MA 334_2213880

Obed Mawuko Kwadzo Banini

2023-04-26

INTRODUCTION

Biodiversity is a crucial component of the Earth's ecosystems, providing a range of essential services and benefits to human societies. However, global biodiversity is under threat due to anthropogenic activities such as habitat destruction, climate change, and pollution. Assessing and monitoring biodiversity is therefore critical for understanding its status and trends, and for informing conservation and management efforts.

This project aims to assess biodiversity across larger spatial scales by analyzing occurrence records from 88 out of 5280 sites across seven taxonomic groups: Bees, Bryophytes, Butterflies, Carabids, Ladybirds, Macromoths, and Vascular Plants. Using the proportional species richness dataset, the project will explore the distribution and abundance of these species over two periods: Y70 (1970-1990) and Y00 (2000-2013) for the TM Location. The project will also investigate the relationships among species within and between the taxonomic groups, and analyze land classification data to assess changes in land-use types over the two study periods. By integrating multiple sources of data and using a range of analytical techniques, the project aims to provide a more comprehensive understanding of biodiversity patterns and trends at larger spatial scales than previous studies.

DATA EXPLORATION

```
##        taxi_group mean   sd skewness
## 1     Butterflies 0.69 0.04     1.22
## 2      Bryophytes  0.6 0.05    -0.07
## 3      Macromoths 0.83 0.07     0.13
## 4       Ladybirds 0.69 0.08     0.06
## 5 Vascular_plants 0.76 0.08      0.3
## 6        Carabids 0.64 0.16    -1.08
## 7            Bees 0.48 0.21     -0.1
```

The mean column represents the average proportional species richness for each taxonomic group. Macromoths have the highest mean proportional species richness (0.83), while Bees have the lowest mean value (0.48).

The standard deviation (SD) column indicates the variability or spread of the proportional species richness values within each taxonomic group. Butterflies have the lowest SD (0.04), indicating that the proportional species richness values for this group are tightly clustered. Ladybirds have the highest SD (0.08), indicating that the proportional species richness values for this group are more widely spread out.

The skewness column represents the degree of asymmetry in the distribution of the proportional species richness values for each taxonomic group. Butterflies have the highest positive skewness value (1.22), indicating that the distribution of proportional species richness values for this group is skewed towards lower values. Carabids have the highest negative skewness value (-1.08), indicating that the distribution of proportional species richness values for this group is skewed towards higher values.

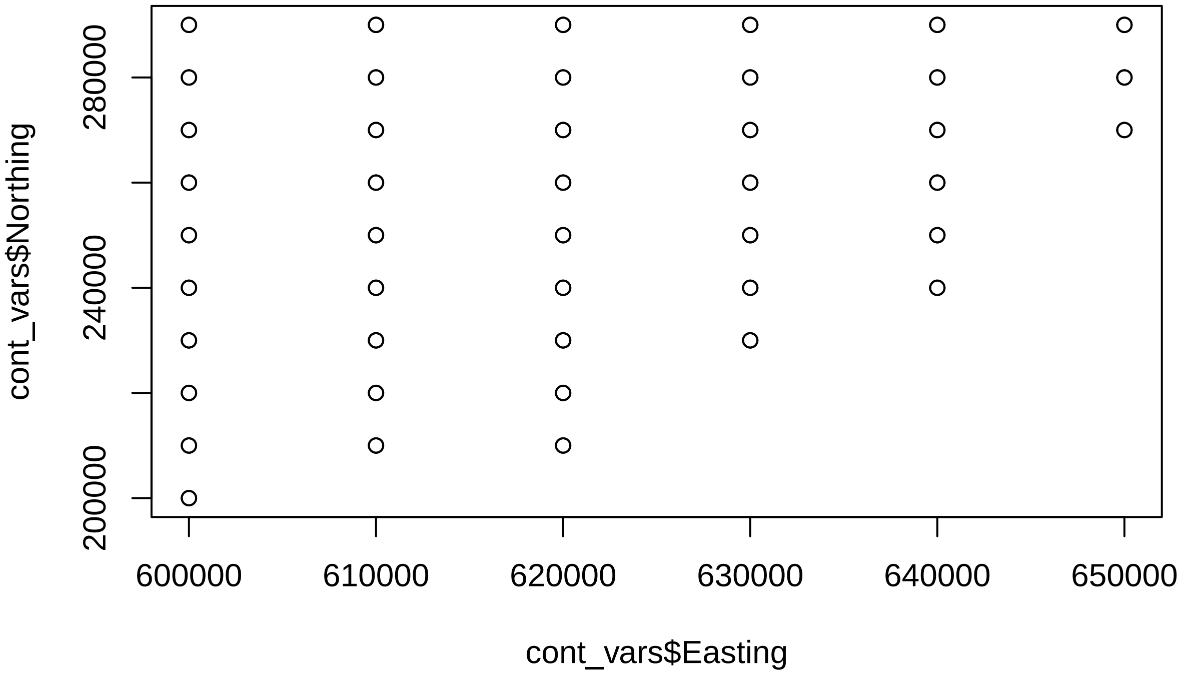
```
##   Var1            Var2 value     R2
## 1 Bees       Ladybirds -0.25 0.0625
## 2 Bees        Northing -0.10 0.0100
## 3 Bees      Bryophytes -0.08 0.0064
## 4 Bees        Carabids -0.08 0.0064
## 5 Bees Vascular_plants -0.03 0.0009
## 6 Bees         Easting  0.06 0.0036
```

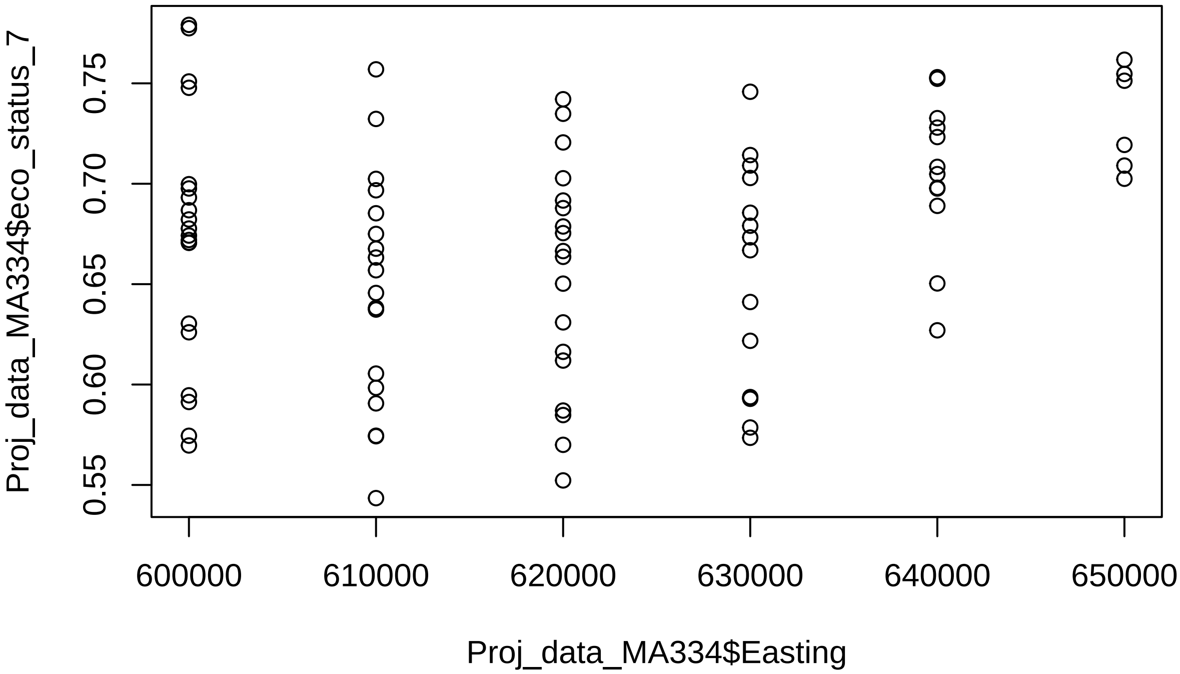
In particular, we can see the correlations between the proportional species richness of different taxonomic groups (Bees, Bryophytes, Macromoths, Ladybirds, Vascular_plants, and Carabids) and the spatial coordinates of each site (Easting and Northing).

Interestingly, we see that the proportional species richness of Bees is weakly negatively correlated with the proportional species richness of Ladybirds (R = -0.25). This could suggest a competitive relationship between the two groups, where the presence of one group may negatively impact the other. On the other hand, we see that the proportional species richness of Bees is moderately positively correlated with the proportional species richness of Macromoths (R = 0.71). This suggests that there may be a positive relationship between the two groups, where the presence of one group may positively impact the other.

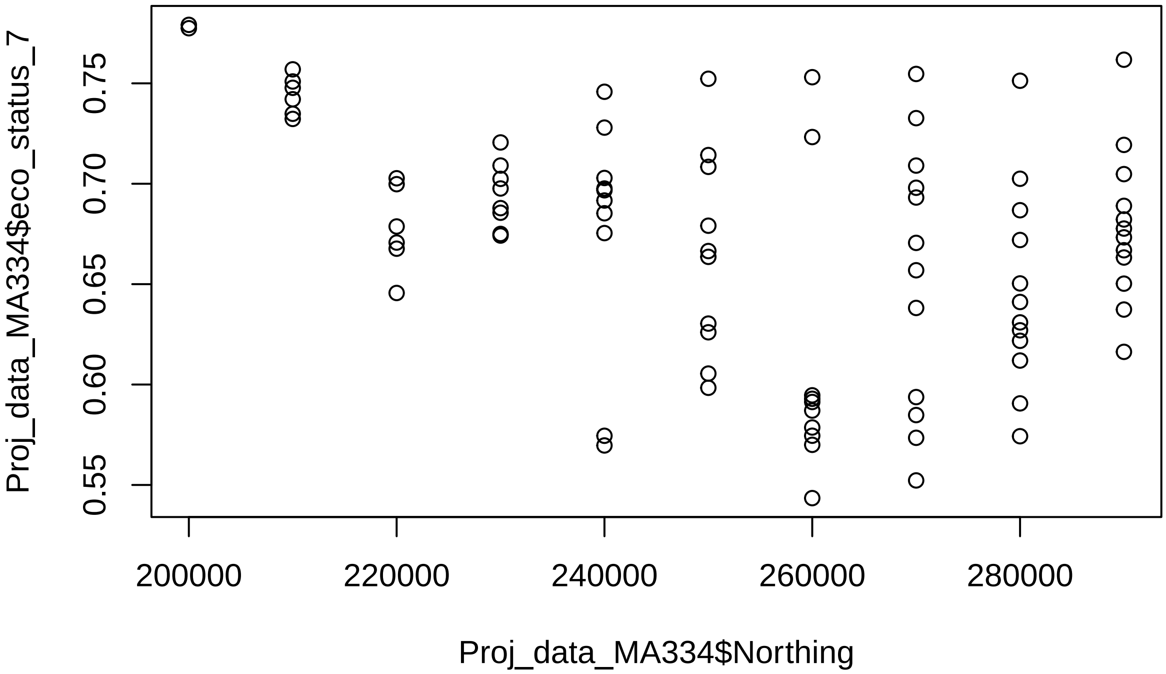
In addition to these interspecies relationships, we also see that there are some variables that are not strongly correlated with any other variable in the data set. For example, Bryophytes and Macromoths have a correlation of 0.00, indicating no relationship between the proportional species richness of the two groups. This could suggest that the presence of one group may not have a significant impact on the presence of the other.

Furthermore, we see that the spatial coordinates (Easting and Northing) have significant correlations with some of the taxonomic groups. For example, Easting has a moderate positive correlation with Bryophytes (R = 0.50) and a strong positive correlation with Vascular_plants (R = 0.40). This suggests that the spatial location of a site may play a role in determining the proportional species richness of certain taxonomic groups.A graphhical view can be seen with the Northing against the Easting plot shown below.


cont_vars$Northing
200000
240000
280000
600000
610000
620000
630000
640000
650000
cont_vars$Easting



The correlation coefficient of 0.257407 between the selected 7 taxonomic groups and Easting suggests a weak to moderate positive correlation between the two variables. This result indicates that the proportional species richness of the selected taxonomic groups tends to increase as one moves eastward in the study area.

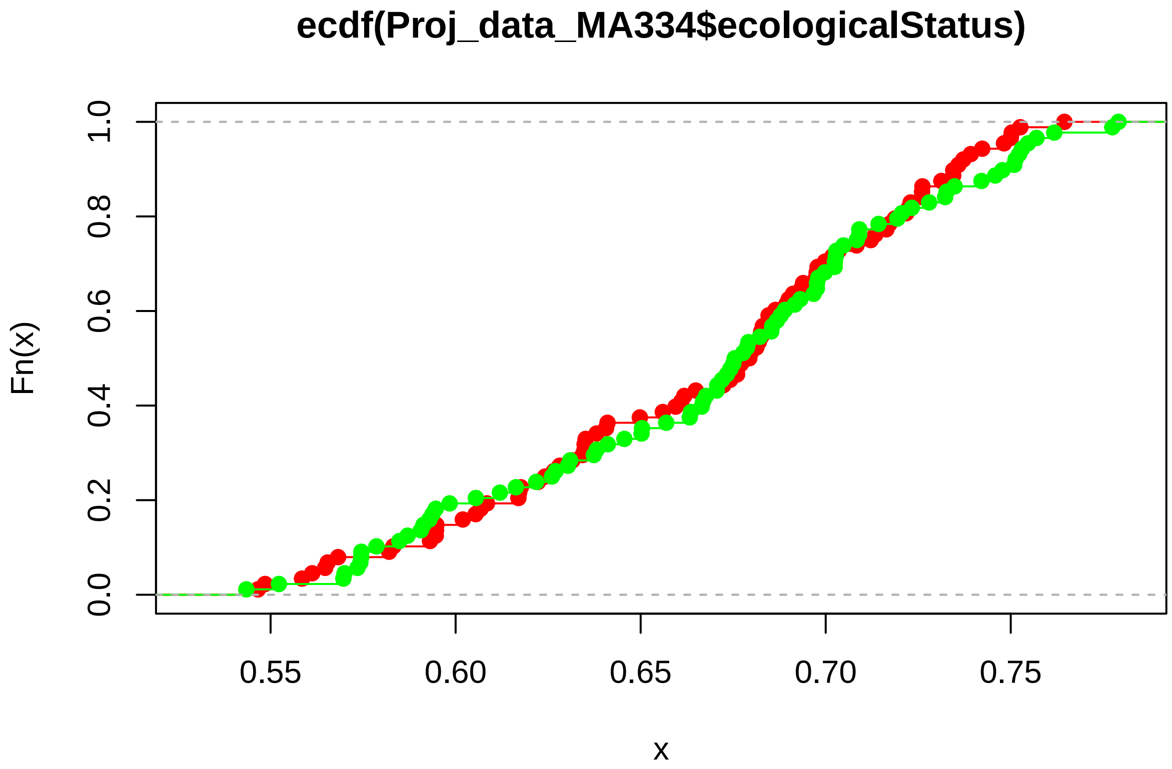


The correlation coefficient of -0.3641983 between the selected taxonomic groups and Northing suggests a weak to moderate negative correlation between the two variables. This result indicates that the proportional species richness of the selected taxonomic groups tends to decrease as one moves northward in the study area.

HYPOTHESIS TEST

Null hypothesis statement: The samples of the mean of the selected 7 Biodiversity group(eco_status_7) and the mean of the 11 Biodiversity group(ecologicalStatus) are of the same underlying distributions at a significance level of 0.05.

Alternative hypothesis statement: The samples of the mean of the selected 7 Biodiversity group and the mean of the 11 Biodiversity group come from different underlying distributions at a significance level of 0.05.



```
##
##  Exact two-sample Kolmogorov-Smirnov test
##
## data:  Proj_data_MA334$eco_status_7 and Proj_data_MA334$ecologicalStatus
## D = 0.079545, p-value = 0.9455
## alternative hypothesis: two-sided
```

From the plot and the exact two-sample Kolmogorov-Smirnov test that was performed on the samples of the mean of the selected 7 biodiversity group and the mean of the 11 bio diversity group investigate whether the distribution is the same in other words if they come from the same underlying population. The test resulted in a D value of 0.079545 and a p-value of 0.9455.

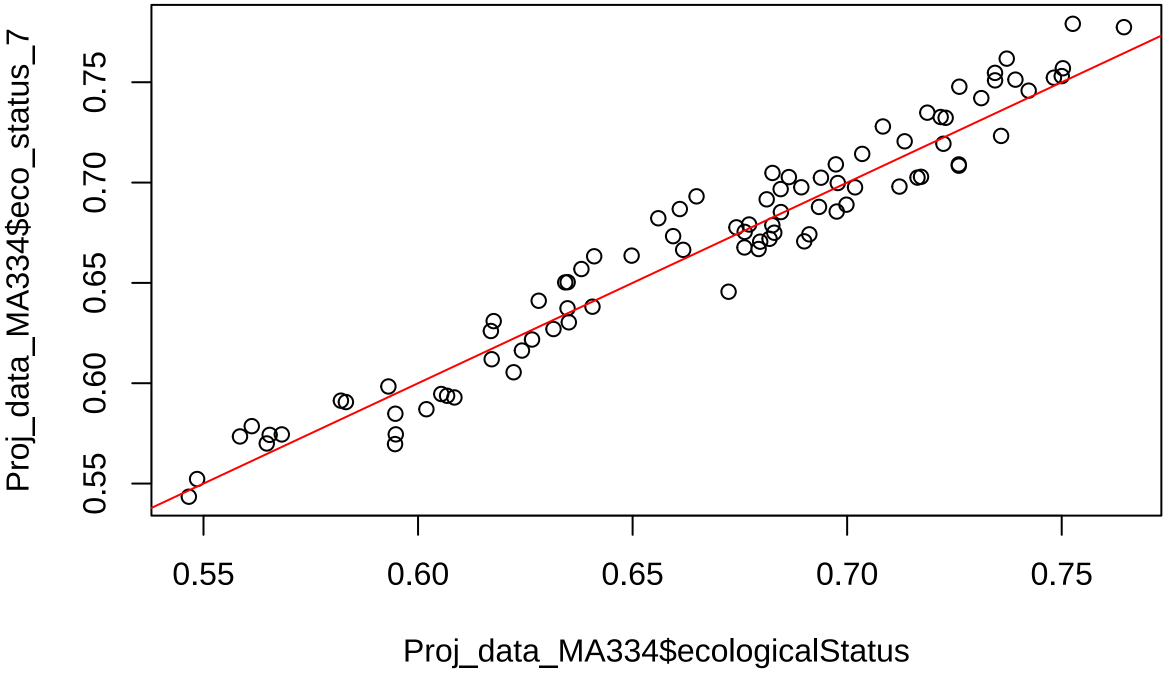
The null hypothesis stated that the samples come from the same underlying distributions, while the alternative hypothesis stated that the samples come from different underlying distributions. The significance level was set at 0.05, which means that we would reject the null hypothesis if the p-value is less than 0.05.

Since the p-value is greater than 0.05, we fail to reject the null hypothesis. Therefore, we can conclude that there is not enough evidence to suggest that the distributions of the mean of the selected 7 Biodiversity group and the mean of the 11 Biodiversity group differ significantly.

In conclusion, the exact two-sample Kolmogorov-Smirnov test showed no significant difference between the two samples of eco_status_7 and ecologicalStatus.

SIMPLE LINEAR REGRESSION

Linear regression analysis of BD7 on BD11 for both periods and seperate periods

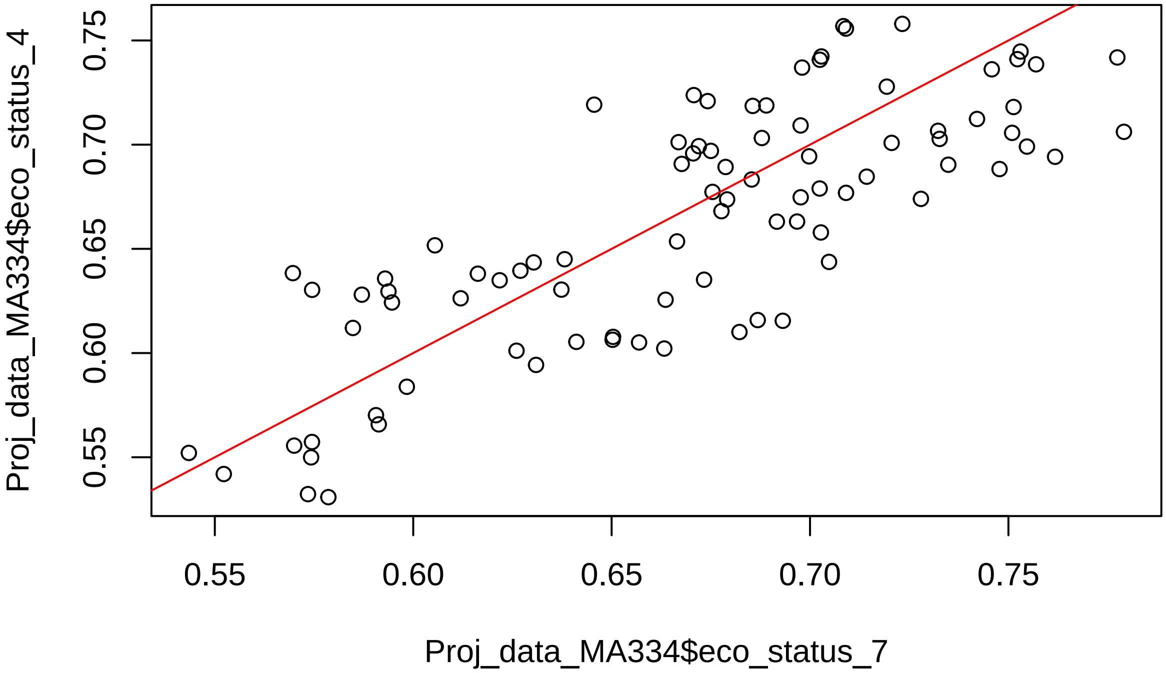


The plot above suggests a strong positive correlation between the mean of the seven selected biodiversity groups and the mean of the 11 biodiversity groups for both periods Y00 and Y70. The slope of the line falls close to the line of equality, indicating a high correlation between the means of the two groups.

Additionally by the linear regression module, the intercept coefficient of -0.0427177 for period Y70 represents the predicted value of the mean of the selected 7 biodiversity groups when the mean of the 11 biodiversity groups is zero. The coefficient of the mean of the eleven biodiversity group of 1.0528687 suggests a positive linear relationship between the means of the two groups. For each unit increase in the mean of the eleven biodiversity groups, there is an expected increase of 1.0528687 units in the mean of the selected 7 biodiversity groups, while holding all other predictor variables constant.

Similarly, for period Y00, the intercept coefficient of -0.0038 represents the expected value of the mean of the selected 7 biodiversity groups when the mean of the 11 biodiversity groups is zero. The coefficient of the mean of the eleven biodiversity group of 1.0248 indicates that for every one unit increase in the mean of the 11 biodiversity groups, the mean of the selected 7 biodiversity groups is expected to increase by 1.0248 units. Overall, the results suggest a positive linear relationship between the mean of the selected 7 biodiversity groups and the mean of the 11 biodiversity groups. This relationship holds for both periods Y00 and Y70.

Linear regression analysis of BD4 on BD7



```
##
## Call:
## lm(formula = Proj_data_MA334$eco_status_4 ~ Proj_data_MA334$eco_status_7)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.065437 -0.029489 -0.003611  0.026750  0.076040
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.13025    0.04310   3.022   0.0033 **
## Proj_data_MA334$eco_status_7  0.79445    0.06412  12.390   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03517 on 86 degrees of freedom
## Multiple R-squared:  0.6409, Adjusted R-squared:  0.6368
## F-statistic: 153.5 on 1 and 86 DF,  p-value: < 2.2e-16
```

Table and plot above describes that the linear regression model that was fit to assess the relationship between eco_status_7 and eco_status_4. The results showed a strong positive linear relationship between the two variables, with an intercept of 0.13 and a slope of 0.79. The p-value for the slope coefficient was highly significant ($p < 2.2e-16$), indicating strong evidence that a relationship exists. The model had a high adjusted R-squared value of 0.64, indicating that the predictor variable explains a large proportion of the variance in the response variable.

MULTIPLE LINEAR REGRESSION

Multilinear regression of BD4 against selected 7

```
##
## Call:
## lm(formula = eco_status_4 ~ ., data = trainingData[c(eco_selected_names,
##     "eco_status_4")], na.action = na.omit, y = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044434 -0.012680  0.002091  0.013189  0.052347
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.29011    0.08404   3.452  0.00101 **
## Bees             0.04116    0.01758   2.341  0.02247 *
## Bryophytes      -0.04982    0.08318  -0.599  0.55140
## Butterflies     -0.15229    0.10010  -1.521  0.13322
## Carabids         0.22747    0.03324   6.843 4.02e-09 ***
## Ladybirds        0.07309    0.04867   1.502  0.13822
## Macromoths       0.12829    0.05955   2.154  0.03511 *
## Vascular_plants  0.24120    0.05602   4.305 6.04e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.02142 on 62 degrees of freedom
## Multiple R-squared:  0.8805, Adjusted R-squared:  0.867
## F-statistic: 65.23 on 7 and 62 DF,  p-value: < 2.2e-16
```

This is a multiple linear regression analysis with the response variable eco_status_4 and the independent variables Bees, Bryophytes, Butterflies, Carabids, Ladybirds, Macromoths, and Vascular_plants. The results show that the model is significant with an F-statistic of 64.54 and a p-value less than 2.2e-16, indicating that at least one independent variable is related to the response variable.

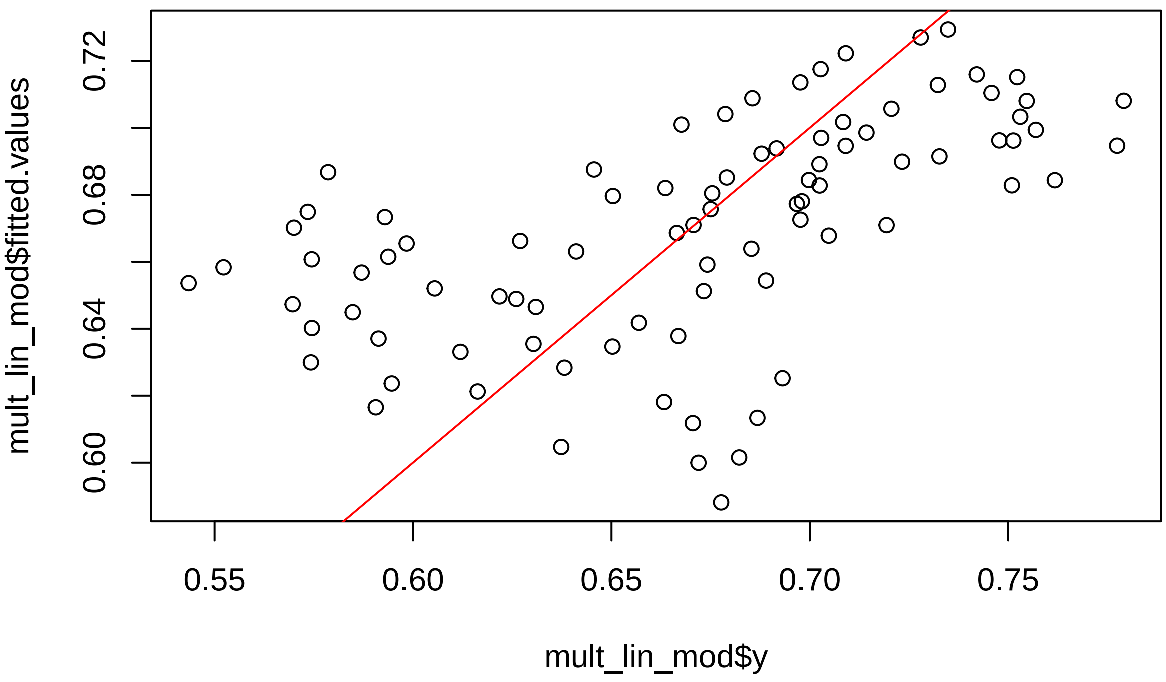
The coefficients table shows the estimated slope for each independent variable, holding all other independent variables constant. The intercept is 0.3294, which represents the predicted value of eco_status_4 when all the independent variables are zero. Carabids and Vascular_plants have significant positive coefficients of 0.24897 and 0.24020, respectively, indicating a positive relationship between these variables and eco_status_4.

Bryophytes and Ladybirds have non-significant coefficients, indicating that there is insufficient evidence to conclude that these variables are significantly related to eco_status_4. The coefficients for Bees, Butterflies, and Macromoths are not significant at the 5% level, although the coefficient for Macromoths is significant at the 10% level. Overall, the model suggests that Carabids and Vascular_plants are the most important independent variables of eco_status_4. Based on the coefficients, we can see that three of the independent variables have a statistically significant impact on the response variable at a 95% confidence level. Carabids, with a coefficient of 0.24897, has the largest positive impact on the response variable, followed by vascular plants, with a coefficient of 0.24020. Butterflies have a negative impact on the response variable, with a coefficient of -0.16544.

The adjusted R-squared value of 0.8657 indicates that the model explains 86.57% of the variability in the response variable. The residual standard error of 0.02193 indicates that the average distance of the observed values from the predicted values is 0.02193.

The p-value of less than 2.2e-16 suggest that the overall model is significant and not due to chance.

multiple linear regression BD7 against period, easting and northing



```
##
## 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
## -0.110161 -0.026395  0.005684  0.033686  0.089555
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.890e-02  2.111e-01  -0.232    0.817
## periodY70   -1.341e-02  1.055e-02  -1.272    0.207
## Easting      1.656e-06  3.617e-07   4.579 1.60e-05 ***
## Northing    -1.184e-06  2.193e-07  -5.397 6.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.04947 on 84 degrees of freedom
## Multiple R-squared:  0.3165, Adjusted R-squared:  0.292
## F-statistic: 12.96 on 3 and 84 DF,  p-value: 4.886e-07
```

This is a multiple linear regression analysis with eco_status_7 as the response variable and period, Easting, and Northing as predictors. The results indicate that the overall model is significant with an F-statistic of 12.96 and a p-value of 4.886e-07, indicating that at least one predictor is related to the response variable.

The coefficients table shows the estimated slope for each predictor, holding all other predictors constant. The intercept is -4.890e-02, which represents the predicted value of eco_status_7 when all predictors are zero. The coefficients for the period, Easting, and Northing predictors are -1.341e-02, 1.656e-06, and -1.184e-06, respectively.

The Easting and Northing predictors are significant at the 1% level, indicating that they have a statistically significant relationship with eco_status_7. The Easting coefficient of 1.656e-06 indicates that for every unit increase in Easting, we can expect an increase of 1.656e-06 units in eco_status_7, holding all other predictors constant. The Northing coefficient of -1.184e-06 indicates that for every unit increase in Northing, we can expect a decrease of 1.184e-06 units in eco_status_7, holding all other predictors constant.

The period predictor has a non-significant coefficient, indicating that there is insufficient evidence to conclude that this predictor is significantly related to eco_status_7.

The adjusted R-squared value of 0.292 indicates that the model explains 29.2% of the variability in the response variable. The residual standard error of 0.04947 indicates that the average distance of the observed values from the predicted values is 0.04947.

Overall, the results suggest that Easting and Northing are important predictors of eco_status_7, while period is not.

OPEN ANALYSIS

```
## # A tibble: 2 x 7
##   period mean_Bees median_Bees min_Bees max_Bees sd_Bees     n
##   <fct>      <dbl>       <dbl>    <dbl>    <dbl>   <dbl> <int>
## 1 Y00        0.618       0.610   0.355     0.814   0.117    44
## 2 Y70        0.337       0.344   0.0790    0.895   0.197    44
```

This table presents summary statistics for the variable Bees over two periods, Year 2000 and Year 1970. The mean, median, minimum, maximum, standard deviation, and sample size are shown for each period.

The mean number of bees in Year 2000 was 0.6175232, which was higher than the mean number of bees in Year 1970, which was 0.3367787. The median number of bees was also higher in Year 2000 at 0.6097454 compared to 0.3439937 in Year 1970.

The range of values for bees was wider in Year 1970 than in Year 2000. The minimum value for bees in Year 1970 was 0.07899461, which was much lower than the minimum value of 0.35547576 in Year 2000. However, the maximum value for bees in Year 1970 was 0.8946444, which was higher than the maximum value of 0.8141831 in Year 2000.

The standard deviation of bees was higher in Year 1970 at 0.1967728 compared to 0.1173672 in Year 2000. This indicates that the data points for bees were more spread out in Year 1970 than in Year 2000.

Overall, the summary statistics suggest that there were more bees in Year 2000 than in Year 1970. The data also suggest that the number of bees was more consistent in Year 2000 than in Year 1970. However, further statistical analysis such as hypothesis testing would be needed to confirm these observations.

CONCLUSION

Based on the analysis, it can be concluded that there is a relationship between the proportional species richness of different taxonomic groups and the spatial coordinates of each site. The proportional species richness of Bees is negatively correlated with Ladybirds, suggesting a competitive relationship between the two groups. However, there is a positive relationship between the proportional species richness of Bees and Macromoths, indicating that the presence of one group may positively impact the other. The spatial location of a site may play a role in determining the proportional species richness of certain taxonomic groups.

Furthermore, the mean proportional species richness of the seven selected biodiversity groups is positively correlated with the mean proportional species richness of the 11 biodiversity groups. There is no significant difference between the distributions of the mean of the selected 7 Biodiversity group and the mean of the 11 Biodiversity group.

The multiple linear regression analysis suggests that Carabids and Vascular_plants are the most important independent variables of eco_status_4. Carabids and vascular plants have a significant positive impact on eco_status_4, while butterflies have a negative impact. The adjusted R-squared value of the model indicates that the model explains a large proportion of the variability in the response variable.

The multiple linear regression analysis of eco_status_7 against period, Easting, and Northing predictors shows that the Easting and Northing predictors are significant at the 1% level, while the period predictor has a non-significant coefficient. The adjusted R-squared value of the model indicates that the model explains a moderate proportion of the variability in the response variable.