Эффекты ареальности в Сибири

Максим Бажуков, Таня Казакова

2023-06-26

Введение

Этот проект — продолжение работы над проектом
“Morphosyntactic complexities in (North-Eastern) Siberia: evidence for contact-induced convergence?”
Alexey Vinyar, Tatiana Kazakova, Alexandra Nogina, Alexey Baklanov, Daria Ignatenko, Ksenia Lapshina & Ivan Stenin

Задача: Определить, являются ли некоторые явления ареальными Ареал: Северо-Восточная Азия

Готовые данные проекта:

3 явления:

- наличие аттенуатива (0/1)
- наличие именного времени (0/1)
- наличие пролатива (0/1)

Это явления, знакомые авторам проекта по предыдущим исследованиям. В будущем планируется добавить и другие (например: показатель перемещения на глаголах, прохибитив и др.).

Рассчитаем сперва базовые статистики для категориальных переменных на небольшой выборке, связанной с Сибирью. Мы построим таблицу сопряжённости, мозаичный график в качестве визуализации этой таблицы, а также проверим распределение по критерию χ^2 и точным тестом Фишера (более верным при малом количестве данных)

```
make_contingency_stats <- function(table, col1, col2){
  formula <- as.formula(
    paste(
      "~", paste(c(col1, col2), collapse = "+"),
      collapse=" "
    )
  )

  cont_table <- xtabs(formula, data=table)

  print(cont_table)

  mosaicplot(cont_table, color=viridis(2))

  print(chisq.test(cont_table))

  print(fisher.test(cont_table))
}
```

Именное время

```
nomtense_syb <- read_csv2("./data/nominal_tense_siberia.csv")
nomtense_syb %>%
  mutate(in_siberia = as.logical(in_siberia),
         family=as.factor(family),
         nominal_tense = as.logical(case_when(
           nominal_tense == "N" ~ 0,
           nominal_tense == "Y" ~ 1
         )),
         tensed_possesion = as.logical(case_when(
           tensed_possesion == "N" ~ 0,
           tensed_possesion == "Y" ~ 1
         ))
  ) ->
  nomtense_syb
```

```
head(nomtense_syb)
```

```
## # A tibble: 6 × 16
##   family              branch    name      x y     id    in_siberia nominal_tense
##   <fct>               <chr>     <chr> <dbl> <lgl> <chr> <lgl>      <lgl>
## 1 Eskimo-Aleut        Eskimo    Old …  64.5 NA    sire… TRUE       FALSE
## 2 Eskimo-Aleut        Eskimo    Cent…  63.4 NA    cent… TRUE       TRUE
## 3 Chukotko-Kamchatkan Chukotian Chuk…  68.6 NA    chuk… TRUE       TRUE
## 4 Chukotko-Kamchatkan Chukotian Alut…  60.4 NA    alut… TRUE       FALSE
## 5 Chukotko-Kamchatkan Kamchatk… West…  56.0 NA    itel… TRUE       FALSE
## 6 Yukaghir            Kolymic   Sout…  64.2 NA    sout… TRUE       TRUE
## # ℹ 8 more variables: tensed_possesion <lgl>, summary_nt <dbl>, ...11 <lgl>,
## #   ...12 <lgl>, ...13 <lgl>, ...14 <lgl>, ...15 <lgl>, ...16 <lgl>
```

Рассмотрим два явления из этой области: собственно маркирование времени на именах и маркирование времени на обладаемом.

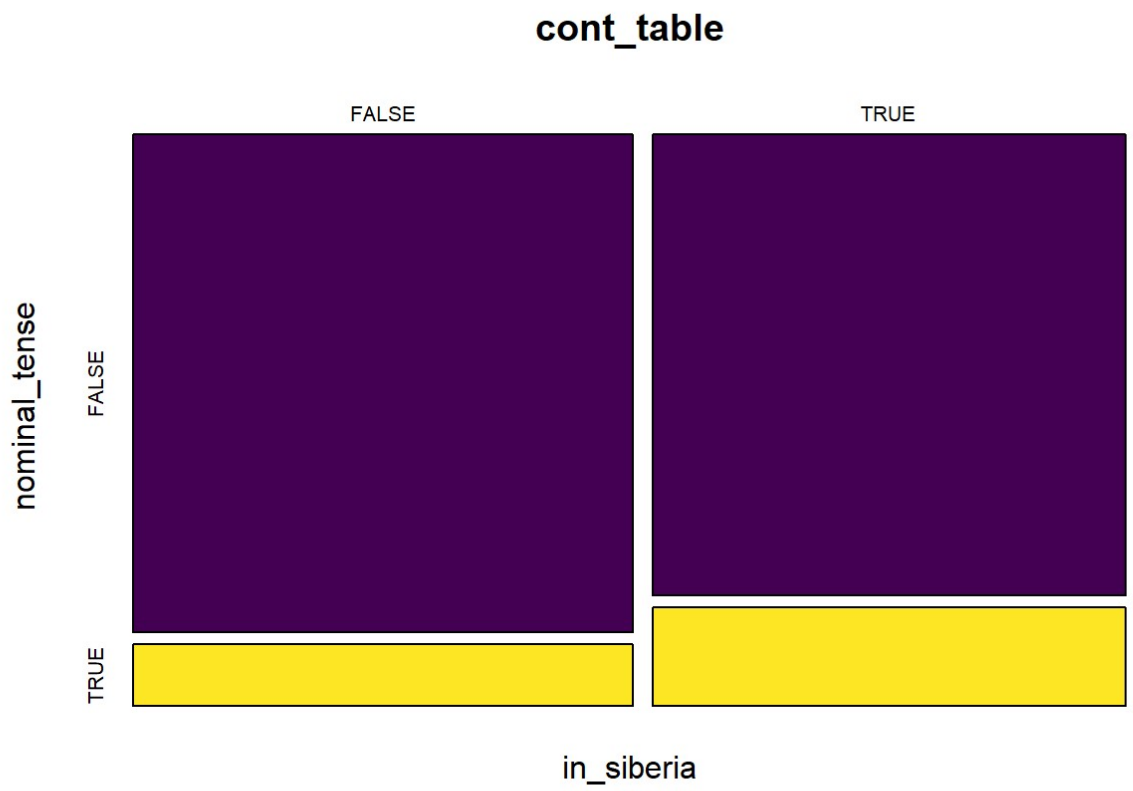
Собственно время

Именное время (Nordlinger, Sadler, 2000) само по себе редкая категория. Согласно данным выборки, оно примерно одинаково редко как в Сибири, так и вне её. Оно вряд ли может быть названо ареальным для Сибири явлением.

```
make_contingency_stats(nomtense_syb, "in_siberia", "nominal_tense")
```

```
##           nominal_tense
## in_siberia FALSE TRUE
##      FALSE    32    4
##      TRUE     28    6
```

```
## Warning in chisq.test(cont_table): аппроксимация на основе хи-квадрат может
## быть неправильной
```



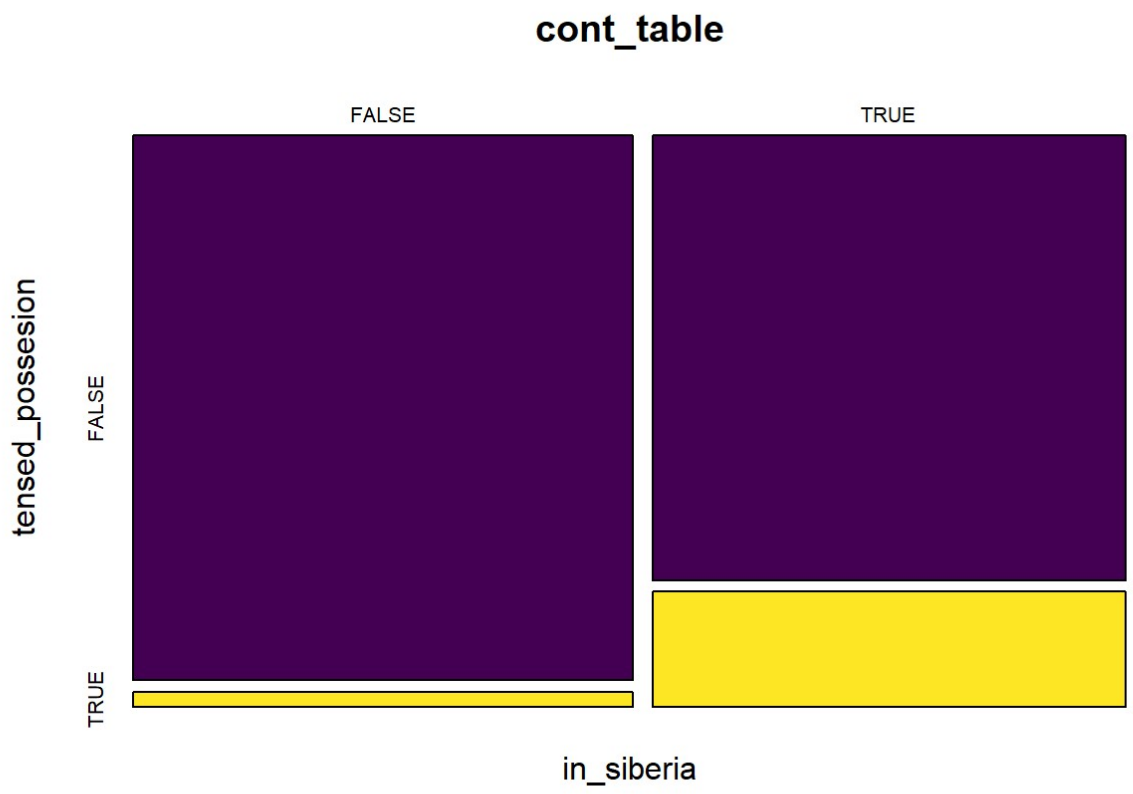

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 0.19301, df = 1, p-value = 0.6604
##
##
##  Fisher's Exact Test for Count Data
##
## data:  cont_table
## p-value = 0.5079
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.360716 9.070604
## sample estimates:
## odds ratio
##   1.701129
```

Посессивность с маркированием времени

```
make_contingency_stats(nomtense_syb, "in_siberia", "tensed_possesion")
```

```
##           tensed_possesion
## in_siberia FALSE TRUE
##      FALSE    35    1
##      TRUE     27    7
```

```
## Warning in chisq.test(cont_table): аппроксимация на основе хи-квадрат может
## быть неправильной
```




```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 3.8613, df = 1, p-value = 0.04941
##
##
##  Fisher's Exact Test for Count Data
##
## data:  cont_table
## p-value = 0.02564
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    1.034693 419.445449
## sample estimates:
## odds ratio
##   8.830951
```

С другой стороны, для не менее редкого явления времени на обладаемом имени видна корреляция с ареалом: в 7 языках Сибири есть это явление, и согласно тесту Фишера, такое совместное распределение можно считать значимо неслучайным (p=0.026)

Полезно может быть также рассмотреть семьи, в языках которых есть это явление и которые относятся к Сибири и сравнить их с другими языками семьи. Это позволит отделить

наследственность от ареальонсти.

Языки, которые относятся к Сибири и в которых есть явление:

```
nomtense_syb %>%
  count(family, tensed_possesion, in_siberia, .drop = FALSE) %>%
  filter(tensed_possesion & in_siberia) ->
  sib_tensed_poss

sib_tensed_poss
```

```
## # A tibble: 2 × 4
##   family   tensed_possesion in_siberia     n
##   <fct>    <lgl>            <lgl>      <int>
## 1 Tungusic TRUE             TRUE           4
## 2 Uralic   TRUE             TRUE           3
```

Посчитаем, сколько в семьях всех остальных языков (сибирские и в них нет явления, либо несибирские с наличием или отсутствием явления).

```
nomtense_syb %>%
  # mutate(tensed_possesion = as.factor(tensed_possesion),
  #        in_siberia = as.factor(in_siberia)) %>%
  count(family, tensed_possesion, in_siberia, .drop = FALSE) %>%
  filter(!(tensed_possesion & in_siberia)) %>%
  group_by(family) %>%
  summarise(n.other = sum(n)) ->
  nonsib_or_nontensed_poss

nonsib_or_nontensed_poss
```

```
## # A tibble: 14 × 2
##    family              n.other
##    <fct>                 <int>
##  1 Ainu                      1
##  2 Chukotko-Kamchatkan       3
##  3 Eskimo-Aleut              7
##  4 Japanese                  1
##  5 Korean                    1
##  6 Mongolic-Kitan            9
##  7 Nivkh                     1
##  8 Russian                   1
##  9 Sino-Tibetan              1
## 10 Tungusic                  2
## 11 Turkic                   17
## 12 Uralic                   15
## 13 Yenisean                  2
## 14 Yukaghir                  2
```

После этого, совместим таблицы и сравним количество языков в семье, которые принадлежат сибирскому ареалу и в которых есть явление и количество всех остальных языков этой семьи.

```
sib_tensed_poss %>%
  left_join(nonsib_or_nontensed_poss, by="family")
```

```
## # A tibble: 2 × 5
##   family   tensed_possesion in_siberia     n n.other
##   <fct>    <lgl>            <lgl>      <int>   <int>
## 1 Tungusic TRUE             TRUE           4       2
## 2 Uralic   TRUE             TRUE           3      15
```

Оказывается, что действительно, в тунгусско-маньчжурской семьяе языков, расположенных в Сибири и имеющих явление больше, чем сибирских языков без явления и несибирских языков вообще. Для уральской семьи в выборке доступно много языков и это не соблюдается. Однако мы знаем, что все остальные языки, это языки без явления (и большинство несибирские):

```
nomtense_syb %>%
  # mutate(tensed_possesion = as.factor(tensed_possesion),
  #        in_siberia = as.factor(in_siberia)) %>%
  count(family, tensed_possesion, in_siberia, .drop = FALSE) %>%
  filter(!(tensed_possesion & in_siberia)) %>%
  filter(family == "Uralic")
```

```
## # A tibble: 2 × 4
##   family tensed_possesion in_siberia     n
##   <fct>  <lgl>            <lgl>      <int>
## 1 Uralic FALSE            FALSE         11
## 2 Uralic FALSE            TRUE           4
```

Это подтверждает утверждение о связи явления с ареалом.

Аттенуатив

Данные по аттенуативу: есть ли в языке особый показатель для *пониженной интенсивности* действия?

```
atten_syb <- read_csv("./data/atten_syb_2.csv")
atten_world <- read_csv("./data/atten_world.csv")
```

```
head(atten_syb)
```

```
## # A tibble: 6 × 7
##   family              language          x     y id    has_attenuative in_siberia
##   <chr>               <chr>         <dbl> <dbl> <chr>           <dbl>      <dbl>
## 1 Eskimo-Aleut        Aleut          52.1 186.  aleu…               1          0
## 2 Chukotko-Kamchatkan Alutor         60.4 166.  alut…               0          1
## 3 Nivkh               Amur Nivkh     52.6 141.  gily…               1          1
## 4 Mongolic-Kitan      Bonan          35.7 103.  bona…               0          0
## 5 Eskimo-Aleut        Central Sibe…  63.4 190.  cent…               1          1
## 6 Turkic              Chagatai       38.2  57.9 chag…               0          0
```

```
head(atten_world)
```

```
## # A tibble: 6 × 4
##   language          glottocode family          has.atten
##   <chr>             <chr>      <chr>           <chr>
## 1 Mpade             mpad1242   Afro-Asiatic    0
## 2 Tiefo-Daramandugu tief1242   Atlantic-Congo  0
## 3 Bambassi          bamb1262   Blue Nile Mao   0
## 4 Ma'di             madi1260   Central Sudanic 0
## 5 Sheko             shek1245   Dizoid          0
## 6 Yanda Dom Dogon   yand1257   Dogon           0
```

Выборки пересекаются лишь по нескольким языкам:

```
in_sib_world <- intersect(atten_syb$id, atten_world$glottocode)
in_sib_world
```

```
## [1] "chuk1273" "halh1238" "ngan1291" "sout2750" "tata1255" "udih1248"
```

Добавим данные по тому, сибирский ли это язык, из первой выборки:

```
atten_world %>%
  mutate(in_siberia = if_else(
    glottocode %in% in_sib_world, 1, 0
  )) %>%
  rename(., has_attenuative=has.atten) %>%
  mutate(family = as.factor(family), in_siberia = as.logical(in_siberia)) ->
  atten_world

atten_world
```

```
## # A tibble: 120 × 5
##    language          glottocode family          has_attenuative in_siberia
##    <chr>             <chr>      <fct>           <chr>           <lgl>
##  1 Mpade             mpad1242   Afro-Asiatic    0               FALSE
##  2 Tiefo-Daramandugu tief1242   Atlantic-Congo  0               FALSE
##  3 Bambassi          bamb1262   Blue Nile Mao   0               FALSE
##  4 Ma'di             madi1260   Central Sudanic 0               FALSE
##  5 Sheko             shek1245   Dizoid          0               FALSE
##  6 Yanda Dom Dogon   yand1257   Dogon           0               FALSE
##  7 Northern Gumuz    gumu1244   Gumuz           0               FALSE
##  8 Sandawe           sand1273   isolate         1               FALSE
##  9 Ts'ixa            tsix1234   Khoe-Kwadi      0               FALSE
## 10 Uduk              uduk1239   Koman           0               FALSE
## # ℹ 110 more rows
```

```
atten_world %>%
  filter(has_attenuative %in% c(0,1)) %>%
  mutate(has_attenuative = as.logical(as.numeric(has_attenuative))) ->
  atten_world
```

```
table(atten_world$has_attenuative, atten_world$in_siberia)
```

```
##
##         FALSE TRUE
##   FALSE    90    1
##   TRUE     14    5
```

```
atten_full <- merge(
  x = atten_syb,
  y = (atten_world %>% filter(!(glottocode %in% in_sib_world))),
  all = TRUE
)

head(atten_full)
```

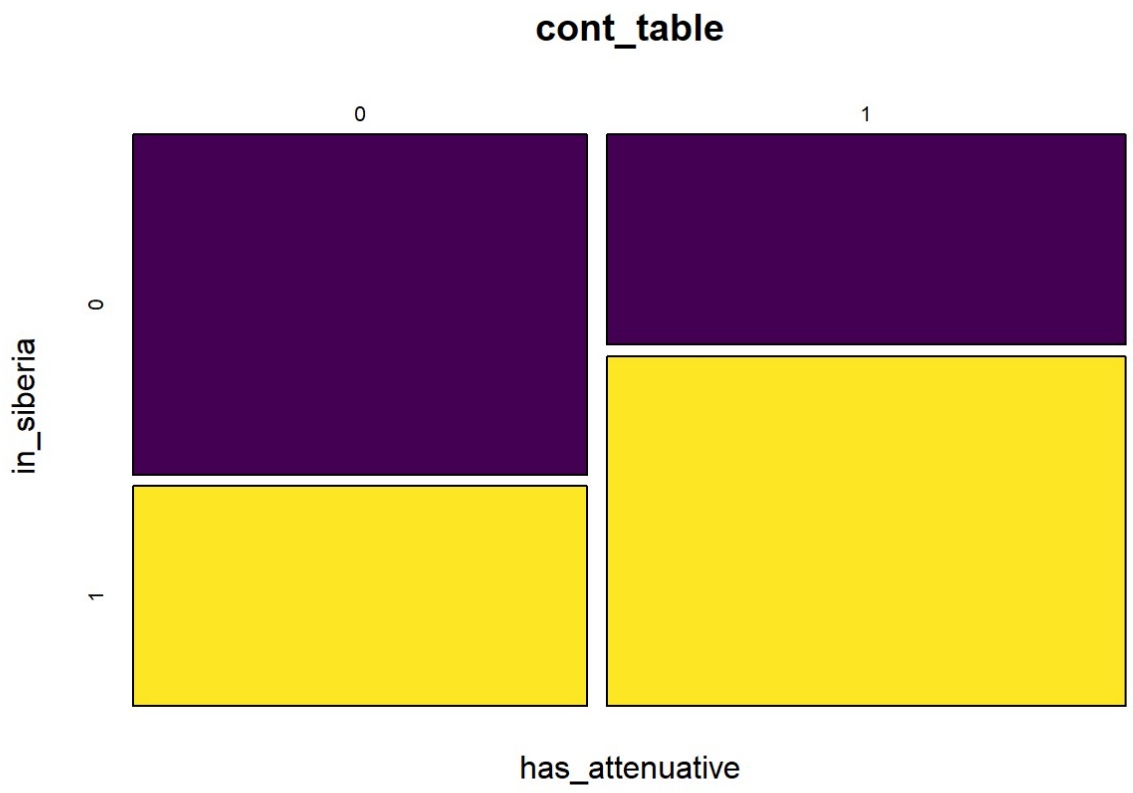
```
##         family            language has_attenuative in_siberia  x  y   id
## 1 Abkhaz-Adyge           Kabardian               0          0 NA NA <NA>
## 2 Afro-Asiatic               Mpade               0          0 NA NA <NA>
## 3         Ainu       Hokkaido Ainu               0          0 NA NA <NA>
## 4        Algic Northwestern Ojibwa               0          0 NA NA <NA>
## 5         Anim              Marind               1          0 NA NA <NA>
## 6   Araucanian          Mapudungun               0          0 NA NA <NA>
##   glottocode
## 1   kaba1278
## 2   mpad1242
## 3   ainu1240
## 4   nort2961
## 5   nucl1622
## 6   mapu1245
```

```
table(atten_full$has_attenuative)
```

```
##
##   0   1
## 118  46
```

```
make_contingency_stats(atten_syb, 'has_attenuative', 'in_siberia')
```

```
##                in_siberia
## has_attenuative  0  1
##               0 17 11
##               1 12 20
```

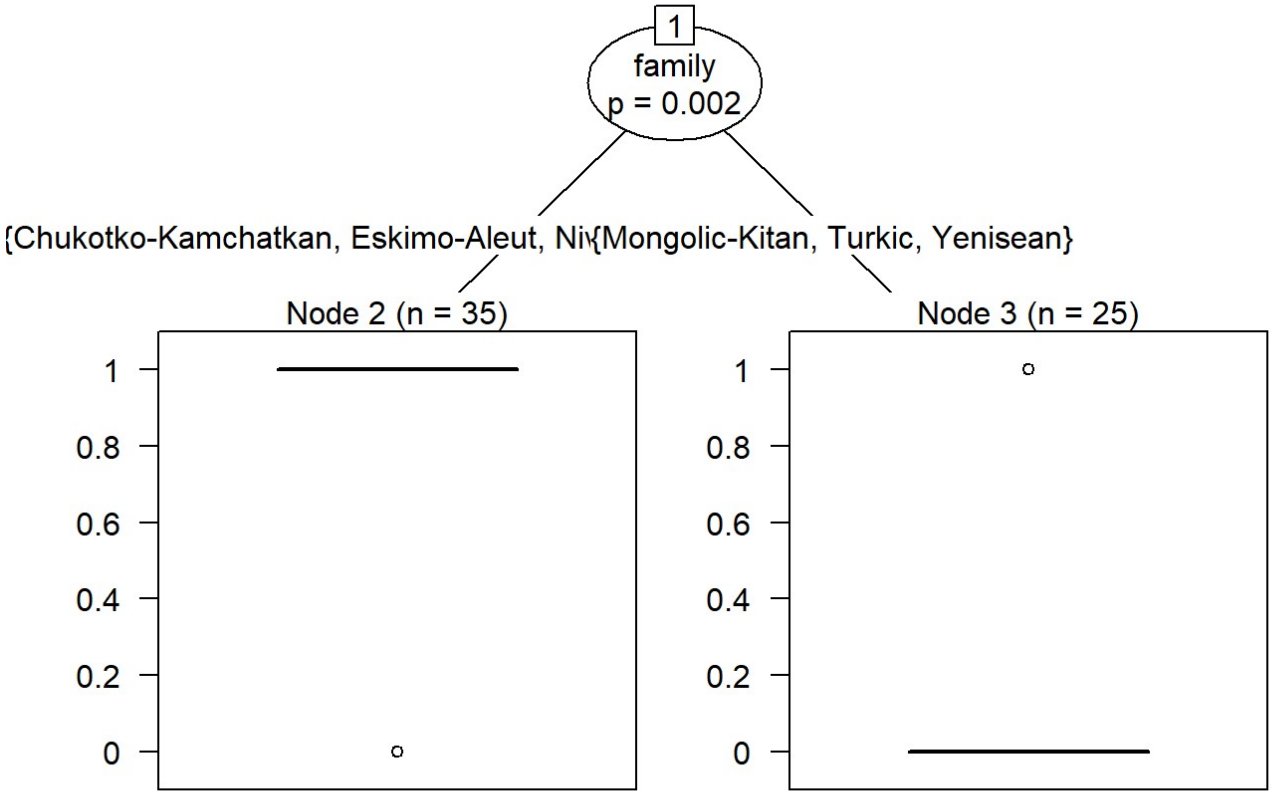



```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 2.3601, df = 1, p-value = 0.1245
##
##
##  Fisher's Exact Test for Count Data
##
## data:  cont_table
## p-value = 0.1197
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.8072183 8.3271693
## sample estimates:
## odds ratio
##   2.533928
```

Согласно графику, действительно в языках, отнесённых к Сибири, аттенуатив чаще присутствует, чем нет, а не в Сибири наоборот. Однако статистические критерии (здесь уместно говорить о точном тесте Фишера) не поддерживают такой вывод.

Построим решающее дерево.

```
set.seed(42)

atten_syb %>%
  mutate(family = as.factor(family), in_siberia = as.logical(in_siberia),
         has_attenuative = as.logical(has_attenuative)) ->
  atten_syb

atten_syb_fit <- ctree(has_attenuative ~ family + in_siberia, data = atten_syb)
plot(atten_syb_fit)
```




```
print(atten_syb_fit)
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  has_attenuative
## Inputs:  family, in_siberia
## Number of observations:  60
##
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Nivkh, Tungusic, Uralic, Yukaghir}; cr
iterion = 0.998, statistic = 25.996
##   2)*  weights = 35
## 1) family == {Mongolic-Kitan, Turkic, Yenisean}
##   3)*  weights = 25
```

```
atten_syb_fit2 <- cforest(has_attenuative ~ family + in_siberia, data = atten_syb,
                          controls=cforest_unbiased(ntree=100, mtry=2))
# plot(atten_syb_fit2)
print(atten_syb_fit2)
```

```
##
##   Random Forest using Conditional Inference Trees
##
## Number of trees:  100
##
## Response:  has_attenuative
## Inputs:  family, in_siberia
## Number of observations:  60
```

Со стандартным значением α=0.05 оказывается в нашем случае, что алгоритм построения дерева выполняет лишь одно деление (https://cran.r-project.org/package=party (https://cran.r-project.org/package=party)). Предикат для разделения оказывается тривиальным (все семьи, где есть аттенуатив против остальных). Однако можно заметить также и выбросы, по одному выбивающемуся наблюдению в каждой группе.

Языки, которые относятся к Сибири и в которых есть явление:

```
atten_syb %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(has_attenuative & in_siberia) ->
  syb_atten

syb_atten
```

```
## # A tibble: 8 × 4
##   family              has_attenuative in_siberia     n
##   <fct>               <lgl>           <lgl>      <int>
## 1 Chukotko-Kamchatkan TRUE            TRUE           2
## 2 Eskimo-Aleut        TRUE            TRUE           2
## 3 Mongolic-Kitan      TRUE            TRUE           1
## 4 Nivkh               TRUE            TRUE           1
## 5 Tungusic            TRUE            TRUE           4
## 6 Turkic              TRUE            TRUE           1
## 7 Uralic              TRUE            TRUE           7
## 8 Yukaghir            TRUE            TRUE           2
```

```
atten_syb %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(!(has_attenuative & in_siberia)) %>%
  group_by(family) %>%
  summarise(n.other = sum(n)) ->
  nonsib_or_nonatten

nonsib_or_nonatten
```

```
## # A tibble: 7 × 2
##   family              n.other
##   <fct>                 <int>
## 1 Chukotko-Kamchatkan       1
## 2 Eskimo-Aleut              4
## 3 Mongolic-Kitan            6
## 4 Tungusic                  2
## 5 Turkic                   15
## 6 Uralic                   10
## 7 Yenisean                  2
```

```
syb_atten %>%
  left_join(nonsib_or_nonatten, by="family")
```

```
## # A tibble: 8 × 5
##   family              has_attenuative in_siberia     n n.other
##   <fct>               <lgl>           <lgl>      <int>   <int>
## 1 Chukotko-Kamchatkan TRUE            TRUE           2       1
## 2 Eskimo-Aleut        TRUE            TRUE           2       4
## 3 Mongolic-Kitan      TRUE            TRUE           1       6
## 4 Nivkh               TRUE            TRUE           1      NA
## 5 Tungusic            TRUE            TRUE           4       2
## 6 Turkic              TRUE            TRUE           1      15
## 7 Uralic              TRUE            TRUE           7      10
## 8 Yukaghir            TRUE            TRUE           2      NA
```

Согласно небольшой выборке, аттенуатив не специфичен для Сибири (ср. уральские и сибирские тюркские.)

```
atten_syb %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(!(has_attenuative & in_siberia)) %>%
  filter(family %in% c("Eskimo-Aleut", "Chukotko-Kamchatkan"))
```

```
## # A tibble: 2 × 4
##   family              has_attenuative in_siberia     n
##   <fct>               <lgl>           <lgl>      <int>
## 1 Chukotko-Kamchatkan FALSE           TRUE           1
## 2 Eskimo-Aleut        TRUE            FALSE          4
```

```
atten_syb %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(!(has_attenuative & in_siberia)) %>%
  filter(family %in% c("Uralic", "Turkic"))
```

```
## # A tibble: 5 × 4
##   family has_attenuative in_siberia     n
##   <fct>  <lgl>           <lgl>      <int>
## 1 Turkic FALSE           FALSE          8
## 2 Turkic FALSE           TRUE           5
## 3 Turkic TRUE            FALSE          2
## 4 Uralic FALSE           FALSE          4
## 5 Uralic TRUE            FALSE          6
```

Данные по миру

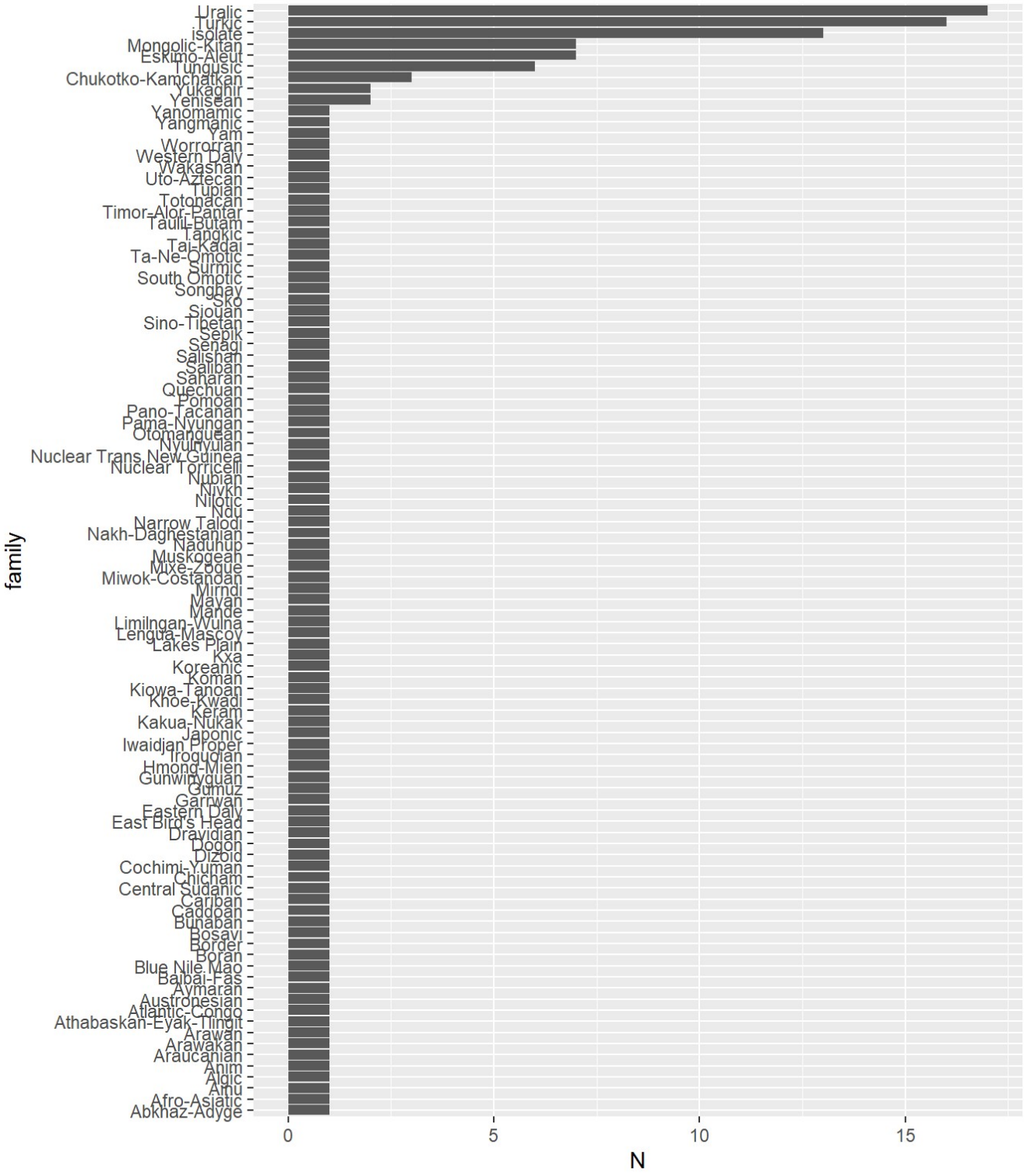
Рассмотрим теперь, как малые данные околосибирской выборки встраиваются в несколько больший мировой контекст.

```
head(atten_full)
```

```
##         family            language has_attenuative in_siberia  x  y   id
## 1 Abkhaz-Adyge           Kabardian               0          0 NA NA <NA>
## 2 Afro-Asiatic               Mpade               0          0 NA NA <NA>
## 3         Ainu       Hokkaido Ainu               0          0 NA NA <NA>
## 4        Algic Northwestern Ojibwa               0          0 NA NA <NA>
## 5         Anim              Marind               1          0 NA NA <NA>
## 6   Araucanian          Mapudungun               0          0 NA NA <NA>
##   glottocode
## 1   kaba1278
## 2   mpad1242
## 3   ainu1240
## 4   nort2961
## 5   nucl1622
## 6   mapu1245
```

В выборке много уральских, тюркских, монгольских, эскимоссо-алеутских за счёт слияния с околосибирской соединения:

```
atten_full %>%
  group_by(family) %>%
  summarise(N = n()) %>%
  mutate(family = fct_reorder(family, N)) %>%
  ggplot(aes(x=family, y=N)) +
  geom_col() +
  coord_flip()
```

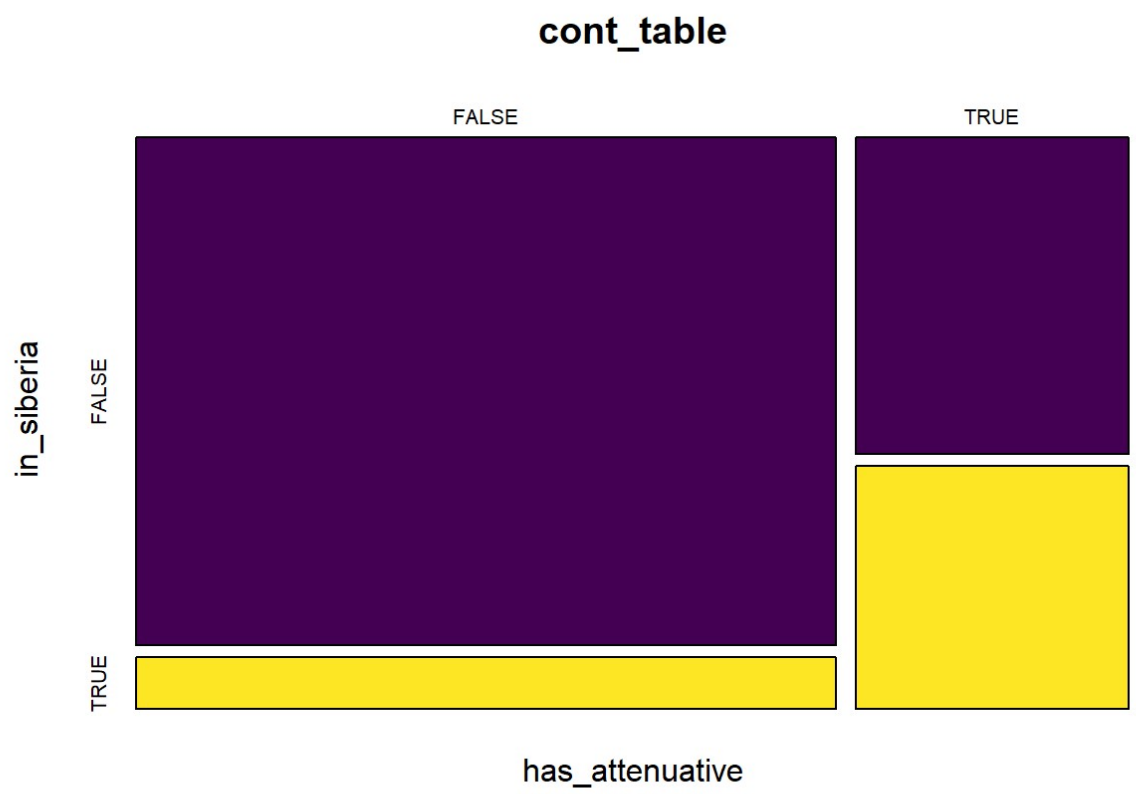

family
Uralic
Turkic
isolate
Mongolic-Kitan
Eskimo-Aleut
Tungusic
Chukotko-Kamchatkan
Yukaghir
Yeniseian
Yanomamic
Yangmanic
Yam
Worrorran
Western Daly
Wakashan
Uto-Aztecan
Tupian
Totonacan
Timor-Alor-Pantar
Tauli-Butam
Tangkic
Tai-Kadai
Ta-Ne-Omotic
Surmic
South Omotic
Songhay
Sko
Siouan
Sino-Tibetan
Sepik
Senagi
Salishan
Saliban
Saharan
Quechuan
Pomoan
Pano-Tacanan
Pama-Nyungan
Otomanguean
Nyulnyulan
Nuclear Trans New Guinea
Nuclear Torricelli
Nubian
Nivkh
Nilotic
Ndu
Narrow Talodi
Nakh-Daghestanian
Naduhup
Muskogean
Mixe-Zoque
Miwok-Costanoan
Mirndi
Mayan
Mande
Limilngan-Wulna
Lengua-Mascoy
Lakes Plain
Kxa
Koreanic
Koman
Kiowa-Tanoan
Khoe-Kwadi
Keram
Kakua-Nukak
Japonic
Iwaidjan Proper
Iroquoian
Hmong-Mien
Gunwinyguan
Gumuz
Garrwan
Eastern Daly
East Bird's Head
Dravidian
Dogon
Dizoid
Cochimi-Yuman
Chicham
Central Sudanic
Carib
Caddoan
Bunaban
Bosavi
Border
Boran
Blue Nile Mao
Baibai-Fas
Aymaran
Austronesian
Atlantic-Congo
Athabaskan-Eyak-Tlingit
Arawan
Arawakan
Araucanian
Anim
Algic
Ainu
Afro-Asiatic
Abkhaz-Adyge
0
5
10
15
N

```
atten_world
```

```
## # A tibble: 110 × 5
##    language          glottocode family          has_attenuative in_siberia
##    <chr>             <chr>      <fct>           <lgl>           <lgl>
##  1 Mpade             mpad1242   Afro-Asiatic    FALSE           FALSE
##  2 Tiefo-Daramandugu tief1242   Atlantic-Congo  FALSE           FALSE
##  3 Bambassi          bamb1262   Blue Nile Mao   FALSE           FALSE
##  4 Ma'di             madi1260   Central Sudanic FALSE           FALSE
##  5 Sheko             shek1245   Dizoid          FALSE           FALSE
##  6 Yanda Dom Dogon   yand1257   Dogon           FALSE           FALSE
##  7 Northern Gumuz    gumu1244   Gumuz           FALSE           FALSE
##  8 Sandawe           sand1273   isolate         TRUE            FALSE
##  9 Ts'ixa            tsix1234   Khoe-Kwadi      FALSE           FALSE
## 10 Uduk              uduk1239   Koman           FALSE           FALSE
## # ℹ 100 more rows
```

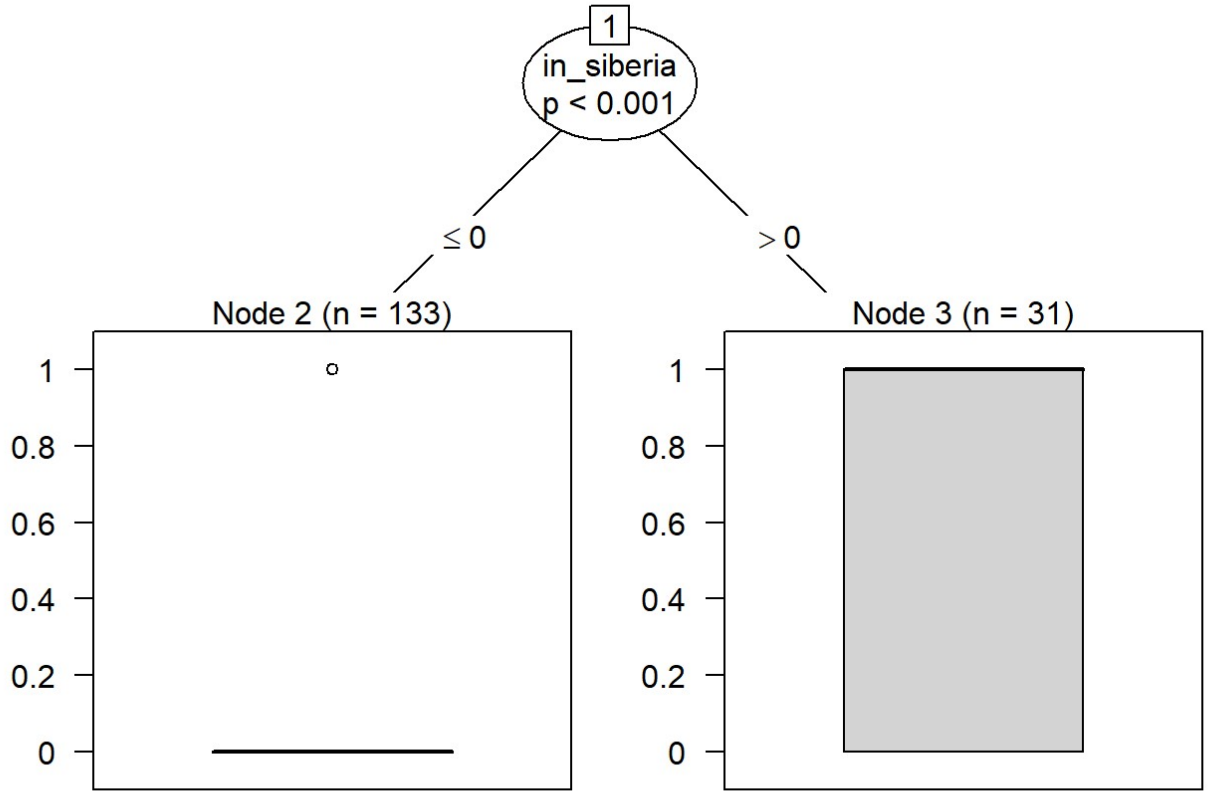
```
atten_full %>%
  mutate(family = as.factor(family), in_siberia = as.logical(in_siberia),
         has_attenuative = as.logical(has_attenuative)) ->
  atten_full

atten_full %>%
  filter(!is.na(has_attenuative)) %>%
  make_contingency_stats("has_attenuative", "in_siberia")
```

```
##                in_siberia
## has_attenuative FALSE TRUE
##           FALSE   107   11
##           TRUE     26   20
```




```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 23.01, df = 1, p-value = 1.612e-06
##
##
##  Fisher's Exact Test for Count Data
##
## data:  cont_table
## p-value = 2.277e-06
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   2.950912 19.338034
## sample estimates:
## odds ratio
##   7.363434
```

```
atten_full %>%
  filter(!is.na(has_attenuative)) %>%
  ctree(has_attenuative ~ family + in_siberia, data = .) ->
  atten_full_fit

plot(atten_full_fit)
```




```
print(atten_full_fit)
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  has_attenuative
## Inputs:  family, in_siberia
## Number of observations:  164
##
## 1) in_siberia <= 0; criterion = 1, statistic = 117.315
##   2)*  weights = 133
## 1) in_siberia > 0
##   3)*  weights = 31
```

На этой выборке, включающей много мировых языков, наиболее важным предикатом оказывается то, в Сибири ли язык! Тут тоже не выполняется дальнейших делений в дереве, как и в ситуации выше, с околосибирской выборкой.

```
atten_full %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(has_attenuative & in_siberia) ->
  world_atten


atten_full %>%
  count(family, has_attenuative, in_siberia, .drop = FALSE) %>%
  filter(!(has_attenuative & in_siberia)) %>%
  group_by(family) %>%
  summarise(n.other = sum(n)) ->
  nonworld_or_nonatten

world_atten
```

```
##                family has_attenuative in_siberia n
## 1 Chukotko-Kamchatkan            TRUE       TRUE 2
## 2        Eskimo-Aleut            TRUE       TRUE 2
## 3      Mongolic-Kitan            TRUE       TRUE 1
## 4               Nivkh            TRUE       TRUE 1
## 5            Tungusic            TRUE       TRUE 4
## 6              Turkic            TRUE       TRUE 1
## 7              Uralic            TRUE       TRUE 7
## 8            Yukaghir            TRUE       TRUE 2
```

```
world_atten %>%
  left_join(nonworld_or_nonatten, by="family")
```

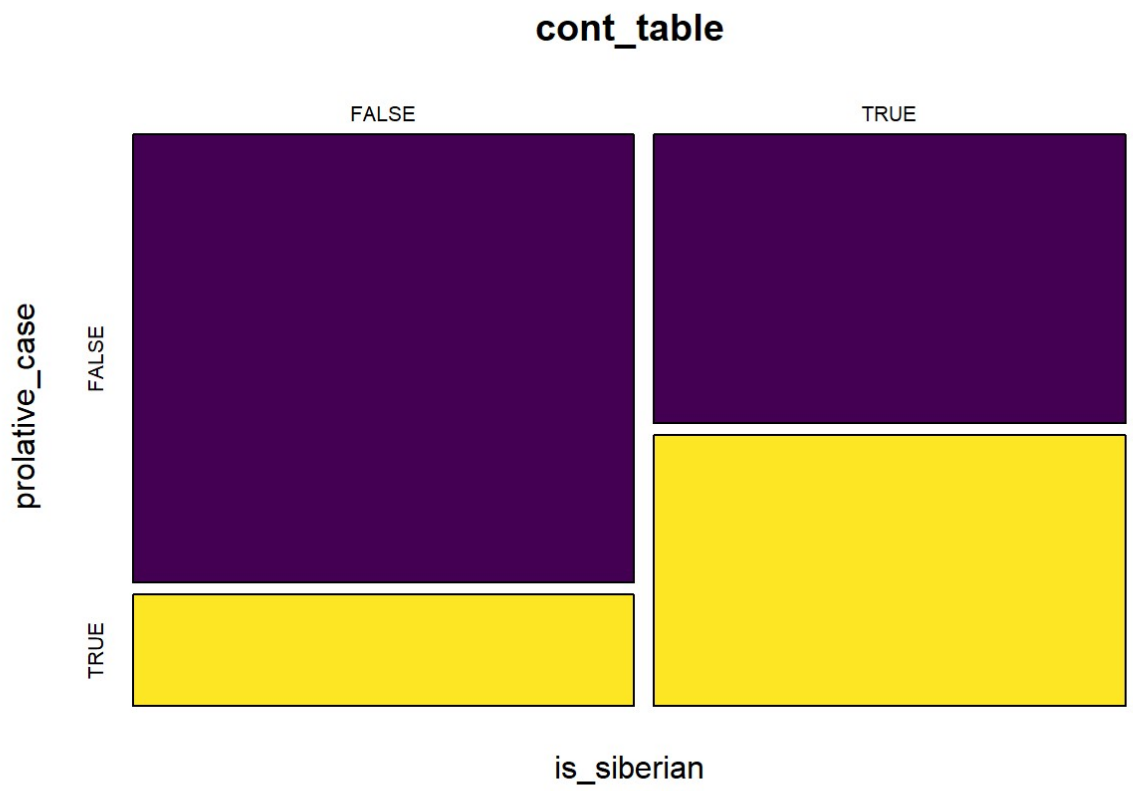
```
##                family has_attenuative in_siberia n n.other
## 1 Chukotko-Kamchatkan            TRUE       TRUE 2       1
## 2        Eskimo-Aleut            TRUE       TRUE 2       5
## 3      Mongolic-Kitan            TRUE       TRUE 1       6
## 4               Nivkh            TRUE       TRUE 1      NA
## 5            Tungusic            TRUE       TRUE 4       2
## 6              Turkic            TRUE       TRUE 1      15
## 7              Uralic            TRUE       TRUE 7      10
## 8            Yukaghir            TRUE       TRUE 2      NA
```

Рассуждения выше про уральские и тюркские остаются актуальны, но действительно на мировой выборке то, сибирский ли язык, оказывается ключевым фактором.

Пролатив

```
make_contingency_stats(prol_syb, "is_siberian", "prolative_case")
```

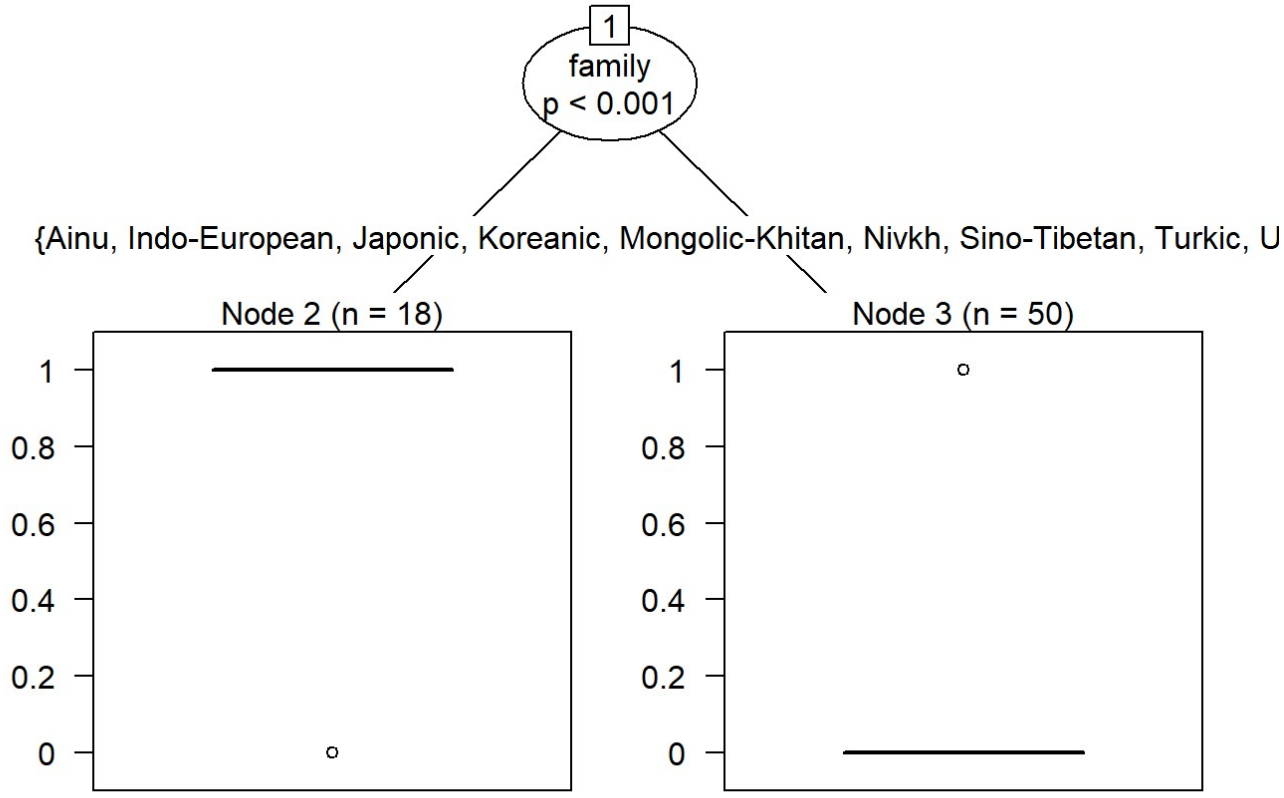
```
##            prolative_case
## is_siberian FALSE TRUE
##       FALSE    28    7
##       TRUE     17   16
```




```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  cont_table
## X-squared = 4.9503, df = 1, p-value = 0.02609
##
##
##  Fisher's Exact Test for Count Data
##
## data:  cont_table
## p-value = 0.02046
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   1.15108 12.95908
## sample estimates:
## odds ratio
##   3.688074
```

Картина по пролативу на небольшой (немного другой) околосибирской выборке более явная, чем была по аттенуативу. Здесь видим, значимые значения критериев χ^2 и теста Фишера ($p=0.026$). Более однозначно выглядит и график: пролатив почти не представлен в несибирских языках, а сибирские языки делятся на две равные группы по его наличию.

```
prol_syb %>%
  filter(!is.na(prolative_case)) %>%
  ctree(prolative_case ~ family + is_siberian, data = .) ->
  prol_syb_fit

plot(prol_syb_fit)
```



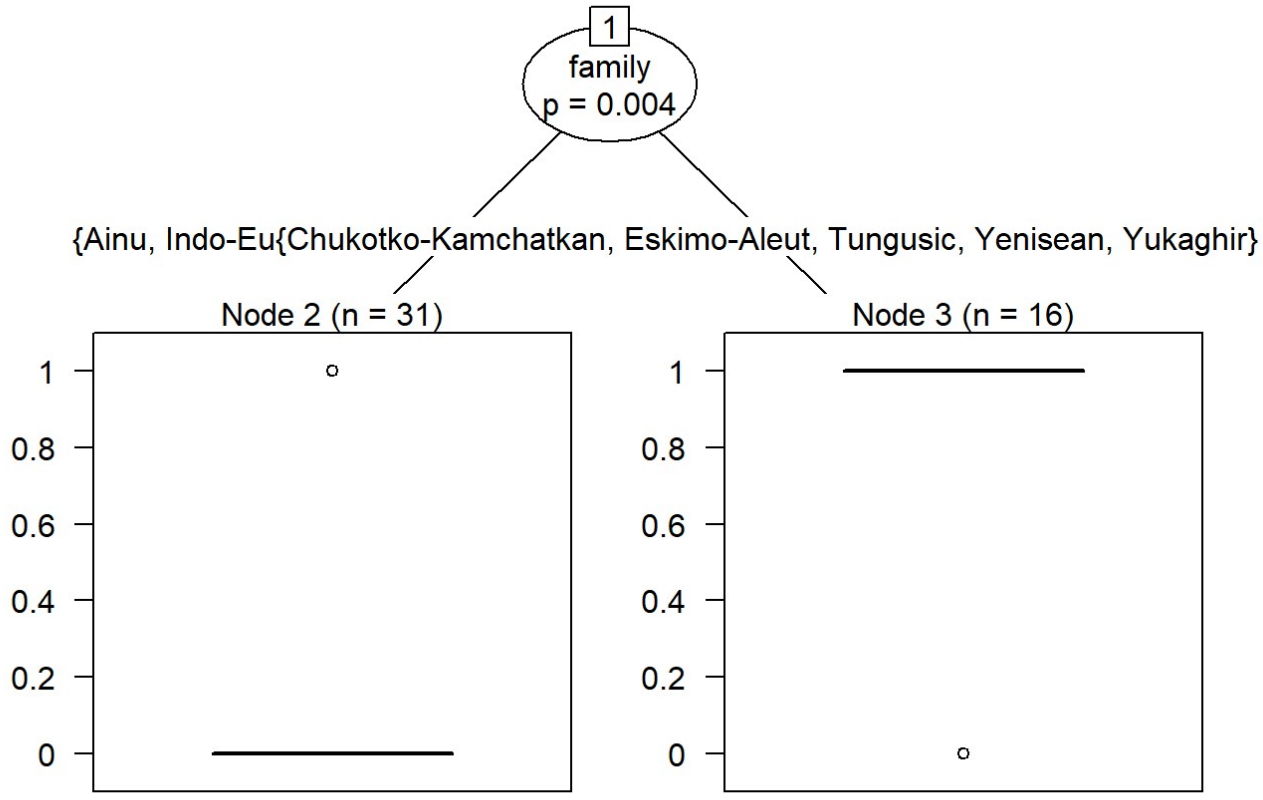

```
print(prol_syb_fit)
```

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  68
##
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Yukaghir}; criterion = 1, st
atistic = 40.85
##   2)*  weights = 18
## 1) family == {Ainu, Indo-European, Japonic, Koreanic, Mongolic-Khitan, Nivkh, Sino-Tibe
tan, Turkic, Uralic, Yenisean}
##   3)*  weights = 50
```

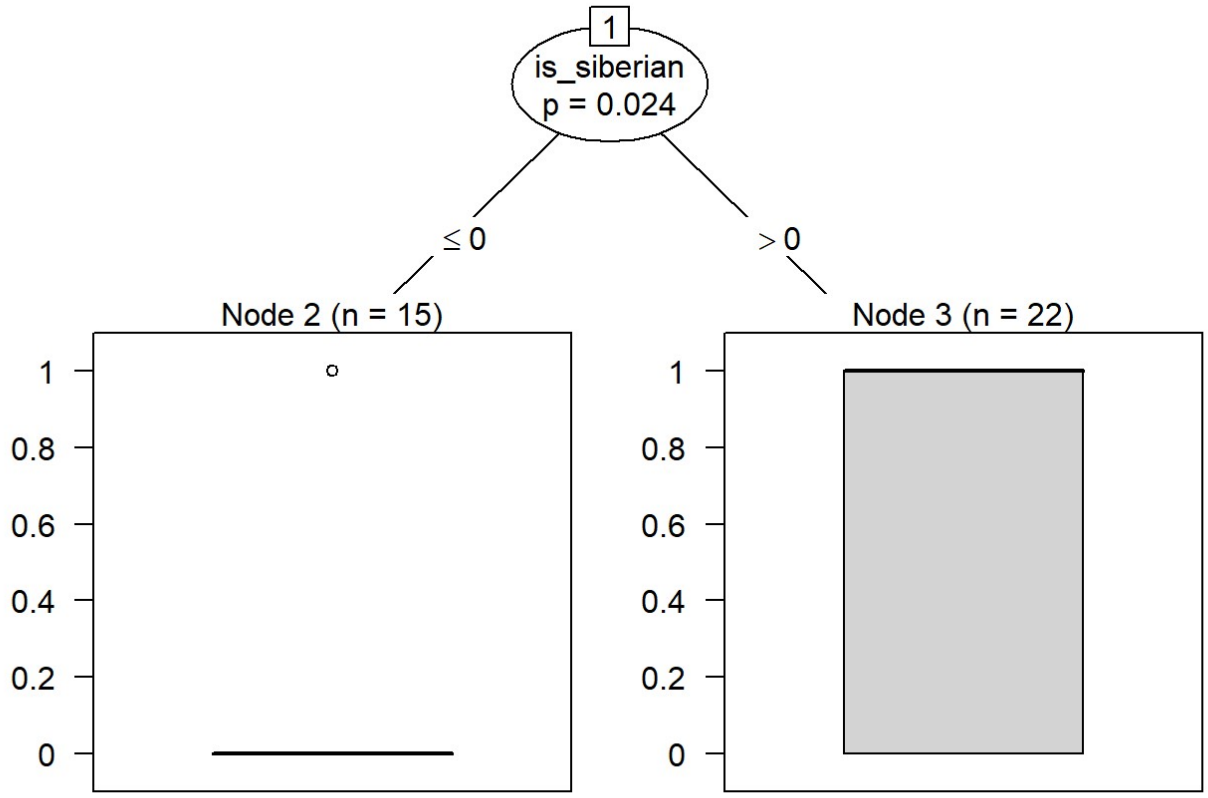
Точно так же, как и выше, алгоритм просто отобрал семьи, где присутствует пролатив.

Ситуация почти не меняется, если семплировать данные и повторять построение. Однако иногда принадлежность к Сибири может служить вторым предикатом для отобранных по семплу семей.

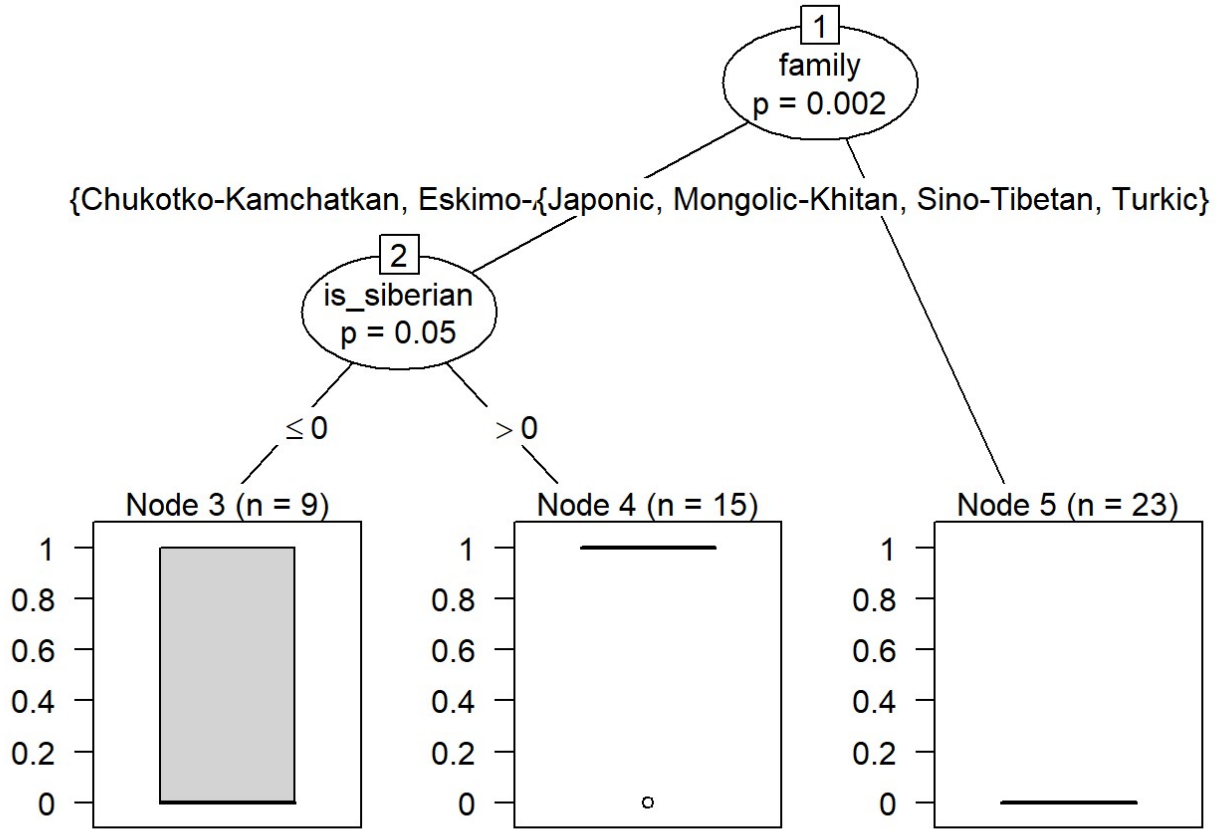
```
set.seed(12)

for(i in 1:7){

nona <- prol_syb %>%  filter(!is.na(prolative_case))

sample <- sample(c(TRUE, FALSE), nrow(nona), replace=TRUE, prob=c(0.65,0.35))

prol_syb_train  <- nona[sample, ]
prol_syb_test   <- nona[!sample, ]

prol_syb_train %>%
  filter(!is.na(prolative_case)) %>%
  ctree(prolative_case ~ family + is_siberian, data = .) ->
  prol_syb_fit2

plot(prol_syb_fit2)
print(prol_syb_fit2)

}
```



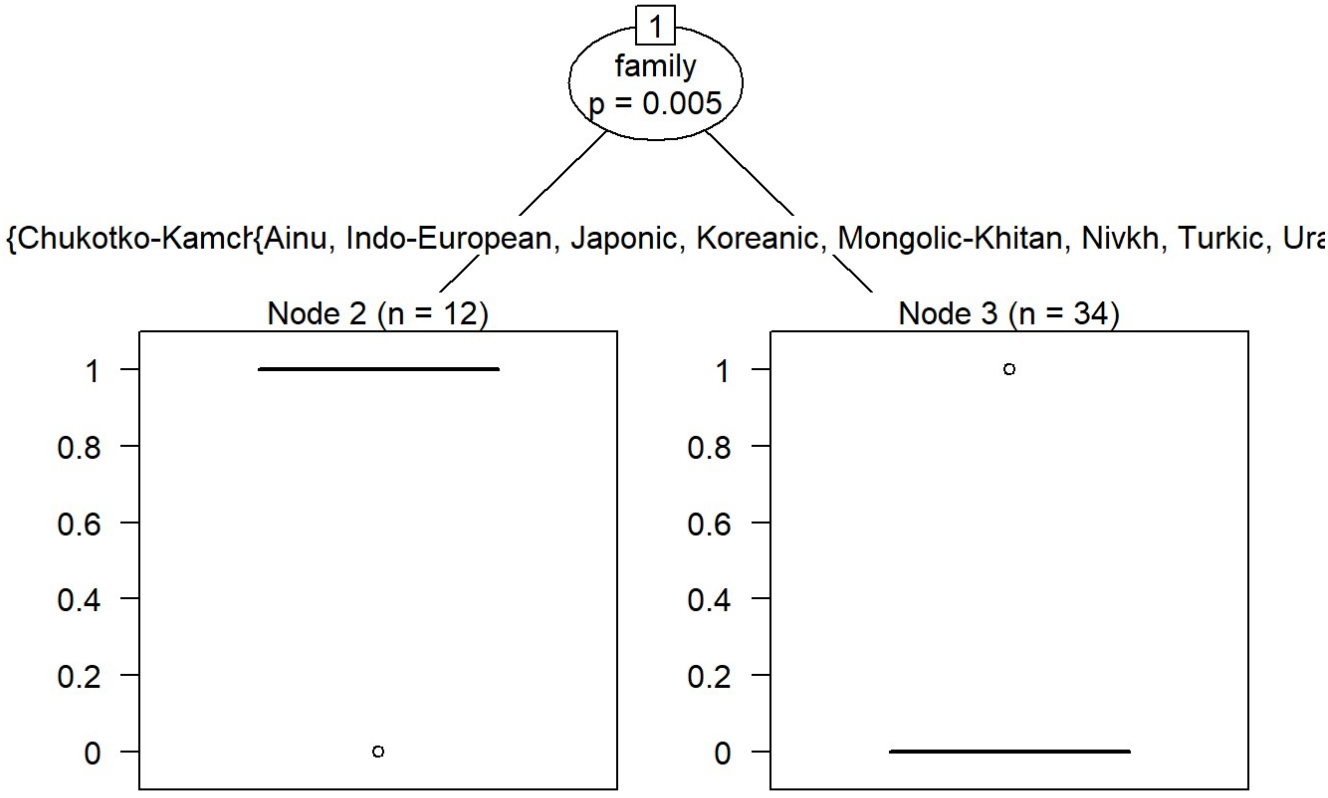
```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  47
##
## 1) family == {Ainu, Indo-European, Mongolic-Khitan, Turkic, Uralic}; criterion = 0.996,
statistic = 26.164
##   2)*  weights = 31
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Yenisean, Yukaghir}
##   3)*  weights = 16
```



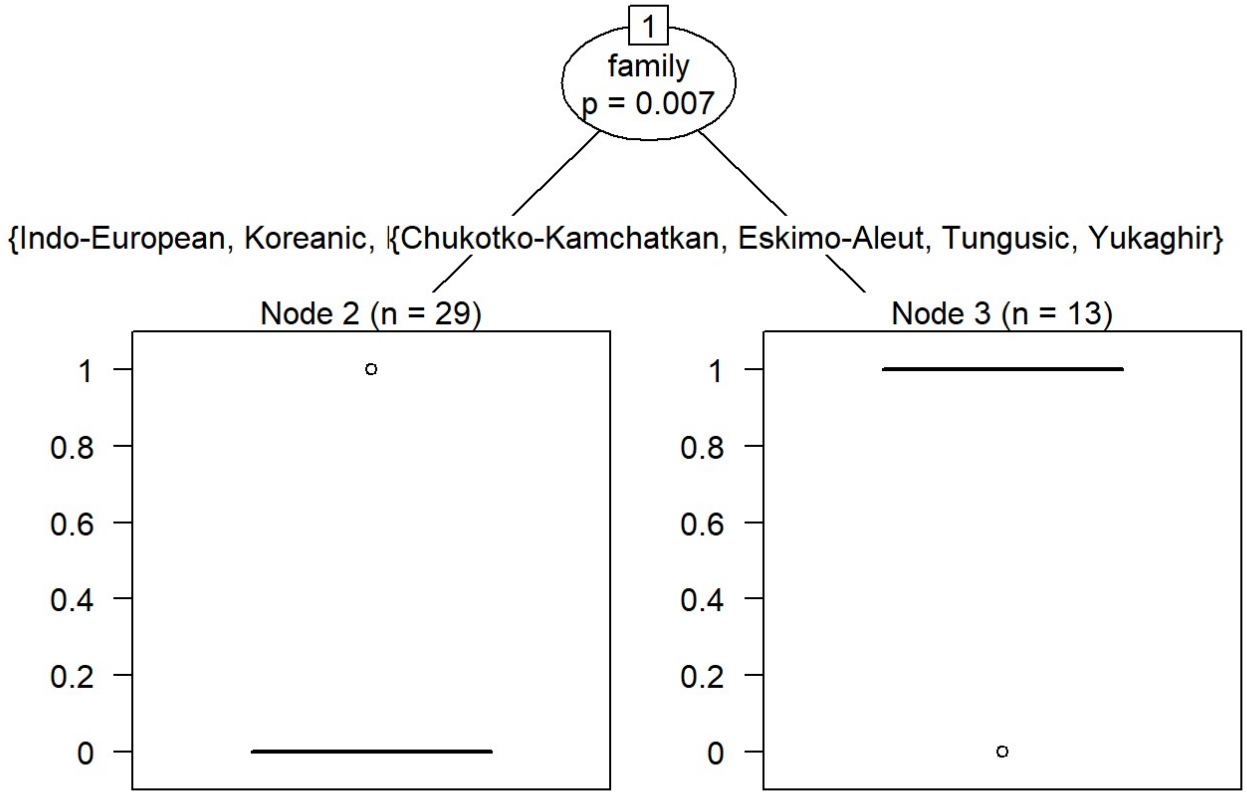

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  37
##
## 1) is_siberian <= 0; criterion = 0.976, statistic = 20.291
##   2)*  weights = 15
## 1) is_siberian > 0
##   3)*  weights = 22
```



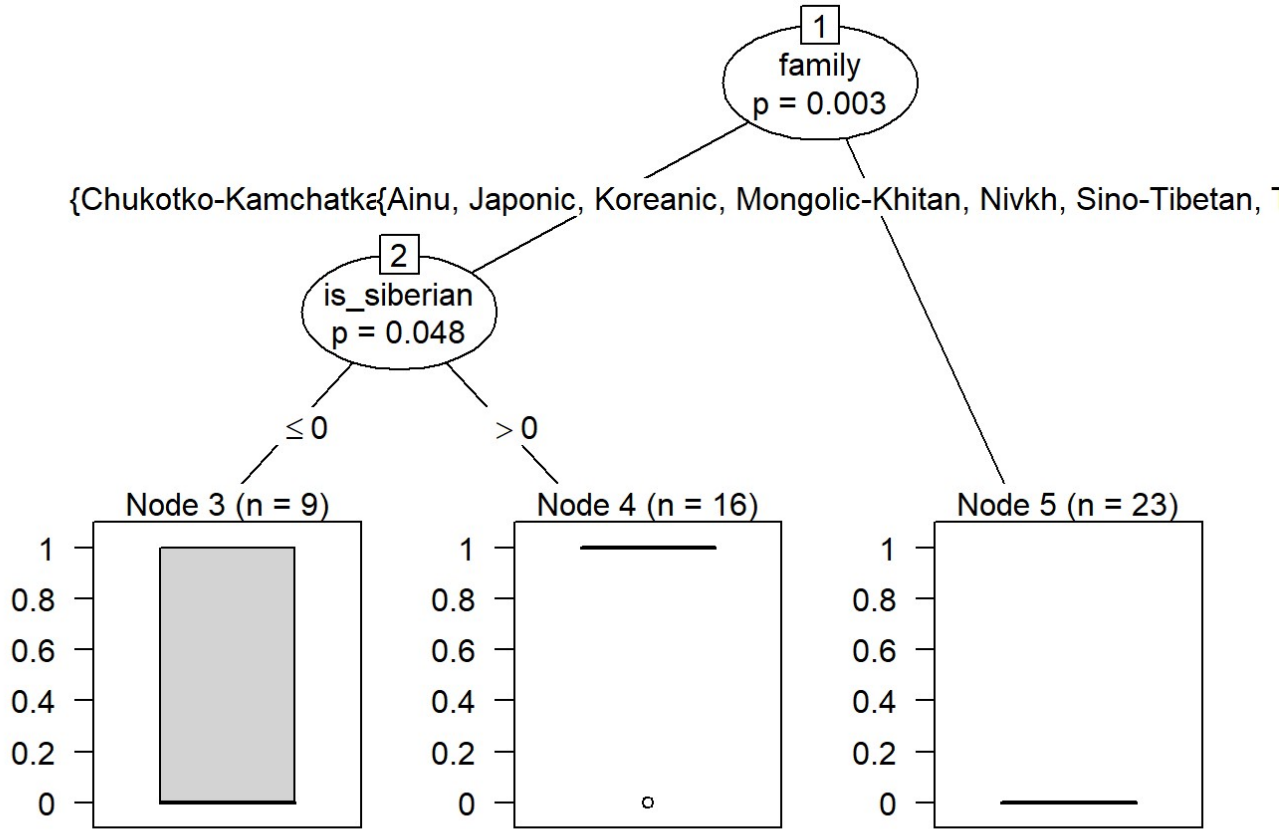

```
##
##   Conditional inference tree with 3 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  47
##
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Uralic, Yenisean, Yukaghir};
criterion = 0.998, statistic = 27.608
##   2) is_siberian <= 0; criterion = 0.95, statistic = 6.304
##     3)*  weights = 9
##   2) is_siberian > 0
##     4)*  weights = 15
## 1) family == {Japonic, Mongolic-Khitan, Sino-Tibetan, Turkic}
##   5)*  weights = 23
```



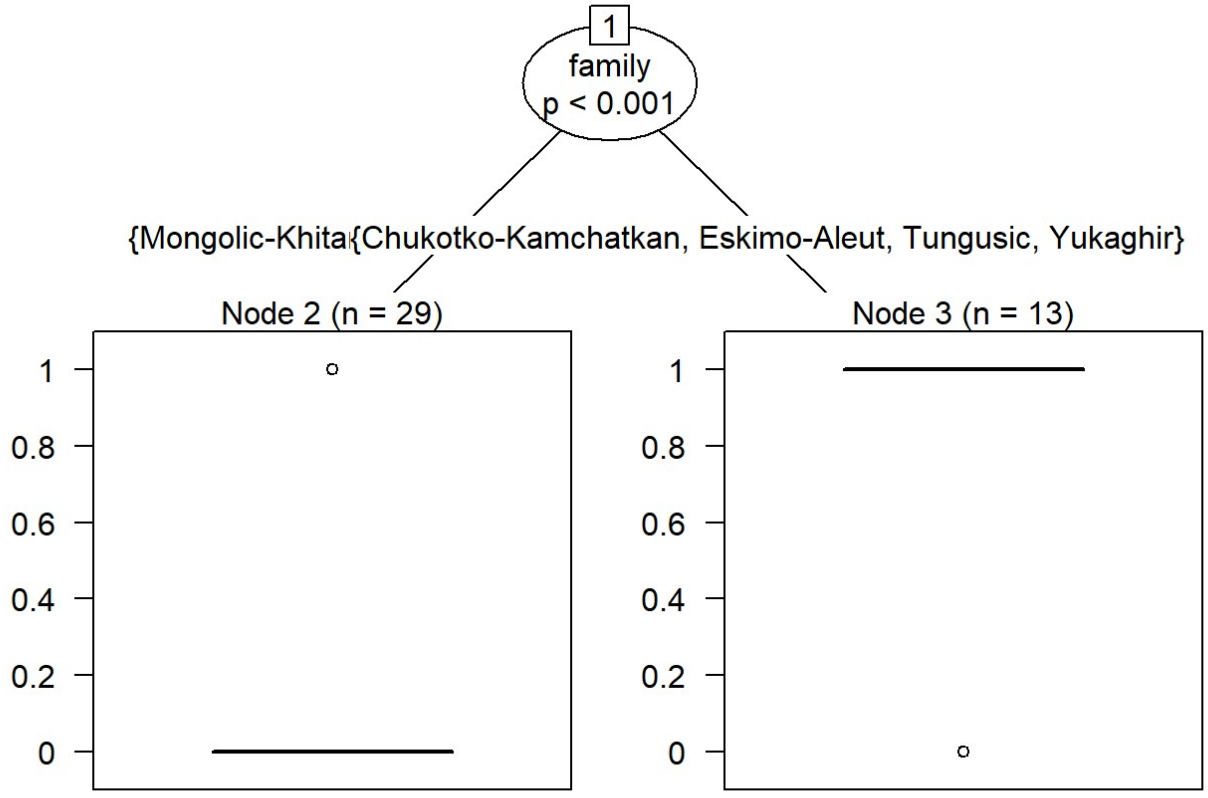

```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  46
##
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Yenisean, Yukaghir}; criteri
on = 0.995, statistic = 30.356
##   2)*  weights = 12
## 1) family == {Ainu, Indo-European, Japonic, Koreanic, Mongolic-Khitan, Nivkh, Turkic, U
ralic}
##   3)*  weights = 34
```




```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  42
##
## 1) family == {Indo-European, Koreanic, Mongolic-Khitan, Turkic, Uralic, Yenisean}; crit
erion = 0.993, statistic = 24.46
##   2)*  weights = 29
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Yukaghir}
##   3)*  weights = 13
```




```
##
##   Conditional inference tree with 3 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  48
##
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Uralic, Yenisean, Yukaghir};
criterion = 0.997, statistic = 32.146
##   2) is_siberian <= 0; criterion = 0.952, statistic = 7.069
##     3)*  weights = 9
##   2) is_siberian > 0
##     4)*  weights = 16
## 1) family == {Ainu, Japonic, Koreanic, Mongolic-Khitan, Nivkh, Sino-Tibetan, Turkic}
##   5)*  weights = 23
```




```
##
##   Conditional inference tree with 2 terminal nodes
##
## Response:  prolative_case
## Inputs:  family, is_siberian
## Number of observations:  42
##
## 1) family == {Mongolic-Khitan, Sino-Tibetan, Turkic, Uralic}; criterion = 1, statistic
= 27.89
##   2)*  weights = 29
## 1) family == {Chukotko-Kamchatkan, Eskimo-Aleut, Tungusic, Yukaghir}
##   3)*  weights = 13
```

```
prol_syb %>%
  count(family, prolative_case, is_siberian, .drop = FALSE) %>%
  filter(prolative_case & is_siberian) ->
  sib_prolatives

sib_prolatives
```

```
## # A tibble: 6 × 4
##   family              prolative_case is_siberian     n
##   <fct>               <lgl>          <lgl>       <int>
## 1 Chukotko-Kamchatkan TRUE           TRUE            3
## 2 Eskimo-Aleut        TRUE           TRUE            2
## 3 Tungusic            TRUE           TRUE            4
## 4 Uralic              TRUE           TRUE            4
## 5 Yenisean            TRUE           TRUE            1
## 6 Yukaghir            TRUE           TRUE            2
```

Посчитаем, сколько в семьях всех остальных языков (сибирские и в них нет явления, либо несибирские с наличием или отсутствием явления).

```
prol_syb %>%
  count(family, prolative_case, is_siberian, .drop = FALSE) %>%
  filter(!(prolative_case & is_siberian)) %>%
  group_by(family) %>%
  summarise(n.other = sum(n))  ->
  nonsib_or_noprolative

nonsib_or_noprolative
```

```
## # A tibble: 12 × 2
##    family          n.other
##    <fct>             <int>
##  1 Ainu                  1
##  2 Eskimo-Aleut          5
##  3 Indo-European         1
##  4 Japonic               1
##  5 Koreanic              1
##  6 Mongolic-Khitan       9
##  7 Nivkh                 1
##  8 Sino-Tibetan          1
##  9 Tungusic              3
## 10 Turkic               17
## 11 Uralic               12
## 12 Yenisean              1
```

После этого, совместим таблицы и сравним количество языков в семье, которые принадлежат сибирскому ареалу и в которых есть явление и количество всех остальных языков этой семьи.

```
sib_prolatives %>%
  left_join(nonsib_or_noprolative, by="family")
```

```
## # A tibble: 6 × 5
##   family              prolative_case is_siberian     n n.other
##   <fct>               <lgl>          <lgl>       <int>   <int>
## 1 Chukotko-Kamchatkan TRUE           TRUE            3      NA
## 2 Eskimo-Aleut        TRUE           TRUE            2       5
## 3 Tungusic            TRUE           TRUE            4       3
## 4 Uralic              TRUE           TRUE            4      12
## 5 Yenisean            TRUE           TRUE            1       1
## 6 Yukaghir            TRUE           TRUE            2      NA
```

Здесь ситуация в целом аналогична ситуации выше с временем на обладаемом: в семьях в целом преобладают интересующие нас языки — сибирские и с явлением. Их больше или столько же, сколько вообще всех остальных языков семьи в выборке. Про эскимосско-алеутские и уральские языки в целом укладываются в картину и будут подробнее обсуждены ниже. Разумные исключения здесь — чукотско-камчатские и юкагирские, среди которых нет несибирских языков (а вообще все языки с явлением).

```
prol_syb %>%
  filter(family %in% c("Chukotko-Kamchatkan", "Yukaghir"))
```

```
## # A tibble: 5 × 23
##   language          family      branch latitude longitude glottocode is_siberian
##   <chr>             <fct>       <chr>     <dbl>     <dbl> <chr>      <lgl>
## 1 Alutor            Chukotko-K… Chuko…     60.4      166. alut1245   TRUE
## 2 Chukchi           Chukotko-K… Chuko…     68.6      170. chuk1273   TRUE
## 3 Northern Yukaghir Yukaghir    <NA>       70.5      158. nort2745   TRUE
## 4 Southern Yukaghir Yukaghir    Kolym…     64.2      154. sout2750   TRUE
## 5 West Itelmen      Chukotko-K… Kamch…     56.0      156. itel1242   TRUE
## # ℹ 16 more variables: prolative_case <lgl>, ...9 <lgl>, ...10 <lgl>,
## #   ...11 <lgl>, ...12 <lgl>, ...13 <lgl>, ...14 <lgl>, ...15 <lgl>,
## #   ...16 <lgl>, ...17 <lgl>, ...18 <lgl>, ...19 <lgl>, ...20 <lgl>,
## #   ...21 <lgl>, ...22 <lgl>, ...23 <lgl>
```

```
prol_syb %>%
  count(family, prolative_case, is_siberian, .drop = FALSE) %>%
  filter(!(prolative_case & is_siberian)) %>%
  filter(family %in% c("Eskimo-Aleut", "Uralic"))
```

```
## # A tibble: 5 × 4
##   family       prolative_case is_siberian     n
##   <fct>        <lgl>          <lgl>       <int>
## 1 Eskimo-Aleut FALSE          FALSE           1
## 2 Eskimo-Aleut TRUE           FALSE           4
## 3 Uralic       FALSE          FALSE           8
## 4 Uralic       FALSE          TRUE            2
## 5 Uralic       TRUE           FALSE           2
```

Видим, что в эскимосско-алеутских в целом распространён пролатив. С другой стороны, в уральских пролатив встречается лишь в 2 несибирских языках и не встречается в 2 сибирских. В целом это укладывается в картину.

Обсуждение

Мы рассмотрели некоторые явления Сибири в связи с идеей о Сибири как ареале.

Для аттенуатива при рассмотрении мировой выборки и для пролатива и времени на обладаемом при рассмотрении уже небольших выборок видна связь наличия явления и того, сибирский ли язык.

Здесь мы прибегаем к более простым моделям, учитывающим прежде всего сам ареал. Однако для пролатива и времени на обладаемом можно говорить как минимум о стабильности (в терминах Nichols) этих явлений в Сибири. Здесь не так важно оказывается рассматривать генетическую принадлежность, поскольку видна распространённость явления в Сибири и меньшая распространённость вне её. То есть Сибирь оказывается островком стабильности для этого явления.

Тем не менее, генетическую близость мы также смогли учесть при анализе.

Мы пробовали строить решающее дерево для оценки важностей переменных. Но во многих случаях алгоритм решающего дерева строил пень, дерево глубиной 1, а предикат был тривиальным: “семья находится среди семей с пролативом”.