MA334 Report

2213186

April 2023

Introduction

This statistical analysis report is conducted on a dataset containing the proportional species richness of 11 taxonomic groups across Great Britain. The data is taken from the research paper "Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain" published by Robert James Dyer, Simon Gillings, Richard F. Pywell, Richard Fox, David B. Roy, and Tom H. Oliver in the Journal of Applied Ecology in 2017. The proportional species richness is considered in categories of location based on the National Grid, the easting and northing values, and the dominant land class. The data was collected over two time intervals: 1970-1990 and 2000-2013.

Data Exploration

Out of the 11 taxonomic groups, seven are selected and subjected to univariate and correlation analyses. The groups being studied are butterflies, carabids, isopods, ladybirds, macromoths, grasshoppers&crickets and vascular plants. The summary function is used to study each group, and their relationships with each other are explored through a correlation matrix and plot. The table below contains the results of the summary statistics conducted for each of the seven groups.

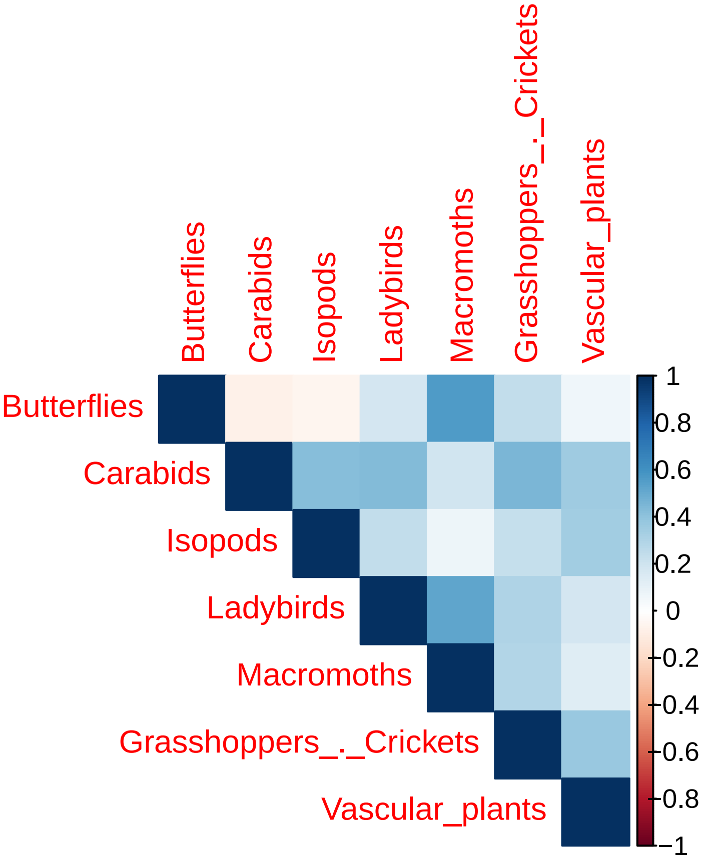
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Butterflies	0.3166667	0.7925509	0.8862745	0.8745706	0.9676818	1.394366
Carabids	0.0115321	0.4753877	0.6355274	0.6070596	0.7616054	1.199766
Isopods	0.0462185	0.3916521	0.5393599	0.5499496	0.7162319	1.257732
Ladybirds	0.0614035	0.4545455	0.6394850	0.6140336	0.7972264	1.840000
Macromoths	0.0894655	0.7855507	0.8766727	0.8492665	0.9415221	1.260447
Grasshoppers_._Crickets	0.0707965	0.4876033	0.6250000	0.6288737	0.7933884	1.593750
Vascular_plants	0.4178992	0.7213031	0.7911554	0.7868766	0.8551347	1.202265

The seven groups are then put into a correlation matrix. A correlation matrix is a table that shows the correlation coefficients between multiple variables. Correlation matrices help easily understand the strength and direction of the relationships between variables. The coefficients range from -1 to 1, with -1 suggesting a perfect negative correlation, 1 suggesting a perfect positive correlation, and 0 suggesting no correlation at all. The correlation matrix of the seven variables considered in this report is as follows:

	Butterflies	Carabids	Isopods	Ladybirds	Macromoths	Grasshoppers_._Crickets	Vascular_plants
Butterflies	1.000	-0.072	-0.051	0.185	0.561	0.243	0.066
Carabids	-0.072	1.000	0.427	0.430	0.193	0.453	0.352
Isopods	-0.051	0.427	1.000	0.242	0.070	0.230	0.340
Ladybirds	0.185	0.430	0.242	1.000	0.523	0.309	0.182

	Butterflies	Carabids	Isopods	Ladybirds	Macromoths	Grasshoppers_._Crickets	Vascular_plants
Macromoths	0.561	0.193	0.070	0.523	1.000	0.294	0.132
Grasshoppers_._Crickets	0.243	0.453	0.230	0.309	0.294	1.000	0.373
Vascular_plants	0.066	0.352	0.340	0.182	0.132	0.373	1.000

This data can be visualized in a correlation plot to easily understand whether the correlations are positive or negative and how strong they are.



With the help of the correlation plot, we can easily identify that the carabids and the isopods have a negative correlation with the butterfly plot.

Hypothesis Tests

In statistics, hypothesis tests are used to examine if a statement or hypothesis regarding a population parameter is likely to be true. This involves comparing a sample statistic to a theoretical distribution and calculating a probability value that represents the likelihood of observing the sample statistic or one more extreme, assuming the null hypothesis is true. The hypothesis tests conducted here are the t-test and the Kolmogorov-Smirnov test (KS test).

A t-test is a statistical test that compares the means of two data sets. It tests whether the means are significantly different from each other, taking into account the sample size, standard deviation, and degrees of freedom. In this case, sample sets are created for each period and tested for each group.

Null hypothesis: There is no significant difference in BD7 between the two time periods.

Alternative hypothesis: BD7 is significantly different between the two time periods.

```
##
##  Welch Two Sample t-test
##
## data:  sample_1$Butterflies and sample_2$Butterflies
## t = -32.671, df = 5269.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1214582 -0.1077073
## sample estimates:
## mean of x mean of y
## 0.8172793 0.9318620

## p-value: 2.304041e-213
##  Reject the null hypothesis. There is a significant difference in BD7 between the two time periods.
```

The p-values for all the seven groups are:

```
##                      Group       p-value
## 1              Butterflies 2.304041e-213
## 2                 Carabids 1.283525e-261
## 3                  Isopods  0.000000e+00
## 4                Ladybirds  1.245247e-07
## 5               Macromoths 5.419426e-136
## 6 Grasshoppers_._Crickets  1.266050e-25
## 7         Vascular_plants 5.045748e-112
```

The p-values being evidently lower than the significance level for all seven groups shows that there is a significant difference between the dependent and independent variables. The low p-values suggest that there is a significant relationship between the dependent variable and the independent variable. Therefore, we can reject the null hypothesis and conclude that there is a significant relationship between each of the independent variables and the dependent variable.

The Kolmogorov-Smirnov or KS test is a non-parametric test in statistics that compares two probability distributions to determine if they are significally different from each other. It compares the empirical distribution of the data to a reference distribution, which can be either a theoretical distribution or another empirical distribution.

```
##
##  Asymptotic two-sample Kolmogorov-Smirnov test
##
## data:  MainData$BD7 and MainData$ecologicalStatus
## D = 0.05322, p-value = 6.401e-07
## alternative hypothesis: two-sided
```

The test statistic (D) is 0.05322 and the p-value is 6.401e-07, which is less than the significance level of 0.05. Therefore, we can reject the null hypothesis that there is a statistically significant difference between the distributions of BD7 and BD11.

Simple Linear Regression

Linear regression is a modelling technique in statistics a single dependent variable and an independent variable. It involves fitting a straight line line through a set of data points in order to predict the value of the dependent variables based on the values of the independent variables. The summary of the simple linear regression performed on BD7 (mean of 7 groups selected) and BD11(ecologicalStatus) is as follows :

```
##
## Call:
## lm(formula = ecologicalStatus ~ BD7, data = MainData)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.120482 -0.022516 -0.001847  0.019584  0.242798
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.087784   0.002871   30.57   <2e-16 ***
## BD7         0.894594   0.004040  221.46   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.03368 on 5278 degrees of freedom
## Multiple R-squared:  0.9028, Adjusted R-squared:  0.9028
## F-statistic: 4.904e+04 on 1 and 5278 DF,  p-value: < 2.2e-16
```

The coefficients show that there is a positive relationship between BD7 and BD11. The p-values are extremely small, indicating that the relationship is statistically significant. The high values of the R-squared suggests that the model explains a large proportion of the variance in BD11 and is a good fit for the data. Overall, this model suggests that BD7 is a significant predictor of BD11.

The linear regression model summaries for each period are:

Y70 : Years 1970 - 1990

```
##
## Call:
## lm(formula = ecologicalStatus ~ BD7, data = period_70)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.111866 -0.019550 -0.001577  0.018776  0.120089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.053376   0.003877   13.77   <2e-16 ***
## BD7         0.924386   0.005304  174.27   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.02931 on 2638 degrees of freedom
## Multiple R-squared:  0.9201, Adjusted R-squared:  0.9201
## F-statistic: 3.037e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

Y00 : Years 2000 - 2013

```
##
## Call:
## lm(formula = ecologicalStatus ~ BD7, data = period_00)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.091178 -0.021740 -0.003693  0.017739  0.227107
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.090444   0.003727   24.27   <2e-16 ***
## BD7         0.909606   0.005400  168.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03265 on 2638 degrees of freedom
## Multiple R-squared:  0.9149, Adjusted R-squared:  0.9149
## F-statistic: 2.838e+04 on 1 and 2638 DF,  p-value: < 2.2e-16
```

For both models, the intercept and slope estimates are statistically significant with p-values <2e-16, indicating that there is a significant linear relationship between BD11 and BD7. The results confirm that the strong positive relationship concluded from the overall data is consistent across both time periods as well.

Multiple Linear Regression

Mutliple linear regression is the extension of simple linear regression, and provides higher accuracy as multiple independent variables are used. Here, multiple regression is performed on the mean of the remaining four groups against all the seven variables selected.

```
##
## Call:
## lm(formula = BD4 ~ Butterflies + Carabids + Isopods + Ladybirds +
##     Macromoths + Grasshoppers_._Crickets + Vascular_plants, data = MainData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29583 -0.05494 -0.00234  0.05077  0.58129
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.047856   0.011860   4.035 5.54e-05 ***
## Butterflies              0.281738   0.010535  26.742  < 2e-16 ***
## Carabids                 0.092916   0.007083  13.117  < 2e-16 ***
## Isopods                 -0.010662   0.006115  -1.743   0.0813 .
## Ladybirds                0.117191   0.005589  20.968  < 2e-16 ***
## Macromoths               0.153281   0.011626  13.185  < 2e-16 ***
## Grasshoppers_._Crickets  0.052600   0.006724   7.822 6.23e-15 ***
## Vascular_plants          0.202783   0.012961  15.646  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08393 on 5272 degrees of freedom
## Multiple R-squared:  0.5378, Adjusted R-squared:  0.5372
## F-statistic: 876.5 on 7 and 5272 DF,  p-value: < 2.2e-16
```

The coefficients table shows the estimated effect of each predictor on the response variable, as well as their statistical significance. The intercept term is also included. Based on the p-values, all predictor values except isopods are significantly associated with BD4. All six groups except isopods have positive coefficients, indicating that an increase in their abundance is associated with an increase in BD4. Isopodes have a negative coefficient, indicating that an increase in their abundance is associated with a decrease in BD4, although this effect is not statistically significant at the 0.05 level. The adjusted R-squared value of 0.5372 indicates that the model explains approximately 54% of the variance in BD4, which is a moderate-to-good fit.

For the selection of the final model, the step() function is used. Step() is used for stepwise model selection in linear regression to select the best subset of predictor variables for a model based on a chosen criterion such as AIC or adjusted R-squared. The function takes a fitted model object as an input and uses the direction argument to specify whether to perform forward or backward stepwise selection. The scope argument is user to specify the set of models to be considered, while the k argument determines the the penalty term used in the AIC or BIC criterion. The function performs a series of model fits, adding or removing variables one at a time, and returns the final selected model based on the chosen criterion. The selected model is often considered to strike a balance between goodness-of-fit and complexity.

```
## Start:  AIC=-26104.77
## BD4 ~ Butterflies + Carabids + Isopods + Ladybirds + Macromoths + 
##     Grasshoppers_._Crickets + Vascular_plants
## 
##                            Df Sum of Sq    RSS    AIC
## - Isopods                   1    0.0214 37.158 -26110
## <none>                                  37.137 -26105
## - Grasshoppers_._Crickets   1    0.4310 37.568 -26052
## - Carabids                  1    1.2121 38.349 -25944
## - Macromoths                1    1.2245 38.361 -25942
## - Vascular_plants           1    1.7243 38.861 -25874
## - Ladybirds                 1    3.0971 40.234 -25690
## - Butterflies               1    5.0375 42.174 -25442
## 
## Step:  AIC=-26110.29
## BD4 ~ Butterflies + Carabids + Ladybirds + Macromoths + Grasshoppers_._Crickets + 
##     Vascular_plants
## 
##                            Df Sum of Sq    RSS    AIC
## <none>                                  37.158 -26110
## - Grasshoppers_._Crickets   1    0.4317 37.590 -26058
## - Carabids                  1    1.2161 38.374 -25949
## - Macromoths                1    1.2414 38.400 -25945
## - Vascular_plants           1    1.7256 38.884 -25879
## - Ladybirds                 1    3.0759 40.234 -25699
## - Butterflies               1    5.0589 42.217 -25445
```

```
## 
## Call:
## lm(formula = BD4 ~ Butterflies + Carabids + Ladybirds + Macromoths + 
##     Grasshoppers_._Crickets + Vascular_plants, data = MainData)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.29733 -0.05492 -0.00236  0.05031  0.58526 
## 
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)             0.047323   0.011859   3.991 6.68e-05 ***
## Butterflies            0.282233   0.010534  26.794  < 2e-16 ***
## Carabids               0.089553   0.006817  13.137  < 2e-16 ***
## Ladybirds              0.116307   0.005567  20.892  < 2e-16 ***
## Macromoths             0.154181   0.011616  13.273  < 2e-16 ***
## Grasshoppers_._Crickets 0.052642   0.006726   7.827 6.00e-15 ***
## Vascular_plants        0.197739   0.012636  15.648  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08395 on 5273 degrees of freedom
## Multiple R-squared:  0.5376, Adjusted R-squared:  0.537
## F-statistic:  1022 on 6 and 5273 DF,  p-value: < 2.2e-16
```

The final model obtained using stepwise selection based on AIC has BD4 as the response variable and includes six predictor variables: Butterflies, Carabids, Ladybirds, Macromoths, Grasshoppers_._Crickets and Vascular_plants.

The AIC value for this model is -26110.29, which indicates that this model has the best balance of fit and complexity among all the models considered during the selection process. The regression coefficients for the final model show that all six predictor variables have a significant positive effect on the response variable, BD4. Butterflies have the largest coefficient of 0.282, indicating that a unit increase in Butterflies is associated with an average increase in BD4 of 0.282 units, holding all other variables constant. The other significant predictors have smaller coefficients, with Carabids having a coefficient of 0.089, Ladybirds having a coefficient of 0.116, Macromoths having a coefficient of 0.154, Grasshoppers_._Crickets having a coefficient of 0.053, and Vascular_plants having a coefficient of 0.198.

Overall, this final model suggests that the abundance of Butterflies, Carabids, Ladybirds, Macromoths, Grasshoppers_._Crickets, and Vascular_plants are all positively associated with BD4. This information can be useful for predicting and managing the abundance of BD4 in ecosystems.

Open Analysis

For the open analysis, the change in species richness across the two periods is considered. To understand the change in species richness across the two periods, the proportional species richness values for each taxonomic group were calculated by dividing the species richness of each group by the total species richness for each location in each period. The following table contains the percentages of species richness for each time period.

	1970 - 1990	2000 - 2013
proportional_species_richness_Bees	6.099429	8.779770
proportional_species_richness_Bird	11.078217	11.744706
proportional_species_richness_Bryophytes	10.053855	10.414258
proportional_species_richness_Butterflies	10.378284	12.129488
proportional_species_richness_Carabids	8.859339	6.382290
proportional_species_richness_Hoverflies	9.012574	8.130017
proportional_species_richness_Isopods	8.364244	5.544333
proportional_species_richness_Ladybirds	7.310614	7.867473
proportional_species_richness_Macromoths	10.130957	11.583528
proportional_species_richness_Grasshoppers_Crickets	8.261421	7.570835
proportional_species_richness_Vascular_plants	10.451066	9.853302

From the table, we can see that, in general, there is an increase in the proportional species richness for most taxa between the two periods. The groups that show the most significant increase in proportional species richness are Butterflies, Macromoths, and Bryophytes. The taxa that show the most significant decrease in proportional species richness are Isopods, Carabids, and Grasshoppers/Crickets. The proportional species richness of Birds, Ladybirds, and Vascular plants did not show significant changes between the two periods.

In the previous linear regressions, Butterflies showed the highest positive correlation, and Isopodes was one of the groups with a negative correlation against BD4. The recent results reinforce the findings from the linear regression.