Cultural distances: Wikipedia data

Contents

Introduction

We compare cultural distances between socieites with linguistic similarities between societies, controlling for shared history in two ways.

The first test uses mixed effects modelling. The pairing of the language family of each language (according to Glottolog) is used as a random effect. That means that the model can capture the likelihood that two languages from the Indo-European language family will be more similar to each other than two languages from different language families. The same is done with geographic area according to Autotyp.

The second test controls for history using distances from a phylogenetic tree. The tree comes from Bouckaert et al. (2012). Patristic distances between languages are used as a measure of historical distance between societies in a Mantel test. Note that the Mantel test assumes a strict distance metric, which is not necessarily the case with this data, but there are few other ways to deal with continuous pairwise distances.

Load libraries

```
library(ape)
library(ecodist)
library(lme4)
library(sjPlot)
library(ggplot2)
library(igraph)
library(lattice)
```

Parameters (using data from Northuralex and Wikipedia, k=100, unfiltered):

```
datasetName = "wikipedia-main"
lingDistancesFile = "../data/FAIR/nel-wiki-k100-alignments-by-language-pair.csv"
lingDistancesFileNK = "../data/FAIR/nel-wiki-k100-alignments-by-language-pair-without-kinship.csv"
lingDistancesByDomainFile = "../results/EA_distances/nel-wiki-k100_with_ling.csv"
# (generated by ../processing/combineCultAndLingDistances.R)
```

All domains

Load data

Read the cultural distances:

```
cult = read.csv("../results/EA_distances/CulturalDistances_Long.csv", stringsAsFactors = F)
names(cult) = c("l1","l2","cult.dist")
cultLangs = unique(c(cult$Var1,cult$Var2))
```

Add language family:

```
l = read.csv("../data/FAIR_langauges_glotto_xdid.csv", stringsAsFactors = F)
g = read.csv("../data/glottolog-languoid.csv/languoid.csv", stringsAsFactors = F)
l$family = g[match(l$glotto,g$id),]$family_pk
l$family = g[match(l$family,g$pk),]$name
```

Read the semantic distances

```
ling = read.csv(lingDistancesFile, stringsAsFactors = F)
```

There are very few possible comparisons for Slovenian and Northern Sami, so we'll remove these:

```
ling = ling[!(ling$l1=="se" || ling$l2 == "se"),]
ling = ling[!(ling$l1=="sl" || ling$l2 == "sl"),]
```

Combine the lingusitic and cultural distances. Note that we flip the cultural measure from a distance measure to a similarity measure.

```
cult$l1.iso2 = l[match(cult$l1,l$Language2),]$iso2
cult$l2.iso2 = l[match(cult$l2,l$Language2),]$iso2

fairisos = unique(c(ling$l1,ling$l2))
cultisos = unique(c(cult$l1.iso2, cult$l2.iso2))

cult = cult[(cult$l1.iso2 %in% fairisos) & (cult$l2.iso2 %in% fairisos),]
ling = ling[(ling$l1 %in% cultisos) & (ling$l2 %in% cultisos),]

matches = sapply(1:nrow(ling), function(i){
  which(cult$l1.iso2==ling$l1[i] & cult$l2.iso2==ling$l2[i])
})

ling$cult.dist = cult[matches,]$cult.dist
# Flip
ling$cult.dist = 1 - ling$cult.dist
# Scale
ling$cult.dist.center = scale(ling$cult.dist)
cdc.s = attr(ling$cult.dist.center,"scaled:scale")
cdc.c = attr(ling$cult.dist.center,"scaled:center")
ling$cult.dist.center = as.numeric(ling$cult.dist.center)
ling$comparison_count.center =
  scale(ling$comparison_count)

ling$family1 = l[match(ling$l1, l$iso2),]$family
ling$family2 = l[match(ling$l2, l$iso2),]$family
ling$area1 = l[match(ling$l1, l$iso2),]$autotyp.area
ling$area2 = l[match(ling$l2, l$iso2),]$autotyp.area
```

```r
fgroup = cbind(ling$family1,ling$family2)
fgroup = apply(fgroup,1,sort)
ling$family.group = apply(fgroup,2,paste,collapse=":")
agroup = cbind(ling$area1,ling$area2)
agroup = apply(agroup,1,sort)
ling$area.group = apply(agroup,2,paste,collapse=":")

ling$rho.center = scale(ling$local_alignment)
```

Each observation is now assocaited with a language family pair:

```r
head(ling[,c("l1","l2","local_alignment",'family.group')])
```

```
##      l1  l2 local_alignment                    family.group
## 7    ja  ab      0.01930414          Abkhaz-Adyge:Japonic
## 8    ab  zh      0.02225169  Abkhaz-Adyge:Sino-Tibetan
## 10   cv xal      0.02765860             Mongolic:Turkic
## 11  xal  ja      0.02832668             Japonic:Mongolic
## 12  xal  zh      0.02895876        Mongolic:Sino-Tibetan
## 14   bn  ab      0.03192066 Abkhaz-Adyge:Indo-European
```

And the same is true for area:

```r
tail(ling[,c("l1","l2","local_alignment",'area.group')])
```

```
##        l1 l2 local_alignment              area.group
## 2522 fr es       0.3936442            Europe:Europe
## 2524 cs uk       0.4023323        Europe:Inner Asia
## 2528 cs ru       0.4082099        Europe:Inner Asia
## 2529 be ru       0.4129814 Inner Asia:Inner Asia
## 2532 uk be       0.4276664 Inner Asia:Inner Asia
## 2535 uk ru       0.5079911 Inner Asia:Inner Asia
```

Number of observations:

```r
# Number of datapoints:
nrow(ling)
```

```
## [1] 733
```

```r
# Number of unique languages:
length(unique(unlist(ling[,c("l1","l2")])))
```

```
## [1] 40
```

```r
# Number of unique langauge families:
uniqueFamilies = unique(unlist(ling[,c("family1","family2")]))
length(uniqueFamilies)
```

```
## [1] 10
```

```r
# Number of unique areas:
uniqueAreas = unique(unlist(ling[,c("area1","area2")]))
length(uniqueAreas)
```

```
## [1] 6
```

Cross-over between language famlies and areas:

```r
tx = data.frame(lang= c(ling$l1,ling$l2),
        fam = c(ling$family1,ling$family2),
        area= c(ling$area1,ling$area2))
tx = tx[!duplicated(tx),]
table(tx$fam,tx$area)
```

```
## 
##               Europe Greater Mesopotamia Indic Inner Asia N Coast Asia
##   Abkhaz-Adyge      0                   1     0          0           0
##   Afro-Asiatic      0                   1     0          0           0
##   Dravidian         0                   0     3          0           0
##   Indo-European    11                   2     1          5           0
##   Japonic           0                   0     0          0           1
##   Koreanic          0                   0     0          0           1
##   Mongolic          0                   0     0          1           0
##   Sino-Tibetan      0                   0     0          0           0
##   Turkic            0                   1     0          5           0
##   Uralic            1                   0     0          5           0
## 
##               Southeast Asia
##   Abkhaz-Adyge              0
##   Afro-Asiatic             0
##   Dravidian                0
##   Indo-European            0
##   Japonic                  0
##   Koreanic                 0
##   Mongolic                 0
##   Sino-Tibetan             1
##   Turkic                   0
##   Uralic                   0
```

LMER models

Mixed effects model, predicting Linguistic similar{ys from cultural similarity, with random intercept for family and area and random slope for cultural similarity for family and area.

We start with a null model with random intercepts for family and area, and random slopes for cultural similarity by both. We add a fixed effect of the number of comparisons made for each datapoint (number of concepts that were available to compare). Then we add a fixed effect of cultural similarity

```r
m0 = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = ling
)
m0.5 = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = ling
)
m1 = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    cult.dist.center +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = ling
)
anova(m0,m0.5,m1)
```

```
## refitting model(s) with ML (instead of REML)

## Data: ling
## Models:
## m0: rho.center ~ 1 + (1 + cult.dist.center | family.group) + (1 +
## m0:     cult.dist.center | area.group)
## m0.5: rho.center ~ 1 + comparison_count.center + (1 + cult.dist.center |
## m0.5:     family.group) + (1 + cult.dist.center | area.group)
## m1: rho.center ~ 1 + comparison_count.center + cult.dist.center +
## m1:     (1 + cult.dist.center | family.group) + (1 + cult.dist.center |
## m1:     area.group)
##        Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## m0      8 1658.0 1694.7 -820.97   1642.0
## m0.5    9 1298.0 1339.3 -639.98   1280.0 361.979      1  < 2.2e-16 ***
## m1     10 1285.3 1331.3 -632.65   1265.3  14.675      1  0.0001277 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cultural similarity is significantly correlated with Linguistic similarity. Here are the model estimates:

```r
summary(m1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rho.center ~ 1 + comparison_count.center + cult.dist.center +
##     (1 + cult.dist.center | family.group) + (1 + cult.dist.center |
```

```
##     area.group)
##    Data: ling
##
## REML criterion at convergence: 1278.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6298 -0.6127  0.0942  0.6554  4.7348
##
## Random effects:
##  Groups       Name            Variance  Std.Dev. Corr
##  family.group (Intercept)     0.1688145 0.41087
##               cult.dist.center 0.0004627 0.02151  1.00
##  area.group   (Intercept)     0.0464131 0.21544
##               cult.dist.center 0.0029874 0.05466  -1.00
##  Residual                     0.2900042 0.53852
## Number of obs: 733, groups:  family.group, 48; area.group, 20
##
## Fixed effects:
##                          Estimate Std. Error t value
## (Intercept)             -0.39344    0.09035  -4.354
## comparison_count.center  0.60900    0.02697  22.580
## cult.dist.center         0.18268    0.03233   5.651
##
## Correlation of Fixed Effects:
##            (Intr) cmpr_.
## cmprsn_cnt.  0.093
## clt.dst.cnt -0.120 -0.202
```

Plot the estimates, rescaling the variables back to the original units:

```
gx = sjp.lmer(m1,'pred','cult.dist.center',
              prnt.plot = F)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
gx$plot$data$y = gx$plot$data$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
gx$plot$data$resp.y = gx$plot$data$resp.y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
gx$plot$data$x = gx$plot$data$x *
  cdc.s +cdc.c
gx = gx$plot + coord_cartesian(ylim=c(-0.05,0.5),
                               xlim=c(0.15,0.85)) +
  xlab("Cultural similarity") +
  ylab("Semantic alignment") +
  ggtitle("")
gx
```
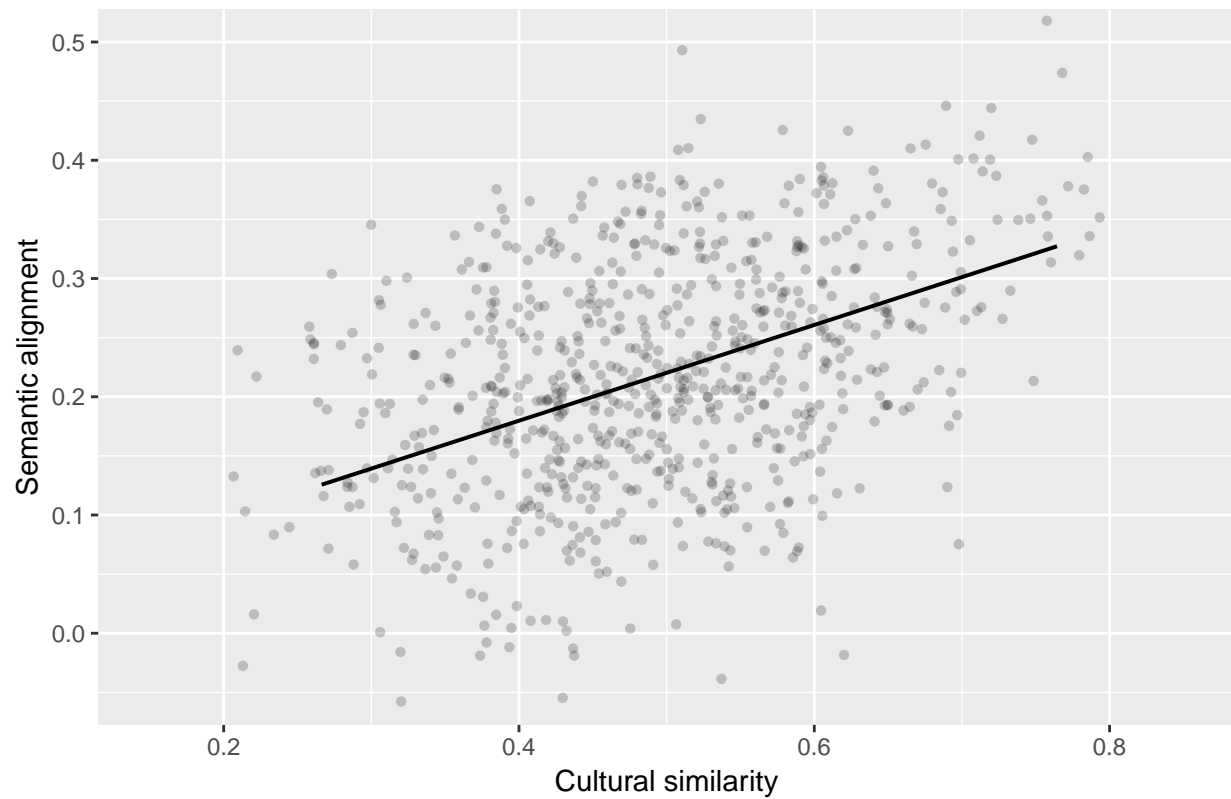
```
pdf(paste0("../results/stats/",datasetName,"/CulturalDistance_Rho_Graph.pdf"),
    height=2.5, width=2.5)
gx
dev.off()
```
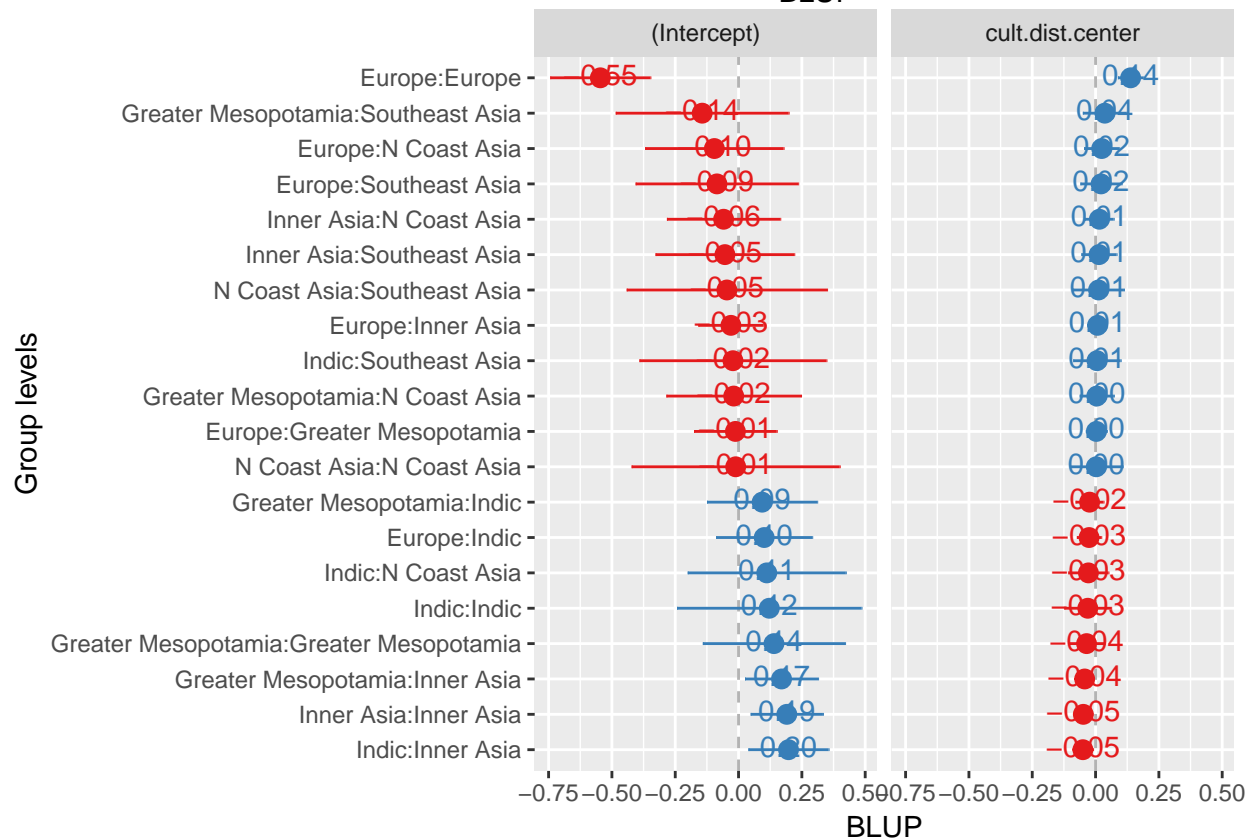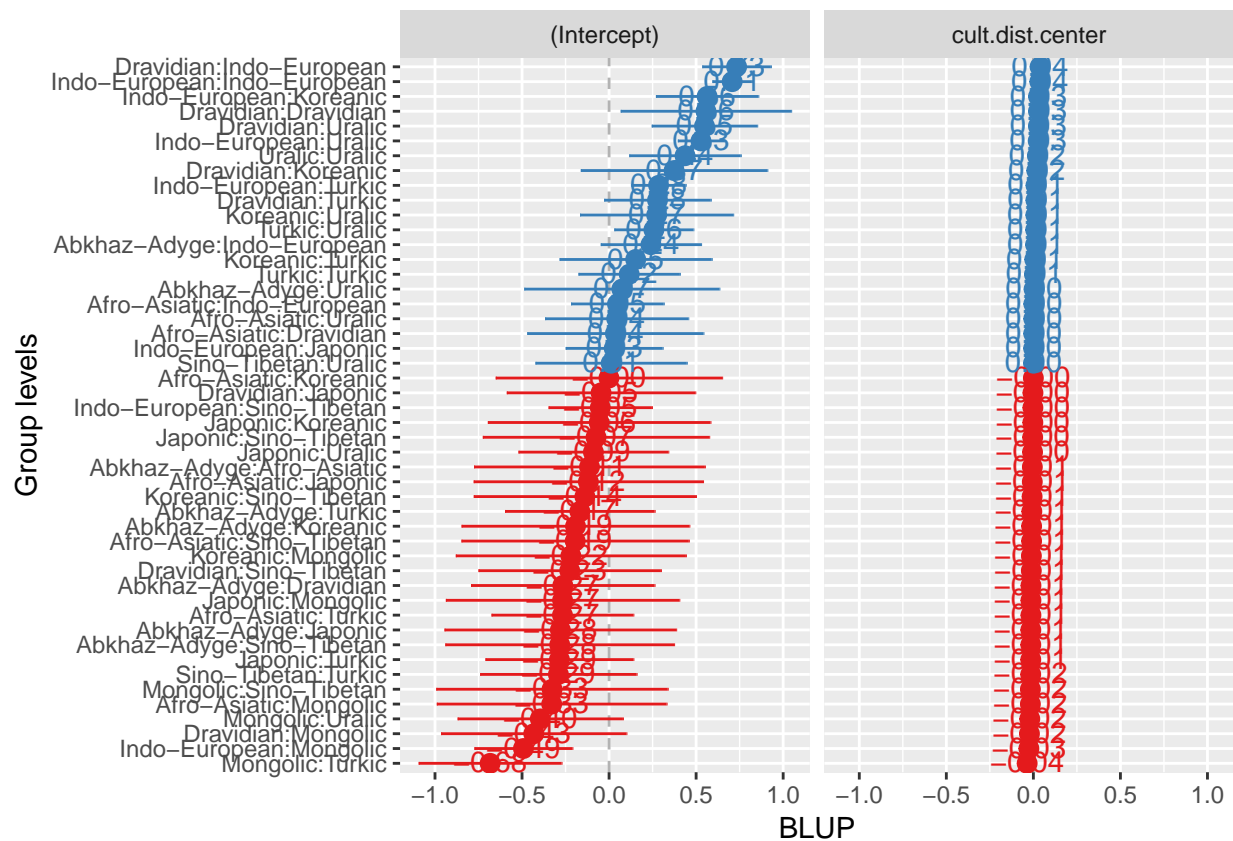
```
## pdf
##   2
```

Plot the random effects:

```
sjp.lmer(m1,'re', sort.est = "cult.dist.center")
```

```
## Plotting random effects...
## Plotting random effects...
```
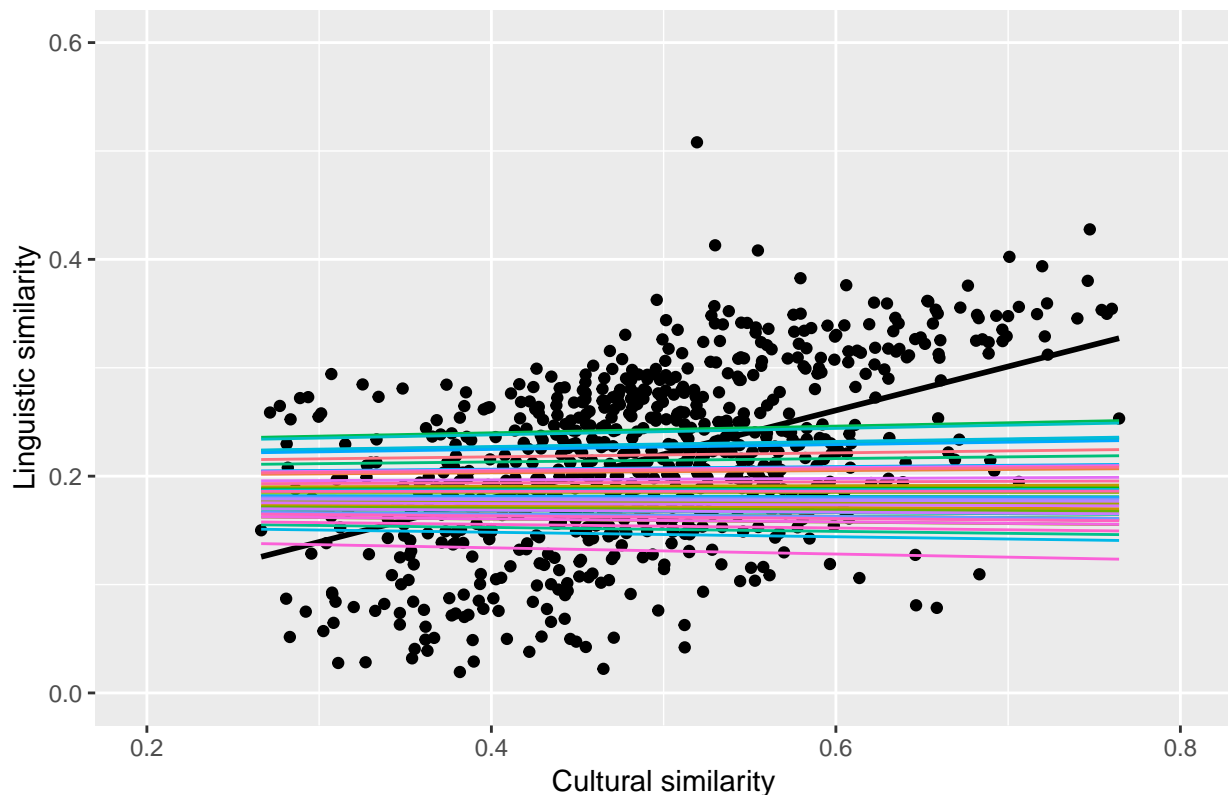
8

```
px = sjp.lmer(m1,'rs.ri', prnt.plot = F)
dx = px$plot[[1]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")

ggplot(dx,aes(x,y)) +
  geom_point(data=ling,
   mapping=aes(x=as.numeric(cult.dist),
               y=as.numeric(local_alignment))) +
  stat_smooth(data=gx$data,method="lm",colour="black",
              se=F)+
  geom_line(mapping = aes(colour=grp)) +
  xlab("Cultural similarity")+
  ylab("Linguistic similarity") +
  ggtitle("Language family pair random effects") +
  coord_cartesian(ylim=c(0.0,0.6),
                  xlim=c(0.2,0.8)) +
  theme(legend.position = "none")
```


Language family pair random effects

```
dx = px$plot[[2]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center,"scaled:scale") +
  attr(ling$rho.center,"scaled:center")
```

```
ggplot(dx,aes(x,y)) +
  geom_point(data=ling,
             mapping=aes(x=as.numeric(cult.dist),
                         y=as.numeric(local_alignment))) +
  stat_smooth(data=gx$data,method="lm",colour="black",
              se=F)+
  geom_line(mapping = aes(colour=grp)) +
  xlab("Cultural similarity")+
  ylab("Linguistic similarity") +
  ggtitle("Area pair random effects") +
  coord_cartesian(ylim=c(0.1,0.6),
                  xlim=c(0.2,0.8))
```
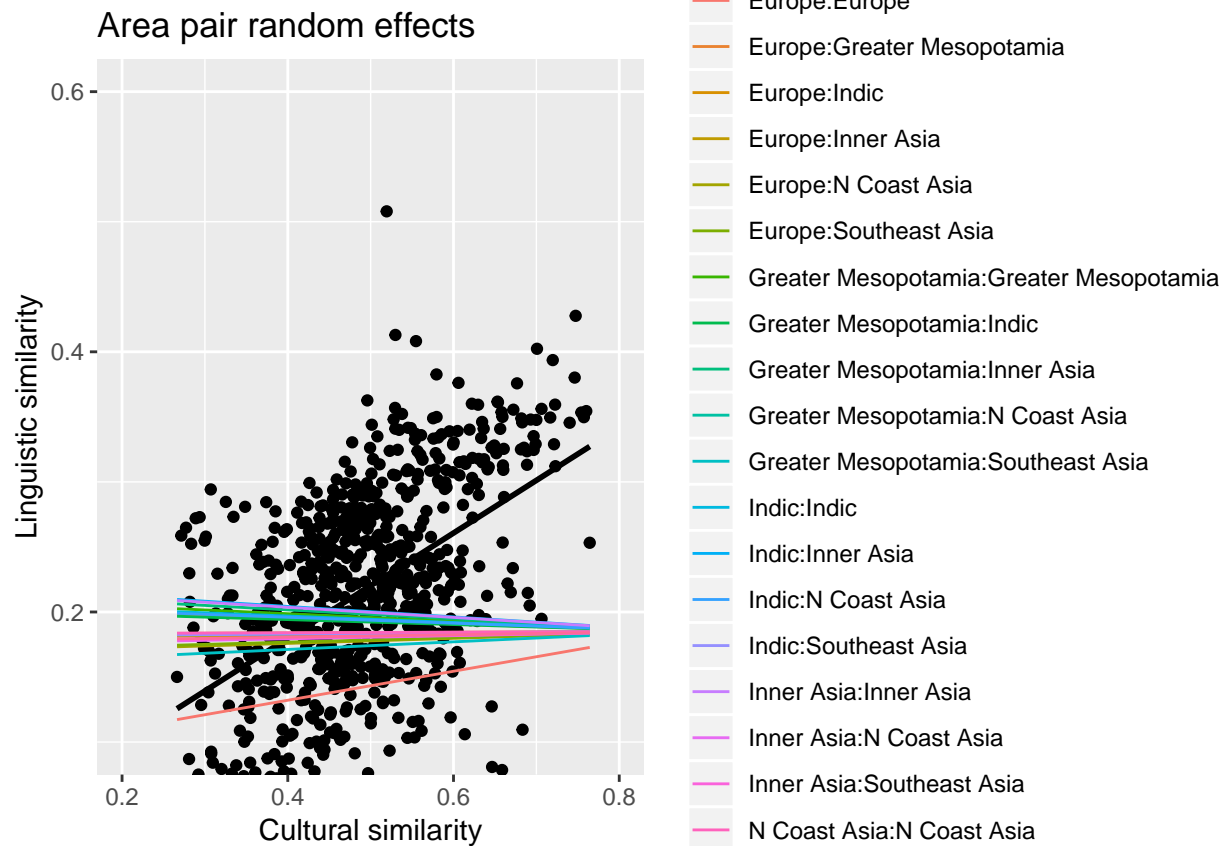
Without Kinship data

The analyses below show that the strongest relationship is with Kinship. Here we run the analysis as above, but using semantic distances computed without concepts that relate to kinship. Note that the local alignment values correlate with r > 0.99.

Code for constructing the data is hidden, but it is the same as above and available in the Rmd file:

Run the lmer models:

```
mONK = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = lingNK
)
m0.5NK = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = lingNK
)
m1NK = lmer(
  rho.center ~ 1 +
    comparison_count.center +
    cult.dist.center +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = lingNK
)
anova(mONK,m0.5NK,m1NK)
```

```
## refitting model(s) with ML (instead of REML)

## Data: lingNK
## Models:
## mONK: rho.center ~ 1 + (1 + cult.dist.center | family.group) + (1 +
## mONK:     cult.dist.center | area.group)
## m0.5NK: rho.center ~ 1 + comparison_count.center + (1 + cult.dist.center |
## m0.5NK:     family.group) + (1 + cult.dist.center | area.group)
## m1NK: rho.center ~ 1 + comparison_count.center + cult.dist.center +
## m1NK:     (1 + cult.dist.center | family.group) + (1 + cult.dist.center |
## m1NK:     area.group)
##          Df    AIC    BIC  logLik deviance   Chisq Chi Df Pr(>Chisq)
## mONK      8 1658.0 1694.7 -820.97   1642.0
## m0.5NK    9 1298.0 1339.3 -639.98   1280.0 361.979      1  < 2.2e-16 ***
## m1NK     10 1285.3 1331.3 -632.65   1265.3  14.675      1  0.0001277 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1NK)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rho.center ~ 1 + comparison_count.center + cult.dist.center +
##     (1 + cult.dist.center | family.group) + (1 + cult.dist.center |
##     area.group)
```

```
##     Data: lingNK
##
## REML criterion at convergence: 1278.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6298 -0.6127  0.0942  0.6554  4.7348
##
## Random effects:
##  Groups       Name             Variance  Std.Dev. Corr
##  family.group (Intercept)      0.1688145 0.41087
##               cult.dist.center 0.0004627 0.02151   1.00
##  area.group   (Intercept)      0.0464131 0.21544
##               cult.dist.center 0.0029874 0.05466  -1.00
##  Residual                      0.2900042 0.53852
## Number of obs: 733, groups:  family.group, 48; area.group, 20
##
## Fixed effects:
##                         Estimate Std. Error t value
## (Intercept)             -0.39344    0.09035  -4.354
## comparison_count.center  0.60900    0.02697  22.580
## cult.dist.center         0.18268    0.03233   5.651
##
## Correlation of Fixed Effects:
##            (Intr) cmpr_.
## cmprsn_cnt.  0.093
## clt.dst.cnt -0.120 -0.202
```

MRM

Use multiple regression on distance matrices to do the same test as above. Here, the distance between two languages in terms of family is just whether they are part of the same family. Note that this does not take into account particular values for particular families, nor the rnadom slopes within families.

```r
# Use graph method to make distance matrix
grph <- graph.data.frame(ling[,c("l1",'l2','local_alignment')], directed=FALSE)
# add value as a weight attribute
ling.m = get.adjacency(grph, attr="local_alignment", sparse=FALSE)
rownames(ling.m) = l[match(rownames(ling.m),l$iso2),]$Language2
colnames(ling.m) = l[match(colnames(ling.m),l$iso2),]$Language2
# Same for comparison_count.center
grph <- graph.data.frame(ling[,c("l1",'l2','comparison_count')], directed=FALSE)
# add value as a weight attribute
cc.m = get.adjacency(grph, attr="comparison_count", sparse=FALSE)
rownames(cc.m) = l[match(rownames(cc.m),l$iso2),]$Language2
colnames(cc.m) = l[match(colnames(cc.m),l$iso2),]$Language2

cult.m = read.csv("../results/EA_distances/CulturalDistances.csv", stringsAsFactors = F)
rownames(cult.m) = cult.m[,1]
cult.m = cult.m[,2:ncol(cult.m)]
cult.m = as.matrix(cult.m)
# Flip cultural value to distance
cult.m = 1-cult.m
mx = match(rownames(ling.m),rownames(cult.m))
cult.m = cult.m[mx,mx]

family.matrix = l[match(rownames(ling.m),l$Language),]$family
family.matrix = outer(family.matrix,family.matrix,"!=") *1

# center and scale
ling.m =  matrix(scale(as.vector(ling.m)),nrow=nrow(ling.m))
cc.m =  matrix(scale(as.vector(cc.m)),nrow=nrow(cc.m))
cult.m = matrix(scale(as.vector(cult.m)),nrow=nrow(cult.m))

ecodist::MRM(as.dist(ling.m) ~
             as.dist(cult.m) + as.dist(family.matrix) +
             as.dist(cc.m),nperm = 10000)
```

```
## $coef
##                        as.dist(ling.m)    pval
## Int                          0.1694732  0.0794
## as.dist(cult.m)              0.2629547  0.0064
## as.dist(family.matrix)      -0.1627927  0.1913
## as.dist(cc.m)                0.6686839  0.0001
##
## $r.squared
##        R2       pval
## 0.6740298 0.0001000
##
## $F.test
##        F    F.pval
## 534.8618   0.0001
```

14

Mantel tests

Read the historical distances for Indo-European, based on the phylogenetic distances.

Data prep

Load historical distances:

```
hist = read.csv("../data/trees/IndoEuropean_historical_distances.csv", stringsAsFactors = F)
hist = hist[!duplicated(hist[,1]),!duplicated(hist[,1])]
rownames(hist) = hist[,1]
hist = hist[,2:ncol(hist)]
hist.m = as.matrix(hist)
colnames(hist.m) = rownames(hist.m)
hist.m = hist.m/max(hist.m)
```

Read the cultural distance as a matrix:

```
cult.m = read.csv("../results/EA_distances/CulturalDistances.csv", stringsAsFactors = F)
rownames(cult.m) = cult.m[,1]
cult.m = cult.m[,2:ncol(cult.m)]
```

Flip the cultural distance into a cultural similarity measure:

```
cult.m = 1-cult.m
```

Convert the linguistic similarities to a matrix. This uses `igraph` to make an undirected graph from the long format with `local_alignment` as the edge weights, then output a matrix of adjacencies.

```
grph <- graph.data.frame(ling[,c("l1",'l2','local_alignment')], directed=FALSE)
# add value as a weight attribute
ling.m = get.adjacency(grph, attr="local_alignment", sparse=FALSE)
rownames(ling.m) = l[match(rownames(ling.m),l$iso2),]$Language2
colnames(ling.m) = l[match(colnames(ling.m),l$iso2),]$Language2
```

Load the geographic distances:

```
geoDist = read.csv("../data/GeographicDistances.csv",stringsAsFactors = F)
geoDist.m = as.matrix(geoDist)
# Convert to log distance
geoDist.m = log(geoDist.m)
geoDist.m[is.infinite(geoDist.m)] = 0
rownames(geoDist.m) = colnames(geoDist.m)
```

Match the distance matrices

```
in.analysis = intersect(rownames(ling.m),rownames(cult.m))
in.analysis = intersect(in.analysis, rownames(hist.m))
cult.m2 = cult.m[in.analysis,in.analysis]
ling.m2 = ling.m[in.analysis,in.analysis]
hist.m2 = hist.m[in.analysis,in.analysis]
geo.m2 = geoDist.m[in.analysis,in.analysis]
```

Note that there are only 19 languages with data on lingusitic, cultural and historical distance.

```
plot(as.dist(cult.m2),as.dist(ling.m2),
     xlab="Cultural similarity",
     ylab="Linguistic similarity")
```
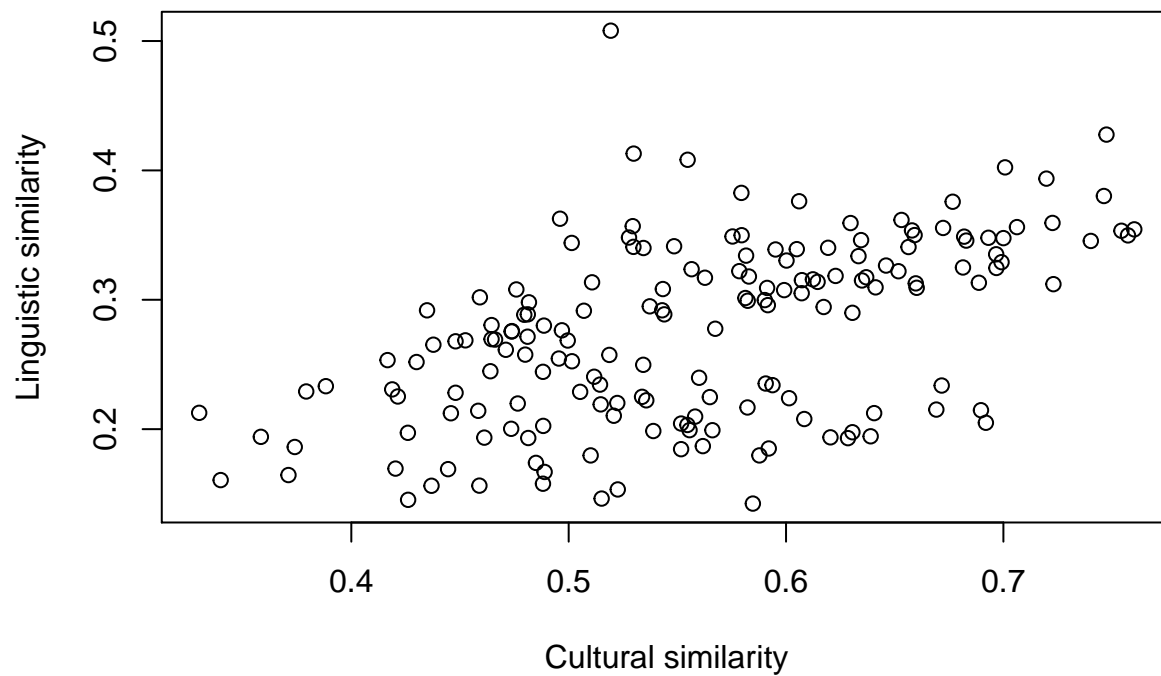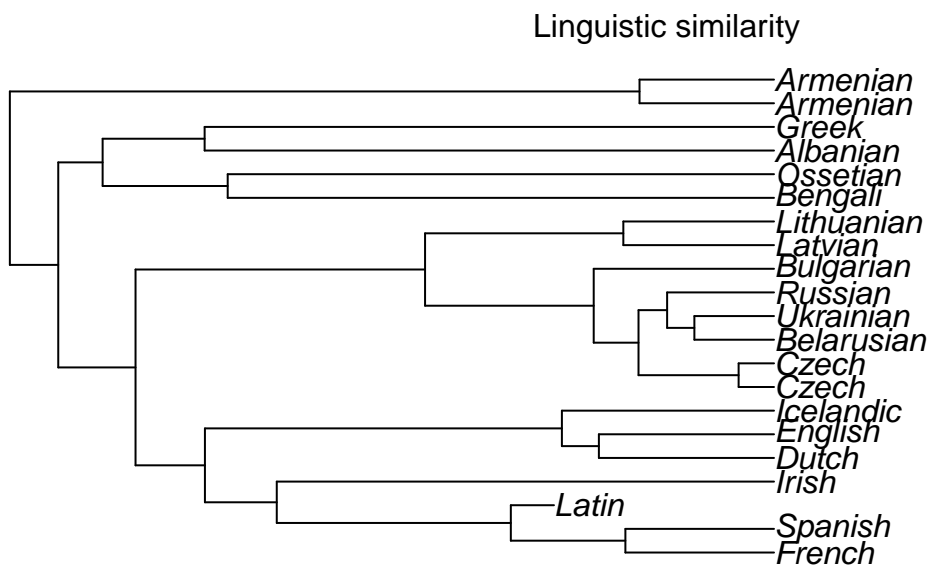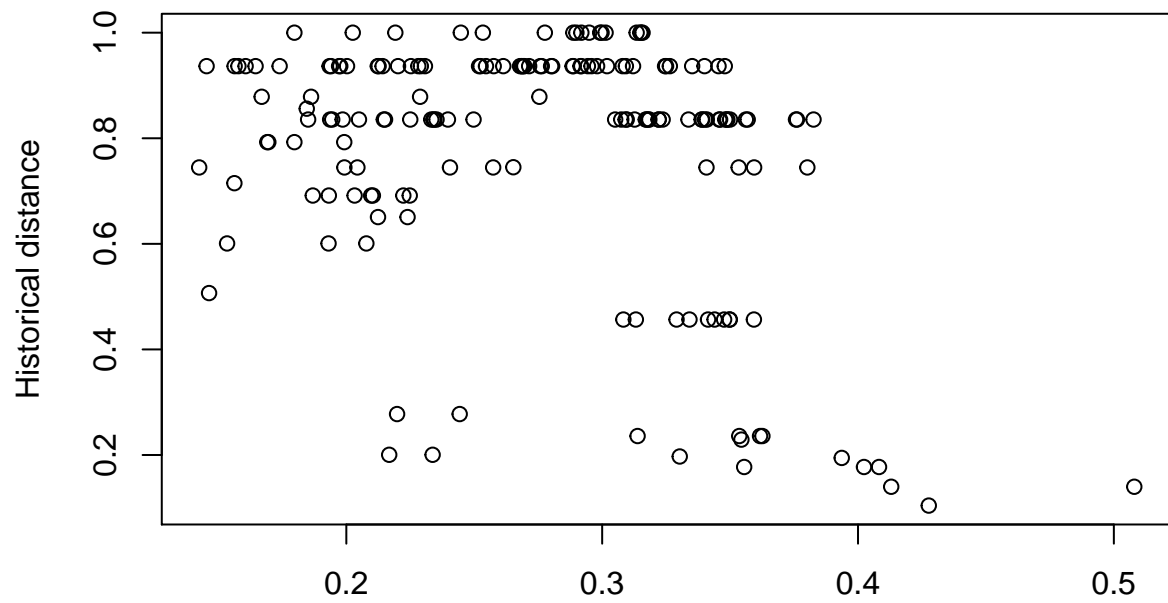
```
plot(as.dist(cult.m2),as.dist(hist.m2),
     xlab="Cultural similarity",
     ylab="Historical distance")
```
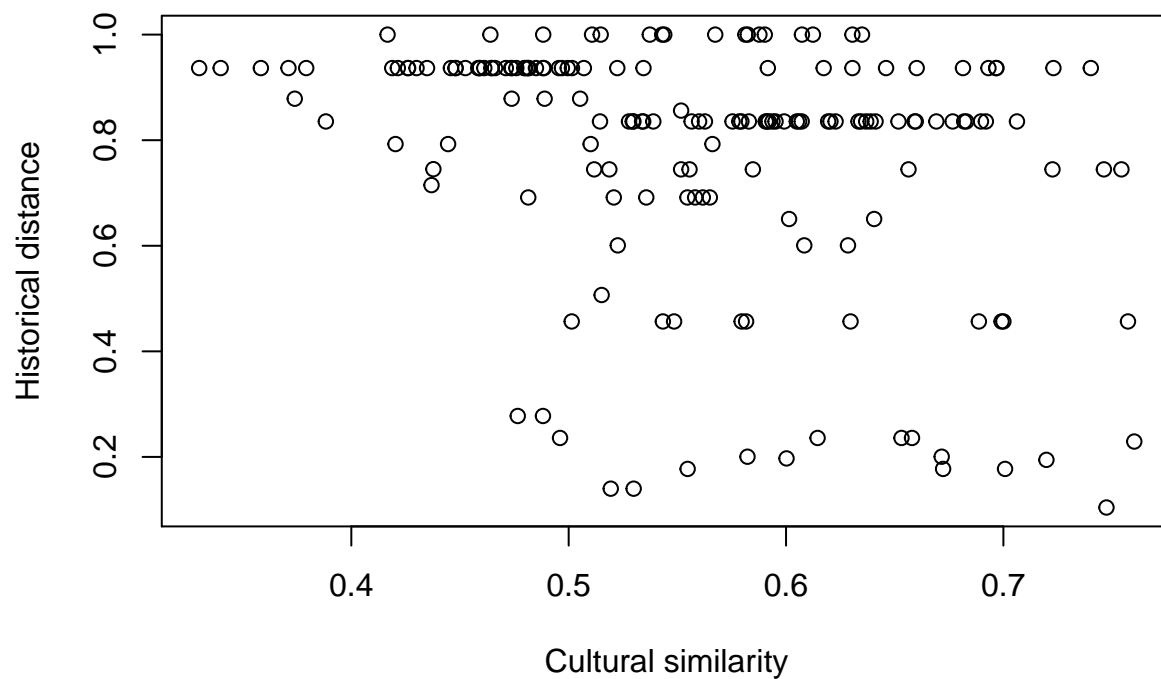


```
plot(as.dist(ling.m2),as.dist(hist.m2),
     xlab="Linguistic similarity",
     ylab="Historical distance")
```

Historical distance

Linguistic similarity

Armenian
Armenian
Greek
Albanian
Ossetian
Bengali
Lithuanian
Latvian
Bulgarian
Russian
Ukrainian
Belarusian
Czech
Czech
Icelandic
English
Dutch
Irish
Latin
Spanish
French

Tests

The results of the test list the following measures:

- mantelr: Mantel correlation coefficient.
- pval1: one-tailed p-value (null hypothesis: r <= 0).
- pval2: one-tailed p-value (null hypothesis: r >= 0).
- pval3: two-tailed p-value (null hypothesis: r = 0).
- llim: lower confidence limit for r.
- ulim: upper confidence limit for r.

```r
set.seed(1498)
```

```r
distms = list("Cultrual"= cult.m2,
              "Linguistic" = ling.m2,
              "Historical" = hist.m2,
              "Geographic" = geo.m2)
for(i in 1:3){
  for(j in (i+1):4){
    print(paste("Correlation between",
                names(distms)[i],"and",names(distms)[j]))
    print(ecodist::mantel(as.dist(distms[[i]]) ~
                as.dist(distms[[j]]),
                nperm = 100000))
  }
}
```

```
## [1] "Correlation between Cultrual and Linguistic"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##   0.5243289   0.0050000   0.9950100   0.0050300  0.3796035  0.6586819
## [1] "Correlation between Cultrual and Historical"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##  -0.3243830   0.9871000   0.0129100   0.0138900 -0.4402666 -0.2385575
## [1] "Correlation between Cultrual and Geographic"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##  -0.4495398   0.9967200   0.0032900   0.0032900 -0.5754918 -0.3109193
## [1] "Correlation between Linguistic and Historical"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##  -0.3372882   0.9859600   0.0140500   0.0167300 -0.5019408 -0.1639425
## [1] "Correlation between Linguistic and Geographic"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##  -0.2594386   0.9182700   0.0817400   0.1195200 -0.3694719 -0.1840035
## [1] "Correlation between Historical and Geographic"
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##   0.4210629   0.0004100   0.9996000   0.0004100  0.3313578  0.5176683
```

Run a mantel test comparing the Linguistic alignment to the cultural similarity, controlling for the historical distance between languages:

```r
ecodist::mantel(as.dist(ling.m2)~
                as.dist(cult.m2) +
                as.dist(hist.m2),
              nperm = 100000)
```

```
##     mantelr       pval1       pval2       pval3  llim.2.5% ulim.97.5%
##   0.4659407   0.0100000   0.9900100   0.0107800  0.3408500  0.5938397
```

Main Test: Run a mantel test comparing the Linguistic alignment to the cultural similarity, controlling for the historical distance and geographic distance between languages:

```
ecodist::mantel(as.dist(ling.m2)~
                as.dist(cult.m2) +
                as.dist(hist.m2) +
                as.dist(geo.m2),
              nperm = 100000)
```

```
##     mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
##   0.4508309  0.0114200  0.9885900  0.0119500  0.2962271  0.5993660
```

MRM

Perform the main test, but using multiple regression on distance matrices (MRM).

Lichstein, J. W. (2007). Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecology, 188(2), 117-131.

```
ecodist::MRM(as.dist(ling.m2)~
                as.dist(cult.m2) +
                as.dist(hist.m2) +
                as.dist(geo.m2), nperm=10000)
```

```
## $coef
##                    as.dist(ling.m2)    pval
## Int                      0.098589462 0.8440
## as.dist(cult.m2)         0.350192033 0.0130
## as.dist(hist.m2)        -0.059044640 0.1639
## as.dist(geo.m2)          0.003551452 0.8186
##
## $r.squared
##        R2       pval
## 0.3072419 0.0074000
##
## $F.test
##        F    F.pval
## 24.68846  0.00740
```

Comparison between domains

The code that produce the results of this section can be found in `analysis/compareDomains.R`.

Part 1: Compare each linguistic domain to the overall cultural similarity

We fit a mixed effects model to compare the linguistic similarity in a given domain to the overall cultural distance. The linguistic similarity for the given domain is the dependent variable. There are random intercepts for language family and area pairs, and random slopes for overall cultural similarity by language family and by area. The `comparison_count` variable is added as a fixed effect. This null model is compared to a model with an additional fixed effect for the overall cultural similairty.

There are 21 linguistic domains with enough data. All correlations are positive and 16 are significant at the 0.05 level.

The full results are in the file:

`../results/stats/wikipedia-main/Cor_LingAlignmentByDomains_vs_OverallCulturalSimilarity.csv`

Summary:

`p1res`

```
##                                Domain       Beta            p  Adjusted p sig
## 2                      Food and drink 0.29038732 3.842274e-08 8.068775e-07   *
## 6       Miscellaneous function words 0.31349261 1.431794e-07 3.006768e-06   *
## 9                            The body 0.23184641 8.711593e-07 1.829434e-05   *
## 13                            Animals 0.26468547 4.281952e-06 8.992099e-05   *
## 21                               Time 0.26704310 3.341177e-05 7.016471e-04   *
## 3        Agriculture and vegetation 0.21315112 4.954909e-05 1.040531e-03   *
## 16                       Modern world 0.15391498 2.860946e-04 6.007988e-03   *
## 14                  The physical world 0.15534299 6.771587e-04 1.422033e-02   *
## 11                   Spatial relations 0.11188011 1.323355e-03 2.779045e-02   *
## 20                            Kinship 0.25409044 1.332699e-03 2.798669e-02   *
## 7              Clothing and grooming 0.16482738 2.245060e-03 4.714625e-02   *
## 10                  Sense perception 0.11167650 2.806430e-03 5.893504e-02
## 15 Social and political relations 0.10210510 6.603817e-03 1.386802e-01
## 1                           The house 0.10780689 1.485121e-02 3.118755e-01
## 18                           Quantity 0.13804391 1.691864e-02 3.552914e-01
## 8                 Speech and language 0.11367022 3.209804e-02 6.740588e-01
## 19    Basic actions and technology 0.06996425 7.317704e-02 1.000000e+00
## 17                          Cognition 0.06410616 9.294337e-02 1.000000e+00
## 12              Emotions and values 0.06324335 9.437249e-02 1.000000e+00
## 5                         Possession 0.07833760 1.102507e-01 1.000000e+00
## 4                             Motion 0.05542835 2.537090e-01 1.000000e+00
```

Part 2: Compare each linguistic domain to the cultural similarity of each original D-PLACE domain

The method is the same as for part 1, except the cultural distance for a particular cultural domain is used instead of the overall cultural distance.

The full results are in the file:

`../results/stats/wikipedia-main/Cor_LingAlignmentByDomains_vs_DPlaceCulturalDomains.csv`

The graph below shows the mixed effects model coefficient estimate for the relationship between each linguistic domain and each cultural domain. Pink colours indicate positive correlations and blue colours indicate negative correlations. Stronger colours indicate stronger correlations. An asterisk indicates that the correlation is stronger than would be expected by chance, when adjusting the p-value for multiple comparisons.

The insert in the top left shows the distribution of Beta values.

The domains are clustered using higherarchical clustering. This is for visualisaiton and reflects similarity in the numeric relations, not history or conceptual hierarchies.

List of significant correlations (after adjusting p-value for multiple comparisons):

```
##                            Ling Domain Cult Domain      Beta   Adjusted p
## 19       Agriculture and vegetation  Settlement 0.2036154 7.336590e-03
## 11                    Food and drink  Settlement 0.2246202 6.885552e-03
## 43    Miscellaneous function words  Settlement 0.2879246 1.378491e-05
## 83                  Spatial relations  Settlement 0.1189770 8.020950e-03
## 59               Speech and language  Settlement 0.1848472 3.806376e-03
## 17       Agriculture and vegetation Subsistence 0.2792663 3.749181e-05
## 97                           Animals Subsistence 0.2991488 7.642513e-04
## 49             Clothing and grooming Subsistence 0.2718306 2.096807e-05
## 89             Emotions and values Subsistence 0.1392281 3.317802e-02
## 9                     Food and drink Subsistence 0.3682025 8.136554e-08
## 41    Miscellaneous function words Subsistence 0.3353682 2.648483e-05
## 121                     Modern world Subsistence 0.1792931 8.993827e-04
## 73                  Sense perception Subsistence 0.1665568 1.865977e-02
## 113 Social and political relations Subsistence 0.1332652 7.808655e-03
## 81                  Spatial relations Subsistence 0.1303419 4.923756e-02
## 65                         The body Subsistence 0.2701108 1.258912e-06
## 105               The physical world Subsistence 0.1949683 1.148414e-02
## 161                             Time Subsistence 0.3140335 5.332731e-04
```
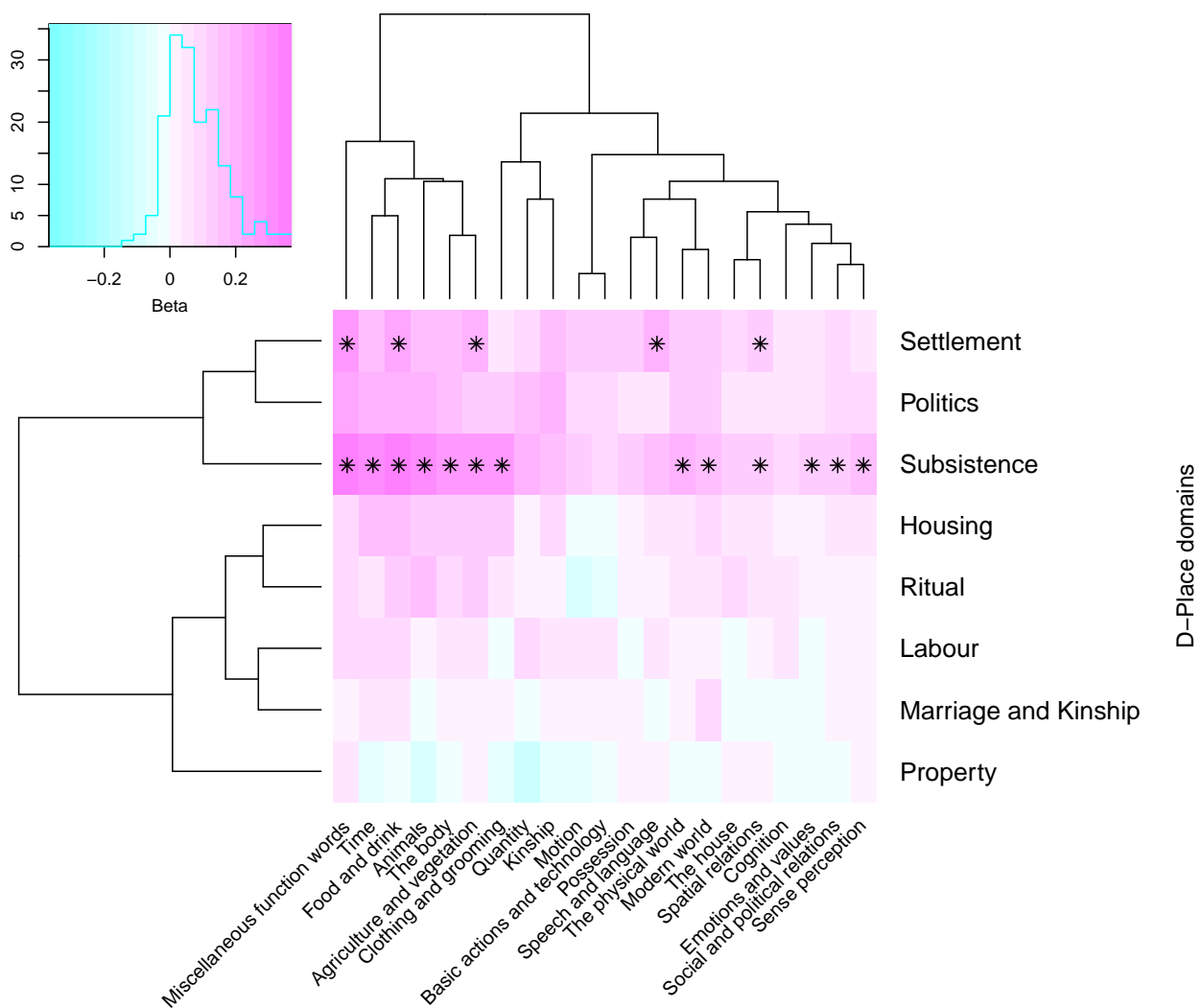
Figure 1:

Part 3: Compare each linguistic domain to the phylogenetic and geographic distance

This test compares each linguistic similarity scores to each of three target distances: the cultural distance, the historical distance and the geographic distance. We use a partial Mantel test (from the package `ecodist`) to estimate the strength of the relationship between the linguistic domain and the target distance, while controlling for the other two distances. The test uses 100,000 permutations.

The full results are in the file:

`Cor_LingAlignmentByDomains_vs_HistoricalAndGeographicalDistance.csv`

The graph below shows the results. Point estimates are the estimated Mantel R. The error bars show the 95% confidence intervals from the permutation test.
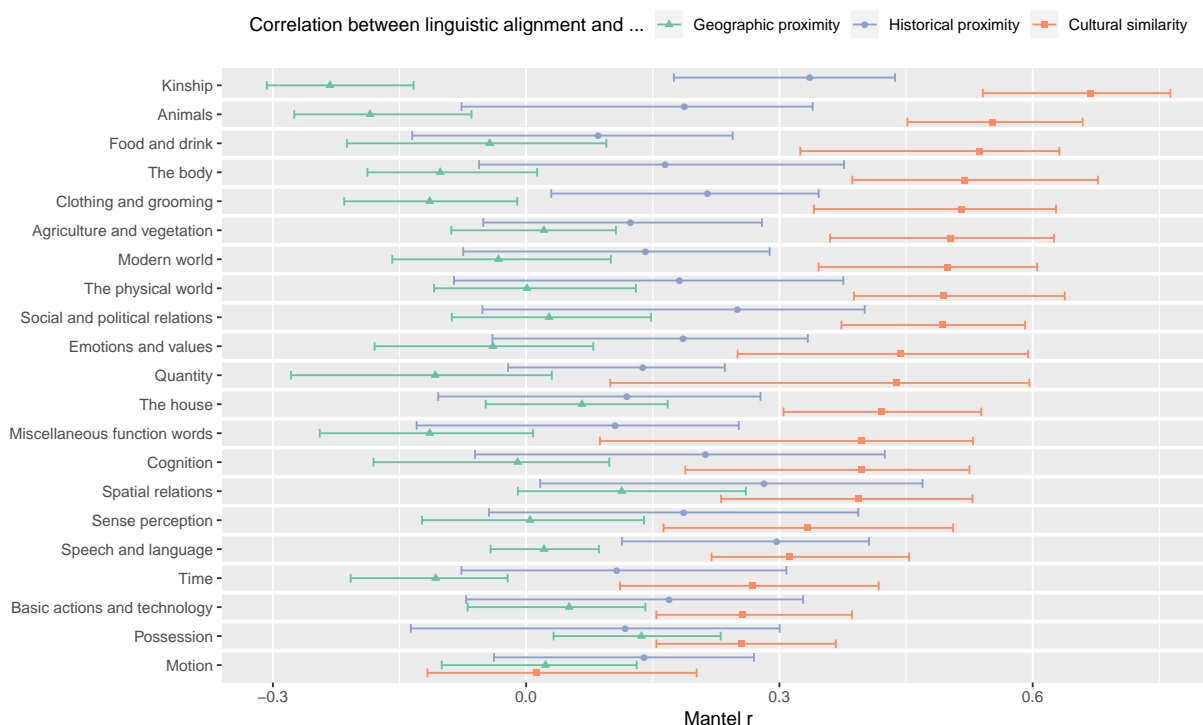

Figure 2:

There appears to be a trade-off: The stronger the relationship with geographic distance, the weaker the relationship with cultural distance (r = -0.529, t = -2.72, df=19, p = 0.014). This does not hold for historical and cultural distance (r = 0.27, t = 1.22, df=19, p = 0.24).

Note that, after controlling for multiple comparisons, only 2 domains are significant:

```
##     domain comparison    mantelr     lower     upper   pval3 p.adjusted
## 37 Animals  lingVCult  0.5518312 0.4515039 0.6591382 0.00129    0.02709
## 58 Kinship  lingVCult  0.6687835 0.5407906 0.7629987 0.00012    0.00252
```

References

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012). Mapping the origins and

expansion of the Indo-European language family. Science, 337(6097), 957-960.