Composite Monsters

Contents

Introduction

Load libraries

```
library(ggplot2)
library(maps)
library(mapproj)
library(party)
library(lattice)
library(rsq)
library(car)
library(lme4)
library(MuMIn)
library(sjPlot)
```

```
library(glmmTMB)
library(lmtest)
library(ggpubr)
library(ape)
library(caper)
library(phytools)
library(stringr)
library(polycor)
library(forcats)
library(fields)
library(RColorBrewer)
library(brms)
```

Load data

Load the main data and make sure the categorical variables are treated appropriately.

```
d = read.csv("../data/clean/monsters.csv",stringsAsFactors = F)
```

For convenience, make a labelled monster variable.

```
d$monster_present2 = factor(d$monster_present,
                            labels=c("Monster Absent","Monster Present"))
```

Make sure class and caste are categorical variables

```
d$class.stratified = factor(d$class.stratified, labels = c("Not stratified", "Class stratified"))
d$caste.stratified = factor(d$caste.stratified, labels = c("Not stratified", "Caste stratified"))
```

Urbanization as ordered category

```
d$urban = factor(d$urban, ordered = TRUE,
                 labels = c("< 100 persons",
                            "100-199 persons",
                            "200-399 persons",
                            "400-999 persons",
                            "1000+ persons"))
```

Agriculture as ordered category

```
d$ag = factor(d$ag, ordered = TRUE,
                 labels = c("None",
                            "10% food supply",
                            "10 %; secondary",
                            "Primary; not intensive",
                            "Primary; intensive"))
```

Population Density as ordered category

```
d$popdens = factor(d$popdens, ordered = TRUE,
                 labels = c("< 1 person / sq. mile",
                            "1-5 persons / sq. mile",
                            "5.1-25 persons/ sq. mile",
                            "26-100 persons  / sq. mile",
                            "100 persons / sq. mile"))
```

Fixity of Residence as ordered category

```
d$fixity = factor(d$fixity, ordered = TRUE,
                  labels = c("Nomadic",
                             "Seminomadic",
                             "Semisedentary",
                             "Sedentary; impermanent",
                             "Sedentary"))
```

Land Transport as ordered category

```
d$land = factor(d$land, ordered = TRUE,
                  labels = c("Human only",
                             "Pack Animals",
                             "Draft Animals",
                             "Animal-drawn vehicles",
                             "Automotive vehicles"))
```

Contact as an ordered category

```
d$contact = factor(d$contact, ordered = TRUE,
                  labels = c("Rare or never",
                             "Occasional",
                             "Frequent"))
```

Money as ordered category

```
d$money = factor(d$money, ordered = TRUE,
                  labels = c("None",
                             "Domestically usable particles",
                             "Alien currency",
                             "Elementary forms",
                             "True money"))
```

Political Integration as ordered category

```
d$politic = factor(d$politic, ordered = TRUE,
                  labels = c("None",
                             "Autonomous local communities",
                             "1 level above community",
                             "2 levels above community",
                             "3 levels above community"))
```

Social Stratification as ordered category

```
d$strata = factor(d$strata, ordered = TRUE,
                  labels = c("Egalitarian",
                             "Wealth Differences or hereditary slavery",
                             "2 social classes, no castes/slavery",
                             "2 social classes, castes/slavery",
                             "3 social classes or castes, w/ or w/out slavery"))
```

Technological Specialization as ordered category

```
d$tech = factor(d$tech, ordered = TRUE,
                  labels = c("None",
                             "Pottery only",
                             "Loom weaving only",
                             "Metalwork only",
                             "Smiths, weavers, potters"))
```

Writing and Records as ordered category

```
d$writing = factor(d$writing, ordered = TRUE,
                   labels = c("None",
                              "Mnemonic devices",
                              "Nonwritten records",
                              "True writing; no records",
                              "True writing, records"))
```

Market Exchange as ordered category

```
d$market = factor(d$market, ordered = TRUE, labels = c("Local",
                                                       "Outside Local",
                                                       "Regional",
                                                       "Supra-Regional"))
```

Collapse land transport categories 4 and 5

```
d$land = forcats::fct_collapse(d$land,
            "Human only"    = c("Human only"),
            "Pack Animals"  = c("Pack Animals"),
            "Draft Animals" = c("Draft Animals"),
            "Wheeled Vehicles"   = c("Animal-drawn vehicles", "Automotive vehicles"))
```

Collapse technological specialization

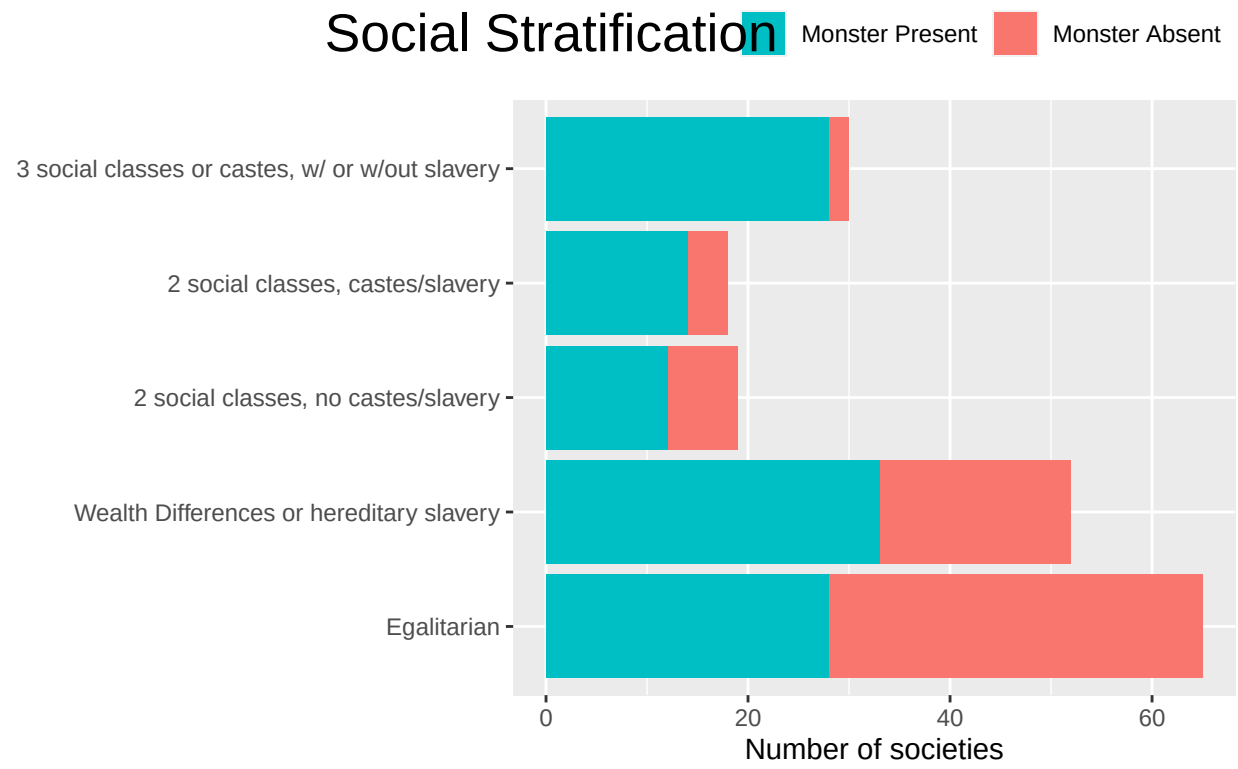
```
d$tech = forcats::fct_collapse(d$tech,
            "None" = c("None"),
            "Pottery or Loom" = c("Pottery only", "Loom weaving only"),
            "Metallurgy" = c("Metalwork only"),
            "Multiple specialists" = c("Smiths, weavers, potters"))
```

Collapse True writing in Writing

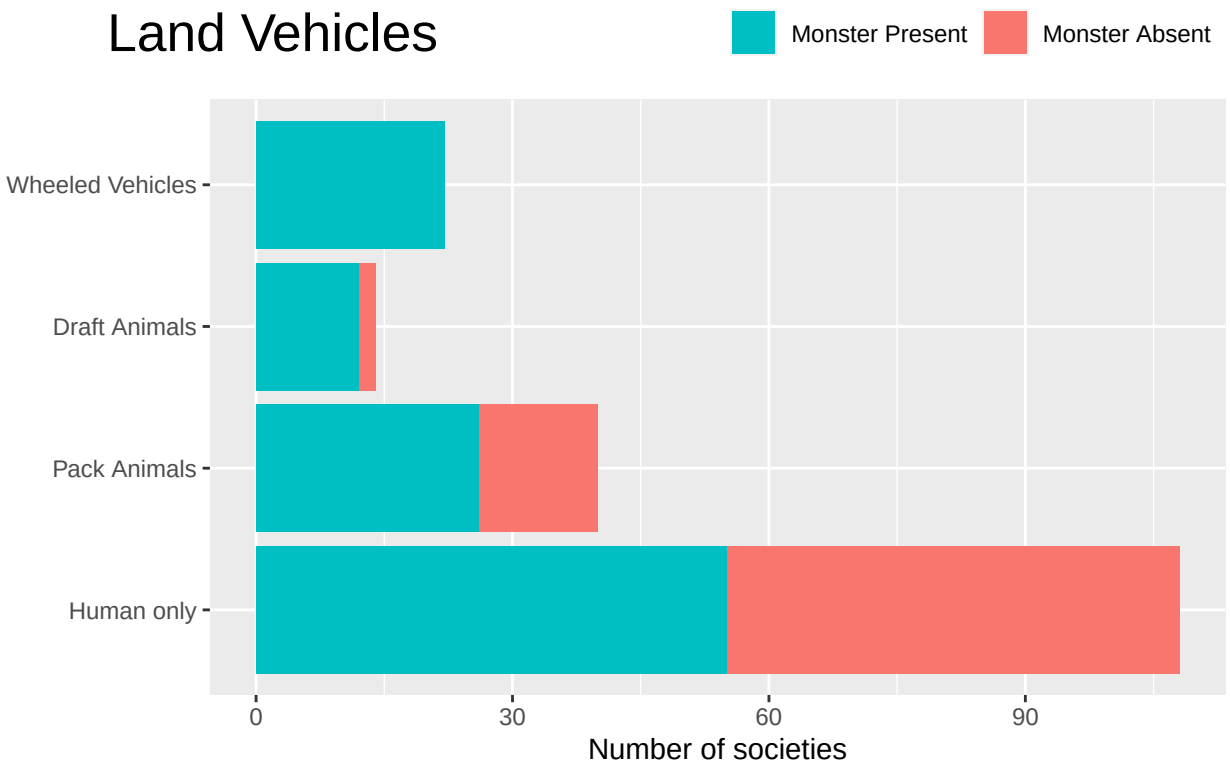
```
d$writing = forcats::fct_collapse(d$writing,
             "None" = c("None"),
             "Mnemonic devices" = c("Mnemonic devices"),
             "Nonwritten records" = c("Nonwritten records"),
             "True writing" = c("True writing; no records", "True writing, records"))
```

Data plots

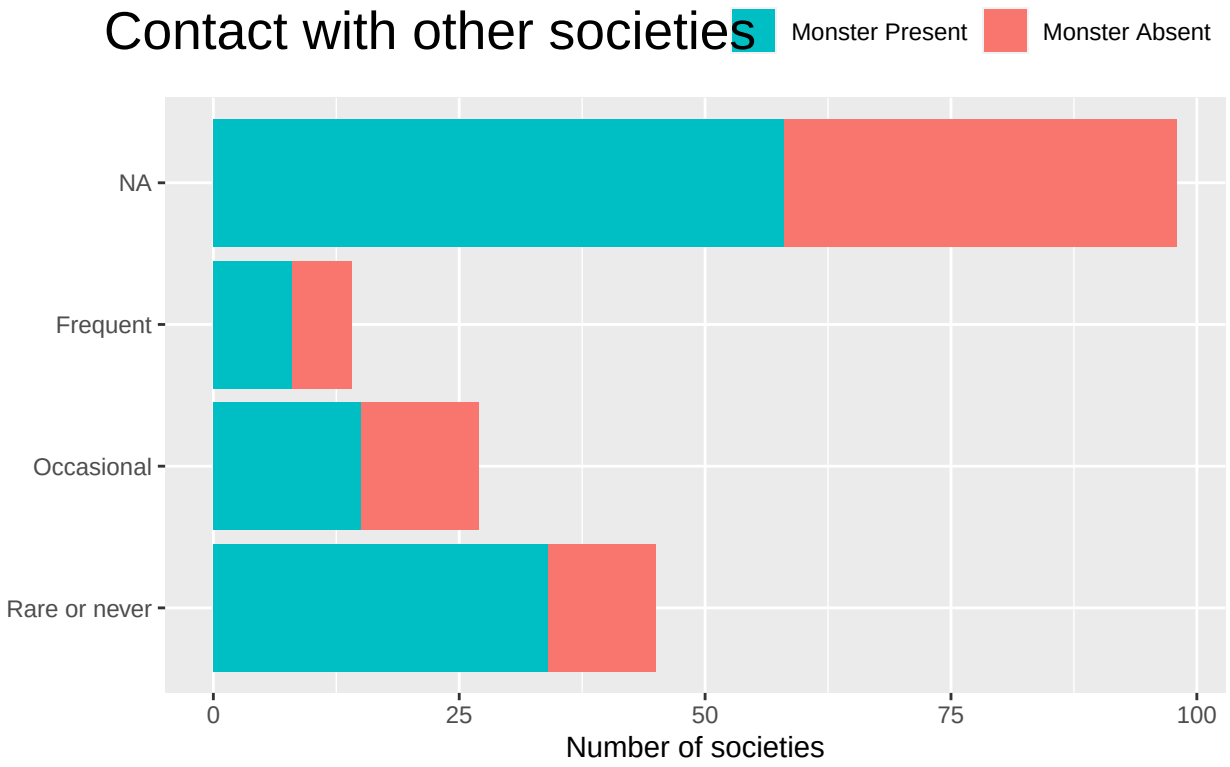
```
stackedBar = function(var,xtitle,hjust=-0.7){
  gx = ggplot(d,aes_string(fill="monster_present2",x=var)) +
    geom_bar() +
    ylab("Number of societies") +
    scale_fill_discrete(guide = guide_legend(reverse = TRUE)) +
    theme(legend.position = "top",
          legend.justification = c(1,0),
          axis.title.y = element_blank(),
          legend.title = element_blank(),
          plot.title = element_text(hjust = hjust,vjust=-6,size = 20)) +
    ggtitle(xtitle)+
    coord_flip()
  fn=paste0("../results/bar/Bar_",gsub(" ","_",xtitle),".pdf")
  pdf(fn,width=8,height=4)
  plot(gx)
  dev.off()
  gx
}

stackedBar("strata","Social Stratification")
```



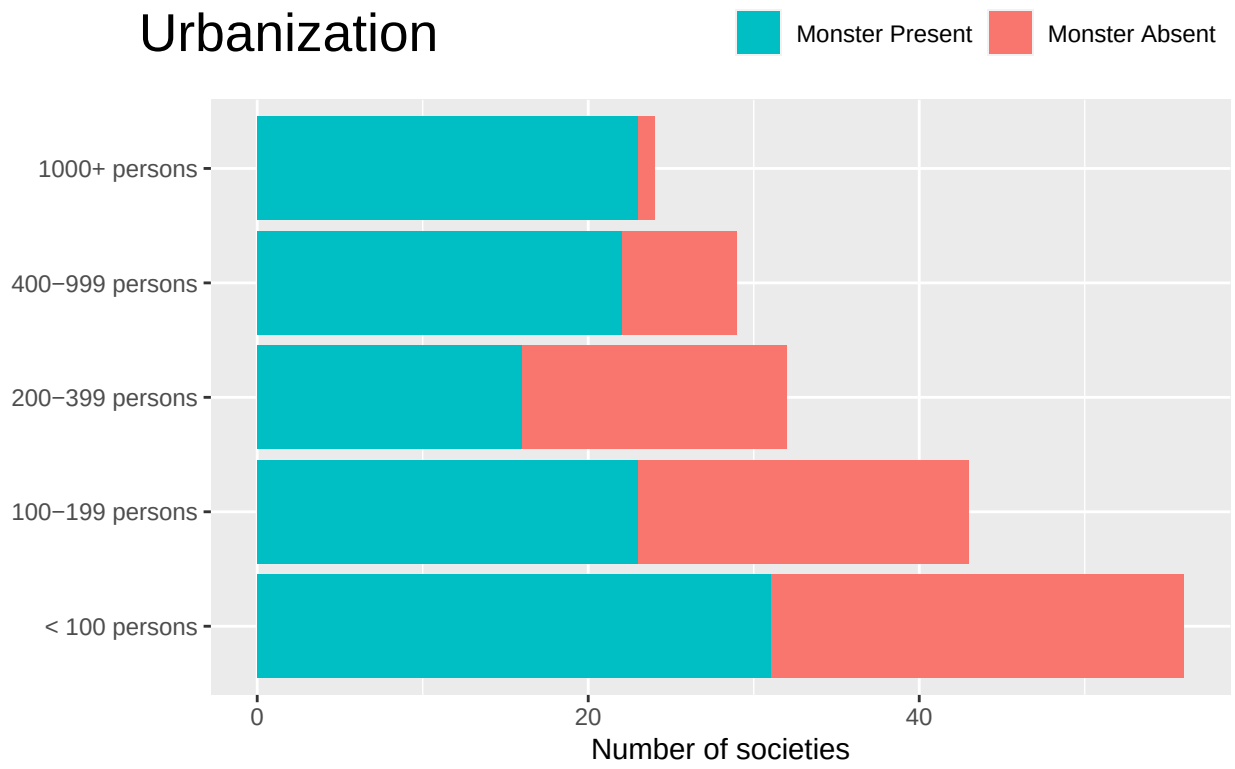

```
stackedBar("land","Land Vehicles",-0.15)
```



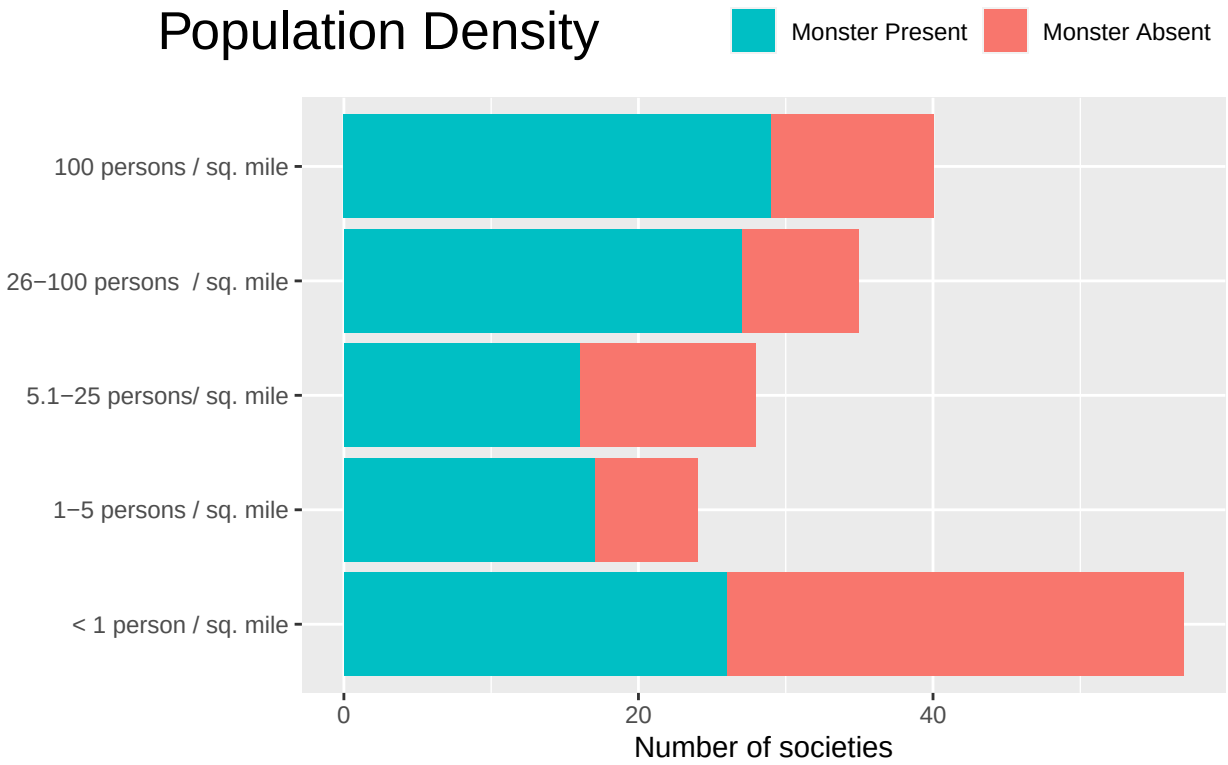

```
stackedBar("contact","Contact with other societies",-0.15)
```



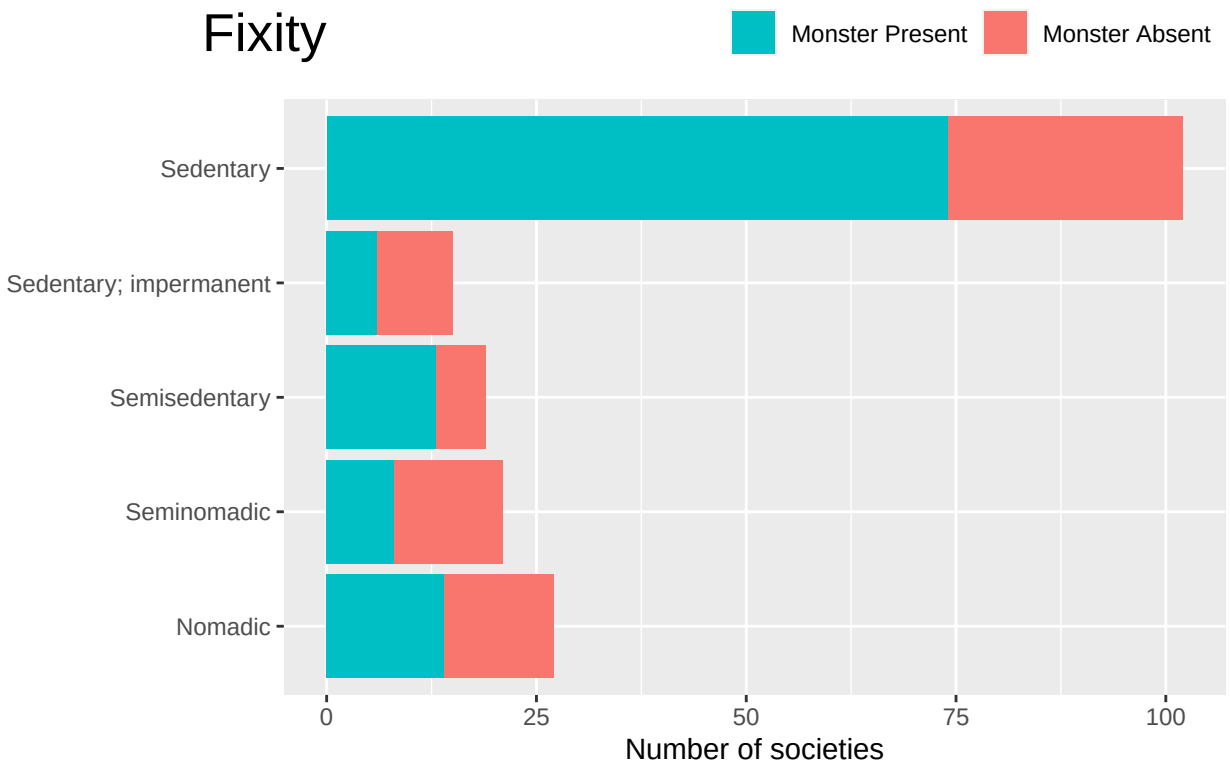

```
stackedBar("urban","Urbanization",-0.1)
```



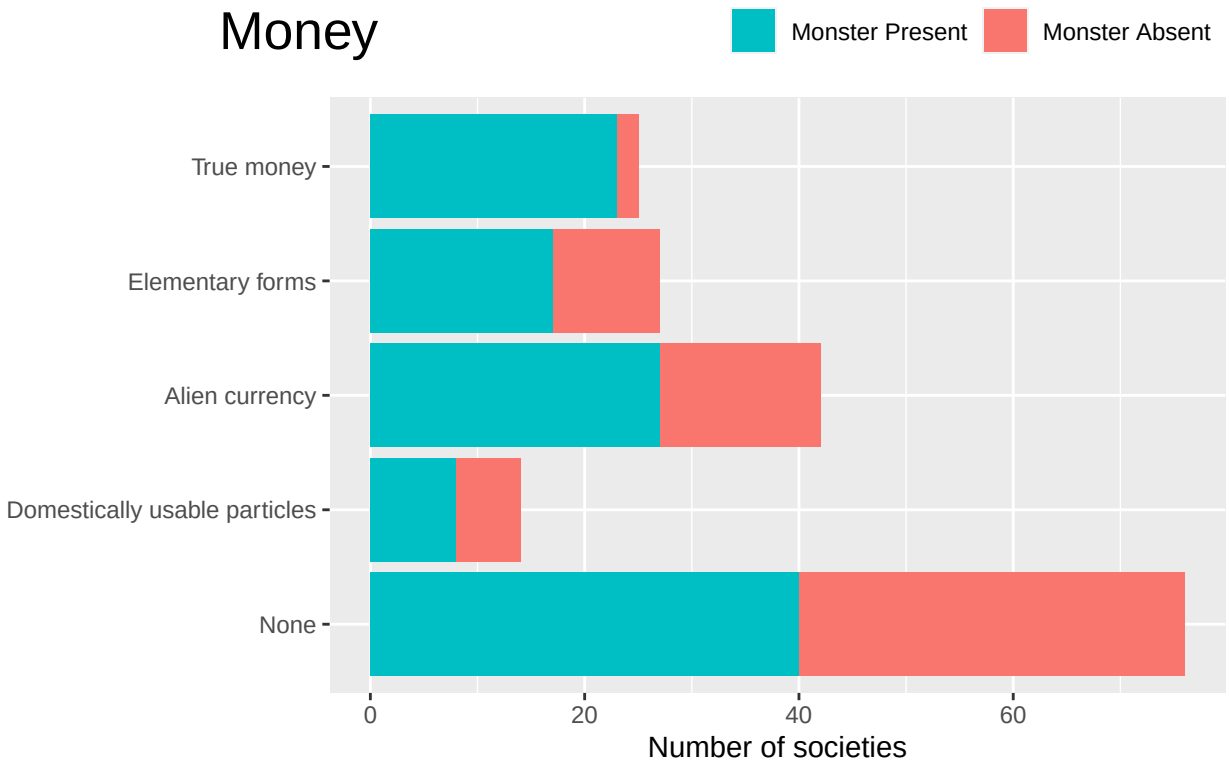

```
stackedBar("popdens","Population Density",-0.3)
```



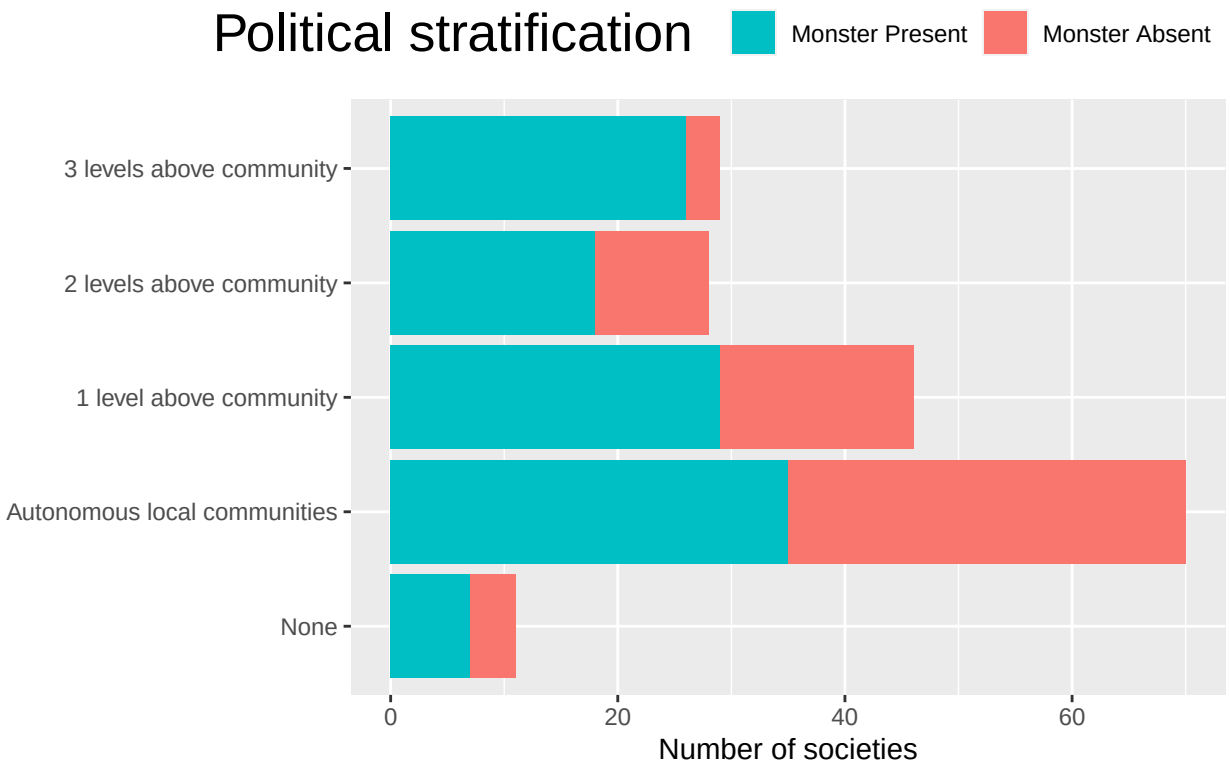

```
stackedBar("fixity","Fixity",-0.1)
```




```
stackedBar("money","Money",-0.15)
```

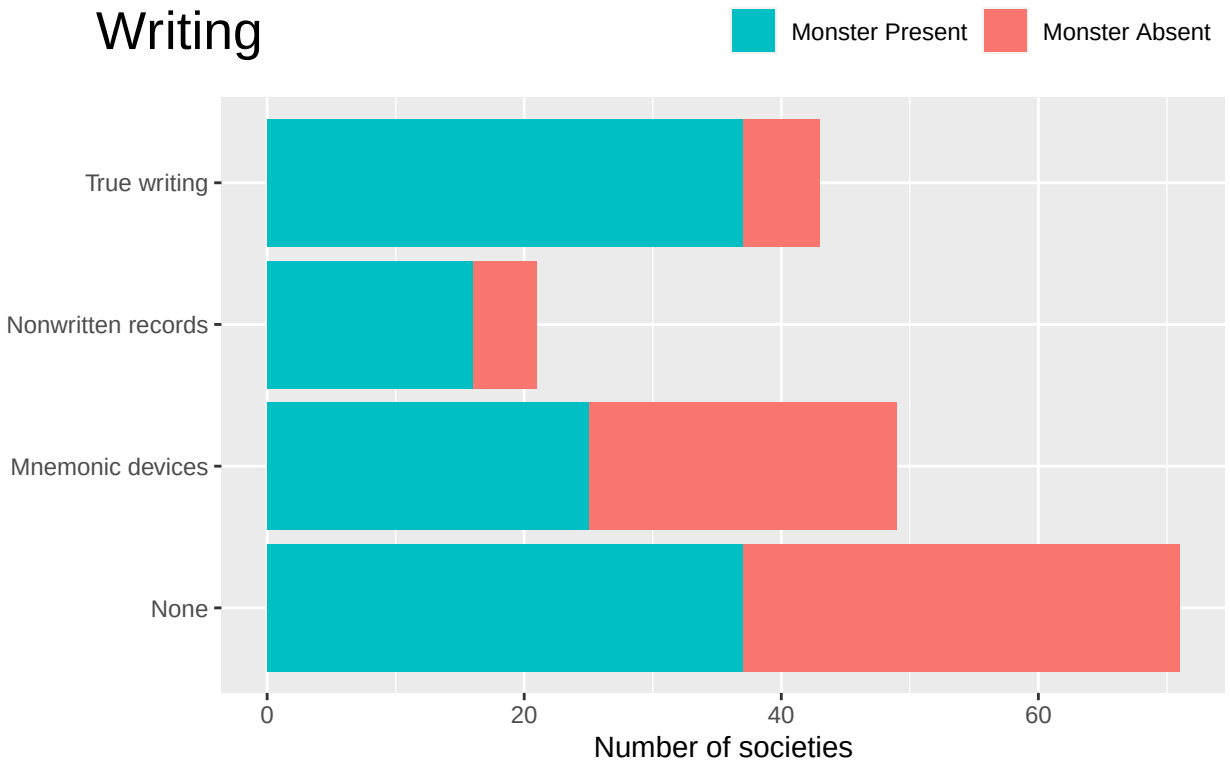



```
stackedBar("politic","Political stratification",-0.35)
```




```
stackedBar("tech","Technical specialists",-0.2)
```


```
stackedBar("writing","Writing",-0.15)
```


Writing
Monster Present
Monster Absent
True writing
Nonwritten records
Mnemonic devices
None
0
20
40
60
Number of societies

Hypothesis tests

Are composite monsters universal?

```
table(d$monster_present)
```

```
##
## FALSE  TRUE
##    69   115
```

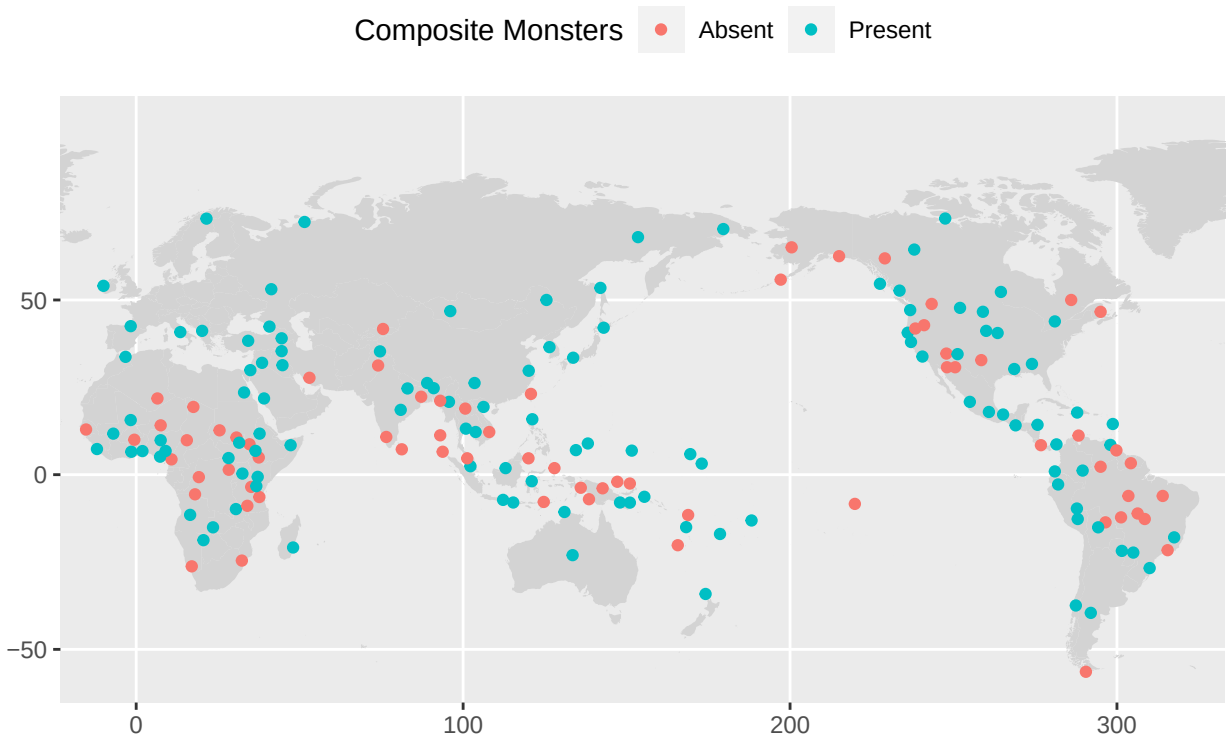
```
prop.table(table(d$monster_present))
```

```
##
## FALSE  TRUE
## 0.375 0.625
```

Composite monsters were present in 115 societies (62.5%) and absent in 69. These represent societies that speak languages from 62 different language families, including 5 isolates.

Plot map of the world:

```
d$lon2 <- ifelse(d$longitude < -25, d$longitude + 360, d$longitude)
mapWorld <- map_data('world', wrap=c(-25,335), ylim=c(-55,75))
d$`Composite Monsters` = factor(d$monster_present,labels=c("Absent","Present"))
monsterMap = ggplot() +
geom_polygon(data = mapWorld, aes(x=long, y = lat, group = group),
             fill="lightgray") +
  geom_point(data = d, aes(x = lon2, y = latitude,
                           colour=`Composite Monsters`)) +
  coord_map(projection = "gall",0) +
  theme(legend.position = "top",axis.title = element_blank())
monsterMap
```




```
pdf("../results/MonsterMap.pdf",width=8,height=5)
monsterMap
dev.off()
```

```
## pdf
##   2
```

There does not appear to be a high degree of geographic clustering. Indeed, running Moran's I test for geospatial autocorrelation shows that there is no significant spatial clustering:

```
# Distance between points
mdists = rdist.earth(cbind(d$longitude, d$latitude),miles = F)
# Invert
mdists.inv <- 1/mdists
diag(mdists.inv) <- 0
Moran.I(d$monster_present, mdists.inv)
```

```
## $observed
## [1] 0.001120952
##
## $expected
## [1] -0.005464481
##
## $sd
## [1] 0.0101295
##
## $p.value
## [1] 0.5156121
```

Phylogenetic signal

The SCCS sample aims for maximum diversity across the world. This is not ideal for testing whether the presence of composite monsters has a phylogenetic signal. Still, some formal tests can be run.

The raw data folder includesn phylogenetic trees based on linguistic properties, downloaded from https://github.com/D-PLACE/dplace-data/tree/master/phylogenies/. The D-PLACE database links these languages to SCCS societies. In the tests below, we look at Austronesian (Gray et al., 2009) and Bantu (Grollemund et al., 2015). Similar phylogenies are available for Indo-Europen, Semitic and Uto-Aztecan languages, but these only contain 3 or fewer societies in the SCCS.

Two formal tests of phylogenetic signal are computed. First, the D statistic for binary traits (Fritz & Purvis, 2010). The statistic reflects the estimated number of changes along the phylogeny to produce the observed data at the tips. A statistic near 0 suggests that the trait is highly conserved (strong signal), and a near 1 suggests the trait is similar to a random trait (no phylogenetic signal). Values above 1 indicate that related societies are more different than would be expected by chance (overdispersion). This statistic is compared to two models of cultural evolution: random distribution (where the tips are shuffled and the trait is re-calcualted), and Brownian motion (a continuous trait which takes a random walk is simulated and converted to a binary one).

The second statistic is Pagel's lambda (Pagel, 1999). It indicates the extent of the scaling of the branch lengths that would be required for the data to fit a model of Brownian motion. Values close to 0 have low phylogenetic signal, and values close to 1 have high phylogenetic signal.

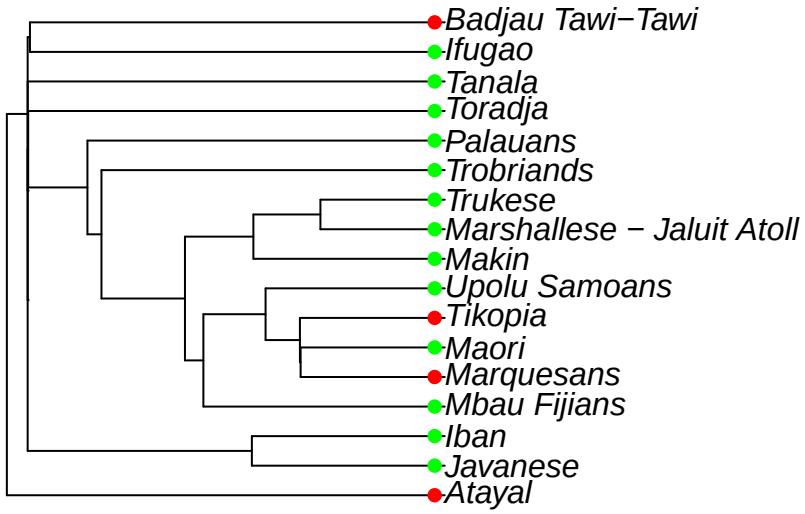
```
# Function to run phylogenetic tests
runPhylogenyTest = function(treeFile, treeLabelFile,plotType="flat"){
# Load trees
  tree = read.nexus(treeFile)
  treeLabels = read.csv(treeLabelFile,stringsAsFactors = F)

  # Taxa data has multiple codes per locus
  # So identify SCCS ID
  treeLabels$SCCSID = str_extract(treeLabels$soc_ids,"SCCS[0-9]+")

  # Set labels of tree to be D-PLACE xd-ids
  tree$tip.label = treeLabels[match(tree$tip.label,treeLabels$taxon),]$SCCSID

  # Remove tips that aren't in D-PLACE
  # (remove tips that are now named "")
  tree = drop.tip(tree,"")

  # Find tips in tree that aren't in our data:
  notInMonsters = tree$tip.label[!tree$tip.label %in% d$id]
  # Remove these tips from the tree
  tree = drop.tip(tree, notInMonsters)

  # subset of data in tree
  dx = d[d$id %in% tree$tip.label,]

  # Choose colours for the tips
  chosenVariableToPlot = "monster_present"
  tipColours = c("red", "green")[1 + dx[,chosenVariableToPlot]]
  names(tipColours) = dx$id
  tipColours = tipColours[tree$tip.label]

  # Convert labels back to names
```

```
  tree$tip.label = dx[match(tree$tip.label,dx$id),]$pref_name_for_society

  if(plotType=="flat"){
    plot(tree)
    tiplabels(pch=16,col=tipColours,adj = 0.4)
  } else{
    tree$tip.label = iconv(tree$tip.label,sub='')
    plot.phylo(tree,align.tip.label=0)
    tiplabels(pch=16,col=tipColours)
    dx$pref_name_for_society = iconv(dx$pref_name_for_society,sub="")
  }

  tree = di2multi(tree)

  # Calculate phylogenetic signal
  # (Phylo - D)
  pd = phylo.d(data=dx[,c("pref_name_for_society","monster_present")],
          phy=tree,
          names.col = pref_name_for_society,
          binvar = monster_present,permut = 100000)
  print(pd)
  # Pagel's Lambda
  x = dx$monster_present
  names(x) = dx$pref_name_for_society
  print(phylosig(tree,x,method="lambda"))
}
```

Austronesian

Tree from Gray et al. (2009)

```
set.seed(2389)
runPhylogenyTest("../data/raw/trees/Austronesian/summary.trees",
                 "../data/raw/trees/Austronesian/taxa.csv")
```

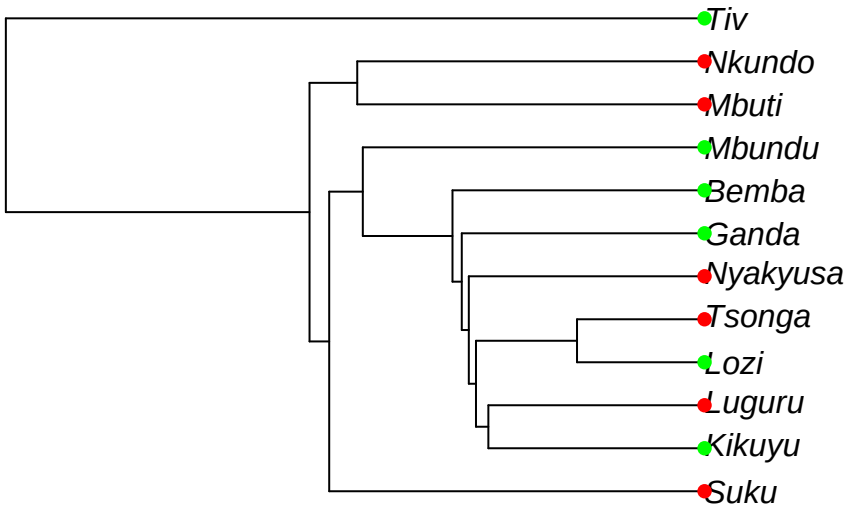



```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  data
##   Binary variable :  monster_present
##   Counts of states:  FALSE = 4
##                      TRUE = 13
##   Phylogeny :  phy
##   Number of permutations :  1e+05
##
## Estimated D :  0.8878351
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.41429
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.35646
##
##
## Phylogenetic signal lambda : 6.6107e-05
## logL(lambda) : -9.54299
```

Bantu

Tree from Grollemund et al. (2015)

```
runPhylogenyTest("../data/raw/trees/Bantu/summary.trees",
                 "../data/raw/trees/Bantu/taxa.csv")
```

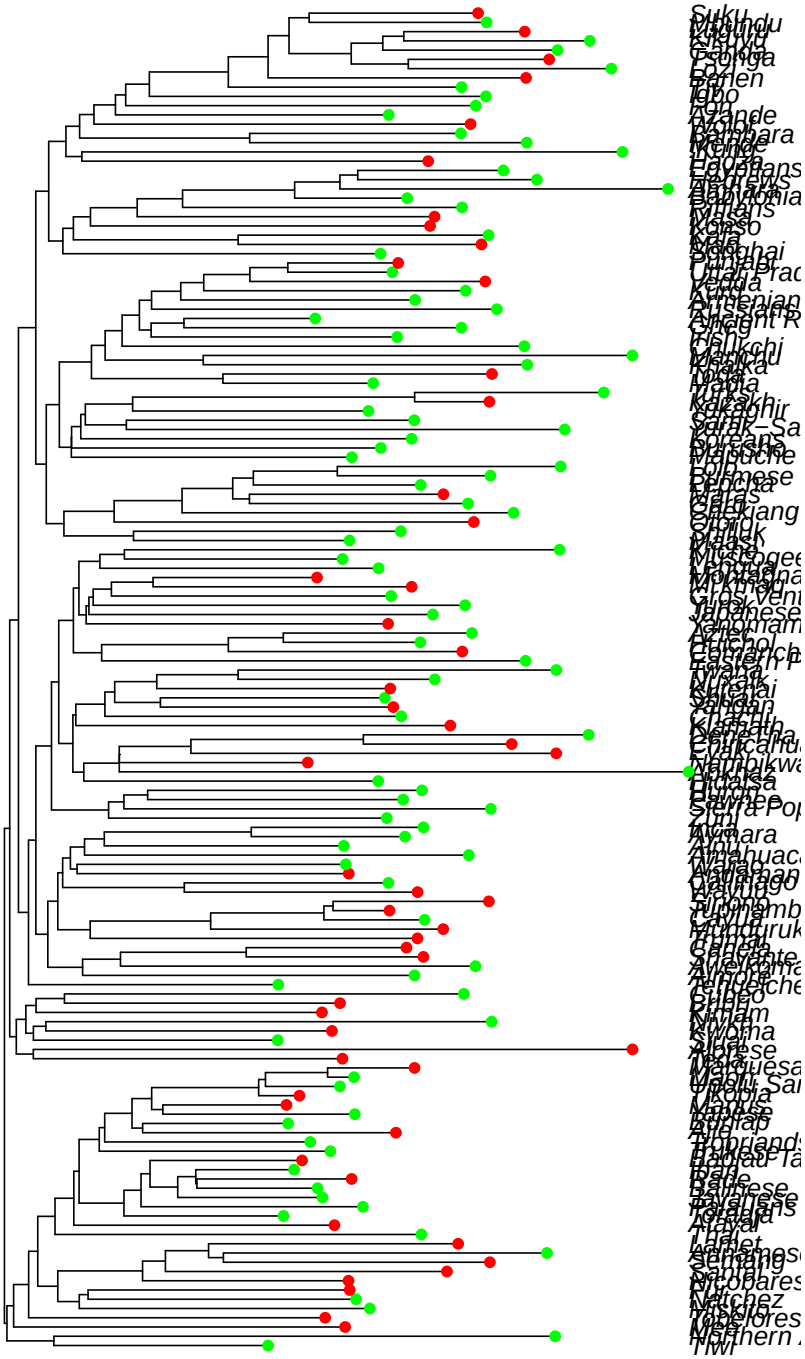



```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  data
##   Binary variable :  monster_present
##   Counts of states:  FALSE = 6
##                      TRUE = 6
##   Phylogeny :  phy
##   Number of permutations :  1e+05
##
## Estimated D :  2.115693
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.61609
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.26137
##
##
## Phylogenetic signal lambda : 8.08671e-05
## logL(lambda) : -8.70973
```

Global

Tree from Jäger (2018). This is a global tree of languages, calculated from an analysis of the forms of basic vocabulary words. While this may not be ideal for the current questions, it's the best currently available global tree of languages.

```
runPhylogenyTest(
  "../data/raw/trees/Global/JaegerGlobalTree_SCSS.tree",
  "../data/raw/trees/Global/taxa.csv", plotType = "big")
```



```
##
## Calculation of D statistic for the phylogenetic structure of a binary variable
##
##   Data :  data
##   Binary variable :  monster_present
##   Counts of states:  FALSE = 52
##                      TRUE = 93
##   Phylogeny :  phy
##   Number of permutations :  1e+05
##
## Estimated D :  1.136271
## Probability of E(D) resulting from no (random) phylogenetic structure :  0.67589
## Probability of E(D) resulting from Brownian phylogenetic structure    :  0.00122
##
##
## Phylogenetic signal lambda : 6.6107e-05
## logL(lambda) : -101.456
```

Summary

In all cases above, both tests suggest a low phylogenetic signal. There is some evidence of overdispersion. However, the current sample of cultures is not ideal to test whether the presence of composite monsters follows laws of cultural evolution. Rather, the conclusion at this point is that there is little historical signal to control for, which is expected given that we're using the SCCS sample.

Explanatory factors

Ethnographic coverage

Coding presence and absence from ethnographic sources can be affected by the amount of ethnographic materials available. Although we use multiple sources, HRAF was the main source, and indeed, we found that the presence of monsters can be predicted by a greater number of sources and source pages in HRAF:

```
t.test(d$HRAFNumPages~d$monster_present)
```

```
##
##  Welch Two Sample t-test
##
## data:  d$HRAFNumPages by d$monster_present
## t = -6.1379, df = 149.87, p-value = 7.135e-09
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -1844.8263  -946.3041
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            1139.043            2534.609
```

```
t.test(d$HRAFNumSources~d$monster_present)
```

```
##
##  Welch Two Sample t-test
##
## data:  d$HRAFNumSources by d$monster_present
## t = -4.9589, df = 181.98, p-value = 1.618e-06
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -12.787657  -5.507996
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            11.39130            20.53913
```

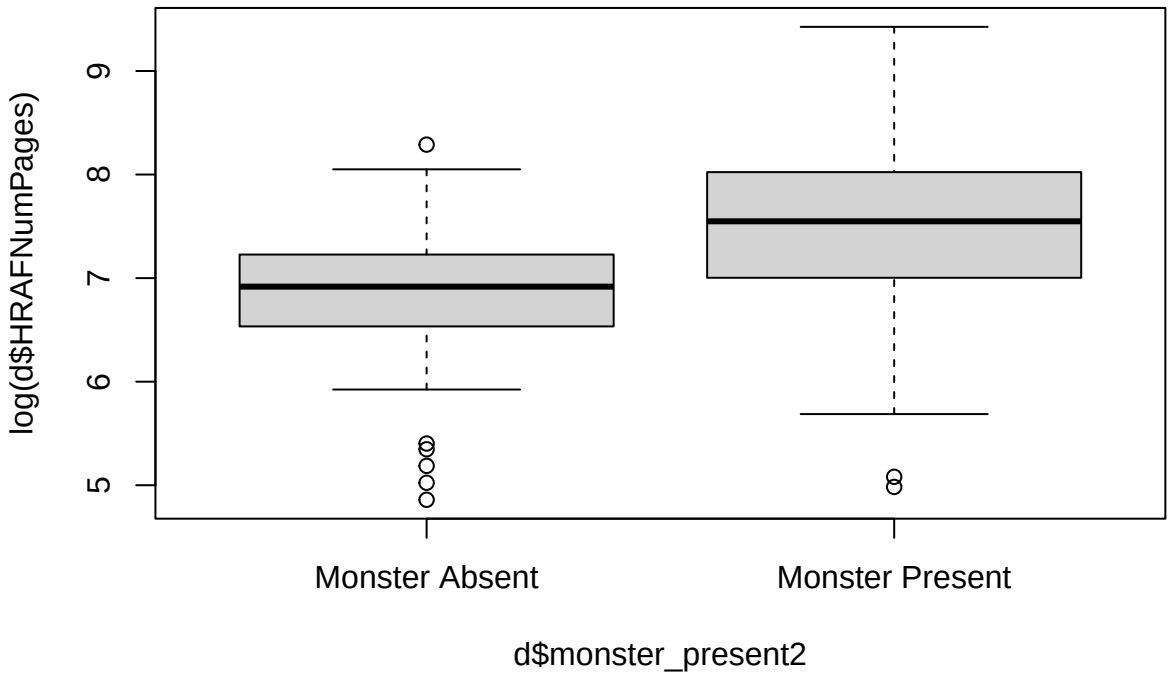
On average, societies with monsters present were covered by sources 20.5 and 2534.6 pages. Societies with monsters absetn were coded by 11.4 sources and 1139 pages.

This is an alternative explanation for the presence and absence of monsters, so we control for this in the explantory model below. The number of pages and number of sources are highly correlated, so we choose just one variable. The number of pages is a stronger predictor than the number of sources, and using the log of the number of pages makes this variable normally distributed and still stronger associated:

```
t.test(log(d$HRAFNumSources)~d$monster_present)
```

```
##
##  Welch Two Sample t-test
##
## data:  log(d$HRAFNumSources) by d$monster_present
## t = -5.5784, df = 148, p-value = 1.123e-07
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -0.7837906 -0.3737391
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2.202489            2.781254
```

```
boxplot(log(d$HRAFNumPages)~d$monster_present2)
```



So we will control for log number of pages in HRAF.

```
d$HRAFNumPages.log = log(d$HRAFNumPages)
```

Social stratification

Is the presence of composites associated with increased social stratification?

Fisher's exact test of the association between monster presence and class stratification as an ordered variable

```
tx = table(d$strata,d$monster_present2)
tx
```

```
##
##                                                 Monster Absent
##   Egalitarian                                               37
##   Wealth Differences or hereditary slavery                  19
##   2 social classes, no castes/slavery                        7
##   2 social classes, castes/slavery                           4
##   3 social classes or castes, w/ or w/out slavery            2
##
##                                                 Monster Present
##   Egalitarian                                               28
##   Wealth Differences or hereditary slavery                  33
##   2 social classes, no castes/slavery                       12
##   2 social classes, castes/slavery                          14
```

```
##   3 social classes or castes, w/ or w/out slavery              28
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                                                  Monster Absent
##   Egalitarian                                             56.92
##   Wealth Differences or hereditary slavery                36.54
##   2 social classes, no castes/slavery                     36.84
##   2 social classes, castes/slavery                        22.22
##   3 social classes or castes, w/ or w/out slavery          6.67
##
##                                                  Monster Present
##   Egalitarian                                             43.08
##   Wealth Differences or hereditary slavery                63.46
##   2 social classes, no castes/slavery                     63.16
##   2 social classes, castes/slavery                        77.78
##   3 social classes or castes, w/ or w/out slavery         93.33
```

```
# Rank correlation
cor(as.numeric(d$strata),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.3217255
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 2.236e-05
## alternative hypothesis: two.sided
```

The strata variable is derived from two earlier measures of class stratification and caste stratification. These variables are highly correlated, but since 'strata' has more ordered levels and therefore a more specific hypothesis, we use this instead of the other two. We note below that the association with composite monsters is significant for the class stratificaiton variable, but not significant for caste stratification variable.

Fisher test of the association between monster presence and class stratification:

```
tx = table(d$class.stratified,d$monster_present2)
print(tx)
```

```
##
##                    Monster Absent Monster Present
##   Not stratified               41              35
##   Class stratified             28              80
```

```
print(round(100*prop.table(tx,1),2))
```

```
##
##                    Monster Absent Monster Present
##   Not stratified            53.95           46.05
##   Class stratified          25.93           74.07
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
```

```
##
## data:  tx
## p-value = 0.000183
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.713883 6.549615
## sample estimates:
## odds ratio
##   3.323246
```

Test caste stratification:

```
tx = table(d$caste.stratified,d$monster_present2)
tx
```

```
##
##                     Monster Absent Monster Present
##   Not stratified               60              92
##   Caste stratified              9              19
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                     Monster Absent Monster Present
##   Not stratified            39.47           60.53
##   Caste stratified          32.14           67.86
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.5304
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.5484573 3.6902188
## sample estimates:
## odds ratio
##   1.374433
```

Increased travel

Do composite monsters reflect an increased interest in intersociety cultural differences accompanying increased long-distance travel? Murdock and Provost's (1973) scale of Land Transport serves as our proxy measure for increased long-distance travel in SCCS societies.

```
tx = table(d$land,d$monster_present2)
tx
```

```
##
##                     Monster Absent Monster Present
##   Human only                    53              55
##   Pack Animals                  14              26
##   Draft Animals                  2              12
##   Wheeled Vehicles               0              22
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                     Monster Absent Monster Present
##   Human only                 49.07           50.93
##   Pack Animals               35.00           65.00
##   Draft Animals              14.29           85.71
##   Wheeled Vehicles            0.00          100.00
```

```
# Rank correlation
cor(as.numeric(d$land),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.3114226
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 4.865e-06
## alternative hypothesis: two.sided
```

```
tx = table(d$fixity,d$monster_present2)
tx
```

```
##
##                           Monster Absent Monster Present
##   Nomadic                             13              14
##   Seminomadic                         13               8
##   Semisedentary                        6              13
##   Sedentary; impermanent               9               6
##   Sedentary                           28              74
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                           Monster Absent Monster Present
##   Nomadic                          48.15           51.85
##   Seminomadic                      61.90           38.10
##   Semisedentary                    31.58           68.42
##   Sedentary; impermanent           60.00           40.00
##   Sedentary                        27.45           72.55
```

```
cor(as.numeric(d$fixity),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2038183
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.005584
## alternative hypothesis: two.sided
```

Contact with other societies

From Ross (1983)

```
tx = table(d$contact,d$monster_present2)
tx
```

```
##
##                 Monster Absent Monster Present
##   Rare or never             11               34
##   Occasional                12               15
##   Frequent                   6                8
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                 Monster Absent Monster Present
##   Rare or never          24.44           75.56
##   Occasional             44.44           55.56
##   Frequent               42.86           57.14
```

```
cor(as.numeric(d$contact),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] NA
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.1576
## alternative hypothesis: two.sided
```

Urbanisation

In the case of the Bronze Age Near East and Mediterranean, Wengrow (2013, 74) has argued that "urban and state-like societies" provided a setting conducive for composites. If this is generalizable to a global ethnographic sample, we ought to find composite beings associated with an increased Urbanization and/or an increased Level of Political Integration (Murdock and Provost 1973).

Test urbanization:

```
tx = table(d$urban,d$monster_present2)
tx
```

```
##
##                   Monster Absent Monster Present
##   < 100 persons               25               31
##   100-199 persons             20               23
##   200-399 persons             16               16
##   400-999 persons              7               22
##   1000+ persons                1               23
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                   Monster Absent Monster Present
##   < 100 persons            44.64           55.36
```

```
##   100-199 persons          46.51           53.49
##   200-399 persons          50.00           50.00
##   400-999 persons          24.14           75.86
##   1000+ persons             4.17           95.83
```

```
# Rank correlation
cor(as.numeric(d$urban),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2084927
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.0003196
## alternative hypothesis: two.sided
```

Test political integration:

```
tx = table(d$politic,d$monster_present2)
tx
```

```
##
##                                Monster Absent Monster Present
##   None                                      4               7
##   Autonomous local communities             35              35
##   1 level above community                  17              29
##   2 levels above community                 10              18
##   3 levels above community                  3              26
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                                Monster Absent Monster Present
##   None                                  36.36           63.64
##   Autonomous local communities          50.00           50.00
##   1 level above community               36.96           63.04
##   2 levels above community              35.71           64.29
##   3 levels above community              10.34           89.66
```

```
# Rank correlation
cor(as.numeric(d$politic),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.206354
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.004399
## alternative hypothesis: two.sided
```

Technology

Wengrow (2014) sees Bronze Age composite figures as operating according to the same underlying logic as particular crafting technologies and bureaucratic technologies of writing and record-keeping (the latter also highlighted by Graeber (2015, xvii)). Are composite beings globally associated with greater specialization in crafting or recording-keeping technologies?

Technology specialists:

```
tx = table(d$tech,d$monster_present2)
tx
```

```
##
##                         Monster Absent Monster Present
##   None                              23               16
##   Pottery or Loom                   23               33
##   Metallurgy                        19               37
##   Multiple specialists               4               29
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                         Monster Absent Monster Present
##   None                           58.97            41.03
##   Pottery or Loom                41.07            58.93
##   Metallurgy                     33.93            66.07
##   Multiple specialists           12.12            87.88
```

```
# Rank correlation
cor(as.numeric(d$tech),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2715461
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.0003985
## alternative hypothesis: two.sided
```

Writing technology:

```
tx = table(d$writing,d$monster_present2)
tx
```

```
##
##                       Monster Absent Monster Present
##   None                            34               37
##   Mnemonic devices                24               25
##   Nonwritten records               5               16
##   True writing                     6               37
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                       Monster Absent Monster Present
##   None                         47.89            52.11
```

```
##   Mnemonic devices            48.98           51.02
##   Nonwritten records          23.81           76.19
##   True writing                13.95           86.05
```

```
# Rank correlation
cor(as.numeric(d$writing),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2497125
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.0002826
## alternative hypothesis: two.sided
```

Other measures

Money

```
tx = table(d$money,d$monster_present2)
tx
```

```
##
##                                Monster Absent Monster Present
##   None                                     36              40
##   Domestically usable particles             6               8
##   Alien currency                           15              27
##   Elementary forms                         10              17
##   True money                                2              23
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                                Monster Absent Monster Present
##   None                                  47.37           52.63
##   Domestically usable particles         42.86           57.14
##   Alien currency                        35.71           64.29
##   Elementary forms                      37.04           62.96
##   True money                             8.00           92.00
```

```
# Rank correlation
cor(as.numeric(d$money),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2055975
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.006705
## alternative hypothesis: two.sided
```

Agriculture

```
tx = table(d$ag,d$monster_present2)
tx
```

```
##
##                           Monster Absent Monster Present
##   None                                18              19
##   10% food supply                      9               8
##   10 %; secondary                      4               6
##   Primary; not intensive              26              37
##   Primary; intensive                  12              45
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                           Monster Absent Monster Present
##   None                             48.65           51.35
##   10% food supply                  52.94           47.06
##   10 %; secondary                  40.00           60.00
##   Primary; not intensive           41.27           58.73
##   Primary; intensive               21.05           78.95
```

```
# Rank correlation
cor(as.numeric(d$ag),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.2054279
```

```
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.02321
## alternative hypothesis: two.sided
```

Population density

```
tx = table(d$popdens,d$monster_present2)
tx
```

```
##
##                              Monster Absent Monster Present
##   < 1 person / sq. mile                  31              26
##   1-5 persons / sq. mile                  7              17
##   5.1-25 persons/ sq. mile               12              16
##   26-100 persons  / sq. mile              8              27
##   100 persons / sq. mile                 11              29
```

```
round(100*prop.table(tx,1),2)
```

```
##
##                              Monster Absent Monster Present
##   < 1 person / sq. mile                54.39           45.61
##   1-5 persons / sq. mile               29.17           70.83
##   5.1-25 persons/ sq. mile             42.86           57.14
##   26-100 persons  / sq. mile           22.86           77.14
##   100 persons / sq. mile               27.50           72.50
```

```r
# Rank correlation
cor(as.numeric(d$popdens),
    as.numeric(d$monster_present2),method="kendall")
```

```
## [1] 0.20032
```

```r
fisher.test(tx)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tx
## p-value = 0.01286
## alternative hypothesis: two.sided
```

Combined model

Various measures above co-occur with the presence of composite monsters. However, many sociocultural variables correlate with each other, making it unclear which variables are directly associated. How might we determine which of these variables are most important when accounting for the presence of composite beings in our global ethnographic sample?

Below, we use a machine learning method (decision trees and random forests) to identify the most efficient combination of variables to predict the presence of composite monsters. We then entered these variables into a predictive regression model to test significance.

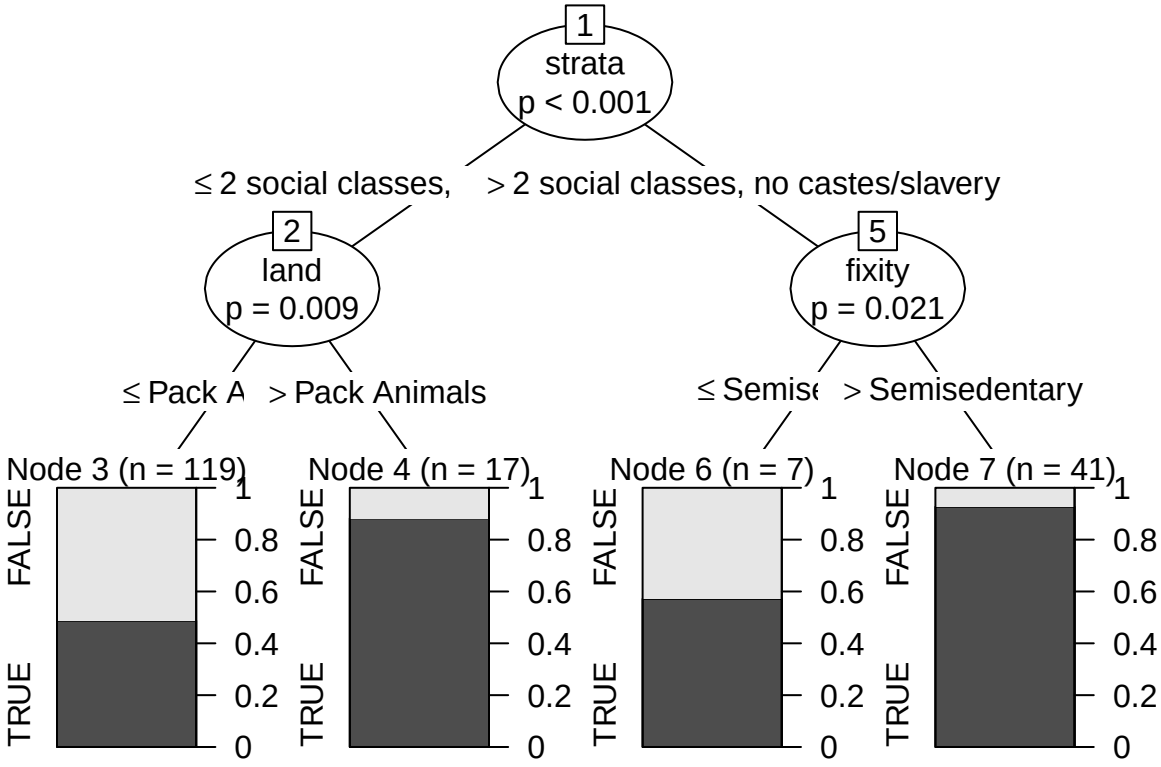
Decision trees are a computational method of making predictions by dividing the data into sub-sets (see Strobl, Malley & Tutz, 2009). The algorithm works out the most efficient series of binary questions to ask about a set of independent variables in order to make a guess about the dependent value of a data point. The method is robust to correlations between independent variables and to small sub-sample sizes, and it can detect interactions and non-linearities in the data.

A single tree is the most efficient set of questions for the given dataset. However, small differences in the sample could lead to very different trees. One way of evaluating the relative importance of variables is to calculate a large number of decision trees using random sub-samples of data and independent variables. Each variable receives an "importance" score based on how many trees it is selected for and how high in the tree the variable is placed.

For use of decision trees in social science, see Roberts, Torreira & Levinson (2015).

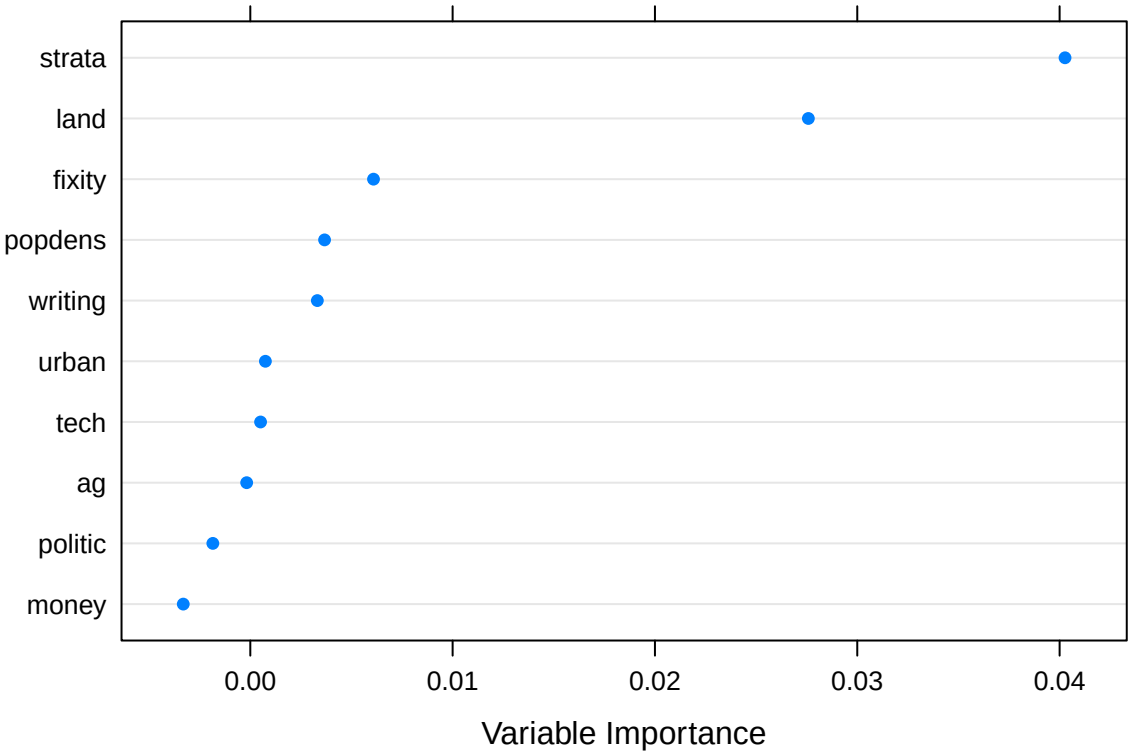
Decision tree for selecting variables

```
decisionTree = ctree(factor(monster_present) ~
                         strata +
                         urban + ag + popdens + fixity +
                         land + money + politic +
                         tech + writing, data = d)
plot(decisionTree)
```



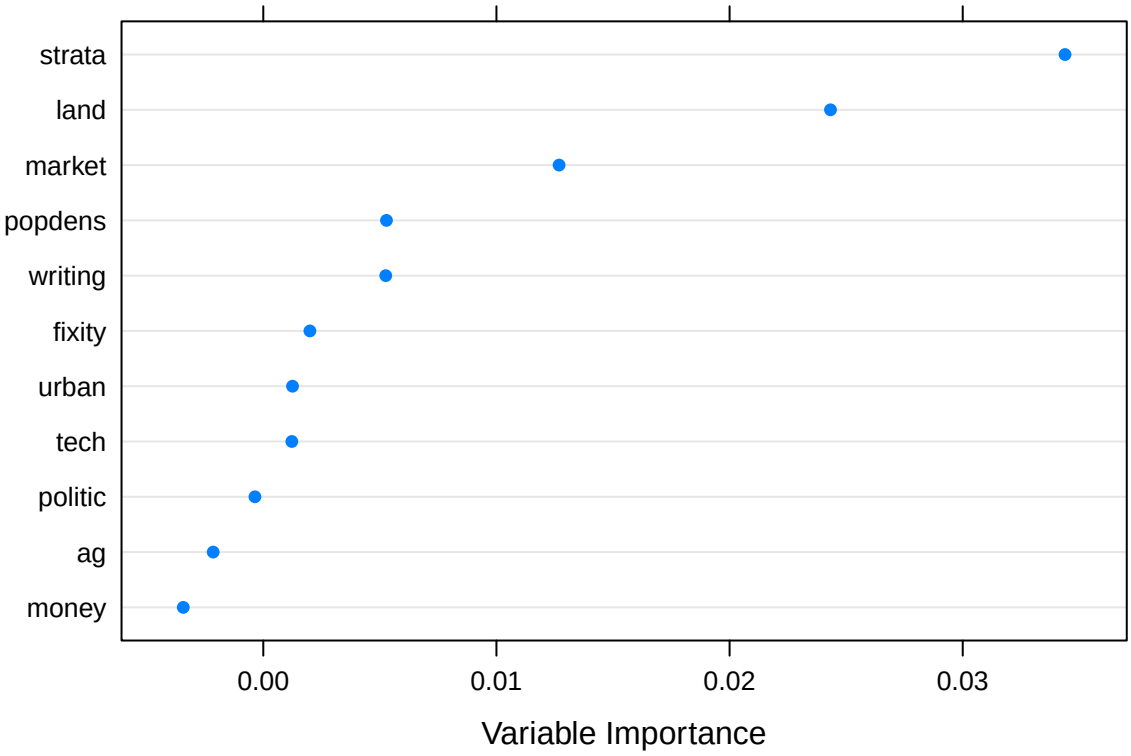
Now we calculate the importance scores with a random forest. The variable 'market' has many missing values, limiting the number of complete observations. So the first forest is run without this variable:

```
randomForest = cforest(factor(monster_present) ~
                         strata +
                         urban + ag + popdens + fixity +
                         land + money + politic +
                         tech + writing, data = d)
importance = varimp(randomForest)
dotplot(sort(importance),xlab="Variable Importance")
```



Random forest including market:

```
randomForest = cforest(factor(monster_present) ~
                         market +
                         strata +
                         urban + ag + popdens + fixity +
                         land + money + politic +
                         tech + writing, data = d)
importance = varimp(randomForest)
dotplot(sort(importance),xlab="Variable Importance")
```



The "market" variable seems relatively important, though strata and land are still the most important. Since including 'market' limits the data sample size, we will not include it in the combined model.

Predictive model

Below we fit a binomial regression model, predicting whether a monster is present or absent. The independent variables are added to the model one by one in the order identified by the random forests procedure. In order to obtain a relatively simple model that is not over-fitted, we stop when adding variables does not significantly increase the fit of the model to the data. This is evaluated using a likelihood ratio test.

For the ordered factors, use (reverse) Helmert coding, so each coefficient compares the current level to the previous level of the variable.

```
contrasts(d$strata) = contr.helmert(5)
contrasts(d$land) = contr.helmert(4)
contrasts(d$fixity) = contr.helmert(5)
```

Build a series of models, each adding a variable. Start with the baseline model, using just the log number of pages of documentation:

```
m0 = glm(factor(monster_present) ~ 1 + HRAFNumPages.log,
         family="binomial",data=d)
summary(m0)
```

```
##
## Call:
## glm(formula = factor(monster_present) ~ 1 + HRAFNumPages.log,
##     family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9277  -1.1566   0.6247   0.8994   1.9705
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -7.026      1.623  -4.330 1.49e-05 ***
## HRAFNumPages.log     1.051      0.227   4.632 3.63e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 243.46  on 183  degrees of freedom
## Residual deviance: 216.10  on 182  degrees of freedom
## AIC: 220.1
##
## Number of Fisher Scoring iterations: 4
```

Add the cultural variables in the order suggested by the random forests analysis:

```
m1 = update(m0, ~.+strata)
m2 = update(m1, ~.+land)
m3 = update(m2, ~.+fixity)
```

Use a likelihood ratio test to test whether adding each variable increases the fit of the model

```
lrtest(m0,m1,m2,m3)
```

```
## Likelihood ratio test
##
## Model 1: factor(monster_present) ~ 1 + HRAFNumPages.log
```

```
## Model 2: factor(monster_present) ~ HRAFNumPages.log + strata
## Model 3: factor(monster_present) ~ HRAFNumPages.log + strata + land
## Model 4: factor(monster_present) ~ HRAFNumPages.log + strata + land +
##     fixity
##   #Df   LogLik Df   Chisq Pr(>Chisq)
## 1   2 -108.048
## 2   6  -98.776  4 18.5446  0.0009655 ***
## 3   9  -91.645  3 14.2632  0.0025680 **
## 4  13  -88.743  4  5.8037  0.2142917
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including `strata` and `land` improve the model, but `fixity` does not.

Test if the interaction between strata and land is significant:

```
m4 = update(m2, ~.+ strata:land)
lrtest(m2,m4)
```

```
## Likelihood ratio test
##
## Model 1: factor(monster_present) ~ HRAFNumPages.log + strata + land
## Model 2: factor(monster_present) ~ HRAFNumPages.log + strata + land +
##     strata:land
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   9 -91.645
## 2  20 -89.471 11 4.3471     0.9586
```

No, so stick with model 2 (strata and land). The summary below shows how each level of strata and land change the estimate of the likelihood of monsters being present. Each estimate compares the current level to the prior level. It appears as if the critical threshold for strata is between Egalitarian and other societies. For `land`, the critical threshold is between those with only human or pack animal transport, and other societies. These results reflect the decision tree.

```
summary(m2)
```

```
##
## Call:
## glm(formula = factor(monster_present) ~ HRAFNumPages.log + strata +
##     land, family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2020  -0.9245   0.2582   0.8527   2.2348
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.00702  320.70264  -0.003 0.997495
## HRAFNumPages.log   0.88462    0.25915   3.414 0.000641 ***
## strata1            0.40201    0.21616   1.860 0.062916 .
## strata2            0.05579    0.18875   0.296 0.767572
## strata3            0.29020    0.15977   1.816 0.069322 .
## strata4            0.19904    0.17218   1.156 0.247684
## land1              0.14775    0.22529   0.656 0.511953
## land2              0.62914    0.28226   2.229 0.025820 *
## land3              4.08881  320.69720   0.013 0.989827
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 243.46  on 183  degrees of freedom
## Residual deviance: 183.29  on 175  degrees of freedom
## AIC: 201.29
##
## Number of Fisher Scoring iterations: 17
```

F-test to confirm the claims above for each variable:

```
car::Anova(m2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: factor(monster_present)
##                   LR Chisq Df Pr(>Chisq)
## HRAFNumPages.log   13.882  1  0.0001947 ***
## strata              9.746  4  0.0449313 *
## land               14.263  3  0.0025680 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pseudo R-squared for model:

```
MuMIn::r.squaredLR(m2)
```

```
## [1] 0.2789094
## attr(,"adj.r.squared")
## [1] 0.3801417
```

Plotting the model results

We want the plot to reflect the confidence intervals around the estimates.

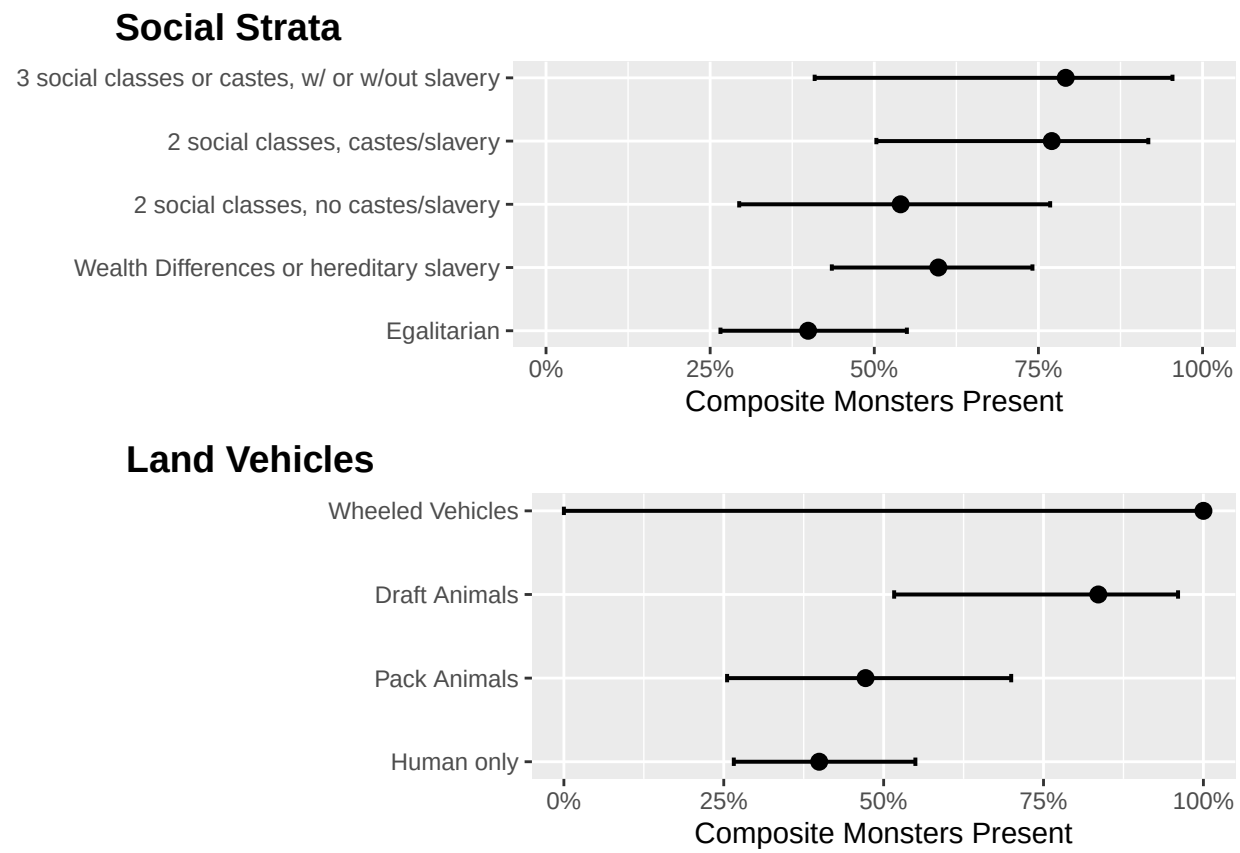
Plot the model:

```
plt = plot_model(m2,"pred")
```

```
## Data were 'prettified'. Consider using `terms="HRAFNumPages.log [all]"` to get smooth plots.
```

```
p1 = plt[[2]] + coord_flip(ylim=c(0,1)) +
  ggtitle("")+
  theme(axis.title.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  ylab("Composite Monsters Present")
p2 = plt[[3]] + coord_flip(ylim=c(0,1)) +
  ggtitle("")+
  theme(axis.title.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  ylab("Composite Monsters Present")

bigplot = ggarrange(p1,
          ggarrange(ggplot() + theme_void(),
                    p2,nrow=1,widths = c(1,3),
                    labels=c("","")),
          ncol = 1, labels=c("Social Strata","Land Vehicles"))
```

```
bigplot
```




```
pdf("../results/BigPlot.pdf",width=6,height=4)
bigplot
dev.off()
```

```
## pdf
##   2
```

Below we fit a mixed effects model with random effects for language family. This strategy is used to control for the historical relatedness of societies. However, the result is a fit that essentially ignores the random effects and is almost identical to the fixed-effects model above. In other words, the relatedness of societies does not seem to affect the results.

```
mMF = glmer(factor(monster_present) ~  1 + HRAFNumPages.log + strata + land +
              (1 | language_family),
            family="binomial", data = d)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## Model failed to converge: degenerate Hessian with 1 negative eigenvalues
```

```
summary(mMF)
```

```
## Warning in vcov.merMod(object, use.hessian = use.hessian): variance-covariance matrix computed from
## not positive definite or contains NA values: falling back to var-cov estimated from RX
```

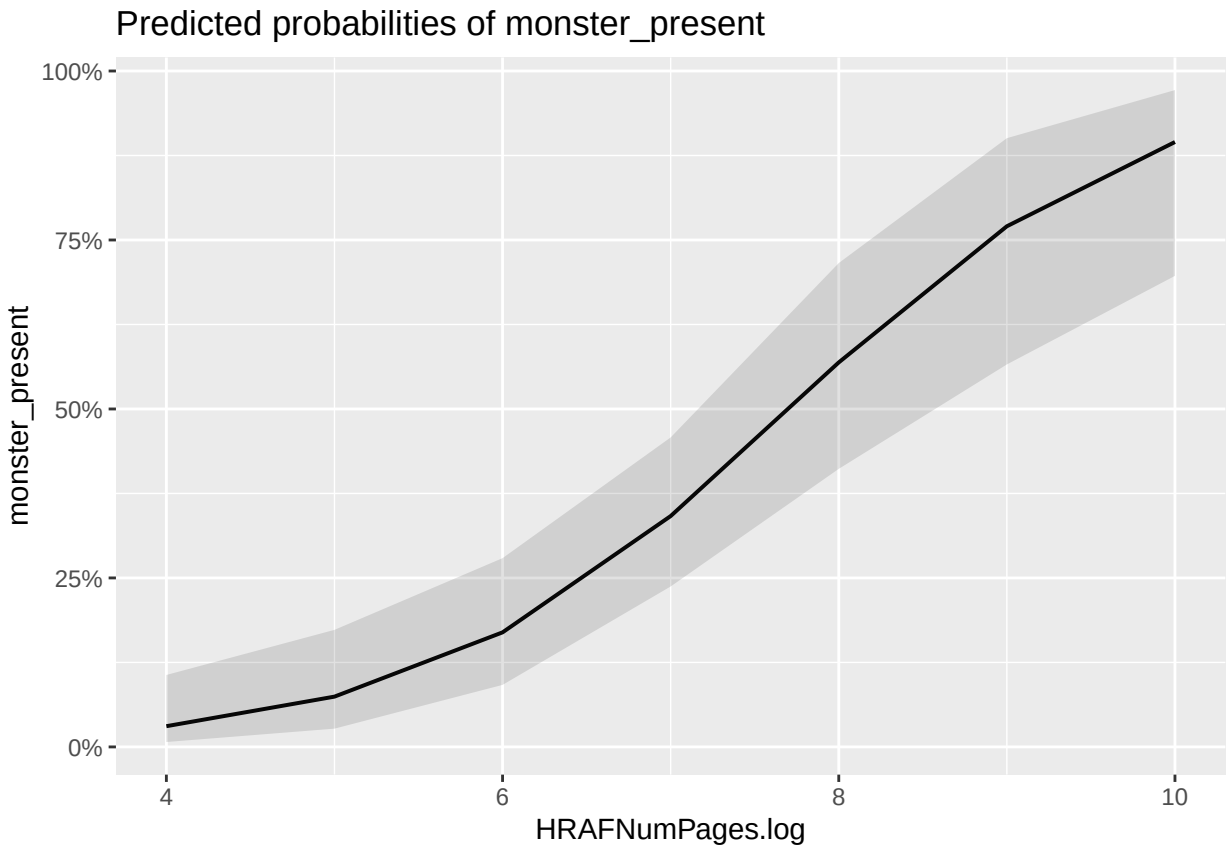
```
## Warning in vcov.merMod(object, correlation = correlation, sigm = sig): variance-covariance matrix co
## not positive definite or contains NA values: falling back to var-cov estimated from RX

## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: factor(monster_present) ~ 1 + HRAFNumPages.log + strata + land +
##     (1 | language_family)
##    Data: d
##
##      AIC      BIC   logLik deviance df.resid
##    203.3    235.4    -91.6    183.3      174
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.1293 -0.7269  0.1809  0.6543  3.3084
##
## Random effects:
##  Groups          Name        Variance Std.Dev.
##  language_family (Intercept) 0.02674  0.1635
## Number of obs: 184, groups:  language_family, 81
##
## Fixed effects:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.02089  554.96872  -0.002 0.998532
## HRAFNumPages.log   0.89145    0.25991   3.430 0.000604 ***
## strata1            0.40540    0.21700   1.868 0.061738 .
## strata2            0.05654    0.19084   0.296 0.767018
## strata3            0.29159    0.16010   1.821 0.068563 .
## strata4            0.19997    0.17287   1.157 0.247365
## land1              0.14809    0.22707   0.652 0.514289
## land2              0.63508    0.28417   2.235 0.025425 *
## land3              4.10854  554.96556   0.007 0.994093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) HRAFNP strat1 strat2 strat3 strat4 land1  land2
## HRAFNmPgs.l -0.003
## strata1      0.000 -0.107
## strata2      0.001 -0.140 -0.001
## strata3      0.000  0.090 -0.015 -0.178
## strata4      0.000 -0.083  0.041 -0.086 -0.086
## land1        0.000  0.126 -0.133 -0.007 -0.030 -0.311
## land2        0.000 -0.027  0.208  0.071  0.048  0.088 -0.171
## land3        1.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## unable to evaluate scaled gradient
## Model failed to converge: degenerate  Hessian with 1 negative eigenvalues
```

Bayesian estimation

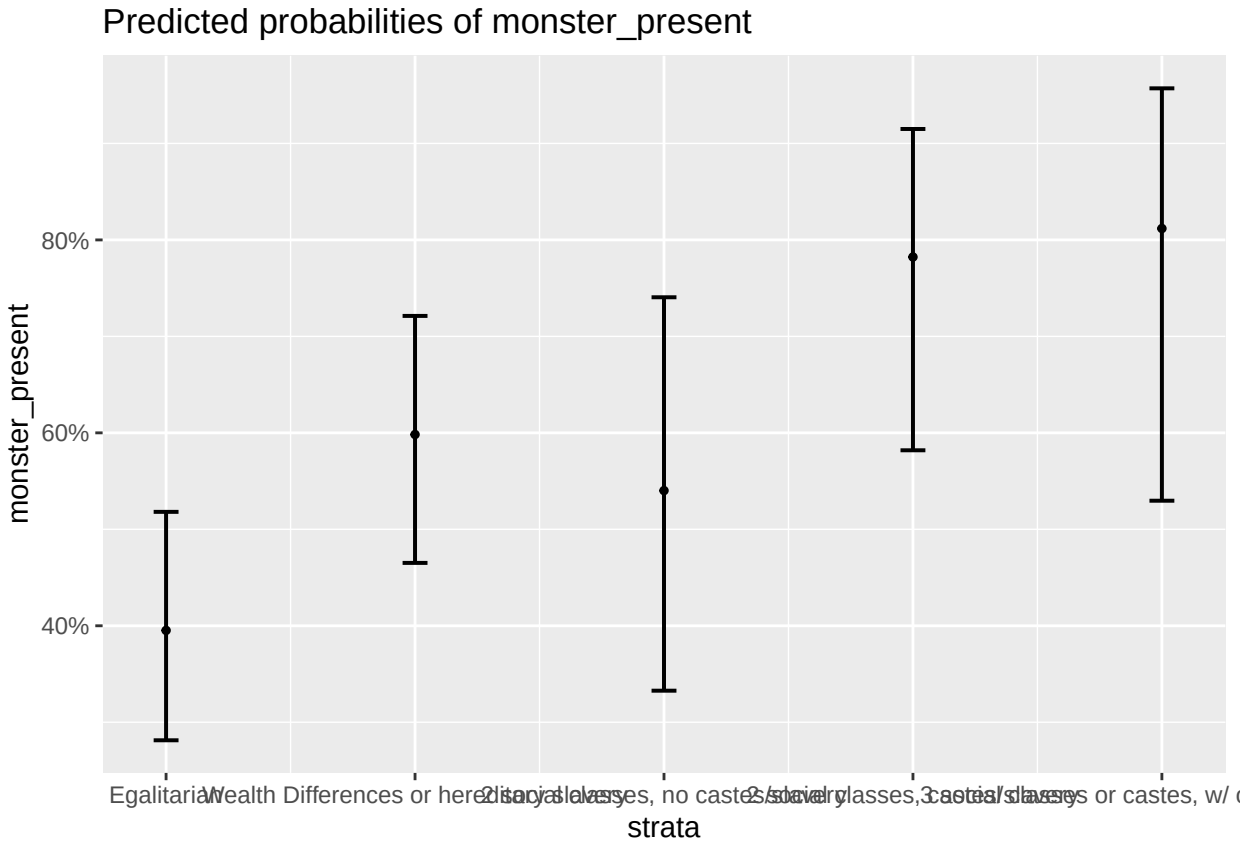
The estimation above has a large uncertainty for the probability of composite monsters for societies with wheeled vehicles. This is despite 100% of societies with wheeled vehicles having composite monsters. To test whether this is due a to poor model convergence, we use a different framework for estimating the model parameters. Below we use a Bayesian estimation using *brms*, using unifnromative priors. We demonstrate that the coefficient estimates are almost identical, but that the estimate for the confidence interval for societies with wheeled vehicles is much narrower, which fits better with the overall data.

```
set.seed(2189)
warmupN = 1000
mB = brm(as.numeric(monster_present) ~
           HRAFNumPages.log +
           strata+land,warmup = warmupN,
         iter = 100000,cores = 4,chains=8,
         family="bernoulli",data=d)
plot_model(mB,"pred")
```

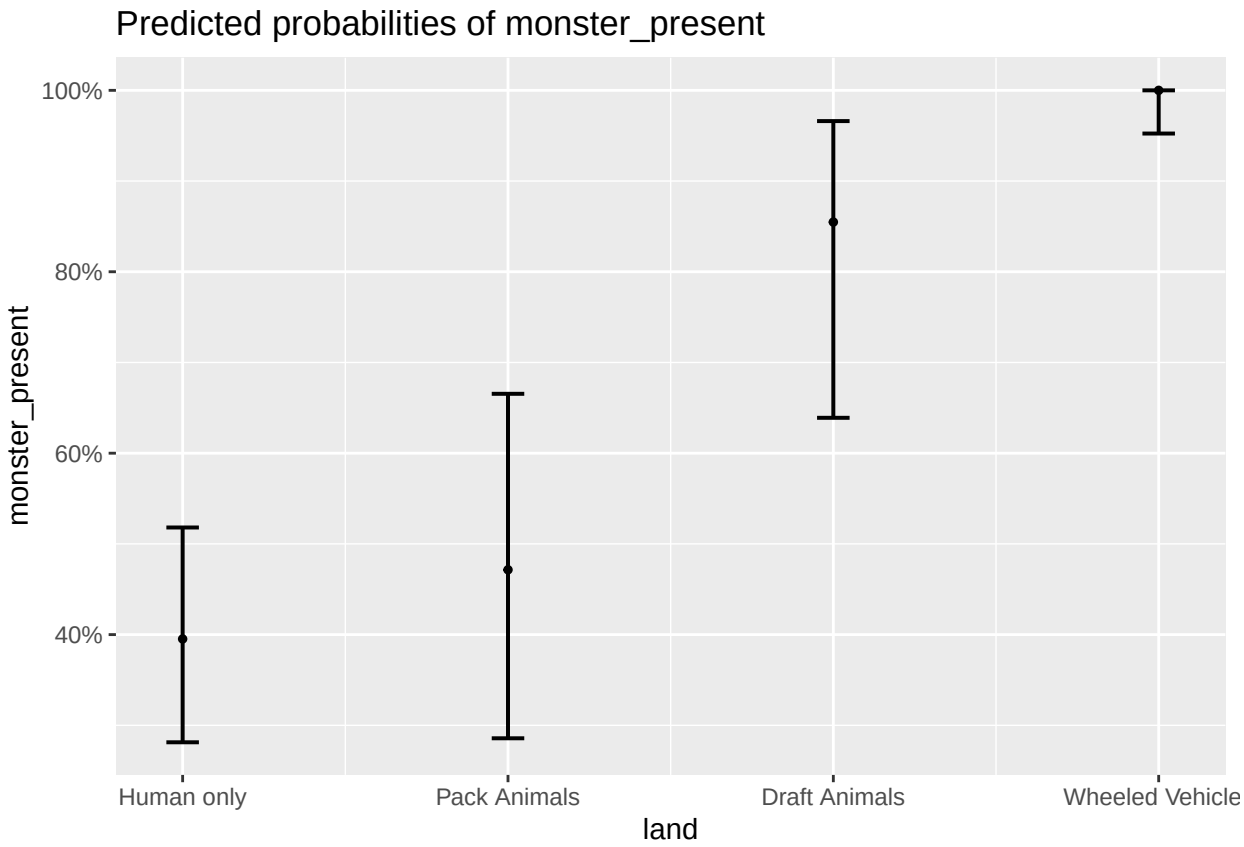
```
## $HRAFNumPages.log
```




```
##
## $strata
```


Predicted probabilities of monster_present
monster_present
80%
60%
40%
Egalitarian
Wealth Differences or hereditary classes, no castes/slavery
2 social classes, castes/slavery
3 social classes or castes, w/
strata

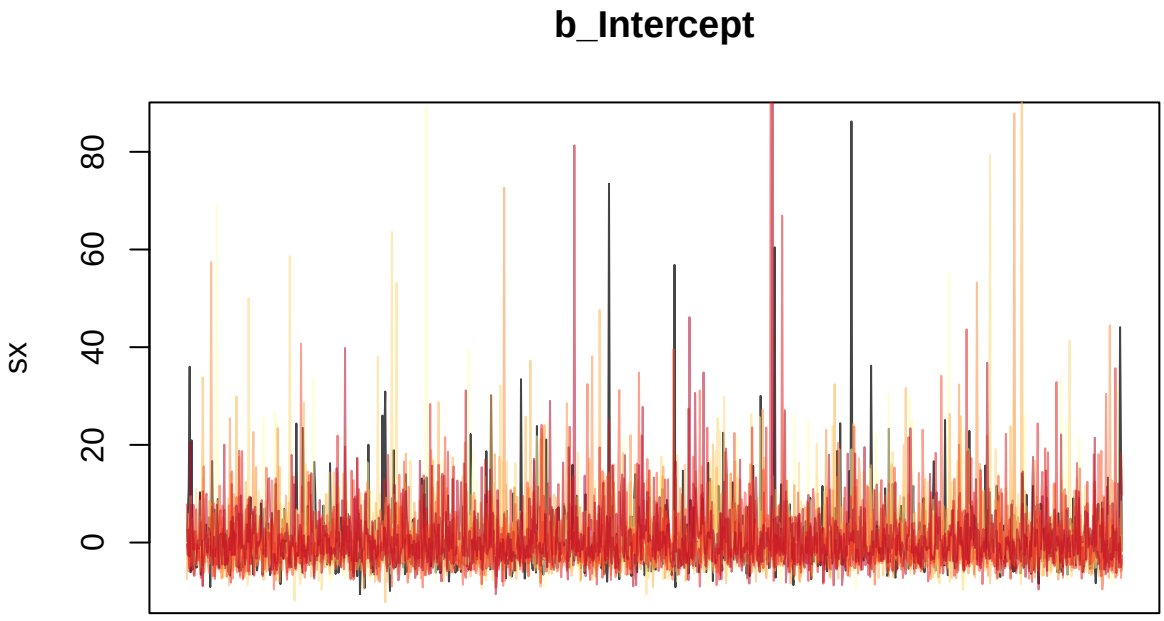
```
##
## $land
```

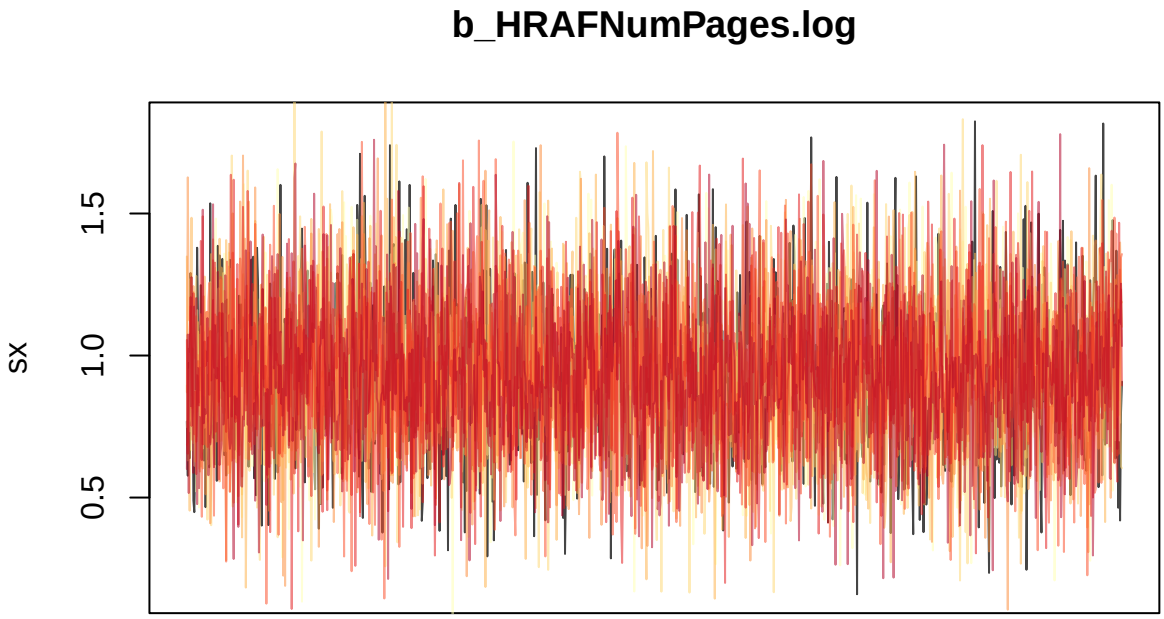


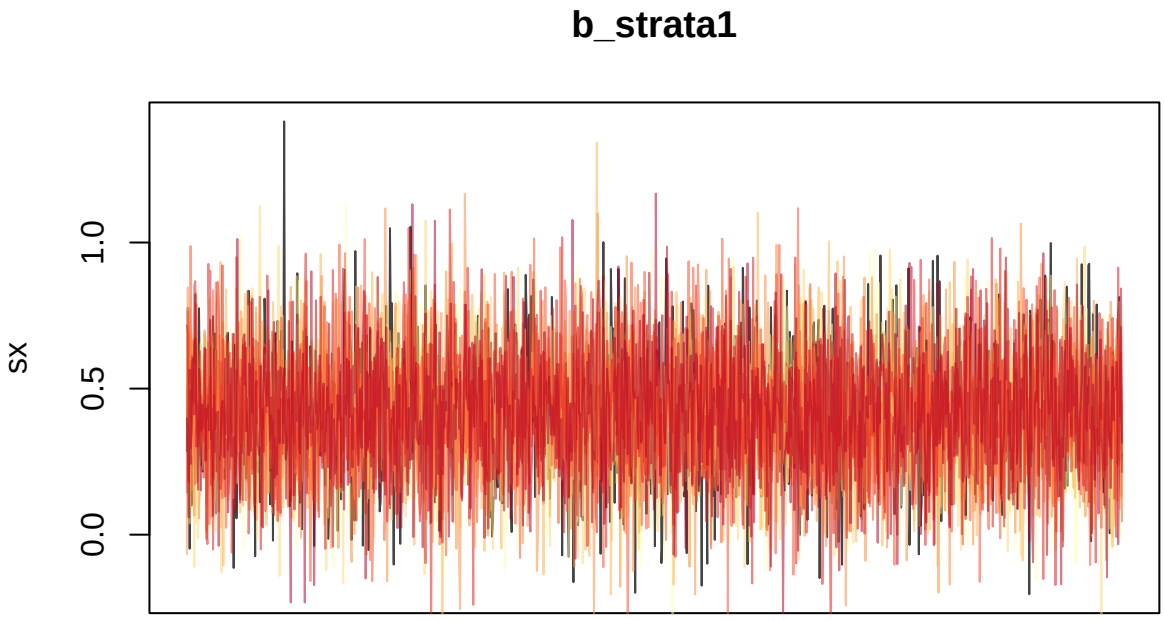

```
summary(mB)
```

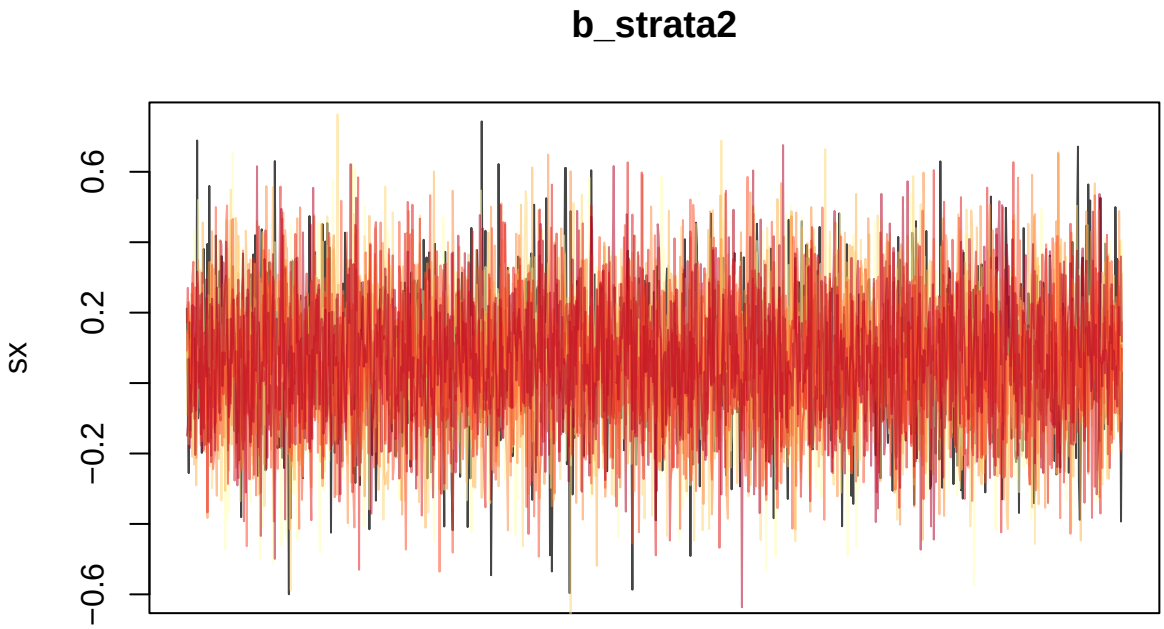
```
##  Family: bernoulli
##   Links: mu = logit
## Formula: as.numeric(monster_present) ~ HRAFNumPages.log + strata + land
##    Data: d (Number of observations: 184)
##   Draws: 8 chains, each with iter = 1e+05; warmup = 1000; thin = 1;
##          total post-warmup draws = 792000
##
## Population-Level Effects:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           0.59      7.63    -7.05    17.92 1.00   136274    77692
## HRAFNumPages.log    0.94      0.27     0.44     1.49 1.00   653689   527827
## strata1             0.41      0.22    -0.01     0.85 1.00   674775   553332
## strata2             0.06      0.20    -0.32     0.45 1.00   652492   536109
## strata3             0.32      0.17     0.00     0.66 1.00   672869   500784
## strata4             0.23      0.19    -0.10     0.64 1.00   601915   450238
## land1               0.16      0.23    -0.30     0.61 1.00   614183   554478
## land2               0.71      0.31     0.16     1.38 1.00   633173   444496
## land3               5.95      7.38     0.38    23.02 1.00   135096    75405
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

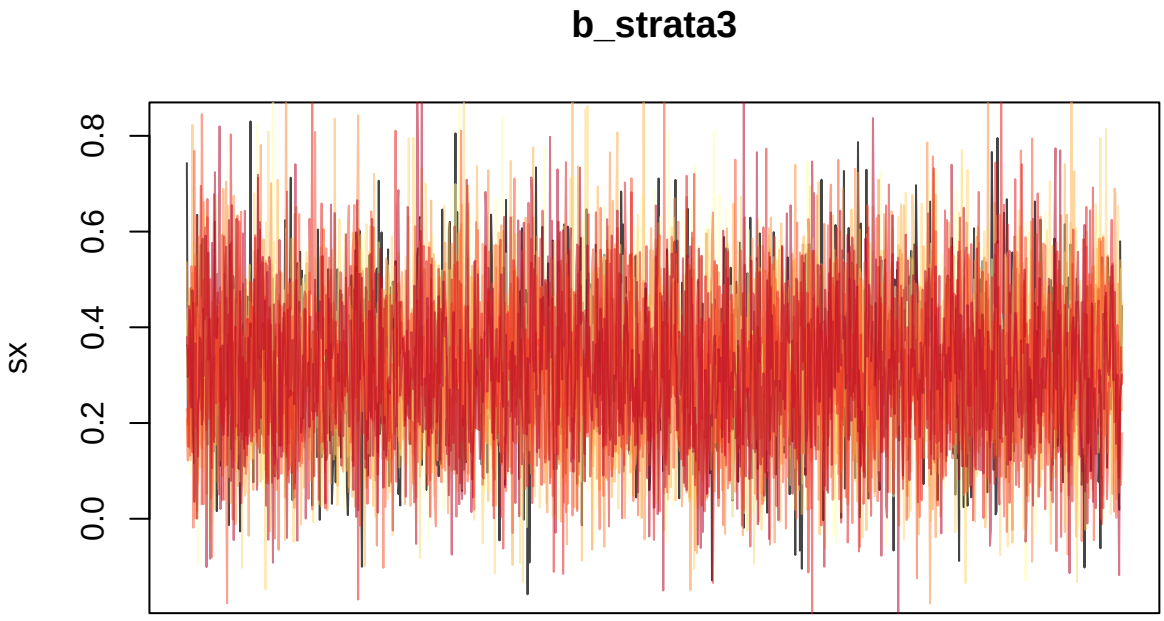
Plot model convergence:

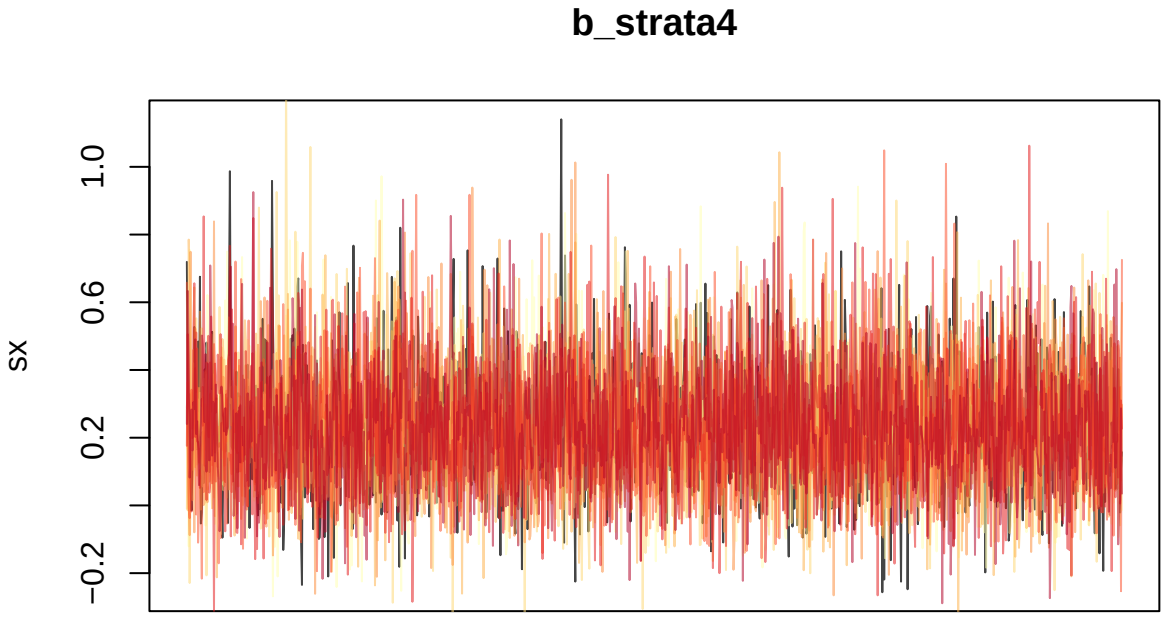

b_Intercept
sx
0
20
40
60
80

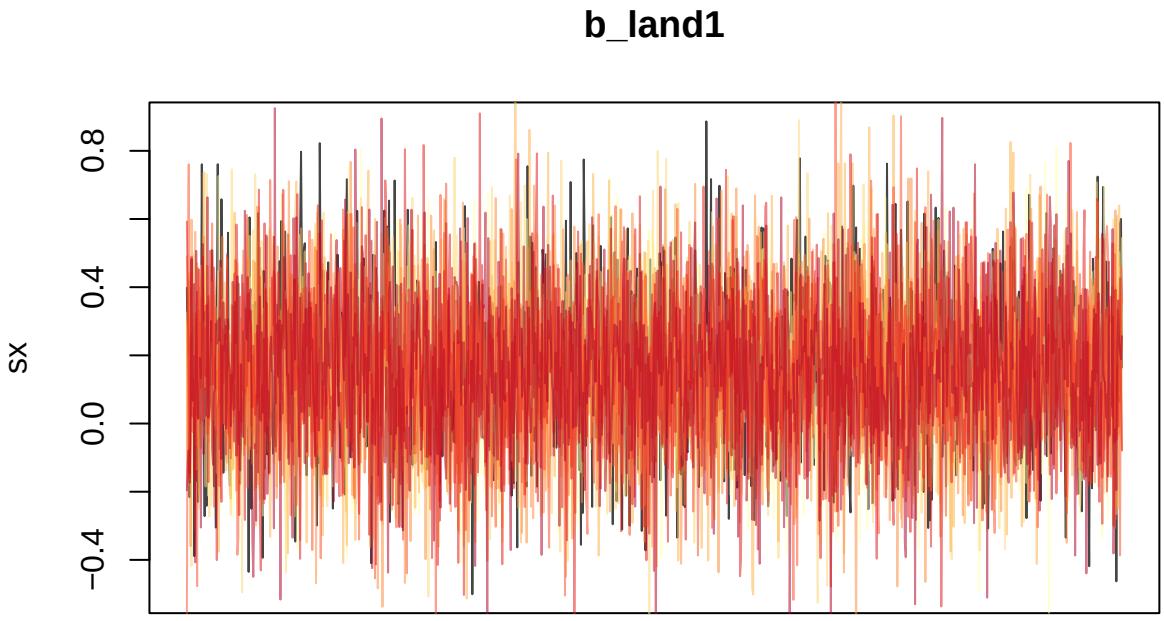

b_HRAFNumPages.log
sx
0.5
1.0
1.5

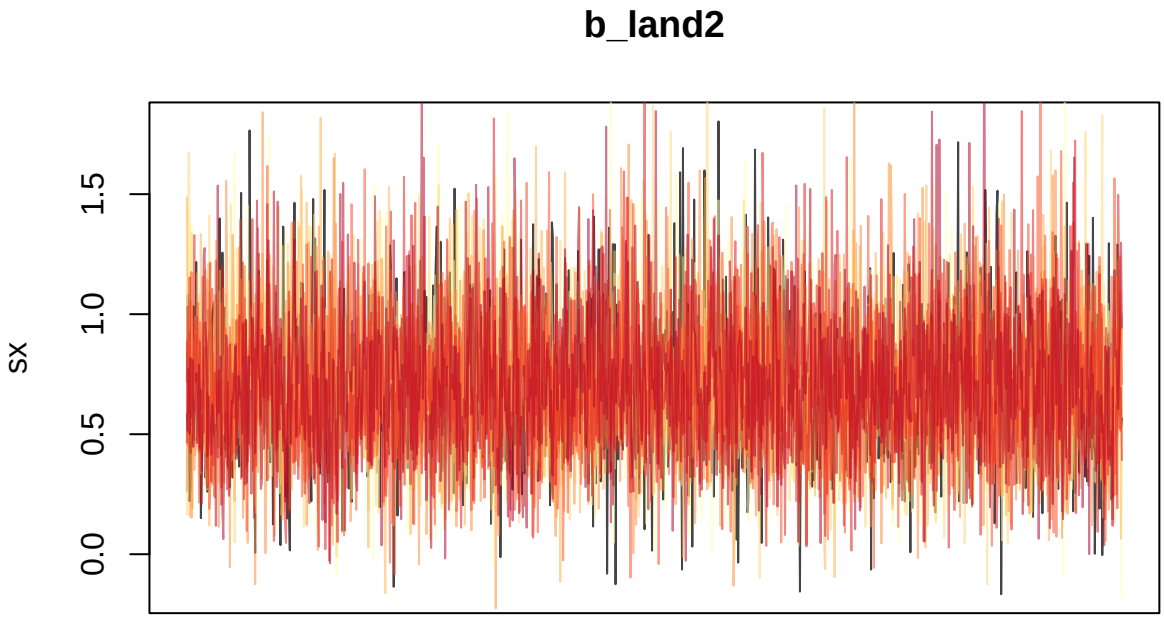

b_strata1
sx
0.0
0.5
1.0

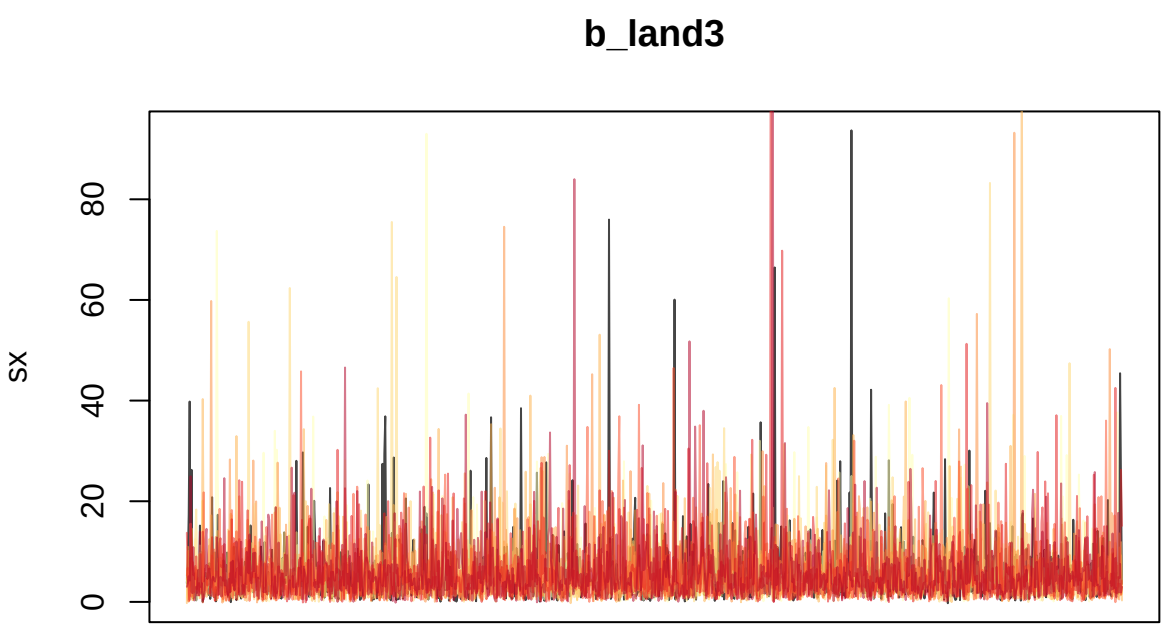

b_strata2
sx
0.6
0.2
−0.2
−0.6


b_strata3
sx
0.0
0.2
0.4
0.6
0.8


b_strata4
sx
1.0
0.6
0.2
−0.2


b_land1
sx
0.8
0.4
0.0
−0.4


b_land2
sx
0.0
0.5
1.0
1.5

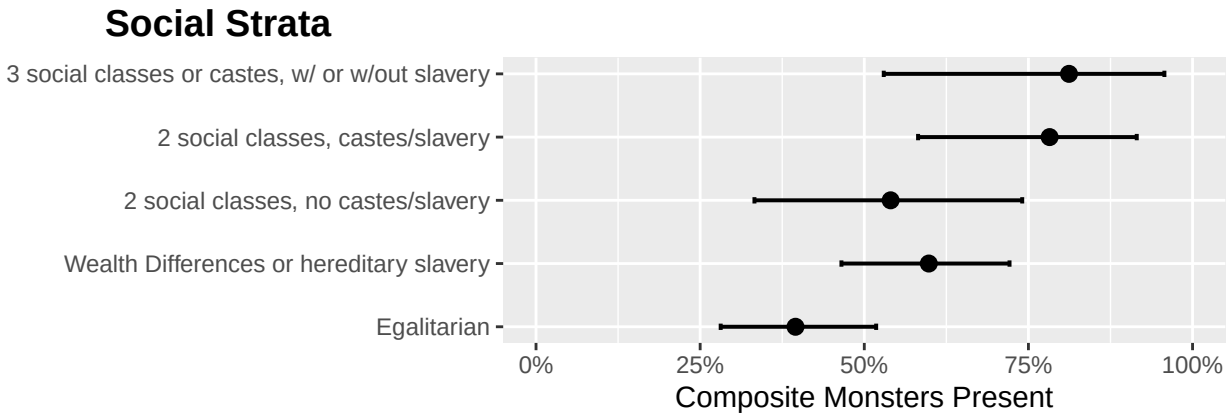


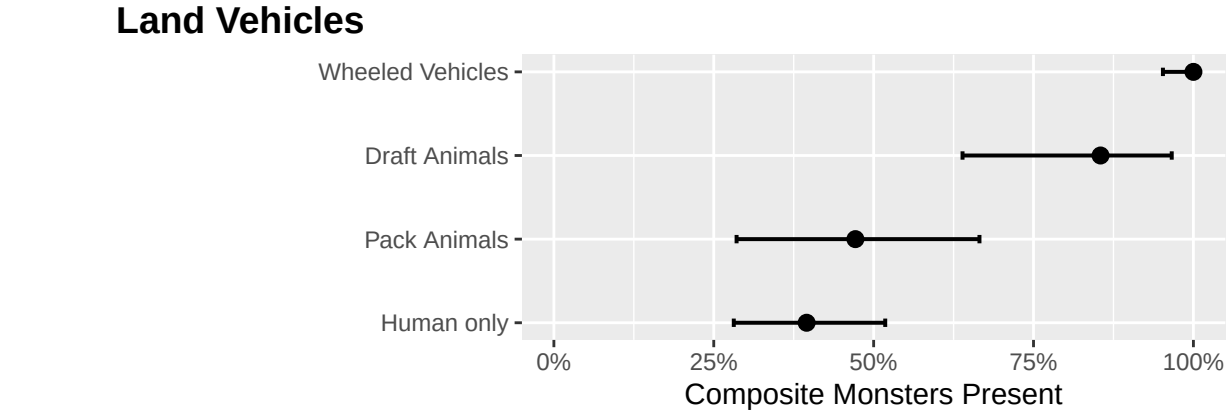
Plot the results:

```
pltB = plot_model(mB,"pred")
```

```
## Data were 'prettified'. Consider using `terms="HRAFNumPages.log [all]"` to get smooth plots.
```

```
p1B = pltB[[2]] + coord_flip(ylim=c(0,1)) +
  ggtitle("")+
  geom_point(size=2.5)+
  theme(axis.title.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  ylab("Composite Monsters Present")
p2B = pltB[[3]] + coord_flip(ylim=c(0,1)) +
  ggtitle("")+
  geom_point(size=2.5)+
  theme(axis.title.y = element_blank(),
        panel.grid.minor.y = element_blank()) +
  ylab("Composite Monsters Present")

bigplotB = ggarrange(p1B,
          ggarrange(ggplot() + theme_void(),
                    p2B,nrow=1,widths = c(1,3),
                    labels=c("","")),
          ncol = 1, labels=c("Social Strata","Land Vehicles"))
bigplotB
```



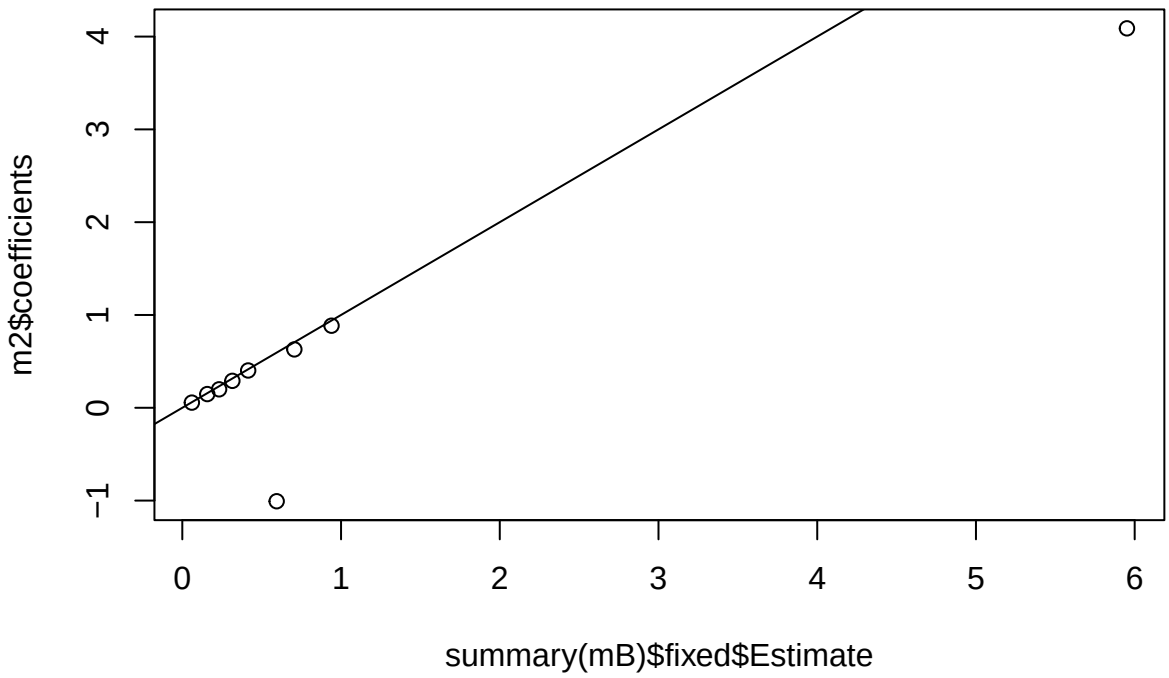



```
pdf("../results/BigPlot_Bayesian.pdf",width=6,height=4)
bigplotB
dev.off()
```

```
## pdf
##   2
```

Compare estimates between the standard and Bayesian models, the major difference is for wheeled vehicles:

```
plot(summary(mB)$fixed$Estimate,
     m2$coefficients)
abline(0,1)
```




```
cor(summary(mB)$fixed$Estimate,
     m2$coefficients)
```

```
## [1] 0.9325302
```

Unclear cases

In nine cases, possible composite beings were identified, but either the available descriptions were too vague or else the sources were an insufficient match to the SCCS time period or community. In the analyses above, these cases were assumed to have composite beings. Here we test whether that assumption is affecting the results.

Since the number of unclear cases is small, we can run the test with all possible combinations of presence and absence (there are 512 possible combinations), re-run the test, and see if the result changes. Note that controls for multiple comparisons are not needed: we're not claiming that any one significant result proves the hypothesis: we're expecting that ALL tests below should be significant in order for the hypothesis to be robust.

```
# Order by unclear so unclear cases are at the top
d = d[order(d$monster_unclear,decreasing = T),]
# Generate all possible combinations of presence/absence for
# the nine cases
possibleCombinationsOfUnclearCases =
  expand.grid(c(T,F),c(T,F),c(T,F),c(T,F),c(T,F),c(T,F),c(T,F),c(T,F),c(T,F))

testUnclear = function(varToTest){
    print(paste("*",varToTest,"*"))
    fx1 = fisher.test(
      table(d[!d$monster_unclear,]$monster_present,
            d[!d$monster_unclear,varToTest]))
    print("Excludnig unclear cases:")
    print(fx1$p.value)
    # For each possible combination ...
    rangeOfFisherPValues =
    apply(possibleCombinationsOfUnclearCases,1,
        function(X){
          # ... replace the nine cases with the possible combination
          mPresent = d$monster_present
          mPresent[1:9] = X
          # Run the fisher test again
          fx = fisher.test(table(mPresent,d[,varToTest]))
          fx$p.value
        })
    # Return the range of values
    print("Range of possible values:")
    range(rangeOfFisherPValues)
}

testUnclear("strata")
```

```
## [1] "* strata *"
## [1] "Excludnig unclear cases:"
## [1] 8.634055e-06
## [1] "Range of possible values:"
## [1] 3.930482e-07 5.713377e-04
```

```
testUnclear("land")
```

```
## [1] "* land *"
## [1] "Excludnig unclear cases:"
## [1] 1.108596e-05
```

```
## [1] "Range of possible values:"
## [1] 1.543862e-06 4.087067e-05
testUnclear("urban")
## [1] "* urban *"
## [1] "Excludnig unclear cases:"
## [1] 0.001089126
## [1] "Range of possible values:"
## [1] 0.0001685333 0.0034510031
testUnclear("politic")
## [1] "* politic *"
## [1] "Excludnig unclear cases:"
## [1] 0.007202879
## [1] "Range of possible values:"
## [1] 0.001197838 0.020952802
testUnclear("tech")
## [1] "* tech *"
## [1] "Excludnig unclear cases:"
## [1] 0.0009060922
## [1] "Range of possible values:"
## [1] 0.0000791471 0.0092628589
testUnclear("writing")
## [1] "* writing *"
## [1] "Excludnig unclear cases:"
## [1] 4.373475e-05
## [1] "Range of possible values:"
## [1] 3.461887e-06 1.627860e-03
```

All main tests are still significant when excluding the 9 unclear cases. In addition, for all main variables, there is no combination of the 9 unclear cases that increases the p-value above 0.05. This suggests the main qualitative conclusions are not affected by the unclear cases.

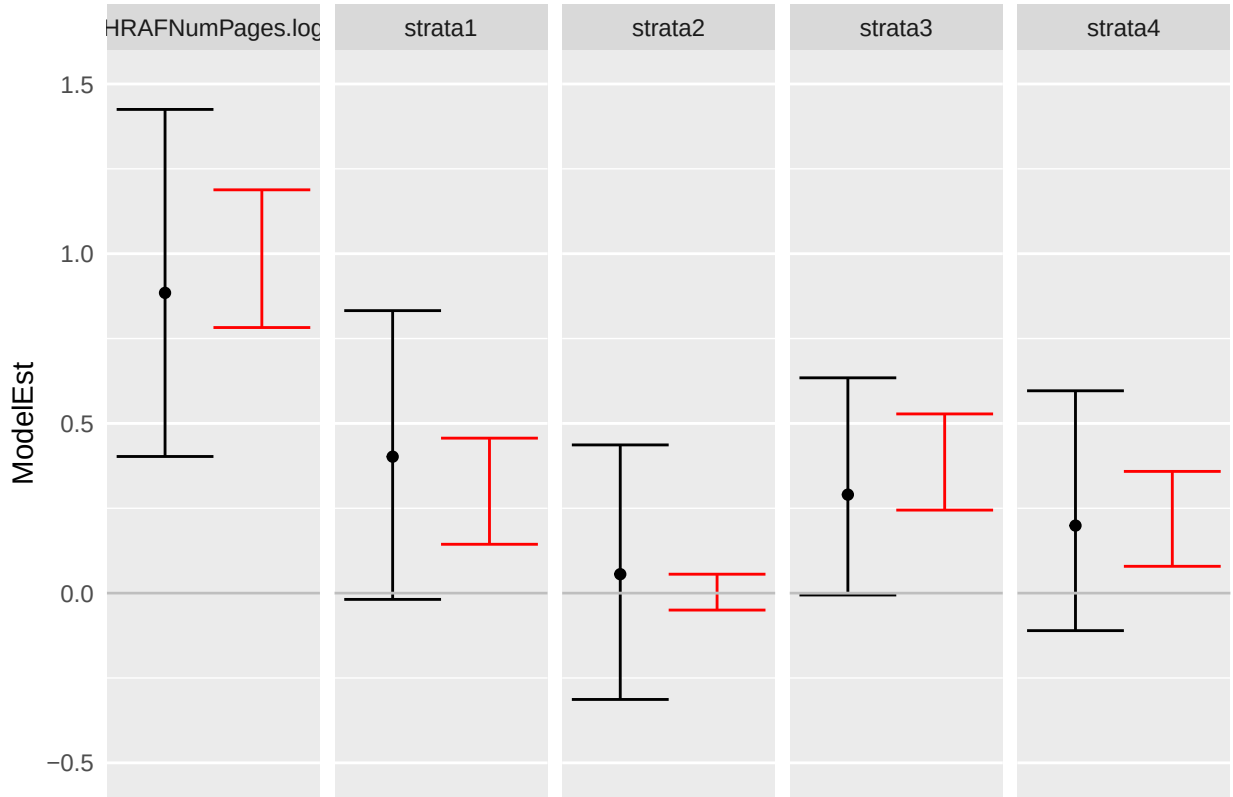
In the code below, we re-run the combined model (m2) with all possible combinations of the uncertain data.

```
# For each possible combination ...
rangeOfValues =
apply(possibleCombinationsOfUnclearCases,1,
    function(X){
      # ... replace the nine cases with the possible combination
      mPresent = d$monster_present
      mPresent[1:9] = X
      # Run the fisher test again
      m2x = glm(factor(mPresent) ~
                  HRAFNumPages.log + strata+land, family="binomial",data=d)
      return(m2x$coefficients)
    })
coefRanges = cbind(
  t(apply(rangeOfValues,1,range)),
  coef(m2), confint(m2))
```

```
coefRanges = as.data.frame(coefRanges)
names(coefRanges) = c("unclearLow","unclearHigh","ModelEst",
                      "ModelConfLow","ModelConfHigh")
coefRanges$Var = rownames(coefRanges)
```

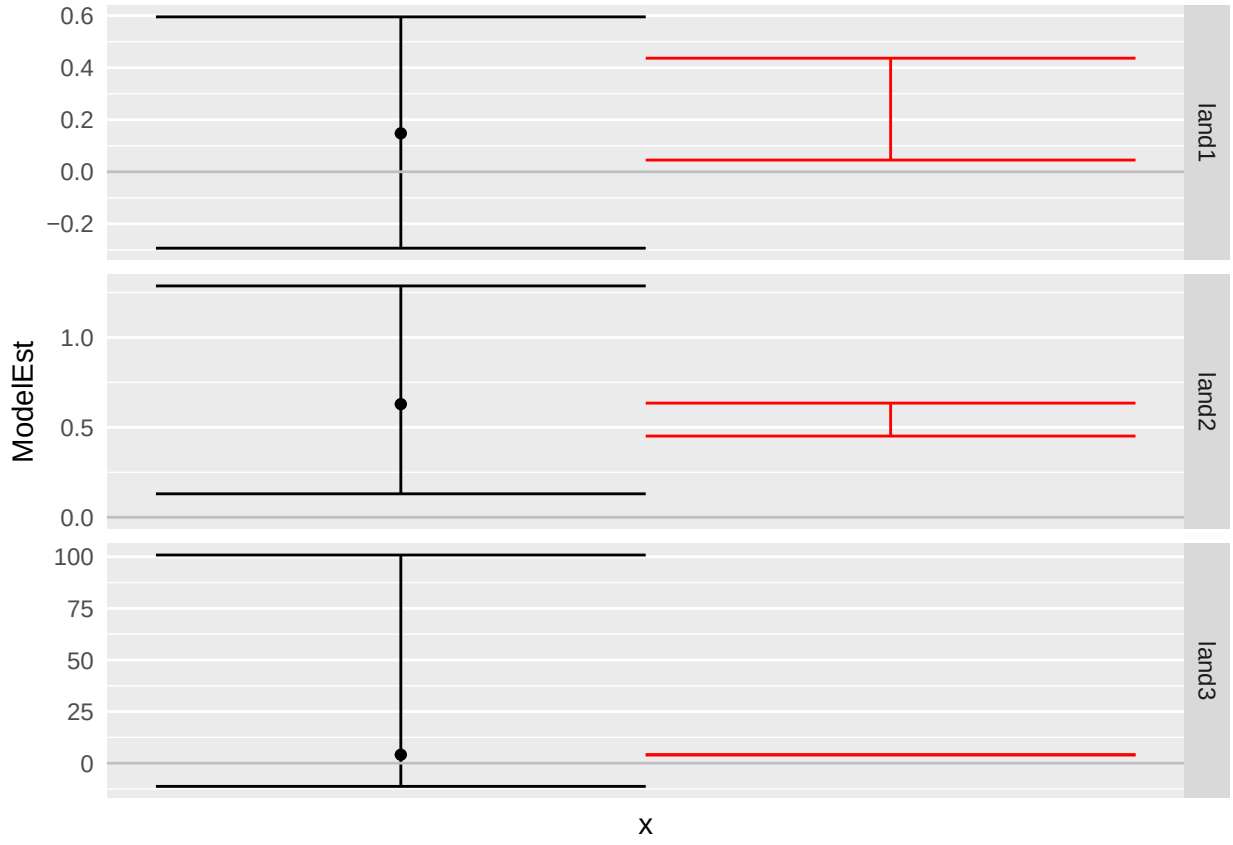
The results show that the range of coefficient estiamtes from the possible models above (red lines) are within the 95% confidence intervals of the coefficients for the main models (black lines):

```
ggplot(coefRanges[2:6,],aes(x=1)) +
  geom_point(aes(y=ModelEst,x=0.95)) +
  geom_errorbar(aes(ymin=ModelConfLow,ymax=ModelConfHigh,
                    x=0.95), width=0.1) +
  geom_errorbar(aes(ymin=unclearLow,ymax=unclearHigh,
                    x=1.05),colour="red",width=0.1) +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())+
  geom_hline(yintercept = 0, colour="gray")+
  facet_grid(cols=vars(Var)) +
  xlab("")+
  coord_cartesian(ylim=c(-0.5,1.5))
```




```
ggplot(coefRanges[7:9,],aes(x=1)) +
  geom_point(aes(y=ModelEst,x=0.95)) +
  geom_errorbar(aes(ymin=ModelConfLow,ymax=ModelConfHigh,
                    x=0.95), width=0.1) +
```

```
  geom_errorbar(aes(ymin=unclearLow,ymax=unclearHigh,
                    x=1.05),colour="red",width=0.1) +
  theme(axis.text.x = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major.x = element_blank(),
        panel.grid.minor.x = element_blank())+
  geom_hline(yintercept = 0, colour="gray")+
  facet_grid(rows=vars(Var),scales = "free")
```



We run a similar test below where all possible combinations are run, and we calculate the p-value from the F-test of the main combined model (m2).

```
rangeOfFTestProbs =
apply(possibleCombinationsOfUnclearCases,1,
    function(X){
      # ... replace the nine cases with the possible combination
      mPresent = d$monster_present
      mPresent[1:9] = X
      # Run the fisher test again
      m2x = glm(factor(mPresent) ~
                  strata+land, family="binomial",data=d)
      ft = car::Anova(m2x)
      return(ft$`Pr(>Chisq)`)
    })
```

Range of p-values for land show that all are below 0.05:

```
range(rangeOfFTestProbs[2,])
```

```
## [1] 0.0002909898 0.0034112405
```

The range of p-values for Strata show that 5 out of 512 tests have p-values above 0.05:

```
range(rangeOfFTestProbs[1,])
```

```
## [1] 8.990634e-05 7.418215e-02
```

We can see which settings of these tests lead to non-significance:

```
# Table of strata for the uncertain condition
d[1:9,c("pref_name_for_society","strata")]
```

```
##     pref_name_for_society                                        strata
## 34                 Wolof 3 social classes or castes, w/ or w/out slavery
## 58                Kazakh                2 social classes, castes/slavery
## 97               Alorese        Wealth Differences or hereditary slavery
## 112                 Lesu                                     Egalitarian
## 126              Mi'kmaq                                     Egalitarian
## 128                Aleut                2 social classes, castes/slavery
## 138            Havasupai                                     Egalitarian
## 150       Tohono O'odham                                     Egalitarian
## 163           Saramaccan                                     Egalitarian
```

```
# Table where each column is a different combination
#  of settings for the monster_present variable
cbind( d[1:9,c("pref_name_for_society")],
    t(possibleCombinationsOfUnclearCases[
      which(rangeOfFTestProbs[1,]>0.05),]))
```

```
##                       36      40      48      56      296
## Var1 "Wolof"          "FALSE" "FALSE" "FALSE" "FALSE" "FALSE"
## Var2 "Kazakh"         "FALSE" "FALSE" "FALSE" "FALSE" "FALSE"
## Var3 "Alorese"        "TRUE"  "FALSE" "FALSE" "FALSE" "FALSE"
## Var4 "Lesu"           "TRUE"  "TRUE"  "FALSE" "TRUE"  "TRUE"
## Var5 "Mi'kmaq"        "TRUE"  "TRUE"  "TRUE"  "FALSE" "TRUE"
## Var6 "Aleut"          "FALSE" "FALSE" "FALSE" "FALSE" "FALSE"
## Var7 "Havasupai"      "TRUE"  "TRUE"  "TRUE"  "TRUE"  "TRUE"
## Var8 "Tohono O'odham" "TRUE"  "TRUE"  "TRUE"  "TRUE"  "TRUE"
## Var9 "Saramaccan"     "TRUE"  "TRUE"  "TRUE"  "TRUE"  "FALSE"
```

The uncertain cases include 5 egalitarian societies and 3 that have at least 2 social classes with castes/slavery. The non-significant results emerge when at least 4 of the egalitarian societies these have composite monsters, while the 3 more hierarchical societies have none. That is, when the data is exactly against the predicted direction.

References

Fritz, S. A. and Purvis, A. (2010). Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. Conservation Biology, 24(4):1042-1051.

Gray RD, Drummond AJ, & Greenhill SJ 2009. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science, 323(5913), 479-483.

Grollemund R, Branford S, Bostoen K, Meade A, Venditti C & Pagel M. 2015. Bantu expansion shows habitat alters the route and pace of human dispersals. Proceedings of the National Academy of Sciences of

the USA, 112(43), 13296-13301.

Jäger, G., 2018. Global-scale phylogenetic linguistic inference from lexical resources. Scientific Data, 5(1), pp.1-16.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401, 877.

Roberts, S.G., Torreira, F. and Levinson, S.C., 2015. The effects of processing and sequence organization on the timing of turn taking: a corpus study. Frontiers in psychology, 6, p.509.

Ross, M. H. (1983). Political decision making and conflict: Additional cross-cultural codes and scales. Ethnology, 22(2), 169-192.

Strobl, C., Malley, J. and Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological methods, 14(4), p.323.