Univariate analysis and basic R programming

1. a) Summary for all my 5 variables
("Butterflies","Hoverflies","Ladybirds","Grasshoppers_._Crickets","Vascular_plants")

My-group	**Min**	**Quantile_first**	**Median**	**Mean**	**Quantile_third**	**Max**
Butterflies	0.32	0.79	0.89	0.87	0.97	1.39
Grasshoppers-Crickets	0.07	0.49	0.62	0.63	0.79	1.59
Hoverflies	0.12	0.57	0.7	0.68	0.81	1.15
Ladybirds	0.06	0.45	0.64	0.61	0.8	1.84
Vascular_plants	0.42	0.72	0.79	0.79	0.86	1.2

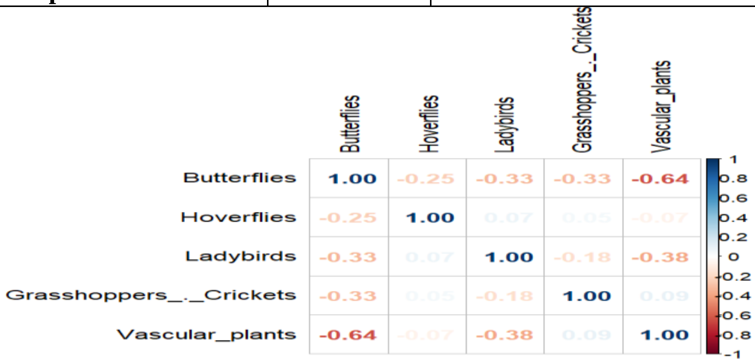
b) **20% Winsorized mean**

My-group	**Min**	**Quantile_first**	**Median**	**Mean**	**Quantile_third**	**Max**	**Winsorized_Mean**
Butterflies	0.32	0.79	0.89	0.87	0.97	1.39	0.88
Grasshoppers-Crickets	0.07	0.49	0.62	0.63	0.79	1.59	0.69
Hoverflies	0.12	0.57	0.7	0.68	0.81	1.15	0.62
Ladybirds	0.06	0.45	0.64	0.61	0.8	1.84	0.64
Vascular_plants	0.42	0.72	0.79	0.79	0.86	1.2	0.79

The Given summary of the major summary statistics for the variables in BD5 group. They include Min, First Quartile, Median, Mean, Third Quartile, and Maximum. It also introduced into the branch a 20% Winsorized Mean statistic in so as to offer a very strong measure of central tendency, and take care of any possible outliers. This table ensures a comprehensive understanding of the nature of properties of each variable as well as gives an insight into the nature of the distributions of the variables. Winsorized mean for Butterflies (0.88) lies just below the standard mean value (0.87), thereby indicating possible sway of outliers when central tendency is considered. Grasshoppers/Crickets have a wider spread (0.07 to 1.59) that typifies more variation harbored within the biodiversity. Ladybirds show a smaller range of 0.06 to 1.84 as compared to hoverflies thus indicating possibility for stability. The higher mean hoverflies 0.68 compared to ladybirds 0.61 indicates larger density of hoverflies present which help in intensifying the richness of biological landscape and thus provide more preferred habitat for hoverflies. Winsorized mean is developed to reduce the impact of the outliers so as to deliver a robust measure of central tendency. Such an alternative improves the conventional summary statistics since, in essence, it provides the user with a reliable understanding of variable features.

2) **Correlation between 5 variables**: - a) **Table** b) **Corr-plot**

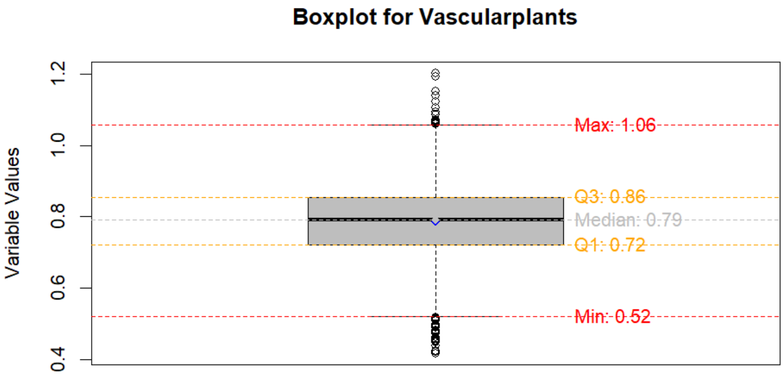
	Butterflies	**Grasshoppers-Crickets**	**Hoverflies**	**Ladybirds**	**Vascular-plants**
Butterflies	1.00000000	0.2797315	0.1851943	0.2425145	0.06627743
Grasshoppers-Crickets	0.27973153	1.0000000	0.3989119	0.4223757	0.33977413
Hoverflies	0.18519428	0.3989119	1.0000000	0.3092311	0.18153709
Ladybirds	0.24251447	0.4223757	0.3092311	1.0000000	0.37269174
Vascular-plants	0.06627743	0.3397741	0.1815371	0.3726917	1.00000000



Cor-relation value Range	Relationship	Variable pairs information
-1 to -0.7	Strong-negative linear relationship	Butterflies and vascular plants (-0.64)
-0.7 to -0.3	Moderate-negative linear relationship	Butterflies with hoverflies (-0.25), ladybirds (-0.33), grasshoppers/crickets (-0.33),
-0.1 to +0.1	Negligible-linear relationship	Ladybirds with grasshoppers/crickets (-0.18)
+0.1 to +0.3	Weak-positive linear relationship	Hoverflies with ladybirds (0.07) and grasshoppers/crickets (0.05); ladybirds with vascular plants (0.09)
+0.3 to +0.7	Moderate-positive linear relationship	None observed in this table
+0.7 to +1.0	Strong-positive linear relationship	None observed in this table

The correlation table brings out the relationships that are between every pair of variables in this BD5 group. And the linear relations and their direction and strength are presented through use of the correlation coefficients which generally will vary between -1 and 1. One of the most important findings is about the strong negative relation of butterflies to the vascular plants and in contrast among other variables the correlations being frequent weak to moderate. It helps understand how the variables in the BD5 group are related. Since generally correlations between those variables are weak to moderately highlights absence of a strong linear relationship tendency between them.

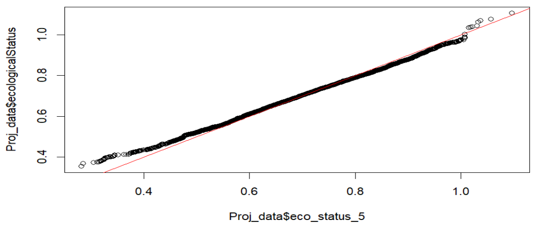
3) **Box-plot for Variable Butterflies:**

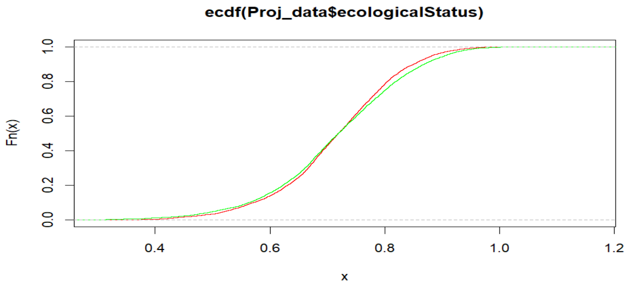


Hence likely a boxplot graphically represents the distribution of vascular plants in group BD5. With the two main points represented by the boxplot being that the whiskers could be seen where the median value of 0.79 denotes central tendency, and the rest of the diagram's farthest spread. Since by the length of upper whisker is greater than lower one, therefore suggesting skew to the right. In boxplot of vascular plants, there is quartiles, central tendencies and existence of outliers. A larger concentration of values on the top end of the distribution is indicated by the skewness to the right.

Hypothesis tests

Asymptotic two-sample Kolmogorov-Smirnov test:

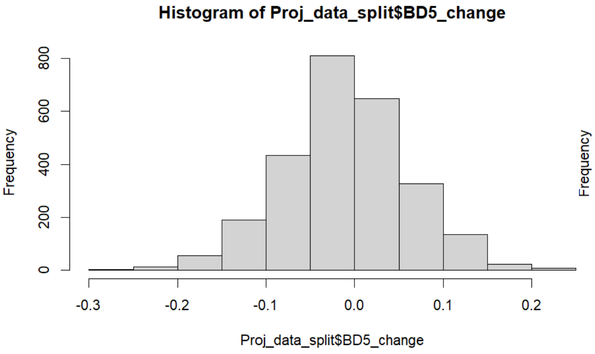


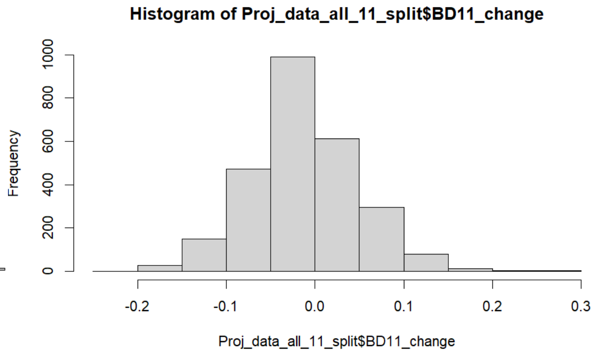


The most paths of the QQ-plot give the normal distribution and in few paths discrepancies that are assumed to be slightly deviated with some outliers, Empirical Cumulative Distribution Function. The distributions between BD5 to BD11 ecological status compared with the spread of the dataset are closely almost similar.

Conclusion: Thus, the KS test is a nonparametric test which tells about both sensitivity towards differences of location as well as shape of empirical cumulative distribution functions. Given, p-value - 0.0001922, while the p-value is less significance level (ex. 0.05) there is an evidence to reject a null hypothesis which is lack consideration difference between two distributions statement. The break on histogram's BD5 will suggest that as most of the changes are happening at about 0.0, at values approaching −0.3 or 0.2 fewer such change is being executed.

One sample T-Test :-





This would mean implying that the changes between one year and another are roughly symmetrical in nature but slightly positively skewed. The histogram is usually in a symmetric nature and the changes are distributed in a broad cast of values between -0.2 to 0.3 though many of the peaks tend to be concentrated mainly between 0.0 and 0.1 for BD11. This shows positive changes of at least as large magnitude in the former and negative changes for the latter relative to zero. Outliers on the extremes indicate exceptional data points falling in the tails of the average range. General information overall obtainable in the histogram were whether it is either symmetric or skewed, spread and more so outliers in data.

Conclusion: One-sample t-test is used for establishing if the mean of a single sample significantly differs as compared to an already known or assumed population mean after going through the process and application. From One sample T-test, respective p-value = 1.868e-13 (BD5) and p-value < 2.2e-16(BD11) respectively both Reject null hypothesis. There is evidence on significance difference existence between respective two distribution.

Contingency table/comparing categorical variables:

BD11 and BD5 are tables that give me a clear breakdown of the count of locations either based on an increase or No increase in Bio-Diversity which is shown below.

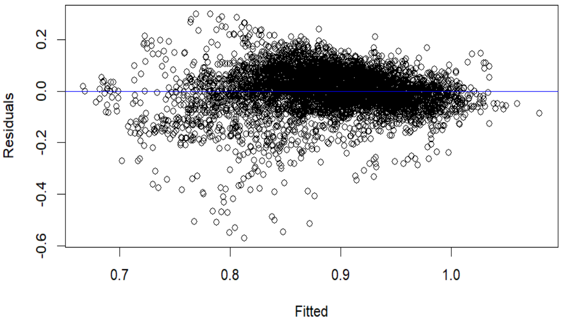
Bio-diversity increase (BD11up)	Count	Bio-diversity increase (BD5up)	Count
No Increase (0)	1638	No Increase (0)	1502
Increase (1)	1002	Increase (1)	1138

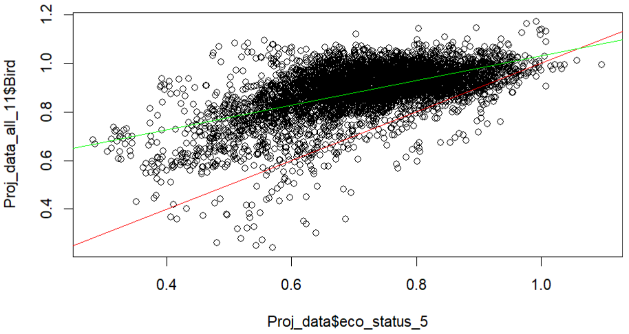
	Biodiversity 'BD5up' - No Increase	Biodiversity 'BD5up' - Increase
Biodiversity 'BD11up' - No Increase	1254	384
Biodiversity 'BD11up' – Increase	248	754

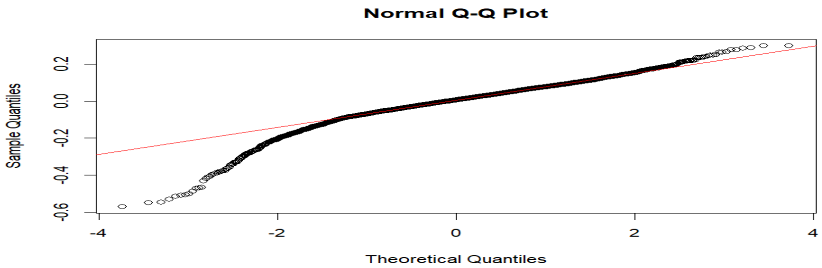
The above contingency table looks into joint distribution of location with respect to whether biodoversity increased or did not change for 'BD11' and 'BD5' variables.

The independence of study between the two categorical variables was tested using the chi-squared test or equivalently log-likelihood ratio test. The corresponding statistic is chi-squared with 1 degree of freedom and has the value 678.25 that leads to the p-value less than 2.2e-16. So, one can make a conclusion that strong dependence among various statuses of increasing biodiversity ‘BD11’ and ‘BD5’ exist.
Odds Ratio: For every unit increase in both BD11 and BD5, the odds that an event will occur is almost 9.93 times higher than not occurring.
Sensitivity: The predictive value for a positive rising both BD11 and BD5 is about 75.25% on all cases.
Specificity: 76.56% of the time, the model properly detects negative instances (both BD11 and BD5 not growing).
Youden's Index: The summary measure of diagnostic ability of the test. This corresponds to a sensitivity and specificity below 70% while for the value of 0.52, however, points at a relatively good separated overall performance of the model in separating between the two classes.
Conclusions: The model odds ratio would imply statistically significant association exists when both rise in BD5 and BD11 are on the rise.The overall model reflects good ability in identification of case when BD11 is on rise as well where it didn't rise(sensitivity) and also does not rise at all (specificity). Youden's Index, which considers both sensitivity and specificity, points to the model's strong overall performance.

Simple linear regression





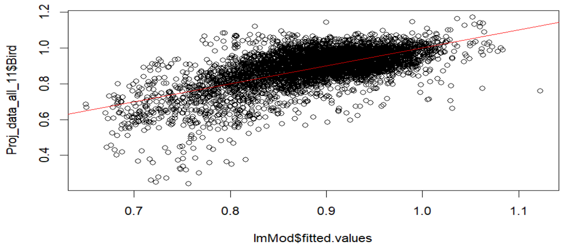


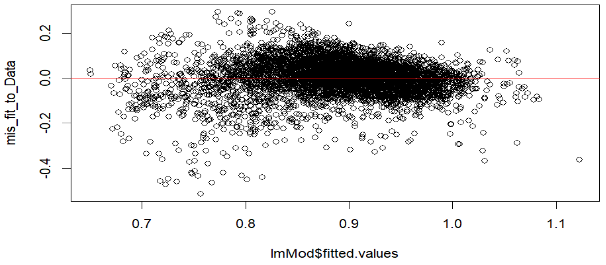
The scatterplot points out the association between Bird and eco_status_5. It helps illustrate how Bird variable changes with to eco_status_5 variable. It is clearly shown as eco_status_5 increases, so does Bird in a positive correlation. A line of linear regression (green line) has been fitted to the data. In this instance, the line of best fit looks like the line of least possible discrepancy between predicted and observed values. The estimate for the intercept (0.5248) shows the estimated value of Bird when eco_status_5 equals zero, whereas slope estimate (0.5056) indicates a change in Bird that can be attributed to a one-unit rise in eco_status_5. Expected value of Bird w.r.t eco_status_5 equals to zero. The slope - the change in Bird with respect to one unit change in eco_status_5.

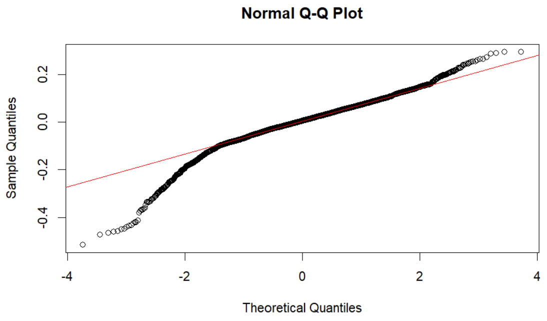
Interpretation: Given the positive slope (0.5056), Eco_status_5 and Bird are positively correlated. The Bird is expected to increase by about 0.5056 for every one unit of increase in the eco_status_5. The R-squared value (0.3248) gives a measure of how well the model fits the data and proportion of the variation in the Bird, which can be explained by eco_status_5 as about 32.48%. Given p-value = 2.2e-16, The model is statistically significant as both coefficients are very significant ($p < 0.05$). The residuals summarize the spread between observed and expected Bird values by standard deviation, 0.08753 over 5278 degrees of freedom.

Conclusion: The basic linear regression model shows strong positive between eco_status_5 with Bird. About moderate variation of Bird is explained by the basic linear regression model based on eco_status_5. The correctness of the regression analysis is backed by the diagnostic charts as they give an understanding how the model assumptions are correct.

Multi linear regression







The purpose of this Multi Linear Regression Analysis is to model the Bird species richness (BD1) using account for abundances with five various taxonomic groups i.e. butterflies, hoverflies, ladybirds, grasshoppers/crickets and vascular plants. LmMod — AIC's very low chi-square value of -11370.58 shows that the model fitting to the data is good enough.

Interpretation: Relationship between BD1 and ranked abundances in taxonomic categories and species richness of upland birds. The intercept (0.51) in base terms represents estimated BD1 when all predictors are zero for purposes of understanding favorable coefficients exactly display favorable relationships with BD1 among taxonomic groupings. Precisely, one-unit increase relates with the projected increase on butterflies (0.16), hoverflies (0.12), ladybird (0.17) or vascular plants (0.08) with projected increase on BD1. Whilst BD1 decrease one- unit increase grasshoppers/crickets (-0.03) mild relates.

This is clearly brought out in analyzing the model variance whatever perceives from within models early examined, that is how well the former explains BD1 variance observed from distribution of their residuals. From the residuals scatter plots and quantile-quantile plots does not show visible mean patterns thus meet the model assumptions.

For instance in Multiple R-squared, the model describes almost 40.25% of the variation in BD1 but then further shows that the F-statistic is highly significant ($p<2.2\times10-16$) thus indicating that atleast one predictor variable significantly explains BD1. With the introduction of multiple linear regression analysis, this provides the predictive mechanism by which that Bird species richness (BD1) can actually be calculated from these independant variables as well as an insight into how the different taxonomic grouping contribute to the overall BD1.

Comparing in LmMod Reduced every test the full model (lmMod) against reduced model (lmMod_reduced) where it was omitted one particular variable. This could help to understand if that omitting variable turned out degrading greatly our model fit.

Test-3 (Omit "Ladybirds"):

Summarized, that reduction removed model into "Ladybirds" further decreases cumulated AIC in total means that through this model goodness of fit improvement in comparison with full one. It is indicative that through the exclusion of the "Ladybirds," the increase of the model fit is achieved.

	DF	**AIC**
lmMod_reduced	6	-10161.40
LmMod	7	-11370.58

Test-3 (Interaction between "Ladybirds* Grasshoppers-Crickets"):

LM-MOD-INTERACTION:-

Interpretation: This interaction model " with Ladybirds* Grasshoppers-Crickets" has the lower AIC better fit compared to full model. Hence, this shows that the interaction model "with Ladybirds* Grasshoppers-Crickets" enhances the fitness of the model.

	DF	**AIC**
lmMod	7	-11370.58
lmMod_reduced	6	-11016.54
lmMod_interaction	8	-11383.11

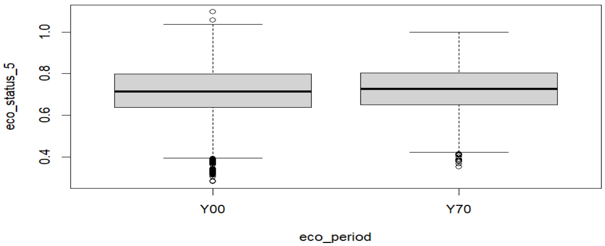
Training-Test set:-

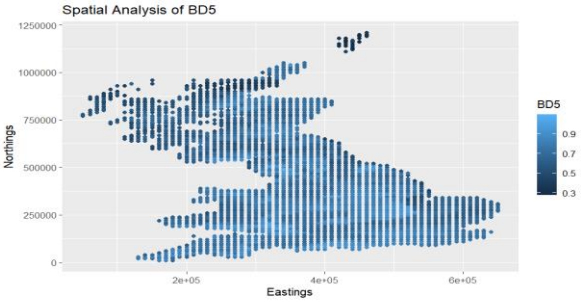
The dataset has two levels of the 'period' variable, Y00 and Y70 each with 2640 observations and it is all balanced. Justification to this can be attributed by data representation in similar periods hence much analysis can be reached. The whole dataset rows sum up to 5280 and this represents a large sample size to run diverse analyses.

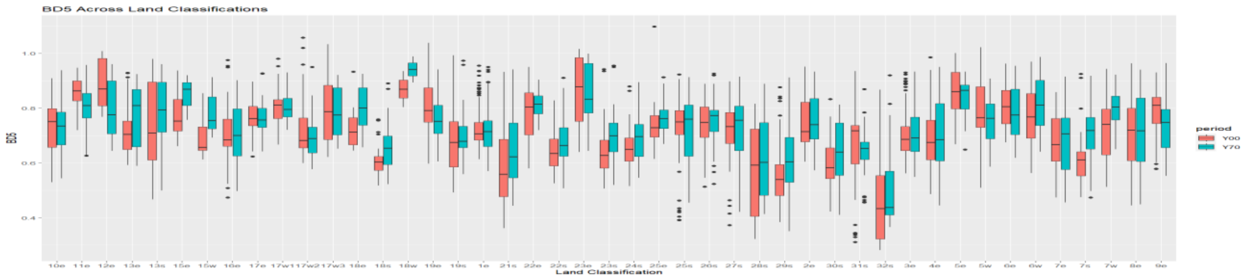
The following analysis fits a linear regression model with Y70 period serving as the training set and the Y00 era as holdout sample. The diagnostic plots help in shedding light on the assumptions that guide the fitted model. These plots include the residuals vs. fitted values plot and the QQ plot for normality. This is so because first, the points are located near the red line that shows normalcy assumptions in the QQ plot and secondly, plot called residuals vs. fitted values plot contain a read line at zero only, helps to reveal any pattern in residuals.

To evaluate the how well does the model prediction on the test set (Y00), the scatter of predicted value (Predict_00) against the actual values is plotted. For a perfect prediction within this graphic model, the red line should have reached on actual values and forecasted ones at the same time. Mean Squared Error (MSE) values point as well in that the model does great. MSE at the test set 0.005865441(Y00) is slightly larger than on the training set (Y70), and it is projected as 0.008571187. MSE computes the average squared difference between predicted values and actual values, so a higher value suggests that in total maybe the predictions would have been slightly less on target.

OPEN ANALYSIS







Period Analysis: Comparing the median eco_status_5 between Y00 and Y70, one sees that it is only modestly lower for Y00 than for Y70 suggesting perhaps that the overall ecological condition declined somewhat from 1970–1990 to 2000–2013. Though not by much. The difference is not significant as the box of the Y00 is wider than that of Y70 indicating more spread on eco_status_5 in this latter period. This means that there could have been improvements in some sectors while in others decreases could have occurred.

Spatial Analysis: x-represent Easting and the y- representing Northing- distance north from a fixed parallel. The mean value of each of the five BD5 variables at each location is represented by the color gradient on the graph. The number of butterfly, hoverfly and ladybird species is highest mean located in southern and eastern part of England and Wales and it reduces rapidly as moves to north and west for Grasshoppers/Crickets, Vascular Plants mean number is highest located in southern and eastern part of England and Wales and declines steadily in northwards and westwards. With this analysis, we can identify priority areas for the protection of species.

Land-Class Analysis: A land class analysis is as follows 45 land classes are made a total of the land classification, whereby from England there stands 21, from Scotland 16, from Wales 8 and both periods (1970–1990)x(2000–2013) compared to my five factors. There is also the suggestion that there has been an increase in suitable habitat between 1970 and 2013 as median land categorization value for butterflies tends to be slightly higher in Y00 than it is in Y70. Thereafter, however, the Y00 box widens, which therefore means that the land classes were more variable.

The hoverflies follow a similar trend, possibly with slightly higher median land categorisation value in Y00 than the butterflies suggesting possible tendency towards increasing appropriate habitat. To shed light upon this, it may be mentioned that the bigger Y00 box indicates more variability in hoverfly land classifications for the later period.

The case is however not the same for the ladybirds. Between 1970 and 2013, there could have been a decline in suitable habitat as depicted by the lower medians land categorization value of Y00 compared to Y70. Less variation in land classifications for ladybirds during the latter time is suggested by the smaller Y00 box.

The median of the land classification value here, however less in Y00 suggests a possible reduced habitat. The observation of ladybirds follow the same pattern to that of grasshoppers/crickets but the inclusion box with reference to ladybirds has a wider Y00 as compared to that of grasshoppers.

The median land classification value for the vascular plants of Y00 is little less than that for Y70 suggesting a drop in the quantile indicating possible decline in suitable habitat between the period 1970 and 2013. The box for Y00 is wider thus emphasizing higher nativity in land classifications for the later period.