MA334: Final Project
2201687

Introduction

Biodiversity is an important area of study, especially given a documented trend of decline in biodiversity across the UK. This project was conceived with the aim of analysing the biodiversity within the UK. The measures of biodiversity are standardised so differing groups can be compared effectively against each other. Additionally, data has been collected over two differing time periods, referred to as Y70 and Y00 - signifying a period round 1970 and between 2000 and 2013 respectively. This allows for analysis over time, and comparing biodiversity measures between the two periods.
Data within this set has been divided into 3 groups: biodiversity 11 (BD11), BD7, and BD4. These refer to the mean biodiversity of all taxonomic groups, a selection of 7 taxonomic groups from the original BD11, and the remaining 4 taxonomic groups left out of BD7 respectively. These will be used for comparisons and to aid in regression analysis regarding biodiversity and changes in it. An aim of this report is to establish if using a subdivision of taxonomic groups is an effective method of establishing overall biodiversity, or if analysis using all 11 groups is required to fully understand biodiversity within the UK. Furthermore, key questions this project seeks to answer include how BD7 differs from BD11 and BD4, and how these change over time.
The chosen 7 taxonomic groups of BD7 are Bees, Bryophytes, Butterflies, Isopods, Macromoths, Grasshoppers and Crickets, and Vascular plants; this leaves Birds, Carabids, Hoverflies, and Ladybirds in BD4.

Exploratory Data Analysis

With the aim of better understanding the data set being used in this project some exploratory data analysis will be conducted with a focus on exploring the chosen 7 as single variables, and exploring correlations between them and other features within the data set.

Looking at the species richness of the BD7 individually it can be seen that Butterflies are the most common across all hectads, with a standardised richness of 0.87 with a standard deviation (SD) of 0.14; meanwhile Isopods have the lowest level of richness at 0.55 with an SD of 0.22. Bees have the most variation, with a SD of 0.31, while Vascular plant have the lowest variation with an SD of 0.1. Finally, Macromoths have the most skewed distribution with a skewness of -1.14, while Isopods have the lowest at 0.05. These figures highlight that there is noticeable variance between the chosen 7 groups within BD7.

A quick analysis of the individual change in species richness for the chosen 7 taxonomic groups between the two periods reveals substantial changes for some and little change for others. For the chosen 7, the groups with the most variations

over time are -35% for Isopods, and +41% for Bees across all hectads. This level of change sharply moves towards the mean with a -9.1% change for Grasshoppers and Crickets, and a +14% change for butterflies. Overall, the mean species richness change for the chosen 7 is +2.26%. However, very few of the chosen 7 have a change close to this mean; Bryophytes are the closest with a 0.7% change, but all other groups within the chose 7 have much larger changes in species richness between the two periods.

Doing a basic correlation matrix of the continuous variables highlights strong correlations between differing species and their location via Easting or Northing. This suggests certain species become more common the further east or north you go. For example, Bryophytes have a strong negative correlation (-0.91) with Easting, suggesting they are noticeably more common in the west, becoming less common the further east travelled. Meanwhile, macromoths have a strong negative correlation (0.76) with Northing, suggesting they are most common in the southern hectads. While there are noticeable correlations between various different taxonomic groups, interestingly when correlating Isopods with Macromoths there is no correlation at all, with a correlation of 0; this is the only pair with no correlation at all, although not the only low correlation. Vascular plants appear to have low correlations with any taxonomic group capable of flight, with correlations of 0.01, -0.07, and -0.02 for Bees, Butterflies, and Macromoths respectively. Although vascular plants correlate reasonably well - correlations of 0.43 and 0.46 respectively - with animals such as grasshoppers and crickets, and isopods, suggesting that these likely share similar land classes.
Taxonomic groups capable of flight appear to have strong species richness correlations, with bees, butterflies, and macromoths all having correlations with each other between 0.61 and 0.77.

A correlation analysis between BD7 and Easting and Northing shows there is a negative correlation between BD7 and Northing (-0.34), while effectively no correlation between BD7 and Easting at 0.02. This signifies that changes in latitude have a much larger impact on biodiversity, which is to be expected as the environment sees much larger shifts in type travelling north to south than east to west in the UK thanks to the geographical shape of the country. What is interesting is that travelling northwards in the UK is mildly correlated with a decline in biodiversity, suggesting Scotland would have a lower biodiversity than the other areas of the UK.

Open Analysis

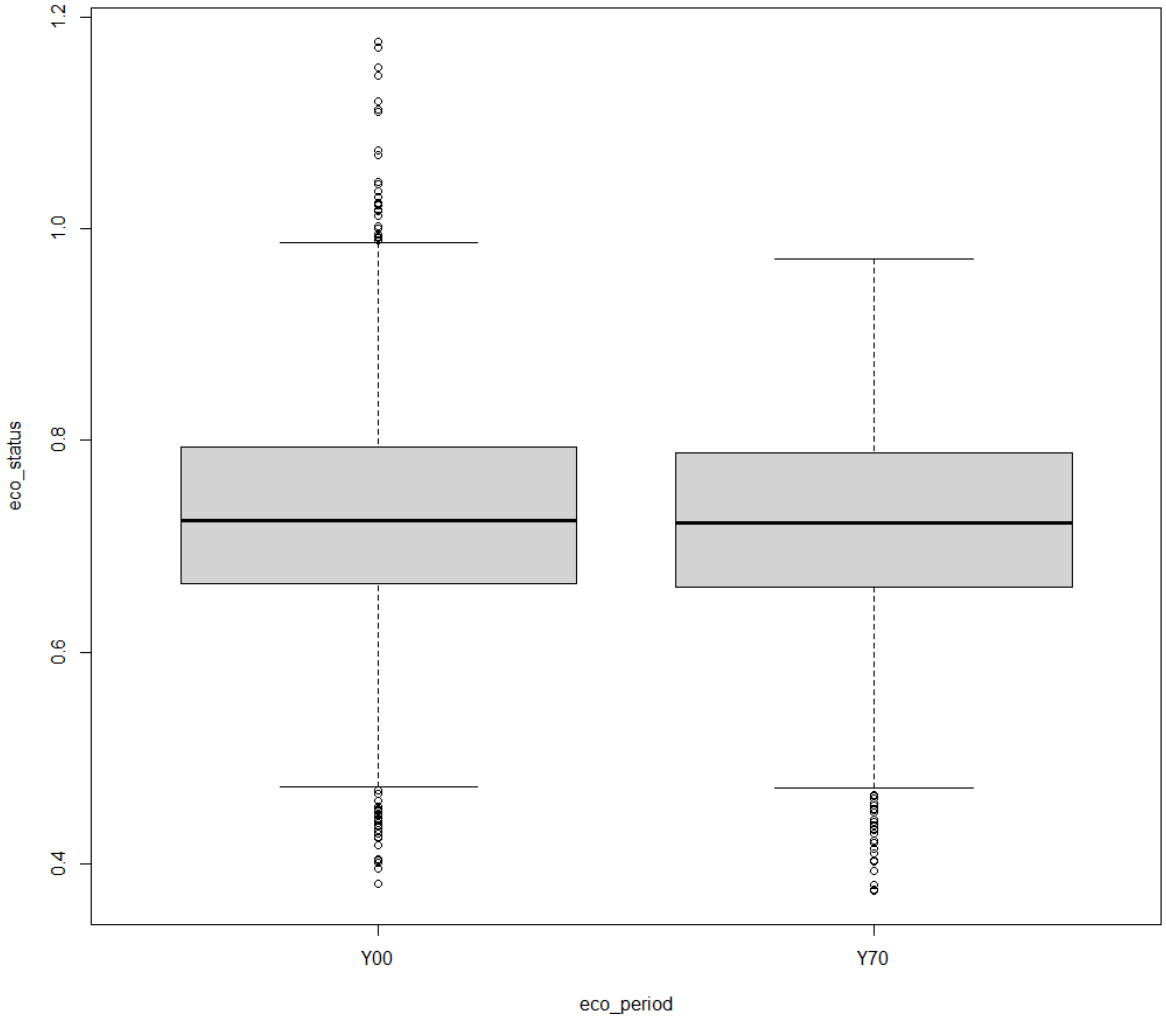
Some quick analyses reveals that Scotland is indeed the country with the lowest biodiversity within the UK, while out of England, Wales, and Scotland, Wales has the highest level of biodiversity. If northing is as good an indicator of biodiversity as suspected, then the UK subdivision with the least variation in northing - in this case Wales - is expected to have higher levels of biodiversity.
Using the assumptions made from analysis of the previous correlation matrix, lets test the hypothesis that biodiversity decreases the further north travelled. Firstly, by using northing as the predictor variable against BD7 in a linear model it can be seen that there is strong evidence for northing being a good predictor of BD7. With an F-

statistic of 685 and a p-value of < 2.2e-16 it is clear there is a relationship between the two. Although the R-squared value of 0.11 suggests while there is a relationship between the two variables, northing predicts little of the variance found within BD7. This is further seen by the number of data points with residuals larger than 0.4 - while this is still not a massive residual, it is clear it is not as effective a model as could be found. This likely means that while northing has an impact on biodiversity – probably due to the more drastic changes in temperatures and land classes – it is not the main feature influencing biodiversity, which makes logical sense.

Following on from earlier testing, it was discovered that northing is a good predictor for BD7, now lets expand that and see if UK subdivision is an effective predictor for BD7. Through regression analysis it can be seen that all 3 subdivisions - England, Scotland, and Wales - are effective predictors, sharing a relationship with BD7, as all of them have p-values close to 0. However, analysis of the R-squared values shows that UK subdivision explains even less variation within BD7 than northing, with England and Wales both having values of 0.01, while Scotland is slightly better at 0.05.
An AIC analysis of the three models shows that Scotland is the best predictor for BD7 with an AIC score of -9250 compared with -9085 and -9073 for England and Wales respectively. While it is unwise to draw conclusions regarding the reason for this, out of the three subdivisions, Scotland had the lowest biodiversity by a noticeable amount, despite its size compared to Wales. This is surprising given Scotland has a greater number of land classes than Wales which might be assumed to have more of an impact on biodiversity than northing.
Completing a series of Kolmogorov-Smirnov tests shows that the different biodiversity scores by subdivision do not share similar distributions, with incredibly low p-values for all three Kolmogorov-Smirnov tests run. This suggests there is a noticeable difference in how the taxonomic groups are spread across the UK and within its subdivisions.

Now we are looking at using the differing land classes as predictors for BD7 and BD11, this will be localised to Scotland as it is the area with the least biodiversity of the three subdivisions, while having a wide variety in its different land classes. From the analysis it can be seen that land class is better as a predictor variable for BD11 than BD7, as it has a higher F-statistic at 28.41 to 26.42, although both have the same very low p-value (< 2.2e-16), suggesting it is a good predictor variable for both. Continuing, only one of the differing land classes within Scotland has p-value above 0.05 for BD11, but there are three land classes with p-values above 0.05 for BD7, likely contributing to the lower f-statistic seen. Therefore, while land class within Scotland is a good predictor variable, by removing certain land classes from the model - namely 26s for BD11, and 26s, 23s, and 7s for BD7 - the AIC score for the model can likely be reduced below the current best of -3415. Finally, a quick look at the residuals shows there is very little difference with BD7 at 0.35 maximum away from 0 compared to 0.36 for BD11. The same can be said for the R-squared score of 0.17 to 0.18 for BD7 and BD11 respectively.

The change in biodiversity over time



These two boxplots show the differences to be found in BD7 between Y70 and Y00. Comparing the two it is clear to see there is a greater level of variance in Y00 than in Y70, with a large number of outliers both above and below the tails of the boxplot. Unlike in Y70, there are also a large number of outliers above the tail for the boxplot, suggesting some areas have seen a noticeable increase in the level of biodiversity compared with Y70. This is further seen in the slight increase in the mean, interquartile range, and upper tail for biodiversity in Y00, although this increase is very slight. This is to be expected as there has been noticeable declines in biodiversity across many different land classes and taxonomic groups, as well as increases in a few select land classes as well. [1]

Conducting a few statistical tests confirms the interpretations of the boxplots. There was a slight increase in the mean biodiversity in Y00 over Y70, though this is very slight at roughly 1%. This was discovered using a T-test on the change between the two periods; a p-value of 2.582e-12 provides strong evidence the true mean in biodiversity change is not 0, confirming there was a change in biodiversity between the two periods, with the expected mean being 0.01.

Simple Linear Regression

This section will use simple linear regression to analyse how BD7 matches BD11. This will be done over the two periods Y70 and Y00, as well as a simple comparison using all data from both periods; this will be to see if there is a noticeable difference between the different periods or if it is relatively unchanging.

The results of a simple linear regression analysis of BD7 against BD11 show that BD11 is a strong indicator of BD7 with an incredibly small p-value of < 2.2e-16. Additionally, with an R-squared value of 0.86 BD11 can be used to explain much of the variation found with BD7.
Further study of the residuals shows they are very close to 0, with all residuals being within 0.16 of 0. These results suggest this is an effective model for estimating BD7. Looking at the Q-Q plot it appears there is a distribution close to normal, but with a slight skew to the data.
This helps show that using a reduced number of biodiversity markers can be used well in determining local area biodiversity scores, as this model clearly shows itself to be effective for prediction of biodiversity.

Further deconstructing BD7 and BD11 into the two different time periods and running another simple linear regression model on BD7 against BD11 offers little change in the p-values, confirming what would have been suspected from the previous regression analysis - that BD11 is a good predictor for BD7. However, the difference comes when looking at the multiple R- squared value. In Y70 BD11 served as a better indicator of BD7 than in Y00, with BD11 explaining 0.89% of the variance in Y70, but only 0.85% in Y00. This suggests that there has been a greater level of divergence between the two, as BD11 explains less of the variance in BD7 roughly 30 years later. Of course while there has been some change over the two periods, it is clear BD11 is still a very effective predictor of BD7, and this is clearly an effective model, since the change is not drastic.

Kolmogorov-Smirnov Test

Looking at the Quantile-Quantile plot of the distributions for BD7 and BD11 shows they do not follow a normal distribution. Instead, it appears the distributions are slightly skewed, likely with thin tails.
Conducting a Kolmogorov-Smirnov test on BD7 and BD11 to establish the similarity in distributions between the two groups provides strong evidence against their similarity. With a very low p-value of 8.138e-06 the alternate hypothesis of a two-sided distribution must be accepted, suggesting these two groups likely did not come from the same sample. Given the increasing level of divergence over time between the two, as seen in previous tests, this may explain some of the difference in distributions.

Doing basic linear regression of BD4 against BD7 shows that BD7 is an effective predictor for BD4 with a very small p-value, however BD7 is not as good a predictor of BD4 as BD11 is for BD7, with a much higher - though still statistically significant -

p-value of 0.02 for the intercept. Additionally, while BD7 is a good predictor for BD4, it does not explain the majority of the variance in BD4; the multiple R-squared value for this model is 0.41.
Looking at the residuals we can see a higher level of variance than we have come to expect from out previous models. Variance here peaks at -0.48 which is noticeably higher than many of the previous models used. This likely suggests that BD7 is not as good a predictor for BD4 as BD11 was for BD7 previously. This also reflects the higher p-value for the intercept found within this model when compared to prior ones. And a quick look at the residuals on a Q-Q plot shows a slight skew to the data, confirming it does not follow an exact normal distribution.

Multiple Linear Regression

Building a model to predict the missing BD4 values from BD7 data through multiple linear regression shows that the individual taxonomic groups from BD7 can be used effective for this purpose. First splitting the data into a train and test set with an train/test split of 80/20%. Then using the individual variables from BD7 - from the training set - as the predictor variables for BD4 finds that each of them is an effective predictor with high t-values and very low p-values (<2e-16). Additionally, the adjusted R-squared value is 0.58, suggesting these predictor variables are effective at explain the majority of the variance with the dependent variable. Overall, the model has a very high F-statistic of 839.5 and a very low p-value of < 2.2e-16, suggesting the null should be rejected and conclude there is strong evidence for a relationship existing between BD4 and the chosen 7. An analysis of the residuals of this model further evidences its accuracy. All residuals bar one lie within 0.3 of 0, and the one outlier is only 0.4 away from 0.

Looking at the correlations between the predictions for BD4 based on actual values in the training and test sets shows strong positive correlations - 0.76 for the training set and 0.72 for the test set. This further evidences the conclusion that there is a relationship between BD4 and the chosen 7, and that this is an effective model for predicting BD4 using the constituents parts of BD7.

Continuing, by regressing the predicted values of BD4 against the actual values of BD4 in the test set it can be seen that the model is effective at predicting BD4. With a high F-statistic of 1166 and a p-value of < 2.2e-16 it is clear the null should be rejected and accept the H1 of there being a strong relationship between BD4 predictions and the actual values. The multiple R-squared value of 0.53 shows that the predictor variable explains the majority of the variance in out dependent variable. Looking at the residuals for the prediction of BD4 against the actual values of BD4 shows little distance from 0, with the peak distance for any residual being 0.28. And finally, looking at the Q-Q plot of the residuals shows a slight leftward skew to the data distribution.

Looking to improve the model used earlier, more biodiversity indicators were added, using all taxonomic groups, rather than simply the chosen 7. This was done as earlier analysis showed that all chosen 7 were effective within the model, leading to the conclusion that the model would not necessarily be improved by removing any of them. Thus, adding to the model appeared to be the logical conclusion. Using all 11

taxonomic groups massively improved up the AIC value, reducing it from -7704 for the chosen 7, to -281684 for all 11. However, when using all 11 several of the taxonomic groups now show p-values above 0.05 suggesting the model could be improved through their removal.
However, even when removing the taxonomic groups with p-values above 0.05 from the model, the AIC score could not be improved. Using the newly selected taxonomic groups instead of all 11 gave a higher AIC score of -281098 instead of -281684.

This clearly highlights that using a smaller selection of taxonomic groups is not as effective as using all of them. Use of a reduced number has proven relatively effective throughout previous testing, however, ultimately it still offers the chance to miss out essential information which may lead to a noticeably less effective model for biodiversity classification.

Conclusion

To conclude, throughout this report it has been evidenced that using the chosen 7 as a predictor for biodiversity across more or less taxonomic groups is effective when compared with using all 11. However, this is not without caveats. Use of all 11 taxonomic groups is noticeably more effective as a predictor than using a subset of 7, and a subset of 4 is also limited. Although this may be the case use of the 7 proved effective for estimations of biodiversity for all 11. Additionally, the chosen 7 also proved effective for estimating the other 4 groups not included in the 7, even when testing for potentially missing data.
Moreover, splitting our analysis along the lines of UK subdivisions proved noticeable differences between the three subdivisions - England, Scotland, and Wales. When expanded upon Scotland's hectads proved to be the best predictor. Once again use of BD11 instead of BD7 proved to be more effective, giving more evidence to the idea that using a subset of taxonomic groups, while still effective, offers a more limited view of biodiversity and may miss important aspects.
And finally, changes in BD7 between the two periods proved to fairly limited with a mean change of 1%, and although there was some change in the extent of the variation of biodiversity between the two time periods, it was not enough to warrant continuing to divide biodiversity exploration by the two periods.

Bibliography

[1]: Dyer, R. et al. Developing a biodiversity-based indicator for large-scale environmental assessment: a case study of proposed shale gas extraction sites in Britain. *Journal of Applied Ecology.* doi: 10.1111/1365-2664.12784