PNMS BRUV steps

Loaded data into R – for initial data checks and prep.

Outliers – two huge biomass outliers were detected:

1. Tiger shark in Sample PAL22\_059 – W2
2. Mola Mola in PAL22\_137 – S1

Using the Tukey Interquartile Range (IQR) method:

Calculate quartiles (Q1 and Q3) of biomass data by determining the IQR as Q3-Q1 and then establishing outlier threshold rules using 1.5 x IQR rule:

* Lower threshold: Q1 - (1.5 × IQR)
* Upper threshold: Q3 + (1.5 × IQR)

**Biomass\_kg:**

Using this approach there are 21 outliers.

Focusing on extreme outliers using multiplying with 3, there are 7 outliers.

**MaxN:**

Using this approach there are 28 outliers.

Focusing on extreme outliers using multiplying with 3, there are 23 outliers.

Many of these are ecological important and real observations though.

\*\*So – will probably not be using this approach to determine outliers.

**Final decision with regards to outliers:**

Considering that large schools of decapterus and mahi mahi were observed across multiple strings and are characteristic of the pelagic environment, as well as a few observations of large billfish – it was decided to keep all of these observations.

Only two outliers were removed:

The single observation of a tiger shark and mola mola.

All further analyses etc. were done with these two outliers removed.

**Sample aggregations**

We have 28 strings with 5 non-independent samples on each string.

So – we will focus on the following metrics:

**Abundance as MaxN:**

Species/group specific Max N per string (This uses the maximum count of each species across the samples within a string, recognizing that the 5 samples on a string are not independent observations). We have 2-3 strings per site – so we can get an average MaxN for a species/group in each site based on 2-3 data points.

**Biomass:**

Here we will follow a similar approach as with MaxN – considering the issue of double counting fish across samples from the same string.

So, we will use the maximum biomass for a species/group found across the 5 samples – and use as the biomass estimate for that species/group in that string. We then again have 2-3 biomass estimates per species/group for each site.

**Diversity**

This is straightforward enough – and will be the maximum number of different species/groups across all samples in a string. We will then get diversity per string and can average it across 2-3 strings in each site.

**Observation summary**

Of the 28 strings deployed, fish were not observed in 3 strings:

* PAL22\_3 N1 PNMS North
* PAL22\_8 N4 PNMS North
* PAL22\_21 S3 PNMS south

Successful observations on 89% of deployments

**Plankton data**

Sorting notes from Sarah:

*Additionally, to that point, the column titled “Total Count” relates the portion of the sample that was counted to the sample as a whole. In other words, if I split the sample in half, three times, the first portion of the sample that I sorted/counted was 1/16th of the entire sample. If I reached a total of 70 or more individuals of a given morphotype in that proportion (this would be the initial count), the multiplication factor (of 16) was entered, and the “Total Count” column represents the extrapolated total of said morphotype in the entire sample. Rare morphotypes were counted across the whole sample (i.e. initial count=total count), while more common morphotypes were counted until the sample fraction that contained the threshold of 70+ individuals, and then multiplied accordingly to achieve “Total Count” (i.e. initial count X multiplication factor = total count). Does this make sense? Hopefully the short methods section I contributed communicates that process succinctly and successfully. Worth also noting that the counts have yet to be corrected for volume of water sampled. I can work on adding a column with corrected counts over the course of the next week if you would like.*

Went back to the Plankton data – and it seems the following samples are missing:

String 19, 22, 25, 28

Went back on emails and saw that Sarah edited the RAW data to remove the following:

* Cephalopods: SUPER low abundance and uncertain IDs
* Ascidean Tadpoles+Larvaceans: lumped into Other Gelatinous
* Copepod 10: singleton with failed ID, uncertain of its copepod-ness
* Deciduous Larvae: Not abundant AT ALL.
* A handful of Echinoderms that were not abundant or uncertain of ID
* Nemerteans: Uncertain of IDs and not abundant at all
* Oligochaete: Singleton
* Digeneans: shouldn’t have been free floating in the lifestage found/shouldn’t be counted as part of the plankton population/abundance. Though a relevant part of the ecosystem.
* Porifera: very not abundant
* Some uncertain Sipunculan IDs
* Tick: singleton

We also removed:

1. All the single specimens from a single site that cannot be grouped into a 'larger' classification

2. All those types that were not quantified (e.g. ciliates and bryozoans - is that right?)

3. All the unknown specimens

Checked for outliers – and it seems there are only 2 major outliers – not sure I want to remove them though?

Edited spreadsheet shared by Sarah – with the different groupings will be used for analyses.

Need to format this spreadsheet – and run in R.

Did some basic Plankton check and visualisations.

Also included BRUV string in the plankton data for each plankton sample – so that the data sets can be grouped later on.

COMBINED data:

**1. Combined\_Plankton\_BRUV\_All\_Records.csv**

* **Size**: 14,518 rows × 22 columns
* **Aggregation Method**: No aggregation - direct many-to-many joining
* **Linking Variables**: Joined by BRUVString column
* **Content**:
  + Contains every possible combination of plankton and fish records that share the same BRUVString
  + Preserves all original detail from both datasets
  + Column names have .x and .y suffixes to distinguish source (e.g., Sample.x from plankton, Sample.y from BRUV)

**2. Combined\_Plankton\_BRUV\_Summarized.csv**

* **Size**: 3,979 rows × 19 columns
* **Aggregation Method**: Plankton data summarized by taxonomic grouping before joining – so the larger groups.
* **Linking Variables**: Joined by BRUVString column
* **Content**:
  + Plankton data aggregated by BRUVString, Zone, Site, Type, Plankton\_Category and Specimen\_Type
  + For each plankton group, shows total specimens and unique morphotype count
  + All fish data preserved at original detail level (Family, Binomial, etc.)
* **Best Used For**:
  + Analysis of relationships between plankton taxonomic groups and fish observations
  + More manageable dataset size while maintaining fish taxonomic detail
  + Plankton community composition analysis alongside fish data

**3. Combined\_Plankton\_BRUV\_Site\_Level.csv**

* **Size**: 11 rows × 7 columns
* **Aggregation Method**: Highest level of aggregation - both datasets summarized to site level
* **Linking Variables**: Joined by Site and Zone columns
* **Content**:
  + One row per sampling site (11 sites total)
  + Summarized metrics for each site:
    - Plankton: total count and unique morphotype count
    - Fish: total MaxN, total biomass, and species richness
* **Best Used For**:
  + Site-level ecological analyses
  + Broad patterns and correlations between plankton and fish communities
  + Statistical testing of hypotheses about site characteristics
  + Visualization of spatial patterns across the study area
  + Quick overview of biodiversity and abundance patterns