Marco Rampin 12236811

2023S - Practical data analysis in ecology, biodiversity and zoology

Final version

Exercise 1

Are the observed local abundances, across all 157 surveys, of Coenonympha pamphilus, Maniola jurtina, Melanargia galathea, Pieris rapae and Polyommatus icarus related to each other?

You can explore potential relationships among observed local abundances through a correlation analysis. Using PAST (Hammer et al., 2001), I generated a correlation table, which was adjusted using Bonferroni corrections. The results reveal a significant positive correlation of moderate strength between the abundances of Coenonympha pamphilus and Polyommatus icarus (Figure 1). It's important to note that the correlation analysis assumes normally distributed residuals. However, as depicted in Figure 2, this assumption is not met. Nevertheless, due to the high significance of the correlation (p: <0.0001), I would continue to consider it statistically meaningful.

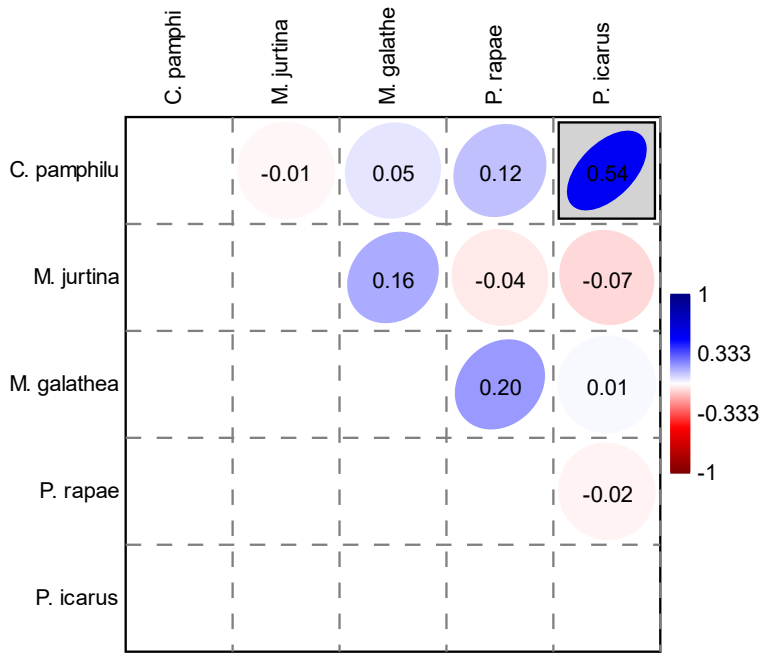


Figure 1: Correlation matrix of species abundances across all surveys. The numbers within each cell represent the correlation coefficients, depicted visually as ellipses with varying color saturation. More saturated colors indicate stronger correlations, with blue indicating positive correlations and red indicating negative ones. Grey boxed cells represent values with a p-value < 0.05 after Bonferroni correction.

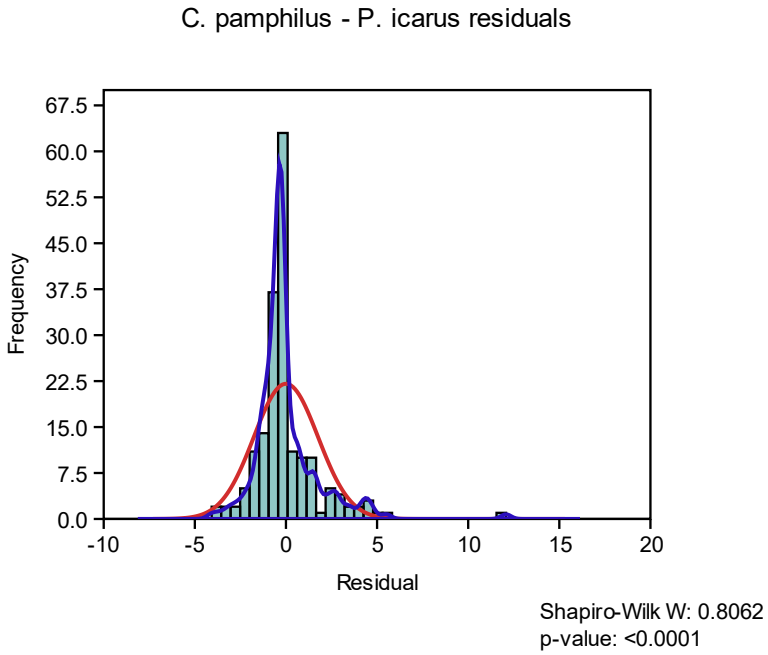


Figure 2: Residuals distribution of the correlation between C. pamphilus and P. icarus. The plot includes the red curve representing the normal distribution, the dark-blue curve depicting the kernel distribution, and the Shapiro-Wilk normality test results in the bottom-right corner

Could transformation of data improve the outcome of your analysis?

Therefore, I explored multiple transformations, including square root, logarithmic, and Box-Cox, with the latter proving to be the most effective in achieving data distribution closest to normal (Figure 3). This transformation also exhibited the highest correlation with formal normality tests (Table 1) and the fewest significant outliers.

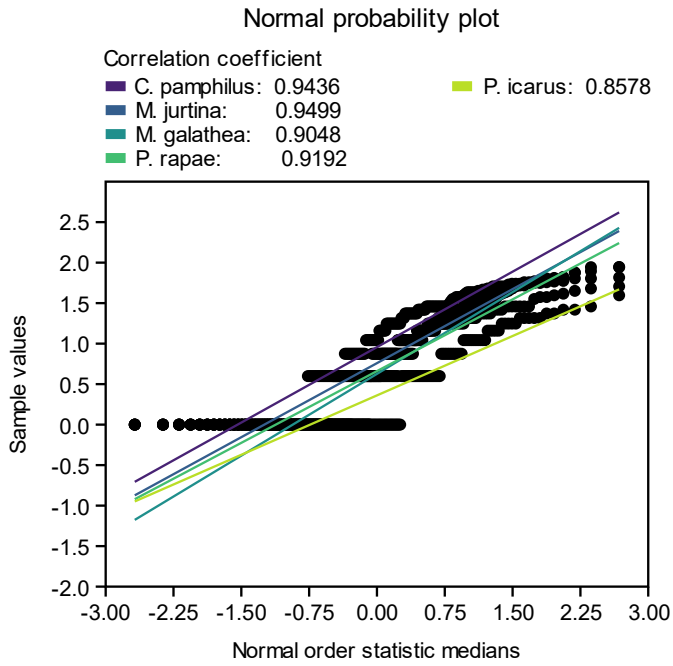


Figure 3: Correlation of transformed (Box-Cox) species abundances with normal distribution.

	C. pamphilus	M. jurtina	M. galathea	P. rapae	P. icarus
Shapiro-Wilk	0.8835	0.896	0.8125	0.8384	0.7317
p	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001

Table 1: Shapiro-Wilk test to determine if the species distribution data is significantly different from normal distribution.

Upon repeating the correlation analysis (Figure 4), it's apparent that the correlation between Coenonympha pamphilus and Polyommatus icarus (p: <0.0001) has marginally decreased compared to the initial analysis. However, a new significant relationship has emerged between M. galathea and P. rapae (p: <0.0001).

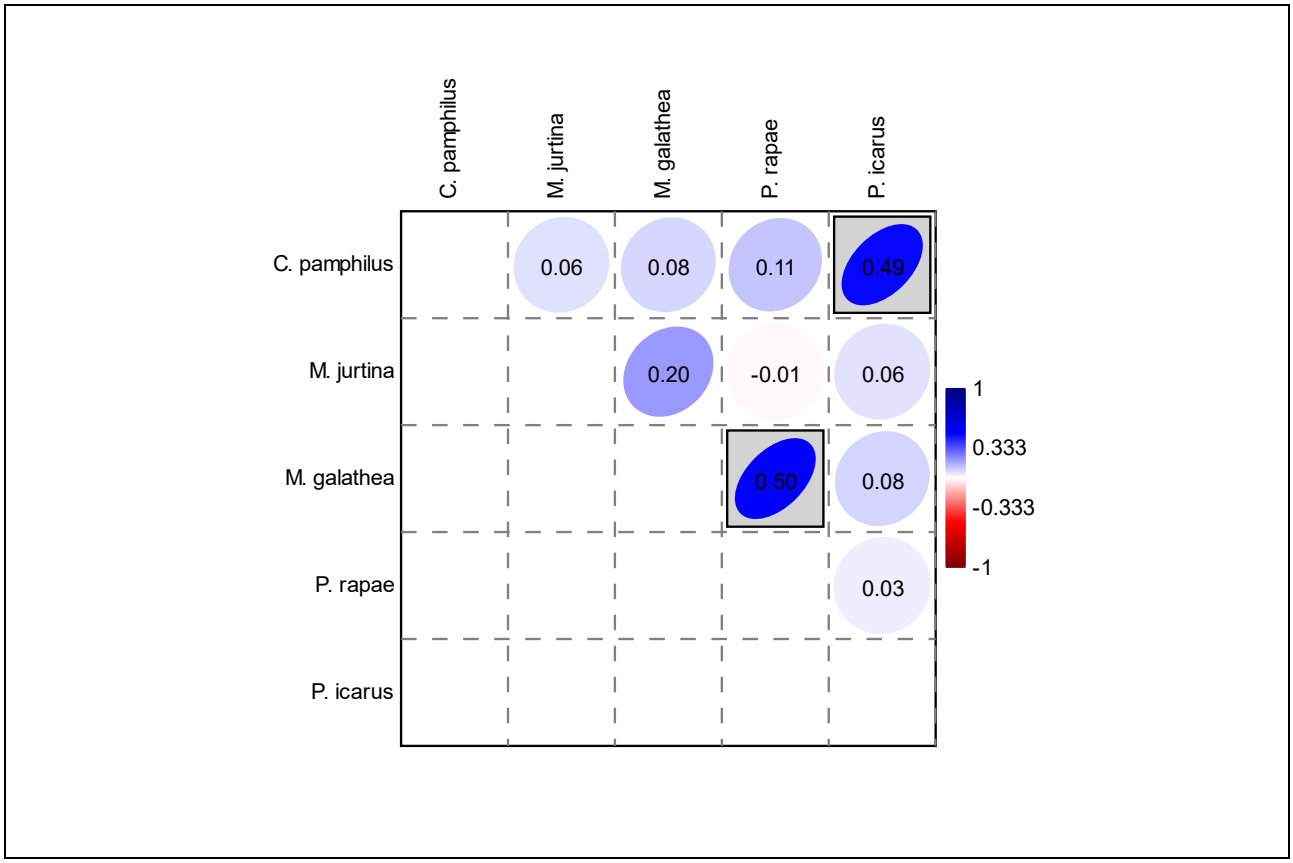
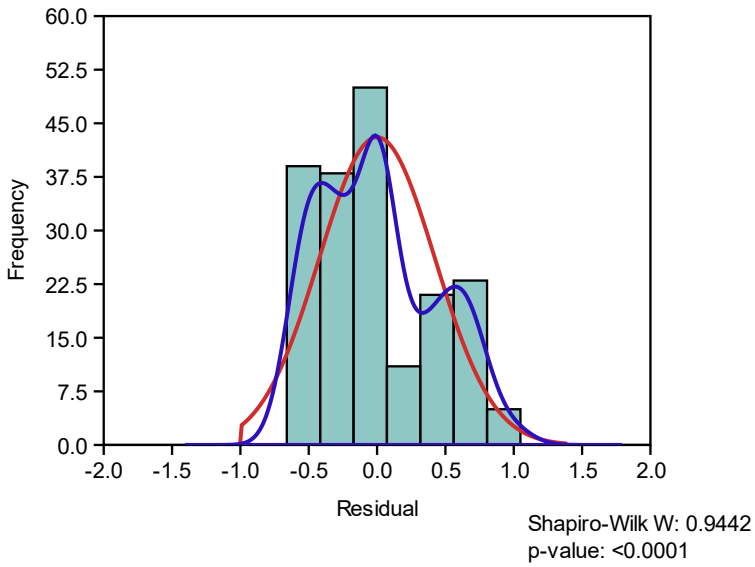


Figure 4: Correlation matrix of species abundances (abundances transformed with Box-Cox) across all surveys. The numbers within each cell represent the correlation coefficient, visually depicted as ellipses with varying saturation. Strong correlations are indicated by more saturated colors, with blue representing positive correlations and red indicating negative ones. Grey boxes highlight values with a p-value < 0.05 after Bonferroni correction.

Upon analyzing the residuals, it is evident that their distribution still significantly differs from normality. Although the curves are closer to normality, they exhibit multimodal distributions (Figure 5).

A



B

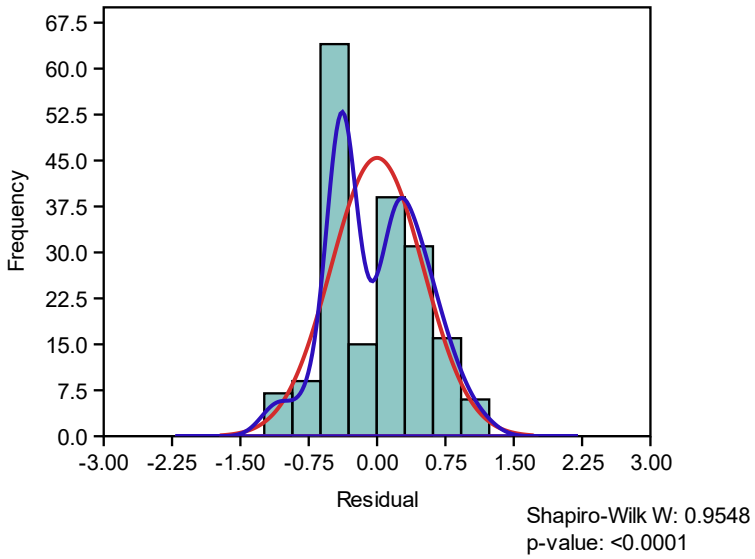


Figure 5: Residual distribution of the correlations between C. pamphilus and P. icarus (A) and between M. galathea and P. rapae (B) species abundances. The red curve represents a normal distribution, while the dark-blue curve illustrates the kernel distribution. The Shapiro-Wilk normality test results are provided in the bottom-right corners.

For C. pamphilus and P. rapae, also set up a contingency table summarizing the incidence (rather than abundance) of their occurrences across the 3 land-use types. Are there any differences, beyond random noise, between the two species in that regard?

I have compiled two tables that aggregates the incidence data for C. pamphilus and P. rapae categorizing them based on land use (table 2):

(A) C. pamphilus	present	absent
extensive	54	9
fallow	49	15
pasture	43	17

(B) P. rapae	present	absent
extensive	31	32
fallow	47	17
pasture	36	24

Table 2: Contingency tables of C. pamphilus (A) and P. rapae (B) incidences organized by land-use type.

Chi-squared resulted not significant for C. pamphilus and significant for P. rapae (Table 3):

(A) C. pamphilus
Degrees freedom: 2
Chi2: 3.6727 p (no assoc.): 0.1594
Monte Carlo p : 0.1662
Fisher´s exact
p (no assoc.): 0.1457

(B) P. rapae
Degrees freedom: 2
Chi2: 7.8673 p (no assoc.): 0.0196
Monte Carlo p : 0.0174
Fisher´s exact
p (no assoc.): 0.0194

Table 3: Chi squared results for the C. pamphilus (A) and P. rapae (B) incidence across three different land use types.

Finally, when conducting a post-hoc analysis on P. rapae incidence (Table 4), it's worth noting that although the Chi-squared analysis initially yielded overall significance, the significance diminishes when applying Bonferroni correction to the p-values.

(A) Corrected p-values	Present	Absent
Extensive	0.113	0.113
Fallow	0.0699	0.0699
pasture	1	1

(B) Standardized residuals	Present	absent
Extensive	-1.1951	1.4935
Fallow	1.2782	-1.5973
pasture	-0.0955	0.1193

Table 4: Post-hoc analysis of the Chi-squared analysis on P. rapae incidence across three different land use types. Panel (A) displays the p-values indicating significantly different incidences (corrected with Bonferroni), while panel (B) showcases standardized residuals, highlighting whether observed values are higher (positive) or lower (negative) than expected.

What can you conclude from these analyses about potential habitat preferences of the 5 target species?

Based on the previous analysis, it appears that only a few species have a distribution that is partially interrelated among themselves. However, when compared to different land-use types, no statistical significance is observed. This suggests a high likelihood that other environmental factors besides land use influence their distribution. Furthermore, these related species may have partially overlapping niches, allowing them to coexist within the same sampled areas.

Exercise 2

Compare the observed abundances per site of Maniola jurtina across the 3 land-use types and 5 sampling years. Does that tell something about possible preferences in regard to land-use? Is any temporal pattern apparent in the data?

To assess preferences for land-use and temporal patterns, a two-way ANOVA was conducted. Ideally, this analysis requires data that are normally distributed and exhibit similar variances. However, when sample sizes are equal, the impact of these assumptions is less critical, and homoscedasticity can be checked using the Levene test.

In this case, the data were transformed using the Box-Cox method, and for both land-use types and years, the variances appeared similar. Specifically, the Levene test yielded p-values of 0.2914 for land-use types and 0.3359 for years.

As a result, I proceeded with the two-way ANOVA to examine whether there are significant differences in the abundance of M. jurtina among the various land-use types and years. Both land-use and year were found to be statistically significant factors, whereas their interaction did not exhibit significance (Table 5).

FIXED-EFFECTS TWO-WAY ANOVA (unbalanced, Model II)

	Sum of sqrs	df	Mean square	F	p (same)
Land-Use:	3.10556	2	1.55278	6.016	0.003
Year:	16.9946	5	3.39891	13.17	<0.0001
Interaction:	4.0883	10	0.40883	1.584	0.115
Within:	43.6194	169	0.258103		
Total:	67.7878	186			

Table 5: Two-way ANOVA using Land-use and years as fixed-effects for M. jurtina abundances.

Subsequent post-hoc analysis using the Tukey's test revealed that the abundance in the extensive land-use type significantly differed from the other two. However, no significant difference was observed between the fallow and pasture land-use types (Table 6).

	pasture	extensive	fallow
pasture		0.0041	0.7995
extensive	0.0041		0.0238
fallow	0.7995	0.0238	

Table 6: post-hoc analysis to determine the land-use types with significantly different M. jurtina abundances. Light-red cells are the statistically significant values ($p < 0.05$).

Indeed, a violin plot visually illustrates that the pasture and fallow land-use types exhibit quite similar means and distributions, while the extensive land-use type consistently shows higher abundance levels (Figure 6).

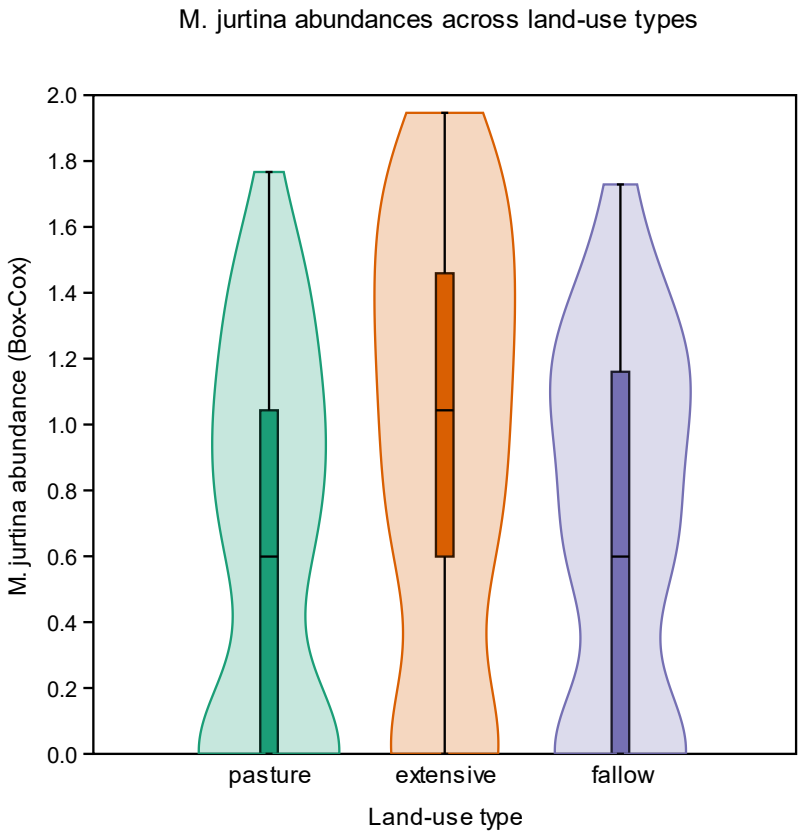


Figure 6: M. jurtina abundances across land-use types.

Meanwhile, during the Tukey's post hoc analysis for the years, the most significant differences become apparent in 2007 and 2011 (Table 7).

	2005	2006	2007	2008	2009	2011
2005		1	<0.0001	1	0.111	0.03
2006	1		0.0003	1	0.1771	0.0242
2007	<0.0001	0.0003		0.0002	0.2887	<0.0001
2008	1	1	0.0002		0.1603	0.0213
2009	0.1113	0.1771	0.2887	0.1603		<0.0001
2011	0.03	0.0242	<0.0001	0.0213	<0.0001	

Table 7: post-hoc analysis to determine the years with significantly different M. jurtina abundances. Light-red cells are the statistically significant values ($p < 0.05$).

When graphed, it becomes more evident that 2007 exhibited lower abundance, while 2011 showed higher abundance compared to the other years (Figure 7).

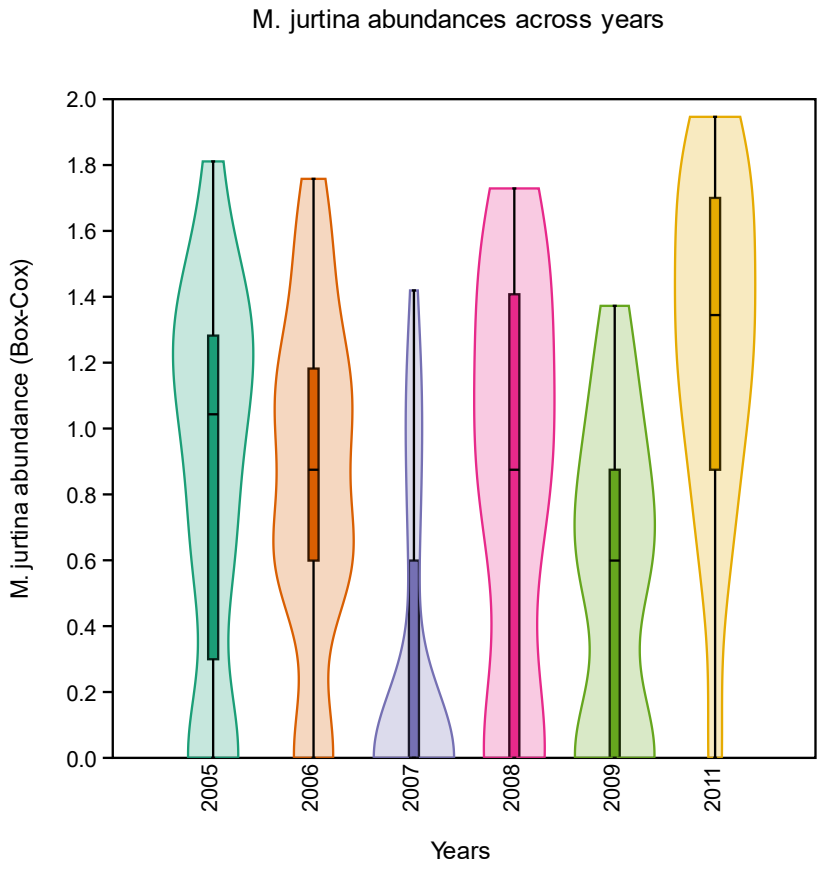


Figure 7: M. jurtina abundances across years.

How do results change if you model year as a random (instead of a fixed) factor?

When considering the year variable as random effects, the significance of both land-use and years marginally decreased. However, they still maintain statistical significance, and there are no discernible differences in the post-hoc analysis (Table 8).

RANDOM-EFFECTS TWO-WAY ANOVA (unbalanced, Model II)

	Sum of sqrs	df	Mean square	F	p (same)
Land-Use:	3.10556	2	1.55278	3.798	0.0244
Year:	16.9946	5	3.39891	8.314	<0.0001
Interaction:	4.0883	10	0.40883	1.584	0.115
Within:	43.6194	169	0.258103		
Total:	67.7878	186			

Table 8: Two-way ANOVA using years as a random-effect for M. jurtina abundances across samples.

Repeat the analyses for the relative contributions of this species to the respective sum of butterfly sightings per survey locality. Do emergent patterns look different than with absolute observation records? What would you infer from that?

Following the transformation of the observed abundances of M. jurtina into relative contributions, I experimented with various transformations to enhance normality. Ultimately, the square root transformation proved to be the most effective in approximating a normal distribution (Figure 8).

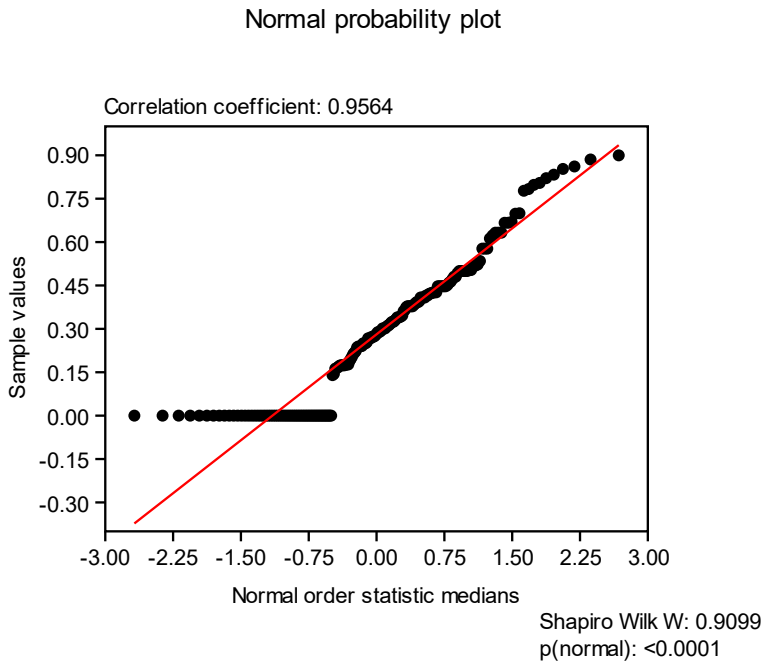


Figure 8: Correlation of transformed (sqrt) M. jurtina abundance data with normal distribution. In the bottom-right corner the Shapiro-Wilk normality test results.

Also in this case I checked the Levene tests and meanwhile for land-use type it wasn't significant (0.1066), it was when considering years (0.0108). Nonetheless, when performing a two-way ANOVA with fixed factors, the results are almost identical to the ANOVA performed previously on observed abundances; the same is true also when observing to the Tukey's post hoc and at the random factor analysis (Table 9).

(A) FIXED-EFFECTS TWO-WAY ANOVA (unbalanced, Model II)

	Sum of sqrs	df	Mean square	F	p (same)
Land-Use:	0.384426	2	0.192213	5.25	0.0061
Year:	3.72638	5	0.745276	20.36	<0.0001
Interaction:	0.58572	10	0.058572	1.6	0.1103
Within:	6.18709	169	0.03661		
Total:	10.8836	186			

(B) RANDOM-EFFECTS TWO-WAY ANOVA (unbalanced, Model II)

	Sum of sqrs	df	Mean square	F	p (same)
Land-Use:	0.384426	2	0.192213	3.282	0.04
Year:	3.72638	5	0.745276	12.72	<0.0001
Interaction:	0.58572	10	0.058572	1.6	0.1103
Within:	6.18709	169	0.03661		
Total:	10.8836	186			

Table 9: Two-way ANOVA using Land-use and years as fixed-effects (A) and random-effect (B) for M. jurtina relative contribution across samples.

This example suggests that utilizing relative contributions doesn't reveal any significant differences compared to using observed abundances for analysis. Nevertheless, it's not entirely ruled out as a potential method for improving data distribution normality. This approach may be particularly useful when achieving a normal distribution remains challenging, even after employing various transformations. Notably, relative contributions inherently mitigate the impact of extreme significant values and could serve as a means to standardize species abundances, particularly in cases where sampling effort varies among samples.

Exercise 3

Can the total number of butterfly sightings per survey, aggregated across all species, be modelled using local nectar flower abundance as predictor?

I aggregated the abundances of all species and applied a Box-Cox transformation to achieve a normal distribution of the data (Shapiro-Wilk W: 0.9971 p: 0.979). I decided not to transform the local nectar flower abundance since I observed no improvement in its distribution, and it resulted in a poorer residual distribution when I performed a linear regression with the total abundance. You can see the scatterplot generated from this data using ordinary least squares (OLS) regression in Figure 9, and Figure 10 illustrates its residuals.

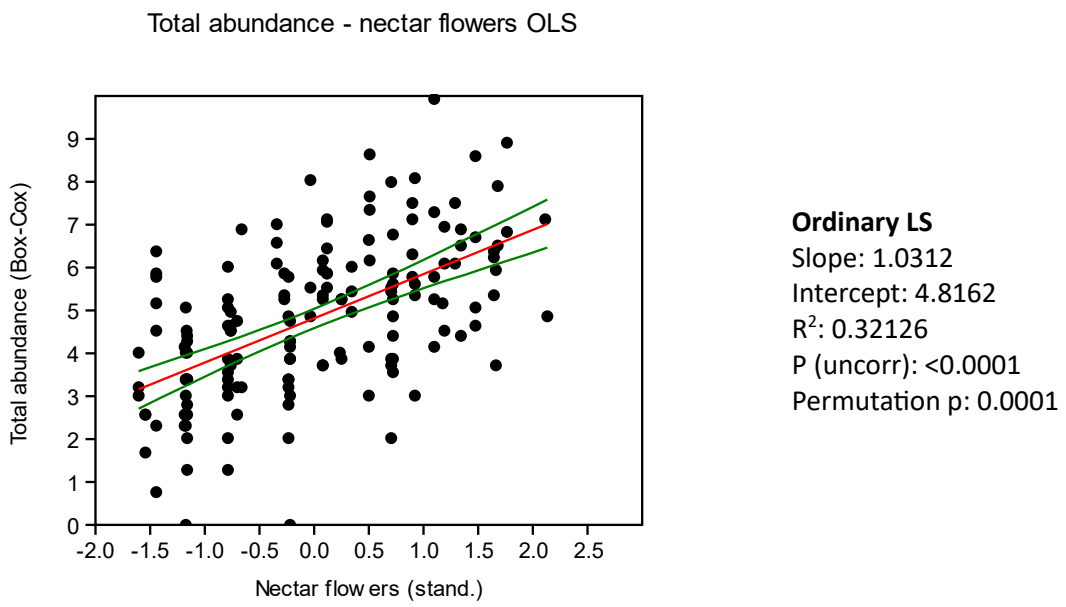


Figure 9: Linear regression (OLS) between the Box-Cox transformed total abundance and the standardized data of Nectar flowers. The red line represents the regression line, while the green lines depict the 95% confidence interval, obtained through bootstrapping.

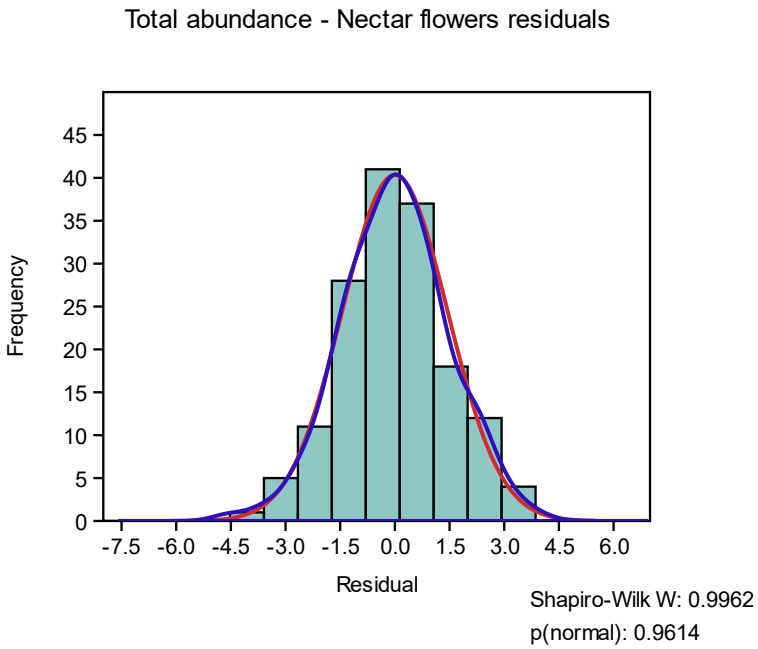


Figure 10: Residual distribution of the total abundance - nectar flowers OLS. The red curve represents the normal distribution, while the dark-blue curve shows the kernel distribution. In the bottom-right corner, you can find the results of the Shapiro-Wilk normality test.

As observed in the analysis, the correlation between nectar flower abundance and total butterfly abundance is indeed highly significant. However, it accounts for only 32% of the total butterfly abundance.

Check whether a relationship between abundance and nectar availability also holds true at the level of two common individual species (viz. P. rapae and M. galathea).

To ensure that the residuals of the linear regression approximated a normal distribution as closely as possible (compared to using log and Box-Cox transformations), I applied a square root transformation to the data for M. galathea and P. rapae.

For M. galathea, nectar flower abundance is a less potent predictor, explaining only one-third of the variability compared to total abundance (Figure 11). For P. rapae, instead, nectar flower abundance is not a significant predictor at all (Figure 13). It's essential to note that both of these analyses are less reliable than those involving total abundance because the residuals of the linear regressions do not follow a normal distribution (Figure 12 and 14).

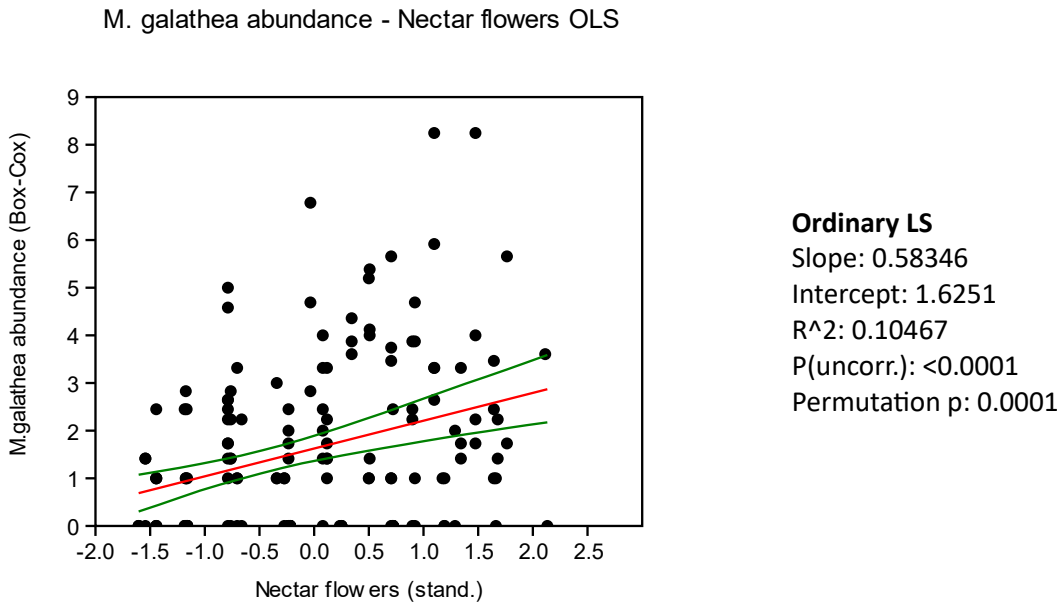


Figure 11: Linear regression (OLS) between the square root transformed M. galathea abundance and the standardized data of Nectar flowers. The red line represents the regression line, while the green lines depict the 95% confidence level obtained from bootstrapping.

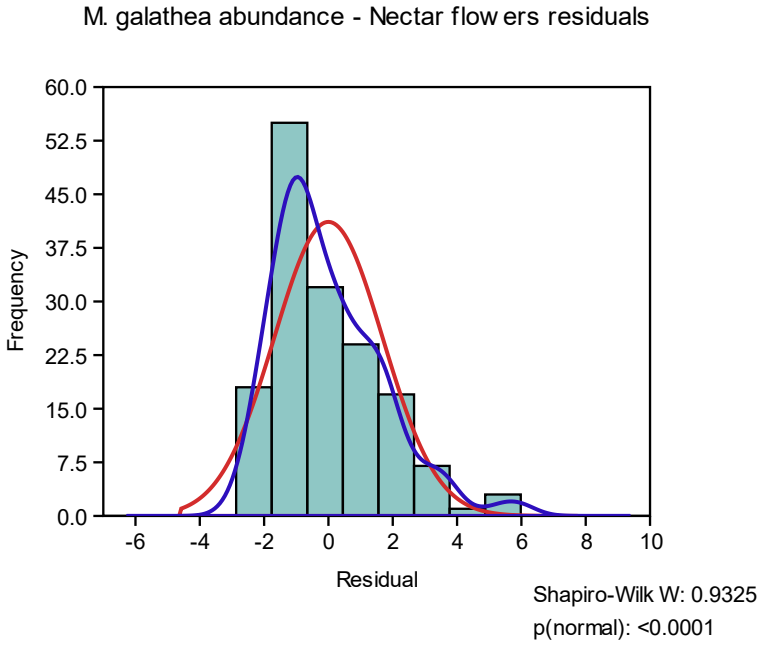


Figure 12: Residual distribution of the M. galathea abundance - nectar flowers OLS. The red curve represents the normal distribution, while the dark-blue curve depicts the kernel distribution. In the bottom-right corner, you can find the Shapiro-Wilk normality test results.

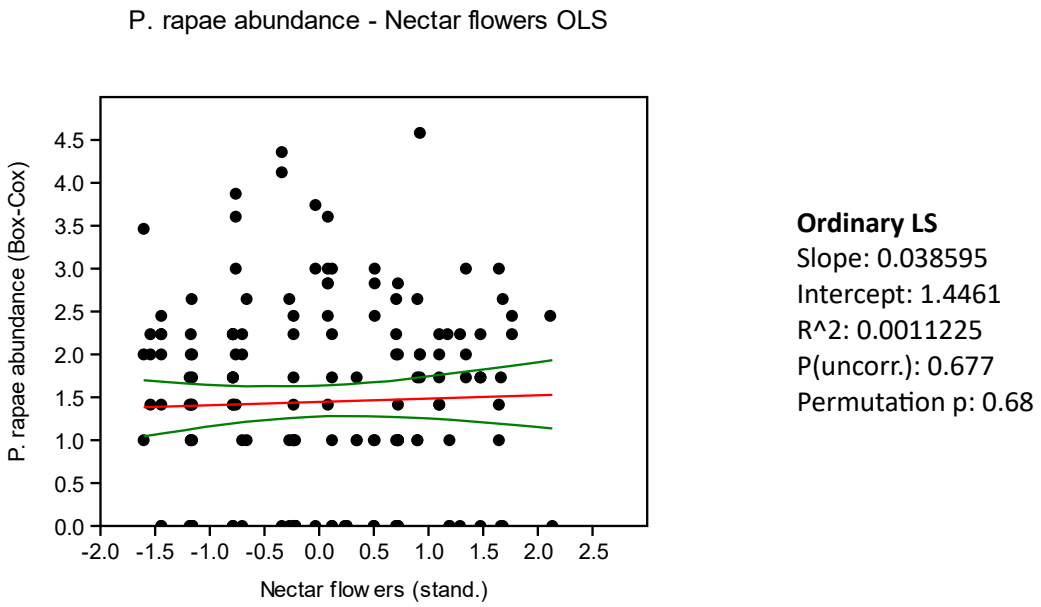


Figure 13: Linear regression (OLS) between the square root transformed P. rapae abundance and the standardized data of Nectar flowers. The red line represents the regression line, while the green lines represent the 95% confidence level obtained from bootstrapping.

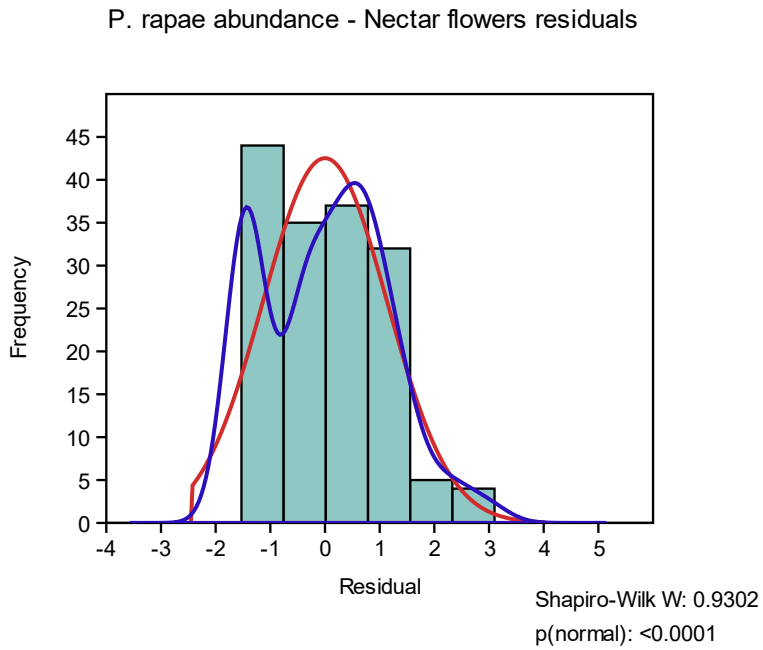


Figure 14: Residual distribution of the P. rapae abundance - nectar flowers OLS. The red curve represents the normal distribution, while the dark-blue curve represents the kernel distribution. In the bottom-right corner, you can find the results of the Shapiro-Wilk normality test.

Due to the high amounts of zero values, which may be influenced by various factors, I conducted a new analysis by excluding them. Interestingly, for both species, this adjustment led to a decrease in the slope of the regression line, bringing it closer to zero. Additionally, the R^2 values and significance levels substantially decreased. This could mean that the frequency of the zero values is higher at low Nectar flowers values, thus enhancing the correlation between species abundances and Nectar flowers availability.

Did nectar flower abundance consistently differ between the 3 habitat types over the 5 study years for which data are available?

The data and residuals of the ANOVA do not exhibit a normal distribution (Figure 15). However, the variances are consistent, and there are highly significant differences between land-use types (Table 10). Post-hoc Tukey's tests reveal that only pasture types significantly differ from the others, while fallow and extensive types are similar (Table 11). This distinction is evident in the violin plot, where pastures generally have lower nectar flower values compared to the other two types (Figure 16).

	Sum of sqrs	df	Mean square	F	p (same)
Between groups:	19.1362	2	9.56812	11.09	<0.0001
Within groups:	132.863	154	0.862744	Permutation p (n=99999)	
Total:	151.999	156	0.00004		
omega2:	0.1139				
Levene´s test for homogeneity of variance, from means p (same):					0.0837

Table 10: one-way ANOVA testing Nectar flower abundance differences between land-use types.

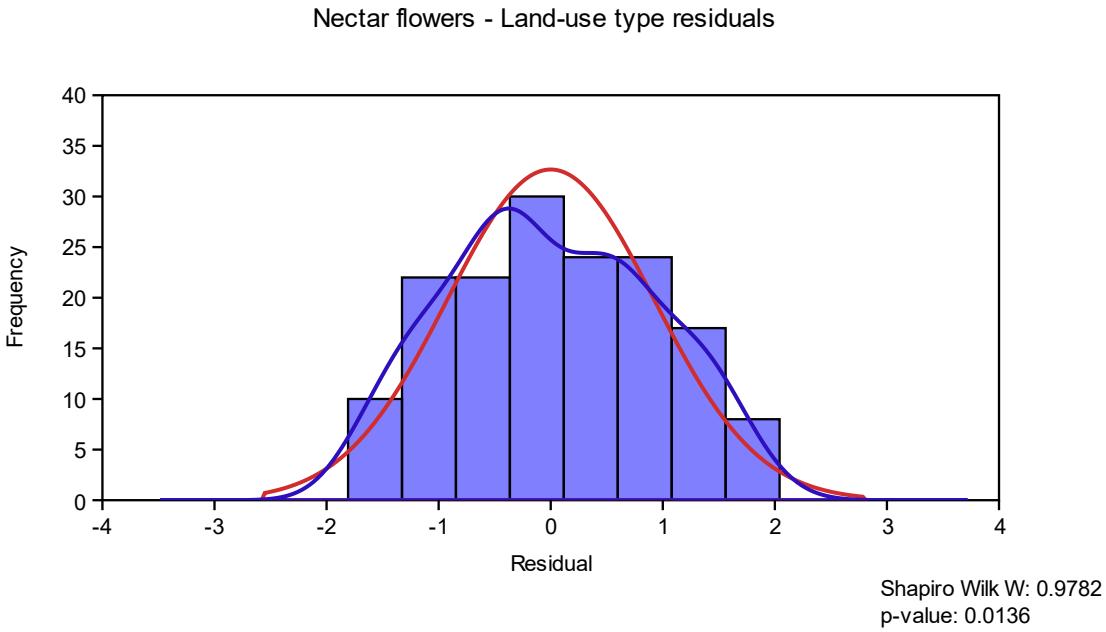


Figure 15: Residual distribution of the Nectar flower abundance – land-use type ANOVA. The red curve depicts the normal distribution, while the dark-blue curve illustrates the kernel distribution. In the bottom-right corner, you can find the results of the Shapiro-Wilk normality test.

	pasture	extensive	fallow
Pasture		<0.0001	0.0056
Extensive	6.542		0.283
fallow	4.45	2.154	

Table 11: Tukey's pairwise post-hoc analysis on Nectar flowers between land-use types. Below the diagonal, you can see the Tukey's Q value, and above the diagonal, the corresponding p-values. The light-red cells indicate statistically significant values ($p < 0.05$).

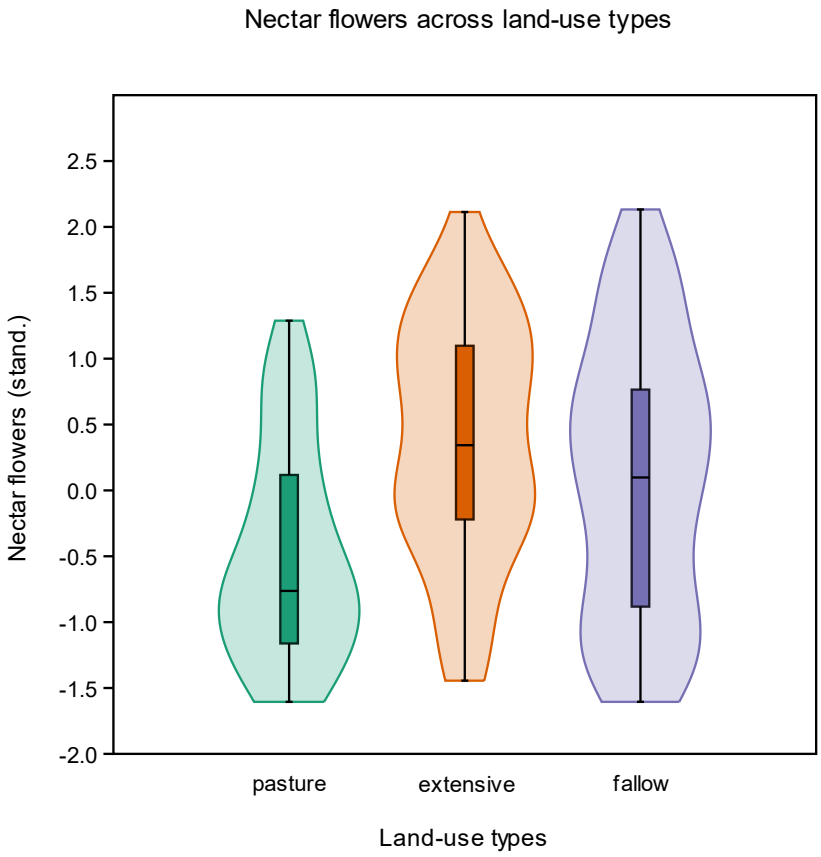


Figure 16: Nectar flower abundance across Land-use types.

Was the number of observed butterfly species (instead of individuals) related to nectar flower availability?

Incidence data underwent a Box-Cox transformation to achieve a more normal distribution (Figure 17).

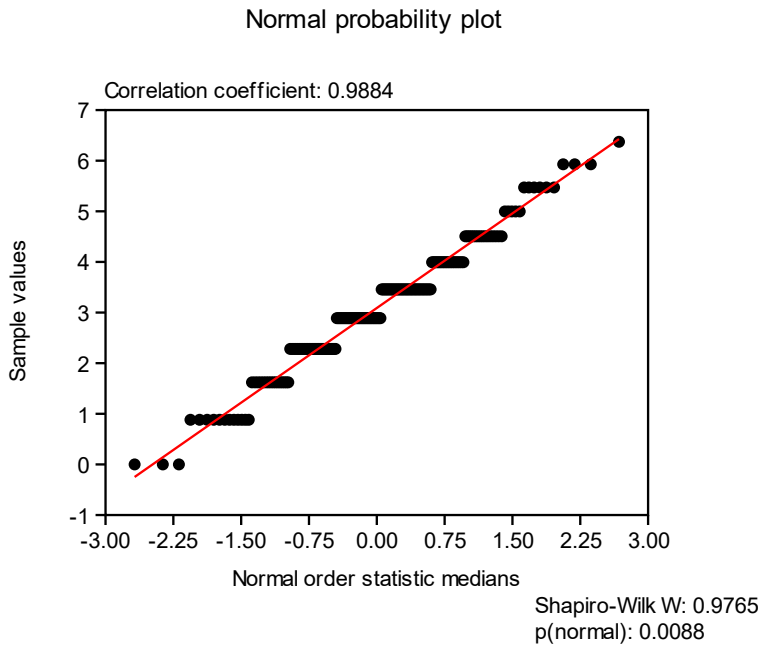


Figure 17: Correlation of transformed (Box-Cox) total incidence data with normal distribution. In the bottom-right corner the Shapiro-Wilk normality test results.

Subsequently, a linear regression was employed to compare this data with nectar flower abundance (Figure 18) and its residuals were checked to ensure their normality (Figure 19).

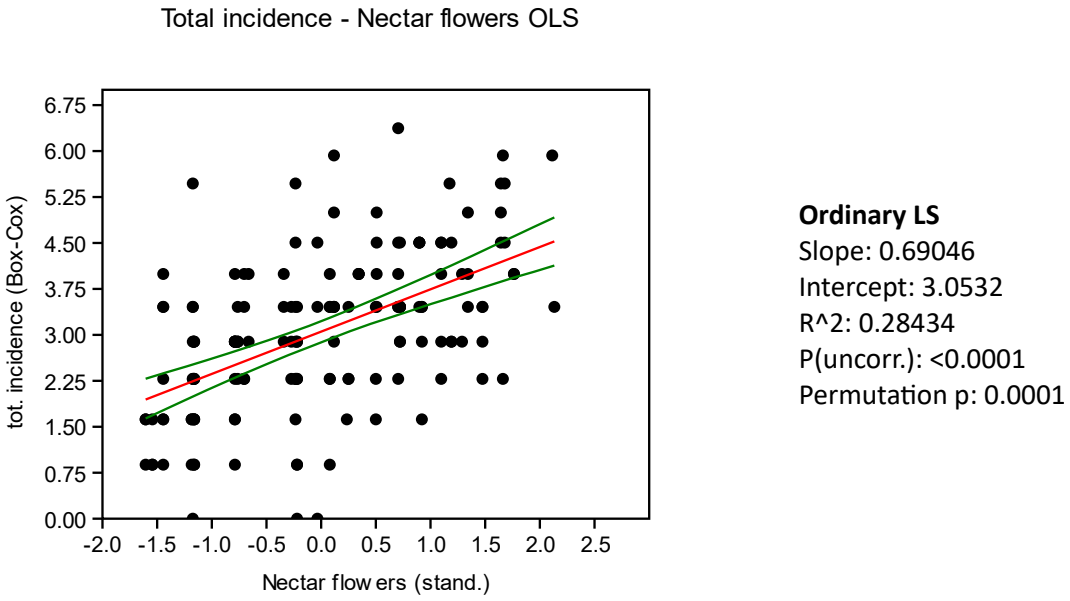


Figure 18: Linear regression (OLS) between the Box-Cox transformed total incidence and the standardized data of Nectar flowers. The red line represents the regression line, while the green lines represent the 95% confidence interval obtained from bootstrapping.

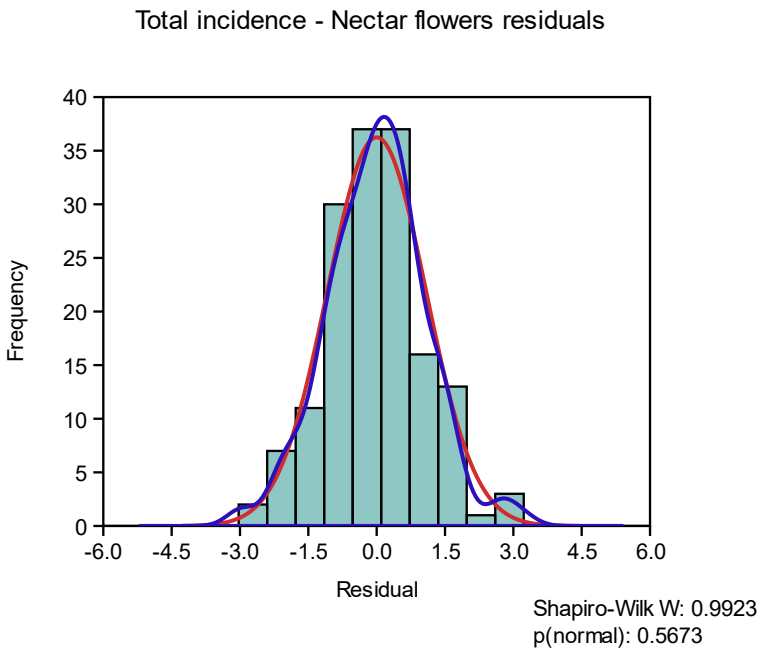


Figure 19: Residual distribution of the total incidence - nectar flowers OLS. The red curve represents the normal distribution, while the dark-blue curve represents the kernel distribution. In the bottom-right corner, you can find the Shapiro-Wilk normality test results.

The relationship between nectar flower abundance and total incidence is comparable to the one between nectar flower abundance and total abundances. However, the latter exhibits a slightly stronger correlation, with a slope very close to 1 and a slightly higher R-squared value.

Exercise 4

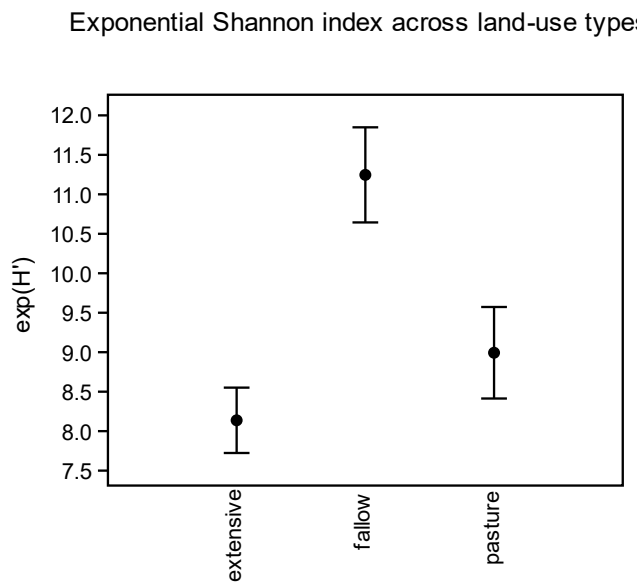
Aggregate replicate butterfly samples for the three land-use types (i.e. ignoring possible differences between years) and then use these data to calculate the following (bias-corrected) statistical attributes of butterfly assemblages in each grassland type: exponential Shannon diversity; Berger-Parker dominance; Fishers's alpha; and the Chao1 estimator of total species richness per land-use type. Which was the dominant species in either grassland type?

When the abundances of each species were aggregated, the results indicated that Maniola jurtina was the most dominant species in the extensive land-use samples, with a total of 552 observations. In contrast, for the fallow and pasture land-use types, Coenonympha pamphilus emerged as the most dominant species, with 368 and 411 total observations, respectively.

Produce appropriate graphs of these 'diversity indexes', including their 95% confidence intervals. What can you infer from these analyses?

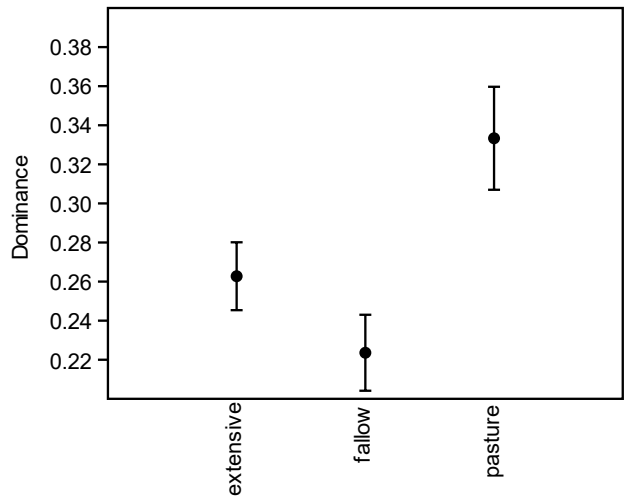
In figure 20 all the indices have been calculated with CI type: "boot, centered":

(A)



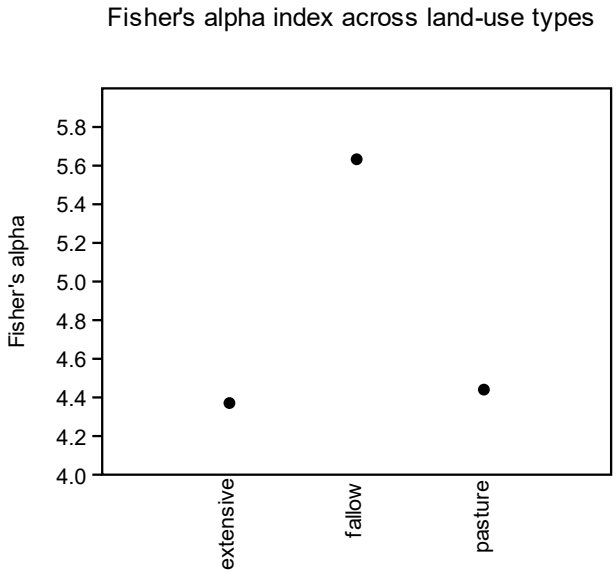
Extensive: 2.107 (2.056 2.157)
Fallow: 2.436 (2.383 2.489)
Pasture: 2.212 (2.148 2.277)

(B)



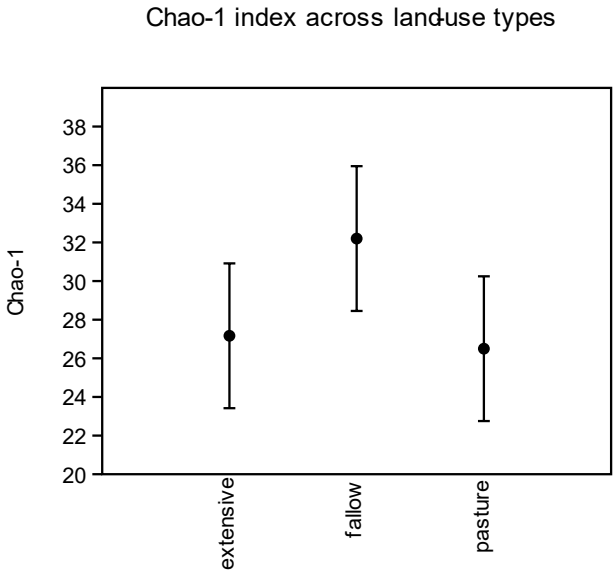
Extensive: 0.2627 (0.2449 0.2806)
Fallow: 0.2236 (0.2047 0.2424)
Pasture: 0.3333 (0.3066 0.3601)

(C)



Extensive: 4.371 (4.371 4.371)
Fallow: 5.633 (5.633 5.633)
Pasture: 4.44 (4.44 4.44)

(D)



Extensive: 27.17 (22.17 32.16)
Fallow: 32.2 (28.45 35.95)
Pasture: 26.5 (22.75 30.25)

Figure 20: Exponential Shannon diversity (A), Berger-Parker dominance (B), Fisher's alpha (C) and Chao-1 (D) species richness indexes per Land-use types.

Shannon diversity, Fisher's alpha, and Chao1 indices all yield consistent results, indicating that the fallow land-use type tends to exhibit higher species richness, while the extensive and pasture land-use types demonstrate relatively similar values. Chao1, which provides an estimate of species diversity, offers a more straightforward interpretation, while the other indices provide values with a more abstract meaning. Conversely, the Berger-Parker dominance index presents an opposing trend, with the fallow land-use type showing the lowest dominance. This suggests that the higher richness in the fallow type prevents any single species from dominating to the same extent as in the other land types. The pasture land type, despite having a similar species richness to the extensive type according to the other indices, displays higher dominance, potentially indicating the presence of a single species that dominates the community to a greater degree.

Finally, produce one graph comparing the individual-based species accumulation curves between the three land-use types (with confidence limits). What can you infer about possible differences in Lepidopteran gamma diversity between the 3 land-use types from this approach?

In the case of the species accumulation plot, I employed the iNEXT online tool (Chao et al., 2016) to generate the species abundance plot. As a preliminary step, I verified that the sample coverage was approximately 1, which it was indeed for extensive, fallow, and pasture land-use types (0.9991, 0.9988, and 0.9976, respectively). This suggests that the abundance plot can be considered comprehensive (Figure 21).

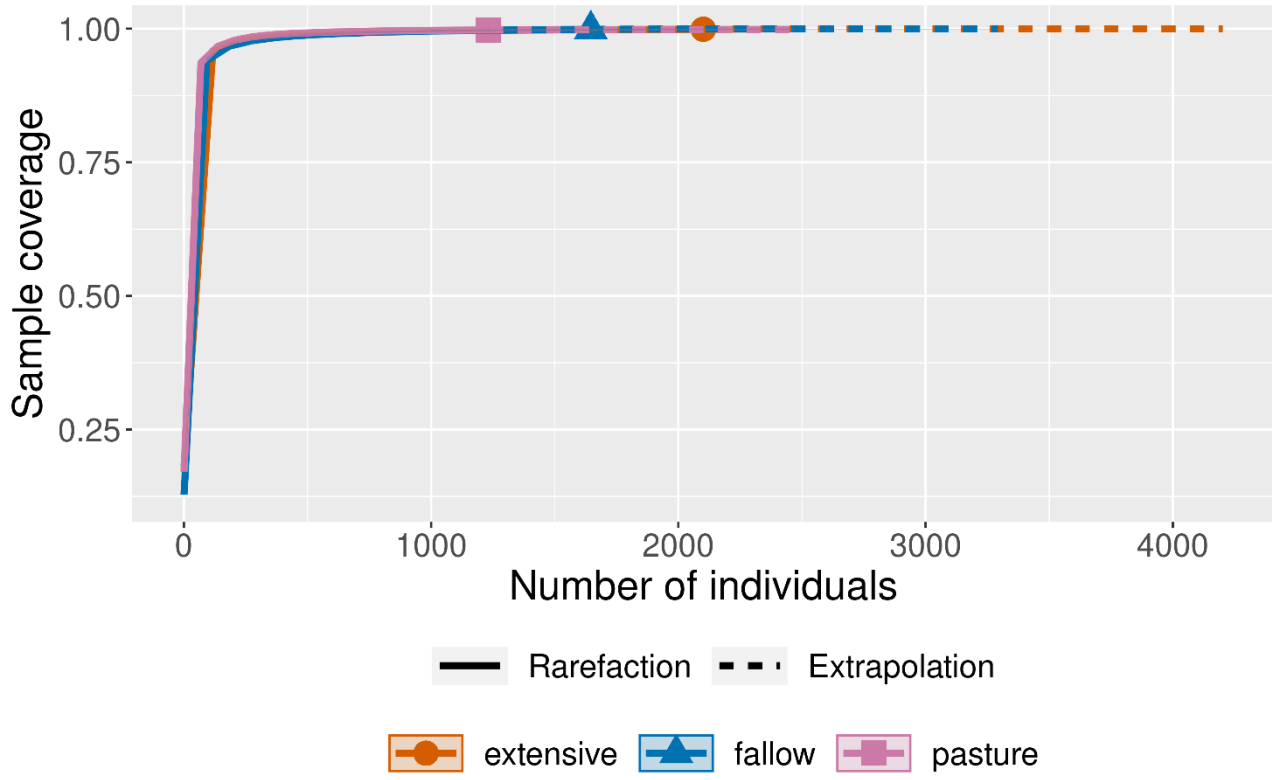


Figure 21: Sample coverage plot generated from the species abundance data across the various land-use types.

Next, I assessed species diversity using the Chao-1 estimator, revealing that the fallow land-use type generally demonstrates higher species diversity compared to the other two types (Figure 22). However, this difference reaches statistical significance only when the number of sampled individuals is rarified to match the number of individuals sampled in the pasture habitat. Conversely, when the samples are extrapolated to include approximately 2000 individuals, the confidence limits no longer support the assertion that gamma diversity differs significantly between fallow and pasture land types. Moreover, the differences between extensive and pasture types become significant by only a small margin.

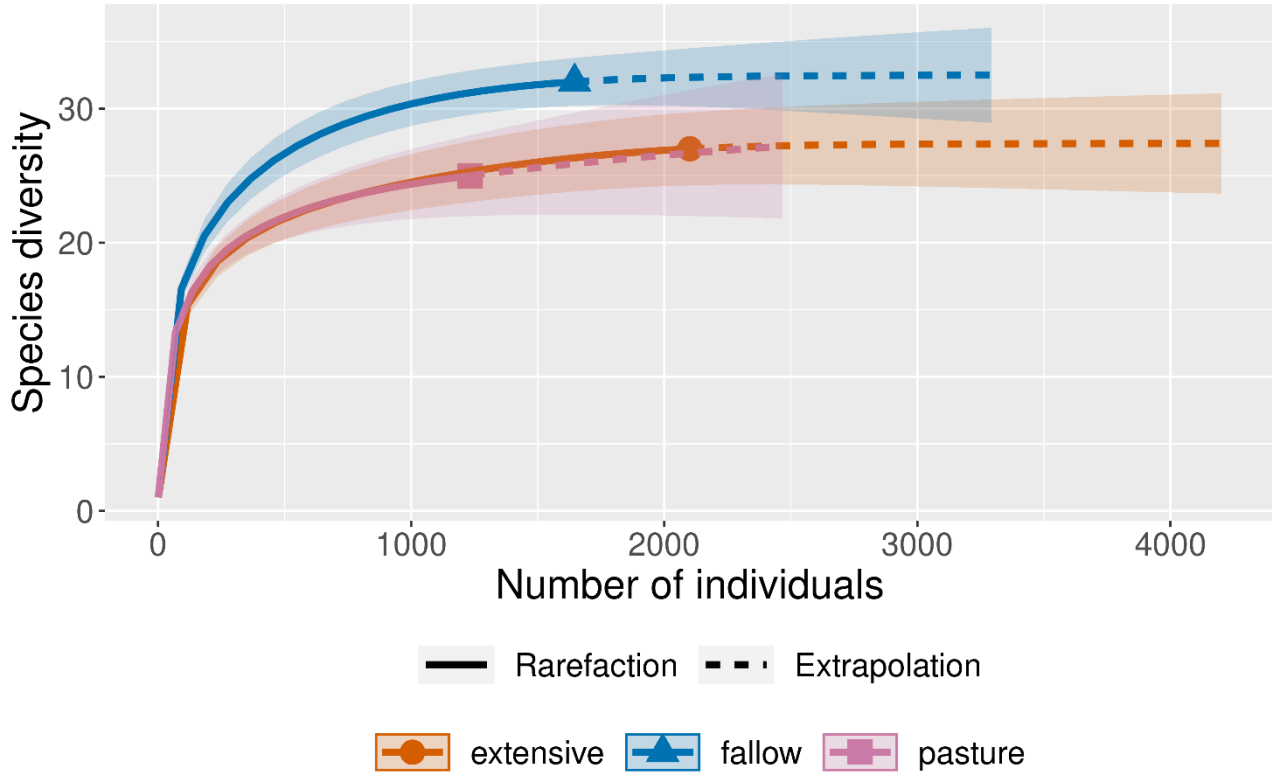


Figure 22: Species accumulation curves across the various land-use types.

Exercise 5

Explore the quantitative relationships among the site descriptors. Take care of appropriate transformations (if applicable). What can you infer from these outcomes on the environmental gradients under study?

I began by examining the linearity and normality of the data. For the variables with skewed distributions, I applied different transformation methods, such as square root or Box-Cox, selecting the most effective transformation. Following this, I conducted a z-transformation on the data and proceeded with a Principal Component Analysis (PCA) to visualize patterns and relationships within the data groups. Upon evaluating the eigenvalues of the components, the percentage of explained variance (Table 12), and reviewing the scree plot (Figure 23), I decided to retain the first four principal components.

PC	Eigenvalue	% variance
1	5.34569	38.183
2	3.30393	23.6
3	1.42961	10.211
4	1.08415	7.7439
5	0.761008	5.4358
6	0.705795	5.0414
7	0.474257	3.3876
8	0.384469	2.7462
9	0.265072	1.8934
10	0.0946601	0.67614
11	0.0889232	0.63517
12	0.0347129	0.24795
13	0.0187777	0.13413
14	0.00894628	0.063902

Table 12: Eigenvalues and % of variance explained by each principal component.

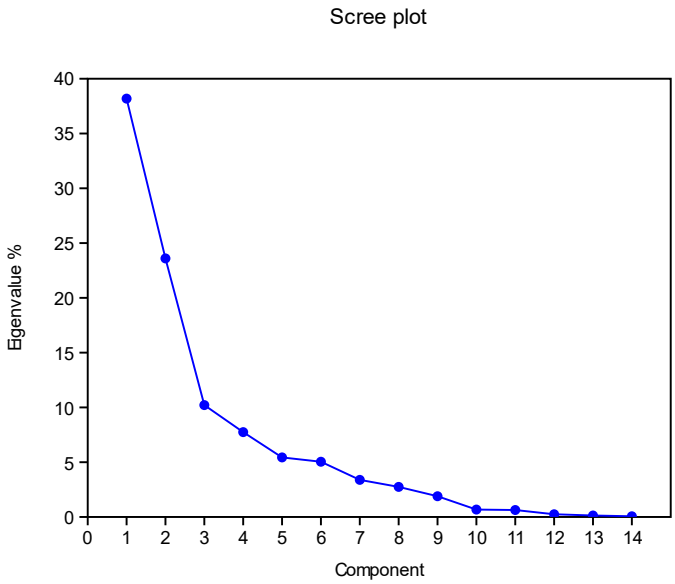


Figure 23: Scree plot obtained by the PCA made on the site descriptors.

Then it's possible to visualize the newly created principal components (Figure 24).

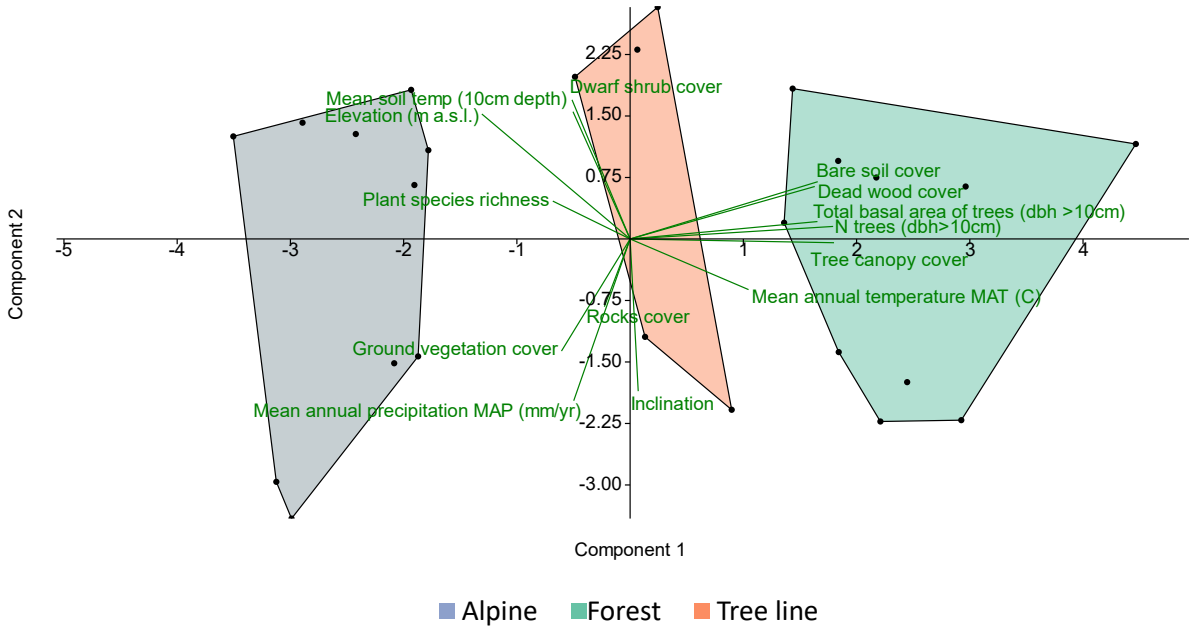


Figure 24: Plot of PC1 and PC2, displaying vectors for each site descriptor and dots representing different sites grouped by habitat type.

Next, I examined the PCA loadings and loadings plots to gain insights into what each of the different components was summarizing about the sites:

PC1 appears to represent a gradient from alpine to forest environments, as evidenced by its strong positive correlations with total basal area (0.93), number of trees (0.86), tree canopy cover (0.95), bare soil cover

(0.83), and dead wood cover (0.89). It also moderately correlates with mean annual temperature (0.5) but negatively correlates with elevation (-0.73), dwarf shrub cover (-0.38), and plant species richness (-0.375).

On the other hand, PC2 is more complex. It exhibits positive correlations with elevation (0.60), mean soil temperature (0.67), dwarf shrub cover (0.62), and dead wood cover (0.89), while negatively correlating with inclination (-0.76), mean annual precipitation (-0.72), ground vegetation cover (-0.61), mean annual temperature (-0.34), and rocks cover (-0.31). This suggests that PC2 likely represents a combination of factors related to tree-line habitats, including higher elevations and substantial dead wood cover.

PC3 is primarily correlated with plant species richness (0.697), ground vegetation cover (0.51), rocks cover (0.369), and mean soil temperature (0.357), with a negative correlation with dwarf shrub cover (-0.539).

Lastly, PC4 shows positive correlations with rocks cover (0.658) and mean annual precipitation (0.427), while negatively correlating with mean annual temperature (-0.48) and ground vegetation cover (-0.34).

After considering the correlations between PCs and variables, as well as the angles and lengths between them, I selected the following variables for further analysis: elevation, mean annual temperature, mean soil temperature, mean annual precipitation, dead wood cover, and plant species richness. These variables exhibit a reasonable degree of independence from each other, collectively capture diverse facets of climate and topographic influences on vegetation, and adequately span the different directions represented by the first four principal components (Figure 25).

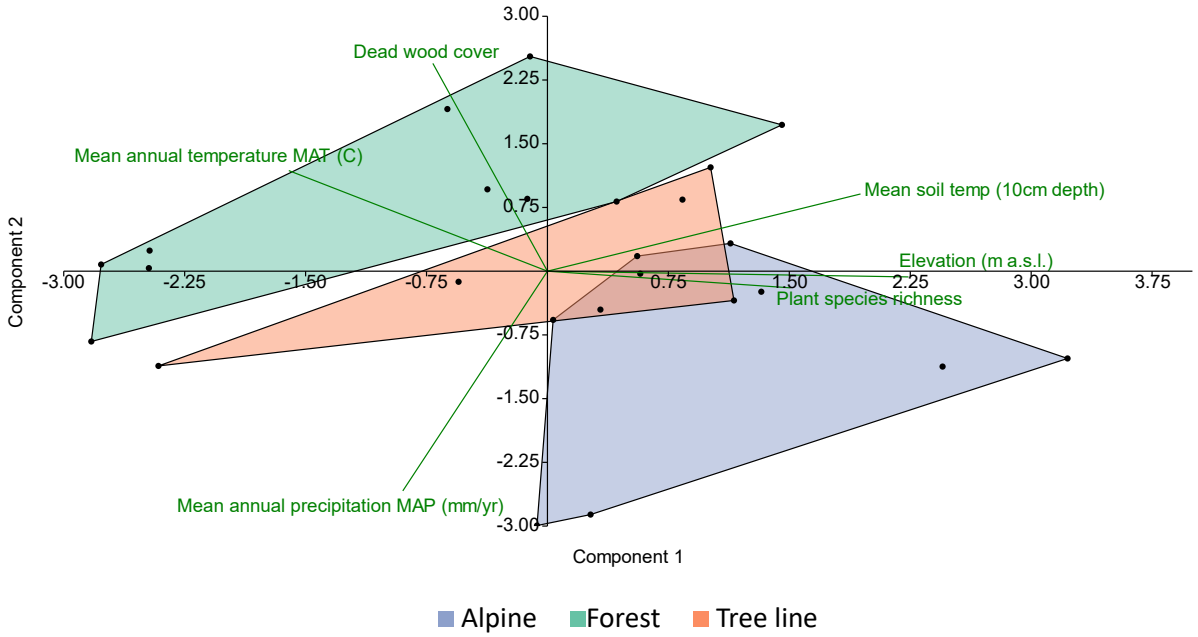


Figure 25: PC1 and PC2 plot, with the most descriptive vectors, each representing a site descriptor and the various sites as dots grouped by habitat type.

Did ant species richness per site differ between the 3 habitat types?

As there were no data of the observed abundances for each species in the different sites, I used the number of sampling methods that produced records for the respective ant species as incidence data, so that it could be used to estimate the species richness in each habitat with the Chao2 estimator (Table 13) and then visually represented by using the iNext online tool (Figure 26).

	Alpine	Tree-line	Forest
Observed	9	9	12
Chao2	9	9	12.45
SD	0.1335	0.3414	1.1882
Sample coverage	1.0000	1.0000	0.9695

Table 13: Table showing the observed, Chao-2 estimated, Standard deviation and sample coverage species richness in each habitat.

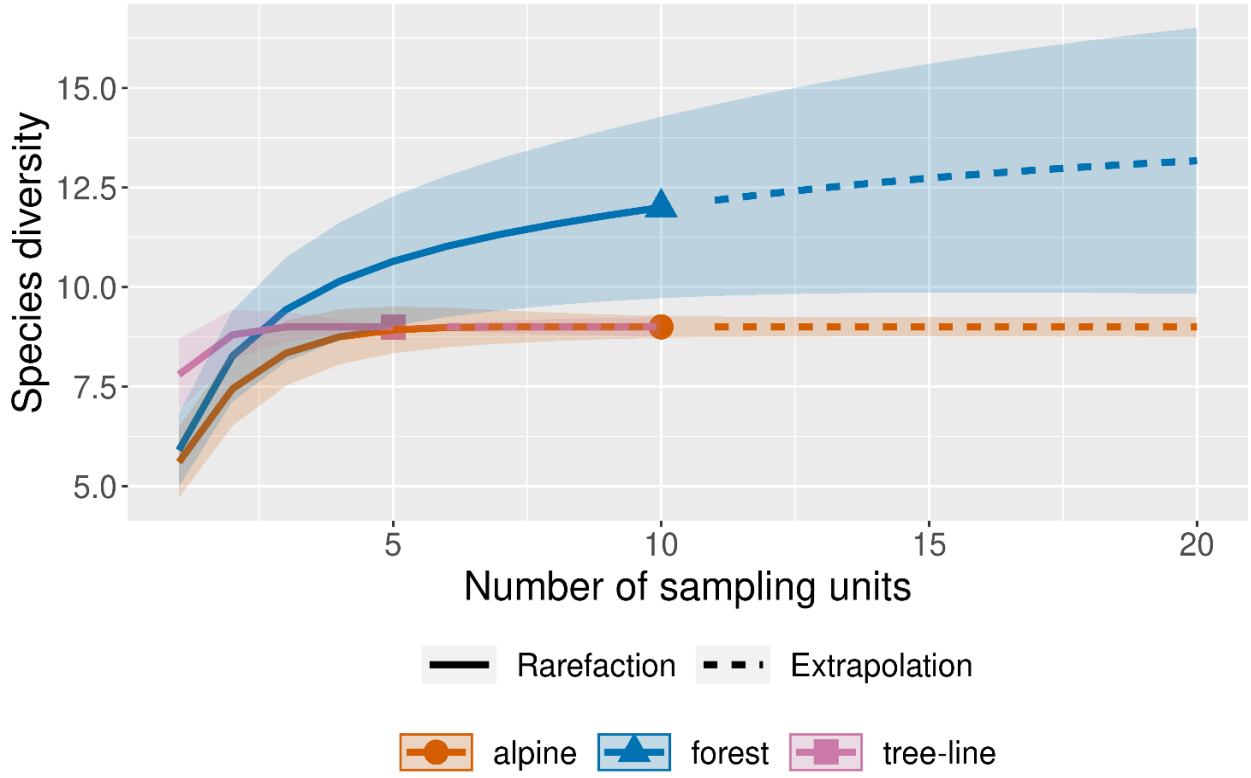


Figure 26: Rarefaction and extrapolation of species diversity across different habitats.

The analysis revealed that all the samples exhibit a sample coverage close to 1. Interestingly, tree-line and alpine environments demonstrated approximately equal species richness, while the forest exhibited significantly higher species richness to around 12 for the sampling units examined, but possibly as high as 13 in the case of 20 sampling units.

Was local ant species richness related to any of the site characteristics?

Upon visualizing the correlation between incidence data of each sampled plot alongside the six previously selected site descriptors, it became evident that all relationships were weak and lacked statistical significance (Bonferroni corrected).

Can you model the species composition of local ant assemblages using a reasonable selection of habitat descriptors, or alternatively by using composite variables aggregated through an appropriate multivariate exploration?

To build a model relating species composition to habitat descriptors, I employed NMDS (Non-Metric Multidimensional Scaling) and utilized the Bray-Curtis similarity index for individual species incidence data.

The site descriptors selected for the previous PCA analysis were incorporated as environmental variables in this analysis (Figure 27).

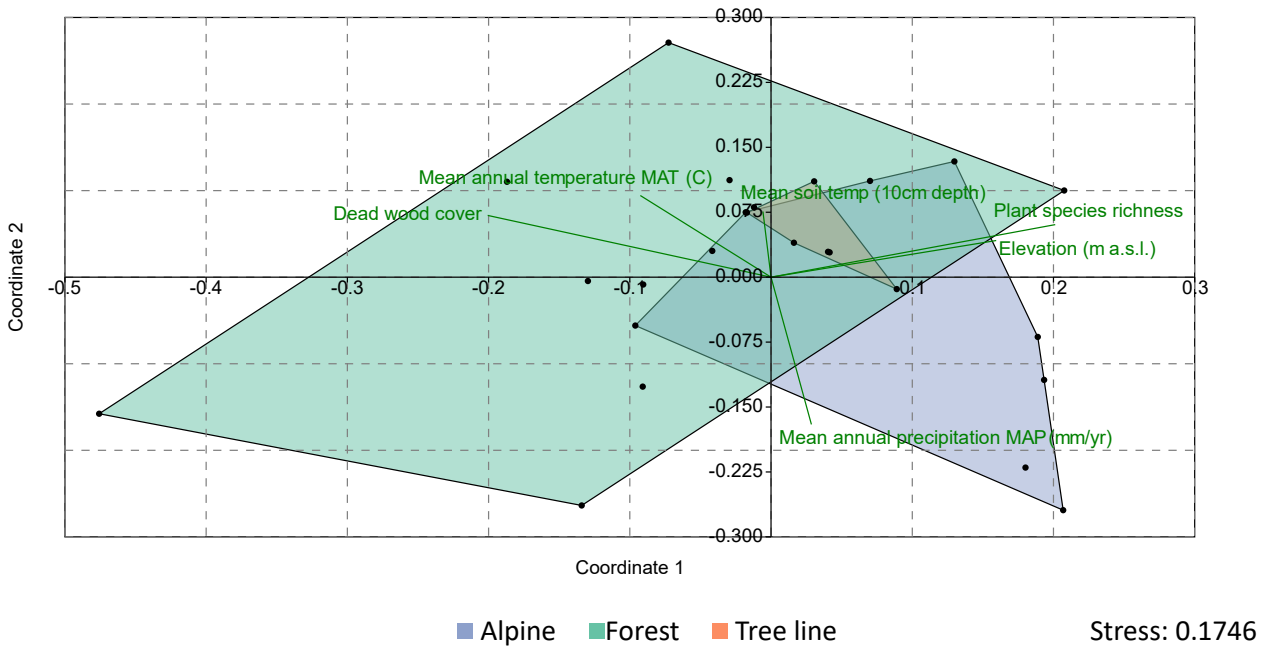


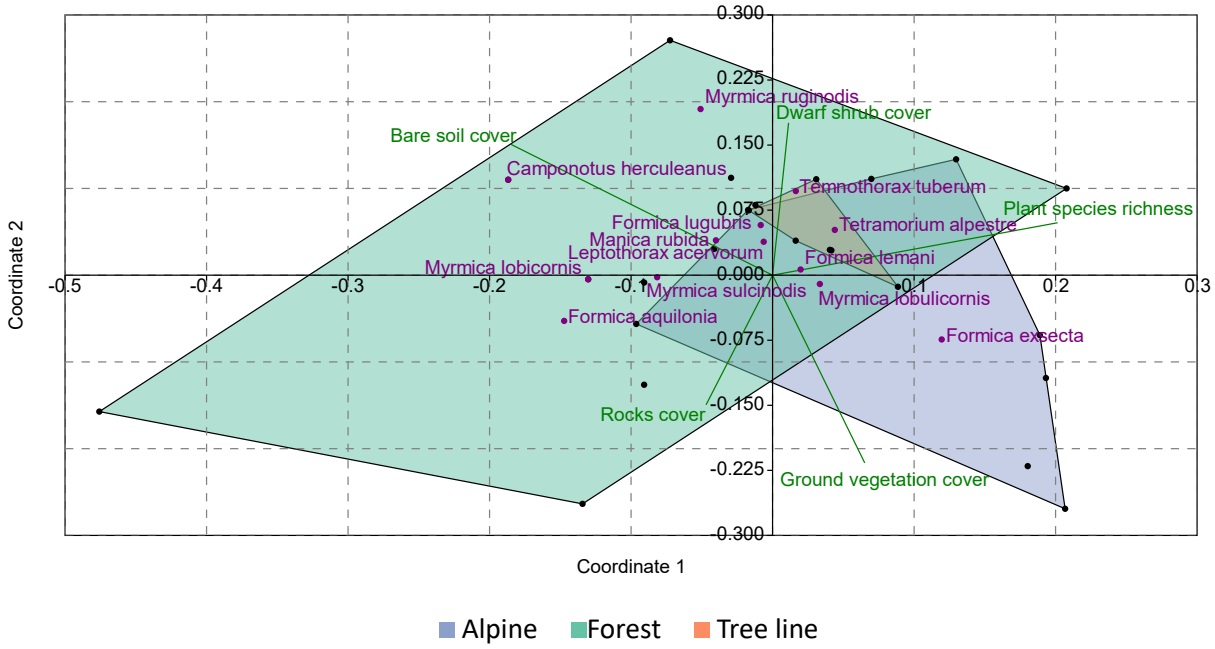
Figure 27: NMDS plot generated using species incidence data and selected site descriptors as environmental variables, with habitat types grouped and differentiated by colors.

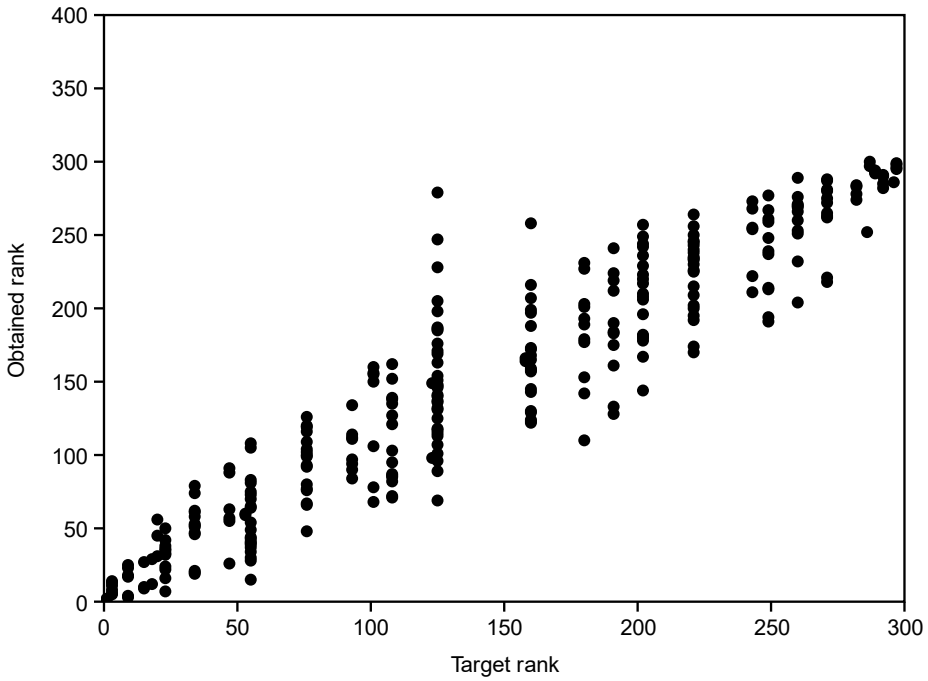
In this instance, the stress value is below 0.2, and the Shepard plot exhibited relatively low scattering, indicating the NMDS plot's high accuracy. Regarding the selected site descriptors, some of them display complete overlap (e.g., elevation and plant species). Conversely, it demonstrates how forested and alpine sites partially overlap, with tree-line samples positioned within this overlapping space.

Alpine assemblages are positively influenced by factors such as elevation, plant species richness and mean annual precipitation. Conversely, forests are positively associated with variables like dead wood cover, mean annual temperature, and mean soil temperature. Tree-line ecosystems represent a mixture of the two preceding habitats, with positive associations with mean soil temperature, elevation, and species richness, albeit to a lesser extent than alpine habitats.

To explore if there are superior site descriptors for modeling site assemblages, I conducted another NMDS analysis using all descriptors and ultimately selected the following factors: plant species richness, ground vegetation cover, dwarf shrub cover, rocks cover, and bare soil cover (Figure 28). In this case I also added each species name with its coordinate, so that it could be visible how the different site descriptors could influence species presence or absence.

The selection process involved visually identifying the longest vectors with the most wide angles between them. Additionally, I considered factors that could comprehensively represent and have connections to various other factors. This approach allows for the combination of two or more site descriptor vectors through vector summation, enabling the estimation of species composition at any site based on its descriptive values.





Stress: 0.1748

Figure 28: NMDS obtained by species incidence data and selected site descriptors as environmental variables; habitat types are grouped and showed with different colors (A). Shepard plot of the NMDS (B).

The Chao2 estimator showed that there are significantly different species richness across land-use types and the previous NMDS analysis showed that species had different coordinates across the land-uses, but to

determine if the assemblages of those land-use types are significantly different it's necessary to perform a PERMANOVA (Table 14):

(A) PERMANOVA	
Permutation N:	9999
Total sum of squares:	2.043
Within-group sum of squares:	1.615
F:	2.914
p (same):	0.0157

(B) Pairwise	alpine	tree line	forest
alpine		1	0.0234
tree line	1		0.2838
forest	0.0234	0.2838	

Table 14: PERMANOVA results demonstrating statistically significant differences in species assemblages among various habitats (A). Pairwise tests with Bonferroni correction highlighting the habitats with significantly distinct assemblages; light-red cells indicate statistically significant values ($p < 0.05$) (B).

Furthermore, I attempted to plot the incidence data using the PCA axes derived from the consideration of all site descriptors. This visualization aimed to demonstrate how principal components can predict different habitat types (Figure 29). PC1 and PC4 exhibit positive correlations with forested sites and negative correlations with alpine and tree-line sites, respectively. PC2 traverses all three habitat types, characterizing alpine sites with lower values, tree-line sites with intermediate values, and forests with high values. Lastly, PC3 shows a positive correlation with alpine habitats and a negative correlation with forested ones.

Given that these components are made by composite variables, predicting local assemblages becomes more challenging, making it essential to sample real data to understand and anticipate local patterns accurately.

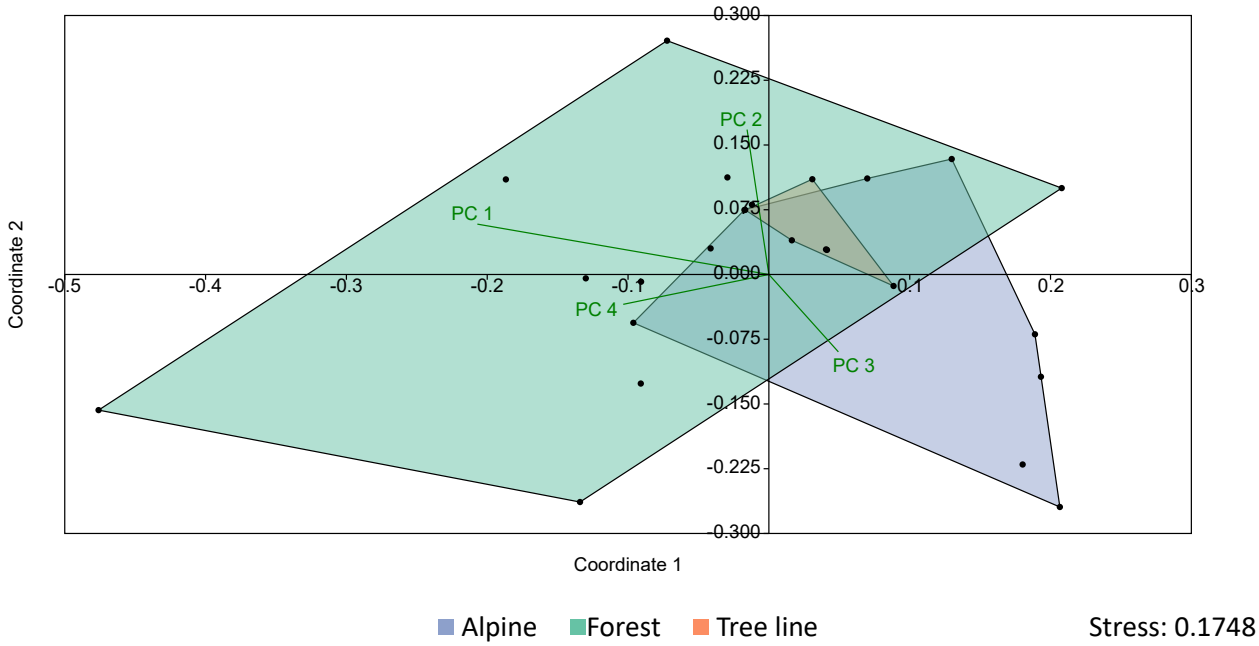


Figure 29: NMDS obtained by species incidence data and with four principal components, obtained by all the site descriptors as environmental variables; habitat types are grouped and showed with different colors.

Acknowledgments

This document underwent text refinement and readability enhancements using ChatGPT (OpenAI, 2023) an AI-based tool, which contributed to improve clarity and coherence.

References

Chao, A., Ma, K. H., and Hsieh, T.C. (2016). iNEXT (iNterpolation and EXTrapolation) Online: Software for Interpolation and Extrapolation of Species Diversity. Program and User's Guide published at http://chao.stat.nthu.edu.tw/wordpress/software_download/inext-online/

Hammer, Ø., Harper, D.A.T., Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4(1): 9pp. http://palaeoelectronica.org/2001_1/past/issue1_01.htm

OpenAI. (2023). GPT-3.5 (Version). [Computer software]. Retrieved from https://www.openai.com