Supplementary Material: R script for generating plots and overviews

Mixtec Sound Change Database

Sandra Auderset

16 November, 2023

This script provides the code for generating the plots in the paper and aggregating the data for the use cases described there. All maps are created with the package 'ggmap' (Kahle & Wickham 2013) using Stamen maps from Stadia (https://stadiamaps.com/attribution/).

Preparation

First, we load the necessary packages (cf. the Rmd file for details).

We read in the files from the Mixtec Sound Change Database repository. Next, we check the coverage of variables and languages and exclude those who have very low coverage from further analysis. Low coverage for varieties is defined here as having NA for more than a third of the variables. This is not currently the case in our database, so we can keep all the languages. Low coverage for variables is defined here the same way, i.e. as having NA for more than one third of the varieties. For the current study, this leads to the exclusion of 29 variables (see details in code). We create a file for analysis that excludes these low coverage variables.

```
# read in variable file
var_seg_all <-
read_tsv("https://raw.githubusercontent.com/SAuderset/MixteCaSo/main/variables/variables_segments.tsv")
```

```
## Rows: 105 Columns: 252
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (252): DOCULECT, A01, A02, A03, A04, A05, A06, A07, A08, A09, A10, A11, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# summarize language coverage
total_number_changes <- ncol(select(var_seg_all, -DOCULECT))
# check for number of NA
coverage_lang <- var_seg_all %>%
  mutate(na_lang = rowSums(is.na(.))) %>%
  select(DOCULECT, na_lang) %>%
  arrange(desc(na_lang)) %>%
  mutate(exclude = if_else(na_lang>(total_number_changes/3), "yes", "no"))
# none of the languages needs to be excluded

# summarize variable coverage
total_number_languages <- nrow(var_seg_all)
# exclude variables that cover less than 1/3rd of varieties
```

```
coverage_var <- var_seg_all %>%
  summarize(across(matches("[[:upper:]]{1}\\d{2}"), ~sum(is.na(.)))) %>%
  pivot_longer(everything())
# get names of low coverage variables; list them
low_coverage_var <- coverage_var %>%
  filter(value>(total_number_languages/3)) %>%
  pull(name)
low_coverage_var
```

```
##  [1] "A06" "A07" "A09" "A10" "A11" "I09" "I17" "I19" "I21" "J05" "K01" "K03"
## [13] "Q02" "O06" "T05" "T08" "T11" "U08" "U11" "U12" "U13" "U19" "U27" "U28"
## [25] "U33" "W04" "W14" "W15" "X07"
```

```
# create data set for analysis excluding low coverage variables
var_seg <- var_seg_all %>%
  select(!c(low_coverage_var))
```

Map of the sample with subgroup/dialect area membership and inset

We set up base map of the Mixtec region for plotting with the ggmap package (Kahle & Wickham 2013) with a stamen map. We will exclude the diaspora variety Abosolo del Valle, located in Veracruz, from all maps, because including it would lead to less visibility for other varieties on the map. In the paper we will mention if it diverges from the rest of Group 71 (Mixtepec) to which it belongs. The map also includes important towns of the region which were not sampled to be better orient the reader in space.

```
# read in map data
map_data <-
read_tsv("https://raw.githubusercontent.com/SAuderset/MixteCaSo/main/data/metadata.tsv")
%>%
  filter(MapAbbr!="abas")

# get min/max values for map box
summary(map_data$Latitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.13   17.02   17.31   17.27   17.59   18.63
```

```
summary(map_data$Longitude)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -99.09  -98.05  -97.84  -97.78  -97.50  -96.76
```

```
# set up map with stamen
mixtec_base <- get_stadiamap(bbox = c(left=-99.2, bottom=16, right=-96.55, top=18.75),
maptype = "stamen_toner_background", zoom = 10, crop = TRUE)

# set up map with labels as function, to use as a base for plotting variables
map_all_labels <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(map_data, is.na(DOCULECT)), size = 2.5, fontface = "bold") +
  geom_label_repel(aes(x = Longitude, y = Latitude, label = MapAbbr), data =
  filter(map_data, !is.na(DOCULECT)), size = 3.5, fontface = "bold", family = "Linux
  Libertine", max.overlaps = 40, box.padding = 0, point.padding = 0.1, label.padding =
  0.1, color = "white", bg = "black", alpha = 0.5) +
```

```
  guides(color="none") +
  theme_map()
```

Next, I will set up a custom color scheme for the subgroups. For this, I first create a new column in the metadata with subgroups that shows lower-levels only for the very large Group 7.

```
metadata <-
read_tsv("https://raw.githubusercontent.com/SAuderset/MixteCaSo/main/data/metadata.tsv")
%>%
  mutate(AudersetGroupC = case_when(AudersetGroup=="Group 7" & !is.na(AudersetGroupSub) ~
  AudersetGroupSub,
                                    TRUE ~ AudersetGroup), .before = AudersetGroup) %>%
  mutate(AudersetGroupC = factor(AudersetGroupC, levels = c("Unclear", "Group 1", "Group
  2", "Group 3", "Group 4", "Linkage 5", "Group 6", "Group 7", "Group 71", "Group 72",
  "Group 73", "Group 74", "Linkage 75", "Group 76")))
```

```
## Rows: 105 Columns: 16
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (14): DOCULECT, VillageName, Abbreviation, MapAbbr, AudersetGroup, Auder...
## dbl  (2): Latitude, Longitude
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# custom color scale based on viridis palettes for 14 levels
ccol_14 <- c(
  "#6D4C3D",
  "#90d743",
  "#2B4570",
  "#35b779",
  "#43BCCD",
  "#4f772d",
  "#F9C80E",
  "#6e0930",
  "#e36596",
  "#bc3908",
  "#832161",
  "#c9071b",
  "#ba63ce",
  "#f26419"
 )
```

We also set up a custom function to count number of changes easily and a custom template for the maps, using a unicode font.

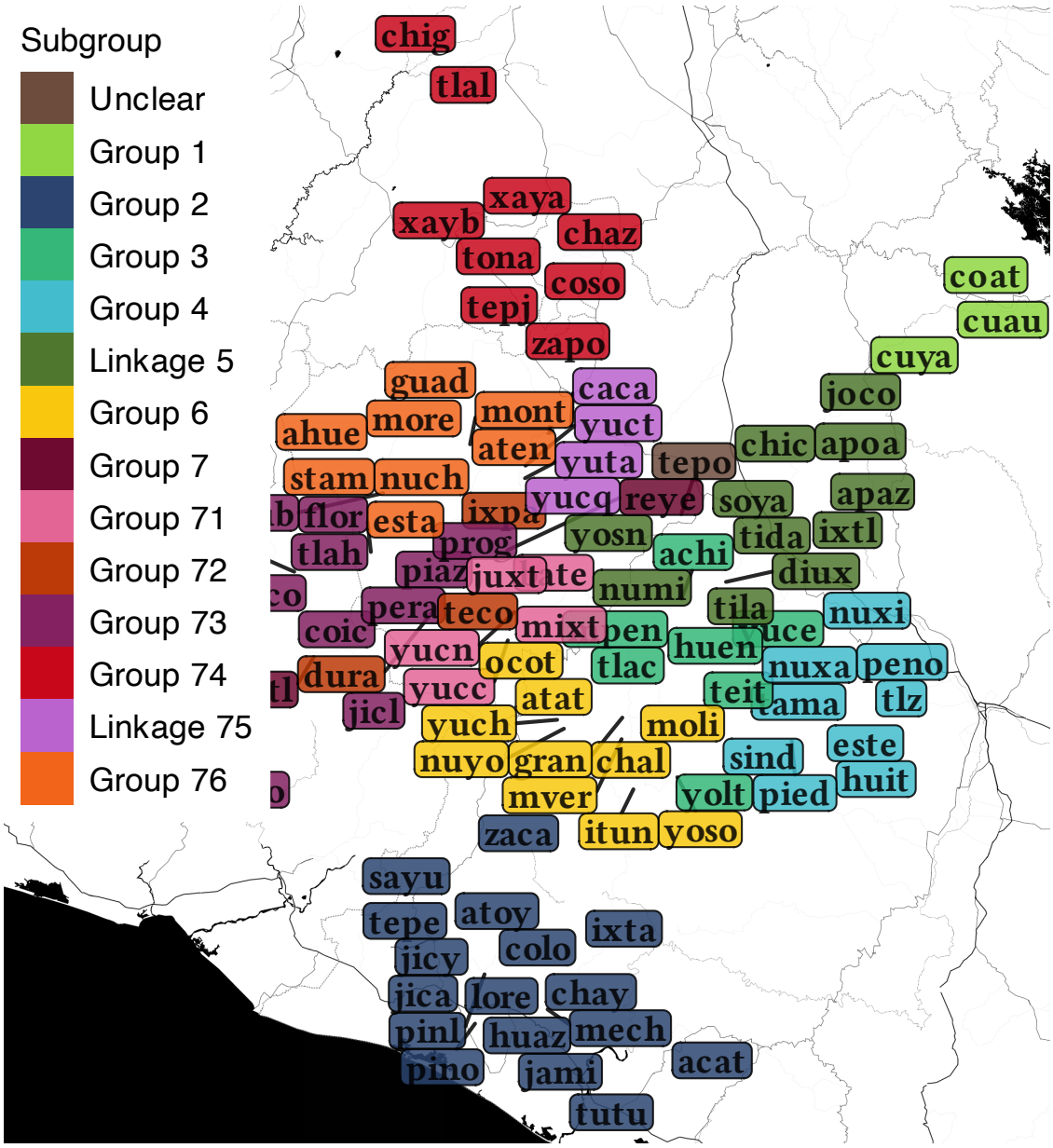
```
# function to calculate number of changes, number of NA, and ratio per variety
changes_calc <- function(c){
  c %>%
  group_by(DOCULECT) %>%
  mutate(n_change = sum(Value, na.rm = TRUE)) %>%
  mutate(n_na = sum(is.na(Value))) %>%
  mutate(r_change = round((100/(length(unique(Variable))-n_na))*n_change)) %>%
  ungroup() %>%
  distinct(DOCULECT, .keep_all = TRUE) %>%
```

```
  select(-Value) %>%
  mutate(Variable = str_match(Variable, "[[:upper:]]+")) %>%
  relocate(n_change:r_change, .after = DOCULECT)
}

# set up ggplot theme with unicode font and appropriate font size for reuse
theme_scmap <- theme(plot.title=element_markdown(family = "Noto Sans", size = 11, color =
"white", fill = "black", linewidth = 5, margin = margin(t=40, b=-30, l=50), padding =
unit(c(4, 4, 4, 4), "pt")),
        legend.title = element_blank(),
        legend.text = element_text(family = "Noto Sans", size = 10),
        legend.position = c(0, 0.5))
```

Now we can display a map of the sample colored by subgroups.

```
# map with subgroups
map_subgroups <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(metadata, is.na(AudersetGroupC)), size = 2.5, fontface = "bold") +
  geom_label_repel(aes(x = Longitude, y = Latitude, label = MapAbbr, fill =
  AudersetGroupC), data = filter(metadata, !is.na(AudersetGroupC)), size = 3.5, fontface
  = "bold", family = "Linux Libertine", max.overlaps = 40, box.padding = 0, point.padding
  = 0.1, label.padding = 0.1, alpha = 0.85) +
  scale_fill_manual(values = ccol_14) +
  guides(fill = guide_legend(title = "Subgroup", override.aes = aes(label = "", alpha =
  1))) +
  theme_map() +
  theme(legend.title=element_text(size=9),
        legend.text=element_text(size=9),
        legend.key.size = unit(1,"line"),
        legend.position = c(0, 1),
        legend.justification = c(0, 1),
        legend.box.background = element_rect(color = "white"))
map_subgroups
```



We create a map of Mesoamerica with the sample map highlighted as an inset.

```
# mesoamerica map
mesoamerica_map <- get_stadiamap(bbox = c(left = -112, bottom = 7.7, right = -82, top =
30), zoom = 5, maptype = "stamen_toner_lite")
```

```
## i © Stadia Maps © Stamen Design © OpenMapTiles © OpenStreetMap contributors.
```

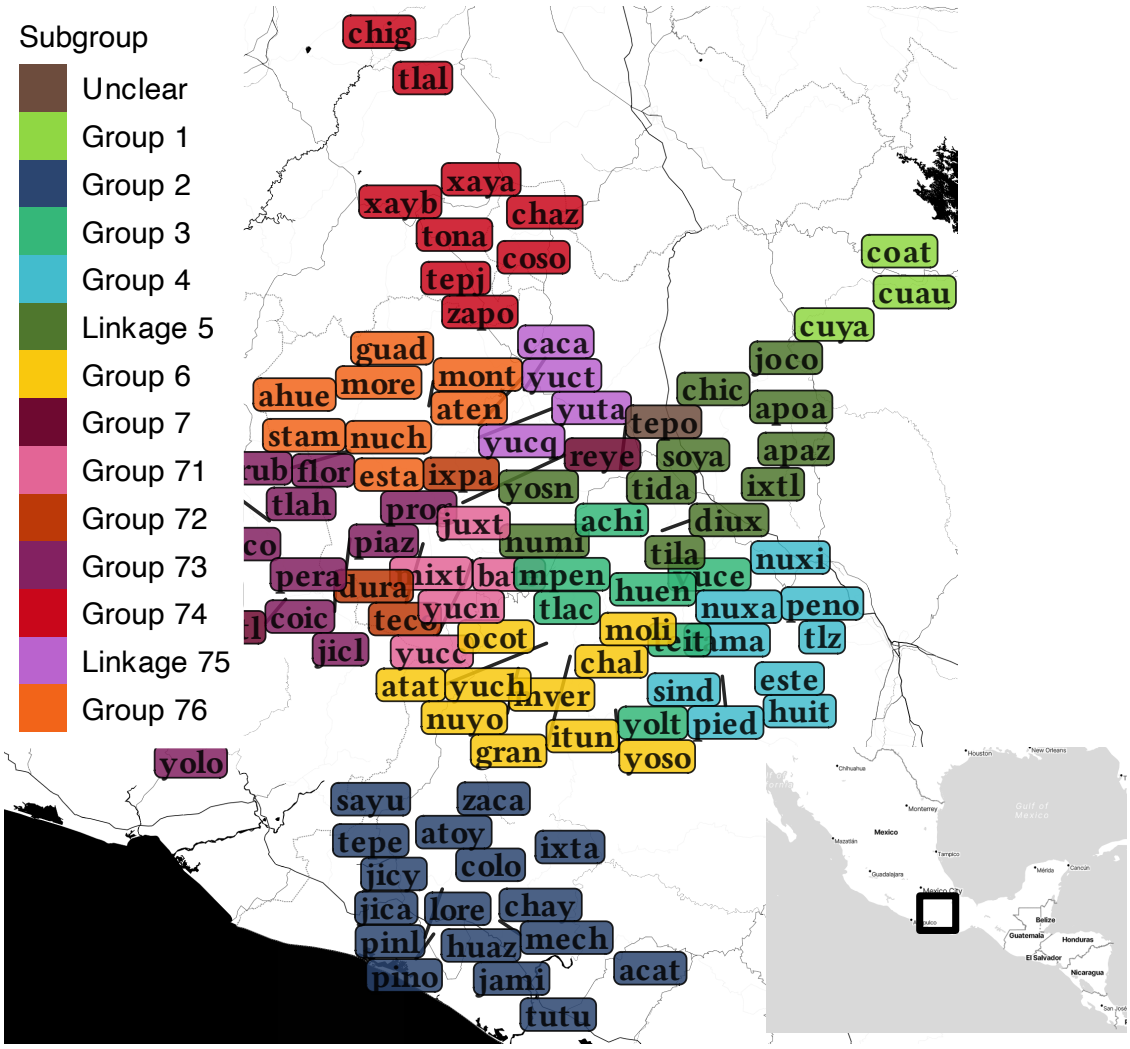
```
# make square for where the inset is
square <- data.frame(
  lon = c(-99.2, -99.2, -96.2, -96.2, -99.2),  # Adjust these values for your square
  lat = c(16, 18.75, 18.75, 16, 16)  # Adjust these values for your square
)

# map with square
mesoamerica_inset <- ggmap(mesoamerica_map) +
```

```
  geom_polygon(data = square, aes(x = lon, y = lat), fill = NA, color = "black", size =
  1) +
  theme_map()
mesoamerica_inset
```

Now we can combine the two maps.

```
map_subgroups_inset = ggdraw() +
  draw_plot(map_subgroups) +
  draw_plot(mesoamerica_inset, x = 0.65, y = 0.01, width = 0.3, height = 0.3, valign = 0,
  halign = 0)
map_subgroups_inset
```



Overview numbers

I generate some overview numbers for the paper regarding number of cognate sets, proto-forms, sound changes etc.

```
# number of reconstructed proto-forms, number of new reconstructions, number of
classifier morphemes
# add variable that encodes whether the form was reconstructed by a previous source
protoforms <-
read_tsv("https://raw.githubusercontent.com/SAuderset/MixteCaSo/main/data/protoforms.tsv")
%>%
  mutate(previous_source = if_all(ends_with("ID"), ~is.na(.)))
```

```
## Rows: 236 Columns: 13
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (8): MEANING, SPANISH, PMX, PMX_Josserand1983, PMX_Durr1987, PMX_Swanton...
## dbl (5): COGIDS, JosserandID, DurrID, SwantonMendoza_ID, SwantonID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# lexical entries in cognate sets
# set up vector with sample languages
sample_languages <- unique(var_seg_all$DOCULECT)
cognate_db <-
read_tsv("https://raw.githubusercontent.com/SAuderset/mixtecan-cognate-database/main/mixtecan_cognate_d
%>%
  filter(DOCULECT %in% sample_languages)
```

```
## Rows: 16158 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (10): CONCEPT, GLOSS, COGIDS, DOCULECT, SOURCE_ORIGINAL, SOURCE_ORTHOGRA...
## dbl  (1): ID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# combine
numbers_overview <- tibble(
  pf_total = nrow(protoforms),
  pf_new = nrow(filter(protoforms, previous_source==TRUE)),
  pf_clf = nrow(filter(protoforms, str_detect(MEANING, "CLF"))),
  cognates = nrow(cognate_db),
  sc_total = ncol(var_seg_all),
  sc_vowels = ncol(select(var_seg_all, matches("A|E|I|O|U|Y"))),
  sc_cons = ncol(var_seg_all)-ncol(select(var_seg_all, matches("A|E|I|O|U|Y"))),
  sc_excluded = length(low_coverage_var)
)
numbers_overview
```

```
## # A tibble: 1 x 8
##   pf_total pf_new pf_clf cognates sc_total sc_vowels sc_cons sc_excluded
##      <int>  <int>  <int>    <int>    <int>     <int>   <int>       <int>
## 1      236     63      2    15116      252       133     119          29
```

Use cases

For the analysis and plotting, we create a data set that combines the change variables, metadata, and definitions.

```
# definition file
changes_def <-
read_tsv("https://raw.githubusercontent.com/SAuderset/MixteCaSo/main/definitions/changes_segments.tsv")
```

```
## Rows: 250 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## chr (11): ID, SOUND_FROM, SOUND_TO, ENVIRONMENT_LEFT, ENVIRONMENT_RIGHT, NOT...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# combine, recode to numeric
changes_combined <- var_seg %>%
  pivot_longer(-DOCULECT, names_to = "ID", values_to = "Value") %>%
```

```
  left_join(., select(changes_def, ID:ENVIRONMENT_RIGHT)) %>%
  left_join(., select(metadata, DOCULECT, MapAbbr:AudersetGroupC, Latitude:Longitude))
  %>%
  mutate(Value = case_when(Value=="yes" ~ 1,
                           Value=="no" ~ 0,
                           TRUE ~ as.numeric(Value)))
```

```
## Joining with `by = join_by(ID)`
## Joining with `by = join_by(DOCULECT)`
```

```
glimpse(changes_combined)
```

```
## Rows: 23,310
## Columns: 11
## $ DOCULECT          <chr> "AbasoloValleMixtec", "AbasoloValleMixtec", "Abasolo~
## $ ID                <chr> "A01", "A02", "A03", "A04", "A05", "A08", "A12", "A1~
## $ Value             <dbl> 0, NA, NA, NA, NA, 0, NA, NA, 1, 0, 0, 0, 0, 0, 0, 0~
## $ SOUND_FROM        <chr> "*a", "*a", "*a", "*a", "*a", "*a", "*a", NA, "* d",~
## $ SOUND_TO          <chr> "ja", "e", "i", "e", "i", "o", "o", NA, " d ", "n", ~
## $ ENVIRONMENT_LEFT  <chr> "#", NA, "aj", NA, NA, "a w", "k ", NA, NA, NA, NA, ~
## $ ENVIRONMENT_RIGHT <chr> "s", "ja", NA, " je", " je", NA, "#", NA, "i", "it {~
## $ MapAbbr           <chr> "abas", "abas", "abas", "abas", "abas", "abas", "aba~
## $ AudersetGroupC    <fct> Group 71, Group 71, Group 71, Group 71, Group 71, Gr~
## $ Latitude          <dbl> 17.78287, 17.78287, 17.78287, 17.78287, 17.78287, 17~
## $ Longitude         <dbl> -95.54381, -95.54381, -95.54381, -95.54381, -95.5438~
```

We also set up two custom functions for map plotting.

```
# custom map function for displaying variables
map.part.variable <- function(p, v){
  ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(metadata, is.na(AudersetGroupC)), size = 2.5, fontface = "bold") +
  geom_label_repel(data = p, aes(x = Longitude, y = Latitude, label = MapAbbr, fill = v),
  size = 3.5, fontface = "bold", family = "Linux Libertine", max.overlaps = 40,
  box.padding = 0, point.padding = 0.1, label.padding = 0.1, color = "white")
}
# custom map theme
map.sc.theme <- function(){
  theme_map() +
  theme(legend.title=element_blank(),
        legend.text=element_text(size=9),
        legend.key.size = unit(1,"line"),
        legend.position = c(0, 0.6))
}
```

Summarizing reflexes of a proto-sound and conditioning environments

As an example for how reflexes and conditioning environments of a proto-sound can be summarized, we look at the developments from proto-Mixtec *s. We subset our data set to just the changes from *s and look at the distinct reflexes. We exclude the special cases of loss and metathesis, as we are interested in the primary reflexes. Retention is not coded as such in the database, because that does not constitute a change. However, the varieties that retain *s are those that do not have any changes from *s. We extract those from the database and create a new variable summarizing the main reflexes and retention. We plot this on a map.

```r
# subset to s, keep only those that apply
pm_s <- changes_combined %>%
  filter(str_detect(ID, "S")) %>%
  select(DOCULECT:AudersetGroupC, SOUND_FROM:ENVIRONMENT_RIGHT, Latitude:Longitude)
# overview of modern reflexes
pm_s %>%
  distinct(SOUND_TO)
```

```
## # A tibble: 6 x 1
##   SOUND_TO
##   <chr>
## 1 ð
## 2 h
## 3
## 4 ø
## 5 i
## 6 is
```

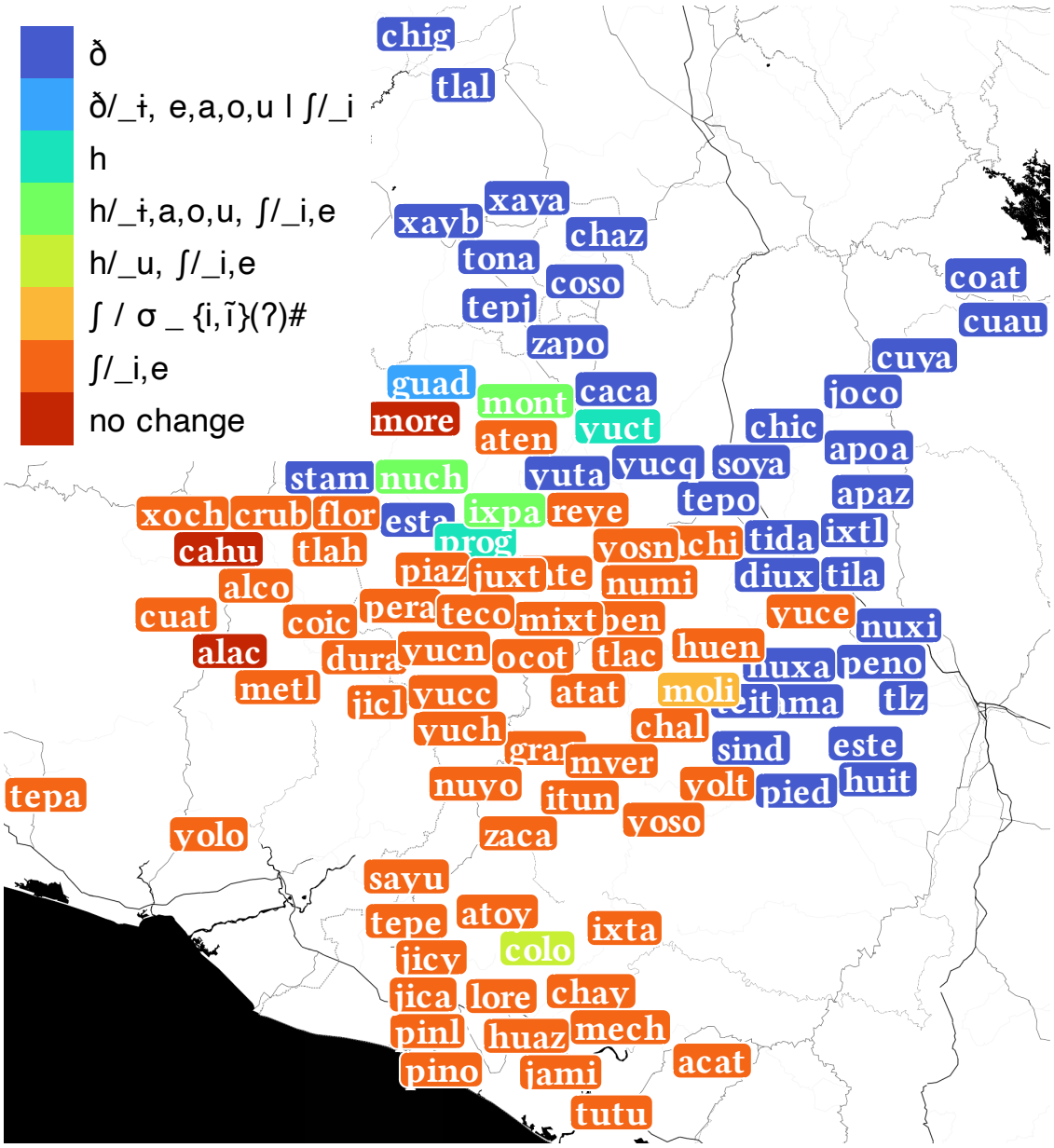
```r
# recode for global changes excluding metathesis and initial loss
# get retention first
pm_s_retention <- pm_s %>%
  filter(!is.na(Value)) %>%
  group_by(DOCULECT) %>%
  mutate(s_reflexes = sum(Value)) %>%
  distinct(DOCULECT, s_reflexes, .keep_all = TRUE) %>%
  filter(s_reflexes==0) %>%
  mutate(s_reflexes = if_else(s_reflexes==0, "no change", "x"))

# changes
pm_s_global <- pm_s %>%
  filter(SOUND_FROM!="*si") %>%
  filter(SOUND_TO!="ø") %>%
  group_by(DOCULECT) %>%
  mutate(s_total = sum(Value)) %>%
  filter(Value==1) %>%
  mutate(s_reflexes = paste(SOUND_TO, "/", ENVIRONMENT_LEFT, "_", ENVIRONMENT_RIGHT)) %>%
  mutate(s_reflexes = str_remove_all(s_reflexes, "NA")) %>%
  distinct(DOCULECT, s_reflexes, .keep_all = TRUE) %>%
  bind_rows(pm_s_retention) %>%
  group_by(DOCULECT) %>%
  mutate(s_reflexes = paste(s_reflexes, collapse = ", ")) %>%
  distinct(DOCULECT, s_reflexes, .keep_all = TRUE) %>%
  mutate(s_reflexes_short = case_when(
    s_reflexes==" /  _ i,e,ĩ,ẽ,  /  _ {i,ĩ}( )#" ~ " /_i,e",
    s_reflexes=="ð /  _  ,e,a,o,u,˜,ẽ,ũ, ð /  _ i,ĩ" ~ "ð",
    s_reflexes=="ð /  _  ,e,a,o,u,˜,ẽ,ũ,  /  _ i,e,ĩ,ẽ,  /  _ {i,ĩ}( )#" ~ "ð/_ , e,a,o,u
| /_i",
    s_reflexes=="h /  _  ,a,o,˜, h /  _ i,e,ĩ,ẽ, h / # _ u,ũ" ~ "h",
    s_reflexes=="h / # _ u,ũ,  /  _ i,e,ĩ,ẽ,  /  _ {i,ĩ}( )#" ~ "h/_u, /_i,e",
    s_reflexes=="h /  _  ,a,o,˜, h / # _ u,ũ,  /  _ i,e,ĩ,ẽ,  /  _ {i,ĩ}( )#" ~
"h/_ ,a,o,u, /_i,e",
    TRUE ~ s_reflexes)) %>%
  ungroup() %>%
  mutate(s_reflexes_short = fct_relevel(s_reflexes_short, "no change", after = Inf)) %>%
```

```
  arrange(s_reflexes_short)
glimpse(pm_s_global)
```

```
## Rows: 105
## Columns: 14
## $ DOCULECT          <chr> "CosoltepecMixtec", "CuyamecalcoVillaZaragozaMixtec"~
## $ ID                <chr> "S01", "S01", "S01", "S01", "S01", "S01", "S01", "S0~
## $ Value             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ SOUND_FROM        <chr> "*s", "*s", "*s", "*s", "*s", "*s", "*s", "*s", "*s"~
## $ SOUND_TO          <chr> "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð~
## $ ENVIRONMENT_LEFT  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ ENVIRONMENT_RIGHT <chr> " ,e,a,o,u,˜,ẽ,ũ", " ,e,a,o,u,˜,ẽ,ũ", " ,e,a,o,u,˜,ẽ~
## $ MapAbbr           <chr> "coso", "cuya", "nuxi", "yuta", "huit", "soya", "xay~
## $ AudersetGroupC    <fct> Group 74, Group 1, Group 4, Linkage 75, Group 4, Lin~
## $ Latitude          <dbl> 18.14291, 17.96525, 17.23883, 17.60626, 16.94760, 17~
## $ Longitude         <dbl> -97.79059, -96.83107, -97.10497, -97.89428, -97.1236~
## $ s_total           <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~
## $ s_reflexes        <chr> "ð /  _  ,e,a,o,u,˜,ẽ,ũ, ð /  _ i,ĩ", "ð /  _  ,e,a,~
## $ s_reflexes_short  <fct> "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð", "ð~
```

```
# map showing distribution of global reflexes
s_refl_map <- map.part.variable(pm_s_global, pm_s_global$s_reflexes_short) +
  scale_fill_viridis(discrete = TRUE, option = "turbo", begin = 0.1, end = 0.9) +
  guides(fill = guide_legend(override.aes = aes(label = ""))) +
  map.sc.theme()
s_refl_map
```




```
# look at reflexes per group
s_refl_group <- pm_s_global %>%
  group_by(AudersetGroupC, s_reflexes) %>%
  summarize()
```

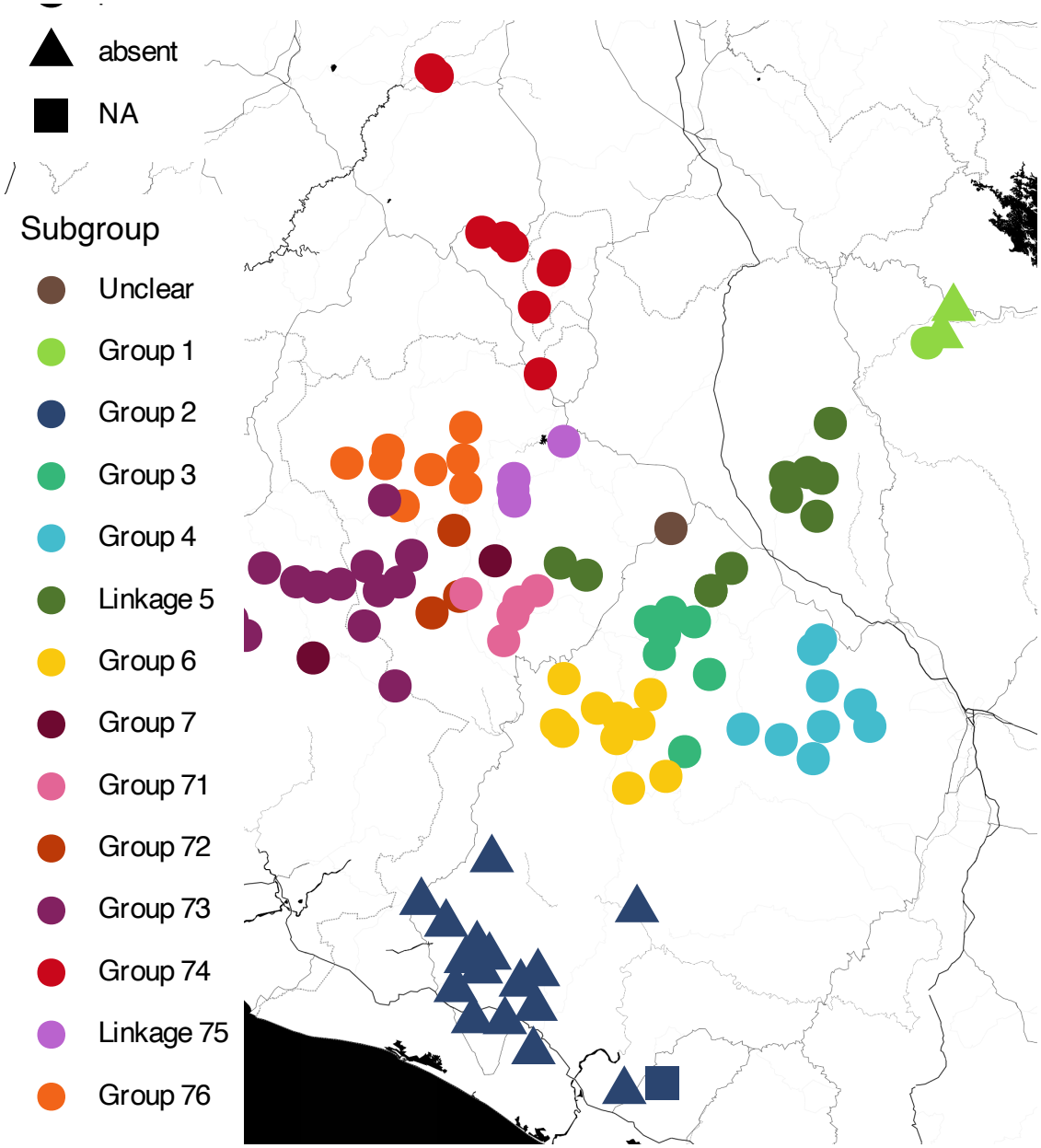
```
## `summarise()` has grouped output by 'AudersetGroupC'. You can override using
## the `.groups` argument.
```

Looking at specific types of changes

The database can also be used to look at specific types of changes, such as the loss of a sound or palatalization of different types of stops. We illustrate this here with the palatalization of proto-Mixtec *t before i. This is coded as one change, so we subset the database to just that. Because we are interested in the distribution of this change across subgroups, we recode it as present/absent and use color to represent subgroups.

```r
# subset
t_i_values <- changes_combined %>%
  filter(ID=="T01") %>%
  mutate(Value = as_factor(case_when(Value==0 ~ "absent",
                                     Value==1 ~ "present",
                                     TRUE ~ "NA"))) %>%
  mutate(Value = fct_relevel(Value, "present", "absent", "NA"))

# plot on map with subgroups
t_i_map <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(map_data, is.na(DOCULECT)), size = 2.5, fontface = "bold") +
  geom_point(aes(x = Longitude, y = Latitude, color = AudersetGroupC, shape = Value),
  data = t_i_values, size = 4) +
  scale_color_manual(values = ccol_14) +
  guides(shape = guide_legend(title = "*t > t  /_i"), color = guide_legend(title =
  "Subgroup", override.aes = list(size = 3), direction = "vertical", legend.position =
  "right")) +
  theme_map() +
  theme(legend.title=element_text(size=9),
        legend.text=element_text(size=8),
        #legend.position = c(0, 1),
        #legend.justification = c("left", "top")
        )
t_i_map
```



Summarizing changes across varieties

We count the total number of sound changes per variety to provide an overview of which varieties are particularly conservative and particularly innovative. We do the same for consonants and vowels separately to check for possible differences between these two sound classes. We do this by aggregating over the variables file with the numeric values (0 and 1).

```
# merge variable file (numeric) with metadata
metadata_sub <- select(metadata, DOCULECT, MapAbbr:JosserandAreaSub, Latitude, Longitude)
# create numeric version of variables
var_num <- var_seg %>%
  mutate(across(matches("^[[:upper:]]{1}\\d{2}$"), ~case_match(.x,
                                                               "yes" ~ 1,
                                                               "no" ~ 0))) %>%
```

```
  left_join(., metadata_sub)
```

```
## Joining with `by = join_by(DOCULECT)`
```

```
# count overall number of changes per variety; add percentage taking into account the NAs
total_changes <- ncol(var_seg[, -1])
changes_total <- var_num %>%
  rowwise() %>%
  mutate(NumAll = sum(across(matches("^[[:upper:]]{1}\\d{2}$")), na.rm = TRUE)) %>%
  mutate(NAAll = sum(is.na(across(matches("^[[:upper:]]{1}\\d{2}$"))))) %>%
  mutate(PropAll = round(100/(total_changes-NAAll)*NumAll, 1)) %>%
  mutate(NumVow = rowSums(across(matches("^[AEIOUY]{1}\\d{2}$")), na.rm = TRUE)) %>%
  mutate(NAVow = sum(is.na(across(matches("^[AEIOUY]{1}\\d{2}$"))))) %>%
  mutate(PropVow = round(100/(total_changes-NAVow)*NumVow, 1)) %>%
  mutate(NumCons = rowSums(across(matches("^[DJKNSTWX]{1}\\d{2}$")), na.rm = TRUE)) %>%
  mutate(NACons = sum(is.na(across(matches("^[DJKNSTWX]{1}\\d{2}$"))))) %>%
  mutate(PropCons = round(100/(total_changes-NACons)*NumCons, 1)) %>%
  select(DOCULECT, NumAll:PropCons, MapAbbr:Longitude) %>%
  ungroup() %>%
  arrange(desc(PropAll))

# summary overview of range
changes_total %>%
  select(NumAll:PropCons) %>%
  summary()
```

```
##      NumAll          NAAll           PropAll          NumVow
##  Min.   :35.00   Min.   : 0.000   Min.   :16.80   Min.   :17.00
##  1st Qu.:49.00   1st Qu.: 3.000   1st Qu.:22.80   1st Qu.:29.00
##  Median :54.00   Median : 6.000   Median :25.60   Median :34.00
##  Mean   :53.88   Mean   : 7.981   Mean   :25.18   Mean   :33.61
##  3rd Qu.:59.00   3rd Qu.: 9.000   3rd Qu.:27.50   3rd Qu.:38.00
##  Max.   :73.00   Max.   :66.000   Max.   :33.00   Max.   :46.00
##      NAVow           PropVow         NumCons          NACons
##  Min.   : 0.000   Min.   : 8.00   Min.   : 7.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:13.60   1st Qu.:14.00   1st Qu.: 1.000
##  Median : 3.000   Median :15.50   Median :19.00   Median : 2.000
##  Mean   : 5.057   Mean   :15.48   Mean   :18.38   Mean   : 2.771
##  3rd Qu.: 6.000   3rd Qu.:17.60   3rd Qu.:22.00   3rd Qu.: 4.000
##  Max.   :42.000   Max.   :20.80   Max.   :27.00   Max.   :22.000
##     PropCons
##  Min.   : 3.200
##  1st Qu.: 6.800
##  Median : 8.600
##  Mean   : 8.386
##  3rd Qu.:10.000
##  Max.   :12.400
```

```
# table of overall changes per variety
changes_total_table <- changes_total %>%
  select(DOCULECT, matches("Num|Prop")) %>%
  arrange(PropAll) %>%
  kable(., booktabs = TRUE, longtable = TRUE, escape = FALSE) %>%
  kable_styling(full_width = FALSE, font_size = 11)
```

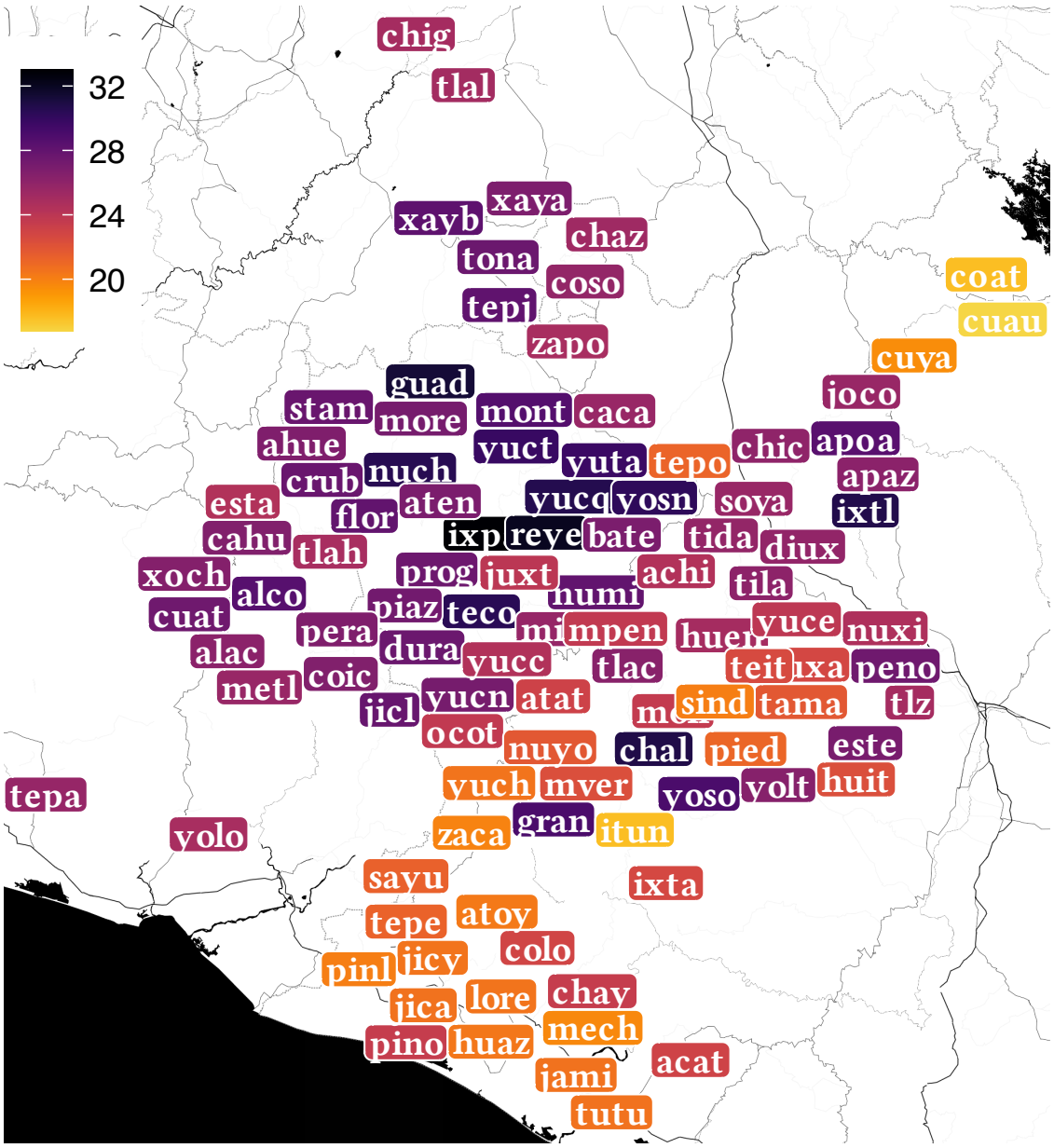
```
changes_total_table
```

DOCULECT	NumAll	PropAll	NumVow	PropVow	NumCons	PropCons
SantaAnaCuauhtemocMixtec	35	16.8	20	9.4	13	6.0
SanJuanCoatzospamMixtec	37	17.7	17	8.0	17	7.8
SantaCruzItundujiaMixtec	39	17.7	21	9.5	16	7.3
CuyamecalcoVillaZaragozaMixtec	42	19.5	21	9.7	19	8.6
SantaCatarinaMechoacanMixtec	43	19.7	28	12.8	13	5.9
SantaMariaZacatepecMixtec	43	19.9	34	15.5	9	4.1
PinotepaDonLuisMixtec	44	20.1	34	15.5	8	3.6
SanMateoSindihuiMixtec	41	20.1	18	8.5	21	9.8
SanPedroAtoyacMixtec	44	20.3	34	15.7	8	3.6
SantaMariaHuazolotitlanMixtec	44	20.4	29	13.4	12	5.4
SanLorenzoMixtec	45	20.5	30	13.6	12	5.4
SanPedroJicayanMixtec	44	20.5	35	16.3	7	3.2
SantaMariaYucuhitiMixtec	43	20.5	24	11.1	17	7.9
SanPedroTututepecMixtec	45	20.6	28	12.8	15	6.8
SantaMariaJicaltepecMixtec	45	20.6	36	16.5	7	3.2
SantiagoJamiltepecMixtec	45	20.6	30	13.7	13	5.9
SanMiguelPiedrasMixtec	46	21.1	20	9.0	23	10.5
SanPedroySanPabloTeposcolula1600Mixtec	47	21.2	29	13.1	16	7.2
SanAntonioTepetlapaMixtec	46	21.3	34	15.7	10	4.5
SanFranciscoSayultepecMixtec	46	21.4	35	16.1	10	4.5
SanJuanTamazolaMixtec	45	21.8	22	10.3	21	9.8
SantiagoNuyooMixtec	47	21.8	26	11.9	19	8.6
SanAntonioHuitepecMixtec	47	22.3	24	11.1	19	8.8
SantaLuciaMonteverdeMixtec	48	22.3	25	11.5	21	9.5
SanJuanTeitaMixtec	49	22.6	30	13.7	17	7.7
SantiagoIxtayutlaMixtec	50	22.7	34	15.5	14	6.3
SantoDomingoNuxaaMixtec	50	22.8	28	12.7	20	9.1
SanJuanColoradoMixtec	50	22.9	34	15.5	14	6.3
SantaMariaAcatepecMixtec	50	23.0	28	12.8	20	9.0
SanEstebanAtatlahucaMixtec	50	23.1	29	13.2	19	8.7
SanAgustinChayucoMixtec	51	23.5	29	13.2	19	8.6
SantiagoPinotepaNacionalMixtec	50	23.5	35	16.4	12	5.4
SanPedroMolinosMixtec	51	23.6	34	15.5	15	6.8
MagdalenaPenascoMixtec	52	23.7	34	15.3	16	7.3
SantoTomasOcotepecMixtec	52	23.7	31	14.0	20	9.1
SanBartolomeYucuaneMixtec	52	23.9	33	15.0	17	7.7
SantiagoJuxtlahuacaMixtec	53	24.0	35	15.8	16	7.2
SanMiguelAchiutlaMixtec	52	24.4	31	14.3	19	8.7
SanMartinEstadoMixtec	54	24.5	39	17.6	14	6.3
SantaMariaYucunicocoMixtec	52	24.5	38	17.5	13	6.0
SanAndresNuxinoMixtec	52	24.6	28	13.1	22	10.0
TlahuapaMixtec	54	24.9	39	17.9	14	6.3

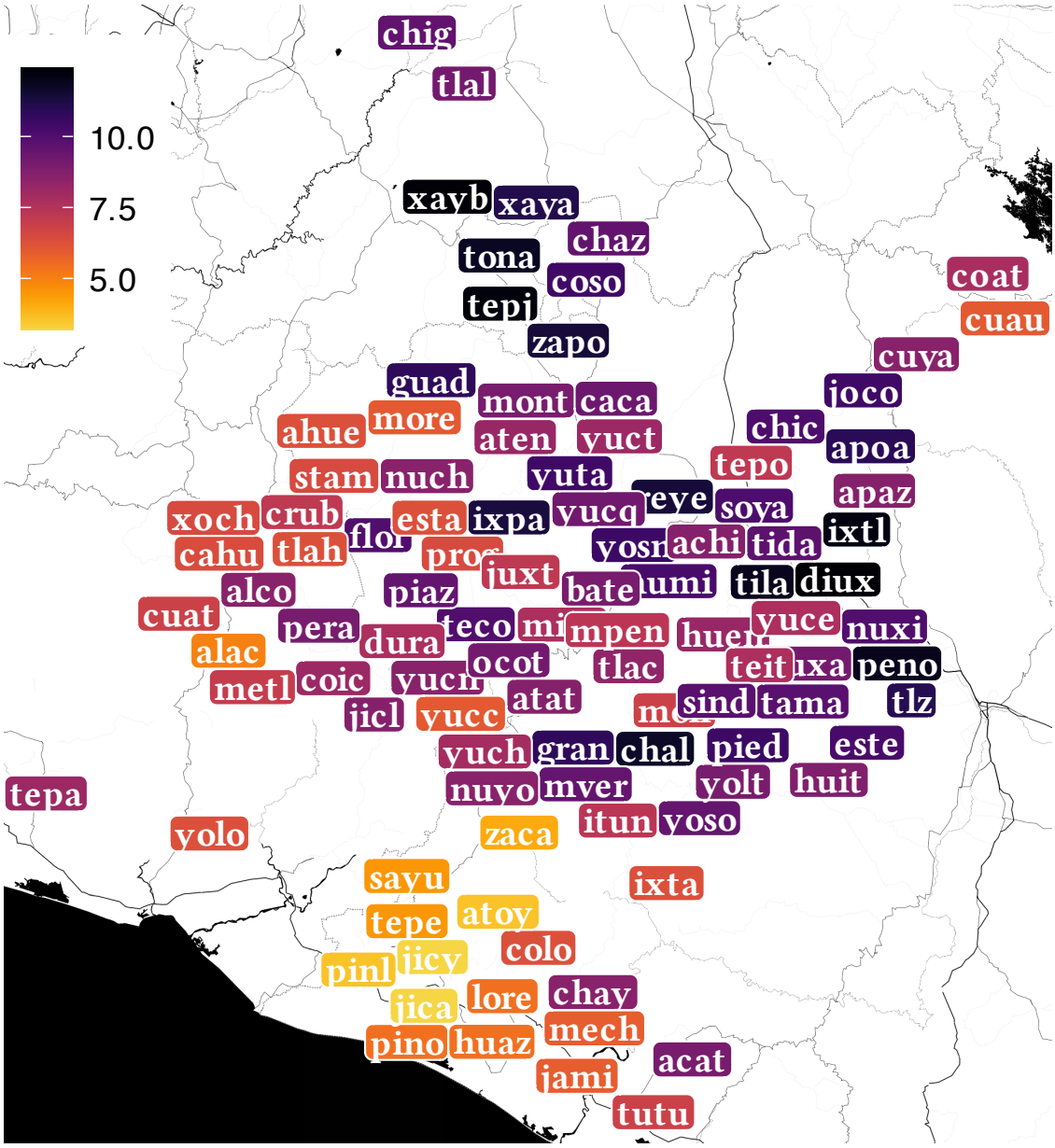
YoloxochitlMixtec	55	24.9	40	18.0	14	6.3
ZapotitlanPalmasMixtec	53	24.9	27	12.5	25	11.4
SantaMariaChigmecatitlanMixtec	54	25.0	32	14.6	21	9.6
SantiagoTlazoyaltepecMixtec	53	25.0	27	12.5	24	11.0
SantoDomingoHuendioMixtec	54	25.0	34	15.5	18	8.2
SantaCatarinaTlaltempanMixtec	47	25.1	27	13.5	19	9.1
SantiagoChazumbaMixtec	55	25.3	33	14.9	21	9.6
MetlatonocMixtec	40	25.6	25	13.9	14	7.0
SanBartoloSoyaltepecMixtec	53	25.6	29	13.6	22	10.1
SanPedroJocotipacMixtec	56	25.6	30	13.6	23	10.4
SantiagoCacaloxtepecMixtec	53	25.6	31	14.6	20	9.3
SanJuanMixtepecMixtec	56	25.7	37	16.8	17	7.7
TepangoMixtec	57	25.7	39	17.6	18	8.1
CosoltepecMixtec	57	25.8	33	14.9	23	10.4
SanPedroTidaaMixtec	56	25.8	33	14.9	21	9.7
SanJuanDiuxiMixtec	52	25.9	24	11.7	27	12.4
SanMiguelChicahuaMixtec	54	26.0	29	13.6	22	10.1
SantiagoTilantongoMixtec	55	26.1	27	12.6	26	11.9
AlacatlatzalaMixtec	56	26.2	44	20.2	11	5.0
SantaMariaApazcoMixtec	56	26.2	34	15.7	19	8.7
SanAgustinTlacotepecMixtec	56	26.3	36	16.3	18	8.4
SantaMariaYolotepecMixtec	57	26.4	35	16.2	20	9.0
XochapaMixtec	52	26.5	37	18.4	14	6.5
CahuatacheMixtec	58	26.6	43	19.5	14	6.4
CoicoyanlasFloresMixtec	59	26.6	40	18.0	18	8.1
SanMartinPerasMixtec	59	26.7	38	17.2	20	9.0
SanMiguelAhuehuetitlanMixtec	59	26.7	43	19.4	14	6.3
SanJeronimoXayacatlanMixtec	52	26.8	27	13.4	24	11.1
LaBateaMixtec	59	26.9	37	16.8	20	9.0
SanAgustinAtenangoMixtec	58	26.9	38	17.4	18	8.2
SantaCatarinaEstetlaMixtec	57	27.0	33	15.1	22	10.2
SanLuisMoreliaMixtec	55	27.1	40	19.2	13	6.0
YucunaniMixtec	59	27.1	37	16.9	19	8.6
PiedraAzulMixtec	60	27.3	38	17.3	21	9.5
CuatzoquitengoMixtec	60	27.4	44	19.8	15	6.8
SantaMariaPenolesMixtec	59	27.4	32	14.5	26	11.9
SanJeronimoProgresoMixtec	56	27.5	39	18.6	14	6.5
SanMartinDuraznosMixtec	61	27.6	43	19.5	17	7.7
SantoDomingoTonahuixtlaMixtec	60	27.6	33	15.1	26	11.8
SantaCruzBravoMixtec	61	27.7	44	19.9	16	7.2
SantiagoTamazolaMixtec	57	27.7	40	18.9	14	6.5
ElJicaralMixtec	60	27.8	40	18.2	19	8.7
SanJuanNumiMixtec	59	28.0	36	16.6	22	10.2
SanMarcoslaFlorMixtec	61	28.0	38	17.4	22	10.0
AbasoloValleMixtec	59	28.1	37	17.4	20	9.1
TepejilloMixtec	62	28.1	33	14.9	27	12.2

XayacatlanBravoMixtec	61	28.2	33	15.1	27	12.3
SantiagoApoalaMixtec	61	28.4	34	15.7	24	10.9
AlcozaucaGuerreroMixtec	63	28.5	43	19.4	19	8.6
SanSebastianMonteMixtec	61	28.6	38	17.8	20	9.0
SanMiguelElGrandeMixtec	64	29.0	38	17.2	24	10.8
SantiagoYosonduaMixtec	64	29.0	41	18.5	21	9.5
SanAndresYutatioMixtec	65	29.7	39	17.8	23	10.4
YucunutiBenitoJuarezMixtec	64	29.8	43	19.8	18	8.1
SanPedroYosonamaMixtec	55	30.2	32	16.3	22	10.5
SanSebastianTecomaxtlahuacaMixtec	65	30.4	41	19.0	22	10.0
SanJorgeNuchitaMixtec	67	30.5	45	20.3	19	8.6
YucuquimiOcampoMixtec	60	30.6	37	18.5	20	9.2
SantiagoIxtaltepecMixtec	66	30.7	37	16.7	26	12.1
ChalcatongoHidalgoMixtec	68	30.8	40	18.0	26	11.7
GuadalupeVillahermosaMixtec	68	31.3	41	18.9	24	10.8
SantosReyesTepejilloMixtec	70	32.0	42	19.0	25	11.4
IxpantepecNievesMixtec	73	33.0	46	20.8	25	11.3

```
# map with proportion of changes
map_changes_overall <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(metadata, is.na(AudersetGroupC)), size = 2.5, fontface = "bold") +
  geom_label_repel(aes(x = Longitude, y = Latitude, label = MapAbbr, fill = PropAll),
  data = changes_total, size = 3.5, fontface = "bold", family = "Linux Libertine",
  max.overlaps = 40, box.padding = 0, point.padding = 0.1, label.padding = 0.1, color =
  "white") +
  scale_fill_viridis(option = "inferno", end = 0.9, direction = -1) +
  theme_map() +
  theme(legend.title=element_blank(),
        legend.text=element_text(size=9),
        legend.key.size = unit(1,"line"),
        legend.position = c(0, 0.7))
map_changes_overall
```

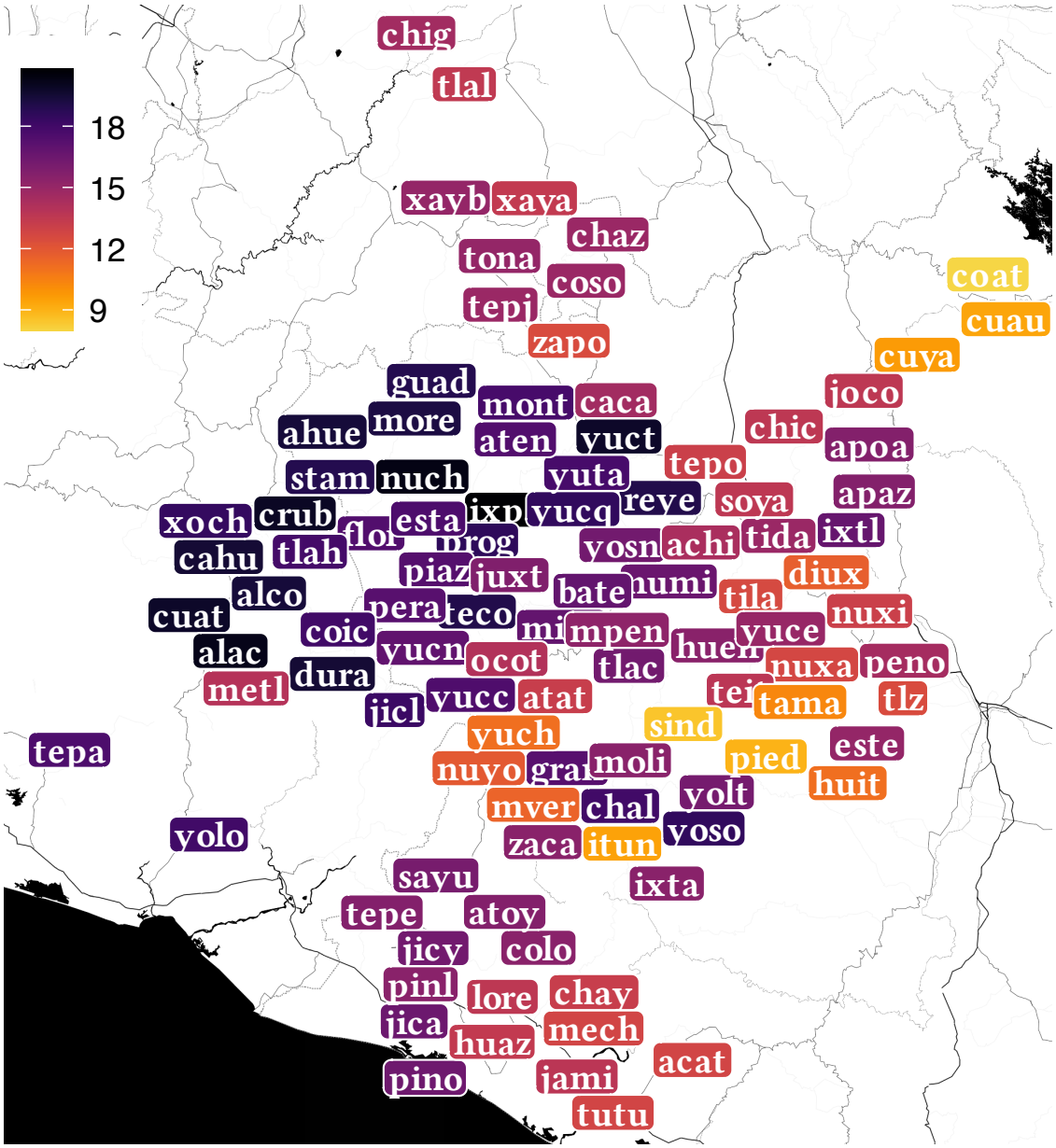



```
# map with total number of changes in consonants
map_changes_cons <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(metadata, is.na(AudersetGroupC)), size = 2.5, fontface = "bold") +
  geom_label_repel(aes(x = Longitude, y = Latitude, label = MapAbbr, fill = PropCons),
  data = changes_total, size = 3.5, fontface = "bold", family = "Linux Libertine",
  max.overlaps = 40, box.padding = 0, point.padding = 0.1, label.padding = 0.1, color =
  "white") +
  scale_fill_viridis(option = "inferno", end = 0.9, direction = -1) +
  theme_map() +
  theme(legend.title=element_blank(),
        legend.text=element_text(size=9),
        legend.key.size = unit(1,"line"),
        legend.position = c(0,0.7))
map_changes_cons
```




```
# map with total number of changes in vowels
map_changes_vow <- ggmap(mixtec_base) +
  geom_text(aes(x = Longitude, y = Latitude, label = Abbreviation), data =
  filter(metadata, is.na(AudersetGroupC)), size = 2.5, fontface = "bold") +
  geom_label_repel(aes(x = Longitude, y = Latitude, label = MapAbbr, fill = PropVow),
  data = changes_total, size = 3.5, fontface = "bold", family = "Linux Libertine",
  max.overlaps = 40, box.padding = 0, point.padding = 0.1, label.padding = 0.1, color =
  "white") +
  scale_fill_viridis(option = "inferno", end = 0.9, direction = -1) +
  theme_map() +
  theme(legend.title=element_blank(),
        legend.text=element_text(size=9),
        legend.key.size = unit(1,"line"),
```

```
        legend.position = c(0,0.7))
map_changes_vow
```



References

Kahle, David & Wickham, Hadley (2013). ggmap: Spatial Visualization with ggplot2. The R Journal 5(1), pp. 144–161. https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf