Biodiversity Measure

Atharva Rodge

2024-01-11

Introduction

The study below performs the statistical analysis and data analysis on the data-set having the taxonomic groups with species richness and the dominant land class. Starting with a selection of 5 from the 11 taxonomic groups for the study. Five taxonomic group which were select for this analysis following Bees, Birds, Hoverflies, Isopods, Grasshoppers_._Crickets. These five taxonomic groups, terming as selected_BD5 in the study. Then, univariate analysis performing on the selected_BD5 that focuses on calculating the characteristics of each selected variable from the selected 5. It involves measures like mean, median, minimum, maximum, 1st quantile, 2nd quantile and winsorized mean. Furthermore, the analysis concludes with the calculation of a correlation matrix between all of the selected BD5, which provides insight into how variables are related to one another. A box plot is then created for the variable 'Isopods'. In addition, the code performs hypothesis tests, such as 'T-Test' and 'KS-Test,' which generate p-values. Following that, a contingency test is run, comparing selected BD5 with actual BD11. Through this contingency table, the code calculates various parameters, such as Odds ratio, Sensitivity, Specificity, and Youden's index. These values provide understanding of the relationships and associations between the variables considered in the analysis. Afterward, the program conducts linear regression and multiple linear regression. The code's primary purpose is to explain the relationships between species and dominating land classes, detect trends or patterns, and investigate potential causes of changes in biodiversity indices across time.

Univariate Analysis

Selected BD5 Summary

Names	**Minim…**	**Q1**	**Medi…**	**M…**	**Q2**	**Maxi…**	**Winsorized**
<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>	<dbl>
Bees	0.03	0.35	0.59	0.61	0.82	3.31	0.60
Bird	0.24	0.85	0.9	0.89	0.96	1.17	0.89
Hoverflies	0.12	0.57	0.7	0.68	0.81	1.15	0.68
Isopods	0.05	0.39	0.54	0.55	0.72	1.26	0.55
Grasshoppers_._Crickets	0.07	0.49	0.62	0.63	0.79	1.59	0.63

5 rows

The above table presents summary statistics for five group's Bees, Bird, Hoverflies, Isopods, Grasshoppers_._Crickets in the BD5 group. It includes traditional values and an additional statistic – the 20% winsorized mean. These mean values provide a summary measure of central

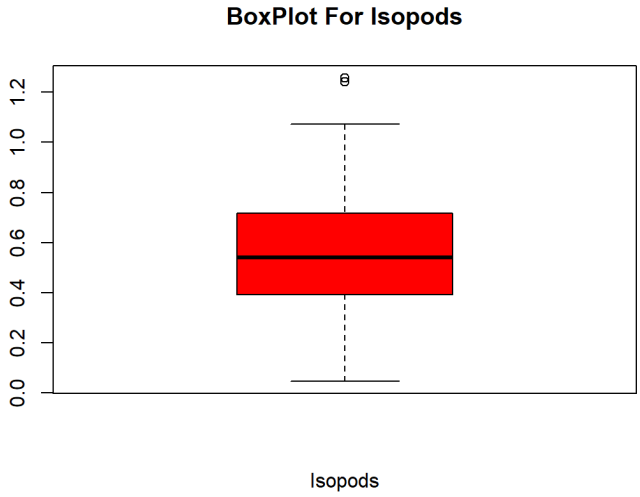
tendency for each variable across the specified species. As in the table, we can see that the mean mirrors for every species expect Bees with a very small value so we can say that there is no variability in the dataset and between the species. The median value, 5 variables is 0.62 we measured median the for central tendency and is the middle point of the data set, indicating that about 50% of the values are below 0.62 and 50% are above this value 0.62. The quarter's 1 and 2 provide insights into the spread of the data, and the minimum and maximum values help identify the range of observations.

Correlation Matrix Between selected BD5

```
##                          Bees      Bird Hoverflies   Isopods Grasshoppers_._Cri
## Bees                1.0000000 0.7309609  0.7650558 0.5728611          0.4793253
## Bird                0.7309609 1.0000000  0.6283667 0.4283062          0.4896522
## Hoverflies          0.7650558 0.6283667  1.0000000 0.7112050          0.3599762
## Isopods             0.5728611 0.4283062  0.7112050 1.0000000          0.3781288
## Grasshoppers_._Cri 0.4793253 0.4896522  0.3599762 0.3781288          1.0000000
```

The correlation matrix shows the pairwise relation between different variable pairs. Each value in the matrix represents the correlation coefficient between each pair of variables ranging from 0 to 1. Finally, we can say that 'Bees' and 'Bird', 'Bees' and 'Hoverflies', and 'Hoverflies' and 'Isopods' these species have strong and positive correlation between each other. Followed by 'Bees' and 'Isopods', 'Bees' and 'Grasshopper_._Crickets', 'Bird' and 'Isopods', & 'Hoverflies' and 'grasshoppers_._Crickets' these species show a weak correlation among themselves. The most strong correlation is between 'Bees' and 'Hoverflies' and the most weakest correlation is between 'Hoverflies' and 'Grasshoppers_._Cricket'. Moreover, as we can see the diagonal values are interpreted as 1.00 which is the correlation of each variable with itself, which is always 1.00 it never changes. The matrix provided us insights of the relationship between each variable and we can find the weak and strong correlation between the variables.

Box Plot



The boxplot of Isopods denotes the maximum, minimum, median, and quantile values. The minimum value of Hoverflies is close to 0.05 and maximum value is around 1.2 from the graph. The, black line inside the red rectangular box denotes median which is around 0.5 from the box

plot. The horizontal line of the red rectangular denotes 1st and 3rd quantile, it also shows that the 1st quantile is somewhere around 0.4 and 3rd quantile is somewhere below 0.8. This fact shows us that median lies in between 0.4 and 0.8, which is 0.5 approximately.

Hypothesis tests

Hypothesis T-test

```
## p-value For Hypothesis T-test: 0.02055546
```

The hypothesis, T-test is performed for dominant land class “Scotland” between two species Bees and Isopods. The p-value, which is calculated from t-test is observed as 0.02055546 which is less than a threshold of significance of 0.05. It shows ample proof against the null hypothesis. suggesting a large difference in means between ‘Bees’ and ‘Isopods’. This conclusion indicates that the observed data gives enough proof to indicate that there is a significant difference among ‘Bees’ and ‘Isopods’, and the null hypothesis is rejected using the hypothesis t-test.

Hypothesis Ks-test

```
## p-value For Hypothesis Ks-test: 1.110223e-16
```

The hypothesis ks-test is performed for dominant land class “England” between two species Bees, Isopods. The result of p-value from the ks-test is 1.110223e-16 which is less the 0.05 and it shows that the null hypothesis is rejected based on the performed Ks-Test. With such a small p-value (1.110223e-16), there’s strong validation to reject the null hypothesis. Thus, you can conclude that the distributions of ‘Bees’ and ‘Isopods’ significantly differ based on the Asymptotic Two-Sample Kolmogorov-Smirnov test.

Contingency table

Contingency table for Independent Model

```
##        Decrease Increase
## BD5up      1730      910
## BD11up     1638     1002
```

The contingency table of independent model shows the counts of occurrence for each combination of BD5up and BD11up. For BD5up there are 1730 times where biodiversity decreases and 910 times increases after calculating the change in two periods Y00 and Y70 periods of BD5, following same with BD11up there are 1638 instances where biodiversity decreases and 1002 instances where it increased. The table helps us understand the distribution of biodiversity changes independently in BD5 and BD11up. The numbers in each cell reflect the number of cases that fall within the given combination of BD5up and BD11up categories.

Contingency Table for BD11up against BD5up

```
##
##            Decrease Increase
##   Decrease     1462      176
##   Increase      268      734
```

The contingency table, above displays the number of occurrences for each combination of BD5up and BD11up categories, with a special focus on instances in which both BD5 and BD11 biodiversity change. In both BD5up and BD11up, there are 1462 cases of reduced biodiversity. There are 176 cases in which BD5 declines and BD11 increases. There are 268 cases in which BD5 grows and BD11 declines. There are 734 instances when both BD5 and BD11 increases. This table focuses on conditions where both BD5 and BD11 are tested. For example, 734 instances show an increase in both BD5 and BD11. It allows us to analyse the joint distribution of changes in BD5 and BD11.

Likelihood ratio test

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  independent_model
## G = 6.9419, X-squared df = 1, p-value = 0.00842
```

```
## Likelihood ratio test p-value for independent model  0.008420129
```

The likelihood ratio test for independent model gives us the p.value as 0.008420129 for further analysis.

```
##
##  Log likelihood ratio (G-test) test of independence without correction
##
## data:  contingency_observed
## G = 1119.5, X-squared df = 1, p-value < 2.2e-16
```

```
## Likelihood ratio test p-value for contingency table  0
```

The likelihood ratio test for contingency table gives us the p.value as 0.00 for further analysis.

```
## For Independent model table having p-value - 0.008420129 - Reject Null Hypothesis
```

```
## For Contingency Table Having p value - 0 - Reject Null Hypothesis
```

The G-test of independence, also known as the log likelihood ratio test, is a statistical test used to assess the independence between two categorical variables in a contingency table. As we can see the output of the likelihood ratio test the p. values for both the table are 0.008420129 and 0 which means that we have to reject the null hypothesis, since both the values are lower than of the confidence level, we would reject the null hypothesis in each case.

Odds Ratio (OR)

```
## Odds Ratio: 22.75076
```

An odds ratio of 22.75076 implies a strong positive association between the increase and decrease in the contingency table. The probabilities of the event occurring in an increase are more than 22 times than in a decrease, showing a considerable and statistically significant link between the variables.

Sensitivity

```
## Sensitivity: 0.8065934
```

A sensitivity of 0.8065934 shows that the test correctly detected the problem in almost 80% of situations when it was present. It illustrates how well the test captures and keeps track of actual positive cases, proving its capacity to identify the condition properly.

Specificity

```
## Specificity: 0.8450867
```

This high specificity indicates that the test is effective at avoiding false positives while correctly identifying true negatives.

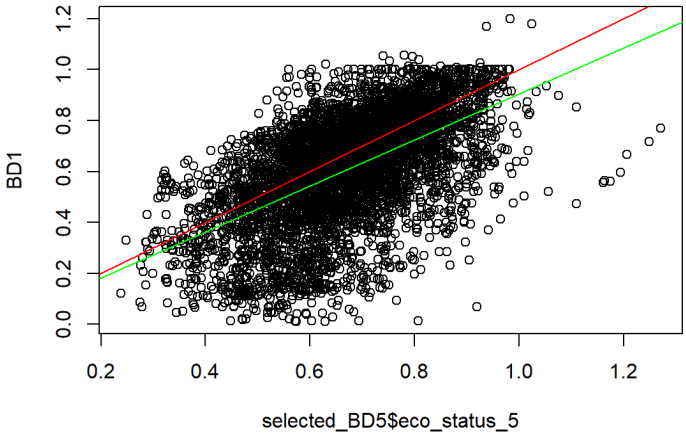
Youden’s Index

```
## Youden’s Index: 0.6516801
```

The high Youden’s Index further supports the overall good performance of the diagnostic test. The value represents the test’s overall performance, taking sensitivity and specificity under consideration. A higher Youden’s Index shows that the test has better overall discriminating ability. In our case, a Youden’s Index of 0.6516801 suggests that your test achieves an acceptable balance between sensitivity and specificity.

Simple Linear Regression

The simple linear regression is done on variable ‘Carabids’ which is selected from BD11 other than selected BD5 this variable is denoted as BD1 for the further analysis against ecological status of the 5 selected variables named eco_status_5. IN the above scatter plot the x-axis shows the mean ecological status of 5 selected variables and y-axis show the abundance of ‘Carabids’’ The pattern shows us there is a positive linear relation between these variables. The red line is indicated as regression line and is easily observed that the points are following a linear pattern with rising pattern i.e as eco_status_5 rises there is a rise in BD1 ‘Carabids’. A green line’s positive slope is significantly distinct from zero, which supports this association. The above figure depicts the relationship between the two variables and how strongly they are linked they have a positive correlation between the two variables. The additional line which is green line represents the best fit line or alternative regression line as well.



Slope

```
## Estimated Slope: 0.901095
```

The slope of the regression line 0.901095 indicates high strength between BD1 and eco_status_5. A positive slope indicates positive relation between the variables i.e if eco_status_5 increases BD1 increases and have strong relation.

Multiple Linear Regression

AIC (Initial Model)

```
## AIC initial model: -3802.147
```

```
## Maximun p.value for reduced model: 0.06344796
```

The AIC also known as Akaike Information Criterion helps us too measure the relative quality of a statistical model for a given set of data. It balances The level of fit among the model & the complexity of the model, penalizing models with more parameters. The lower the AIC, the better the model is considered. The output value of -3802.1 suggests that the initial MLR model has a good balance between fitting the data well and avoiding over fitting.

```
##
## Call:
## lm(formula = BD1 ~ ., data = selected_BD5[c(species_5)], y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71546 -0.09924  0.02634  0.11700  0.50895
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.097567   0.020221  -4.825 1.44e-06 ***
## Bees                     0.015951   0.008592   1.856   0.0634 .
## Bird                     0.215568   0.025074   8.597  < 2e-16 ***
## Hoverflies               0.279383   0.016773  16.657  < 2e-16 ***
## Isopods                  0.263000   0.011802  22.285  < 2e-16 ***
## Grasshoppers_._Crickets  0.269129   0.012662  21.254  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1687 on 5274 degrees of freedom
## Multiple R-squared:  0.3847, Adjusted R-squared:  0.3841
## F-statistic: 659.4 on 5 and 5274 DF,  p-value: < 2.2e-16
```

AIC reduced model

Next for feature selection selection we have to eliminate one variable considering its p.value from the summary of the AIC initial model. The p value suggests us that it failed to reject the null hypothesis from the data output of the summary. so as the p.value of the “Bees” is 0.06 which is above our significance level so we remove the species variable Bees for AIC of reduced. Now, as we can see that the AIC value of reduced model is slightly lower than initial model which indicates us a better fitted model therefore the reduced model is preferred over initial model.

	df <dbl>	**AIC** <dbl>
initial_mlr_model	7	-3802.147
initial_mlr_model_reduced	6	-3800.698

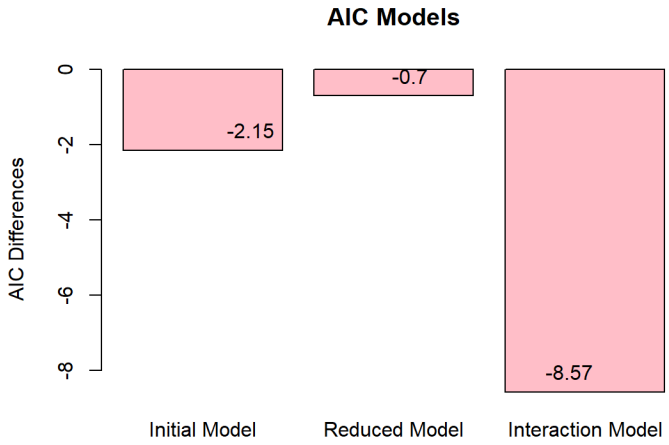
2 rows

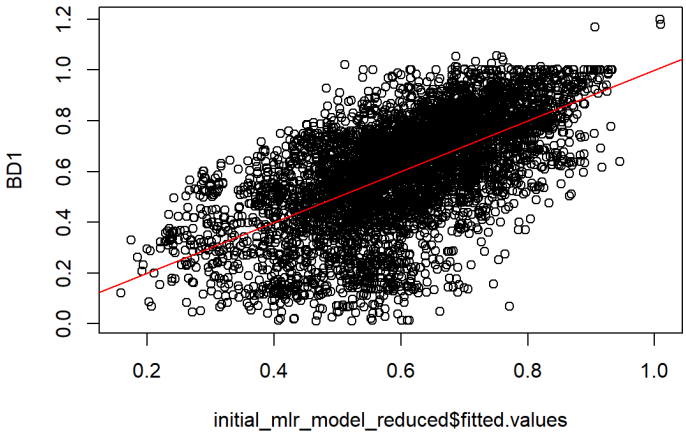
AIC interaction model

The next step is to perform the AIC for interaction model where we multiply two variables. The first variable we took is ‘Bees’ which we eliminated earlier and for the second variable to multiply with ‘Bees’ we will analyse the summary of the reduced MLR model the p.values for all the variables are extremely small, which indicates that each predictor is likely to have a statistically significant effect on response variable so all the variables equally contributes to the model. So for interaction we took ‘Grasshoppers_._Crickets’ as a multiplier variable with ‘Bees.’

	df <dbl>	AIC <dbl>
initial_mlr_model	7	-3802.147
initial_mlr_model_reduced	6	-3800.698
initial_mlr_model_interaction	8	-3808.573

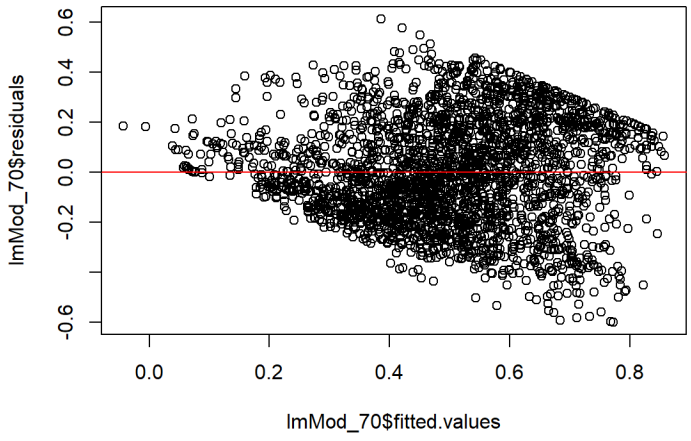
3 rows

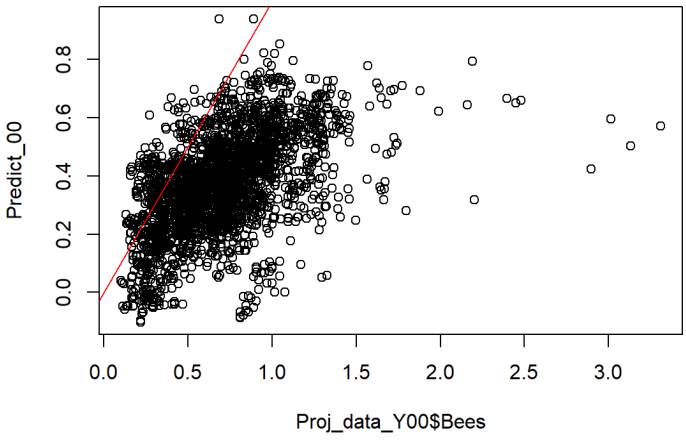




For the above bar plot we can observe the values of all the AIC models and we can conclude that the lower AIC is of reduced. and we have performed a Linear regression model test for reduced model above. Moreover, there is a residual plot of the best AIC model which is reduced AIC model as it is the best fitted model. As observed the plot is linear and the values are properly fitted around the red line.

MSE (Mean Square Error)






```
## Mean square error on train set: 0.04329819
```

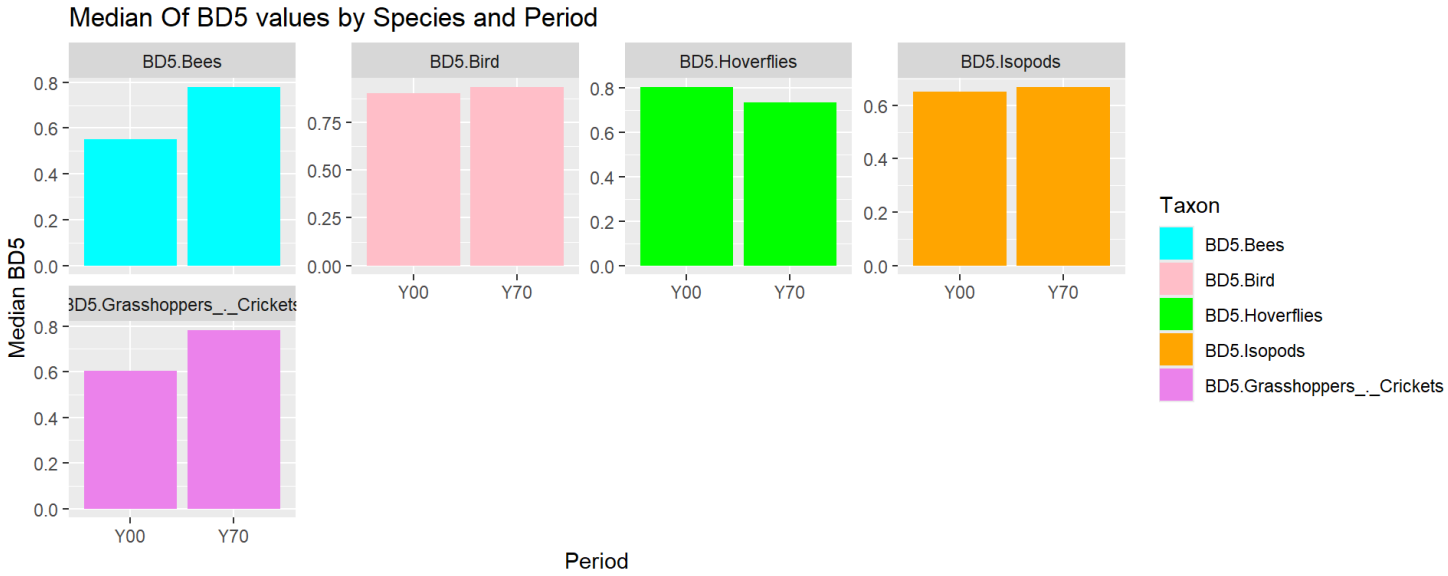
```
## Mean square error on test data set: 0.188409
```

The project data is divided into two sub data sets considering two periods Y70 and Y00. Filtered period data of Y70 is taken for training and Y00 is taken for testing. Then, MSE is calculated on both the training and testing data. MSE on the training set measures how well the model fits the data it was trained on. A lower MSE indicates better model performance. In our case, the MSE is lower on the training set 0.04329819 compared to the test set 0.188409 i.e the the training set fits the training data well. MSE on the test set measures the model's performance on new, unseen data. A lower MSE on the test set indicates better generalization, but our data might indicate that the model doesn't generalize well to new data.

Open Analysis

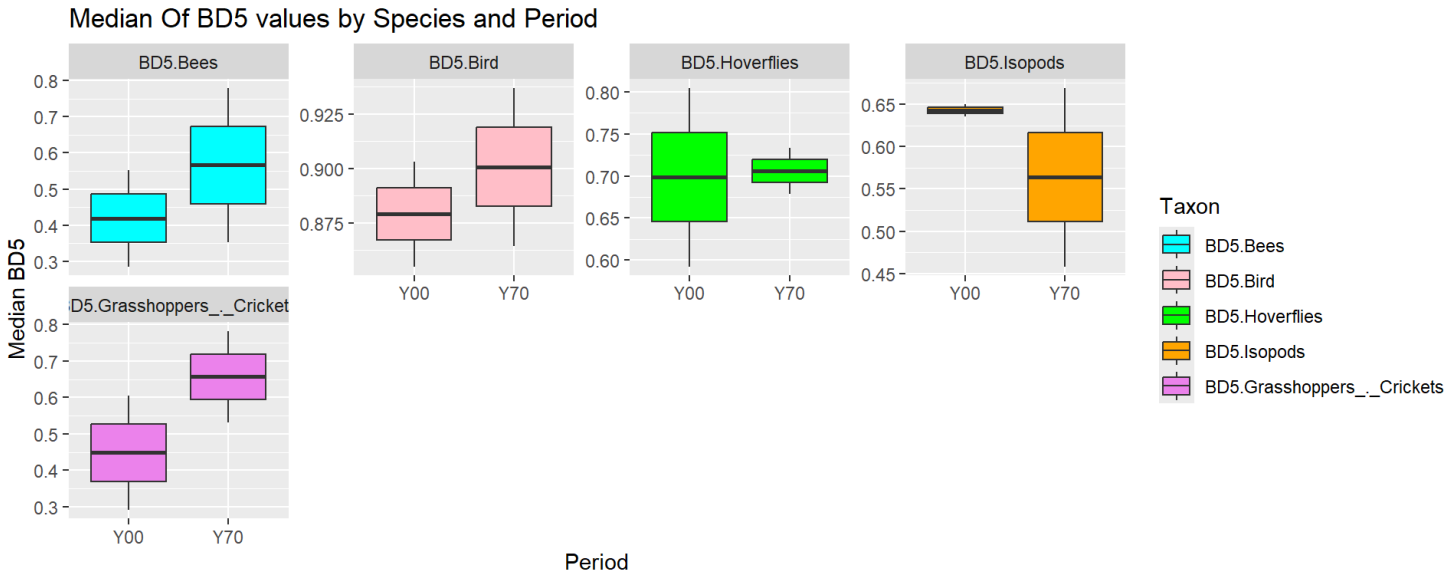
Bar Plot

The code first combines the time period and dominating land class of the specified five groups BD5. A new data is created with median values for each land type and two time period. With the help of ggplot it then generates the bar plot, where the block is of one variable of 5 group. These bars are represented by a distinct color. The Bar plot illustrates the median values for period Y00 and period Y70 for land class "Coastal plains/soft coasts, S-W Scotland", "Isolated hills/mountain summits, W Scotland". The 5 bar plots are for 5 different species which were selected at the start of the project. Plotted for Median of BD5 against two periods. The bar plot as we can see above shows which period of particular variable has greater median value.



Box Plot

Again, the Box plot is created which illustrates the median values for period Y00 and period Y70 for land class "Coastal plains/soft coasts, S-W Scotland", "Isolated hills/mountain summits, W Scotland". The box plot can show us the minimum, maximum, 1st quantile, 2nd quantile and mean/median. The 5 bar plots are plotted each for one specie of BD5. The box is plotted with median of BD5 on y-axis and two periods Y00, Y70 on x-axis. For example, if we look for the box plot of "Bees" we can find that the minimum value is somewhere around 0.3, the 1st quantile is approximately 0.35, median is somewhere near 0.42, 2nd quantile near 0.5 and the max value 0.55. Considering these we can interpret the summary of the data from the box plot.



Conclusion

The study performs various tests and helps us to analyse the data executing BD5 summary i/e of 5 species and correlation of these species. We determined that the strong relation is between “Bees” and “Hover flies”. Then by the use of box plot we determined various values of specie “Isopods” eg min, max, median, quantiles which we can also calculate for other species using box plot. The code then follows hypothesis testing performing t.test and Ks-test. The T-test gives us evidence for the null hypothesis. Moreover, the simple and linear regression is performed on specie ‘Carabids’ against selected 5 species. the graph shows us the strong relation between the species and indicates positive relation between the variables.the plot shows reduced AIC model as it is the best fitted model. the plot is linear and the values are properly fitted around the red line. In conclusion. the study helps us developing idea to analyse the graphs and relation between different variables and the plot assist us for further analysis.