Data Analysis and Statistics

2212188

2023-04-20

INTRODUCTION

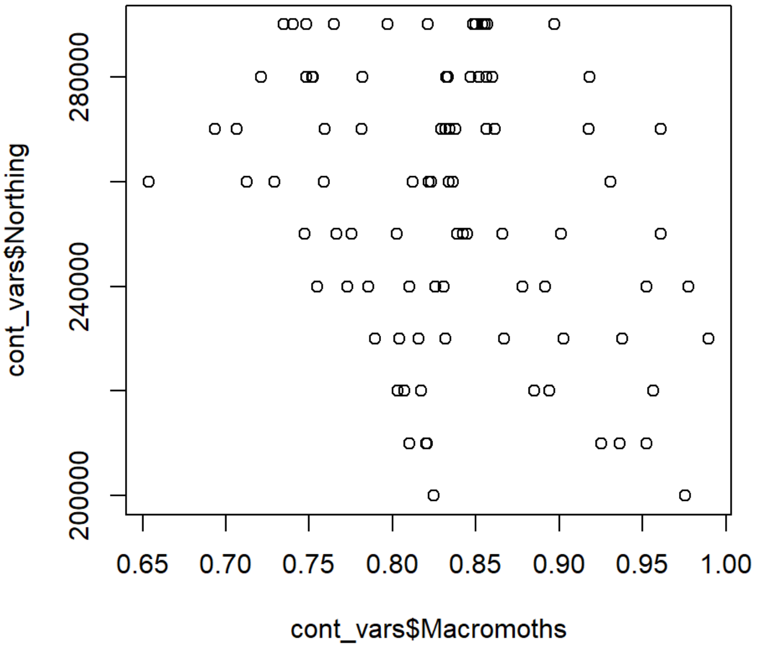
Conservation biologists and policymakers face significant challenges in quantifying and prioritising biodiversity. Increasing demands on land use, such as food and energy security, as well as housing expansion, are creating a greater need for methods to identify and prioritise regions of 'high' ecological value. There are several techniques based on land cover extent that may be used as a proxy for biodiversity, but they make a number of assumptions and either demonstrate inadequate match to empirical data or have not yet been adequately tested.

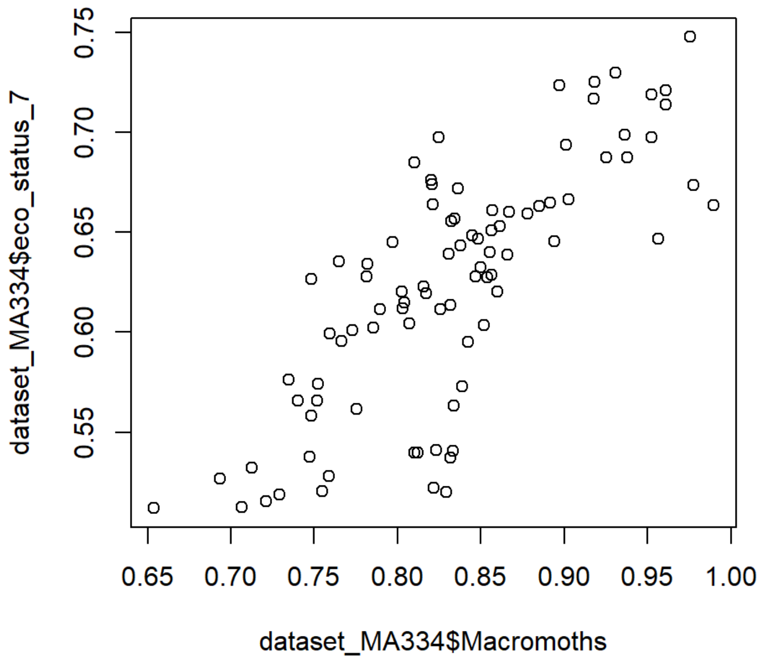
The environmental effects of changing land use on possible development sites are now predicted in Great Britain (GB) using environmental impact assessments (EIAs) and strategic environmental assessments (SEAs). The low priority accorded to biodiversity in general and the assessment's narrow emphasis on a limited group of priority species and habitats are two major flaws. The Biodiversity Action Plan , which defined priority species and priority habitats in GB up to 2010, before these lists became devolved to separate countries, is one example of how threatened species, threatened habitats, and Sites of Special Scientific Interest (SSSI) are used to assess biodiversity.However, because they only represent a small percentage of total biodiversity, they may not accurately reflect the spatial patterns and temporal trends in this 'wider' biodiversity, even though their usage can be useful for prioritising and conservation at the local level. For instance, the needs of species like the Great Crested Newt Triturus cristatus, a species listed as a European Protected Species under the Habitats Directive, may not coincide with those of other species that might profit from specific interventions like the creation of green infrastructure. Reporting on the condition of historically widespread species is critical, especially as they may support important ecological services outside of the narrow selection of officially protected species.

Data Exploration

With macromoths on the x-axis and eco_status_7 on the y-axis, the scatter plot depicts the association between the biodiversity measure (eco_status_7) and the number of macromoths detected in the dataset. These two variables have a 0.82 correlation coefficient, indicating a significant positive relationship.

The plot shows a cluster of locations near 0.8 macromoths and 260000 northing. This could imply that there is a specific habitat or environmental situation that promotes both high biodiversity and high macromoth numbers. More research into this subject may clarify what elements contribute to this tendency.





Hypothesis tests

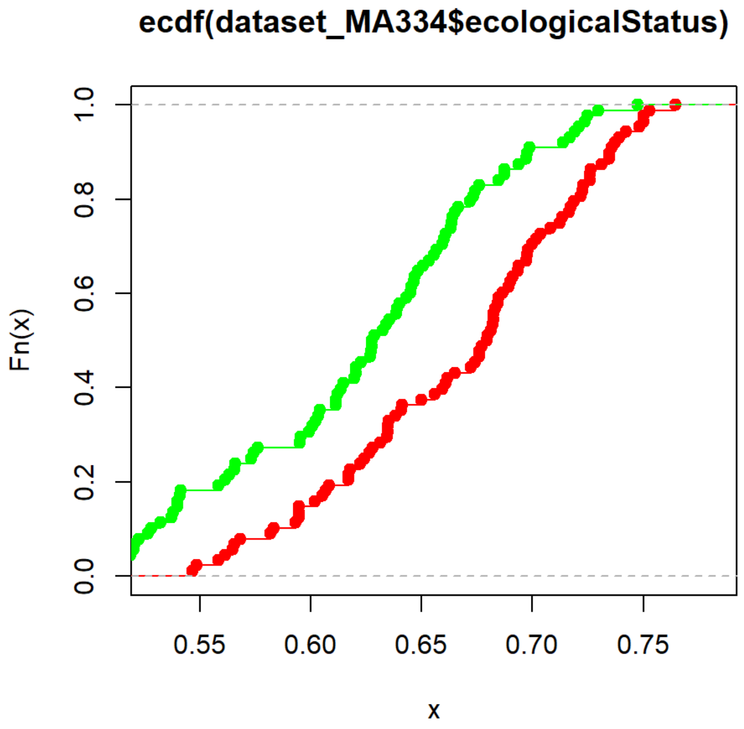
The graph depicts the empirical cumulative distribution functions (ECDF) for two variables from the dataset dataset_MA334, eco_status_7 and ecologicalStatus. The x-axis represents variable values, whereas the y-axis displays ECDF values (Fn(x)).

For each variable, the ecdf() function is used to compute the empirical cumulative distribution function. The green line shows the ECDF for ecologicalStatus, while the red line indicates the ECDF for eco_status_7.

The Kolmogorov-Smirnov test (ks.test()) is then used to determine if the two variables are from the same distribution or not. The test results in a D-statistic value of 0.375 and a very modest p-value (6.695e-06), indicating that we reject the null hypothesis that the two variables come from the same distribution.

The graph shows that the green line (ECDF for eco_status_7) is consistently greater than the red line (ECDF for ecologicalStatus) across a wide range of values. This suggests that eco_status_7 has greater values than ecologicalStatus. The difference between the two ECDFs is most noticeable at the right side of the graph (about x = 0.75). This implies that the eco_status_7 distribution has more high values than the ecologicalStatus distribution.

Overall, the graph and KS test indicate that the two variables are distinct, with eco_status_7 having greater values than ecologicalStatus. hist(dataset_MA334_split$BD7_change) # the distribution of the BD7 change

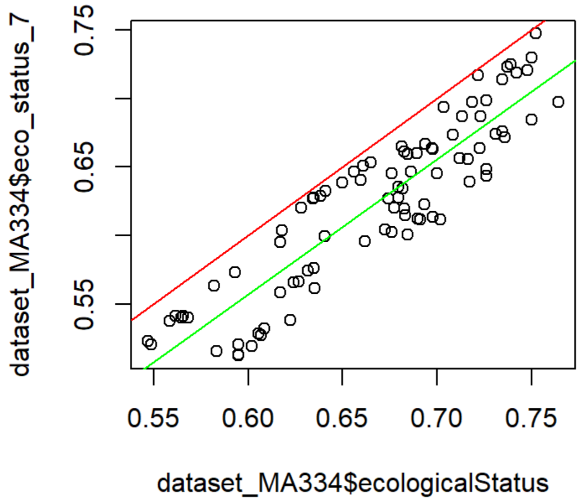


Simple linear regression

The scatter plot depicts the association between ecologicalStatus and eco_status_7. Each dot in the graph represents a single observation from the dataset. The horizontal axis (x-axis) represents the variable ecologicalStatus, and the vertical axis (y-axis) represents the variable eco_status_7.

The regression line in the plot illustrates the linear relationship between the two variables.Because the data points do not lie exactly on a straight line, the regression line does not touch the majority of the scattered points, indicating that the relationship between the two variables may be variable. The regression line begins at 0.55 on the x-axis and 0.54 on the y-axis, suggesting that when the ecologicalStatus value is 0.55, the projected value for eco_status_7 is 0.54. Similarly, the regression line terminates at 0.75 on the x-axis and 0.76 on the y-axis, suggesting that when the ecologicalStatus value is 0.75, the projected value for eco_status_7 is 0.76.

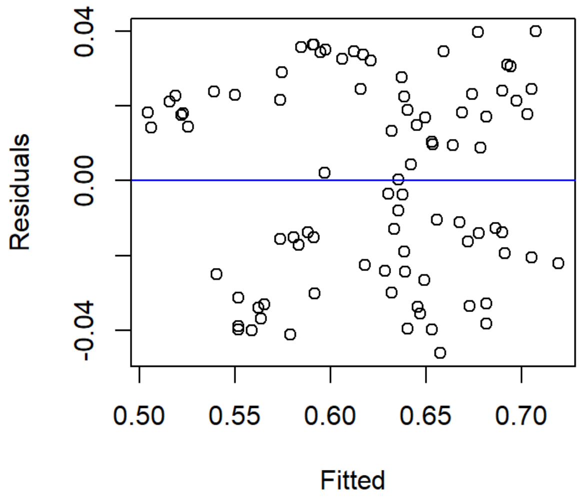
The green line represents the best-fitting linear regression line to the data. It traverses the scattered sites and estimates the link between the two variables. The x-axis shows ecologicalStatus values ranging from 0.55 to 0.75, and the y-axis shows eco_status_7 values ranging from 0.50 to 0.76.



The residuals (the difference between the observed values and the values predicted by the linear regression model) are plotted against the fitted values (the values predicted by the model) in this code.

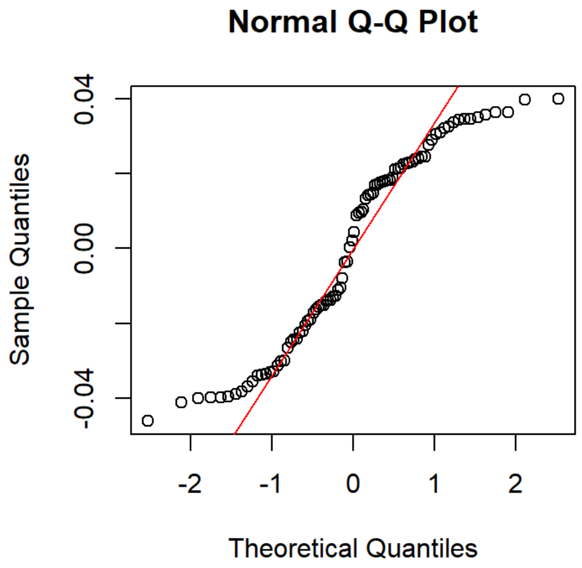
The fitted values are shown on the x-axis and range from 0.50 to 0.70 in 0.05 increments. The residuals are shown on the y-axis, and they range from -0.04 to 0.04 in 0.02 increments. The jitter() function is used to add a little amount of random noise to the fitted values, making any patterns in the data easier to notice.

The blue line shows the line of zero residual error and is a horizontal line at y=0. It is useful to understand how far the real residuals stray from the zero error line.



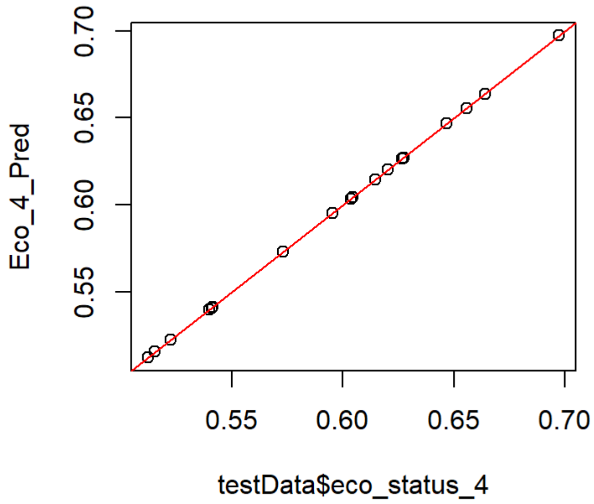
The function qqline() is used to add a reference line to the plot. This line is drawn on the same scale as the x and y axes and aids in identifying departures from normality. In this scenario, the line is coloured red and goes through the scattered points from x-axis -2 to y-axis -0.020, ending at x-axis and y-axis 0.015.

If the data has a normal distribution, the points on the QQ plot should be near to the reference line. If the points stray greatly from the line, it may suggest that the data is not regularly distributed. In this scenario, the residuals appear to follow a reasonably normal distribution, with minor aberrations around the tails.

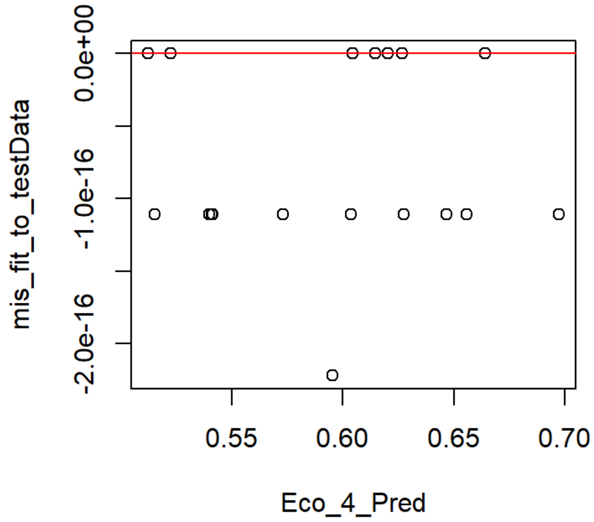


Multiple Linear Regression

The code generates a scatter plot comparing the anticipated values Eco_4_Pred to the actual values testData$eco_status_4. The x-axis shows the actual values, and the y-axis shows the expected values. The red line is a reference line with a slope of one and an intercept of zero, indicating a flawless forecast. If the predicted and actual values are exactly correlated, then all of the distributed points will fall on this line.

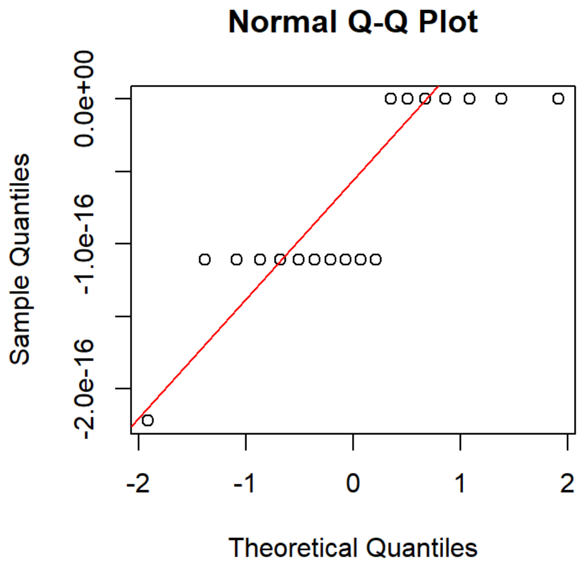


In this code, we plot the model’s misfits to the test data against the anticipated values. The x-axis shows the expected values for eco_status_4, and the y-axis shows the misfits, which are the disparities between the predicted and actual values for eco_status_4 in the test data. The red line in the plot depicts the ideal fit with no misfits, with a slope of 0 and an intercept of 0. The red line goes through the origin because if the expected and actual values are the same, the misfit is 0. The plot reveals that the misfits are relatively minor and scattered randomly around the horizontal line, indicating that the model fits the test data well.



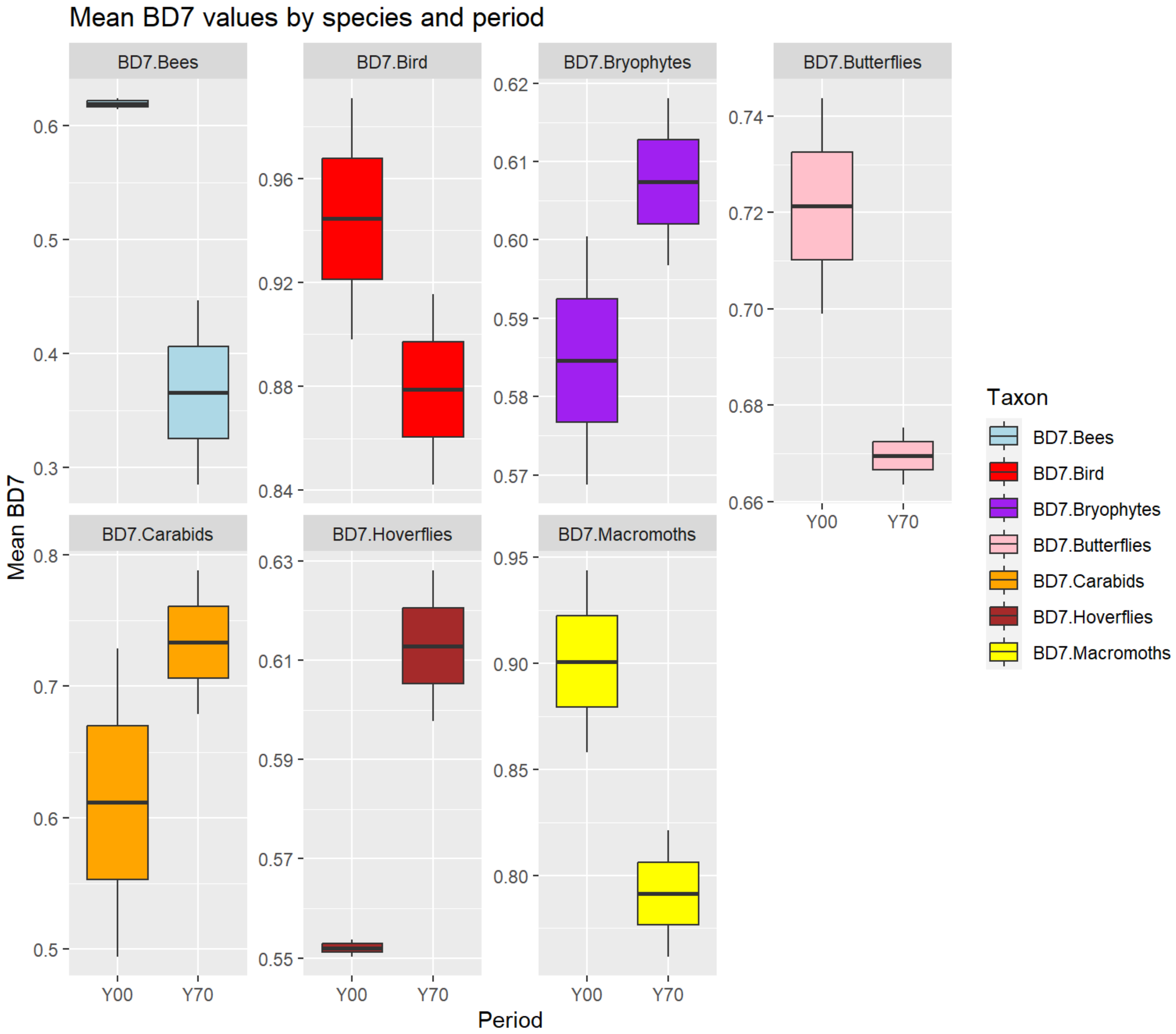
In prediction, this function checks the residuals for normalcy. The qqnorm function is used to generate a normal probability map of the residuals, where the x-axis represents the theoretical quantiles of a normal distribution with mean zero and standard deviation one, and the y-axis represents the residual sample quantiles. The qqline function is used to draw a reference line through the first and third quartiles of the normal distribution.

The figure in this case demonstrates that the residuals are nearly normally distributed, as most of the dots follow the diagonal line. The red line in the centre of the plot indicates that the residuals are roughly regularly distributed.



Open Analysis

The code generates seven box plots, one for each taxon, with the x-axis indicating the two periods of interest (Y70 and Y00) and the y-axis representing the mean BD7 value for each taxon for those times. The fill colour is used to distinguish between the various taxon categories in the plot.



Conclusion

According to the findings, there appears to be a considerable difference in biodiversity based on 7 and 11 taxonomic groups. The t-test result indicates that the mean difference in biodiversity between the two groups is statistically significant. Furthermore, the Kolmogorov test result indicates that the biodiversity distributions based on 7 and 11 taxonomic groups are not the same.

The correlation matrix analysis indicates some strong relationships between the selected variables. For example, there is a positive link between biodiversity and macromoths, which suggests that macromoths could be a useful predictor of biodiversity.

Further investigation and analysis may be required to properly comprehend the links and patterns in the data.

##References

1. https://moodle.essex.ac.uk/course/view.php?id=15074 (https://moodle.essex.ac.uk/course/view.php?id=15074)
2. https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/Journal%20of%20Applied%20Ecology%20-%202016%20-%20Dyer%20-%20Developing%20a%20biodiversity%E2%80%90based%20indicator%20for%20large%E2%80%90scale%20environmental.pdf (https://moodle.essex.ac.uk/pluginfile.php/2009058/mod_folder/content/0/Materials/Journal%20of%20Applied%20Ecology%20-%202016%20-%20Dyer%20-%20Developing%20a%20biodiversity%E2%80%90based%20indicator%20for%20large%E2%80%90scale%20environmental.pdf)