# ArchaeoGLOBE trend analysis

*Nick Gauthier*

*Last knit on: 08 November, 2018*

Sample analysis code for the ArchaeoGlobe database. Here we fit 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 the 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 "wiggliness" 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 diviations from the global trend, use a k-means clustering alogirithm 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(vows)
library(sf)

#install patchwork from github
#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/Survey_scrubbed_Aug20_IDs.csv')
regions <- st_read('data/Simplified_Regions2.shp', quiet = TRUE)
```

## 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

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 analyize
      family = ocat(R = n_cats), # ordered categorical with n levels
      # final 3 arguments just speed up the model fitting
      method = 'fREML',
      discrete = TRUE,
      nthreads = cores)
}
```

This function extracts the fitted splines for each region, ignoring factors such as the global trend and region and contributor specific intercepts so that the focus is on the shape of the local trends. Then it clusters these local deviations from the global trend into discrete clusters.

```
extract_trends <-function(mod, n_clusters = 6){
  set.seed(1000)
  archaeoglobe %>%
    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') %>%
    mutate(preds = predict(mod, .)) %>%
      mutate(preds = plogis(preds)) %>%
  spread(time, preds) %>%
  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 goal, 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.

```
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))
```

```
trend_dat %>%
  filter(prefix %in% c('HUNT','EXAG','INAG','PAST')) %>%
  select(prefix, data) %>%
  unnest
```

```
## # A tibble: 28,280 x 11
##    prefix CONTRIBUTR WORLD_ID WORLD_LAB REGION_ID REGION_LAB TOT_AREA
##    <chr>  <fct>         <int> <fct>         <int> <fct>         <dbl>
##  1 HUNT   CQIIK             1 Northern~         1 Alaska      1499260
##  2 HUNT   CYRBU             1 Northern~         1 Alaska      1499260
```

```
##  3 HUNT   CMNXE             1 Northern~         1 Alaska      1499260
##  4 HUNT   CDTDK             1 Northern~         1 Alaska      1499260
##  5 HUNT   CYQUC             1 Northern~         1 Alaska      1499260
##  6 HUNT   CIGSJ             1 Northern~         1 Alaska      1499260
##  7 HUNT   CYRBU             1 Northern~         2 Yukon Ter~   482548
##  8 HUNT   CHBZL             1 Northern~         2 Yukon Ter~   482548
##  9 HUNT   CIGSJ             1 Northern~         2 Yukon Ter~   482548
## 10 HUNT   CYRBU             1 Northern~         3 Northwest~  1342540
## # ... with 28,270 more rows, and 4 more variables: LAND_AREA <dbl>,
## #   time <dbl>, value <ord>, cat_num <dbl>
```

```
library(psych)
archaeoglobe %>%
  gather(key, value, EXP_10000: URBN_00150) %>%#HUNT_10000:PAST_00150) %>%
  separate(key, into = c('type', 'time')) %>%
  spread(type, value) %>%
  select(DQ:URBN) %>%
  mutate(DQ = ordered(DQ, levels = c('Unknown', 'Low', 'Moderate', 'Good')),
         EXP = ordered(EXP, levels = c('None', 'Low', 'High')),
         EXAG = ordered(EXAG, levels = c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'
         HUNT = ordered(HUNT, levels = c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'
         INAG = ordered(INAG, levels = c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'
         PAST = ordered(PAST, levels = c('none', 'minimal (<1%)', 'common (1-20%)', 'widespread (>20%)'
         URBN = ordered(URBN, levels = c('Absent', 'Present'))) %>%
  select(-c(DQ,EXP)) %>%
  mutate_all(as.numeric) %>%
  as.matrix %>%
  fa(nfactors = 1, rotate = 'none',cor = 'poly')
```

```
## Warning in matpLower(x, nvar, gminx, gmaxx, gminy, gmaxy): 4 cells were
## adjusted for 0 values using the correction for continuity. Examine your
## data carefully.
```

```
## Factor Analysis using method =  minres
## Call: fa(r = ., nfactors = 1, rotate = "none", cor = "poly")
## Standardized loadings (pattern matrix) based upon correlation matrix
##        MR1   h2    u2 com
## EXAG  0.74 0.55 0.446   1
## HUNT -0.48 0.23 0.771   1
## INAG  0.94 0.89 0.108   1
## PAST  0.83 0.68 0.318   1
## URBN  0.95 0.91 0.092   1
##
##                 MR1
## SS loadings    3.26
## Proportion Var 0.65
##
## Mean item complexity =  1
## Test of the hypothesis that 1 factor is sufficient.
##
## The degrees of freedom for the null model are  10  and the objective function was  3.71 with Chi Squa
## The degrees of freedom for the model are 5  and the objective function was  0.1
##
```

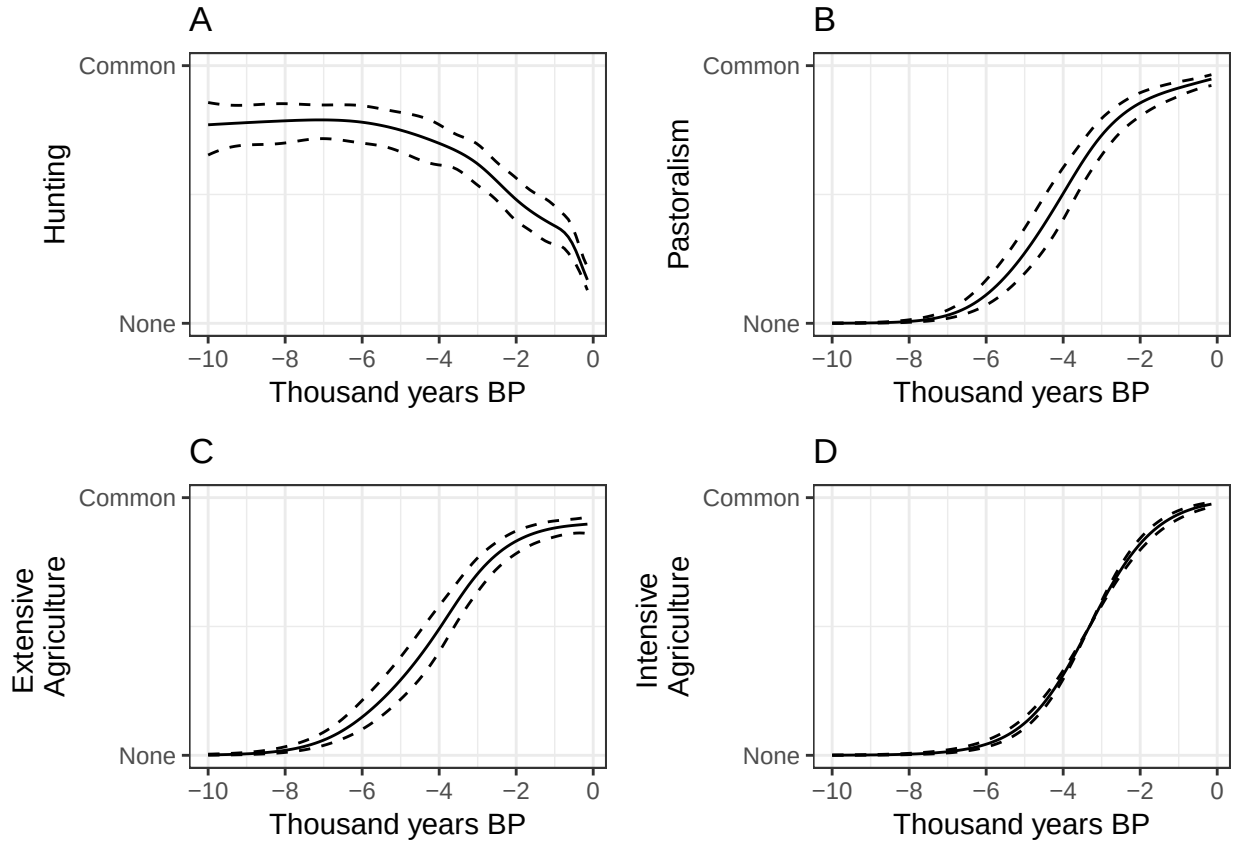
```
## The root mean square of the residuals (RMSR) is  0.04
## The df corrected root mean square of the residuals is  0.06
##
## The harmonic number of observations is  7070 with the empirical chi square  245.31  with prob <  5.6
51
## The total number of observations was  7070  with Likelihood Chi Square =  734.03  with prob <  2.1e-
156
##
## Tucker Lewis Index of factoring reliability =  0.944
## RMSEA index =  0.144  and the 90 % confidence intervals are  0.135 0.152
## BIC =  689.71
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                 MR1
## Correlation of (regression) scores with factors   0.98
## Multiple R square of scores with factors          0.96
## Minimum correlation of possible factor scores     0.92
```

## Results

```
t1 <- trend_dat[3,]$mod[[1]] %>%
    plot(select = 0) %>% # selects the global trend
    .[1] %>%
    map(~tibble(time = .$x, fit = c(.$fit), se = .$se)) %>%
    .[[1]] %>%
    ggplot(aes(time / 1000, plogis(fit)))+
    geom_line() +
    geom_line(aes(y = plogis(fit + 2 * se)), linetype = 2) +
    geom_line(aes(y = plogis(fit - 2 * se)), linetype = 2) +
      scale_x_continuous(breaks = c(-10, -8,-6,-4,-2,0)) +
    scale_y_continuous(breaks = c(0,1), labels = c('None', 'Common'), limits = c(0,1)) +
    labs(title = 'A', x = 'Thousand years BP', y = 'Hunting') +
    theme_bw()
t2 <- trend_dat[6,]$mod[[1]] %>%
    plot(select = 0) %>% # selects the global trend
    .[1] %>%
    map(~tibble(time = .$x, fit = c(.$fit), se = .$se)) %>%
    .[[1]] %>%
    ggplot(aes(time / 1000, plogis(fit)))+
    geom_line() +
    geom_line(aes(y = plogis(fit + 2 * se)), linetype = 2) +
    geom_line(aes(y = plogis(fit - 2 * se)), linetype = 2) +
      scale_x_continuous(breaks = c(-10, -8,-6,-4,-2,0)) +
    scale_y_continuous(breaks = c(0,1), labels = c('None', 'Common'), limits = c(0,1)) +
    labs(title = 'B', x = 'Thousand years BP', y = 'Pastoralism') +
    theme_bw()
t3 <- trend_dat[4,]$mod[[1]] %>%
    plot(select = 0) %>% # selects the global trend
    .[1] %>%
    map(~tibble(time = .$x, fit = c(.$fit), se = .$se)) %>%
    .[[1]] %>%
    ggplot(aes(time / 1000, plogis(fit)))+
    geom_line() +
```

```
    geom_line(aes(y = plogis(fit + 2 * se)), linetype = 2) +
    geom_line(aes(y = plogis(fit - 2 * se)), linetype = 2) +
      scale_x_continuous(breaks = c(-10, -8,-6,-4,-2,0)) +
    scale_y_continuous(breaks = c(0,1), labels = c('None', 'Common'), limits = c(0,1)) +
    labs(title = 'C', x = 'Thousand years BP', y = 'Extensive \nAgriculture') +
    theme_bw()
t4 <- trend_dat[5,]$mod[[1]] %>%
    plot(select = 0) %>% # selects the global trend
    .[1] %>%
    map(~tibble(time = .$x, fit = c(.$fit), se = .$se)) %>%
    .[[1]] %>%
    ggplot(aes(time / 1000, plogis(fit)))+
    geom_line() +
    geom_line(aes(y = plogis(fit + 2 * se)), linetype = 2) +
    geom_line(aes(y = plogis(fit - 2 * se)), linetype = 2) +
      scale_x_continuous(breaks = c(-10, -8,-6,-4,-2,0)) +
    scale_y_continuous(breaks = c(0,1), labels = c('None', 'Common'), limits = c(0,1)) +
    labs(title = 'D', x = 'Thousand years BP', y = 'Intensive \nAgriculture') +
    theme_bw()

t1+t2+t3+t4
```




```
ggsave('figures/global_trends.png',height = 8, width = 12)
```

## Hunting

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.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
trend_dat[3, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                      cluster = recode_factor(cluster,
                                      `6` = '1', `5` = '2',`1` = '3',
                                      `2` = '4', `3` = '5', `4` = '6')))) %>%
                                      #`4` = '1', `5` = '2',`6` = '3',
                                      #`3` = '4', `1` = '5', `2` = '6')))) %>%
plot_trends2('Hunting')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning in xList[i] <- valueList: number of items to replace is not a
## multiple of replacement length
```

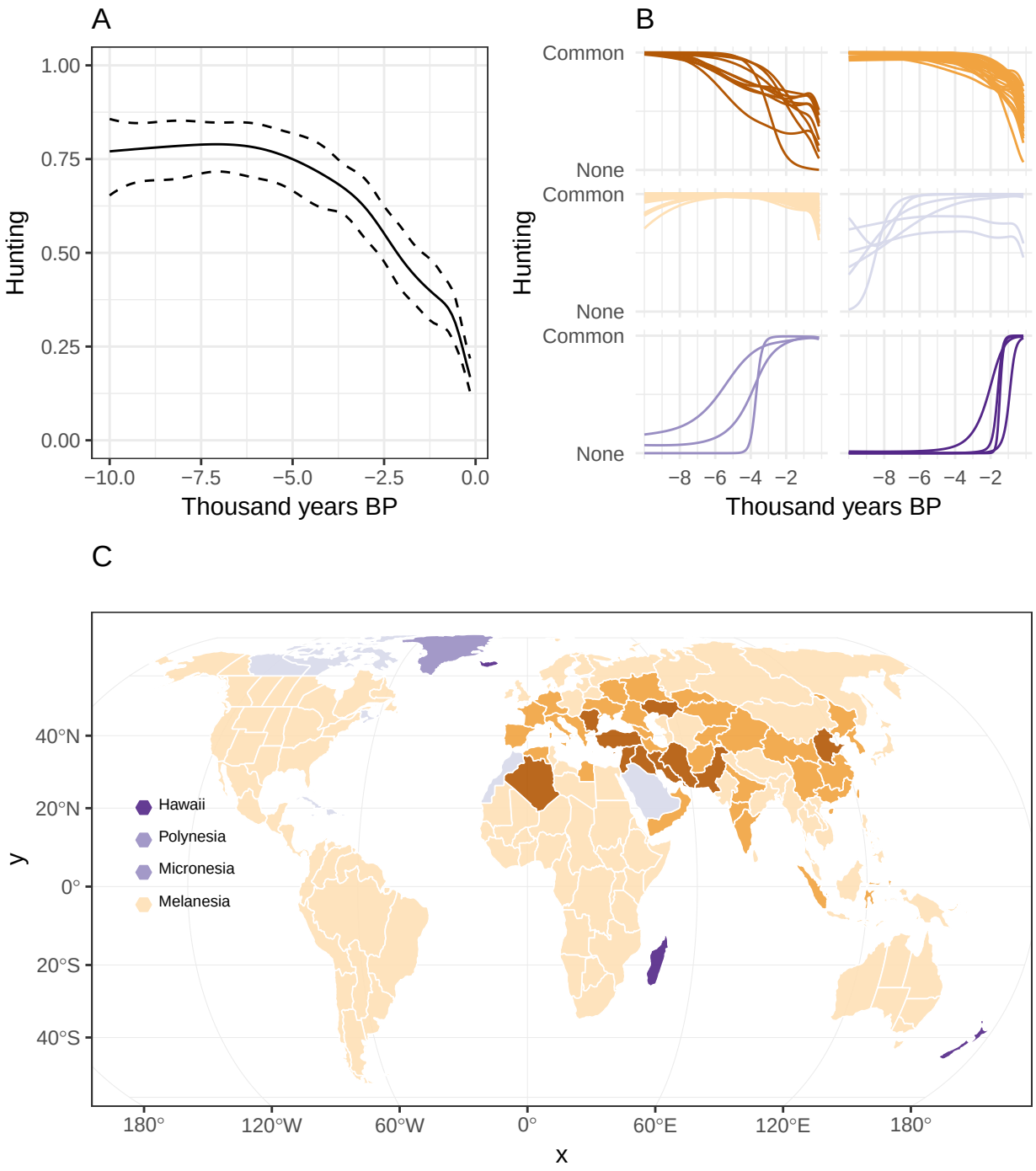
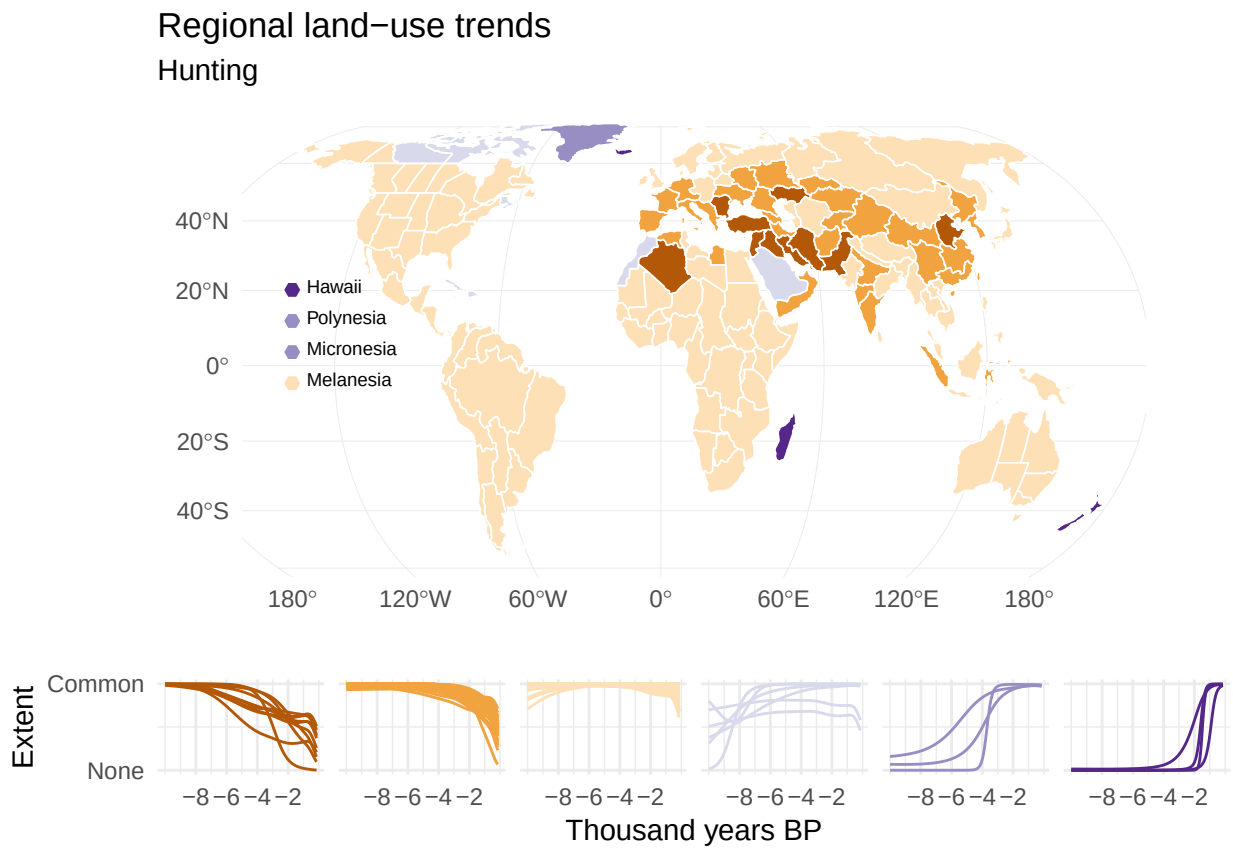



Figure 1: Global and regional trends in the areal extent of hunting. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.




```
ggsave('figures/trends_local_hunting.png', height = 8, width = 12)
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: number of items to replace is not a multiple of replacement length
```

## Extensive Agriculture

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.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
trend_dat[4, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                      cluster = recode_factor(cluster,
```

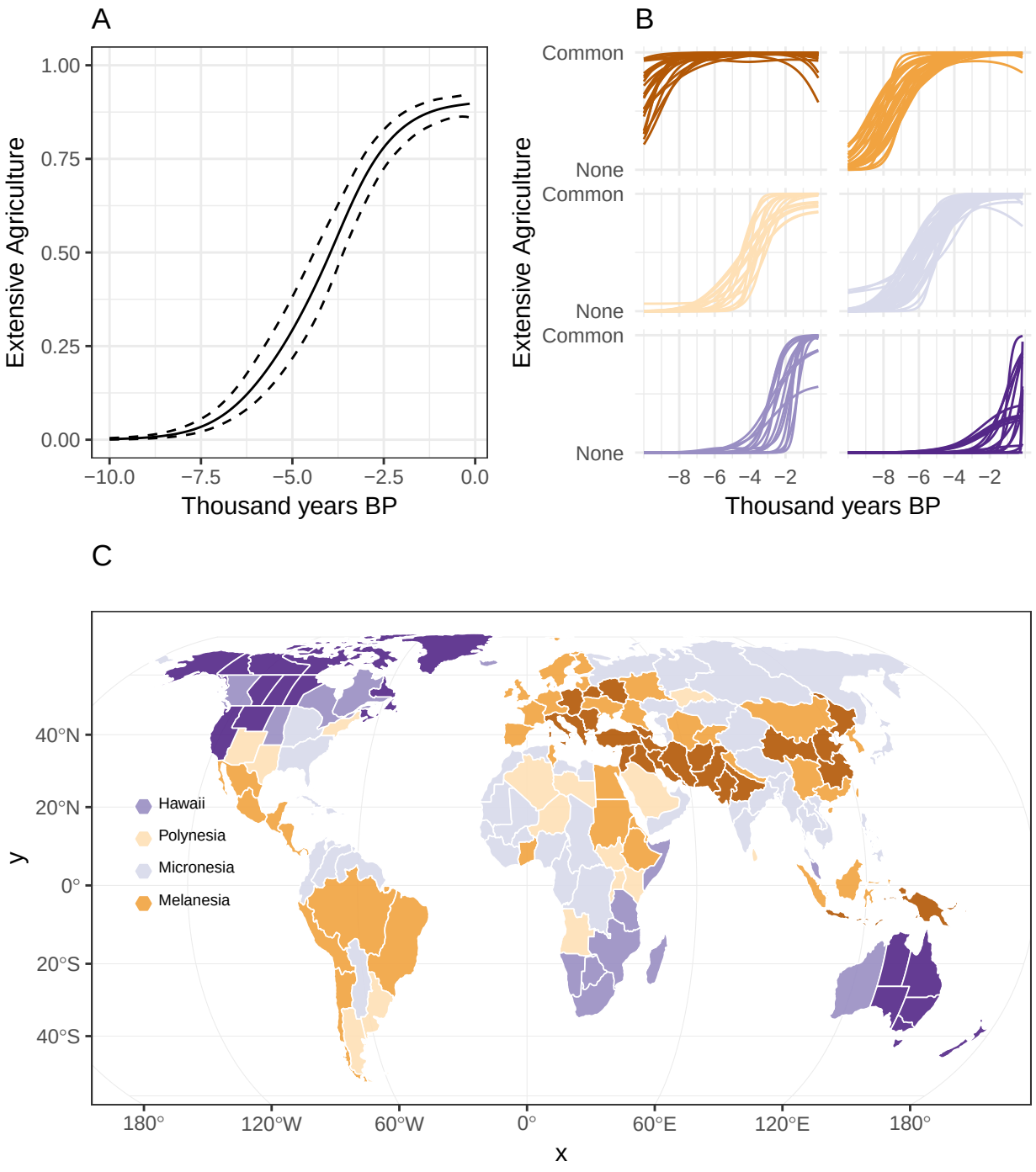


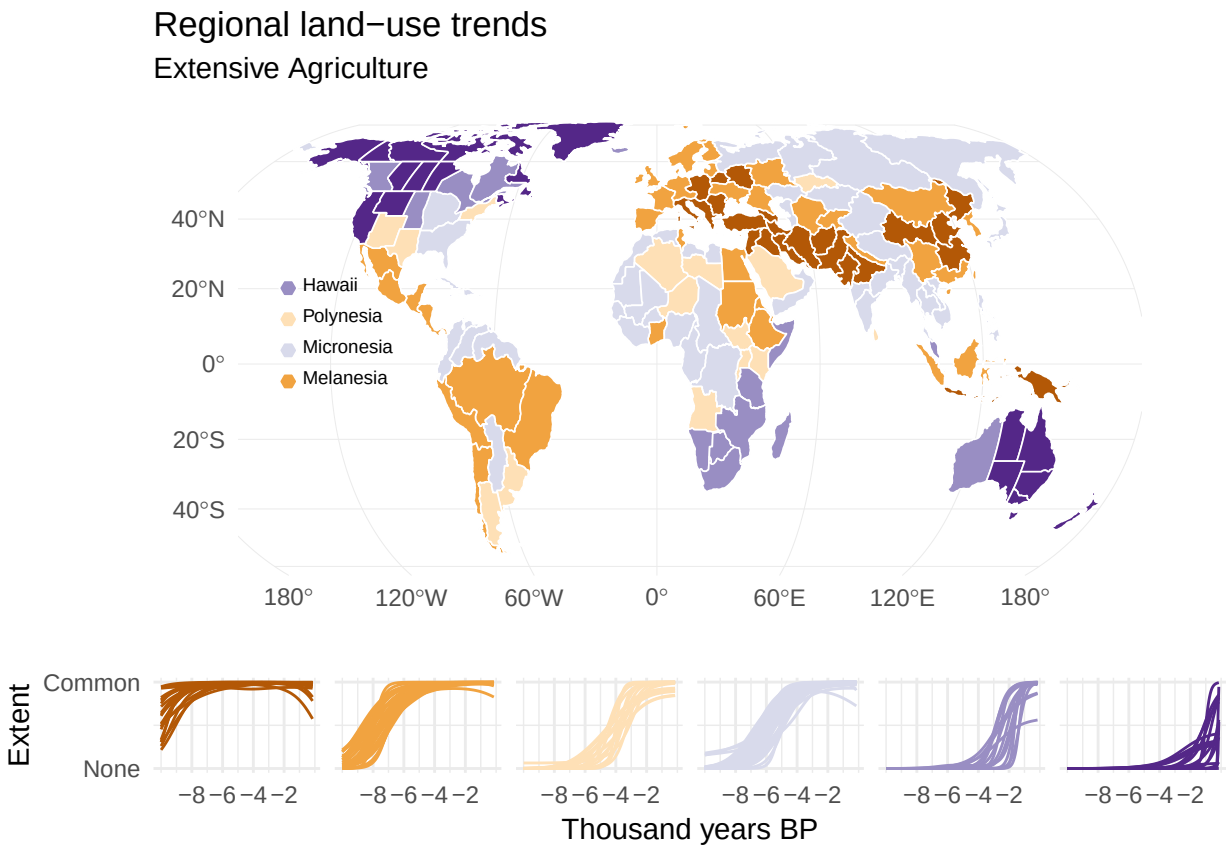

Figure 2: Global and regional trends in the areal extent of extensive agriculture. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.

```
                                `4` = '1', `2` = '2',`1` = '3',
                                `5` = '4', `3` = '5', `6` = '6')))) %>%
                              # `6` = '1', `1` = '2',`3` = '3',
                                #`4` = '4', `5` = '5', `2` = '6')))) %>%
plot_trends2('Extensive Agriculture')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning in xList[i] <- valueList: number of items to replace is not a
## multiple of replacement length
```




```
ggsave('figures/trends_local_extensive_ag.png', height = 8, width = 12)
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: number of items to replace is not a multiple of replacement length
```

## Intensive Agriculture

See above.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

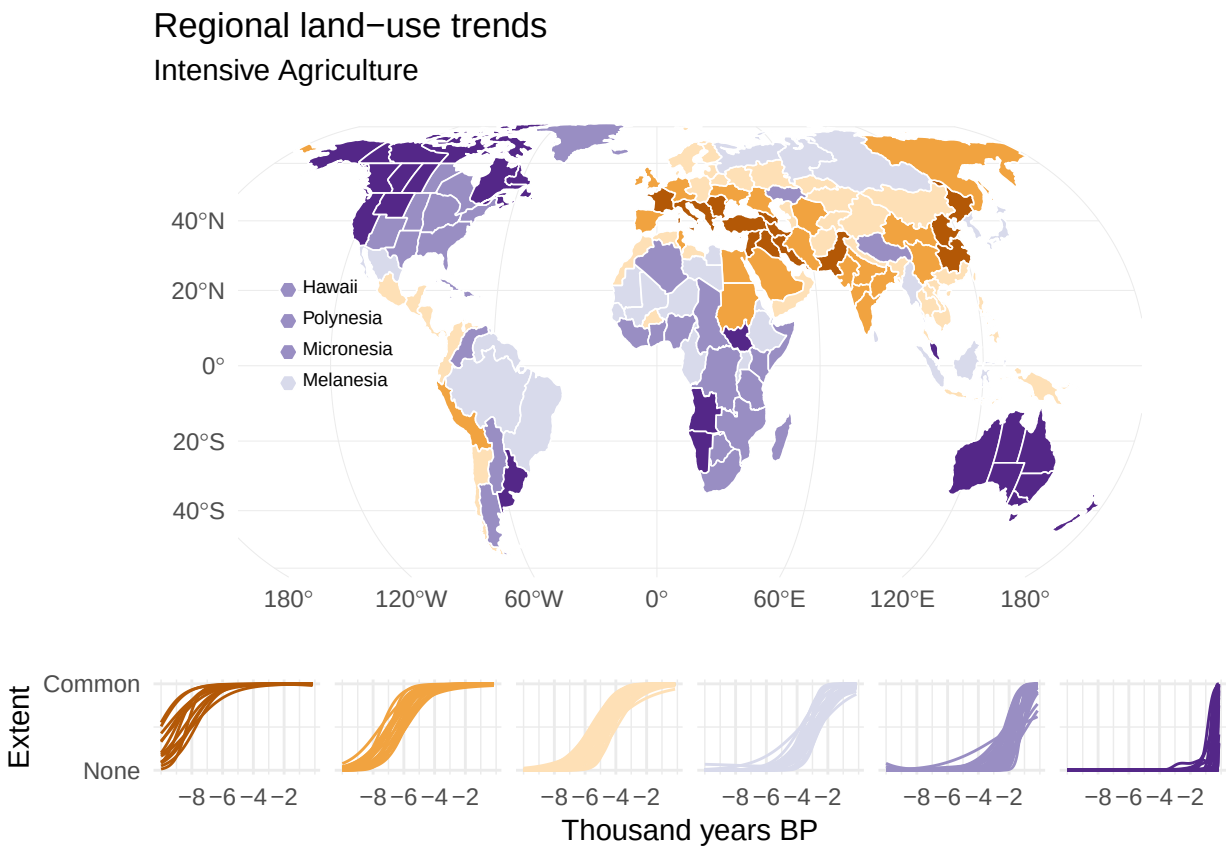
```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
trend_dat[5, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                      cluster = recode_factor(cluster,
                                     `6` = '1', `3` = '2',`4` = '3',
                                      `5` = '4', `1` = '5', `2` = '6')))) %>%
                                      #`1` = '1', `2` = '2',`5` = '3',
                                      #`4` = '4', `3` = '5', `6` = '6')))) %>%
plot_trends2('Intensive Agriculture')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning in xList[i] <- valueList: number of items to replace is not a
## multiple of replacement length
```



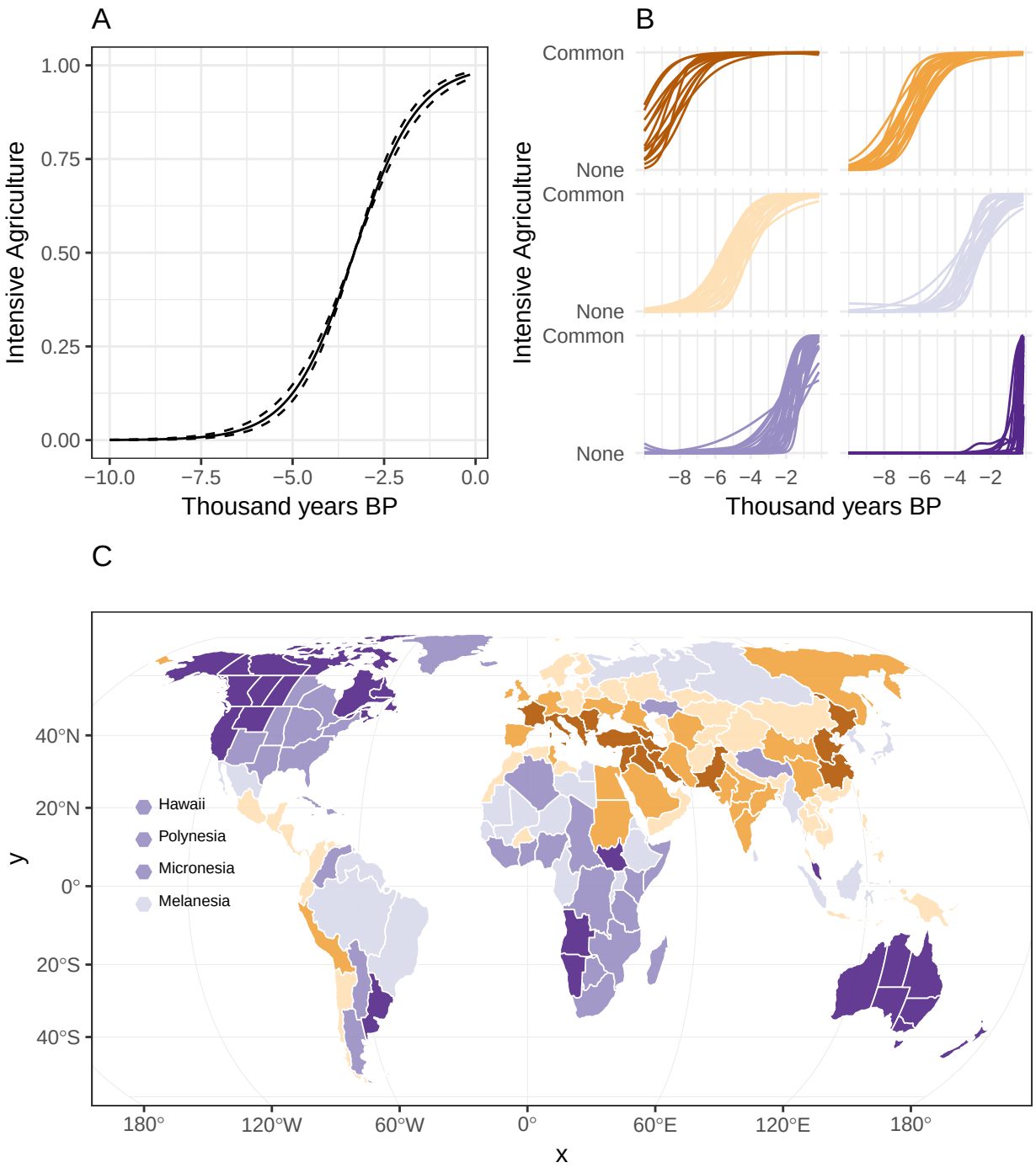



Figure 3: Global and regional trends in the areal extent of intensive agriculture. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.

```
ggsave('figures/trends_local_intensive_ag.png', height = 8, width = 12)
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: number of items to replace is not a multiple of replacement length
```

### Pastoralism

See above.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
trend_dat[6, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                cluster = recode_factor(cluster,
                                `6` = '1', `5` = '2',`3` = '3',
                                `1` = '4', `4` = '5', `2` = '6')))) %>%
plot_trends2('Pastoralism')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning in xList[i] <- valueList: number of items to replace is not a
## multiple of replacement length
```

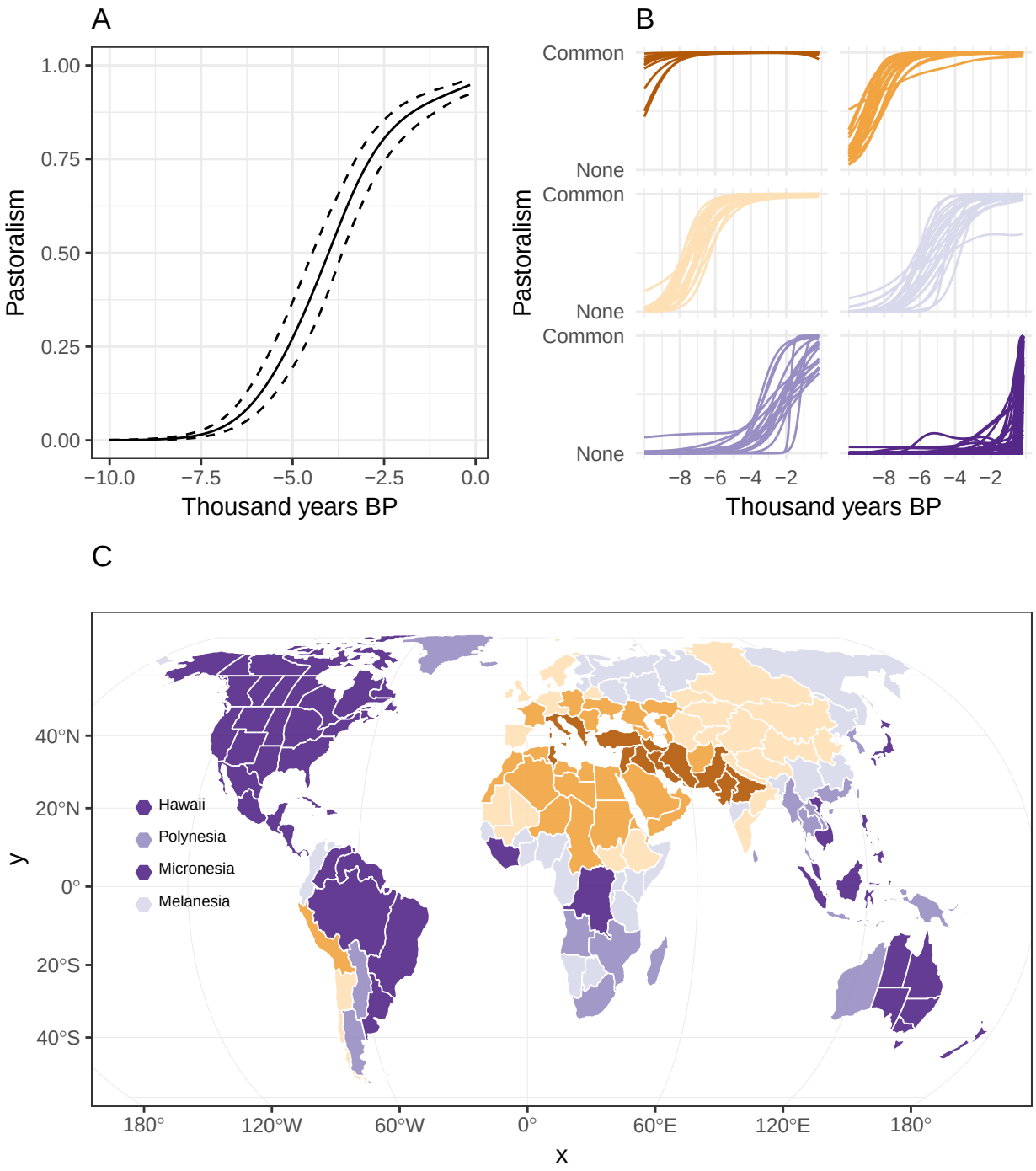
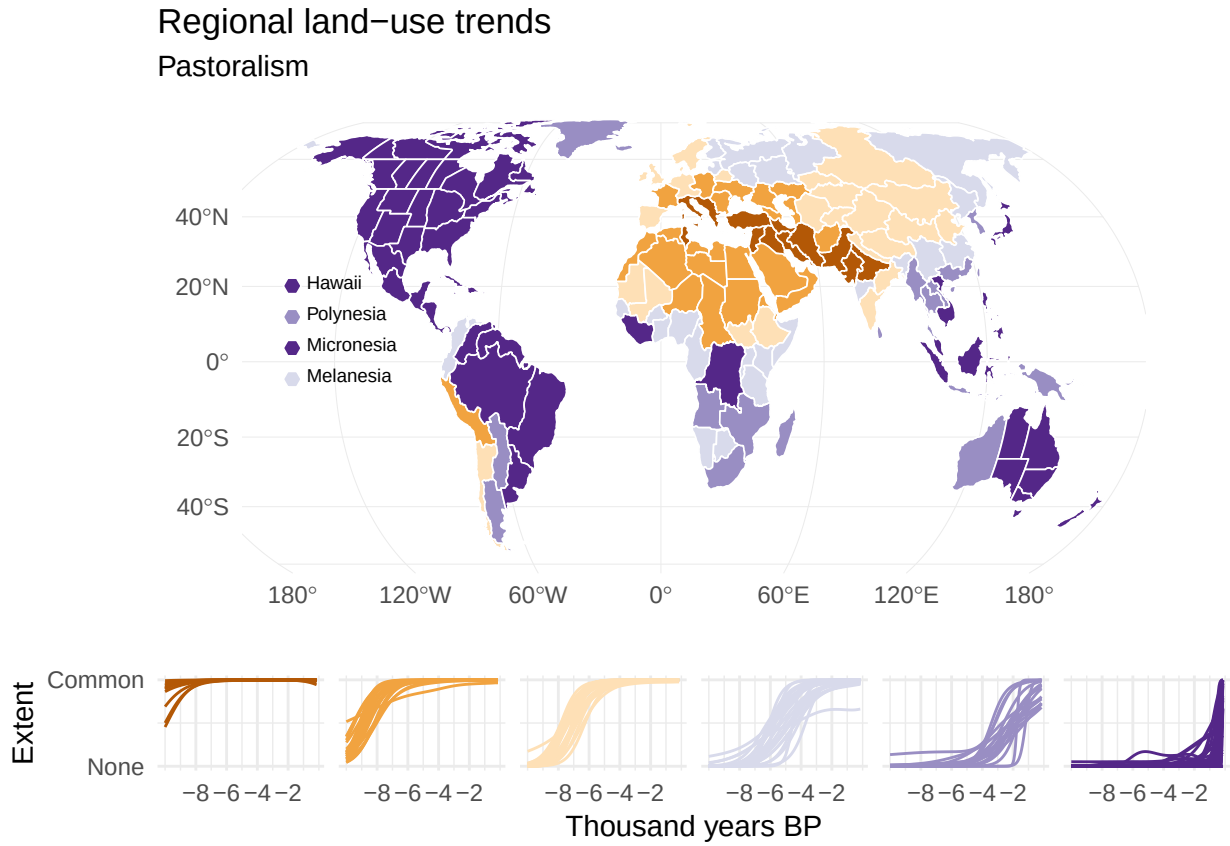



Figure 4: Global and regional trends in the areal extent of pastoralism. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.




```
ggsave('figures/trends_local_pastoralism.png', height = 8, width = 12)
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: number of items to replace is not a multiple of replacement length
```

## Urbanism

See above.

```
trend_dat[7, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                      cluster = recode_factor(cluster,
                                      `2` = '1', `6` = '2',`5` = '3',
                                      `4` = '4', `1` = '5', `3` = '6')))) %>%
plot_trends('Urbanism')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
ggsave('figures/trends_urbanism.png')
```

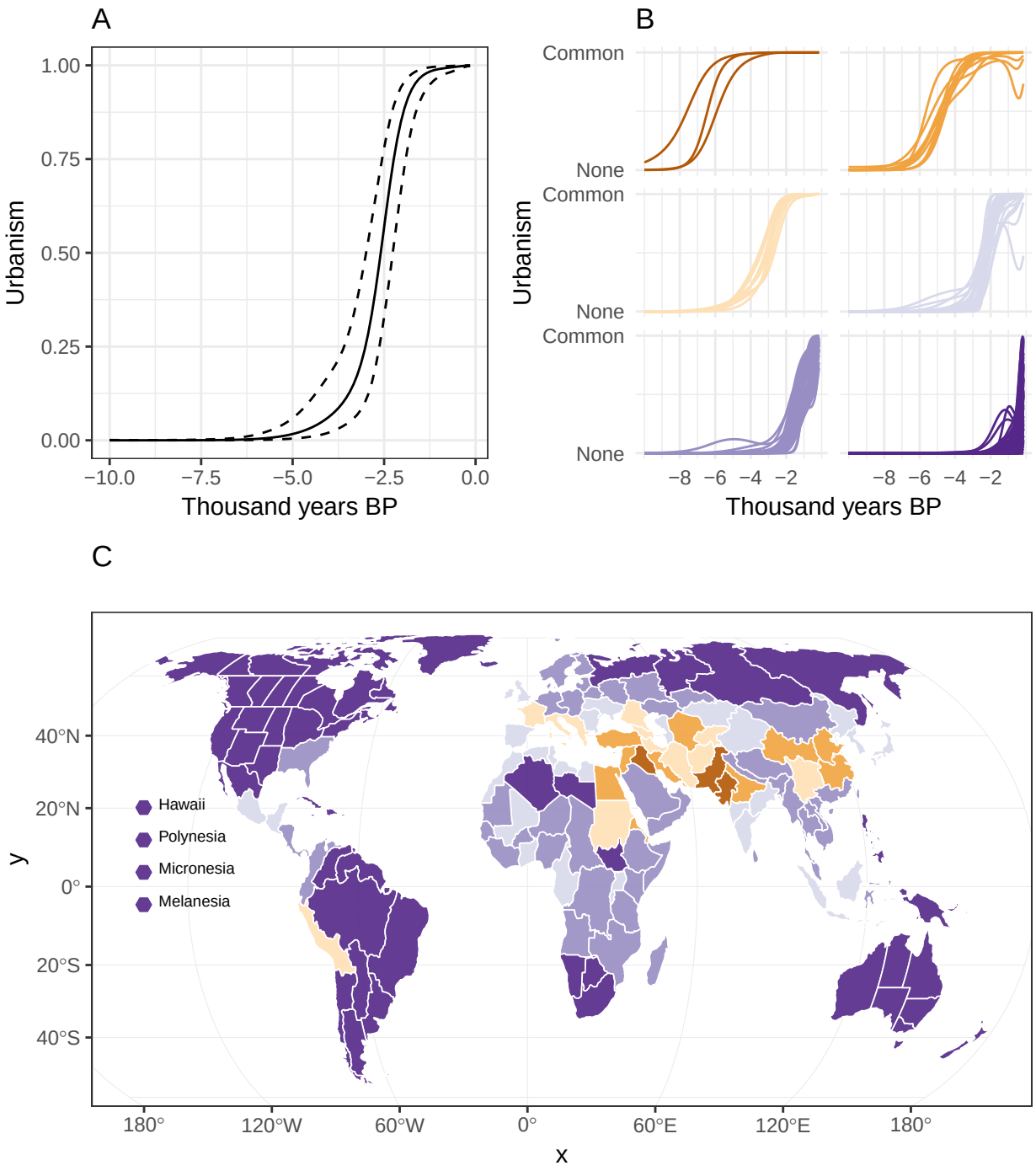



Figure 5: Global and regional trends in the presence of urban centers. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.

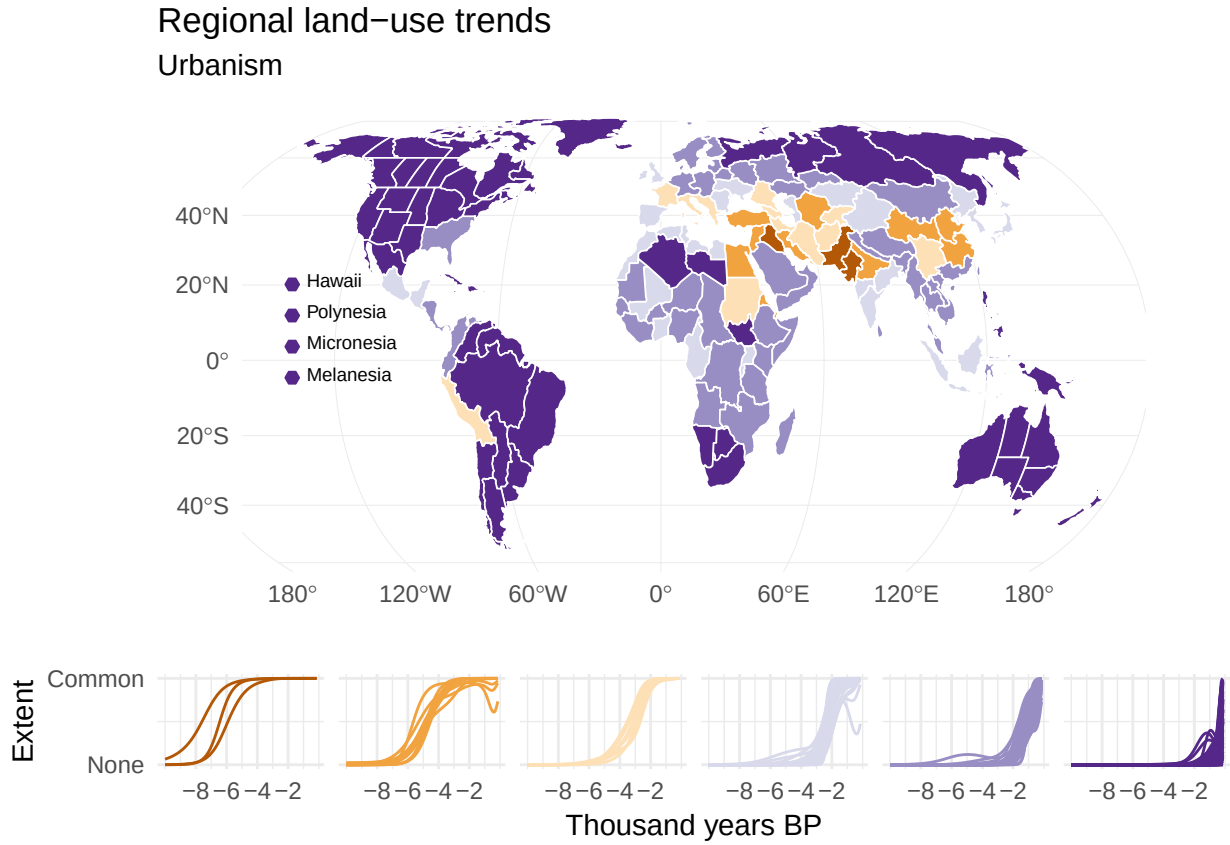
```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
trend_dat[7, ] %>%
  mutate(trends = map(trends, ~mutate(.,
                                      cluster = recode_factor(cluster,
                                      `2` = '1', `6` = '2',`5` = '3',
                                      `4` = '4', `1` = '5', `3` = '6')))) %>%
plot_trends2('Urbanism')
```

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning in xList[i] <- valueList: number of items to replace is not a
## multiple of replacement length
```




```
ggsave('figures/trends_local_urbanism.png', height = 8, width = 12)
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: number of items to replace is not a multiple of replacement length
```

## Expertise

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.

Now we cluster together the local deviations from the global trend using a k-means algorithm. The selection of 6 clusters is somewhat arbitrary, and is made simply based on visual comparisons of different cluster solutions with the goal making the results visually interpretable. The trajectories in these clusters are deviations from the global trend, so a horizontal line would indicate no deviation from the global trend.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

## Data Quality

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.

```
## Warning: Column `REGION_LAB` joining character vector and factor, coercing
## into character vector
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

```
## Warning: Removed 142 rows containing missing values (geom_text).
```

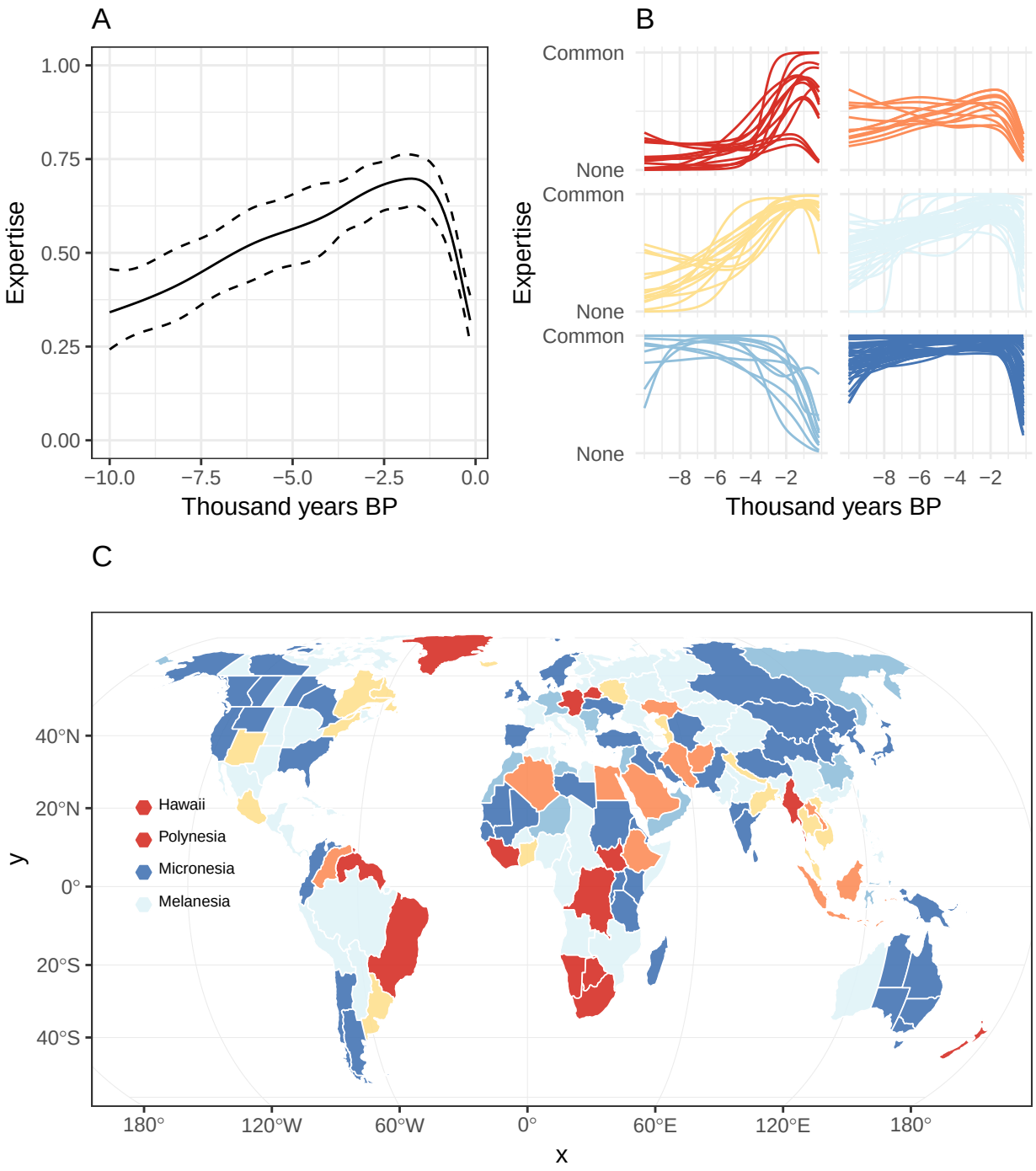



Figure 6: Global and regional trends in self-reported expertise. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.

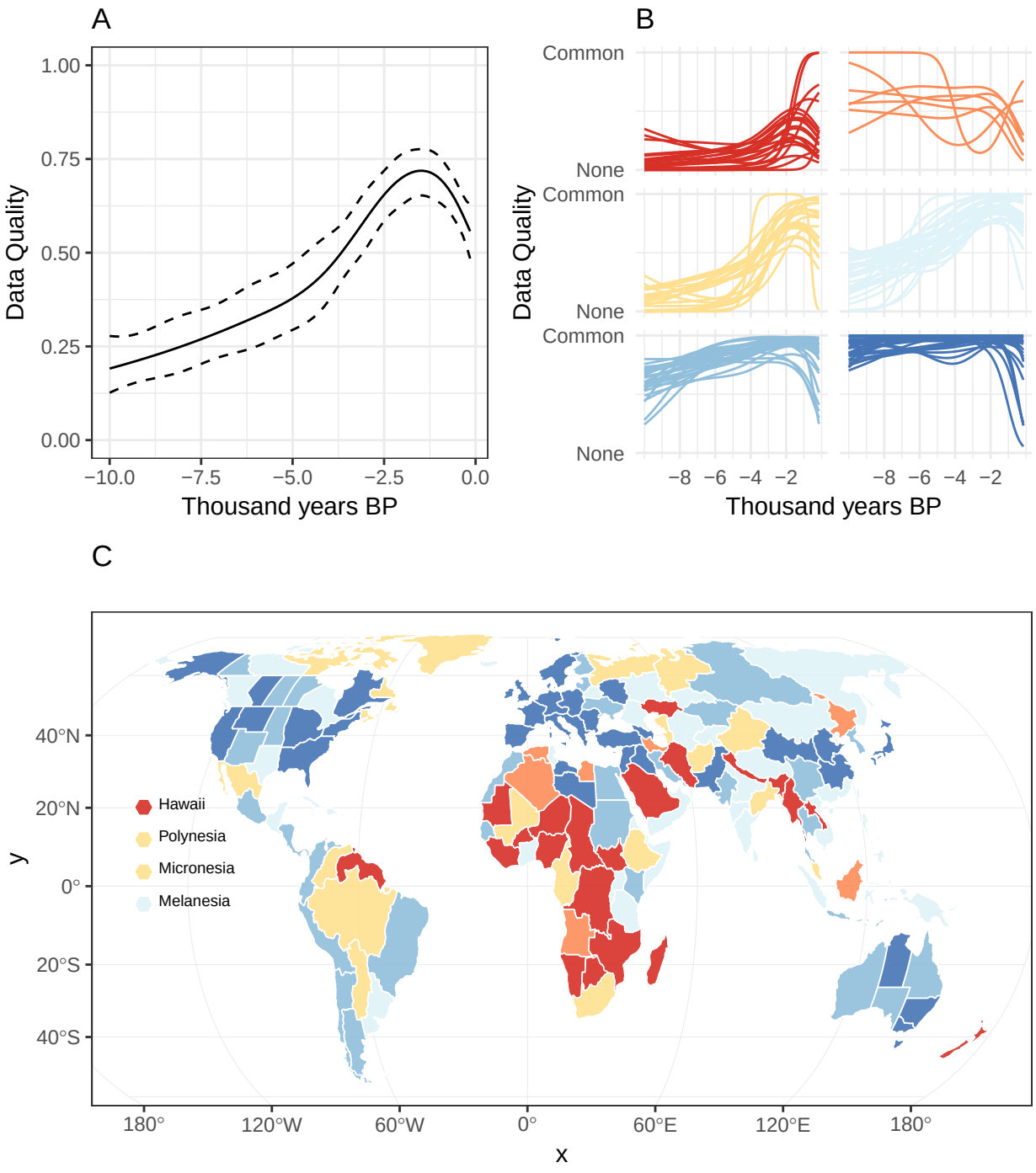



Figure 7: Global and regional trends in perceived data quality. (A) Global trend (all regions) with 95% confidence interval. (B) Regional deviations from global trend, clustered via k-means. (C) Map of the local deviations from the global trend, same clusters as in B.