Дополнительные материалы к статье

Т.Д. Зинченко, В.К. Шитиков, Л.В. Головатюк. ДОННЫЕ СООБЩЕСТВА И АБИОТИЧЕСКИЕ ФАКТОРЫ: АНАЛИЗ СТАТИСТИЧЕСКОЙ СВЯЗИ С ИСПОЛЬЗОВАНИЕМ ИНДЕКСА НЕСТАБИЛЬНОСТИ И МЕТОДА ВИРТУАЛЬНЫХ ВИДОВ

// Журнал Сибирского федерального университета. Биология

1. Загрузка и анализ исходных данных

Таблицы исходных данных были сформированы по результатам многолетнего изучения донных сообществ малых и средних рек бассейна Средней и Нижней Волги, а также мониторинга абиотических факторов в этом регионе. Подробности формирования таблиц, их характеристики и результаты статистического анализа иными методами представлены как в материалах статьи, так и предыдущих сообщениях блога:

- https://stok1946.blogspot.com/2020/09/blog-post.html - **Интерполяция и визуализация пространственных данных;**
- https://stok1946.blogspot.com/2020/11/sdm.html - **Модели пространственного распределения видов.**

Комплект исходных данных по теме статьи включает три таблицы:

- `TAXA` – усредненные по количеству сделанных проб численности 147 видов, наблюдаемых в 132 реках (участках) изучаемого региона;
- `VAR` – список из 132 рек, включающий их наименование, тип, географические координаты, а также 8 отобранных гео-, метео- и гидрохимических показателей;
- `Species147` – список из наименований 147 видов, включая принадлежность к подсемействам (`Com` – отмеченная встречаемость в реках из обследованных 132).

Таблицы размещены в файле "InStab_dat.RData", который необходимо загрузить с общедоступного ресурса http://www.ievbras.ru/ecostat/Kiril/R/Blog/InStab_dat.RData и поместить в рабочий каталог среды R:

```
#  Загрузка исходных данных из файла
load(file="InStab_dat.RData")
ls()
[1] "Species147" "TAXA"       "VAR"
#  Список анализируемых видов
head(Species147)
      Code Com           Наименование Подсемейство
1 AtAth.ib  21           Atherix ibis  Athericidae
2 BiEug.a.  27      Euglesa acuminata   Euglesidae
3 BiEug.sp  71            Euglesa sp.   Euglesidae
4 BiHen.h.  27 Henslowiana henslowana   Euglesidae
5  BiPis.a  33       Pisidium amnicum   Pisidiidae
6 BiPis.i.  17      Pisidium inflatum   Pisidiidae
#  Список обследованных рек
head(VAR[,1:9])
        Name    Type        X        Y NamRiver Ground MTemp PrecDQ Alt
1    Актушка   малая 49.00051 53.41473     Акту      1    47     82  98
2     Аманак   малая 51.98063 53.74324     Аман      3    43     88  88
3       Анлы   малая 52.33430 53.93778     Анлы      5    36     91 193
4 Б. Вязовка   малая 50.18092 52.52049     Б.Вя      5    47     86 110
5 Б. Глушица   малая 50.84062 52.15519     Б.Гл      3    50     74  66
6   Б. Иргиз верхнее 50.72543 52.26324   Б.Ир_в      3    50     77  64
```

Распределение численностей сильно асимметрично (левое плечо меньше правого в 1000 раз), поэтому выполняем преобразование исходных значений по алгоритму «Хи-квадрат», рекомендуемому П.Лежандром с соавторами и представленному в пакете `vegan`.

```
#    Преобразование значений средних численностей
quantile(TAXA[TAXA!=0])
        0%        25%        50%        75%       100%
    1.0000    30.0000    80.0000   203.6667 81143.3333
library(vegan)
TAXA.Ch <- decostand(TAXA, method="chi.square")
quantile(TAXA.Ch[TAXA.Ch!=0])
          0%          25%          50%          75%         100%
6.484131e-04 6.697773e-02 1.569950e-01 3.699117e-01 1.601506e+01
```

Определим комплект исходных данных в виде набора четырех объектов (таблицы, векторов и переменной), которые будут ниже использоваться в последовательности шагов вычислений. Для выполнения расчетов по любым иным данным необходимо в правую часть выражений подставить другие соответствующие по смыслу компоненты:

```
data <- cbind(VAR[,-(1:4)], TAXA.Ch)  # Объединенная таблица данных
Samples <- "NamRiver"           # Поле со списком рек
Taxa <- colnames(TAXA.Ch)       # Список видов
(Variables <- colnames(VAR[,-(1:5)]))  #  Список абиотических переменных
[1] "Ground" "MTemp"  "PrecDQ" "Alt"    "TRI"    "Miner"  "NH4"    "O2"
```

Уточним также список обозначений абиотических показателей: метеорологических – среднегодовая температура `MTemp`, осадки самого засушливого квартала `PrecDQ`; геоморфологических – высота `Alt`, индекс шероховатости рельефа `TRI`; гидрохимических – минерализация воды `Miner`, аммонийный азот `NH4`, насыщение кислородом `O2` и категории грунтов `Ground`

2. Расчет индексов нестабильности и анализ их важности

Блоки кода R, использованные ниже, являются фрагментами скрипта, присланного нам К.Гисандом частным письмом и относящегося к функции FCA из пакета, находящегося на стадии отладки. Первый шаг скрипта выполняет проверку комплектности данных и стандартизацию переменных на интервале 0-1.

```
###  Должно быть больше одного вида
if(length(Taxa)==1){
stop("The algorithm only works with a community, so more than one species")
                    }
### Пропущенные значения для видов конвертируются в 0
datosT<-data.frame(subset(data, select=Taxa))
datosT[is.na(datosT)]<-0
###   Проверка и отбор заданных переменных
datosT<-data.frame(subset(data, select=Samples), datosT, subset(data,
                    select=Variables))
datos<-na.exclude(datosT)
remove(datosT)
### Стандартизация данных от 0 до 1
selection <-datos[,-1]
a <- dim(selection)
datosE <- selection
for (z in 1:a[2]){
    matrixE <- matrix(c(0, 1, min(selection[,z],na.rm=TRUE),
               max(selection[,z],na.rm=TRUE)), nrow = 2 , ncol = 2)
    reg <- lm(matrixE[,1] ~ matrixE[,2])
    datosC <- reg$coefficients[1] + selection[,z]*reg$coefficients[2]
    datosE <- cbind(datosE,datosC)
    }
datosE <- datosE[,-c(1:a[2])]
colnames(datosE) <- colnames(selection)
datos <- cbind(datos[,1],datosE)
names(datos) <- c(Samples,names(selection))
```

```
remove(datosE)
remove(selection)
```

Индекс нестабильности I_{ij} каждого вида i (i = 1…147) на каждом участков отбора проб j , j = *1…m, m* = 132, рассчитывается по формуле дивергенции энтропии Кульбака-Лейблера

$$I_{ij} = p_{ij} \log_2 p_{ij} / p_{im},$$

где p_{ij} – доля численности вида i на j-м участке по отношению к его сумме на всех участках, p_{im} – средняя доля численности вида i на всех m участках.

```
###  Расчет индексов нестабильности для видов
datosE <- data.frame(subset(datos, select=Taxa))
a <- dim(datosE)
for (z in 1:a[2]){
    sum <- sum(datosE[,z],na.rm=TRUE)
    datosE[,z] <- datosE[,z]/sum
   }
media <- apply(X = datosE , MARGIN = 2 , FUN = mean , na.rm=TRUE)
for (z in 1:a[2]){
   val <- datosE[,z]/media[z]
   datosE[,z] <- abs(datosE[,z]*log(val,base=2))
   }
suma1 <- apply(X = datosE , MARGIN = 1 , FUN = sum , na.rm=TRUE)
final2 <- cbind(datos[,1],datosE)
names(final2)<-c(Samples,Taxa)
remove(datosE)
```

Индекс нестабильности I_{ij} каждой абиотической переменной i (i = 1…8) на каждом участков отбора проб j , j = *1…m, m* = 132, рассчитывался по формуле дивергенции энтропии Кульбака-Лейблера

$$I_{ij} = p_{ij} \log_2 p_{ij} / p_{im},$$

где p_{ij} – доля значений показателя i в его сумме на всех участках, p_{im} – среднее p_{ij} для m участков.

```
###  Расчет индексов нестабильности для переменных
datosE <- data.frame(subset(datos, select=Variables))
a <- dim(datosE)
for (z in 1:a[2]){
   sum <- sum(datosE[,z],na.rm=TRUE)
   datosE[,z]<-datosE[,z]/sum
   }
media <- apply(X = datosE , MARGIN = 2 , FUN = mean , na.rm=TRUE)
for (z in 1:a[2]){
   val <- datosE[,z]/media[z]
   datosE[,z] <- abs(datosE[,z]*log(val,base=2))
   }
suma2 <- apply(X = datosE , MARGIN = 1 , FUN = sum , na.rm=TRUE)
final3 <- cbind(datos[,1],datosE)
names(final3) <- c(Samples,names(datosE))
remove(datosE)
```

Внимание: При расчете индексов имеют место несколько сообщений об ошибке

```
In log(val, base = 2): created NaN
```

Поскольку данные стандартизируются от 0 до 1, то каждый вектор должен иметь хотя бы один 0. Это - вполне допустимое значение, которое здесь просто теряется при попытке его прологарифмировать. Необходимо изменить способ стандартизации данных при последующей доработке метода. Но пока надо освободиться от символов недопустимых значений.

```
### Конвертация в 0 отсутствующих значений
final2[is.na(final2)] <- 0
final3[is.na(final3)] <- 0
final1 <- data.frame(datos[,1], suma1,suma2)
names(final1) <- c(Samples, "InstabilityTaxa", "InstabilityEnvironment")
final1[is.na(final1)] <- 0
```

Результаты расчетов помешаются в 5 текстовых (`csv`) файлов, что дает возможность дополнительного их анализа, например, средствами Excel. В файлы `file2` и `file3` загружаются оба подмножества индексов – по видам и по физическим переменным, а в файл `file1` - суммарные их значения для каждой реки:

```
### Сохранение индексов нестабильности в файлах
na = "NA"
row.names = FALSE
file1 = "Instability indices.csv"
file2 = "Instability of each taxon.csv"
file3 = "Instability of environmental variables.csv"
file4 = "Contribution of variables to instability of taxa.csv"
file5 = "Contribution of variables to instability of environment.csv"
write.csv(x=final1,file = file1, fileEncoding = "", row.names=row.names,
                      na=na)
write.csv(x=final2,file = file2, fileEncoding = "", row.names=row.names,
                      na=na)
write.csv(x=final3,file = file3, fileEncoding = "", row.names=row.names,
                      na=na)
Ind <- read.csv(file = file1)
    NamRiver InstabilityTaxa InstabilityEnvironment
1       Акту       3.9370756              0.01785801
2       Аман       2.3265259              0.03011839
3       Анлы      10.9172690              0.05426322
4       Б.Вя       3.1911714              0.03207529
5       Б.Гл       1.9488967              0.02412576
6     Б.Ир_в       5.2038865              0.01877871
```

Теперь зададимся следующим вопросом: на нестабильность каких видов сильнее воздействует нестабильная величина абиотического фактора? Или тот же вопрос в другой плоскости: как нестабильность того или иного вида связана с нестабильностями переменных среды. Напомним, что под нестабильностью мы понимаем исключительно степень отклонения рассматриваемого показателя от его среднего значения.

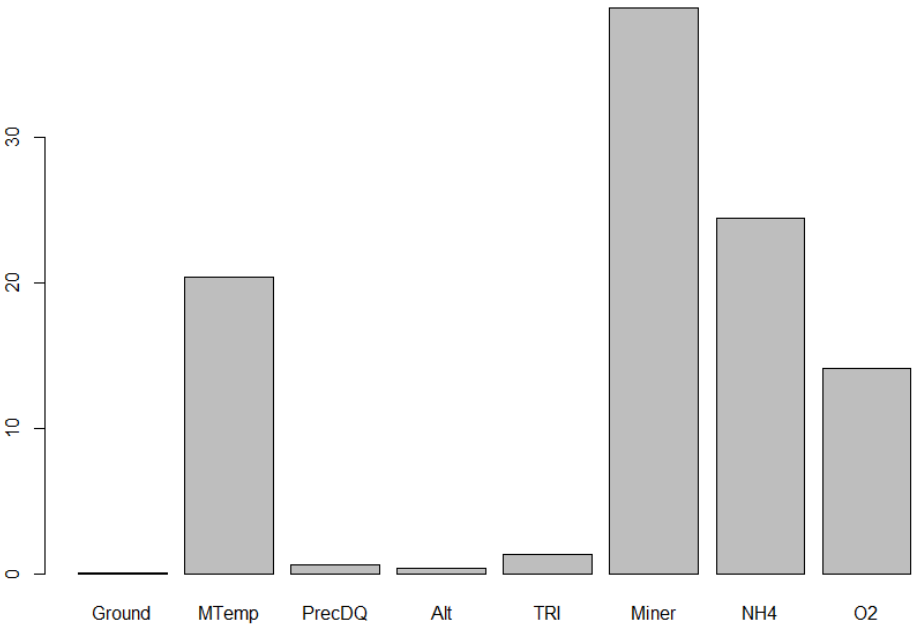
Для оценки связи между подмножествами индексов I_{TAXA} и I_{VAR} будем строить модели множественной регрессии методом случайного леса (*Random Forest*), в результате чего для каждого таксона устанавливается величина вклада W_i каждого фактора среды, пропорционального его важности (*importance*). Оценку последней рассчитывали по среднему снижению точности предсказания на оставшихся данных после исключения тестируемого показателя.

```
library (randomForest)
data <- cbind(final2[,-1],final3[,-1])
lenv <- length(Variables)
lens <- length(Taxa)
# Оценка важности для таксонов
data1 <- data[,Variables]
envvar <- names(data1)
zz=1
for(zz in 1:lens){
  sp <- Taxa[zz]
  data2 <- data.frame(data[,sp], data1)
  names(data2) <- c("VAR", envvar)
  data2 <- na.exclude(data2)
```

```
  ### Random forest
  reg <- randomForest(VAR ~ ., data = data2, importance = TRUE)
  vari <- reg$importance[,1] *100/sum(reg$importance[,1] )
  if(zz==1){
      Importance<-data.frame(Variables,vari)
  }
  else{
      Importance<-cbind(Importance,vari)
  }
}   ###  Конец цикла для списка видов
colnames(Importance)[2:ncol(Importance)] <- Taxa
head(Importance[,2:8])
       AtAth.ib  BiEug.a.  BiEug.sp   BiHen.h.    BiPis.a  BiPis.i.   BiPis.sp
Ground 109.795129 10.983470  5.086206   7.858105  0.8866538  9.796841 -16.602552
MTemp    3.445119  5.306429 22.828758  16.170710 13.5169851  3.297726  61.826886
PrecDQ   7.173756 28.297486  7.364909  57.781996  8.1629842  7.132784  -9.136864
Alt     80.529183 16.621699 -6.556878  17.851142  0.9693920 33.992806  30.664416
TRI    -12.041435  8.334923 15.223465  30.983760  2.5874728 -2.149690   1.118508
Miner   -6.874645 26.235657  1.143253 -18.459229 33.7994179 39.854502 -28.613630
###  Сохраняем таблицу важностей для списка видов/переменных
write.csv(x = Importance, file = file4, fileEncoding = "",
                                        row.names=row.names, na=na)
```

Оценим теперь, каков средний вклад каждой из 8 абиотических переменных в общую нестабильность на каждом участке рек.

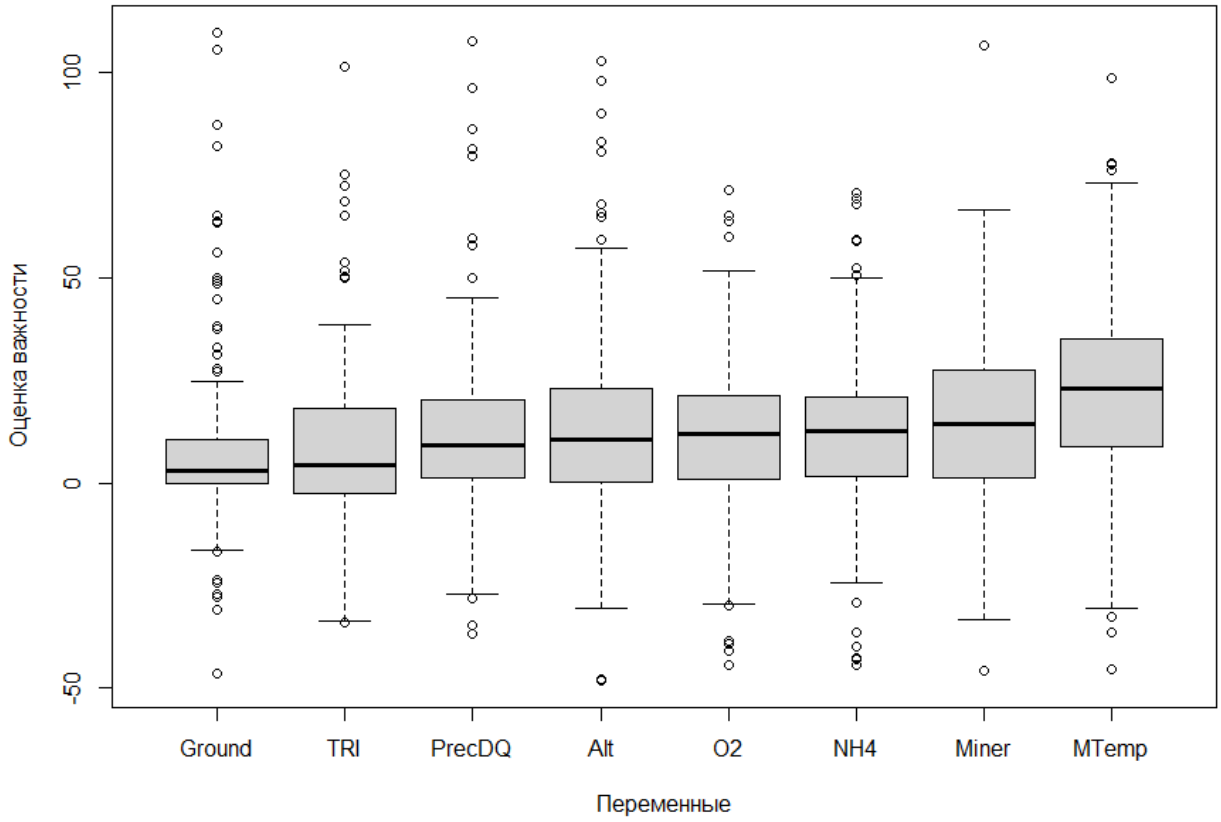
```
###  Оценка важности для переменных среды
Datav <- data.frame(final1[,3],final3[,-1])
names(datav) <- c("Inst",Variables)
reg <- randomForest(Inst ~ ., data = datav, importance = TRUE)
(mat <- abs(reg$importance[,1]) *100/sum(abs(reg$importance[,1] )))
   Ground     MTemp    PrecDQ       Alt       TRI     Miner       NH4        O2
 0.026548 20.360633  0.608919  0.353304  1.321295 38.839534 24.400016 14.089748
barplot(height=mat)
```




```
mat <- data.frame(names(mat),mat)
names(mat) <- c("Variables", "Contribution")
write.csv(x = mat, file = file5, fileEncoding = "", row.names=row.names,
                                        na=na)
```

Рассмотрим также характер распределения важности каждой абиотической переменной относительно нестабильности видов на каждом участке рек.

```
### Оценка важности каждой переменной для каждого вида
for(zz in 2:(lens+1)){
  Value <- Importance[,zz]
  Box <- data.frame(Variables,Value)
  if(zz==2)  datos <- box else datos <- rbind(datos,box)
}
#  Отбраковка очень больших и малых значений
datos[datos$Value < (-50)  & complete.cases(datos),] <- NA
datos[datos$Value > 110  & complete.cases(datos),] <- NA
#  Отрисовка графика бокс с усами
varX <- "Variables"
varY = "Value"
datos[,varX] <- factor(datos[,varX], levels = unique(datos[,varX]),
                       labels = unique(datos[,varX]))
median <- tapply(datos[complete.cases(datos), varY],
                 datos[complete.cases(datos), varX], median)
LabelCat <- names(sort(median, decreasing=FALSE))
datos[,varX] <- factor(datos[,varX], levels = LabelCat, labels = LabelCat)
XLABE = "Переменные"
YLABE = "Оценка важности"
boxplot( datos[,varY]~datos[,varX], xlab=XLABE, ylab=YLABE,
                                                col = "lightgrey" )
```

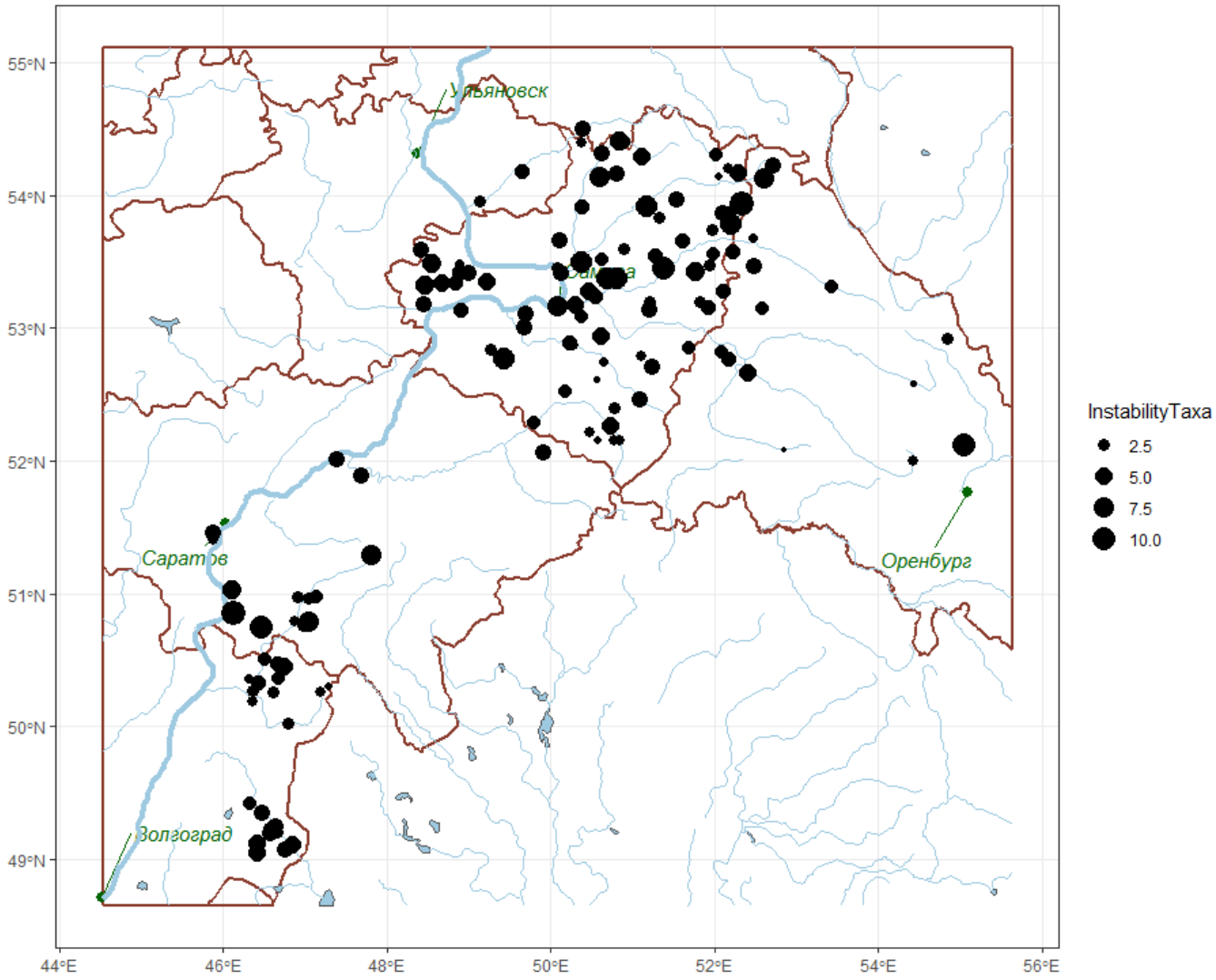


Внимание: Вычисленные нормированные оценки важности в обычной интерпретации являются положительными числами от 0 до 100%. Однако при выполнении расчетов величина `importance` для некоторых переменных среды неожиданно принимала отрицательные значения (т.е. средняя точность предсказания после исключения тестируемого показателя не снижается, а наоборот, увеличивается). При дальнейшей доработке метода необходимо внимательно рассмотреть обоснованность такой ситуации.

3. Использование индексов нестабильности для картографирования

В предыдущем сообщении блога https://stok1946.blogspot.com/2020/09/blog-post.html была сформирована компьютерная карта региона для визуализации результатов интерполяции. Объект ggplot2, воспроизводящий эту карту, представлен в файле по адресу: http://www.ievbras.ru/ecostat/Kiril/R/Blog/WB_map.RData. Его надо скачать и поместить в рабочий каталог R.

```
# Добавляем к суммарным значениям индесов географические координаты
Ind$X <- VAR$X
Ind$Y <- VAR$Y
# Грузим слои карты региона
load(file="WB_map.RData")
library(ggplot2)
Basemap +
  geom_point(data = Ind, aes(x = X , y = Y, size=InstabilityTaxa))  +
  theme_bw()
```



На построенной карте диаметр кружков в каждом районе отбора гидробиологических проб пропорционален суммарному индексу нестабильности всех обнаруженных там видов.

Использование метода виртуальных видов для оценки индексов пригодности среды обитания ($H \in [0,1]$ – *environmental suitability*) и его последующая картографическая интерпретация в виде ареалов подробно описана в разделе 4 сообщения https://stok1946.blogspot.com/2020/11/sdm.html. Там приведен подробный пример на основе *Procladius ferrugineus,* но вычислительные процедуры построения модели для Prodiamesinae используются совершенно аналогично.