

ArchaeoGLOBE trend analysis

Nick Gauthier

Last knit on: 25 January, 2019

Sample analysis code for the ArchaeoGlobe database. Here we use Generalized Additive Models (GAMs), a flexible form of nonlinear regression model capable of fitting smooth, time-varying trends to the ordered categorical ArchaeoGLOBE response data.

We model ordered categorical data using a latent variable following a logistic distribution. The model identifies a series of cut points, which correspond to the probabilities of the latent variable falling within each of our categories.

We fit two sets of trends. One trend is fitted to all the data simultaneously, representing the global trend across all archaeological regions. Then we fit region-level trends, which represent the deviation of each region from the global trend. By penalizing the “wiggleness” of the trend lines, we allow regional trends that don’t significantly deviate from the global trend to be penalized to 0, effectively reducing that particular region to the global trend. This is a form of partial pooling, allowing the model to share information between groups and in so doing make the results less sensitive to regions with exceptionally low response rates.

After fitting the model, we can extract the region-specific deviations from the global trend, use a k-means clustering algorithm to group together regions with similar trends, and map the results. We repeat this analysis for both self-reported expertise and perceived data quality.

Setup

Import packages needed for analysis. We’ll use packages from the `tidyverse`, such as `readr`, `dplyr`, and `ggplot2` for data import, processing, and plotting. We’ll also use `mgcv` for fitting nonlinear trends to the data. We’ll use the `sf` package to help us plot shapefiles in a tidy context. Finally, we’ll use `patchwork` to combine multiple ggplots in the same image.

```
library(tidyverse)
library(mgcv)
library(sf)

# install patchwork from github if needed
# devtools::install_github('thomasp85/patchwork')
library(patchwork)
```

Data import

Read in the latest version of the ArchaeoGLOBE database and the regions shapefile.

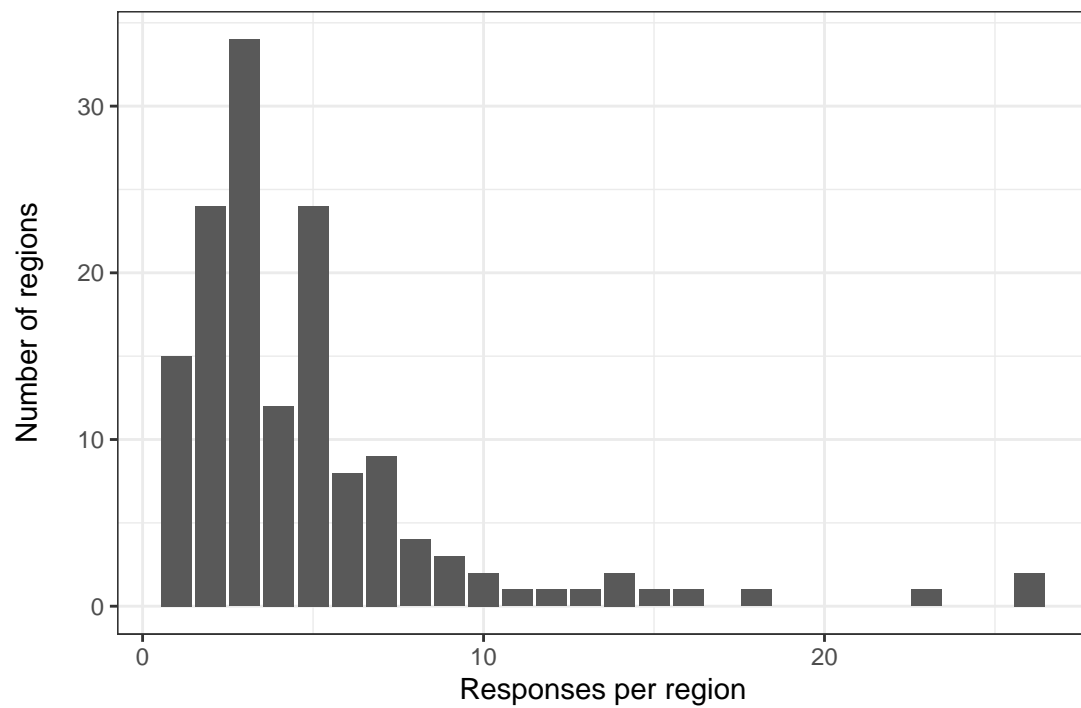
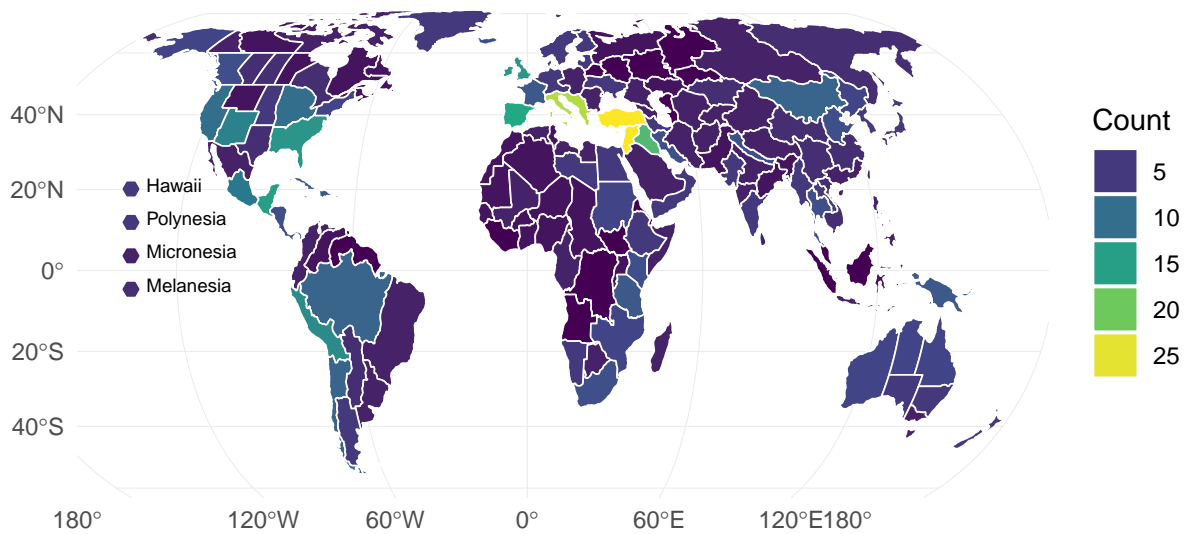
```
archaeoglobe <- read_csv('data/ARCHAEOGLOBE_PUBLIC_DATA.csv')
regions <- st_read('data/Simplified_Regions2.shp', quiet = TRUE)
```

Exploratory plots

Before running any analyses, let’s look at the data. How many responses do we have per region?

```
response_counts <- archaeoglobe %>%
  group_by(REGION_ID) %>%
  count
```

Total responses per region



Analysis functions

Define some analysis functions that we'll be using repeatedly in the analysis, so that we don't have to keep copying and pasting the same lines of code.

This function subsets the data to highlight a variable of interest, and converts it from a wide to a long "tidy" format to make analysis and plotting easier.

```
preprocess <- function(prefix, categories){
  archaeoglobe %>% # start with the full ArcheoGlobe data
    # drop columns not related to the variable of interest
    select(c(CONTRIBUTR:LAND_AREA, starts_with(prefix))) %>%
    gather(time, value, starts_with(prefix)) %>% # one value per row
    mutate(time = parse_number(time) * -1, # convert time period labels to years
           value = ordered(value, levels = categories),
           cat_num = as.numeric(value)) %>%
    mutate_if(is.character, as.factor) # convert characters to factors
}
```

This function takes a data frame produced by the above function and fits GAM to the global trend and local deviations for each region, accounting for inter-observer variability. This function takes as arguments a preprocessed data frame containing time slices, regions, contributors, and the ordered categorical response variable transformed to a numeric vector.

```
cores <- max(parallel::detectCores() / 2, 1) # physical cores for parallelization
cl <- parallel::makeCluster(cores)
```

```
fit_gam <- function(x, n_cats){
  bam(cat_num ~
    # this spline is for the global trend
    s(time, bs = 'cr', m = 2) +
    # region-specific trends. bs = 'ts' and m = 1
    # help penalize deviation from the global model
    s(time, by = REGION_LAB, bs = 'cs', m = 1) +
    # add back in region-specific intercepts
    REGION_LAB +
    # model contributor as a random effect
    s(CONTRIBUTR, bs = 're'),
    data = x, # data frame to analyze
    family = ocat(R = n_cats), # ordered categorical with n levels
    # final 3 arguments just speed up the model fitting
    method = 'fREML',
    discrete = TRUE,
    cluster = cl)
}
```

This function extracts the estimated trends for each region, incorporating the global and regional splines as well as the region and contributor specific intercepts. Then it clusters these trends into 6 discrete clusters using k-means. The choice of 6 clusters is somewhat arbitrary, and is made simply based on visual comparisons of different cluster solutions with the goal of ensuring visually interpretable results.

```
extract_trends <-function(mod, n_clusters = 6){
  set.seed(1000) # set seed for reproducability of clusters
```

```

archaeoglobe %>% # create dummy data for prediction in the following lines
  select(REGION_LAB) %>%
  group_by(REGION_LAB) %>%
  slice(1) %>%
  slice(rep(1:n(), each = 198)) %>%
  ungroup %>%
  mutate(time = rep_len(seq(-10000, -150, 50), n()),
         CONTRIBUTR = 'CYRBU') %>% # select an arbitrary contributor
  mutate(preds = predict(mod, .)) %>% # estimate trend lines
  mutate(preds = plogis(preds)) %>% # transform responses to [0,1] scale
  spread(time, preds) %>%
  # next is the actual kmeans clustering code
  mutate(cluster = kmeans(., -c(1,2)), n_clusters, iter.max = 100, nstart = 100)$cluster)
}

```

Analysis

Now we use the functions defined above on the ArchaeoGlobe data. For convenience, first define a data frame that lists the prefixes of the variables we are interested in (e.g. “EXP” for expertise) and the levels of the ordered factors associated with each variable. This will make it easier to quickly focus on a specific variable. The `tribble` command is simply a way to make a data frame by row rather than column, which makes the code easier to read.

```

response_levels <- tribble(
  ~prefix, ~categories,
  'EXP', c('None', 'Low', 'High'),
  'DQ', c('Unknown', 'Low', 'Moderate', 'Good'),
  'HUNT', c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'),
  'EXAG', c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'),
  'INAG', c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'),
  'PAST', c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'),
  'URBN', c('Absent', 'Present')
)

```

Now map each of the above functions to each variable. This allows us to run the analysis for all variables of interest in a single step, and save all the outputs in a tibble format for easy plotting. This will take a long time, so the results are cached by default for future use. If you’re running this for the first time, it should take about 40 minutes to run on a Intel NUC with a 5th-gen Intel Core i7 processor and 16gb of RAM running Linux.

```

trend_dat <- response_levels %>%
  mutate(data = map2(prefix, categories, ~preprocess(.x,.y)),
         n_cats = map_dbl(categories, length),
         mod = map2(data, n_cats, fit_gam),
         trends = map(mod, extract_trends))

```

Results

First we plot out the global trends for each land use type. Then we plot out the local (regional trends) for all land use types, and map out their associated clusters. Please refer to the `.rmd` source file for the code to make the plots.

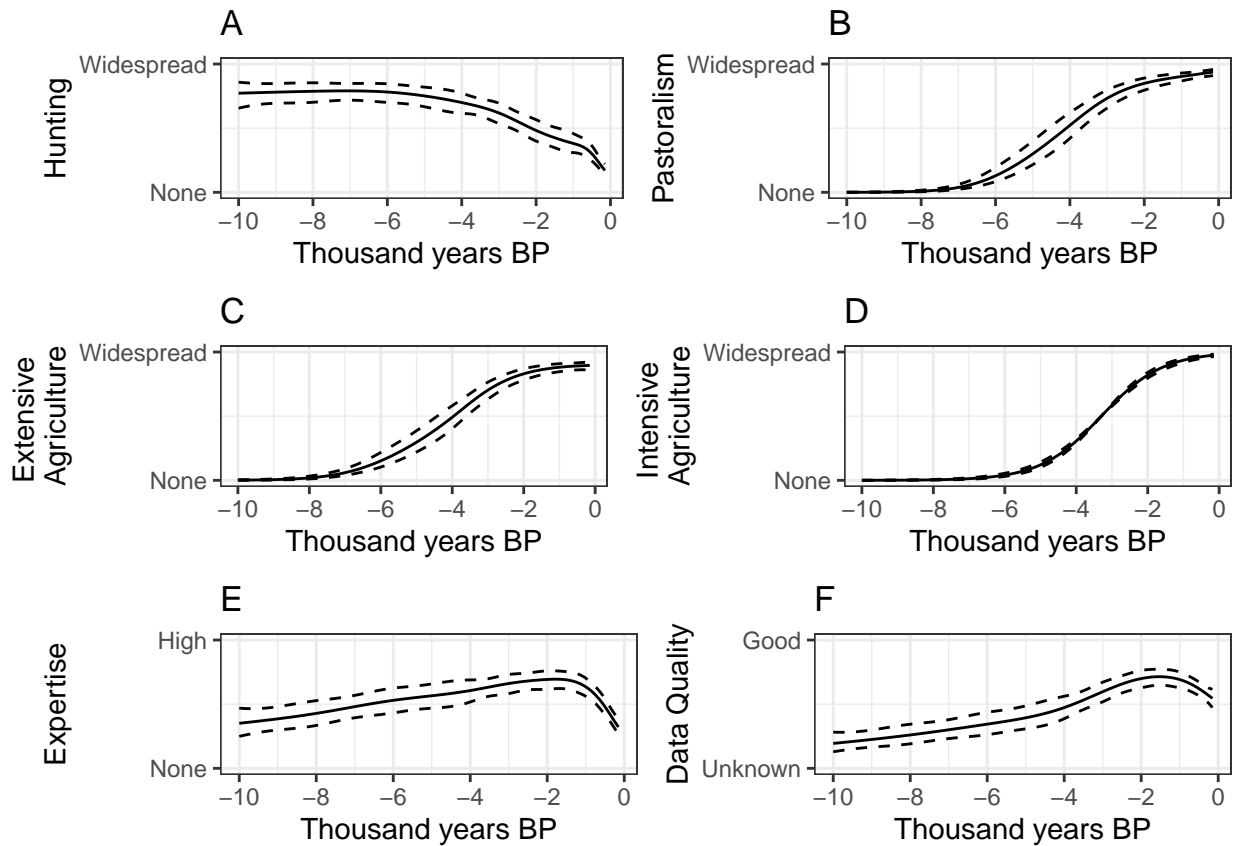
The global trend in hunting shows constant high prevalence until around 6,000 years ago, after which there is a smooth decline until the present day when it is very rare. Mapping out the clusters reveals a clear east-west divide, which regions in Afro- Eurasia seeing hunting earlier then the global mean, and regions in the Americas and Oceania seeing later peaks in hunting.

The global trends in the prevalence of pastoralism, extensive and intensive agriculture, and urbanism all follow a sigmoidal curve, which means the trend is linear on the scale of the linear predictor (the ordered categorical GAM uses a logit transform as a latent link function). This means that there is a simple increase in the probability of each land use type being prevalent over time.

How does self-professed level of expertise vary in each region over time? The global trend is a roughly linear increase in self-reported expertise from 10ka BP up to 2ka BP, then a falloff continuing to the present day. The present day expertise values are approximately the same as at 10ka BP. This makes sense, as it points to both the increased frequency of preserved archaeological materials with time as well as the reduction in archaeological attention in periods with extensive historical records.

The global trend in data quality is more or less the same as the expertise data, with the peak in data quality occurring more recently than for expertise and with a less dramatic falloff leading to the present day. Unlike expertise, which reaches the same values at 10ky BP and present, data quality in the present day remains high in spite of the falloff in the last 2 millennia. Also note the confidence interval for the global trend is generally wider than for the expertise responses.

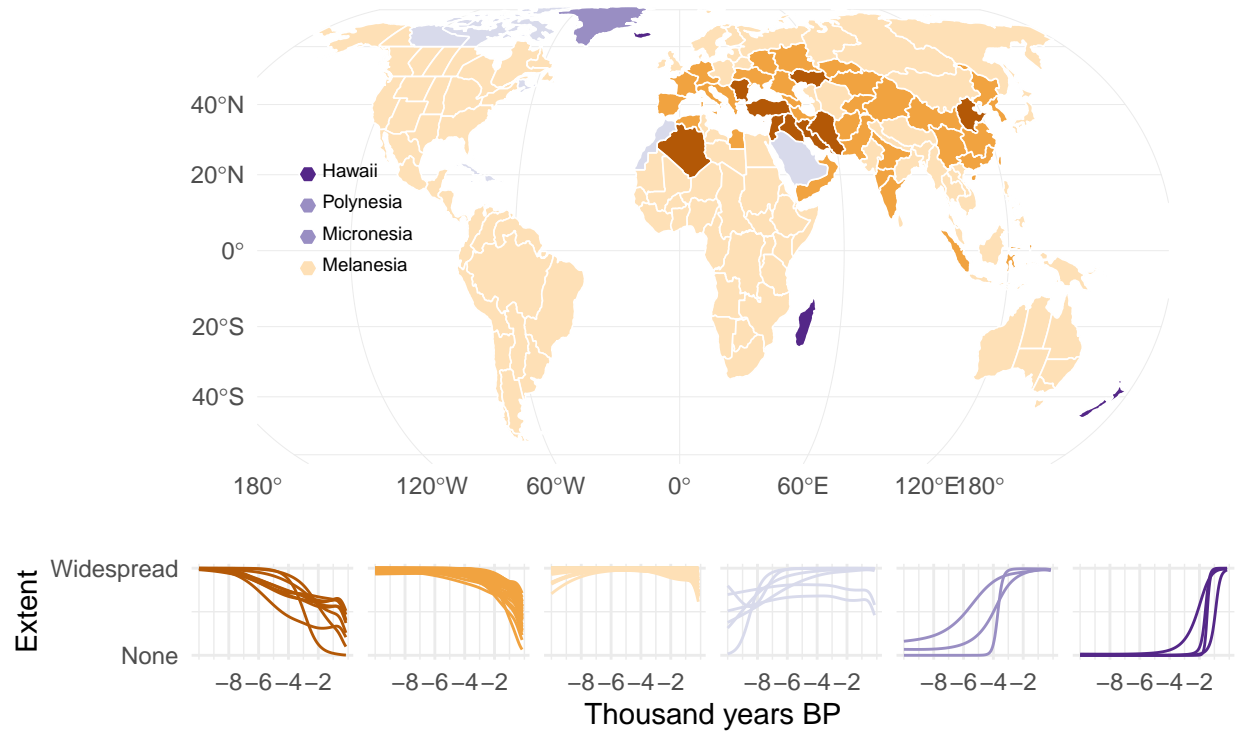
Global Trends



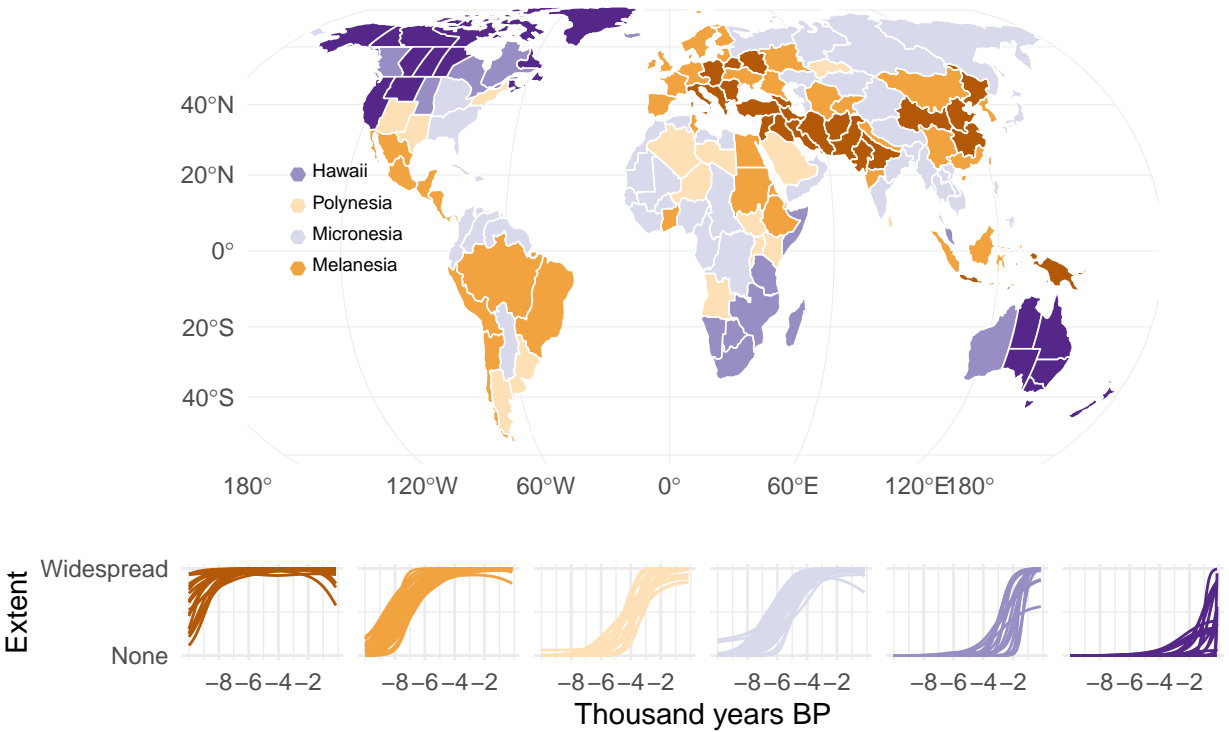
Regional Trends

Regional land-use trends

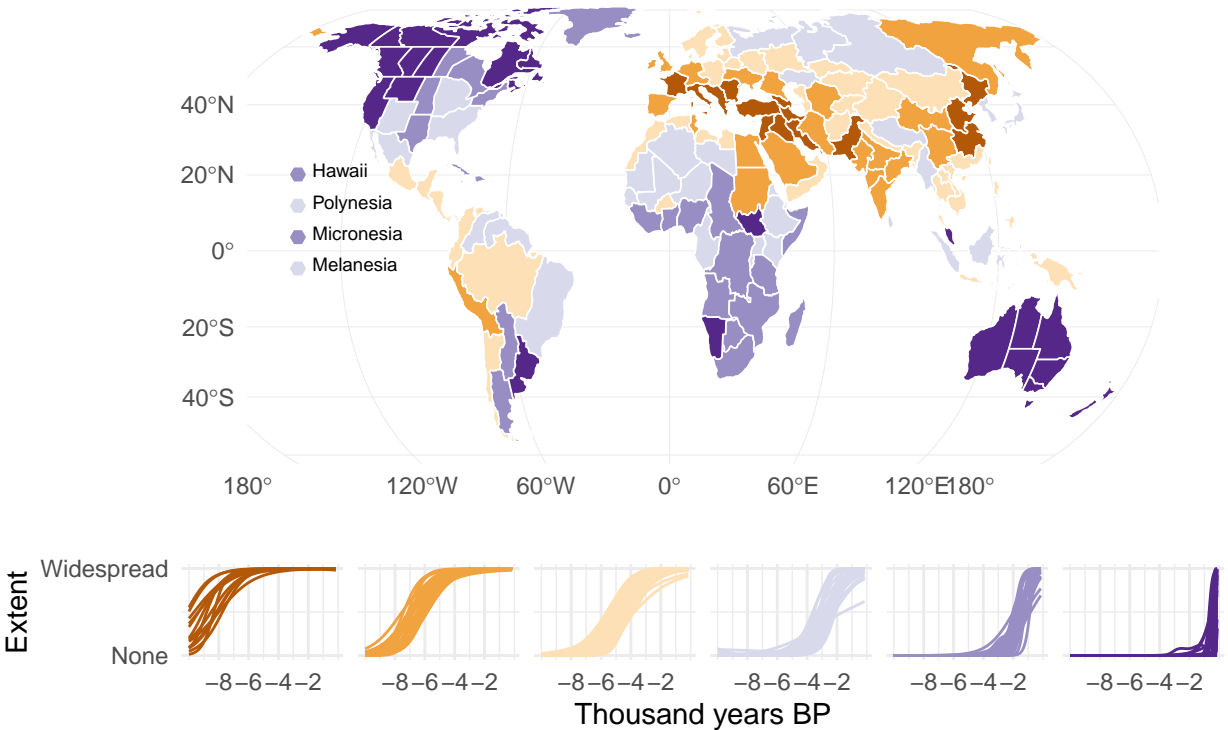
Hunting



Regional land–use trends
Extensive Agriculture

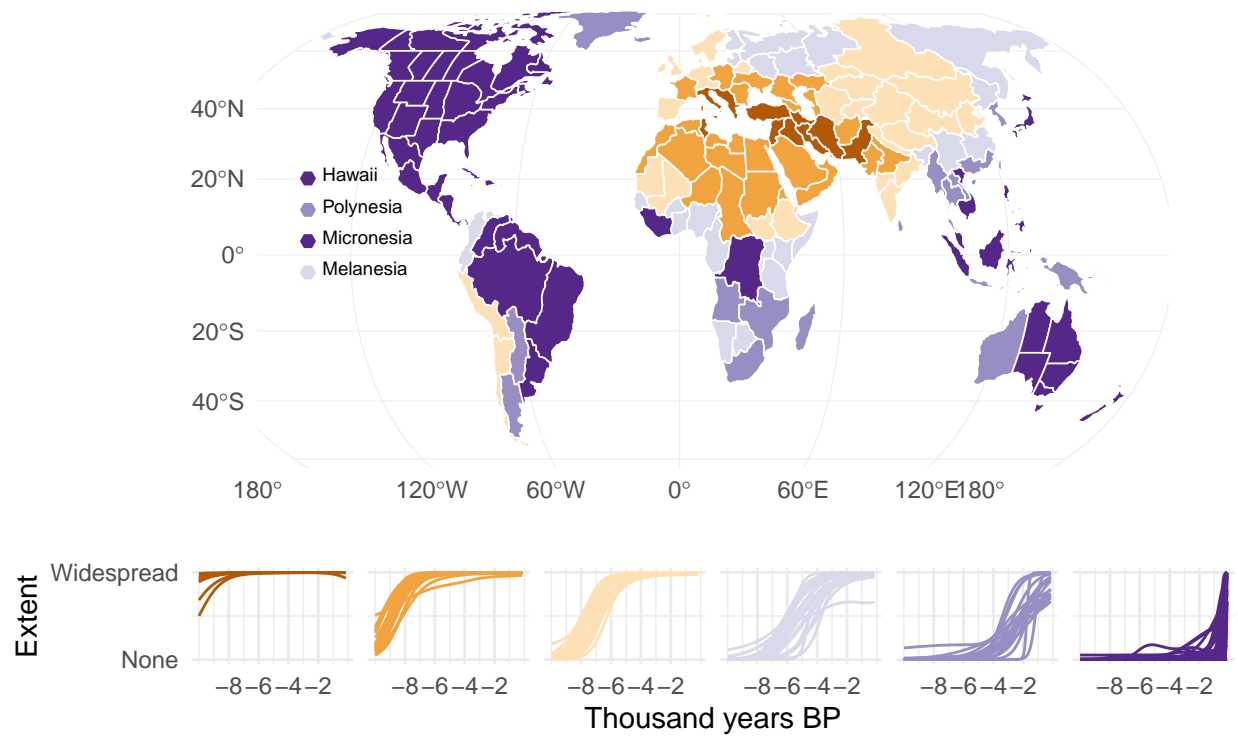


Regional land–use trends
Intensive Agriculture



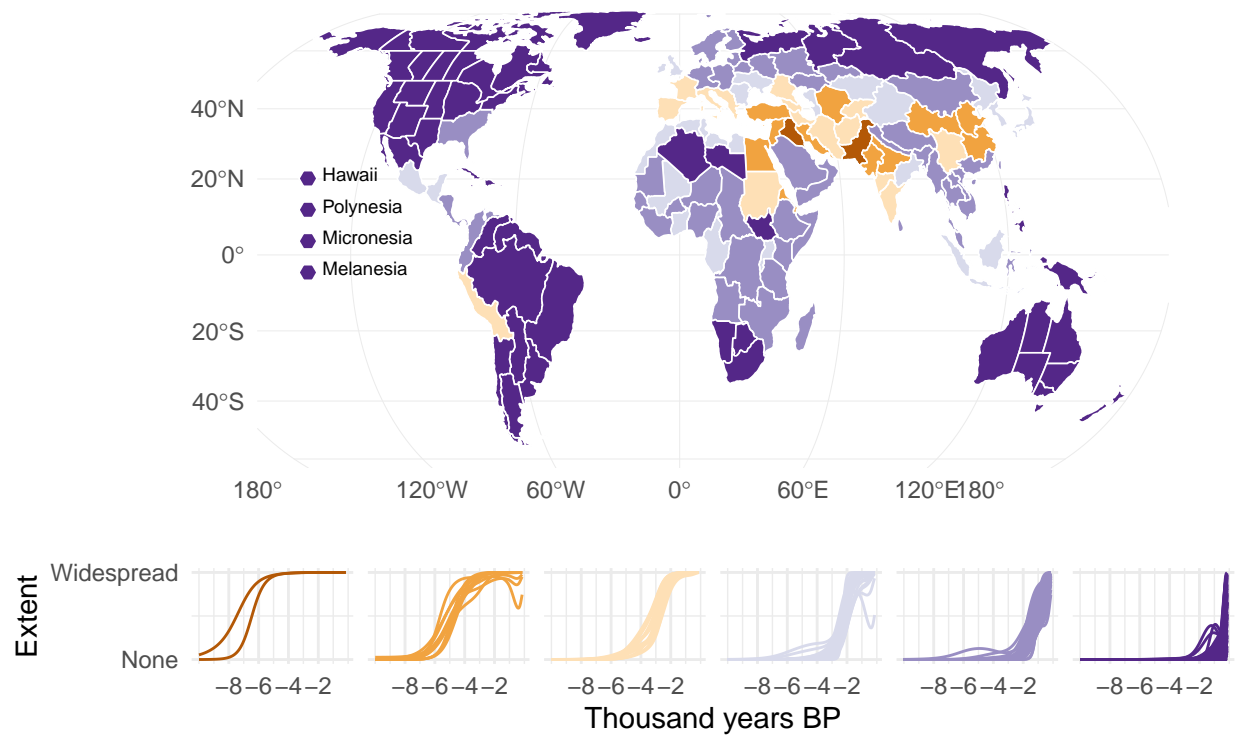
Regional land–use trends

Pastoralism



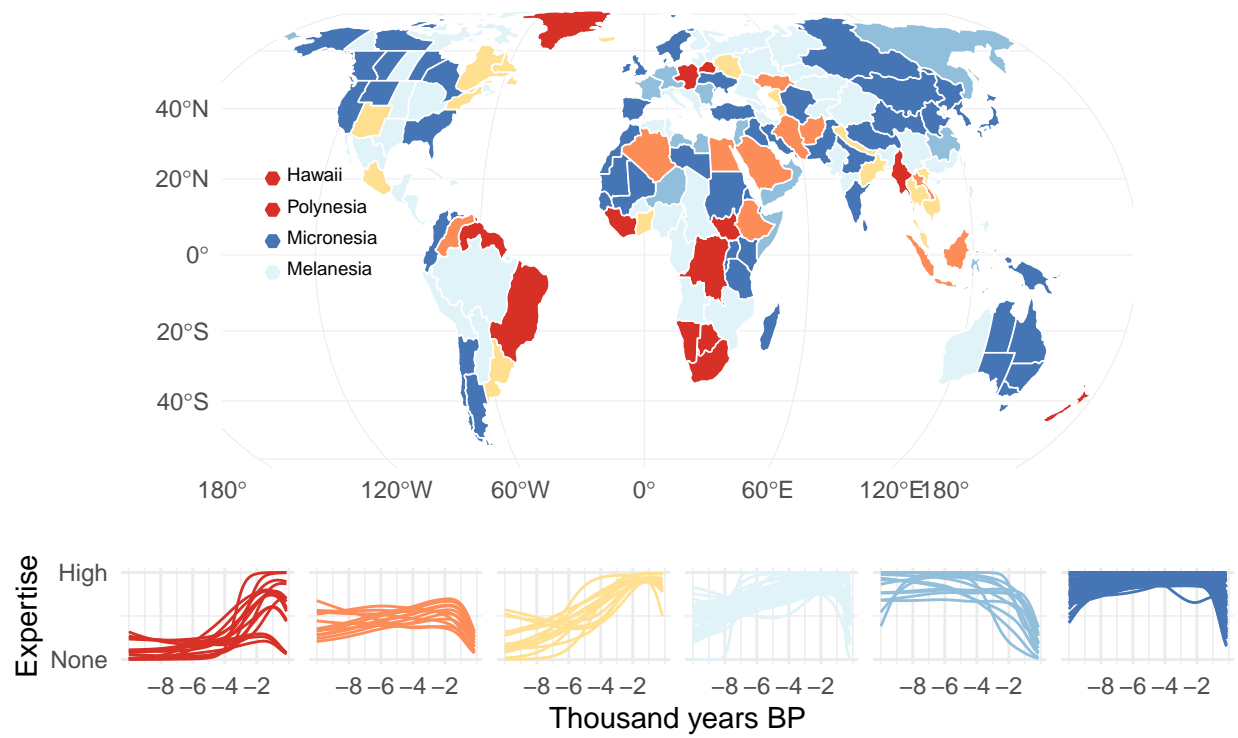
Regional land-use trends

Urbanism



Archaeological trends

Expertise



Archaeological trends

Data Quality

