Data, data documentation and analysis scripts for

Lexical differences between Tuscan dialects and standard Italian: Accounting for geographic and socio-demographic variation using generalized additive mixed modeling

Martijn Wieling[1,2] & Simonetta Montemagni[3] & John Nerbonne[1,4] & R. Harald Baayen[2,5]

[1]University of Groningen, the Netherlands & [2]University of Tübingen, Germany & [3]National Research Council, Italy & [4]Freiburg Institute for Advanced Studies, Germany & [5]University of Alberta, Canada

Journal: **Language** (accepted for publication, October 2013)
Preprint: `http://martijnwieling.nl/files/WielingMontemagniNerbonneBaayen2013.pdf`

Abstract

This study uses a generalized additive mixed-effects regression model to predict lexical differences in Tuscan dialects with respect to standard Italian. We used lexical information for 170 concepts used by 2060 speakers in 213 locations in Tuscany. In our model, geographical position was found to be an important predictor, with locations more distant from Florence having lexical forms more likely to differ from standard Italian. In addition, the geographical pattern varied significantly for low versus high frequency concepts and older versus younger speakers. Younger speakers generally used variants more likely to match the standard language. Several other factors emerged as significant. Male speakers as well as farmers were more likely to use a lexical form different from standard Italian. In contrast, higher educated speakers used lexical forms more likely to match the standard. The model also indicates that lexical variants used in smaller communities are more likely to differ from standard Italian. The impact of community size, however, varied from concept to concept. For a majority of concepts, lexical variants used in smaller communities are more likely to differ from the standard Italian form. For a minority of concepts, however, lexical variants used in larger communities are more likely to differ from standard Italian. Similarly, the effect of the other community- and speaker-related predictors varied per concept. These results clearly show that the model succeeds in teasing apart different forces influencing the dialect landscape and helps us to shed light on the complex interaction between the standard Italian language and the Tuscan dialectal varieties. In addition, this study illustrates the potential of generalized additive mixed-effects regression modeling applied to dialect data.

Keywords: Tuscan dialects, Lexical variation, Generalized additive modeling, Mixed-effects regression modeling, Geographical variation.

1 Packages and functions

```
library(mgcv)
library(Hmisc)
require(parallel)

R.Version()$version.string

## [1] "R version 3.0.2 (2013-09-25)"

packageVersion("mgcv")

## [1] '1.7.27'

packageVersion("Hmisc")

## [1] '3.12.2'

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

2 Data set

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

Legenda tuscan (384454 observations of 32 variables):

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

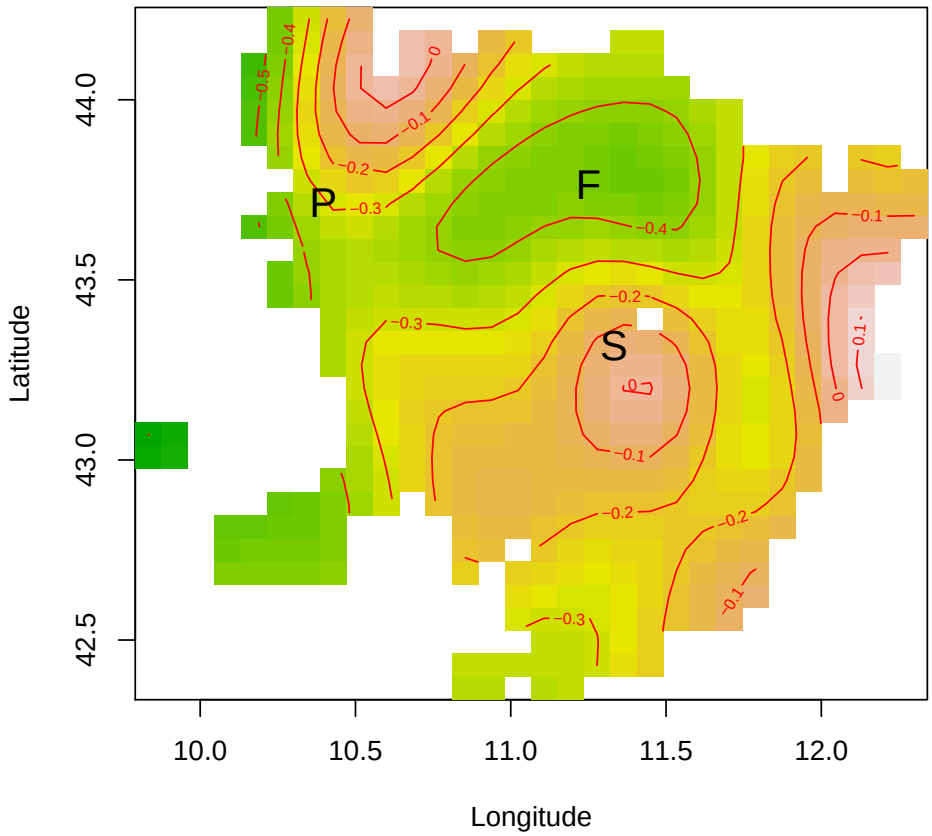
1. Concept : the concept for which lexical forms were obtained
2. Location : the location in which speakers were asked for their lexical form
3. Speaker : the speaker whose lexical variants were obtained
4. NormalizedLexicalVariant : the normalized lexical variant of the speaker
5. NormalizedVariantUnequalToStd_noMorphVariation : binary value indicating if the normalized form (excluding possible morphological variation) is different from the standard Italian form (1) or equal (0)
6. NormalizedVariantUnequalToStd_inclMorphVariation : binary value indicating if the normalized form (including morphological variation) is different from the standard Italian form (1) or equal (0)
7. Longitude : longitude of the location
8. Latitude : latitude of the location
9. ConceptFreq.log : frequency of the concept (log-transformed) from Web 1T 5-gram corpus
10. SpeakerBirthYear : year of birth of the speaker
11. SpeakerEduLevel.log : education level of the speaker (1: illiterate to 6: university degree)
12. SpeakerIsMale : gender of speaker (1: male, 0: female)
13. SpeakerJob_JOBNAME : binary value indicating if the speaker had the job JOBNAME (1) or not (0). Possible job names: Farmer, Craftsman, Trader_Businessman, Executive_AuxiliaryWorker, KnowlegeWorker_Manager_Nurse, Teacher_Freelance, CommonLaborer_Apprentice, SkilledWorker_QualifiedWorker, NonProfessional
14. CommunitySize.log : number of inhabitants in the location (log-transformed)
15. CommunityAvgAge.log : average age of inhabitants in the location (log-transformed)
16. CommunityAvgIncome.log : average income of inhabitants in the location (log-transformed)
17. CommunityRecordingYear : year of recording for the location

3 Analysis and results

```
geo <- bam(NormalizedVariantUnequalToStd_noMorphVariation ~ s(Longitude,Latitude),
           family="binomial", gc.level=2, data=tuscan)

summary(geo)

##
## Family: binomial
## Link function: logit
##
## Formula:
## NormalizedVariantUnequalToStd_noMorphVariation ~ s(Longitude,
##     Latitude)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.24796    0.00326     -76   <2e-16
##
## Approximate significance of smooth terms:
##                         edf Ref.df Chi.sq p-value
## s(Longitude,Latitude) 27.5   28.6   1580  <2e-16
##
## R-sq.(adj) =  0.00412   Deviance explained = 0.306%
## fREML score = 5.4556e+05  Scale est. = 1         n = 384454
```

Visualization of the simple generalized additive model

```
vis.gam(geo, view=c("Longitude","Latitude"), plot.type="contour",
             color="terrain", too.far=0.045, main="")
text(10.397, 43.716, "P",cex=1.5) # Pisa label
text(11.333, 43.320, "S",cex=1.5) # Siena label
text(11.250, 43.767, "F",cex=1.5) # Florence label
```


Latitude
44.0
43.5
43.0
42.5
10.0
10.5
11.0
11.5
12.0
Longitude
P
F
S
−0.6
−0.4
0
−0.1
−0.2
−0.3
−0.4
−0.2
−0.3
−0.1
−0.2
−0.2
−0.1
−0.3
−0.1
0
0.1

Complete mixed-effects regression model

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

modelTuscan <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          te(Longitude,Latitude,ConceptFreq.log.z,SpeakerBirthYear.z,d=c(2,1,1)) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerIsMale + s(Speaker,bs="re") + s(Location,bs="re") +
          s(Concept,bs="re") + s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
      ) # 11 hours

summaryModelTuscan <- summary(modelTuscan) # 47 minutes

save(modelTuscan,file='results/modelTuscan.rda')
save(summaryModelTuscan,file='results/summaryModelTuscan.rda')
```

```
load('results/summaryModelTuscan.rda')
summaryModelTuscan

##
## Family: binomial
## Link function: logit
##
## Formula:
## NormalizedVariantUnequalToStd_noMorphVariation ~ te(Longitude,
##     Latitude, ConceptFreq.log.z, SpeakerBirthYear.z, d = c(2,
##         1, 1)) + CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
##     SpeakerIsMale + s(Speaker, bs = "re") + s(Location, bs = "re") +
##     s(Concept, bs = "re") + s(Concept, CommunityRecordingYear.z,
##     bs = "re") + s(Concept, CommunitySize.log.z, bs = "re") +
##     s(Concept, CommunityAvgIncome.log.z, bs = "re") + s(Concept,
##     CommunityAvgAge.log.z, bs = "re") + s(Concept, SpeakerJob_Farmer,
##     bs = "re") + s(Concept, SpeakerJob_Executive_AuxiliaryWorker,
##     bs = "re") + s(Concept, SpeakerEduLevel.log.z, bs = "re") +
##     s(Concept, SpeakerIsMale, bs = "re")
##
## Parametric coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.4188     0.1266   -3.31  0.00094
## CommunitySize.log.z      -0.0584     0.0224   -2.60  0.00923
```

```
## SpeakerJob_Farmer       0.0460     0.0169    2.72  0.00648
## SpeakerEduLevel.log.z  -0.0686     0.0126   -5.44  5.2e-08
## SpeakerIsMale           0.0379     0.0128    2.96  0.00308
##
## Approximate significance of smooth terms:
##                                                    edf Ref.df Chi.sq p-value
## te(Long.,Lat.,ConceptFreq.log.z,SpeakerBirthYear.z) 225.9    271   3295 < 2e-16
## s(Speaker)                                          104.3   2005    114  0.0057
## s(Location)                                         174.2    209   5407 < 2e-16
## s(Concept)                                          167.1    168 435155 < 2e-16
## s(Concept,CommunityRecordingYear.z)                 158.8    170 155912 < 2e-16
## s(Concept,CommunitySize.log.z)                      149.9    169  30166 < 2e-16
## s(Concept,CommunityAvgIncome.log.z)                 158.0    170 142938 < 2e-16
## s(Concept,CommunityAvgAge.log.z)                    154.4    170 110096 < 2e-16
## s(Concept,SpeakerJob_Farmer)                         86.1    169  26433 8.9e-08
## s(Concept,SpeakerJob_Executive_AuxiliaryWorker)      53.3    170   3185  0.0015
## s(Concept,SpeakerEduLevel.log.z)                    139.0    169   9206 < 2e-16
## s(Concept,SpeakerIsMale)                             85.2    169 111167 4.6e-11
##
## R-sq.(adj) =  0.301   Deviance explained = 25.6%
## fREML score = 5.3589e+05  Scale est. = 1         n = 379496
```

Effect sizes

```
effectSizes = t(data.frame(
    getEffectSize.gam(tuscan,summaryModelTuscan,"CommunitySize.log.z"),
    getEffectSize.gam(tuscan,summaryModelTuscan,"SpeakerEduLevel.log.z"),
    getEffectSize.gam(tuscan,summaryModelTuscan,"SpeakerIsMale"),
    getEffectSize.gam(tuscan,summaryModelTuscan,"SpeakerJob_Farmer")
))
```

```
##                       Effect size
## CommunitySize.log.z       -0.36180
## SpeakerEduLevel.log.z     -0.27571
## SpeakerIsMale              0.03795
## SpeakerJob_Farmer          0.04601
```

Standard deviations of random effects

```
load('results/modelTuscan.rda')
coefs = coef(modelTuscan)
stdevs = t(data.frame(
    getSD.gam(coefs, "Speaker", "Intercept"),
    getSD.gam(coefs, "Location", "Intercept"),
    getSD.gam(coefs, "Concept", "Intercept"),
    getSD.gam(coefs, "Concept", "CommunityRecordingYear.z"),
    getSD.gam(coefs, "Concept", "CommunitySize.log.z"),
    getSD.gam(coefs, "Concept", "CommunityAvgIncome.log.z"),
    getSD.gam(coefs, "Concept", "CommunityAvgAge.log.z"),
    getSD.gam(coefs, "Concept", "SpeakerJob_Farmer"),
    getSD.gam(coefs, "Concept", "SpeakerJob_Executive_AuxiliaryWorker"),
    getSD.gam(coefs, "Concept", "SpeakerEduLevel.log.z"),
    getSD.gam(coefs, "Concept", "SpeakerIsMale")
))
```

```
##                                                 Std. dev.
## s(Speaker)                                        0.01004
## s(Location)                                       0.18742
## s(Concept)                                        1.62054
## s(Concept,CommunityRecordingYear.z)               0.28283
## s(Concept,CommunitySize.log.z)                    0.17685
## s(Concept,CommunityAvgIncome.log.z)               0.26565
## s(Concept,CommunityAvgAge.log.z)                  0.23998
## s(Concept,SpeakerJob_Farmer)                      0.10326
## s(Concept,SpeakerJob_Executive_AuxiliaryWorker)   0.06498
## s(Concept,SpeakerEduLevel.log.z)                  0.12552
## s(Concept,SpeakerIsMale)                          0.07974
```

Model performance

```
tuscanNM = tuscan[!is.na(tuscan$SpeakerEduLevel.log.z), ] # remove missing values
somers2(fitted(modelTuscan), tuscanNM$NormalizedVariantUnequalToStd_noMorphVariation)

##         C       Dxy         n   Missing
## 8.190e-01 6.381e-01 3.795e+05 0.000e+00
```

Testing the inclusion of geography

```
cl = makeCluster(4) # 4 cores used in calculating the models below

modelTuscanNoGeo <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerIsMale + s(Speaker,bs="re") + s(Location,bs="re") +
          s(Concept,bs="re") + s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
      ) # duration: 410 minutes

modelTuscanGeoSimple <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          s(Longitude,Latitude) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerIsMale + s(Speaker,bs="re") + s(Location,bs="re") +
          s(Concept,bs="re") + s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
      ) # duration: 393 minutes

modelTuscanGeoAge <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          te(Longitude,Latitude,SpeakerBirthYear.z,d=c(2,1)) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerIsMale + s(Speaker,bs="re") + s(Location,bs="re") +
          s(Concept,bs="re") + s(Concept,CommunityRecordingYear.z,bs="re") +
```

```
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
     ) # duration: 455 minutes

modelTuscanGeoFreq <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          te(Longitude,Latitude,ConceptFreq.log.z,d=c(2,1,1)) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerIsMale + s(Speaker,bs="re") + s(Location,bs="re") +
          s(Concept,bs="re") + s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
     ) # duration: 451 minutes

# models are saved as they take a long time to compute
save(modelTuscanNoGeo,file='results/modelTuscanNoGeo.rda')
save(modelTuscanGeoSimple,file='results/modelTuscanGeoSimple.rda')
save(modelTuscanGeoAge,file='results/modelTuscanGeoAge.rda')
save(modelTuscanGeoFreq,file='results/modelTuscanGeoFreq.rda')
```

```
load('results/modelTuscanNoGeo.rda')
load('results/modelTuscanGeoSimple.rda')
load('results/modelTuscanGeoAge.rda')
load('results/modelTuscanGeoFreq.rda')

AIC(modelTuscanNoGeo)

## [1] 393242

AIC(modelTuscanGeoSimple)

## [1] 393238

AIC(modelTuscanGeoAge)

## [1] 392727
```

```
AIC(modelTuscanGeoFreq)

## [1] 391041

AIC(modelTuscan)

## [1] 390479
```

Minimal influence of abstracting away from morphological variation

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

modelTuscanAlt <-
  bam(NormalizedVariantUnequalToStd_noMorphVariation ~
          te(Longitude,Latitude,ConceptFreq.log.z,SpeakerBirthYear.z,d=c(2,1,1)) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerJob_Teacher_Freelance + SpeakerIsMale +
          s(Speaker,bs="re") + s(Location,bs="re") + s(Concept,bs="re") +
          s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
      ) # 653 mins.

summaryModelTuscanAlt <- summary(modelTuscanAlt) # 49 mins.

modelTuscanMorph <-
  bam(NormalizedVariantUnequalToStd_inclMorphVariation ~
          te(Longitude,Latitude,ConceptFreq.log.z,SpeakerBirthYear.z,d=c(2,1,1)) +
          CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
          SpeakerJob_Teacher_Freelance + SpeakerIsMale +
          s(Speaker,bs="re") + s(Location,bs="re") + s(Concept,bs="re") +
          s(Concept,CommunityRecordingYear.z,bs="re") +
          s(Concept,CommunitySize.log.z,bs="re") +
          s(Concept,CommunityAvgIncome.log.z,bs="re") +
          s(Concept,CommunityAvgAge.log.z,bs="re") +
          s(Concept,SpeakerJob_Farmer,bs="re") +
          s(Concept,SpeakerJob_Executive_AuxiliaryWorker,bs="re") +
          s(Concept,SpeakerEduLevel.log.z,bs="re") +
          s(Concept,SpeakerIsMale,bs="re"), data=tuscan,
          family="binomial", method="fREML",
          gc.level=2, cluster=cl
      ) # 802 mins.

summaryModelTuscanMorph <- summary(modelTuscanMorph) # 49 mins.
```

```r
# saved as the calculations take a long time
save(modelTuscanAlt,file='results/modelTuscanAlt.rda')
save(modelTuscanMorph,file='results/modelTuscanMorph.rda')
save(summaryModelTuscanAlt,file='results/summaryModelTuscanAlt.rda')
save(summaryModelTuscanMorph,file='results/summaryModelTuscanMorph.rda')
```

```r
load('results/summaryModelTuscanAlt.rda')
load('results/summaryModelTuscanAlt.rda')

# SpeakerJob_Teacher_Freelance is not significant for this dependent variable
summaryModelTuscanAlt
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## NormalizedVariantUnequalToStd_noMorphVariation ~ te(Longitude,
##     Latitude, ConceptFreq.log.z, SpeakerBirthYear.z, d = c(2,
##         1, 1)) + CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
##     SpeakerJob_Teacher_Freelance + SpeakerIsMale + s(Speaker,
##     bs = "re") + s(Location, bs = "re") + s(Concept, bs = "re") +
##     s(Concept, CommunityRecordingYear.z, bs = "re") + s(Concept,
##     CommunitySize.log.z, bs = "re") + s(Concept, CommunityAvgIncome.log.z,
##     bs = "re") + s(Concept, CommunityAvgAge.log.z, bs = "re") +
##     s(Concept, SpeakerJob_Farmer, bs = "re") + s(Concept, SpeakerJob_Executive_Au...
##     bs = "re") + s(Concept, SpeakerEduLevel.log.z, bs = "re") +
##     s(Concept, SpeakerIsMale, bs = "re")
##
## Parametric coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -0.4162     0.1266   -3.29   0.0010
## CommunitySize.log.z           -0.0580     0.0224   -2.59   0.0096
## SpeakerJob_Farmer              0.0458     0.0169    2.71   0.0067
## SpeakerEduLevel.log.z         -0.0641     0.0129   -4.97  6.6e-07
## SpeakerJob_Teacher_Freelance  -0.0354     0.0229   -1.54   0.1227
## SpeakerIsMale                  0.0366     0.0128    2.85   0.0044
##
## Approximate significance of smooth terms:
##                                                    edf Ref.df Chi.sq p-value
## te(Long.,Lat.,ConceptFreq.log.z,SpeakerBirthYear.z) 227.5    273   3272 < 2e-16
## s(Speaker)                                          102.0   2004    112 0.00669
## s(Location)                                         174.2    209   5395 < 2e-16
## s(Concept)                                          167.1    168 436226 < 2e-16
## s(Concept,CommunityRecordingYear.z)                 158.8    170 155723 < 2e-16
## s(Concept,CommunitySize.log.z)                      149.9    169  29820 < 2e-16
## s(Concept,CommunityAvgIncome.log.z)                 158.0    170 142478 < 2e-16
## s(Concept,CommunityAvgAge.log.z)                    154.4    170 110127 < 2e-16
## s(Concept,SpeakerJob_Farmer)                         86.1    169  26135 1.5e-07
## s(Concept,SpeakerJob_Executive_AuxiliaryWorker)      53.0    170   3256 0.00061
```

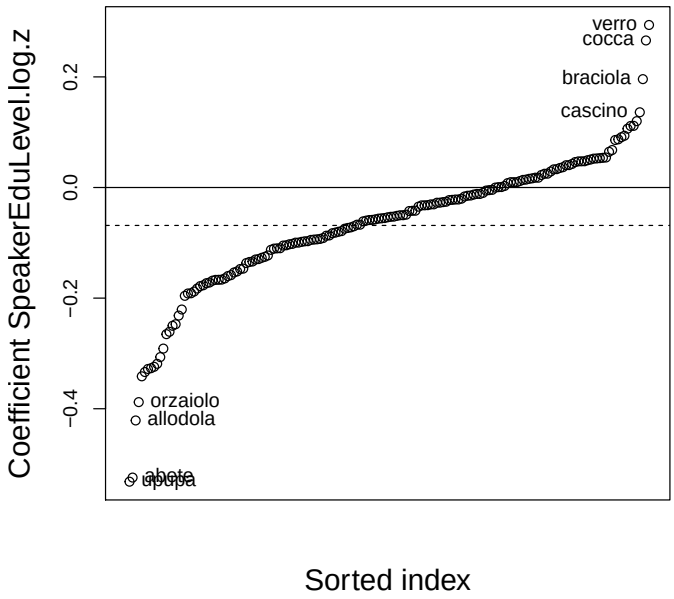
```
## s(Concept,SpeakerEduLevel.log.z)                    139.0    169   9187 < 2e-16
## s(Concept,SpeakerIsMale)                             85.2    169 110342 6.4e-11
##
## R-sq.(adj) =  0.301   Deviance explained = 25.6%
## fREML score = 5.3589e+05  Scale est. = 1         n = 379496
```

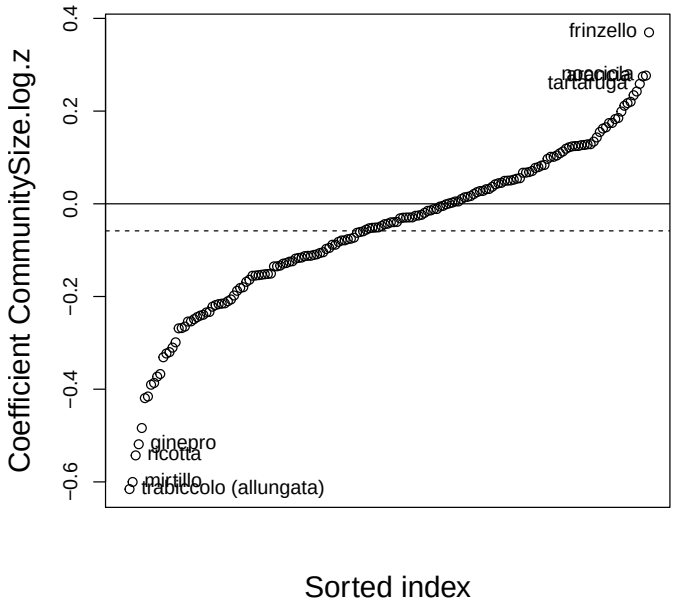
```
# SpeakerJob_Teacher_Freelance is significant for this dependent variable
summaryModelTuscanMorph
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## NormalizedVariantUnequalToStd_inclMorphVariation ~ te(Longitude,
##     Latitude, ConceptFreq.log.z, SpeakerBirthYear.z, d = c(2,
##         1, 1)) + CommunitySize.log.z + SpeakerJob_Farmer + SpeakerEduLevel.log.z +
##     SpeakerJob_Teacher_Freelance + SpeakerIsMale + s(Speaker,
##     bs = "re") + s(Location, bs = "re") + s(Concept, bs = "re") +
##     s(Concept, CommunityRecordingYear.z, bs = "re") + s(Concept,
##     CommunitySize.log.z, bs = "re") + s(Concept, CommunityAvgIncome.log.z,
##     bs = "re") + s(Concept, CommunityAvgAge.log.z, bs = "re") +
##     s(Concept, SpeakerJob_Farmer, bs = "re") + s(Concept, SpeakerJob_Executive_Au...
##     bs = "re") + s(Concept, SpeakerEduLevel.log.z, bs = "re") +
##     s(Concept, SpeakerIsMale, bs = "re")
##
## Parametric coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    0.4237     0.1211    3.50  0.00047
## CommunitySize.log.z           -0.0692     0.0261   -2.65  0.00795
## SpeakerJob_Farmer              0.0683     0.0183    3.73  0.00019
## SpeakerEduLevel.log.z         -0.0909     0.0140   -6.47  9.7e-11
## SpeakerJob_Teacher_Freelance  -0.0605     0.0274   -2.21  0.02719
## SpeakerIsMale                  0.0493     0.0146    3.38  0.00071
##
## Approximate significance of smooth terms:
##                                                    edf Ref.df Chi.sq p-value
## te(Long.,Lat.,ConceptFreq.log.z,SpeakerBirthYear.z) 209.6    248   2677 < 2e-16
## s(Speaker)                                          628.1   2004   1265 < 2e-16
## s(Location)                                         177.5    209  12901 < 2e-16
## s(Concept)                                          167.0    168 354198 < 2e-16
## s(Concept,CommunityRecordingYear.z)                 158.5    170 149871 < 2e-16
## s(Concept,CommunitySize.log.z)                      150.1    169  28885 < 2e-16
## s(Concept,CommunityAvgIncome.log.z)                 155.6    170  82640 < 2e-16
## s(Concept,CommunityAvgAge.log.z)                    152.4    170  79026 < 2e-16
## s(Concept,SpeakerJob_Farmer)                         78.3    169  18208 3.2e-05
## s(Concept,SpeakerJob_Executive_AuxiliaryWorker)      44.9    170   2071  0.0084
## s(Concept,SpeakerEduLevel.log.z)                    141.8    169  11938 < 2e-16
## s(Concept,SpeakerIsMale)                             91.6    169 125887 1.7e-12
##
## R-sq.(adj) =  0.279   Deviance explained = 23.7%
## fREML score = 5.3671e+05  Scale est. = 1         n = 379496
```

By-word random slopes

```
par(mfrow=c(2,1))

plotSlope.gam(modelTuscan,tuscan,"Concept","SpeakerEduLevel.log.z")
plotSlope.gam(modelTuscan,tuscan,"Concept","CommunitySize.log.z")
```





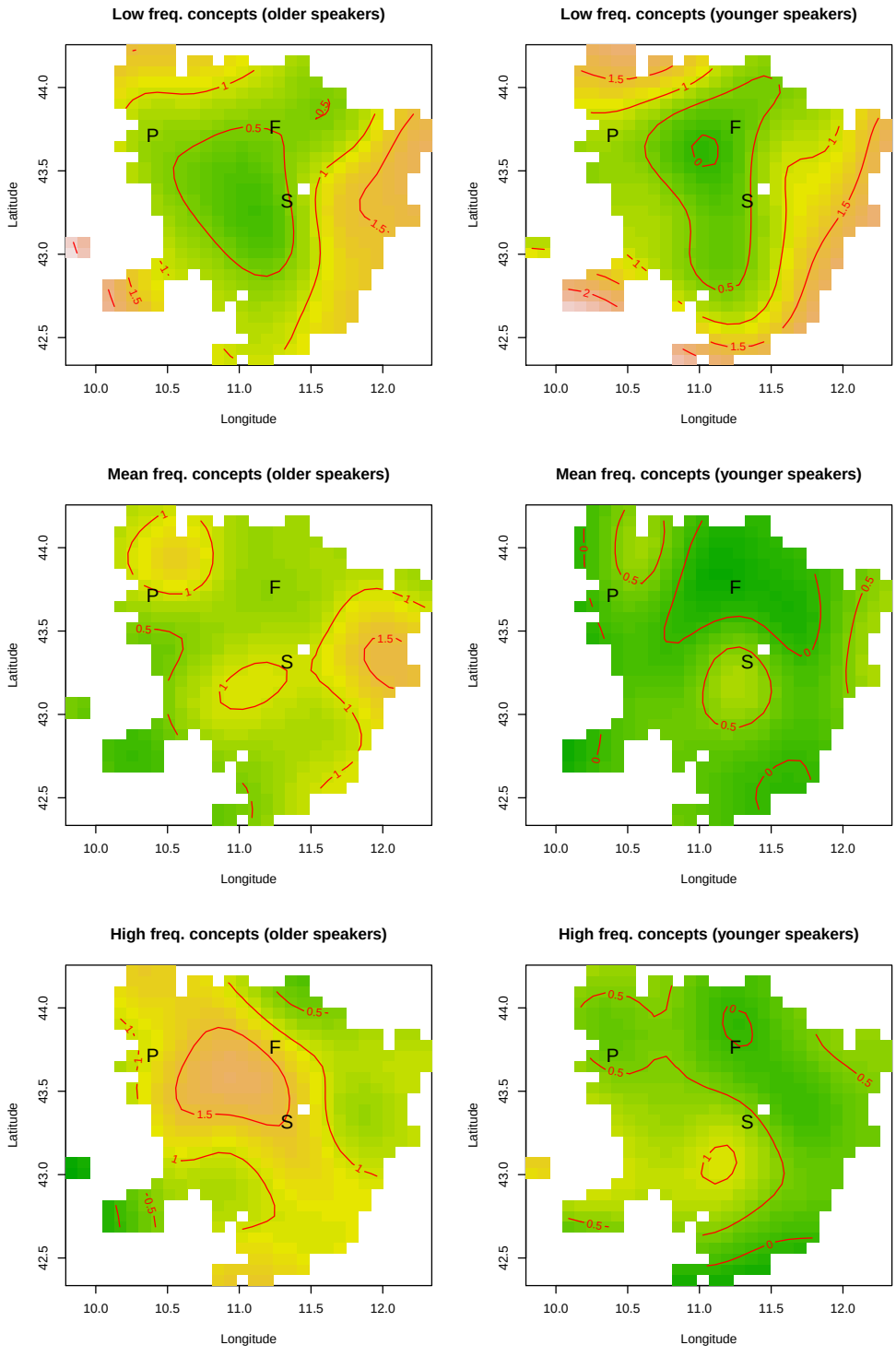
Year of birth of older and younger speakers used in the plots below:

```
# mean year of birth
round(mean(tuscan$SpeakerBirthYear))

## [1] 1931

# older speakers
mean(tuscan$SpeakerBirthYear) - 2*sd(tuscan$SpeakerBirthYear)

## [1] 1888

# younger speakers
mean(tuscan$SpeakerBirthYear) + 2*sd(tuscan$SpeakerBirthYear)

## [1] 1974
```

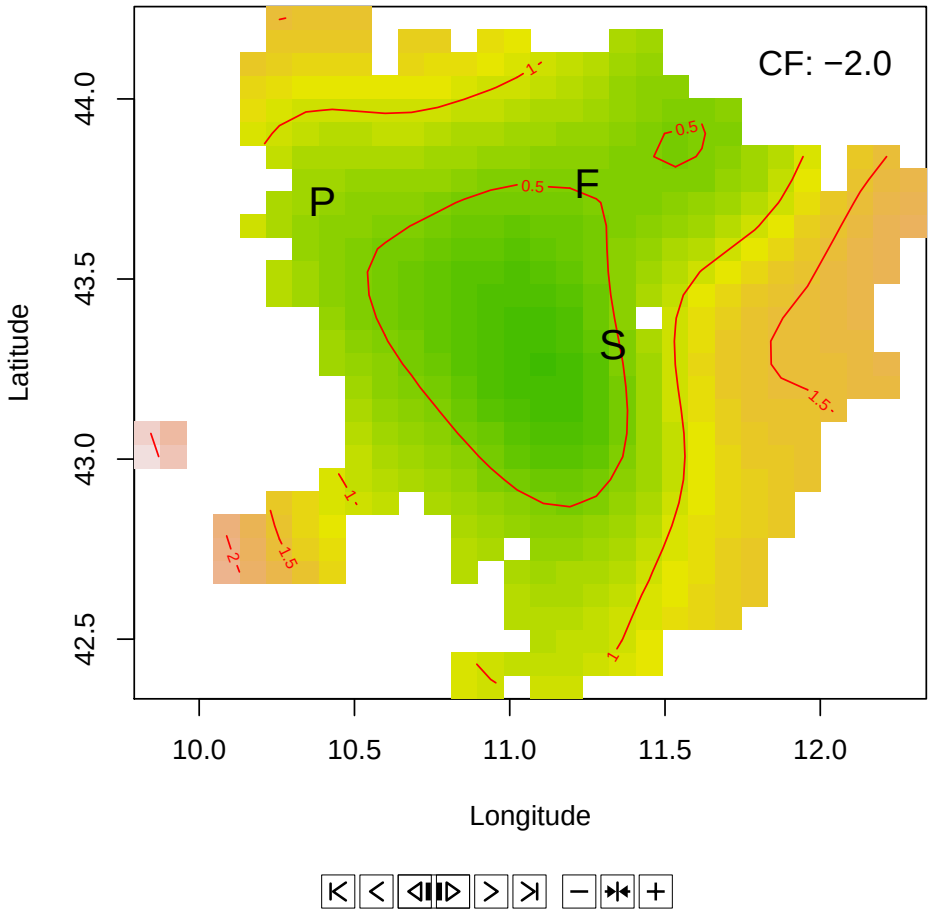
Visualization of the geographical patterns

```
fixedVals = list(CommunitySize.log.z=0, SpeakerJob_Farmer=0.5,
                 SpeakerJob_Executive_AuxiliaryWorker=0.5,
                 SpeakerEduLevel.log.z=0, SpeakerIsMale=0.5)

par(mfrow=c(3,2))
vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=-2,SpeakerBirthYear.z=-2),fixedVals),
        main="Low freq. concepts (older speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label

vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=-2,SpeakerBirthYear.z=+2),fixedVals),
        main="Low freq. concepts (younger speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label

vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=0,SpeakerBirthYear.z=-2),fixedVals),
        main="Mean freq. concepts (older speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label
```

```
vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=0,SpeakerBirthYear.z=+2),fixedVals),
        main="Mean freq. concepts (younger speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label

vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=2,SpeakerBirthYear.z=-2),fixedVals),
        main="High freq. concepts (older speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label

vis.gam(modelTuscan,view=c("Longitude","Latitude"), plot.type="contour",
        color="terrain", too.far=0.045,
        cond=c(list(ConceptFreq.log.z=2,SpeakerBirthYear.z=+2),fixedVals),
        main="High freq. concepts (younger speakers)", zlim=c(-0.5,2.8))

text(10.397, 43.716, "P", cex=1.5) # Pisa label
text(11.333, 43.320, "S", cex=1.5) # Siena label
text(11.250, 43.767, "F", cex=1.5) # Florence label
```


Low freq. concepts (older speakers)
Low freq. concepts (younger speakers)
Mean freq. concepts (older speakers)
Mean freq. concepts (younger speakers)
High freq. concepts (older speakers)
High freq. concepts (younger speakers)
P
F
S
Latitude
Longitude
42.5
43.0
43.5
44.0
10.0
10.5
11.0
11.5
12.0

Animated geographical pattern for increasing concept frequency (older speakers)



Animated geographical pattern for increasing concept frequency (younger speakers)