Data, data documentation and analysis scripts for

Border effects among catalan dialects

Martijn Wieling[(1)] & Esteve Valls[(2)] & R. Harald Baayen[(3,4)] & John Nerbonne[(1,5)]

[1]University of Groningen, the Netherlands & [2]University of Barcelona, Spain & [3]Eberhard Karls University, Germany & [4]University of Alberta, Canada & [5]University of Freiburg, Germany

Book: D. Speelman et al. (eds.) **Mixed Effects Regression Models in Linguistics** (acc.)
Preprint: `http://martijnwieling.nl/files/Wieling-etal-2015-LSD.pdf`

Abstract

In this study, we investigate which factors influence the linguistic distance of Catalan dialectal pronunciations from standard Catalan. We use pronunciations from three regions where the northwestern variety of the Catalan language is spoken (Catalonia, Aragon and Andorra). In contrast to Aragon, Catalan has an official status in both Catalonia and Andorra, which likely influences standardization. Because we are interested in the potentially large range of differences that standardization might promote, we examine 357 words in Catalan varieties and in particular their pronunciation distances with respect to the standard. In order to be sensitive to differences among the words, we fit a generalized additive mixed-effects regression model to this data. This allows us to examine simultaneously the general (i.e. aggregate) patterns in pronunciation distance and to detect those words that diverge substantially from the general pattern. The results reveal higher pronunciation distances from standard Catalan in Aragon than in the other regions. Furthermore, speakers in Catalonia and Andorra, but not in Aragon, show a clear standardization pattern, with younger speakers having dialectal pronunciations closer to the standard than older speakers. This clearly indicates the presence of a border effect within a single country with respect to word pronunciation distances. Since a great deal of scholarship focuses on single segment changes, we compare our analysis to the analysis of three segment changes that have been discussed in the literature on Catalan. This comparison shows that the pattern observed at the word pronunciation level is supported by two of the three cases examined. As not all individual cases conform to the general pattern, the aggregate approach is necessary to detect global standardization patterns.

Keywords: Dialectometry, Catalan dialects, border effects, generalized additive modeling, mixed-effects regression.

1 Packages and functions

```
require(parallel)
library(mgcv)

R.Version()$version.string

## [1] "R version 3.2.2 (2015-08-14)"

packageVersion("mgcv")

## [1] '1.8.8'

source('functions/functions.R') # custom functions
```

2 Complete data set

```
load("data/catalan.rda")
```

Legenda tuscan (112608 of 39 variables):

Note that the columns with a suffix of `.c` or `.z` are not described here. These are simply a centered (mean equals zero) or standardized (mean equals zero and standard deviation equals 1) version of the corresponding variables shown below.

1. Word : the word for which pronunciations were obtained
2. Speaker : the speaker whose pronunciations were obtained
3. Location : the location in which speakers were asked for their pronunciations
4. PronDistStdCatalan : Pronunciation distance from standard Catalan
5. Longitude : longitude of the dialect location
6. Latitude : latitude of the dialect location
7. Region : location in Catalonia (C), Andorra (A) or Aragon (L)
8. InAragon : binary value indicating if the location is in Aragon (1) or not (0)
9. IsUrban : binary value indicating if the location is urban (1) or rural (0)
10. CommunitySize.log : number of inhabitants in the location (log-transformed)
11. CommunityAvgAge : average age in the location
12. CommunityAvgIncome : average income in the location (log-transformed)
13. CommunityRelTouristBeds.log : relative number of tourist beds in the location (log-transformed)
14. SpeakerBirthYear : year of birth of the speaker
15. YBInCatAnd : year of birth of the speaker (only in Catalonia and Andorra, 0 otherwise; also for the z-transformed variant)
16. YBInAragon : year of birth of the speaker (only in Aragon, 0 otherwise; also for the z-transformed variant)
17. SpeakerIsMale : binary value indicating if the speaker is male (1) or not (0)
18. SpeakerEduLevel : education level of the speaker from low to high (0: no schooling, 1: primary school, 2: high school, 3: baccalaureate, 4: technical school, 5: university)
19. SpeakerRecordingYear : year when the speaker's pronunciation was recorded
20. WordLength : the number of sounds in the standard pronunciation of the word

21. WordRefVowelRatio : the relative number of vowels in the standard pronunciation of the word

22. WordCategory : the word category: verbs (V), possessives (P), personal pronouns (PP), locatives (L), demonstratives (D), other demonstratives (O), articles (A), clitics (C)

23. WordCategoryIsACD : binary value indicating if the word's category is article (A), clitic (C) or demonstrative (D)

3 Analysis and results of complete data set

3.1 Correlations

```
cor(catalan$WordCategoryIsACD,catalan$WordLength.z)

## [1] -0.7729993

cor(catalan$CommunitySize.log.z,catalan$CommunityAvgAge.z)

## [1] -0.8077254

cor(catalan$CommunitySize.log.z,catalan$CommunityAvgIncome.z)

## [1] 0.6565238

cor(catalan$CommunityAvgIncome.z,catalan$CommunityAvgAge.z)

## [1] -0.7623128

cor(catalan$SpeakerEduLevel.z,catalan$SpeakerBirthYear.z)

## [1] 0.3083282
```

3.2 Final GAM

```
cl = makeCluster(36) # 4 cores used in calculating the model

modelCatalanFinal <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      YBInAragon.z + YBInCatAnd.z +
      s(Longitude, Latitude) +
      s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge.z, bs = "re") +
      s(Word, CommunitySize.log.z, bs = "re") +
      s(Word, CommunityAvgIncome.z, bs = "re") +
      s(Word, SpeakerEduLevel.z, bs = "re") +
      s(Word, YBInCatAnd.z, bs = "re") +
      s(Word, YBInAragon.z, bs = "re") +
      s(Speaker, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re") +
      s(Speaker, WordLength.z, bs = "re") +
      s(Location, bs = "re") +
      s(Location, YBInCatAnd.z, bs = "re") +
      s(Location, WordRefVowelRatio.z, bs = "re") +
      s(Location, WordCategoryIsACD, bs = "re") +
      s(Location, WordLength.z, bs = "re"), data=catalan,
```

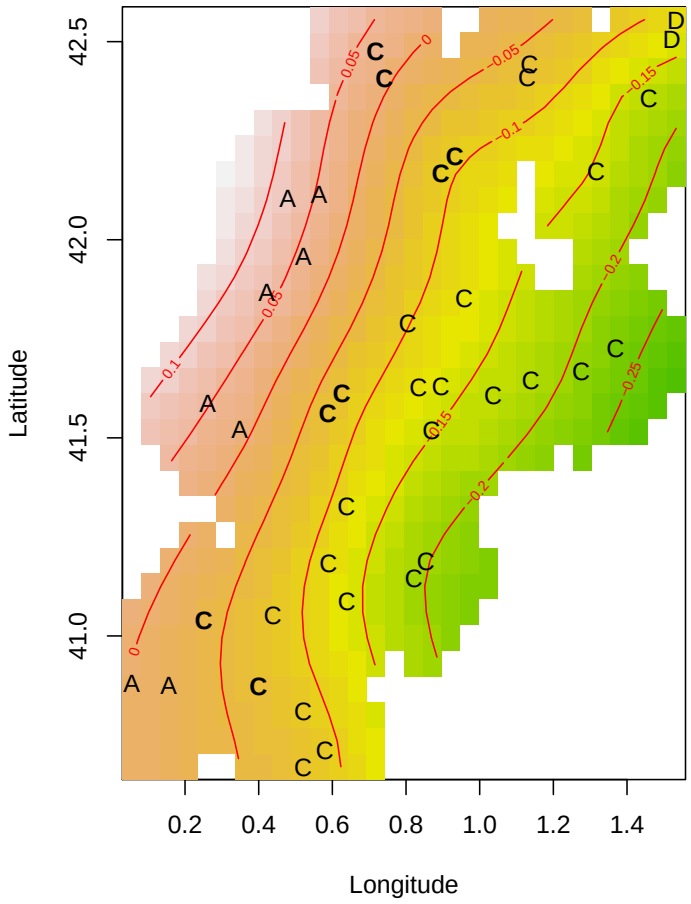
```
      cluster=cl)

smry <- summary(modelCatalan)

# model saved as calculations take about 20 minutes on 36 cores
save(modelCatalanFinal,smry,file='results/modelCatalanFinal.rda')
```

```
load('results/modelCatalanFinal.rda')
smry

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     YBInAragon.z + YBInCatAnd.z + s(Longitude, Latitude) + s(Word,
##     bs = "re") + s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
##     s(Word, CommunityAvgAge.z, bs = "re") + s(Word, CommunitySize.log.z,
##     bs = "re") + s(Word, CommunityAvgIncome.z, bs = "re") + s(Word,
##     SpeakerEduLevel.z, bs = "re") + s(Word, YBInCatAnd.z, bs = "re") +
##     s(Word, YBInAragon.z, bs = "re") + s(Speaker, bs = "re") +
##     s(Speaker, WordRefVowelRatio.z, bs = "re") + s(Speaker, WordCategoryIsACD,
##     bs = "re") + s(Speaker, WordLength.z, bs = "re") + s(Location,
##     bs = "re") + s(Location, YBInCatAnd.z, bs = "re") + s(Location,
##     WordRefVowelRatio.z, bs = "re") + s(Location, WordCategoryIsACD,
##     bs = "re") + s(Location, WordLength.z, bs = "re")
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.032913   0.017535  -1.877  0.06052
## WordRefVowelRatio.z  0.109073   0.014067   7.754 8.97e-15
## WordCategoryIsACD    0.101042   0.034055   2.967  0.00301
## YBInAragon.z         0.004650   0.004322   1.076  0.28196
## YBInCatAnd.z        -0.011576   0.005263  -2.199  0.02786
##
## Approximate significance of smooth terms:
##                                        edf Ref.df        F  p-value
## s(Longitude,Latitude)                11.95  12.02   16.981  < 2e-16
## s(Word)                             353.33 354.00 1747.855  < 2e-16
## s(Word,CommunityRelTouristBeds.log.z) 297.21 357.00   19.385  < 2e-16
## s(Word,CommunityAvgAge.z)            275.14 357.00  106.766  < 2e-16
## s(Word,CommunitySize.log.z)          235.18 357.00   50.689  < 2e-16
## s(Word,CommunityAvgIncome.z)         287.33 357.00  110.505  < 2e-16
## s(Word,SpeakerEduLevel.z)            190.64 357.00    3.025  < 2e-16
## s(Word,YBInCatAnd.z)                 310.81 356.00   30.615  < 2e-16
## s(Word,YBInAragon.z)                 189.25 356.00    2.290  < 2e-16
## s(Speaker)                           223.25 315.00   21.037  0.00041
## s(Speaker,WordRefVowelRatio.z)       164.35 319.00    1.803  < 2e-16
## s(Speaker,WordCategoryIsACD)         135.46 319.00   10.034  < 2e-16
```

```
## s(Speaker,WordLength.z)               183.92 320.00    5.472 5.50e-09
## s(Location)                            23.86  37.00  478.024 6.12e-05
## s(Location,YBInCatAnd.z)               25.99  31.00  550.143  < 2e-16
## s(Location,WordRefVowelRatio.z)        33.90  39.00  127.233 1.44e-09
## s(Location,WordCategoryIsACD)          36.86  39.00 7419.638  < 2e-16
## s(Location,WordLength.z)               37.10  40.00 6916.814  < 2e-16
##
## R-sq.(adj) =  0.735   Deviance explained = 74.2%
## fREML = -28237  Scale est. = 0.032891  n = 112608
```

3.3 Visualization of the non-linear effect of geography

```
vis.gam(modelCatalanFinal, view = c("Longitude","Latitude"),
        cond = list(WordRefVowelRatio.z=0, WordCategoryIsACD=0,
        CommunitySize.log.z=0, YBInAragon.z=0, YBInCatAnd.z=0),
        color='terrain', plot.type='contour',
        too.far=0.12, main="", zlim = c(-0.4,0.15))

inBF = c("Vilaller","El Pont de Suert","Salas de Pallars","Tremp","Lleida",
         "Montoliu de Lleida","Caseres","Alfara de Carles") # printed in boldface

addLabels(catalan,inBF,col='black') # adds labels to plot
```



3.4 Effect sizes

```
effectSizes = t(data.frame(
    getEffectSize.gam(catalan, smry, "WordRefVowelRatio.z"),
    getEffectSize.gam(catalan, smry, "WordCategoryIsACD"),
    getEffectSize.gam(catalan, smry, "YBInCatAnd.z"),
    getEffectSize.gam(catalan, smry, "YBInAragon.z"),
    getEffectSizeSpline.gam(modelCatalanFinal, "s(Longitude,Latitude)")
))
effectSizes
```

```
##                         Effect size
## WordRefVowelRatio.z           0.674
## WordCategoryisACD             0.101
## YBInCatAnd.z                 -0.034
## YBInAragon.z                  0.014
## s(Longitude,Latitude)         0.310
```

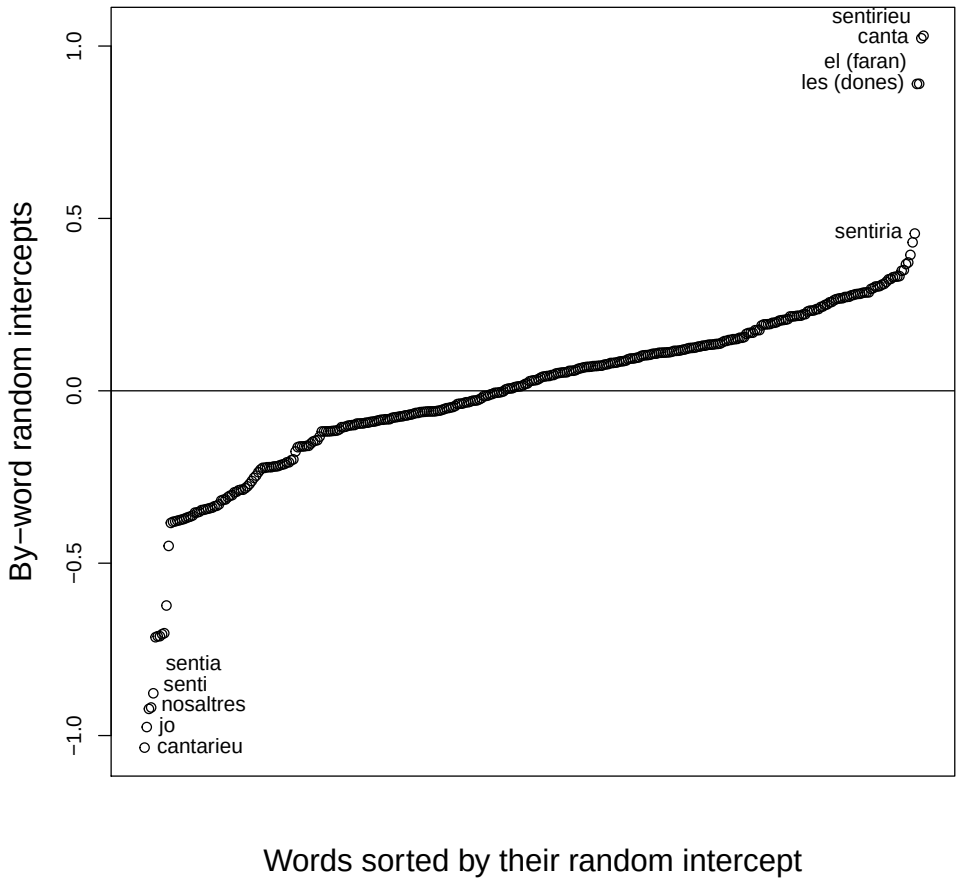
3.5 Standard deviations of random effects

```
coefs = coef(modelCatalanFinal)
stdevs = t(data.frame(
    getSD.gam(coefs, "Word", "Intercept"),
    getSD.gam(coefs, "Word", "CommunityRelTouristBeds.log.z"),
    getSD.gam(coefs, "Word", "CommunityAvgAge.z"),
    getSD.gam(coefs, "Word", "CommunitySize.log.z"),
    getSD.gam(coefs, "Word", "CommunityAvgIncome.z"),
    getSD.gam(coefs, "Word", "SpeakerEduLevel.z"),
    getSD.gam(coefs, "Word", "YBInCatAnd.z"),
    getSD.gam(coefs, "Word", "YBInAragon.z"),
    getSD.gam(coefs, "Speaker", "Intercept"),
    getSD.gam(coefs, "Speaker", "WordRefVowelRatio.z"),
    getSD.gam(coefs, "Speaker", "WordCategoryIsACD"),
    getSD.gam(coefs, "Speaker", "WordLength.z"),
    getSD.gam(coefs, "Location", "Intercept"),
    getSD.gam(coefs, "Location", "YBInCatAnd.z"),
    getSD.gam(coefs, "Location", "WordRefVowelRatio.z"),
    getSD.gam(coefs, "Location", "WordCategoryIsACD"),
    getSD.gam(coefs, "Location", "WordLength.z")

))
stdevs
```

```
##                                       Std. dev.
## s(Word)                                   0.258
## s(Word,CommunityRelTouristBeds.log.z)     0.025
## s(Word,CommunityAvgAge.z)                 0.031
## s(Word,CommunitySize.log.z)               0.020
## s(Word,CommunityAvgIncome.z)              0.032
## s(Word,SpeakerEduLevel.z)                 0.009
## s(Word,YBInCatAnd.z)                      0.029
## s(Word,YBInAragon.z)                      0.019
## s(Speaker)                                0.025
## s(Speaker,WordRefVowelRatio.z)            0.009
## s(Speaker,WordCategoryIsACD)              0.018
## s(Speaker,WordLength.z)                   0.013
## s(Location)                               0.026
## s(Location,YBInCatAnd.z)                  0.021
## s(Location,WordRefVowelRatio.z)           0.015
## s(Location,WordCategoryIsACD)             0.071
## s(Location,WordLength.z)                  0.037
```

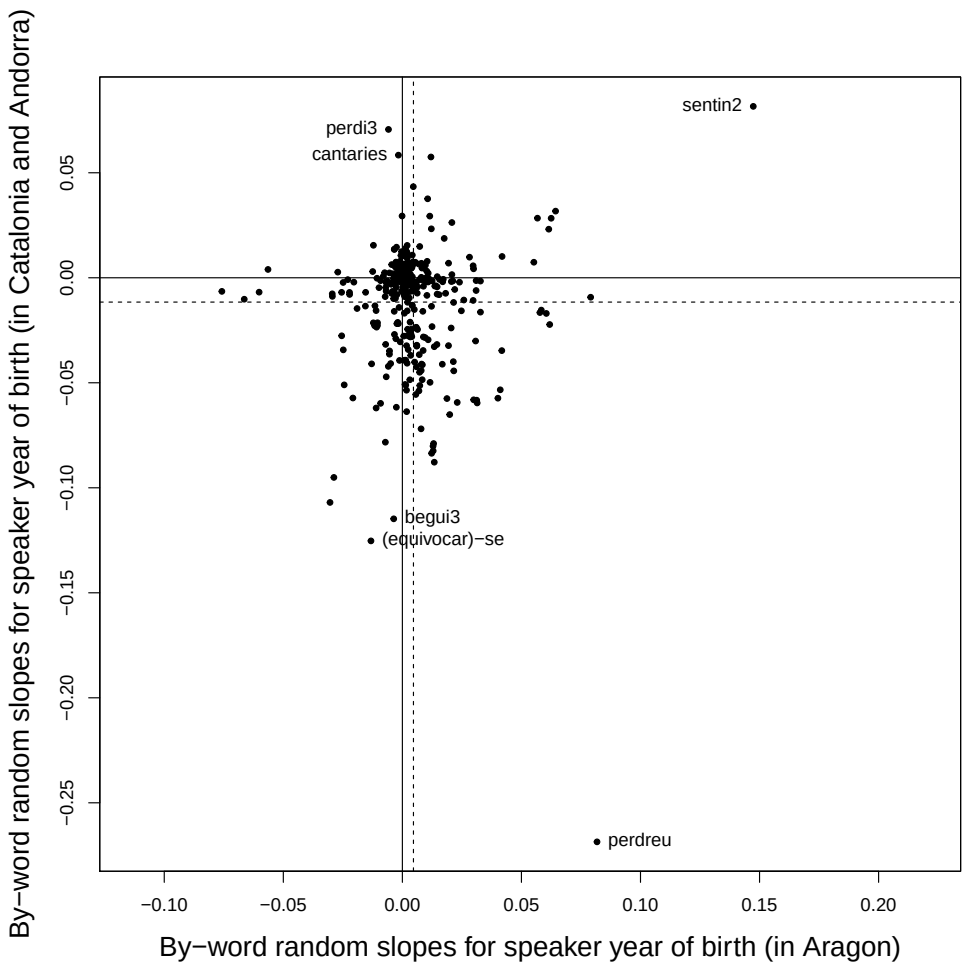
3.6 By-word random intercepts

```
plotSlope.gam(modelCatalanFinal, catalan, "Word", "Intercept",
              xlabel="Words sorted by their random intercept",
              ylabel="By-word random intercepts")
```



3.7 By-word random slopes

```
xlab = "By-word random slopes for speaker year of birth (in Aragon)"
ylab = "By-word random slopes for speaker year of birth (in Catalonia and Andorra)"

plotSlopes.gam(modelCatalanFinal,catalan,"Word","YBInAragon.z",
               "YBInCatAnd.z", xlab, ylab, nrWords=3)
```



3.8 GAM showing explained variance by word only

```
modelCatalanWord <-
  bam(PronDistStdCatalan.c ~ s(Word, bs = "re"), data=catalan)

summary(modelCatalanWord)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ s(Word, bs = "re")
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.003953   0.015032  -0.263    0.793
##
```

```
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Word) 355.3    356 534.5  <2e-16
##
## R-sq.(adj) =  0.628   Deviance explained = 62.9%
## fREML = -12228  Scale est. = 0.046186  n = 112608
```

3.9 GAM showing explained variance by fixed effects only

```
modelCatalanFixef <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      YBInAragon.z + YBInCatAnd.z + s(Longitude,Latitude), data=catalan)

summary(modelCatalanFixef)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     YBInAragon.z + YBInCatAnd.z + s(Longitude, Latitude)
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.028656   0.001141 -25.115   <2e-16
## WordRefVowelRatio.z  0.113625   0.001007 112.799   <2e-16
## WordCategoryIsACD    0.100416   0.002144  46.827   <2e-16
## YBInAragon.z         0.001821   0.002204   0.827    0.408
## YBInCatAnd.z        -0.013087   0.001075 -12.170   <2e-16
##
## Approximate significance of smooth terms:
##                          edf Ref.df     F p-value
## s(Longitude,Latitude) 28.25  28.96 258.8  <2e-16
##
## R-sq.(adj) =   0.16   Deviance explained =   16%
## fREML =  32647  Scale est. = 0.10438   n = 112608
```

3.10 GAM assessing differences in effect of year of birth

```
modelCatalanFinalContrast <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      SpeakerBirthYear.z + SpeakerBirthYear.z:InAragon +
      s(Longitude,Latitude) +
      s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge.z, bs = "re") +
```

```
      s(Word, CommunitySize.log.z, bs = "re") +
      s(Word, CommunityAvgIncome.z, bs = "re") +
      s(Word, SpeakerEduLevel.z, bs = "re") +
      s(Word, YBInAragon.z, bs = "re") +
      s(Word, YBInCatAnd.z, bs = "re")+
      s(Speaker, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re") +
      s(Speaker, WordLength.z, bs = "re") +
      s(Location, bs = "re") +
      s(Location, YBInCatAnd.z, bs = "re") +
      s(Location, WordRefVowelRatio.z, bs = "re") +
      s(Location, WordCategoryIsACD, bs = "re") +
      s(Location, WordLength.z, bs = "re"), data=catalan,
      cluster=cl)

smry <- summary(modelCatalanFinalContrast)

# model saved as calculations take about 20 minutes on 36 cores
save(smry,modelCatalanFinalContrast,file='modelCatalanFinalContrast.rda')
```

```
load('results/modelCatalanFinalContrast.rda')
smry # note that the interaction (i.e. the difference) is significant

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     SpeakerBirthYear.z + SpeakerBirthYear.z:InAragon + s(Longitude,
##     Latitude) + s(Word, bs = "re") + s(Word, CommunityRelTouristBeds.log.z,
##     bs = "re") + s(Word, CommunityAvgAge.z, bs = "re") + s(Word,
##     CommunitySize.log.z, bs = "re") + s(Word, CommunityAvgIncome.z,
##     bs = "re") + s(Word, SpeakerEduLevel.z, bs = "re") + s(Word,
##     YBInAragon.z, bs = "re") + s(Word, YBInCatAnd.z, bs = "re") +
##     s(Speaker, bs = "re") + s(Speaker, WordRefVowelRatio.z, bs = "re") +
##     s(Speaker, WordCategoryIsACD, bs = "re") + s(Speaker, WordLength.z,
##     bs = "re") + s(Location, bs = "re") + s(Location, YBInCatAnd.z,
##     bs = "re") + s(Location, WordRefVowelRatio.z, bs = "re") +
##     s(Location, WordCategoryIsACD, bs = "re") + s(Location, WordLength.z,
##     bs = "re")
##
## Parametric coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -0.032870   0.017535  -1.875  0.06085
## WordRefVowelRatio.z         0.109073   0.014067   7.754 8.97e-15
## WordCategoryIsACD           0.101042   0.034055   2.967  0.00301
## SpeakerBirthYear.z         -0.011576   0.005263  -2.199  0.02786
## SpeakerBirthYear.z:InAragon  0.016226   0.006802   2.385  0.01706
```

```
##
## Approximate significance of smooth terms:
##                                       edf Ref.df        F  p-value
## s(Longitude,Latitude)               11.94  12.02   16.981  < 2e-16
## s(Word)                            353.33 354.00 1747.856  < 2e-16
## s(Word,CommunityRelTouristBeds.log.z) 297.21 357.00   19.385  < 2e-16
## s(Word,CommunityAvgAge.z)          275.14 357.00  106.766  < 2e-16
## s(Word,CommunitySize.log.z)        235.18 357.00   50.689  < 2e-16
## s(Word,CommunityAvgIncome.z)       287.33 357.00  110.505  < 2e-16
## s(Word,SpeakerEduLevel.z)          190.64 357.00    3.025  < 2e-16
## s(Word,YBInAragon.z)               189.25 356.00    2.290  < 2e-16
## s(Word,YBInCatAnd.z)               310.81 356.00   30.615  < 2e-16
## s(Speaker)                         223.25 315.00   21.037  0.00041
## s(Speaker,WordRefVowelRatio.z)     164.35 319.00    1.803  < 2e-16
## s(Speaker,WordCategoryIsACD)       135.46 319.00   10.034  < 2e-16
## s(Speaker,WordLength.z)            183.92 320.00    5.472 5.50e-09
## s(Location)                         23.86  37.00  478.025 6.12e-05
## s(Location,YBInCatAnd.z)            25.99  31.00  550.143  < 2e-16
## s(Location,WordRefVowelRatio.z)     33.90  39.00  127.233 1.44e-09
## s(Location,WordCategoryIsACD)       36.86  39.00 7419.639  < 2e-16
## s(Location,WordLength.z)            37.10  40.00 6916.822  < 2e-16
##
## R-sq.(adj) =  0.735   Deviance explained = 74.2%
## fREML = -28237  Scale est. = 0.032891  n = 112608
```

3.11 GAM assessing border effect (without geographical smooth)

```
modelCatalanNoGeo  <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      YBInAragon.z + YBInCatAnd.z + InAragon +
      s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge.z, bs = "re") +
      s(Word, CommunitySize.log.z, bs = "re") +
      s(Word, CommunityAvgIncome.z, bs = "re") +
      s(Word, YBInCatAnd.z, bs = "re") +
      s(Word, YBInAragon.z, bs = "re") +
      s(Speaker, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re") +
      s(Speaker, WordLength.z, bs = "re") +
      s(Location, bs = "re") +
      s(Location, YBInCatAnd.z, bs = "re") +
      s(Location, WordRefVowelRatio.z, bs = "re") +
      s(Location, WordCategoryIsACD, bs = "re") +
      s(Location, WordLength.z, bs = "re"),
      data=catalan, cluster=cl
     )
```

```r
smry <- summary(modelCatalanNoGeo)

# model saved as calculations take about 20 minutes on 36 cores
save(modelCatalanNoGeo,smry,file='modelCatalanNoGeo.rda')
```

```r
load('results/modelCatalanNoGeo.rda')
smry # note that the difference between regions (InAragon) is significant
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     YBInAragon.z + YBInCatAnd.z + InAragon + s(Word, bs = "re") +
##     s(Word, CommunityRelTouristBeds.log.z, bs = "re") + s(Word,
##     CommunityAvgAge.z, bs = "re") + s(Word, CommunitySize.log.z,
##     bs = "re") + s(Word, CommunityAvgIncome.z, bs = "re") + s(Word,
##     YBInCatAnd.z, bs = "re") + s(Word, YBInAragon.z, bs = "re") +
##     s(Speaker, bs = "re") + s(Speaker, WordRefVowelRatio.z, bs = "re") +
##     s(Speaker, WordCategoryIsACD, bs = "re") + s(Speaker, WordLength.z,
##     bs = "re") + s(Location, YBInCatAnd.z, bs = "re") + s(Location,
##     WordRefVowelRatio.z, bs = "re") + s(Location, WordCategoryIsACD,
##     bs = "re") + s(Location, WordLength.z, bs = "re") + s(Location,
##     bs = "re")
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.037345   0.025552  -1.461 0.143888
## WordRefVowelRatio.z  0.114824   0.014317   8.020 1.09e-15
## WordCategoryIsACD    0.151450   0.040441   3.745 0.000181
## YBInAragon.z         0.004044   0.003736   1.082 0.279080
## YBInCatAnd.z        -0.020350   0.007242  -2.810 0.004956
## InAragon             0.111511   0.026794   4.162 3.16e-05
##
## Approximate significance of smooth terms:
##                                        edf Ref.df         F  p-value
## s(Word)                            352.126    354  2207.922  < 2e-16
## s(Word,CommunityRelTouristBeds.log.z) 271.275    357    44.950  < 2e-16
## s(Word,CommunityAvgAge.z)              282.976    357   943.543  < 2e-16
## s(Word,CommunitySize.log.z)            177.587    357    66.085  < 2e-16
## s(Word,CommunityAvgIncome.z)           312.636    357   839.266  < 2e-16
## s(Word,YBInCatAnd.z)                   247.241    356     8.841  < 2e-16
## s(Word,YBInAragon.z)                   208.179    356     2.666  < 2e-16
## s(Speaker)                              91.375    124    21.854 0.010649
## s(Speaker,WordRefVowelRatio.z)          76.032    127     3.003 5.68e-07
## s(Speaker,WordCategoryIsACD)            51.630    127     4.329 0.000687
## s(Speaker,WordLength.z)                 52.203    128     1.600 4.61e-05
## s(Location,YBInCatAnd.z)                 5.277      7   232.507 0.000284
## s(Location,WordRefVowelRatio.z)         11.289     15   146.622 2.47e-05
```

```
## s(Location,WordCategoryIsACD)          13.837     15 23523.238 1.43e-13
## s(Location,WordLength.z)               14.911     16 23554.622 3.55e-11
## s(Location)                            13.323     14 12739.877  < 2e-16
##
## R-sq.(adj) =  0.769   Deviance explained =   78%
## fREML = -14407  Scale est. = 0.026753  n = 44761
```

3.12 GAM assessing border effect on sites close to the border

```
# subset to all Aragonese locations and eight locations in Catalonia close
# to the border with Aragon
catalanBorder = catalan[catalan$InAragon == 1 |
                        catalan$Location %in% c("Vilaller","El Pont de Suert",
                        "Salas de Pallars","Tremp","Lleida",
                        "Montoliu de Lleida","Caseres","Alfara de Carles"),]
catalanBorder = droplevels(catalanBorder)

modelCatalanBorderNoGeo  <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      YBInAragon.z + YBInCatAnd.z + InAragon +
      s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge.z, bs = "re") +
      s(Word, CommunitySize.log.z, bs = "re") +
      s(Word, CommunityAvgIncome.z, bs = "re") +
      s(Word, YBInCatAnd.z, bs = "re") +
      s(Word, YBInAragon.z, bs = "re") +
      s(Speaker, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re") +
      s(Speaker, WordLength.z, bs = "re") +
      s(Location, bs = "re") +
      s(Location, YBInCatAnd.z, bs = "re") +
      s(Location, WordRefVowelRatio.z, bs = "re") +
      s(Location, WordCategoryIsACD, bs = "re") +
      s(Location, WordLength.z, bs = "re"), data=catalanBorder,
      cluster=cl
     )

smry <- summary(modelCatalanBorderNoGeo)

# model saved as calculations take about 10 minutes on 36 cores
save(modelCatalanBorderNoGeo,smry,file='modelCatalanBorderNoGeo.rda')
```

```
load('results/modelCatalanBorderNoGeo.rda')
smry # note that the difference between regions (InAragon) is significant

##
```

```
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     YBInAragon.z + YBInCatAnd.z + InAragon + s(Word, bs = "re") +
##     s(Word, CommunityRelTouristBeds.log.z, bs = "re") + s(Word,
##     CommunityAvgAge.z, bs = "re") + s(Word, CommunitySize.log.z,
##     bs = "re") + s(Word, CommunityAvgIncome.z, bs = "re") + s(Word,
##     YBInCatAnd.z, bs = "re") + s(Word, YBInAragon.z, bs = "re") +
##     s(Speaker, bs = "re") + s(Speaker, WordRefVowelRatio.z, bs = "re") +
##     s(Speaker, WordCategoryIsACD, bs = "re") + s(Speaker, WordLength.z,
##     bs = "re") + s(Location, bs = "re") + s(Location, YBInCatAnd.z,
##     bs = "re") + s(Location, WordRefVowelRatio.z, bs = "re") +
##     s(Location, WordCategoryIsACD, bs = "re") + s(Location, WordLength.z,
##     bs = "re")
##
## Parametric coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.037345   0.025552  -1.461 0.143888
## WordRefVowelRatio.z  0.114824   0.014317   8.020 1.09e-15
## WordCategoryIsACD    0.151450   0.040441   3.745 0.000181
## YBInAragon.z         0.004044   0.003736   1.082 0.279080
## YBInCatAnd.z        -0.020350   0.007242  -2.810 0.004956
## InAragon             0.111511   0.026794   4.162 3.16e-05
##
## Approximate significance of smooth terms:
##                                         edf Ref.df         F  p-value
## s(Word)                             352.126    354  2207.922  < 2e-16
## s(Word,CommunityRelTouristBeds.log.z) 271.275    357    44.950  < 2e-16
## s(Word,CommunityAvgAge.z)           282.976    357   943.543  < 2e-16
## s(Word,CommunitySize.log.z)         177.587    357    66.085  < 2e-16
## s(Word,CommunityAvgIncome.z)        312.636    357   839.266  < 2e-16
## s(Word,YBInCatAnd.z)                247.241    356     8.841  < 2e-16
## s(Word,YBInAragon.z)                208.179    356     2.666  < 2e-16
## s(Speaker)                           91.375    124    21.854 0.010649
## s(Speaker,WordRefVowelRatio.z)       76.032    127     3.003 5.68e-07
## s(Speaker,WordCategoryIsACD)         51.630    127     4.329 0.000687
## s(Speaker,WordLength.z)              52.203    128     1.600 4.61e-05
## s(Location)                          13.323     14 12739.877  < 2e-16
## s(Location,YBInCatAnd.z)              5.277      7   232.507 0.000284
## s(Location,WordRefVowelRatio.z)      11.289     15   146.622 2.47e-05
## s(Location,WordCategoryIsACD)        13.837     15 23523.238 1.43e-13
## s(Location,WordLength.z)             14.911     16 23554.622 3.55e-11
##
## R-sq.(adj) =  0.769   Deviance explained =   78%
## fREML = -14407  Scale est. = 0.026753  n = 44761
```

3.13 GAM assessing if older people in urban sites are more standard-like

```
modelCatalanUrban <-
  bam(PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
      IsUrban*YBInAragon.z + IsUrban*YBInCatAnd.z +
      s(Longitude, Latitude) +
      s(Word, bs = "re") +
      s(Word, CommunityRelTouristBeds.log.z, bs = "re") +
      s(Word, CommunityAvgAge.z, bs = "re") +
      s(Word, CommunitySize.log.z, bs = "re") +
      s(Word, IsUrban, bs = "re") +
      s(Word, CommunityAvgIncome.z, bs = "re") +
      s(Word, SpeakerEduLevel.z, bs = "re") +
      s(Word, YBInCatAnd.z, bs = "re") +
      s(Word, YBInAragon.z, bs = "re") +
      s(Speaker, bs = "re") +
      s(Speaker, WordRefVowelRatio.z, bs = "re") +
      s(Speaker, WordCategoryIsACD, bs = "re") +
      s(Speaker, WordLength.z, bs="re") +
      s(Location, bs="re") +
      s(Location, YBInCatAnd.z, bs="re") +
      s(Location, WordRefVowelRatio.z, bs = "re") +
      s(Location, WordCategoryIsACD, bs = "re") +
      s(Location, WordLength.z, bs="re"), data=catalan,
      gc.level=2
     )

smry <- summary(modelCatalanUrban)

# model saved as calculations take about 10 minutes on 36 cores
save(modelCatalanUrban,smry,file='modelCatalanUrban.rda')
```

```
load('results/modelCatalanUrban.rda')
smry # the interactions are not significant: older urban people not more std.

##
## Family: gaussian
## Link function: identity
##
## Formula:
## PronDistStdCatalan.c ~ WordRefVowelRatio.z + WordCategoryIsACD +
##     IsUrban * YBInAragon.z + IsUrban * YBInCatAnd.z + s(Longitude,
##     Latitude) + s(Word, bs = "re") + s(Word, CommunityRelTouristBeds.log.z,
##     bs = "re") + s(Word, CommunityAvgAge.z, bs = "re") + s(Word,
##     CommunitySize.log.z, bs = "re") + s(Word, IsUrban, bs = "re") +
##     s(Word, CommunityAvgIncome.z, bs = "re") + s(Word, SpeakerEduLevel.z,
##     bs = "re") + s(Word, YBInCatAnd.z, bs = "re") + s(Word, YBInAragon.z,
##     bs = "re") + s(Speaker, bs = "re") + s(Speaker, WordRefVowelRatio.z,
##     bs = "re") + s(Speaker, WordCategoryIsACD, bs = "re") + s(Speaker,
##     WordLength.z, bs = "re") + s(Location, bs = "re") + s(Location,
##     YBInCatAnd.z, bs = "re") + s(Location, WordRefVowelRatio.z,
##     bs = "re") + s(Location, WordCategoryIsACD, bs = "re") +
```

```
##     s(Location, WordLength.z, bs = "re")
##
## Parametric coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -0.028763   0.019082  -1.507  0.13173
## WordRefVowelRatio.z   0.109501   0.014500   7.552 4.32e-14
## WordCategoryIsACD     0.100502   0.034963   2.875  0.00405
## IsUrban              -0.007899   0.013696  -0.577  0.56412
## YBInAragon.z          0.005147   0.005928   0.868  0.38529
## YBInCatAnd.z         -0.016957   0.007237  -2.343  0.01913
## IsUrban:YBInAragon.z -0.001248   0.008157  -0.153  0.87835
## IsUrban:YBInCatAnd.z  0.010712   0.009946   1.077  0.28146
##
## Approximate significance of smooth terms:
##                                          edf Ref.df         F  p-value
## s(Longitude,Latitude)                  11.46  11.53    17.006  < 2e-16
## s(Word)                               352.13 354.00 10572.021  < 2e-16
## s(Word,CommunityRelTouristBeds.log.z) 297.13 357.00    20.003  < 2e-16
## s(Word,CommunityAvgAge.z)             288.10 357.00   265.867  < 2e-16
## s(Word,CommunitySize.log.z)           241.31 357.00   188.870  < 2e-16
## s(Word,IsUrban)                       290.29 356.00  7525.943  < 2e-16
## s(Word,CommunityAvgIncome.z)          283.51 357.00   181.522  < 2e-16
## s(Word,SpeakerEduLevel.z)             193.36 357.00     3.852  < 2e-16
## s(Word,YBInCatAnd.z)                  311.70 356.00    34.354  < 2e-16
## s(Word,YBInAragon.z)                  184.21 356.00     2.007  < 2e-16
## s(Speaker)                            222.70 312.00    21.560 0.000327
## s(Speaker,WordRefVowelRatio.z)        163.55 319.00     1.755  < 2e-16
## s(Speaker,WordCategoryIsACD)          137.66 319.00    10.598  < 2e-16
## s(Speaker,WordLength.z)               184.69 320.00     5.673 5.30e-09
## s(Location)                            23.58  36.00   434.060 0.000245
## s(Location,YBInCatAnd.z)               25.09  30.00   570.581  < 2e-16
## s(Location,WordRefVowelRatio.z)        33.82  39.00   136.152 4.14e-07
## s(Location,WordCategoryIsACD)          36.63  39.00  7670.972  < 2e-16
## s(Location,WordLength.z)               36.94  40.00  7148.463  < 2e-16
##
## R-sq.(adj) =   0.74   Deviance explained = 74.7%
## fREML = -28699  Scale est. = 0.032353  n = 112608
```

4 Data set for individual linguistic variables

```
load("data/catalanVars.rda")
```

Legenda catalanVars (11516 observations of 15 variables):

Note that the columns with a suffix of `.z` are not described here. This is simply a standardized (mean equals zero and standard deviation equals 1) version of the corresponding variables shown below.

1. Word : the word for which responses (with respect to the linguistic variables) were collected
2. Speaker : the speaker whose responses were collected
3. Location : the location where the speaker originates from
4. Var1NonStd : first linguistic variable: replacement of [ʎ] (standard) by [j] (non-standard)
5. Var2NonStd : second linguistic variable: variation in the final morphemes for the Present Subjunctive ([i]: standard, other vowels: non-standard)
6. Var3NonStd : third linguistic variable: use of [β] as opposed to another consonant (mainly [w]) within the possessive adjectives
7. Longitude : longitude of the dialect location
8. Latitude : latitude of the dialect location
9. InAragon : binary value indicating if the location is in Aragon (1) or not (0)
10. Region : location in Catalonia or Andorra (CataloniaAndorra) or Aragon (Aragon)
11. SpeakerIsMale : binary value indicating if the speaker is male (1) or not (0)
12. SpeakerEduLevel : education level of the speaker from low (0) to high (5)
13. SpeakerBirthYear : year of birth of the speaker

5 Analysis and results of individual variables

5.1 Model of the first linguistic variable: [ʎ] vs. [j]

```
subsetV1 = droplevels(catalanVars[!is.na(catalanVars$Var1NonStd), ])
dim(subsetV1)

## [1] 3200   15

modelV1 <-
  bam(Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Speaker, bs="re"), data=subsetV1,
                   family="binomial"
      )

summary(modelV1)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -4.95149    0.59110  -8.377  < 2e-16
## SpeakerIsMale                             1.06708    0.61857   1.725   0.0845
## SpeakerEduLevel.z                         0.01848    0.36460   0.051   0.9596
## RegionCataloniaAndorra:SpeakerBirthYear.z  3.13321    0.56493   5.546 2.92e-08
## RegionAragon:SpeakerBirthYear.z           6.44254    1.16902   5.511 3.57e-08
##
## Approximate significance of smooth terms:
##                            edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)    9.426  10.75  34.98 0.00018
## s(Speaker)             115.415 313.00 539.62 < 2e-16
##
## R-sq.(adj) =  0.938   Deviance explained = 92.3%
## fREML = 3227.1  Scale est. = 1         n = 3200
```

5.2 Model of the second linguistic variable: [i] vs. other vowel

```
subsetV2 = droplevels(catalanVars[!is.na(catalanVars$Var2NonStd), ])
dim(subsetV2)

## [1] 6400   15

modelV2 <-
```

```
  bam(Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Word, bs="re") +
                   s(Speaker, bs="re"), data=subsetV2,
                   family="binomial"
      )

summary(modelV2)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Word, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                 1.1301     0.3922   2.882  0.00395
## SpeakerIsMale                              -0.3956     0.4172  -0.948  0.34301
## SpeakerEduLevel.z                          -0.4341     0.2036  -2.132  0.03297
## RegionCataloniaAndorra:SpeakerBirthYear.z  -1.0215     0.2107  -4.849 1.24e-06
## RegionAragon:SpeakerBirthYear.z             0.3132     0.8165   0.384  0.70129
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)   20.50  22.04  176.3  <2e-16
## s(Word)                 17.36  19.00  423.4  <2e-16
## s(Speaker)             177.45 313.00 1594.4  <2e-16
##
## R-sq.(adj) =  0.818   Deviance explained = 79.1%
## fREML = 7355.4  Scale est. = 1         n = 6400
```

5.3 Model of the third linguistic variable: [β] vs. other consonant

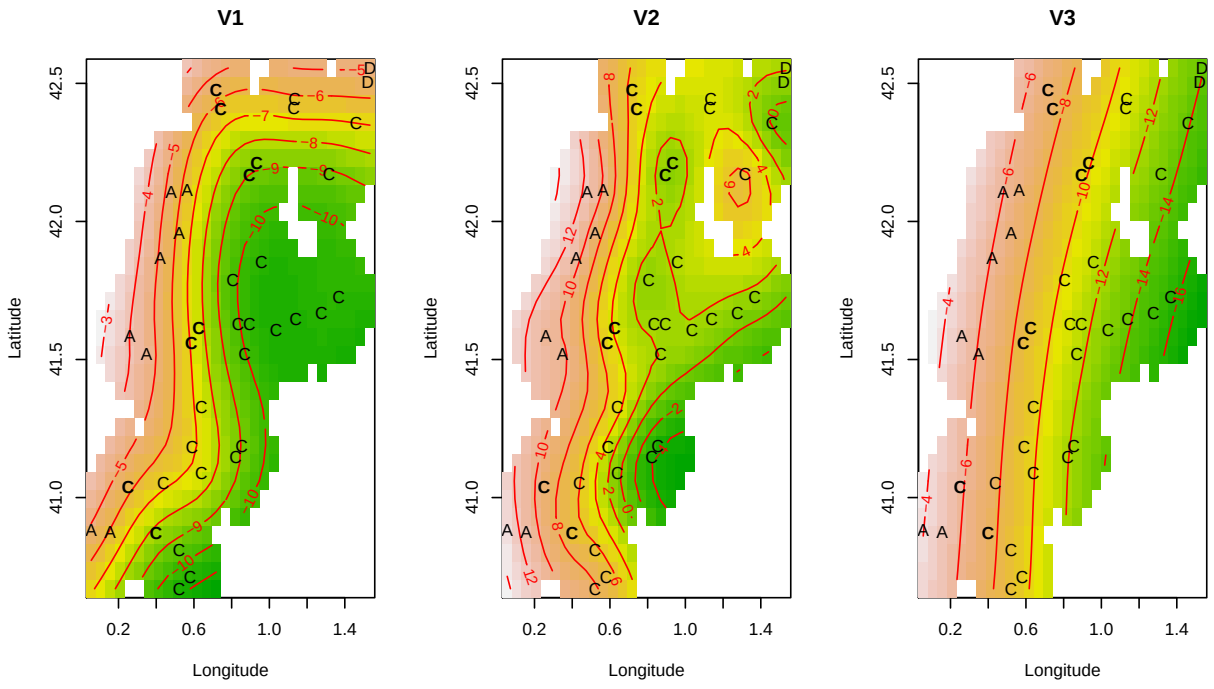
```
subsetV3 = droplevels(catalanVars[!is.na(catalanVars$Var3NonStd), ])
dim(subsetV3)

## [1] 1916   15

# Aragon: only non-standard forms - no variation...
table(subsetV3$Var3NonStd,subsetV3$Region)

##
##     CataloniaAndorra Aragon
##   0              531      0
##   1             1005    380
```

```r
# Add dummy word which is equal to standard for everybody
# In this way Aragon also has some variation, otherwise standard
# errors cannot be determined
subsetV3$Word = as.character(subsetV3$Word)
mv = subsetV3[subsetV3$Word == 'meves',]
mv$Word = 'dummy'
mv$Var3NonStd = 0 # dummy word which is equal to standard
subsetV3 = rbind(subsetV3,mv)
subsetV3$Word = as.factor(subsetV3$Word)

# Aragon: now also has standard pronunciations
table(subsetV3$Var3NonStd,subsetV3$Region)

##
##     CataloniaAndorra Aragon
##   0              787     64
##   1             1005    380

modelV3 <-
  bam(Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Location, bs="re") +
                   s(Word, bs="re") + s(Speaker, bs="re"), data=subsetV3,
                   family="binomial"
      )

summary(modelV3)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region:SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Location, bs = "re") + s(Word,
##     bs = "re") + s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                0.7610     1.8758   0.406    0.685
## SpeakerIsMale                              0.1941     0.4653   0.417    0.677
## SpeakerEduLevel.z                         -0.3789     0.2339  -1.620    0.105
## RegionCataloniaAndorra:SpeakerBirthYear.z -1.3745     0.2522  -5.451    5e-08
## RegionAragon:SpeakerBirthYear.z            0.2531     1.1686   0.217    0.829
##
## Approximate significance of smooth terms:
##                          edf  Ref.df  Chi.sq  p-value
## s(Longitude,Latitude)   3.81   4.112   45.02 5.44e-09
## s(Location)            19.96  37.000  375.58 1.79e-05
## s(Word)                 5.88   6.000   96.79  < 2e-16
## s(Speaker)            110.73 313.000 3112.30  < 2e-16
##
```

```
## R-sq.(adj) =  0.925   Deviance explained = 90.1%
## fREML = 2369.4  Scale est. = 1         n = 2236
```

5.4 Visualization of the geographical pattern per linguistic variable

```
fixedVars = list(SpeakerIsMale=0, SpeakerEduLevel.z=0, SpeakerBirthYear.z=0)
par(mfrow=c(1,3))
vis.gam(modelV1, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.12, cond=fixedVars, main="V1")

inBF = c("Vilaller","El Pont de Suert","Salas de Pallars","Tremp","Lleida",
         "Montoliu de Lleida","Caseres","Alfara de Carles") # printed in boldface

addLabels(catalan,inBF) # adds labels to plot

vis.gam(modelV2, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.12, cond=fixedVars, main="V2")

addLabels(catalan,inBF) # adds labels to plot

vis.gam(modelV3, view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.12, cond=fixedVars, main="V3")

addLabels(catalan,inBF) # adds labels to plot
```



5.5 Assessing significant differences in the effect of year of birth

```
# none of the differences in the effect of year of birth are significant,
# since the interactions are not significant

modelV1a <-
  bam(Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Speaker, bs="re"), data=subsetV1,
                   family="binomial"
      )

summary(modelV1a)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.745315   0.759309  -7.567 3.83e-14
## SpeakerIsMale                   1.147917   0.627817   1.828   0.0675
## SpeakerEduLevel.z               0.005322   0.380012   0.014   0.9888
## RegionAragon                    3.264825   1.621534   2.013   0.0441
## SpeakerBirthYear.z              3.428404   0.650923   5.267 1.39e-07
## RegionAragon:SpeakerBirthYear.z 2.346001   1.308588   1.793   0.0730
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq p-value
## s(Longitude,Latitude)   8.875   10.1  29.09 0.00115
## s(Speaker)            112.672  312.0 530.98 < 2e-16
##
## R-sq.(adj) =  0.938   Deviance explained = 92.4%
## fREML = 3222.1  Scale est. = 1         n = 3200

modelV2a <-
  bam(Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Word, bs="re") +
                   s(Speaker, bs="re"), data=subsetV2,
                   family="binomial"
      )

summary(modelV2a)

##
## Family: binomial
## Link function: logit
##
```

```
## Formula:
## Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Word, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      0.2704     0.6335   0.427   0.6695
## SpeakerIsMale                   -0.3869     0.4215  -0.918   0.3586
## SpeakerEduLevel.z               -0.4482     0.2066  -2.169   0.0301
## RegionAragon                     5.3950     3.2672   1.651   0.0987
## SpeakerBirthYear.z              -1.0157     0.2120  -4.790 1.66e-06
## RegionAragon:SpeakerBirthYear.z  1.3463     0.9431   1.428   0.1534
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq  p-value
## s(Longitude,Latitude)  18.82  20.17  116.5 2.32e-15
## s(Word)                17.36  19.00  424.5  < 2e-16
## s(Speaker)            176.08 312.00 1595.2  < 2e-16
##
## R-sq.(adj) =  0.818   Deviance explained = 79.2%
## fREML = 7350.8  Scale est. = 1         n = 6400

modelV3a <-
  bam(Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Longitude, Latitude) + s(Location, bs="re") +
                   s(Word, bs="re") + s(Speaker, bs="re"), data=subsetV3,
                   family="binomial"
      )

summary(modelV3a)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Longitude, Latitude) + s(Location, bs = "re") + s(Word,
##     bs = "re") + s(Speaker, bs = "re")
##
## Parametric coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      0.5919     1.9713   0.300    0.764
## SpeakerIsMale                    0.1984     0.4688   0.423    0.672
## SpeakerEduLevel.z               -0.3825     0.2358  -1.622    0.105
## RegionAragon                     1.0610     2.0154   0.526    0.599
## SpeakerBirthYear.z              -1.3718     0.2538  -5.404 6.51e-08
## RegionAragon:SpeakerBirthYear.z  1.6239     1.3249   1.226    0.220
##
## Approximate significance of smooth terms:
```

```
##                            edf  Ref.df  Chi.sq  p-value
## s(Longitude,Latitude)    3.835   4.108   31.49 2.99e-06
## s(Location)             19.402  36.000  364.51 2.88e-05
## s(Word)                  5.854   6.000   69.61 2.79e-12
## s(Speaker)             109.883 312.000 3251.21  < 2e-16
##
## R-sq.(adj) =  0.926   Deviance explained = 90.1%
## fREML = 2366.6  Scale est. = 1         n = 2236
```

5.6 Assessing border effect (without geographical smooth)

```
# RegionAragon is significant everywhere

modelV1b <-
  bam(Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Speaker, bs="re") + s(Location,bs='re'), data=subsetV1,
                   family="binomial"
      )

summary(modelV1b)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Speaker, bs = "re") + s(Location, bs = "re")
##
## Parametric coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -5.50288    0.72579  -7.582 3.40e-14
## SpeakerIsMale                    1.09861    0.61115   1.798  0.07224
## SpeakerEduLevel.z                0.02078    0.36955   0.056  0.95516
## RegionAragon                     3.72923    1.16052   3.213  0.00131
## SpeakerBirthYear.z               3.11654    0.59870   5.205 1.93e-07
## RegionAragon:SpeakerBirthYear.z  2.73231    1.29095   2.117  0.03430
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq p-value
## s(Speaker)  111.10    314  655.6  0.0322
## s(Location)  17.23     38  381.3  0.5431
##
## R-sq.(adj) =  0.938   Deviance explained = 92.3%
## fREML = 3234.6  Scale est. = 1         n = 3200

modelV2b <-
  bam(Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Word, bs="re") + s(Location,bs='re') +
```

```
                  s(Speaker, bs="re"), data=subsetV2,
                  family="binomial"
      )

summary(modelV2b)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Word, bs = "re") + s(Location, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.3892     0.6111  -0.637   0.5242
## SpeakerIsMale                   -0.4343     0.4155  -1.045   0.2960
## SpeakerEduLevel.z               -0.4480     0.2088  -2.145   0.0319
## RegionAragon                     8.1161     1.6161   5.022 5.11e-07
## SpeakerBirthYear.z              -1.0060     0.2064  -4.874 1.09e-06
## RegionAragon:SpeakerBirthYear.z  1.1381     0.9347   1.218   0.2234
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq  p-value
## s(Word)      17.35     19   1442  < 2e-16
## s(Location)  29.73     38  22211 1.85e-10
## s(Speaker)  168.81    314  28206 7.69e-15
##
## R-sq.(adj) =  0.818   Deviance explained = 79.1%
## fREML = 7346.2  Scale est. = 1         n = 6400

modelV3b <-
  bam(Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                  s(Location, bs="re") + s(Word, bs="re") + s(Speaker, bs="re"),
                  data=subsetV3, family="binomial"
      )

summary(modelV3b)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Location, bs = "re") + s(Word, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
```

```
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -0.2997     1.9059  -0.157  0.87507
## SpeakerIsMale                     0.1916     0.4607   0.416  0.67744
## SpeakerEduLevel.z                -0.3970     0.2313  -1.716  0.08617
## RegionAragon                      5.4867     1.8896   2.904  0.00369
## SpeakerBirthYear.z               -1.3341     0.2469  -5.402 6.58e-08
## RegionAragon:SpeakerBirthYear.z   1.5929     1.2928   1.232  0.21791
##
## Approximate significance of smooth terms:
##                 edf Ref.df  Chi.sq  p-value
## s(Location)  28.528     38 2556.58  < 2e-16
## s(Word)       5.865      6   69.67 7.76e-12
## s(Speaker)  108.273    314 3571.58  < 2e-16
##
## R-sq.(adj) =  0.925   Deviance explained =   90%
## fREML = 2380.5  Scale est. = 1         n = 2236
```

5.7 Assessing border effect on sites close to the border

```
# RegionAragon is (again) significant everywhere

subsetV1Border = subsetV1[subsetV1$Region == "Aragon" |
                          subsetV1$Location %in% c("Vilaller","El Pont de Suert",
                          "Salas de Pallars","Tremp","Lleida",
                          "Montoliu de Lleida","Caseres","Alfara de Carles"),]
subsetV1Border = droplevels(subsetV1Border)

modelV1c <-
  bam(Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Speaker, bs="re") + s(Location,bs='re'), data=subsetV1Border,
                   family="binomial"
      )

summary(modelV1c)

##
## Family: binomial
## Link function: logit
##
## Formula:
## Var1NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Speaker, bs = "re") + s(Location, bs = "re")
##
## Parametric coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.2673     1.6963  -3.695  0.00022
## SpeakerIsMale                     1.5287     0.9924   1.540  0.12345
## SpeakerEduLevel.z                -0.6411     0.6431  -0.997  0.31881
## RegionAragon                      4.2814     1.8146   2.359  0.01830
```

```
## SpeakerBirthYear.z                4.4906     1.6751   2.681  0.00735
## RegionAragon:SpeakerBirthYear.z   1.7181     2.0888   0.823  0.41076
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq  p-value
## s(Speaker)  45.643    122 257.14 1.15e-07
## s(Location)  2.354     14  14.88    0.689
##
## R-sq.(adj) =  0.946   Deviance explained =   93%
## fREML = 1292.8  Scale est. = 1         n = 1280
```

```
subsetV2Border = subsetV2[subsetV2$Region == "Aragon" |
                          subsetV2$Location %in% c("Vilaller","El Pont de Suert",
                          "Salas de Pallars","Tremp","Lleida",
                          "Montoliu de Lleida","Caseres","Alfara de Carles"),]
subsetV2Border = droplevels(subsetV2Border)

modelV2c <-
  bam(Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Word, bs="re") + s(Location,bs='re') +
                   s(Speaker, bs="re"), data=subsetV2Border,
                   family="binomial"
      )

summary(modelV2c)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Var2NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Word, bs = "re") + s(Location, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.0033     1.0603   0.946 0.344021
## SpeakerIsMale                    -0.5279     0.7982  -0.661 0.508409
## SpeakerEduLevel.z                -0.2388     0.4194  -0.569 0.569112
## RegionAragon                      6.6051     1.6992   3.887 0.000101
## SpeakerBirthYear.z               -1.0622     0.4301  -2.470 0.013522
## RegionAragon:SpeakerBirthYear.z   1.1197     1.0180   1.100 0.271364
##
## Approximate significance of smooth terms:
##                edf Ref.df   Chi.sq  p-value
## s(Word)     11.361     19   39.04 1.11e-06
## s(Location)  8.212     14 4701.01  0.00159
## s(Speaker)  47.078    122 1599.10  0.02787
##
## R-sq.(adj) =   0.85   Deviance explained = 83.9%
## fREML = 2698.2  Scale est. = 1         n = 2560
```

```
subsetV3Border = subsetV3[subsetV3$Region == "Aragon" |
                          subsetV3$Location %in% c("Vilaller","El Pont de Suert",
                          "Salas de Pallars","Tremp","Lleida",
                          "Montoliu de Lleida","Caseres","Alfara de Carles"),]
subsetV3Border = droplevels(subsetV3Border)

modelV3c <-
  bam(Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region*SpeakerBirthYear.z +
                   s(Location, bs="re") + s(Word, bs="re") + s(Speaker, bs="re"),
                   data=subsetV3Border, family="binomial"
      )

summary(modelV3c)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## Var3NonStd ~ SpeakerIsMale + SpeakerEduLevel.z + Region * SpeakerBirthYear.z +
##     s(Location, bs = "re") + s(Word, bs = "re") + s(Speaker,
##     bs = "re")
##
## Parametric coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      2.0730     2.1418   0.968  0.33310
## SpeakerIsMale                   -0.5096     0.8908  -0.572  0.56724
## SpeakerEduLevel.z               -0.8815     0.5190  -1.698  0.08944
## RegionAragon                     3.8960     1.7875   2.180  0.02929
## SpeakerBirthYear.z              -2.2696     0.6901  -3.289  0.00101
## RegionAragon:SpeakerBirthYear.z  2.7519     1.6292   1.689  0.09120
##
## Approximate significance of smooth terms:
##                edf Ref.df Chi.sq  p-value
## s(Location)  4.267     14  83.58    0.011
## s(Word)      5.792      6  66.42 1.39e-09
## s(Speaker)  26.432    122 540.69  < 2e-16
##
## R-sq.(adj) =  0.957   Deviance explained = 93.8%
## fREML = 893.31  Scale est. = 1         n = 892
```