Report HOPE Hypothesis 1 methodology & results

Contents

Preface

The main scope of this report is to give you a detailed overview of the data input, data process, and analyses that has been done to find out if human activity have changed fundamental ecological processes over time (Hypothesis 1 in the HOPE project).

The report contain the workflow of methodological steps, description, and results. The complexity of the data processing prior to the main analyses makes it difficult to present orally without authors having relatively good understanding of the workflow, to be able to give (critical) comments (as to if major changes are needed). The hope is that this then will make the foundation of the methodology and result part of a manuscript.

The next step is to write a manuscript, but before we start it would be good to know what kind of involvement you would like to have, if you prefer to write certain sections etc. It is all open for suggestions and feedback. A meeting will be schedule to discuss these matters.

Things that would be good to make clear for everyone involved:

1. Authorship:
    - Authorships
    - Data contributors
2. Target journal -> important to decide now as to how to structure a manuscript
3. Interesting highlights of our results - the main story
4. Feedback and comments to the methodology? Major changes needed?
5. Progress plan

(Disclaimer: Many of the images and figures are quickly made, or older figures/images are included for help in visualising variables for the sake of completeness of the report (for HOPE members), I did not focus on good scientific writing either because that takes too much time for me at this stage, so please do not share this beyond the group).

Part I: Methodology

Main data aquisition

FOSSILPOL

Aquiring the main pollen data for our analyses is done a priori. Raw pollen datasets are carefully selected using the *R-Fossilpol* package and the guidelines to the workflow are well described in Flantua et al. 2023 and our website Fossilpol project. Most of the compilation of datasets are available through the Neotoma

Paleoecology Database. Though some additional datasets are from private owners in regions with lacking data, and these have restricted access for use. We do not have the intellectual property rights to make them public available. Only the derivative of the analyses will be possible to share openly. Table 1 provide the overview of the the settings used in the FOSSILPOL workflow to get the standardised datasets for this project.

Harmonisation tables

An important step in FOSSILPOL to get the standardisation of pollen records within and across regions is harmonisation of pollen types. There are different analyst with different schools and background, and the nomenclature can vary widely. To be able to make numerical comparisons across different pollen records, the level of pollen taxonomy should be similar. As a result, pollen harmonisation tables are produced for different regions, attempting to minimise biases related to this. The regional harmonisation tables used in our project are for Europe, Levant, Siberia, Southern Asia, Northern America, Latin America, Africa, and the Indo-Pacific region (Birks et al. harmonisation paper). These tables can be downloaded from (xxx), and are used as additional input in the Fossilpol workflow above (see Fossilpol step_by_step guide).

Workflow for HOPE Hypothesis 1

We use the targets package in R to produce a reproducible workflow for all data processing, analyses, and visulisation of the results from this project. For this we created a data analysis project called HOPE_hypothesis1 in Github. This contain all data, metadata, and R functions needed to run this project, and it provides full transparency in all steps from data input, data processing, statistical analysis, and the derived results.

Our output of targets is set up with in an external storage at OneDrive where we save the main targets folder with data and meta data, while the target script and functions are saved in our Github project. Note that first time running targets without access to our main data folder will take time. The targets are split up in many steps with several specialised functions that is made to loading necessary data, to get estimates of various variables or structuring various subsets of data step wise. This is to avoid the need for re-running major parts that take too long when the data is already processed. There is a dependency in the list of targets that the in the end to run the final analyses, all other targets need to be up-to-date. Targets will detect any changes made in the functions, if this happens the targets will automatically rerun the parts that are dependent on this change, but at the same time skip all the parts that are up-to-date.

The file structure in Github is

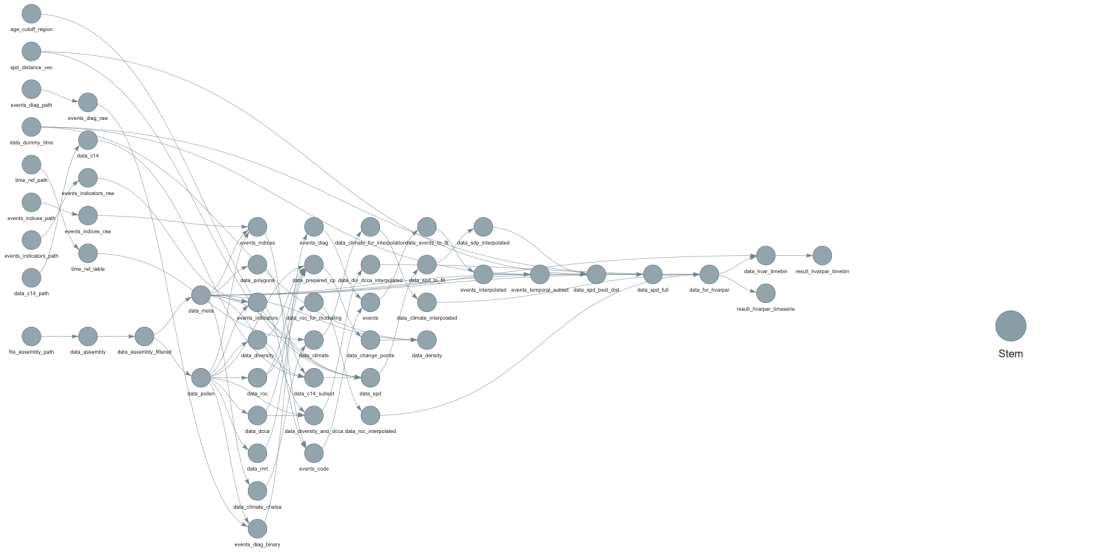
```
##                     levelName
## 1  HOPE_Hypothesis1
## 2   ¦--___Init_project___.R
## 3   ¦--_targets.R
## 4   ¦--_targets_packages.R
## 5   ¦--HOPE_Hypothesis1.Rproj
## 6   ¦--R
## 7   ¦   °--functions
## 8   ¦       ¦--climate
## 9   ¦       ¦--data_wrangling
## 10  ¦       ¦--events
## 11  ¦       ¦--hvarpart
## 12  ¦       ¦--modelling
## 13  ¦       ¦--PAPs
## 14  ¦       ¦--spd
## 15  ¦       °--visualisation
## 16  °--renv
```

All that is needed when the project is set up, is to run the scripts `___init_project.R___` and `_targets.R`. This install and load all the packages that is needed and functions that have been created to run the targets

Table 1: Table 1: Selection of settings applied in FOSSILPOL

Setting type	Selection
General	
geography_criteria	-180
long_max	180
lat_min	-90
lat_max	90
alt_min	NA
alt_max	NA
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	NA
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

or tasks for this project. The functions are divided into different folders to make it easier to find the functions for subtasks instead of collecting them in one script. The functions are specialised wrapper functions for our data structure, but they depend on costume made functions that are gathered in the `R-Ecopol` package. The figure below show the network of targets.



The targets are arranged in a order to prepare all the data needed for the main analysis in the end. First, it is setting the options for data interpolation to get all samples on equal time spacing, loading the FOSSILPOL dataset, creating dummy tables and variables needed for input in the specialised functions to process data. This is followed by the data processing steps for each of the explanatory and response variables that is needed in this project. The largest pre-processing of data starts with preparation of the explanatory variable detecting past human presence and impact. This is a major analyses in itself. This follows by 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. It is followed by different targets that process all estimates of the response variables selected to get measures of pollen assemblage properties. In general, all data such as raw estimates and interpolated data are kept to allow careful checking and validation of output. In the end, data is structure to fit the main analysis. This analysis is divided into two parts which we call the 1) the spatial (within core) analysis, and 2) the temporal analysis (a *spatial* or between core/sample analysis per timestep ca. every 500 years). Several choices are made, which are described in the detail in the text below.

Data filtering

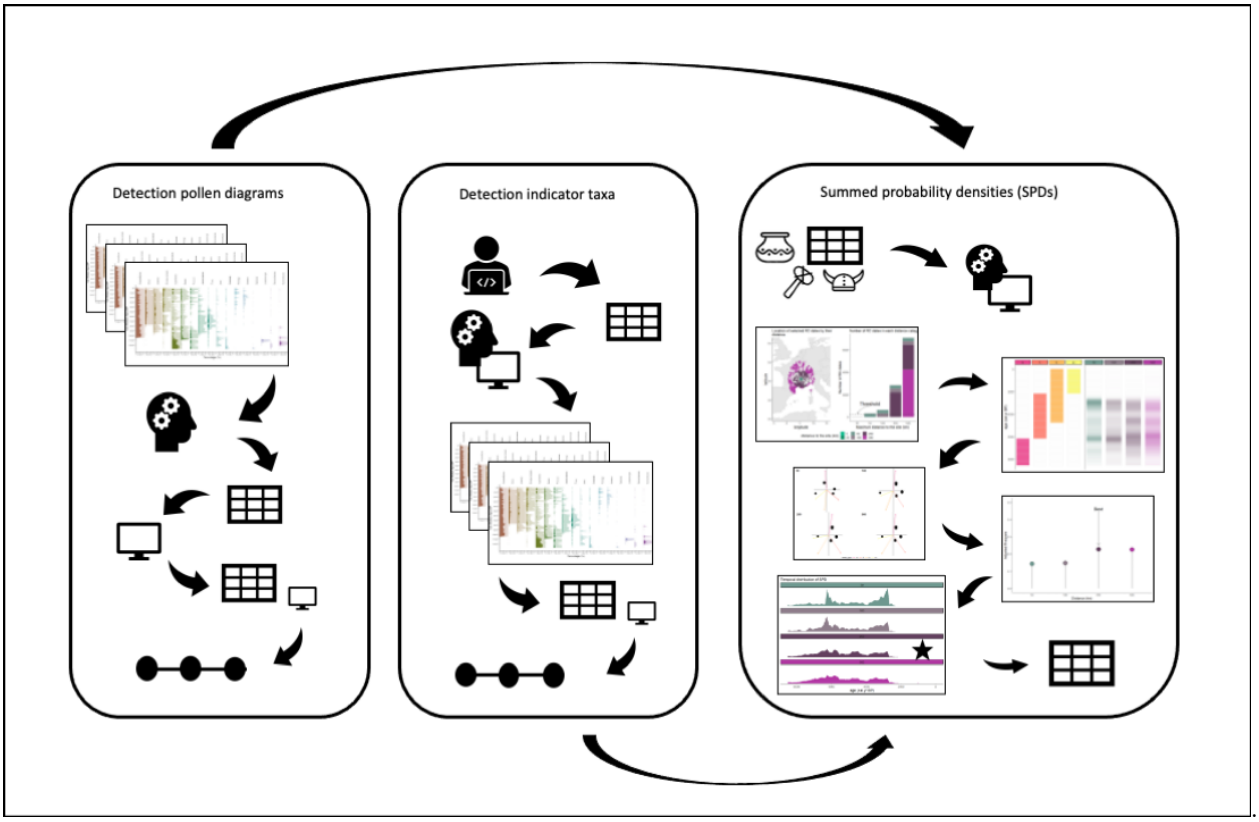
The main data are divided into `data_assembly` which store all the pollen records and chronologies, and `data_meta` which contain the general site information. An extra data filtering step is done on the `data_assembly` to get as high data quality as possible to be able compare the numerical estimates on standardised datasets. These extra filtering criterias are removing potentially duplicated pollen records, sorting levels (samples) by ages, filtering out levels (samples) based on a threshold of total number pollen grains counted (= pollen sum), filtering out sequences (pollen records) based on age limits (minimum and maximum age ranges), filtering out levels (samples) by the last control point, filtering out samples beyond the age limits of interest, and filtering out pollen records based on the total number of samples (N).

This filtering is done on the chronologies , raw pollen counts, harmonised pollen counts, and the age uncertainties related to the chronologies. 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, if less than 50 % of the samples had a low count, as low pollen sum can be due to varying pollen concentrations in different parts of a sequence. This allow us to keep more datasets, but in these cases the pollen records have a low pollen sum, we acknowledge that the estimates of pollen assemblage properties (PAPs) are less robust (check/report how many cases..). The maximum age beyond extrapolation is set to 3000 years. Ages extrapolated beyond this threshold is considered highly uncertain. Also all pollen records with less than 5 samples are removed for further analyses. The data used further in our project is called the `data_assembly_filtered`.

Detection of past human presence

In order to detect the impact of past humans on fundamental ecosystem properties we needed to develop indicators of past human presence and activity. This led to the development of a new approach where we use detection of human events identified from pollen records based on expert knowledge, in combination with the methodology of quantifying presence of humans based on radiocarbon dates from archaeological artefacts and Summed Probability Densities (SPD) (Bird et al. 2022). In our view this solved some issues where we can use one standardised variable as an indicator of past human impact, and partly avoid the difficulties of creating standardised variables for detecting human disturbance events in different regions and across continent, and reduce the circularity of detecting human events on the same pollen records as the estimates of ecosystem properties.



Detection of human events For each pollen record, we have detected periods of human events 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 from 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: Table 2: Type of human events identified in pollen diagrams

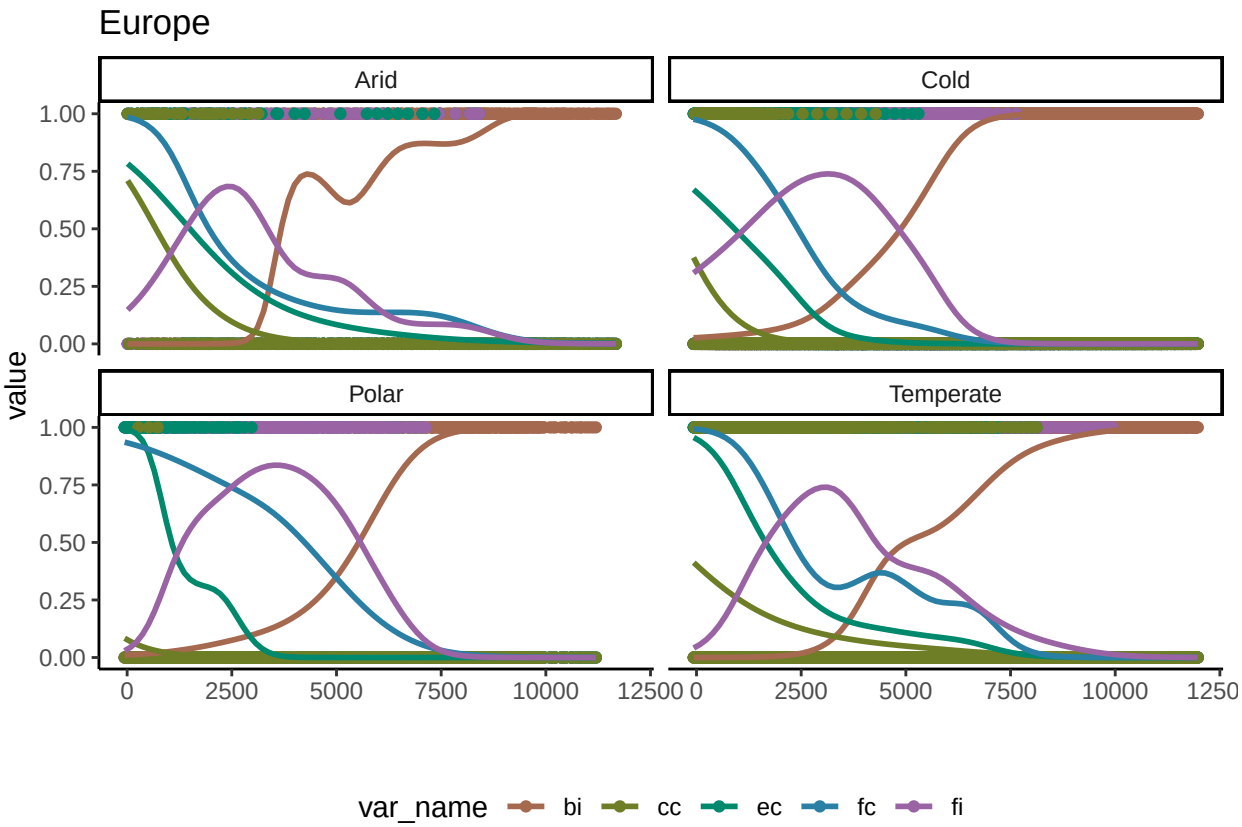
Region	Event.type
North America	BI = Before Impact; FC = First Cultivation; ES = European Settlement
Europe	BI = Before Impact; FI = First Indication; FCu = First Cultivation; EC = Extensive Clearance; CC =
Asia	BI = Before Impact; FI = First Indication; FCu = First Cultivation; EI = Extensive Impact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

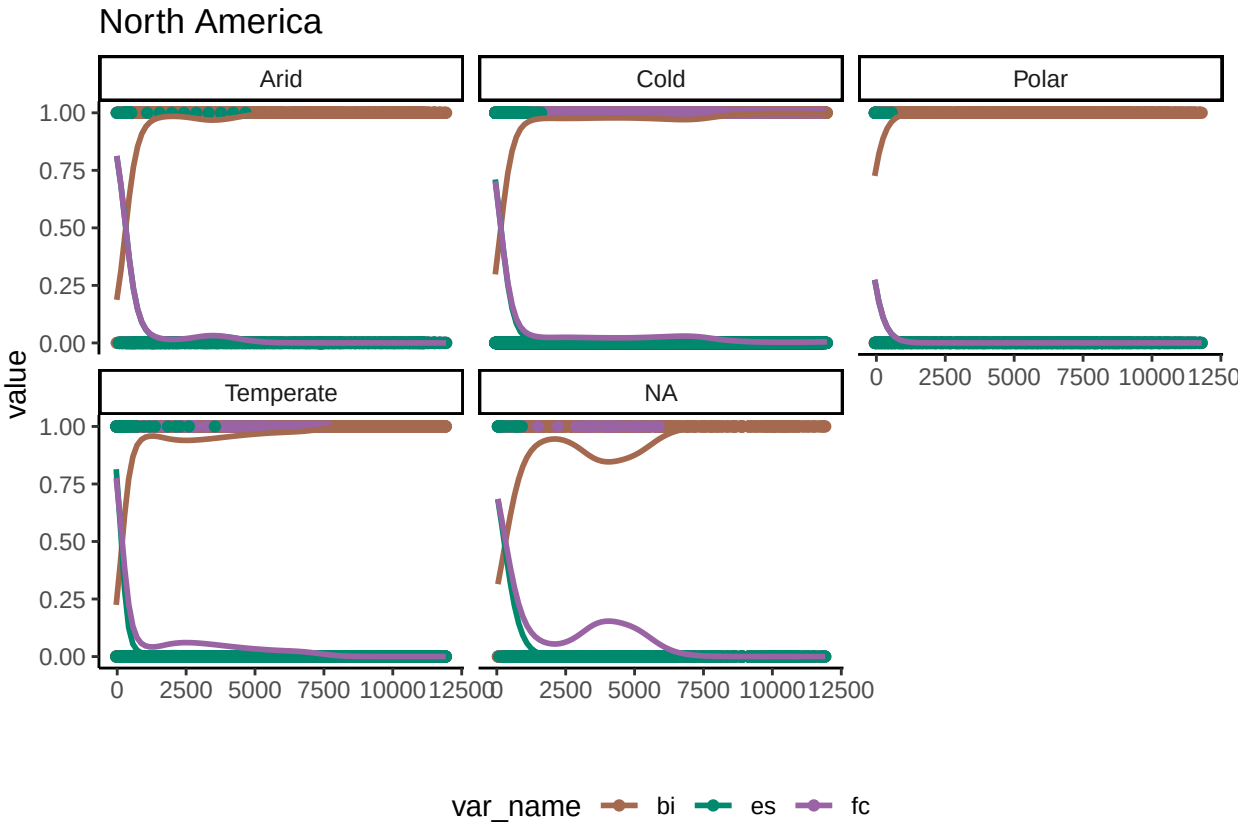
Add more text about how the event types have been defined. . . ..

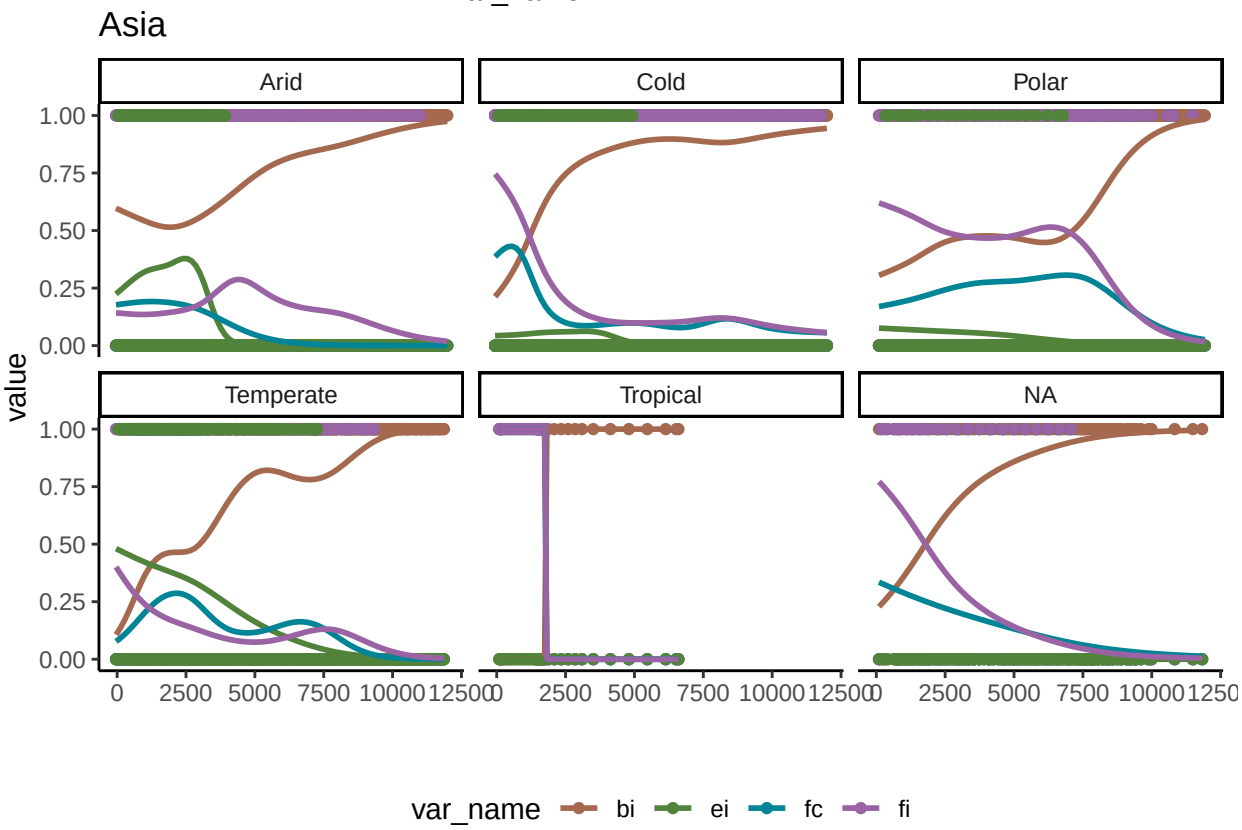
Note that the event types are uniquely defined within continents, and event types with the same name have different meanings between continents.

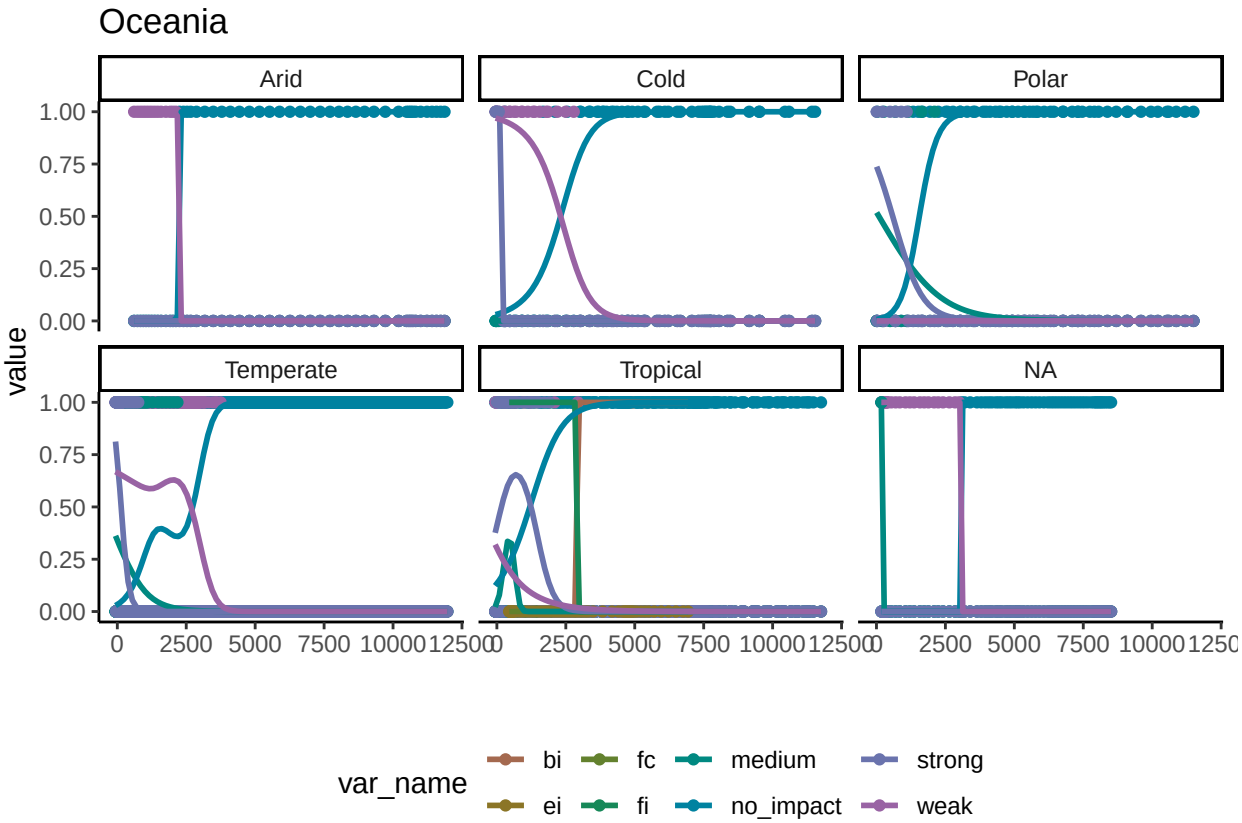
Second, an algorithm is made to get the binary variables (0/1) per event type identified in each pollen record. A new vector with values of average ages in between levels (samples) of the identified event type was created, then a matrix of 'age vector x event types' were made with binary value (0/1) assign to each value depending on the age of the event type (the event type is present (1) if age > detected age). Finally, logical rules were applied and the binary values were adjusted for each region to get the event data for each pollen record. If there is no human event identified, it does not mean humans were absent, but that is was not possible to identify human activity from the pollen records.

Below is a summary figure for the events aggregated within different regions and ecozones using the raw event data. The different colouring display the event types with each region. The lines represent a simple binomial GAM model only for visualising the main trends of changes in different human events identified in the pollen diagrams over time in each region. The raw data is the 0/1 dots.




North America
Arid
Cold
Polar
Temperate
NA
value
1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
10000
12500
var_name
bi
es
fc


Asia
Arid
Cold
Polar
Temperate
Tropical
NA
value
1.00
0.75
0.50
0.25
0.00
0
2500
5000
7500
10000
12500
var_name
bi
ei
fc
fi



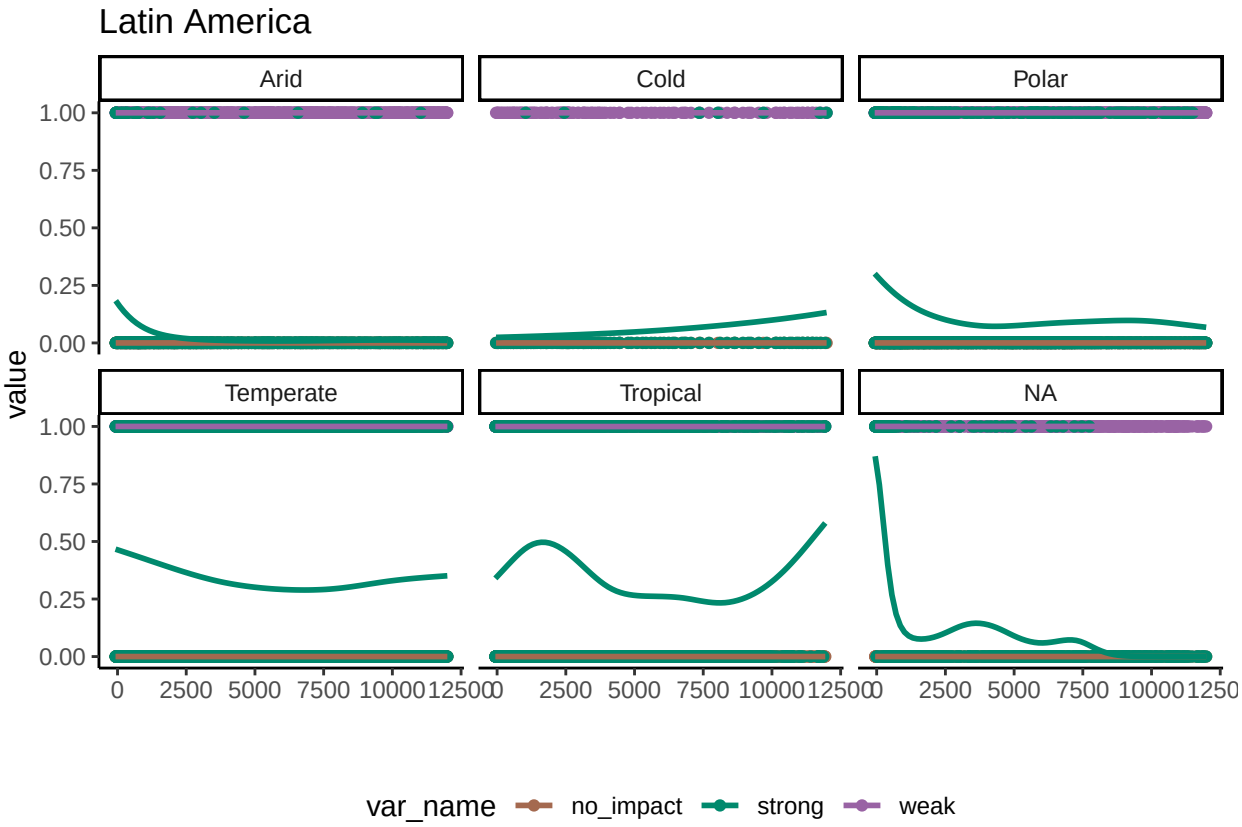
In North America, Asia, and Oceania there is an extra panel entitled `NA`. It contains a few datasets in these regions, most likely because they have not the geographical location within a defined ecozone in the shapefile in FOSSILPOL. In addition, Oceania contains also the var_name of the events type identified for Asia. This might be because some of the sites in Asia have been reassigned to the region Oceania.

Detection using indicator taxa An extensive literature review has been done to detect fossil pollen taxa that can be associated with human activity (Add the two tables of groups of indictors and single indictors?). The aggregation of indicator groups vary across the continent, both between regions and countries, and the combination of indicator groups for pollen records within the area of expertise. An algorithm has been made to assess each pollen records, which goes through the presence of taxa in combinations of the recorded indicator groups, and detect if the signal of pollen indicator is based on a defined threshold.

(Add tables with scrollbars here?)

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" (disregard of the source of the classification). Specifically, at least 1 pollen grain of indicators have to present as 'weak' and more than 1 has to present as 'strong'. In addition, the pollen-type *Pinus* was only considered as a human indicator outside of its native distribution (see XXX?). 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:



Archeological 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 require 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 filtered out as the first step. 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 (see above). Radiocarbon dates were calibrated using `calibrate` function from `rcarbon` package with apropriate 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.

In order to select distance from each pollen record, which will limit the area of human activity relevant to that pollen record, we used an expert-based detection of human events from pollen records 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 for 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

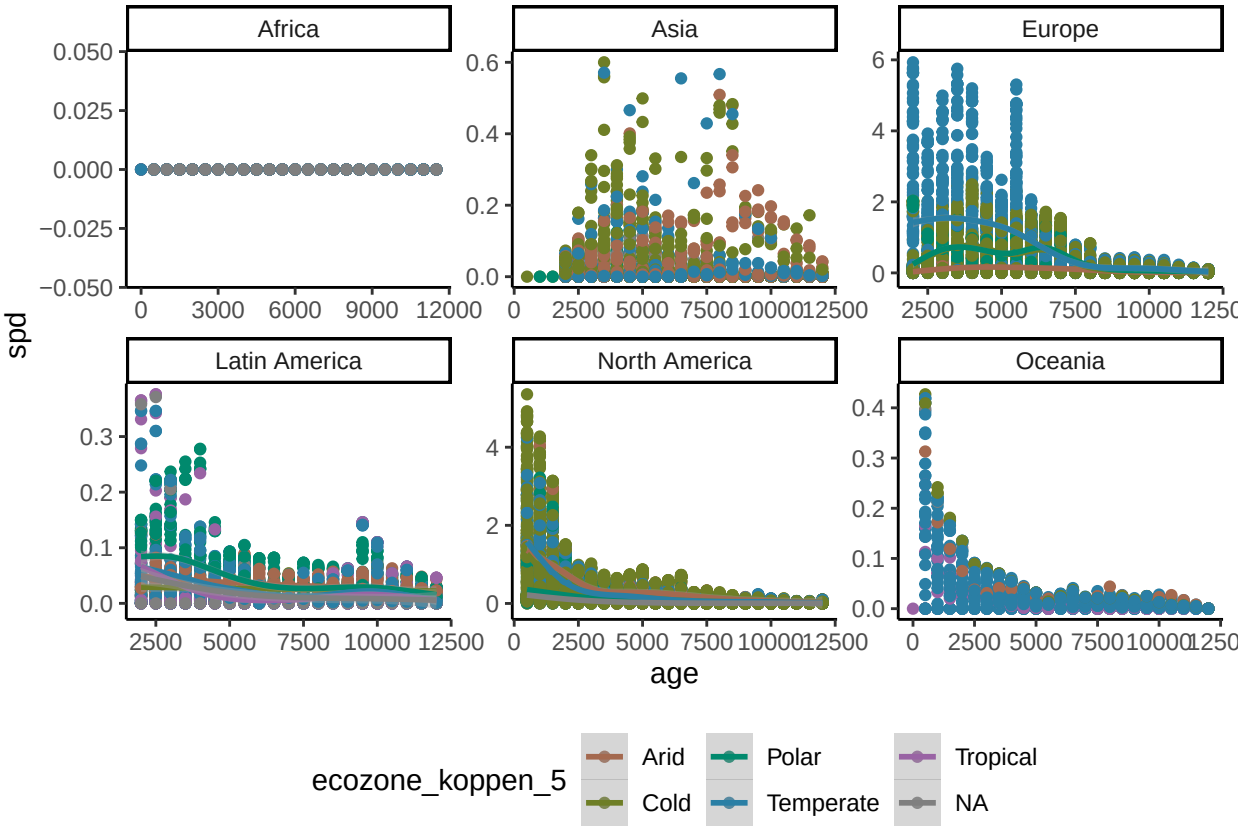
Table 3: Table 3: Minimum threshold age for different regions

region	age_from
Europe	2000
Latin America	2000
Asia	2000
Africa	2000
North America	500
Oceania	500

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 sites. See demonstration of this method Detection of past humans.

Summary figure of SPDs in time for different regions and ecozones:



As you can see from the figure of the different regions, we have no data for past human influence in Africa, so for the main analyses, this region has been filtered out.

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 and to be able to analyse vegetation changes in time due to 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 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:

$$^{q}D = (\sum_{i=1}^{S} p_i^q)^{1/(1-q)}$$

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).

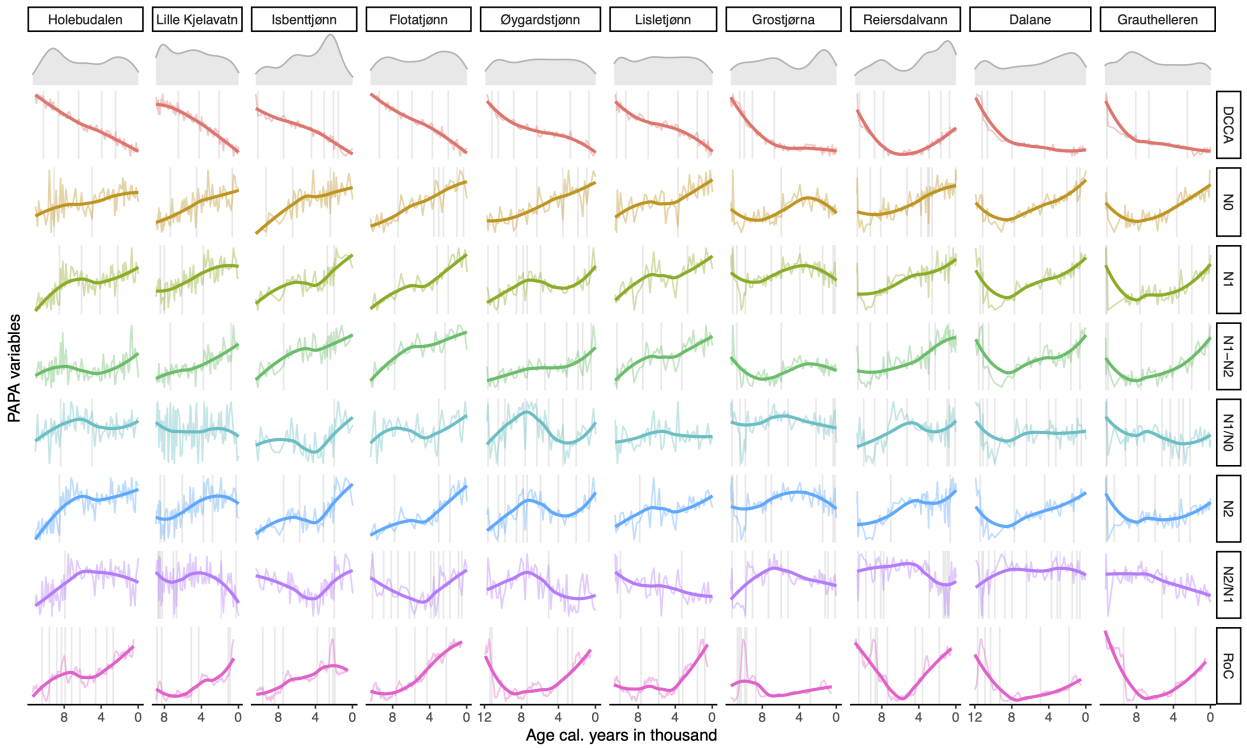
Compositional turnover Compositional turnover is estimated using detrended canonical correspondence analysis (DCCA) with age as the explanatory variable (ter Braak and Smilauer 2007?). Changes in Weighted average (WA) sample scores (CaseR scores sensu ter Braak and Smilauer 2012) are measures of compositional 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 (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 density is calculated for each of these variables using a Gaussian kernel, re-scaling them to each of the records specific age ranges (i.e. minimum and maximum ages). To solve the boundary issue in density estimation the data is reflected to 0. Hierarchical generalised additive models (HGAM) are then used to produce two different fitted variables where the first density get the common pattern of the density estimates of all the variables reflecting significant changes in richness, diversity, and evenness, and the second variable get the common pattern of the densities of the variables reflecting significant change in pollen assemblages (MRT, RoC, DCCA1).

Example of ten sites and PAP estimations:



Hierarchical variation partitioning

All response variables, except the two fitted diversity and compositional density variables, have been estimated using the harmonised pollen counts, and then the raw estimates have been linearly interpolated to get the data points on equal time spacing of 500 years. Equal time spacing is necessary for the second temporal-spatial analysis (below). Comparison of interpolation methods using either linear or generalised additive models (GAMs) showed that linear interpolations resulted in correlations between the response variables that are more similar to the correlations between the variables of the raw estimates of data. The univariate GAM models could show unexpected patterns in single variables that changed these multivariate correlations. Since we cannot individually assess single GAM models for each variable for all the sites, we choose the simplest interpolation method to keep the data as close as possible to the raw estimates before using them in the multivariate models. In addition, we remove Africa as a region, because as you can see from section about *past human detection* above, we do not have any data of *past human presence*.

To test if ecological processes have changed due to past human activity within cores, 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 estimate 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 we are dealing with a mix of numerical responses variables. These are the different PAPs with varying units. Depending on the type of analyses (see below), the explanatory variables are SPDs and palaeoclimatic variables. For within core analysis (see below) we also include time as an explanatory variable. SPD is then the important 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 differences in climatic conditions within all the regions (regarding temperatures, seasonality and aridity). Time is represented by the ages of each pollen record, however, this is more difficult to interpret. We assume age may represent time dependent vegetation changes such as natural successions and/or ecological changes due to interaction between taxa.

To do this, we have coded two functions one generalised called `get_varhp` and a specialised function that perform the analysis on our dataset called `run_hvar`. The parameters `data_source` is a combined pre-

structure `tibble` that take the data stored in the `data_merge` column. This contain all preprepared variables for each dataset that will be input into the analysis. Because this contain all variables, there are two parameters `response_vars` and `predictor_vars` where you can select the variables you want to include as responses and predictors in the hierarchical variation partitioning analysis. The predictor variables can be applied either as individual predictors or as groups as 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. (Though the overall results does not change much). The advantage using palaeoclimate as single predictors is that it is possible to get which of climatic variables are important (high adj R2) or statistically significant, if significant testing is turned on. Statistical testing can be applied in two ways, both which shuffle the predictor variables. Using `time_series` set as TRUE, statistical testing is run 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, predictor variables will be randomly shuffled as many times set in the parameter `permutations` (hence it use the default permutation in the `rdcca.hp` package).

The analysis is run in two different ways to analyse: - 1) to analyse *spatial changes* which run the hierarchical variation partitioning within single cores or site, and - 2) to analyse the *temporal patterns* in space for each of the 500 year time steps. In this latter analysis, we restructure the data as samples across space within selected Koppen-Geiger division of five major ecozones on each continent, whereas the analysis is run per time bin. The predictor groups are past human presence and the matrix of palaeoclimatic variables. For this, we filter out time bins which have less than 5 samples. For some bins, if all the spds equal zero, the analysis will fail. These cases indicate insufficient numbers of predictors and will return NA for these specific time bins.

Part II: Results

The results of each dataset/subset from which the multivariate model is run provide a list of outputs which contain:

- a vector of total explained variation
- the full output of the hierarchical variation partitioning analysis (figure below display the overall results of unique variation in fractions and percentage, but the it contains a list of all the output)
- a summary table with the results of the hierarchical variation partitioning showing the unique, average shared and individual variation (individual variation = if the model is run with only that specific predictor), and the percentage of the individual variation. Our main interest is the predictor human. When this is zero it means this variable was not present in the subset, so we added the zero values as human predictor would not be able to explain any variation.
- a vector with summmary variation, which is the total eigenvalue, constrained eigenvalue, and unconstrained eigenvalue from a model (without variation partitioning) using db-RDA with Gower distances including all predictors. The reason for this, was because the variation partitioning focus on explained variation, and the idea was that this then add some potential extra information about the unconstrained variation.

Visual example from the output of a model:

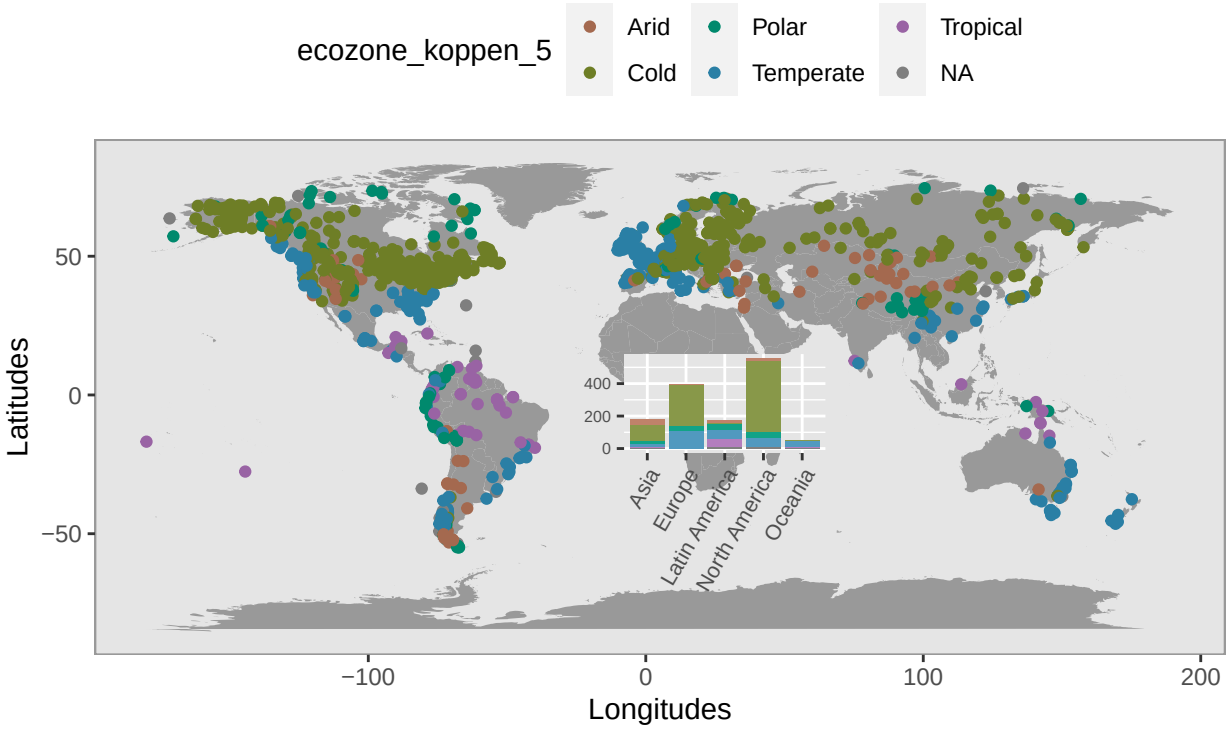
```
> data_hvar$varhp[[1]]
$varhp_output
$Method_Type
[1] "dbRDA" "adjR2"

$Total_explained_variation
[1] 0.228

$Var.part
                                  Fractions   % Total
Unique to human                     -0.0177     -7.79
Unique to climate                    0.0869     38.14
Unique to time                      -0.0130     -5.70
Common to human, and climate         0.0103      4.50
Common to human, and time            0.0167      7.34
Common to climate, and time          0.1080     47.39
Common to human, climate, and time   0.0367     16.11
Total                                0.2279    100.00

$Hier.part
         Unique Average.share Individual I.perc(%)
human   -0.0177        0.0257     0.0080      3.51
climate  0.0869        0.0714     0.1583     69.43
time    -0.0130        0.0746     0.0616     27.02

attr(,"class")
[1] "rdaccahp"

$summary_table
# A tibble: 3 × 6
  predictor  Unique Average.share Individual `I.perc(%)` `Pr(>I)`
  <chr>       <dbl>         <dbl>      <dbl>       <dbl> <chr>
1 human     -0.0177        0.0257     0.008         3.51 "0.5    "
2 climate    0.0869        0.0714     0.158        69.4  "0.14   "
3 time      -0.013         0.0746     0.0616       27.0  "0.59   "

$summary_variation
# A tibble: 1 × 3
  Total_eig Constrained_eig Unconstrained_eig
      <dbl>           <dbl>             <dbl>
1      2.25            1.03              1.22
```

The output is similar for the second *temporal* analysis, though only human and climate are the predictors within each timebin. The subset of this analysis is aggregated samples within ecozones across the different regions. Records from Africa have been filtered out.

The example output is included for you to have an example of what information there is, and to help you

think about what is relevant and interesting to summarise in this study beyond the examples made further in this result part.

In total there are 1350 datasets/sites after sites from Africa have been filter out (represent 20 lost sites). The distribution of sites across ecozones and region is shown in figure 1. In addition, there are 20 datasets within Asia, Latin America, Oceania, and North America that has missing (NA) values for ecozones, meaning they have not been assigned to an ecozone, and therefore not included in the summary results for simplicity. This has been fixed in FOSSILPOL, and need to be updated in our dataset. The preliminary results are then including the 1330 remaining sites.



Summary results for the spatial changes (within core analysis)

Figure 2 - violin plots of total explained variation per region across ecozones Figure 3 - pie charts of uniely percentage explained variation of total constrained
Figure 3 - uniquely explained variation of past human impact across the world - maps Figure 4 -