

# Анализ статистической связи структуры экологических сообществ с абиотическими факторами на основе индексов нестабильности

Автор: Владимир Шитиков  
<https://stok1946.blogspot.com/>

## 1. Введение

Современной парадигмой количественной оценки связи между структурой сообществ и окружающей средой является моделирование пространственно-временного распределения видов в природе (SDM – *Species Distribution Models*, Peterson et al., 2011). За последние 20 лет этой тематике посвящены сотни публикаций – см. обзоры (Franklin, 2009; Guisan et al., 2017; Лисовский и др., 2020) –, однако до сих пор в полной мере отсутствует не только единая теория, но и конкретные практические рекомендации. Это обусловлено как объективно существующим многообразием изучаемых экологических сообществ, природно-климатических зон, жизненных форм и техник проведения наблюдений, так и большим арсеналом разработанных методов компьютерной обработки и верификации моделей, выбор которых в значительной мере определяется субъективными взглядами исследователей. В частности, развернутый анализ результатов использования 33 моделей SDM на сообществах птиц, бабочек, деревьев и травянистой растительности показал (Norberg et al., 2019), что успех моделирования зависит от природы изучаемого сообщества на 36%, постановки задачи (интерполяция или экстраполяция) на 26%, выбранного алгоритма на 33% и объема выборки – только на 2%.

Построение моделей SDM требует высокой квалификации аналитика. В частности, в экологических исследованиях для большинства организмов и ситуаций достоверно можно оценить только наличие того или иного вида в точках отбора проб, тогда как надежные данные об "ареалах отсутствия" не могут быть эмпирически обоснованы. Это обусловило разработку различных алгоритмов, основанных на многократном случайном выборе подмножества точек, где, как предполагается, вид отсутствует (*pseudo-absence*, или "*background*" *points*). Из таких алгоритмов можно отметить часто используемый метод максимальной энтропии, реализованный в программе MaxEnt (Phillips et al., 2006). Однако показано (Hastie, Fithian, 2013), что эта процедура оценивает не столько искомую вероятность присутствия вида, сколько неоднородность использованных эмпирических данных, а местоположение и количество точек "псевдо-отсутствия" вида сильно зависят от протяженности и характера доступной географической территории. Отмечается также (Лисовский, Дудов, 2020), что работа алгоритма во многом зависит от выбора формы частных функций, предварительной фильтрации исходных данных, использования слоев коррекции и др., а в отсутствие объективных критериев проверки результатов прогноза кажущаяся успешность моделирования может легко ввести в заблуждение.

Перечисленные проблемы побудили разработку математически "простых" процедур, позволяющих выбрать минимальное количество некоррелированных факторов среды, определяющих с высокой вероятностью экологические предпочтения вида, роль которых также может быть обоснована и биологическими причинами. Например, экологический анализ факторов ниши (ENFA; Hirzel et al., 2002) основан на оценке маргинальности  $M$  (то есть экологического расстояния между оптимумом вида, полученным из данных его изучения, и средним значением в пределах эталонной области) и специализации  $S$  (отношения глобальной дисперсии анализируемого фактора к дисперсии в точках, где обнаружен вид).

Сходный смысл имеет индекс флуктуации Дюбуа (Dubois, 1973), который был в различных вариантах модифицирован под названием индекса нестабильности  $I$

(Guisande et al., 2006, 2017). Применение этого индекса для отбора подмножества информативных предикторов позволяет свести к минимуму влияние коллинеарности на результат любых моделей SDM, но целенаправленно он используется в составе модулей вычислительной среды ModestR (García-Roselló et al., 2014). Этот свободно распространяемый пакет программ обеспечивает комплексную обработку данных о встречаемости видов и абиотических факторов, включая доступ к всемирным базам географической, климатической и ресурсной информации и создание комплекта тематических карт. Включенная в его состав процедура NOO (*niche of occurrence* – García-Roselló et al., 2019) выполняет интуитивно понятное моделирование вероятностного распределения видов на основе выборочных данных об их присутствии, блокируя попытки выхода прогнозов за пределы доступной области SDM. Важным компонентом ModestR является база по пресноводным ресурсам мира, на основании которой построены карты распределения рыбных сообществ (González-Vilas et al., 2016).

В настоящем сообщении мы проводим анализ результатов апробации индекса нестабильности на примере донных сообществ малых и средних рек в бассейне Средней и Нижней Волги для ранжирования факторов среды и оценки их роли в формировании экологических предпочтений различных таксонов гидробионтов.

## **2. Сбор и предварительная обработка 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. Всего было собрано 1400 проб с идентификацией 740 видов и таксонов рангом выше вида, которые для дальнейшего анализа были ограничены списком из 147 видов встретившихся не менее, чем в 15 пробах или на 10 участках из 132.

На станциях отбора проб параллельно проводился мониторинг 30 факторов среды, включающих гидрологические параметры водотоков, показатели качества воды и содержание основных химических ингредиентов (состав донных грунтов, насыщение воды кислородом, минерализация и др.). Растровые таблицы, содержащие основные метеорологические и геоморфологические показатели для региона исследований с разрешением 2.5', загружались с сервера свободно распространяемой информации WorldClim.

Поскольку между всеми этими переменными наблюдалась сильная корреляционная связь, проводили анализ индексов инфляции дисперсии VIF (*Variance Inflation Factor* – Джеймс и др., 2016) и избыточные предикторы со значениями VIF более 30 удалялись из рассмотрения. В результате было отобрано 8 факторов среды, коллинеарность которых считалась приемлемой: среднегодовая температура MTemp, °C; осадки самого засушливого квартала PrecDQ, мм; высота Alt, м; индекс шероховатости рельефа TRI; минерализация воды Miner, мг/л; аммонийный азот NH<sub>4</sub>, мг/л; насыщение кислородом O<sub>2</sub>, мг/л и категория грунтов Ground в баллах от 1 до 6.

Подробности формирования таблиц исходных данных, некоторые их характеристики и результаты статистического анализа иными методами представлены в предыдущих сообщениях нашего 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 сообщения включает три таблицы:

- ТАХА – 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](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" "ТАХА"      "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
```

В таблице VAR абиотические показатели соответствуют следующим именам столбцов: *метеорологические* – среднегодовая температура MTemp, осадки самого засушливого квартала PrecDQ; *геоморфологические* – высота Alt, индекс шероховатости рельефа TRI; *гидрохимические* – минерализация воды Miner, аммонийный азот NH<sub>4</sub>, насыщение кислородом O<sub>2</sub> и категории грунтов Ground.

Необходимо отметить, что распределение удельной численности (экз/м<sup>2</sup>) видов макрозообентоса сильно асимметрично относительно среднего: левое плечо меньше правого в 1000 раз. Поэтому выполним предварительно нормирующее преобразование, с использованием  $\chi^2$ -дистанции, рекомендуемое П.Лежандром с соавторами (Legendre, Gallagher, 2001) и представленное в пакете vegan. Эта трансформация данных позволяет достичь, по всей вероятности, наиболее разумного компромисса при учёте, как роли ведущих компонент, так и вклада редких или малочисленных таксонов.

```
# Преобразование значений средних численностей
quantile(ТАХА[ТАХА!=0])
  0%      25%      50%      75%      100%
1.0000  30.0000  80.0000  203.6667  81143.3333
library(vegan)
ТАХА.Ch <- decostand(ТАХА, method="chi.square")
quantile(ТАХА.Ch[ТАХА.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"
```

### 3. Расчет индексов нестабильности и анализ их важности

Индекс нестабильности<sup>1</sup>  $I_{ij}$  произвольного показателя  $x_i$  для каждого участка отбора проб  $j$ ,  $j = 1 \dots m$ , рассчитывается по формуле дивергенции энтропии Кульбака-Лейблера

$$I_{ij} = p_{ij} \log_2 p_{ij} / p_{im}, \quad (1)$$

где  $p_{ij}$  – доля  $x_{ij}$  в сумме значений этого показателя на всех участках,  $p_{im}$  – среднее  $p_{ij}$  для  $m$  участков (Guisande et al., 2006). Значение  $I_{ij}$  велико при больших отклонениях  $x_i$  от своего среднего и равняется 0 при  $p_{ij} = p_{im}$ .

Индексы нестабильности вычисляются как для каждого таксона TAXA, так и каждой переменной VAR, а их суммы  $I_{TAXAj}$  и  $I_{VARj}$  отражают соответственно нестабильность донного сообщества и нестабильность воздействия факторов окружающей среды, наблюдаемые на участке  $j$ .

Блоки кода R, использованные ниже, являются фрагментами скрипта, присланного нам К.Гисандом частным письмом, и относятся 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)
```

<sup>1</sup> Использование термина "нестабильность" в этом контексте следует признать неудачным, поскольку традиционно в экологии под нестабильностью сообщества или факторов окружающей среды понимается их неустойчивость или изменчивость во времени. Однако прямой перевод *Instability index* не оставляет иных возможностей.

Индекс нестабильности  $I_{ij}$  каждого вида  $i$  ( $i = 1 \dots 147$ ) на каждом участках отбора проб  $j$ ,  $j = 1 \dots m$ ,  $m = 132$ , рассчитывался по формуле (1), 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))
}
sumal <- 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  $i$  ( $i = 1 \dots 8$ ) на каждом участках отбора проб  $j$ ,  $j = 1 \dots m$ ,  $m = 132$ , рассчитывался по формуле (1), 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 имело 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], sumal, suma2)
names(final1) <- c(Samples, "InstabilityTaxa", "InstabilityEnvironment")
final1[is.na(final1)] <- 0
```

Результаты расчетов помещаются 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 абиотическим 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 понимается исключительно степень отклонения рассматриваемого показателя от его среднего значения.

Для оценки связи между подмножествами индексов  $I_{\text{TAXA}}$  и  $I_{\text{VAR}}$  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.

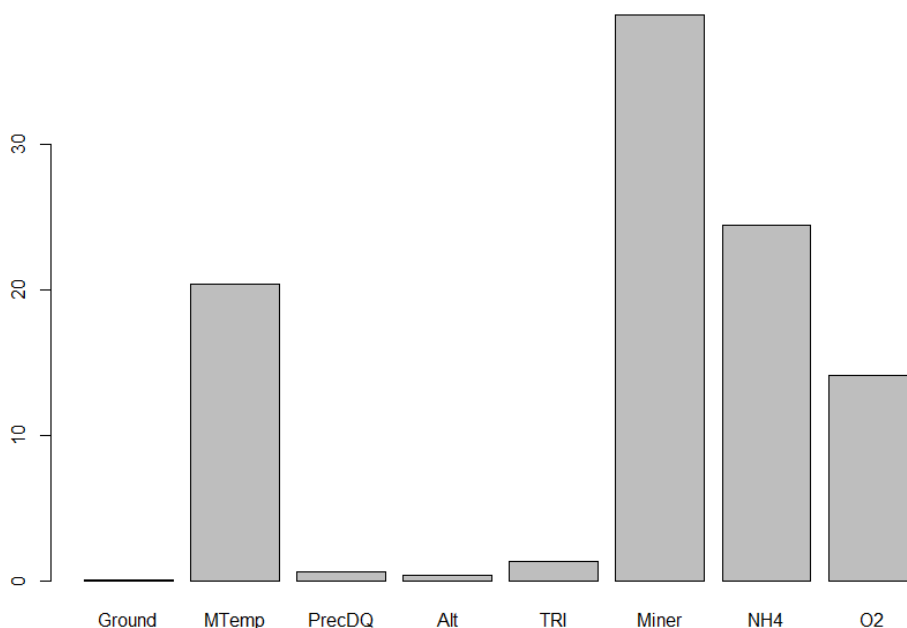
```
library (randomForest)
data <- cbind(final2[, -1], final3[, -1])
lenv <- length(Variables)
lens <- length(Taxa)
# Оценка важности для таксонов
data1 <- data[, Variables]
envvar <- names(data1)
zz=1
for(zz in 1:lens){
  sp <- Taxa[zz]
  data2 <- data.frame(data[, sp], data1)
  names(data2) <- c("VAR",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

Однако если проанализировать все 147 частных регрессионных моделей, связывающих подмножества индексов  $I_{TAXA}$  и  $I_{VAR}$  и обобщающих влияние факторов среды на нестабильность численности каждого из видов донных сообществ, то обнаруживается существенно более сложная картина. Оценки вклада  $W$  (%) абиотических переменных в вариацию обилия некоторых таксонов макрозообентоса, представленные в табл. 1, показывают, что экологические предпочтения видов чаще всего определяются не одним каким-то фактором, а их различными нестационарными комбинациями. В частности, отдельные виды имеют высокую статистическую зависимость от количества осадков или геоморфологических показателей.

**Таблица 1.** Оценки вклада  $W$  (%) в нестабильность численности некоторых таксонов макрозообентоса различных факторов окружающей среды ( $m$  – количество проб, в которых встретился вид, из общего числа 1400)

Наименование	Подсемейство	$m$	MTe mp	PrecD Q	Alt	TRI	Miner	NH4	O2	Group nd
<i>Euglesa acuminata</i>	Euglesidae	27	10.82	29.98	15.02	1.65	23.10	8.05	5.20	6.18
<i>Pisidium amnicum</i>	Pisidiidae	33	18.89	8.68	0.29	1.77	22.99	33.85	11.04	2.50
<i>Ablabesmyia monilis</i>	Tanypodinae	78	21.93	3.42	11.48	2.83	-2.06	16.48	39.47	6.46
<i>Chironomus salinarius</i>	Chironomini	148	32.66	12.36	-0.46	-1.60	22.39	17.84	12.62	4.18
<i>Cladotanytarsus mancus</i>	Tanytarsini	323	5.57	14.02	-9.26	16.96	19.89	32.41	13.29	7.13
<i>Cricotopus salinophilus</i>	Orthoclaadiinae	211	21.71	-1.83	0.23	2.65	32.70	41.21	3.89	-0.55
<i>Dicrotendipes nervosus</i>	Chironomini	90	20.19	7.20	3.14	-0.14	40.89	19.45	9.22	0.06
<i>Paratendipes albimanus</i>	Chironomini	61	33.37	0.05	51.33	3.22	6.52	2.12	3.13	0.25
<i>Polypedilum nubeculosum</i>	Chironomini	448	6.30	8.05	29.58	1.87	31.89	9.68	11.55	1.09
<i>Procladius ferrugineus</i>	Tanypodinae	395	22.89	15.78	11.93	-6.85	29.63	5.15	7.07	14.41
<i>Psectrocladius sordidellus</i>	Orthoclaadiinae	75	36.88	0.42	2.97	-8.96	17.81	26.38	21.63	2.86
<i>Paratanytarsus confusus</i>	Tanytarsini	123	32.35	15.59	-2.49	-0.99	13.10	18.69	23.23	0.52
<i>Stictochironomus crassiforceps</i>	Chironomini	112	37.54	12.53	2.22	-9.19	10.18	12.26	16.22	18.24
<i>Tanypus punctipennis</i>	Tanypodinae	112	21.23	13.24	18.44	27.72	3.29	-1.40	11.65	5.82
<i>Haliphus ruficollis</i>	Haliplidae	29	38.18	-3.89	3.21	-1.69	13.03	29.61	18.12	3.44
<i>Asellus aquaticus</i>	Asellidae	86	31.19	11.31	-3.88	-0.56	10.00	6.07	39.69	6.17
<i>Cloeon dipterum</i>	Baetidae	35	46.61	5.98	25.40	12.85	-1.88	7.57	4.03	-0.55
<i>Caenis robusta</i>	Caenidae	59	48.90	6.53	10.25	15.31	7.72	5.04	4.03	2.21
<i>Ephemerella ignita</i>	Ephemerellidae	32	-0.27	56.81	-5.96	23.75	2.58	1.74	17.70	3.65
<i>Plea minutissima</i>	Pleidae	27	29.99	0.51	2.17	1.21	21.07	12.56	32.53	-0.04
<i>Sialis sordida</i>	Sialidae	30	26.37	5.70	10.34	0.89	12.95	9.54	35.26	-1.05
<i>Gomphus vulgatissimus</i>	Gomphidae	30	21.60	6.66	20.16	0.19	8.92	25.27	17.37	-0.19
<i>Ischnura elegans</i>	Coenagrionidae	20	52.74	0.41	34.77	8.01	4.08	7.48	-7.41	-0.09
<i>Limnodrilus hoffmeisteri</i>	Tubificidae	444	34.59	1.42	0.22	2.24	24.79	23.72	10.79	2.23
<i>Stylaria lacustris</i>	Naididae	56	39.18	10.00	4.57	16.48	13.09	8.49	7.49	0.69
<i>Tubifex tubifex</i>	Tubificidae	272	2.95	-2.97	9.44	21.25	28.82	9.93	9.18	21.41
<i>Ecnomus tenellus</i>	Ecnomidae	61	25.14	15.18	6.92	3.81	9.73	26.28	10.48	2.45
<i>Prodiamesa sp.</i>	Prodiamesinae	227	21.78	18.84	36.05	8.16	6.70	1.21	1.11	6.15

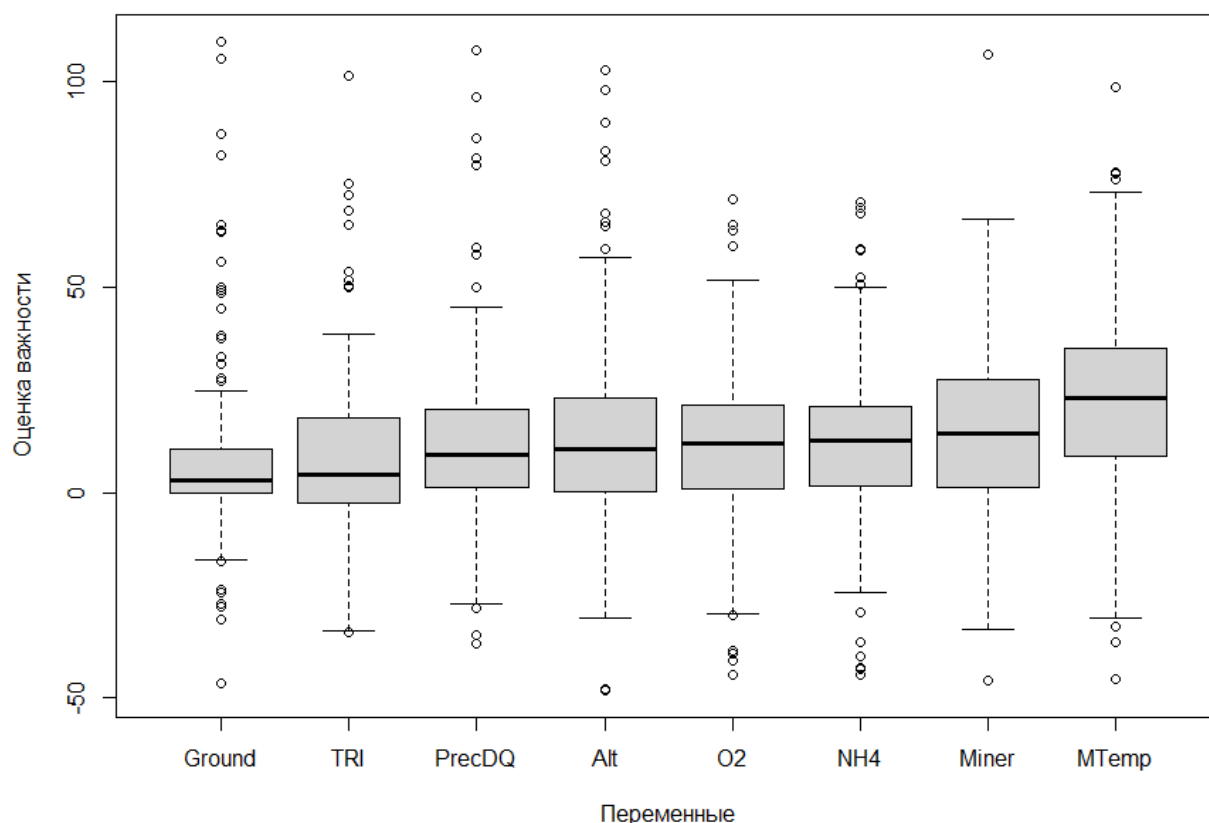
**Внимание:** Вычисленные нормированные оценки важности в обычной интерпретации являются положительными числами от 0 до 100%. Однако при выполнении расчетов величина importance для некоторых переменных среды неожиданно принимала отрицательные значения (т.е. средняя точность предсказания после исключения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



При дальнейшей доработке метода необходимо внимательно рассмотреть обоснованность такой ситуации.

Рассмотрим теперь 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" )
```



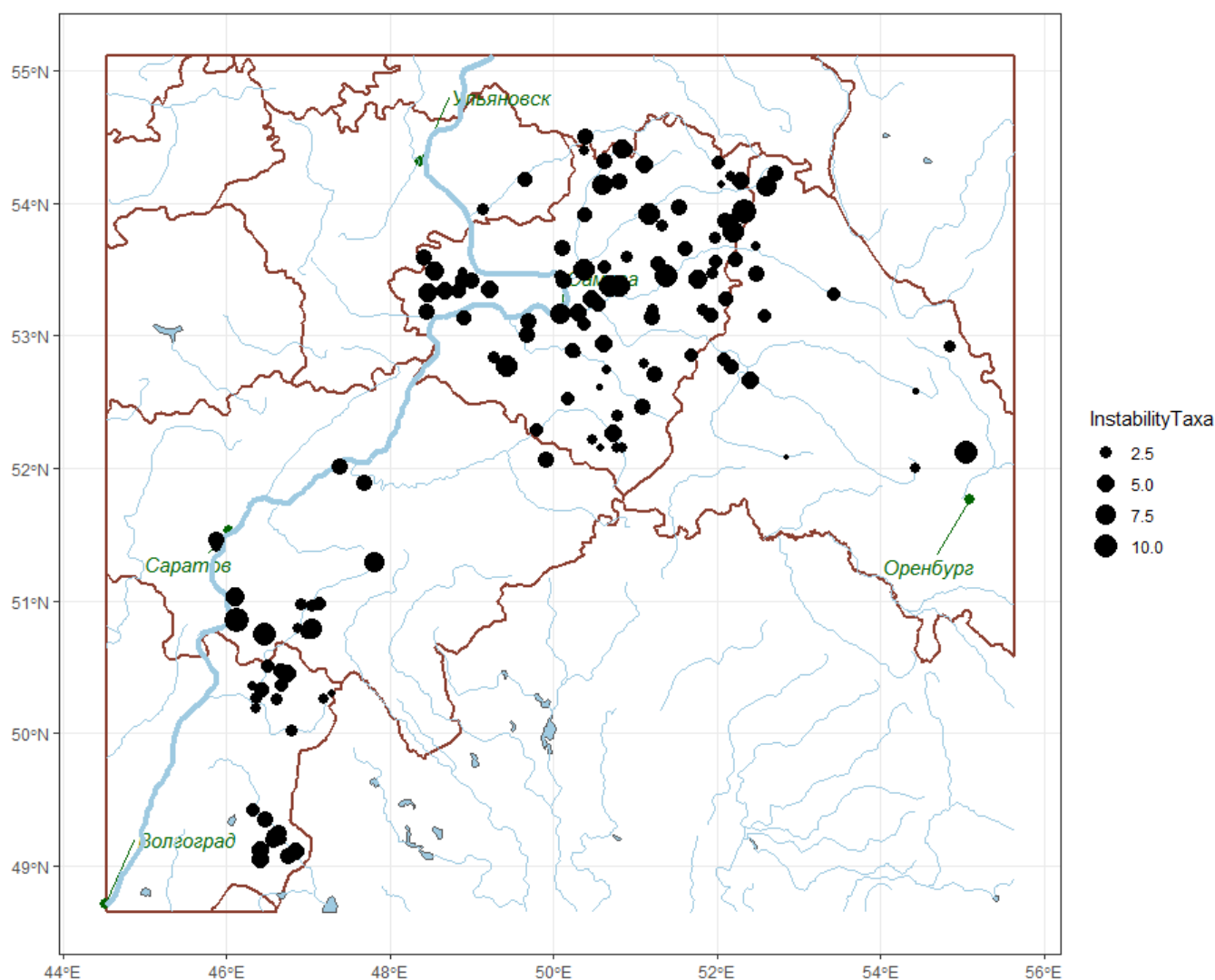
В результате обобщения этой информации по всем 147 видам вариация вклада  $W$  каждой из абиотических составляющих в изменчивость обилия различных таксонов, как видно из графика, имеет весьма широкие пределы. Если ориентироваться на медианные значения, то структура видового состава донных сообществ по региону определяется, в первую очередь, среднегодовой температурой (MTemp,  $W = 22,8$ ),

минерализацией (Miner,  $W = 18,9$ ) и  $\text{NH}_4^+$  ( $W = 17,5$ ), в отличие от расчлененности рельефа рек (TRI,  $W = 1,2$ ).

#### **4. Использование индексов нестабильности для картографирования**

В одном из сообщений 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](http://www.ievbras.ru/ecostat/Kiril/R/Blog/WB_map.RData). Его надо скачать и поместить в рабочий каталог R.

```
# Добавляем к суммарным значениям индексов географические координаты
Ind$X <- VAR$X
Ind$Y <- VAR$Y
# Грузим слой 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.

Вклады индексов нестабильности предлагается использовать (Guisande et al., 2017) в качестве исходных данных для метода виртуальных видов. В разделе 4 сообщения <https://stok1946.blogspot.com/2020/11/sdm.html> подробно описано использование этого метода для оценки индексов пригодности среды обитания ( $H \in [0,1]$  – *environmental suitability*) с последующей картографической интерпретацией и построением ареалов видов.

Однако для метода виртуальных видов необходимо также указать, каков характер связи между каждым абиотическим фактором и встречаемостью вида – прямо или обратно пропорциональный. Это ни по индексам нестабильности, ни по оценки их важности установить нельзя. В частности, как показано в таблице 1, численность подсемейства Prodiamesinae в составе трех видов *Monodiamesa bathyphila*, *Odontomesa fulva* и *Prodiamesa olivacea*, которые были обнаружены на 41 участке рек из 132, в наибольшей степени определяются четырьмя геофизическими показателями. Если по этим предикторам построить модель логистической регрессии, то она имеет вид:

```
mlog1 <- glm(V~MTemp+PrecDQ+Alt+TRI, Prodi, family = "binomial")
summary(mlog1)
Coefficients:
      Estimate Std. Error z value Pr(>|z|)
MTemp  -0.069332  0.023838  -2.909  0.00363 **
PrecDQ   0.007493  0.016500   0.454  0.64975
Alt      0.009893  0.004735   2.089  0.03670 *
TRI      0.011781  0.005937   1.984  0.04721 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 182.99  on 132  degrees of freedom
Residual deviance: 120.33  on 128  degrees of freedom
AIC: 128.33
```

Согласно этой модели вероятность встречаемости продиамезин  $V$  прямо зависит от количества летних осадков PrecDQ, высоты Alt и шероховатости рельефа TRI, но обратно пропорциональна среднегодовой температуре MTemp. Таким образом, использование индексов нестабильности для определения функций виртуальных видов generateSpFromFun (...) не всегда корректно, и в этом случае лучше использовать коэффициенты логистической регрессии.

## 5. Заключение

После появления индекса видового разнообразия Шеннона использование информационных показателей постоянно в фокусе экологии сообществ (Шитиков и др., 2005). Дивергенция Кульбака-Лейблера имеет смысл меры различия энтропии двух эмпирических распределений, а в частном случае индексов нестабильности отражает тот информационный выигрыш, который исследователь получает, рассматривая конкретную реализацию случайной величины вместо ее выборочного среднего значения. Вычисление сумм  $I_{ij}$  для отдельных участков по всем факторам среды или численностям видов в теории информации считается корректным.

Сравнительный анализ индексов нестабильности позволяет рассмотреть нетривиальные научные гипотезы о роли переменных среды или характерных особенностях таксономической структуры отдельных местообитаний. Моделирование пространственного распределения индексов позволяет выделить в составе изученной территории области со стабильным или нестабильным уровнем популяционной плотности произвольного таксона или величины фактора среды.

В рамках геостатистического анализа мы рассматривали также наличие тренда, степень автокорреляции и характер функции приращения в зависимости от расстояния между опорными точками. Однако статистически значимых закономерностей

распределения точек на карте региона обнаружить не удалось. Это приводит к выводу, что значимые однородные кластеры или иные общие тенденции изменчивости структуры сообществ в рамках всего региона не обнаруживаются, что не отрицает существования интересных феноменов в пределах отдельных рек или их локальных сетей.

Как показано выше, величина индекса нестабильности пропорциональна абсолютному отклонению показателя в конкретной точке пространства от его среднего значения в пределах всей изучаемой области. При этом остается за рамками характер зависимости, т.е. экстремальным значениям индекса может соответствовать как минимум, так и максимум анализируемого фактора, что часто имеет решающее значение. Поэтому использовать непосредственно индексы нестабильности для построения карт виртуальных видов, как это предложено в статье (Guisande et al., 2017), следует, лишь только убедившись в одинаковом характере функции связи каждого предиктора с индексом пригодности среды обитания виртуального таксона.

Тем не менее, использование индексов нестабильности в модификации К.Гисанда с соавторами (Guisande et al., 2017) позволяет редуцировать многомерные данные о видовой структуре сообществ и абиотических факторов в наборы соизмеримых стандартизованных переменных и оценить тесноту связи между ними. Метод может применяться для предварительного разведывательного анализа компонентов и связей различных экосистем, в том числе, для ранжирования местообитаний по степени нестабильности окружающей среды и/или видовых ассоциаций, а также отбора наиболее информативных абиотических переменных, определяющих популяционную плотность таксонов.

Нам остается только выразить признательность проф. К.Гисанду (С. Guisande, University of Vigo, Spain) за методическую поддержку и важные замечания по ходу подготовки сообщения.

### **Список литературы**

*Джеймс Г., Уиттон Д., Хастис Т., Тибишрани Р.* (2016) Введение в статистическое обучение с примерами на языке R. М., ДМК Пресс, 450 с.

*Лисовский А.А., Дудов С.В.* (2020) Преимущества и ограничения методов экологического моделирования ареалов. 2. MaxEnt. Журнал общей биологии, 81(2): 135–146.

*Лисовский А.А., Дудов С.В., Оболенская Е.В.* (2020) Преимущества и ограничения методов экологического моделирования ареалов. 1. Общие подходы. Журнал общей биологии, 81(2): 123–134.

*Пузаченко Ю.Г.* (2004) Математические методы в экологических и географических исследованиях. М., Академия, 416 с.

*Шутиков В.К., Мاستицкий С.Э.* (2017) Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Электронная книга: 351 с. URL: <https://stok1946.blogspot.com>.

*Шутиков В.К., Розенберг Г.С., Зинченко Т.Д.* (2005) Количественная гидроэкология: методы, критерии, решения: в 2-х кн. Москва, Наука, Кн. 1: 281 с.; Кн. 2: 337 с.

*Dubois D.M.* (1973) An index of fluctuations, Do, connected with diversity and stability of ecosystems: applications in the Lotka-Volterra model and in an experimental distribution of species. Rapport de sythèse III, Programme National sur l'environnement Physique et Biologique, Project Mer. Commission Interministérielle de la Politique Scientifique. Liège

*Franklin J.* (2009) Mapping species distributions: spatial inference and prediction. Cambridge, Cambridge University Press, 320 p.

*García-Roselló E., Guisande C., González-Vilas L., González-Dacosta J., Heine J., Pérez-Costas E., Lobo J.* (2019) A simple method to estimate the probable distribution of species. *Ecography*, 42(9): 1613-1622

*García-Roselló E., Guisande C., Heine J., Pelayo-Villamil P., Manjarrés-Hernández A., González-Vilas L., González-Dacosta J., Vaamonde A., Granado-Lorencio C.* (2014) Using ModestR to download, import and clean species distribution records. *Methods in Ecology and Evolution*, 5(7): 708–713

*González-Vilas L., Guisande C., Vari R., Pelayo-Villamil P., Manjarrés-Hernández A., García-Roselló E., González-Dacosta J., Heine J., Pérez-Costas E., Granado-Lorencio C., Palau-Ibars A., Lobo J.M.* (2016) Geospatial data of freshwater habitats for macroecological studies: an example with freshwater fishes. *International Journal of Geographical Information Science*, 30(1): 126-141

*Guisan A., Thuiller W., Zimmermann N.E.* (2017) *Habitat suitability and distribution models: with applications in R*. Cambridge, Cambridge University Press, 478 p.

*Guisande C., Barreiro A., Maneiro I., Riveiro I., Vergara-Castano A.R., Vaamonde A.* (2006) *Tratamiento de datos*. Madrid, Ediciones Diaz de Santos, 365 p.

*Guisande C., Garcia-Rosello E., Heine J., Gonzalez-Dacosta J., Gonzalez-Vilas L., Garcia-Perez B.J., Lobo J.M.* (2017) SPEDInstabR: an algorithm based on a fluctuation index for selecting predictors in species distribution modeling. *Ecological Informatics*, 37: 18–23

*Hastie T., Fithian W.* (2013) Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8): 864–867

*Hirzel A.H., Helfer V., Metral F.* (2001) Assessing habitat-suitability models with a virtual species. *Ecological Modelling*, 145(2-3): 111–121

*Hirzel A.H., Hausser J., Chessel D., Perrin N.* (2002) Ecological-niche factor analysis: how to compute habitat suitability maps without absence data? *Ecology*, 83(7): 2027–2036

*Legendre P., Gallagher E.* (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2): 271–280

*Leroy B., Meynard C.N., Bellard C., Courchamp F.* (2016) *virtualspecies*, an R package to generate virtual species distributions. *Ecography*, 39(6): 599-607

*Norberg A., Abrego N., Blanchet F.G., Adler F.R., Anderson B.J., Anttila J., Araújo M.B., Dallas T., Dunson D., Elith J., Foster S.D., Fox R., Franklin J., Godsoe W., Guisan A., O'Hara B., Hill N.A., Holt R.D., Hui F.K.C., Husby M., Kålås J.A., Lehtikoinen A., Luoto M., Mod H.K., Newell G., Renner I., Roslin T., Soininen J., Thuiller W., Vanhatalo J., Warton D., White M., Zimmermann N.E., Gravel, D., Ovaskainen O.* (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, 89(3): e01370

*Peterson A.T., Soberón J., Pearson R.G., Anderson R.P., Martínez-Meyer E., Nakamura M., Araújo M.* (2011) *Ecological niches and geographic distributions (MPB-49)*. Princeton, Princeton University Press, 328 p.

*Phillips S.J., Anderson R.P., Schapire R.E.* (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3–4): 231–259.