Multivarijatne metode II

Još primjera, metode temeljene na udaljenostima

Primjer – Ponašanje Dinarskog voluhara (*Dinaromys bogdanovi*) u okolinama različite složenosti

14 jedinki dinarskog voluhara (6m i 8ž) – u zatočeništvu (Zoo Zagreb)

Svaka je puštena u 4 okoline različite složenosti

Mjerena su vremena i frekvencije ponašanja

Category	Behavior (variable)	Definition
Entrance	Entrance	Instantaneous event. Every time the animal's hind legs enter the terrarium. If the animal remains at the door, this is not considered entrance but instead peeping
Space	Peeping	The animal is at the door of the *experimental* terrarium looking into or out of the *keeping* terrarium for a minimal duration of 0.8 s before entering, or is touching and sniffing the connecting door. If the animal does not enter but only looks into the *experimental* terrarium for as long as 0.4 s this is also considered peeping
	Closer half	The animal is in the closer half of the terrarium, at least with its hind legs. The only exception is when the animal is peeping into its *keeping* terrarium. The closer half is defined as the half from the entrance door to the center of the *experimental* terrarium
	Remote half	The animal is in the remote half of the *experimental* terrarium, i.e., its hind legs cross the center of the *experimental* terrarium. The remote half is defined as the half furthest from the entrance and starting at the center of the experimental terrarium
	Net	The animal has climbed or jumped onto the wire net at the top of terrarium and is not touching the floor or rocks of the terrarium. Its body may be touching the side walls. This behavior finishes when the animal jumps back onto the closer or remote half
Rock	Rock	The animal is standing on the rock with all four legs or with hind legs only. If the animal is fixed between rock and glass with one side touching the rock and the other the wall that is also considered as stay on the rock as described in the ethogram by Malenica (2011)
Rearing	Rearing	The animal is standing on its hind legs with its front paws either leaning against the walls or rocks, or not (in air)
Movement	B0 (staying still)	The animal is not moving forward or standing on its hind legs for a duration of 1 s or more but may be engaged in grooming or may be sniffing any surface
	Moving	The animal is moving. While moving the animal may also be sniffing rocks or the floor
Digging	Digging	The animal is notably moving the sawdust with its front legs
Jump	Jump	The animal is rising above the floor with all four legs
Grooming	Grooming	Self-grooming of any kind, fur maintenance, cleaning of body surface
Sniffing	Total sniffing	The animal is sniffing the floor (tiles or sawdust), walls of the terrarium, doors, or any part of the rocks
	Sniffing rock/novel object	The animal is sniffing rock (in setup 1) or of novel object (in novel object setup (4))
Novel object contact	Novel object	Every contact with the novel object (touching, biting, or carrying the novel object)

Setup 1 was a simple, open space with only one central rock (similar to the size of the animal), mimicking the lowest level of environmental complexity.

Setup 2 was more complex than setup 1, with six to eight rocks (similar to the size of the animal), separated just so that the animal could walk between them. This setup represented moderate complexity, enabling the animal to move or jump on rocks but leaving little possibility to hide.

Setup 3 was the most complex setup with many rocks, several times larger than the animal, covering most of the floor of the terrarium and set to create tunnel-like structures, giving the animals the opportunity of complete cover.

Setup 4 was designed similarly to setup 1, which served as its control, but instead of one centrally positioned rock, a 4-cm long piece of rubber tile was added, which was by smell, shape, and touch a completely new and unfamiliar object to the animals.

Behavioral response of the endemic Martino's vole *Dinaromys bogdanovi* (Martino 1922) to environmental complexity

Maša Ljuština[1] · Lea Vidatić[1] · Toni Safner[2] · Igor Ivanek[3] · Ivan Budinski[4] · Maja Damjanović[3] · Maja Lukač[5] · Duje Lisičić[1]

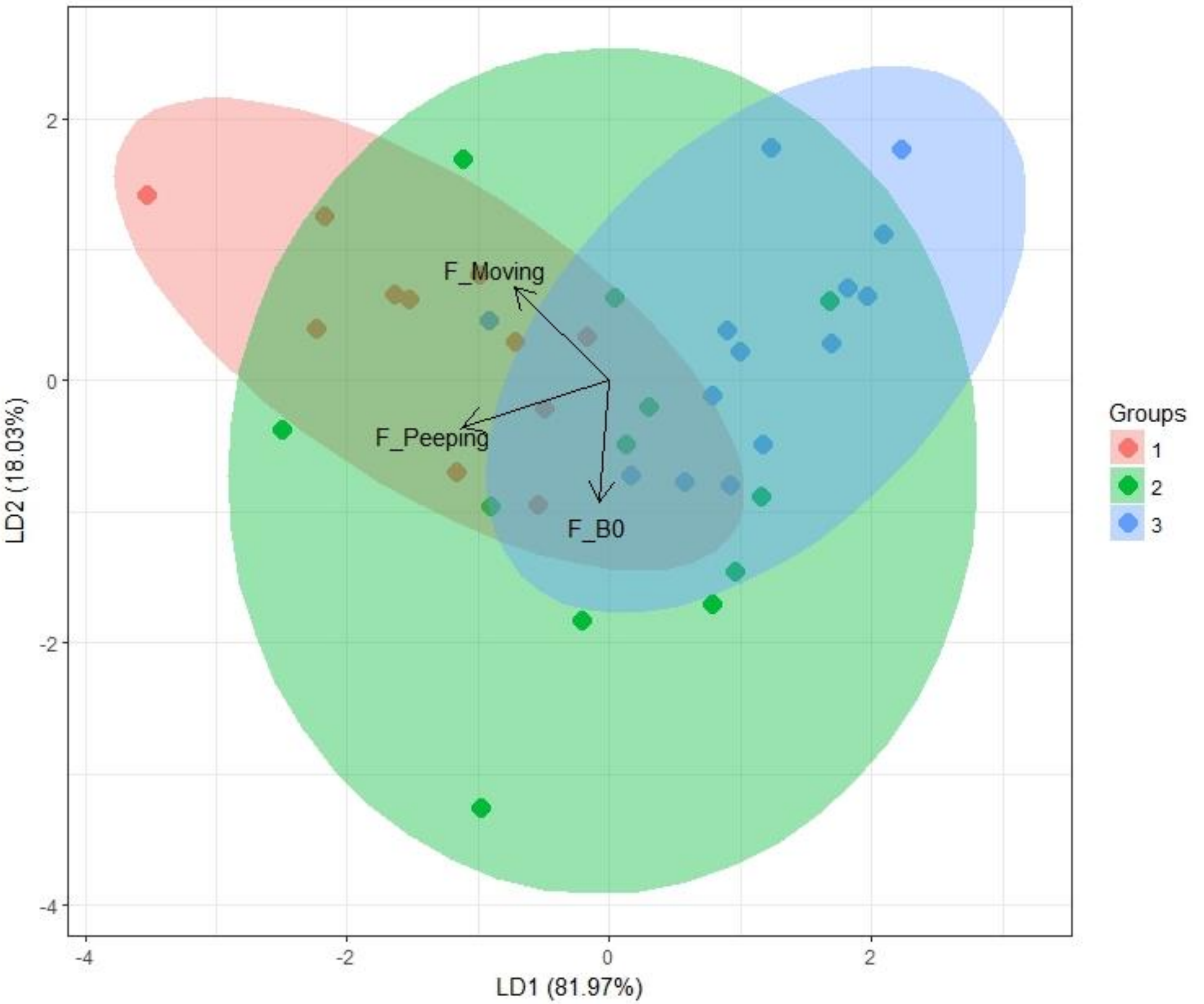
To summarize, the goals of this study were to determine if (1) differences in environmental complexity influence exploratory behavior, (2) the presence of a novel object influences exploratory behavior, and (3) exploratory behavior show seasonal variations.

LDA – odabrati varijable (ponašanja) koja najjače „razdvajaju" prve 3 okoline

Za početnu analizu je zadržano 8 varijabli (ponašanja) – je li to dovoljno?

Od tih 8 odabran minimum koji ima isti učinak kao svih 8 (stepwise addition/substraction)

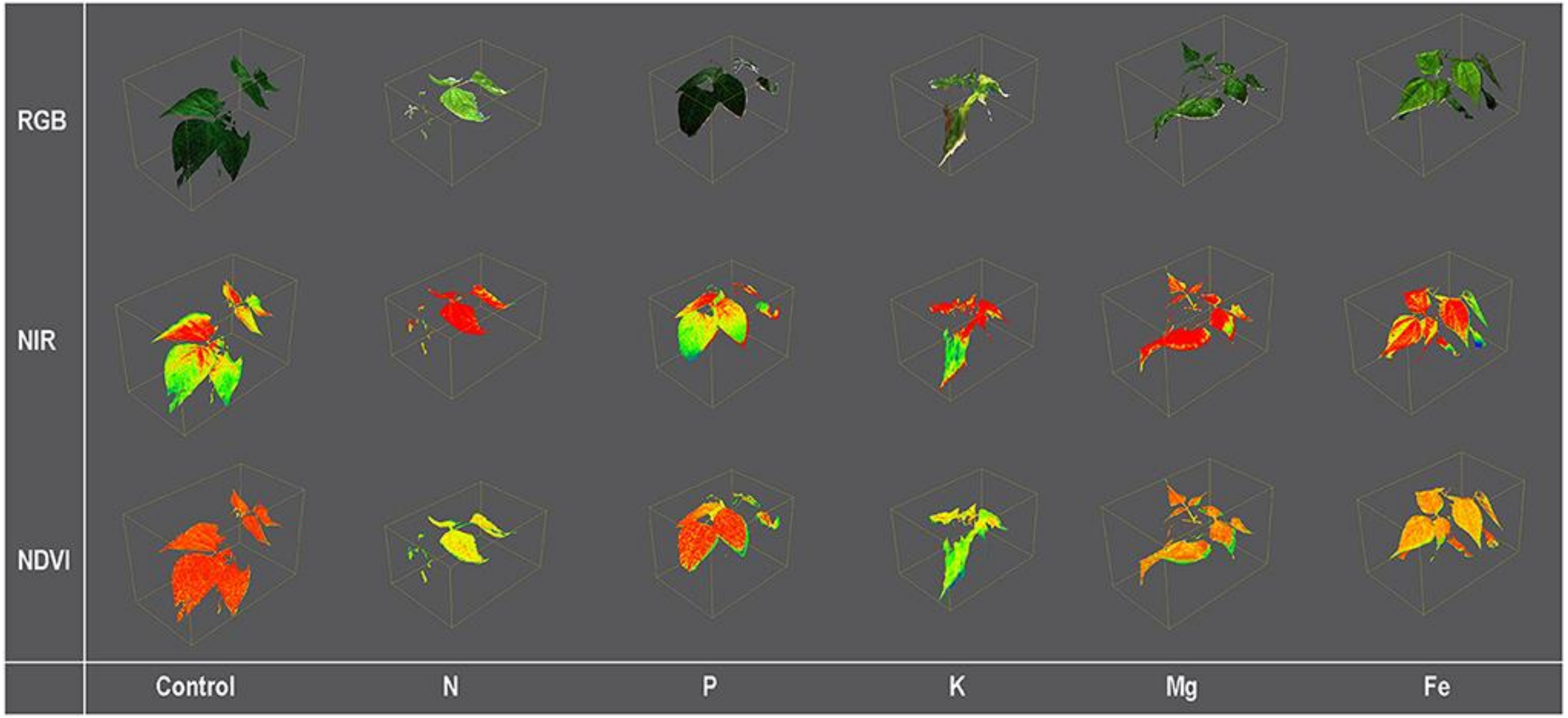
Rezultat pokazao da su 3 varijable dovoljne (Peeping, Moving i mirovanje)
LDA sa 3 varijable – grafički prikaz



Of the eight analyzed frequencies of behaviors, three were chosen by stepwise discriminant analysis as the best discriminating factor between three different setups. Wilks' λ was 0.6 ($p < 0.0001$) for frequency of peeping, 0.44 ($p < 0.0001$) for frequency of moving, and 0.38 ($p < 0.0001$) for frequency of not moving (B0). The most important factor contributing to the discrimination of behaviors between setups was the frequency of peeping (partial $R_2 = 0.4$), followed by the frequency of moving (partial $R_2 = 0.27$) and the frequency of not moving (B0) (partial $R_2 = 0.15$).
Canonical discriminant analysis based on the three selected behaviors revealed that the first canonical discriminant variate (LD1) explained 81.97% of the variation between behaviors in different setups (Fig. 1), and LD2 explained the remaining 18.03%. Discriminant function successfully classified 55% of animals in setup 1, 50% of animals in setup 2, and 71% of animals in setup 3, with a total classification success of 60%. Multivariate Wilks' $\lambda = 0.38$ ($p < 0.0001$) indicated that the centroids of behaviors in different setups were significantly different. Setups 1 and 3 were almost completely separated, while setup 2 overlapped with setups 1 and 3, with frequencies of moving and peeping highest in setup 1 (Fig. 1).

Primjer – Otkrivanje nedostatka hraniva kod graha

60 biljaka graha uzgojeno je u klima komorama u hidroponima
Nakon 10 dana podijeljene su u grupe po 10 biljaka kojima su u hranjivim otopinama ukinuta neka hranjiva (N, P, K, Mg, Fe i kontrolna skupina)
4 puta (svaka 3 dana nakon tretmana) mjereni su setovi morfoloških (8), multispektralnih (12) i 14 svojstava klorofilne fluorescencije.



This study aims (i) to compare groups of traits obtained by high-throughput phenotyping techniques (chlorophyll fluorescence, multispectral traits, and morphological traits) in their efficiency to discriminate among deficiency symptoms of five different plant nutrients [nitrogen (N), phosphorus (P), potassium (K), magnesium (Mg), and iron (Fe)], (ii) to generate rules, based on the variables within each group of traits, for classification of plants with different nutrient deficit treatments at early and prolonged nutrient deficiency stress, and (iii) to compare the efficiency of two different classification methods (LDA and recursive partitioning) for the classification of plants based on the specific nutrient deficiency.

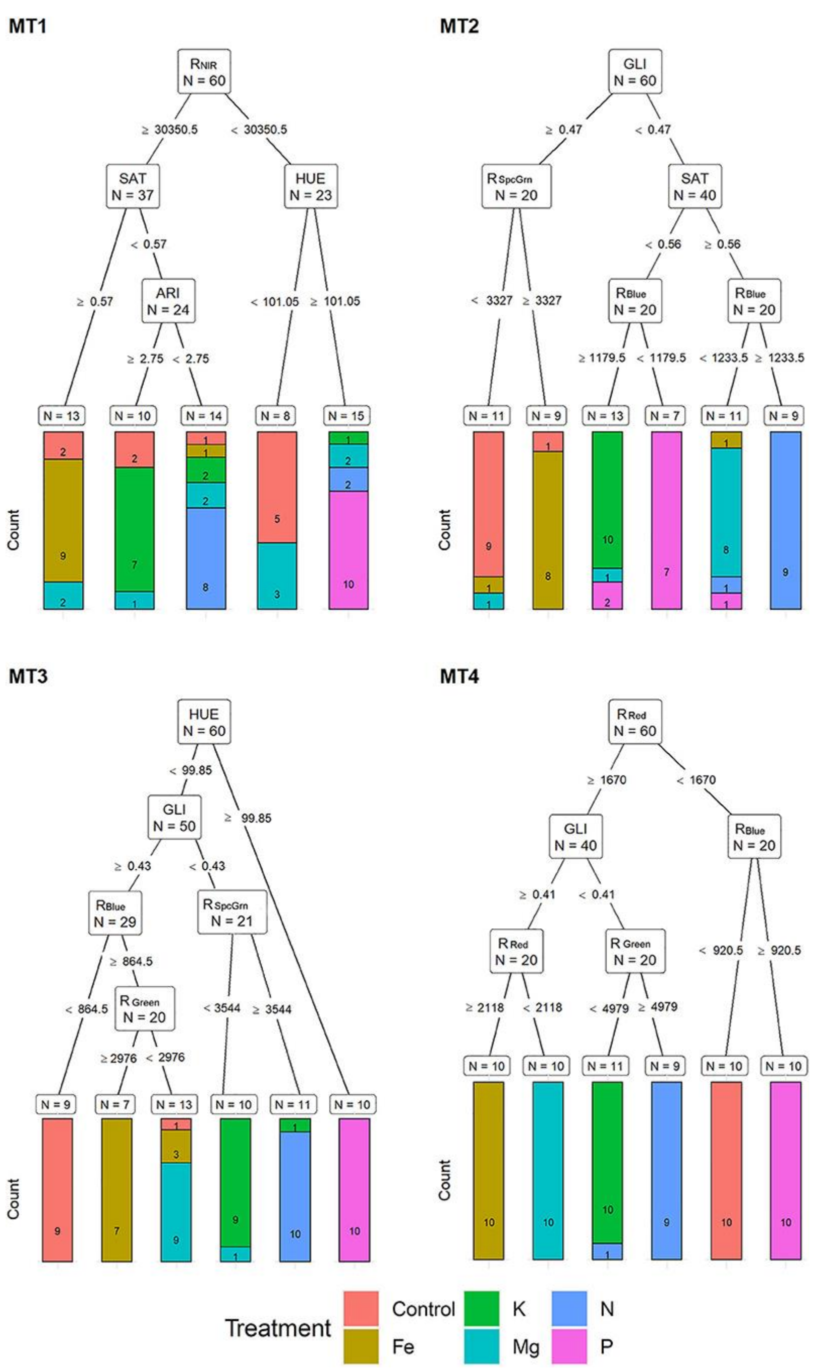
LDA za svaki set svojstava – koliko točno se po ovim svojstvima mogu prepoznati nedostaci hranjiva; koja svojstva su za to najvažnija

lda.model<- lda(treatment~., data = df)

Recursive partitioning (decision tree) – metoda strojnog učenja, robusna na odstupanje od pretpostavki normalnosti… funkcija rpart() u R-u (paket rpart)

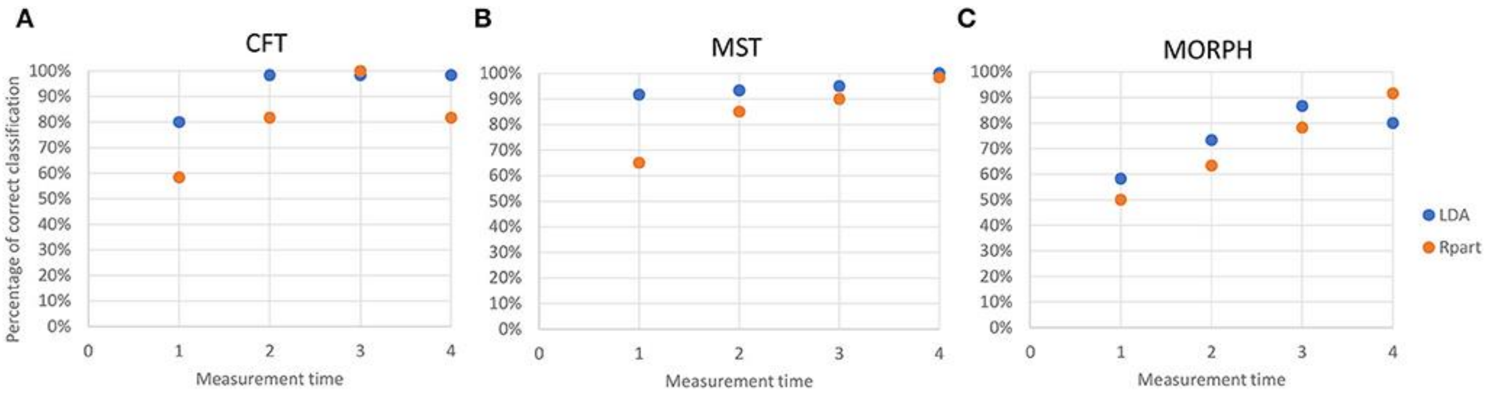
part.model<- rpart(treatment~., data = df)

This method is used to build classification or regression models using a multi-stage procedure where the resulting models can be represented as decision trees, making them easy to visualize. At each stage, the variable that best separates the data into groups defined by the classification variable (treatment) is selected, and data are subsequently split until the model reaches the best possible classification of data into predefined groups (treatments).



Visualization of classification tree for multispectral traits (MST). Each node shows the variable chosen as the best for the split in the data and the number of observations at that node (N). On the edges, between nodes, are threshold values of the split variables. Bar charts at each terminal node (leaf) represent the numbers of observations classified into each treatment (indicated by different colors). MT1 to MT4 represent measurement times.

Discriminant analysis based on MST correctly classified 91.6% plants at MT1, 93.3% plants at MT2, 95% plants at MT3, and 100% plants at MT4 into their respective nutrient deficiency groups



Multivarijatne metode temeljene na udaljenostima (distances; dissimilarities)

Indeksi udaljenosti mjere koliko su observacije (jedinke) mulitvarijatno različite (ako je svaka varijabla jedna dimenzija, onda je to multidimenzionalna udaljenost)

Indeksi sličnosti – mjere koliko su observacije slične (sličnost dvije jedinke sa istim vrijednostima svih varijabli=1)

Indeksa (mjerila) udaljenosti ima jako puno, razlikuju se u svojstvima (paket vegan) – odabir ovisi o vrsti varijable i namjeni (primjer Simple matching vs Jaccard u genetici!)

Za indekse koji su namijenjeni izračunavanju sličnosti (Jaccard, Simpson, Bray-Curtis) udaljenost se može izračunati kao: $D=1-S$

Postoje indeksi koji nisu metrike!!

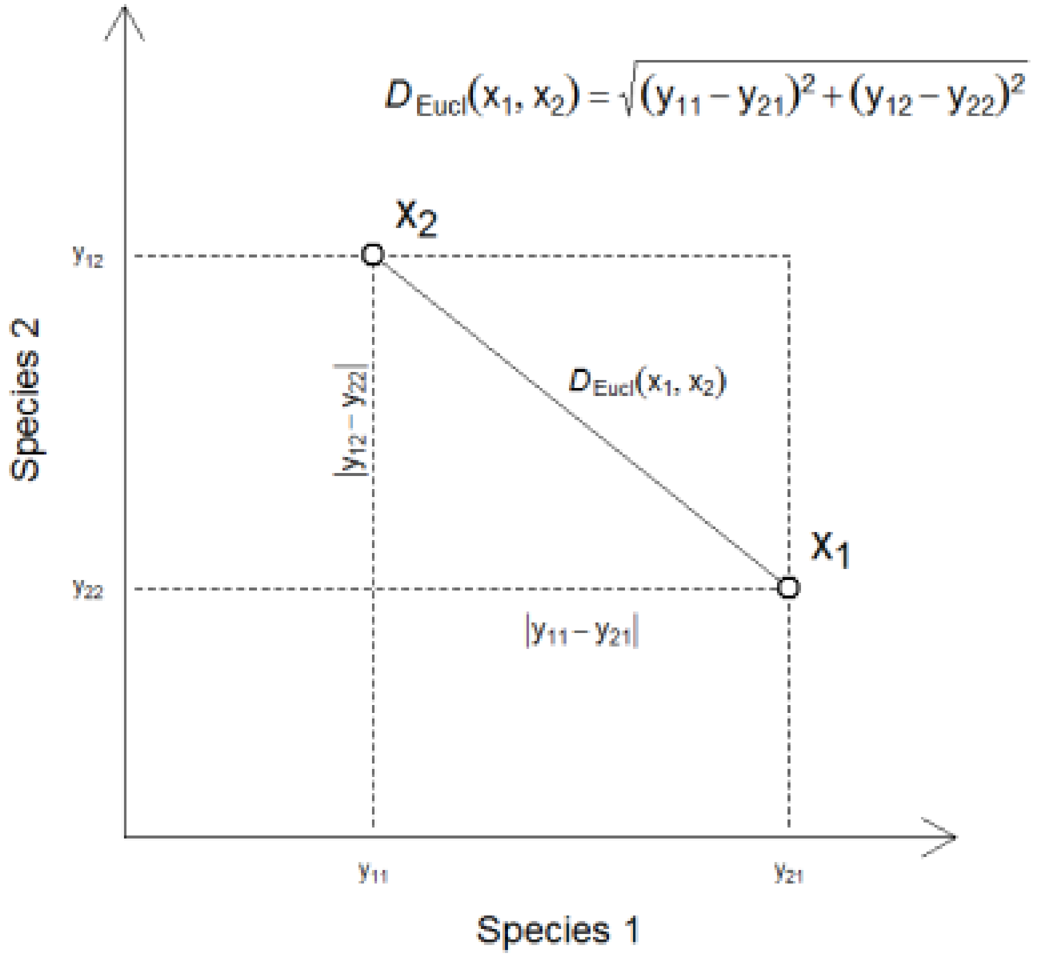
Metrike imaju sljedeća 4 svojstva: 1) Najmanja udaljenost je 0; 2) Udaljenosti su uvijek pozitivne (ili 0); 3) Udaljenost od a do b je jednaka kao i od b do a; 4) Pravilo trokuta – suma 2 stranice je uvijek veća od treće stranice

Euklidska distanca – prikaz za 2 (kvantitativne) varijable

Udaljenost točaka x1 i x2 je hipotenuza trokuta koju možemo izračunati korištenjem pitagorinog teorema

Duljine „kateta” možemo izračunati kao razlike vrijednosti dvaju varijabli između 2 jedinke.

Isti princip primjenjujemo i za više (kvantitativnih) varijabli



Neka od mjerila (indeksa) udaljenosti za kvantitativne varijable

Dissimilarity	Equation
Minkowski	$\left(\sum_{j=1}^{p} \lvert y_{1j} - y_{2j} \rvert^{\lambda}\right)^{1/\lambda}$
Euclidean ($\lambda = 2$)	$\sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$
City block (Manhattan: $\lambda = 1$)	$\sum_{j=1}^{p} \lvert (y_{1j} - y_{2j}) \rvert$
Canberra	$\frac{1}{p-q} \sum_{j=1}^{p} \frac{\lvert y_{1j} - y_{2j} \rvert}{(y_{1j} + y_{2j})}$
Bray–Curtis (Czekanowski)	$1 - \frac{2\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{1j} + y_{2j})} = \frac{\sum_{j=1}^{p} \lvert y_{1j} - y_{2j} \rvert}{\sum_{j=1}^{p} (y_{1j} + y_{2j})}$
Kulczynski	$1 - \frac{\left(\frac{\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{1j})} + \frac{\sum_{j=1}^{p} \min(y_{1j}, y_{2j})}{\sum_{j=1}^{p} (y_{2j})}\right)}{2}$
Chi-square	$\sqrt{\sum_{j=1}^{p} \frac{\left(y_{1j} \Big/ \sum_{j=1}^{p} y_{1j} - y_{2j} \Big/ \sum_{j=1}^{p} y_{2j}\right)^2}{\sum_{i=1}^{n} y_i}}$

Funkcija vegdist() iz paketa vegan računa velik broj indeksa udaljenosti (i sličnosti)

https://rdrr.io/cran/vegan/man/vegdist.html

Za binarne varijable (0,1), uglavnom se koriste indeksi sličnosti (simillarity)

a – obje jedinke imaju 1
b i c – jedna jedinka ima 1 a druga 0
d – obje jedinke imaju 0 (je li to sličnost?)

Coefficients	Similarity expression	Source
1.Simple matching (SM)	$\frac{a + d}{a + b + c + d}$	Sokal and Michener, 1958
5. Jaccard (J)	$\frac{a}{a + b + c}$	Jaccard, 1901
6. Sorensen-Dice (SD)	$\frac{2a}{2a + b + c}$	Sorensen, 1948; Dice, 1945

Ako među podatcima imamo različite tipove varijabli (kvantitativne, kategorijske, binarne…)

Gowerova distanca – prosjek distanci izračunatih za svaku pojedinačnu varijablu (izračuna se distanca za svaku varijablu – različitim indeksima – te se iz svih izračuna prosjek)

$$\frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}}$$

Gdje je w=1 ako se udaljenost za varijablu j može izračunati a 0 ako se ne može

ODABIR INDEKSA UDALJENOSTI MOŽE UTJECATI NA REZULTATE ANALIZA ZA KOJE SE TE UDALJENOSTI KORISTE!!

PODATKE JE POTREBNO STANDARDIZIRATI (center and scale)

Za svaki multivarijatni set podataka, bez obzira na tip varijabli, može se izračunati matrica udaljenosti (distanci) – prisutnost vrsta, obilježja staništa, morfološki i genetski biljezi…

	1	2	3	4	5	6	7	8	9	10
1	0	87.85	26.99	28.15	60.87	60.01	59.2	59.46	17.39	34.63
2	87.85	0	74.39	101.5	75.22	77.78	76.07	74.92	79.92	71.65
3	26.99	74.39	0	42.47	45.83	44.32	46.36	45.47	21.09	22.2
4	28.15	101.5	42.47	0	80.5	80.14	79.05	79.12	29.28	55.91
5	60.87	75.22	45.83	80.5	0	34.71	38.11	35.27	54.3	39.63
6	60.01	77.78	44.32	80.14	34.71	0	28.93	26.35	56.33	33.38
7	59.2	76.07	46.36	79.05	38.11	28.93	0	23.34	55.06	36.84
8	59.46	74.92	45.47	79.12	35.27	26.35	23.34	0	55.05	37.38
9	17.39	79.92	21.09	29.28	54.3	56.33	55.06	55.05	0	30.27
10	34.63	71.65	22.2	55.91	39.63	33.38	36.84	37.38	30.27	0

Hijerarhijsko klasteriranje – hclust()

Skupina metoda za grupiranje observacija (ili varijabli) u hijerarhijske grupe (klastere) prema sličnosti. Polazište je matrica sličnosti (ili udaljenosti). Grafički prikaz - dendrogram

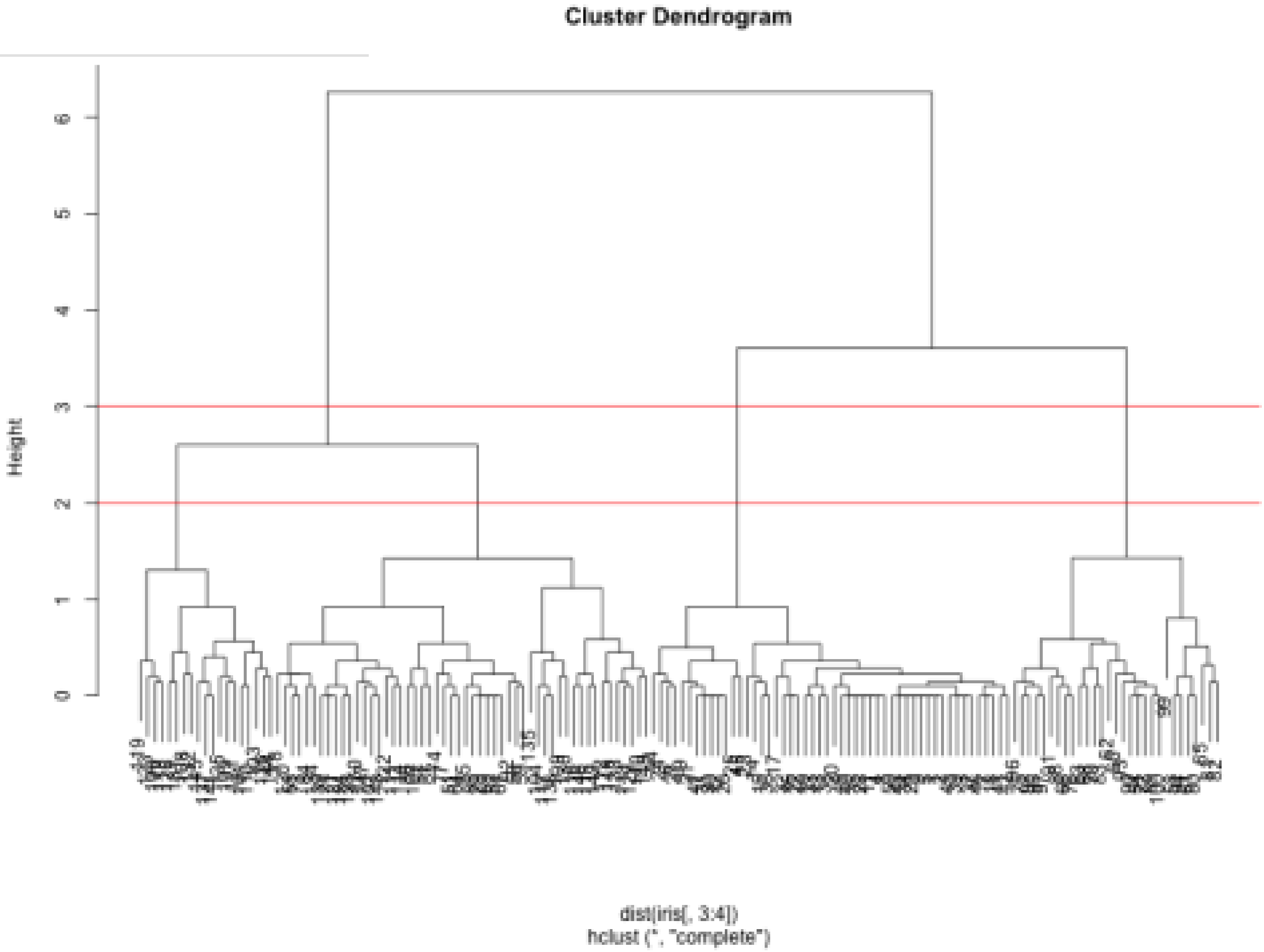
Aglomerativni i divizivni algoritmi

Aglomerativni algoritmi kreću od pozicije gdje je svaka jedinka jedan klaster, a zatim dvije najsličnije observacije spaja u jedan klaster te preračunava sličnosti svih preostalih jedinki sa novim klasterom. Ovaj postupak se nastavlja dok sve observacije nisu uključene u neki od klastera. Algoritmi se razlikuju u načinu kako mjere sličnosti jedinki sa postoječim klasterima: Single linkage (nearest neighbour), Complete linkage (furthest neighbour), average linkage (UPGMA)…

Divizivni algoritmi kreću od pozicije da su sve jedinke u istom klasteru, kojeg zatim razdvaja na manje grupe sve dok ne razdvoji sve jedinke (TWINSPAN…)

Interpretacija je problematična i nejednoznačna – nema jasno definiranog broja clustera

```
clusters <- hclust(dist(iris[, 3:4]))
plot(clusters)
```



Dendrogram često loše predstavlja odnose iz matrice distanci – MDS je metoda koju bi trebalo odabrati za potrebe prikaza odnosa

Taksonomija i slične potrebe – očekuju se hijerarhijski odnosi, svakako potrebno hijerarhijsko klasteriranje

https://www.r-bloggers.com/2016/01/hierarchical-clustering-in-r-2/
https://www.davidzeleny.net/anadat-r/doku.php/en:hier-agglom_examples

Particioniranje matrice udaljenosti – ANOSIM i PERMANOVA

Uspoređuje udaljenosti između observacija unutar istih predefiniranih grupa sa udaljenostima između observacija u različitim grupama- analogno ANOVA-i (ona uspoređuje varijabilnost a ne udaljenosti)

H0 – rang udaljenosti svih parova observacija u različitim grupama jednak je rangu udaljenosti parova observacija unutar istih grupa (otprilike – jedinke u različitim grupama jednako su slične kao jedinke u istim grupama)

$$R = \frac{\bar{r}_{\mathrm{B}} - \bar{r}_{\mathrm{W}}}{n(n-1)/4}$$

$R>0$ – Jedinke se više razlikuju između nego unutar grupa

$R=0$ – H0 se ne može odbaciti

$R<0$ – Jedinke unutar istih grupa su različitije nego jedinke između grupa

Particioniranje matrice udaljenosti – ANOSIM i PERMANOVA

Test je permutacijski (jedinke se nasumično raspoređuju u grupe, izračuna se R za svaki od nasumičnih rasporeda, usporede se R-ovi nasumičnih rasporeda sa „pravim” Rom)

Složeni dizajni (više grupirajućih varijabli, interakcije…) se teško mogu testirati na ovaj način

PERMANOVA – radi na sličan način ali sa sumama kvadrata udaljenosti

https://jkzorz.github.io/2019/06/11/ANOSIM-test.html

`vegan` : functions for community composition analysis

- `mantel()`
- `protest()`
- `anosim()`
- `adonis()` for PERMANOVA
- `betadisper()` for testing homogeneity of within-group variances

Primjeri 14-17

https://mjkeough.github.io/examples.html