Bio-diversity measure of species richness in the United Kingdom

2214122

Abstract—The development of land use and other human activities have a negative effect on biodiversity. The purpose of this study is to examine the species richness of the taxonomic categories over two time periods and determine whether such human activities have had an impact on biodiversity and smaller species. Various statistical analyses have been performed on the data. Species richness estimates have been seen to generally go down over the course of the two time periods.

Index Terms—biodiversity, taxonomic groups, ecological status, mean, species richness, time period, regression, coefficients, and p-value.

1 Introduction

The term "biological diversity" refers to the variety of life on earth, including the diversity within and between species as well as ecosystems. The impact of various human activities on this biodiversity must be screened. Also, it is necessary to research the direct and indirect effects of various ongoing projects on biodiversity in various geographic locations. In Great Britain, environmental impact assessments (EIAs) and strategic environmental assessments (SEAs) are used to determine the impact of various land-use developments on various species. [1]However it was found to be not so effective in determining the effect on a varied range of biodiversity and mainly focused on some particular species and the geographical areas where they dwell. [2] It does not concentrate on other smaller species because its major focus is on threatened species and habitats. These other smaller species also contribute significantly to various ecosystems and should be researched. [3] Dyer, Gillings et al in their research developed a biodiversity indicator and a new method named Frescalo to assess the impact and variations of environmental changes in shale gas extraction sites across 11 taxonomic groups consisting of 5553 species across two time periods which is outside the scope of this paper. [3]

In this study, we will concentrate on the Estimated Species Richness dataset, which spans the time periods of 1970–1990 and 2000–2013 and includes the 11 taxonomic groups Bees, Birds, Bryophytes, Butterflies, Carabidae, Hoverflies, Isopoda, Ladybirds, Moths, Grasshopper and Crickets, and Vascular plants. Initially, we'll examine the seven taxonomic groups that have been allotted to us: bryophytes, butterflies, carabids, hoverflies, isopods, grasshoppers_crickets, and vascular plants. We'll follow up with an analysis of all 11 groups during the two time periods. We will report the changes in the seven groups across the two time periods and also report on how the mean of the seven groups differs from the total mean of all 11 groups.

2 Methodology

2.1 Data-set explanation

The UK species richness data set has been taken from the Environmental Information Data Centre and it has been collected from Biological Records Centre (BRC). Data was collected at 10km2 scale squares(location column in the dataset) for two time periods of 1970-1990 and 2000-2013. There are 45 land classifications present spread across England, Scotland, and Wales. There are 21 different land classifications in England, 16 for Scotland and 8 in Wales. [4]

Data Description: The national grid map of Great Britain is covered by grid squares measuring 100 kilometres across. Each grid square is identified by two letters, such as HW, HX, NM, NO, SJ, SU. [5]

1) Location:- 10km x 10km square scale identified by two letters and assigned to an environmental zone, determined by land cover type, climate, geology, and topography.
2) Bees:- Species richness value
3) Bird:- Species richness value
4) Bryophytes:- Species richness value
5) Butterflies:- Species richness value
6) Carabids:- Species richness value
7) Hoverflies:- Species richness value
8) Isopods:- Species richness value
9) Ladybirds:- Species richness value
10) Macromoths:- Species richness value
11) Grasshoppers_Crickets:- Species richness value
12) Vascular_plants:- Species richness value
13) Easting:- The vertical lines on the national grid map. They increase in value as we move eastwards on the map. [5]
14) Northing:- The horizontal lines on the grid map. They increase in value as we move up north on the map. [5]
15) dominantLandClass:- 45 land classes taken from 2007 ITE land classification. They were used to assign specific environmental zones to the hectad

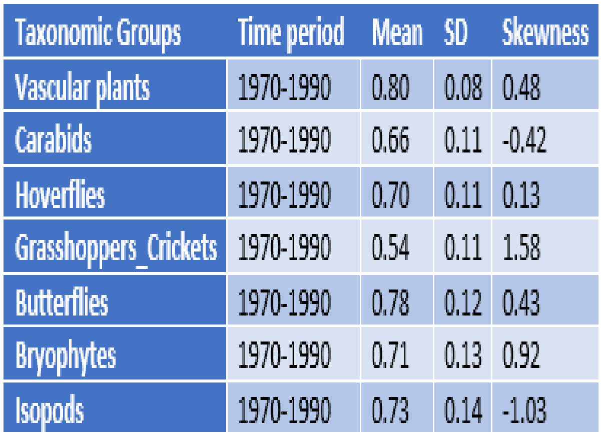
areas under similar ecosystems or habitats (Bunce et al. 2007) [6]

16) ecologicalStatus:-calculated using a relative measure of estimated species richness
17) period:- Y70 indicates the time period 1970-1990 and Y00 indicates the time period of 2000-2013
In the next section, we will review the dataset and data analysis methods used in the project.

2.2 Data exploration

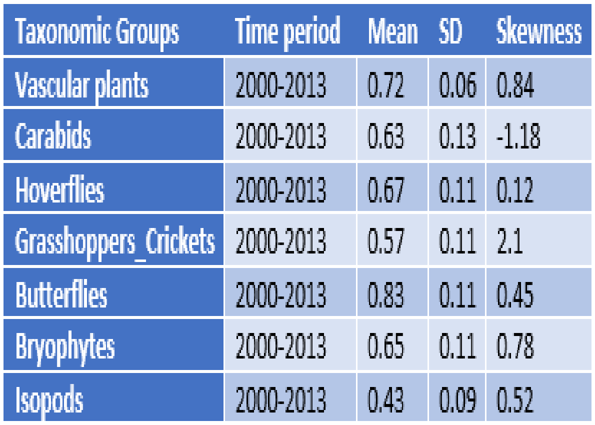
The dataset was checked for null values and all such values were eliminated and the location squares were selected which had data for all taxonomic groups for both time periods. This version of data with 5280 records has been used for analysis. We have selected the land classification of 3e(Flat/gently undulating plains, E Anglia/S England) with 346 records for the analysis of the seven allocated taxonomic groups.

- **Uni-variate analysis** The seven taxonomic groups are bryophytes, butterflies, carabids, hoverflies, isopods, grasshoppers_crickets, and vascular plants. The mean, standard deviation, and skewness of the seven groups have been calculated and compared for the two time periods. In Table 1 and Table 2 we can see the mean values for only butterflies and grasshopper_crickets have increased in the new time period with the standard deviation being almost the same and having a low value indicating that the data points are close to the mean value. The positive skewness indicates an excess of low values. For isopods, it has decreased drastically from 0.73 to 0.43 followed by carabids and bryophytes. There is not much change in standard deviation. For isopods the negative skewness of -1.03 changes to the positive skewness of 0.52 in the new time period.So the distribution changes from an excess of high values to an excess of low values in the new time period.



Taxonomic Groups	Time period	Mean	SD	Skewness
Vascular plants	1970-1990	0.80	0.08	0.48
Carabids	1970-1990	0.66	0.11	-0.42
Hoverflies	1970-1990	0.70	0.11	0.13
Grasshoppers_Crickets	1970-1990	0.54	0.11	1.58
Butterflies	1970-1990	0.78	0.12	0.43
Bryophytes	1970-1990	0.71	0.13	0.92
Isopods	1970-1990	0.73	0.14	-1.03

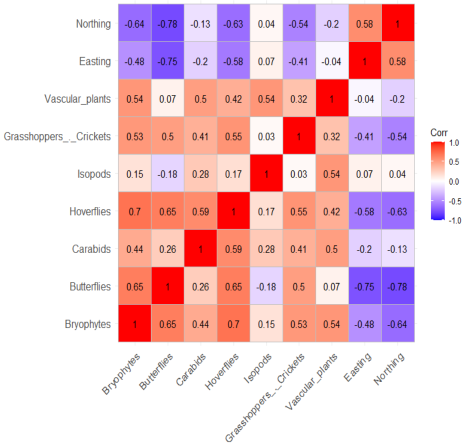
Fig. 1. Table1: Mean, Standard Deviation and Skewness for the time period 1970-1990 and



Taxonomic Groups	Time period	Mean	SD	Skewness
Vascular plants	2000-2013	0.72	0.06	0.84
Carabids	2000-2013	0.63	0.13	-1.18
Hoverflies	2000-2013	0.67	0.11	0.12
Grasshoppers_Crickets	2000-2013	0.57	0.11	2.1
Butterflies	2000-2013	0.83	0.11	0.45
Bryophytes	2000-2013	0.65	0.11	0.78
Isopods	2000-2013	0.43	0.09	0.52

Fig. 2. Table 2: Mean, Standard Deviation and Skewness for the time period 2000-2013

- **Correlations between the seven variables** The correlation matrix in Figure 2 will help us to identify how strongly any two of the variables are related to each other. Only Isopods have a low positive correlation of 7% and 4% with Eastings and Northings respectively. Bryophytes, Butterflies, Hoverflies, and Grasshopper_Crickets have a high negative correlation while carabids and vascular plants have a low negative correlation with Eastings and Northings. Most of the seven species have a medium to high positive correlation with each other. The only exception is Butterflies and Isopods having a negative correlation of 18%.
The second image in Figure 2 shows the species distribution across Eastings and Northings. Most of the species are concentrated as we travel more east and higher up north. Then there is a drop off of the species after 600000 Eastings. Next, we will observe the seven species with the Eastings and Northings separately.



	Bryophytes	Butterflies	Carabids	Hoverflies	Isopods	Grasshoppers_._Crickets	Vascular_plants	Easting	Northing
Northing	-0.64	-0.78	-0.13	-0.63	0.04	-0.54	-0.2	0.58	1
Easting	-0.48	-0.75	-0.2	-0.58	0.07	-0.41	-0.04	1	0.58
Vascular_plants	0.54	0.07	0.5	0.42	0.54	0.32	1	-0.04	-0.2
Grasshoppers_._Crickets	0.53	0.5	0.41	0.55	0.03	1	0.32	-0.41	-0.54
Isopods	0.15	-0.18	0.28	0.17	1	0.03	0.54	0.07	0.04
Hoverflies	0.7	0.65	0.59	1	0.17	0.55	0.42	-0.58	-0.63
Carabids	0.44	0.26	1	0.59	0.28	0.41	0.5	-0.2	-0.13
Butterflies	0.65	1	0.26	0.65	-0.18	0.5	0.07	-0.75	-0.78
Bryophytes	1	0.65	0.44	0.7	0.15	0.53	0.54	-0.48	-0.64

Fig. 3. Correlation matrix of the seven taxonomic groups, Eastings, and Northings.

- **Relation between the ecological status of 7 taxonomic groups and Eastings** : We examine the resid-

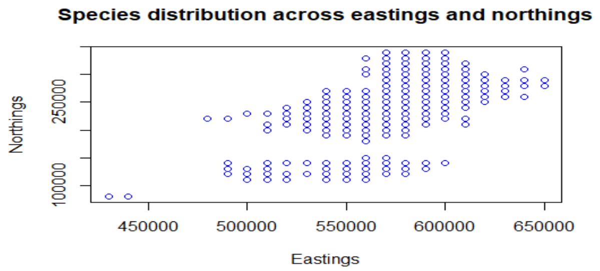


Fig. 4. Species distribution of the seven groups across Eastings and Northings

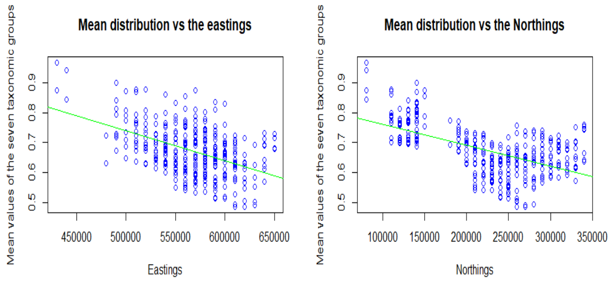
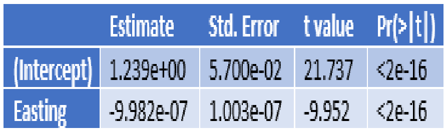


Fig. 5. Mean of 7 species against Eastings and Mean of 7 species against Northings

ual plot after fitting the regression model to make sure the errors don't exhibit any patterns. The residuals are dispersed randomly across the entire range of fitted values, as is seen from the plot in Figure 7. It can be observed from Table 3 that Eastings has a p-value of less than 2.2e-16 which is below the 0.05 level of significance, indicating that it is a significant variable in the regression model. Also, there is a negative correlation of 0.47 between the mean of the seven groups and Eastings as can be seen in Figure 5, where, as the Eastings increases the mean value decreases. The Q-Q plot for the linear regression model in Figure 7 shows the points forming a line that's roughly straight so we can assume that the data is normally distributed.

- **Relation between the ecological status of 7 taxonomic groups and Northings** Similarly, for the Northings, the residuals are randomly dispersed(Figure 9) and there is no discernible pattern. Table 4 demonstrates that Northings can also be employed as a predictor in the model because its p-value less than 2.2e-16 is below the 0.05 level of significance. Moreover, there is a negative correlation of 0.56 between the mean of the seven groups and the Northings.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.239e+00	5.700e-02	21.737	<2e-16
Easting	-9.982e-07	1.003e-07	-9.952	<2e-16

Fig. 6. Table 3:Coefficients of linear regression for BD7 and the Eastings

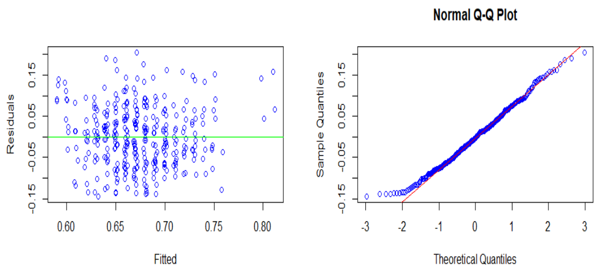
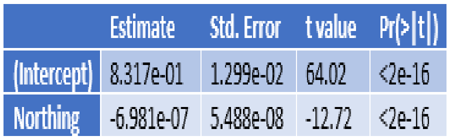


Fig. 7. Residuals vs Fitted values and Normal Q-Q plot for linear regression model of 7 taxonomic groups and the Eastings



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.317e-01	1.299e-02	64.02	<2e-16
Northing	-6.981e-07	5.488e-08	-12.72	<2e-16

Fig. 8. Table 4:Coefficients of linear regression for BD7 and the Northings

The points practically fall on a straight line in the Q-Q plot(Figure 9) for this model, which supports the notion that the data has a normal distribution. Finally, the ecological status of the seven species has been compared between the two time periods. It is clear from the box plots in Figure 10 that for the older time period(1970-1990), the median value is greater and the values are also more dispersed.We also observe some outliers for both time periods.

2.3 Hypothesis Testing

1) **One Sample t-test** : We first determine the change in the mean of the seven taxonomic groupings across the two time periods of 1970–1990 and 2000–2013 and examine if it is a normal distribution for the one-sample t–test (Figure 11). Then we will define the question and the null and alternate hypotheses. Question: Is the mean value of the sample significantly larger (or smaller) than the mean value of the population?
Null Hypothesis HO: True mean is equal to zero Alternative hypothesis H1: True mean is not equal to 0

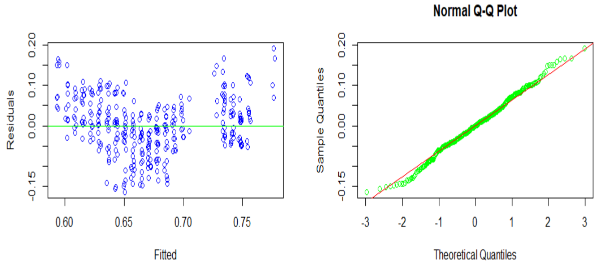


Fig. 9. Residuals vs Fitted values and Normal Q-Q plot for linear regression model of 7 taxonomic groups and the Northings

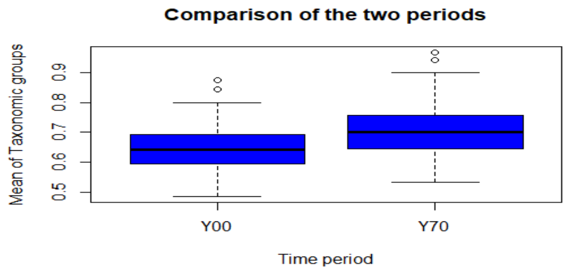


Fig. 10. Comparison of the two time periods

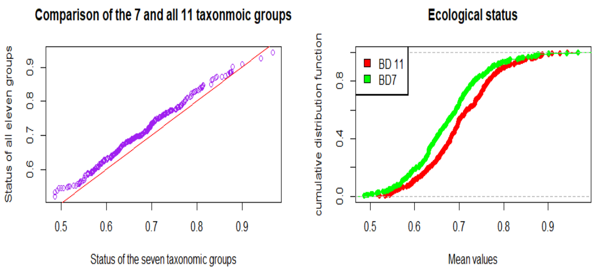
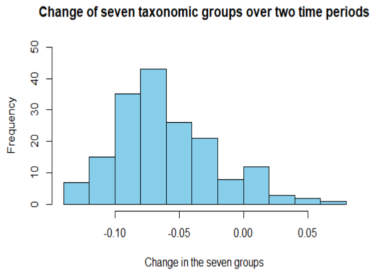
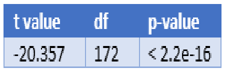


Fig. 12. Q-Q plot of the 7 and 11 taxonomic groups and cdf comparison for Kolmogorov-Smirnov test





t value	df	p-value
-20.357	172	< 2.2e-16

Fig. 11. Distribution of the change in the mean of the seven taxonomic groups between the two time periods and Table 5:Values of the one-sample t-test

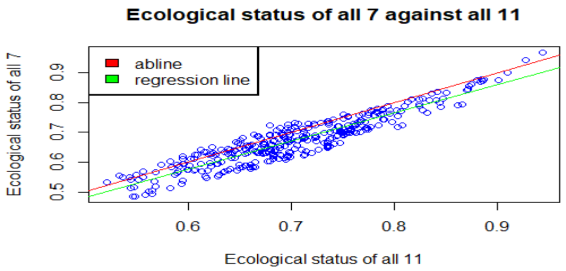


Fig. 13. Linear regression of BD7 and BD11

How many values can vary independently is indicated by the number of degrees of freedom(df). The 95% confidence interval is calculated as [-0.06531853 -0.05377140] and the mean of the sample estimates is -0.05. From the values in Table 5, it is observed that the p-value less than 2.2e-16 which is much less than the significance level of 0.05. Thus the null hypothesis can most certainly be rejected.

2) **Asymptotic two-sample Kolmogorov-Smirnov test** The Q-Q plot in Figure 12 shows that the center of the plot is away from the central line thus suggesting that the two samples are from different populations. However, this is not a test and just a visual representation of the data. [7] The two-sample Kolmogorov-Smirnov test will evaluate the difference between the cumulative distribution function(cdf) of the distribution of the ecological status of all 11 taxonomic groups and the seven taxonomic groups. (Figure 12)
Null Hypothesis HO: The two samples are drawn from the same continuous distribution, suggesting that the mean ecological status of the seven groups must have a distribution function that is equal to that of the eleven taxonomic groups.
Alternative Hypothesis H1: Two-sided, the two samples are from different continuous distributions. D is the maximal distance between the two cdfs seen in Figure 12. From the Kolmogorov-Smirnov test, we get the values of **D = 0.16763** and **p-value = 0.0001198**. The p-value is much smaller than the significance level of 0.05. Thus we can reject the null hypothesis that the two samples have come from the same population.

2.4 Linear regression

In this section, we are going to analyze the linear relationship between the allocated seven taxonomic groups and all of the eleven taxonomic groups present in the dataset. The seven groups denoted as BD7 are the dependent variables and the eleven groups are the independent variables. The ecological status values of BD7 and BD11 are plotted(Figure 13) and the linear regression line is drawn showing a positive linear relationship between the two groups indicating that with the increase in values of the independent variables, the value of the dependent variable also increases. The residuals plot in Figure 14 demonstrates that the values are arbitrarily distributed across the whole range of fitted values, demonstrating that there are no clear patterns in the errors. The Q-Q plot(Figure 14) displays that most of the data points are on a straight line indicating that the data is normally distributed.

We can observe the linear regression of BD7 and BD11 performed independently for the two time periods of 1970–1990 and 2000–2013 in Figure 15. Both time periods show a positive linear relationship. The intercept values for both time periods are different(Table 7 and Table 8) and we can see in Figure 15 that the regression line for the time period 2000-2013 is shifted up on the Y-axis compared to the regression line for the time period 1970-1990. The slope value in Table 7 for the period 1970–1990 is 0.907355, which means that when the mean ecological status value for all 11 taxonomic groups increases by one unit, the mean of the BD7 group increases on average by roughly 0.907355 units (Figure 17).Similar to the previous example, during the time period 2000–2013, the slope is 0.95190 as shown in Table

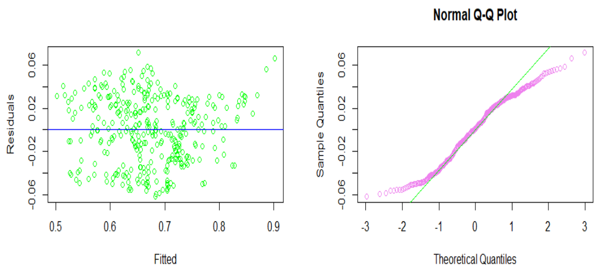


Fig. 14. Residuals plot and Q-Q plot for linear regression of BD7 against BD11

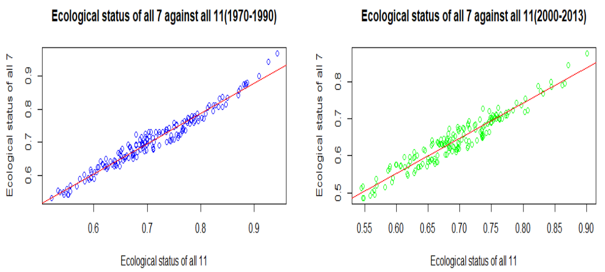
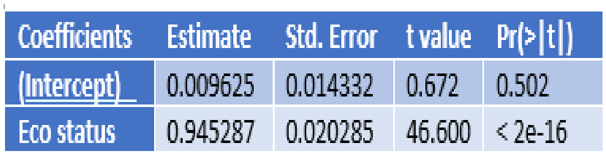


Fig. 15. Linear regression of BD7 and BD11 for the time periods 1970-1990 and 2000-2013

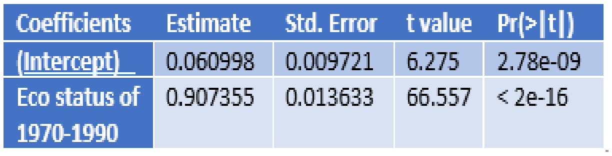
8(Figure 18), suggesting that when the mean ecological status value increases by 1 unit for all 11 taxonomic groups, it raises the mean of the BD7 group on average by roughly 0.95190 units.As a result, for every unit change in the mean of BD11 in the latter era, the mean of BD7 changes by a larger percentage.The slope of 0.945287 is in between the values of the two time periods determined individually and is seen in Table 6's linear regression results for the total time periods for BD7 and BD11.Additionally, the p values are same for both time periods and when combined.

2.5 Multiple Linear Regression



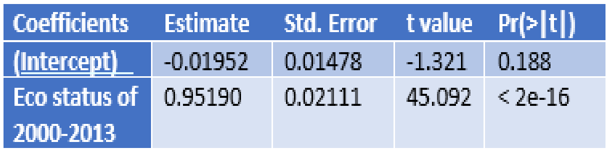
Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.009625	0.014332	0.672	0.502
Eco status	0.945287	0.020285	46.600	< 2e-16

Fig. 16. Table 6:Linear Regression of BD7 and BD11 for both time periods



Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.060998	0.009721	6.275	2.78e-09
Eco status of 1970-1990	0.907355	0.013633	66.557	< 2e-16

Fig. 17. Table 7:Linear Regression of BD7 and BD11 for time period 1970-1990



Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.01952	0.01478	-1.321	0.188
Eco status of 2000-2013	0.95190	0.02111	45.092	< 2e-16

Fig. 18. Table 8:Linear Regression of BD7 and BD11 for time period 2000-2013

Firstly we will interpret the simple linear regression relationship between the remaining four taxonomic groups (BD4) and the allocated seven taxonomic groups(BD7) out of total 11 groups. The seven groups denoted as BD7 are the dependent variables and the four groups are the independent variables. The ecological status values of BD7 and BD4 are plotted(Figure 19) and the linear regression line is drawn showing a positive linear relationship between the two groups indicating that with the increase in values of the independent variables, the value of the dependent variable also increases. The residuals plot in Figure 20 demonstrates that the values are arbitrarily distributed across the whole range of fitted values, demonstrating that there are no clear patterns in the errors. The Q-Q plot(Figure 20) displays that most of the data points are on a straight line indicating that the data is normally distributed.

Next we will perform a multiple linear regression. We will split the data into 80:20 for training and testing purposes respectively.So initially we have seven predictors(BD7) and BD4 as the response variable.We now consider the multiple linear regression model with all 7 predictor variables.We can observe the results for this model in Table 9(Figure 21).We observe that the predictor variables Butterflies,Hoverflies and Vascular plants have p value more than 0.05 .Also the model has an Akaike information criterion(AIC) score of -810.9778 and the correlation value is 0.8054359 between predicted and test data showing a high correlation.Our aim is to improve the performance of the model and get a lower AIC score to justify removal of certain predictors from the model.First we remove the Butterflies predictor since it has the highest p value out of all seven and we see that the AIC score gives down to -812.9672 .

So we decide to remove all three predictors butterflies,hoverflies and vascular_plants with p-value more than 0.05 and again test the model. It receives a low AIC score of -815.402 suggesting it is most likely to be the best model after performing feature selection. The correlation between predicted and test data is 0.8048384 .

Table 10 contains the values of regression coefficients for the final model.The final predictors for the model are Bryophytes,Carabids,Isopods and Grasshopper_Crickets.All the predictor variables have p value less than 0.05 and thus are significant to the model.Only Isopods have a negative slope value indicating a negative linear relationship and the rest of the predictors have positive slope values ranging from 0.26 to 0.35 .

Figure 23 shows a fairly linear relationship between the actual and predicted values,the residuals appear scattered indicating no significant pattern and the center of the Q-Q plot is away from the central line with a few points lying on the sloped reference line suggesting that the residuals of prediction data may not be normally distributed.

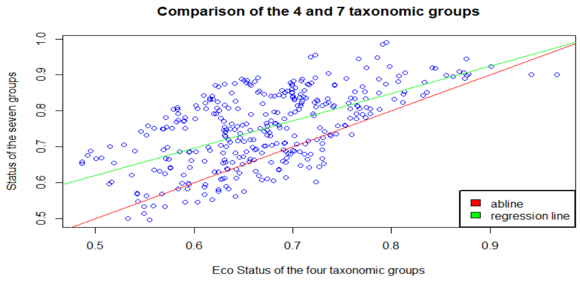


Fig. 19. Linear regression of BD4 and BD7

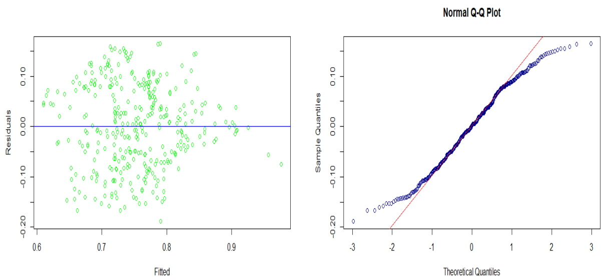
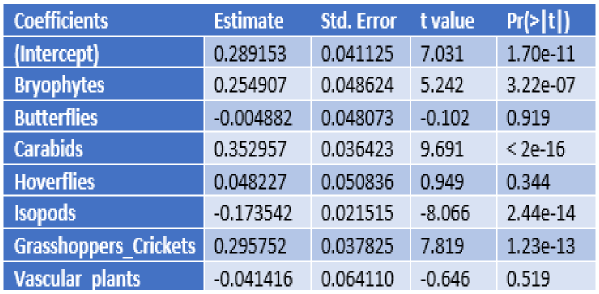
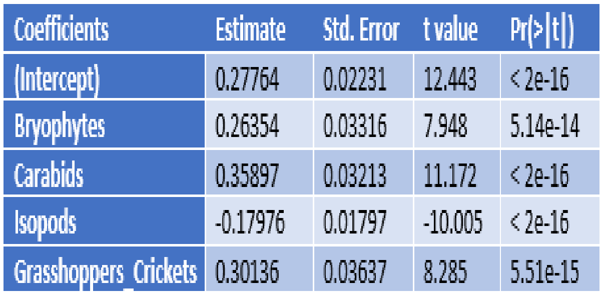


Fig. 20. Residuals plot and Q-Q plot for linear regression of BD4 against BD7



Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.289153	0.041125	7.031	1.70e-11
Bryophytes	0.254907	0.048624	5.242	3.22e-07
Butterflies	-0.004882	0.048073	-0.102	0.919
Carabids	0.352957	0.036423	9.691	< 2e-16
Hoverflies	0.048227	0.050836	0.949	0.344
Isopods	-0.173542	0.021515	-8.066	2.44e-14
Grasshoppers_Crickets	0.295752	0.037825	7.819	1.23e-13
Vascular_plants	-0.041416	0.064110	-0.646	0.519

Fig. 21. Table 9: Multiple regression values for initial model



Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.27764	0.02231	12.443	< 2e-16
Bryophytes	0.26354	0.03316	7.948	5.14e-14
Carabids	0.35897	0.03213	11.172	< 2e-16
Isopods	-0.17976	0.01797	-10.005	< 2e-16
Grasshoppers_Crickets	0.30136	0.03637	8.285	5.51e-15

Fig. 22. Table 10 :Multiple regression values for final model

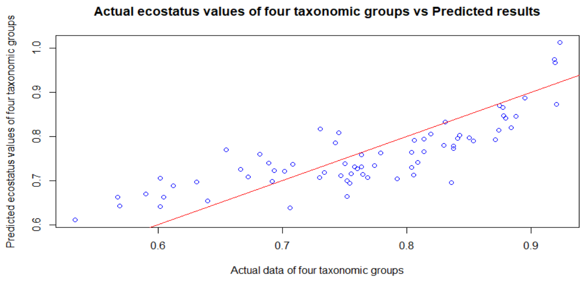


Fig. 23. Actual and Predicted data values of BD4

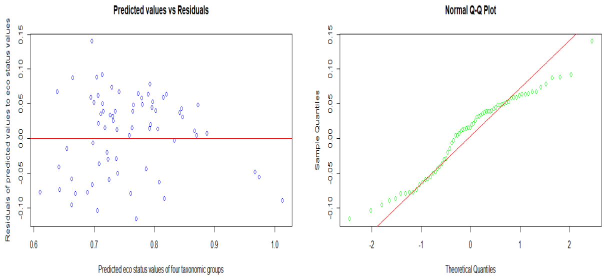


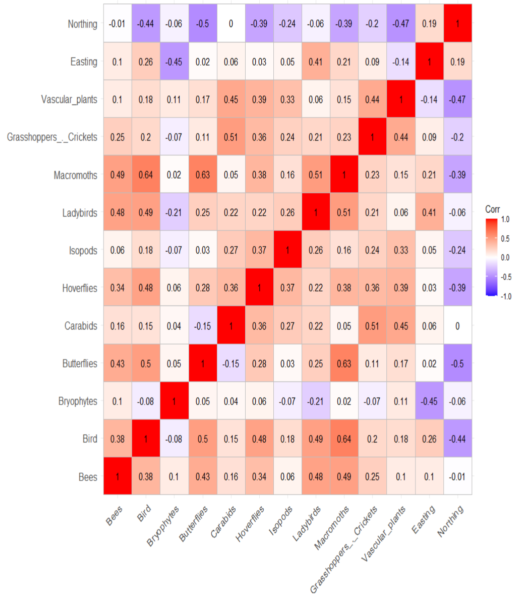
Fig. 24. Residuals plot and Q-Q plot for multiple linear regression model

2.6 Open Analysis

In the open analysis section we will perform a comparison of the eleven taxonomic groups between the two time periods. We have selected all the land classification codes for Scotland for our analysis.

- **Correlation Analysis**: The correlation matrix in Figure 25 shows that most of the 11 groups have a positive correlation with each other.Birds have a negative correlation with Bryophytes,Bryophytes have a negative correlation with Isopods,Ladybirds and Grasshopper_Crickets.Butterflies have negative correlation with Carabids. The highest positive correlation value is of 0.64 between Bird and Macromoths, and 0.63 between Butterflies and Macromoths.
- **Ecological status value of 11 taxonomic groups versus Eastings and Northings** : The image in figure 25 shows the species mean status values across Eastings and Northings.We have a few outliers for high values of Northings.Other than that most of the data is concentrated below 1100000 Northings.
- **Box plot comparison of the two periods**: The boxplot comparisons(Figure 26) demonstrate that the median value for the older time period is higher than that for the current time period.Few outliers exist for both time periods.The range of values appears to be the same for both times, though.
- **Hypothesis Testing**

  1) **One Sample t-test** : Prior to doing the one-sample t-test, we first calculate the change in the mean of the 11 taxonomic categories across the two time periods of 1970–1990 and 2000–2013 (Figure 27). Then we will define the



	Bees	Bird	Bryophytes	Butterflies	Carabids	Hoverflies	Isopods	Ladybirds	Macromoths	Grasshoppers_._Crickets	Vascular_plants	Easting	Northing
Northing	-0.01	-0.44	-0.06	-0.5	0	-0.39	-0.24	-0.06	-0.39	-0.2	-0.47	0.19	1
Easting	0.1	0.26	-0.45	0.02	0.06	0.03	0.05	0.41	0.21	0.09	-0.14	1	0.19
Vascular_plants	0.1	0.18	0.11	0.17	0.45	0.39	0.33	0.06	0.15	0.44	1	-0.14	-0.47
Grasshoppers_._Crickets	0.25	0.2	-0.07	0.11	0.51	0.36	0.24	0.21	0.23	1	0.44	0.09	-0.2
Macromoths	0.49	0.64	0.02	0.63	0.05	0.38	0.16	0.51	1	0.23	0.15	0.21	-0.39
Ladybirds	0.48	0.49	-0.21	0.25	0.22	0.22	0.26	1	0.51	0.21	0.06	0.41	-0.06
Isopods	0.06	0.18	-0.07	0.03	0.27	0.37	1	0.26	0.16	0.24	0.33	0.05	-0.24
Hoverflies	0.34	0.48	0.06	0.28	0.36	1	0.37	0.22	0.38	0.36	0.39	0.03	-0.39
Carabids	0.16	0.15	0.04	-0.15	1	0.36	0.27	0.22	0.05	0.51	0.45	0.06	0
Butterflies	0.43	0.5	0.05	1	-0.15	0.28	0.03	0.25	0.63	0.11	0.17	0.02	-0.5
Bryophytes	0.1	-0.08	1	0.05	0.04	0.06	-0.07	-0.21	0.02	-0.07	0.11	-0.45	-0.06
Bird	0.38	1	-0.08	0.5	0.15	0.48	0.18	0.49	0.64	0.2	0.18	0.26	-0.44
Bees	1	0.38	0.1	0.43	0.16	0.34	0.06	0.48	0.49	0.25	0.1	0.1	-0.01

Fig. 25. Correlation matrix of the 11 taxonomic groups,Eastings, and Northings

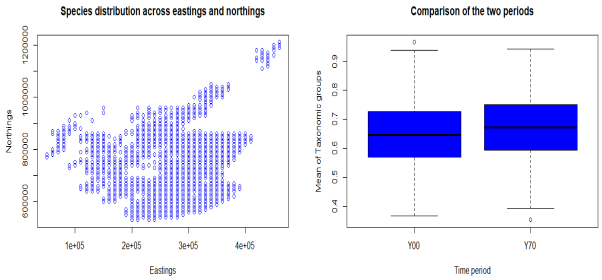


Fig. 26. BD11 for Eastings and Northings and Box plot comparison of the two time periods

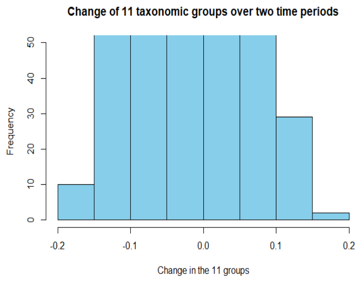
question and the null and alternate hypotheses.

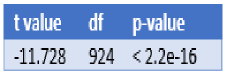
Question: Is the mean value of the sample significantly larger (or smaller) than the mean value of the population?

Null Hypothesis HO: True mean is equal to zero Alternative hypothesis H1: True mean is not equal to 0

How many values can vary independently is indicated by the number of degrees of freedom(df). The 95% confidence interval is calculated as [-0.02660037 -0.01897419] and the mean of the sample estimates is -0.02278728 . From the values in Table 11, it is observed that the p-value less than 2.2e-16 which is much less than the significance level of 0.05. Thus the null hypothesis can most certainly be rejected.

2) **Asymptotic two-sample Kolmogorov-Smirnov test** The Q-Q plot in Figure 28 shows that the center of the plot is away from the central line thus suggesting that the two samples are from different populations.





t value	df	p-value
-11.728	924	< 2.2e-16

Fig. 27. Distribution of the change in the mean of the 11 taxonomic groups between the two time periods and Table 11:Values of the one-sample t-test

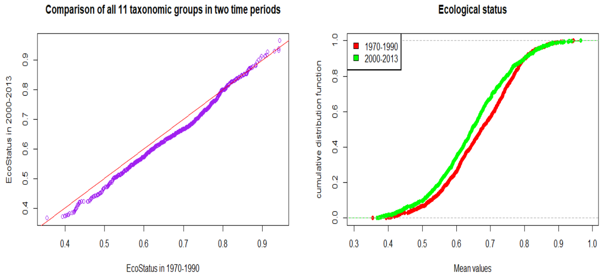


Fig. 28. Q-Q plot of the 11 taxonomic groups and cdf comparison across two time periods for Kolmogorov-Smirnov test

The two-sample Kolmogorov-Smirnov test will evaluate the difference between the cumulative distribution function(cdf) of the distribution of the ecological status of all 11 taxonomic groups in 1970-1990 and 2000-2013. (Figure 28)

Null Hypothesis HO: The two samples are drawn from the same continuous distribution, suggesting that the mean ecological status of the 11 groups must have a distribution function that is equal in both time periods. Alternative Hypothesis H1: Two-sided, the two samples are from different continuous distributions.

D is the maximal distance between the two cdfs seen in Figure 28. From the Kolmogorov-Smirnov test, we get the values of **D = 0.11568** and **p-value = 8.426e-06**. The p-value is much smaller than the significance level of 0.05. Thus we can reject the null hypothesis that the two samples have come from the same population.

- **Simple linear regression of the two time periods** The linear relationship between the two time periods for the eleven taxonomic groups will then be examined. The linear regression line between the two groups has a positive linear association (Figure 29), showing that the mean ecological status value rises linearly during both time periods. There are no apparent patterns in the residuals, as shown by the residuals figure in Figure 30. The majority of

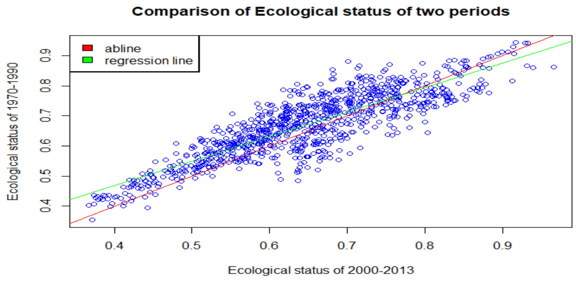


Fig. 29. Linear Regression of BD11 for both time periods

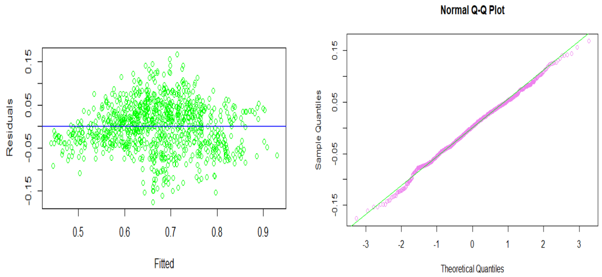
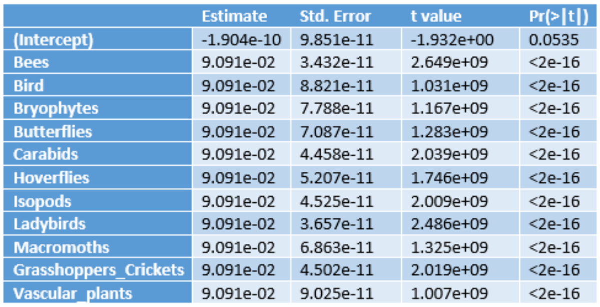


Fig. 30. Residuals plot and Q-Q plot for simple linear regression model

the data points on the Q-Q plot (Figure 30) are on a straight line, which indicates that the data are normally distributed.

- **Multiple Linear Regression** To create the model, we select 80% of the training data from Scotland and 20% of the Test data from England to build the model. There are 11 predictors(BD11) and the mean ecological status value is the response variable. In Table 12 (Figure 31), the coefficient values for this model are displayed. The p values for each predictor are all under 2e-16 and the slope values are also the same. Additionally, the model's Akaike information criterion (AIC) score is -60704.58 and the correlation between predicted and test data is 1, demonstrating a perfect correlation. So, it's fascinating that when we use training and test data from two separate land classifications, we end up with an almost perfect model! We have also included the plots for the relation between the actual values and predicted values for the model in Figure 32 showing the data to be almost in a perfectly linear relationship and the Q-Q plot showing normalized data.(Figure33)

  Multiple linear regression of the 11 taxonomic groups against the time periods, Eastings and Northings: We construct a regression model for the BD11 and time period, Eastings and Northings. Table 12 lists the coefficients for the model. The model's AIC score is -883.5546. We eliminate the predictor time period because it has a p-value of 0.124656, which is higher than 0.05, and evaluate the model's AIC score once more; this time, it reads -883.1657. Since the model's AIC score is unchanged, we choose to keep the predictor time period in place, and our original model is chosen as the final model.



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.904e-10	9.851e-11	-1.932e+00	0.0535
Bees	9.091e-02	3.432e-11	2.649e+09	<2e-16
Bird	9.091e-02	8.821e-11	1.031e+09	<2e-16
Bryophytes	9.091e-02	7.788e-11	1.167e+09	<2e-16
Butterflies	9.091e-02	7.087e-11	1.283e+09	<2e-16
Carabids	9.091e-02	4.458e-11	2.039e+09	<2e-16
Hoverflies	9.091e-02	5.207e-11	1.746e+09	<2e-16
Isopods	9.091e-02	4.525e-11	2.009e+09	<2e-16
Ladybirds	9.091e-02	3.657e-11	2.486e+09	<2e-16
Macromoths	9.091e-02	6.863e-11	1.325e+09	<2e-16
Grasshoppers_Crickets	9.091e-02	4.502e-11	2.019e+09	<2e-16
Vascular_plants	9.091e-02	9.025e-11	1.007e+09	<2e-16

Fig. 31. Table11:Coefficient values for the model

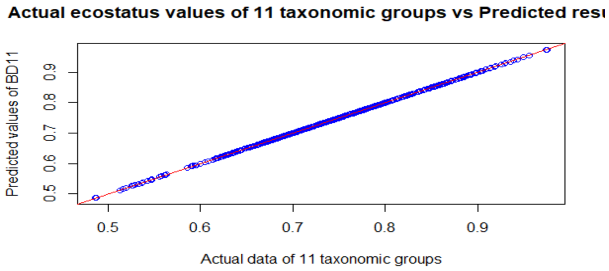


Fig. 32. Actual and Predicted values

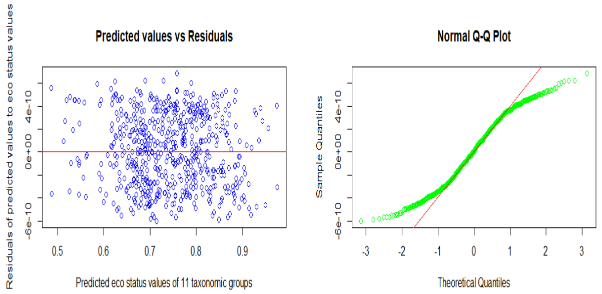
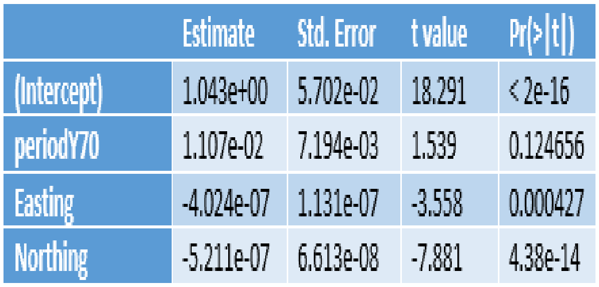


Fig. 33. Residuals plot and Q-Q plot for multiple linear regression model



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.043e+00	5.702e-02	18.291	< 2e-16
periodY70	1.107e-02	7.194e-03	1.539	0.124656
Easting	-4.024e-07	1.131e-07	-3.558	0.000427
Northing	-5.211e-07	6.613e-08	-7.881	4.38e-14

Fig. 34. Table 12: Coefficient values of Multiple linear regression of BD11 against time period,Eastings,Northings

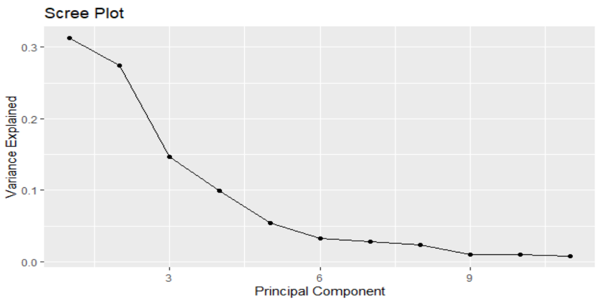


Fig. 35. Scree plot

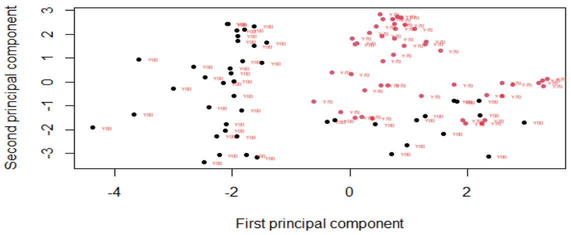


Fig. 36. Scatter plot for land classification 7w

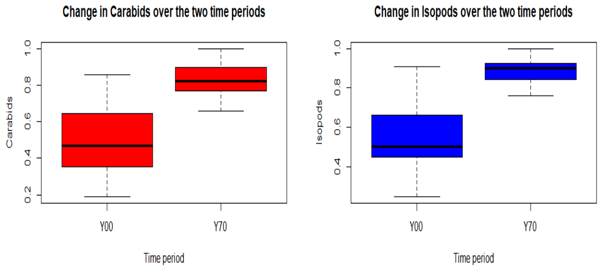


Fig. 37. Box plot comparison of Carabids and Isopods for two time periods for 7w

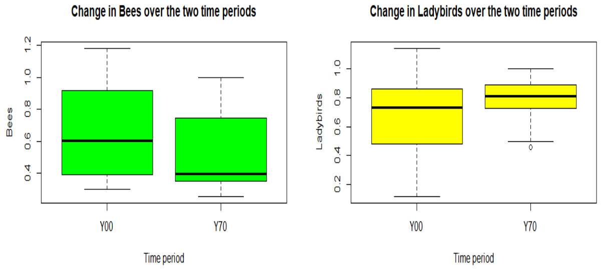


Fig. 38. Box plot comparison of Bees and Ladybirds for two time periods for 7w

- **Principal Component Analysis(PCA)**: We will use Principal Component Analysis or PCA to reduce the dimensionality of the dataset of eleven taxonomic groups into a smaller dataset. We have chosen a set of three different land classifications for our analysis which are Sea Cliffs/Hard Coast, Wales(7w), Sea Cliffs/Hard Coast, England(7e), which are similar, and Complex valley systems/Table Lands, Wales(6w), which is entirely different than the other two. The aim is to study the variations in the mean ecological status of the 11 groups for the two time periods and determine how they affect both comparable and different land classifications.
  - **Sea Cliffs/Hard Coast, Wales(7w)**: First, we construct the scree plot(Figure 35) for the 11 components and find that most of the data can be selected by using the first two components. Next, we create a scatter plot(Figure 36), the red dots are labeled by the time period Y70 and the black dots are for the time period Y00.It shows that the points moved from right to left indicating a decrease in values from the older to the newer time period.

    Box plots were used to track the change in some of the species over the two time periods. Carabids, Isopods(Figure 37), and ladybirds(Figure 38) decreased from the later to the new time period which is not a good sign for the species. Only bees(Figure 38) were found to increase in the new time period.
  - **Sea Cliffs/Hard Coast, England(7e)**: Similarly, we create a scatter plot for the first two principal components(Figure 39) where the pattern shows the points of the new time period moving up from the old time period data points. However, some values moved to the left indicating a decrease in values, while some data points are found to the right indicating an increase in values for some species. From the box plots, we observe that unfortunately in this land classification also the carabids and isopods have decreased(Figure 40) over time. However, there has been an increase in bees and ladybirds. (Figure 41)
  - **Complex valley systems/Table Lands, Wales(6w)** The scatter plot for land classification 6w(Figure 43) does not exhibit a good trend following a top-down pattern from the later years to the newer time period of 2000-2013. On the second principal component, all of the data points for the years 2000 to 2013 have decreased in value, and while some of them are still on the right for

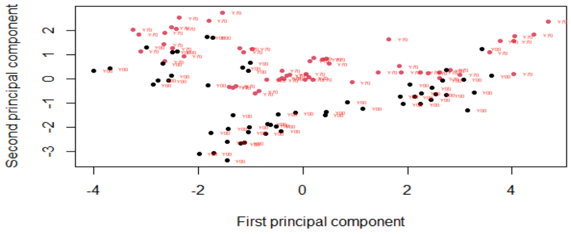


Fig. 39. Scatter plot for land classification 7e

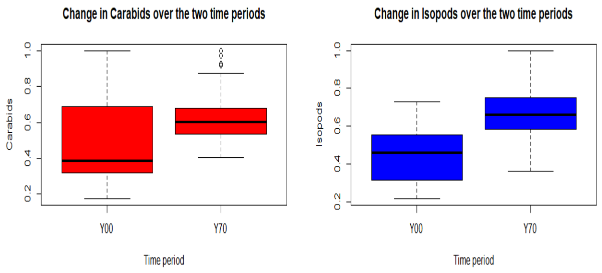


Fig. 40. Box plot comparison of Carabids and Isopods for two time periods for 7e

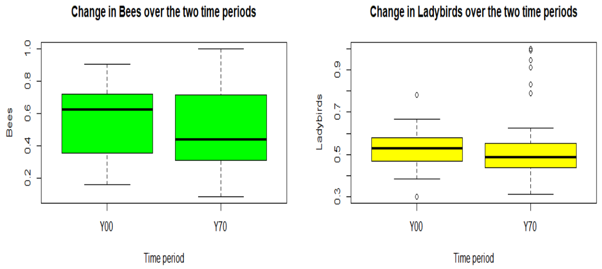


Fig. 41. Box plot comparison of Bees and Ladybirds for two time periods for 7e

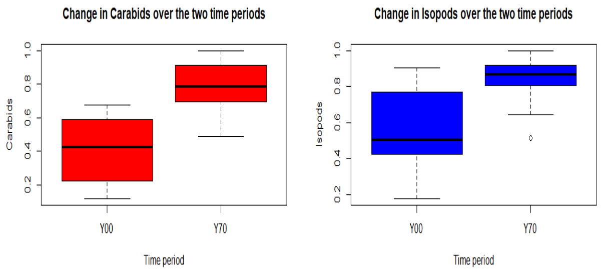


Fig. 43. Box plot comparison of Carabids and Isopods for two time periods for 6w

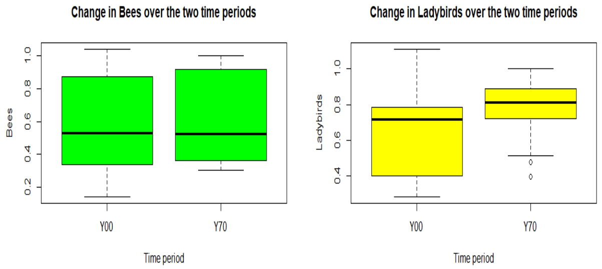


Fig. 44. Box plot comparison of Bees and Ladybirds for two time periods for 6w

the current period, the majority of them have migrated to the left, showing a general decline in values. The carabids and isopods have decreased(Figure 43) in this land classification too, and the bees did not increase and the ladybirds also decreased. (Figure 44)

3 Discussions and Conclusion

Various types of analysis have been conducted on the data set of the 11 taxonomic groups for some of the land classification codes selected from the list made available to us, the results of which have been discussed at length in our report. We see a general decline in the mean ecological status values for the taxonomic categories over the two time periods, indicating that the shale gas extraction sites have had a detrimental influence on biodiversity. In several land classes, certain species have declined, including carabids and isopods. While declining in some locations, ladybird populations have increased in others. Most of the time, we observe an upward tendency for bee populations, however, in other land classes, this trend is negative. In conclusion, we observe a loss in species richness for these 11 taxonomic categories and believe that conservation efforts for diverse biodiversity and these smaller species, which also have a significant impact on our ecology, should receive more attention.

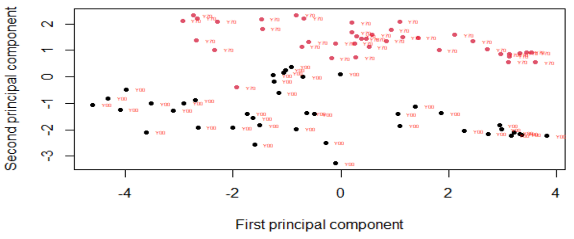


Fig. 42. Scatter plot for land classification 6w

References

[1] Roel Slootweg and Arend Kolhoff. A generic approach to integrate biodiversity considerations in screening and scoping for eia. *Environmental Impact Assessment Review*, 23(6):657–681, 2003.

[2] Vinod Mathur and Asha Rajvanshi. Integrating biodiversity into environmental impact assessment. 01 2012.

[3] Robert Dyer, Simon Gillings, Richard Pywell, Richard Fox, D.B. Roy, and Tom Oliver. Developing a biodiversity-based indicator for large-scale environmental assessment: A case study of proposed shale gas extraction sites in britain. *Journal of Applied Ecology*, 54, 09 2016.

[4] T. Dyer, R.;Oliver. Uk ecological status map version 2, 2016.

[5] Grid reference finder, os getoutside, mar. 2023, [online].

[6] C.J.;Clarke R.T.;Howard D.C.;Scott W.A. Bunce, R.G.H.;Barr. Ite land classification of great britain 2007, 2007.

[7] G. Upton and D. Brawn. In *Data Analysis: A Gentle Introduction for Future Data Scientists.*, 2023.