MA334_2200694

ANALYSIS OF PROPORTIONAL SPECIES RICHNESS

INTRODUCTION

In this report, it is specifically based on the exploration of different taxonomic species based on their proportional specie richness in different areas at two time periods. This would help see the environmental impact (biodiversity) to typically rare and threatened species.

METHOD

The data used for this project contains 5280 observation of species richness in each hectads (location) and 17 variables. This data (known as BD11) is divided into two groups namely, BD7 (for my allocated 7 taxonomic group) and BD4 (for the remaining taxonomic group). The data has a class 'period' having two classes Y00 and Y70, the BD7 will be analysed with respect to these two periods. The data exploration phase will consider each feature as a univariate variable to develop my hypothesis question and then perform a hypothesis test to find facts to either support the null hypothesis or reject it. Then I would create a simple linear regression model to find out if BD7 matches with BD11 making use of their proportional species richness as the variables for our model. Next a multiple linear regression will be carried out using the BD4 as the response variable and selected 7 (BD7) as the predictor.

DATA EXPLORATION AND ANALYSIS

<u>Univariate Analysis</u>

Firstly, I explored selected BD7 species as univariate variables (Bees and Grasshopper/Cricket). Without handling outliers or normalizing the figures, the median value for the bees is 0.58869 and the mean is 0.60502. The mean and median values are a bit far apart which is because the bee's variables are right skewed measured at 0.958063. Assuming the data had a normal distribution, the mean should be around 0.5889555.

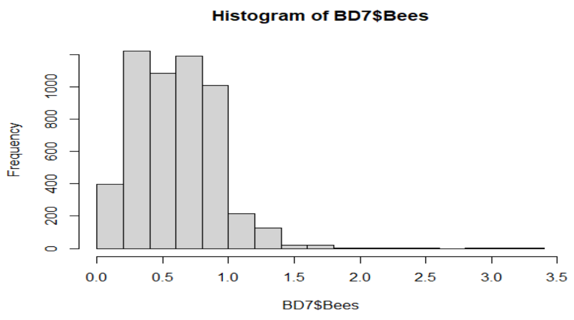


Fig. 1. Showing the distribution for Bees for all periods.

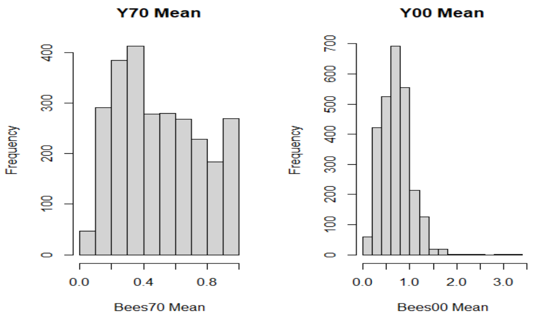


Fig. 2. Difference in Bees by period.

The periodical difference for the bee's species was reported as a 19.6% increase in mean from Y70 to Y00.

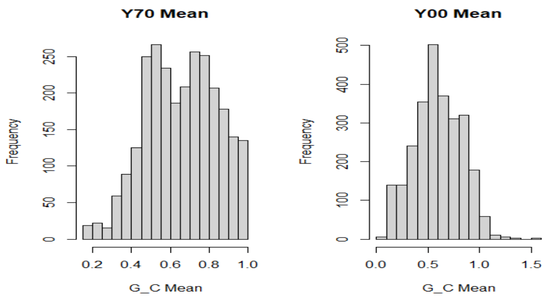


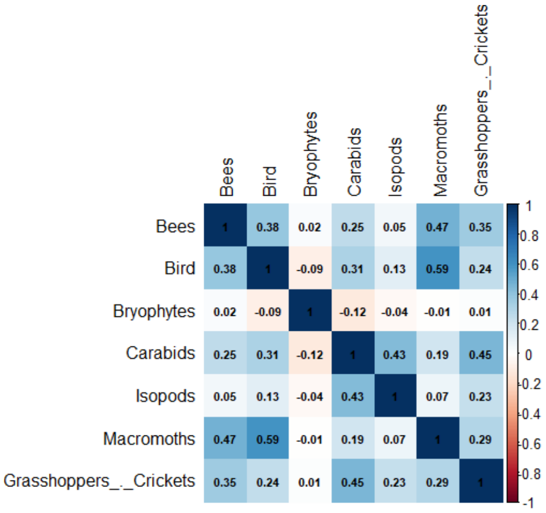
***Fig. 3.** Showing the difference in distribution between the two time periods for the Grasshopper/Cricket specie.*

Grasshopper/Cricket specie shows the difference in its distribution between the two time periods. For period Y00 the species fall between 0.1 and 1.3 with a few extreme values and Y70 fall between 0.2 and 1.0. Checking the difference in mean for the two periods, the grasshopper/cricket specie had a decrease of -6.1%.

<u>Correlation Analysis for all 7 species</u>

Considering the correlation between all 7 species, you can see that there isn't so much correlation between the species and the highest correlation value of 0.59 occurring between the bird and the macromoths. The number of specie occurrences by land classification is equal for both periods.

***Fig. 4.** Showing the correlation between all seven selected taxonomic group. Each box is coloured according to how negative or positive correlated they are to each other.*



	Bees	Bird	Bryophytes	Carabids	Isopods	Macromoths	Grasshoppers_._Crickets
Bees	1	0.38	0.02	0.25	0.05	0.47	0.35
Bird	0.38	1	-0.09	0.31	0.13	0.59	0.24
Bryophytes	0.02	-0.09	1	-0.12	-0.04	-0.01	0.01
Carabids	0.25	0.31	-0.12	1	0.43	0.19	0.45
Isopods	0.05	0.13	-0.04	0.43	1	0.07	0.23
Macromoths	0.47	0.59	-0.01	0.19	0.07	1	0.29
Grasshoppers_._Crickets	0.35	0.24	0.01	0.45	0.23	0.29	1

The ecological status of BD7 was explored by land class for each period, reporting that 35 out of the 45 unique land classes decreased by 77.78% and the remaining 10 increased by 22.22%.

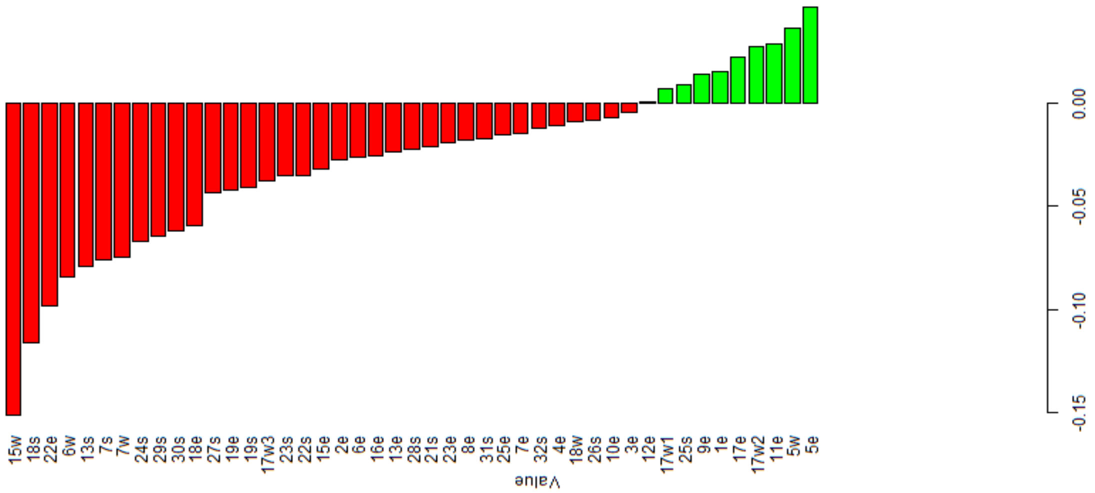


Fig.5. *Temporal ecological status change in biodiversity for each zone between the period, Y00 and Y70. Red colour indicates a decrease in status and green an increase in status.*

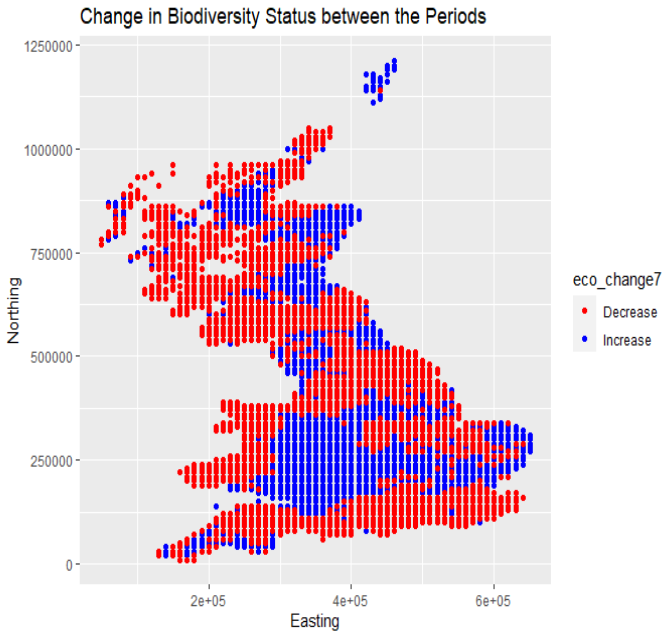


Fig.6. *Map showing the spatial patterns of the ecological status mean for each hectad. Red indicating area with decrease in ecological status and blue indicating area of high ecological status.*

On a smaller scale, I discovered that 1,652 hectads, which is more than half of the total hectad, had low difference in ecological status and about 988 hectads with high ecological status between the period. An example of the zones with low mean ecological statues as seen in the fig 5 is 15w (Flat River valleys/lower hill slopes, Wales) had the lowest difference of -0.152.

HYPOTHESIS TEST

From the exploratory analysis, the hypothesis question is said to be "Is there a difference in the ecological status of BD7 between the two time periods?", the null hypothesis is therefore that the difference in the mean of the two period and the difference in variance for the two periods are equal to 0 and 1 respectively. Two hypothesis tests were conducted, T-test to compare the mean and F-test to compare the variance. Comparing the difference between the mean using the T-test method, the p-value of the test is 3.209e-12, which is very small and indicates strong inference against the null hypothesis that the mean ratio between the two groups Y00 and Y70 is equal to 0. At a 95% confidence interval, the mean difference was reported as (0.01532260, 0.02727968) further supporting the rejection of the null hypothesis.

For the F test result, the p-value is 1.681e-08, which is less than 0.05 therefore indicating strong statistical inference to reject the null hypothesis that the ratio of variance between the two samples Y00 and Y70 is equal to 1. There ratio of variance reported in this test is 0.8026024, which is not 1.

MODEL AND RESULTS

In this phase, I made use of the Linear regression model. First performing a simple linear regression to see how well the model would perform on our data and see how BD7 matches with BD11 for each period. Then, performing a multiple linear regression and using the AIC method to find the best performing model.

Simple Linear Regression

Parsing BD7 as the dependent variable and BD11 as the independent variable, the p-value was 2.2e-16 (p<0.001) which means that there is a relationship between the two variables and therefore the null value is rejected. The intercept estimate is -0.0137, meaning that the model predicted the value -0.0137 for BD7 when BD11 is 0. The coefficient of BD11 is 1.0004 for every one-unit increase. The R-squared value shows that 94.34% of BD7 variance can be explained by BD11, thus proving that the model is able to predict the BD7 values at a 94.34% accuracy.

After splitting the two groups by period, the results gotten suggests that BD11's Y00 (Y00_11) is a strong predictor of BD7's Y00 (Y00_7), and the linear regression model is a good fit for the data. The coefficient estimates of Y00_11 is 1.009929, meaning that for every one-unit increase in Y00_11, the predicted value of Y00_7 increases by 1.009929. This is also the case for the BD11's Y70 (Y70_11) and BD7's Y70 (Y70_7) which has the same p-value as of 2.2e-16 but a coefficient estimates of 0.984297 which also means that for a one-unit increase in Y70_11, the predicted value for Y70_7 increase by 0.984297. Overall, we can say that there is a good positive linear relationship between BD7 and BD11 for both periods.

Multiple Linear Regression

Using the ecological status of BD4 as the dependent variable and the BD7 species as the independent variable, the p-value were the same for all BD7 species with a value of 2e-16 which is lower than the threshold of 0.05 or (p<0.001) and can be generally considered statistically significant to the dependent variable. With this, I performed a feature selection using AIC to get the best model for each selection. I ran the selection in a for loop that ran seven times and gave an AIC performance values at every selection iteration. The best model gave an AIC value of -11124.279 with all seven feature and the poorest model gave a value of -7771.674 with one feature. The final model gave a p-value of 2e-16 which has been earlier reported. The coefficient of each feature tells that for each unit increase in each feature, the predicted BD4 ecological status is expected to increase by the estimated value. For example, in every one-unit increase in Bees, there is expected to be an estimated increment in BD4 mean ecological statues by 0.093636.

CONCLUSION

In conclusion, analysis of the change in ecological status for BD7 between the two-time period recognized an overall decrease in ecological status. 62.5% of hectad (1,652 out of 2,640) has a decrease in ecological status. The was a large decrease observed in 35 environmental zones (land classes) and an increase in 10 zones.