MA334 FINAL PROJECT

REGN NO: 2311465

INTRODUCTION

In the final project for MA334 we will be exploring the change in biodiversity of 11 different taxonomic groups in the UK across two periods (1970 and 2000). The data varies according to location and with the different dominant land classes throughout the UK which is segmented using the UK National Grid Scale. Out of the 11 different taxonomic groups available to us, we will be analyzing 5 which are:

- Bird
- Bryophytes
- Hoverflies
- Isopods
- Ladybirds

Our analysis will include proportional species richness of each group, their correlation with each other, two separate hypothesis tests, reporting the p-values and their interpretation. Additionally, we will also be applying regression models to deduce the relationship between variables and to access their predictive powers. Finally, we will be conducting an open analysis on the change of biodiversity in mountainous and the coastal regions of the UK over the two time periods.

UNIVARIATE ANALYSIS

Univariate analysis technique is used to explore each variable in the data set independently and give meaningful insights regarding that particular variable in relation to the whole dataset by looking atthe range of values and the central tendencies.

The 7 basic summary statistics

Tax Group	Mean	Win Mean	SD	Min	Max	1st Quantile	Median	3rd Quantile
Bird	0.89	0.90	0.11	0.24	1.17	0.85	0.90	0.96
Bryophytes	0.79	0.79	0.13	0.39	1.17	0.69	0.80	0.89
Hoverflies	0.68	0.69	0.18	0.12	1.15	0.57	0.70	0.81
Isopods	0.55	0.55	0.22	0.05	1.26	0.39	0.54	0.72
Ladybirds	0.61	0.61	0.27	0.06	1.84	0.45	0.64	0.8

Table 1: Univariate Analysis

Table 1 shows the result of the univariate analysis for each of the 5 taxonomic groups. Some interesting aspects of the table are as follows: -

a. The mean environmental status for the groups is between 0.55 to 0.89 with Isopods the lowest and Bird the highest. A similar trend can be seen in the Winsorized Mean (Win Mean) which ranges from 0.55 (lowest for Isopods) to 0.90 (highest for Bird)
b. The Standard Deviation (SD) for the taxonomic group ranges from 0.11 (Bird having the lowest variability) to 0.27 (Ladybirds having the highest variability)

Correlation Matrix

A correlation matrix is used to display the pairwise relation between every element of the 5 taxonomic groups. It ranges from -1 to +1 where the higher the value the better the correlation is.

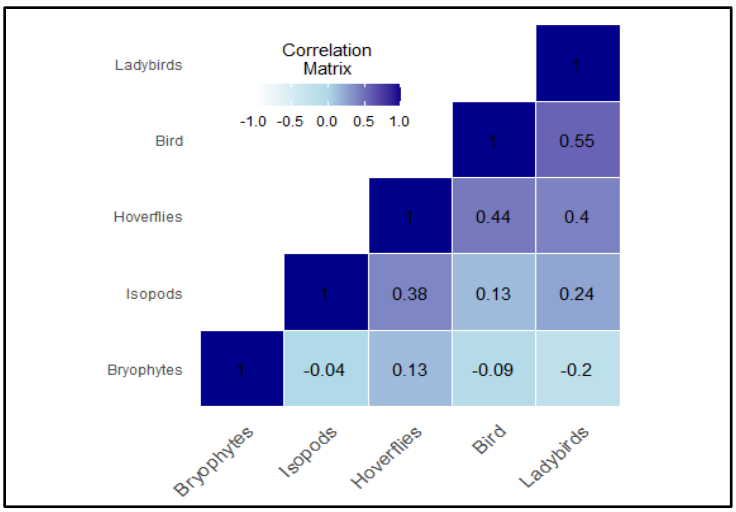

Figure 1: Correlation Matrix of 5 taxonomic Groups

From the pairwise correlation matrix given in Figure 1, we can draw the following useful insights: -

a. There is a moderate positive correlation (0.55) between Ladybirds and Birds which means that any area containing a high number of Ladybirds is also expected to have a higher diversity of Birds. There are also a similar moderate positive correlation values between Hoverflies and Birds (0.44) and Hoverflies and Ladybirds (0.4) indicating same pattern.
b. There is a low positive correlation between the species of Isopods and Birds (0.13) and Bryophytes and Hoverflies (0.13) which suggests that the areas containing a high number of Isopods or Bryophytes would also have a slightly diverse colony of Birds or Hoverflies respectively.
c. There are net weak negative correlations between Ladybirds and Bryophytes (-0.2) indicating that areas with high number of Ladybirds is likely to have lower diversity of Bryophytes.
d. There are overall weak negative correlations between Bird and Bryophytes (-0.09) and Isopods and Bryophytes (-0.04) meaning that a higher diversity of Birds or Isopods in areas will lead to a slightly lower diversity of Bryophytes respectively.

Box Plot

The boxplot shown is of the taxonomic group Isopods. As is visible in the boxplot, the overall diversity value for isopods has decreased in the UK from year 1970 to the year 2000 represented by the shift in median from 0.65 in 1970 to 0.4 in the year 2000 with the IQRs remaining the same. However, there are a few outliers in the year 2000 with the max value of 1.2 indicating slight abnormalities in that year.

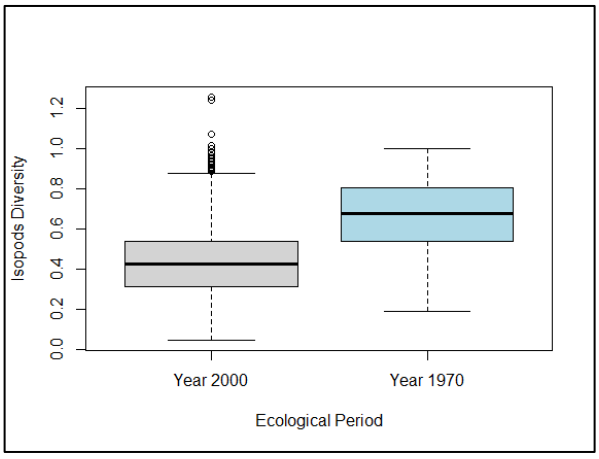

Figure 2: Box plot of Isopods

HYPOTHESIS TESTS

We will do a total of two hypothesis tests on our given 5 taxonomic groups. The tests to be conducted are the following: -

- T – Test
- KS Test

T – Test

A T – test on our eco_status_5 variable will give the difference between the biodiversity measure in the year 2000 and the year 1970 (Y00 – Y70). The null hypothesis of this test states that the population mean change in the BD5 will be equal to 0. Subsequently, the alternate hypothesis will be that the mean change in the BD5 is not equal to 0. Upon performing the T – Test we get the following results in R-Studio.

One Sample t-test
data: BD5_change
t = -34.183, df = 2639, **p-value < 2.2e-16**
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval: -0.05011487 -0.04467718
sample estimates:
mean of x
-0.04739602

Table 2: Output of T – Test from R-Studio

Based on this result in table 2, the p – value < 2.2 e-16 and are considered to be equal to zero the t – value of -34.183 is also insignificant. So, in this case the T – Test proves that the null hypothesis is rejected and the alternative hypothesis stands true.

KS – Test

The Kolmogorov-Smirnov test is used to test the null hypothesis that the sample data set comes from a normally distributed set of values. The values range 0.0 to 1.0 and the higher the vale the better the fit of the data set into a normally distributed set of values. Using the ECDF plots for both the BD5 data set (denoted in green) and the BD11 data set (denoted in red), we can see that the means of both groups differs slightly.

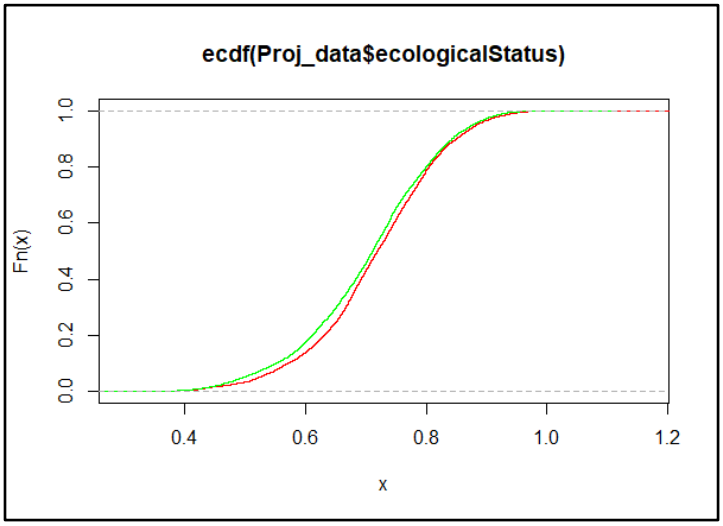

Figure 3: CDFs for BD5 and BD11

Asymptotic two-sample Kolmogorov-Smirnov test

data: Proj_data$eco_status_5 and Proj_data$ecologicalStatus
D = 0.055492, p-value = 1.737e-07
alternative hypothesis: two-sided

Table 3: Output of KS – Test from R-Studio

Running the KS – Test in R-Studio, we can see that the p – value is equal to zero so the null hypothesis is rejected and we can confidently say that these are two separate distributions. Moreover, the D value of 0.0554 indicates that there is a vertical distance between the ECDF of the sample data set and the CDF of the population data set further confirming our claim.

CONTINGENCY TABLES AND COMPARISON VARIABLES

Creating a Contingency Table is helpful in analyzing the frequency distribution of two categorical variables. In our case the variables are BD5 and BD11. We want to see whether both the variables have seen an increase or a decrease over the time period from 1970 to 2000.:

BD11 / BD5	Contingency Table			Independent Model		
	Decrease	**Increase**	**Total**	**Decrease**	**Increase**	**Total**
Decrease	1466	172	1638 (0.62)	1258.6	106.64	1365.24
Increase	564	438	1002 (0.37)	751.1	162.06	913.16
Total	2030	610	2640	2009.7	268.7	2278.4

Table 4: Contingency Table & Independent Model of BD11 / BD5

From Table 4 we can draw the following conclusions:

- The observed values in the contingency table do not form a close association with the expected counts in the independent model table as there are more cases of both BD11 and BD5 increasing (438) compared to the expected count in the independent model (162.06) suggesting a potential positive correlation between the increases in both BD11 and BD5.
- Alternatively, there are fewer instances of BD11 increasing and BD5 decreasing (564) compared to the expected count in the independent model (751.5) suggesting a potential negative correlation between both values.

Likelihood – Ratio Statistic

To find the Likelihood – Ratio Statistic we will perform a G-Test in R-Studio on the Contingency In our case we get the following result once the G-Test is performed:

Log likelihood ratio (G-test) test of independence without
correctiondata: Contingency_Table

G = 380.39, X-squared df = 1, p-value < 2.2e-16

Table 5: Result of G-Test from R-Studio

The larger the G-test value the stronger the evidence there is to reject the null hypothesis so in our case the value of 380.39 gives a very strong reason for us to reject the hypothesis, furthermore a p-value equivalent to zero also gives a strong reason for the null hypothesis to be rejected. This result gives the conclusion that the observed data is unlikely to be under the assumption of independence.

Odds Ratio, Sensitivity, Specifity and Youden's Index

Calculating the given parameters in R-Studio for the Contingency Table gave us the values as expressed in Table 5.

- The Odds ratio gives a measure of associativity between BD11 and BD5. A positive association can be seen with an odds ratio of 6.619 which means that the odds of seeing both BD11 and BD5 increasing are 6.6 times more than the odds of both of them decreasing at the same time.
- The sensitivity, also called the true positive rate gives a measure of all the correctly identified positive values. In our context a value of 0.43 shows that the model is correct in predicting an increase in BD11 43% of the time.

Parameter	Value
Odds Ratio	6.619124
Sensitivity	0.4371257
Specifity	0.8949939
Youden's Index	0.3321196

Table 6: Comparison values

- Specifity is the opposite of Sensitivity such that it tries to correctly identify the total number of negatives. In our dataset a value of 0.89 shows that the model is correct in predicting a decreasing value of BD11 89% of the time.
- Youden's Index is a combined statistic which gives insight to the performance of the model as a whole as it combines both the values of Sensitivity and Specifity. It ranges from -1 to +1 and the higher the value the better the performance of the model. In our case the value 0.33 suggests a moderately positive performance of the model.

SIMPLE LINEAR REGRESSION

In our analysis we used the taxonomic group "Carabids" as our BD1 and applied the linear regression model on it against the original 5 taxonomic groups BD5. The output of the linear regression model shows that the model is highly significant (indicated by the p-value and the value of the F- statistic). The R squared value of 0.311 represents a good fit of the model and the coefficients show the relationship between the predictor variable (BD1) and the response variable (BD5).

Call:

lm(formula = Proj_data_all_11$Carabids ~ Proj_data$eco_status_5)

Residuals:

Min	1Q	Median	3Q	Max
-0.84625	-0.09988	0.02756	0.12671	0.50902

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.14682	0.01564	-9.388	<2e-16	***
Proj_data$eco_status_5	1.07169	0.02196	48.807	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1784 on 5278 degrees of freedom

Multiple R-squared: 0.311, Adjusted R-squared: 0.3108

F-statistic: 2382 on 1 and 5278 DF, p-value: < 2.2e-16

Table 7: Output of Simple Linear Regression from R-Studio

The residuals represent the difference between the predicted values and the observed values after applying the model and since they are near to zero this indicates a good fit for the model. They have been plotted in the qq-plot having a straight line indicating the residuals form a part of the normal distribution The following results were obtained on R-Studio after applying the linear regression model. Figure 4

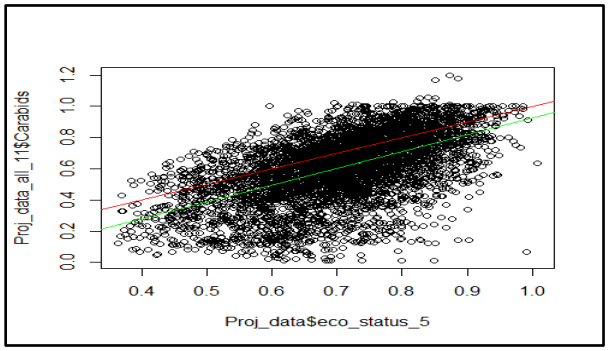

Figure 4: LM on BD1 vs BD5

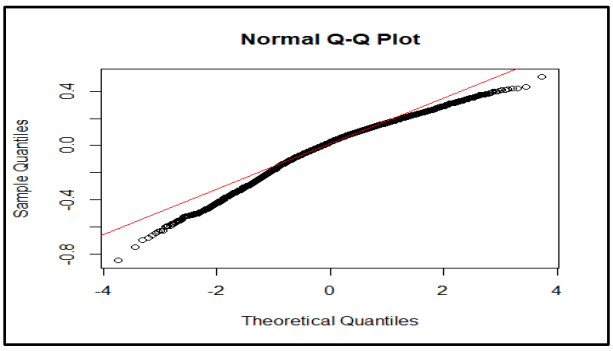

Figure 5: Normality QQ Plot

represents a scatter plot between BD5 and BD1 while the red and green lines represent the best fit line on the BD5 and the BD1 datasets respectively. As is visible, there is a positive linear relationship between both the data sets hence an increment of one will lead to an increment of the other. From Table 4, we can see that the coefficient of that increment will be 1.07. Figure 5 is a QQ-plot that assess the residual distribution's normality. A probability plot is generated to represent the actual values of BD1 and then an ideal line is pasted on the plot to represent the expected values. The actual values are either on top or nearabout the expected outcome showing that the residuals are distributed approximately normally with slight deviations. In summary, it is safe to assume that the residuals are relatively well approximated by a normal distribution.

MULTIPLE LINEAR REGRESSION

Call:

lm(formula = Proj_data_all_11$Carabids ~ ., data = Proj_data[c(eco_selected_names)], y = TRUE)

Residuals:

Min	**1Q**	**Median**	**3Q**	**Max**
-0.92442	-0.09206	0.01380	0.11335	0.54190

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.21691	0.02604	8.328	<2e-16	***
Bird	**0.02283**	**0.02760**	**0.827**	**0.408**	
Bryophytes	-0.17775	0.01873	-9.490	<2e-16	***
Hoverflies	0.39832	0.01646	24.194	<2e-16	***
Isopods	0.24359	0.01192	20.429	<2e-16	***
Ladybirds	0.17115	0.01118	15.315	<2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.17 on 5274 degrees of freedom

Multiple R-squared: 0.3748, Adjusted R-squared: 0.3742

F-statistic: 632.4 on 5 and 5274 DF, p-value: < 2.2e-16

Table 8: Multiple Linear Regression on Carabids vs 5 Taxonomic Groups

Table 8 represents the output of a multiple linear regression model applied on BD1 vs the BD5. All the taxonomic groups except for the Bird (p value 0.408) are statistically significant and the the R-Squared value of 0.37 represents that 37% of the variation in the response variable (Carabids) can be successfully explained by the model. Furthermore, the F-Statistic of 632.4 is highly significant and paired with a zero p-value gives the conclusion that the model is statistically significant in explaining the response variable.

Feature Selection and AIC Models

We will now remove the feature **"Bird"** from the dataset and run a reduced LM on the data. The justification of removing the feature is that it is the only non-significant group so we can drop it based off of its p-value. After dropping the variable, we see an improvement of F-statistic of (762.4) which is higher than the previous value of 632.4, showing that this model is a better fit for explaining the behavior of our BD1. Furthermore, the AIC and DF values before and after removing a predictor variable are given in table 9. These further justify the removal of the feature variable as correct as they lead to better results. Now, using the two predictor variables of "Ladybirds" and "Bryophytes" based on the least correlation between them (from the correlation matrix given in figure 1) we created an interaction model and got the lowest possible value for the AIC, resulting in the best fit model out of the three. This is due to the fact that when variables are weakly correlated, it results in a model which is not overly complex and has fewer parameters. It also removes the issue of "overfitting" where the model fits the data too closely and can capture noise.

	Degree of Freedom	AIC Value
Original Linear Model	7	-3718.119
Reduced Linear Model	6	-3719.434
Interaction Linear Model	8	-3772.693

Table 9: DF and AIC Values of LM, LM Reduced and Interaction LM

Mean Square Error Test

Dividing the data set into two parts we get the following distribution:

- Y70 as the training set
- Y00 as the test set

The scatter plot represents the relation between the Test value Y00 (independent value) and the predicted values of Y00. Applying the best fit line on the dataset we get the plot as shown in Figure 6. As is visible many of the data points are lying on the best fit line or are slightly around it. This shows that the model is moderately accurate in its prediction when compared to the actual values. The points that are above the best fit line indicate that the actual values are higher than the predicted values. There are a few outliers as well around the 0.2 and less mark which can be looked into further to determine their cause.

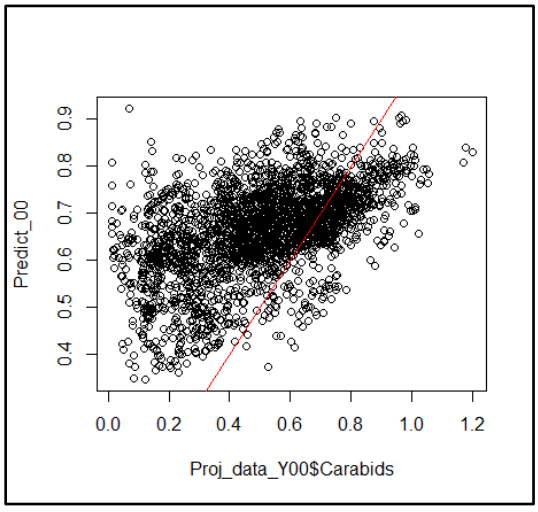

Figure 6: Test vs Predicted Scatter Plot

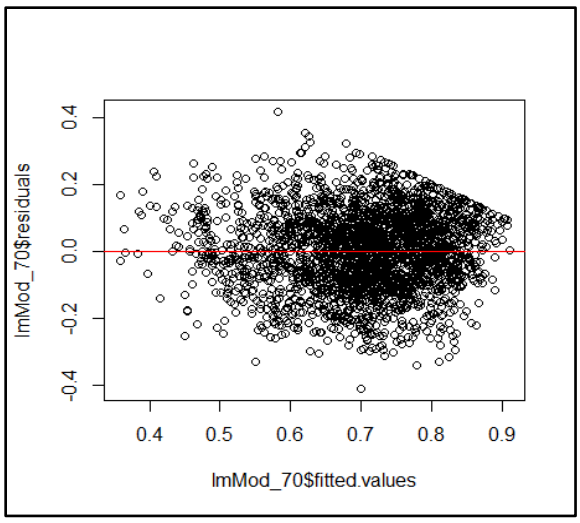

Figure 7: Residuals Scatter Plot

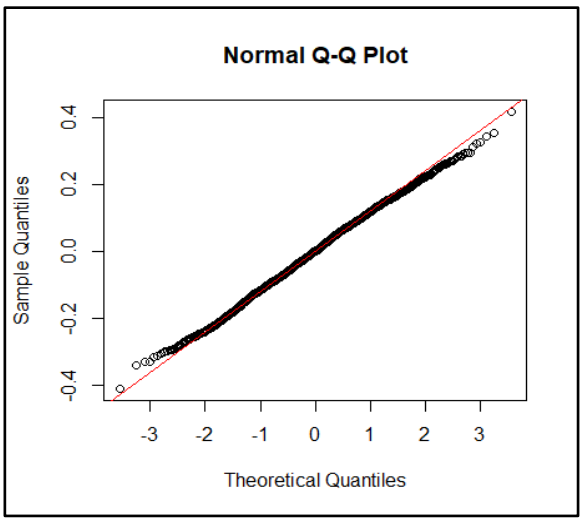

Figure 8: Normal Distribution QQ Plot

Figure 7 and Figure 8 is looking at the residuals of the model. The scatter plot shows that approximately 80% of the values fall under the range of 0.5 – 0.9 which means that the model was successful in capturing the underlying patterns in the data. This is further strengthened by the QQ-plot in figure 8 where we can see that the majority of the data points are lying on the red best fit line indicating a near perfect normal distribution of data. Performing the Mean Square Error tests on both the Test and the Training Data yields the results in Table 10. It can be seen that the MSE values for the test set are higher than those of the training set due to the fact that the model may be suffering from overfitting where the model struggles to generalize to new and unseen data. Moreover, the model may also struggle to account for the ecological changes between the two eras.

	MSE Value
Test Set	0.06047982
Training Set	0.01345989

Table 10: Mean Square Error Values

OPEN ANALYSIS SECTION

For the open analysis section, we wanted to check the effect of biodiversity change over both the hilly/mountainous regions and the coastal areas of the UK. To achieve that we first had to make a table of the dominant land classes from the paper's extra resources and them make a list of all the land classes pertaining to mountains and to the coasts. We came up with the following two sets of data:

- Mountains – '18e','22e','23e','18s','19s','21s','22s','23s','24s','17w1','17w2','17w3','18w'
- Coasts – '7e','8e','7s','29s','30s','31s','32s','7w'

Plotting them next to each other as shown in figures shows a quite diverse result as the coastal areas seem to have suffered a decline in the biodiversity measures, hitting a peak at 0.5 in the year 1970. This is totally opposite to the trend followed by BD5 values in the Mountainous regions of the UK where the values showed an increase from the year 1970 to the year 2000, hitting a peak at around 0.4 in the year 2000. This shows that the taxonomic groups have thrived in these areas since the turn of the century while they are seeing a decline in the coastal regions.

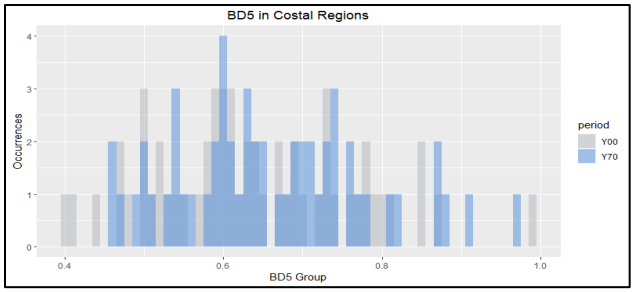

Figure 9: BD5 values in Coastal Regions

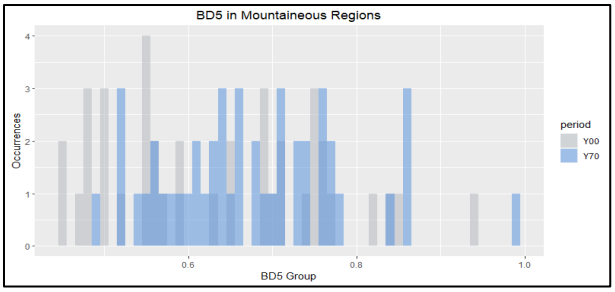

Figure 10: BD5 Values in Mountains

This may be down to a few reasons, one of them being urbanization of the coasts [2] and fragmentation of habitats which has led to an increased difficulty for wildlife to find shelter, feeding and breeding locations.

Another factor may be the increase of pollution in these areas. Coastal areas are usually more susceptible to industrial waste discharge and agricultural runoffs which affects the health of both marine and land ecosystems. Furthermore, coastal areas are more in danger of suffering from invasive species through shipping and trades. These species can cause an imbalance the native flora and fauna and cause a decline in the areas' natural biodiversity.

Moreover, the effects of global warming have hit the coastal areas the hardest while locations that are higher in altitude are yet to receive as adverse effects. These changes effect the habitat and availability of food resources for the wildlife living there. What this data shows is that special care and efforts need to be put into saving wildlife that lives next to the coastlines of the UK as they are the ones being affected the most since the turn of the century.

CONCLUSION

The given dataset covering the 5 taxonomic groups over two time periods show a close and generally positive correlation with the original 11 taxonomic groups. We interpreted that the data does not change dramatically but there is overall decline in biodiversity from Y70 to Y00. We then examined the correlation among the pairs themselves with majority of them showing a positive correlation. Finally, we found using regression models that behavior of another taxonomic group can be predicted based on our current data with moderate accuracy and they can be used as predictors of the area's ecological status. Another interesting finding is that the amount of specie biodiversity differs over the two-time period which brings to light the requirement of monitoring the biodiversity changes to support the ecosystem and prevent mass extinctions.

REFERENCES

Dyer RJ, Gillings S, Pywell RF, Fox R, Roy DB, Oliver TH. Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain. Collen B, editor. Journal of Applied Ecology. 2016 Oct 6;54(3):872–82.

Office for national statistics. Coastal towns in England and Wales - Office for National Statistics [Internet]. www.ons.gov.uk. 2020. Available from: https://www.ons.gov.uk/businessindustryandtrade/tourismindustry/articles/coastaltownsinenglandandwales/2020-10-06