www.emu.ee
Eesti Maaülikool
Estonian University of Life Sciences

Mitmemõõtmeline skaleerimine (*multidimentional scaling*, MDS)

Klasteranalüüs (*cluster analysis*)

Tanel Kaart

Mitmemõõtmelise statistika koolitus
Jaanuar 2016, Tartu

Mitmemõõtmeline skaleerimine

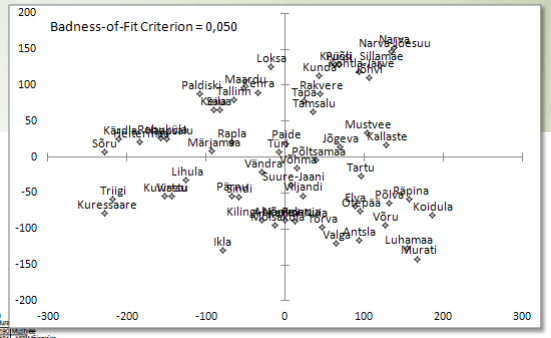
(*multidimentional scaling*, MDS)

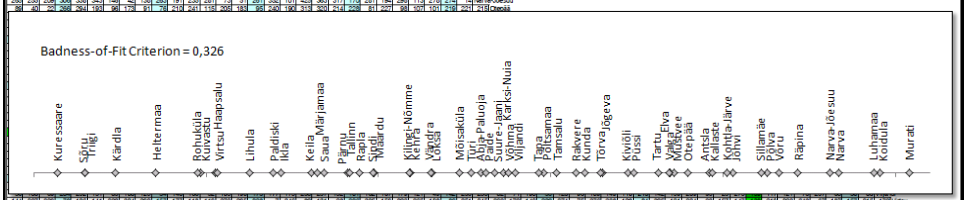
Mitmedimensionaalse skaleerimise eesmärgiks on objektide erinevuste või sarnasuste teisendamine distantsiks mitmemõõtmelises ruumis ja saadu esitamine kaardina objektide omavahelisest paiknemisest.

Kasutusala: kus iganes (geneetikas, ökoloogias, majandusteaduses, sotsioloogias, …)

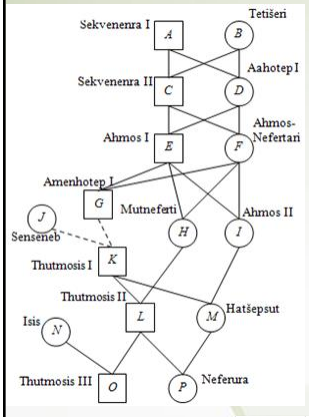
Mitmedimensionaalne skaleerimine

Eesti linnadevahelised kaugused

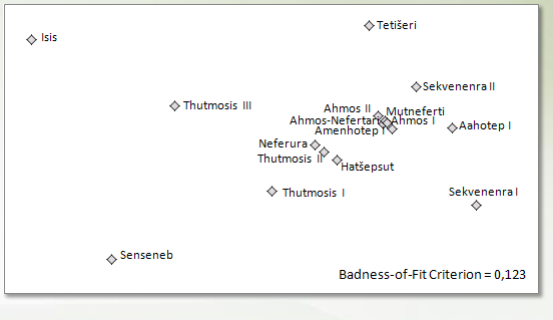




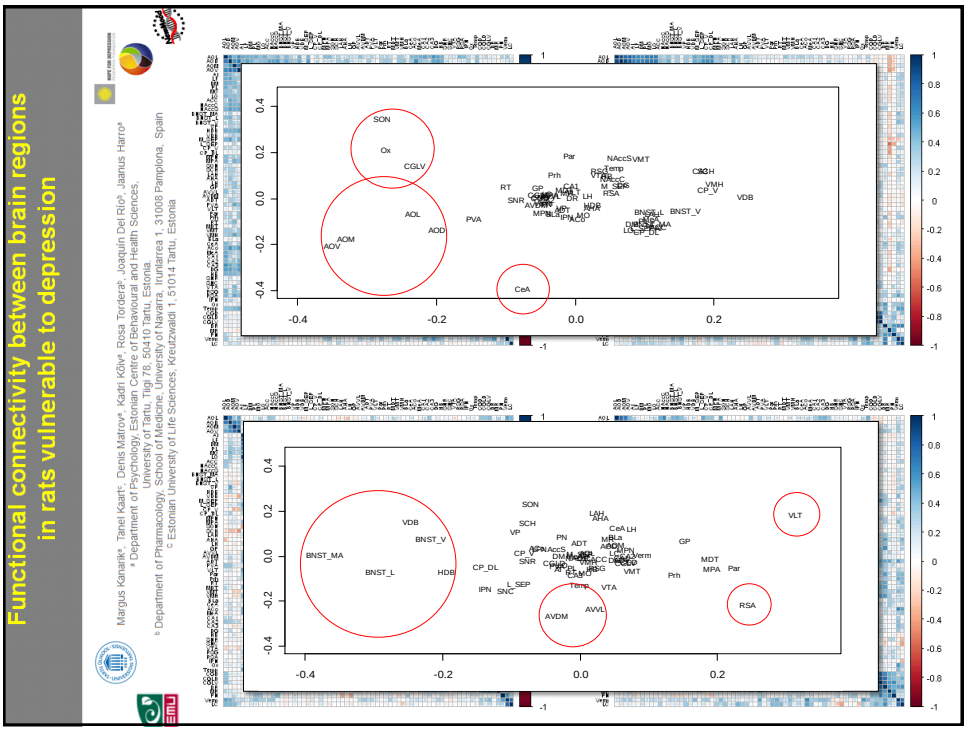
Mitmedimensionaalne skaleerimine



Egiptuse vaaraode 18. dünastia sugupuu algus



Vaarao	Sekvenenra_I	Tetisheri	Sekvenenra_II	Aahotep	Ahmos_I	Ahmos_Nefertari	Amenhotep_I	Mutneferti	Ahmos_II	Senseneb	Thutmosis_I	Thutmosis_II	Hatshepsut	Isis	Thutmosis_III	Neferura
Sekvenenra_I	1	0	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0	0,25	0,375	0,375	0	0,1875	0,37
Tetisheri	0	1	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0	0,25	0,375	0,375	0	0,1875	0,37
Sekvenenra_II	0,5	0,5	1	0,5	0,75	0,75	0,75	0,75	0,75	0	0,375	0,5625	0,5625	0	0,28125	0,562
Aahotep	0,5	0,5	0,5	1	0,75	0,75	0,75	0,75	0,75	0	0,375	0,5625	0,5625	0	0,28125	0,562
Ahmos_I	0,5	0,5	0,75	0,75	1,25	0,75	1	1	1	0	0,5	0,75	0,75	0	0,375	0,7
Ahmos_Nefertari	0,5	0,5	0,75	0,75	0,75	1,25	1	1	1	0	0,5	0,75	0,75	0	0,375	0,7
Amenhotep_I	0,5	0,5	0,75	0,75	1	1	1,375	1	1	0	0,6875	0,84375	0,84375	0	0,421875	0,8437
Mutneferti	0,5	0,5	0,75	0,75	1	1	1	1,375	1	0	0,5	0,9375	0,75	0	0,46875	0,8437
Ahmos_II	0,5	0,5	0,75	0,75	1	1	1	1	1,375	0	0,5	0,75	0,9375	0	0,375	0,8437
Senseneb	0	0	0	0	0	0	0	0	0	1	0,5	0,25	0,25	0	0,125	0,2
Thutmosis_I	0,25	0,25	0,375	0,375	0,5	0,5	0,6875	0,5	0,5	0,5	1	0,75	0,75	0	0,375	0,7
Thutmosis_II	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,9375	0,75	0,25	0,75	1,25	0,75	0	0,625	
Hatshepsut	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,75	0,9375	0,25	0,75	0,75	1,25	0	0,375	
Isis	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0,5	
Thutmosis_III	0,1875	0,1875	0,28125	0,28125	0,375	0,375	0,421875	0,46875	0,375	0,125	0,375	0,625	0,375	0,5	1	0
Neferura	0,375	0,375	0,5625	0,5625	0,75	0,75	0,84375	0,84375	0,84375	0,25	0,75	1	1	0	0,5	1,37



Klasteranalüüs

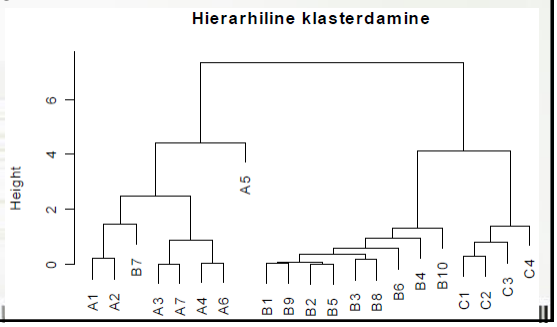
(*cluster analysis*)

Klasteranalüüsi eesmärgiks on kas tunnuste või uuritavate objektide rühmitamine. Kaks peamist algoritmi:

- hierarhiline klasterdamine;
- k-keskmiste klasterdamine.

Hierarhiline klasterdamine

- Loodusuurija sõitis Asustamata Saarele ja avastas seal suure hulga seni teadusele tundmatuid putukaid.
- Ta mõõtis oma lühikese saarelviibimise jooksul tervel hunnikul seninägematutel putukatel igasuguseid näitajaid (igatsorti pikkuseid ja mustrielementide arvu ja paljut muudki).
- Järgmiseks soovis loodusuurija määratleda, mitmesse alamliiki leitud putukad võiksid kuuluda.
- Et saada esimest ligikaudset lähendit, kust oma uurimistööga pihta hakata, soovis ta leida sarnaste putukate rühmad – kes oleksid siis alamliikide kanditaatideks.
- Selleks söötis ta oma andmed klasteranalüüsi teostavasse programmi, mis joonistas järgmise pildi:



Märt Mölsi loengukonspektist

Hierarhiline klasterdamine

... on hästi kasutatav siis, kui meil on suhteliselt vähe objekte või kui on oodata, et klastrid suhteliselt selgelt üksteisest eristuvad.

Hierarhiline klasteranalüüs põhineb väga lihtsal algoritmil:

samm-sammult hakatakse omavahel kokku panema kõige sarnasemaid objekte. Näiteks, kui leidub kaks täpselt ühesuguste tulemustega objekti, siis liidetakse nad esimesel sammul üheks klastriks, peale seda võrreldakse kõiki üksikobjekte ja juba tekkinud klastreid ja liidetakse jälle kõige sarnasemad omavahel jne.

Vaatluste omavahelise kauguse määramine:

Eukleidese kaugus: $d\big((x_1, y_1), (x_2, y_2)\big) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

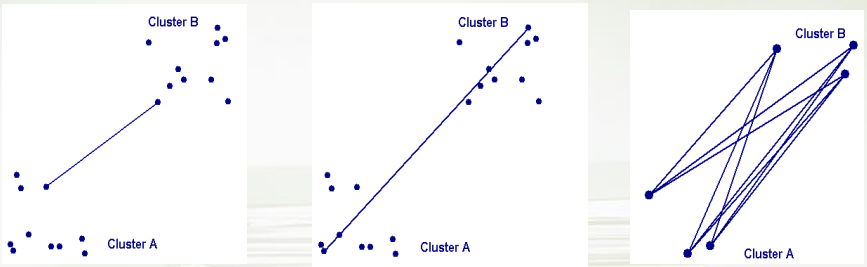
Manhattani kaugus: $d\big((x_1, y_1), (x_2, y_2)\big) = |x_1 - x_2| + |y_1 - y_2|$

...

Hierarhiline klasterdamine

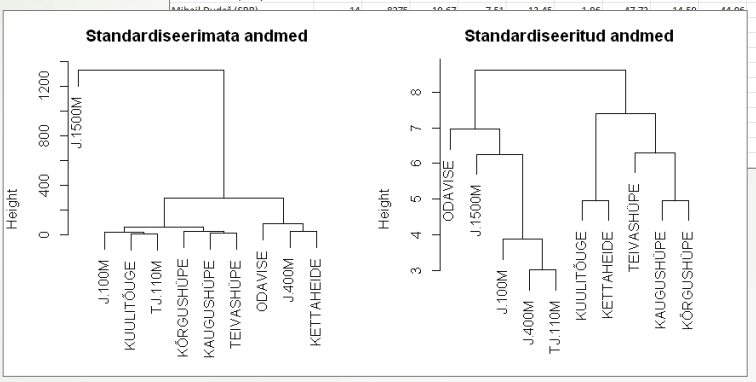
Klastritevahelise kauguse määramine:

“single” “complete” “average”



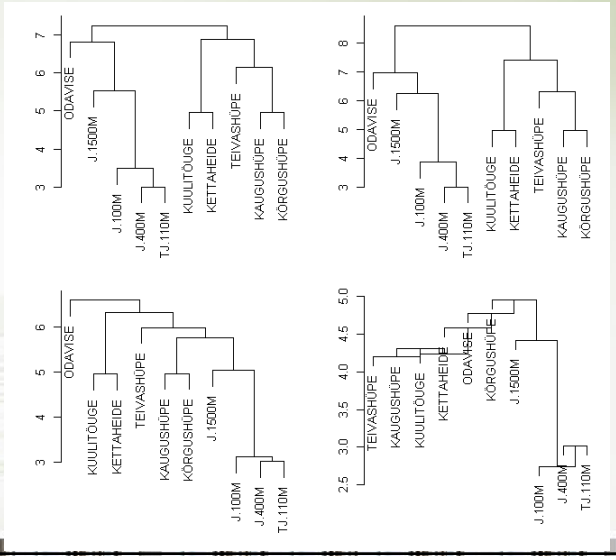
Näide: kergejõustiku MM 2013, meeste 10-võistlus

NIMI	KOHT	KOKKU	J 100M	KAUGUSH	KUULITÕU	KÕRGUSH	J 400M	TJ 110M	KETTAHEI	TEIVASHÜ	ODAVISE	J 1500M
Ashton Eaton (USA)	1	8809	10.35	7.73	14.39	1.93	46.02	13.72	45	5.2	64.83	269.8
Michael Schrader (GER)	2	8670	10.73	7.85	14.56	1.99	47.66	14.29	46.44	5	65.67	265.4
Damian Warner (CAN)	3	8512	10.43	7.39	14.23	2.05	48.41	13.96	44.13	4.8	64.67	270
Kevin Mayer (FRA)	4	8446	11.23	7.5	13.76	2.05	49.53	14.21	45.37	5.2	66.09	265
Eelco Sintnicolaas (NED)	5	8391	10.85	7.65	14.08	2.02	48.25	14.18	39.21	5.3	56.75	264.6
Carlos Chinin (BRA)	6	8388	10.78	7.54	14.49	1.96	48.8	14.05	45.84	5.1	59.98	276
Rico Freimuth (GER)	7	8382	10.6	7.22	14.8	1.99	48.05	13.9	48.74	4.9	56.21	277.8
Ilya Shkurenev (RUS)	8	8370	10.97	7.35	13.88	2.05	48.39	14.34	44.06	5.4	59.46	277
Willem Coertzen (RSA)	9	8343	10.95	7.44	13.88	2.05	48.32	14.3	43.25	4.5	69.35	264.6
Leonel Suárez (CUB)	10	8317	11.07	7.33	14.2	1.9	48.21	14.62	46.41	4.9	68.61	263.9
Pascal Behrenbruch (GER)	11	8316	10.95	7.19	15.86	1.99	48.4	14.46	45.66	4.7	67.07	277.2
Andrei Krauchanka (BLR)	12	8314	11.19	7.39	14.84	2.11	49.65	14.44	46.12	5.1	59.98	279.6
Gunnar Nixon (USA)	13	8312	10.84	7.8	14.68	2.14	48.56	14.57	42.38	4.6	57.97	275.8
										4.9	59.06	266.6
										5.1	65.31	277.9
										5	61.83	273
										5	62.89	273.2
										4.6	50.74	273.7
										4.9	59.63	288.3
										4.5	69.38	286.4
										5.1	64.38	275.9
										4.9	67.65	278.9
										4.5	57.3	278.5
										4.5	54.86	260.1



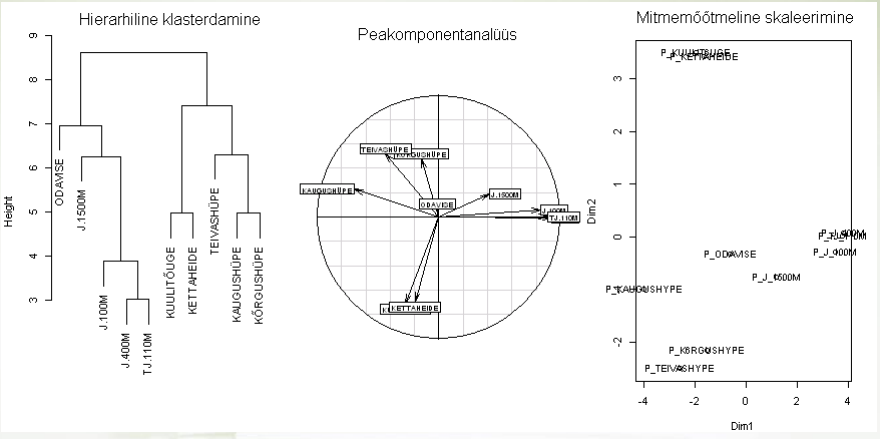
Näide: kergejõustiku MM 2013, meeste 10-võistlus

Erinevad klastrite vahelise kauguse määramise meetodid võivad anda erinevad tulemused!



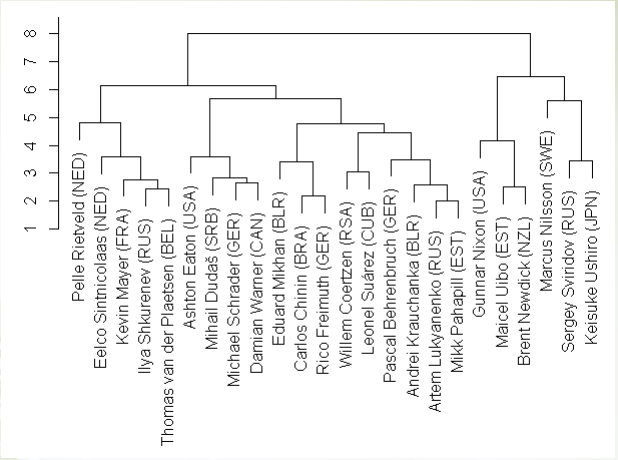
Näide: kergejõustiku MM 2013, meeste 10-võistlus

Grupeerimiseks/klasterdamiseks võib kasutada ka teisi meetodeid.



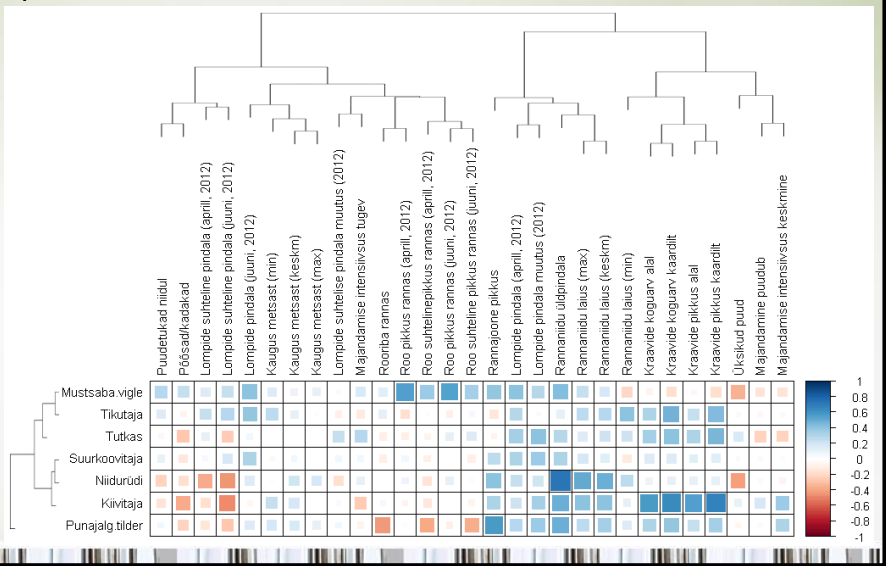
Näide: kergejõustiku MM 2013, meeste 10-võistlus

Klasterdada võib ka andmebaasi ridu.



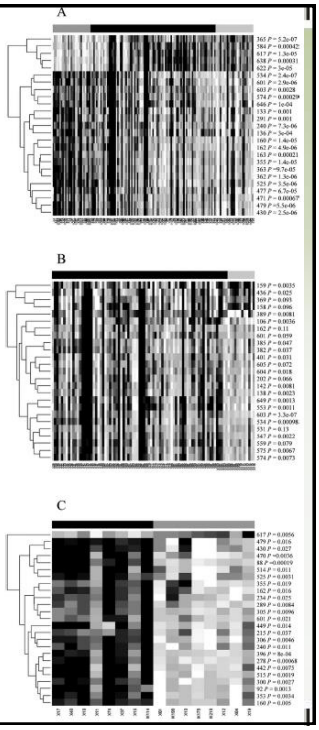
Klasteranalüüs tunnuste järjestamismeetodina

Eesmärgiks tuua visuaalselt selgemalt välja andmetes peituvad sarnasusmustrid.



Heatmap

Figure 3. (A) Heat map of noncoagulating (dark gray bar on top left), intermediate (black bar on top center), and well-coagulating (light gray bar on top right) milk samples. Statistical comparison with Student *t*-test was calculated between well-coagulating and noncoagulating groups. The 25 most significantly different m/z values with respective *P*-values can be seen in the right-hand side. Rows are arranged into clusters (dendrogram on left-hand side) based on pattern similarities. The lighter the shading of a data point, the more intense the signal is compared with the mean value for the respective m/z among all analyses. Gray points are close to the mean and dark tones indicate signal intensities below the mean; (B) heat map of milks with poor (black bar) and good (gray bar) coagulation ability; (C) heat map of noncoagulating (black bar) and well-coagulating (gray bar) milks.



J. Dairy Sci. TBC:1–8
doi:10.3168/jds.2011-4468

Comparison of the metabolic profiles of noncoagulating and coagulating bovine milk

H. Harzia,*†‡[1] K. Kilk,† I. Jõudu,*‡ M. Henno,* O. Kärt,* and U. Soomets†
*Institute of Veterinary Medicine and Animal Sciences, Estonian University of Life Sciences, Kreutzwaldi 46, Tartu 51006, Estonia
†Department of Biochemistry, Medical Faculty, University of Tartu, the Centre of Excellence for Translational Medicine, Ravila 19, 50411 Tartu, Estonia
‡Bio-Competence Centre of Healthy Dairy Products, Kreutzwaldi 1, 51014 Tartu, Estonia

k-keskmiste klasterdamine

... sobib grupeerimise meetodiks siis, kui objekte on nii palju, et hierarhilise klasteranalüüsi tulemus muutub ebaülevaatlikuks, kuid ka siis kui me oskame meile sobivat klastrite arvu ligilähedaselt ennustada ning ühtlasi soovime saada ka tekkivate klastrite kirjelduse nende tunnuste osas, mis on grupeerimise aluseks.

Algoritm:

- kõigepealt tuleb määrata klastrite arv, siis
- jagada objektid esialgsetesse klastritesse,
- arvutada välja klastrite keskpunktid ning
- hakata võrdlema igat objekti kõigi klastrite keskpunktidega; kui osutub, et objekti kaugus mõne muu klastri keskpunktist on väiksem kui selle klastri keskpunktist,milles ta parasjagu asub, siis tuleb objekt teise klastrisse ümber tõsta;
- peale objekti ümbertõstmist tuleb pöörduda uuesti sammu 3 juurde ja jätkata protsessi niikaua kui kõik objektid on klastris, mille keskpunktile nad kõige lähemal asuvad.

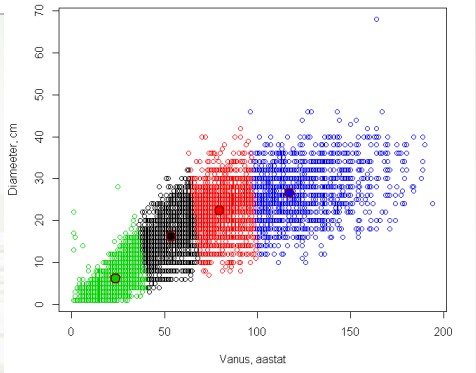
Näide: puude klasterdamine

ID	KUIV	PE	H100	KKT	ARENGUK	D	H	A
410057	0	MA	15.9	MS	K	14	12	55
410082	0	KU	23.6	JM	V	28	23	90
410111	0	KS	16.3	TR	K	14	13	60
410151	0	MA	14.8	MS	L	12	11	55
410173	0	MA	17.5	KN	L	6	4	29
410179	0	MA	22	MS	K	26	22	100
410202	0	MA	17.2	MS	K	16	15	70
410230	0	MA	24	JM	V	26	23	85
410248	1	MA	19.5	KM	K	24	20	110
410251	1	MA	17.5	SN	N	4	5	26

- 29539 puud,
- eesmärk jagada puud klastritesse diameetri (D) ja vanuse (A) alusel.

Näide: puude klasterdamine

```
> table(puu_kmean1$cluster) # klastrite suurused

   1    2    3    4
8964 7493 5941 3026
> # klasterdamise aluseks olnud tunnuste keskmised klastrite kaupa
> puu_kmean1$centers
       [,1]     [,2]
1  53.52443 16.21586
2  79.33298 22.34472
3  23.54385  6.14543
4 117.20787 26.56345
```



Miks tundub, et klastrid on moodustatud eelkõige puu vanuse järgi?

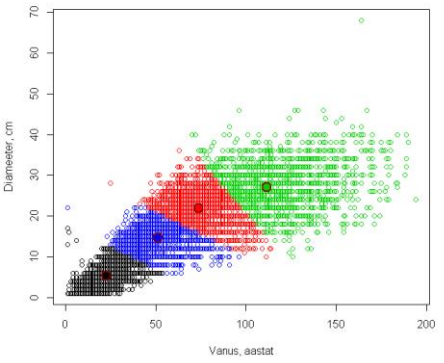
Vastus – puude vanust märkivad arvud (eelkõige nende varieeruvus) on suuremad, kui diameetrit märkivad arvud.
Soovides, et klasterdamisalgoritm peaks puu diameetrit sama oluliseks kui vanust, tuleb klasterdamiseks kasutatavad tunnused eelnevalt standardiseerida.

Näide: puude klasterdamine

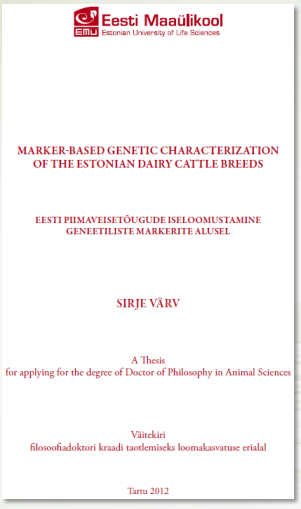
Klasterdamise tulemused peale vanuse ja diameetri standardiseerimist

```
> table(puu_kmean2$cluster) # klastrite suurused

   1    2    3    4
5346 8295 3836 7947
>
> # klasterdamise aluseks olnud tunnuste keskmised klastrite kaupa
> puu_kmean2$centers
        [,1]        [,2]
1 -1.2838316 -1.4539809
2  0.3980414  0.6267682
3  1.6430013  1.2736169
4 -0.3449028 -0.2908839
>
>  # Mis tüüpi puud mingitesse klastritesse kuuluvad?
> table(puu_kmean2$cluster, puud1$ARENGUKL)

      A    K    L    N    S    V    Y
  1   1  148 2480 2566    3  110   38
  2   0 5003    1    0    0 1592 1699
  3   0  895    0    1    3  697 2240
  4   0 5825 1529    0    1  358  234
```

Kuidas on puud klasterdunud sõltuvalt arenguklassist (A - lage, N - noorendikud, L - latimets, K - keskealised, V - valmiv, Y - küps, S - selguseta)?



Näide: Sirje Värv, veiste klasterdamine geneetiliste markerite alusel


Eesti Maaülikool
Estonian University of Life Sciences

MARKER-BASED GENETIC CHARACTERIZATION OF THE ESTONIAN DAIRY CATTLE BREEDS

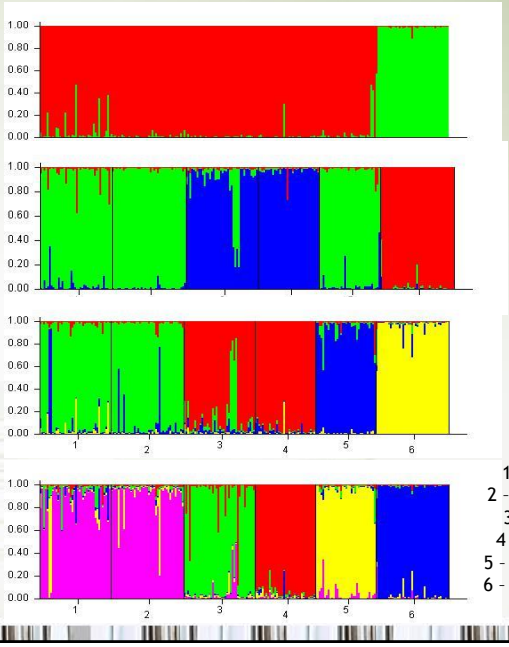
EESTI PIIMAVEISETÕUGUDE ISELOOMUSTAMINE GENEETILISTE MARKERITE ALUSEL

SIRJE VÄRV

A Thesis
for applying for the degree of Doctor of Philosophy in Animal Sciences

Väitekiri
filosoofiadoktori kraadi taotlemiseks loomakasvatuse erialal

Tartu 2012



1 - EN,
2 - WFC,
3 - ER,
4 - EHF,
5 - SwRP,
6 - DkJer