Numeral Classifiers in Yucatec Maya: Microvariation and syntactic change

Supplementary Material

Barbara Blaha and Stavros Skopeteas

September 28, 2024

Contents

1 Global

1.1 Libraries

data processing

```
  library(readxl)
  library(writexl)
  library(reshape2)
  library(scales)
  library(dplyr)
  library(stringr)
  library(showtext)#fonts
  library(kableExtra)
  library(broom)
```

graphs and maps

```
  library(ggplot2)
  library(ggmap)
  library(ggspatial)
```

statistics

```
  library(mgcv)
  library(mgcViz)
  library(itsadug)
```

1.2 Settings

```
#setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
theme_set(theme_bw())
#setting the size of annotations
annosize = 6
#downloads google font for graphs
font_add_google("Martel Sans", family = "special")
showtext_auto()
```

1.3 Loading data

```
load(file='YUC-CL.rda')
```

2 Base map

This function creates a base map of Yucatán, based on the GADM data (https://gadm.org)

```r
yuc.map.bare <- ggplot(data = gadm) +
  geom_sf(fill = "#faf9f8",color = "grey30",size=.1) + #FloralWhite
  theme(text = element_text(family = "special"),
        axis.text = element_text(size = 18, color="grey40"),
        axis.title = element_text(size = 22))+
  coord_sf(xlim = c(-91, -86.7), ylim = c(18.1, 21.8), expand = T)+
  scale_x_continuous(labels = abs,breaks = seq(-180,180,by=1),
                     minor_breaks = seq(-180,0,by=.1))+#,breaks = seq(-91,-87,1),
  scale_y_continuous(labels = abs,breaks = seq(0, 100, by=1),
                     minor_breaks = seq(0,100,by=0.2))+
  annotation_north_arrow(location = "br", which_north = "true",
                         height = unit(0.7, "cm"),
                         width = unit(0.7, "cm"),
                         pad_x = unit(0.02, "in"),
                         pad_y = unit(0.1, "in"),
                         style = north_arrow_fancy_orienteering(line_width = 0.4,
                                                                line_col = "black",
                                                                fill = c("white", "grey30"),
                                                                text_col = "black",
                                                                text_size = 12))+
annotation_scale(location = "br",
                 width_hint = 0.12,
                 line_width = 0.2,
                 bar_cols = c("grey30", "white"),
                 pad_y = unit(0.1, "cm"),
                 pad_x = unit(0.1, "cm"),
                 text_cex = 1,
                 height = unit(0.1, "cm"))+
  labs(x="Longitude (°W)",y="Latitude (°N)")

yuc.map <- yuc.map.bare+
  annotate("text",x=-90.4,y=20.05, label = "Camino\nReal",
           size = annosize, lineheight = 0.3,family="special")+
  annotate("text",x=-88,y=21.1, label = "Northeast",
           size = annosize, family="special")+
  annotate("text",x=-89.9,y=20.8, label = "Nortwest",
           size = annosize, family="special")+
  annotate("text",x=-89.9,y=19.2, label = "Los Ch'e'enes",
           size = annosize, family="special")+
  annotate("text",x=-89.1,y=20, label = "Southern\nYucatán",
           lineheight = 0.3,size = annosize)+
  annotate("text",x=-88.4,y=19.6, label = "Central\nQR",
           lineheight = 0.3,size = annosize)+
  annotate("text",x=-88.3,y=18.2, label = "BELIZE",size = annosize)+
  annotate("text",x=-88.6,y=21.3, label = "YUCATAN",size = annosize)+
  annotate("text",x=-88.4,y=19, label = "QUINTANA ROO",size = annosize)+
  annotate("text",x=-90,y=18.8, label = "CAMPECHE",size = annosize)
yuc.map
```

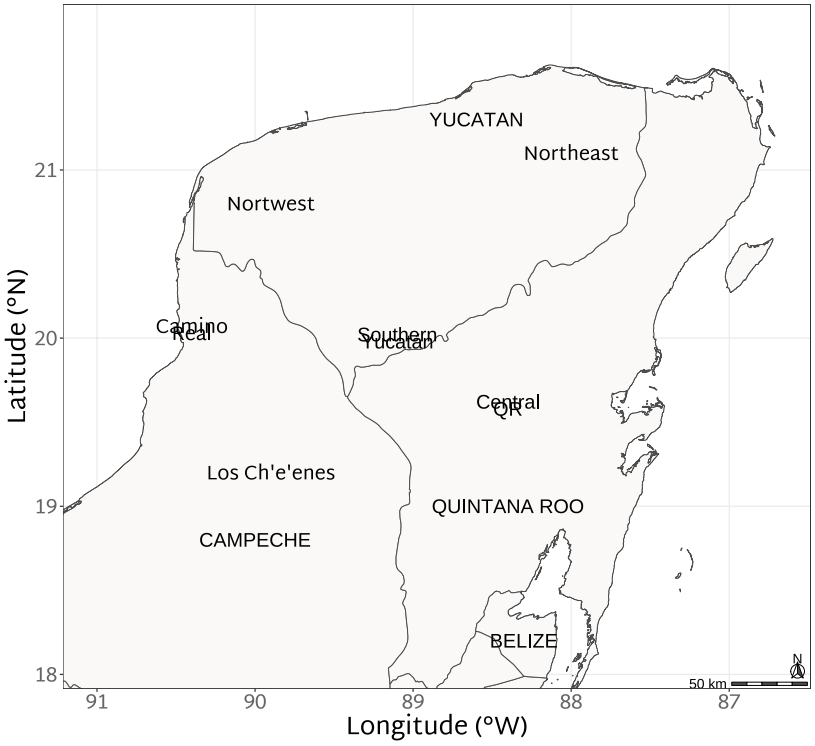

Figure 1: Figure: Base map

3 Functions

3.1 Map graphs

The following function creates plots of gradient features on the base map. The maps contain 5 colors on a scale. The use of colors depends on the distribution of the data points in the dataset (see median-based distribution in p1...p5)

```
plot.map.gradient <- function(data)
  {
       caption <- substitute(data)
       data$Value.p <- rescale(data$Value.p)
       p5 <- 1
       p4 <- 1-(1-median(data$Value.p))/2
       p3 <- median(data$Value.p)
       p2 <- median(data$Value.p)/2
       p1 <- 0
       sum.plot <- yuc.map +
         theme(text = element_text(family = "special",size=14))+
               geom_point(data = data,
                  aes(x = Long, y = Lat,fill=Value.p),
                  size=3,
                  alpha=.9,
                  shape=21)+
         theme(legend.key.size = unit(0.1, 'lines'),
               legend.text = element_text(size=18,vjust= 1,lineheight = 0),
```

```
              legend.title = element_text(lineheight = 0.4, size=19,angle=0,hjust= 1),
              legend.position="top",
              legend.justification="right",
              legend.box.background = element_rect(color="grey30", size=.1)
              )+
        scale_fill_gradientn(values=c(p5, p4, p3, p2, p1),
                        colours=c("#663300", "#996633", "#CCFFCC", "#99FF33", "#66CC00"),
                        guide = guide_colourbar(direction = "horizontal",
                                                frame.colour="grey",
                                                barheight = .7,
                                                barwidth = 4,
                                                ticks = T,
                                                title.position = "left"),
                           breaks=c(0,1),
                           name="% general classifier,\n rescaled to [0,1]"
                          )


    print(sum.plot)
}
```

3.2 Statistics

3.2.1 Plotting GAM coefficients

```
      my.cex = 2
      my.cex.lab = 3

plot.gam <- function(model) {
    caption = substitute(model)
      vis.plot <- vis.gam(model,
                          plot.type="contour",
                          color="terrain",
                          too.far=0.15,
                          view=c("Long","Lat"),
                          main="",
                          cex.axis=2,
                          cex.lab=my.cex.lab
                          )
    text(-89.8,21, label = "Nortwest",cex=my.cex)
    text(-90.3,y=20.05, labels = "Camino Real", xpd=TRUE,cex=my.cex)
    text(-88,21.2, labels = "Northeast", xpd=TRUE,cex=my.cex)
    text(-89.8,19.2, labels = "Los Ch'e'enes", xpd=TRUE,cex=my.cex)
    text(x=-89.1,y=20, labels = "Southern Yucatán", xpd=TRUE,cex=my.cex)
    text(x=-88.4,y=19.6, labels = "Central QR", xpd=TRUE,cex=my.cex)
    text(-88.3,18.5, labels = "Belize", xpd=TRUE,cex=my.cex)

    print(vis.plot)
}
```

Plotting GAM co-efficients with interactions of the smooth term

```
plot.gam.2 <- function(model,condition="") {
    caption = substitute(model)
    caption.condition = substitute(condition)
        vis.plot <- vis.gam(model,
                            plot.type="contour",
                            color="terrain",
                            cond=list(Subtype=condition),
                            too.far=0.15,
                            view=c("Long","Lat"),
                            main="",
                            cex.axis=3,
                            cex.lab=4
                            )
                        text(-89.8,21, label = "Nortwest",cex=my.cex)
                        text(-90.3,y=20.05, labels = "Camino Real", xpd=TRUE,cex=my.cex)
                        text(-88,21.2, labels = "Northeast", xpd=TRUE,cex=my.cex)
                        text(-89.8,19.2, labels = "Los Ch'e'enes", xpd=TRUE,cex=my.cex)
                        text(x=-89.1,y=20, labels = "Southern Yucatán", xpd=TRUE,cex=my.cex)
                        text(x=-88.4,y=19.6, labels = "Central QR", xpd=TRUE,cex=my.cex)
                        text(-88.3,18.5, labels = "Belize", xpd=TRUE,cex=my.cex)
    print(vis.plot)
}
```

3.2.2 Density graph

```
plot.density <- function(data,xlab = "n responses by speaker",bin = 2,xlim1=0,xlim2)
{
  caption <- substitute(data)
  density.plot <- ggplot(data,aes(Freq)) +
    geom_histogram(aes(y=..density..),binwidth=bin, colour="black", fill="grey")+
    xlim(xlim1,xlim2)+
    geom_density(alpha=.1, fill="blue")+
    geom_vline(aes(xintercept=mean(Freq)),
               color="darkred", linetype="dashed", size=1)+
    labs(x=paste0(xlab, " (mean = ",
                  round(mean(data$Freq)),")"))+
    theme(text = element_text(size = 30, family = "special"))

  print(density.plot)
}
```

3.2.3 Scatter plots

```
plot.scatter <- function(data)
  {
  myplot <- ggplot(result.type.2, aes(x=sum, y=percent,fill=Subtype)) +
    geom_smooth(method=lm, aes(color=Subtype,fill=Subtype),
                level=.50,linetype="dotted",se=F)+
    geom_jitter(aes(fill=Subtype),size=3, shape=21, width=0.5,height=0.5,alpha=0.7)+
    scale_x_continuous(trans = 'log2')+
```

```
    scale_colour_manual(values=c("grey","black","lightblue"),name = "Measure Type")+
    scale_fill_manual(values=c("grey","black","lightblue"),name = "Measure Type")+
    ylab("% multiple classifiers")+
    xlab("Frequency per classifier (log-scale)")+
    theme(axis.text = element_text(lineheight = 0.5, size=20),
          axis.title = element_text(lineheight = 0.5, size=24))+
    theme(legend.text = element_text(lineheight = 0,size=20),
        legend.title = element_text(size=24))+
    theme(legend.position="right",
          legend.box.background = element_rect(colour = "black"))
    print(myplot)

}
```

4 Data processing

For the following analysis, all relevant data was merged in the dataframe "dta":

4.1 Labels

4.1.1 Prompts

- **Question_ID**: question identifier
- **Question**: Form of the question
- **Q_Label**: label of the question (including Question_ID)
- **Translation**: translation of the question

4.1.2 Prompt Groups

- **Type**: division of prompts in: sort | measure
- **TypePS**: division of prompts in: sort | part | portion & sum
- **TypePP**: division of prompts in: sorts | part & portion | portion
- **Subtype**: division of prompts in: sorts | part | portion | sum

4.1.3 Speakers

- **Speaker_ID**: speaker identifier
- **Birthyear**: year of speaker's birth
- **Gender**: M=male, F=female
- **Spanish_odds**: Spanish:Maya ratio in speaker's data (in AYM)
- **Long.jitter**: latitude of Location with jitter for each speaker (to avoid overplotting speakers of the same location)
- **Lat.jitter**: latitude of Location with jitter for each speaker (to avoid overplotting speakers of the same location)

4.1.4 Locations

- **Location**: speaker's place of living
- **Lat**: latitude of Location

- **Long**: longitude of Location
- **Pop.size**: population size of Location
- **P.yuc**: p of indigenous speakers in Location

4.1.5 Responses

- **Form**: transcribed response
- **Value**: classifier
- **Variant**: variant. Annotations of the numeric expression:
  - **CL.Gen**: general classifier p'éel;
  - **CL.Spec**: specific classifier, e.g., *túul* 'CL.ANIM', *kúul* 'CL.PLANT', *ts'íit* 'CL.LONG', *maataj* 'stem' (measure noun from Spanish), *che'* 'tree', *xéek* 'CL.FOOT_OF_TREE';
  - **CL.Gen CL.Spec**: general classifier followed by specific classifiers;
  - **NV**: non-valid;
  - **NV_noData**: no data available;
  - **NV_noNoun**: no noun available;
  - **NV_noNum**: no numeric expression, e.g., un hombre rendered as *e máako'* ; intended numeric expressions translated with the vague quantifier *ya'ab* 'much/many';
  - **NV_noClass**: no classifier available;
  - **NV_Spanish**: Spanish syntax;
  - The same lexical item may occur as CL.Spec and as N. It is coded as CL.Spec if it appears in a classifier construction, i.e., it is accompanied by an N without bearing possessive morphology, e.g., *jum p'éel maata wayúum* (one CL.UNIT stem huaya) 'one stem huaya'; the same lexical item is the head N of an NP if it bears possessive morphology, e.g., *jum p'éel u maata-il wayúum* (one CL.UNIT A.3 stem-REL huaya) 'one stem of huaya'.
  - In case of multiple responses: the annotation corresponds to the last valid answer.
- **Comment**: explanations for Non-Valid data
- **Structure**: structural representation [Annotation of the structure of the numeral expression at issue:
  - **A**: set A person marker;
  - **Adj**: Adjective (annotated in cases that it needs to be distinguished from classifiers), e.g., Q167 *um p'ée wolis sakan* (one CL.UNIT round dough); also deverbal adjectives in *-bil* and *-Vkbal*, which are not classifiers
  - **Adj.Size**: size modifier (e.g., chan): only these modifiers are annotated, to test how often they intervene between numerals and classifiers;
  - **CL.Gen**: general classifier *p'éel* (CL.UNIT);
  - **CL.Spec**: specific classifier (sortal/mensural) or measure noun;
  - **D**: Determiner;
  - **N**: noun;
  - **Num**: numeral;
  - **(Num)**: phonologically zero numeral (only applies to 'one').
  - **Rel**: relationalizer *-il/-ij/-i* (evidence for a possessive construction);
  - In case of multiple responses: the annotation corresponds to the last valid answer; : preposition;
- **Origin**: origin of the classifier (Maya, Spanish) [Origin of classifier: this annotation refers to the leftmost specific classifier:
  - **Spanish**: Spanish origin, e.g., *maataj* 'stem', *fiilaj* 'row';
  - **Maya**: Maya origin, e.g., *p'éel* 'CL.Unit', *túul* 'CL.Anim',
  - **NV**: not applicable if no specific classifier appears.

4.2 Data

```
kable(head(dta)) %>%
  kable_styling("striped", full_width = F) %>%
  scroll_box(width = "100%", height = "400px")
```

Question_ID	Form	Speaker_ID	Question	Birthyear	G
Q111	bisa' jun tu' mak' ich sajka'	S023	llevaron un hombre a una cueva	1933	M
Q111	bisa' u p'éej máak te' kueeva	S115	llevaron un hombre a una cueva	1985	F
Q111	bisa'a' jun túu máak ti' um p'ée saajkab	S149	llevaron un hombre a una cueva	1965	M
Q111	bisa'a' p'e winik ti' u p'e sajka'	S003	llevaron un hombre a una cueva	1931	M
Q111	bisa'a' u p'ée máak ti' u p'ée kueeva	S074	llevaron un hombre a una cueva	1947	M
Q111	bisa'a' u túu máak' te' sajka'	S168	llevaron un hombre a una cueva	1960	M

4.3 Sample

```
speakers.df <- unique(select(dta,Speaker_ID,Birthyear,Gender))
speakers.n <- nrow(speakers.df)

#locations
locations.n <- length(unique(dta$Location))

#age
speakers.sum <- summary(speakers.df$Birthyear)
speakers.gender.sum <- summary(as.factor(speakers.df$Gender))

#summary
speakers.summary <- data.frame(info=c("total speakers",
                                      "female speakers",
                                      "locations",
                                      "birthyear: min",
                                      "birthyear: max",
                                      "birthyear: mean",
                                      "birthyear: median"),
          n=round(c(speakers.n,speakers.gender.sum[2],locations.n,
                    speakers.sum[1],speakers.sum[6],speakers.sum[4],speakers.sum[3]),0))

  kable(speakers.summary) %>%
  kable_styling("striped", full_width = F)
```

info	n
total speakers	173
female speakers	116
locations	85
birthyear: min	1906
birthyear: max	1987
birthyear: mean	1953

birthyear: median 1953

4.4 Responses

```
questions.n <- length(unique(dta$Question_ID))
speakers.n <- length(unique(dta$Speaker_ID))
permutations.n <- speakers.n * questions.n
responses.n <- nrow(dta)
responses.p <- responses.n*100/permutations.n

#table
responses.summary <- data.frame(info=c("n questions","n speakers","n questions*speakers",
                                       "n responses","% responses (out of n permutations)"),
           n=round(c(questions.n,speakers.n,permutations.n,responses.n,responses.p),0))

  kable(responses.summary) %>%
  kable_styling("striped", full_width = F)
```

info	n
n questions	23
n speakers	173
n questions*speakers	3979
n responses	3446
% responses (out of n permutations)	87

4.4.1 Non-valid data

```
nv <- data.frame(xtabs(~Variant+Comment,dta))

  kable(nv) %>%
  kable_styling("striped", full_width = F)
```

Variant	Comment	Freq
CL.Gen	Construction at issue: measure not as classifier	0
CL.Gen CL.Spec	Construction at issue: measure not as classifier	0
CL.Spec	Construction at issue: measure not as classifier	0
NV	Construction at issue: measure not as classifier	49
CL.Gen	Construction at issue: no overt noun	0
CL.Gen CL.Spec	Construction at issue: no overt noun	0
CL.Spec	Construction at issue: no overt noun	0
NV	Construction at issue: no overt noun	225
CL.Gen	Construction at issue: Spanish priming	0
CL.Gen CL.Spec	Construction at issue: Spanish priming	0
CL.Spec	Construction at issue: Spanish priming	0
NV	Construction at issue: Spanish priming	9

CL.Gen	Content at issue: deviating from prompt	0
CL.Gen CL.Spec	Content at issue: deviating from prompt	0
CL.Spec	Content at issue: deviating from prompt	0
NV	Content at issue: deviating from prompt	48
CL.Gen	Content at issue: no expression of measure	0
CL.Gen CL.Spec	Content at issue: no expression of measure	0
CL.Spec	Content at issue: no expression of measure	0
NV	Content at issue: no expression of measure	29
CL.Gen	Content at issue: no num-classifier construction	0
CL.Gen CL.Spec	Content at issue: no num-classifier construction	0
CL.Spec	Content at issue: no num-classifier construction	0
NV	Content at issue: no num-classifier construction	282

```
nv.n <-  sum(nv$Freq)
nv.content.n <-  sum(nv[grepl("Content at issue", nv$Comment),]$Freq)
nv.construction.n <-  sum(nv[grepl("Construction at issue", nv$Comment),]$Freq)

nv.summary <- data.frame(info=c("n non-valid",
                                "n non-valid: content",
                                "n non-valid: construction"),
           n=round(c(nv.n,nv.content.n,nv.construction.n),0))

  kable(nv.summary) %>%
  kable_styling("striped", full_width = F)
```

info	n
n non-valid	642
n non-valid: content	359
n non-valid: construction	283

excluding non valid data; dataset contains:

```
dta.v1 <- droplevels(subset(dta,Variant != "NV"))
  nrow(dta.v1)
```

```
## [1] 2803
```

responses with Spanish classifiers

```
#spanish
dta.spa <- droplevels(subset(dta.v1,Origin == "Spanish"))
  nrow(dta.spa)
```

```
## [1] 152
```

excluding responses with Spanish classifiers

```
dta.v2 <- droplevels(subset(dta.v1,Origin != "Spanish"))
```

excluding speakers who do not have at least one valid token for sortal and one valid token for mensural classifiers

```
  speakers.n <- length(unique(dta.v2$Speaker_ID))
  v.speaker <- data.frame(xtabs(~Type+Speaker_ID,dta.v2))
  v.speaker.cast <- dcast(v.speaker, ... ~ Type,value.var = "Freq")
  nv.speakers <- v.speaker.cast[v.speaker.cast$sort == 0 |
                                  v.speaker.cast$measure == 0, ]$Speaker_ID

dta.v <- droplevels(subset(dta.v2,!(Speaker_ID %in% nv.speakers)))
```

4.4.2 Valid data

n and % valid

```
  valid.n <- nrow(dta.v)
  valid.prop <-   100*valid.n/responses.n
  valid.n
```

```
## [1] 2645
```

n valid responses per question

```
  questions.valid <- data.frame(xtabs(~Question_ID,dta.v))
  summary(questions.valid$Freq)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    64.0    88.0   127.0   115.0   136.5   166.0
```

```
  plot.density(questions.valid, xlab = "n responses by question",bin = 10,xlim2=180)
```

n valid responses per speaker

```
  speakers.valid <- data.frame(xtabs(~Speaker_ID,dta.v))
  summary(speakers.valid$Freq)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00   13.00   16.00   15.65   19.00   23.00
```

```
  plot.density(speakers.valid,xlim2=25)
```

4.4.3 classes of prompts

Prompt labels are listed for Appendix I

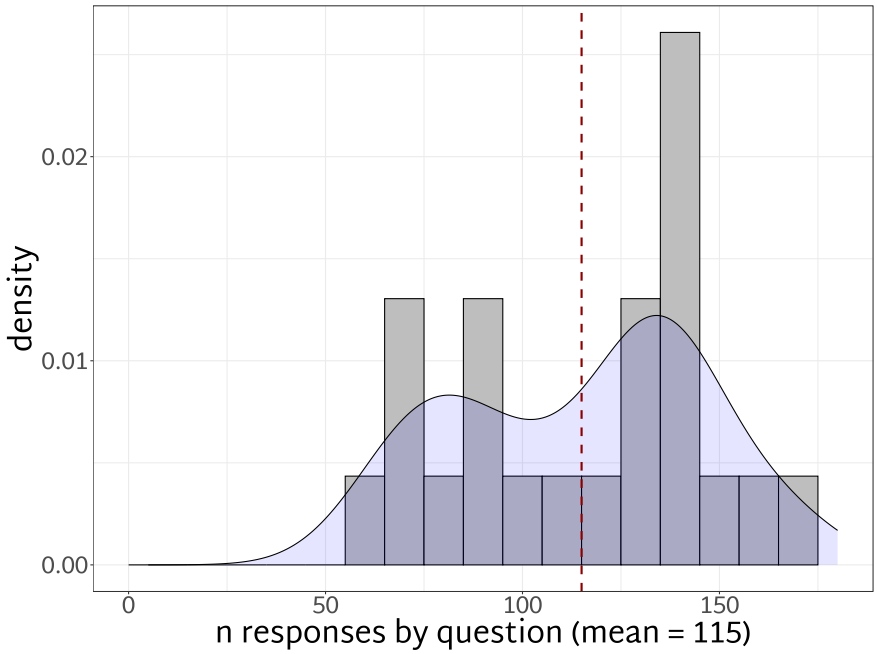

Figure 2: Figure: Histogramm per questions

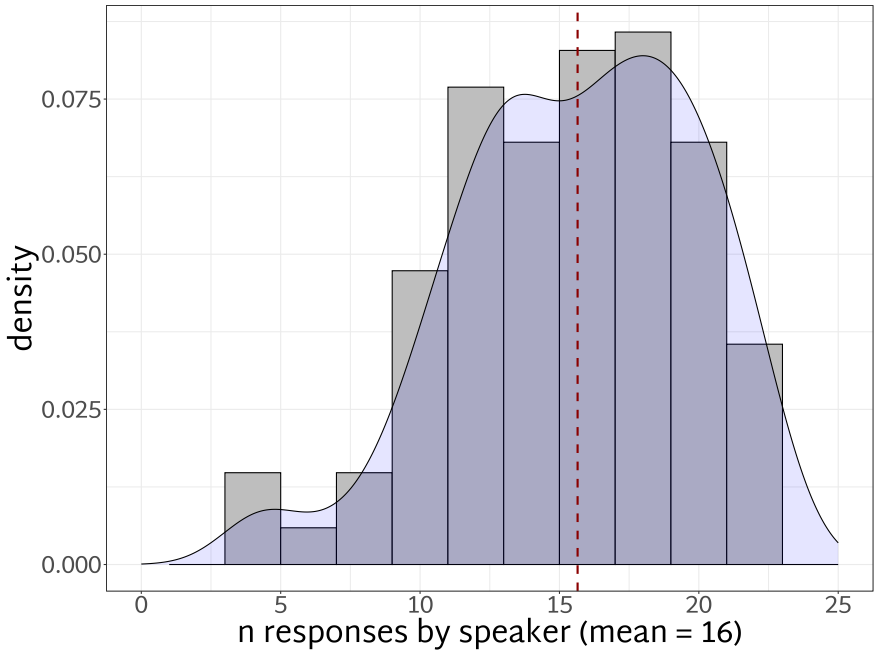

Figure 3: Figure: Histogramm per speakers

```
kable(unique(dta.v[dta.v$Subtype =="sort",]$Question_Label)) %>%
  kable_styling("striped", full_width = F)
```

x
Q111: llevaron un hombre a una cueva (they took a man to a cave)
Q163: un cochino (one/a pig)
Q165: una mata de huaya (one/a huaya plant)
Q166: una vela (one/a candle)

```
kable(unique(dta.v[dta.v$Subtype =="portion",]$Question_Label)) %>%
  kable_styling("striped", full_width = F)
```

x
Q167: una bola de masa (recado) (one/a ball of dough)
Q170: un puño de cemento, fertilizante (one/a fistful of cement, fertilizer)
Q178: un montón (de excremento) (one/a pile (of excrement))
Q180: dos tragos de agua (two shots of water)
Q187: una gota de agua (one/a drop of water)

```
kable(unique(dta.v[dta.v$Subtype =="part",]$Question_Label)) %>%
  kable_styling("striped", full_width = F)
```

x
Q171: una mitad de sandía (one/a half of watermelon)
Q172: un pedazo de piedra, vidrio, plástico (one/a piece of stone, glass, plastic)
Q173: un pedazo quebrado de madera (one/a broken piece of wood)
Q177: un pedazo rasgado de ropa (one/a torn piece of clothing)
Q1831: dos rebanadas de queso (two slices of cheese)
Q185: un pedazo de ropa (one/a piece of clothing)
Q186: un pedazo cortado de madera (one/a cut piece of wood)

```
kable(unique(dta.v[dta.v$Subtype =="sum",]$Question_Label)) %>%
  kable_styling("striped", full_width = F)
```

x
Q169: una fila de piedras (one/a row of stones)
Q174: un montón, estiba de papel (one/a pile, stack of paper)
Q179: una carga de leña (one/a load of firewood)
Q181: un manojo de hierbas (one/a bunch of herbs)
Q182: una doblada de ropas (one/a folded bundle of clothes)
Q184: dos pesadas de frijol (two weights of beans)
Q272: un amarre (arrollo) de frijol (a mooring (arrollo) of beans)

5 Descriptive stats

Creating two subsets of the data – for sortal and mensural classifiers:

```
  dta.v.sort <- dta.v[dta.v$Type =="sort",]
  dta.v.meas <- dta.v[dta.v$Type =="measure",]
```

5.1 Sortal classifiers

5.1.1 Frequencies

reporting frequencies of sortal classifiers

```
  sortal_table <- dta.v.sort %>%
    group_by(Variant) %>%
    summarise(frequency = n()) %>%
    mutate(percentage = round(frequency*100 / sum(frequency),1))

kable(sortal_table) %>%
  kable_styling("striped", full_width = F)
```

Variant	frequency	percentage
CL.Gen	261	45.3
CL.Gen CL.Spec	8	1.4
CL.Spec	307	53.3

excluding multiple classifiers from this analysis

```
  dta.v.sort <- droplevels(subset(dta.v.sort, Variant != "CL.Gen CL.Spec"))
  dta.v.sort$Variant <- factor(dta.v.sort$Variant, levels = c("CL.Spec","CL.Gen"))
```

5.1.2 Use of the general classifier

Aggregating by speaker and by location and ploting

```
# per speaker
  dta.v.sort$Substitute.p <- as.numeric(as.factor(dta.v.sort$Variant))-1
  dta.v.sort.aggr.speaker <-aggregate(Substitute.p ~ Speaker_ID+Long+Lat,
                                data= dta.v.sort[,c("Speaker_ID","Lat","Long",
                                                    "Substitute.p")],FUN=mean)

# per location
  dta.v.sort.aggr <-aggregate(Substitute.p ~ Location+Long+Lat,
                                data= dta.v.sort[,c("Location","Lat","Long",
                                                    "Substitute.p")],FUN=mean)
  summary(dta.v.sort$Substitute.p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4595  1.0000  1.0000
```

```r
nrow(dta.v.sort.aggr)
```

```
## [1] 85
```

```r
dta.v.sort.aggr$Value.p <- dta.v.sort.aggr$Substitute.p
```

```r
  plot.map.gradient(dta.v.sort.aggr)
```

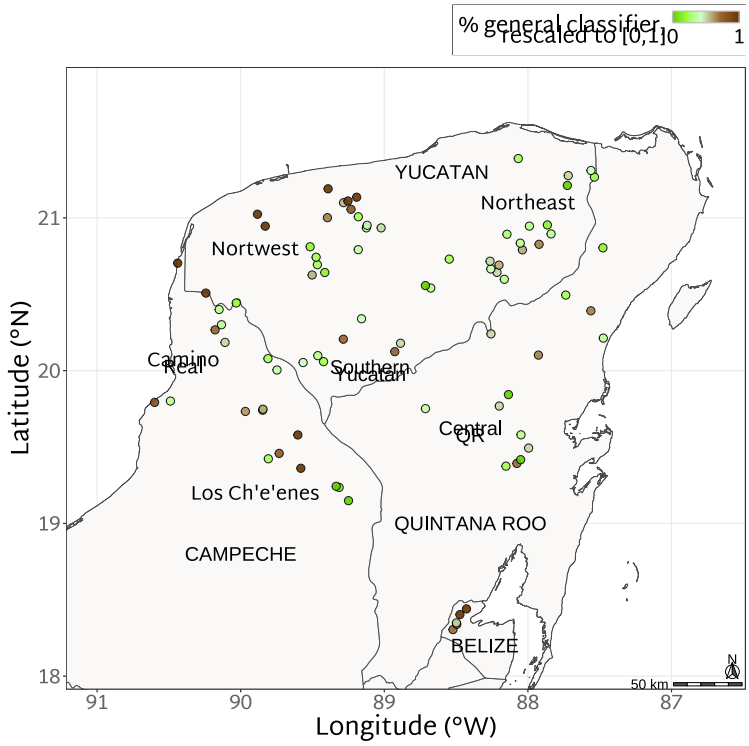

Figure 4: Figure: Sortal classifiers on base map

5.1.3 List of sortal classifiers

```r
dta.v.sort$sortalCL <- dta.v.sort$Value
dta.v.sort$sortalCL <- gsub(" .*","",dta.v.sort$sortalCL)

sortal.classifiers <- as.data.frame(xtabs(~sortalCL, dta.v.sort))
kable(sortal.classifiers) %>%
kable_styling("striped", full_width = F)
```

sortalCL	Freq
che’	1
kúul	60
p’éel	261

ts'íit	43
túul	200
xéek	3

```
  sum(sortal.classifiers$Freq)
```

```
## [1] 568
```

5.2 Mensural classifiers

5.2.1 Data

Creating a frequency table for mensural classifiers

```
  nrow(dta.v.meas)
```

```
## [1] 2069
```

```
## prompts
  mensural_table <- dta.v.meas %>%
    group_by(Variant) %>%
    summarise(frequency = n()) %>%
    mutate(percentage = round(frequency*100 / sum(frequency),1))
kable(mensural_table) %>%
  kable_styling("striped", full_width = F)
```

Variant	frequency	percentage
CL.Gen CL.Spec	283	13.7
CL.Spec	1786	86.3

Creating a frequency table for mensural classifiers by measure type

```
  subtype_table <- dta.v.meas %>%
    group_by(Subtype,Variant) %>%
    summarise(frequency = n())
  subtype.df <- dcast(data.frame(subtype_table), ... ~Variant)
  subtype.df$sum <- subtype.df$'CL.Gen CL.Spec'+ subtype.df$CL.Spec
  subtype.df$pCL.Gen <- round(100*subtype.df$'CL.Gen CL.Spec'/subtype.df$sum,1)
  subtype.df$pCL.Spec <- round(100*subtype.df$CL.Spec/subtype.df$sum,1)

  kable(subtype.df) %>%
  kable_styling("striped", full_width = F)
```

Subtype	CL.Gen CL.Spec	CL.Spec	sum	pCL.Gen	pCL.Spec
part	33	819	852	3.9	96.1
portion	83	410	493	16.8	83.2

sum	167	557	724	23.1	76.9

5.2.2 Use of multiple classifiers

Aggregating by speaker and by location and ploting

```
  dta.v.meas$Variant <- factor(dta.v.meas$Variant, levels = c("CL.Spec","CL.Gen CL.Spec"))

  dta.v.meas$Insert.p <- as.numeric(dta.v.meas$Variant)-1
  dta.v.meas.aggr.speaker <-aggregate(Insert.p ~ Speaker_ID+Long.jitter+Lat.jitter,
                                      data= dta.v.meas[,c("Speaker_ID",
                                                          "Lat.jitter",
                                                          "Long.jitter",
                                                          "Insert.p")],
                                      FUN=mean)
  dta.v.meas.aggr <-aggregate(Insert.p ~ Long+Lat,
                                      data= dta.v.meas[,c("Lat","Long","Insert.p")],
                                      FUN=mean)
  dta.v.meas.aggr$Value.p <- dta.v.meas.aggr$Insert.p
  dta.v.meas.aggr.speaker$Value.p <- dta.v.meas.aggr.speaker$Insert.p
  summary(dta.v.meas.aggr$Value.p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.05263 0.11111 0.15661 0.25000 0.80000
```

```
  summary(dta.v.meas.aggr.speaker$Value.p)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 0.00000 0.09091 0.14561 0.25000 0.84615
```

```
  plot.map.gradient(dta.v.meas.aggr)
```

Entering the p of substituting sortal by general to the mensural classifier table (for statistic evaluations)

```
  dta.v.meas <- merge(dta.v.meas,
                      dta.v.sort.aggr.speaker[,c("Speaker_ID","Substitute.p")],
                      by="Speaker_ID",
                       all.x = T)
```

5.2.3 List of mensural classifiers

Creating lists of classifiers for the Appendix

```
  dta.v.meas$mensuralCL <- dta.v.meas$Value
  dta.v.meas$mensuralCL <- str_remove(dta.v.meas$mensuralCL, "p'éel ")
  dta.v.meas$mensuralCL <- gsub(" .*","",dta.v.meas$mensuralCL)

  result.type <- as.data.frame(xtabs(~mensuralCL+Variant+Subtype, dta.v.meas))
  result.type.2 <- dcast(result.type, mensuralCL + Subtype ~ Variant)
```

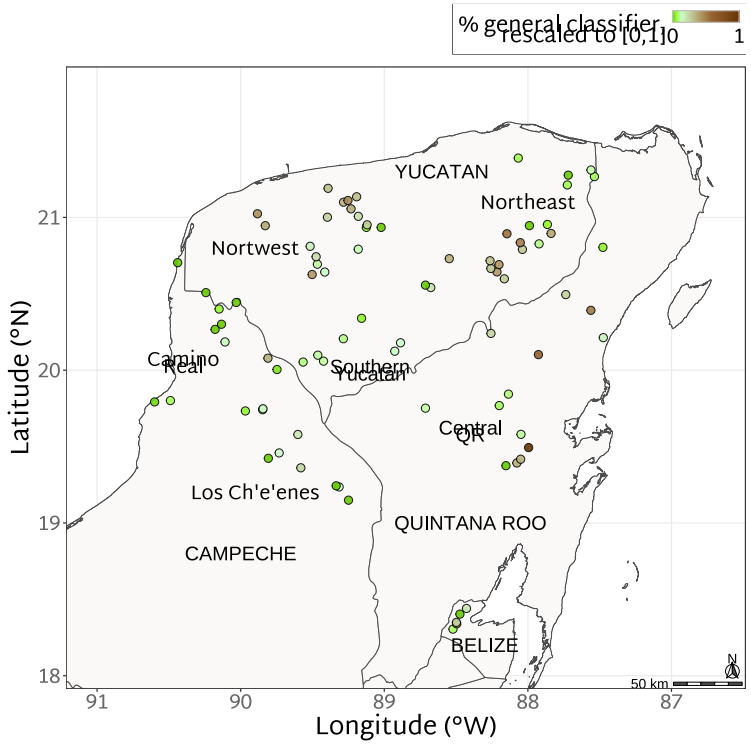

Figure 5: Figure: Mensural classifiers on base map

```
result.type.2$sum <- result.type.2$'CL.Gen CL.Spec'+result.type.2$CL.Spec
result.type.2$percent <- 100*result.type.2$'CL.Gen CL.Spec'/result.type.2$sum
result.type.2 <- droplevels(subset(result.type.2,percent != "NaN"))
kable(result.type.2) %>%
kable_styling("striped", full_width = F) %>%
scroll_box(width = "100%", height = "200px")
```

	mensuralCL	Subtype	CL.Spec	CL.Gen CL.Spec	sum	percent
3	beel	sum	3	0	3	0.000000
4	búuj	part	75	0	75	0.000000
8	ch'áaj	portion	88	14	102	13.725490
10	ch'áak	part	1	0	1	0.000000
14	ch'ooj	portion	15	4	19	21.052632
18	ch'uuy	sum	2	0	2	0.000000
19	cháach	part	1	0	1	0.000000
20	cháach	portion	2	1	3	33.333333
21	cháach	sum	82	18	100	18.000000
24	chooj	sum	1	0	1	0.000000
26	chúuch	portion	1	0	1	0.000000
27	chúuch	sum	3	2	5	40.000000
30	chúuj	sum	0	1	1	100.000000
32	chuuk'	portion	1	0	1	0.000000
34	jáat	part	47	1	48	2.083333
39	jaats	sum	0	1	1	100.000000

41	jeneb	portion	41	1	42	2.380952
45	jiil	sum	3	1	4	25.000000
48	juuts’	sum	2	0	2	0.000000
51	k’áax	sum	80	16	96	16.666667
53	k’ab	portion	2	0	2	0.000000
54	k’ab	sum	1	1	2	50.000000
55	káach	part	49	4	53	7.547170
58	kóots	part	10	0	10	0.000000
59	kóots	portion	1	0	1	0.000000
63	kúuch	sum	114	27	141	19.148936
65	láap’	portion	28	6	34	17.647059
66	láap’	sum	9	4	13	30.769231
68	lóob	portion	0	1	1	100.000000
71	lóoch’	portion	7	4	11	36.363636
72	lóoch’	sum	2	1	3	33.333333
74	lóot	portion	3	0	3	0.000000
75	lóot	sum	3	1	4	25.000000
77	luuch	portion	2	0	2	0.000000
80	luuk’	portion	87	23	110	20.909091
83	maach	portion	1	0	1	0.000000
84	maach	sum	18	1	19	5.263158
87	meek’	sum	1	0	1	0.000000
89	múuch’	portion	10	0	10	0.000000
90	múuch’	sum	3	3	6	50.000000
92	múul	portion	3	0	3	0.000000
93	múul	sum	1	1	2	50.000000
96	muut	sum	3	0	3	0.000000
98	nikib	portion	3	0	3	0.000000
100	p’aay	part	10	0	10	0.000000
105	p’iis	sum	42	13	55	23.636364
108	p’óoch	sum	0	2	2	100.000000
110	p’u’uk	portion	6	0	6	0.000000
112	p’úuy	part	1	0	1	0.000000
113	p’úuy	portion	0	1	1	100.000000
117	paak	sum	34	12	46	26.086956
119	t’aaj	portion	6	0	6	0.000000
121	t’i’il	part	1	0	1	0.000000
126	t’i’in	sum	0	1	1	100.000000
129	t’o’ol	sum	8	5	13	38.461539
132	t’úul	sum	0	1	1	100.000000
133	táaj	part	3	0	3	0.000000
136	táan	part	5	2	7	28.571429
141	to’	sum	1	0	1	0.000000
144	ts’áam	sum	1	0	1	0.000000
147	ts’ap	sum	62	23	85	27.058823
149	ts’úuk	portion	1	1	2	50.000000
151	tséej	part	4	0	4	0.000000
154	tsíil	part	1	0	1	0.000000
159	tsóol	sum	54	26	80	32.500000
161	tukub	portion	7	2	9	22.222222

162	tukub	sum	3	0	3	0.000000
163	wáat	part	1	0	1	0.000000
168	waats’	sum	1	1	2	50.000000
170	wóoch’	portion	1	0	1	0.000000
171	wóoch’	sum	1	0	1	0.000000
173	wóok	portion	8	1	9	11.111111
174	wóok	sum	3	0	3	0.000000
175	wóol	part	0	1	1	100.000000
176	wóol	portion	74	20	94	21.276596
177	wóol	sum	6	1	7	14.285714
180	wuuts’	sum	10	4	14	28.571429
182	xa’ak	portion	0	1	1	100.000000
184	xéek	part	1	0	1	0.000000
187	xéet’	part	503	23	526	4.372624
190	xóot’	part	106	2	108	1.851852
194	xuuch	portion	12	3	15	20.000000

5.2.4 Frequency of mensural in multiple classifiers

create scatter plot

```
1  plot.scatter(result.type.2)
```

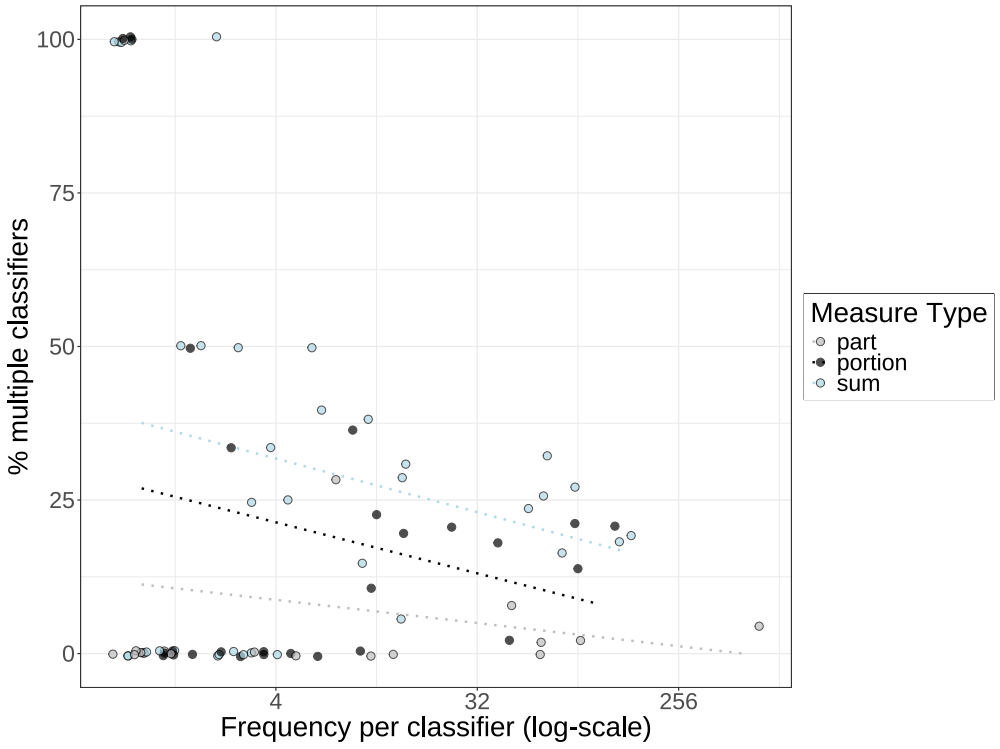

Figure 6: Figure: Scatter plot for Measure types

5.3 Inferential stats: Generalized additive models (GAM)

5.3.1 Sortal Classifiers

Maximal model:

```
gam.sortal.0 = mgcv::gam((Variant=="CL.Gen") ~
                           Spanish_odds +
                           Birthyear +
                           Pop.size +
                         s(Long, Lat, bs="tp",k=15),
                       data=dta.v.sort,method="ML")
print(summary(gam.sortal.0))
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (Variant == "CL.Gen") ~ Spanish_odds + Birthyear + Pop.size +
##     s(Long, Lat, bs = "tp", k = 15)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.986207   2.580205  -1.545 0.122939
## Spanish_odds  0.445919   0.120023   3.715 0.000224 ***
## Birthyear     0.002237   0.001327   1.685 0.092461 .
## Pop.size     -0.002844   0.014700  -0.193 0.846657
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df     F  p-value
## s(Long,Lat) 11.94  13.43 3.782 7.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.14   Deviance explained = 16.3%
## -ML =  380.6  Scale est. = 0.21396   n = 568
```

Model comparisons:

```
gam.sortal.1 <- update(gam.sortal.0, . ~ . -Pop.size)
    compareML(gam.sortal.0,gam.sortal.1,suggest.report = T)
```

```
## gam.sortal.0: (Variant == "CL.Gen") ~ Spanish_odds + Birthyear + Pop.size +
##     s(Long, Lat, bs = "tp", k = 15)
##
## gam.sortal.1: (Variant == "CL.Gen") ~ Spanish_odds + Birthyear + s(Long, Lat,
##     bs = "tp", k = 15)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.sortal.0 is [not sig
## -----
```

```
##          Model    Score Edf Difference    Df p.value Sig.
## 1 gam.sortal.1 380.6154   6
## 2 gam.sortal.0 380.5968   7      0.019 1.000   0.847
##
## AIC difference: 1.79, model gam.sortal.1 has lower AIC.
```

```
gam.sortal.2 <- update(gam.sortal.1, . ~ . -Birthyear)
    compareML(gam.sortal.1,gam.sortal.2,suggest.report = T)
```

```
## gam.sortal.1: (Variant == "CL.Gen") ~ Spanish_odds + Birthyear + s(Long, Lat,
##     bs = "tp", k = 15)
##
## gam.sortal.2: (Variant == "CL.Gen") ~ Spanish_odds + s(Long, Lat, bs = "tp",
##     k = 15)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.sortal.1 is [not si
## -----
##          Model    Score Edf Difference    Df p.value Sig.
## 1 gam.sortal.2 382.0136   5
## 2 gam.sortal.1 380.6154   6      1.398 1.000   0.094
##
## AIC difference: -1.76, model gam.sortal.1 has lower AIC.
```

```
gam.sortal.3 <- update(gam.sortal.2, . ~ . -Spanish_odds)
    compareML(gam.sortal.2,gam.sortal.3,suggest.report = T)
```

```
## gam.sortal.2: (Variant == "CL.Gen") ~ Spanish_odds + s(Long, Lat, bs = "tp",
##     k = 15)
##
## gam.sortal.3: (Variant == "CL.Gen") ~ s(Long, Lat, bs = "tp", k = 15)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.sortal.2 is [signif
## -----
##          Model    Score Edf Difference    Df   p.value Sig.
## 1 gam.sortal.3 388.9529   4
## 2 gam.sortal.2 382.0136   5      6.939 1.000 1.950e-04  ***
##
## AIC difference: -7.76, model gam.sortal.2 has lower AIC.
```

Coefficients of winner model

```
  print(summary(gam.sortal.2))
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (Variant == "CL.Gen") ~ Spanish_odds + s(Long, Lat, bs = "tp",
##     k = 15)
##
## Parametric coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.35946    0.03266  11.005  < 2e-16 ***
## Spanish_odds  0.45508    0.11940   3.811 0.000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                edf Ref.df     F  p-value
## s(Long,Lat) 11.83  13.37 3.818 5.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.137   Deviance explained = 15.7%
## -ML = 382.01  Scale est. = 0.21467   n = 568
```

Coefficients of smooth term of Space

```
plot.gam(gam.sortal.2)
```

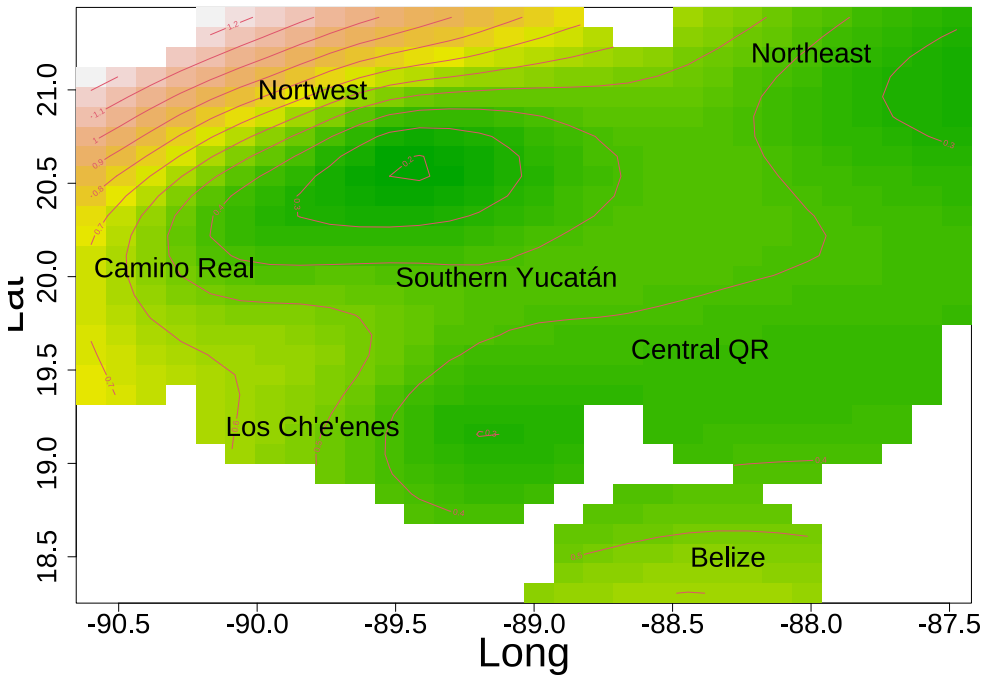

Figure 7: Figure: Space coefficients for Sorts

5.3.2 Mensural classifiers

Contrast coding: successive differences of subtypes

```
dta.v.meas$Subtype <- factor(dta.v.meas$Subtype)
contrasts(dta.v.meas$Subtype) <- MASS::contr.sdif(3)
```

model.0: Maximal model

```
gam.mensural.0 = mgcv::gam((Variant=="CL.Gen CL.Spec") ~
                           Substitute.p +
                           Subtype +
                           Spanish_odds +
                           Birthyear +
                           Pop.size +
                           s(Long, Lat, bs="tp",k=15,by=Subtype),
                         data=dta.v.meas,method="ML")
print(summary(gam.mensural.0))
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + Pop.size + s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## Parametric coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.999e-01  9.297e-01   0.323 0.747020
## Substitute.p  9.627e-02  2.523e-02   3.816 0.000139 ***
## Subtype2-1    1.405e-01  1.801e-02   7.804 9.46e-15 ***
## Subtype3-2    6.067e-02  1.855e-02   3.271 0.001090 **
## Spanish_odds  2.418e-01  4.298e-02   5.626 2.10e-08 ***
## Birthyear    -1.240e-04  4.776e-04  -0.260 0.795235
## Pop.size      4.322e-05  4.875e-03   0.009 0.992928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                                edf Ref.df     F p-value
## s(Long,Lat):Subtypepart      2.002  2.004 4.995 0.00683 **
## s(Long,Lat):Subtypeportion   5.779  7.760 2.754 0.00582 **
## s(Long,Lat):Subtypesum      11.359 13.066 7.057 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.151   Deviance explained = 16.2%
## -ML = 567.91  Scale est. = 0.10025   n = 2069
```

Removing SPACE : MEASURE TYPE

```
gam.mensural.0.interaction = mgcv::gam((Variant=="CL.Gen CL.Spec") ~
                           Substitute.p +
                           Subtype +
                           Spanish_odds +
                           Birthyear +
                           Pop.size +
                           s(Long, Lat, bs="tp",k=15),
                         data=dta.v.meas,method="ML")
print(summary(gam.mensural.0.interaction))
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + Pop.size + s(Long, Lat, bs = "tp", k = 15)
##
## Parametric coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.741580   0.952816   0.778   0.4365
## Substitute.p  0.103979   0.026644   3.903 9.82e-05 ***
## Subtype2-1    0.142363   0.018144   7.846 6.84e-15 ***
## Subtype3-2    0.058663   0.018667   3.143   0.0017 **
## Spanish_odds  0.238931   0.046470   5.142 2.98e-07 ***
## Birthyear    -0.000345   0.000490  -0.704   0.4814
## Pop.size     -0.001577   0.005386  -0.293   0.7697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F p-value
## s(Long,Lat) 10.31  12.32 6.966  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.137   Deviance explained = 14.4%
## -ML = 580.25  Scale est. = 0.10189   n = 2069
```

Model comparison, winner: model.0

```
compareML(gam.mensural.0,gam.mensural.0.interaction,suggest.report = T)
```

```
## gam.mensural.0: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + Pop.size + s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.0.interaction: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + Pop.size + s(Long, Lat, bs = "tp", k = 15)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.0 is [signi
## -----
##                         Model    Score Edf Difference    Df   p.value Sig.
## 1 gam.mensural.0.interaction 580.2455  10
## 2             gam.mensural.0 567.9083  16     12.337 6.000 3.923e-04  ***
##
## AIC difference: -22.25, model gam.mensural.0 has lower AIC.
```

Removing POPULATION SIZE

Model comparison: non-significant

```
gam.mensural.1 <- update(gam.mensural.0, . ~ . -Pop.size)
  compareML(gam.mensural.0,gam.mensural.1,suggest.report = T)
```

```
## gam.mensural.0: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + Pop.size + s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.1: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.0 is [not s
## -----
##             Model    Score Edf Difference    Df p.value Sig.
## 1 gam.mensural.1 567.9084  15
## 2 gam.mensural.0 567.9083  16      0.000 1.000   0.993
##
## AIC difference: 2.02, model gam.mensural.1 has lower AIC.
```

Removing TIME

Model comparison: non-significant

```
gam.mensural.2 <- update(gam.mensural.1, . ~ . -Birthyear)
  compareML(gam.mensural.1,gam.mensural.2,suggest.report = T)
```

```
## gam.mensural.1: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     Birthyear + s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.1 is [not s
## -----
##             Model    Score Edf Difference    Df p.value Sig.
## 1 gam.mensural.2 567.9422  14
## 2 gam.mensural.1 567.9084  15      0.034 1.000   0.795
##
## AIC difference: 2.06, model gam.mensural.2 has lower AIC.
```

Removing SPANISH ODDS

Model comparison: significant

```
gam.mensural.3 <- update(gam.mensural.2, . ~ . -Spanish_odds)
  compareML(gam.mensural.2,gam.mensural.3,suggest.report = T)#win 2
```

```
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.3: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + s(Long,
##     Lat, bs = "tp", k = 15, by = Subtype)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.2 is [signi
## -----
##             Model    Score Edf Difference    Df   p.value Sig.
## 1 gam.mensural.3 582.5011  13
## 2 gam.mensural.2 567.9422  14     14.559 1.000 6.811e-08  ***
##
## AIC difference: -17.61, model gam.mensural.2 has lower AIC.
```

Removing MEASURE TYPE

Model comparison: significant

```
  gam.mensural.4 <- update(gam.mensural.2, . ~ . -Subtype)
    compareML(gam.mensural.2,gam.mensural.4,suggest.report = T)#win 2
```

```
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.4: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.2 is [signi
## -----
##             Model    Score Edf Difference    Df  p.value Sig.
## 1 gam.mensural.4 647.9576  12
## 2 gam.mensural.2 567.9422  14     80.015 2.000  < 2e-16  ***
##
## AIC difference: -160.28, model gam.mensural.2 has lower AIC.
```

Collapsing Parts and Sums

Model comparison: significant

```
  gam.mensural.41 <- update(gam.mensural.4, . ~ . +TypePS)
    compareML(gam.mensural.2,gam.mensural.41,suggest.report = T)#win 2
```

```
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.41: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype) + TypePS
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.2 is [signi
## -----
##              Model    Score Edf Difference    Df p.value Sig.
## 1 gam.mensural.41 573.3127  13
## 2  gam.mensural.2 567.9422  14      5.371 1.000   0.001  **
##
## AIC difference: -8.77, model gam.mensural.2 has lower AIC.
```

Collapsing Parts and Portions

Model comparison: significant

```
  gam.mensural.42 <- update(gam.mensural.4, . ~ . +TypePP)
    compareML(gam.mensural.2,gam.mensural.42,suggest.report = T)#win 2
```

```
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.42: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Spanish_odds +
```

```
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype) + TypePP
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.2 is [signi
## -----
##             Model    Score Edf Difference    Df   p.value Sig.
## 1 gam.mensural.42 598.1196  13
## 2  gam.mensural.2 567.9422  14     30.177 1.000 7.921e-15  ***
##
## AIC difference: -58.61, model gam.mensural.2 has lower AIC.
```

Removing Likelihood of Substitution

Model comparison: significant

```
gam.mensural.5 <- update(gam.mensural.2, . ~ . -Substitute.p)
  compareML(gam.mensural.2,gam.mensural.5,suggest.report = T)
```

```
## gam.mensural.2: (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## gam.mensural.5: (Variant == "CL.Gen CL.Spec") ~ Subtype + Spanish_odds + s(Long,
##     Lat, bs = "tp", k = 15, by = Subtype)
##
## Report suggestion: The Chi-Square test on the ML scores indicates that model gam.mensural.2 is [signi
## -----
##            Model    Score Edf Difference    Df   p.value Sig.
## 1 gam.mensural.5 575.1302  13
## 2 gam.mensural.2 567.9422  14      7.188 1.000 1.497e-04  ***
##
## AIC difference: -10.15, model gam.mensural.2 has lower AIC.
```

Coefficients of winner model

```
summary(gam.mensural.2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## (Variant == "CL.Gen CL.Spec") ~ Substitute.p + Subtype + Spanish_odds +
##     s(Long, Lat, bs = "tp", k = 15, by = Subtype)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.05814    0.01326   4.386 1.21e-05 ***
## Substitute.p  0.09612    0.02521   3.813 0.000142 ***
## Subtype2-1    0.14053    0.01800   7.808 9.17e-15 ***
## Subtype3-2    0.06072    0.01854   3.276 0.001072 **
## Spanish_odds  0.24170    0.04291   5.632 2.03e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Approximate significance of smooth terms:
##                                edf Ref.df     F p-value
## s(Long,Lat):Subtypepart      2.001  2.003 5.102 0.00615 **
## s(Long,Lat):Subtypeportion   5.849  7.841 2.755 0.00556 **
## s(Long,Lat):Subtypesum      11.379 13.076 7.071 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.152   Deviance explained = 16.2%
## -ML = 567.94  Scale est. = 0.10015   n = 2069
```

Coefficients of the smooth term of space

```
plot.gam.2(gam.mensural.2,"part")
```

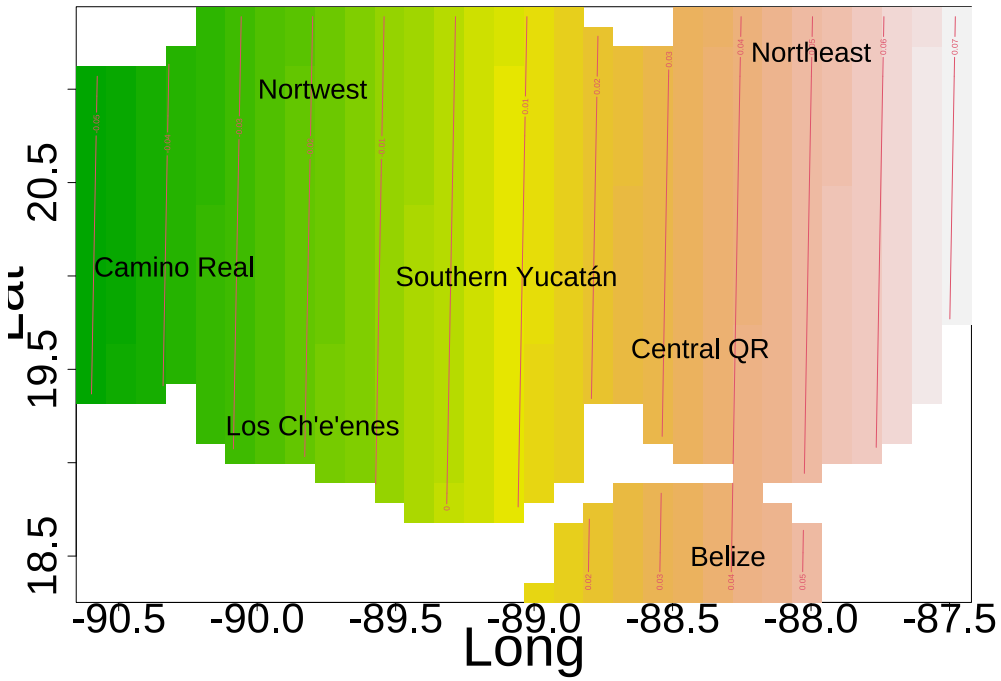

Figure 8: Figure: Space coefficients for Parts

```
plot.gam.2(gam.mensural.2,"sum")
```

```
plot.gam.2(gam.mensural.2,"portion")
```

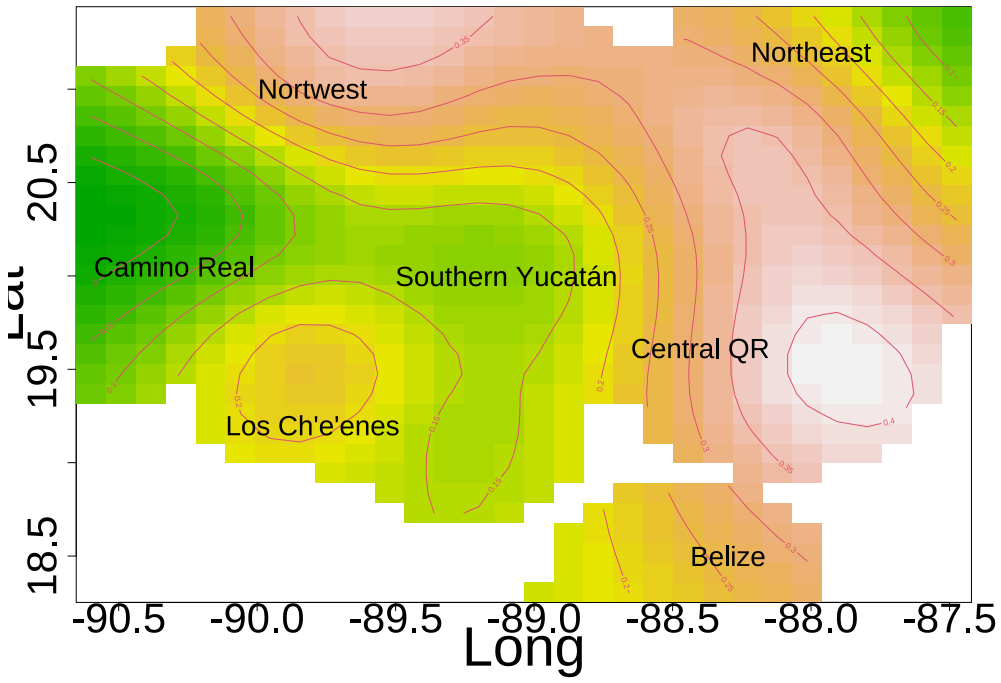

Figure 9: Figure: Space coefficients for Sums

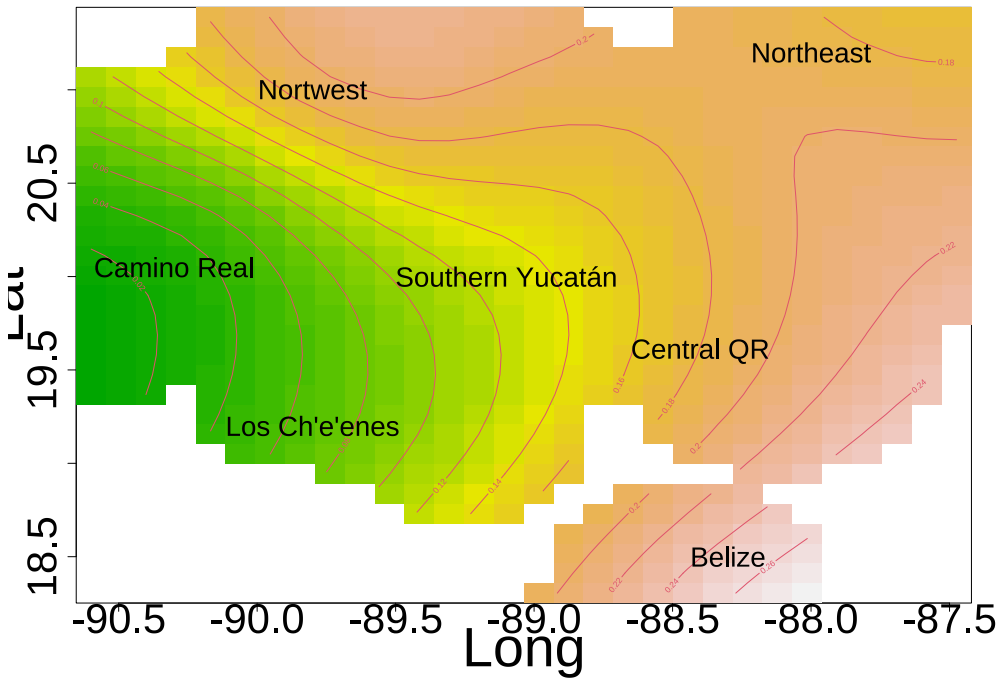

Figure 10: Figure: Space coefficients for Portions

6 References

6.1 R and R Version

```
version[['version.string']]
```

```
## [1] "R version 4.4.1 (2024-06-14 ucrt)"
```

```
print(citation())
```

```
## To cite R in publications use:
##
##   R Core Team (2024). _R: A Language and Environment for Statistical
##   Computing_. R Foundation for Statistical Computing, Vienna, Austria.
##   <https://www.R-project.org/>.
##
## Ein BibTeX-Eintrag für LaTeX-Benutzer ist
##
##   @Manual{,
##     title = {R: A Language and Environment for Statistical Computing},
##     author = {{R Core Team}},
##     organization = {R Foundation for Statistical Computing},
##     address = {Vienna, Austria},
##     year = {2024},
##     url = {https://www.R-project.org/},
##   }
##
## We have invested a lot of time and effort in creating R, please cite it
## when using it for data analysis. See also 'citation("pkgname")' for
## citing R packages.
```

6.2 Packages

```
packages <- c("readxl","writexl","reshape2","scales","dplyr",
              "stringr","showtext","ggplot2","ggmap","ggspatial",
              "mgcv","mgcViz","itsadug")

knitr::write_bib(c(packages), width = 60)
```

```
## @Manual{R-dplyr,
##   title = {dplyr: A Grammar of Data Manipulation},
##   author = {Hadley Wickham and Romain François and Lionel
##     Henry and Kirill Müller and Davis Vaughan},
##   year = {2023},
##   note = {R package version 1.1.4},
##   url = {https://dplyr.tidyverse.org},
## }
##
## @Manual{R-ggmap,
```

```
##   title = {ggmap: Spatial Visualization with ggplot2},
##   author = {David Kahle and Hadley Wickham and Scott
##     Jackson},
##   year = {2023},
##   note = {R package version 4.0.0},
##   url = {https://github.com/dkahle/ggmap},
## }
##
## @Manual{R-ggplot2,
##   title = {ggplot2: Create Elegant Data Visualisations
##     Using the Grammar of Graphics},
##   author = {Hadley Wickham and Winston Chang and Lionel
##     Henry and Thomas Lin Pedersen and Kohske Takahashi and
##     Claus Wilke and Kara Woo and Hiroaki Yutani and Dewey
##     Dunnington and Teun {van den Brand}},
##   year = {2024},
##   note = {R package version 3.5.1},
##   url = {https://ggplot2.tidyverse.org},
## }
##
## @Manual{R-ggspatial,
##   title = {ggspatial: Spatial Data Framework for ggplot2},
##   author = {Dewey Dunnington},
##   year = {2023},
##   note = {R package version 1.1.9},
##   url = {https://paleolimbot.github.io/ggspatial/},
## }
##
## @Manual{R-itsadug,
##   title = {itsadug: Interpreting Time Series and
##     Autocorrelated Data Using GAMMs},
##   author = {Jacolien {van Rij} and Martijn Wieling and R.
##     Harald Baayen},
##   year = {2022},
##   note = {R package version 2.4.1},
##   url = {https://CRAN.R-project.org/package=itsadug},
## }
##
## @Manual{R-mgcv,
##   title = {mgcv: Mixed GAM Computation Vehicle with
##     Automatic Smoothness Estimation},
##   author = {Simon Wood},
##   year = {2023},
##   note = {R package version 1.9-1},
##   url = {https://CRAN.R-project.org/package=mgcv},
## }
##
## @Manual{R-mgcViz,
##   title = {mgcViz: Visualisations for Generalized Additive
##     Models},
##   author = {Matteo Fasiolo and Raphael Nedellec},
##   year = {2023},
##   note = {R package version 0.1.11},
##   url = {https://github.com/mfasiolo/mgcViz},
```

```
## }
##
## @Manual{R-readxl,
##   title = {readxl: Read Excel Files},
##   author = {Hadley Wickham and Jennifer Bryan},
##   year = {2023},
##   note = {R package version 1.4.3},
##   url = {https://readxl.tidyverse.org},
## }
##
## @Manual{R-reshape2,
##   title = {reshape2: Flexibly Reshape Data: A Reboot of the
##     Reshape Package},
##   author = {Hadley Wickham},
##   year = {2020},
##   note = {R package version 1.4.4},
##   url = {https://github.com/hadley/reshape},
## }
##
## @Manual{R-scales,
##   title = {scales: Scale Functions for Visualization},
##   author = {Hadley Wickham and Thomas Lin Pedersen and Dana
##     Seidel},
##   year = {2023},
##   note = {R package version 1.3.0},
##   url = {https://scales.r-lib.org},
## }
##
## @Manual{R-showtext,
##   title = {showtext: Using Fonts More Easily in R Graphs},
##   author = {Yixuan Qiu and authors/contributors of the
##     included software. See file AUTHORS for details.},
##   year = {2024},
##   note = {R package version 0.9-7},
##   url = {https://github.com/yixuan/showtext},
## }
##
## @Manual{R-stringr,
##   title = {stringr: Simple, Consistent Wrappers for Common
##     String Operations},
##   author = {Hadley Wickham},
##   year = {2023},
##   note = {R package version 1.5.1},
##   url = {https://stringr.tidyverse.org},
## }
##
## @Manual{R-writexl,
##   title = {writexl: Export Data Frames to Excel xlsx
##     Format},
##   author = {Jeroen Ooms},
##   year = {2024},
##   note = {R package version 1.5.0},
##   url = {https://docs.ropensci.org/writexl/},
## }
```

```
##
## @Article{ggmap2013,
##   author = {David Kahle and Hadley Wickham},
##   title = {ggmap: Spatial Visualization with ggplot2},
##   journal = {The R Journal},
##   year = {2013},
##   volume = {5},
##   number = {1},
##   pages = {144--161},
##   url =
##     {https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf},
## }
##
## @Book{ggplot22016,
##   author = {Hadley Wickham},
##   title = {ggplot2: Elegant Graphics for Data Analysis},
##   publisher = {Springer-Verlag New York},
##   year = {2016},
##   isbn = {978-3-319-24277-4},
##   url = {https://ggplot2.tidyverse.org},
## }
##
## @Article{mgcv2011,
##   title = {Fast stable restricted maximum likelihood and
##     marginal likelihood estimation of semiparametric
##     generalized linear models},
##   journal = {Journal of the Royal Statistical Society (B)},
##   volume = {73},
##   number = {1},
##   pages = {3-36},
##   year = {2011},
##   author = {S. N. Wood},
## }
##
## @Article{mgcv2016,
##   title = {Smoothing parameter and model selection for
##     general smooth models (with discussion)},
##   author = {S.N. Wood and {N.} and {Pya} and B. S{"a}fken},
##   journal = {Journal of the American Statistical
##     Association},
##   year = {2016},
##   pages = {1548-1575},
##   volume = {111},
## }
##
## @Article{mgcv2004,
##   title = {Stable and efficient multiple smoothing
##     parameter estimation for generalized additive models},
##   journal = {Journal of the American Statistical
##     Association},
##   volume = {99},
##   number = {467},
##   pages = {673-686},
##   year = {2004},
```

```
##   author = {S. N. Wood},
## }
##
## @Book{mgcv2017,
##   title = {Generalized Additive Models: An Introduction
##     with R},
##   year = {2017},
##   author = {S.N Wood},
##   edition = {2},
##   publisher = {Chapman and Hall/CRC},
## }
##
## @Article{mgcv2003,
##   title = {Thin-plate regression splines},
##   journal = {Journal of the Royal Statistical Society (B)},
##   volume = {65},
##   number = {1},
##   pages = {95-114},
##   year = {2003},
##   author = {S. N. Wood},
## }
##
## @Article{mgcViz2020,
##   title = {Scalable visualisation methods for modern
##     Generalized Additive Models.},
##   journal = {Journal of the Royal Statistical Society (B)},
##   volume = {29},
##   number = {1},
##   pages = {78-86},
##   year = {2020},
##   author = {{Fasiolo} and {Matteo} and {Nedellec} and
##     {Rapha{"e}l} and {Goude} and {Yannig} and {Wood} and
##     Simon N},
## }
##
## @Article{reshape22007,
##   title = {Reshaping Data with the {reshape} Package},
##   author = {Hadley Wickham},
##   journal = {Journal of Statistical Software},
##   year = {2007},
##   volume = {21},
##   number = {12},
##   pages = {1--20},
##   url = {http://www.jstatsoft.org/v21/i12/},
## }
```

— end of documentation —