# Анализ связи между наборами данных

## Анализ и визуализация многомерных данных с использованием R

Вадим Хайтов, Марина Варфоломеева

# Вы сможете

- Отразить связь между nMDS и значениями признаков объектов, которые не были использованы при построении ординации.
- Количественно оценить степень взаимосвязи между несколькими наборами данных.
- Найти ортимальное сочетание признаков, не вошедших в ординацию, которые "объясняют" характер взаиморасположения точек на ординации

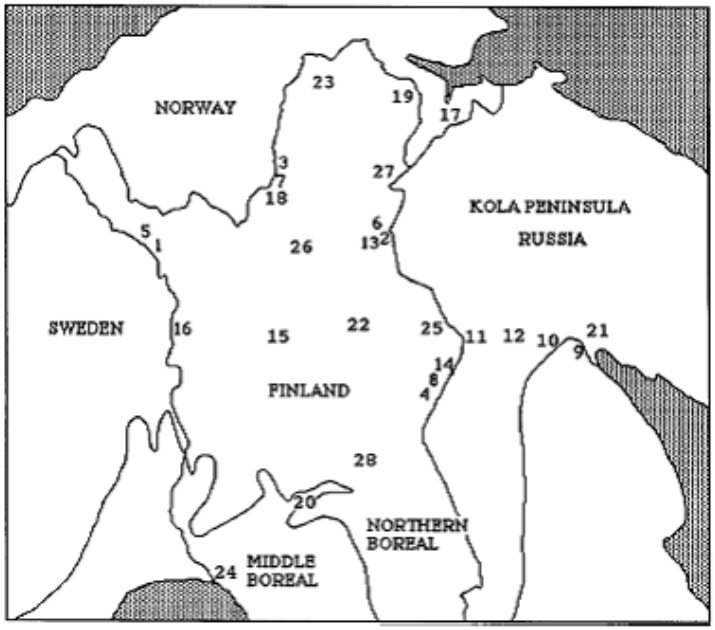
# Постановка проблемы

Мы имеем набор объектов, охарактеризованных двумя сопряженным наборами переменных

- Обилия видов ($M$ видов $\times$ $N$ описаний) и параметры среды ($K$ параметров $\times$ $N$ описаний)
- Морфометрические показатели ($M$ признаков $\times$ $N$ объектов) и генетические признаки (Экспрессия $K$ генов $\times$ $N$ объектов)
- Признаки хозяина ($M$ признаков $\times$ $N$ объектов) и призники паразита ($K$ признаков $\times$ $N$ объектов)

# Ординация растительности на пастбищах северных оленей

Väre, H., Ohtonen, R. and Oksanen, J. (1995) Effects of reindeer grazing on understorey vegetation in dry Pinus sylvestris forests. Journal of Vegetation Science 6, 523–530.




из Väre, Ohtonen & Oksanen (1995)

```
library(vegan)
data(varespec)
data(varechem)
```

Два набора данных:

- `varespec` - Описание растительности (обилия отдельных видов)

# Часть 1. Выявление связи ординации объектов и значений конкретных факторов

# Задание

1. Постройте ординацию описаний растительности в осях MDS.
2. Вычислите величину стресса
3. Раскрасьте точки в соответствии с концентрацией Al

Hint. В качестве меры различия используйте коэффициент Брея-Куртиса

# Решение

```
## Square root transformation
## Wisconsin double standardization
## Run 0 stress 0.184
## Run 1 stress 0.252
## Run 2 stress 0.228
## Run 3 stress 0.222
## Run 4 stress 0.235
## Run 5 stress 0.23
## Run 6 stress 0.185
## ... procrustes: rmse 0.0494  max resid 0.158
## Run 7 stress 0.197
## Run 8 stress 0.208
## Run 9 stress 0.196
## Run 10 stress 0.212
## Run 11 stress 0.183
## ... New best solution
## ... procrustes: rmse 0.0417  max resid 0.152
## Run 12 stress 0.21
## Run 13 stress 0.242
## Run 14 stress 0.198
## Run 15 stress 0.206
## Run 16 stress 0.198
## Run 17 stress 0.214
## Run 18 stress 0.196
## Run 19 stress 0.236
## Run 20 stress 0.197
```

# Анализ связи с переменными с помощью функции envfit()

```
env_fit <- envfit(veg_ord, varechem)
env_fit
```

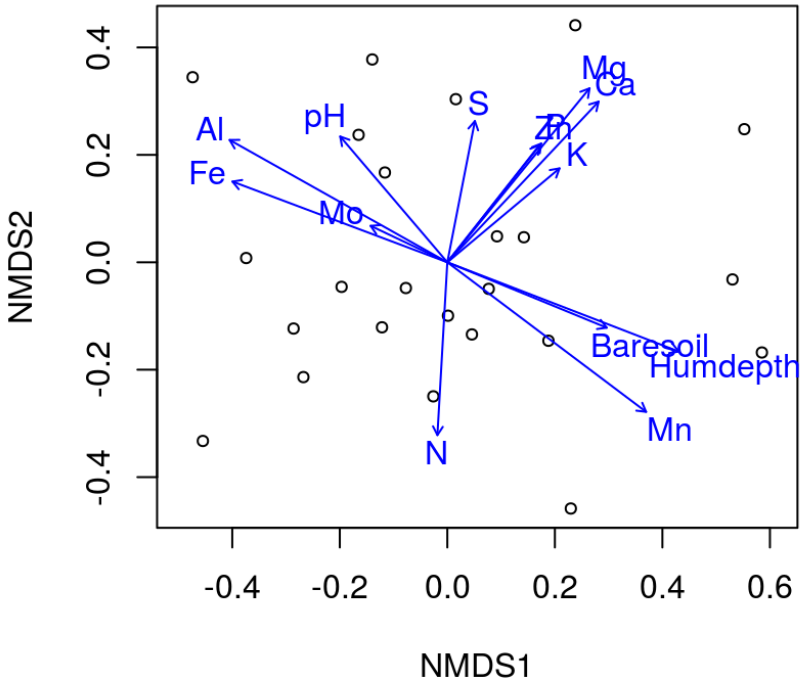
```
##
## ***VECTORS
##
##            NMDS1  NMDS2   r2 Pr(>r)
## N         -0.057 -0.998 0.25  0.051 .
## P          0.620  0.785 0.19  0.115
## K          0.766  0.642 0.18  0.118
## Ca         0.685  0.728 0.41  0.006 **
## Mg         0.632  0.775 0.43  0.003 **
## S          0.191  0.982 0.18  0.147
## Al        -0.872  0.490 0.53  0.001 ***
## Fe        -0.936  0.352 0.45  0.002 **
## Mn         0.799 -0.602 0.52  0.001 ***
## Zn         0.618  0.787 0.19  0.124
## Mo        -0.903  0.429 0.06  0.515
## Baresoil   0.925 -0.380 0.25  0.061 .
## Humdepth   0.933 -0.360 0.52  0.002 **
## pH        -0.648  0.762 0.23  0.068 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

В основе работы функции лежит регрессионный анализ

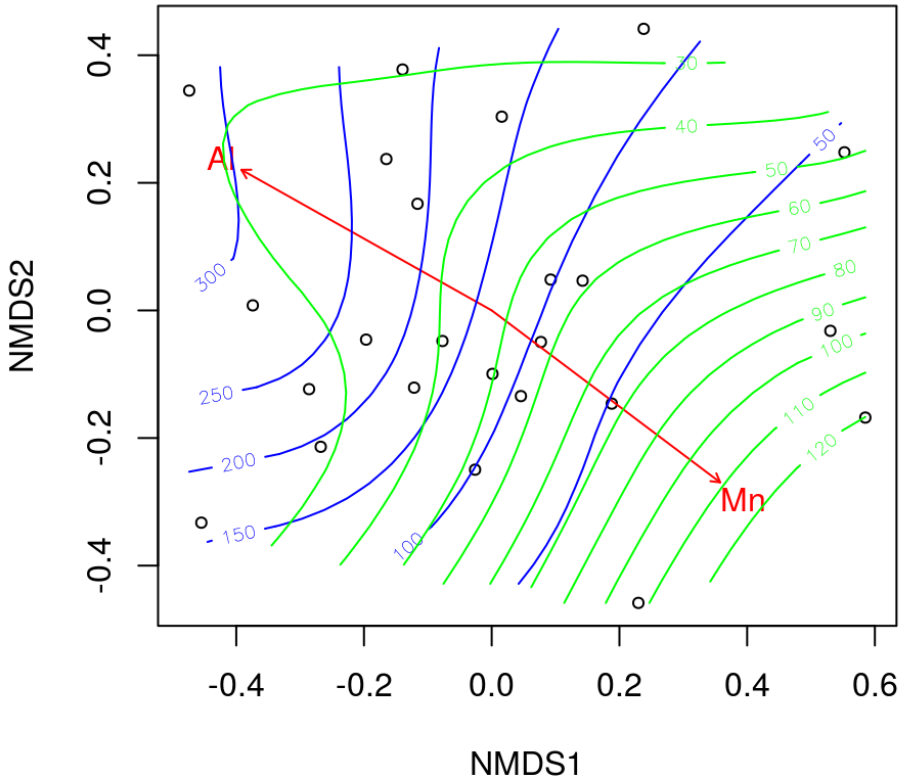
- Колонки `MDS1` и `MDS2` содержат косинусы углов (пропорциональны коэффициентам частной корреляции)

# Визуализация результатов

```
plot(veg_ord, display = "site")
plot(env_fit)
```



# Анализ связи с переменными с помощью функции ordisurf()




```
env_fit2 <- envfit(veg_ord ~ Al + Mn, data = varechem)
plot(veg_ord, display = "site")
plot(env_fit2, col = "red")
ordisurf(veg_ord, varechem$Al, add = TRUE, col="blue")
ordisurf(veg_ord, varechem$Mn, add = TRUE, col="green")
```

# Задание:

Отразите связь ординации растительности со значениями концентрации гумуса.

# Часть 2. Тест Мантела

# Постановка проблемы

Нам необходимо оценить связаны ли, вцелом, два набора данных и оценить силу этой связи

Зависит ли растительность от параметров среды?Связаны ли морфологические признаки и экспрессия генов?Связаны ли характеристики паразитов и хозяев?

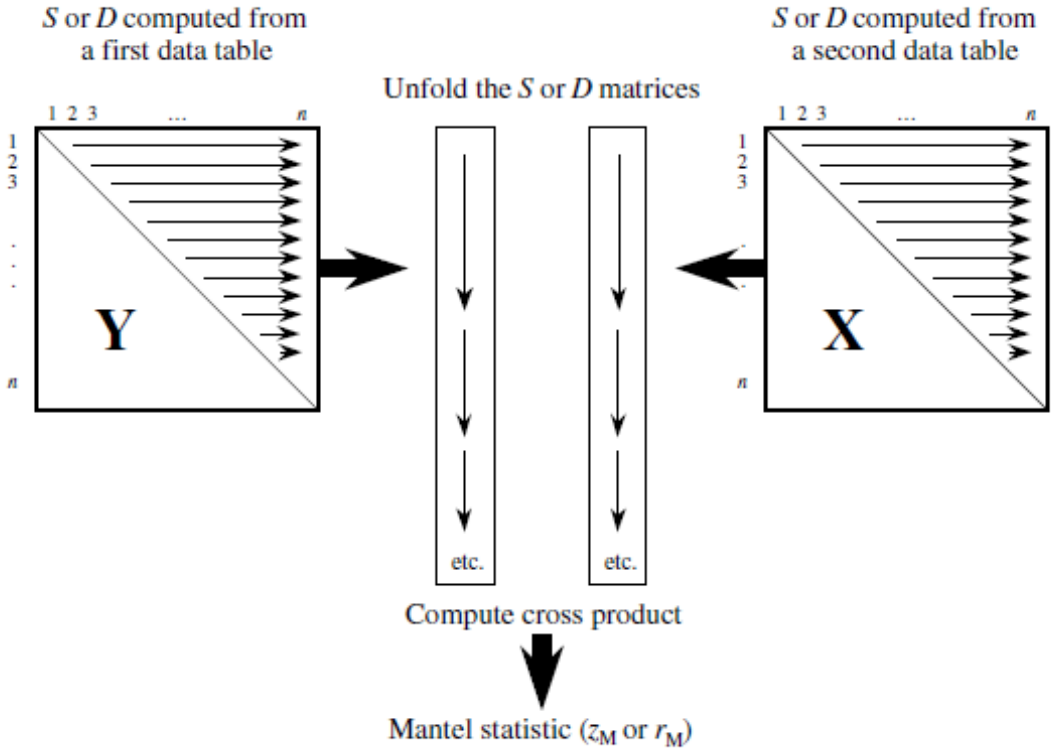
и т.п.

# Метод сравнения сопряженных матриц, описывающих попарные расстояния (или сходства), был предложен Натаном Мантелом

Если две матрицы сопряжены, то меры сходства/различия в одной матрице должны быть подобны мерам сходства/различия в другой матрице

```
dist_com <- vegdist(varespec, method = "bray")
dist_chem <- vegdist(varechem, method = "euclidean")
```

# Корреляция матриц сходства/различия




Внимание! Достоверность этой корреляции нельзя оценивать как обычную корреляцию, например функцией `cor.test()` или по таблице пороговых значений коэффициента корреляции.

# Проверка достоверности Мантеловской корреляции

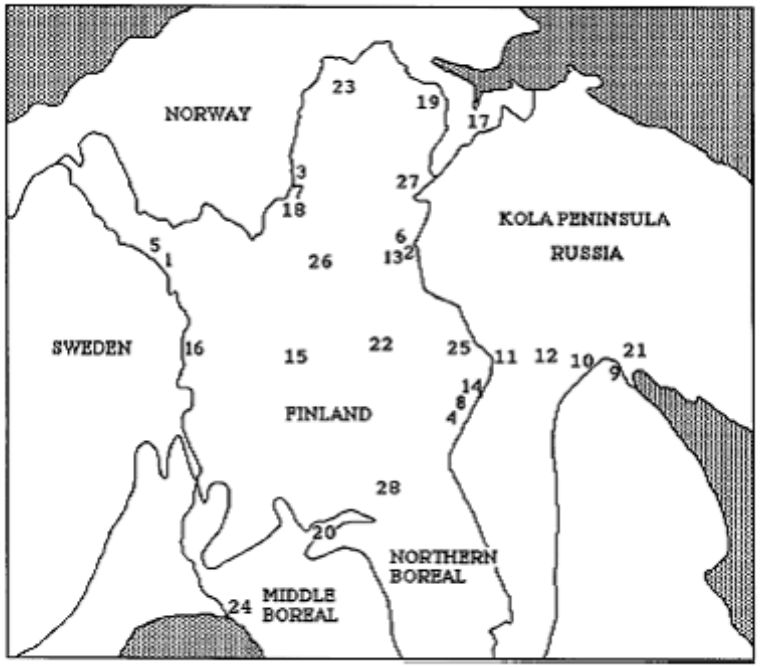
Для оценки достоверности Мантеловской корреляции применяется пермутационная процедура. Эта процедура реализована в функции mantel()

```
options(digits=4)
mant <- mantel(dist_com, dist_chem, method="spearman", permutations = 999)
mant
```

```
##
## Mantel statistic based on Spearman's rank correlation rho
##
## Call:
## mantel(xdis = dist_com, ydis = dist_chem, method = "spearman",      permutations = 999)
##
## Mantel statistic r: 0.224
##       Significance: 0.009
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.105 0.141 0.173 0.210
## Permutation: free
## Number of permutations: 999
```

Вероятность наблюдать такое значение при условии, что верна $H_0$, равна 0.009

# Частная Мантеловская корреляция




из Väre, Ohtonen & Oksanen (1995)

В материале есть одна проблема

- сходство между отдельными описаниями может быть обусловлено не только их биологическими свойствами, но и тем, что они просто располагаются ближе друг к другу в пространстве.
- Корреляция между биологическими признаками и химическими должна оцениваться при учете еще одной матрицы - **Матрицы географических расстояний**

# Частная Мантеловская корреляция

```
mantel_partial <- mantel.partial(dist_com, dist_chem, dist_geo, method = "pearson",
permutations = 9999)
mantel_partial
```

```
##
## Partial Mantel statistic based on Pearson's product-moment correlation
##
## Call:
## mantel.partial(xdis = dist_com, ydis = dist_chem, zdis = dist_geo,      method =
"pearson", permutations = 9999)
##
## Mantel statistic r: 0.182
##       Significance: 0.019
##
## Upper quantiles of permutations (null model):
##   90%   95% 97.5%   99%
## 0.113 0.147 0.173 0.206
## Permutation: free
## Number of permutations: 9999
```

# Часть 3. Подбор оптимальной модели: процедура BIO-ENV

# Постановка задачи

Необходимо выбрать предикторы, которые наилучшим образом объясняют поведение биологической системы.

$NB$! Эта задача аналогична задачам, ставящимся в регрессионном анализе.

К. Кларком и М. Эинсвортом был предложен метод BIO-ENV (Clarke, Ainsworth, 1993). Это непараметрический аналог пошагового регрессионного анализа.

# Процедура BIO-ENV

В этом анализе есть две сцепленные матрицы:

- Зависимая матрица (BIO) - матрица геоботанических описаний.
- Матрица-предиктор (ENV) - матрица химических параметров.

# Алгоритм процедуры BIO-ENV

- Вычисляется матрица взаимных расстояний между объектами для зависимой матрицы $D_{BIO}$. Используются все ее перменные.
- Матрица-предиктор имеет $p$ переменных. Вычисляются все возможные матрицы взаимных расстояний между объектами для всех возможных комбинаций признаков матрицы ENV - $D_{ENV_i}$. ВНИМАНИЕ! Таких матриц будет $2^p - 1$.
- Между каждой из матриц $D_{ENV_i}$ и матрицей $D_{BIO}$ вычисляется мантеловская корреляция.
- Находится матрица $D_{ENV_i}$, имеющая максимальное значение мантеловской корреляции.
- Выводятся признаки матрицы ENV, на основе которых получена максимально подобная матрица $D_{ENV_i}$.

# Функция bioenv() из пакета vegan

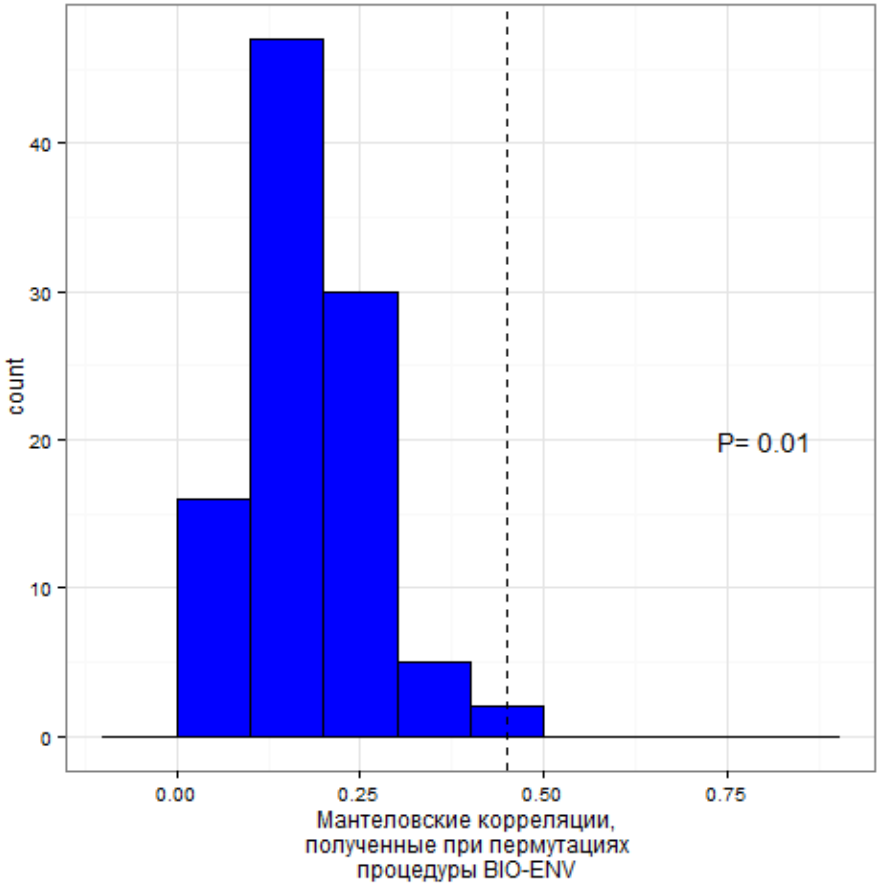
```
BioEnv <- bioenv(varespec, varechem, method = "spearman", index = "bray")

## 16383 possible subsets (this may take time...)

BioEnv

##
## Call:
## bioenv(comm = varespec, env = varechem, method = "spearman",      index = "bray")
##
## Subset of environmental variables with best correlation to community data.
##
## Correlations:    spearman
## Dissimilarities: bray
## Metric:          euclidean
##
## Best model has 5 parameters (max. 14 allowed):
## N P Al Mn Baresoil
## with correlation  0.4494
```

Внимание! Не надо оценивать достоверность результата процедуры BIO-ENV путем оценки достоверности мантеловской корреляции между $D_{BIO}$ и матрицей, полученной в результате применения BIO-ENV $D_{ENV}$. **Это будет жульничеством**, так как это уже максимально подобная матрица.

Для оценки достоверности полученного результата применяется пермутационный метод, основанный на **многократном повторении самой процедуры BIO-ENV**.

Внимание! Это занимает очень много времени

# Алгоритм оценки достоверности применения процедуры BIO-ENV



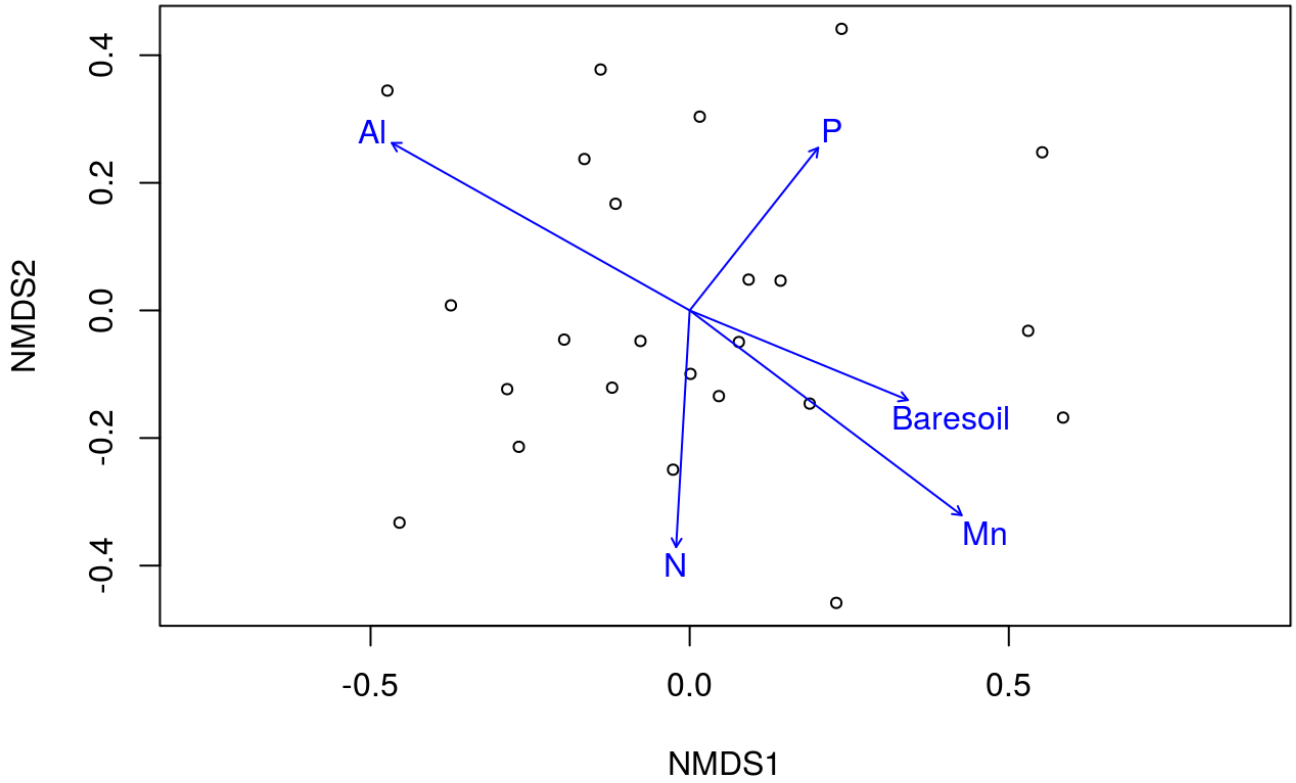

1. Применяем процедуру BIO-ENV и находим лучшее сочетание переменных в матрице-предикторе (ENV).
2. Пермутируем зависимую матрицу (BIO).
3. Применяем процедуру BIO-ENV к пермутированной матрице и вновь находим наилучшее сочетание и записываем значение мантеловской корреляции.

# Трактовка результатов BIO-ENV?

Задание: Постройте ординацию описаний в осях nMDS и отразите на этой диаграмме вектора, соответствующие результатам процедуры `BIO-ENV`

# Решение

```
plot(veg_ord, display = "site")
plot(envfit(veg_ord ~ N + P + Al + Mn + Baresoil, data = varechem ))
```



# Summary

- Оценку связи между ординацией объектов и значениями признаков, не использованных в ординации, можно осуществлять с помощью процедур `envfit()` и `ordisurf()`
- Степень сопряженности двух наборов пизнаков можно оценивать с помощью теста Мантела.
- Оценка достоверности теста Мантела и корреляций с признаками, вычисленными в процедуре `envfit()` проводится пермутационным методом
- С помощью процедуры `BIO-ENV` можно выявить набор переменных в матрице-предикторе, которые обеспечивают наибольшее сходство с завивисимой матрицей.

# Другое программное обеспечение

PRIMER 6.

Здесь реализована расширенная процедура BEST.
Она позволяет проводить не только полный перебор всех переменных в матрице-предикторе (Bio-Env), но и оптимизировать эту процедуру (BVStep). Кроме того, есть возможность оценивать достоверность результатов анализа. Но работает так же медленно.

# Что почитать


- Oksanen, J., 2011. Multivariate analysis of ecological communities in R: vegan tutorial. R package version 2–0.
- Clarke, K. R & Ainsworth, M. 1993. A method of linking multivariate community structure to environmental variables. Marine Ecology Progress Series, 92, 205–219.
- 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. (В этом издании приводятся ссылки на реализацию методов в R)