Report HOPE methodology & results

Table of Contents

The objective of this report is to give an overview of data inputs, data processing steps, data analysis, and results to determine whether human activity has changed the fundamental ecological processes over the last 8500 years (Hypothesis 1), and to further assess if interrelationship between fundamental ecosystem properties have significantly changed during this time (Hypothesis 2).

# Part I: Methodology

## Main data aquisition

### FOSSILPOL

The compilation of a pollen dataset for our analysis is performed a priori. Raw pollen datasets are carefully selected with the *R-Fossilpol* package, and the guidelines to the workflow are well described in Flantua et al. 2023 and in our website [Fossilpol project](https://hope-uib-bio.github.io/FOSSILPOL-website/about.html). Most datasets are obtained from the [Neotoma Paleoecology Database](https://www.neotomadb.org). Some additional data are from private owners in areas with data gaps, which we have limited access to use. We do not have the intellectual property rights to make these data public available. Therefore, only the derivatives of the analysis can be publicly shared. Table 1 provides a summary of the settings used in the FOSSILPOL workflow to obtain a standardised project dataset. This input data are further filter during the data processing steps in HOPE to get the final collection of a standardized dataset of high data quality we can use further in our data analyses.

Table 1: Selection of settings applied in FOSSILPOL

| **Settings** | **Setting type** | **Selection** |
| --- | --- | --- |
| General |  |  |
|  | geography\_criteria | -180 |
|  | long\_max | 180 |
|  | lat\_min | -90 |
|  | lat\_max | 90 |
|  | alt\_min |  |
|  | alt\_max |  |
|  | private\_data | TRUE |
| Neotoma |  |  |
|  | dataset\_type | pollen |
|  | sel\_var\_element | pollen |
|  | chron\_order.type1 | Varve years BP |
|  | chron\_order.type2 | Calibrated radiocarbon years BP |
|  | chron\_order.type3 | Calendar years BP |
|  | chron\_order.type4 | Radiocarbon years BP |
|  | chron\_order.type5 | Calendar years AD/BC |
|  | chron\_order.type6 |  |
| Age-depth  models |  |  |
|  | min\_n\_of\_control\_points | 3 |
|  | default\_thickness | TRUE |
|  | default\_error | 100 |
|  | max\_age\_error | 3000 |
|  | guess\_depth | 10 |
|  | default\_iteration | 10000 |
|  | default\_burn | 2000 |
|  | default\_thin | 8 |
|  | iteration\_multiplier | 5 |
| Site filtering |  |  |
|  | pollensum.filter\_by\_pollen\_sum | TRUE |
|  | pollensum.min\_n\_grains | 25 |
|  | pollensum.target\_n\_grains | 150 |
|  | pollensum.percentage\_samples | 50 |
|  | filter\_by\_age\_limit | TRUE |
|  | extrapolation.filter\_by\_extrapolation | TRUE |
|  | extrapolation.maximum\_age\_extrapolation | 3000 |
|  | extrapolation.filter\_by\_interest\_region | TRUE |
|  | extrapolation.n\_levels.filter\_by\_number\_of\_levels | TRUE |
|  | extrapolation.n\_levels.min\_n\_levels | 5 |
|  | extrapolation.use\_age\_quantiles | TRUE |
|  | extrapolation.use\_bookend\_level | TRUE |
| Data download: 23-05-22 | | |

### Harmonisation tables

An important step in FOSSILPOL to obtain a standardised pollen data set within and across regions is the harmonisation of pollen types. Different analysts have different backgrounds and schools using different nomenclature, and the level of pollen taxonomic identifications and names can vary widely. To be able to make numerical comparisons of different pollen records, the level of pollen taxonomy should be similar. Consequently, pollen harmonisation tables have been produced for different regions to try to minimise biases related to this. The regional harmonisation tables created in our project are for Europe, Levant, Siberia, Southern Asia, Northern America, Latin America, and the Indo-Pacific region (Birks et al. harmonisation paper). These tables are used as input in the Fossilpol workflow above ([see Fossilpol step\_by\_step guide](https://hope-uib-bio.github.io/FOSSILPOL-website/step_by_step_guide.html)).

## Workflow for HOPE Hypothesis

We use the [targets](https://books.ropensci.org/targets/) R-package to generate a reproducible workflow for all data processing steps, analysis of hypothesis 1 and 2, and visualisation of results in the project. We have created a Github repository of the data analysis R-project called [HOPE\_hypothesis1 in Github](https://github.com/HOPE-UIB-BIO/HOPE_Hypothesis1), which contains all the data, metadata, and R functions required to run this R-project.

We configured the targets folders with data to be saved in a shared folder on Google drive, while all the scripts and functions for running the targets are stored in our Github repository. The targets are divided in small steps with specialized functions to load necessary data, estimate the different variables, or to structure the different subsets of the data. This will avoid the need to rerun major parts between us that take too long and has already been processed. If any changes are made in the functions, targets will automatically detect it and rerun the parts that depends on this change, and skip the up-to-date analysis.

The file structure in Github is set up for the two main hypotheses in HOPE:

## levelName  
## 1 HOPE\_Hypothesis1   
## 2 ¦--HOPE\_Hypothesis1.Rproj   
## 3 ¦--\_targets.yaml   
## 4 ¦--README.md   
## 5 ¦--R   
## 6 ¦ ¦--branching\_pipeline   
## 7 ¦ ¦--general\_patterns\_of\_predictors  
## 8 ¦ ¦--hypothesis\_1   
## 9 ¦ ¦--hypothesis\_2   
## 10 ¦ ¦--report\_methodology   
## 11 ¦ ¦--report\_validation   
## 12 ¦ ¦--reports   
## 13 ¦ ¦--working\_scripts   
## 14 ¦ ¦--\_\_\_Init\_project\_\_\_.R   
## 15 ¦ ¦--00\_Config\_file.R   
## 16 ¦ °--functions   
## 17 ¦ ¦--climate   
## 18 ¦ ¦--data\_wrangling   
## 19 ¦ ¦--events   
## 20 ¦ ¦--hvarpart   
## 21 ¦ ¦--modelling   
## 22 ¦ ¦--PAPs   
## 23 ¦ ¦--procrustes   
## 24 ¦ ¦--spd   
## 25 ¦ ¦--validation   
## 26 ¦ °--visualisation   
## 27 °--renv

When the R-project is set up, you need to run the scripts \_\_\_init\_project.R\_\_\_ and 00\_Config\_file.R. This will install and load the R packages for used in this project, and provide the paths needed.

Two main targets pipelines are created to run hypothesis 1 and hypothesis 2, and output and meta data are saved in two separate targets folders. The pipelines are arranged in a order to prepare the input data needed for the main analysis in the end. The pipeline in hypothesis 1 contain the data processing steps. The final data input in hypothesis 1 is then used in hypothesis 2 pipeline.

1. The first steps of the workflow imports and load all data, vectors, and settings needed for the following pre-processing data steps. The first filtering of datasets depend for example on number of sample numbers, number of pollen grains, chronology quality, and duration of sequences.
2. This is followed by data processing for each of the explanatory and response variables for the hypotheses 1 and 2 to be explored:
   1. Preparation of the variable needed to detect past human presence and impact. This is a major analyses in itself.
   2. Data extraction of palaeo-climate from the CHELSA paleoclimate database. This is modeled palaeo-climate data for each of the geographical location of the pollen records. First time the function is run, it will download the data from a URL connection, and extract data for the climatic variables selected, and deleted the data that is not needed to save storage in local computer.
   3. Different targets that process all the estimates of pollen assemblage properties.
3. In the end, we have a standardized and filter dataset to explore the two hypothesis in HOPE:

Hypothesis 1 exists of two analysis called - a) the spatial (within core) analysis, and - b) the temporal analysis (a *spatial* or between core/sample analysis per timestep ca. every 500 years).

Hypothesis 2 exists on one major analysis.

Several choices are made, which are described in the detail in the text below.

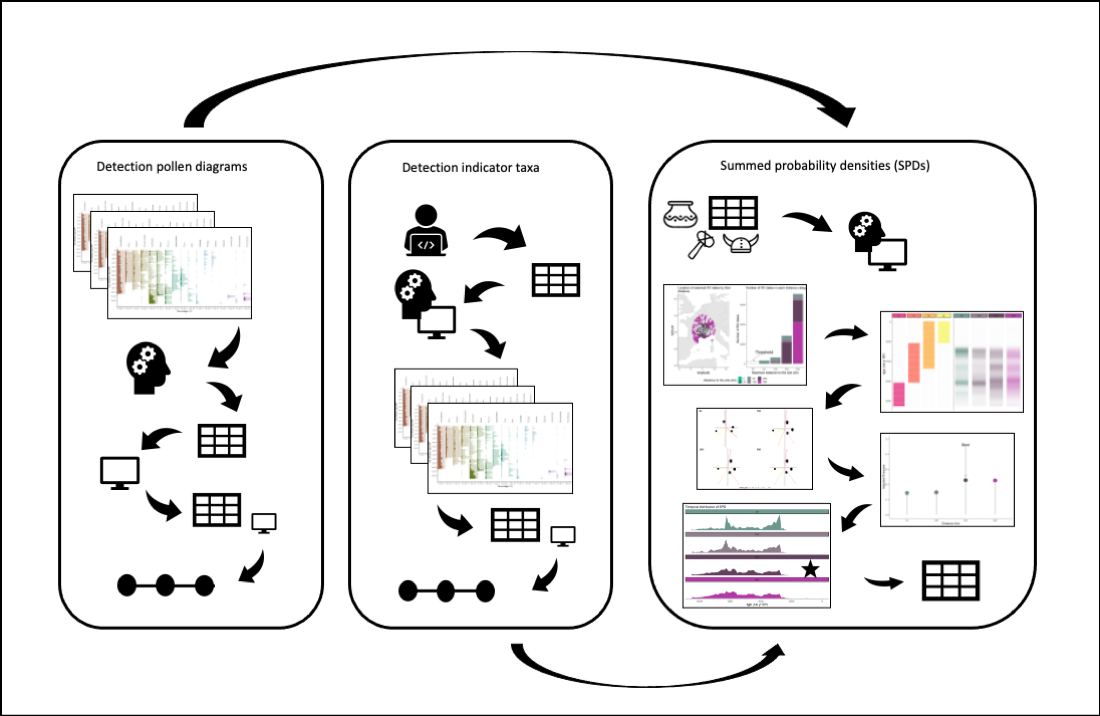
### Data filtering

The main data is divided into the targets named data\_assembly, which stores all pollen records and chronology tables, and the targets named data\_meta, which contain general site information. We have applied a number of filtering criteria to obtain as high a data quality as possible so that we can compare the numerical estimates on standardised data sets. These filtering criteria are: remove potentially duplicated pollen records, sorting levels (samples) by age, remove levels (samples) lower than a threshold of total number pollen grains counted (= pollen sum), remove pollen records based on age (minimum and maximum age ranges), remove levels (samples) depending on the age of the last control point, remove samples beyond the age ranges of interest, and remove pollen records if the total number of samples (N) is too low.

This filtering is done on the chronologies, raw pollen counts, harmonised pollen counts, and the age uncertainties from the age-depth models (Bchron). The preferable number of minimum pollen grains is set to 150, but this led to a great loss of datasets in regions with less data coverage, and we therefore reduced this number to 25 with the condition that less than 50 % of the samples must have a low pollen sum. This allow us to keep more datasets, but in the cases pollen records have a low minimum pollen sum, we acknowledge that the estimates of pollen assemblage properties (PAPs) are less robust. The maximum age beyond extrapolation is set to 3000 years because ages extrapolated beyond this threshold is considered highly uncertain. Finally, pollen records with less than 5 samples are removed for further analyses. The final HOPE dataset compilation used further in our project is called the data\_assembly\_filtered.

### Detection of past human presence

To determine the impact of past humans on fundamental ecosystem properties, we need to develop indicators of past human presence and activity. This led to the development of a new method (see next figure) where we use human event detection and indicators identified from pollen records based on expert knowledge (left and middle), combined with the method for quantifying human presence based on radiocarbon dates derived from archaeological artifacts and Summed Probability Densities (SPD, right) (Bird et al. 2022). We believe that this solves the issue that we can use a standardised variable as indicator of past human impact, and partially avoids the difficulty of creating standardised variables to detect human disturbance events across different regions and continents. This may reduce the potential circularity of human detection events derived from the same pollen records as the estimates of ecosystem properties.



#### *Detection of human events*

For each pollen record, we have detected periods of human presence from the pollen data. Two methods have been used: i) detection from pollen diagrams (North America, Europe, Asia, Indopacific; ii) detection using indicator taxa (Latin America).

#### *Detection of human events in pollen diagrams*

First, a pollen diagram of each pollen record has been examined by a regional expert and the age of each event type has been recorded.

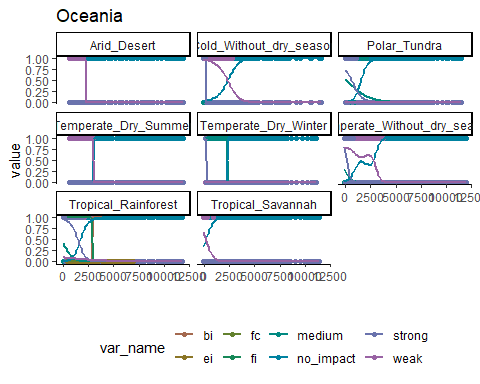
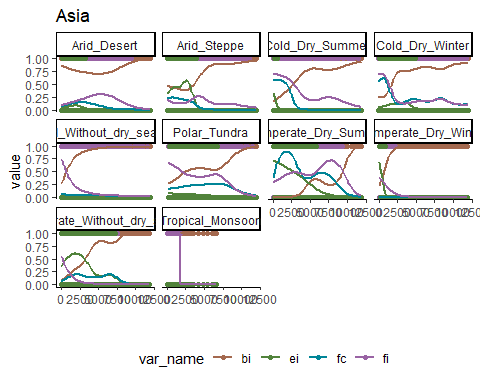
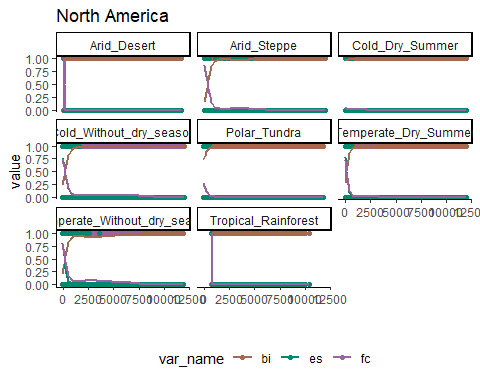
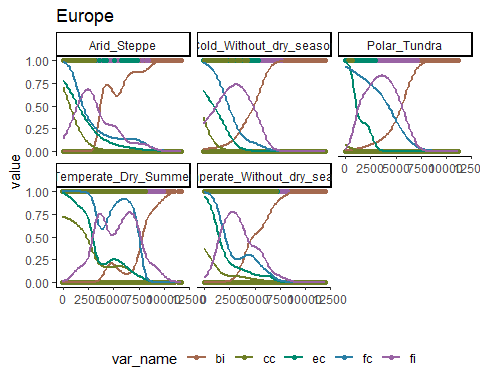
Table 2: Type of human events identified in pollen diagrams

| **Region** | **Event.type** |
| --- | --- |
| North America | BI = Before Impact; FC = First Cultivation; ES = European Settlement |
| Europe | BI = Before Impact; FI = First Indication; FCu = First Cultivation; EC = Extensive Clearance; CC = Complete Clearance |
| Asia | BI = Before Impact; FI = First Indication; FCu = First Cultivation; EI = Extensive Impact |
| Indopasific | no\_impact = No Impact; weak = Weak Impact; medium = Medium Impact; strong = Strong Impact |

Note that the event types are uniquely defined within continents, and event types with the same name have different meanings between continents.

Second, an algorithm is made to obtain the binary variables (0/1) associated with each event type that is identified in each pollen record. A new vector with the average ages in between levels (samples) of the identified event type was created because the time of the events is assumed to have occurred prior to the changed event. A new matrix was created that uses the new age vector with the events type. The different event types are assigned binary values (0/ 1) depending on if the event type is present (i.e. when the age of the specific event is detected). Ultimately, logical principles were followed and the binary values were adjusted to get the event data for each pollen record. If no human event is recognised, it does not necessarily mean that humans were absent, but instead that there was not enough information to identify human activity in the pollen records.

Below is a figure that intends to represent the underlying raw data associated with human events. To simplify the data, they are aggregated by ecozones and regions, instead of individual datasets. The different colours represent the differnet types of human events in each region. The raw event types is represented by the data that indicates whether or not the event occured (the circle dots on the y-axis are 0 or 1). The smooth trend lines represent simple binomial GAM models for each ecozone that show the main trends and alteration in the timing of events over time. For example to interpret this figure, it demonstrate that in the Cold ecozone of Europe, human impact (bi = before impact) was not observed in any of the datasets before around 7000 years BP, however, *bi* then declined steadily over time, other events that were identified began to become visible. The first indication (fi) of human interference begins to increase around 7500 years BP when *bi* is dropping. The first indication of humans increase steadily in this ecozone up to around 2500 years BP before it begins to decline. After that, the other relevant events types become apparent, such as the first cultivation (fc) increase, followed by an increase in signs of early clearance (around 3000 years BP). Complete clearance is only documented in the latter part, around 2000 years, when it increased to the present time (but this is not extensive across the pollen records, the highest proportion is around 0.25.

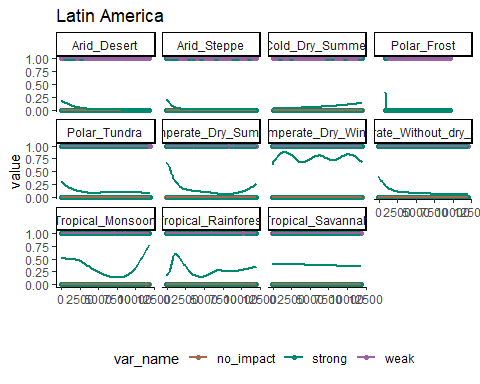


#### *Detection using indicator taxa*

The methodology for identifying past human activity in Latin America apply a different automated process to get past human events. An extensive literature review has been devoted to the detection of pollen types that are associated with human activity. The composition of indicator groups varies across the continent, regions, and countries. A computer algorithm has been created to assess each pollen record, this algorithm traverses the combinations of taxa that are recorded along with the specific indicator groups, and determined if the signal of a pollen indicator for human activity is weak or strong based on a defined threshold.

Two categories were created: i) ‘human event indicators’: a single taxa associated with human activity classified as ‘weak’ or ‘strong’ impact; ii) ‘human event indices’: a combination of particular taxon is classified as ‘weak’ or ‘strong’ impact. For each level of each pollen record, all indicators and indices were tested for its presence in the level and whole level we classified as “no impact”, “weak impact” or “strong impact”. Specifically, when one pollen grain of a human indicator is present, it is assigned as ‘weak’ and when more than 1 pollen grain of a strong impact indicator is present we assign it as ‘strong’. In addition, the pollen-type *Pinus* was only considered as a human indicator outside of its native distribution. For indices, any number of pollen grain needs to present for that specific combination.

Summary figure of the events using indicator taxa in different ecozones in Latin America:



#### *Arcehological artefacts and Summed Probability Densities*

We used the global dataset of radicarbon dates (RC dates) of archeological artefacts from Bird et al. 2022. The quantification of SPD requires a distance to be selected around each site location to collect the relevant dates of archaeological artefacts around it. This will limit the area of human presence and indirectly the amount of human activity relevant to pollen records from each site.

Only RC dates with valid geographical location (longitude and latitude), and ‘LocAccuracy’ > 0 were selected. For each pollen record, RC dates were classified by the geographical distance to the pollen record. The chosen distance classes were: 5, 25, 50, 100, 250, 500 km.

For each pollen record, one variable of SPD´s in time was calculated for each distance class. Radiocarbon dates were calibrated using calibrate function from rcarbon package with appropriate calibration curves (“IntCal20”, “ShCal20”, “mixed”). Calibration curves were obtained rcarbon package and “mixed” was created using rcarbon::mixCurves function with ‘p’ = 0.5. Calibration curves were assigned by their geographical location following Hua et al., 2013.

SPD is estimated using spd function from rcarbon package for each distance class for each year between a minimum threshold age and 12 ka. However, distance class with less than 50 RC dates is filtered out in order to maintain robust SPD estimation. The minimum threshold ages are different for different regions and are decided based on the availability of radiocarbon dating for different regions. In general there is a bias that radiocarbon dating is rather limited on younger material where in many regions there are a lack of C14 data during the last 2000 years. Table 5 show the ages where data younger than age\_from where removed.

Table 5: Minimum ages above which C14 data was removed (age\_from) for different regions

| **region** | **age\_from** |
| --- | --- |
| Europe | 2000 |
| Latin America | 2000 |
| Asia | 2000 |
| North America | 500 |
| Oceania | 500 |

In order to select distance from each pollen record, which will limit the area of human activity relevant to that pollen record, we used the expert-based detection of human events and the human impact detection from indicators of pollen types to inform the estimation of SPD:

For each distance class of SPD of each pollen record, one Redundancy Analysis (RDA) is estimated using vegan::rda function with event types as responses (binary) and SPD values as predictors. Next, R2 is estimated using vegan::RsquareAdj function for each distance class. Finally, the distance class with the highest R2 is selected as the representation of human presence, and indirectly human activity, for that pollen record.

This approach is not perfect and it is neglecting topographic differences and presence of water bodies. However, this was selected as a balance between simplicity and generality, and to avoid unnecessary increase of complexity in choosing the distance from the records. See demonstration of this method [Detection of past humans](https://github.com/HOPE-UIB-BIO/HOPE_Archaeo_C14/blob/master/R/07_Methodology%20example.md).

We had to remove Africa as a region, because we do not have data of human events to estimate *past human presence*.

### Paleo Climate

Paleoclimate from the CHELSA-TraCE21k downscaling algorithm is downloaded from the CHELSA database (Karger et al. 2021, Karger et al. 2021). The selected bioclimatic variables are annual mean temperatures ℃ (bio1), minimum temperatures of coldest month ℃ (bio6), annual precipitation kg m-2 year-1 (bio12), precipitation seasonality (bio15), precipitation of warmest quarter kg m-2 quarter-1 (bio18) and precipitation of coldest quarter kg m-2 quarter-1 (bio19), where we extracted climate values for the coordinates for each dataset\_id retrieving the full time series of every 100 years. In addition, we downloaded the monthly climatology for daily maximum near-surface temperature K/10 (tasmin).

### Pollen assemblage properties (PAP) estimation

To prepare the response variables of our main pollen dataset compilation and to be able to analyse fundamental ecosystem properties, we prepared the standard estimates of pollen assemblage properties (PAP) (Bhatta et al. 2023). The PAP estimations provide different aspects of pollen assemblage diversity which includes palynological richness, diversity and evenness, compositional change and turnover, and Rate-of-Change (RoC).

These response variables are calculated using the newly developed [R-Ecopol package](https://github.com/HOPE-UIB-BIO/R-Ecopol-package) that contain all the functions needed to estimate PAPs for our pollen data assembly. The base functions used in this package are derived from other dependency packages such as mvpart package (Therneau et al. 2014) to estimate pollen zonations with multivariate regression trees, vegan (Oksanen et al. 2022) for other mutivariate techniques and dissimilarity indices, R-Ratepol (Mottl 2021) to get the estimates of RoC, functions from iNext (Chao et al. 2014) that have been modified to extract interpolated Hill numbers based on a minimum sample size, and newly developed R functions to run DCCA using Canoco 4.5 (ter Braak xxxx) to list a few, among other, dependency packages.

#### *Pollen richness, diversity, and evenness*

The different aspects of palynological diversity are estimated using Hill´s effective species numbers N0, N1, N2, and the associated evenness ratios of N2/N1 and N1/N0. These are combined through one equation where the effective species numbers differ mainly in how the rare taxa are weighted in the parameter q:

When q is 0, rare and abundant taxa have equal weight and the number is simply the number of taxa in the sample. The equation is not possible to define for q = 1, but as it approaches 1, it is equal to the exponential of the well-known Shannon index and reports the number of equally common taxa. When q = 2, it is the same as the inverse Simpson diversity index and provides the number of equally abundant taxa with a low weight on rare taxa. The advantage of using effective species numbers is that they provide easily interpretable units and contain the doubling effect. To standardize the sample sizes, we use the rarefaction approach developed by Chao et al. These estimates are rarefied to the number of n = 150 grains, or in some cases to a lower sum (minimum n = 25). Some pollen records were only available as pollen percentages, and as the sample size is unknown, these are then rarefied to the minimum sum of percentages. The evenness ratios will be 1 if all taxa are equally abundant, and the ratios hence indicate changes in abundances between the numbers of rare, equally common, and abundant taxa.

We acknowledge that even though attempts are made to standardise richness and diversity estimates based on standard sample size, there are additional biases that are not taken into consideration such as differences in total pollen production and pollen representation (Odgaard 1998, 2001). In some cases, the total pollen sum is also too low to be considered a robust estimate, but it was a choice made on balancing loosing too much information from geographical areas with less data coverage (see data filtering above).

#### *Compositional change*

Compositional change is calculated using multivariate regression trees (MRT) with age as the constraining variable. MRT is in general a robust tool to explore and predict changes in multivariate data sets using environmental predictor variables (De´ath, Simpson and Birks 2012). This technique has been adopted in palaeoecology to detect major zones in pollen diagrams or shifts between periods of homogeneous vegetation in time (Simpson and Birks 2012). We use the pollen taxa in percentages without any data transformations as the response and the median ages derived from the age-depth model as the constraining variable. The recursive partitioning are based on chi-square distances between pollen samples constrained by time. The number of cross-validation is set to 1000, and the optimal sized tree is chosen based on the 1SD rule (Simpson and Birks 2012).

#### *Compositional turnover*

Compositional turnover is estimated using detrended canonical correspondence analysis (DCCA) with age as the explanatory variable (ter Braak and Smilauer 2007?). Changes in Weighted average (WA) sample scores (CaseR scores sensu ter Braak and Smilauer 2012) are measures of compositional turnover in standard deviation (SD) units (Birks 2007). The WA scores are regressed with time using a second-order polynomial (age+age^2) to allow more flexibility in the turnover pattern within a pollen record. Total compositional turnover is a measure of the total length of CaseR scores along the DCCA axis 1, whereas the pattern within a record is the measures between the individual samples along the DCCA axis 1. The response data are pollen percentages without any transformation to maintain the chi-square distances between samples, whereas the ages are the median ages derived from the age-depth model for each site.

#### *Rate-of-change*

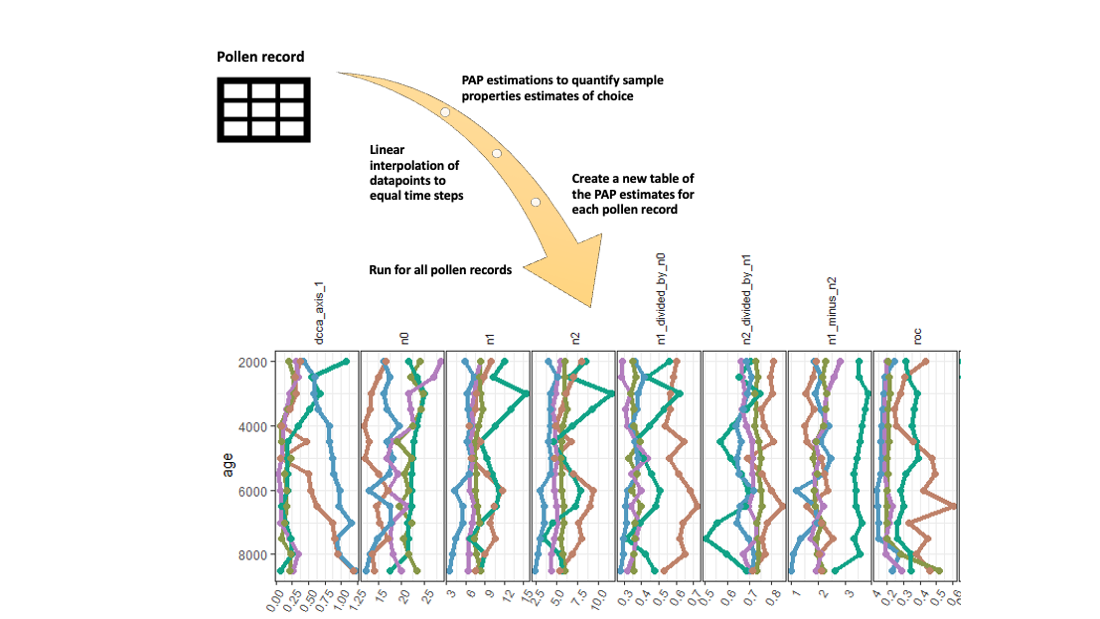
Rate-of-change for the pollen assemblages in the pollen records are quantified using the novel [R-Ratepol package](https://github.com/HOPE-UIB-BIO/R-Ratepol-package) (Mottl et al. 2021). RoC is estimated using moving windows of 500 years´time bins of five number of windows shifts where samples are randomly selected for each bin. This approach is shown to increase the correct detection of RoC peak-points than the more traditional approaches (Mottle et al. 2021). RoC are reported as dissimilarity per 500 years using the Chord dissimilarity coefficient. Sample size is standardized in each working unit to 150 grains or the lowest number detected in each dataset. We use only the RoC scores further in the analyses.

#### *Change-points detection and density estimates*

Change-points detection of all the PAP variables are calculated using conventional regression trees (RT) for single variables with Euclidean distances. An algorithm is made to detect the transitions between the resulting groups (or zones) per variable, and these are coded as new binary (0/1) variables. A change-point is defined as 1, where the mean ages between the two consecutive samples are used as the timing of this significant change. This is done individually for each PAP variable.

The significant change points of the richness, diversity, and evenness variables are combined into one variable, and the significant change points of compositional turnover, compositional change, and rate-of-change is combined in a second variable. The density of these two variables are calculated using a Gaussian kernel,and re-scaled to each of specific age ranges for each individual pollen record (i.e. minimum and maximum ages). To solve the boundary issue in density estimation the data is reflected to 0. We extract the interpolated values at every 500 years time step.

Conceptual figure to show the process:



#### Equal spacing of variables

All response variables have been estimated using the harmonised pollen records for each location. To obtain estimates of equal spacing of 500 years, we used linear interpolation. In the context of the temporal analysis of hypothesis 1 and hypothesis 2 we analyse samples distributed in space across time, and equal time steps are necessary.

In order to choose a method of interpolation to obtain data on equal time steps, we compared generalise additive models (GAM), hierarchical generalised additive models (HGAM), and simple linear interpolation. By applying linear interpolation, we found that the correlation structure between the multivariate response variables are more similar to the original estimates without equal spacing than applying a GAM or HGAM. The GAM or HGAM models sometime showed unexpected patterns in single PAP estimate that changed these correlations. Since we cannot individually assess all single models for each of the variable in all of the records (>1000), we choose the simplest linear interpolation method. Similar issues were detected when estimating the density variables of changes points. The first approach was to estimate densities of the individual change points, and then use hierarchical generalised additive models (HGAM) to find the common pattern between the two groups representing significant changes in richness, diversity, and evenness, or change in pollen assemblages (MRT, RoC, DCCA1). As some of the models did not converge and showed inconsistent patterns we use the density estimates on the combined variables directly and extracted the interpolated values at every 500 years step.

### Numerical analysis of hypothesis 1

#### Hierarchical variation partitioning

To test if the ecological processes have changed due to past human activity in single records, we use reduced rank multivariate regression. This is also known as distance-based redundancy analysis (db-RDA). We used the R package rdacca.hp to run hierarchial variation partitioning with several predictors. This estimates the variation per variables in different combinations to get the average variable importance independent of the order of predictors. db-RDA was performed using *Gower-distances* adding a constant because our response data, the PAPS, is a mixture of different units.

Depending on the type of spatial or temporal analysis for hypothesis 1, the explanatory variables are either past human impact (SPDs), palaeoclimatic variables, and/or time (see below).

SPD is the variable of main interest as it represent past human presence. The palaeoclimate is a matrix of summer precipitation, winter precipitation, annual temperatures, and winter temperatures. These are selected as we considered them most relevant to represent major differences in climatic conditions in all the regions (in respect to differences in warm, cold, dry, wet, or regions with high seasonality). Time is represented by the ages of each pollen record, however, this is more difficult to interpret. We assume age may represent time dependent changes such as natural successions and/or ecological changes due to interaction between taxa.

In the hierarchichal variation partitioning analysis, the predictor variables can be applied either as individual predictors or as groups of predictors. In our case, we run the analysis with *groups of predictors*. This means that the palaeoclimatic variables are included as one matrix and not assessed as individual predictors. (The overall results is not very different from using individual predictors).

Statistical testing can be applied in two ways, both which shuffle the predictor variables. Using time\_series set as TRUE, statistical testing is performed with restricted Monte Carlo permutation for time series analysis (sensu ter Braak XXXX) which has a cyclic behavior of samples to keep the stratigraphical order intact. If time\_series is set as FALSE, the predictor variables will be randomly shuffled as many times set in the parameter permutations (= the default permutation in the rdcca.hp package).

The analysis is run in two different ways:

1. to analyse *spatial changes* which run the hierarchical variation partitioning within single record.
2. to analyse the *temporal patterns* in space for each region of the 500 year time steps.

In this latter analysis, we restructure the data so that each analysis is run per time bin for each continent. First we did it for each ecozone on a continent, but the results is relatively similar so we use the continental scale. The predictor groups in each time bin are the past human presence and the matrix of palaeoclimatic variables. It was necessary to filter out time bins which have less than 5 samples, and for some bins, if all the spds equal zero, the analysis will fail. In this cases there is an insufficient numbers of predictors and the analysis will return NA for these specific time bins.

We use the adjusted R2 to assess the goodness of fit of models. Adjusted R2 is the modified version of R2 that correct for number of samples and predictors in the model. It is calculated using the Ezekiel formula adjR2 = 1 - (1-R2)\*(n-1)/ (n-m-1), where n = number of samples in the dataset, and m is the number of variables. In vegan these are so called semipartial R2 (Legendre et al. 2011). Adjusted R2 can be negative,in case the adjR2 for predictors are negative we replace the values with 0 as these can be ignore.

### Numerical analysis of hypothesis 2

To test if the interrelationship between the pollen assemblage properties have changed through time, we performed a principal component analysis (PCA) with scaling between variables for samples within each time step in different ecozones. We use the global Koppen-Geiger Ecozones division of 15 groups. Within each ecozone we perform procrustes analysis to compare the changes in PCA between all time steps. We create a distance matrix based on the sum-of-squares with symmetric scaling (m2). The focus is on scores from the pollen assemblage properties (= species scores). The m2 will represent a matrix of dissimilarity distances between the different time steps which we use further as response matrix in the hierarchical variation partitioning.

Hierarchical generalised additive models (HGAM) are used to create the general patterns of past human presence (SPDs) and palaeoclimate based on the individual sequences we have from the different ecozones. The predictions represent the general pattern of these variables in each ecozone which we then use further as explanatory variables in the hierachical variation partitioning to assess how much past humans, climate or time dependent processes explains the variation of changes among the correlations between ecosystem properties in the past (hence pollen assemblage properties).

# Part II: Results

In total, we have 1263 pollen records in the HOPE standardised dataset distributed across the different regions of North America, Europe, Asia, Latin America, and Oceania included in the analysis.

Inspection of the models from spatial models of hypothesis 1 show there are models with a very poor fit with negative total adjR2 (= 13 models). These models are removed. There are still some models that have an adjR2 close to zero, which may give very high unique percentages above 100 (range of adjR2 is 0.005 to 1). To remove the uncertain data points, we defined a minimum threshold of the adjR2 using the lower 5 % quantile of all adjR2. This removes models with an adjR2 < 0.121 from the summary (= 63 models). This leaves us with 1184 models that are considered good enough to be included in the the overall summary.

The unique and average shared variances are recalculated as percentages of the total, and we summarise the results using the median unique and average shared percentages for the predictors in each ecozone in each region. If an ecozone have less than 5 records, it is not included in the summary.

The temporal analysis is conducted on the regional level, showing the unique and average shared percentages for the predictors of total variance over each 500 years time steps up to 8500 years.

In the hypothesis 2 we have one model per ecozone, therefore we use directly the unique and avereage shared percentages of the total.

Figure 1-5 provide the regional summary of the results in the order North America, Europe, Asia (northern hemisphere), Latin America, and Oceania (southern hemisphere). The colors of the circular bar chart correspond to the ecozones displayed on the regional map (for h1 at the moment). The clear (darker) color represent the unique median percentage, and the shaded (lighter) color represent the average shared variability between the predictors.Together, each bar represent the median percentage individual contribution for each ecozone. When the shaded color is large(r) than the clear color it indicate that most of the variability explained by this predictor is shared among all the predictors. The second figure show the unique and average shared contribution for humans and climate over the time steps in each region. If there is no shaded color it means the average shared contribution is low. Ecozones with less than 5 samples are not included as the summary statistics results will be based on very local signals (only 2-4 models/records).

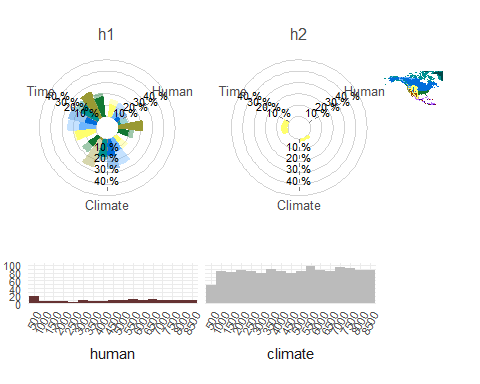


Figure 1: The summary results for North America.

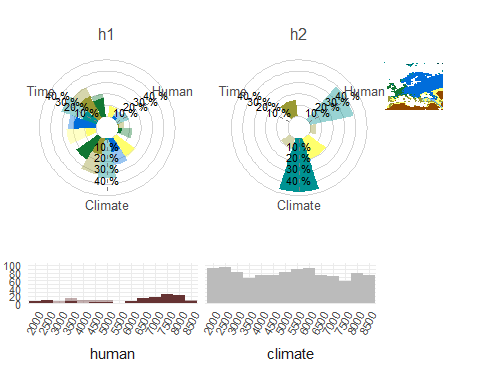


Figure 2. The summary results for Europe.

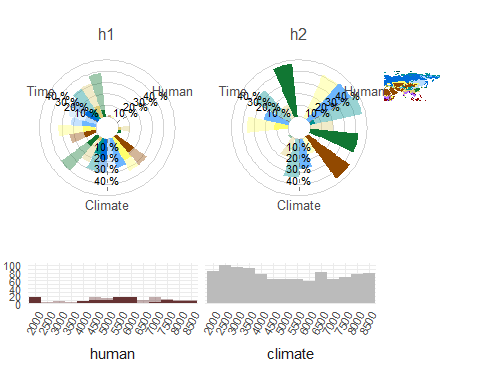


Figure 3: The summary results for Asia

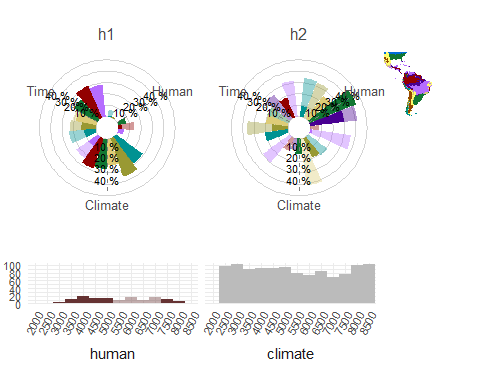


Figure 4: The summary results for Latin America.

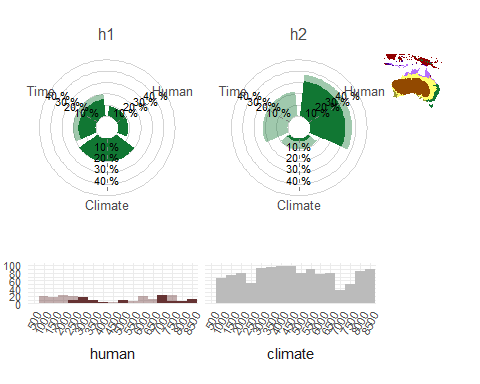
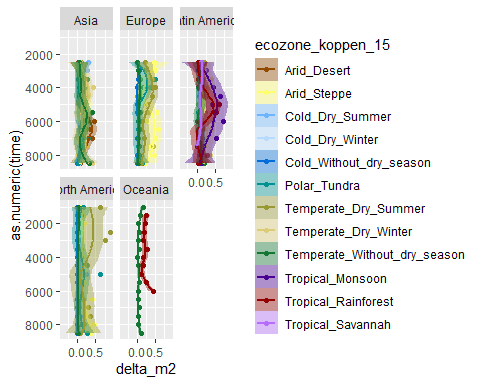


Figure 5: The summary results for Oceania.

Explorative figures of temporal changes in procrustes sum-of-square (m2):



# Part III: Validation of methodology

Our results show that humans have little effect on composition of ecosystem properties over time and the conclusion based on the our results are that past humans have not significantly changed ecosystem properties in the past.

We like to find out: 1. How does the composition of ecosystem properties look like, and are there any visual changes in the composition of ecosystem properties over time? 2. How do the spd look like, and are they useful indices of past humans? 3. Are our results robust?

## Explorative data analysis

To answer these questions, we did some explorative analyses of the ecosystem properties to try to get a better visual understanding of the data.

To address the first question how the composition of ecosystem properties look like and to find out if there are any changes in composition of ecosystem properties, we have performed unconstrained PCOA (or db-PCA) using Gower distances, and plotted the first PCOA axis with time for the individual records (but aggregated within Ecozones to avoid too many figures). We also explore the spds by plotting them with time to see how they look like.

To find out if spds are able to explain any changes in the data, we use multivariate regression trees (MRT). The response data is the composition of ecosystem properties transformed with Gower- distances and the constraining variables are time and spd. We used MRTs as these are easy tools that can handles non-linear patterns and missing data. We combine the results of the MRT partition with the PCOA (unconstrained data) to explore the unconstrained data and get a better understanding of the differences in the ecosystem properties.

We use only the subset of Europe for exploration. The analyses are run three times on different scales: 1) for all records individually (396 individual records), 2) for records nested within ecozones (4 subsets), and 3) whole region (the full subset).

The results of the individual PCOA analysis of each record is not shown. Instead, the first PCOA axis is plotted against time to see if there are changes in the composition of ecosystem properties over time. The overall results of the MRTs for individual records are summaried in tables.

For the analysis within ecozones in regions, we extract the significant partitions from the MRT analyses, and explore the results using ordination diagrams. In addition, we passively fit the explanatory variables, time and spds, to the ordination models to help explore the results.

## Results

### 1. Results from the analyses of individual records

The summary results of the first PCOA axis against time show there are changes in the composition of ecosystem properties over time for individual sites (Figure 1). To see how spds represented as past human population vary across records, the individual curves are plotted against time between 2000 and 8500 years BP (Figure 2). First inspections looks reasonable.

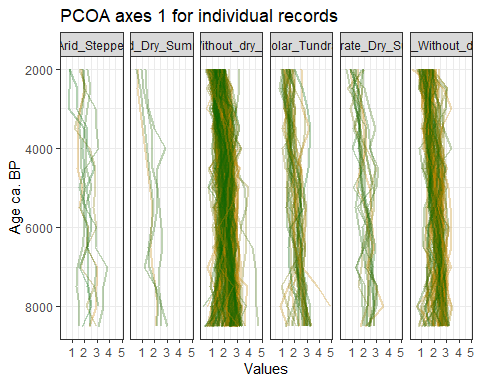


Figure 1: Plot of the individual PCOA axis 1 for all records against time.

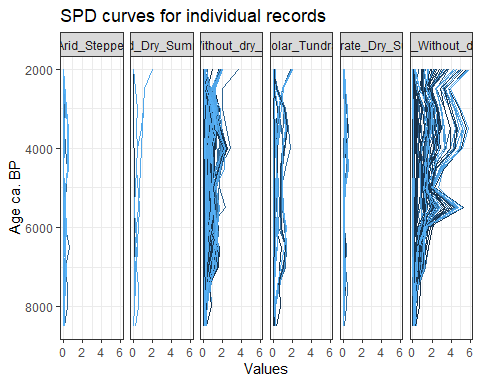


Figure 2: Plot of the individual spd curves representing changes in past human population for all records over time.

The results of the MRT are summaried in one figure and two tables below. Overall the results of the MRT analyses show that it is possible to detect major changes in the composition of ecosystem properties explained by time and humans. Only one record show no change, most records have either two or three major significant divisions, whereas a few records show 4 significant divisions (Figure 3).

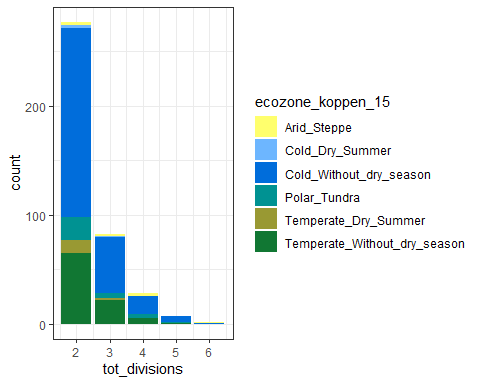


Figure 3: Number of records with different divisions representing signifcant changes in ecosystem properties explained either by time, humans or both. The records are coloured by ecozones.

Two tables below are created to give a simple summary of the MRT results, first table summarise the counts in total whereas the second table show the numbers within ecozones.

The total number of records are counted for:

* spd explains the significant partitions (spd\_true)
* age explains the significant partitions (age\_true)
* only spd explains the significant partition (spd\_unique)
* only age explains the significant partition (age\_unique)
* where both age and humans explains the significant partitions (spd\_age\_true)

Table 1: Summary of the total number and percentage for the whole region Europe of records analyses where age, humans, or both variables explains most of the changes as observed the MRTs.

| **tot\_records** | **variable** | **counts** | **percent** |
| --- | --- | --- | --- |
| 396 | spd\_true | 68 | 17.171717 |
| 396 | age\_true | 360 | 90.909091 |
| 396 | spd\_unique | 36 | 9.090909 |
| 396 | age\_unique | 328 | 82.828283 |
| 396 | spd\_age\_true | 32 | 8.080808 |

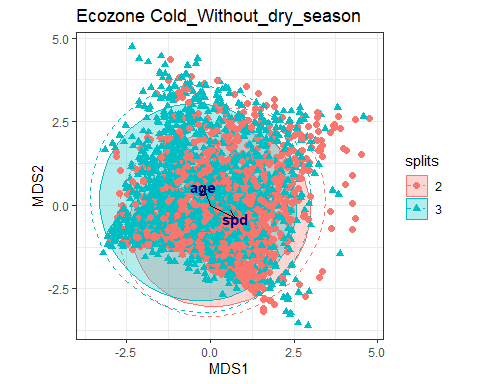
Table 2: Summary of the total number and percentages for records in the different ecozones where age, humans, or both variables explains most of the changes as observed the MRTs.

| **ecozone\_koppen\_15** | **tot\_records** | **variable** | **counts** | **percent** |
| --- | --- | --- | --- | --- |
| Arid\_Steppe | 7 | spd\_true | 1 | 14.285714 |
| Arid\_Steppe | 7 | age\_true | 7 | 100.000000 |
| Arid\_Steppe | 7 | spd\_unique | 0 | 0.000000 |
| Arid\_Steppe | 7 | age\_unique | 6 | 85.714286 |
| Arid\_Steppe | 7 | spd\_age\_true | 1 | 14.285714 |
| Cold\_Dry\_Summer | 4 | spd\_true | 2 | 50.000000 |
| Cold\_Dry\_Summer | 4 | age\_true | 2 | 50.000000 |
| Cold\_Dry\_Summer | 4 | spd\_unique | 2 | 50.000000 |
| Cold\_Dry\_Summer | 4 | age\_unique | 2 | 50.000000 |
| Cold\_Dry\_Summer | 4 | spd\_age\_true | 0 | 0.000000 |
| Cold\_Without\_dry\_season | 248 | spd\_true | 41 | 16.532258 |
| Cold\_Without\_dry\_season | 248 | age\_true | 228 | 91.935484 |
| Cold\_Without\_dry\_season | 248 | spd\_unique | 20 | 8.064516 |
| Cold\_Without\_dry\_season | 248 | age\_unique | 207 | 83.467742 |
| Cold\_Without\_dry\_season | 248 | spd\_age\_true | 21 | 8.467742 |
| Polar\_Tundra | 30 | spd\_true | 3 | 10.000000 |
| Polar\_Tundra | 30 | age\_true | 28 | 93.333333 |
| Polar\_Tundra | 30 | spd\_unique | 2 | 6.666667 |
| Polar\_Tundra | 30 | age\_unique | 27 | 90.000000 |
| Polar\_Tundra | 30 | spd\_age\_true | 1 | 3.333333 |
| Temperate\_Dry\_Summer | 14 | spd\_true | 1 | 7.142857 |
| Temperate\_Dry\_Summer | 14 | age\_true | 13 | 92.857143 |
| Temperate\_Dry\_Summer | 14 | spd\_unique | 1 | 7.142857 |
| Temperate\_Dry\_Summer | 14 | age\_unique | 13 | 92.857143 |
| Temperate\_Dry\_Summer | 14 | spd\_age\_true | 0 | 0.000000 |
| Temperate\_Without\_dry\_season | 93 | spd\_true | 20 | 21.505376 |
| Temperate\_Without\_dry\_season | 93 | age\_true | 82 | 88.172043 |
| Temperate\_Without\_dry\_season | 93 | spd\_unique | 11 | 11.827957 |
| Temperate\_Without\_dry\_season | 93 | age\_unique | 73 | 78.494624 |
| Temperate\_Without\_dry\_season | 93 | spd\_age\_true | 9 | 9.677419 |

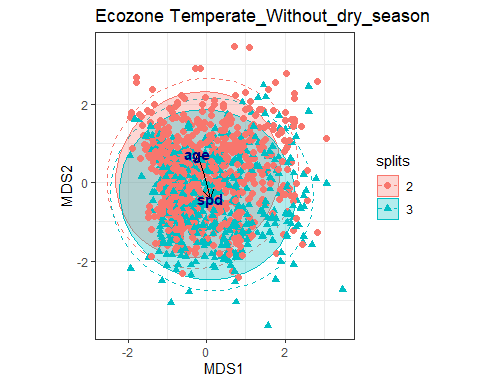
### 2. Results for the analyses with records nested in ecozones

Below figures display the ordination diagrams of the unconstrained PCOA using Gower-distances for each of the ecozones in Europe. The samples are coloured by the significant MRT partitions for the same dataset. The constraining variables age and humans are then passively fitted to the PCOA model using envfit. This means it has no direct effect on the sample scores, but shows how well correlated these variables are with PCOA axis 1 and 2. The length and direction of the arrow indicate which axes it has best correlation with.

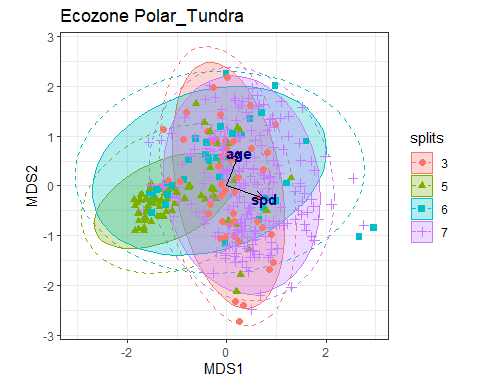
## [[1]]



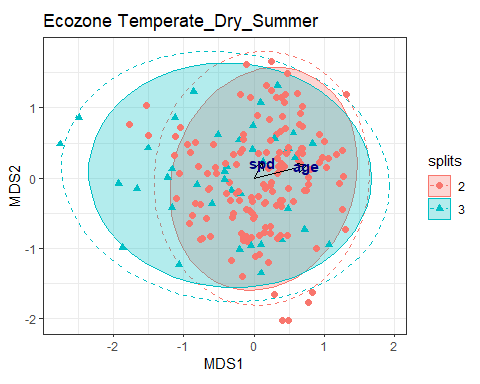
##   
## [[2]]



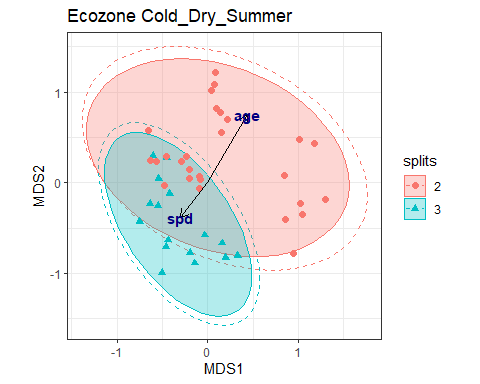
##   
## [[3]]



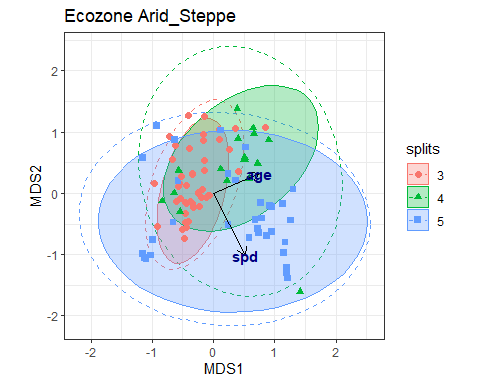
##   
## [[4]]



##   
## [[5]]



##   
## [[6]]

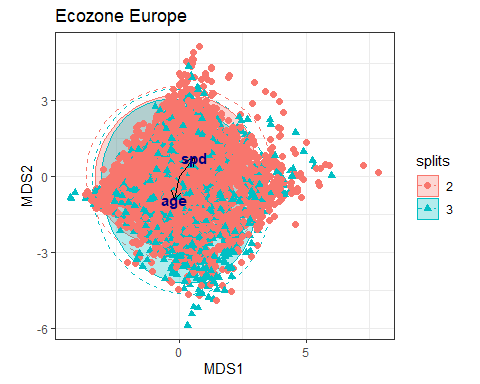


The following list show the results of the envfit with the PCOA model and age and time.

## [[1]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age -0.33487 0.94227 0.0281 0.001 \*\*\*  
## spd 0.89199 -0.45206 0.0531 0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999  
##   
##   
##   
## [[2]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age -0.28767 0.95773 0.1001 0.001 \*\*\*  
## spd 0.22539 -0.97427 0.0268 0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999  
##   
##   
##   
## [[3]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age 0.34987 0.93680 0.0778 0.001 \*\*\*  
## spd 0.94841 -0.31703 0.0957 0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999  
##   
##   
##   
## [[4]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age 0.97425 0.22546 0.0810 0.001 \*\*\*  
## spd 0.47397 0.88054 0.0092 0.409   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999  
##   
##   
##   
## [[5]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age 0.50185 0.86495 0.5713 0.001 \*\*\*  
## spd -0.61780 -0.78634 0.1829 0.015 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999  
##   
##   
##   
## [[6]]  
##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age 0.91662 0.39976 0.0765 0.016 \*   
## spd 0.46035 -0.88774 0.1489 0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999

### 3. Results for the analyses with records nested in Europe

The next figure show the results for all records in one analyses for Europe.



The results of the envfit of the PCOA model with time and humans.

##   
## \*\*\*VECTORS  
##   
## MDS1 MDS2 r2 Pr(>r)   
## age -0.19617 -0.98057 0.0265 0.001 \*\*\*  
## spd 0.62666 0.77929 0.0290 0.001 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
## Permutation: free  
## Number of permutations: 999

### Summary

For the summary of the individual records, past humans (spd) is important explaining MRT partitions in 17.2 % of the records, whereas in 9.85 % of the record, it is uniquely explaining the partitions. Time is explaining most of the MRT partitions in the individual records, and it is uniquely explaining the variation in 82.6 % of the individual records (for comparison with numbers in the different ecozones, see Table 2).

When analysing the data on larger spatial scale as records nested within the ecozones, the results show that past humans have affected the fundamental ecosystem properties in the Cold and Polar ecozones. Changes in the Polar region is more complex, and time dependent factors are also important in the partitioning, whereas in the Temperate and Arid ecozones, time is the most important variable for the variation of the data.

When putting all the records together, it is still possible to distinguish differences in composition of ecosystem properties in Europe explained by past humans. Past human activity seems to have started a shift in the composition of ecosystem properties in the another direction, and this is determined by the average threshold spd>= 0.2555 & spd < 0.2555.

This is interesting, because changes explained by humans in smaller percent of the records (local), have had an effect on ecosystem properties that seems to be possible to be scaled up to a larger scale.

Though past humans have an observed impact on the composition of past ecosystem properties, the MRTs have a high prediction error (CV error ~ 0.9). You can see this in the PCOA diagram where most of the samples overlap. Hence the results indicate that past human have not statistically significantly alter the fundamental ecosystem properties of the past.

Summary:

1. The results show that there are changes in the composition of ecosystem properties over the time period of interest
2. The index of past humans looks reasonably - there are changes in spd and this vary with different records
3. Past human (spds) are able to explain changes in the composition of ecosystem properties using MRTs
4. The MRTs indicate that past humans have had an effect on ecosystem properties, but the results have a low prediction power
5. It seems that MRTs are useful tools for summarising results on different spatial scales.