[redacted] [redacted]

[redacted] Data Analysis and Statistics with R

INTRODUCTION

The study starts by undertaking an in-depth analysis of the changing patterns of biodiversity illustrated in a dataset with a wide range of variables. The dataset, which encompasses geographical areas, ecological indicators, and taxonomic categories, includes essential elements such as bees, birds, bryophytes, butterflies, carabids, hoverflies, isopods, ladybirds, macromoths, grasshoppers_._Crickets, and vascular plants. The data set gives a solid basis for examining the complex relationships among ecosystems because it comprises geographical coordinates (Easting and Northing) as well as additional factors including dominantLandClassification, ecological condition, and measurement periods. The primary objective of this project is to statistically analyze Biodiversity Measure of BD5, which is the computed mean of proportional species values across five randomly selected taxonomic ("Bird", "Bryophytes", "Butterflies", "Ladybirds", "Macromoths").

Insights into the distribution of species richness can be retrieved by computing descriptive statistics for each of the five taxonomic groups. Furthermore, in order to reduce the impact of outliers, a winsorized mean is determined. The correlation matrix is analyzed in order to gain insight into the associations amongst the chosen variables. The analysis further encompasses graphical representations, such as box plots, to display the distribution and correlation patterns. The proportional species richness of birds and macromoths is contrasted in hypothesis testing, which is carried out using t-tests and Kolmogorov-Smirnov tests. The results indicate that there was a very significant variance among the two groups.

The correlation between changes in ecological status (BD11up and BD5up) throughout various time periods has been examined employing contingency tables. The significance of this relationship is evaluated using a likelihood-ratio statistics using G-test. Youden's index, sensitivity, specificity, and odds ratio are calculated to give an in-depth understanding of the interaction. Regression analysis is examined extensively in the sections that follow, starting with simple linear regression on the Hoverflies variable. Species richness and diverse predictor variables are having to be modeled in the regression analysis. P-values and AIC criteria are used for feature selection, and interactions between predictor variables are explored. In order to evaluate the predicted performance of the model, the dataset is subsequently split into training and test sets based on the period.

UNIVARIATE ANALYSIS AND BASIC R PROGRAMMING

Univariate analysis includes analysis and interpreting the distribution and characteristics of five taxonomic variables through descriptive statistics, 20% winsorized mean, boxplot and correlations between all pairs of variables in BD5.

1.Statistical Analysis

```
                  Min        Q1    Median      Mean        Q3      Max
Bird        0.24151709 0.8461695 0.9038163 0.8871739 0.9570151 1.171986
Bryophytes  0.39406158 0.6885744 0.7993481 0.7865969 0.8855011 1.174597
Butterflies 0.31666667 0.7925509 0.8862745 0.8745706 0.9676818 1.394366
Ladybirds   0.06140351 0.4545455 0.6394850 0.6140336 0.7972264 1.840000
Macromoths  0.08946553 0.7855507 0.8766727 0.8492665 0.9415221 1.260447
            Winsorized_Mean
Bird              0.8952239
Bryophytes        0.7874146
Butterflies       0.8760949
Ladybirds         0.6096508
Macromoths        0.8567578
```

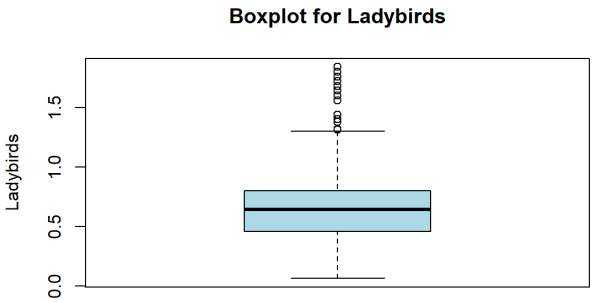
The table demonstrate standard statistics Min, 1st Quarter, Median, Mean, 3 rd Quarter, Max and 20% winsorized mean providing information on each group's distribution and central tendency with "Ladybirds" have substantial variability, whereas "Bird" has the highest mean.

2.Correlations Estimation

```
                  Bird   Bryophytes Butterflies  Ladybirds   Macromoths
Bird         1.00000000 -0.085146093   0.3422573  0.5496637  0.594111785
Bryophytes  -0.08514609  1.000000000   0.1930644 -0.2025997 -0.009209195
Butterflies  0.34225731  0.193064377   1.0000000  0.1851943  0.560875283
Ladybirds    0.54966367 -0.202599732   0.1851943  1.0000000  0.523412724
Macromoths   0.59411179 -0.009209195   0.5608753  0.5234127  1.000000000
```

The table gives the positive correlations and associations between "Bird" and "Macromoths" and "Ladybirds" and "Bird" which determines potential ecological relationships, while the negative correlations and inverse association between "Bryophytes" and "Ladybirds".

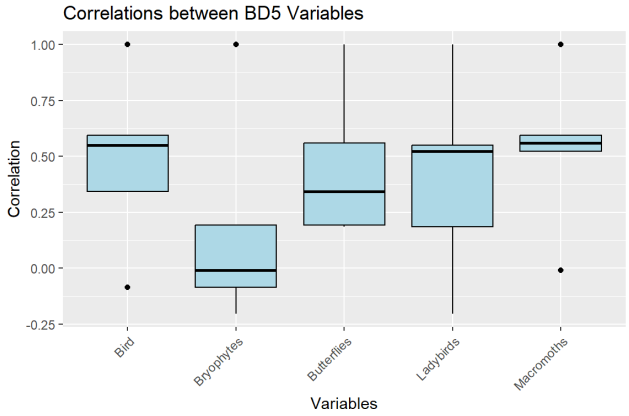
3. Boxplot



[redacted]

Choosing “Ladybirds” for the boxplot determines the exploration of different proportional species richness, contributing to a more detailed understanding of the ecological relationships within BD5.

4.Conclusion



The results highlight potential variations in relative species richness, such as “Birds” have the highest mean and “Ladybirds” have the lowest mean with significant variability. The “Ladybirds” with winsorized Mean (0.6097) determines strong insights with central tendency. Positive correlations, specifically between “Bird” and “Macromoths,” proves ecological relationships. Ecological trade-offs are demonstrated by the negative correlation (−0.20) between “Bryophytes” and “Ladybirds”. The boxplot highlights “Ladybirds” is in line with the species’ lowest mean and greater variability, which makes it easier to identify outliers and visualise patterns for biodiversity.

HYPOTHESIS TESTS

```
    Welch Two Sample t-test

data:  BD5$Bird and BD5$Macromoths
t = 15.614, df = 9836.3, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03314845 0.04266643
sample estimates:
mean of x mean of y
0.8871739 0.8492665
```

```
[1] "P-value for T-test: 2.61986073814515e-54"
```

```
    Asymptotic two-sample Kolmogorov-Smirnov test

data:  BD5$Bird and BD5$Macromoths
D = 0.15549, p-value < 2.2e-16
alternative hypothesis: two-sided
```

```
[1] "P-value for KS-test: 0"
```

The “Bird” and “Macromoths” among BD5 variables were chosen for hypothesis testing due to their ecological significance. Prominent evidence for significant variations in mean of proportional species richness and distribution charachetristics was obtained from the T-test and Kolmogorov-Smirnov test results.The KS-test determined p-value less than 2.2e-16, whereas the t-test determined highly significant p-value of 2.62e-54. The null hypothesis was rejected and it is signified with results of the t-test, with the confidence interval (0.033 to 0.043). These analysis, with the biological evidence for variable selection, emphasises ecological variations in proportional species richness between “Bird” and “Macromoths”.

CONTINGENCY TABLE / COMPARING CATEGORICAL VARIABLES

1.Contingency Table

```
   0    1
1638 1002
```

```
   0    1
 363 2277
```

```
                 BD5up decreases BD5up increases
  BD11up decreases             298            1340
  BD11up increases              65             937
```

There is a positive correlation between the changes in BD5 and BD11, as the contingency table shows a significant number of cases where, the contingency table illustrates the number of observations classified by variations in the biodiversity measures influencing BD11 and BD5. Remarkably, 1340 observations showed an increase in BD5up and a decrease in BD11, indicating an association between the two biodiversity measures. The positive correlation has been observed between BD5 and BD11 with increases or decreases in one biodiversity measure often being reflected in the other.

2. Likelihood-Ratio Statistic

```
Number of cases in table: 2640
Number of factors: 2
Test for independence of all factors:
    Chisq = 71.83, df = 1, p-value = 2.342e-17
```

```
[1] "Likelihood-Ratio Statistic: 78.9936895082649"
```

```
[1] "BD11up and BD5up exhibit a statistically significant association."
```

The likelihood-ratio statistic is used to evaluate the independence between changes in BD5 and BD11, and a computed value of 78.99 was determined. The difference in likelihood among the actual contingency table and the table expected under the independence assumption is calculated by this statistic. The likelihood-ratio test's p-value (about 2.342e-17) rejects the null hypothesis of independence. The strong correlation between the changes in BD5 and BD11 is highlighted by the confidence interval for this likelihood-ratio statistic, which adds an additional reason for rejection of independence.

3.Computation Odds Ratio, Sensitivity, Specificity, and Youden's Index

```
[1] "Odds Ratio: 3.20580941446613"
```

```
The odds ratio shows that BD11up and BD5up are positively correlated.
```

```
[1] "Sensitivity: 0.181929181929182"
```

```
[1] "Specificity: 0.935129740518962"
```

```
With a higher specificity, the test performs better in preventing false positive results.
```

```
[1] "Youden's Index: 0.117058922448144"
```

```
A positive Youden's index indicates generally good diagnostic performance.
```

A significant correlation was observed between BD11up and BD5up by proportional analysis, which indicated significant statistical association. A significant correlation was further justified by the odds ratio of 3.21, illustrating that an increase in BD11 has been associated with a greater likelihood of an increase in BD5. Nevertheless, the sensitivity of 0.18 wasn't very high, indicating that the ability to precisely identify increases in BD11 was constrained. However, the test's high specificity of 0.94 showed an excellent ability to prevent false positives, which added to its dependability. Regardless of the trade-off between sensitivity and specificity, the positive Youden's score of 0.12 suggested overall satisfactory diagnostic performance.

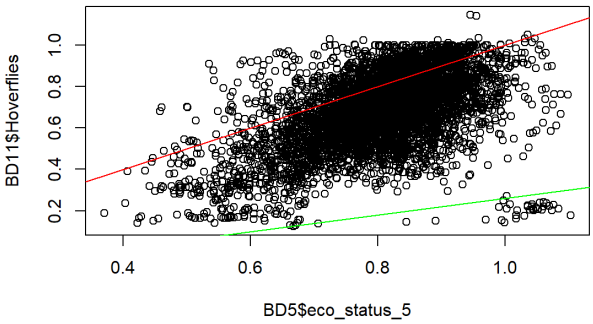
SIMPLE LINEAR REGRESSION

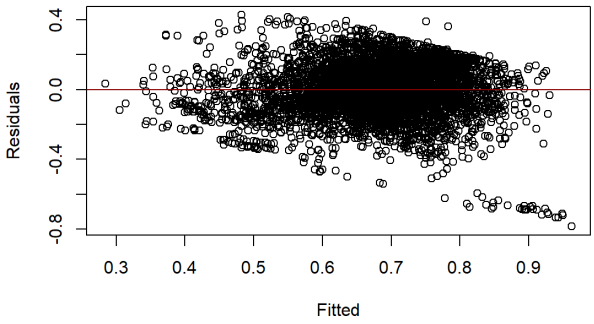
```
Call:
lm(formula = BD1_linear ~ BD5$Bird + BD5$Bryophytes + BD5$Butterflies +
    BD5$Ladybirds + BD5$Macromoths)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78498 -0.09010  0.01003  0.10323  0.42753

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.14500    0.02289  -6.334 2.58e-10 ***
BD5$Bird         0.40494    0.02583  15.674  < 2e-16 ***
BD5$Bryophytes   0.25665    0.01652  15.535  < 2e-16 ***
BD5$Butterflies  0.08697    0.01857   4.682 2.91e-06 ***
BD5$Ladybirds    0.16598    0.01004  16.537  < 2e-16 ***
BD5$Macromoths   0.10056    0.02202   4.567 5.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1503 on 5274 degrees of freedom
Multiple R-squared:  0.2811,    Adjusted R-squared:  0.2804
F-statistic: 412.4 on 5 and 5274 DF,  p-value: < 2.2e-16
```






```
Slope: 0.4049407
```

```
P-value for the slope: 3.726436e-54
```

```
The determined slope is statistically significant.
```

The significant positive correlation between the BD5 group and the hoverflies (BD1) was determined by simple linear regression analysis. Approximately 28.1% of the variability in hoverflies is explained by this model. In particular, the increase in Bird demonstrated a statistically significant impact on the growth in Hoverflies, which is shown by an increasing slope (p-value: 3.73e-54). There is a significantly higher slope that illustrates there is a strong linear relationship between Hoverflies and Birds. Significant insights regarding the ecological relationships among the dataset are given by this analysis.

MULTIPLE LINEAR REGRESSION

```
[1] "Correlation between predicted values and real response for the inital MLR model: 0.530185687201775"
```

There are statistically significant positive relationships between each predictor and Hoverflies(BD1) in the initial multiple linear regression (MLR) model that uses BD1 as the response variable and BD5 predictor variable. Overall, the model fits the data significantly ($p < 2.2e-16$), accounting for about 28.11% of the variation in BD1 and correlation coefficient of 0.53 between the actual and predicted values.

1.Estimation of AIC for the Initial MLR Model

```
Call:
lm(formula = BD1_linear ~ ., data = BD5[c(BD5_variables)], y = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-0.78498 -0.09010  0.01003  0.10323  0.42753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.14500    0.02289  -6.334 2.58e-10 ***
Bird         0.40494    0.02583  15.674  < 2e-16 ***
Bryophytes   0.25665    0.01652  15.535  < 2e-16 ***
Butterflies  0.08697    0.01857   4.682 2.91e-06 ***
Ladybirds    0.16598    0.01004  16.537  < 2e-16 ***
Macromoths   0.10056    0.02202   4.567 5.07e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1503 on 5274 degrees of freedom
Multiple R-squared:  0.2811,    Adjusted R-squared:  0.2804
F-statistic: 412.4 on 5 and 5274 DF,  p-value: < 2.2e-16
```

```
[1] "Estimation of AIC for the initial MLR model: -5022.99851314175"
```

The MLR model's estimated AIC of -5022.999 implies a good trade-off between complexity and model fit, indicating its efficiency. The corrected R-squared of 0.2804 validates the significant predictors' contribution for explaining BD1 variation.

2.Feature selection based on p-values and AIC

```
Start:  AIC=-20008.99
BD1_linear ~ Bird + Bryophytes + Butterflies + Ladybirds + Macromoths

             Df Sum of Sq    RSS    AIC
<none>                     119.08 -20009
- Macromoths  1    0.4709 119.56 -19990
- Butterflies 1    0.4950 119.58 -19989
- Bryophytes  1    5.4495 124.53 -19775
- Bird        1    5.5473 124.63 -19771
- Ladybirds   1    6.1750 125.26 -19744
```

The MLR model was subjected to the backward stepwise feature selection approach. BD5 predictors have been preserved in the reduced model, indicating its statistical significance in interpreting BD1 variation. The model's efficiency with the chosen predictors is proven by the adjusted R-squared (0.2804) and AIC, and eliminating them does not enhance the model's performance.

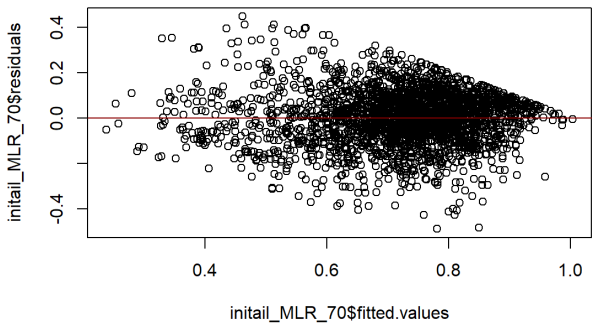
3.Interaction of any two predictor variables in BD5 group

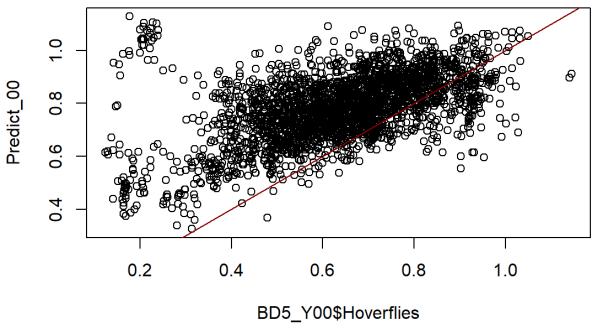
```
                       df       AIC
initail_MLR             7 -5022.999
initail_MLR_reduced     7 -5022.999
initail_MLR_interaction 8 -5075.642
```

```
[1] "Correlation of inital MLR model interaction between predicted values and real response: 0.537120536179244"
```

The linear regression model with an interaction between Macromoths and Butterflies was built to compare the model's fit to the initial and reduced models. The interaction model exhibited statistical significance in its coefficients, and it introduced an additional variable (Butterflies:Macromoths). The interaction model's AIC was -5075.642, which was higher than the AICs of -5022.999 for the initial and reduced models. Although the model complexity increased, the interaction model did not result in a reduced AIC, suggesting that the explanatory power of the model was not considerably enhanced by the addition of an interaction factor. Furthermore, the interaction model's correlation coefficient between predicted and observed values was 0.54, a slightly higher than the initial model's correlation coefficient of 0.53. As a result, the initial linear regression model without the interaction term remains the favoured approach for predicting Hoverflies in the BD5 group based on taxonomic data.

4.Division of one period Y70 as the training set and another Y00 as the test set





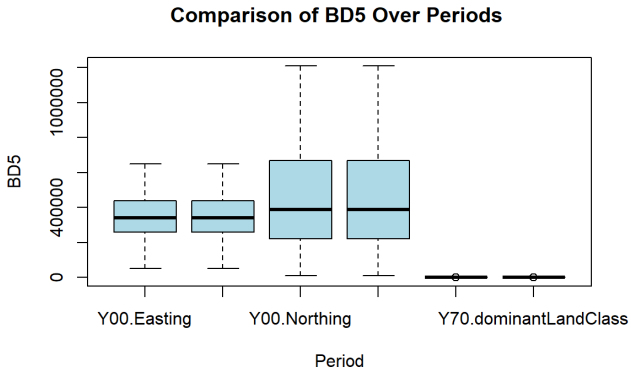

```
[1] "The Mean Square Error on the trainiing set:  0.0141500190660869"
```

```
[1] "The Mean Square Error on the testing set:  0.044018818561058"
```

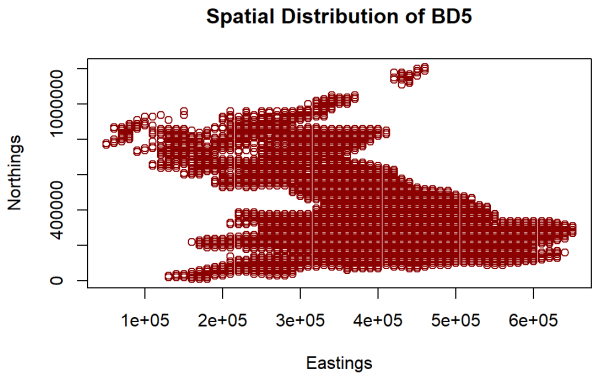
The training set mean squared error (MSE) of 0.01415 suggested that the linear regression model trained on the Y70 subset had a good fit with significant coefficients and appropriate diagnostic tests. On the other hand, the model exhibited a larger MSE on the test set (0.04402) when applied to the Y00 subset, showing lower predicted accuracy for the subsequent time period. The MSE differences that have been observed provide strong statistical evidence for the concept that due consideration should be taken when applying ecological models throughout a range of time periods. This indicates that the model may not generalise well to varying periodic instances, highlighting the significance of integrating time dynamics in ecological modelling.

OPEN ANALYSIS

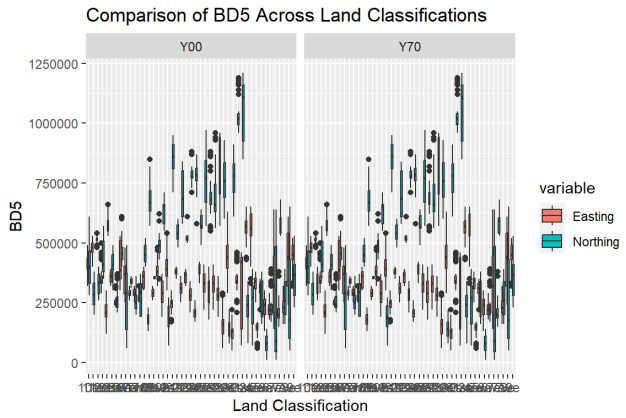
```
Exploratory Data Analysis - Comparision of BD5 over periods
```




```
Spatial Distribution of BD5
```




```
Linear Regression Analysis for BD5 group
```



The linear regression analysis conducted on BD5, incorporating variables such as period, dominatingLandClass, Easting, and Northing into account, illustrates evidence that is statistically significant of the factors influencing BD5. From period Y70 to Y00, the regression model shows a substantial decrease in BD5 (coefficient = -0.057), indicating a periodic shift in BD5 values along with variations in dominant land classes on BD5. Boxplots give a visual illustration of BD5 variations over various land classifications and time periods, whereas the spatial distribution plot emphasises geographical differences over BD5. The entire model, which is demonstrated by the high multiple R-squared value (0.4953), indicates that the incorporated factors together explain a significant part of the variance in BD5. The study is further supported by the coefficients for Easting and Northing, which emphasise their spatial significance and show positive associations with BD5. To summarise, the statistical evidence from the regression model, spatial distribution plot, and boxplots highlights the complex interactions of period and spatial factors influencing BD5. This in-depth open analysis provides significant insights into the ecological dependencies on BD5, consequently enabling well-informed decision-making about ecological relationships.

CONCLUSION

In summary, the analysis uses statistical methods to illustrate correlations and relationships by offering an in-depth analysis of biodiversity patterns in various contexts. In the selected taxonomic categories, descriptive statistics, correlation matrices, and graphical representations provide data regarding the distribution and interdependence of species richness. Hypothesis tests indicate significant differences in proportionaL species richness, highlighting the need of integrating specific taxonomic groups in biodiversity evaluations. The relationship between variations in ecological status over time has been clarified due to the contingency table analysis. The associations between species richness and predictor factors can be determined by regression analysis, which includes both simple and multiple linear regression. The improvement of the models can be assisted by feature selection based on AIC and p-values; interaction variables are additionally taken into account to identify more complex correlations. The analysis concludes with a split-sample approach to assess the predictive performance of the regression models. This approach ensures the generalizability of the models to new data.

[redacted] [redacted]

Overall, the analysis contributes valuable insights into the biodiversity dynamics of the studied taxonomic groups, offering a nuanced understanding of their patterns and relationships over time. The findings and methodologies employed in this analysis can serve as a foundation for further investigations and informed conservation efforts.

REFERENCES

[1] moodle.essex.ac.uk. (n.d.). Moodle: Log in to the site. [online] Available at: https://moodle.essex.ac.uk/course/view.php?id=15198 (https://moodle.essex.ac.uk/course/view.php?id=15198)

[2] Irizarry, R.A. (n.d.). Chapter 1 Getting started with R and RStudio | Introduction to Data Science. [online] rafalab.dfci.harvard.edu. Available at: https://rafalab.dfci.harvard.edu/dsbook/getting-started.html (https://rafalab.dfci.harvard.edu/dsbook/getting-started.html)

[3] moodle.essex.ac.uk. (n.d.). Moodle: Log in to the site. [online] Available at: https://moodle.essex.ac.uk/pluginfile.php/2016015/mod_resource/content/15/Guideline%20code%20for%20the%20MA334%20assignment%20Autumn%202 (https://moodle.essex.ac.uk/pluginfile.php/2016015/mod_resource/content/15/Guideline%20code%20for%20the%20MA334%20assignment%20Autumn%20

[4] James, R., Richard, F., David, B. and Tom, H., 2017. Developing a biodiversityâ based indicator for largeâ scale environmental assessment: a case study of proposed shale gas extraction sites in Britain. Journal of applied ecology, 54.

[5] National Grid. Available at: https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/guide-to-nationalgrid.pdf (https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/guide-to-nationalgrid.pdf)

[6] Land Classification. Available at: https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/Land%20Classification%20codes%20and%20explanations.pdf (https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/Land%20Classification%20codes%20and%20explanations.pdf) (Accessed: 15 January 2024).

[redacted]