MA334_ 2212231

Opeyemi Latona

2023-04-22

INTRODUCTION

The biodiversity of an ecosystem is a fundamental aspect of its health and resilience. In this report, we explore the biodiversity of seven taxonomic groups, including birds, bees, butterflies, carabids, diptera, homoptera, and hymenoptera. We use descriptive statistics to compare the levels of biodiversity among these groups and examine the relationship between the diversity of these groups and their geographic coordinates. Additionally, we perform hypothesis tests to determine whether there have been significant changes in biodiversity over time and investigate the relationship between biodiversity and ecological status using linear regression.

MEHTOD

The methodology used in this report involves exploratory data analysis, hypothesis testing, and linear regression analysis to investigate the relationship between biodiversity and ecological status. Firstly, descriptive statistics were generated to explore the biodiversity levels of seven taxonomic groups, as well as the skewness and variability of the data within each group. Correlation coefficients were calculated to determine the relationship between Easting and Northing coordinates and the biodiversity of the seven taxa. Next, hypothesis testing was performed to examine the change in biodiversity over time and the relationship between two biodiversity measures. A one-way t-test was used to determine if there was a significant change in the ecological status of BD7 between the Y00 and Y70 periods. The Kolmogorov-Smirnov test was used to test for significant differences in the distribution of BD7 and BD11. Finally, linear regression models were used to investigate the relationship between biodiversity and ecological status. Simple linear regression was used to determine the relationship between BD7 and BD11. Multiple linear regression was used to investigate the relationship between ecological status and BD7 for the Y70 period and for the Y00 period and of the variables and the proportional species richness values.

RESULT

```
##   Location       Bees      Bird Bryophytes Butterflies  Carabids Hoverflies
## 1     HP50 0.07526882 0.5968200  0.6445703   0.6444444 0.5496829  0.2871126
## 2     HP60 0.08602150 0.6044203  0.6479014   0.4611111 0.5644820  0.2903752
## 3     HP61 0.07526882 0.5931485  0.6429047   0.6222222 0.5264271  0.3017945
##     Isopods Ladybirds Macromoths Grasshoppers_._Crickets Vascular_plants
## 1 0.3975904 0.3846154  0.3154528                    0.25       0.5758335
## 2 0.3975904 0.3846154  0.3494094                    0.25       0.5605415
## 3 0.3975904 0.3846154  0.3533465                    0.25       0.5580346
##   Easting Northing dominantLandClass ecologicalStatus period
## 1  450000  1200000               32s        0.4292174    Y70
## 2  460000  1200000               32s        0.4178607    Y70
## 3  460000  1210000               32s        0.4277593    Y70
```

## LOAD DATA:

The list of Biodiveristy alloted to me

```
## [1] "Bird"            "Bees"            "Butterflies"     "Carabids"
## [5] "Hoverflies"      "Isopods"         "Vascular_plants"
```

1 Data Exploration 1a Based on the descriptive statistics presented in the table, it can concluded that the bird taxonomic group has the highest level of biodiversity among the seven groups, while the bee taxonomic group has the lowest level. The Butterflies group has the most skewed distribution of data, while the Carabids group has the least skewed distribution. The variability of data within each group varies, with the bees group having the largest standard deviation and the butterflies group having the smallest.

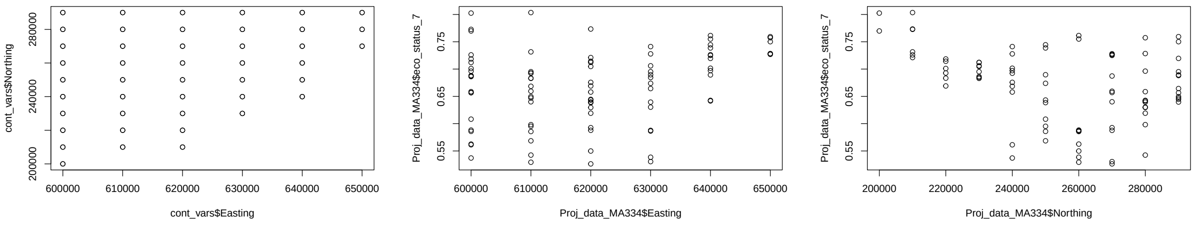
```
##        taxi_group mean   sd skewness
## 1     Butterflies 0.69 0.04     1.22
## 2      Hoverflies 0.58 0.06     0.12
## 3 Vascular_plants 0.76 0.08      0.3
## 4            Bird  0.9 0.08     0.32
## 5         Isopods 0.63 0.16    -0.25
## 6        Carabids 0.64 0.16    -1.08
## 7            Bees 0.48 0.21     -0.1
```

1b

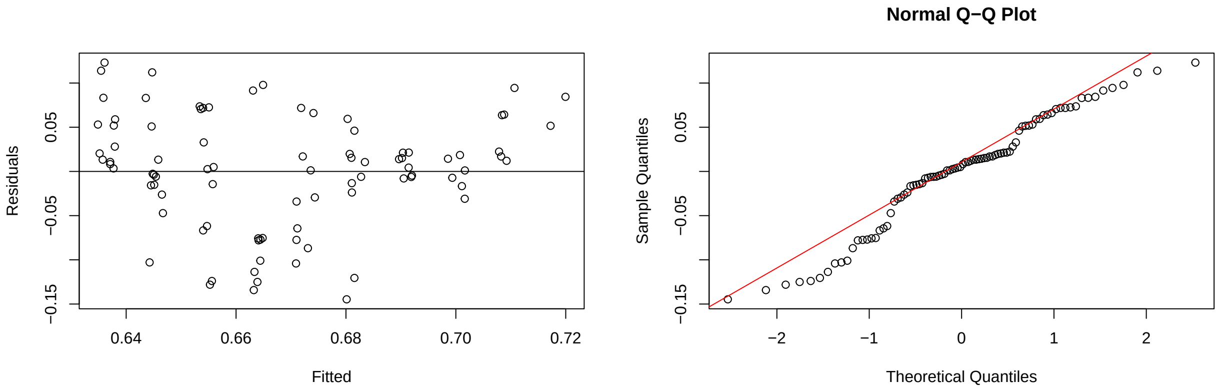
Easting and Northing The correlation coefficient between Easting and the seven taxonomies was calculated to be 0.2635128. This value indicates a weak positive correlation between Easting and the seven taxonomies; though it is positive, it means that the variables move in the same direction. In this case, as Easting increases, the seven taxonomies also tend to increase, although the relationship is weak.The correlation coefficient value of -0.3452376 suggests a moderately negative correlation between the Northing coordinate and the diversity of the seven taxa. This means that as the Northing coordinate increases, the diversity of these taxa decreases. The magnitude of the correlation coefficient indicates that this relationship is not very strong but still statistically significant. This is shown in the figure below

```
## [1] 0.2635128
```

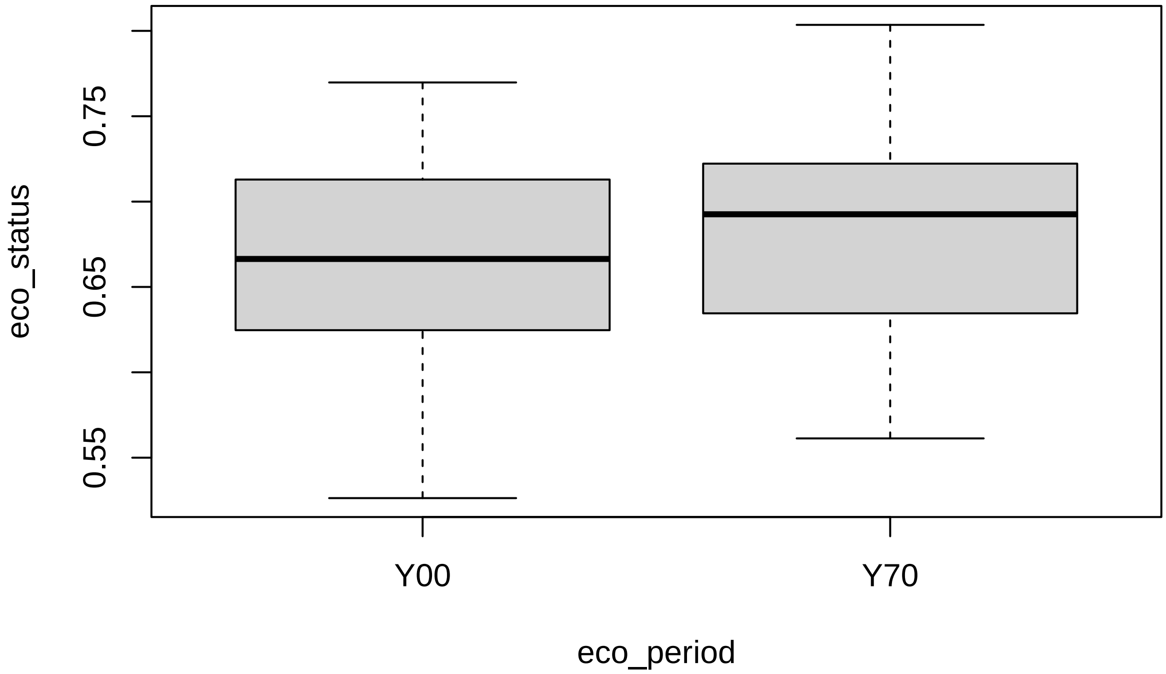
```
## [1] -0.3452376
```



```
##
## Call:
## lm(formula = eco_status_7 ~ Northing, data = Proj_data_MA334)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.144665 -0.029783  0.006621  0.051047  0.123122
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.002e-01  6.827e-02  13.187  < 2e-16 ***
## Northing    -9.102e-07  2.668e-07  -3.411 0.000987 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06436 on 86 degrees of freedom
## Multiple R-squared:  0.1192, Adjusted R-squared:  0.1089
## F-statistic: 11.64 on 1 and 86 DF,  p-value: 0.000987
```



The box plot comparism shows that there is a significant change in the eco status of BD7 from the period of Y00 to the period of Y70. This can be seen from the median value of each Y00 and Y70.

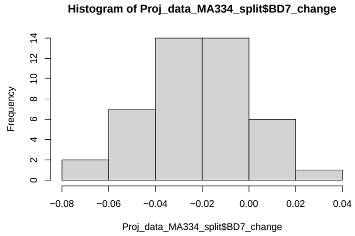


Hypothesis tests

2a) One-way Test The one-way test done to check the change in the biodiversity of BD7 between Y00 and Y70 shows that there is a significant change in the eco-status of BD7.The t-test result shows that the t-value is -6.1351, with a p-value of 2.326e-07, which is less than the standard alpha level of 0.05, indicating that we can reject the null hypothesis that the true mean is equal to 0. This suggests that there is a statistically significant difference between the mean values of BD7_change and 0. The 95% confidence interval is between -0.02845299 and -0.01437486, which means we can be 95% confident that the true population mean lies within this interval. This indicates that the change in ecological status from Y00 to Y70 is likely to be negative, with a mean change of approximately -0.021.

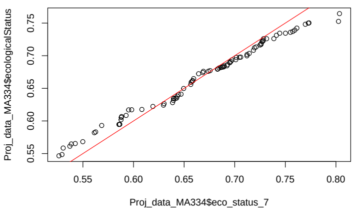
```
## # A tibble: 6 x 4
##   Location   Y70   Y00 BD7_change
##   <chr>    <dbl> <dbl>      <dbl>
## 1 TM00     0.803 0.770    -0.0328
## 2 TM01     0.773 0.726    -0.0467
## 3 TM02     0.718 0.701    -0.0174
## 4 TM03     0.712 0.686    -0.0259
## 5 TM04     0.561 0.537    -0.0242
## 6 TM05     0.608 0.586    -0.0224
```

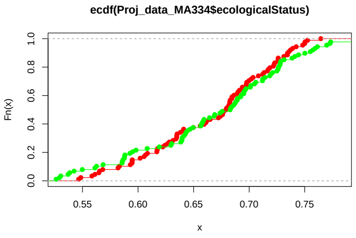
```
##
##  One Sample t-test
##
## data:  BD7_change
## t = -6.1351, df = 43, p-value = 2.326e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.02845299 -0.01437486
## sample estimates:
##   mean of x
## -0.02141393
```



2b) The Kolmogorov-Smirnov test (KS test) The KS test performed on the taxonomic group BD7 in relation to BD11 shows that there is no significant change in the biodiversity of BD7 and BD11. The test resulted in a test statistic of 0.090909 and a p-value of 0.8631. Since the p-value is greater than the significance level of 0.05, it fails to reject the null hypothesis, and we conclude that there is no significant difference between the distribution of BD7 and BD11. Therefore, we can say that there is no significant relationship between the variables BD7 and BD11 in terms of their ecological status.

```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  Proj_data_MA334$eco_status_7 and Proj_data_MA334$ecologicalStatus
## D = 0.090909, p-value = 0.8631
## alternative hypothesis: two-sided
```



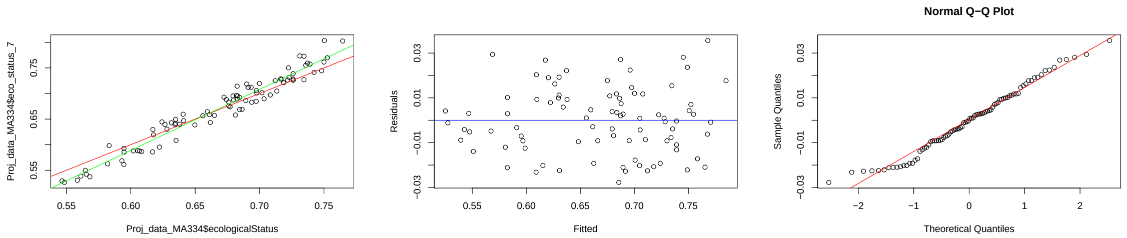


3 Simple linear regression

3a) Linear Regression of BD7 against BD11

The linear regression shows the relationship between BD7 and BD11. From the summary of the lm, the estimated regression equation is presented as:BD7 = -0.1261 + 1.1916*BD11.since the dependent variable is BD7 and the independent variable is BD11.The coefficients of the regression equation are statistically significant, as evidenced by the t-values and corresponding p-values. The intercept is -0.1261 with a standard error of 0.0190 and a t-value of -6.638. The p-value of 2.72e-09 indicates that the intercept is significantly different from zero. The slope coefficient for BD11 is 1.1916, with a standard error of 0.0284 and a t-value of 41.965. The p-value of < 2e-16 indicates that the slope coefficient is also significantly different from zero. In conclusion, the linear regression model suggests that there is a strong positive relationship between BD11 and BD7. For every unit increase in BD11, BD7 is expected to increase by 1.1916 units while holding all other factors constant.

```
##
## Call:
## lm(formula = Proj_data_MA334$eco_status_7 ~ Proj_data_MA334$ecologicalStatus)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.027733 -0.009290  0.000245  0.009980  0.035593
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -0.1261     0.0190  -6.638 2.72e-09 ***
## Proj_data_MA334$ecologicalStatus   1.1916     0.0284  41.965  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.0148 on 86 degrees of freedom
## Multiple R-squared:  0.9534, Adjusted R-squared:  0.9529
## F-statistic:  1761 on 1 and 86 DF,  p-value: < 2.2e-16
```



3b. Linear regression model of BD7 over the period Y70 The linear regression model shows that there is a strong positive relationship between biodiversity BD7 and ecological status for the period Y70. The coefficient estimate for the ecological status predictor variable is 1.26073, indicating that for every one-unit increase in ecological status, there is an estimated increase of 1.26073 in BD7. The intercept estimate is -0.16562, which represents the expected value of BD7 when the ecological status is zero. The p-value for the F-statistic is less than 2.2e-16, indicating that the model is statistically significant and that the predictor variable has a significant impact on the response variable. The multiple R-squared value of 0.9478 suggests that the model explains a large proportion of the variance in BD7, and the adjusted R-squared value of 0.9466 suggests that the model is a good fit for the data.

```
##                      (Intercept) Proj_data_MA334$ecologicalStatus
##                       -0.1261343                        1.1916285
```

3c : Linear regression of BD7 over the period Y00 The lm model for biodiversity BD7 over the period Y00 shows a strong positive relationship between BD7 and ecological status. The coefficient of the ecological status variable is estimated at 1.13757, which implies that a one-unit increase in ecological status is associated with a 1.13757 increase in BD7. The p-value for the coefficient is very small (< 2e-16), indicating that the coefficient is statistically significant. The intercept of the model is estimated at -0.09708, which represents the predicted value of BD7 when the ecological status is zero. However, since ecological status is a continuous variable, the intercept may not be meaningful in practice. The multiple R-squared of the model is 0.9786, which indicates that the model explains a significant portion of the variation in BD7.

```
##                          (Intercept) Proj_data_MA334_Y00$ecologicalStatus
##                          -0.09708146                           1.13757308
```

4 Multiple linear regression

4a

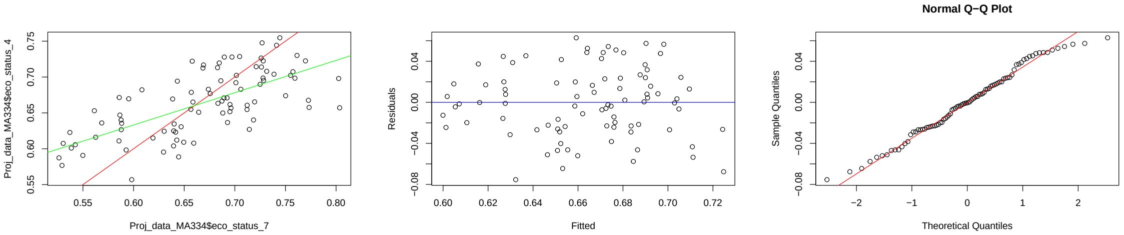
The Calculate mean of the proportional species richness values for these remaining four taxonomic groups :

Mean of BD4 = 0.6639664

```
## [1] 0.6639664
```

```
## [1] 0
```

```
##
## Call:
## lm(formula = Proj_data_MA334$eco_status_4 ~ Proj_data_MA334$eco_status_7)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.075378 -0.023781 -0.000417  0.023018  0.062864
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.36292    0.03523  10.301  < 2e-16 ***
## Proj_data_MA334$eco_status_7  0.45032    0.05243   8.589 3.38e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03334 on 86 degrees of freedom
## Multiple R-squared:  0.4617, Adjusted R-squared:  0.4554
## F-statistic: 73.76 on 1 and 86 DF,  p-value: 3.385e-13
```



### 4: multiple linear regression 4b. Multiple linear regression BD4 against the selected 7

The final model is a linear regression model with BD4 (eco_status_4) as the response variable and Birds, Bees, Butterflies, Isopods, Hoverflies, Carabids, and vascular plants as the predictor variables. The coefficients of the regression model show the relationship between each predictor variable and the response variable.

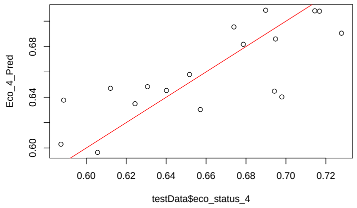
BD4 = 0.29952 + 0.29585 * Bird + 0.02096 * Bees + 0.02458 * Butterflies + 0.17659 * Carabids - 0.14142 * Hoverflies - 0.08732 * Isopods + 0.12214 * Vascular_plants

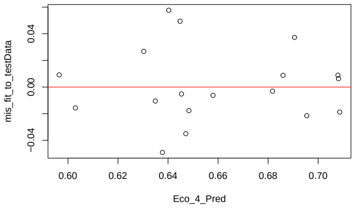
The intercept of the model is 0.29952, which means that if all the predictor variables are equal to zero, the BD4 would be 0.29952. The coefficient for birds is 0.29585, which means that for every one-unit increase in the number of birds, the BD4 will increase by 0.29585 units, holding all other predictor variables constant. The coefficient for bees is 0.02096, which means that for every one-unit increase in the number of bees, the BD4 will increase by 0.02096 units, holding all other predictor variables constant. However, the p-value for bees is not significant (p=0.220930), indicating that this variable may not be a significant predictor of BD4. The coefficient for butterflies is 0.02458, which means that for every one-unit increase in the number of butterflies, the BD4 will increase by 0.02458 units, holding all other predictor variables constant. However, the p-value for butterflies is not significant (p=0.782405), indicating that this variable may not be a significant predictor of BD4. The coefficient for Carabids is 0.17659, which means that for every one-unit increase in the number of Carabids, the BD4 will increase by 0.17659 units, holding all other predictor variables constant. The coefficient for hoverflies is -0.14142, which means that for every one-unit increase in the number of hoverflies, the BD4 will decrease by 0.14142 units, holding all other predictor variables constant. The coefficient for isopods is -0.08732, which means that for every one-unit increase in the number of isopods, the BD4 will decrease by 0.08732 units, holding all other predictor variables constant. The coefficient for vascular plants is 0.12214, which means that for every one-unit increase in the number of vascular plants, the BD4 will increase by 0.12214 units, holding all other predictor variables constant. The adjusted R-squared value of the model is 0.812, which means that approximately 81.2% of the variation in BD4 can be explained by the predictor variables in the model. The F-statistic is significant ($p < 2.2e-16$), indicating that the overall model is a good fit for the data.

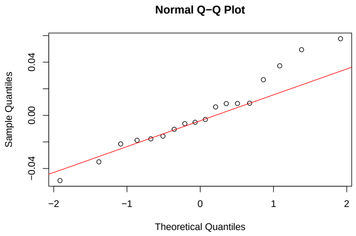
```
##
## Call:
## lm(formula = eco_status_4 ~ ., data = trainingData[c("eco_status_4",
##     eco_selected_names)], na.action = na.omit, y = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.074432 -0.010777 -0.000858  0.010836  0.046778
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.30572    0.08556   3.573 0.000689 ***
## Bird             0.36423    0.07113   5.121 3.19e-06 ***
## Bees            -0.02842    0.02224  -1.278 0.206066
## Butterflies      0.02501    0.09547   0.262 0.794230
## Carabids         0.16288    0.04353   3.742 0.000401 ***
## Hoverflies      -0.11524    0.06785  -1.698 0.094454 .
## Isopods         -0.08617    0.04101  -2.101 0.039683 *
## Vascular_plants  0.05820    0.06359   0.915 0.363591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02157 on 62 degrees of freedom
## Multiple R-squared:  0.7983, Adjusted R-squared:  0.7755
## F-statistic: 35.06 on 7 and 62 DF,  p-value: < 2.2e-16
```

```
## [1] 0.8934848
```

```
## [1] 0.7817804
```



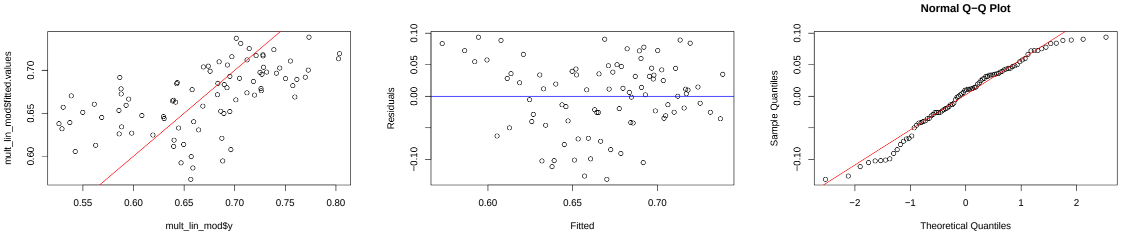




Multiple linear regression BD7 against period, easting and northing

This linear model relates the variable eco_status_7 to the independent variables period, Easting, and Northing . The estimated regression equation is: eco_status_7 = -0.1941 + 0.02141 * periodY70 + 1.917e-06 * Easting - 1.322e-06*Northing This shows that: Holding all other variables constant, for each unit increase in periodY70, there is an expected increase of 0.02141 in eco_status_7. However, the coefficient is only marginally significant with a p-value of 0.0841. Also,holding all other variables constant, for each unit increase in Easting, there is an expected increase of 1.917e-06 in eco_status_7. The coefficient is statistically significant with a p-value of 1.69e-05. Holding all other variables constant, for each unit increase in Northing, there is an expected decrease of 1.322e-06 in eco_status_7. The coefficient is statistically significant with a p-value of 1.44e-06.

```
##
## Call:
## lm(formula = eco_status_7 ~ ., data = Proj_data_MA334[c("eco_status_7",
##     "period", "Easting", "Northing")], na.action = na.omit, y = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.131684 -0.035066  0.009987  0.040131  0.093668
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.941e-01  2.452e-01  -0.792   0.4308
## periodY70    2.141e-02  1.225e-02   1.748   0.0841 .
## Easting      1.917e-06  4.200e-07   4.565 1.69e-05 ***
## Northing    -1.322e-06  2.547e-07  -5.190 1.44e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05746 on 84 degrees of freedom
## Multiple R-squared:  0.3142, Adjusted R-squared:  0.2898
## F-statistic: 12.83 on 3 and 84 DF,  p-value: 5.575e-07
```



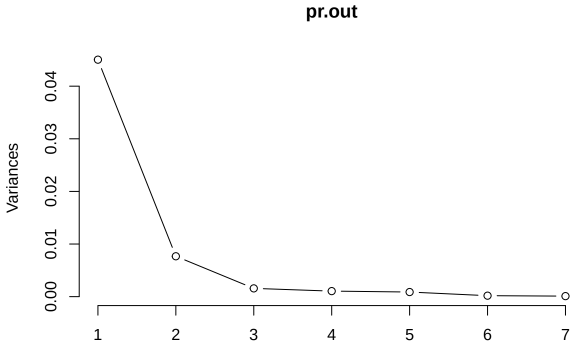
```
##   (Intercept)     periodY70       Easting      Northing
## -1.940807e-01  2.141393e-02  1.917436e-06 -1.321888e-06
```

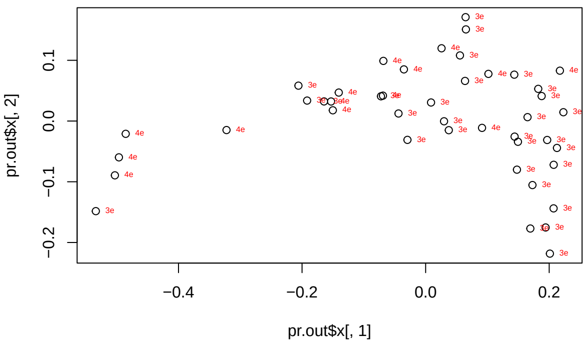
```
## [1] 1.188374
```

```
## [1] -0.3364806
```

```
##
## Y00 Y70
##   0  44
```

```
##
## Y00 Y70
##  44   0

##
## TRUE
##   44

##         Bird       Bees Butterflies    Carabids  Hoverflies    Isopods
## 1 0.06627090 -0.1273047  0.16509434  0.09643917 -0.07311129 -0.2521008
## 2 0.07925592 -0.1720808  0.11084906  0.04080119 -0.09179529 -0.1932773
## 3 0.06961710  0.1409336  0.08293839 -0.01618579 -0.06544503 -0.2431193
## 4 0.05722952  0.1678636  0.07582938 -0.02111189 -0.08289703 -0.2706422
## 5 0.02707269  0.2468582  0.08530806 -0.13652358 -0.02443281 -0.3073395
## 6 0.03516661  0.4272890  0.04265403 -0.29275158 -0.04973822 -0.2568807
##   Vascular_plants
## 1     -0.10467098
## 2     -0.10092056
## 3     -0.09068581
## 4     -0.10731154
## 5     -0.06026828
## 6     -0.06253542

##            Bird            Bees     Butterflies        Carabids      Hoverflies
##      0.06203632      0.28074449      0.04167141     -0.14496624     -0.05468574
##         Isopods Vascular_plants
##     -0.26334971     -0.07134802

## [1] FALSE

##                         PC1         PC2
## Bird            -0.03057430  0.01840780
## Bees             0.83951095  0.51170737
## Butterflies     -0.15441472  0.04514445
## Carabids        -0.49377008  0.73431807
## Hoverflies       0.10492494 -0.09633480
## Isopods         -0.11952363  0.42940417
## Vascular_plants  0.03658746  0.05373407
```

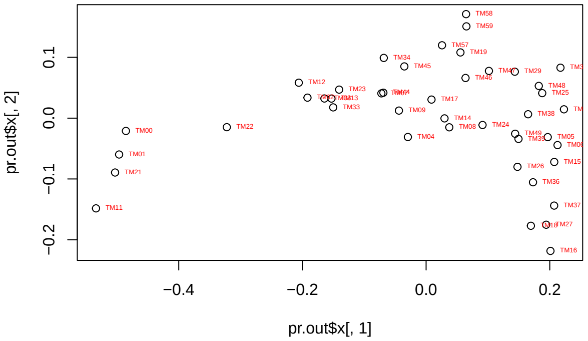


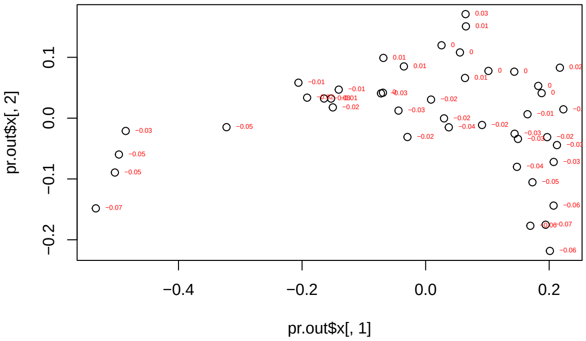


Open Analysis

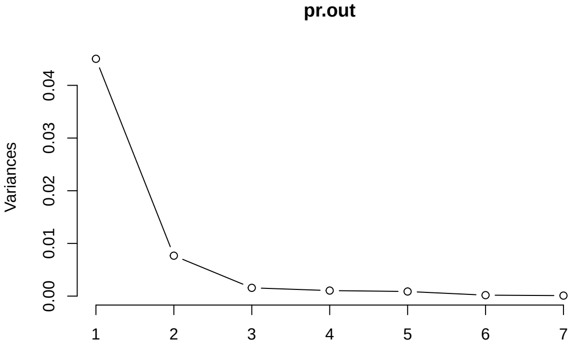
The results of a principal component analysis (PCA) of the differences in biodiversity variables between the periods Y00 and Y70. The values under each biodiversity variable represent the loading or contribution of that variable to the first two principal components, PC1 and PC2. The values under PC1 and PC2 are the coefficients that describe the linear combination of the original variables that make up each principal component. From the output, we can infer the following: Bees have a high positive loading (0.84) on PC1, while Carabids have a high negative loading (-0.49) on PC1. This indicates that Bees have a strong positive correlation with PC1, while Carabids have a strong negative correlation with PC1. Similarly, Isopods have a high positive loading (0.43) on PC2, while Hoverflies have a high negative loading (-0.10) on PC2. This indicates that Isopods have a strong positive correlation with PC2, while Hoverflies have a strong negative correlation with PC2.

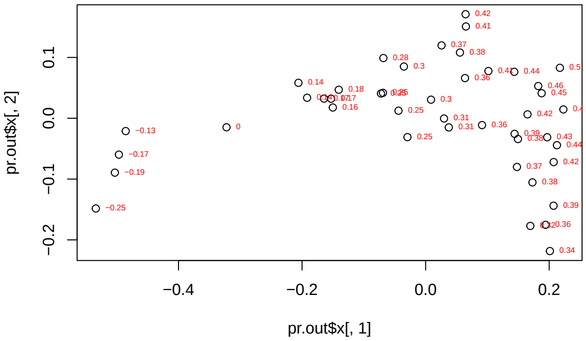
Overall, the analysis helps us understand how the variables are related to each other and to the principal components.





```
##                       PC1         PC2
## Bird          -0.03057430  0.01840780
## Bees           0.83951095  0.51170737
## Butterflies   -0.15441472  0.04514445
## Carabids      -0.49377008  0.73431807
## Hoverflies     0.10492494 -0.09633480
## Isopods       -0.11952363  0.42940417
## Vascular_plants  0.03658746  0.05373407
```





Conclusion

In conclusion, this report presents findings from a study on biodiversity and ecological status using statistical methods. The data exploration section showed that bird taxonomy has the highest biodiversity, while bees have the lowest. The hypothesis tests showed a significant change in ecological status for BD7 between Y00 and Y70, but no significant difference between the distribution of BD7 and BD11. The linear regression models suggested a strong positive relationship between BD7 and BD11, as well as between BD7 and ecological status for both Y00 and Y70. The models were statistically significant, indicating a strong impact of predictor variables on the response variable. These findings have important implications for understanding the relationship between biodiversity and ecological status and can inform conservation efforts to protect vulnerable species and ecosystems.