MA334-Report: 2212064

Report on Biodiversity-based indicator for large-scale environmental Assessment

Rahul Kithalamane Basavaraj

Introduction

Biodiversity assessment is one of the very important factor in determining the effect it has caused on the environment due to large spatial extents taking place in that region. This report specifically determines the extent up-to which the **land use patterns** have caused tremendous effect on the nature's Biodiversity. The assessment records here are based on 11 major taxonomic groups but in this report we consider the 7 taxonomic groups assigned to us individually which is represented as **BD7** in codes and compare that with results presented by the author. These major taxonomic group contains 5599 rows and by removing NA's we get **5281 rows** of species data across 7 major taxonomic groups at different locations inside **Great Britain** and the result is analysed between two time periods **1970-1990** and **2000-2013**. The taxonomic groups were selected based on data which covered 50% of the total hectads sampled across Great Britain.

Increasing need of land use for basic facilities such as food, housing and energy has led to a major challenge in prioritization of biodiversity. The guide to the National grid in the Ordnance survey shows exactly how the country is divided, each one of them have unique reference system. The country is divided into 100 by 100 square kilometers and each of these squares are further divided horizontally and vertically into smaller squares of 10 kilometer spacing.The vertical lines are called **eastings** when you travel east of the map and the horizontal lines are called **northings** as you travel north of the map where the variable are given in meters. We have another piece of data column represented by dominantLandClass which contains **45 different land classification** codes. The first column represents the hectad code where each variable represents one of the hectads within that square under consideration. The standardization of taxonomic groups are based on species richness and the value is obtained by dividing maximum species richness in each taxonomic group and land classification for 1970. The next variable is the authors defined biodiversity measure by taking the mean of all taxonomic groups represented by ecologicalStatus.

In this report we find the mean ecological status across 7 taxonomic groups from the wider GB biodiversity and look into the **correlation** and some interesting highlights between them in the proportional species richness category, perform some hypothesis tests, report on how linear regression of BD7 matches that with BD11, perform multiple linear regressions of BD4 against all 7 of my proportional species values including interpretation of regression coefficients and some open analysis.

Data Exploration

The number of species recorded for all the taxonomic groups after removal of NA's is 5280 where each taxonomic group species falls under 45 dominant land classifications. In this group, **dominant class 3e** which is Flat/gently undulating plains, E Anglia/S England region contains the **maximum** species of 346 and the least number of taxonomic group species are found in Upland valley sides/low mountains, Wales region. We further filter out the data and select the 7 taxonomic groups assigned to us randomly.

The estimated species richness for any given hectad was then compared as a proportion of the total species richness in the most species rich hectad of the relevant environmental zone. The mean ecological status between 1970 and 1990 was 0.68 and between 2000 and 2013 it was 0.61, so in order to compare the **summary statistics** for each taxonomic group we calculated ecological status from the latter period (2000-2013) relative to the species richness maximums from the earlier period (1970-1990). The **mean** value represents the average richness of that taxonomic group across all the study region. Here based on the result vascular plants are found in abundance, with a value of 0.79 per unit area in the study sites. The **Standard Deviation sd** here represents the spread of the richness value for each taxonomic group, the result shows that butterflies are the most tightly clustered species in the region having the value 0.14. The **skewness** here represents the symmetry of distribution across each taxonomic group. Negatively skewed taxonomic groups represents that there are more sites with less species and positively skewed taxonomic groups that there are less sites with abundance of species. Here the result shows that Carabids are the least abundant species across the region.

```
               taxi_group mean   sd skewness
1         Vascular_plants 0.79  0.1    -0.13
2              Bryophytes 0.79 0.13     -0.2
3             Butterflies 0.87 0.14    -0.36
4 Grasshoppers_._Crickets 0.63 0.21    -0.09
5                Carabids 0.61 0.21    -0.49
6                 Isopods 0.55 0.22     0.05
7               Ladybirds 0.61 0.27     0.03
```

Here we made a study of **correlation** between continuous variables. The plot shows that there is a strong positive correlation between eastings and northings as the value of northing increases eastings also increase. On the analysis of correlation between ecological species andeastings it was observed that there was a strong positive correlation based on positive correlation coefficient value 0.166567 but the plot indicated not very significant increase of ecological status with the rise in Easting value but the relation is quite **strong**.

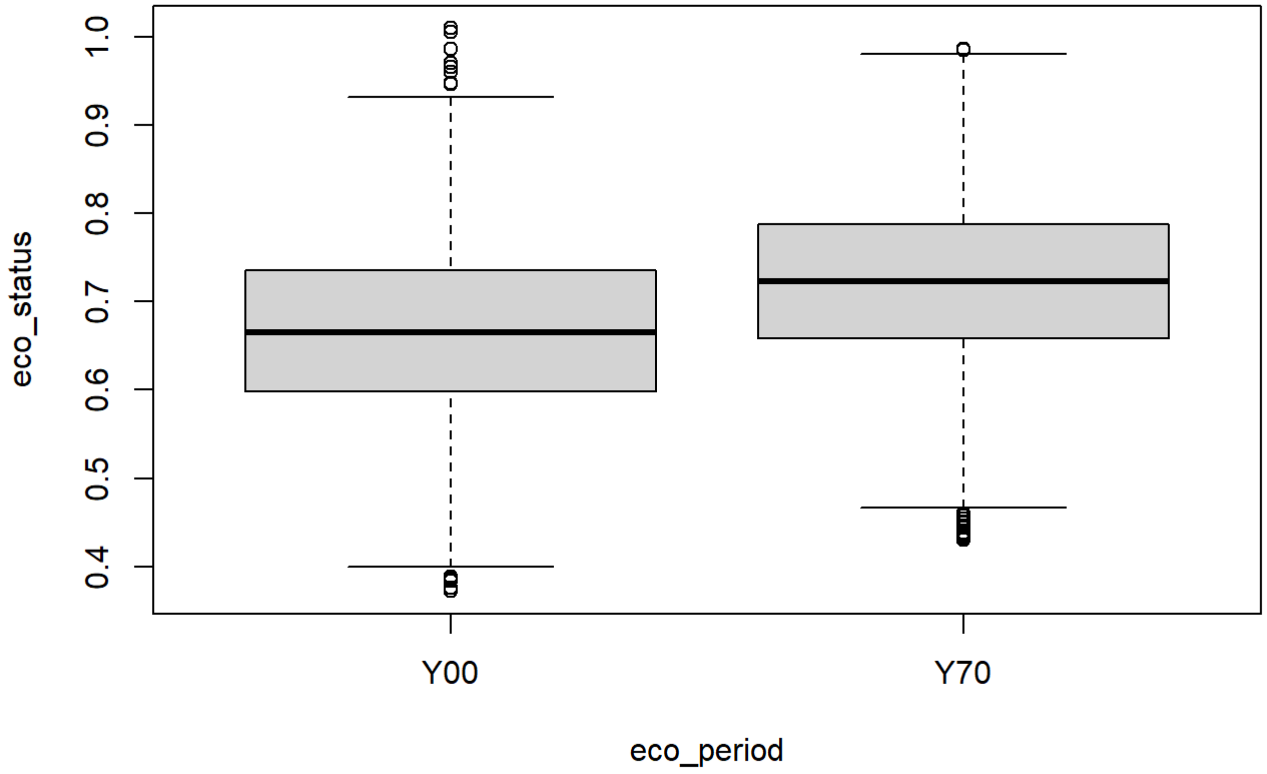
On the analysis of correlation between ecological species and Northing it was observed that there was a negative correlation based on the obtained correlation coefficient value of -0.4056032 but the plot also indicated a negative correlation between these two variables. There is a tendency of biodiversity measure to decrease with the increase in value of Northing. It can be said that the relation between these variables is not very strong.

Hypothesis Tests

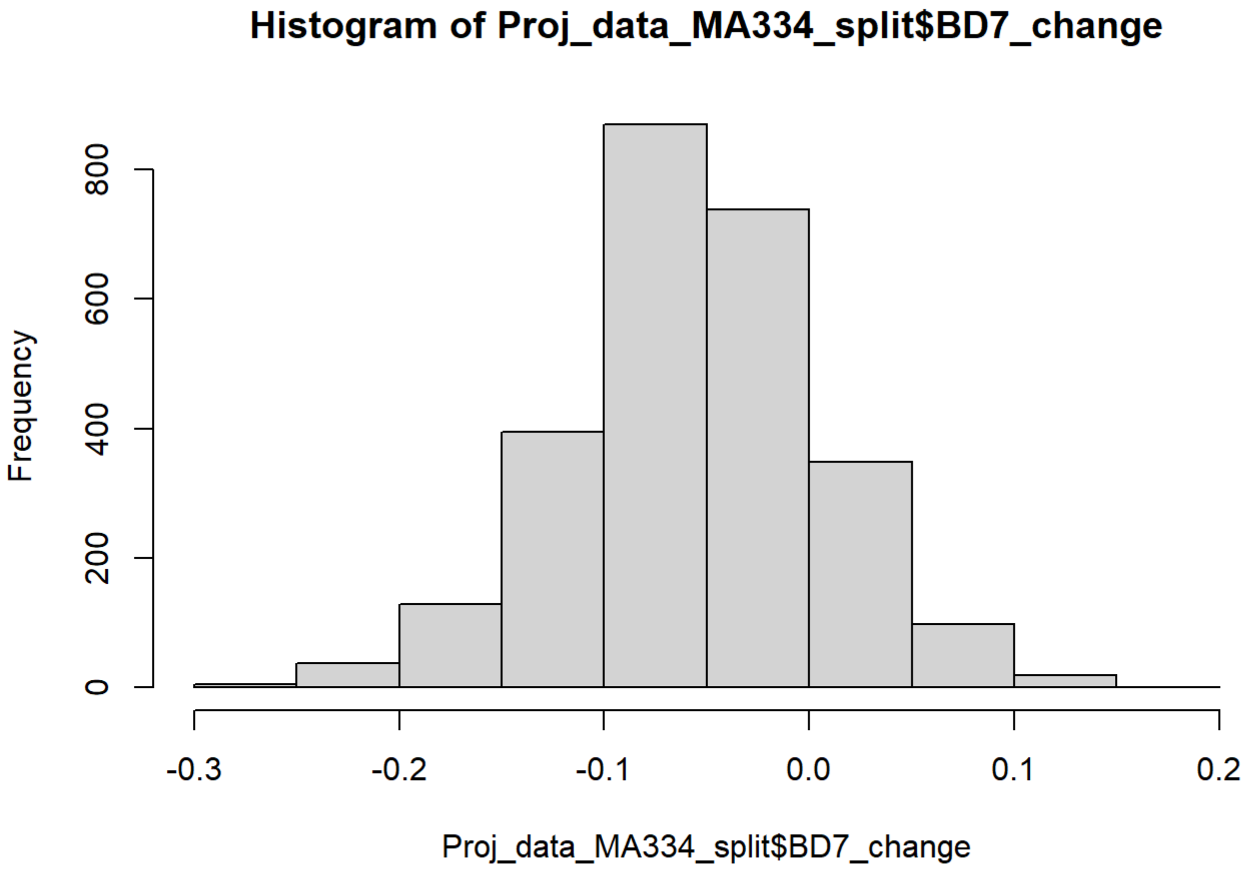
In this analysis we fit **linear regression model** with eco_status_7 as response variable and Northing as predictor variable. We find the summary of this model which executes Coefficients, R-squared values, F-statistic and p-values. Here the **p-value** is the observed difference in the outcome measure and the value here is less than 0.05 which indicates there is a stronger evidence to reject null hypothesis. In this model the R square value is 0.1644 which states that only **16.4%** of the eco_status_7 variability in the model can be explained by Northing predictor variable and is adjusted to 16.51% by penalizing those models which do not improve the overall fit. The residual plot states that the values are having a equal spread and they follow a normal distribution. While the points on the QQ plot will fall approximately on the straight line, this means that they follow a **normal distribution**. The assumptions of linearity, constant variance and normality are met looking into them.

```
Call:
lm(formula = Proj_data_MA334$eco_status_7 ~ Proj_data$Northing)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.298238 -0.067824 -0.000699  0.070969  0.302534 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         7.638e-01  2.571e-03  297.03   <2e-16 ***
Proj_data$Northing -1.575e-07  4.886e-09  -32.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09562 on 5278 degrees of freedom
Multiple R-squared:  0.1645,	Adjusted R-squared:  0.1644 
F-statistic:  1039 on 1 and 5278 DF,  p-value: < 2.2e-16
```

The **box plot** comparison between eco_status and eco_period shows the distribution of a data set for the two periods. The **mean** of the eco_status in the period 2000-2013 is **0.63** and the data ranges between 0.45 - 0.72 whereas the mean of the eco_status in the period **1970-1990** is 0.69 and the data ranges between 0.60 - 0.75.

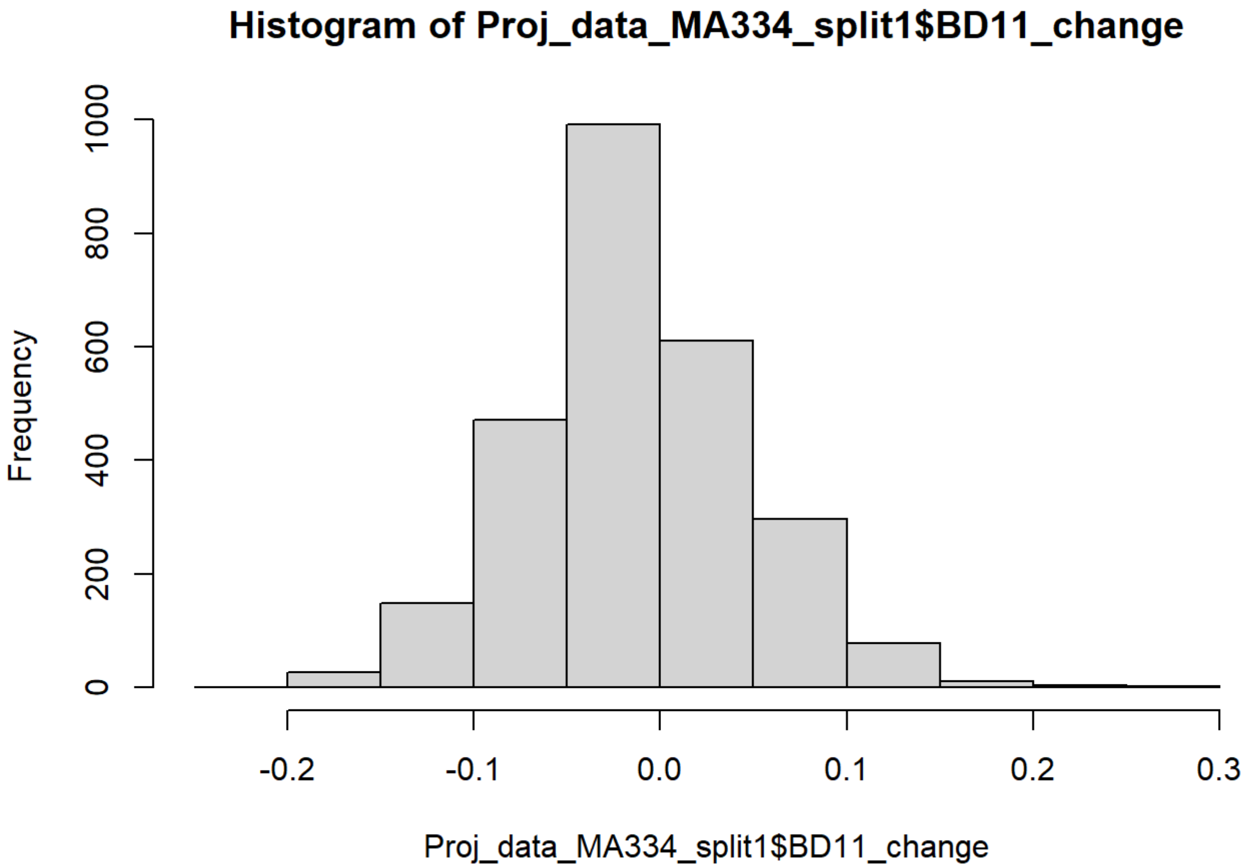


The **histogram** below shows the distribution of biodiversity change for the two time periods 2000-2013 and 1970-1990 according to the location. Here we try to conduct t test analysis to determine whether the mean of the ecological status between 2 time periods is equal to zero. On analysis we found out that t = -45.878, df = 2639 and p-value = 2.2e-16 indicating that mean of BD_change is not equal to zero and hence we reject null hypothesis. 95 percent confidence interval means that 95% of the BD7_change values falls between these 2 ranges: -0.05785916 and -0.05311595.



```
	One Sample t-test

data:  BD7_change
t = -45.878, df = 2639, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.05785916 -0.05311595
sample estimates:
  mean of x 
-0.05548756 
```

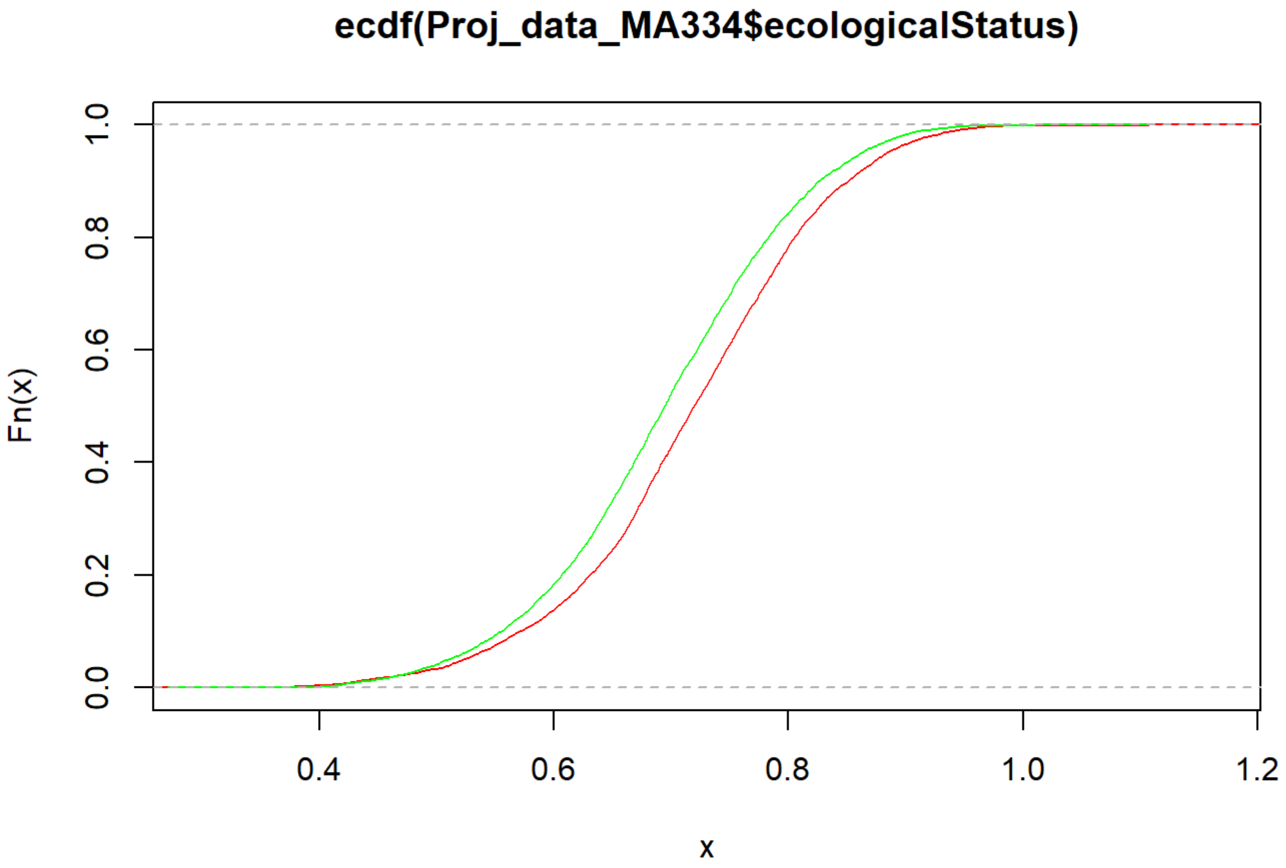
The **histogram** below shows the distribution of biodiversity change for the two time periods 2000-2013 and 1970-1990 according to the location. Here we try to conduct t test analysis to determine whether the mean of the ecological status between 2 time periods is equal to zero. On analysis we found out that t = -10.955, df = 2639 and p-value = 2.2e-16 indicating that mean of BD_change is not equal to zero and hence we reject null hypothesis. 95 percent confidence interval means that 95% of the BD7_change values falls between these 2 ranges: -0.01494000 and -0.01040361



```

	One Sample t-test

data:  BD11_change
t = -10.955, df = 2639, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.01494000 -0.01040361
sample estimates:
  mean of x 
-0.01267181 
```

After rejecting the null hypothesis we then check the distribution of biodiversity division based on 7 and 11 taxonomic groups. On the basis of the QQ plot we can conclude that the distribution of eco_status between 2 different sets of biodiversity groups are not similar since the point don't fall on the same line.

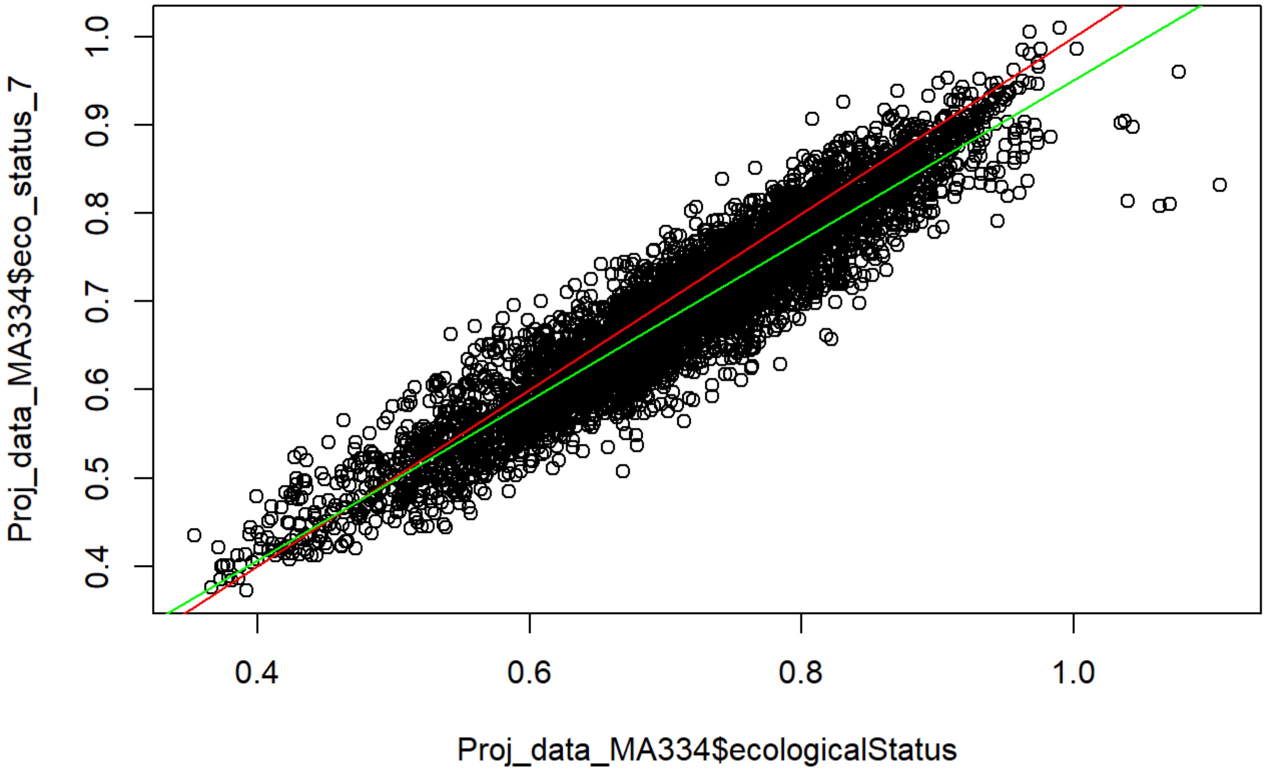
Here we then try to visualize the distribution of data points and its variables in the empirical cumulative distribution function and find how different they are by conducting **Kolmogorov-Smirnov test**. The test result shows that the maximum difference between the two data sets is 0.098674 and the p-value is 2.2e-16 which is indicative of the fact that it is more that 0.05 meaning we cannot reject null hypothesis and hence we can conclude that their is no significant difference in the ecological species data of these two taxonomic groups.



```

	Asymptotic two-sample Kolmogorov-Smirnov test

data:  Proj_data_MA334$eco_status_7 and Proj_data_MA334$ecologicalStatus
D = 0.098674, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Simple linear regression

In the plot of eco_status_7 against ecological status based on all 11 we found a line of equality where eco_status_7 equals ecologicalStatus in the 1st phase and to find the relationship between them we found a linear regression model of eco_status_7 as a function of ecologicalStatus and found the best fit line which best represents the relationship between these 2 variables in a scatter plot.



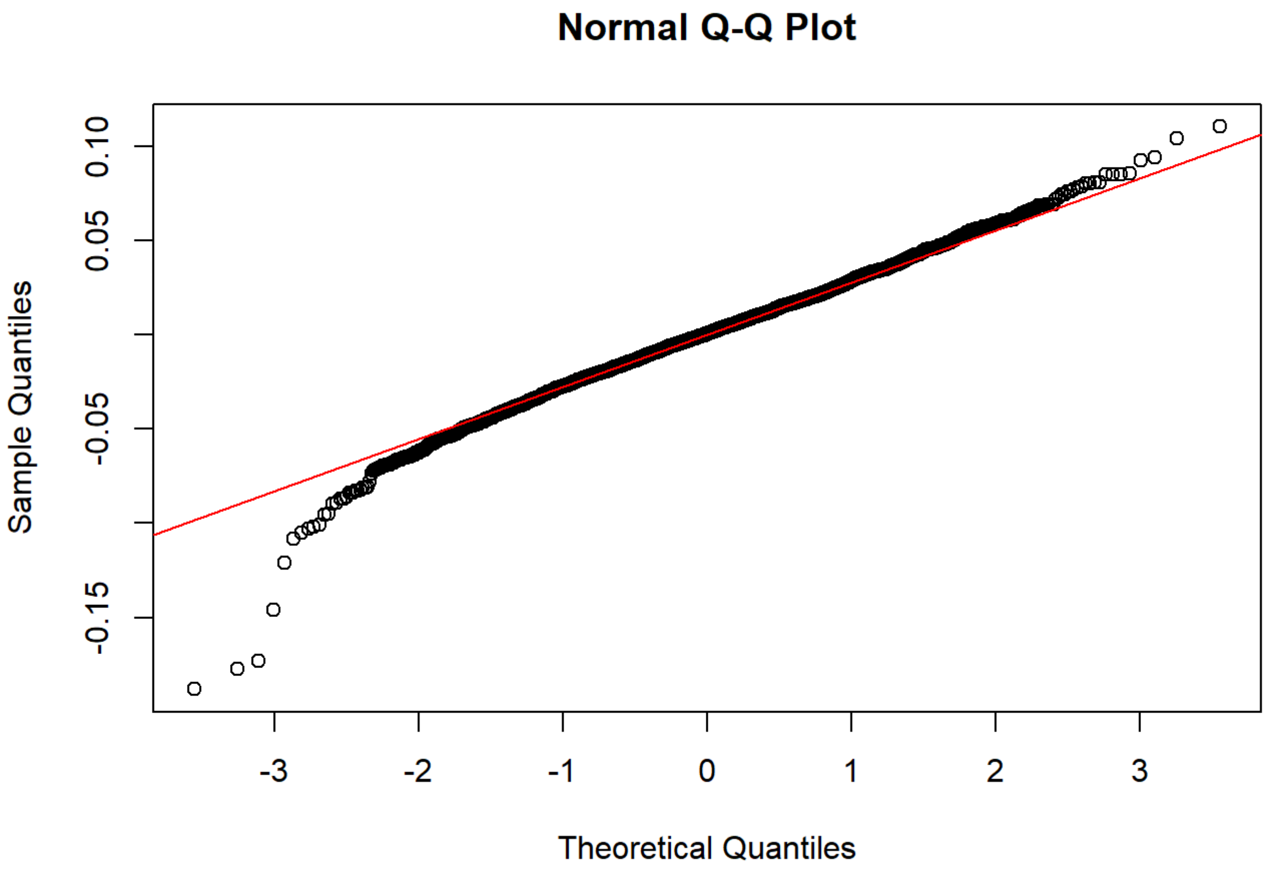
After finding the relationship based on linear regression model, below in this plot we try to analyse the data of overlapping where the noise and spread of residuals is visualized on the plot. Based on the analysis of the plot we find that the residuals are randomly scattered around the horizontal line which represents the linear regression model is a good fit for this data. We then analysed the data through QQ plot to check whether the residuals of linear regression model are normally distributed. On reading the plot it was found that the points are slightly deviated from the line which is an indication that it is not a perfect normal distribution curve.

In this we report on the difference in **coefficients** for each time periods between eco_status_7 against ecological status based on all 11. Based on the 1970-1990 period result the intercept value is 0.2264454 and the coefficient of ecological status is 0.6813281 which means that there will be an increase of 0.681 in biodiversity index for each unit increase in ecological status, in the period 2000-2013 the intercept value is -0.1127431 and the coefficient of ecological status is 1.0991741 which means that 1.0991741 increase in biodiversity index for each unit increase in ecological status.

In this analysis of biodiversity measure we took the mean of 7 taxonomic groups against the mean of 4 taxonomic groups. A line of equality was drawn where eco_status_4 equals eco_status_7 in the 1st phase, later we created a linear regression model of eco_status_4 as a function of eco_status_7 and found the best fit line which best represents the relationship between these 2 variables in a scatter plot.

On analysis we fit linear regression model between eco_status_4 and eco_status_7. We find the summary of this model which executes Coefficients, R-squared values, F-statistic and p-values. The coefficient for eco_status_7 states for every 1 unit increase in eco_status_7 there is expected change in eco_status_4. Here the p-value is the observed difference in the outcome measure and the value here is greater than 0.05 which indicates we cannot reject null hypothesis. In this model the R square value is 0.03 which states that only small proportion of variability in eco_status_4 is explained by eco_status_7 in this model. The residual plot states that the values are having a equal spread approximately and they follow a normal distribution.

After finding the relationship based on linear regression model, below in this plot we try to analyse whether the residuals are randomly scattered and by how much do they deviate from zero or is it following any pattern. Based on the analysis of the plot we find that the residuals are randomly scattered around the horizontal line which represents the linear regression model is a good fit for this data. We then analysed the data through QQ plot to check whether the residuals of linear regression model are normally distributed. On reading the plot it was found that the most of the points are are falling in the straight which is an indication that it follows a path of normal distribution curve.



Multiple Linear Regression

Here we perform **multiple linear regression** of BD4 against all seven of the selected proportional species values. We find the summary of this model which outputs Coefficient values for individual variables, R-squared values, F-statistic and p-values. The residual plot states that the values are having a mean value equal to 0 indicating that the model is a good fit. The Residual standard error: 0.08564 suggest the distance the data points fall into the regression line. From this data we are able to suggest that 63.67% of the changes we observe in eco_status_4 can be explained by the selected 7 biodiversity species. The Adjusted R-squared value of 0.6367 states that 63.67% of the data sets are spread out in eco_status_4. The F statistics in this biodiversity measure explains how independent variables can explain the dependent variables and here since the value of p is 2.2e-16 which is very less, it means that the model is significant and at least 1 of the independent variables in the biodiversity model is a significant predictor of the dependent variable.

The output correlation coefficient here suggest how well the data is fitting into the model. Since the value of correlation is 0.7982905 which is on higher side, it indicates a strong positive correlation between predicted and actual values (lmMod_train$y) in the training data. The correlation coefficient between the predicted and actual values (eco_status_4) is 0.7917619 which indicates that the linear relationship is moderately strong. This plot indicates that the model is not perfectly accurate in predicting the output variable eco_status_4. This plot in the biodiversity species measure is indicate of the fact whether the residuals are scattered around the red line and is not following any pattern. Hence the model need not be further improved for predicting values.

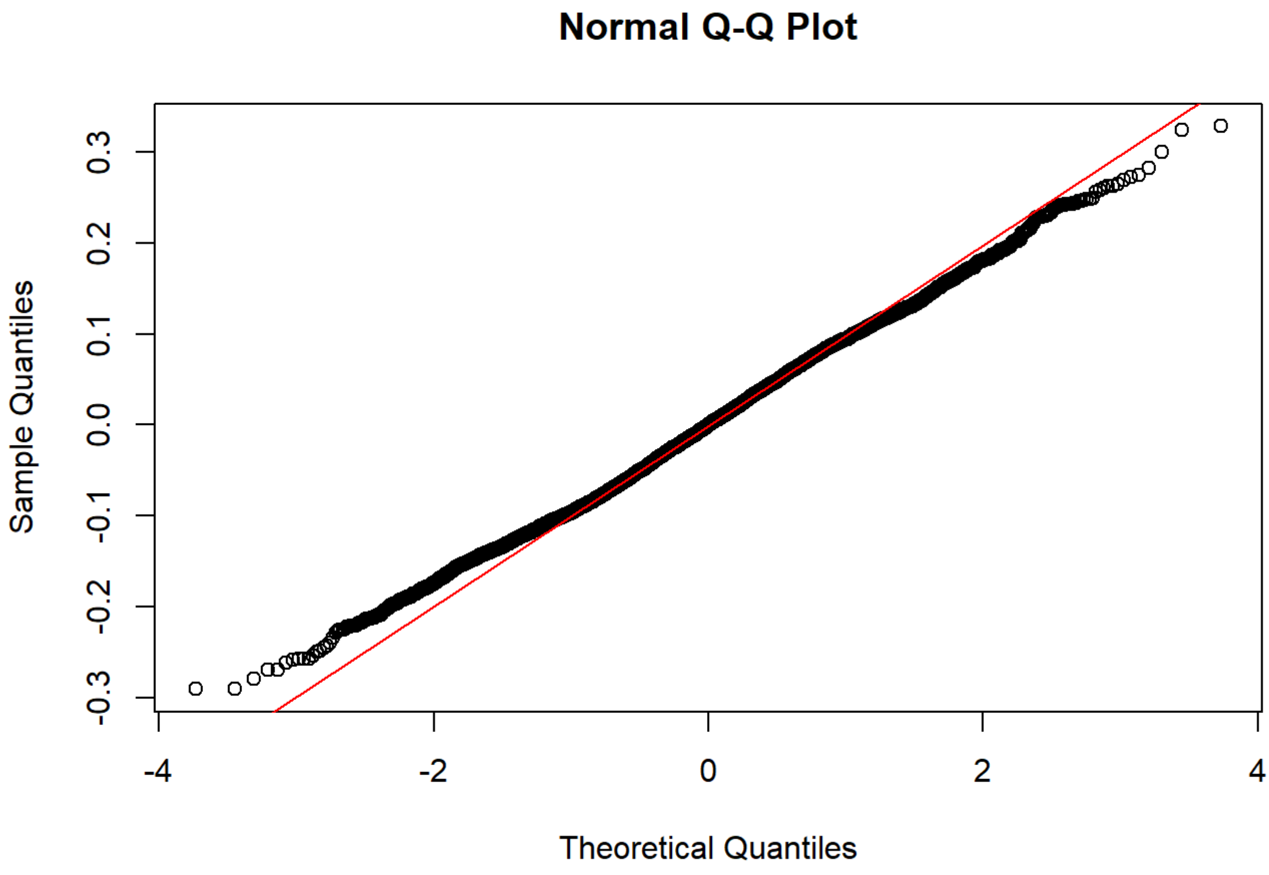
We then analysed the data through QQ plot to check whether the residuals of linear regression model are normally distributed. On reading the plot it was found that the points are slightly close to the reference line which is an indication that it is approximately a normal distribution curve.

Here we perform multiple linear regression of BD7 against period, easting and northing. We find the summary of this model to display the output Coefficient values for their standard errors, R-squared and Adjusted R-squared values, F-statistic and p-values. The residual plot states that the values are having a mean value equal to 0 indicating that the model is a good fit. The **Residual standard error**: 0.09152 suggest that the model explains about 54.58% of the variability in eco_status_7. The Adjusted R-squared value of 0.2345 states that 23.45% of the data sets are spread out in eco_status_7. The **F statistics** in this biodiversity measure explains how independent variable is significant in determining dependent variables and here since the value of p is 2.2e-16 which is very less, it means that the model is significant in predicting at least 1 of the independent variable.

This plot indicates that the predicted value differs from actual values if the points are farther away from the red line and if they are close it means that the model is performing well at predicting values.

Below in this plot we try to analyse whether the predicted values and the residuals of the multiple linear regression model are randomly scattered and by how much do they deviate from zero or is it following any pattern. Based on the analysis of the plot we find that the residuals are randomly scattered around the horizontal line which represents that the model is a good fit for this data.

We then analysed the data through **QQ plot** to check whether the residuals of the multiple linear regression model are normally distributed. On reading the plot it was found that the points are mostly close to the reference line which is an indication that it is approximately a normal distribution curve.



```
Start:  AIC=-20735.39
eco_status_4 ~ Bryophytes + Butterflies + Carabids + Isopods + 
    Ladybirds + Grasshoppers_._Crickets + Vascular_plants

                          Df Sum of Sq    RSS    AIC
<none>                                 31.056 -20735
- Isopods                  1    0.0360 31.092 -20733
- Vascular_plants          1    0.1554 31.212 -20716
- Bryophytes               1    0.2133 31.270 -20709
- Grasshoppers_._Crickets  1    0.5771 31.633 -20660
- Carabids                 1    1.7232 32.780 -20509
- Butterflies              1   11.6009 42.657 -19397
- Ladybirds                1   12.1485 43.205 -19343
```

```


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

Coefficients:
            (Intercept)               Bryophytes              Butterflies  
                0.03403                  0.06051                  0.41375  
               Carabids                  Isopods                Ladybirds  
                0.12359                 -0.01540                  0.23782  
Grasshoppers_._Crickets          Vascular_plants  
                0.06794                  0.07137  
```

Here we are comparing the effect of each significant coefficient to that of period variable. This gives us an idea of how much the change in eco_status_7 is seen with one unit increase in Eastings and Northings.

```
  (Intercept)      periodY70       Easting      Northing 
 7.381922e-01  5.548756e-02 -4.859019e-09 -1.584821e-07 
```

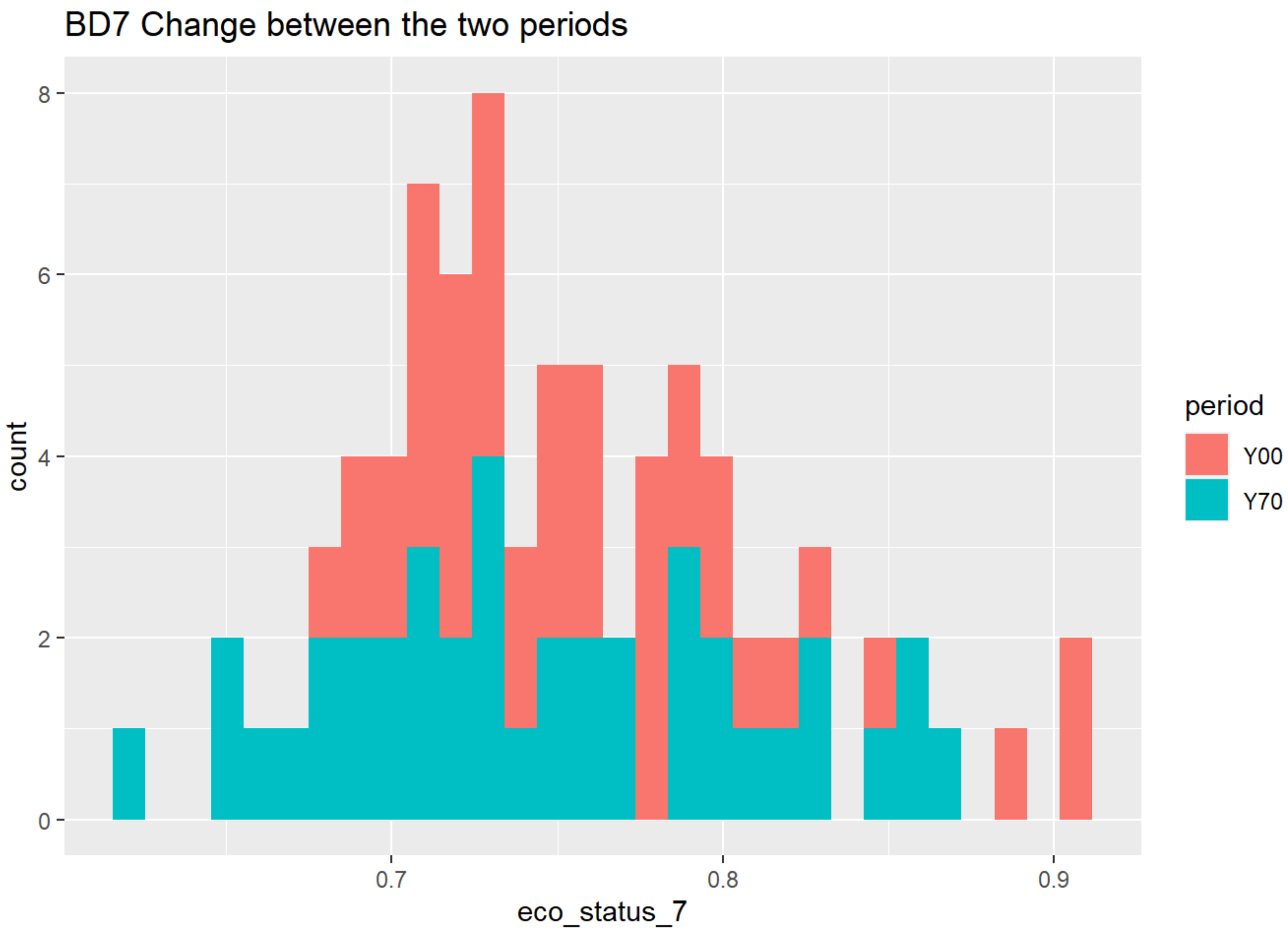
The following PCA method helps us in visualizing the multi-dimensional spread of biodiversity values for the two separate time periods corresponding to the location. The Principal Component analysis are done on the eco_difference data. We then use screen plot function to plot the variances in decreasing order. This plot shows the contribution of each component to the total variances in the data set.

This is a scatter plot which provides the analysis of first and second principal component analysis of the biodiversity data. The class found in these two principal components are labelled for each data point. We found on labeling that 3e Flat/gently undulating plains, E Anglia/S England is the dominant land classification across the study area. This is also a scatter plot which provides the analysis of first and second principal component analysis of the biodiversity data. The locations for these two principal components are labelled for each data point. The PCA analysis here is the biodiversity differences between the two time periods where each point represents a location. The differences between these two time periods are labeled using text which are useful for identifying the change in patterns of biodiversity.

```
                              PC1         PC2
Bryophytes            -0.03876021  0.02294282
Butterflies           -0.11706392  0.04221565
Carabids               0.57949445 -0.01481220
Isopods               -0.53935122  0.04475990
Ladybirds              0.28073091  0.90744012
Grasshoppers_._Crickets  0.52604338 -0.41374690
Vascular_plants        0.05048411  0.02888850
```

Open Analysis

In this analysis we considered the dominant land class as Upland valleys/rounded hill sides of England represented by 17e. We compared this with ecological status of 7 species between 2 time periods 1970-1990 and 2000-2013. Here we find the species ecological status distribution over a period and identify during which time they were highly spread. The maximum species richness was found between the periods 2000-2013 as compared to 1970-1990. These BD7 species ecological status maximum was in the range of 0.7 to 0.8. These species were plotted over a histogram to identify the maximum ecological status a species occur.



Result

The number of species recorded for the 11 taxonomic groups were 5281. The species richness depends on the biodiversity and land cover composition. Ecological status was calculated as the proportion of total species richness in a given hectad relative to the most species rich hectad in the given environmental zone. The mean ecological status across all hectads and environmental zones is 0.70. A strong positive correlation was observed between BD7 and BD4 ecological status. The area across the GB include large proportion of individual environmental zones. For any national scale project this methods to identify the species richness in the region gives us a preliminary stage assessment tool. We also analysed that there is a tendency of biodiversity measure to decrease with increase in value of northing. The future changes are observed based on ecological status in these biodiversity zones. The largest declines were observed in Carabids and Isopods.

References

https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.12784 (https://besjournals.onlinelibrary.wiley.com/doi/10.1111/1365-2664.12784)