Long-term Dynamics and Peasant Autonomy in the Italian Countryside

Roberto Ragno

Table of contents

Overview

AFFILIATION

Dipartimento di Ricerca e Innovazione Umanistica, Università di Bari Aldo Moro, Bari, Italy

EMAIL

roberto.ragno@uniba.it

This research aims to diachronically trace the patterns of subsistence, economy, and environmental change in relation to regional patterns of Italian peasantry in the 1st millennium CE. Scholarly debate on the dynamics of human-nature interaction in Italy during the transition from Late Antiquity to the early Middle Ages still leaves today some open questions. Were the former Roman economic structures and farming practices completely abandoned or did they find continuity during this turbulent time? Which agricultural strategies did peasants develop to cope with political, demographic and climatic change? Past work on agricultural production has been based upon literary sources and field surveys which identify boundaries; a multi-source archaeological study is absent from the discourse. Drawing on environmental proxies, such as animal, seed, and pollen remains, this project uses data from >261 sites [1] (172 botanical and 379 faunal samples[2]) and statistical/dimensionality reduction methods (CA, NCA, nMDS). Specific emphasis is placed on evaluating agricultural strategies during the shift from the Roman Empire to the politically fragmented landscape of early medieval Italy, to assess the role of political organization, economy, culture, and environment in the configuration of agricultural regimes and animal/plant husbandry selection. The integration of different sources is fundamental to casting new light on peasants' lifeways, which may be obscured in the textual records, privileging elite contexts and high-status transactions. The Early Middle Ages mark a period of fundamental change when different geographical regions (and potentially micro-regions) developed their own political and economic frameworks for agricultural production. Bioarchaeological evidence can help visualise the landscape in which these changes were taking place, as archive collections of samples are subjected to new analyses in a holistic context, through the database I have constructed. For example, one can analyse differences in diet and production between elite, religious, urban and rural sites, and conduct regional pattern analysis. Preliminary results show indeed higher degrees of regionalization in agricultural strategies during the EMA. My dissertation will contextualize these findings against textual sources to assess how people used the Italian agricultural landscape,

[1] Last update: July 20th 2022.
[2] Pollen tbd.

and suggest which agents were responsible for changes. I argue that the discourse on Northern Italian/French early Medieval production has heavily influenced interpretations of the agrarian economies of the rest of the peninsula.

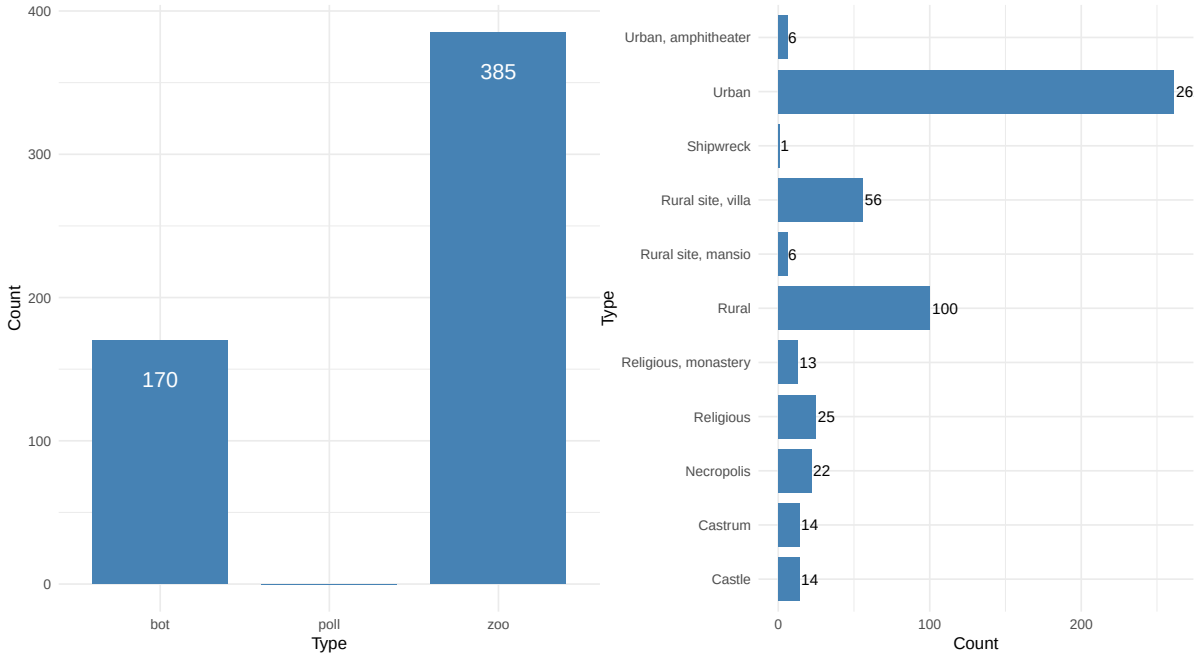
Database status

Distribution map

The map below shows the distribution of the samples used for this project.

Counts

The graphs below (Figure 0.1) provide counts of the (a) sample types[3] and (b) site types in the database.



(a) Data type.

(b) Site type.

Figure 0.1: Database status

[3]Pollen tbd.

1 Introduction

! Page under construction

2 Literature Review

! Page under construction

2.1 Introduction

2.2 Continuity vs rupture

2.3 Environmental archaeology

Maybe move the processual arch. part somewhere here or create a general 'science in arch' section

2.3.1 Archaeobotany

Name origin: Archaeobotany, paleobotany, paleoethnobotany...

2.3.2 Zooarchaeology

Classical intro with name zooarchaeology vs archaeozoology

2.4 Landscape studies

2.5 Diet studies

2.6 Big data in archaeology

2.7 Statistical analysis in archaeology

General introduction (Shennan?)

In archaeology, multivariate statistics has been applied since the mid-60s, when the spread of computers and statistical packages made these methods more easily applicable. The growing popularity of computational archaeology in this period also owes a great debt to the New Archaeology movement (or Processual archaeology). New Archaeology emphasized the application of rigorous scientific analysis at the expense of the cultural historical approach which focused on artifacts cataloging based on ethnic grouping (Binford and Binford (1968)).

Include Baxter and Drennan

2.8 Conclusions

3 Materials

> **!** Page under construction
>
> This page will probably be merged with “Methods”

3.1 Introduction

> **!** Section in progress

3.2 Archaeobotany

> **!** Section in progress

Should the maps be here?

3.2.1 Carpology

3.2.1.1 Database storing procedures

3.2.2 Palinology

3.2.2.1 Database storing procedures

3.3 Zooarchaeology

> ! Section in progress

Should the maps be here?

3.3.1 Database storing procedures

4 Methods

> ! Page under construction

4.1 Introduction

This chapter presents the methodology employed to carry out this study. The first section discusses statistical computing programming languages such as R and Python. The second section introduces databases and their employment in archaeology, then outlines the construction of the database used for this research. GitHub as a hosting service for this research is discussed in section three. The fourth section describes the logic behind the choice of the temporal boundaries, the creation of chronologies and presents a possible solution to mitigate the problem of chronological fuzziness of samples.

4.2 Statistical computing

Statistical computing, often referred to as computational statistics, is a branch of statistics that uses computational approaches to solving statistical problems. Traditionally, the term *statistical computing* places more emphasis on numerical methods, while *computational statistics* refer to topics such as resampling methods, exploratory data analysis, neural networks, etc. (Rizzo (2019)). Specifically, computational statistics deals with methods "unthinkable before the computer age" (Lauro (1996)). In Archaeology, mainly two programming languages are used today for exploring large sets of data: R and Python. Both programming languages can be used for statistical modeling, although R was designed explicitly for statistical computing and Python as an object-oriented language for more general purposes. Python is one of the most used and fastest-growing programming languages in the world, and its package number is extremely high compared to R. While R benefits from a strong academic community, Python relies on developers to add new packages almost continuously. R was used for the majority of this project due to its familiarity, while Python was used for certain algorithms that were easier to implement.

Figure 4.1: Logos of R and Python

4.2.1 The *R* programming language

R, as described on the R-Project FAQs is a "system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files". Increasingly popular for data scientists, R is based on S and provides its own IDE (the R GUI), although RStudio[1] is the most popular IDE for computing the R language. For this project, RStudio was the standard IDE.

4.2.1.1 R Packages

In addition to base R, several packages enhance its performances and offer more tools to users. The packages are distributed by the official CRAN repository, which counts more than 18,452 packages[2].

4.2.1.1.1 tidyverse

The **tidyverse** ecosystem (Wickham et al. (2019)) is a core set of packages for R, maintained by Hadley Wickham for importing, tidying, transforming and visualising data which includes packages such as—ggplot2, dplyr, tidyr, stringr, tibble, forcats, purr, readr.

4.2.1.1.2 ggplot2

ggplot2 is the most common data visualisation package for R, included in the tidyverse environment. The package substitutes the R base graphics and allows visualisation of single and multiple components (Wickham (2016)).

[1] RStudio announced on July 2022 that its name will change to Posit, as it is expanding to Python and Visual Studio Code.

[2] Updated on August 16th, 2022.

4.2.1.1.3 knitr

The **knitr** engine enables the integration of R with HTML, Markdown and LaTeX. The package allows reproducible research (Xie (2021)) and was used for generating dynamic reports and documentation for this thesis.

4.2.1.1.4 vegan

The **vegan** package (Oksanen et al. (2022)) is designed for ordination methods, diversity analysis and multivariate analysis in ecology.

4.2.2 The *Python* programming language

Python is a popular high-level general-purpose programming language that supports object-oriented, procedural and functional programming. Several IDE support Python, including JupyterLab (open source), RStudio, PyCharm and Visual Studio Code. In recent years, many statistical packages have been developed to compete with R in the data science field. This project uses Python in addition to R as its environment offers well-designed and intuitive packages for some of the data-reduction methods used on the datasets.

4.2.2.1 Python packages

Python's motto "batteries included" refers to its comprehensive standard library. However, over 390,000 packages are contained in the PyPi repository providing users with many options for coding. In particular, this project uses some data science libraries (set of packages) and packages listed below.

4.2.2.1.1 Pandas

Pandas is the most used package for data handling, manipulation and analysis (Reback et al. (2022)).

4.2.2.1.2 Matplotlib

Matplotlib is Python's standard data visualisation library (Hunter (2007)).

4.2.2.1.3 Seaborn

Based on Matplotlib, **Seaborn** is a data visualisation library that allows users to create more complex graphs (Waskom (2021)).

4.2.2.1.4 Scikit-learn

Scikit-learn is the most comprehensive Python library for machine learning, including methods for classification, clustering, data reduction and regression (Pedregosa et al. (2011)).

4.2.2.1.5 NumPy

The **NumPy** library is the main tool for array programming and it includes several functions for numerical analysis (Harris et al. (2020)).

4.2.2.1.6 SciPy

The **SciPy** library provides tools for scientific computing (data integration, interpolation, optimization, linear algebra, etc.) and it works with NumPy multidimensional arrays (Virtanen et al. (2020)).

4.3 GitHub: hosting the project

> **!** Section under construction
>
> This section will include a general intro to GitHub and why I chose to host data there.

4.4 Database

4.4.1 What is a database?

Databases are increasingly being used in archaeology to archive and collect data digitally. There are mainly two types of databases—relational and non-relational. A **relational database** is more appropriate for well-defined data structures which can be linked through a mutual attribute. It is built and maintained with Structured Query Language (SQL), which allows the user to interrogate the database through queries (Gattiglia (2018)). A SQL database consists of several tables containing information in columns (variables) and rows (entries). Each row is defined by an unique key. Examples of relational database management systems include—MySQL, PostgreSQL, MariaDB, Microsoft SQL Server and Oracle Database. A **non-relational database** (NoSQL) is advantageous in the case of unstructured data, as data is archived as a single document rather than in a table. This structure allows much more flexibility, although NoSQL databases can be harder to use by non-specialists. Among the NoSQL database management systems, MongoDB is the most widely used.

4.4.2 Databases in archaeology

Databases are used in archaeology primarily for two reasons—data management and data sharing. Most excavations are now working with databases, where the information concerning each stratigraphic unit and finds is recorded. Databases can also be linked and interact with Geographical Information Systems (GIS), that allow researchers to work with a spatial component. In many cases, this data remains private, even after the publication of the excavation results, although increasingly more teams are also making their data available to the public. The growing popularity of open data has also created the need for standardisation. Since the 1970s, researchers started working on *thesauri* (Figuera (2018)), dictionaries and guides for the correct archival of archaeological information (e.g. pottery classes, context types, chronologies, etc.). A shared system of naming practices is essential for sharing, integrating and analysing data. Recently, more standardised databases and repositories are being created and openly published:

- Archaeological Data Service (ADS). A non-profit organisation, based at the University of York. The website provides a large repository of downloadable archaeological data (mostly from the British isles) (Richards (2021)).
- ARIADNEplus, a Horizon 2020 project funded by the European Commission aiming to build an integrated european archaeological data structure. Over 2 million datasets are part of this project, with the original data still managed by the original creators (Niccolucci (2020)).

Most of the archaeological databases are based on SQL, as the visual relationships between different data structures are easier to understand.

4.4.3 Creating an Environmental Archaeology database

An integrated database with environmental archaeology data is still missing in Italy. A first step towards botanical data digitalisation has been the BRAIN project (Botanical Records of Archaeobotany Italian Network), a census of the Italian excavations that reported archaeobotanical data (Mercuri et al. (2015)). The website, although not providing raw data, has been useful to the bibliographical research for this project. For what concerns faunal remains, a database is missing, although a pilot project was started at the University of Siena by Boscato et al. (2007). The database was created using FileMaker Pro and was likely never published.

For the scope of this project, it was thus necessary to create a database that contained raw environmental data. The creation of a database for this research responded to the need of a systematic approach in storing environmental data in a common format and in a way that is convenient for querying, rather than merely archiving the information. The goal is to have data readily available for exploratory data analysis, in an automated process that does not

require to adapt the query and manually wrangle data each time a new sample is added to the collection. This project uses MariaDB, a fast and stable fork of MySQL.

The database was structured with the creation of 21 tables. The table `site_list` is the core table, from which most of the other tables in the schema depend. The entries in the table are based on the chronology of a single sample. The chronology is defined both by the macroperiod[3] and by the centuries, with two columns: `startcentury` and `endcentury`. If a context has been sampled more than once, there will be as many entries as many chronologies. If a context has been sampled both for seeds and bones, the sample IDs will be recorded on the same entry (if the chronologies match).

Context	Sample #ID
Imola, villa Clelia, 6th c.	15
Imola, villa Clelia, 10th-11th c.	16

Since the table is context based, each site has been provided with an unique ID, so that if the site has been sampled in different areas the database can be still queried using the site ID[4]. In addition to the reference to the samples table, the table `site_list` also contains references to child tables with information about: site type, geography, altitude, coordinates, and region. The samples are organized in three tables, depending on the type of remain that has to be recorded in the database:

- `plant_remains`
- `pollen_remains`
- `faunal_remains`

Each entry in the samples tables contains information about chronology, sampling type and notes, reference to the publication, reference to the main site, and the raw data. The tables `faunal_chrono_weights` and `plants_chrono_weights` contain the automated calculation of the samples weights (refer to Section 4.5.1). The Figure 4.2 shows the schema with the complete list of tables in the database.

The data was stored in the database after a thorough bibliographical research of the excavated sites (with a chronology pertaining to the 1st millennium CE) where environmental analyses have been undertaken. As in most of the cases the material was retrieved from physical publications, not digitised, the process of data entry was manual and could not be automatised. A list of the publication types where data was retrieved from can be visualised in Figure 4.3.

[3] See Section 4.5

[4] If the site is very large and the samples were numerous (e.g. Pompeii, Rome), the sites were recorded with different IDs.

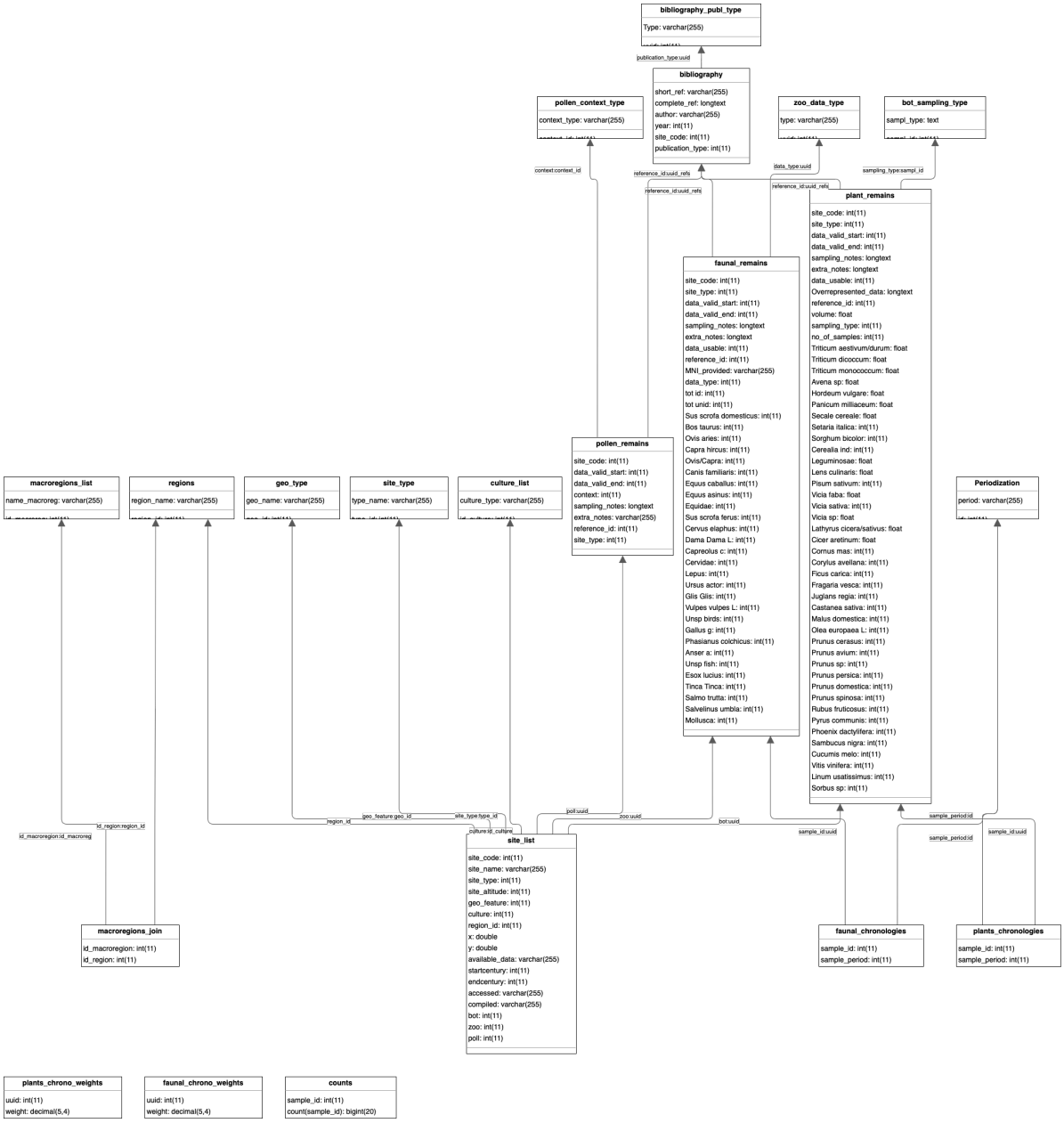


Figure 4.2: Schema of the database, with the table `site_list` at the core.

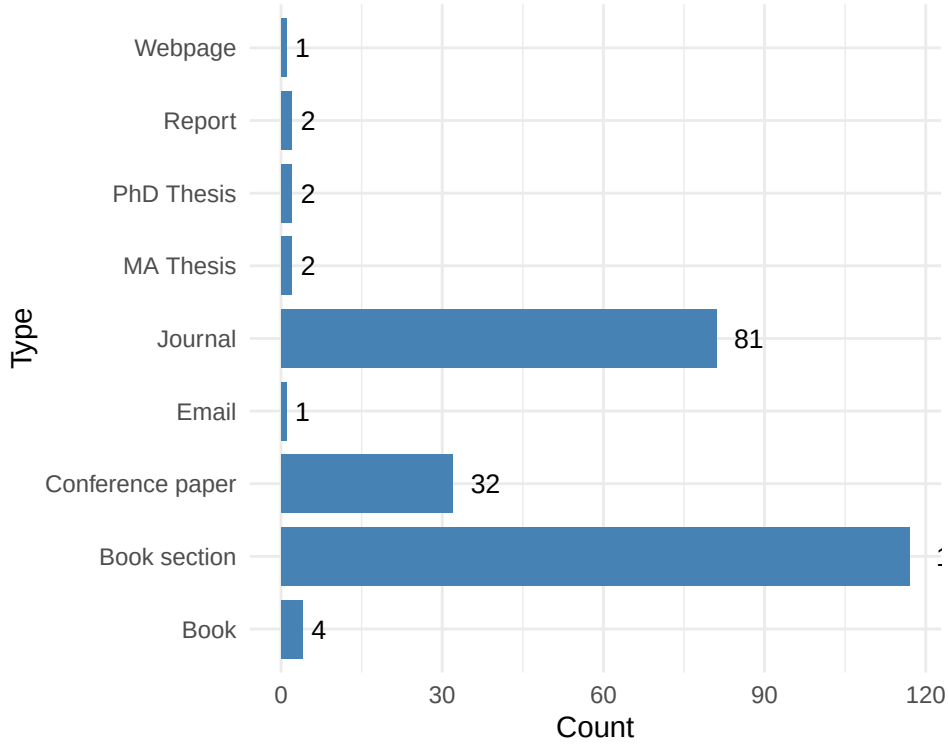


Figure 4.3: Count of the publication types in the database.

4.5 Periodization

Some of the statistics performed on the dataset have been based on sample periodization, from the Roman age to the Medieval age. The chronologies have been defined as follows:

- [**R**] **Roman**: from the 1st century BCE to the 2nd century CE.
- [**LR**] **Late Roman**: from the 3rd to the 5th century CE.
- [**EMA**] **Early Middle Ages**: from the 6th to the 10th century CE.
- [**Ma**] **Middle Ages:** from the 11th century CE onwards.

In the database, the tables `faunal_chronologies` and `plants_chronologies` connect each bioarchaeological sample to another table with the identification numbers for the periods (e.g. Sample 1 = ID 1). If a sample has a chronology ranging between two periods, two separate entries will be recorded on the database (e.g. Sample 1 – 2nd to 3rd c. CE = Periods: Roman, Late Roman) with the result of the sample being repeated in both periods.

Figure 4.4: Periodization schema.

4.5.1 Chronological fuzziness

One of the methodological issues affecting this project is that of chronological fuzziness. Dating plant and animal remains using radiocarbon is very rare, at least in the samples recorded in the database. Most of these samples are dated using ceramics, and chronologies can range between one century or several. Taking this into account, I weighted each sample as follows:

$$W = \frac{1}{(C_{end} - C_{start} + 1)}$$

Where:

- C_end is the terminus ante quem.
- C_start is the terminus post quem.

- 1 has been summed to the denominator to avoid 0 values.

The imported tables already contain a column of weights, as this operation has been performed on the database prior the export. A diachronic table of means for both datasets has been generated using the functions:

- `zooarch_tables` (custom)
- `Bot_mean_table` (custom)
- `Bot_mean_fun` (custom)
- `weighted.median` (from the package `spatstat`)

The weight can be used for weighted means and medians, with samples with larger chronologies (hence less precise/fuzzy) weighting less. This method provides each sample with a weight proportional to the length of its chronology so that lower weight values have a smaller impact on the computations.

4.6 Multivariate statistics

> ! Section in progress

This research uses multivariate analysis to explore possible relationships within the sets of environmental data under investigation. Univariate analyses can be easily plotted and visualized with bar charts, histograms and density curves. A scatterplot (or scattergram) shows the relationship between two variables by plotting their values on the axes of a diagram using Cartesian coordinates. This relationship can also be mathematically measured by calculating a distance between two points in the graph. However, the relationship between more than two variables is much harder to read on a scatterplot, as it would require as many axes as the number of variables. If our analysis is exploratory, we might not know yet where to look for correlations. A possible solution is analysing each combination of two variables in the dataset and measure their correlation. This would require too much computation time if the dataset had a large number of variables, that is to say if we are dealing with big data. In addition, we would only gather information about the relationship between two variables, when there might be another factor influencing a trend or phenomenon. Multivariate statistics has a wide range of applications, including grouping and multidimensional scaling, as well as in machine learning and predictive analysis (Fletcher and Lock (2005)). In this research, multivariate methods have been applied both to the botanical and faunal datasets, mainly with the objective of *(i)* testing hypotheses between multiple variables and *(ii)* reducing the dimensionality of data to—assess which variables are the main drivers of change in the datasets and examine relationships between these variables.

4.6.1 Statistical hypothesis testing

Explain what is hypothesis testing

4.6.1.1 PERMANOVA

Permutational multivariate analysis of variance (PERMANOVA) is a non-parametric multivariate statistical test used to compare group of objects. By using measure space, the null hypothesis that the centroids and dispersion of groups are identical is tested. The null hypothesis is rejected if either the centroid or the spread of the objects differs between the groups. A prior calculation of the distance between any two objects included in the experiment is used to determine whether the test is valid or not. (Anderson (2017)).

4.6.2 Dimensionality reduction and ordination

Dimensionality reduction techniques transform high-dimension datasets into a low-dimension ordination space, with the intention of maintaining the geometry and data structure as close as possible to the original dataset. The **dimension** of a dataset is given by the number of variables (*i.e.* the columns in the table). As anticipated in Section 4.6, as each variable is graphically represented by an axis, it would be virtually impossible to represent more than three axes in a graph. Ordination allows to reduce data dimensionality to usually one to three (at most) dimensions. Moreover, focusing on a reduced number of dimensions reduces the amount of "noise" that can mislead the interpretation (Gauch (1982)). The points generated through ordination techniques (the objects in our dataset) can eventually be plotted in a scatterplot. In most of the ordination methods, points plotting closer together in graph are more similar, whereas points far apart from each other are more dissimilar (Shennan (1997, p. 197)). For instance, one could perform an ordination on a group of burials in a cemetery where each point represents a single burial assemblage. After the ordination it is also possible to group the new reduced set of variables, to observe differences between groups and facilitate the interpretation. In the previous example, a group might be represented by burials of the same ethnic group, status, etc.

Many of the ordination techniques described in this chapter developed in fields outside archaeology, and are thus borrowed from disciplines as community ecology. Ecologists regularly apply ordination methods for the study of environmental gradients, so that the term "gradient analysis" is often used interchangeably with "ordination". An environmental gradient refers to variations in site characteristics (*e.g.* time, elevation, temperature, vegetation, etc.), which in turn can affect biotic factors (*e.g.* species abundance, diversity, etc.) (Grebner et al. (2013)). The purpose of ordination is then to identify the cause of ecological processes in the dataset. Generally, it is possible to apply ordination on datasets in which the variables have a cause-and-effect (*e.g.* climate vs. plant species) or mutual influences on each other. There are two main types of ordination, or gradient analysis, techniques (see Table 4.2): **direct** (constrained)

or **indirect** (unconstrained). The objective of indirect (unconstrained) gradient analysis is to identify patterns between samples of 'dependent variables' (*e.g.* which sites are more similar according to their species composition). Conversely, direct gradient (or constrained) analysis includes more information (or tables) in a single analysis—the dependent variables are now constrained to be a function of other sets of 'independent variables' (usually environmental proxies). In short, a constrained analysis uses both datasets to find the best possible mathematical relationship between the dependent and independent variables. In this sense, direct gradient analysis can be considered as an extension of unconstrained analysis (Syms (2008)). The choice between using constrained or unconstrained methods for ordination strictly depends on the research questions and on the researcher's dataset.

Table 4.2: Ordination methods used in gradient analysis. Table after Cuffney et al. (2014, p. 149)

Response model	Indirect gradient analysis	Direct gradient analysis
Linear	Principal component analysis (PCA)	Redundancy analysis (RDA)
Unimodal	Correspondence analysis (CA) Detrended CA (DCA)	Canonical correspondence analysis (CCA) Detrended CCA
Monotonic	non-metric multidimensional scaling (nMDS)	

4.6.2.1 PCA

i Eigenvalues and eigenvectors

Both PCA and CA are methods based on eigenanalysis. The analysis of eigenvalues and eigenvectors will not be described in detail in this chapter, and the meaning of the terms will be only briefly explained in the text. However, some information must be given for the eigenanalysis-based ordination methods:

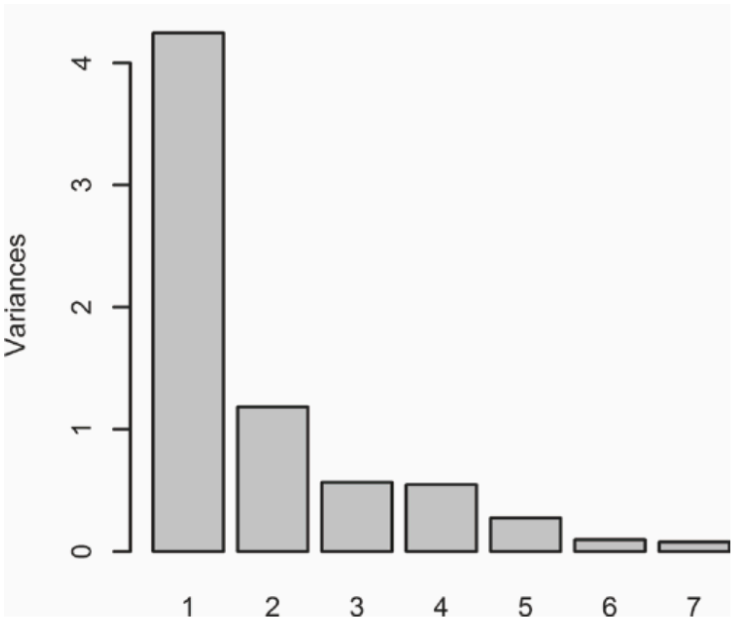
- The methods are performed on a square symmetric matrix (for instance a correlation matrix);
- The order of the variables is not important for the result;
- Axes are ranked according to their eigenvalues (*i.e.* the first axis has the largest eigenvalue, the second the second-largest, etc.)
- Eigenvalues have mathematical meaning that is important for the interpretation of our data.

Principal component analysis (PCA) is an useful tool to identify possible patterns or subgroups in data. The input variables in the dataset must be numerical (or at least dichotomous). If used for machine learning, PCA is in fact an *unsupervised* method, meaning that the data points will be unlabelled after the transformation. As for other dimensionality reduction methods, the algorithm tries to preserve the global structure of data by finding a space in which dimensions can be reduced without losing much information. In other words, PCA optimizes the loss of variance. The first step in PCA is creating a correlation matrix or covariance matrix, to extract the eigenvalues and eigenvectors. The decision between correlation or covariance matrix depends on our data and on our research questions. Correlation matrices are helpful if our variables have different units (*e.g.* lenght, volume, etc.). Covariance matrices can be used if the variables have the same unit and comparable variances. The latter is an important factor to consider before choosing covariance matrices, as the biggest variances will overshadow the others (Carlson (2017, p. 267)). For PCA, the **eigenvalues** are the variances of the principal components (so that the first component has the largest eigenvalue) and the **eigenvectors** are the principal component loadings (the score for each variable in the new dimensional space). After extracting the eigenvalues and eigenvectors, the user must decide how many components he will use. The components (defined as the sum in % of the eigenvalues) are the axes in our new dimensional space. Most of the times, the first two components are enough to explain the variance in our dataset. For instance, if the first two axes explain 80% and 12% of the variance, it is acceptable to use a 2D graph—the graphical representation of the two axes will explain 92% of the variance. It is possible to choose the appropriate number of components by using a scree plot, which shows the amount of variance for each component. Scree plots are an important of PCA, because they allow us to see how many components have a significant variance (eigenvalue). A significant variance must be greater than 1. If more than 2 or 3 components have a variance greater than 1, PCA might not be the best method for reducing our set of data. The example in Figure 4.5 shows that PCA in this case is an appropriate method, as only the first two components have variances greater than 1. Therefore we can discard the other components as they are of little significance.

Most of the statistical packages offer an automatic rotation of the axes for an easier interpretation of the results, which is more important if the data has been standardised prior to the calculations.

4.6.2.1.1 Interpreting a PCA

PCA are often visualised through biplots, which shows at the same time the scores of the observations (rows) and the loadings of the variables (columns). Generally, the x-axis of the plot represents the first component, and the y-axis the second component. Making sense of the plotted **observations** is straightforward—points that are closer together are more similar and they can create clusters. It is important to remember that first component of the PCA lays on the x-axis, so differences in clusters on the first component (*i.e.* the horizontal distance) is actually larger than differences in clusters on the second component (*i.e.* the vertical distance). Interpreting the **variable** loadings can be more difficult. The variables are represented by



plot PCA.png plot PCA.bb

Figure 4.5: Example of a scree plot showing that only the first two components have a variance greater than 1. Figure after Carlson (2017, p. 271)

arrows, implying that the variable increases in the direction of the arrow and decreases in the opposite direction. The direction of the arrow is given by the loadings of the first component—if the loadings are negative the arrows will point to the left. If the signs of the loadings are flipped the result will be the same but the arrows will point towards opposite directions. The arrows show how strongly each variable influences a principal component, with longer arrows being the best described in the dimension. The angle between the arrows also indicate extra information useful for the interpretation:

- If two arrows form a small angle, the two variables are positively correlated.
- If two arrows form a large diverging angle, the two variables are negatively correlated.
- If two arrows form an angle close to 90°, there might be no correlation at all.

Figure 4.6 shows an example of a PCA biplot. All the arrows (variables) point to the same direction, but we can interpret their correlations using the angles as described above. For example, B1 and B2 are not correlated, as they form an angle of 90°. On the contrary, the variables B and L are positively correlated, as they form a very small angle.

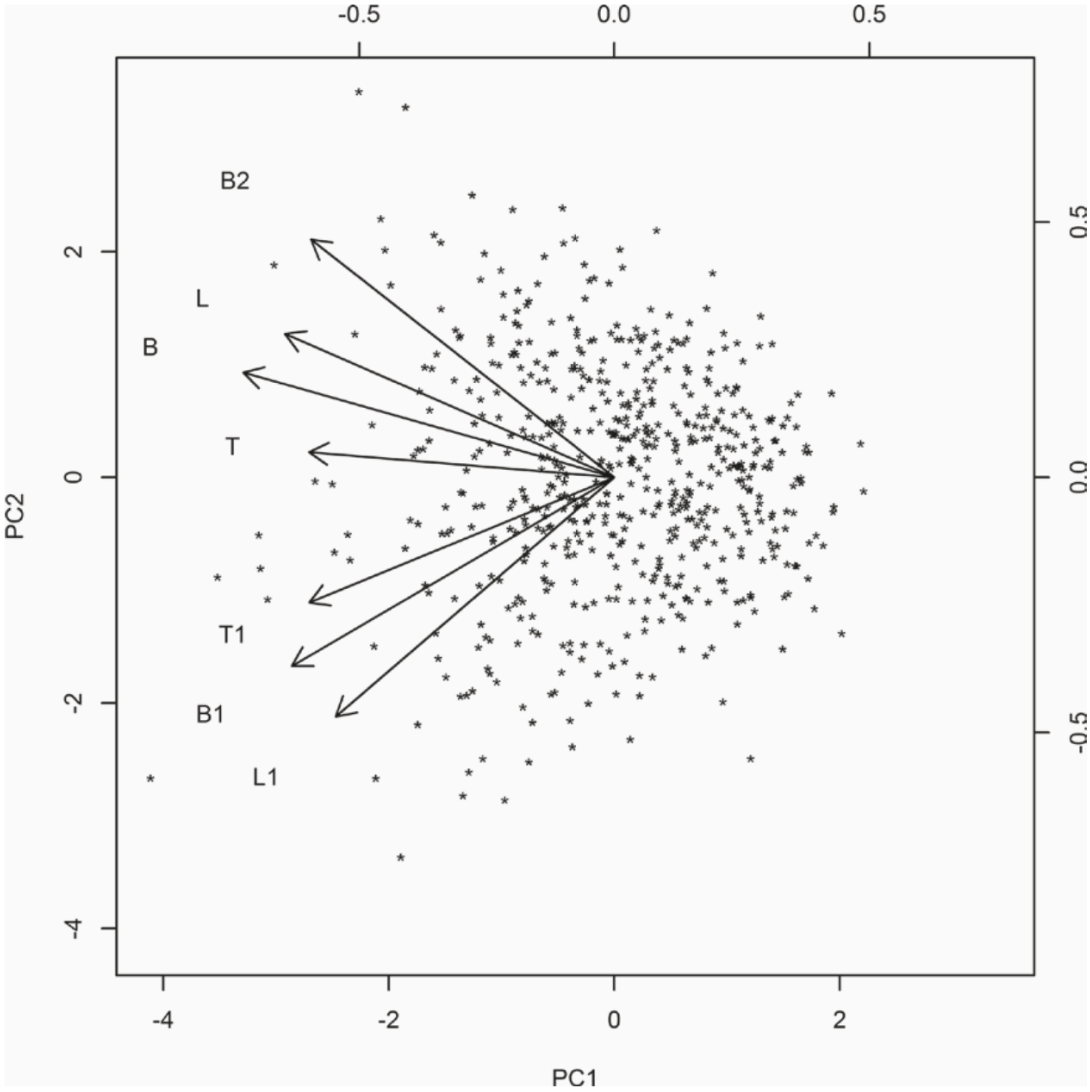


Figure 4.6: PCA Biplot, after Carlson (2017, p. 274)

4.6.2.2 CA

Correspondence analysis (CA) is a dimensionality reduction technique that can be applied to contingency tables. Contingency tables (or crosstabs) are two-way tables used to display frequency data formed by two categorical variables. As for PCA, it is an indirect gradient analysis[5]. Instead of maximizing the variance, CA maximizes the degree of correspondence between rows and columns (*i.e.* observations and variables). CA is appropriate for exploring non-negative data (e.g. relative abundance, counts, presence/absence) in a contingency table, and examining the relationships between cells in a row, in a column, and their interrelationship (Baxter (2015, p. 101)). Moreover, CA is suited for categorical or qualitative data.

On a practical level, CA can be used to compare assemblages from different sites, to answer questions such as "Which sites do certain finds correspond to?" or "Which sites are more similar according to their assemblages?". This technique is extremely useful for archaeologists, even though its reception by the archaeological community has been slow. The first uses of CA in archaeology have developed from the mid-70s in France, but the technique was only introduced to the English-speaking audience in the 80s (**Cite Baxter 2010 + Bolkvien et al 1982**) and became more popular in the 90s (**Baxter 1994**).

eigenvalues are different from pca eigenvalues

Sensitive to outliers and rare species

In `R`, common implementations of this algorithm are included in the packages `MASS`, `FactoMineR` and `vegan`. Correspondence analysis has been used several times in this research using the function `cca()` from the `vegan` package, which was applied to the presence/absence archaeobotanical dataset (**add where else**[6]). The function `cca()` performs a canonical correspondence analysis (CCA), but it can also be used for correspondence analysis.

4.6.2.2.1 Interpreting a CA

CA can also visualised through biplots as for PCA.

4.6.2.3 nMDS

Multidimensional scaling (MDS) is a technique to visualise the level of similarity of individual observations (e.g. sites/cases) in a dataset. MDS works with matrices containing Euclidean distances between each pair of observations. Conversely, **non-metric multidimensional scaling** (nMDS) is a rank-based approach that finds both:

[5] It is important not to confuse correspondence analysis (CA) with canonical correspondence analysis (CCA), which is a direct gradient analysis method.

[6] remember to change this

- A non-parametric monotonic relationship between the items in the dissimilarity matrix and the Euclidean distances.
- The location of items in the low-dimensional space.

The goal of nMDS is to represent the pairwise dissimilarity between items in the matrix as closely as possible. For this reason, it is considered as a good technique for multivariate data visualisation. nMDS can be used on quantitative, qualitative and mixed data. For this project, it has been applied to a subset of the early medieval archaeobotanical dataset that included presence/absence values of cereals. The full breakdown of the process is reported in Section 5.4.2. The `R` function `metaMDS` from the package `vegan` allows users to select the distance metric most appropriate to their data (e.g. Bray-Curtis, Jaccard, etc.). As nMDS is an iterative approach, meaning that the the computations are run until the best solution is found, it can be quite computationally demanding for larger datasets. Although the nMDS algorithm tries to minimize the ordination stress, it is a good practice to compute the **ordination stress** value to judge the reliability of the solution found (goodness-of-fit). Ordination stress indicates how much distorted are the fitted data when compared to the original observed samples. Stress values can also be visualised with the function `stressplot()` (`vegan` package), which produces a Shepard stressplot (Figure 4.7).

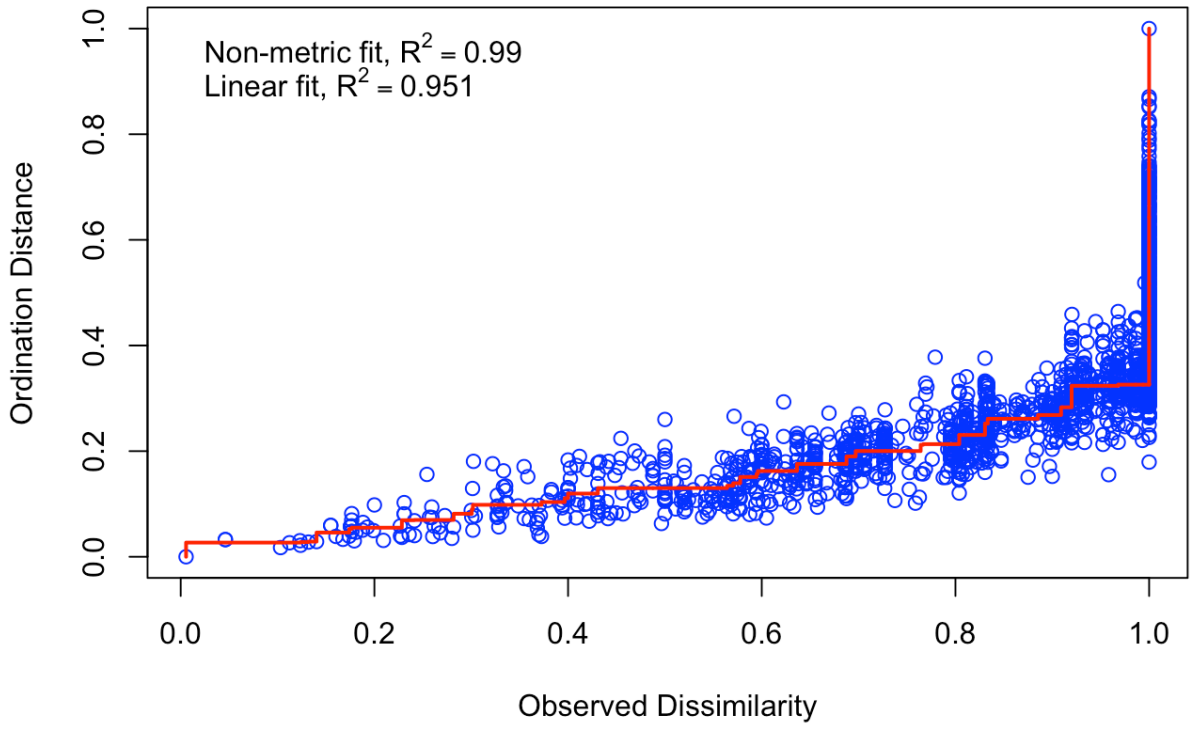


Figure 4.7: A Shepard plot, from R Studio Pubs.

The Shepard plot displays the ordination distance against the observed distance. Ideally, the

higher the points should fall on a monotonic line, where an increased observed distance is related to an increased ordination distance. Moreover, the higher the number of dimensions, the lower the stress value. If interested in choosing the appropriate number of dimensions, it is possible to use a scree plot which shows the number of dimensions against the stress level. Generally, it is possible to interpret stress values following these guidelines (Dexter et al. (2018)):

Table 4.3: Ordination stress for the interpretation of nMDS

Interpretation	Stress level
Excellent	< 0.05
Good	< 0.1
Usable (but caution is required)	> 0.2
Random	> 0.3

If the solution has produced a good stress level for the number of dimensions required, it is possible to plot the nMDS and interpret the results. Points that plot closer together are more similar, while points that are distant one to each other are more different. The nMDS plot can also be useful in recognizing groups (points grouping together and plotting further from other points). An example is provided in Figure 4.8.

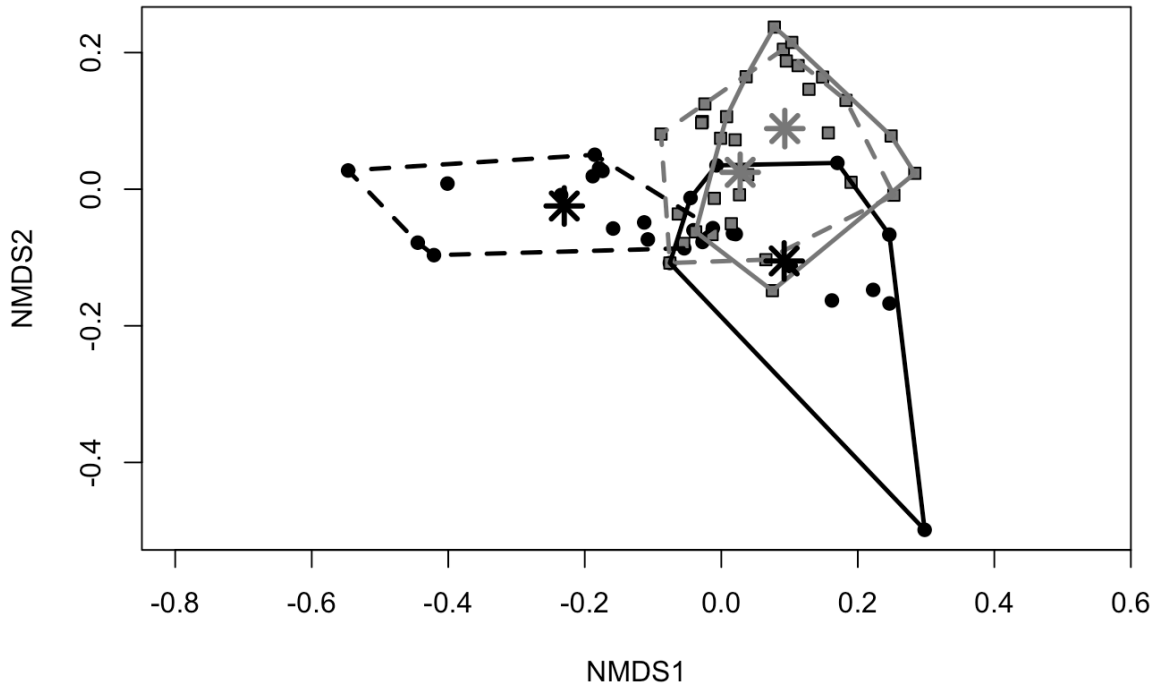


Figure 4.8: A nMDS plot with clusters, from R Studio Pubs.

4.6.2.4 NCA

The **Neighborhood Component Analysis** is a supervised machine learning algorithm for metric learning, developed by Goldberger et al. (2005) at the University of Toronto. The documentation for the `Python` implementation (package `sklearn.neighbors`) reports that:

> It learns a linear transformation in a supervised fashion to improve the classification accuracy of a stochastic nearest neighbours rule in the transformed space.

The algorithm optimizes the overall classification accuracy of multivariate data, and it works in a similar way to the k-Nearest Neighbours (k-NN). NCA is mostly used for distance metric learning, although it can be used as a linear dimensionality reduction technique. Both k-NN and NCA can be applied to non-normal distributions, as they are non-parametric methods (Cardarelli (2022)).

In this project, NCA has been applied to the same subset used for the nMDS (which included presence/absence values of cereals), with the scope of dimensionality reduction. The dataset has been reduced to one dimension to show a greater degree of separation between Northern and Southern Italy in the early Middle Ages. The entire process of data preparation and processing is described in Section 5.4.3.

4.7 Archaeobotany

> ! Section in progress
>
> Introduce the topic

4.7.1 Methodological issues

> ! Section in progress
>
> **Issues**: problems with the data and general problems in sampling
> **BIASES**: Talk about the biases and dataset problems here or do it in the Materials chapter?

The data collection process did not present major problems in taxa comparability, since the names of plant were matching. In some cases, it was only possible for the archaeobotanist to specify the name of the species (e.g. *Avena sp.*), or it was not possible to properly identify the taxa. In the latter case, I reported all the names that can possibly identify the seed (e.g. *Triticum aestivum/durum*). Quantifications on the archaeobotanical dataset were affected

by several biases. The samples in the database have been collected using different variations of visual strategies. Visual sampling occurs when archaeologists can visually see or expect macroremains in the feature they are excavating. Common features from which samples are usually collected include—pits, hearths, filled anforae, etc. But "how can one argue that the contents of a pit reflect activities involving food specific to that context, if one has not examined samples from floor deposits into which the pit was dug, or the deposits overlying it?" (Pearsall (2015) [p.75]). Lennstrom and Hastorf (1995) lament "feature biases" in many archaeological excavations, as paleobotanists are often called after archaeologists have already recovered materials from specific site features. The authors call for a general misconception in the goal of archaeobotany—to collect as many macroremains as possible. This strategy is, in fact, not very informative about the relationships between plant remains and stratigraphic units, and the general deposition patterns in the site. For instance, Jones et al. (1986) were able to reconstruct the functions of structures and the methods of crop storage at Assiros, northern Greece, through a thorough extensive sampling. Pearsall (2015, p. 74) reccommends a "blanket sampling" strategy, which consists in collecting samples for flotation from each stratigraphical unit. The advantages of blanket sampling are manifold:

- If in theory sampling from strategical features (e.g. hearths) maximizes the chances of recovering more macrobotanical remains, in practice this is not always the case. For instance, if a hearth was used on a regular basis, it could possibly have been cleaned of ashes frequently, and the excavator might have more luck sampling around it.
- Including the collection of samples from each layer in the excavation leads an uniform and standardised procedure, reducing the variation across samples.
- Blanket sampling allows an easier reconstruction of deposition patterns and of stratigraphic units relations as stated above. For instance, one can analyse differences in macroremains densities across samples.

While archaeobotanists have advocated for more specific sampling strategies for over 40 years, none of the Italian excavations in this dataset applied blanket sampling, and sampling from features is still the most common practice.

4.7.2 Quantifications

4.7.2.1 Ubiquity

Ubiquity, or presence analysis, is a popular approach in archaeobotanical quantitative analysis. The method is straightforward—the number of sites/contexts where a plant is present is divided by the total number of sites/contexts under examination. If, for instance, an olive pit is present in 3 sites out of 10, the ubiquity for the olive will be 30%. The formula for the calculation is at follows:

$$U_x = (\frac{N_p}{N_{tot}}) \cdot 100$$

where N_p is the number of presences, and N_{tot} is the total number of contexts. The result can be multiplied by 100 to obtain a score in %.

This approach has both advantages and drawbacks. Presence analysis minimizes the impact of outliers (overrepresented plant species) on calculations (Wright (2010), 51-52), but the relative importance of a plant in a particular context is lost. It is also important to keep in mind that taxa richness is influenced by factors including sample size, deposition and preservation modes, and sampling strategies (e.g. sieving methodologies) (Pearsall (2015), 161-2).

- Write about sample size problems and include a graph
- Write about deposition and preservation
- Write about sampling strategies
- Write about the problems of this dataset that led to the choice of ubiquity as parameter

Ubiquity is the best option to immediately read the Italian peninsular botanical dataset. The variability in the seeds/fruits samples is too high, with different species being outliers in different sites. A likely reason for this is probably the poor sampling quality, usually occurring after an agglomerate of seeds is found during excavation. Normally, agglomerates are found in specific storage places or processing areas (e.g. wine/olive processing quarters), skewing the distribution of the curve. Ubiquity overcomes this issue, as it provides a score based on the percentage of presences of a plant species in the samples considered.

In addition to the general calculation of the diachronic ubiquity in the entire peninsula, it is also important to look for regional differences in the archaeobotanical dataset. To do so, I created an `R` function to subset data related to Northern, Central and Southern Italian regions. For a clearer reading of the plot, I divided the plants into–`Cereals`, `Pulses` and `Fruits/Nuts`. The results are in Chapter 5.

4.7.3 Diversity

> **i** Notes:
>
> Introduce the concept of diversity. Why is it useful?

Species richness (S) is the number of species found within a community or ecosystem. The boundaries of the region are defined by the researcher. While ecologists use sampling or census to obtain the richness value, archaeobotanists can only rely on sampling, counting the presence of species in the area under investigation (Moore (2013)). **Species diversity** is a measurement

of species richness combined with **species evenness**, meaning it takes into account not only how many species are present but also how evenly distributed the numbers of each species are.

There are different ways to calculate species diversity…

Shannon-W

4.8 Zooarchaeology

> ! Section in progress

Part I

Results

5 Archaeobotany

> ! Page under construction
>
> The results presented here are preliminary and the chapter has yet to be written.

In this chapter, I will present the macrobotanical data from 170 case studies used to carry on this research (Chapter 3), along with the quantifications performed on the absolute counts. The data will be first presented temporally, and a discussion of the diachronic trends will follow at the end of the chapter.

5.1 Case studies

The following map shows the sites under investigation, divided by chronology. Please select the desired chronology (or chronologies) from the legend on the right.

5.2 Ubiquity

In Chapter 4 ubiquity has been described as the best way to present the archaeobotanical remains from the Italian peninsula, given the numerous biases in the samples. The heatmap below (Figure 5.1) provides a good overview of the temporal trends of presence of cereals, legumes, fruits and nuts in the entire area under examination.

5.2.1 Macroregional differences

The heatmap displayed in Figure 5.1 presents diachronical ubiquity values of the entire peninsula. However, it is also possible to look at the macroregional differences in plants ubiquities. The R function `Ubiquity_macroreg_chrono()` (**?@sec-Ubiquity-macroreg-chrono**) was created to subset data related to (current) Northern, Central and Southern Italian regions. Subsetting the dataset required a larger chronological division to obtain enough sites for a statistical interpretation of the results. The ubiquity values are presented using the variable `Chronology` rather than the individual centuries. For a clearer reading of the plot, the taxa

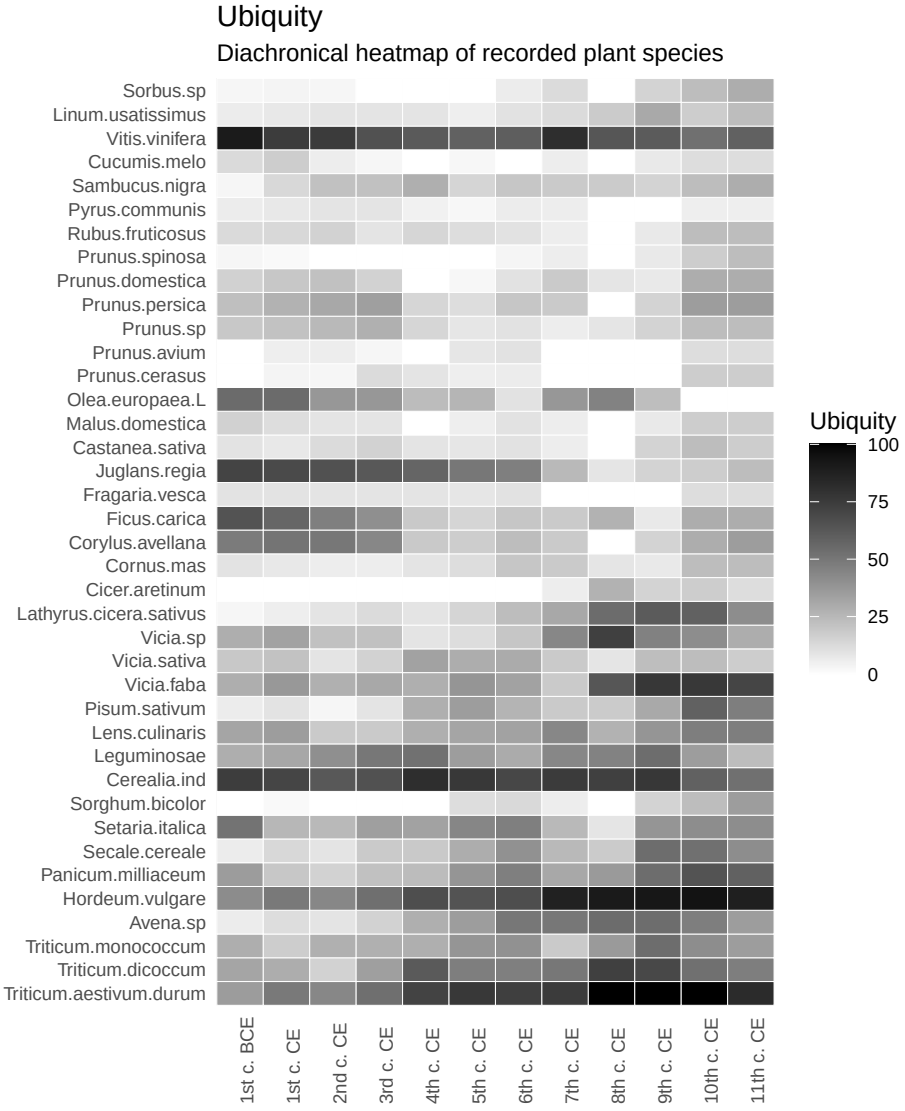


Figure 5.1: Diachronical heatmap of recorded plant species

have been divided into–Cereals, Pulses and Fruits/Nuts. Some taxa have been omitted from the plot.

5.2.1.1 Cereals

It is interesting to notice how in the **Roman age**, cereals are similarly ubiquitous in Southern and Northern Italy, although there are some exceptions (*i.e.* einkorn, rye, oats, proso millet) that can derive from the randomness of samples. Unfortunately, only three sites provided botanical samples for Roman Central Italy and the values have been omitted from the plot. These sites (from the *Roman Peasant Project*, Tuscany) only reported three kinds of cereal: common wheat, emmer, and barley. Similar ubiquity values for the two macroregions under assessment in the Roman age may suggest similar production patterns in the whole peninsula. In the **Late Roman age**, ubiquity data has been calculated for the three macroregions. Three crops are found on 62-75% of the Central Italian sites: common wheat, barley and emmer. Other cereals are present, but less ubiquitously. These three cultivations seem to be diffused in the south as well. Conversely, in Northern Italy common wheat and barley were important cultivations but competed with other cereals including millet, sorghum, and rye (now doubled in presence). The **Early Medieval age** seems to mark a shift in agricultural practices—cereals ubiquities vary more markedly in the three macroregions. In Southern Italy, common wheat and barley were still the predominant cereals. This is true for Central and Northern Italy, however in these regions other cereals are also widely present in a large number of sites. The samples from the **Medieval age** are fewer in number since the upper boundary of this project's chronology is the 11^{th} c. Despite the short chronology, it is possible to make some considerations. Medieval Centraly Italy relied heavily on common wheat, barley and emmer, with other cereals increasingly important. Barley is the most ubiquitous cereal in Northern Italy in this period, followed by common wheat, millets and sorghum.

5.2.1.2 Pulses

In the **Roman Age**, pulses are an important part of the diet and are cultivated both in Northern and Southern Italy. In the latter, vetch/broad beans are present in 22-32% of the samples, and lentils are present in 38% of the sites. In the **Late Roman Age**, broad beans are equally important in Central and Northern Italy, and peas are present in 50% of the Central Italian sites. In the **Earily Medieval Age**, pulses are present in many Central Italian sites, especially blue/red peas, broad beans and other Fabaceae. Lentils and broad beans are also cultivated in almost half of the Northern Italian sites. The importance of pulses in Central Italy is confirmed by the **11^{th} c.** samples, where every specie is present in over 66% of the sites and Fabaceae and blue/red peas are found in every sample. Conversely, in Northern Italy broad bean is found in 66% of the sites.

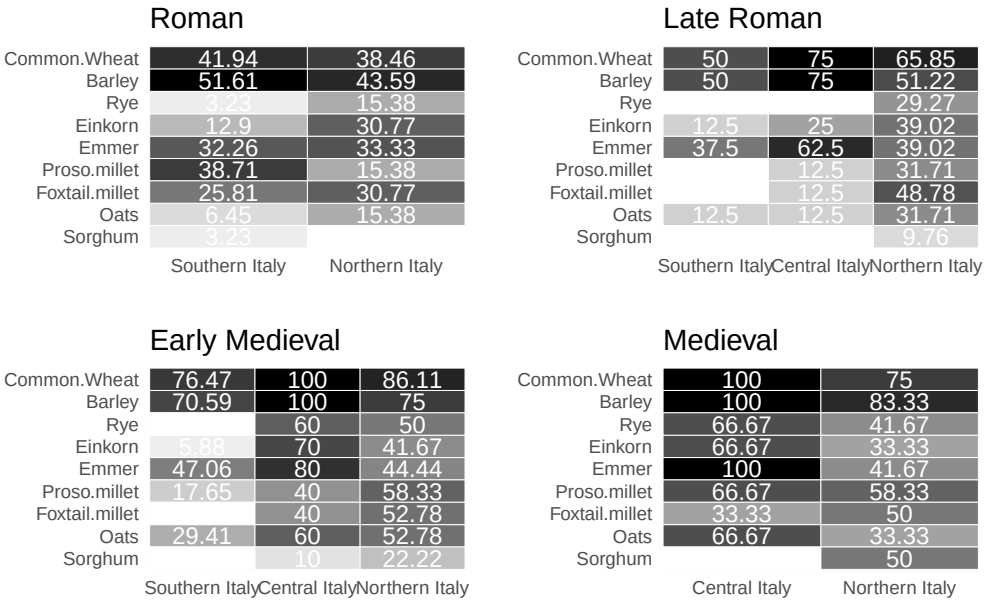


Figure 5.2: Diachronical heatmap of cereals in the Italian macroregions

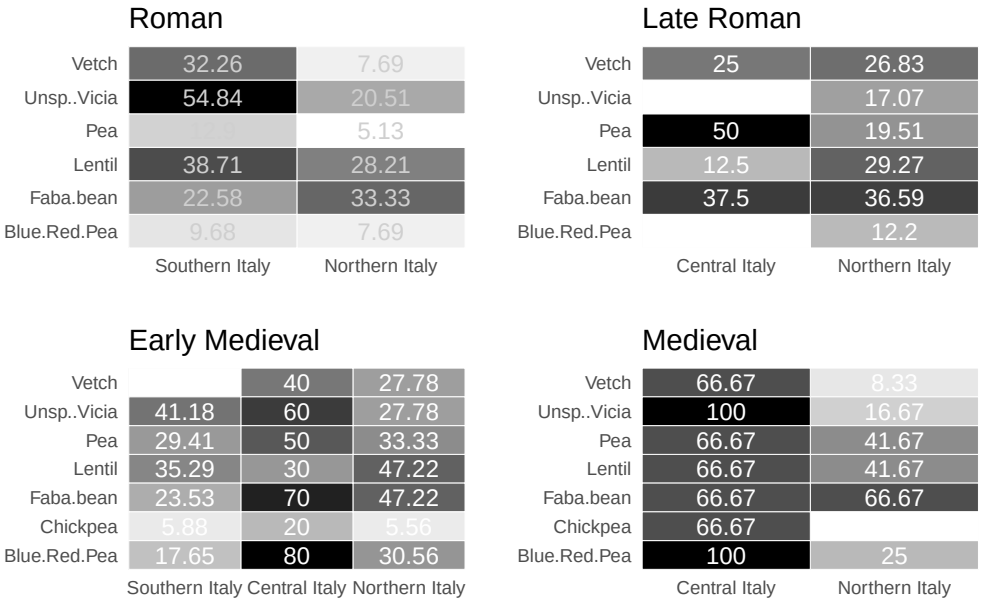


Figure 5.3: Diachronical heatmap of pulses in the Italian macroregions

5.2.1.3 Fruits and nuts

Olive and grape are two essential cultivations in the Italian peninsula. Olive pits, as can be expected, are more ubiquitous in Southern Italy, where in Roman times are present in >87% of the sites and in over 58% of the sites in the following chronologies[1]. Conversely, the grape is important in Central and Northern Italy in the Late Roman, Early Medieval and Medieval ages.

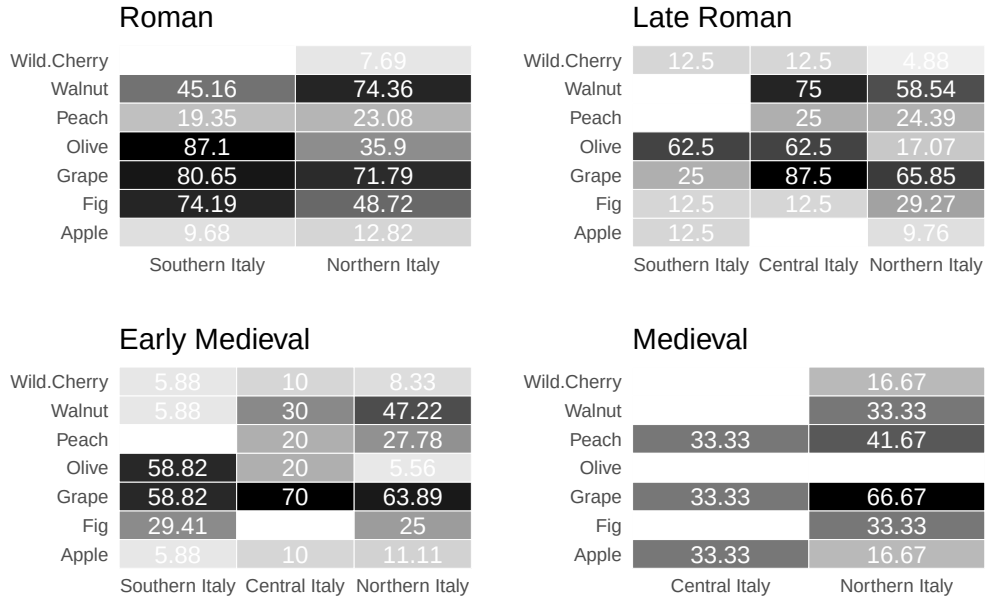


Figure 5.4: Diachronical heatmap of fruits/nuts in the Italian macroregions

5.3 Richness and diversity

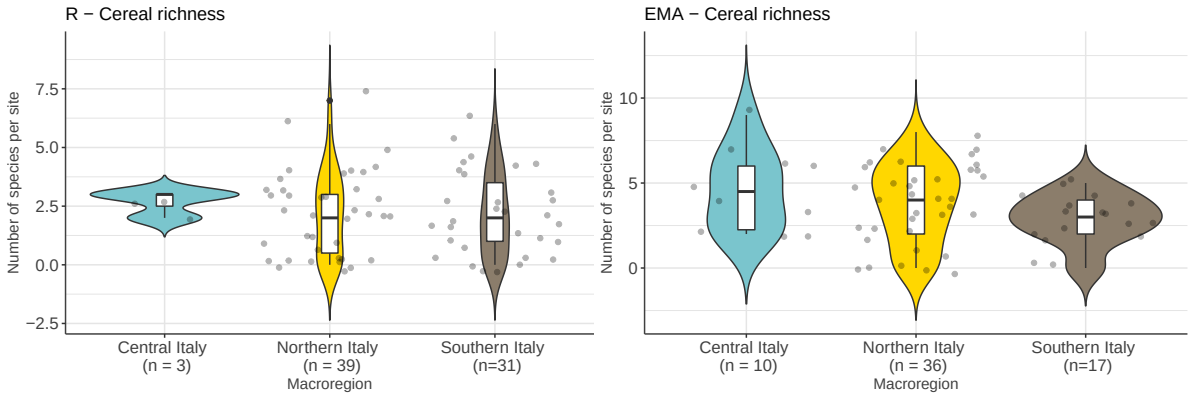
> ! Section in progress

5.3.1 Richness and diversity in the Italian macroregions

Cereals share similar presence values in Roman Northern and Southern Italian sites (Figure 5.5). Central Italy reports higher values, although this is based only on three sites and

[1]The Late Roman values for Southern Italy are only based on 5 samples (3 of which are from the same site, Salapia) so the values are not very trustworthy.

hence it is not reliable. During the Early Middle Ages, Central Italy again is the richest in cereals, closely followed by Northern Italy. Interestingly, Southern Italy still reports values very close to the Roman age. A full list of the Southern Italian EMA sites is reported in Table 5.1.



(a) Roman age.

(b) Early Medieval age.

Figure 5.5: Violin plots of cereal richness in the Italian macroregions. The grey dots (jitters) indicate the value for the single site, while the white boxplot shows the median and the quartile values.

Table 5.1: List of Southern Italian sites with chronology EMA

ID	Site	Region	Geography	Type	Culture/Influence
98	S. Maria in Cività, D85	Molise	Hilltop	Urban	Lombard
107	S. Giovanni di Ruoti, Phase 3A	Basilicata	Mountain	Monastery	Lombard
107	S. Giovanni di Ruoti, Phase 3B	Basilicata	Mountain	Monastery	Lombard
198	Salapia, area botteghe, US 2475	Puglia	Coast/Lagoon	Urban	Lombard
198	Salapia, area botteghe, US 2437	Puglia	Coast/Lagoon	Urban	Lombard
199	Salapia, area conceria, US 2054	Puglia	Coast/Lagoon	Urban	Lombard
199	Salapia, area conceria, US 2211-2217	Puglia	Coast/Lagoon	Urban	Lombard
199	Salapia, area conceria, 8th-9th c.	Puglia	Coast/Lagoon	Urban	Lombard
196	Faragola, wastepit 61	Puglia	Plain	Rural, villa	Lombard

ID	Site	Region	Geography	Type	Culture/Influence
196	Faragola, wastepit 66	Puglia	Plain	Rural, villa	Lombard
234	Colle Castellano, Phase 3-4	Molise	Hill	Urban	Lombard
177	San Vincenzo al Volturno, kitchen area	Molise	Hill	Monastery	Lombard
101	Supersano, loc. Scorpo	Puglia	Plain	Rural	Byzantine
250	Apigliano, 9th-10th c., pits	Puglia	Plain	Rural	Byzantine
250	Apigliano, 10th-11th c., pits	Puglia	Plain	Rural	Byzantine
196	Faragola, granary A7	Puglia	Plain	Rural, villa	Lombard
196	Faragola, granary A8	Puglia	Plain	Rural, villa	Lombard

5.4 Cereals regionality

5.4.1 PERMANOVA

> **i** Notes on terminology: PERMANOVA
>
> Permutational multivariate analysis of variance (PERMANOVA) is a non-parametric multivariate statistical test used to compare group of objects. By using measure space, the null hypothesis that the centroids and dispersion of groups are identical is tested. The null hypothesis is rejected if either the centroid or the spread of the objects differs between the groups. A prior calculation of the distance between any two objects included in the experiment is used to determine whether the test is valid or not[a] (Anderson (2017)). In this context, the *null hypothesis* is that there is no regional difference in the cereals dataset, with cereals being evenly distributed across macroregions and chronologies.
>
> [a]Source: Wikipedia. Change the source later

The suggestion of an Early Medieval shift in cereal farming stated in Section 5.2.1 and Section 5.3.1 needs statistical support. Considering that data is not unimodal and that we are dealing with presence/absence analysis, the best choice is to use a non-parametric test as PERMANOVA on the early medieval botanical dataset. Prior to performing the test, it was necessary to pre-process data by:

- Selecting all the cereals columns of the plant remains table, keeping some categorical variables: `Macroregion`, `Chronology`, `Geography` and `Type`.
- Removing the empty rows (caused by the fact that some sites have seeds/fruits, but not cereals).
- Transforming the raw counts into presence/absence, using the function `decostand()` (`method=pa`) in the `R` package `vegan` (Oksanen et al. (2020)).

After the pre-processing, it was possible to run the PERMANOVA using the function `adonis2()` in the package `vegan`. The function creates a distance matrix and computes an analysis of variance on the matrix. The method chosen to calculate the distance matrix is the `jaccard` distance. The Jaccard distance (Kosub (2019)), based on the Jaccard similarity index, is a value of dissimilarity between sample sets. When compared to other dissimilarity indices, it is more appropriate for presence/absence analyses as it is not based on Euclidean distance.

Results of `adonis2()`.

```
Permutation test for adonis under reduced model
Terms added sequentially (first to last)
Permutation: free
Number of permutations: 10000

adonis2(formula = cer_macroreg_ubiquity_transp.dist ~ Macroregion, data = cer_macroreg_ubiqu
            Df SumOfSqs      R2      F    Pr(>F)
Macroregion  1   1.3024 0.13495 7.3322 9.999e-05 ***
Residual    47   8.3487 0.86505
Total       48   9.6512 1.00000
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

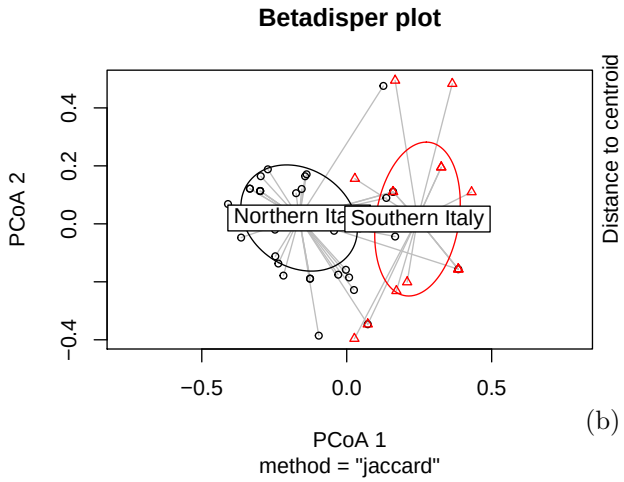
The results of the PERMANOVA indicate that the variable `Macroregion` is highly significant, meaning that we can be 99.9% confident that it is a discriminant in the early medieval dataset.

After running the PERMANOVA, it is necessary to check the homogeneity of variances, to confirm the results (especially when dealing with small groups of data). The function `betadisper()` from the package `vegan` provides the distances of group samples from centroids. If the variation is even, the null hypothesis of no difference in dispersion between groups is accepted. To test the variation, it is possible to use the analysis of variance (ANOVA).

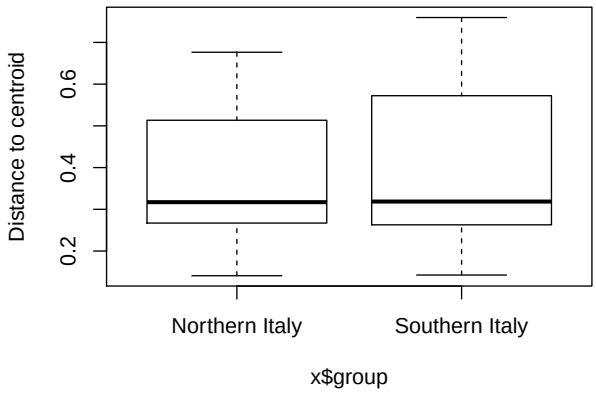
Results of `anova()` on the betadisper.

```r
# We will see that the ANOVA's p-value is not significant meaning that group dispersions a
#("Null hypothesis of no difference in dispersion between groups"; https://www.rdocumentat

anova(cer_macroreg_ubiquity_transp.betadisper) # This should not be significant!
```

```
Analysis of Variance Table

Response: Distances
          Df  Sum Sq   Mean Sq F value Pr(>F)
Groups     1 0.00211 0.0021112  0.0715 0.7903
Residuals 47 1.38696 0.0295098
```



(a) Groups dispersions plot with confidence ellipses.



(b) Boxplot showing equal distances from centroid.

Figure 5.6: Results of the betadisper() (groups dispersions) on the distance matrix calculated with the Jaccard method.

The `betadisper()` graphs (Figure 5.6) show similar distances from the centroids for the categories `Northern Italy` and `Southern Italy`. In addition, the ANOVA on the `betadisper()` shows that the separation is not significant (p-value over the significance treshold), meaning that the groups dispersions are homogeneous. We can now be more confident of the PERMANOVA results and accept the difference between the two groups of sites under investigation. In other words, the Southern and Northern Italian group of sites are different during the Early Middle Ages.

Running the same tests on the Roman sites failed to separate the two groups of sites, confirming that there was not a major difference in the types of cereals cultivated during the Roman age between Northern and Southern Italy.

5.4.2 nMDS

5.4.3 NCA

> **i** Notes on terminology: Wasserstein metric
>
> The Wasserstein distance (or earth's mover distance) is a measure of distance between two probability distributions on a metric space.

In addition to statistically testing the separation between the Northern and Southern Italian early medieval cereals dataset (Section 5.4.1), it is possible to measure the distance between groups of sites both in the Roman and early Middle ages. For this task, a machine learning algorithm for metric learning has been chosen: the *Neighborhood Component Analysis*, from the `Python` package `NeighborhoodComponentAnalysis` (in `sklearn.neighbors`). A more in-depth explanation of the algorithm can be read in Section 4.6.2.4. To work with balanced group of samples, the group sizes have been arbitrarily set to 20 random samples, allowing replacement (meaning that a sample can randomly be picked twice). The `Python` function `sample()` (from the `random` library) was used to select random samples. To avoid fallacy in computations, the macroregion `Central Italy` and the chronologies `LR` (Late Roman) and `Ma` (11^{th} c. onwards) have been excluded from this test—the uneven distribution of the group of samples required a cautious approach. The NCA has been run with a reduction to only one dimension, using KDE plots to visualize the results. Setting the dimension to one allows easier calculations of distance. In Figure 5.7 (a), it is possible to see the NCA performed on the Roman cereals presence/absence dataset. As already pointed out, the PERMANOVA did not produce significant results for this dataset and the Wasserstein distance (calculated with the `wasserstein_distance()` function in the `scipy` library) is indeed shorter for the Roman dataset. For both chronologies there is an overlap in the curves, which is more considerable in the Roman age (indicating that the group of samples are more similar). The overlap for the EMA groups (Figure 5.7, b) is probably due to the fact that the presence of the noble grains is not by itself a 'marker' of Southern Italian sites—these grains are also very common in the North. The difference is that in the South noble grains are not cultivated in conjunction with other grains. The graph for the EMA chronology shows a clearer separation of the macroregional groups, with some minor overlaps. Moreover, the graph also displays variability in the Northern Italian dataset. The variability can also be assessed from the outliers in the boxplots in Figure 5.8.

Plots

```
NCA_KDE_1D, ax = plt.subplots(1, 2, figsize=(10, 5), sharey=True, sharex=True)

sns.kdeplot(data=df_R_merge, x="value", ax=ax[0], hue="Macroregion", fill=True, alpha=.1,
```

```
sns.kdeplot(data=df_EMA_merge, x="value", ax=ax[1], hue="Macroregion", fill=True, alpha=.1

NCA_KDE_1D.text(0.08, 0.03, 'Weighted Wasserstein Distance: W = 67.64 \nPERMANOVA test: p>
NCA_KDE_1D.text(0.68, 0.03, 'Weighted Wasserstein Distance: W = 200.65\nPERMANOVA test: p<
plt.tight_layout()
plt.subplots_adjust(bottom=0.19)
plt.show()
```

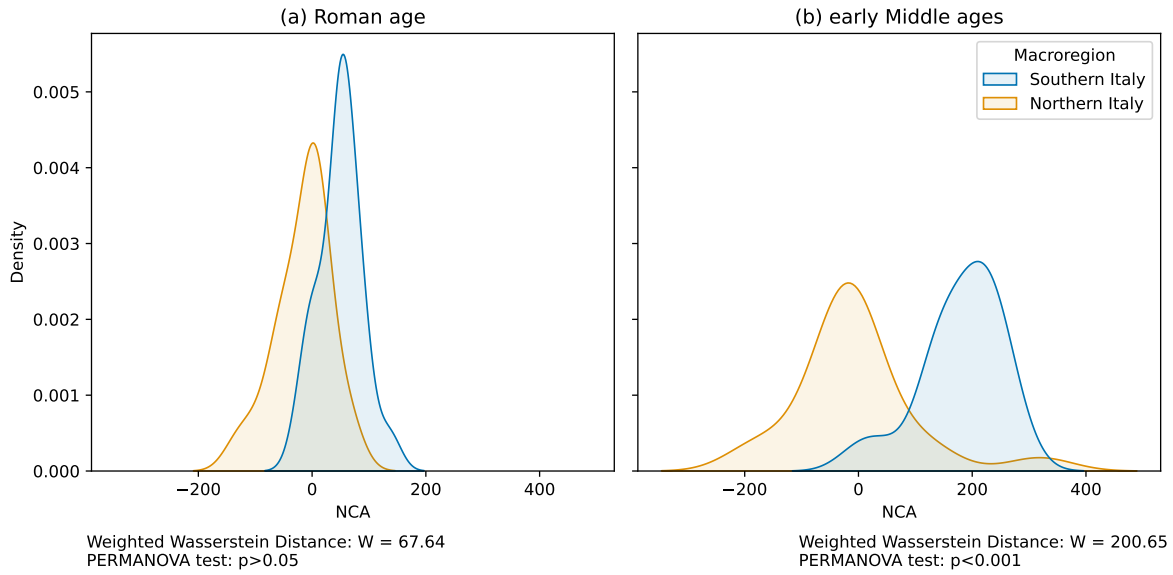


Figure 5.7: One-Dimension NCA on the Presence/Absence Cereals Dataset

5.4.4 Network Analysis of cereals in EMA sites

> ! Section in progress

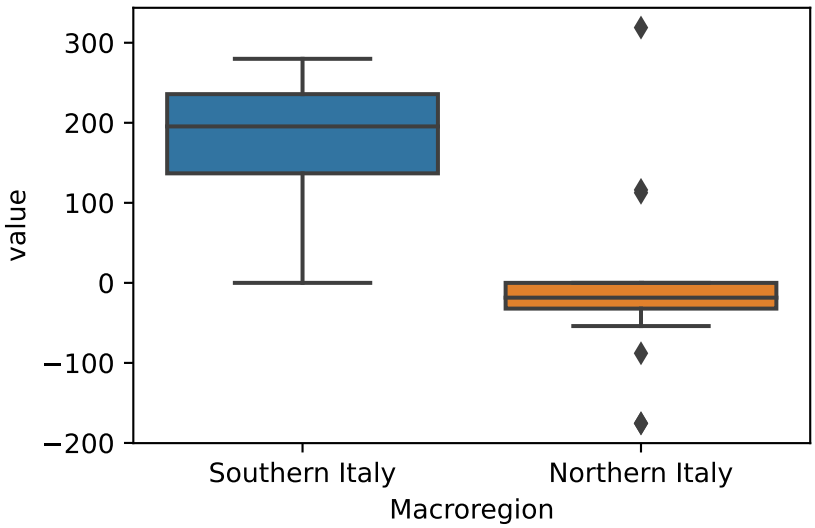


Figure 5.8: Boxplots showing the NCA value for Northern and Southern Early Medieval Italy.

6 Zooarchaeology

> ! Page under construction

6.1 Case studies

The following map shows the sites under investigation, divided by chronology. Please select the desired chronology (or chronologies) from the legend on the right.

Show map

7 Discussion

> ! Page under construction

8 Conclusions

! Page to be added

References

Anderson, M.J., 2017. Permutational Multivariate Analysis of Variance (PERMANOVA). In: Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd, pp. 1–15.

Baxter, M.J., 2015. Exploratory multivariate analysis in archaeology, Foundations of archaeology. Percheron Press, a division of Eliot Werner Publications, Inc, Clinton Corners, New York.

Binford, S.R., Binford, L., 1968. New Perspectives in Archaeology. Aldline Press, Chicago.

Boscato, P., Fronza, V., Salvadori, F., 2007. Proposta di un database per i reperti faunistici. In: Fiore, I., Malerba, G., Chilardi, S. (Eds.), Atti Del 3° Convegno Nazionale Di Archeozoologia. Siracusa 3-5 Novembre 2000. Istituto Poligrafico e Zecca dello Stato, Roma, pp. 1–14.

Cardarelli, L., 2022. A deep variational convolutional Autoencoder for unsupervised features extraction of ceramic profiles. A case study from central Italy. Journal of Archaeological Science 144, 105640.

Carlson, D.L., 2017. Quantitative methods in archaeology using R, Cambridge University Press. ed, Cambridge Manuals in Archaeology. Cambridge.

Cuffney, T.F., Kennen, J.G., Waite, I.R., 2014. Aquatic Ecosystems as Indicators of Status and Trends in Water Quality. In: Comprehensive Water Quality and Purification. Elsevier, pp. 122–156.

Dexter, E., Rollwagen-Bollens, G., Bollens, S.M., 2018. The trouble with stress: A flexible method for the evaluation of nonmetric multidimensional scaling: *The Trouble with Stress*. Limnology and Oceanography: Methods 16, 434–443.

Figuera, M., 2018. Database management e dati archeologici: standardizzazione e applicazione della logica fuzzy alla gestione delle fonti e delle attribuzioni tipologiche. Archeologia e Calcolatori 29, 143–160.

Fletcher, M., Lock, G.R., 2005. Digging numbers: Elementary statistics for archaeologists, 2nd ed. ed, Monograph (Oxford University School of Archaeology). Oxford University Committee for Archaeology, Oxford : Oakville, CT.

Gattiglia, G., 2018. Databases in Archaeology. In: López Varela, S.L. (Ed.), The Encyclopedia of Archaeological Sciences. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp. 1–4.

Gauch, H.G., 1982. Multivariate Analysis in Community Ecology, First. ed. Cambridge University Press.

Goldberger, J., Hinton, G.E., Roweis, S., Salakhutdinov, R.R., 2005. Neighbourhood components analysis. In: Saul, L.K., Weiss, Y., Bottou, L. (Eds.), Advances in Neural Information Processing Systems. Bradford Books, Vancouver, pp. 513–520.

Grebner, D.L., Bettinger, P., Siry, J.P., 2013. Forest Dynamics. In: Introduction to Forestry

and Natural Resources. Elsevier, pp. 243–254.
Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362.
Hunter, J.D., 2007. Matplotlib: A 2D graphics environment. Computing in science & engineering 9, 90–95.
Jones, G., Wardle, K., Halstead, P., Wardle, D., 1986. Crop Storage at Assiros. Scientific American 254, 96–103.
Kosub, S., 2019. A note on the triangle inequality for the Jaccard distance. Pattern Recognition Letters 120, 36–38.
Lauro, C., 1996. Computational statistics or statistical computing, is that the question? Computational Statistics & Data Analysis, Classification 23, 191–193.
Lennstrom, H.A., Hastorf, C.A., 1995. Interpretation in context: Sampling and analysis in paleoethnobotany. American Antiquity 60, 701–721.
Mercuri, A.M., Allevato, E., Arobba, D., Bandini Mazzanti, M., Bosi, G., Caramiello, R., Castiglioni, E., Carra, M.L., Celant, A., Costantini, L., Di Pasquale, G., Fiorentino, G., Florenzano, A., Guido, M., Marchesini, M., Mariotti Lippi, M., Marvelli, S., Miola, A., Montanari, C., Nisbet, R., Peña-Chocarro, L., Perego, R., Ravazzi, C., Rottoli, M., Sadori, L., Ucchesu, M., Rinaldi, R., 2015. Pollen and macroremains from Holocene archaeological sites: A dataset for the understanding of the bio-cultural diversity of the Italian landscape. Review of Palaeobotany and Palynology 218, 250–266.
Moore, J.C., 2013. Diversity, taxonomic versus functional. In: Encyclopedia of Biodiversity. Elsevier, pp. 648–656.
Niccolucci, F., 2020. ARIADNEplus: L'avventura continua. DigItalia 2, 88–95.
Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2020. Vegan: Community ecology package.
Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., O'Hara, R.B., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., Caceres, M.D., Durand, S., Evangelista, H.B.A., FitzJohn, R., Friendly, M., Furneaux, B., Hannigan, G., Hill, M.O., Lahti, L., McGlinn, D., Ouellette, M.-H., Cunha, E.R., Smith, T., Stier, A., Braak, C.J.F.T., Weedon, J., 2022. Vegan: Community ecology package.
Pearsall, D.M., 2015. Paleoethnobotany: A handbook of procedures, Third edition. ed. Left Coast Press Inc, Walnut Creek, California.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.
Reback, J., jbrockmendel, McKinney, W., Bossche, J.V. den, Roeschke, M., Augspurger, T., Hawkins, S., Cloud, P., gfyoung, Sinhrks, Hoefler, P., Klein, A., Petersen, T., Tratner, J.,

She, C., Ayd, W., Naveh, S., Darbyshire, J.H.M., Shadrach, R., Garcia, M., Schendel, J., Hayden, A., Saxton, D., Gorelli, M.E., Li, F., Wörtwein, T., Zeitlin, M., Jancauskas, V., McMaster, A., Li, T., 2022. Pandas-dev/pandas: Pandas 1.4.3.

Richards, J.D., 2021. Archiving Archaeological Data in the United Kingdom. Internet Archaeology 58.

Rizzo, M.L., 2019. Statistical Computing with R, Second. ed. Chapman and Hall/CRC, New York.

Shennan, S., 1997. Quantifying archaeology, 2nd ed. ed. Edinburgh University Press, Edinburgh.

Syms, C., 2008. Ordination. In: Encyclopedia of Ecology. Elsevier, pp. 2572–2581.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., 2020. SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature methods 17, 261–272.

Waskom, M.L., 2021. Seaborn: Statistical data visualization. Journal of Open Source Software 6, 3021.

Wickham, H., 2016. Programming with Ggplot2. In: Wickham, H. (Ed.), Ggplot2: Elegant Graphics for Data Analysis, Use R! Springer International Publishing, Cham, pp. 241–253.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the {tidyverse} 4, 1686.

Wright, P.J., 2010. Methodological Issues in Paleoethnobotany: A consideration of Issues, Methods, and Cases. In: VanDerwarker, A.M., Peres, T.M. (Eds.), Integrating Zooarchaeology and Paleoethnobotany: A Consideration of Issues, Methods, and Cases. Springer New York, New York, NY, pp. 37–64.

Xie, Y., 2021. Knitr: A general-purpose package for dynamic report generation in r.

Custom functions

This section contains the list of custom functions that have been written to prepare, handle and visualize the data exported from the database.

Archaeobotany

archaeobotany_tables()

This function has two arguments:

- a dataframe of the exported table of plants from the database (`view_archaeobot.csv`).

```
 [1] "site_code"              "site_name"
 [3] "type_name"              "region_name"
 [5] "data_valid_start"       "data_valid_end"
 [7] "weight"                 "sampling_notes"
 [9] "extra_notes"            "short_ref"
[11] "culture_type"           "x"
[13] "y"                      "Triticum.aestivum.durum"
[15] "Triticum.dicoccum"      "Triticum.monococcum"
[17] "Avena.sp"               "Hordeum.vulgare"
[19] "Panicum.milliaceum"     "Secale.cereale"
[21] "Setaria.italica"        "Sorghum.bicolor"
[23] "Cerealia.ind"           "Leguminosae"
[25] "Lens.culinaris"         "Pisum.sativum"
[27] "Vicia.faba"             "Vicia.sativa"
[29] "Vicia.sp"               "Lathyrus.cicera.sativus"
[31] "Cicer.aretinum"         "Cornus.mas"
[33] "Corylus.avellana"       "Ficus.carica"
[35] "Fragaria.vesca"         "Juglans.regia"
[37] "Castanea.sativa"        "Malus.domestica"
[39] "Olea.europaea.L"        "Prunus.cerasus"
[41] "Prunus.avium"           "Prunus.sp"
[43] "Prunus.persica"         "Prunus.domestica"
[45] "Prunus.spinosa"         "Rubus.fruticosus"
```

```
 [47] "Pyrus.communis"          "Sambucus.nigra"
 [49] "Cucumis.melo"            "Vitis.vinifera"
 [51] "Linum.usatissimus"       "Sorbus.sp"
```

- the century of interest.

The function `archaeobotany_tables()` can be used to return the ubiquity, relative proportions or a print of the table with the sites from the chosen century. The comments in the code below explain the process.

```
##FUNCTION FOR GENERATING CENTURY BASED
# - UBIQUITY
# - RELATIVE PROPORTIONS
# - A PRINT OF THE TABLE

archaeobotany_tables <- function(x, century) {
  # Load the tidyverse library if it hasn't been loaded in the page before
  library(tidyverse)

  # Remove NAs
  x[is.na(x)] <-0

  # Filter the table for the chosen century
  # package: tidyverse
  x <- filter(x, data_valid_start <= century & data_valid_end >= century)

  # The total of each row is needed to calculate the relative proportions
  # Note: Calculation starts from column 14 because it is the first column with numerical
  Total <- rowSums(x[,14:ncol(x)])

  # Subsetting the given dataframe by creating a new dataframe with fewer columns
  plants <- data.frame(x$site_name, x$type_name,
                       x$data_valid_start, x$data_valid_end,
                       x$culture_type, x[14:ncol(x)],
                       Total
  )

  # Calculating the relative proportions and rounding the results to 2 digits.
  Rel_Prop <- round(((x[14:ncol(x)]/Total)*100), digits=2)

  # Ubiquity:
  #Note: It is given by the no. of sites where the plant is present divided by the total o
  # Note: Total of sites: (No. of rows - header row)
```

```
  # Creating a new dataframe from the Relative Proportions one (Rel_Prop).
  # Note: This can be done also from the original dataframe, it is not important since it
  Pres_Abs <- Rel_Prop

  # If the value is > 0 it means that the plant is present: this line replaces this value
  Pres_Abs[Pres_Abs > 0] <- 1

  # In how many sites is this plant present?
  Tot_sites_present <- colSums(Pres_Abs)

  # Finally calculate ubiquity
  # Note: The score is multiplied by 100 to obtain results in %
  Ubiquity <- (Tot_sites_present / nrow(Pres_Abs))*100

  return(list(
    Ubiquity_exp = Ubiquity,
    Rel_Prop_exp = Rel_Prop,
    Raw_Counts = plants
    ))
}
```

Rel_Prop_per_Century()

This function has two arguments: - a dataframe of the exported table of plants from the database (`view_archaeobot.csv`).

- the century of interest. The function `Rel_Prop_per_Century()` can be used to return the relative proportions of each site from the chosen century. The comments in the code below explain the process.

```
## Convert each site raw data into relative proportions

Rel_Prop_per_Century <- function(x, century) {

  # Remove NAs
  x[is.na(x)] <-0

  # Filter the table for the chosen century
  # package: tidyverse
  library(tidyverse)
```

```
  x <- filter(x, data_valid_start <= century & data_valid_end >= century)

  # Calculate the total of the row and divide each value by the total to get proportions
  # round() is used to get two decimal values
  Total_per_site <- rowSums(x[,14:ncol(x)])
  Rel_Prop_per_site <- round(((x[14:ncol(x)]/Total_per_site)*100), digits=2)

  # Create new dataframe with the information we need
  plants_rel_prop <- data.frame(
    "Site" = x$site_name,
    "Type" = x$type_name,
    "From.Century" = x$data_valid_start,
    "To.Century"= x$data_valid_end,
    "Weight"=x$weight,
    "Culture"=x$culture_type,
    "x"=x$x,
    "y"=x$y,
    Rel_Prop_per_site
  )

  return(plants_rel_prop)
}
```

Ubiquity_macroreg_chrono()

This function has three arguments:

- a dataframe of the exported condensed table of plants from the database (`Archaeobot_Condensed.csv`). It is a table of plants exported with their common English name and with a column of totals for each type of plant (Cereals, Fruit/Nuts, …).
- the macroregion of interest: `Southern Italy`, `Central Italy`, `Northern Italy`.
- the chronology of interest: `R`, `LR`, `EMA`, `Ma`.

```
Ubiquity_macroreg_chrono <- function(df, macroregion, chronology) {

  # Load the tidyverse library if it hasn't been loaded in the page before
  library(tidyverse)

  # Remove NAs
```

```
  df[is.na(df)] <- 0

  # Filter the table for the chosen chronology and macroregion
  # package: tidyverse
  df.chronology <- filter(df, Chronology == chronology & Macroregion == macroregion)

  #Remove useless columns: Tots, unsp.cols
  df.chronology <- df.chronology[-c(23,24,32,33,56)]

  # Create a counts dataframe where the taxa that are present will be stored as 1
  df.counts <- df.chronology[14:ncol(df.chronology)]
  df.counts[df.counts>0] <- 1

  # Create a dataframe with a sum of presences
  df.sites.present <- colSums(df.counts)

  # Calculate ubiquity and round the value to 2 decimals
  Ubiquity <- (df.sites.present / nrow(df.chronology))*100
  Ubiquity <- round(Ubiquity, 2)

  # Add a category that explains what type of plant is it (useful for visualisation)
  Plants_Type <- data.frame(Type=1:38)
  Plants_Type$Type[1:9] <- "Cereals"
  Plants_Type$Type[10:16] <- "Pulses"
  Plants_Type$Type[17:38] <- "Fruits/Nuts"

  # Final dataframe that the function will return
  Ubiquity <- cbind.data.frame("Chronology" = chronology,
                               "Macroregion" = macroregion,
                               "Plant"=names(Ubiquity),
                               "Plant.Type"=Plants_Type$Type,
                               "Ubiquity"= Ubiquity)
  Ubiquity <- data.frame(Ubiquity, row.names = NULL)


  return(Ubiquity)
}
```

Zooarchaeology