Analysis of Biodiversity using proportional species richness

1. Introduction

The document provides data on biodiversity measures for seven taxonomic groups. The biodiversity measure used in this analysis is the mean of the proportional species values for the seven allocated taxonomic groups (BD7). The authors also provided the mean of all 11 taxonomic group proportional species values (BD11) as a point of comparison.

The analysis presented in this document provides insights into the biodiversity measures for the seven taxonomic groups, which can inform conservation efforts and policy decisions aimed at protecting and preserving these ecosystems.

2. Methods

To achieve the Analysis objectives following methods are followed:

2.1 Data cleaning

Aim of data cleaning is to ensure that the data is accurate, reliable, and suitable for analysis.It involves several steps,Which will include : 1.Identifying missing values(NA): Function such as is.na() used to identify them.Later can be removed using na.omit().As the data set used in this analysis has no NA values.Moving to next step. 2.Identifying duplicate values :Function such as unique(), duplicated() can be used to find them.Later can be removed using ifelse(), replace().

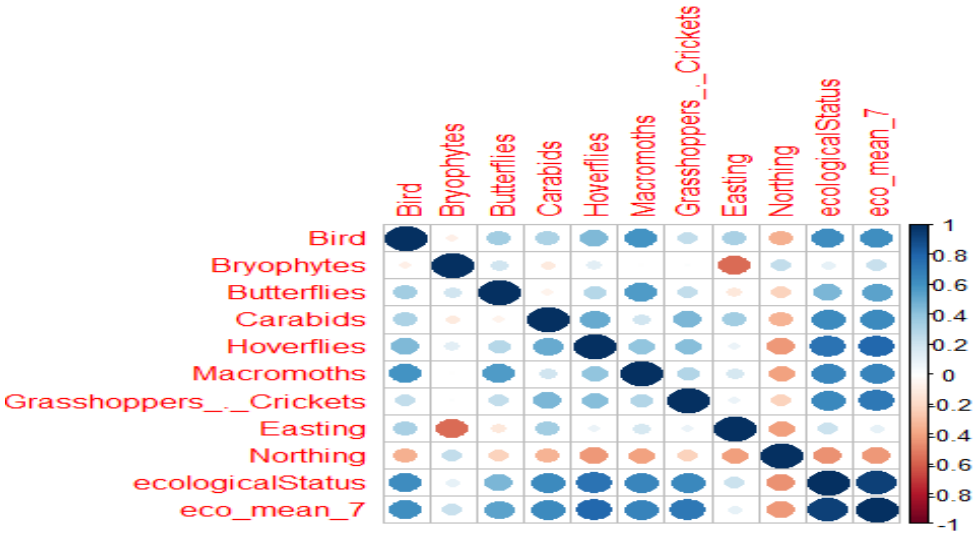
3.Incorrect data types can lead to data analysis errors. So data types can be checked using typeof() or class().Later can be converted using as.factor() , as.numeric() or as.date(). As period and dominantLandClass are characters, they can be converted into factors using as.factor().

4.To improve efficiency in analysis, picking 7 species from group of 11. Which are “Bird”, “Bryophytes”, “Butterflies”, “Carabids”, “Hoverflies”, “Macromoths”, “Grasshoppers_._Crickets”. Later comparing the 7 species statistics with 11 species group.

5.Calculating mean of 7 species(eco_mean_7) which was filtered from 11 species. So that, later it can be compared with mean of 11 species.

2.1 Data exploration

2.1.1 Correlation Analysis



Based on the values from above correlation matrix, it was observed that BD7 and Hoverflies show a positive linear relationship. Similarly, few more positive and negative linear relationship was observed so, Let's draw plots and visualize the relationship between these variables.

Fig 1 Scatter plot

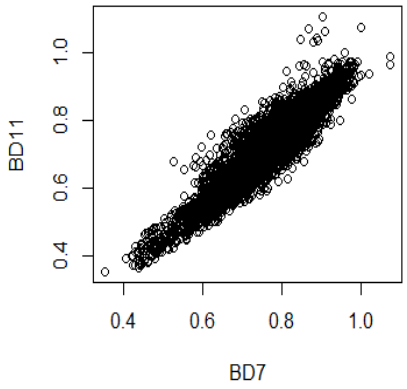
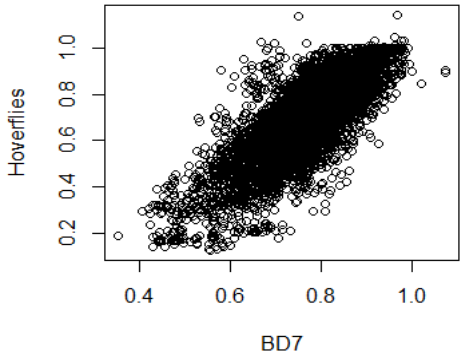
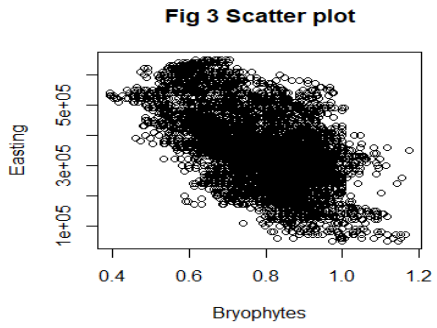
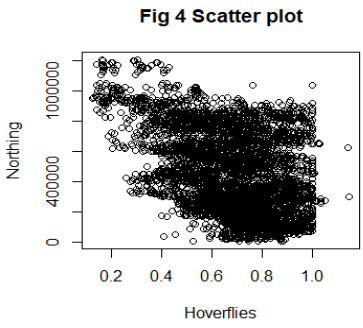


Fig 2 Scatter plot

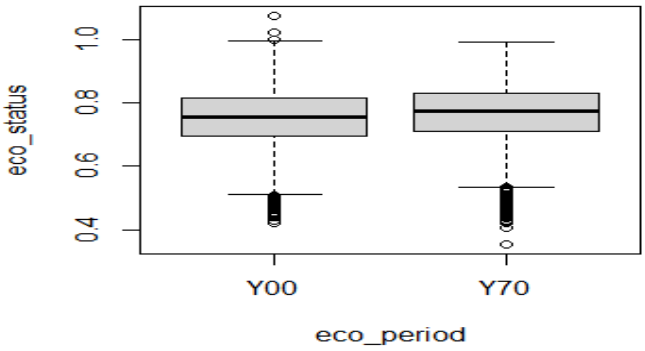


It can be observed that, mean value of 11 species (BD11) has a positive linear relationship (0.92) with mean value of 7 species (BD7). And also mean value of Hover flies has a positive linear relationship (0.78) with mean value of 7 species (BD7).





It can be observed that, mean value of Easting has a negative linear relationship(-0.57) with mean value of Bryophytes .And also mean value of Northing has a negative linear relationship(-0.44) with mean value of Hoverflies.



Above Box plot shows relationship between mean of 7 species and two periods. It can be observed that outliers are less in period Y00 compared to Y70. And also BD7 is more for Y70 compared to Y00.

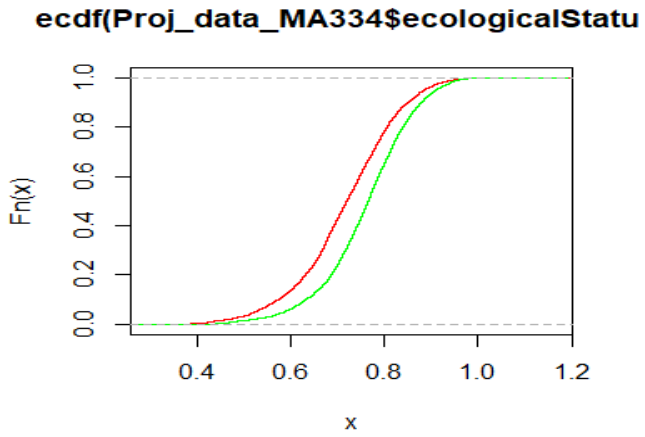
2.2 Hypothesis Test

Firstly selecting 3 columns(Location,period,eco_mean_7) from dataset and storing it in variable Proj_data_MA334_period. Later using pivot_wider() function from the tidyverse package to pivot the data frame Proj_data_MA334_period, converting period column into wider format where period value becomes column name and values of eco_mean_7 is used as value for new column.mutate() adds a new column BD7_change, which is calculated as the difference between the values in the Y00 and Y70 columns. t.test() performs a one-sample t-test, with 'BD7_change' as the input data and 'mu=0' as the stated null hypothesis. t.test() will compare the mean of sample of 'BD7_change' with parameter 'mu'.Which means testing whether the true mean of BD7_change is equal to 0.

```
One Sample t-test
data:  BD7_change
t = -12.456, df = 2639, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01505612 -0.01096043
sample estimates: mean of x
              -0.01300827
```

As shown above output of t-test is displayed, which includes t-statistic, degrees of freedom, p-value, alternative hypothesis,95% confidence interval, and the sample estimate of the mean of BD7_change. The p-value is less than 2.2e-16, which is very small, indicating strong evidence against the null hypothesis.The alternative hypothesis suggests that the true mean of 'BD7_change' is not equal to 0. The 95% confidence interval for the difference between the sample mean and the hypothesized population mean does not contain 0,further supporting the rejection of the null hypothesis.The sample estimate of the mean of 'BD7_change' is -0.01300827.The negative t-value of -12.456 and the large degrees of freedom of 2639 suggest that the sample mean of 'BD7_change' is significantly lower than the specified value of 0,indicating that there is one more evidence to reject the null hypothesis and support the alternative hypothesis that the true mean of BD7_change is not equal to 0.

Performing Kolmogorov-Simirnov test using function ks.test.Before that creating an empirical cumulative distribution function for mean of 7 species(BD7) using ecdf function.Similarly calculating ecdf for mean of 11 species(BD11). Here ecdf is calculated to finally estimate the distribution of the variable.later ploting ecdf function for both variable.Finally performing ks.test to compare the distributions statistically, with the null hypothesis(H0) being that the distributions are the same.




```
Asymptotic two-sample Kolmogorov-Smirnov test
data:  Proj_data_MA334$eco_mean_7 and Proj_data_MA334$ecologicalStatus
D = 0.19299, p-value < 2.2e-16
alternative hypothesis: two-sided
```

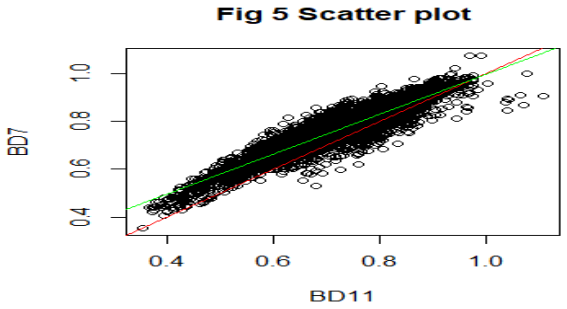
As shown in graph above, there is a gap between the lines of the plots 'BD11_cdf'(represented by red) and 'BD7_cdf'(represented by green),it indicates that ecdf function values of variables eco_mean_7 and ecological Status are not identical.In other words there is difference in the distributions of the two variable. As shown above in the ks.test() output, D = 0.19299 is the maximum absolute difference between the empirical distribution functions of two data frame.p-value is a measure of strength of evidence against the null hypothesis,which states the distributions of the two data sets are the same.In this case the small value of D and the very low p-value(2.2e-16) suggests that there is strong evidence to reject the null hypothesis and conclude that the distributions of eco_mean_7 and ecological status are significantly different from each other.The alternative hypothesis specifies the direction of the difference between the distributions.

In this case, the alternative hypothesis is two-sided, meaning that it allows for differences in either direction (i.e; the distributions can be either greater or less than each other).

2.3 Simple linear regression

Aim is to determine how regressions of mean of 7 species (BD7) matches mean of 11 species (BD11) Here let's start with doing a scatter plot for BD7 vs BD11.

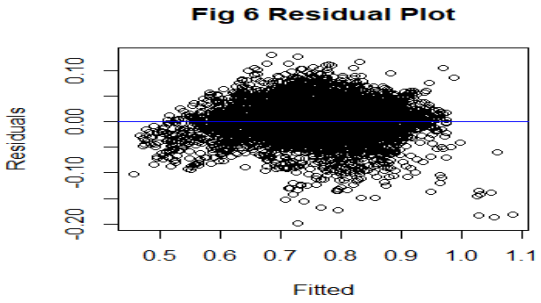
```
Call:
```

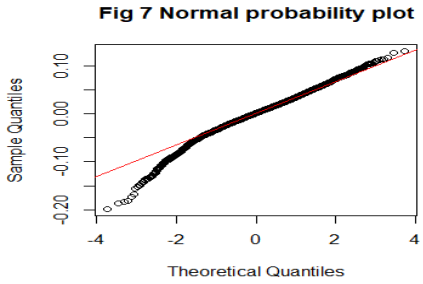



```
lm(formula = Proj_data_MA334$eco_mean_7 ~ Proj_data_MA334$ecologicalStatus)
Coefficients:           (Intercept)        Proj_data_MA334$ecologicalStatus
                           0.161                              0.836
```

Second line of code adds a red diagonal line through the scatter plot.It can be observed that there are few outliers in the beginning, which increases as the value is higher.As shown above in Fig 5, the correlation coefficient of the above model is 0.84, this represents the estimate change in the dependent variable for a one-unit increase in the independent variable.In other words,each unit increase in BD11,the model predicts increase of 0.84 units in BD7.By comparing these two line, lets visually assess how well the linear regression model(green line), approaches the ideal case of a perfect fit(red line).It can be seen that for values ranging from 0 to 0.8 , green line is not much close to red line,indicating predictions have some bias.Where as for values ranging from 0.8 to 1.2 , green line is closer to red line, indicating a good fit.

Now, let's create plot of residuals (difference between the observed and predicted values) vs fitted values(the predicted values).The jitter function is used to add some random noise, helps to visualize overlapping points. Second plot creates normal probability of the residuals.





As shown in Fig 6 Residual plot, blue line represents the zero line.It can be observed that most of the residuals points are centered around zero,indicating that models predictions are unbiased. And also there are few points far away from zero(implies outlier). As shown in Fig 6 Normal probability plot, red line represents expected values of the residuals if they were normally distributed. Till -2 there is a deviation from red line,suggesting departure from normality for those points.But after -2 ,it suggests residuals are distributed normally.

Here aim is to perform same linear model , for each period

```
                    (Intercept)      Proj_data_MA334_Y70$ecologicalStatus
                      0.1466061                                 0.8575741

                    (Intercept)      Proj_data_MA334_Y00$ecologicalStatus
                      0.1737940                                 0.8162085
```

As shown above from the output, the correlation coefficient of the above model for period Y70 is 0.86, this represents the estimate change in the dependent variable for a one-unit increase in the independent variable.In other words,each unit increase in BD11,the model predicts increase of 0.86 units in BD7.And also the correlation coefficient of the above model for period Y00 is 0.82, this represents the estimate change in the dependent variable for a one-unit increase in the independent variable.In other words,each unit increase in BD11,the model predicts increase of 0.82 units in BD7. Overall, 0.86 and 0.82 indicate a strong positive linear relationship between the BD7 and BD11. And also if we compare periods Y70 and Y00, it suggests that relationship between BD11 and BD7 may be stronger during period Y70 compared to Y00, as indicated by higher estimated coefficient and lower intercept for the period Y70 in the above output.

2.4 Multiple linear regression

Lets perform Multiple linear regression of BD4 against BD7 proportional species values.Thus, BD7 with 7 predictors and BD4 as the response variable. Initially, generating training data by randomly sampling 80% of the total number of rows in data set.

Then creating test data by excluding the rows used for training from data set. Function na.omit() is used to remove any rows with missing values from the test data. 1.Firstly, create Training and Test data

2.Build the model and train it using train data Here formula 'eco_mean_4∼.' specified as response variable(BD4).c(eco_selected_names,"eco_mean_4") is used to select BD7 as predictors and BD4 as response variable from training data. "y=TRUE" indicate that response variable is included in the model.Function summary() is used to get information about the coefficients,residuals, and goodness-of-fit measures. cor()is calculating the correlation between the predicted values and the actual response variable(BD4) from the training data.This can be used as a measure of how well the model predicts the response variable.

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.37772 -0.06853 -0.00514  0.06093  0.64873 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -0.065646   0.018475  -3.553 0.000385 ***
Bird                      0.226209   0.020513  11.028  < 2e-16 ***
Bryophytes               -0.048069   0.013344  -3.602 0.000319 ***
Butterflies               0.055148   0.015645   3.525 0.000428 ***
Carabids                  0.193503   0.010115  19.130  < 2e-16 ***
Hoverflies                0.163971   0.012321  13.308  < 2e-16 ***
Macromoths                0.241457   0.016783  14.387  < 2e-16 ***
Grasshoppers_._Crickets   0.095468   0.009524  10.024  < 2e-16 ***

Residual standard error: 0.1073 on 4216 degrees of freedom
Multiple R-squared:  0.5154,    Adjusted R-squared:  0.5146 
F-statistic: 640.5 on 7 and 4216 DF,  p-value: < 2.2e-16

[1] 0.7179089
```

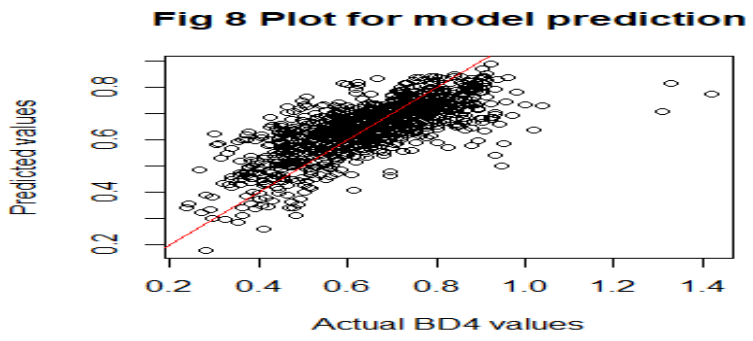
As shown in the above output, Residuals are the difference between the observed values of the response variable and values predicted by the linear regression model.It provides quality of the regression fit.So -0.37 is the smallest residual and 0.65 is largest residual in the data. 0.65 is a large variation in prediction, which may suggest error in model, still it is good to further analyse before coming to conclusion. 25% of residual values are below -0.06 and 75% of residual values are below 0.06. And mid-point of residual is -0.005, Which may suggest that majority(75%) of residuals are close to 0 indicating good fit of the model. And A median close to 0 indicate , models predictions are average, with an equal number of over estimations and underestimations. From positive coefficient estimate for the variable "bird",butterflies,carabids,hoverfies,macromoths and grasshoppers suggest that increase in these predictor variable is associated with an increase in the dependent variable.It may indicate increase in the richness of these species is associated with an increase in the

dependent variable,which could be an indicator of a healthy ecosystem.On the other hand a negative coefficient estimate for the variable bryophytes suggests that an increase in this predictor variable is associated with a decrease in the dependent variable.Overall only bird,carabids,hover flies and macromoths have higher estimate compared to rest.All standard error are small, may suggest more precise estimates.The t-values for most of the variables(Bird,Butterflies,Carabids,Hoverflies,Macromoths,Grasshoppers_crickets) are high,suggesting that these variables have a significant relationship with dependent variables.Asterisks (***) and values less than 2e-16 are used to highlight that the Pr(>|t|) values in the example suggest that each coefficient estimate is likely to be statistically significant at a very low significance level (much smaller than 0.01). This shows that the dependent variable in the regression analysis has a statistically significant association with the relevant predictor variables (Bird, Bryophytes, Butterflies, Carabids, Hoverflies, Macromoths, Grasshoppers, and Crickets).

In the provided output, the RSE is 0.1076, indicating the average residual error in the model is approximately 0.1076 units. A smaller RSE value indicates a better fit of the model to the data, as it implies that the residuals are smaller and closer to zero, which means that the model is better able to explain the variability in the observed data.In the provided output, the multiple R-squared is 0.5127, which means that approximately 51.27% of the variance in the dependent variable can be explained by the regression model with the included independent variables. This indicates that the model accounts for about 51.27% of the total variability in the data, leaving approximately 48.73% unexplained.In the provided output, the adjusted R-squared is 0.5119, which is slightly lower than the multiple R-squared (0.5127). This suggests that the inclusion of the predictors in the model may not be providing a substantial improvement in the explained variance compared to a model with fewer predictors.The F-statistic is 633.8, which indicates that the model has a significant overall fit to the data.The p-value associated with the F-statistic is < 2.2e-16,suggesting the model is statistically significant at a very high level of confidence.In conclusion, the small p-value and the large F-statistic suggest that the regression model is statistically significant and has a good overall fit to the data.

3.Check prediction of model using test data as input. Lets predict the values using a linear regression model(lmMod_train) that has been trained on a separate set of data, then evaluate the predicted values against the actual values of the dependent variable in a test dataset.

```
[1] 0.7266806
```



The trained linear regression model (lmMod_train) and the test dataset (testData) are inputs into the predict () function, which produces predicted values for the dependent variable (Eco_4_Pred). The regression coefficients inferred from the training dataset form the foundation for these anticipated values.The correlation coefficient between the expected values (Eco_4_Pred) and the actual values of the dependent variable (testData eco_{mean_4})iscomputedusingthecor()function. Thedependentvariable'sactualvalues(testData eco_mean_4) are plotted against the anticipated values (Eco_4_Pred) using the plot() function.A red diagonal line is added to the scatter plot using the abline() function.A correlation coefficient of 0.7085936 indicates a moderately strong positive linear relationship between the predicted values (Eco_4_Pred) and the actual values of the dependent variable (testData$eco_mean_4).The magnitude of the correlation coefficient (0.7085936) suggests that the predicted values and actual values are fairly closely related, but not perfectly correlated. As shown above in Fig 7, it can be seen that there is a moderately strong positive linear relationship between the predicted values and the actual values of the dependent variable as suggested by cor() function output.Lets do the further analysis by going through residuals and normality plots.

4.Checking Residual and normality of Model.

Fig 9 Plot for Residuals

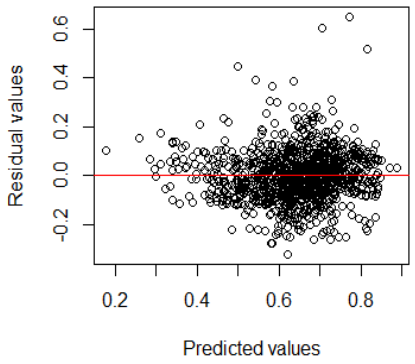
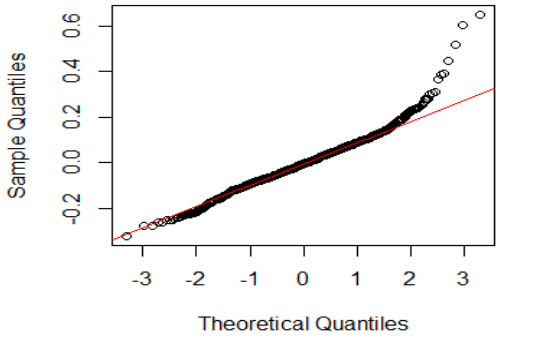


Fig 10 Plot Normality for Residuals



Most of the points of plot Residual values vs Predicted values are close to zero, this indicates that the model is making accurate predictions. Also it can be seen most of the points of residuals fit into Normality,it suggests that the model is performing well and that the errors or deviations from the predicted values are random and evenly distributed around zero.

5.Finally performing feature selection based on p values from the regression.

As it can be seen from all values are significant, still checking with the help of p-values. Removing the variables or predictors whose p-value is less than 0.05. As every variable in BD7 as predictor is significant. None of the variable was removed by the program. Hence there is no requirement of AIC to justify removal.

2.4 Open Analysis

Here data set is examined to get better understanding of its characteristics,structure and patterns.To achieve this , various tasks are performed such as summary, statistics and so on. In below task, data is filtered based on period. Later studying mean,standard deviation and skewness of seven species.

Table 1 Summary Statistics from 1970-1990

Species_group	mean	sd	skewness
Bird	0.87	0.1	-1.9
Butterflies	0.82	0.12	-0.71
Bryophytes	0.78	0.13	-0.31
Macromoths	0.8	0.14	-1.27
Carabids	0.7	0.15	-0.19
Hoverflies	0.72	0.17	-0.63
Grasshoppers_._Crickets	0.66	0.18	-0.13

From Table 1, It can be seen that mean of Bird (0.87), Butterflies (0.82) and Macromoths (0.8) are higher than rest species, suggesting a higher level of biodiversity. While Grasshoppers_crickets(0.66) have the lower level of biodiversity.Grasshoppers_crickets(0.18) has higher sd value compared to all, indicating the variation of biodiversity in different locations.While Bird(0.1) with lower value of sd indicates it has same level of biodiversity in all locations as compared to other species.In terms of skewness,the bird species(-1.9) has most negative skewed distribution,this may indicate biodiversity values of this species group are skewed towards lower value or species richness is concentrated towards higher values. And Grasshoppers_crickets has skewness value of -0.13(almost 0),this may indicate that the biodiversity measure for the species is symmetric or evenly distributed.

Table 2 Summary Statistics from 2000-2013

Species_group	mean	sd	skewness
Bird	0.9	0.11	-1.35
Butterflies	0.93	0.13	-0.37
Macromoths	0.9	0.13	-1.36
Bryophytes	0.79	0.14	-0.12
Hoverflies	0.64	0.17	-0.4
Carabids	0.51	0.22	-0.09
Grasshoppers_._Crickets	0.6	0.23	0.07

From Table 2, It can be seen that mean of Bird (0.9), Butterflies (0.93) and Macromoths (0.9) are higher than rest species, suggesting a higher level of biodiversity. While Carabids(0.51) have the lower level of biodiversity.Grasshoppers_crickets(0.23) has higher sd value compared to all, indicating the variation of biodiversity in different locations.While Bird(0.11) with lower value of sd indicates it has same level of biodiversity in all locations as compared to other species.In terms of skewness,the Macromoths species(-1.36) has most negative skewed distribution,this may indicate biodiversity values of this species group are skewed towards lower value or species richness is concentrated towards higher values. And Grasshoppers_crickets has skewness value of 0.07(almost 0),this may indicate that the biodiversity measure for the species is symmetric or evenly distributed.

Result of open analysis:

Some intriguing insights can be gained from the comparison of Tables 1 and 2. Between 1970 and 2000, the mean biodiversity of a number of species, including birds, butterflies, bryophytes, and Macromoths, increased. This suggests that throughout time, these species may have benefited from conservation initiatives and other environmental measures.

However, at the same time span, species including Carabides, Hoverflies, and Grasshoppers_crickets had a reduction in mean biodiversity. This means that the chances of survival and population increase for these species may be worse.

It's also interesting to observe that whereas Carabides had the lowest mean biodiversity in period "Y00," Grasshoppers crickets had the lowest mean in period "Y70." This can be a sign that the environment is changing or that other things are influencing these species differently through time.

Additionally, every species' skewness value in period "Y00" is less inclined to go in a negative direction than it was in period "Y70," with the exception of Macromoths, whose negative skew value is more extreme in period "Y00" than it was in period "Y70." This implies that while the distribution of biodiversity values for the majority of species has gotten more symmetrical over time, it has become more negatively skewed for Macromoths, pointing to a probable loss in the species' overall biodiversity.