Модельные матрицы, ANOSIM и SIMPER

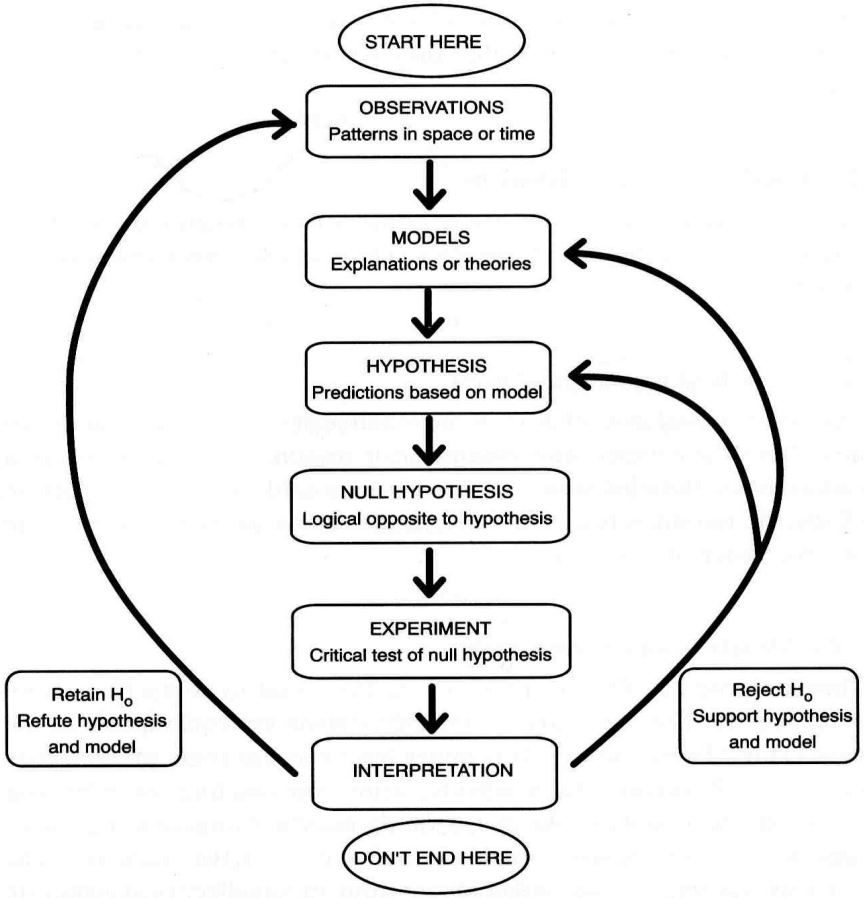
Анализ и визуализация многомерных данных с использованием R

Вадим Хайтов, Марина Варфоломеева

Вы сможете

- Протестировать гипотезу о наличии в данных некоторого специфичского паттерна, используя метод модельных матриц.
- Протестировать гипотезу о различиях между дискретными группами многомерных данных с помощью теста ANOSIM
- Выявить переменные, вносящие наибольший вклад в формирование различий между группами, применив процедуру SIMPER

Вспомним основы



Этапы работы с гипотезами (Underwood, 1997)

Формулировка биологической гипотезы

Численное выражение биологической гипотезы (H)

Формулировка альтернативной гипотезы (H_0 - нулевой гипотезы)

Тестирование гипотезы о соответствии ожидаемому паттерну: метод модельных матриц

Пример: Динамика сообществ мидиевых банок

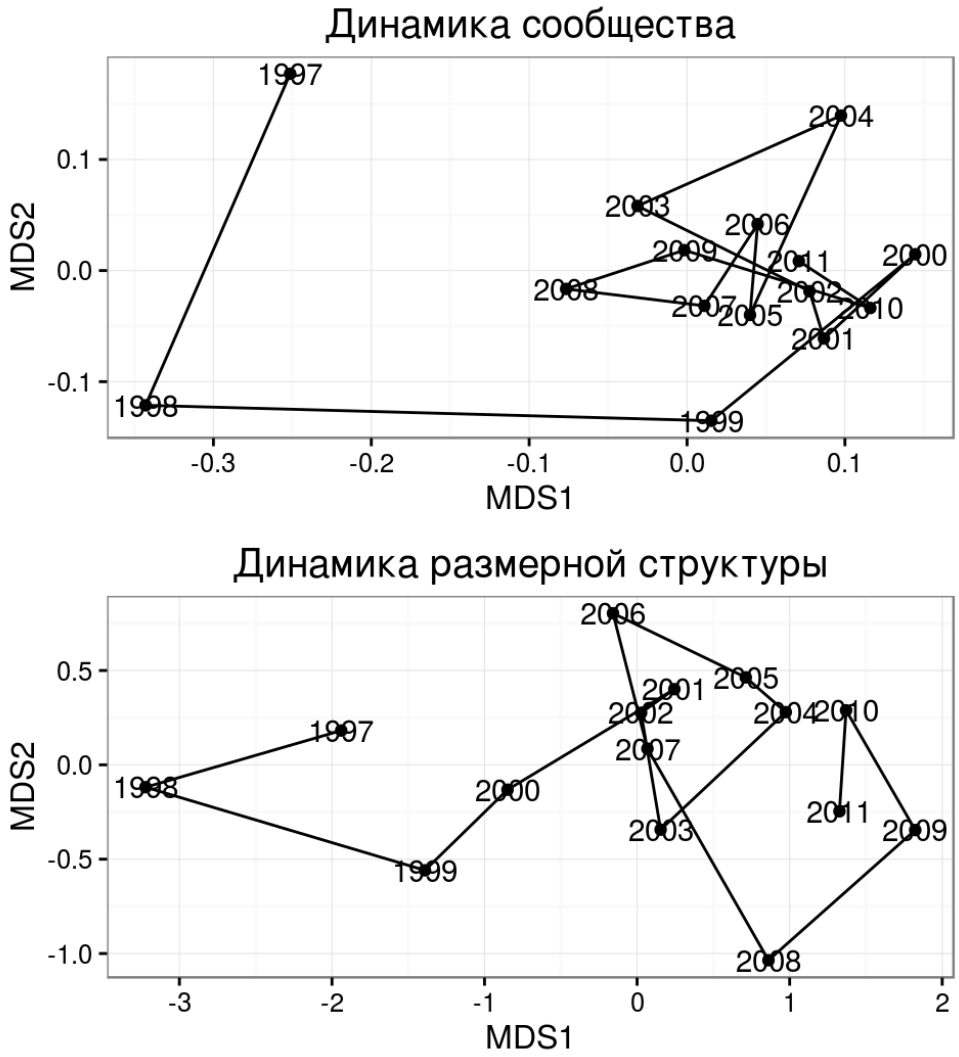
Существуют ли направленные многолетние изменения в размерной структуре поселений мидий и в структуре сообщества (Khaitov, 2013)?

```
com <- read.csv("data/mussel_beds.csv", sep=';', header = T)
ascam <- read.csv("data/ASCAM.csv", sep=';', header = T)
```

`com` — усредненные данные по обилию 12 видов для 3 мидиевых банок (Vor2, Vor4, Vor5).

`ascam` — (averaged size class abundance matrix) средние плотноси поселения мидий разных размеров (6] размерных классов)

Задание



Рассмотрите многолетние изменения структуры сообщества и размерной структуры мидий на мидиевой банке Vor2

Постройте рисунок, аналогичный приведенному на приведенном слайде

Решение

Ординация

```
library(vegan)
log_com <- log(com[,-c(1:3)] + 1)
vor2_log_com <- log_com[com$Bank == "Vor2",]
log_ascam <- log(ascam[, -c(1:2)]+1)
mds_vor2_com <- as.data.frame(metaMDS(vor2_log_com)$points)
vor2_log_ascam <- log_ascam[ascam$Bank == "Vor2",]
mds_vor2_ascam <- as.data.frame(metaMDS(vor2_log_ascam, distance = "euclid" )$points)
```

Решение

График ординации

```
library(ggplot2)
library(gridExtra)
theme_set(theme_bw())

Pl1 <- ggplot(mds_vor2_com, aes(x=MDS1, y=MDS2)) + geom_point() + geom_path() +
geom_text(label = com$Year[com$Bank == "Vor2"]) + ggtitle("Динамика сообщества")

Pl2 <- ggplot(mds_vor2_ascam, aes(x=MDS1, y=MDS2)) + geom_point() + geom_path() +
geom_text(label = com$Year[com$Bank == "Vor2"]) + ggtitle("Динамика размерной структуры")

grid.arrange(Pl1, Pl2)
```

Градиентная модельная матрица

Это матрица Евклидовых расстояний между временными точками.

```
gradient_model <- vegdist(com$Year[com$Bank == "Vor2"], method="euclidian")
gradient_model
```

```
##     1  2  3  4  5  6  7  8  9 10 11 12 13 14
## 2   1
## 3   2  1
## 4   3  2  1
## 5   4  3  2  1
## 6   5  4  3  2  1
## 7   6  5  4  3  2  1
## 8   7  6  5  4  3  2  1
## 9   8  7  6  5  4  3  2  1
## 10  9  8  7  6  5  4  3  2  1
## 11 10  9  8  7  6  5  4  3  2  1
## 12 11 10  9  8  7  6  5  4  3  2  1
## 13 12 11 10  9  8  7  6  5  4  3  2  1
## 14 13 12 11 10  9  8  7  6  5  4  3  2  1
## 15 14 13 12 11 10  9  8  7  6  5  4  3  2  1
```

Тестируем гипотезу о наличии градиента

Протестируем гипотезу о наличии временного градиента с помощью теста Мантела

```
dist_vor2_com <- vegdist(vor2_log_com, method = "bray")
dist_vor2_ascam <- vegdist(vor2_log_ascam, method = "euclidean")
```

1) Наличие градиента в структуре сообщества

```
mantel(dist_vor2_com, gradient_model)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_vor2_com, ydis = gradient_model)
##
## Mantel statistic r: 0.296
##       Significance: 0.02
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.170 0.229 0.271 0.314
## Permutation: free
## Number of permutations: 999
```

Тестируем гипотезу о наличии градиента

2) Наличие градиента в размерной структуре мидий

```
mantel(dist_vor2_ascam, gradient_model)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_vor2_ascam, ydis = gradient_model)
##
## Mantel statistic r: 0.635
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.158 0.213 0.264 0.308
## Permutation: free
## Number of permutations: 999
```

Прослеживается ли связь между размерной структурой мидий и структурой сообщества?

Не самое правильное решение

```
mantel(dist_vor2_com, dist_vor2_ascam)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_vor2_com, ydis = dist_vor2_ascam)
##
## Mantel statistic r: 0.64
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.310 0.406 0.489 0.562
## Permutation: free
## Number of permutations: 999
```

Прослеживается ли связь между размерной структурой мидий и структурой сообщества?

Более корректное решение

```
mantel.partial(dist_vor2_com, dist_vor2_ascam, gradient_model)
```

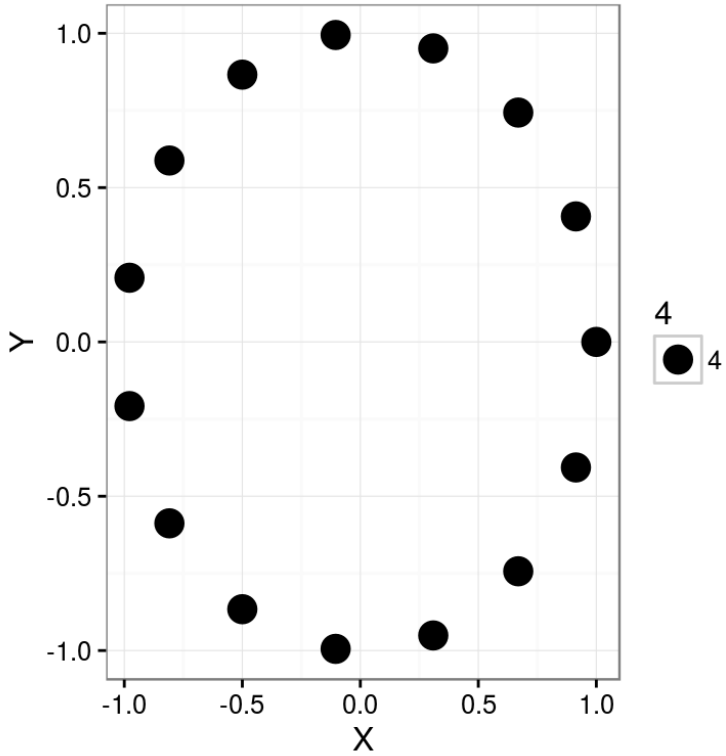
```
##
## Partial Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel.partial(xdis = dist_vor2_com, ydis = dist_vor2_ascam,      zdis = gradient_model)
##
## Mantel statistic r: 0.613
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.339 0.433 0.485 0.529
## Permutation: free
## Number of permutations: 999
```

Задание

1. Выясните есть ли многолетний градиент в динамике размерной струтуры и структуры сообщества на банке Vor4.
2. Оцените связь между размерной структурой мидий и структурой сообщества.

Могут быть и более сложные модельные матрицы

Проверим, нет ли в динамике размерной структуры мидий на банке Vor2 циклических изменений, которые предсказываются теорией динамики плотных поселений (Наумов, 2006)



Циклическая модельная матрица

```
##    1 2 3 4 5 6 7 8 9 10 11 12 13 14
## 2  0
## 3  1 0
## 4  1 1 0
## 5  1 1 1 0
## 6  2 1 1 1 0
```

Выявляется ли циклическая составляющая в динамике размерной структуры?

```
mantel(dist_vor2_ascam, cycl_model)
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = dist_vor2_ascam, ydis = cycl_model)
##
## Mantel statistic r: 0.204
##       Significance: 0.01
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0812 0.1319 0.1623 0.2055
## Permutation: free
## Number of permutations: 999
```

Циклическая составляющая есть, но…

Более корректная оценка

```
mantel.partial(dist_vor2_ascam, cycl_model, gradient_model)
```

```
##
## Partial Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel.partial(xdis = dist_vor2_ascam, ydis = cycl_model, zdis = gradient_model)
##
## Mantel statistic r: -0.202
##       Significance: 1
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.116 0.156 0.190 0.246
## Permutation: free
## Number of permutations: 999
```

Мы не можем говорить о наличии столь длительного цикла.

При данной длине временного ряда нельзя отличить цикл с большим периодом от направленного изменения.

Можно обсуждать только циклы с периодом не более половины длины временного ряда.

ANOSIM: Analysis Of Similarity

Задание

- Постройте ординацию всех описаний датасета `com` (логарифмированные данные) в осях nMDS на основе матрицы Брея-Куртиса
- Раскрасьте разными цветами точки, относящиеся к двум разным группам: "Large-dominated" и "Small-dominated"

Решение

```
library(vegan)
library(ggplot2)
ord_log_com <- metaMDS(log_com, distance = "bray", k=2)
```

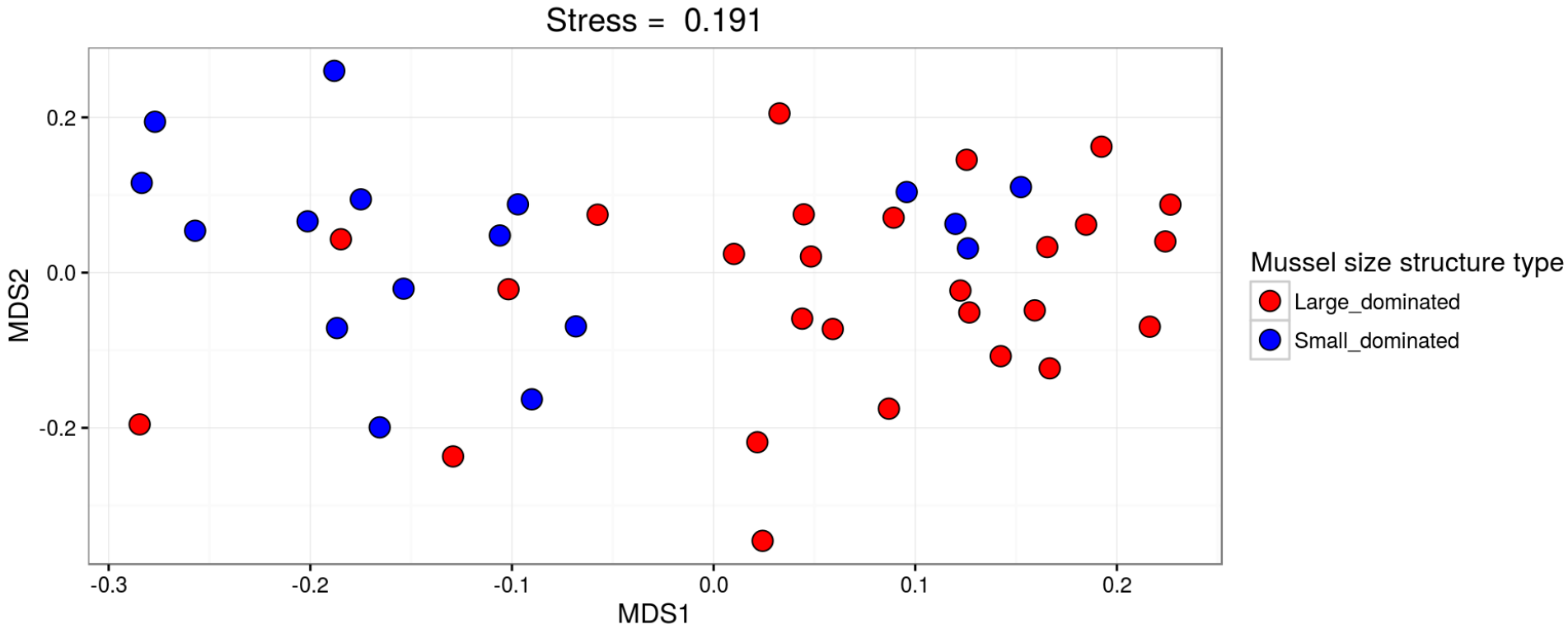
```
## Run 0 stress 0.191
## Run 1 stress 0.201
## Run 2 stress 0.204
## Run 3 stress 0.209
## Run 4 stress 0.205
## Run 5 stress 0.207
## Run 6 stress 0.236
## Run 7 stress 0.204
## Run 8 stress 0.192
## Run 9 stress 0.204
## Run 10 stress 0.192
## Run 11 stress 0.192
## Run 12 stress 0.191
## ... procrustes: rmse 0.0000216  max resid 0.000088
## *** Solution reached
```

```
MDS <- data.frame(ord_log_com$points)
```

Решение

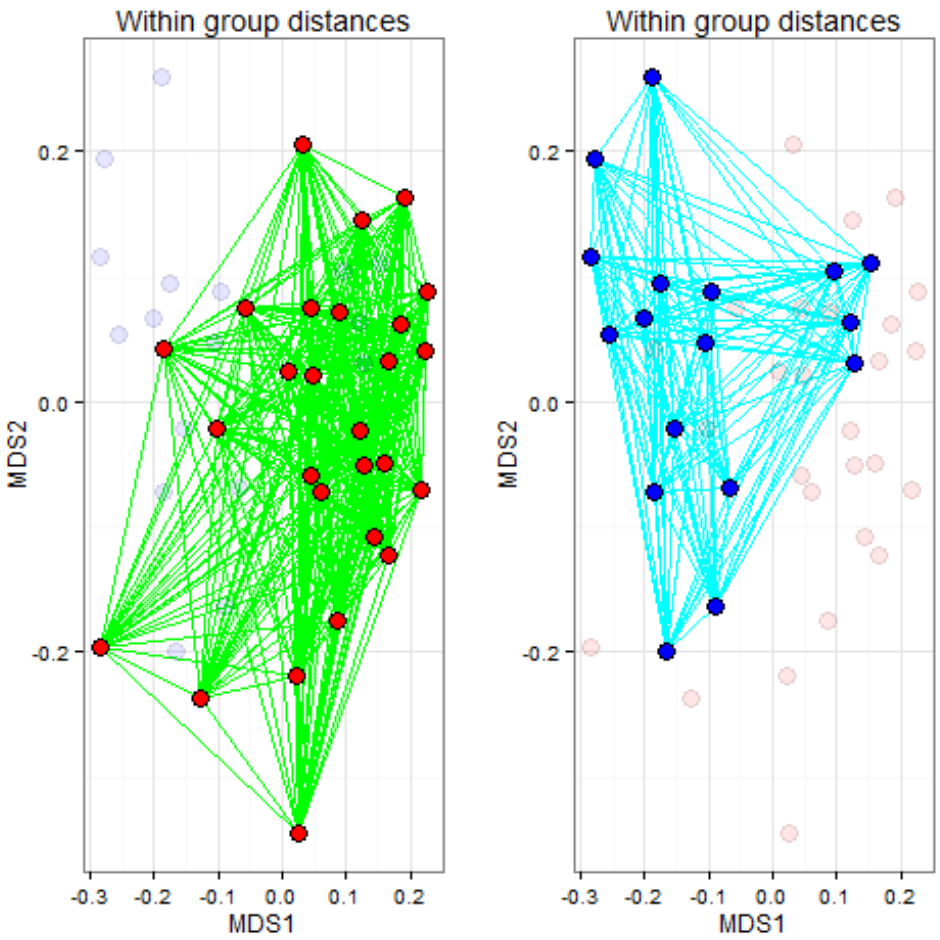
Обратите внимание, здесь есть две группы расстояний между точками

```
ggplot(MDS, aes(x = MDS1, y = MDS2, fill = com$Mussel_size)) + geom_point(shape = 21, size
= 4) + scale_fill_manual(values = c("red", "blue")) + labs(fill = "Mussel size structure
type") + ggtitle(paste("Stress = ", round(ord_log_com$stress, 3), sep = " "))
```



Расстояния между объектами

1. Внутригрупповые расстояния



2. Межгрупповые расстояния

Ранги расстояний

Для работы удобно (но не обязательно!) перейти от исходных расстояний между объектами, к их рангам.

Обозначим внутригрупповые расстояния (ранги), как r_w, а межгрупповые, как r_b.

Вычислим

- средние значения внутригрупповых рангов расстояний R_w
- средние значения межгрупповых рангов расстояний R_b

R - статистика

На основе полученных значений можно построить статистику (Clarke, 1988, 1993)

$$R_{global} = \frac{R_b - R_w}{n(n-1)/4}$$

Эта статистика распределена в интервале -1 < R_{global} < 1

Статистическая значимость этой величины оценивается пермутационным методом

Процедура ANOSIM в пакете vegan

```
com_anosim <- anosim(log_com,
           grouping = com$Mussel_size,
           permutations = 999,
           distance = "bray")
```

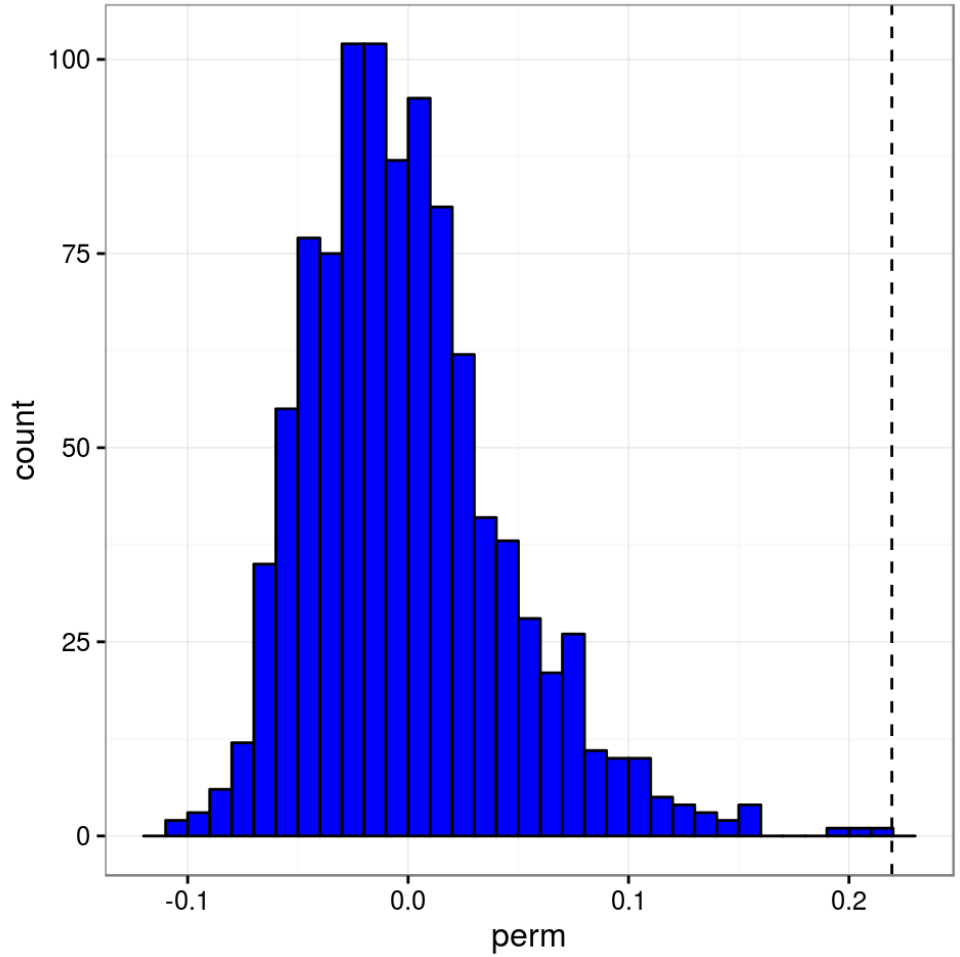
Задание

Изучите структуру объекта `com_anosim` и постройте частотное распределение значений R_{global}, полученных при каждом акте пермутации

Решение

```
anosim_perm <- data.frame(perm = com_anosim$perm)
anosim_perm[(com_anosim$permutations + 1), 1] <- com_anosim$statistic
ggplot(anosim_perm, aes(x = perm)) + geom_histogram(binwidth = 0.01, color = "black", fill
= "blue") + geom_vline(xintercept = com_anosim$statistic, linetype = 2)
```

Результаты оценки статистики R_{global} при пермутациях



Результаты процедуры ANOSIM

```
summary(com_anosim)
```

```
##
## Call:
## anosim(dat = log_com, grouping = com$Mussel_size, permutations = 999,      distance =
"bray")
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.219
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0590 0.0808 0.1051 0.1313
##
## Dissimilarity ranks between and within classes:
##                  0% 25% 50% 75% 100%   N
## Between           1 302 549 766  946 459
## Large_dominated   4 193 360 645  945 351
## Small_dominated   3 275 478 683  938 136
```

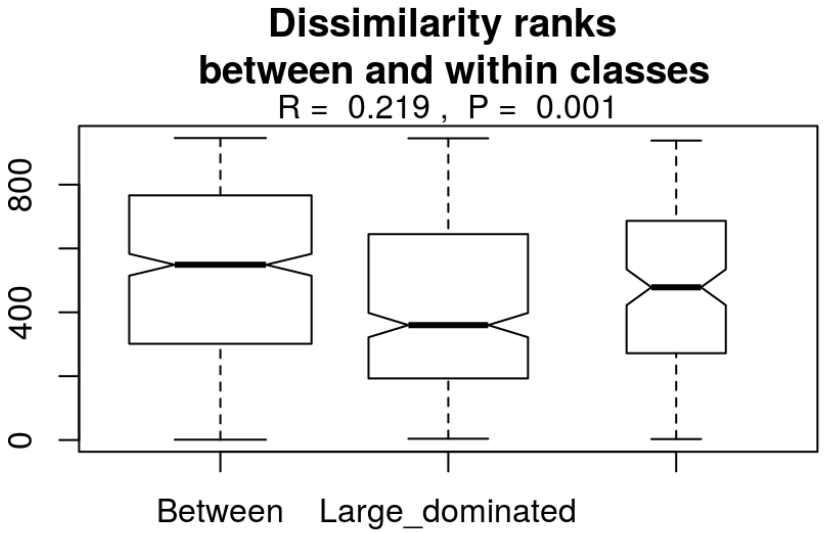
Ограничения (Assumptions) для применения ANOSIM

1) Внутригрупповые расстояния (ранги) должны иметь приблизительно равные медианы и пределы варьирования.

Для проверки этого допущения надо сравнить распределения внутригрупповых и межгрупповых расстояний (рангов)

Распределение расстояний имеет следующий вид

```
plot(com_anosim, main = "Dissimilarity ranks \n between and within classes")
```



ANOSIM позволяет сравнивать одновременно и несколько групп

НО! Есть одно очень важное ограничение:

2) Попарные сравненения групп можно осуществлять только если было показано, что R_{global} достоверно.

Если это условие выполнено, то можно проводить попарные сравнения

Пример

Пусть у нас есть три группы объектов: A, B, C.
Можно вычислить R_{AvsB} , R_{AvsC} , R_{BvsC} .

Но при больших объемах выборки даже незначительные различия будут достоверны. Важно обращать внимание не только на оценку статистической значимости, но и на значения R!

NB! Для сравнения нескольких групп многомерных объектов, есть более мощное средство - PERMANOVA

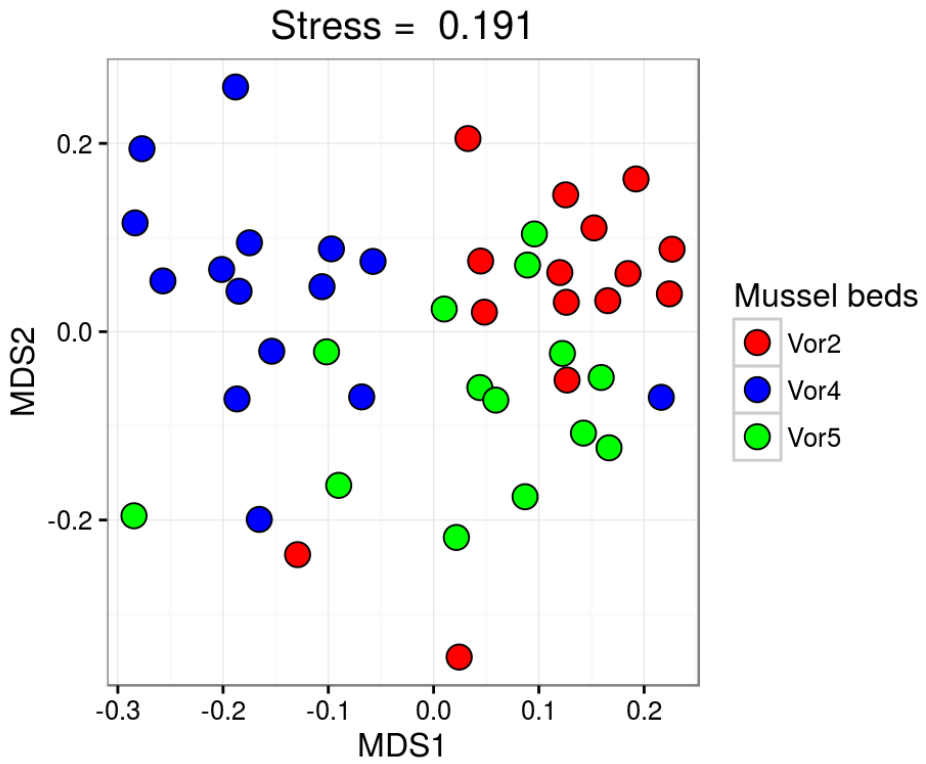
Задание

- Постройте ординацию в осях nMDS, раскрасив точки в разные цвета в зависимости от номера мидиевой банки
- Проверьте гипотезу о различиях в структуре сообществ на разных банках
- Проверьте условия применимости ANOSIM
- Проведите попарное сравнение всех банок

Решение

График ординации

```
ggplot(MDS, aes(x = MDS1, y = MDS2, fill = com$Bank)) + geom_point(shape = 21, size = 4) +
scale_fill_manual(values = c("red", "blue", "green")) + labs(fill = "Mussel beds") +
ggtitle(paste("Stress = ", round(ord_log_com$stress, 3), sep = " "))
```



Решение

Проверка гипотезы о различиях в структуре сообществ на разных банках

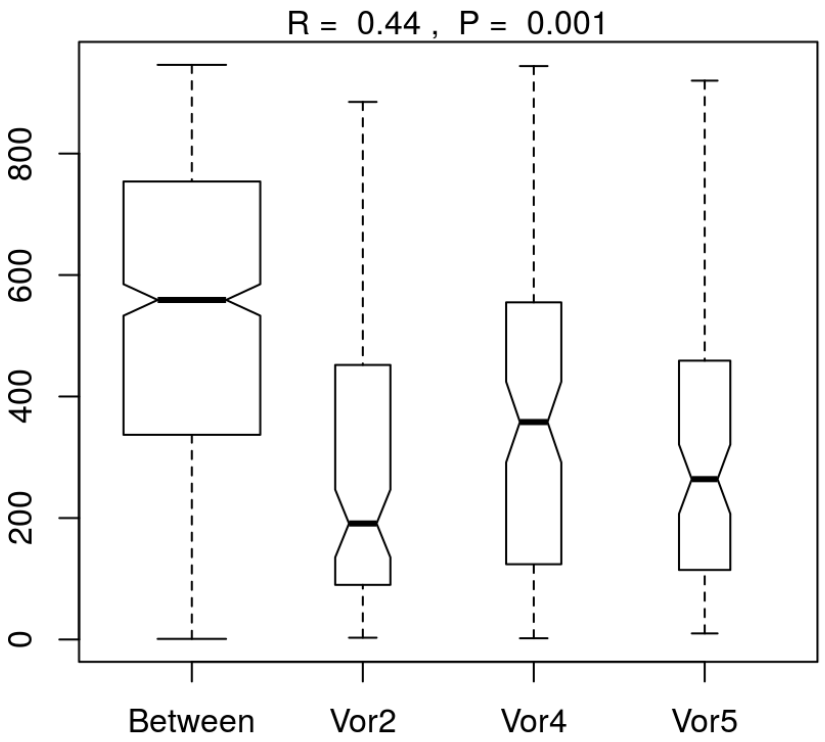
```
bed_anosim <- anosim(log_com, grouping = com$Bank,permutations = 999, distance = "bray")
bed_anosim
```

```
##
## Call:
## anosim(dat = log_com, grouping = com$Bank, permutations = 999,      distance = "bray")
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.44
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
```

Решение

Условия применимости

```
plot(bed_anosim)
```



Решение

Попарные сравнения Vor2 vs Vor4

```
# Vor2 vs Vor4
anosim(log_com [com$Bank %in% c("Vor2", "Vor4"),],
    grouping = com$Bank[com$Bank %in% c("Vor2", "Vor4")])


##
## Call:
## anosim(dat = log_com[com$Bank %in% c("Vor2", "Vor4"), ], grouping = com$Bank[com$Bank
%in%      c("Vor2", "Vor4")])
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.588
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
```

Попарные сравнения Vor2 vs Vor5

```
# Vor2 vs Vor5
anosim(log_com[com$Bank %in% c("Vor2", "Vor5"),], grouping = com$Bank[com$Bank %in%
c("Vor2", "Vor5")])


##
## Call:
## anosim(dat = log_com[com$Bank %in% c("Vor2", "Vor5"), ], grouping = com$Bank[com$Bank
%in%      c("Vor2", "Vor5")])
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.309
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
```

Попарные сравнения Vor4 vs Vor5

```
# Vor4 vs Vor5
anosim(log_com [com$Bank %in% c("Vor4", "Vor5"),], grouping = com$Bank[com$Bank %in%
c("Vor4", "Vor5")])


##
## Call:
## anosim(dat = log_com[com$Bank %in% c("Vor4", "Vor5"), ], grouping = com$Bank[com$Bank
%in%      c("Vor4", "Vor5")])
## Dissimilarity: bray
##
## ANOSIM statistic R: 0.426
##       Significance: 0.001
##
## Permutation: free
## Number of permutations: 999
```

Проблема малых выборок

Мощность ANOSIM невелика.

При малых выборках пермутационная оценка уровня значимости может не выявить достоверных различий, даже при очень высоком значении R.

Модельные матрицы и ANOSIM

При проверке гиптезы о значимости различий между группами можно использовать тест Мантела. В этой ситуации модельная матрица будет содержать 0, если расстояние внутригрупповое, и 1 если рсстояние межгрупповое.

```
m <- vegdist(as.numeric(com$Bank), method = "euclidean")
mm <- m
mm[m > 0] <- 1
mm[m == 0] <- 0
mantel(vegdist(log_com), mm, method = "pearson")
```

```
##
## Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel(xdis = vegdist(log_com), ydis = mm, method = "pearson")
##
## Mantel statistic r: 0.355
##       Significance: 0.001
##
## Upper quantiles of permutations (null model):
##    90%    95%  97.5%    99%
## 0.0343 0.0480 0.0622 0.0823
## Permutation: free
## Number of permutations: 999
```

Значения теста Мантела будут очень близки к R_{global}

SIMPER: Similarity Percentages

Какие признаки зависимой матрицы вносят наибольший вклад в формирование различий между группами?

```
log_com_simper <- simper(log_com, group = com$Mussel_size, permutations = 999)
summary (log_com_simper)
```

```
##
## Contrast: Large_dominated_Small_dominated
##
##                          contr      sd ratio  av.a  av.b cumsum     p
## Polydora_quadrilobata  0.04185 0.02941  1.42 0.862 2.425  0.177 0.002 **
## Hydrobia_ulvae         0.02598 0.01812  1.43 3.259 2.826  0.287 0.879
## Skeneopsis_planorbis   0.02343 0.01722  1.36 0.380 1.282  0.387 0.002 **
## Fabricia_sabella       0.02234 0.01799  1.24 0.921 1.312  0.481 0.334
## Cricotopus_vitripennis 0.02097 0.01512  1.39 1.687 2.221  0.570 0.154
## Onoba_aculeus          0.01879 0.01319  1.42 1.066 1.575  0.650 0.023 *
## Nemertini              0.01718 0.01417  1.21 2.159 2.511  0.722 0.167
## Macoma_balthica        0.01544 0.01299  1.19 2.605 2.678  0.788 0.642
## Littorina_saxatilis    0.01542 0.01003  1.54 2.045 1.689  0.853 0.006 **
## Filamentous_algae      0.01505 0.01145  1.31 0.828 0.629  0.917 0.957
## Gammarus_sp.           0.01145 0.00892  1.28 1.766 1.916  0.965 0.865
## Tubificoides_benedeni  0.00821 0.00613  1.34 4.859 4.755  1.000 0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

Оценка вклада в формирование различий

$$Contr_i = \frac{|y_{i,j} - y_{i,k}|}{\sum (y_{i,j} - y_{i,k})}$$

Величины $Contr_i$ далее усредняются для всех межгрупповых пар

sd - среднеквадратичное отклонение

Отношение $Contr_i/sd$ характеризует дискриминирующую силу переменной

Задание

Выявитие виды, отвечающие за различия в сообществах разых банок

Решение

```
summary(simper(log_com, group = com$Bank, permutations = 999 ))
```

```
##
## Contrast: Vor2_Vor4
##
##                          contr      sd ratio   av.a  av.b cumsum     p
## Polydora_quadrilobata  0.04661 0.02795  1.67 0.6680 2.804  0.180 0.001 ***
## Hydrobia_ulvae         0.03246 0.01559  2.08 3.6840 2.456  0.305 0.004 **
## Skeneopsis_planorbis   0.02820 0.01816  1.55 0.0486 1.460  0.414 0.001 ***
## Fabricia_sabella       0.02229 0.01699  1.31 1.5530 1.095  0.500 0.468
## Nemertini              0.02103 0.01389  1.51 1.9237 2.531  0.582 0.001 ***
## Filamentous_algae      0.01947 0.01314  1.48 1.1917 0.419  0.657 0.003 **
## Onoba_aculeus          0.01845 0.01375  1.34 0.8647 1.416  0.728 0.238
## Littorina_saxatilis    0.01741 0.01013  1.72 2.1832 1.361  0.795 0.001 ***
## Cricotopus_vitripennis 0.01703 0.01182  1.44 2.0702 2.343  0.861 0.985
## Macoma_balthica        0.01426 0.01224  1.16 2.7410 2.767  0.916 0.836
## Gammarus_sp.           0.01218 0.00936  1.30 1.8253 1.994  0.963 0.380
## Tubificoides_benedeni  0.00949 0.00664  1.43 5.0329 4.712  1.000 0.073 .
##
## Contrast: Vor2_Vor5
##
##                         contr      sd ratio   av.a  av.b cumsum     p
## Hydrobia_ulvae         0.0289 0.02376 1.216 3.6840 3.139  0.133 0.165
## Fabricia_sabella       0.0275 0.01847 1.491 1.5530 0.531  0.260 0.003 **
## Polydora_quadrilobata  0.0220 0.02301 0.955 0.6680 0.889  0.361 1.000
## Cricotopus_vitripennis 0.0217 0.01407 1.541 2.0702 1.222  0.461 0.203
## Onoba_aculeus          0.0192 0.01373 1.396 0.8647 1.525  0.550 0.102
## Filamentous_algae      0.0180 0.01220 1.479 1.1917 0.634  0.633 0.023 *
## Macoma_balthica        0.0168 0.01317 1.274 2.7410 2.375  0.710 0.260
## Nemertini              0.0165 0.01208 1.368 1.9237 2.439  0.786 0.570
## Skeneopsis_planorbis   0.0138 0.01094 1.261 0.0486 0.673  0.850 0.996
## Gammarus_sp.           0.0123 0.00893 1.374 1.8253 1.641  0.907 0.360
## Littorina_saxatilis    0.0102 0.00712 1.430 2.1832 2.196  0.953 1.000
## Tubificoides_benedeni  0.0101 0.00690 1.461 5.0329 4.703  1.000 0.008 **
```

Summary

- Можно формулировать гипотезу о существовании некоторого паттерна. Паттерн можно описать модельной матрицей. Проверка соответствия паттерну производится с помощью теста Мантела.
- ANOSIM — простейший вариант сравнения нескольких групп объектов, охарактеризованных по многим признакам.
- С помощью процедуры SIMPER можно оценить вклад отдельных переменных в формирование различий между группами.

Что почитать

- Clarke, K. R., Gorley R. N. (2006) PRIMER v6: User Manual/Tutorial. PRIMER-E, Plymouth.
- Legendre P., Legendre L. (2012) Numerical ecology. Second english edition. Elsevier, Amsterdam.