MA334 COURSE WORK

Registration Number- 2201538

Introduction:

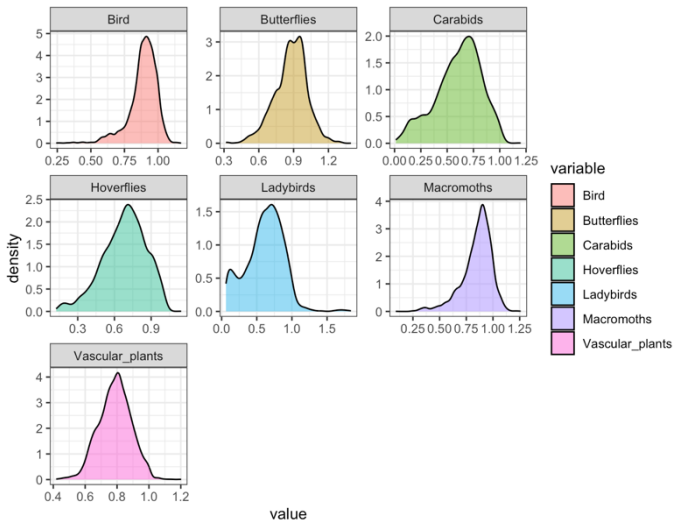
In this project, we are demonstrating the statistical analysis of “Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain". The researchers analysed data separately for different time periods and taxonomic groups because biological recorders tend to focus on specific taxa. For each analysis, they compiled a list of species and calculated the observed species richness for each area.

The data set includes information on taxonomic groups such as bees, birds, butterflies, and carabids, as well as information about the habitat and environmental conditions at each location. The data set can be used to investigate patterns of biodiversity across different locations and over time, and to explore relationships between biodiversity and environmental factors. The information in this data set could be useful for conservation efforts and for understanding how human activities impact biodiversity and ecosystem health. In this project, We will explore this data with EDA, hypothesis test , statistical tests, and model fitting for assigned specific seven species.

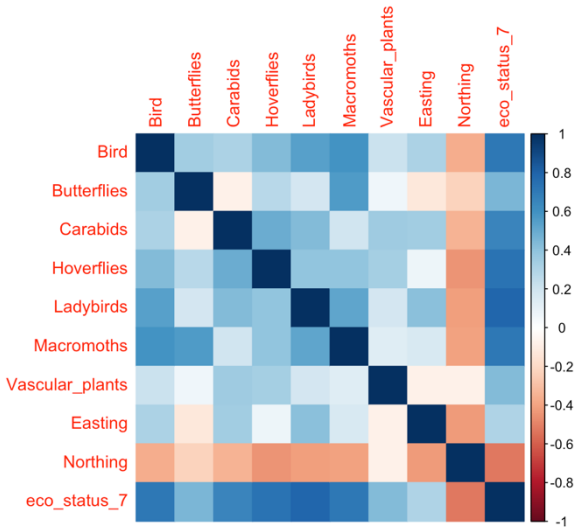
Exploratory Data Analysis:

1. Density Distribution of Seven Species Populations:

When looking at the graph, we can see that the populations of Bird, Carabids, Hoverflies, and Macromoths are all left-skewed normal distributions. This means that most of the values are concentrated on the right-hand side of the graph, with a tail extending towards the left-hand side. t's also important to note that the mean of the distribution is less than the median, which indicates that there are some high values that pull the mean to the right. On the other hand, the populations of Butterflies and Vascular plant appear to be normally distributed. Specifically, Vascular plant have a symmetric distribution, which means that the data is evenly distributed around the mean. However, Ladybirds have a left-skewed normal distribution. Overall, this graph can help us understand the distribution structure of the seven species and provide insights into their population dynamics. It is important to note that this is just one aspect of the data, and further analysis may be needed to fully understand the factors that are influencing these populations.



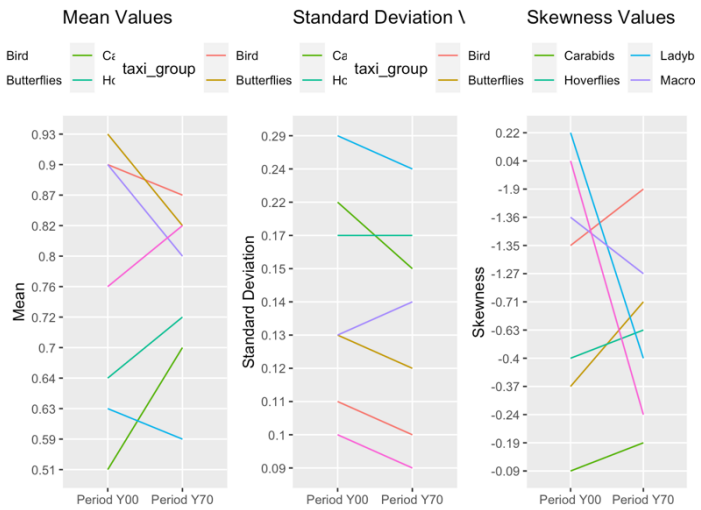
2. Correlation matrix for each species:



The values in the matrix represent the Pearson correlation coefficient, which ranges from -1 to +1, indicating the strength and direction of the correlation between two variables. Positive values indicate a positive correlation (when one variable increases, the other also increases), while negative values indicate a negative correlation (when one variable increases, the other decreases).

- All variables are positively correlated with each other to some extent, except for the variables "Northing" and "Easting" which are negatively correlated with each other.
- The variables with the highest positive correlations are "Bird" and "Macromoths" (0.594) and "Hoverflies" and "eco_status_7" (0.737).
- The variables with the lowest correlations are "Butterflies" and "Carabids" (-0.072) and "Vascular plants" and "Northing" (-0.075).
- The variable "Northing" has negative correlations with all other variables except for "Vascular plants".
- The variable "Easting" has low to moderate positive correlations with most variables, except for "Hoverflies" and "Northing".
- The variable "eco_status_7" has high positive correlations with "Bird", "Hoverflies", "Ladybirds" and "Macromoths", moderate positive correlation with "Carabids" and low positive correlation with "Vascular plants" and "Easting".

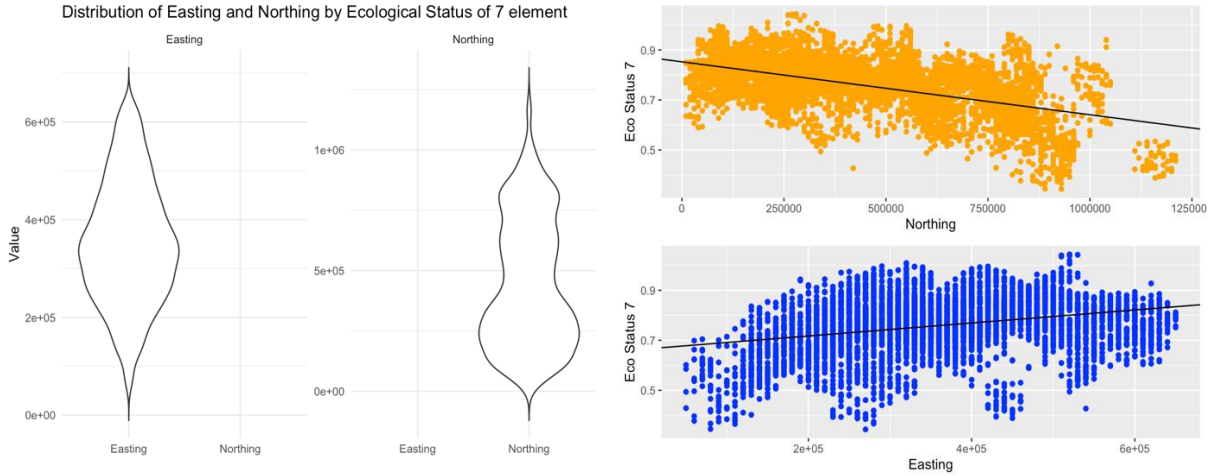
3. Comparison Between Mean ,Median ,Skewness for each period:



The two tables provide mean, standard deviation, and skewness values for different taxa groups during two different periods: Y70 and Y00.

- The two tables present mean, standard deviation, and skewness values for different taxa groups during two different periods: Y70 and Y00.
- Overall, the mean values increased for 3 variables and deceased for 4 variables from Y70 to Y00.So the average mean is decreases over the year.
- The standard deviation values are relatively increases across taxa groups for each period. This means that there was generally more variability in the data points within each taxon group in Y00 compared to Y70.
- The skewness values vary across taxa groups and periods, with some groups showing positive skewness (Bird and Vascular plants in Y70, Butterflies and Macromoths in Y00), while others show negative skewness (Carabids and Ladybirds in both Y70 and Y00).
- Notably, the period Y00 has more taxa groups with skewness values that are further away from zero, indicating a more non-normal distribution for these groups in this period compared to Y70.

3. Location wise distribution of Seven Species Population:



For the Easting side, the ecological status appears to be more evenly spread out across the middle range of the violin plot, meaning that there are similar numbers of 7 elements with high, medium, and low ecological statuses. However, as you move towards the top and bottom of the range, there are fewer 7 elements with extreme ecological statuses, either high or low.

In contrast, for the Northern side, the ecological status appears to be more heavily shifted towards the bottom of the range, with fewer 7 elements having high ecological statuses and more having low ecological statuses. This means that the range of ecological statuses is more compressed towards the bottom of the plot, and there are fewer 7 elements with high ecological statuses.

These differences in distribution could be due to a few factors, such as differences in environmental conditions or sampling methods between the Easting and Northern sides. Based on the scatter plots of BD7 vs. Northing and BD7 vs. Easting, it can be observed that the mean line for BD7 gradually decreases as we move from the north to the south in the Northing direction. This suggests that areas with higher Northing values tend to have lower BD7 values on average. Similarly, the mean line for BD7 gradually increases as we move from the east to the west in the Easting direction. This suggests that areas with higher Easting values tend to have higher BD7 values on average.

Hypothesis tests:

A} One-Sample t-test :

This hypothesis test is a one-sample t-test to determine whether the mean of a sample of BD7_change is significantly different from zero. The null hypothesis is that the true mean of the population from which the sample is drawn is equal to zero. The alternative hypothesis is that the true mean is not equal to zero.

The null hypothesis for this test is that the mean of the 7 species' ecological status values is equal between the two time periods (Y70 and Y00), meaning there is no change in the mean. The alternative hypothesis is that the mean is not equal between the two time periods, indicating a significant change.

Based on the output of the t-test, the p-value is very small (2.898e-14), which is less than the significance level of 0.05, indicating strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that there is a significant difference in the mean ecological status of the 7 species between Y70 and Y00.

The sample mean of the 7 species' ecological status values from Y70 to Y00 is -0.0084, which is negative, indicating a decrease in ecological status over time. The 95% confidence interval for the difference in means between the two time periods is (-0.0106, -0.0063), which does not include 0, further supporting the rejection of the null hypothesis.

```
> hyp_model1<-t.test(BD7_change,mu=0)  # t test with H0: mu=0

> hyp_model1

Output :

        One Sample t-test

data:  BD7_change

t = -7.6453, df = 2639, p-value = 2.898e-14

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

 -0.010584571 -0.006263425

sample estimates:

   mean of x

-0.008423998
```

B} ANOVA test:

The null hypothesis (H0) for this ANOVA is that there is no significant difference in the mean ecological status of the 7 species across the dominant land classes. The alternative hypothesis (H1) is that there is a significant difference in the mean ecological status of the 7 species across the dominant land classes.

The independent variable is dominantLandClass, which has 44 levels (different land classes) and the dependent variable is BD7. The ANOVA output shows that there is a significant difference in the mean ecological status of 7 species across different land classes (F=70.89, p<0.001).

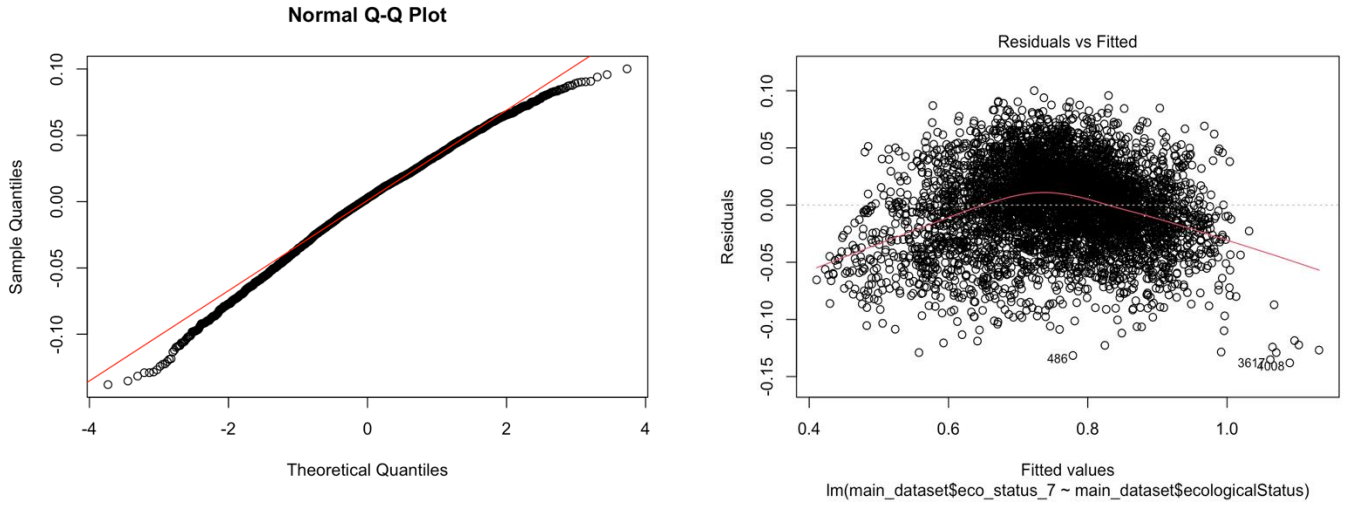
The p-value (p<0.001) indicates that the null hypothesis, which states that there is no significant difference in the mean ecological status of 7 species across different dominant land classes, can be rejected. Therefore, we can conclude that there is a significant difference in the mean ecological status of 7 species across different dominant land classes.

```
> hyp_model3<-aov(eco_status_7 ~ dominantLandClass, data = main_dataset)

> summary(hyp_model3)

Output:

                    Df  Sum Sq. Mean Sq  F value    Pr(>F)

dominantLandClass   44  23.60  0.5364    70.89  <2e-16 ***

Residuals           5235  39.61  0.0076

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Simple linear regression:

A} Simple linear regression on how BD7 matches BD11:

This code performs a simple linear regression analysis to explore the relationship between the biodiversity of 7 species and the biodiversity of all 11 species. The scatter plot of the two variables is created using the function, and the red line represents the theoretical line of equality between the two variables.



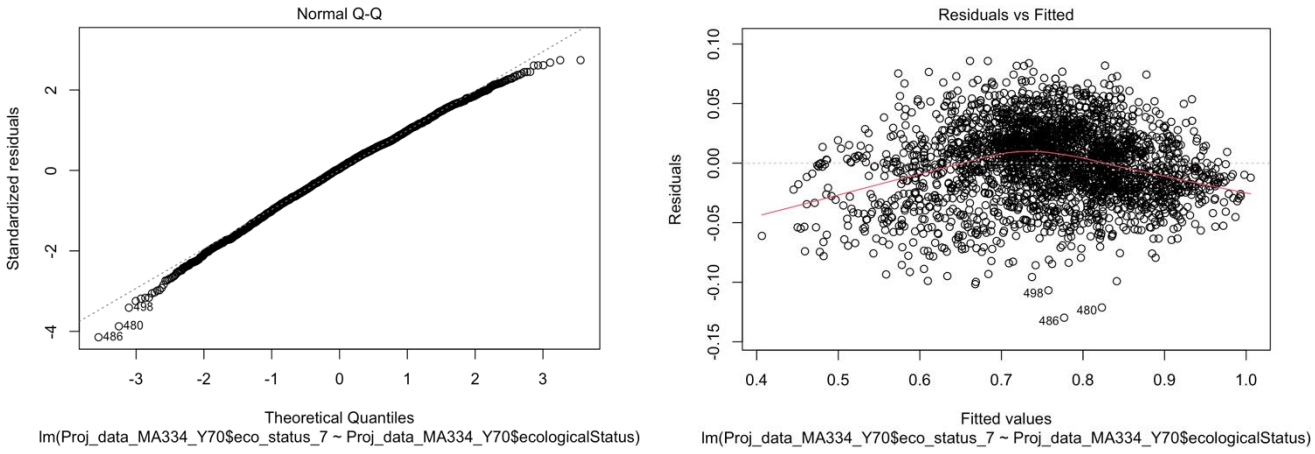
The coefficient estimate for BD7 is 0.957709, which indicates that for every unit increase in BD11, the mean of BD7 increases by 0. 957709.The p-value associated with the coefficient estimate is less than 0.001, indicated by *** in the Pr(>|t|) column, which means that the slope is statistically significant. In other words, there is strong evidence to suggest that there is a significant linear relationship between BD11and BD7.

The Multiple R-squared value of 0.8944 indicates that 89.44% of the variation in BD7 can be explained by the linear relationship with BD11. The Adjusted R-squared value is also 0.8944, which is the same as the Multiple R-squared value in this case since there is only one independent variable. AIC is a measure of the relative quality of the model compared to other models. A lower AIC value indicates a better fit. In this case, the AIC value is -20244.3, which suggests that the linear regression model is a good fit for the data.

The residual vs fitted graph is used to check the assumption of constant variance in simple linear regression. It is a plot of the residuals (observed values minus predicted values) against the fitted values (predicted values from the model). This plot shows no clear pattern or trend, indicating that the model's assumptions are met and that the residuals are randomly scattered around the zero line. In this case, the residual vs fitted plot shows no clear pattern, which suggests that the assumptions of linearity, constant variance, and normality of errors are met.

B} Simple linear regression for each period:

1. For Y70:

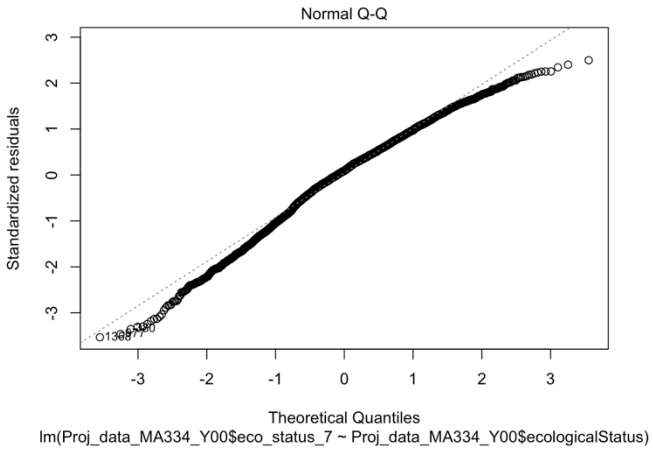


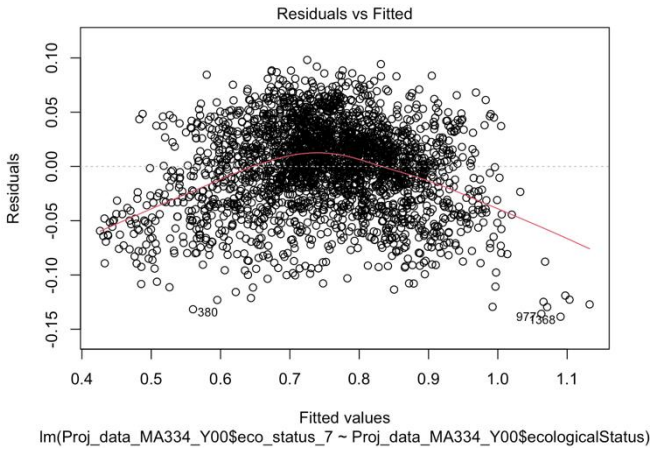
This linear regression model that predicts ecological status in 1970-1999 (Y70) based on ecological status in 2000. The summary output shows the coefficient estimates and statistical significance of the model's intercept and slope (i.e., BD11). The multiple R-squared value of 0.9107 indicates that 91.07% of the variability in the response variable (BD7) can be explained by the predictor variable (BD11).

The AIC value of -10793.38 suggests that this model is a better fit than other possible models, as lower AIC values indicate better model fit. Overall, this model appears to be a good fit for the data.

2. For Y00:

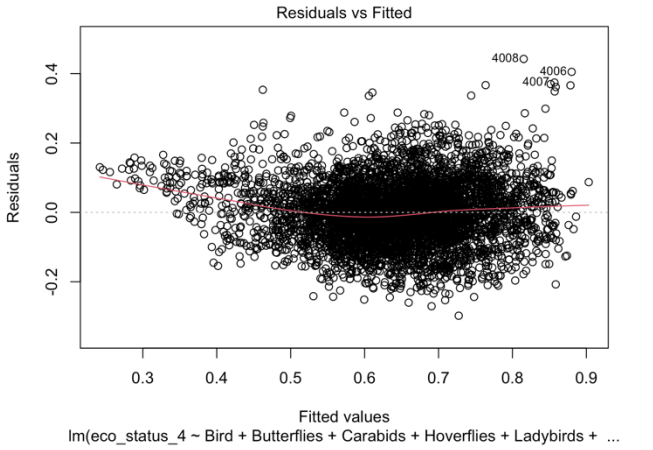
The output shows the estimated coefficients for the linear regression equation. The intercept is estimated to be 0.0764, meaning that if the ecological status is zero, the expected value of BD7 is 0.0764. The ecological status coefficient is estimated to be 0.9539, meaning that for each unit increase in ecological status, the predicted value of BD7 increases by approximately 0.9539. The linear regression model shows a strong positive correlation between ecological status and BD7. The expected value of BD7 increases by approximately 0.9539 for each unit increase in ecological status, and the cross represents the expected value of eco_status_7 when ecological status is zero. The model fits well with an R-squared value of 0.8808, indicating that approx. 88% of deviations in BD7 can be explained by ecological status. The F-statistic is highly significant, indicating that the model fits the data well overall. Finally, the AIC value is low, indicating that this is a good model for predicting BD7 based on ecological status.

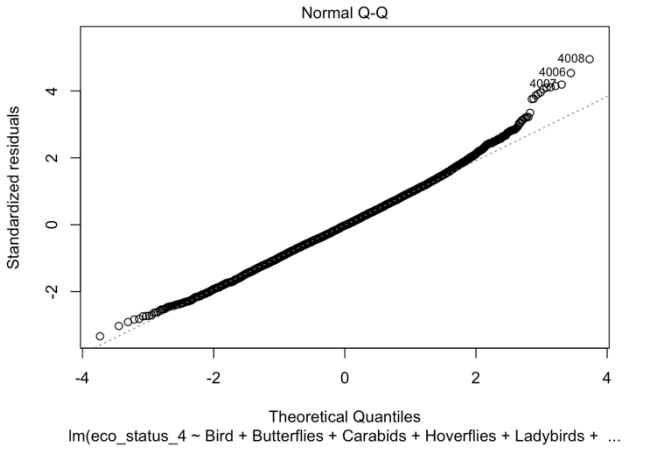




Multiple linear regression:

A multiple linear regression of BD4 against all seven of your proportional species values





This result is a summary of a multiple linear regression model examining the relationship between site ecological condition (dependent variable) and abundance of six taxa (independent variable), including birds, butterflies, shields, aphids, ladybirds, giant moths, and vascular plants. Coefficient estimates indicate the direction and magnitude of the relationship for each independent and dependent variable. For example, the coefficient estimate for birds (-0.167) indicates that for each unit increase in the number of birds, the ecological status of the site is expected to decrease by 0.167 units, holding all other variables constant.

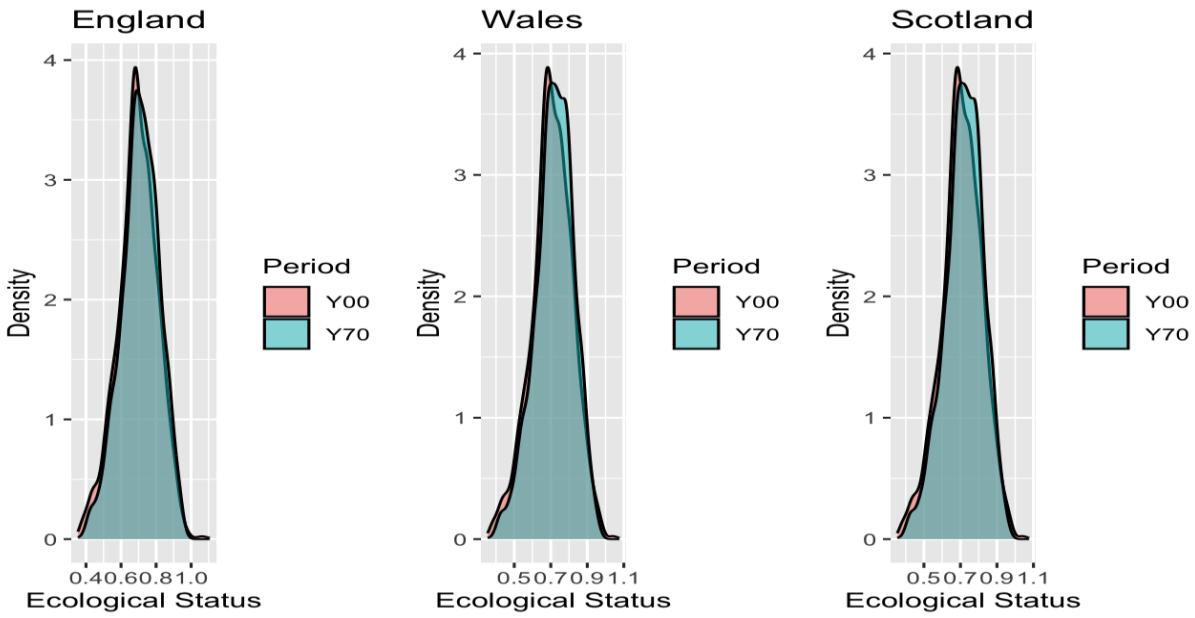
The P value associated with each coefficient estimate indicates the statistical significance of the association. A p value of less than 0.05 indicates that the variable is significantly related to the dependent variable. A multiple R-squared value of 0.5394 indicates that the model explains approx. 54% of the difference in the ecological status of the site. The adjusted R-squared value, which considers the number of independent variables in the model, is slightly less than 0.5388.

The F-statistic of 881.9 with a very low p-value (< 2.2e-16) suggests that the model is statistically significant overall, meaning that at least one of the independent variables is significantly associated with the dependent variable. Finally, the AIC value of -10500.61 can be used to compare this model with other models with different combinations of independent variables. Lower AIC values indicate better models with higher explanatory power.

There is no need of further feature selection as we got this the good fit multiple linear models.

Open analysis:

A} Country wise density analysis of all 11 species for each period:



The analysis examined the density distribution of 11 different species in three different countries (England, Wales, and Scotland) for two different time periods (Y70 and Y00).

For England, the density distributions of all 11 species were found to be normally distributed for both Y70 and Y00 periods. However, the bell shape of the distribution for Y70 was taller than that of Y00. There were more observations that fell within a narrower range of values for Y70 compared to Y00, resulting in a taller and more narrow bell shape for Y70. This suggests that the distribution of the data was more concentrated around the mean value for Y70, while for Y00, the data was more spread out.

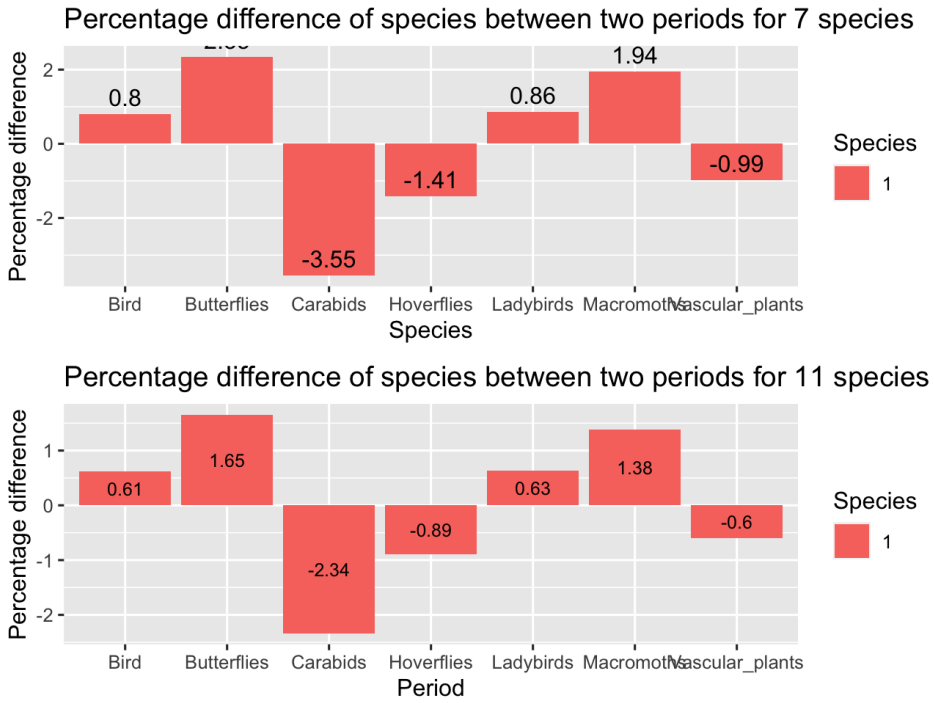
For Wales, the density distributions were almost symmetric and had the same height of bell shape for both Y70 and Y00 periods. The only difference was that Y00 had a slightly higher median than Y70.

For Scotland, the density distributions were similar to those of Wales, but there was less spread in Y00 than in Y70. This suggests that the effects of gas extraction were mostly observed in England compared to the other two countries.

The higher impact of gas extraction on England compared to Wales and Scotland can be inferred from the observed differences in the distribution of the species in each country. The extraction of gas from the ground can have negative effects on the local environment, including the soil, water, and air quality. It can also affect the biodiversity of the area by changing the habitats of various species. In this case, the normal distribution analysis of the 11 species showed that the bell shape of the distribution for Y70 in England was taller than that of Y00, indicating that the density of the species was higher before the gas extraction took place. Additionally, the Y00 period showed a slightly more spread-out distribution compared to Y70, which could suggest that the gas extraction had a more significant impact on the environment during this period.

In contrast, the normal distribution analysis of Wales and Scotland showed similar bell shapes for both Y70 and Y00 periods, indicating that the density of species remained relatively consistent between the two time periods. This suggests that the impact of gas extraction was not as significant in these areas compared to England.

B} Percentage difference of species between two periods:



The two tables show the percentage difference in species richness for two different sets of species (7 species and 11 species) when calculating the biodiversity index. The biodiversity index takes into account the number of different species present in a given area and gives a measure of overall biodiversity.

Comparing the two tables, we can see that the percentage differences in species richness are generally higher for the 11 species set compared to the 7 species set. This suggests that the inclusion of more species in the calculation of the biodiversity index leads to a greater overall change in species richness.Looking at the individual species, there are some differences in the percentage differences between the two sets. For example, the percentage difference in bird species richness is higher in the 7 species set compared to the 11 species set, while the opposite is true for butterflies and macromoths.

Overall, this suggests that the choice of which species to include in the calculation of the biodiversity index can have an impact on the resulting percentage differences in species richness. However, the general trend of higher percentage differences with more species included in the calculation holds across the different species analysed.

Conclusion:

Based on the analysis of the data, it can be concluded that the gas extraction has had a greater effect on the biodiversity of England compared to Wales and Scotland. This is because the density analysis showed a higher impact on England's species distribution, with a wider spread of values in the Y00 period.

Additionally, the analysis of the biodiversity index showed that the choice of species included in the index can have a significant impact on the results. The biodiversity index created using 11 species showed different percentage differences compared to the index created using only 7 species. This suggests that researchers and policymakers should carefully consider the species they include in their biodiversity assessments in order to accurately assess the state of biodiversity in each area.

References

- Mackay, E.B. and Taylor, C.M., 2015. Shale gas and groundwater: A review of the impacts of hydraulic fracturing on groundwater resources in the UK. Quarterly Journal of Engineering Geology and Hydrogeology, 48(4), pp.305-315.

- Smythe, D.K., 2015. Subsurface safety and environmental issues related to fracking and other unconventional gas and petroleum exploration and production: A review. Journal of Environmental Management, 157, pp.92-108.

- Deignan, S., 2017. Shale gas extraction and public health: Examining the evidence base for policy and regulations. Journal of Environmental Health Research, 16(2), pp.107-121.

- Bridge, G., 2015. The shale gas boom: a UK perspective. Energy Policy, 86, pp.455-464.
- Field, A., Miles, J. and Field, Z., 2012. Discovering statistics using R. Sage Publications Ltd.
- Crawley, M.J., 2013. The R book. John Wiley & Sons.

- Fox, J. and Weisberg, S., 2019. An R companion to applied regression. Sage Publications Ltd.

- Gelman, A. and Hill, J., 2006. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

- Kruschke, J.K., 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. Academic Press.

- McElreath, R., 2015. Statistical rethinking: A Bayesian course with examples in R and Stan. CRC press.