level (the same spatial resolution as the scaled sightings) for proper data integration. The aggregation amounts to essentially a weighted geometric mean in which weights are inversely proportional to spatial sampling density on the CalCOFI grid and varied by depth:

$$x_{ij} = \prod_{k=1}^{K_i} \prod_{l=1}^{L_{ik}} \prod_{m=1}^{2} (x_{ijklm})^{w_{klm}}$$

The form in which the calculation is written reflects that weights are determined by first averaging relative abundances across depth (subsurface and Chl-A maximum)

by station, then averaging over stations, and finally averaging over transects. The weights were specified to be inversely proportional to spatial density and maximize α-diversity across depth.[4]

2.2.3 Log-ratio transformations

Data transformations are required to render the average relative abundances in a form suitable for statistical modeling. In particular, 18s rRNA samples are compositional data: the relative abundances of each amplicon sequence variant (ASV) sum to one. As a result, the data exhibit the so-called Aitchison geometry on the simplex (Aitchison, 1982), but most statistical methods are designed for data with Euclidean geometry. Linear transformations through the form of isomorphisms can be used to transform between real space and the Aitchison simplex (Filzmoser et al., 2018).

A centered log-ratio (CLR) transformation is applied to the ASV compositions with D ASVs by dividing each observed value by the geometric mean before taking the log. In detail, with geometric mean $g_i = \sqrt[J]{\prod_{j=1}^{J} x_{ij}}$, the CLR transformation is the log of:

[4]In detail, the weights w_{klm} are specified via the following multi-step calculation on the log scale. First, since two depths are sampled at each station, aggregating over depth by station amounts to taking a convex combination:

$$\log(x_{ijkl}) = \frac{1}{2}\left(c \cdot \log(x_{ijklm}) + (1-c) \cdot \log(x_{ijklm})\right)$$

Then, spatial averaging is done first by transect and then again by cruise:

$$\log(x_{ij}) = \frac{1}{K_i}\sum_{k=1}^{K_i}\left(\frac{1}{L_{ik}}\sum_{l=1}^{L_{ik}}\log(x_{ijkl})\right)$$

All together:

$$\log(x_{ij}) = \frac{1}{K_i}\left\{\sum_{k=1}^{K_i}\left[\frac{1}{L_{ik}}\sum_{l=1}^{2L_{ik}}\left(c \cdot \log(x_{ijkl1}) + (1-c) \cdot \log(x_{ijkl2})\right)\right]\right\}$$

In this case, c is the value determined by brute-force search that maximizes average Shannon α-diversity across cruises of the resulting average relative abundances.

$$z_{ij} = \frac{x_{ij}}{g_i}$$

The resulting quantity z_{ij} represents the relative abundance of ASV j relative to the average relative abundance across all ASVs in the sample. For example, $z_{ij} = 2$ indicates that ASV j is twice as abundant as the average relative abundance in the sample.

2.2.4 Seasonal de-trending

Both datasets exhibit seasonal trends across cruises; marine mammal densities show particularly strong seasonal patterns, with some species observed regularly only at certain times of the year.

To remove seasonal patterns in each dataset, geometric means are used to capture seasonal averages for both whale densities and average relative abundances of ASVs. For ease of notation, the seasonal geometric means are written as functions of the observation index i and vectors of observations $\mathbf{y}$ and $\mathbf{z}_j$ as follows for whale densities and (transformed) ASVs, respectively:

$$g(i, \mathbf{y}) = \left(\prod_{i \in \mathcal{I}(i)} y_i \right)^{1/|\mathcal{I}(i)|}$$
$$g(i, \mathbf{z}_j) = \left(\prod_{i \in \mathcal{I}(i)} z_{ij} \right)^{1/|\mathcal{I}(i)|}$$

The index set $\mathcal{I}(i)$ is the collection of all indices in the same season as observation i. Then, the seasonally-adjusted densities and ASVs are, respectively:

$$\tilde{y}_i = \frac{y_i}{g(i, \mathbf{y})} \qquad \text{and} \qquad \tilde{z}_{ij} = \frac{z_{ij}}{g(i, \mathbf{z}_j)}$$

These form the response and covariates used directly in the model. Those quantities are interpreted as follows:

- $\tilde{y}_i$ represents the observed whale density for cruise i relative to the average seasonal geometric mean density across all cruises in the same season.
- $\tilde{z}_{ij}$ represents the CLR-transformed ASV j composition for cruise i relative to the average seasonal geometric mean across all cruises in the same season.

For contextual examples:

- $\tilde{y}_i = 0.5$ indicates that cruise i has half the observed whale density as the geometric mean across all cruises in that season
- $\tilde{z}_{ij} = 0.5$ indicates that the CLR-transformed composition for cruise i ASV j is half the geometric mean across all cruises in that season.

2.3 Statistical methods

Associations between eDNA and whale abundances were estimated using a log-contrast type model framework using sparse partial least squares (sPLS). This section provides an exposition of the model framework and sPLS estimation method.

2.3.1 Model specification

After adjusting both to remove seasonal patterns, the model relating whale densities to eDNA is:

$$\log\left(\frac{y_i}{g(i,\mathbf{y})}\right) = \beta_0 + \sum_{j=1}^{J} \beta_j \cdot \log\left(\frac{z_{ij}}{g(i,\mathbf{z}_j)}\right) + \epsilon_i \tag{2.1}$$

Expressed in the standard form for a linear model:

$$Y = X\beta + \epsilon$$

Besides the seasonality adjustment, this model is commonly referred to as a log-contrast model (Aitchison and Bacon-Shone, 1984; Combettes and Müller, 2021). The coefficients capture multiplicative changes in (median, if error normality is assumed) scaled sightings associated with multiplicative changes in relative abundances after adjusting for seasonality. Specifically, the model indicates that a k-fold change in the seasonally-adjusted relative abundance of ASV j is associated with a k_j^{β} change in median seasonally-adjusted scaled sightings.

Separate models are specified for each species of baleen whale.

2.3.2 Parameter estimation

In order to fit the model in Equation (2.1), the eDNA data X is first projected onto a low-dimensional subspace of two latent components, producing a latent component matrix:

$$T = XA \tag{2.2}$$

Each column of A defines a linear combination of the seasonally-adjusted, log-transformed observations. The PLS approach estimates the columns of $A = (a_1, a_2)$ so that the latent variables T are maximally correlated with the response. This is achieved through solving successive optimization problems. The general idea is to find a projection that maximizes the covariance:

$$\hat{a}_1 = \text{argmax}_{a_1} \{\text{cov}(Y, Xa_1)\}$$

This gives the first latent variable. The second latent variable is then found by maximizing the residual covariance subject to an orthogonality constraint:

$$\hat{a}_2 = \text{argmax}_{a_2 \perp \hat{a}_1} \{\text{cov}(Y - Xa_1, Xa_2)\}$$

Subsequent components for a higher-dimensional latent space could be estimated similarly. Once the projection is estimated, a regression model is fit in the latent space:

$$Y = \hat{T}\gamma + \epsilon \tag{2.3}$$

Combining Equations 2.1 and 2.3 gives:

$$Y = X\hat{A}\gamma + \epsilon \tag{2.4}$$

Estimates for the coefficients of the model in Equation 2.1 are obtained by back-projecting the least squares estimates $\hat{\gamma} = (T'T)^{-1}T'Y$ to obtain:

$$\hat{\beta} = \hat{A}\hat{\gamma} \tag{2.5}$$

In this approach, the direction vectors of the projection allow noise variables to enter, reducing the interpretability of model estimates. Imposing a sparsity constraint on the projection A improves interpretability by effectively performing automatic variable selection.

Sparse partial least squares (sPLS) consists of fitting the model in Equation 2.1 with a sparsity constraint on the projection in equation 2.2. The general process to estimate parameters is analogous to that outlined above:

1. Initialization of the model in Equation 2.2
2. Finding the first direction vector $\hat{a}_1$
3. Obtaining residuals $Y - X\hat{a}_1$
4. Finding the next direction vector $\hat{a}_2$

Here, only two latent variables are specified. However, estimating subsequent latent variables would continue the process in steps 3 and 4 until the number of desired latent variables is reached. Since the process of estimating direction vectors for the projection is iterative, details are provided here for a single direction vector.

To find a single (sparse) direction vector w that maximizes correlation with the response (or residuals, if finding subsequent direction vectors), the aim is to solve the optimization problem:

$$\max_{w}(w^T M w) \quad \text{such that } w^T w = 1 \quad \text{and } \|w\|_1 \leq s$$

Here, $M = X^T Y Y^T X$ gives the covariance maximization problem, and s determines sparsity. For computational ease, this criterion is reformulated in terms of a "surrogate" direction vector c, assumed to be close to w, so that the L_1 penalty can be

imposed on c but w can be left unconstrained; this allows for the use of standard PLS algorithms with minor modifications to obtain an approximate solution to the problem above. The reformulated criterion is:

$$\min_{w,c} \left(-\frac{w^T M w}{2} + \frac{(c-w)^T M (c-w)}{2} + \eta \|c\|_1 + \lambda \|c\|_2^2 \right), \quad w^T w = 1$$

This problem can be solved by iteratively updating w with fixed c using a standard PLS algorithm (*e.g.*, SIMPLS, NIPALS) and then solving for c with w fixed following results in Chun and Keleş (2010). The single tuning parameter η can be tuned with cross-validation. η acts as a sparsity parameter on all direction vectors simultaneously. $0 \leq \eta \leq 1$ where higher values of η force greater sparsity on the components retained. The L_2 penalty is included for numerical stability. Full computational details are given by Chun and Keleş (2010). At convergence, the sparse surrogate c is taken as the final projection (*i.e.*, a_1).

Chapter 3

RESULTS

Models were fit following the methods outlined in the previous chapter for each of three common whale species: blue whales, fin whales, and humpback whales. This chapter opens with data summaries pertaining to the processing steps, particularly filtering and seasonal adjustments. It then presents results obtained from fitted models, including metrics of fit and prediction, summaries of the selected ASVs for each species by taxonomy, and exploratory summaries of the estimated model coefficients. All statistical analyses were conducted using R Statistical Software (v4.3.3; R Core Team 2023).

3.1 Data summaries

Exploration of the whale abundance estimates and the eDNA read counts prior to model fitting presents interesting insights and provides motivation for several of the processing steps taken in this analysis. Understanding and identifying patterns in the data prior to modeling is critical in properly accounting for fundamental data structures; in this case, seasonality in both the abundance of baleen whales and the rarity of ASVs are key features adjusted for in the analysis.

Abundance patterns differ seasonally for each of the three whale species, with fluctuations over time; seasonality is especially pronounced for blue and fin whales. Figure 3.1 shows scaled sightings for each species of interest over season and year before adjusting for seasonality. Blue whales tend to be observed most frequently in the summer and fall, with minimal sightings during other seasons. Fin whales are also

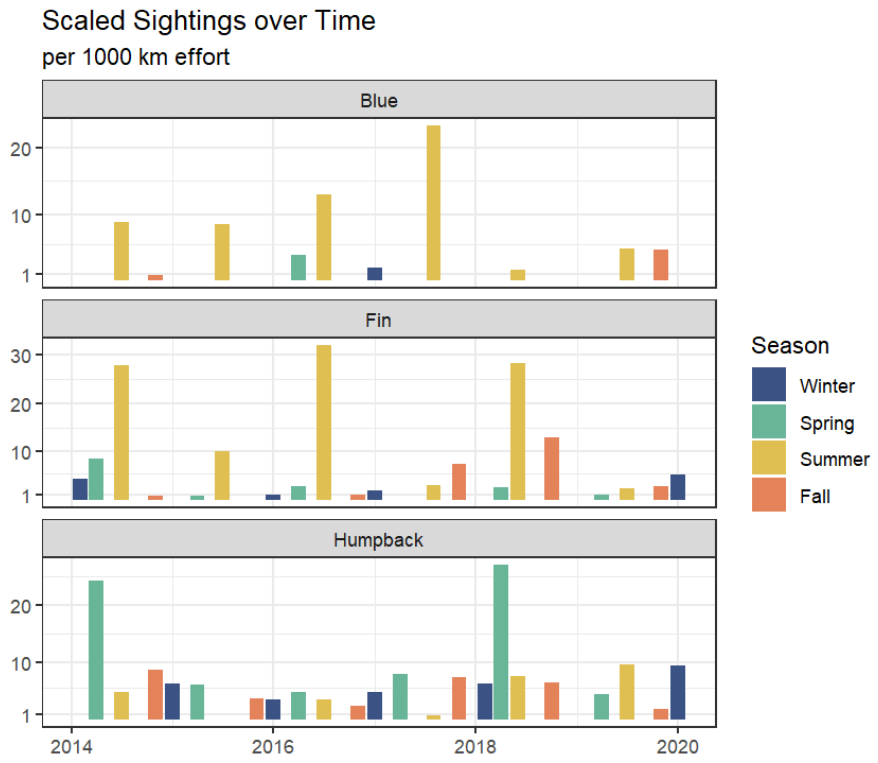

Figure 3.1: Visual whale abundance from winter 2014 to winter 2020 CalCOFI cruises by species, scaled to search effort

most abundant during summer and fall. Humpback whales are sighted less often than blue and fin whales but tend to be seen in greater numbers during spring. Overall, whale abundances tend to be lowest during the winter season. Over time, besides seasonal trends, abundances fluctuate with no apparent pattern.

Figure 3.2 shows seasonally adjusted whale densities over season and year. After the seasonal adjustment, each scaled sighting is expressed as a ratio relative to its respective season's average as measured by the geometric mean across years. In Figure 3.2, the seasonally adjusted scaled sightings are shown on the log scale so that a value of 0 indicates that the scaled sighting was exactly the seasonal mean: $\frac{y_i}{g(i,\boldsymbol{y})} = e^0 = 1$. Thus, a negative value indicates that the observed abundance is lower than the seasonal average (ratio < 1). In contrast, a positive value indicates

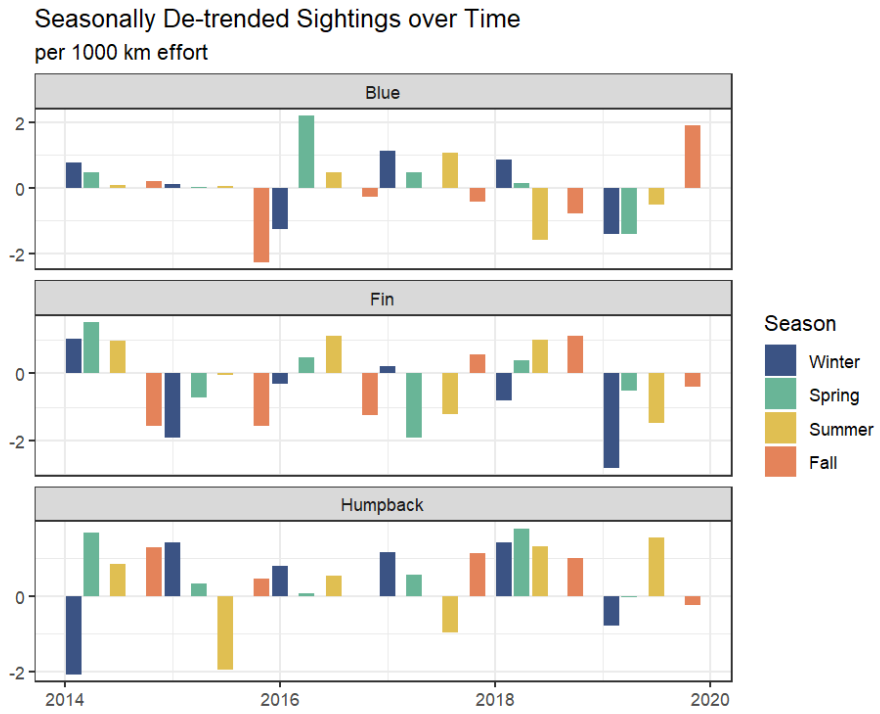

Figure 3.2: Visual whale abundance from winter 2014 to winter 2020 CalCOFI cruises by species, scaled and seasonally detrended

that the observed abundance is higher than the seasonal average (ratio > 1). Despite a few relatively low or high observed abundances, there does not seem to be any apparent residual pattern by season for any of the three whale species, as expected. Fluctuations are present over time, where some years display higher estimated whale densities, and some years display lower densities relative to their seasonal means, so there may be some autocorrelation in the data not explicitly modeled.

Figure 3.3 (left) shows observed abundance by season and species, while Figure 3.3 (right) displays observed abundance relative to the average seasonal abundance, also by season and species. Figure 3.3 (left) shows large differences in the typical observed density across both seasons and species. The highest observed densities for blue and fin whales are in summer, while the highest observed densities for humpback whales are in spring. Blue whales have the lowest average observed abundance in every

season except for spring. In all seasons, except for fall, humpback whales have the highest average observed abundance compared to the fin and blue whales. After removing seasonal patterns from the scaled whale densities, Figure 3.3 (right) shows a lack of associations between abundance, season, and species, as expected; however, interestingly, there are some differences in variability.

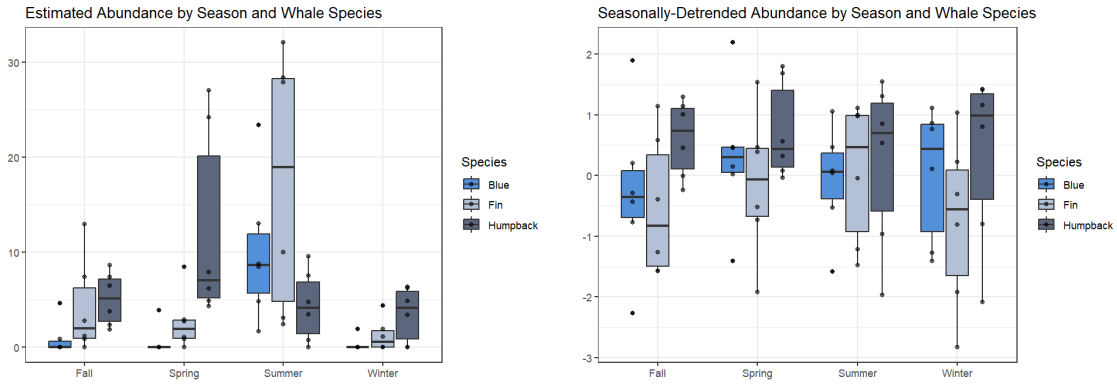

Figure 3.3: Boxplots of visual whale abundance from winter 2014 to winter 2022 CalCOFI cruises by season and species, scaled (left) and seasonally detrended (right)

This presence of seasonal patterns requires adjustment for seasonality in both the response and predictor variables, as described in Section 2.2.4. Should these seasonal patterns go unaccounted for in the modeling stages, the fitted model may identify relationships that rely on overlapping seasonal patterns rather than the ecological associations of interest. Seasonal adjustment makes the results more likely to identify genuine associations between eDNA and whale abundance.

Before fitting the model, ASVs were filtered by rarity and prevalence across samples. The purpose of this filtering was to remove likely insignificant ASVs, allow for manageable modeling, and enhance interpretation. In particular, ASVs that were present in nearly all samples were removed as they would fail to uniquely identify the presence of whales in the observed area. Meanwhile, ASVs that were rarely present were removed to reduce excessive noise and variability from sample to sample.

Table 3.1: Number of ASVs after filtering out rare and pervasive ASVs

Total ASVs		50,408
Common ASVs (in at least 90% of samples)	-	19
Rare ASVs (in 5% or fewer samples)	-	47,141
Remaining ASVs		3,248

Table 3.1 shows that 50,408 unique ASVs were present over all observed samples prior to filtering. Nineteen common ASVs appeared in 90% or more of the samples, which were removed from the dataset. Rare ASVs were defined as those present in 5% fewer samples. Additionally, 47,141 rare ASVs were removed, leaving 3,248 ASVs to model associations with estimated whale abundance. Relative to the 50,408 ASVs available, this represents approximately 6.4% of the initial dataset.

Table 3.2: Number of eDNA samples after filtering and aggregation

eDNA samples from winter 2014 to winter 2020		1,536
Filtering non-genuine samples	-	50
Filtering samples with over 90% rare or common ASVs	-	301
Remaining samples		1,185
Aggregating samples over cruise		25

Table 3.2 shows that of an original 1,536 eDNA samples from CalCOFI stations between winter 2014 and winter 2020, 1,185 samples remained after filtering criteria. These 1,185 samples were then aggregated to the cruise level, resulting in 25 eDNA observations to model associations with estimated whale abundance for each cruise.

3.2 Hyperparameter tuning

Due to the limited number of observations in the integrated data ($n = 25$), leave-one-out cross-validation (LOOCV) was used to assess model fit across a grid of η values for each whale species. Due to high prediction variability, instead of directly optimizing prediction error, selecting an η value for each model was a balanced consideration between selection sparsity, fit, and prediction error. In particular, the analysis sought

to maximize sparsity while retaining a high R^2 coefficient and low prediction error. Visual inspection indicated that an η value of around 0.6 was appropriate for each species.

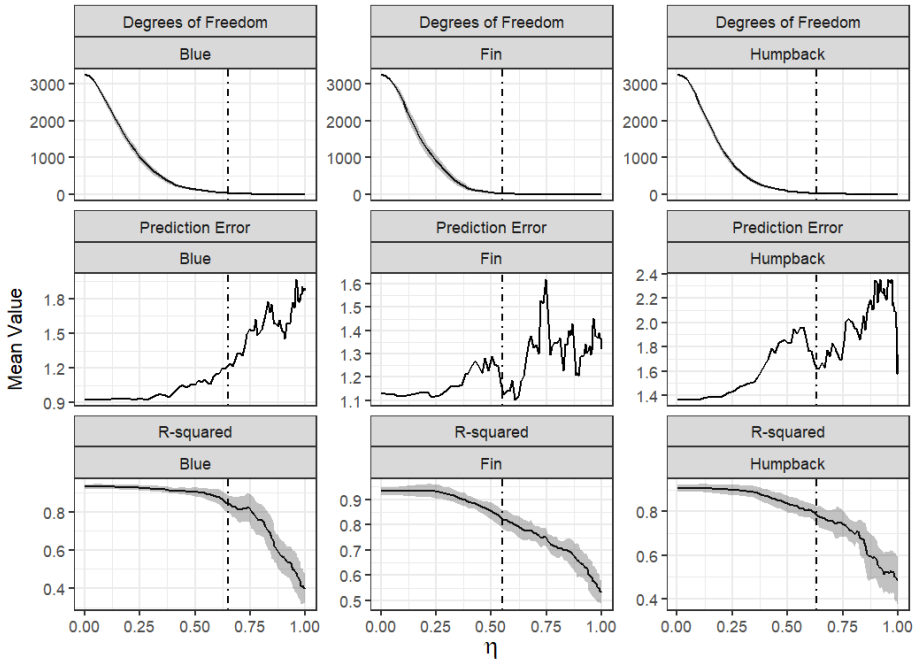

Figure 3.4: Line graphs of degrees of freedom (DF), fit (R^2), prediction error (SPE) values for the selection of η

The exact η values selected for each model are shown in Figure 3.4; the specific fit metrics are shown in Table 3.3.

Table 3.3: Summary of average degrees of freedom (DF), fit (R^2), prediction error (SPE) for each model in optimal η for each whale species

Species	η	DF	R^2	SPE
Blue	0.65	46.56	0.84	1.22
Fin	0.56	31.44	0.82	1.12
Humpback	0.63	35.44	0.79	1.62

3.3 Parameter estimates and model fit

Table 3.4 shows high-level summaries of the models fit for each of the three whale species, reporting (1) the number of ASVs selected via sPLS; (2) the coefficient of determination R^2; (3) an estimate of prediction bias, measured as the ratio of predicted value to observed value; and (4) mean square prediction error on the log scale (*i.e.*, the scale of the response $\log\left(\frac{y_i}{g(i,\mathbf{y})}\right)$).

Table 3.4: Summary of selection size ($ASVs$), fit (R^2), prediction bias, and prediction error ($MSPE$) of sPLS model fitting results for each whale species

Species	# ASVs	R^2	Prediction Bias	MSPE
Blue	52	0.83	1.04	1.22
Fin	27	0.81	1.04	1.12
Humpback	39	0.79	0.96	1.62

All models exhibit strong performance in terms of explanatory power, each explaining around 80% of the variance in seasonally-adjusted whale abundances (on the log scale) using only around 1% of the number of available ASVs. The prediction bias ratio suggests unbiased predictions but indicates slight overprediction for fin and blue whales (bias = 1.04) and slight underprediction for humpback whales (bias = 0.96). The model for fin whales has the smallest mean squared prediction error (MSPE) of 1.12.

Figure 3.5 shows leave-one-out predictions[1] (top) as well as fitted values (bottom) against observed values, and indicates modest predictive capability. In each case, perfect accuracy is indicated by the $y = x$ line. The fitted values indicate the explanatory power of the model by comparing the observed abundance value to the value estimated by the model; these points tend to lie along the $y = x$ line, indicating

[1]Predictions were calculated with a leave-one-out framework where one observation was left out while the model was fit to the remaining 24 observations. A value is then predicted on the data point that was left out.

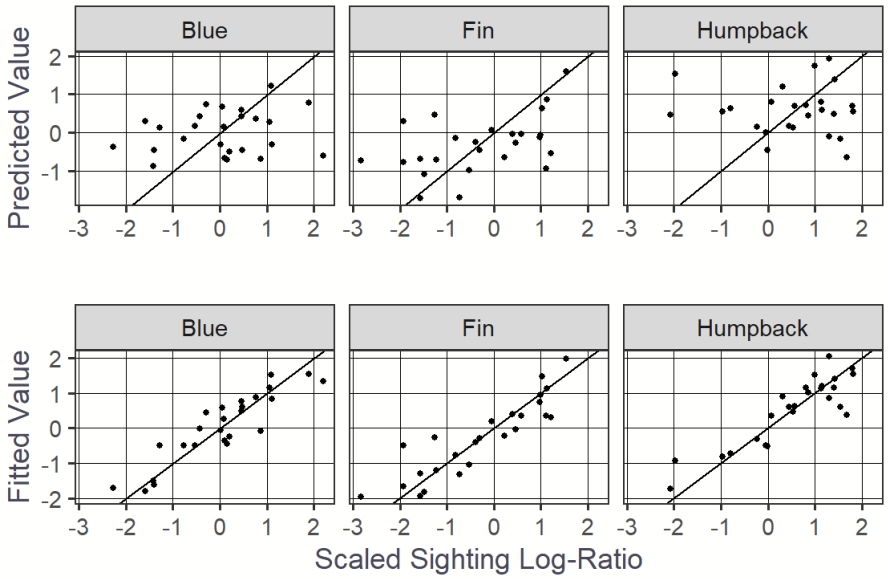

Figure 3.5: Predicted vs. observed response value (top) & model-fitted values vs observed response value (bottom); $y = x$ line indicates perfectly predicted/fitted value

strong explanatory power by each model. By contrast, these predictions are scattered more widely around the $y = x$ line. The strong explanatory power but modest predictive capability may indicate slight overfitting.

3.4 Analysis of selected ASVs

Across the three models, 95 unique ASVs were selected; these appear to be largely species-specific. The taxonomic classifications can be compared for overlap between models (Figure 3.6) or lack thereof. Additionally, the magnitudes of coefficient estimates can be examined to identify particularly strong associations that may point to interesting ecological relationships (Figures 3.7, 3.8).

Figure 3.6 (left) indicates that only four of the 95 selected ASVs (4.2%) were shared among all three whale species, and only 19 (20%) were shared by at least two species. The majority of selected ASVs are unique to just one of the three whale species.

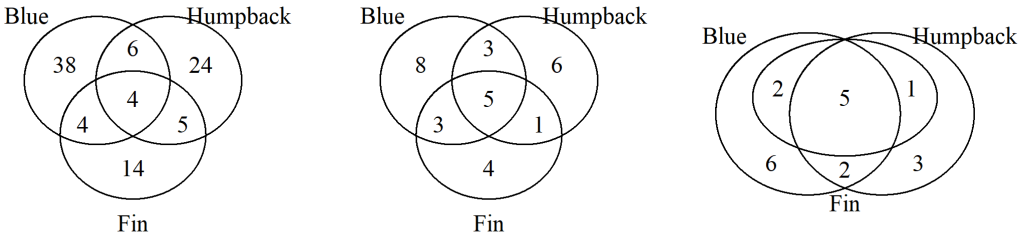

Figure 3.6: Number of unique and intersecting ASVs (left), order (middle), and phylum (right) per whale species

Grouping individual ASVs by their respective taxonomic classifications reveals structural patterns. At the phylum level, the overlapping ovals in Figure 3.6 (right) show that none of the selected phyla are unique to associations with fin whale abundance. Distinction begins at lower taxon levels, as shown at the order level in Figure 3.6 (middle). Order and class levels presented similar groupings per whale species among the ASVs with little difference between class counts and order counts.

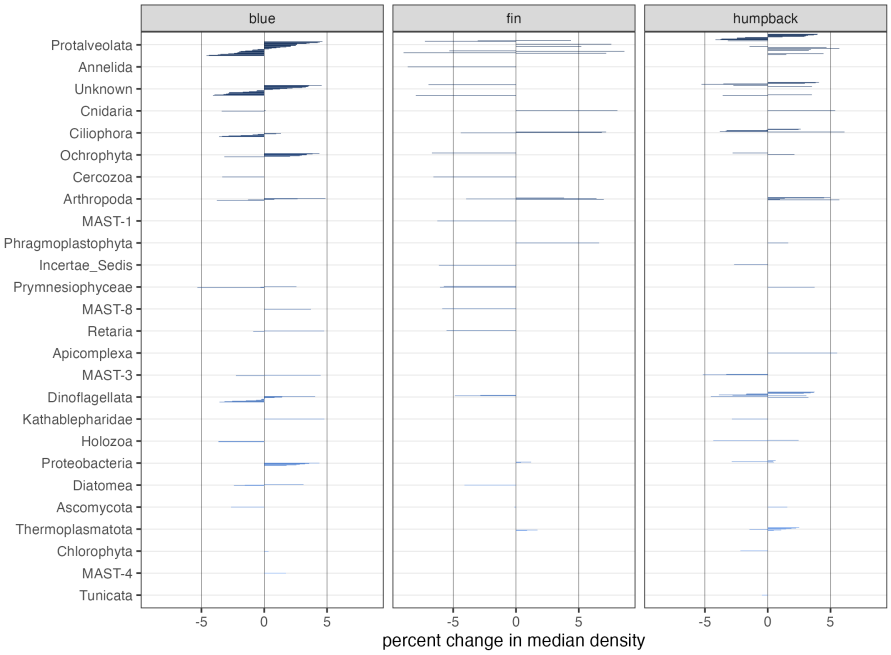

Figure 3.7: Model-fitted ASV coefficients for each whale species grouped by phylum

Figure 3.7 shows the estimated coefficients as a percentage change in median seasonally-adjusted scaled sighting associated with doublings of seasonally-adjusted relative abundances of selected ASVs. ASVs are grouped by phylum across the three whale species. While many phyla have ASVs that have both positive and negative associations with abundance, a few, like cnidaria, have primarily positive associations, while others, like diatomea, have largely negative associations. Associations also can differ in direction between species. Notably, the coefficient magnitudes are generally larger for the model for fin whales, though this model selected the fewest number of ASVs.

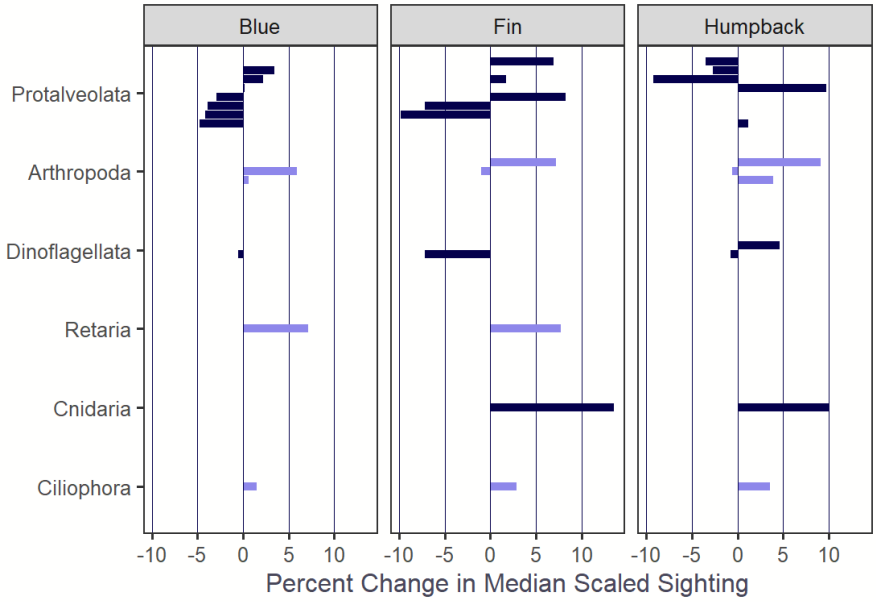

Figure 3.8: Model-fitted ASV coefficients for phylum present across whale species

With the sPLS models fitted for each of the three whale species, ASV coefficient results can be used to identify potential ecological relationships. These models have identified several potentially interesting ASV candidates for possible directions of further study by marine scientists. Figure 3.6 shows that ASVs are largely species-specific. For a given whale species, larger coefficients indicate ASVs that are more strongly associated with the abundance of that species, which may present potential ecological research directions. It may also be interesting to compare ASVs that are

present across whale species. Marine biologists, ecologists, and geneticists can compare the selected ASVs both across and within models, as well as within and across taxonomical groupings to extract ecological insight.

Chapter 4

DISCUSSION

From the 50,408 ASVs initially observed across all samples, the approach developed and applied in this project identified a subset of 95 amplicons (0.188%) that together explain an estimated 80% of variability in seasonally-adjusted abundances of fin whales, blue whales, and humpback whales. Moreover, the selected amplicons were largely species-specific. This degree of model sparsity and explanatory power using an interpretable model framework allows researchers to identify taxa directly from 18S rRNA metabarcoding reads associated with baleen whale abundances and formulate hypotheses about ecological function underlying such associations. More broadly, this sort of analysis has the potential to identify ecological habitats of difficult-to-detect or migratory species. Additional factors not considered here, such as oceanographic conditions or physical properties of water masses, may also play a role in explaining the relationships identified through this analysis.

There are several avenues for future research. For instance, one could assess the predictive capability of the model on data from recent years, provide uncertainty quantification for parameter estimates, incorporate additional oceanographic and/or biological data, or account for autocorrelation in time. However, the limited sample size ($n = 25$ cruises) restricts the extent to which model complexity can be increased. Aggregating data to a smaller spatial scale (e.g., transect) may alleviate this to some extent but would also introduce spatial autocorrelation. These are areas that could be explored in future studies.

More narrowly, several choices made in the analysis presented here could be explored further, particularly the weighting in the spatial aggregation of the eDNA data and the filtering criteria applied to achieve sizeable preliminary reductions in the raw number of ASVs considered for analysis. These choices were made in a way deemed plausible and practical but without rigorous justification, so these choices may impact the analysis in ways that are currently not quantified.

Additionally, well-known biases of eDNA metabarcoding, including eDNA degradation when sampling and sequencing concerns, may not guarantee that the relative abundance of genomic read counts from environmental samples is necessarily reflective of that in the ocean. However, should such biases be assumed to be consistent across time and samples, results may still relate eDNA metabarcodes with whale abundance.

In closing, this project not only demonstrates the statistical feasibility of using environmental DNA in a predictive and/or explanatory capacity, but it also highlights its potential to identify ecological correlations across different levels of organization. This application is distinct from the more common uses of eDNA, such as providing an alternative to direct surveying or a means of rare species detection. One of the key challenges in carrying out correlative analyses like the one demonstrated here is data integration. One novelty of this project lies in its approach to this challenge, combining disparate datasets. However, similar analyses are likely feasible using data from sources that, like CalCOFI, have incorporated eDNA sampling into existing ecological monitoring programs.

BIBLIOGRAPHY

C. I. Adams, M. Knapp, N. J. Gemmell, G.-J. Jeunen, M. Bunce, M. D. Lamare, and H. R. Taylor. Beyond biodiversity: Can environmental dna (edna) cut it as a population genetics tool? *Genes*, 10(3):192, 2019.

J. Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982.

J. Aitchison and J. Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984.

C. S. Baker, D. Steel, S. Nieukirk, and H. Klinck. Environmental dna (edna) from the wake of the whales: Droplet digital pcr for detection and species identification. *Frontiers in Marine Science*, 5:133, 2018.

M. A. Barnes, C. R. Turner, C. L. Jerde, M. A. Renshaw, W. L. Chadderton, and D. M. Lodge. Environmental conditions influence edna persistence in aquatic systems. *Environmental science & technology*, 48(3):1819–1827, 2014.

K. C. Beng and R. T. Corlett. Applications of environmental dna (edna) in ecology and conservation: opportunities, challenges and prospects. *Biodiversity and conservation*, 29(7):2089–2121, 2020.

K. Bohmann, A. Evans, M. T. P. Gilbert, G. R. Carvalho, S. Creer, M. Knapp, W. Y. Douglas, and M. De Bruyn. Environmental dna for wildlife biology and biodiversity monitoring. *Trends in ecology & evolution*, 29(6):358–367, 2014.

A.-L. Boulesteix and K. Strimmer. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in bioinformatics*, 8(1):32–44, 2007.

S. T. Buckland, E. A. Rexstad, T. A. Marques, C. S. Oedekoven, et al. *Distance sampling: methods and applications*, volume 431. Springer, 2015.

K. P. Burnham, D. R. Anderson, and J. L. Laake. Estimation of density from line transect sampling of biological populations. *Wildlife monographs*, (72):3–202, 1980.

B. J. Callahan, J. Wong, C. Heiner, S. Oh, C. M. Theriot, A. S. Gulati, S. K. McGill, and M. K. Dougherty. High-throughput amplicon sequencing of the full-length 16s rrna gene with single-nucleotide resolution. *Nucleic acids research*, 47(18):e103–e103, 2019.

M. L. Calle. Statistical analysis of metagenomics data. *Genomics & informatics*, 17 (1), 2019.

G. S. Campbell, L. Thomas, K. Whitaker, A. B. Douglas, J. Calambokidis, and J. A. Hildebrand. Inter-annual and seasonal trends in cetacean distribution, density and abundance off southern california. *Deep Sea Research Part II: Topical Studies in Oceanography*, 112:143–157, 2015.

H. Chun and S. Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(1):3–25, 2010.

P. L. Combettes and C. L. Müller. Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, 13(2):217–242, 2021.

T. Dejean, A. Valentini, A. Duparc, S. Pellier-Cuit, F. Pompanon, P. Taberlet, and C. Miaud. Persistence of environmental dna in freshwater ecosystems. *PloS one*, 6 (8):e23398, 2011.

N. T. Evans, B. P. Olds, M. A. Renshaw, C. R. Turner, Y. Li, C. L. Jerde, A. R. Mahon, M. E. Pfrender, G. A. Lamberti, and D. M. Lodge. Quantification of mesocosm fish and amphibian species diversity via environmental dna metabarcoding. *Molecular ecology resources*, 16(1):29–41, 2016.

G. F. Ficetola, C. Miaud, F. Pompanon, and P. Taberlet. Species detection using environmental dna from water samples. *Biology letters*, 4(4):423–425, 2008.

P. Filzmoser, K. Hron, M. Templ, P. Filzmoser, K. Hron, and M. Templ. Geometrical properties of compositional data. *Applied Compositional Data Analysis: With Worked Examples in R*, pages 35–68, 2018.

A. D. Foote, P. F. Thomsen, S. Sveegaard, M. Wahlberg, J. Kielgast, L. A. Kyhn, A. B. Salling, A. Galatius, L. Orlando, and M. T. P. Gilbert. Investigating the potential use of environmental dna (edna) for genetic monitoring of marine mammals. 2012.

J. Hinkle and W. Rayens. Partial least squares and compositional data: problems and alternatives. *Chemometrics and intelligent laboratory systems*, 30(1):159–172, 1995.

M. E. Hunter, G. Meigs-Friend, J. A. Ferrante, A. T. Kamla, R. M. Dorazio, L. K. Diagne, F. Luna, J. M. Lanyon, and J. P. Reid. Surveys of environmental dna (edna): a new approach to estimate occurrence in vulnerable manatee populations. *Endangered Species Research*, 35:101–111, 2018.

C. C. James, A. D. Barton, L. Z. Allen, R. H. Lampe, A. Rabines, A. Schulberg, H. Zheng, R. Goericke, K. D. Goodwin, and A. E. Allen. Influence of nutrient supply on plankton microbiome biodiversity and distribution in a coastal upwelling region. *Nature communications*, 13(1):2448, 2022.

A. Kaul, S. Mandal, O. Davidov, and S. D. Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:283205, 2017.

H. Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

T.-Y. Liu, L. Trinchera, A. Tenenhaus, D. Wei, and A. O. Hero. Globally sparse pls regression. In *New Perspectives in Partial Least Squares and Related Methods*, pages 117–127. Springer, 2013.

F. F. Marques and S. T. Buckland. Incorporating covariates into standard line transect analyses. *Biometrics*, 59(4):924–935, 2003.

T. A. Marques, L. Thomas, S. G. Fancy, and S. T. Buckland. Improving estimates of bird density using multiple-covariate distance sampling. *The Auk*, 124(4):1229–1243, 2007.

J.-A. Martín-Fernández, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15(2):134–158, 2015.

K. Morey, T. Bartley, and R. Hanner. Validating environmental dna metabarcoding for marine fishes in diverse ecosystems using a public aquarium. environmental dna, 2 (3), 330–342, 2020.

J. Palarea-Albaladejo and J. A. Martín-Fernández. zcompositions—r package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143:85–96, 2015.

K. M. Parsons, M. Everett, M. Dahlheim, and L. Park. Water, water everywhere: Environmental dna can unlock population structure in elusive marine species. *Royal Society open science*, 5(8):180537, 2018.

V. Pawlowsky-Glahn and A. Buccianti. *Compositional data analysis*. Wiley Online Library, 2011.

H. C. Rees, B. C. Maddison, D. J. Middleditch, J. R. Patmore, and K. C. Gough. The detection of aquatic animal species using environmental dna–a review of edna as a survey tool in ecology. *Journal of applied ecology*, 51(5):1450–1459, 2014.

T. Riaz, W. Shehzad, A. Viari, F. Pompanon, P. Taberlet, and E. Coissac. ecoprimers: inference of new dna barcode markers from whole genome sequence analysis. *Nucleic acids research*, 39(21):e145–e145, 2011.

R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

K. M. Ruppert, R. J. Kline, and M. S. Rahman. Past, present, and future perspectives of environmental dna (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna. *Global Ecology and Conservation*, 17: e00547, 2019.

S. Shokralla, J. L. Spall, J. F. Gibson, and M. Hajibabaei. Next-generation sequencing technologies for environmental dna research. *Molecular ecology*, 21(8):1794–1805, 2012.

P. Suarez-Bregua, M. Alvarez-Gonzalez, K. M. Parsons, J. Rotllant, G. J. Pierce, and C. Saavedra. Environmental dna (edna) for monitoring marine mammals: Challenges and opportunities. *Frontiers in Marine Science*, 9:987774, 2022.

D. Székely, K. M. Cammen, and M. Tange Olsen. Needles in an ocean haystack: using environmental dna to study marine mammals in the north atlantic. *NAMMCO Scientific Publications*, 12, 2021.

P. Taberlet, E. Coissac, M. Hajibabaei, and L. H. Rieseberg. Environmental dna. *Molecular ecology*, 21(8), 2012.

P. F. Thomsen, J. Kielgast, L. L. Iversen, P. R. Møller, M. Rasmussen, and E. Willerslev. Detection of a diverse marine fish fauna using environmental dna from seawater samples. 2012.

M. C. Tsilimigras and A. A. Fodor. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of epidemiology*, 26(5):330–335, 2016.

M. C. Yates, D. J. Fraser, and A. M. Derry. Meta-analysis supports further refinement of edna for monitoring aquatic species-specific abundance in nature. *Environmental DNA*, 1(1):5–13, 2019.

R. Zhou, S. K. Ng, J. J. Sung, W. W. B. Goh, and S. H. Wong. Data pre-processing for analyzing microbiome data–a mini review. *Computational and Structural Biotechnology Journal*, 2023.