MA334_Individual Assignment
Reg_No:2211606

Introduction

This report has examined seven of the eleven taxonomic groups: Bees, Bird, Bryophytes, Butterflies, Carabids, Macromoths and Vascular_plants. It has explored them as univariate variables, their intra-relationships and inter-relationships with associate variables like Northing, Easting and Period through a battery of statistical analysis: correlation, hypothesis testing, simple/ multiple linear regression, feature selection, AIC and principal component analysis.

1) Data Exploration

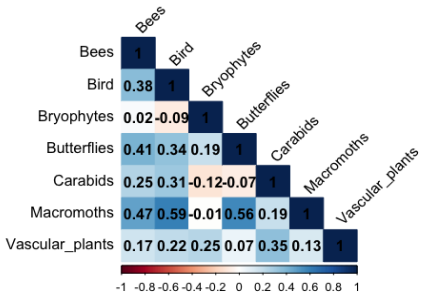
For the select seven variables, their descriptive statistics is presented in Table1. Clearly, the category of Bird has the highest measure of central tendency: mean or median, while the categories of Bees and Carabids have the lowest mean. Notably, Bees display the largest variation in terms of range (i.e., max - min) or standard deviation. The distribution of Vascular_plants is approximately symmetric while those of Bird and Macromoths are highly asymmetric with fatter tails on the left. The distribution of Bird is highly peaked as having the highest kurtosis whereas that of Bryophytes has the lowest kurtosis.

Table 1: Descriptive Statistics

	taxi_group	mean	median	min	max	sd	skewness	kurtosis
1	Bees	0.61	0.59	0.03	3.31	0.31	0.96	6.74
2	Carabids	0.61	0.64	0.01	1.2	0.21	-0.49	2.75
3	Vascular_plants	0.79	0.79	0.42	1.2	0.1	-0.13	3.1
4	Bryophytes	0.79	0.8	0.39	1.17	0.13	-0.2	2.51
5	Macromoths	0.85	0.88	0.09	1.26	0.14	-1.14	5.01
6	Butterflies	0.87	0.89	0.32	1.39	0.14	-0.36	3.52
7	Bird	0.89	0.9	0.24	1.17	0.11	-1.51	7.05

The coefficient of correlation among the seven select variables is presented in Figure 1. Bird and Macromoths have the highest correlation coefficient followed by the pair of Macromoths and Butterflies. All others have coefficient of correlation less than 0.5, that is, weak relationship. The highest negative relationship is observed between Carabids and Bryophytes albeit at a lower magnitude of 0.12. All other negative relationships were even weaker. It may be noted that low absolute value of correlation implies low degree of linear association, low correlation does not imply absence of non-linear association, and that correlation does not imply causality.

Figure 1: Correlation coefficient



The following Figure 2 shows the relationship between Ecological Status and the mean of the seven select variables. As expected, the mean increases with the increase in ecological status, which indicates that the ecosystem has a high abundance of different species, a high diversity of species, and is relatively free of pollution. In the case of Bryophytes, however, they remained largely invariant to the improvements in Ecological Status.

Figure 2: Relation between eco_status_7 and Ecological Status

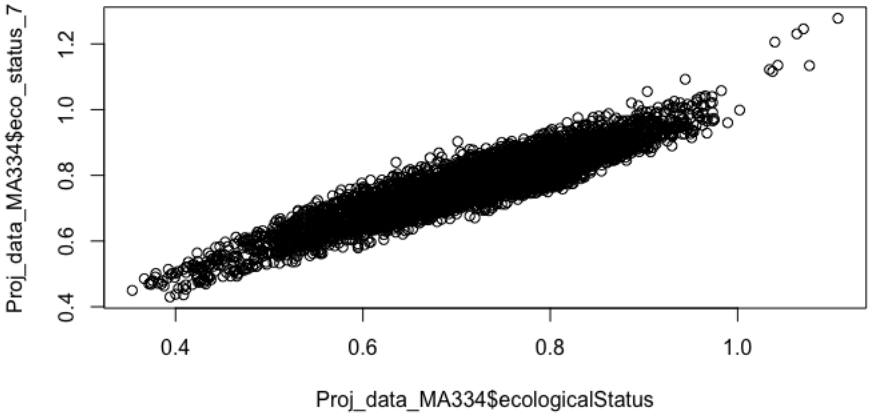
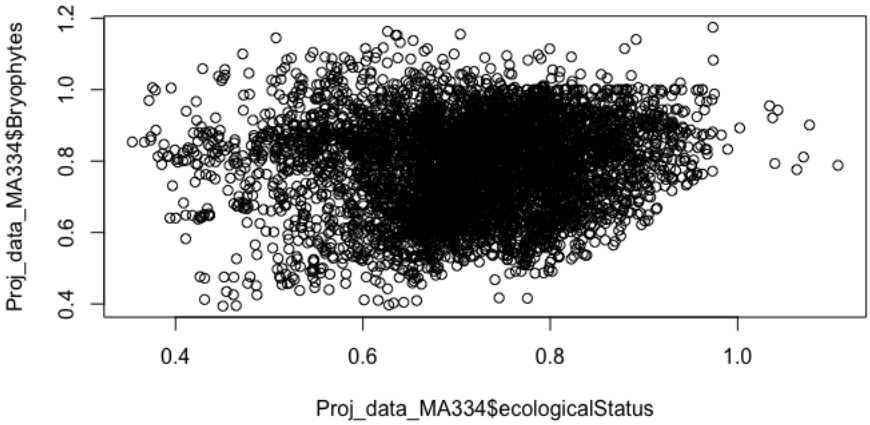


Figure 3: Relation between Bryophytes and Ecological Status



2) Hypothesis Testing

Towards hypothesis testing, we have first bifurcated the data set into two time periods: Y70 and Y00, considering all locations, a sample of which is presented in Figure 4.

Figure 4: Bifurcation according to period

```
> head(Proj_data_MA334_split)
# A tibble: 6 × 4
  Location   Y70   Y00 BD7_change
  <chr>    <dbl> <dbl>      <dbl>
1 HP50     0.486 0.501     0.0150
2 HP60     0.468 0.517     0.0493
3 HP61     0.482 0.538     0.0567
4 HU24     0.500 0.553     0.0532
5 HU25     0.490 0.529     0.0396
6 HU28     0.503 0.539     0.0352
```

Our null hypothesis (H0) is that the mean difference of BD7 is zero across the two time periods. The alternative hypothesis is that mean difference of BD7 is not zero between Y70 and Y00. The t test-statistic turns out to be 22.797 (Figure 5). Because the p-value of our test (2.2e-16) is less than 0.05, we reject the null hypothesis. Therefore, we conclude that the

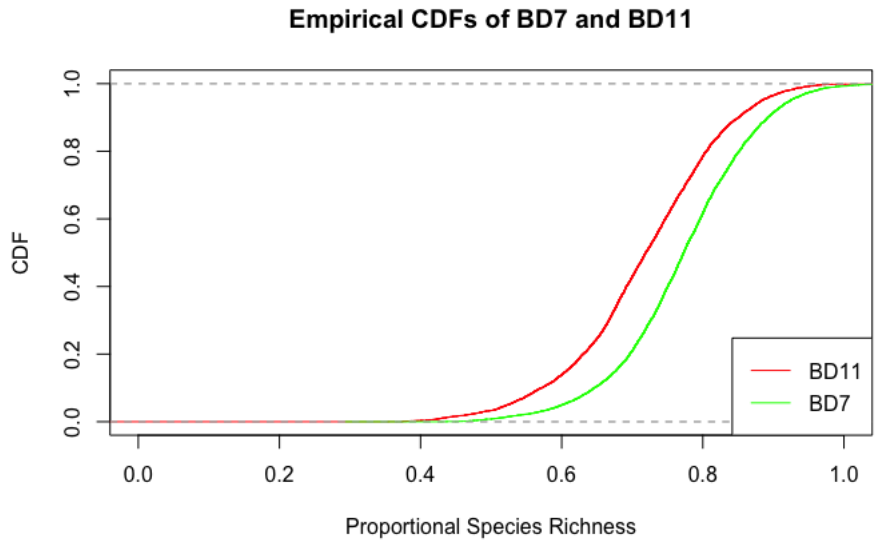
mean values of BD7 between Y70 and Y00 are not equal.

Figure 5: T-test

```
> t.test(BD7_change,mu=0)  # t test with H0: mu=0

        One Sample t-test

data:  BD7_change
t = 22.797, df = 2639, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.02555798 0.03036838
sample estimates:
 mean of x
0.02796318
```

In Figure 6, we have presented the cumulative density functions (cdf) of BD7 and BD11, which are respectively the means of the select seven variables and of the eleven variables, also represented by eco_status_7 and ecologicalStatus. Clearly cdf of BD11 lies above that of BD7 all through. So, in first glance, it appears that the distributions of BD7 and BD11 are different.

Figure 6: cdf of BD7 and BD11



The foregoing statement is examined by conducting a non-parametric Kolmogorov-Smirnov test, which does not require the data set to be normally distributed unlike the t test conducted above. Here the null hypothesis is that the distributions of BD7 and BD11 are the same. The alternative hypothesis is that they are different. Here, the relevant test statistic is D having a value of 0.22727. The p-value being less than 2.2e-16 is far less than 0.05. Therefore, we conclude that BD7 and BD11 have different distributions.

Figure 7: Kolmogorov Test

```
> ks.test(Proj_data_MA334$eco_status_7,Proj_data_MA334$ecologicalStatus)

        Asymptotic two-sample Kolmogorov-Smirnov test

data:  Proj_data_MA334$eco_status_7 and Proj_data_MA334$ecologicalStatus
D = 0.22727, p-value < 2.2e-16
alternative hypothesis: two-sided
```

3) Simple Linear Regression

From the scatter of BD7 and BD11 (Figure 8), it's clear that they share a positive relationship, i.e., they move in the same direction. However, their causal relationship is not clear. Assuming BD7 as the dependent variable and BD11 as the independent variable, we have run a linear regression with a constant (Green line in Figure 8), the results of which are in Figure 9. The regression line has a high R square of 0.8574. The adjusted R square value also remains almost unchanged. As the t-values are very high, both the intercept and BD11 turn out statistically significant. The coefficient of BD11 indicates that one unit change in BD11 leads to a change of 0.839572 in BD7 in the same direction.

Figure 8: Scatter and Regression of BD7 on BD11

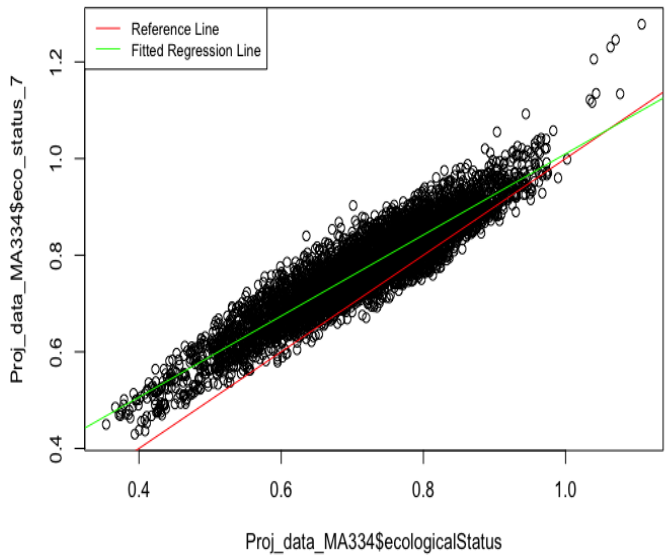


Figure 9: Regression Results of BD7 on BD11

```
lm(formula = Proj_data_MA334$eco_status_7 ~ Proj_data_MA334$ecologicalStatus)

Residuals:
      Min        1Q    Median        3Q       Max
-0.103828 -0.026620 -0.002653  0.024205  0.178194

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.170343   0.003409   49.96   <2e-16 ***
Proj_data_MA334$ecologicalStatus 0.839572   0.004713  178.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.037 on 5278 degrees of freedom
Multiple R-squared:  0.8574,    Adjusted R-squared:  0.8574
F-statistic: 3.174e+04 on 1 and 5278 DF,  p-value: < 2.2e-16
```

From the scatter (Figure 10), BD7 and BD11 obviously share a positive relationship during Y70. Continuing with the earlier assumptions, we find that the regression has higher R square and adjusted R square than in combined time period. As the Pr(>|t|) is less than 0.05, we reject the null hypothesis and find the intercept and BD11 statistically significant. However, lower magnitude of the coefficient of BD11 indicates lower impact on BD7 due to BD11 than for the full period.

Figure 10: Scatter and Regression of BD7 on BD11 in Y70

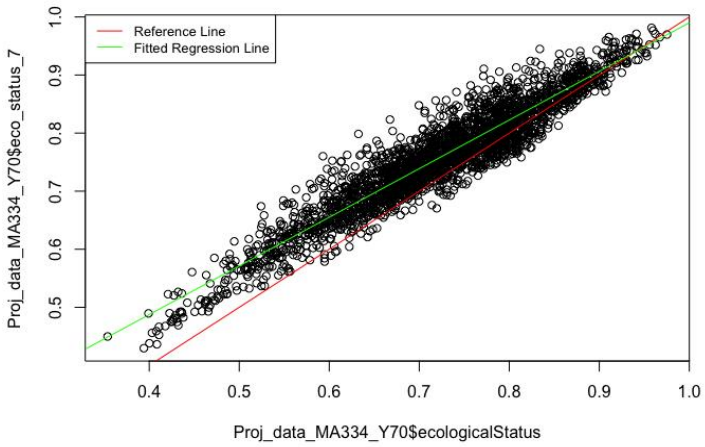


Figure 11: Regression Results of BD7 on BD11 in Y70

```
lm(formula = Proj_data_MA334_Y70$eco_status_7 ~ Proj_data_MA334_Y70$ecologicalStatus)

Residuals:
     Min       1Q   Median       3Q      Max
-0.08453 -0.02013 -0.00238  0.01807  0.11631

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.152506   0.004110    37.1   <2e-16 ***
Proj_data_MA334_Y70$ecologicalStatus 0.837544   0.005638   148.6   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03003 on 2638 degrees of freedom
Multiple R-squared:  0.8932,    Adjusted R-squared:  0.8932
F-statistic: 2.207e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

The positive relationship between BD7 and BD11 remains valid during Y00 (Figure 13). In this case, both intercept and BD11 continue to be statistically significant. Further, both R square and the coefficient of BD11 turn out the highest amongst the three time periods. In other words, the relationship between BD7 and BD11 has strengthened during Y00.

Figure 12: Scatter and Regression of BD7 on BD11 in Y00

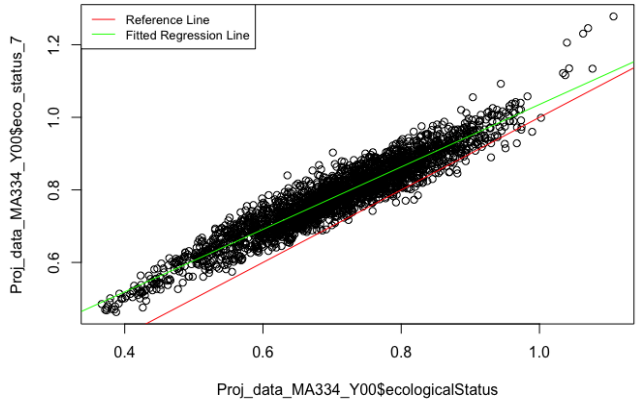


Figure 13: Regression Results of BD7 on BD11 in Y00

```
lm(formula = Proj_data_MA334_Y00$eco_status_7 ~ Proj_data_MA334_Y00$ecologicalStatus)

Residuals:
      Min        1Q    Median        3Q       Max
-0.106402 -0.022584 -0.001665  0.021566  0.150426

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                       0.174563   0.004114   42.43   <2e-16 ***
Proj_data_MA334_Y00$ecologicalStatus 0.860843   0.005731  150.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03295 on 2638 degrees of freedom
Multiple R-squared:  0.8953,    Adjusted R-squared:  0.8953
F-statistic: 2.256e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

4) Multiple Linear Regression

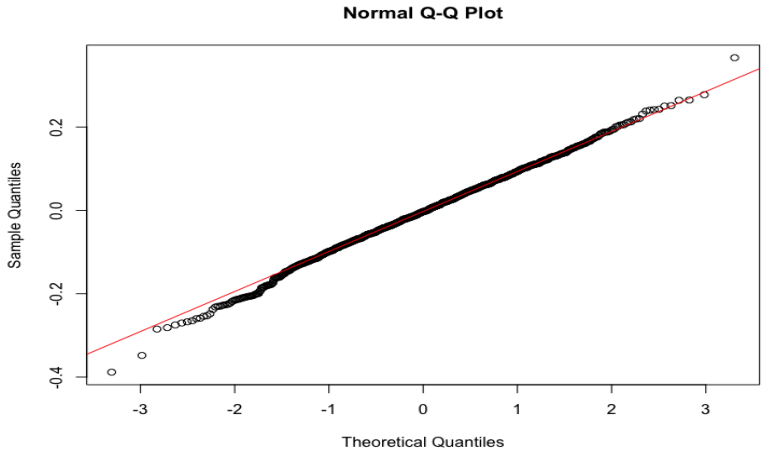
We have first constructed BD4 by taking the mean of four variables (excluding the seven select variables) out of eleven variables: Hoverflies, Isopods, Ladybirds and Grasshoppers_._Crickets. Then, considering 80% of the data set, we have run a regression, taking BD4 as the dependent variable and the select seven as the independent variables (Figure 14). Before interpretating the results of regression, we have checked the predictive power of the regression, using the remaining 20% data, i.e., testing data set. First, we find a high correlation of 0.7 between the predicted BD4 and actual BD4.

Figure 14: Regression Results of BD4 on select 7 variables

```
> summary (lmMod_train)  # model summary

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
-0.38803 -0.06129  0.00267  0.06612  0.27965

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.175629   0.018183  -9.659  < 2e-16 ***
Bees             0.084658   0.005662  14.953  < 2e-16 ***
Bird             0.168698   0.018011   9.367  < 2e-16 ***
Bryophytes      -0.064784   0.012234  -5.295 1.25e-07 ***
Butterflies      0.042987   0.013860   3.101  0.00194 **
Carabids         0.319359   0.008095  39.454  < 2e-16 ***
Macromoths       0.192626   0.015306  12.585  < 2e-16 ***
Vascular_plants  0.317104   0.016571  19.136  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09579 on 4216 degrees of freedom
Multiple R-squared:  0.6058,    Adjusted R-squared:  0.6052
F-statistic: 925.6 on 7 and 4216 DF,  p-value: < 2.2e-16
```

In Figure 15 we have checked the normality of residuals in prediction and found that sample quantiles and theoretical quantiles are intertwined. This implies that the regression model based on 80% data works well for the remaining 20% data as well.

Figure 15: Normality of residuals in prediction



In order to justify the removal or otherwise of variables as predictors, we have carried out a feature selection test: Backward Stepwise Selection based on p values from the regression and AIC (Figure 16). We find that all the seven variables turn out statistically significant and the AIC couldn't be improved. Thus, the Backward Selection exercise has retained and confirmed the regression model based on 80% data. It was also confirmed by the forward selection exercise.

Figure 16: Backward Stepwise Selection

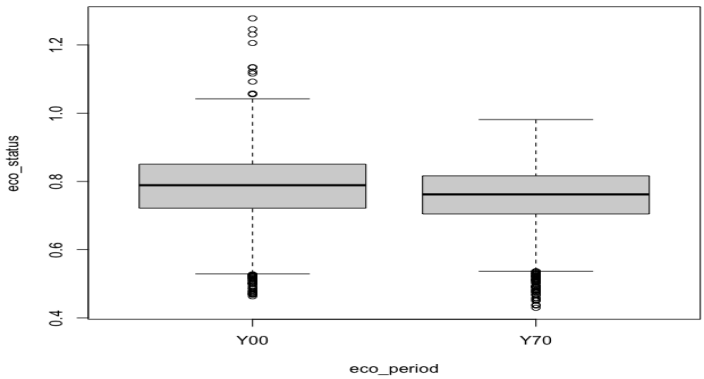
```
> summary(backward)

Call:
lm(formula = eco_status_4 ~ Bees + Bird + Bryophytes + Butterflies +
    Carabids + Macromoths + Vascular_plants, data = trainingData[c(eco_selected_names,
    "eco_status_4")], na.action = na.omit, y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38803 -0.06129  0.00267  0.06612  0.27965

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.175629   0.018183  -9.659  < 2e-16 ***
Bees             0.084658   0.005662  14.953  < 2e-16 ***
Bird             0.168698   0.018011   9.367  < 2e-16 ***
Bryophytes      -0.064784   0.012234  -5.295 1.25e-07 ***
Butterflies      0.042987   0.013860   3.101  0.00194 **
Carabids         0.319359   0.008095  39.454  < 2e-16 ***
Macromoths       0.192626   0.015306  12.585  < 2e-16 ***
Vascular_plants  0.317104   0.016571  19.136  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09579 on 4216 degrees of freedom
Multiple R-squared:  0.6058,    Adjusted R-squared:  0.6052
F-statistic: 925.6 on 7 and 4216 DF,  p-value: < 2.2e-16

> backward$anova
  Step Df Deviance Resid. Df Resid. Dev      AIC
1      NA       NA      4216   38.68677 -19807.4
> backward$coefficients
    (Intercept)            Bees            Bird      Bryophytes     Butterflies
    -0.17562880      0.08465833      0.16869792     -0.06478353      0.04298674
       Carabids      Macromoths Vascular_plants
      0.31935901      0.19262596      0.31710426
```

The regression line has R square of 0.6058. The adjusted R square value at 0.6052 remains almost the same. As the t-values are very high, both the intercept and each of the seven variables turn out statistically significant. While BD4 and each of the seven explanatory variables barring Bryophytes move in the same direction, the strength of the relationship is maximum with Carabids followed by Vascular_plants, and minimum with Butterflies. The coefficients of the variables indicate that one unit change in one of them leads to a smaller change in BD4. Overall, the p value of the model at 2.2e-16 turns out less than 0.05, implying that the regression remains statistically significant.

5) Open Analysis

BD7 is presented in the following boxplot (Figure 17). Clearly BD7 during Y00 is higher than BD7 in Y70. Also, the range of values is wider during Y00.

Figure 17: Box plots showing BD7 in different periods



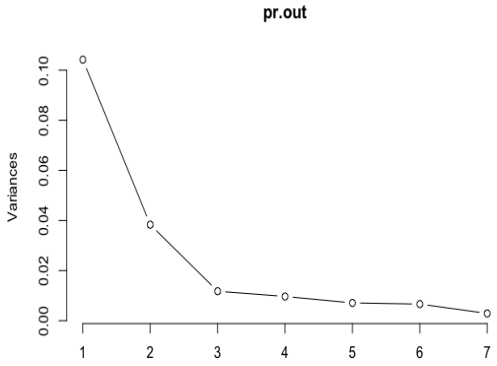
In the following we have run a multiple linear regression of BD7 on easting, northing and a categorical variable: period, i.e., a set of select independent variables other than the proportional species richness values. While the independent variables turn out statistically significant, BD7 is negatively related with periodY70, confirming that BD7 is lower in Y70 than in Y00. Similarly, BD7 is negatively related with Easting, implying that BD7 is lower in areas with higher Easting values. Once again, BD7 is lower in areas with higher Northing values. The low R-squared value of 0.1587 shows that the regression explains 15.87% of the variation in BD7. This implies that factors, not included in the model, predominantly affect BD7. Nonetheless, p-value of < 2.2e-16 being very small, indicates that the model is significant at a very high level. Therefore, periodY70, Easting, and Northing are significant predictors of BD7.

Figure 18: Linear Regression with BD7 on period, easting and northing.

```
> summary(mult_lin_mod)

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
-0.32001 -0.06078  0.00034  0.05839  0.47489

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.593e-01  5.503e-03 156.146  < 2e-16 ***
periodY70   -2.796e-02  2.474e-03 -11.302  < 2e-16 ***
Easting     -3.046e-08  1.082e-08  -2.815  0.00489 **
Northing    -1.407e-07  5.068e-09 -27.763  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08989 on 5276 degrees of freedom
Multiple R-squared:  0.1587,    Adjusted R-squared:  0.1582
F-statistic: 331.8 on 3 and 5276 DF,  p-value: < 2.2e-16
```

Towards reducing the dimensionality of BD7 while retaining the variation in dataset, we have attempted a principal component analysis of the difference (corrected for means) of the seven variables between the two periods: Y00 and Y70. As it is clear from Figure 19, the first two principal components, i.e., PC1 and PC2 account for most part of the variation in data set.

Figure 19: Principal Components and variation in data set.



The loadings of the seven select variables in PC1 and PC2 are shown in Figure 20. Clearly Bees have the highest positive loading in PC1 but small negative loading in PC2. Carabids have the highest positive loading in PC2 and the second highest positive loading in PC1. Macromoths have the highest negative loading in PC1 and the second highest positive loading in PC2. Butterflies have the highest negative loading in PC2. Thus, the dimensionality of the seven select variables is reduced to two variables: PC1 and PC2. Therefore, BD7 in effect can be approximated by considering PC1 and PC2.

Figure 20: Principal Components and loadings

	PC1	PC2
Bees	0.987723558	-0.14864999
Bird	0.020794725	0.02082772
Bryophytes	-0.005522081	-0.07181961
Butterflies	-0.015580792	-0.17061914
Carabids	0.148772408	0.96992120
Macromoths	-0.029345121	0.04695814
Vascular_plants	0.026497800	0.01575572

Concluding observation

The observations presented in the report are as per the questions given and may be taken as preliminary.