Analysis and Visualisation of Complex Agro-Environmental Data

Lesson 06

- Non-parametric hypothesis testing
- Outlier analysis
- Data transformation
- Topics for the final project

U LISBOA | UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR D AGRONOMIA
ISA

Lesson #6

1. Non-parametric hypothesis testing
2. Outlier analysis
3. Data transformation
4. Topics for the final project

Lesson #6

1. Non-parametric hypothesis testing
2. Outlier analysis
3. Data transformation
4. Topics for the final project

Non-parametric hypothesis testing

Randomization or permutation tests

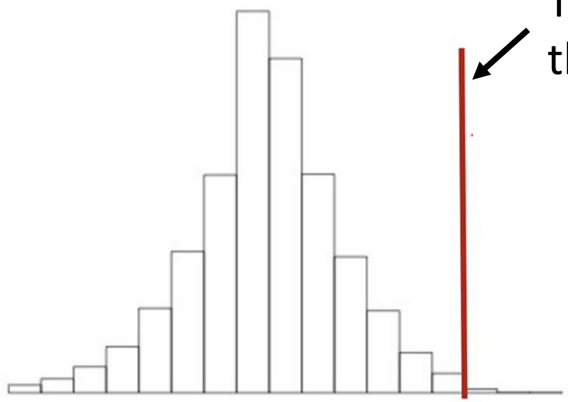
- Based on resampling or reshuffling the original data many times to **generate sample distributions**.

Original data

Create n random (or null) patterns

Compute n test statistic values

Distribution



Test statitistic of the original data

p-value = proportion of test statistics > original value

Non-parametric hypothesis testing

Randomization or permutation tests

Particularly useful when:

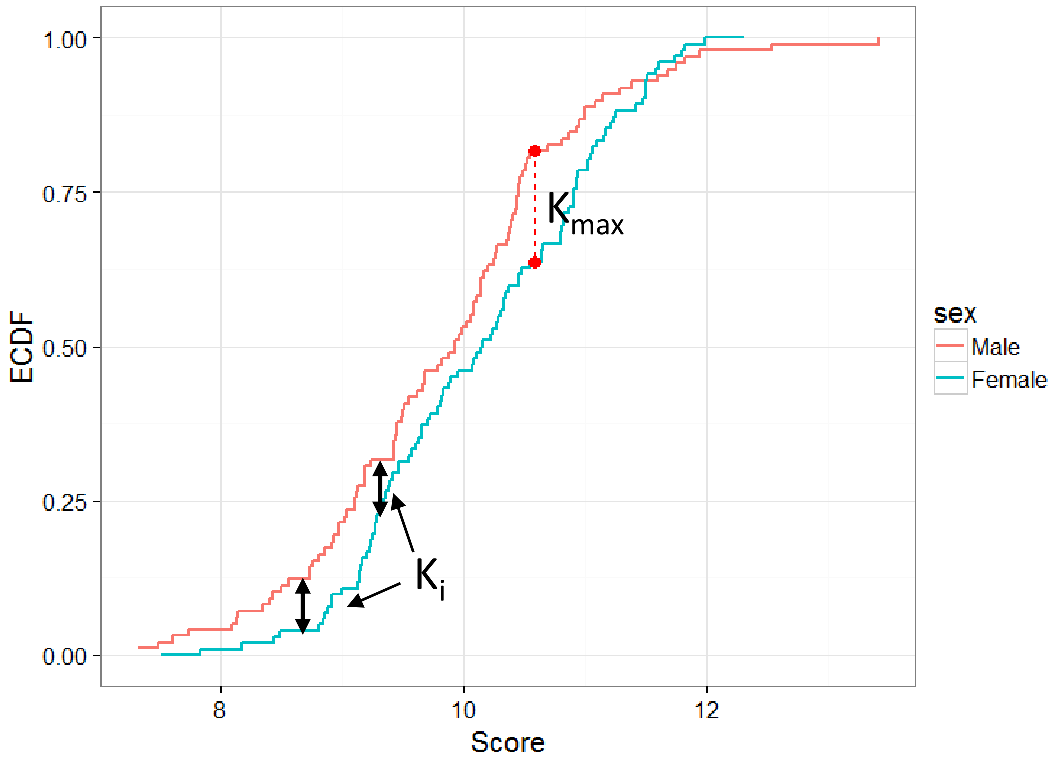
- The **distribution is unknown**
- A **random sampling is not possible** (e.g. data opportunistically collected)
- Other assumptions such as ***iid*** **observations are questionable** (e.g. temporal trends and spatial patterns).

Non-parametric hypothesis testing

Tests based on differences between sample distributions

Kolmogorov-Smirnov test

Based on diferences in the **Empirical Cumulative Probability Function** (ECDF) of two samples (one can be a reference distribution such as the normal distribution – one-sample normality test)



1. Compute K_i (absolute difference)
2. K_{max} will be the test statistic (follows a Kolmogorov distribution)
3. K_{max} > critical K_{df} (table) => H_0 (no diferences) is rejected

Non-parametric hypothesis testing

Rank-based tests

Man-whitney U test (simplest case with no ties)

H_0 – There is no difference in the central tendency between groups

	Gender	Vaue
n_1 = 8	Female	157.522
	Female	201.909
	Female	123.791
	Female	101.078
	Female	106.1
	Female	271.097
	Female	211.835
	Female	186.874
n_2 = 7	Male	106.1
	Male	90.502
	Male	193.807
	Male	95.289
	Male	101.951
	Male	191.076
	Male	99.811

order

Gender	Value
Male	90.502
Male	95.289
Male	99.811
Female	101.078
Male	101.951
Female	106.1
Male	106.1
Female	123.791
Female	157.522
Female	186.874
Male	191.076
Male	193.807
Female	201.909
Female	211.835
Female	271.097

rank

Gender	Rank
Male	1
Male	2
Male	3
Female	4
Male	5
Female	6
Male	7
Female	8
Female	9
Female	10
Male	11
Male	12
Female	13
Female	14
Female	15

Rank sum:

T_1 = 4+6+8+9+10+13+14+15 = 79

T_2 = 1+2+3+5+7+11+12 = 41

Test statistics U:

$$U_1 = n_1 \cdot n_2 + \frac{n_1 \cdot (n_1 + 1)}{2} - T_1 = 13$$

$$U_2 = n_1 \cdot n_2 + \frac{n_2 \cdot (n_2 + 1)}{2} - T_2 = 43$$

Mann-Whitney U = min (U1, U2) = 13

Expected value: $\mu_U = \frac{n_1 \cdot n_2}{4} = 10.5$

- A *p-value* is then obtained for each U and degrees-of-freedom.
- Large sample sizes => z-value can be used

Non-parametric hypothesis testing

Rank-based tests

Wilcoxon signed-rank test (simplest case with no tied ranks)

H_0 – There is no difference in the central tendency between groups

	Time 1	Time 2
Ind 1	157.522	106.1
Ind 2	201.909	90.502
Ind 3	123.791	193.807
Ind 4	101.078	95.289
Ind 5	101.951	106.1
Ind 6	191.076	271.097

Difference and rank

Diff.	Rank	Sign
51.422	3	+
111.407	6	+
-70.016	4	-
5.789	2	+
-4.149	1	-
-80.021	5	-

Rank sum:

$T_{(+)}$ = 3+6+2 = 11

$T_{(-)}$ = 4+1+5 = 10

Test statistics:

$W = min\ (T_{(+)}, T_{(-)}) = 9$

Expected value of W: $\mu_W = \frac{n \cdot (n+1)}{4} = 10.5$

- A *p-value* is then obtained for each U and degrees-of-freedom.
- Large sample sizes => z-value can be used

Non-parametric hypothesis testing

Rank-based tests – multiple samples

Kruskal-Wallis test

- The non-parametric version of one-way ANOVA
- Assesses whether samples belong to the same distribution (tests diferences in the medians).

Number of cases in group i

Mean rank sum in group i

Total sample size

Expected value of the rankings

$$H = \frac{n-1}{n} \cdot \sum_{i=1}^{k} \frac{n_i \cdot (\bar{R} - E_R)^2}{\sigma^2}$$

Rank variance

Friedman test

- the non-parametric alternative to the one-way ANOVA with **repeated measures** (ex. over time)

Sum of the square sum of ranks per group

$$\chi^2 = \frac{12}{N \cdot k \cdot (k+1)} \cdot \sum R^2 - 3 \cdot N \cdot (k+1)$$

Sample size

Number of repetitions

Post-oc or multiple comparisons tests

- A common post-hoc test is the **Dunn's test**

NOTE: In any case, parametric tests should be preferably used over rank-based tests except when **distributions are weird** (and transformations do not help) or **outliers** are present.

Non-parametric hypothesis testing

Tests with categorical variables

Pearson's chi-square test – two categorical variables and large samples

- determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table:

$$\chi^2 = \sum_{i=1}^{n} \frac{(Oi - Ei)^2}{E_i}$$

O_i – Observed values
E_i – Expected values
i – position in the contingency table

- H_0: the two categorical variables are independent

Hypothesis testing in Python

Null Hypothesis	Distributions	SciPy Functions for Test
The population mean has a given value.	Normal distribution (stats.norm), or Student's t distribution (stats.t)	stats.ttest_1samp
The means of two random variables are equal (independent or paired samples).	Student's t distribution (stats.t)	stats.ttest_ind, stats.ttest_rel
The medians of two random variables are equal (independent or paired samples).	Wilcoxon distribution	stats.mannwhitneyu wilcoxon
The distribution of two random variables are equal.	Kolmogorov-Smirnov distribution	stats.kstest
Categorical data occur with given frequency (sum of squared normally distributed variables).	χ2 distribution (stats.chi2)	stats.chisquare
Two categorical variables are independent.	χ2 distribution (stats.chi2)	stats.chi2_contingency
two or more variables have equal variance in samples	F distribution (stats.f)	stats.barlett, stats.levene
Two or more groups have the same population mean (ANOVA).	F distribution	stats.f_oneway, stats.kruskal
Two variables are not correlated.	Beta distribution (stats.beta, stasts.mstats.betai)	stats.pearsonr, stats.spearmanr

Check github: https://github.com/isa-ulisboa/greends-avcad-2025/tree/main/examples

Hypothesis_testing_Non_parametric.ipynb

Lesson #6

1. Non-parametric hypothesis testing
2. Outlier analysis
3. Data transformation
4. Topics for the final project

Outlier analysis

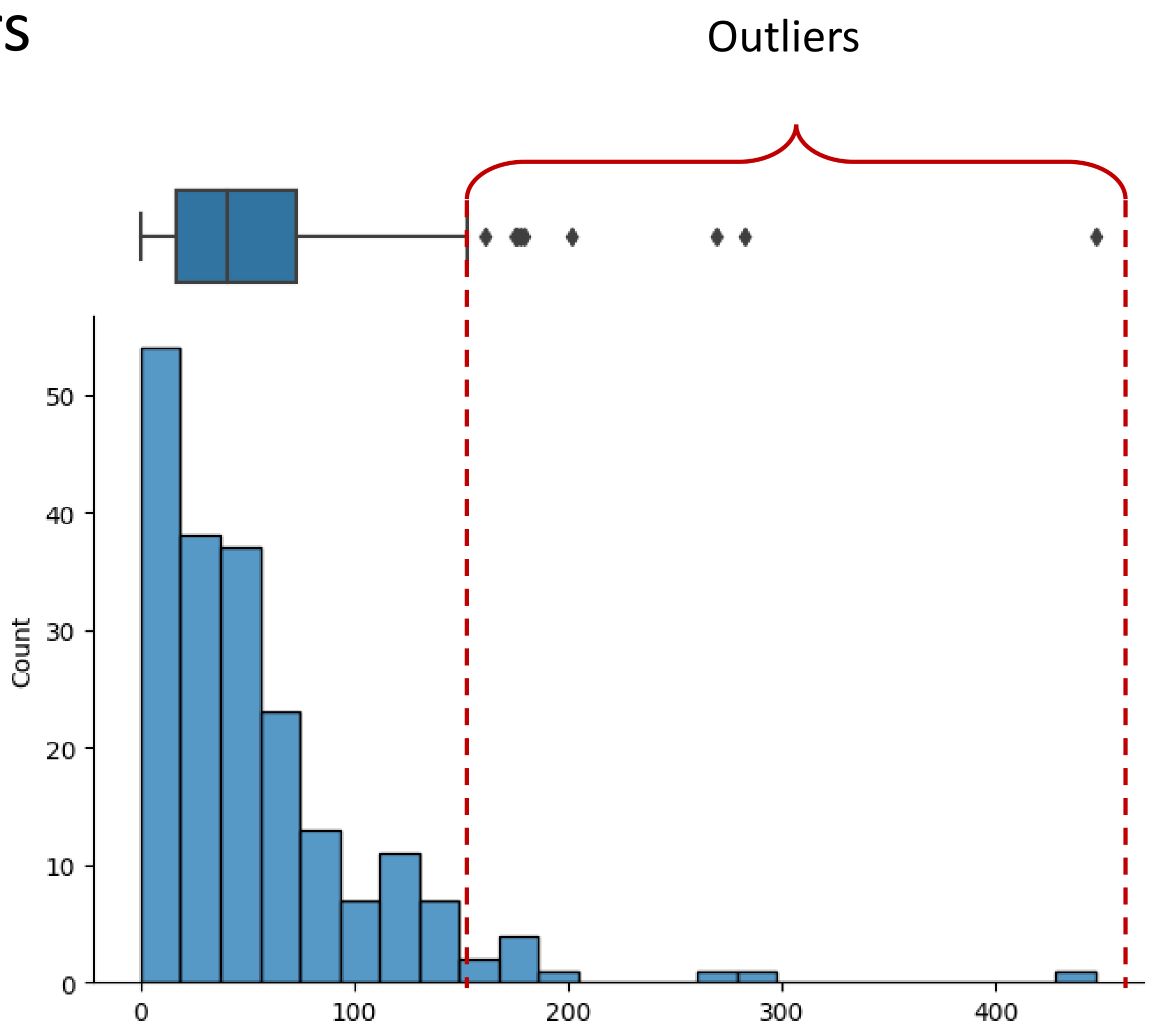
= **outlier detection** = **anomaly detection** (time series): a family of analytical and graphical tools to detect outliers - important step of data mining

Outliers

- A data value *that appears to deviate markedly from other members of the sample in which it occurs* (Grubbs, 1969).
- Impacts on summary statistics, pattern detection and confirmatory analysis
- May be caused by:
  - ✓ Natural variability (e.g. extreme events)
  - ✓ Human error
  - ✓ Faulty equipment
  - ✓ Poor sampling

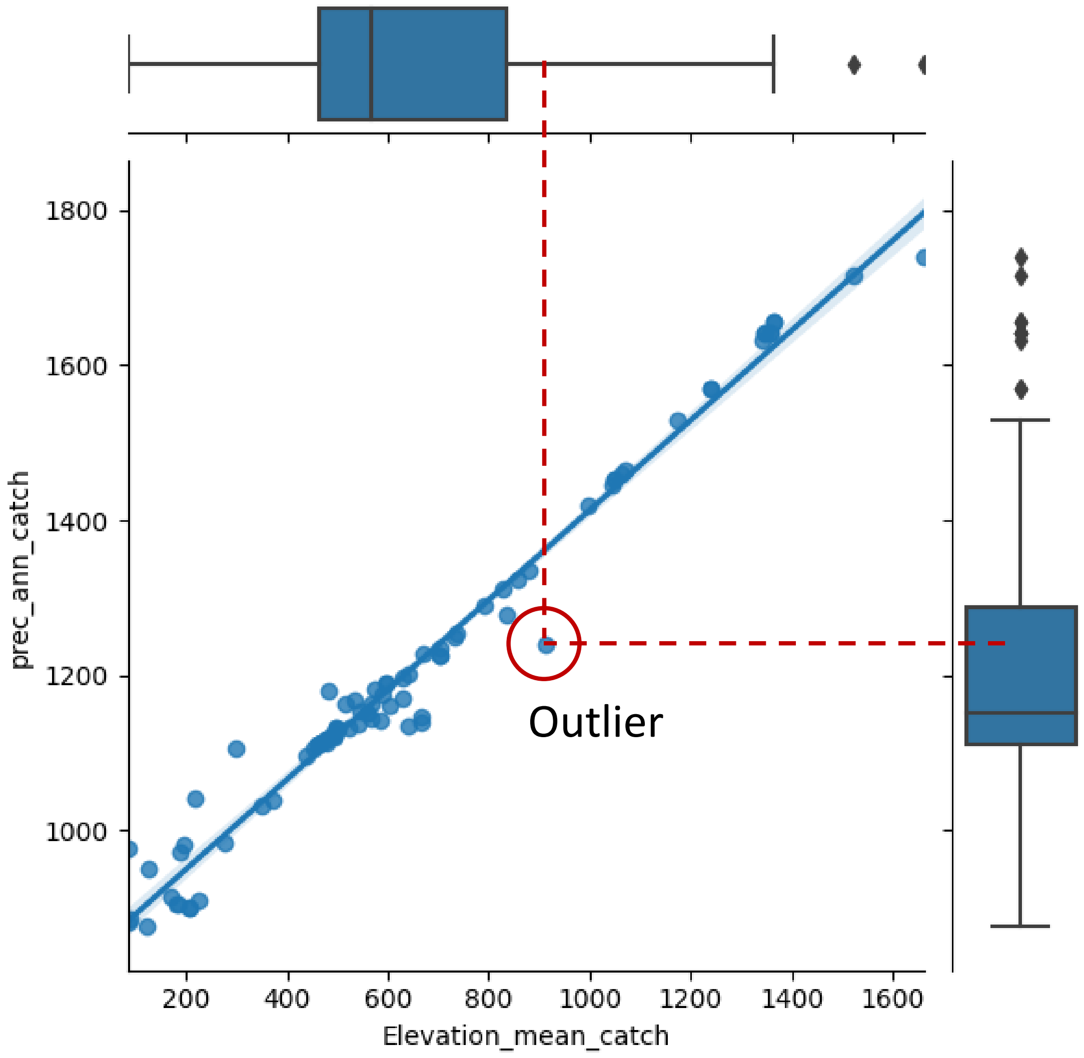
Outlier analysis

Univariate outliers



Outlier analysis

Multivariate outliers



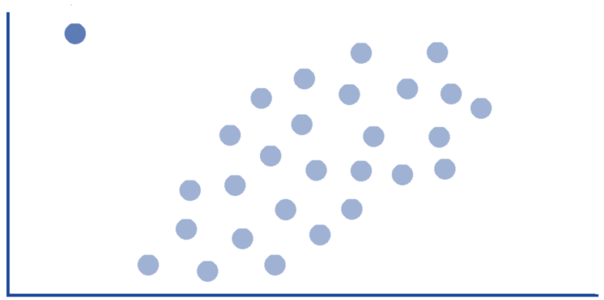
Outliers not detected by univariate plots

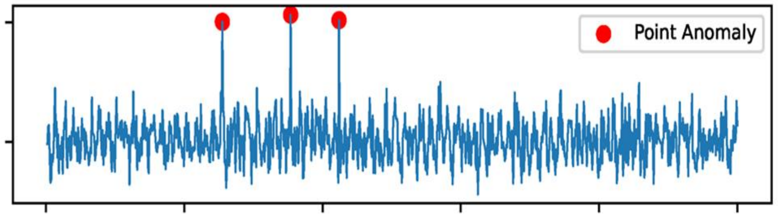
Outlier analysis

3 types of outliers

Type 1: Global outliers
(also called "point anomalies")

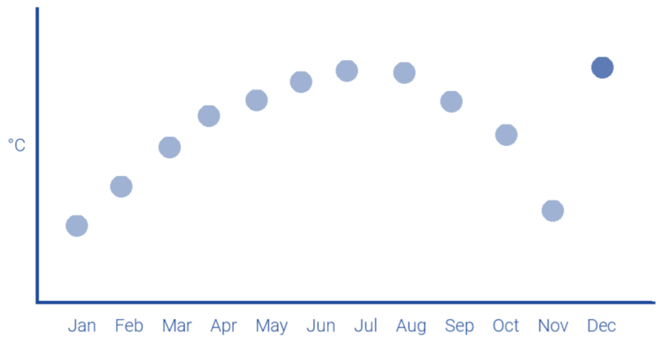
Extreme values at any context

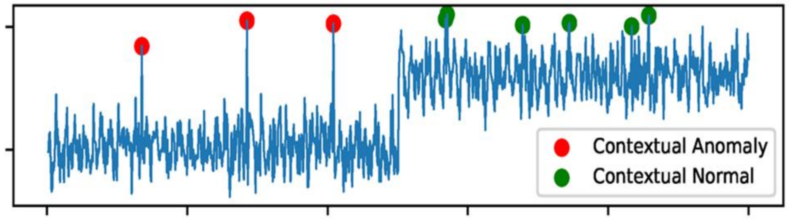




Type 2: Contextual (or conditional) **outliers**

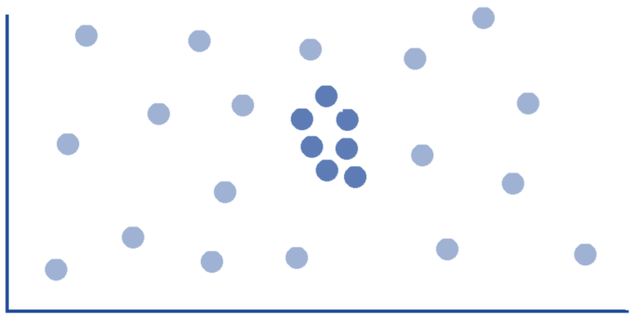
Extreme values that depend on a particular context

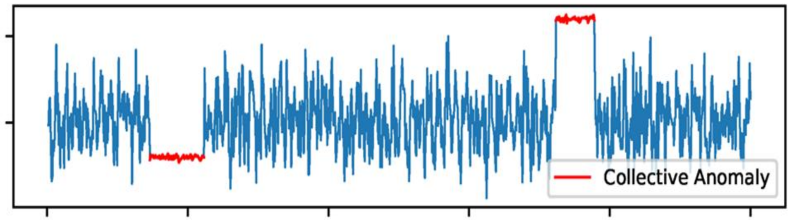




Type 3: Collective outliers

A set of values that induce an anomaly (only) in combination





Outlier analysis

Dealing with outliers

Rule: outliers must be dealt with **only if they are shown to have a significant impact** in the data analysis or model performance.

After detecting outliers, there are four main methods of dealing with them:

- Removing from the dataset
- Reducing the weights of detected outliers
- Changing the values of outliers
  - ✓ Winsorisation - replacing them with the nearest non-outlier values
  - ✓ Imputation - replacing by estimates based on the data (e.g. median, mean, etc)
- Using more robust estimation techniques (e.g. M-estimation for regression).

Outlier analysis

Detecting outliers

Check github: https://github.com/isa-ulisboa/greends-avcad-2025/tree/main/examples

Outlier_analysis.ipynb

Lesson #6

1. Non-parametric hypothesis testing
2. Outlier analysis
3. Data transformation
4. Topics for the final project

Data transformation (*sensu lato*)

Data transformation (*sensu lato*) involves performing different kinds of operations to prepare data to be analyzed, such as:

1. Manipulate the form of the data (data wrangling)
2. Variable standardization and normalization
3. Variable transformation
4. Engineer features in the data

For some of these operations (e.g. 2 and 3), a previous exploratory data analysis must be carried out on the data to assess the necessity and kind of transformation to be carried out.

Data transformation

1. Manipulate the form of the data (data wrangling)

Reshape a data set to make it ready to be analyzed. Examples are:

- Reordering/selecting rows
- Renaming/selecting columns
- Handling missing values
- Transpose data
- Removing duplicate values
- Stack/unstack data
- Pivoting data

Data transformation

2. Variable standardization and normalization

When the analysis needs the data to have similar units, it is a way of rescaling variables to a common scale **without changing the distribution**. Examples are:

- **Data standardization** - involves centering and scaling variables, respectively, to Mean=0 and SD=1:

$$X_{stand} = \frac{x_i - \bar{x}}{s}$$

- **Normalization** - involves for example rescaling variables to values between 0 and 1.

$$X_{norm} = \frac{x_i - min(x)}{max(x) - min(x)}$$

Data transformation

3. Variable transformation (strict sense)

Modify the values of a variable to fulfill certain statistical assumptions, e.g., normality, homogeneity, linearity, ... In this case the distribution of the data is changed

Examples are:

- Logarithmic, square root, cube root, … transformations
- Box-Cox transformation: $$\omega = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda} & when\ \lambda \neq 0 \\ \log(Y) & when\ \lambda = 0 \end{cases}$$

  => involves estimating λ (if λ=1 then Y is already normally distributed)

- Arcsin transformation – for proportions

Data transformation

4. Engineer features in the data

Generate new variables based on existing ones. Examples are:

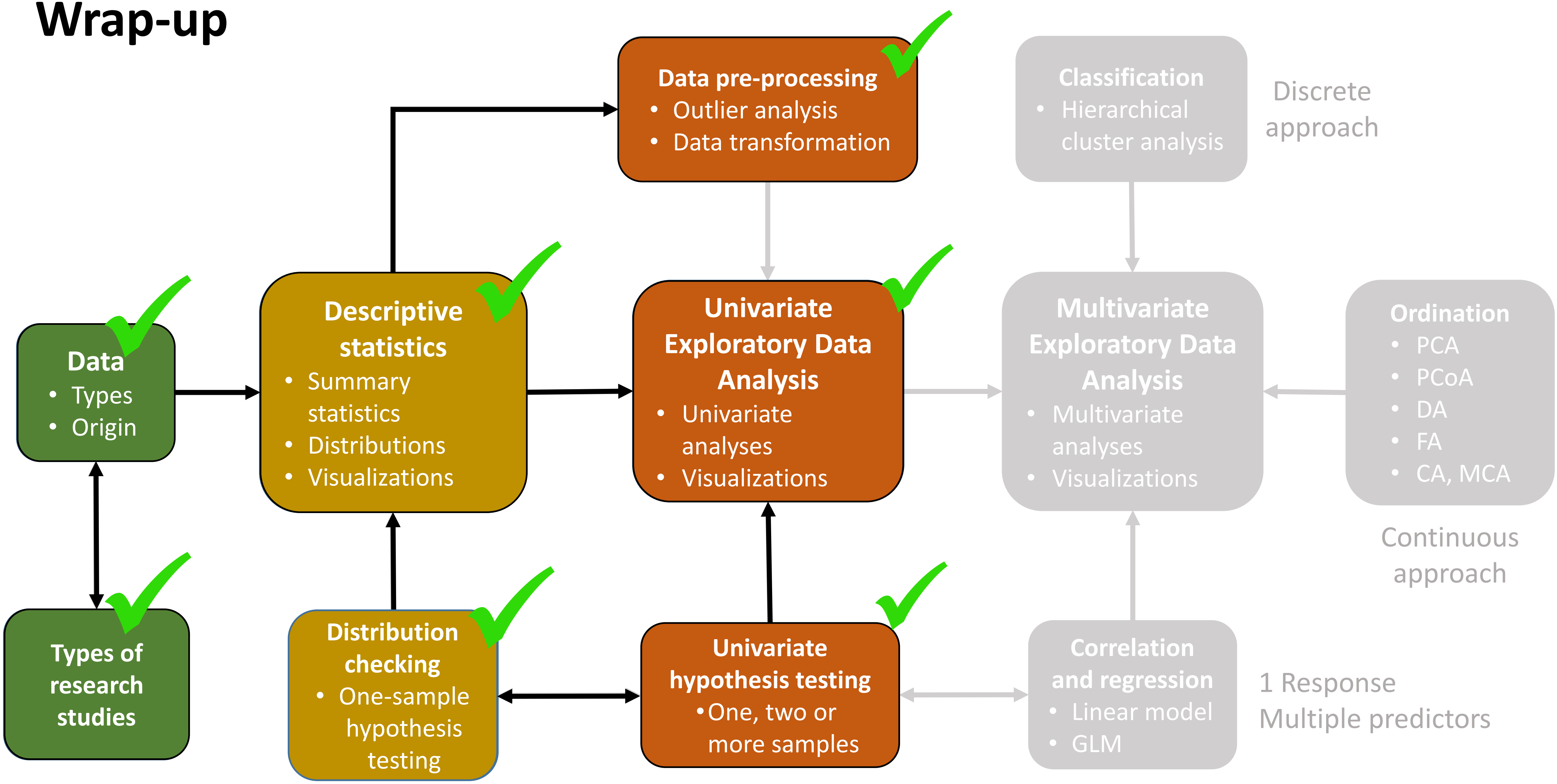
- Aggregate values for each category of a factor, using functions such as the sum, mean, maximum, ...

- Generate new variables by applying a given function (e.g. sum) to a set of columns in a data table - e.g. generate a community species richness variable from presence/absence data of single species.

Some working examples on:
https://github.com/isa-ulisboa/greends-avcad-2025/tree/main/examples

Transformation.ipynb

Wrap-up


Data pre-processing
• Outlier analysis
• Data transformation
Classification
• Hierarchical cluster analysis
Discrete approach
Data
• Types
• Origin
Descriptive statistics
• Summary statistics
• Distributions
• Visualizations
Univariate Exploratory Data Analysis
• Univariate analyses
• Visualizations
Multivariate Exploratory Data Analysis
• Multivariate analyses
• Visualizations
Ordination
• PCA
• PCoA
• DA
• FA
• CA, MCA
Continuous approach
Types of research studies
Distribution checking
• One-sample hypothesis testing
Univariate hypothesis testing
•One, two or more samples
Correlation and regression
• Linear model
• GLM
1 Response
Multiple predictors

Lesson #6

1. Non-parametric hypothesis testing
2. Outlier analysis
3. Data transformation
4. Topics for the final project

Final project

Goal: tell a coherent story from a complex agro-environmental dataset

Steps:
1. Questions/hypothesis definition
2. Preparation of datasets, through database queries and data wrangling
3. Summary statistics
4. Exploratory data analysis
5. Inferential statistics
6. Final visualisation product and storytelling

Assessment:
1. A **live presentation** of your story, a **poster**, or a **1-page interactive dashboard**
2. A short **written report**, including the code as an Appendix.

Groups
- 2-3 students maximum

Final project

Databases:

- INE database
- Other databases of student's interest
- National Forestry Inventory

Report

Should include the following chapters:

1. **Introduction** – Short introduction to the topic, ending with questions/working hypothesis to be addressed and objective (2 pages max.)
2. **Database description** – Short descriptive statistics of the database/tables (2 pages max.)
3. **Exploratory data analysis** – This will be the most important chapter, where you will try to tell a coherent history by means of numerical outputs and visualizations (10 pages max.)
4. **Discussion/Conclusions** – a short discussion/main take home messages/conclusions of the work (2 pages max.).
5. **References**
6. **ANNEX – Python code.**

Final project

INE database

Spatial resolution: civil parishes

Types of variables:

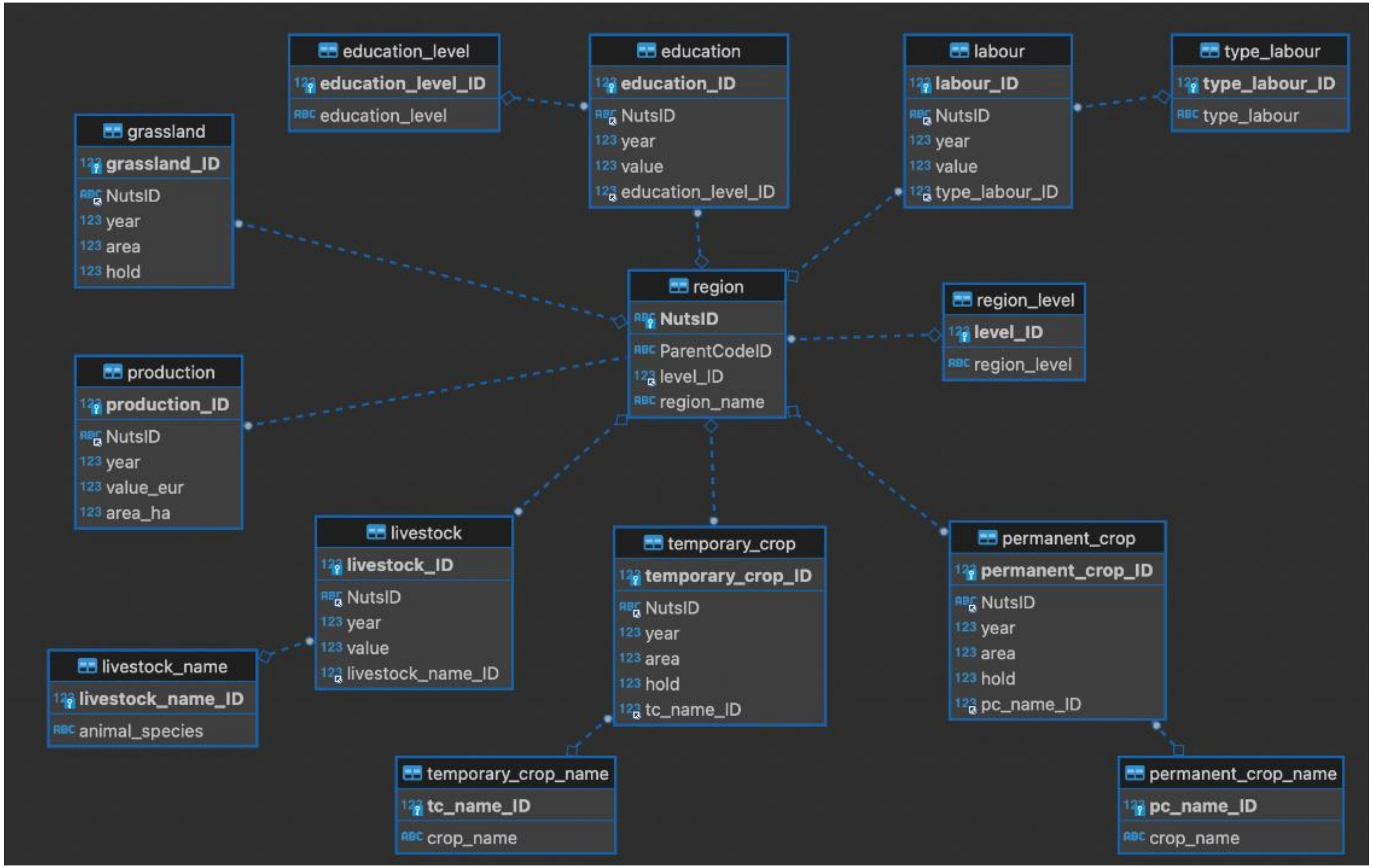
- Education – education levels of agricultural populations
- Workforce - Volume of agricultural labour force (AWU) and Type of labour force
- Production (value per área and total)
- Livestock (number of holdings per species)
- Grasslands (area and number of holdings)
- Temporary crops (area and number of holdings)
- Permanent cultures (area and number of holdings)

National Forestry Inventory

Catographic data on abundance, state and condition of national forest resources.

Final project

Structure of the INE database



Final project

Possible topics

1. Farmland socio-economic analysis: geographical patterns and temporal trends
2. Analysis of livestock activities: geographical patterns and temporal trends
3. Analysis of agricultural activities: geographical patterns and temporal trends
4. Relationship between socio-economic variables and livestock activities
5. Relationship between socio-economic variables and agriculture activities
6. Spatial and temporal analysis of forest resources changes (e.g. differences between NUT II or III regions and across years: 1995, 2005, 2010, 2015).

Exercise 6

1. Using the EFIplus_medit.zip dataset, test if the frequency of sites with presence and absence of *Salmo trutta fario* (Brown Trout) are independent from the country. Please state which is/are the null hypothesis of your test(s). You may try to produce an alluvial plot.
2. Run the non-parametric equivalent of the test you used in exercise 5.3 and compare with the ANOVA test (5.2: Test whether there are differences in the mean elevation in the upstream catchment (Elevation_mean_catch) among the eight most sampled catchments. For which pairs of catchments are these diferences significant? Please state which is/are the null hypothesis of your test(s)).
3. Using the winequality_red.csv file in the examples folder of the github repository, test which wine parameters discriminate the best between wine quality scores categorized into two classes using value 5 as the threshold value (quality>5="good" and quality<5="bad").