REPORT: Ecology in Brazil

Evaluation of the conservation status and monitoring proposal for the coastal reefs of Paraíba, Brazil

Rashmi Konnur (bmat1922)

Sneha K. K. (bmat1934)

OBJECTIVE: The main objective of this study was to evaluate the spatial and temporal variations in macrobenthic populations of the intertidal zones of the coastal reefs in Brazil. However, variations of other ecological variables have also been examined with respect to site and season.

Statistical methods of data analysis used in this project:

The following statistical methods were adopted for data analysis:

1. Two-way ANOVA for the dependence of EEI-c values on site and season
2. Post-hoc Student-Newman-Keuls test
3. Tukey's post-hoc analysis
4. Indicator value (INDVAL) analysis
5. One-way ANOVA for chlorophyll-a concentration in different months

FORMULAE USED

Some useful formulas used: -

ESG I (% cover)=(ESG IA×1)+(ESG IB×0.8)+(ESG IC×0.6)

ESG II (% cover)=(ESG IIA×0.8)+(ESG IIB×1)

EEI c=2+8min⁡(a+b(x/100)+c(x/100)^2+d(y/100)+e(y/100)^2+f(x/100)(y/100)

Where x and y are the % cover of ESG I and ESG II, respectively, and a, b, c, d, e and f are the coefficients of the hyperbole:

a =0.4680, b =1.2088, c = 0.3583, d =1.1289, e =0.5129, f = 0.1869

1.TWO-WAY ANOVA:

A two-way ANOVA was used for the analysis of variation in EEI-c values. The two-way ANOVA not only gives the insights of two separate one-way ANOVA for each factor but it also evaluates how the two factors interact (the effect that one factor has on the other).

NULL HYPOTHESES:

1.The population means of the EEI-c values of the first factor (site) are equal.

2. The population means of the EEI-c values of the second factor (season) are equal.

3.There is no interaction between the two factors, site and season.

ALTERNATE HYPOTHESIS:

The alternate hypothesis for cases 1 and 2 is: the population means are not equal.

The alternate hypothesis for case 3 is: there is an interaction effect between the two factors.

ASSUMPTIONS:

1.The populations from which the samples were obtained must be normally or approximately normally distributed.

2.The samples must be independent.

3.The variances of the populations must be equal.

	Degrees of freedom	Sum of squares	Mean square	F (Test statistic)
Factor A	k-1	SSA	$MSA=\frac{SSA}{(k-1)}$	$F_A=\frac{MSA}{MSE}$
Factor B	l-1	SSB	$MSB=\frac{SSB}{(l-1)}$	$F_B=\frac{MSB}{MSE}$
Interaction A:B	(k-1)(l-1)	SSAB	$MSAB=\frac{SSAB}{(k-1)(l-1)}$	$F=\frac{MSAB}{MSE}$
Error	kl(m-1)	SSE	$MSE=\frac{SSE}{kl(m-1)}$	

k: number of levels of factor A

l: number of levels of factor B

kl: number of treatments

m: number of observations for each treatment

Calculation of EEI-c values

To calculate the percentage coverage for each species in each sample, we divided each cell in the original dataset by 30 and multiplied by hundred(in EXCEL sheet EEICtry). Then we grouped the species into different ESG groups, and calculated the percentage coverage for each ESG group. Then we used the formula

ESG I (% cover) =(ESG IA * 1) + (ESG IB * 0.8) +(ESG IC * 0.6)

ESG II (% cover) = (ESG IIA * 0.8) + (ESG IIB * 1) to calculate percentage coverage of ESG I and ESG II .

We calculated EEI-c using the formula:

EEI-c = 2+8*min(a + b * (x / 100) +c * (x / 100)^ 2+d * (y / 100) + e * (y / 100)^2+ f * (x / 100) * (y / 100),1)

Where a,b,c,d,e, and f are constants:

a =0.4680, b = 1.2088, c =- 0.3583, d = -1.1289, e = 0.5129, f =- 0.1869

Two-way ANOVA in R

The data in EEI-c try was imported and stored in the variable mydata. Two-way ANOVA was performed using the function aov and EEIC~site*season was used to include the interaction effect and the result was stored in the variable taov.

R Output

#	Df	Sum Sq	Mean Sq	F value	Pr(>F)
#site	3	226.3	75.44	35.173	< 2e-16 ***
# season	1	111.8	111.76	52.107	9.34e-13 ***
# site:season	3	46.4	15.45	7.204	8.58e-05 ***
# Residuals	1192	2556.6	2.14		

2.POST-HOC SNK:

Student–Newman–Keuls (or SNK) is a post hoc test for differences in means. Once an ANOVA has given a statistically significant result, you can run a Newman-Keuls to see which specific pairs of means are different. The test is based on the studentized range distribution.

Null hypothesis:

Ho: mean A = mean B

Alternate hypothesis:

Ha: mean A ≠ mean B.

Where A and B could be any possible pair.

Step 1: Order means from largest to smallest.

The next steps are to compare differences between the group with the largest mean to the group with the smallest mean:

Step 2: Calculate the standard error using the mean squared error (from ANOVA output). If the sample sizes are equal, use formula 1. If they are not equal, use formula 2.

$$\textbf{(1)} \quad s_{AB} = \sqrt{\frac{\text{MS Error}}{n}} \qquad \textbf{(2)} \quad s_{AB} = \sqrt{\frac{\text{MS Error}}{2}\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}$$

Step 3: Calculate the q-value using the following formula:

$$q = \frac{(\bar{x}_A - \bar{x}_B)}{s_{AB}}$$

Step 4: Find the q critical value from the q critical value table. The rows are the number of means being compared (i.e. the number of treatments) and the columns are the degrees of freedom.

Step 5: Compare the calculated value (Step 3) with the table value (Step 4). If the calculated value is greater than the table value then reject the null hypothesis that the two means are equal (i.e. the means are therefore not equal). If the calculated value is lower than the table value then the null hypothesis stands that the two means are equal.

If the two means are equal, stop the test. You can conclude that there is no difference between any pairs of means.

If the two means are unequal, repeat Steps 2 through 5 for the next highest mean and the lowest mean. Stop when you find a pair of means that are equal

Post-hoc SNK in R

The function SNK.test in the package agricolae was used for the post-hoc SNK test. The test was performed on the variable taov

R Output

R result
#Student Newman Keuls Test
#for EEIC

#Mean Square Error: 2.144839

#site:season, means

#	EEIC	std	r	Min	Max
#Areia Vermelha:Dry	7.202437	1.7086304	150	1.211692	10
#Areia Vermelha:Rainy	8.265309	1.6388326	150	1.623977	10
#Formosa:Dry	7.900693	2.1433783	150	1.143116	10
#Formosa:Rainy	8.572126	1.6736032	150	1.380973	10
#Picãozinho:Dry	8.952465	0.9635968	150	3.519417	10
#Picãozinho:Rainy	8.933434	0.9309057	150	2.571303	10
#Seixas:Dry	8.089813	1.2303880	150	3.550194	10
#Seixas:Rainy	8.815971	0.9217077	150	3.411704	10

#Alpha: 0.05 ; DF Error: 1192

#Critical Range
#2	3	4	5	6	7	8
#0.3317844	0.3968375	0.4350598	0.4619984	0.4826972	0.4994426	0.5134649

#Means with the same letter are not significantly different.

#	EEIC	groups
#Picãozinho:Dry	8.952465	a
#Picãozinho:Rainy	8.933434	a
#Seixas:Rainy	8.815971	a
#Formosa:Rainy	8.572126	ab
#Areia Vermelha:Rainy	8.265309	bc
#Seixas:Dry	8.089813	c
#Formosa:Dry	7.900693	c
#Areia Vermelha:Dry	7.202437	d

RESULTS AND INFERENCES

A total of 30 macrobenthic species were found (Only those species were considered that had a coverage of >3%)

The p-value obtained in R for two-way ANOVA is very less for all the rows and hence we reject all the null hypotheses and conclude that the variation in EEI-c responded to the interaction between the seasons and reefs.

The significant interaction between the reefs and seasonal factors may be explained by the EEI-c values of Picaozinho during the rainy season. All the other reefs had the lowest EEI-c values during the dry season. Even though the tourist visitation rate is highest during the dry season, the temporal variation in the EEI-c values cannot be attributed to tourism because Formosa Reef is not subjected to tourist impacts and exhibited the same variation. Picaozinho Reef receives the highest tourist visitation during the dry season, but the lowest EEI-c values were found during the rainy season . The temporal variations in the EEI-c values are not responses to the intensity of tourist visitation but rather were impacted by factors not evaluated in the present study.

The post-hoc SNK test helped us compare which pairs of means are significantly different. Means are said to be in a same group if they are not significantly different. The letters indicate whether or not the means belong to the same group. Here mean EEI-c values of Picaozinho:dry , Picaozinho:rainy, Seixas:rainy and Formosa:rainy are not significantly different. The mean EEI-c values of Formosa:rainy and Areia Vermelha:rainy are not significantly different. Similarly the mean EEI-c values of Areia Vermelha:rainy, SeiXas;dry and Formosa:dry are not significantly different. However the mean EEI-c of Areia Vermelha:dry differs significantly from all the others.

The post hoc SNK test indicated that:

1.During the dry season, Areia Vermelha Reef differed from the other reefs because it had the lowest mean EEI-c.This result may be explained by the fact that Areia Vermelha reef is the only reef bordered by a sandbank which generates high silt levels, or by the fact that Areia Vermelha reef is visited more often than the other reefs and is consequently more subject to the impacts of tourism.

2. There were no differences in EEI-c between the Seixas and Formosa reefs in both the seasons.This may be a response to the great proximity of the reefs to the estuaries of the Paraíba and Cabelo rivers, respectively. The proximity of estuaries impacts the physical-chemical patterns of the body of seawater to which the reefs are subjected, consequently impacting the reef community

3. TUKEY'S POST-HOC TEST

Tukey's analysis is a post hoc test is analogous to carrying out t-tests on every pair of treatment to discern which exact pairs have a significant difference amongst them. Just like the SNK post-hoc test this is done after carrying out an ANOVA test on the treatment and the factors.

We define Tukey's criterion as

$$T=q_{\alpha(k,\, n-k)}(MSE/n_k)^{0.5}$$

Where:

α : Confidence level

n: total number of observations

k: total number of treatments

n_k: number of observations per treatment

q: the studentised range statistic, determined by α, the distribution where the q comes from is determined from k, n-k

MSE: mean square error

NOTE: If the number of observations per treatment is not equal, instead of Tukey's post-hoc test, we use Tukey-Kramer post-hoc test. For unequal sample sizes, the confidence coefficient is greater than 1 – α if we use Tukey's post-hoc test. In other words, the Tukey method is conservative when there are unequal sample sizes.

Under this test we can also estimate the confidence interval of the mean differences by the following formula:

$$\bar{y}_{i\bullet} - \bar{y}_{j\bullet} \pm \frac{1}{\sqrt{2}} q_{\alpha;\, r,\, N-r}\ \hat{\sigma}_\epsilon \sqrt{\frac{2}{n}} \quad i, j = 1, \ldots, r;\ i \neq j\,.$$

Where:

N: total number of observations

$\bar{y}_{i\bullet}$: the mean of the i^{th} treatment

r=k: number of treatments

$\hat{\sigma}_\epsilon$: is the standard error of all the observations

n: the number of observations in one treatment.

This is essentially a t-test being carried out for every treatment pair, except that it is correcting for family-wise error rate.

Family Wise Error Rate:

This is the probability of making at least one type I error. In this context it would mean the probability of making at least one type one error while evaluating all the pairs.

Let the number of pairs of treatments be $h(={}^kC_2)$.

When we use a direct t-test for all the pairs instead of the post-hoc, the probability of making a type 1 error in one test is given by α.

The probability that we make at least one type error in the entire testing of pairs is given by

$$P(E)= 1-(1-\alpha)^h$$

As $0 \leq \alpha \leq 1$, we get $0 \leq 1-\alpha \leq 1$. This would mean that $(1-\alpha)^h$ would decrease as h increases.

Hence, we see that as the number of treatments (=k) increase, the family wise error rate increases very rapidly.

Tukey's test corrects for this as it uses the studentised range distribution.

We used Tukey's analysis to as an additional post-hoc test on the same dataset to compare the variation between the EEI-c values of the reefs in different in the dry and rainy season.

ASSUMPTIONS:

1. The observations being tested are independent within and amongst the group.
2. The sample means for each group are normally distributed
3. There is equal within-group variance across the groups associated with each mean in the test

NULL HYPOTHESIS: all means being compared are from the same population (i.e. $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$), hence the means are normally distributed (central limit theorem).

PROCESS: If $|\mu_{treatment1} - \mu_{treatment2}| > T$, the Tukey's criterion, then it can be concluded that there is a significant difference between the means with $(1-\alpha)$ % guarantee. Hence we can conclude that treatment 1 is better than treatment 2, if $\boldsymbol{\mu_{treatment1} > \mu_{treatment2}}$.

R Output:

We ran the TukeyHSD command over an anova on the data of reefs and EEI-c value for each sample for only the dry months first. Codes are in Tukey's post hoc.r

CONCLUSION:

- DRY MONTHS:
  - We see that there is a positive difference between all the other reefs and Areia Vermelha
  - None of the intervals contain 0.
  - All the three have a significant p value ($p<0.01$).

  Hence, we can conclude that Areia Vermelha has a poorer EEIC value than the rest of the reefs, which is in agreement with what we found in the SNK post-hoc test

  - Formosa(the relatively pristine reef) has worse mean EEIc values than Picaozinho and does not differ from Seixas.
- RAINY MONTHS:
  - Here, we find only two pairs with a completely positive interval and a significant adjusted p-value:
    - Seixas-Areia Vermelha
    - Picaozinho-Areia Vermelha
  - Also, there is no significant difference between the EEIC of Formosa and Seixas, just like in dry season, which confirms another one of SNK post-hoc results.
- ALL THE MONTHS COMBINED:

- Surprisingly, Picaozinho outperforms Formosa.
- Areia Vermelha has underperformed with respect to every reef.
- Seixas has lower mean than Picaozinho.

4.INDVAL:

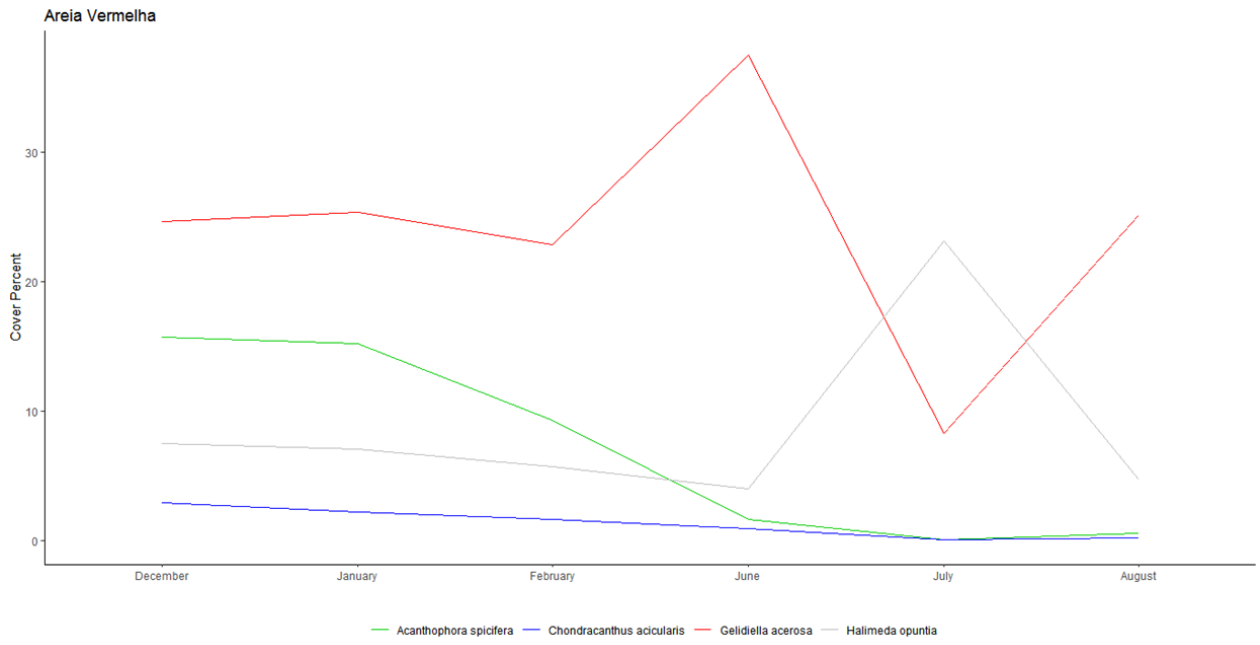
$$IndVal_{ij} = 100 \cdot A_{ij} \cdot B_{ij}$$

A_{ij} is specificity, i.e. the proportion of the individuals of species *i* that are in class *j*

B_{ij} is fidelity, i.e. the proportion of sites in class *j* that contain species *i*

Using this result from Dufrene and Legendre(1997) we obtained the values for all the 30 species for all the four reef.

We plotted the coverage area of top four IndVal species for Areia Vermelha, the R codes of which are in IndVal plot Areia Vermelha.r

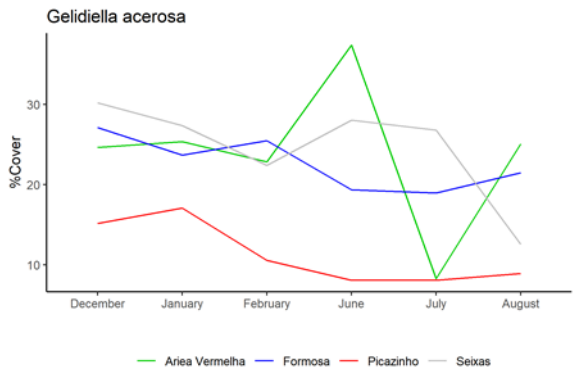


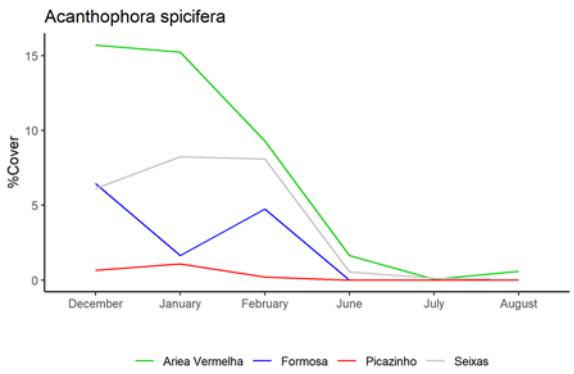
EEIc and IndVal

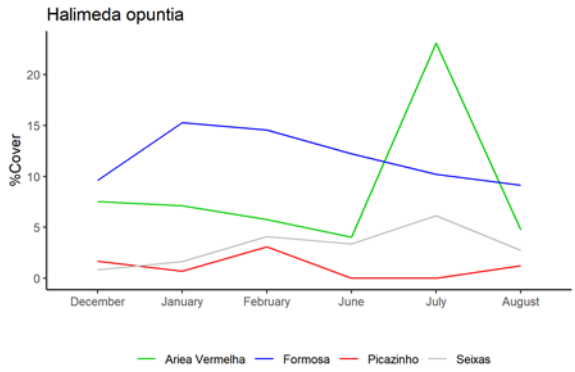
The species would belong to one of ESG1 or ESG2 and it is the abundance or the lack thereof of that determines the EEIc value. Since detrimental species and beneficial species might have the same IndVal, a high ranking IndVal species having an expansive coverage is not indicative of a good EEIc value.

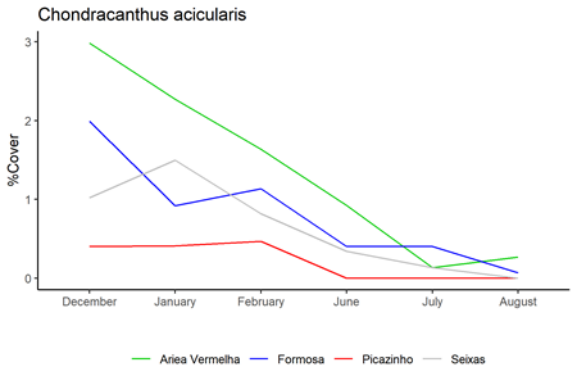
We have graphed the top four IndVal species(in order) for each reef against time. We have also simultaneously plotted their occurrence in other reefs. For the full plots of all thirty species in each reef, refer to Coverplots Stat Plots folder.

AREIA VERMELHA:

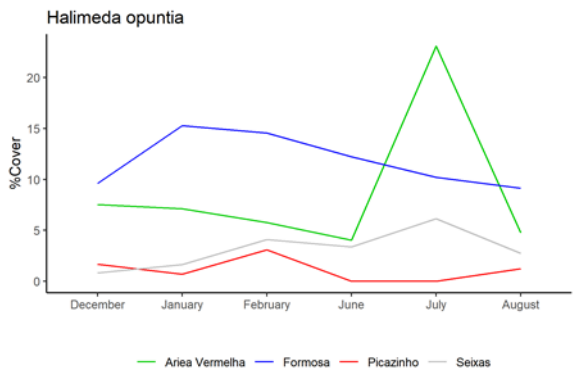


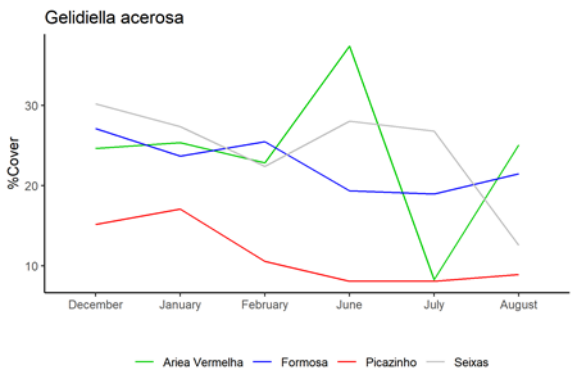


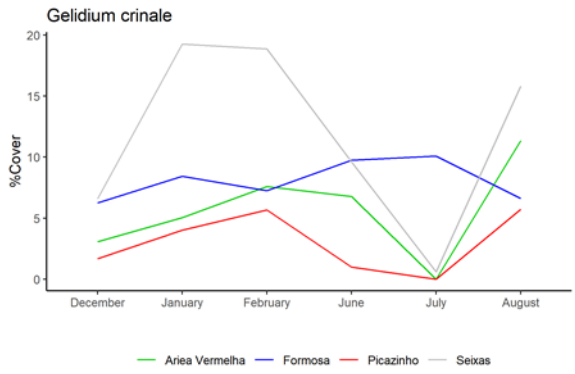


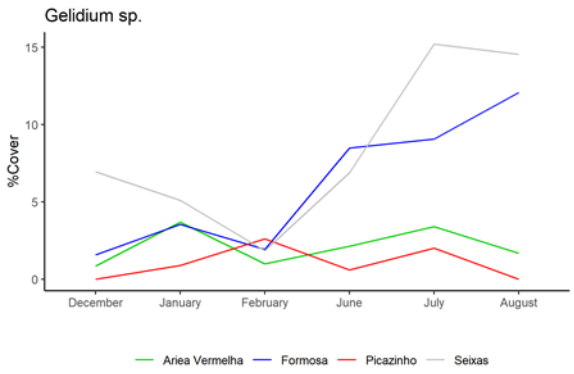


FORMOSA

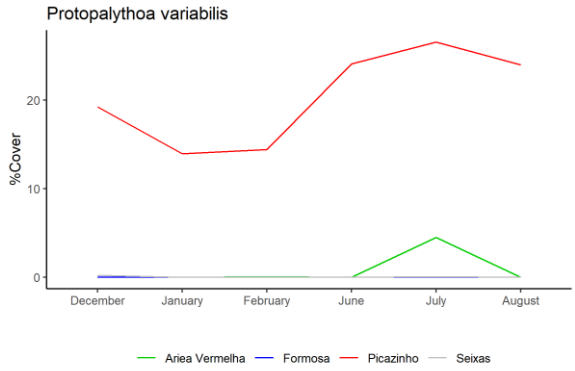


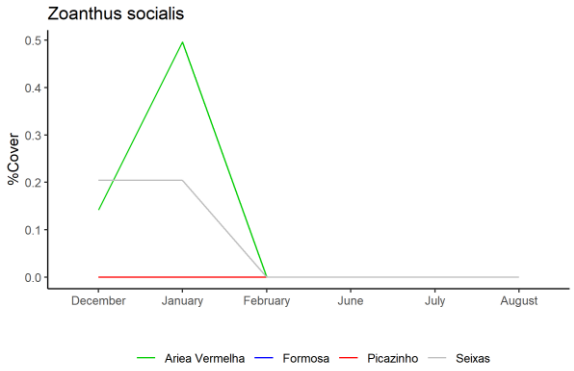


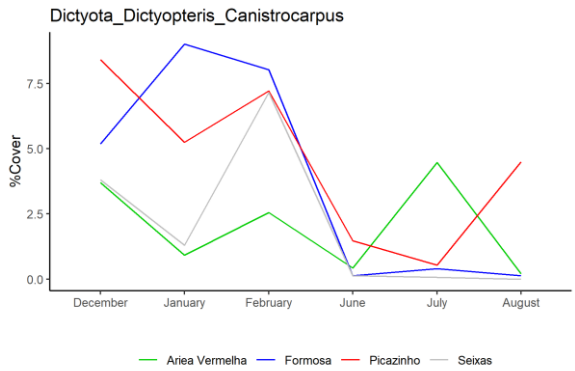


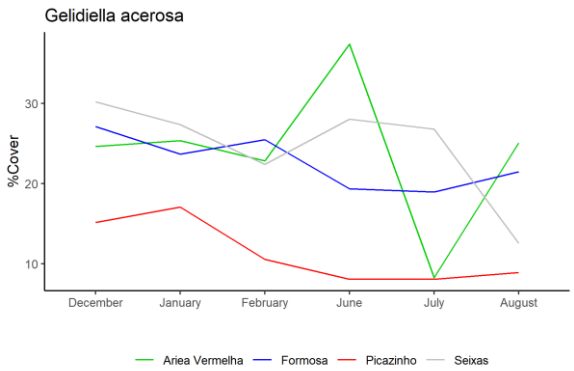


PICAOZINHO

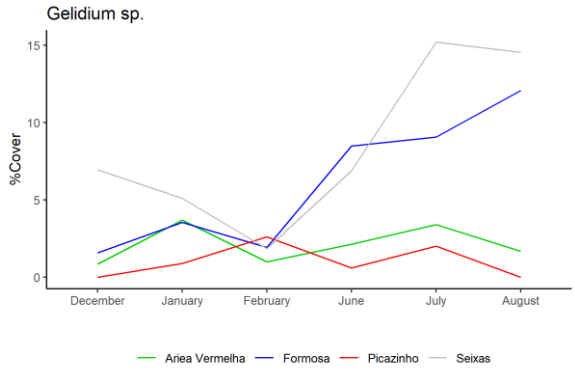


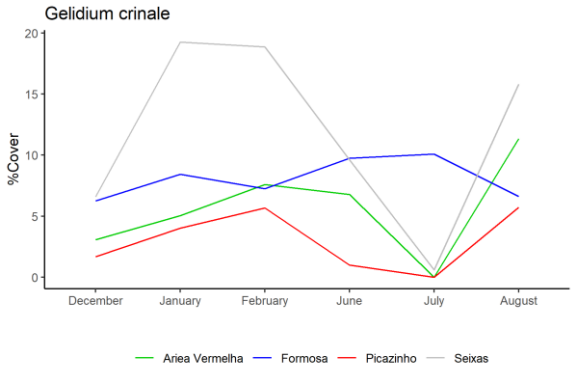


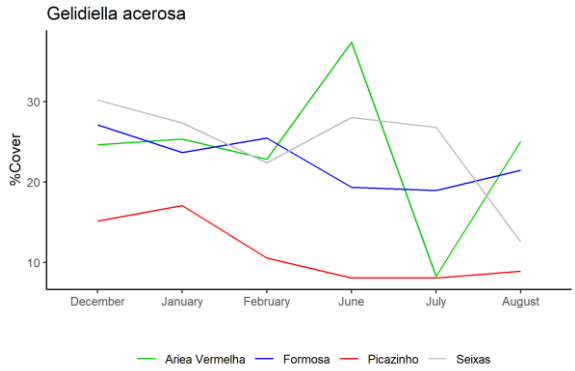


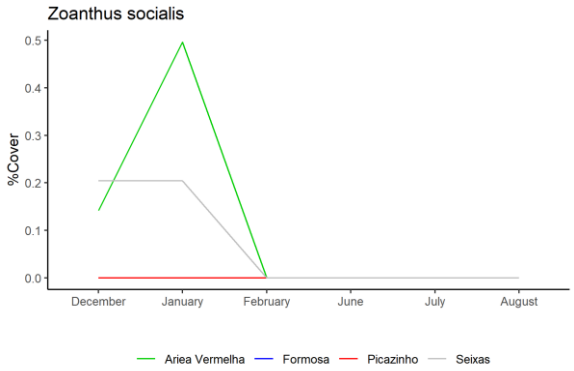


SEIXAS









R codes: Cover plots.r

PROBLEMS WE FACED AND HOW WE HANDLED THEM:

When we first started to calculate the EEI-c values, our results were always less than 1. Then we realized that there is a mistake in the formula given in the pdf and the actual expression is 2+8*min(expression given in pdf,1).This gave us the correct values. We crosschecked this formula with the website referenced in the pdf **https://www.eei.gr/index.html,** and also used their calculator to check our values. The second hurdle was when we couldn't get the interaction effect row in the output of two-way ANOVA , we corrected this by replacing site+season by site*season. However the p value that we obtained for interaction effect was high so we couldn't reject the null hypothesis, and we got that there is no interaction effect , unlike the result given in the pdf. We were stuck at this point for days until we noticed that the degrees of freedom of residuals was very low(16) compared to the value given in the pdf(1192). This was when we realized that the reason for our error is we used average EEI-c values per month instead of per sample in our ANOVA code. We got the interaction effect after we rectified this mistake.

Extra work in this project:

- **Tukey's Post-hoc test:** We re-analysed the EEI-c value Vs. Reefs through an additional post-hoc test- Tukey's post-hoc, to compare our results with SNK post-hoc that the paper had incorporated to analyze the same.
- **Family Wise Error Rate:** We elaborated on FWER to show why using multiple t-tests for post-hoc analysis was not suitable.
- **IndVal:** The paper had simply stated the IndVal of top four species in each reef with no calculation or explanation. We elaborated what IndVal actually signifies and calculated it from the ground up, including the variables necessary in the formula to find it.

Bibliography:

https://people.richland.edu/james/lecture/m170/ch13-2wy.html

https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-6-two-way-analysis-of-variance/

https://www.statisticshowto.com/newman-keuls/

https://www.eei.gr/index.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6193594/

REFERENCES FOR EXTRA WORK:

https://www.youtube.com/watch?v=zQr190cacC0

http://biol09.biol.umontreal.ca/PLcourses/Indicator_species.pdf

https://rfunctions.blogspot.com/2013/02/multivariate-analysis-indicator-value.html

https://en.wikipedia.org/wiki/Indicator_value

https://en.wikipedia.org/wiki/Tukey%27s_range_test

https://en.wikipedia.org/wiki/Family-wise_error_rate

https://methods.sagepub.com/reference/encyc-of-research-design/n478.xml