Klasteranalüüs

Klasteranalüüsi kasutatakse nii objektide kui ka tunnuste grupeerimiseks nende omavahelise sarnasuse alusel.

Objektide grupeerimise ...

... korral on eesmärgiks leida mingite tunnuste alusel sarnased (st sarnaselt vastanud või sarnaselt käituvad) vastajad või muud vaatlusalused objektid ning moodustada nendest grupid e **klastrid**. Näiteks, kui on tegemist küsitlusega, kus uuritakse eestlaste suhtumist erinevates EL'iga seonduvates küsimustes, siis võib meid huvitada, millised on nn tüüpilised inimeste grupid: hüpoteetiliselt võiks olla üks grupp selline, kelle suhtumine on igas aspektis negatiivne; teine selline, kes näeb kõike EL'iga seonduvat positiivsena; kolmas selline, kes on liitumise vastu kuid muidu EL'i suhtes hästi meelestatud; neljas selline, kes on EL'i suhtes neutraalsed kuid peavad liitu astumist paratamatuks jne jne. Saanud klasteranalüüsi tulemusena kätte andmetele põhineva tüpoloogia e sarnaste tulemustega/vastustega objektide grupid, saab hakata edasi uurima, millise "taustaga" vastajad ühte ja teise gruppi kuuluvad.

Klasteranalüüsi puhul võime rääkida kahest erinevast klasterdamise meetodist: **hierarhilisest klasteranalüüs**ist ja **k-keskmiste klasteranalüüs**ist.

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[1] 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 jne. Seega hierarhilise klasterdamise puhul on alguses sama palju klastreid kui uuritavaid objekte ja liitmise protsess lõppeb kui kõik objektid on ühes grupis e klastris. Loomulikult ei huvita meid ei alg- ega lõppseis vaid kogu küsimus peitub selles, et leida nn optimaalne klastrite arv. Siin tuleb uurijal lähtuda eelkõige oma vaistust ja mudeli interpreteerimise võimalustest, kuid selle juures on suureks abiks mõnede matemaatiliste parameetrite jälgimine. Seega *klasteranalüüs on meetod, kus põhimõtteliselt ei ole ühte ja ainuõiget lahendit vaid kus tulemust peab hindama tema interpreteeritavuse ja vahest ka mingile teooriale vastavuse seisukohast.*

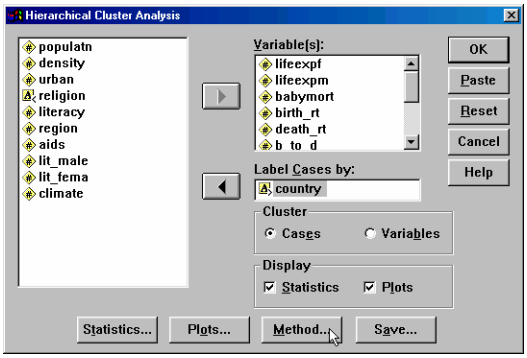
N Järgnevas võtame kasutusele väikese näiteandmestiku, kus on toodud erinevaid arvnäitajaid mõningate riikide kohta. Eesmärgiks on sündivuse, suremuse, keskmise eluea jt sarnaste näitajate alusel riigid klasterdada nii, et ühesuguste näitajatega riigid oleksid ühes grupis.

Hierarhilise klasteranalüüsi tellimine SPSS'is ja **metoodilised valikud**:

- Vali *Analyze/Classify/Hierarchical Cluster...*
- Paiguta klassifitseerimise aluseks olevad tunnused väljale *Variable(s):*

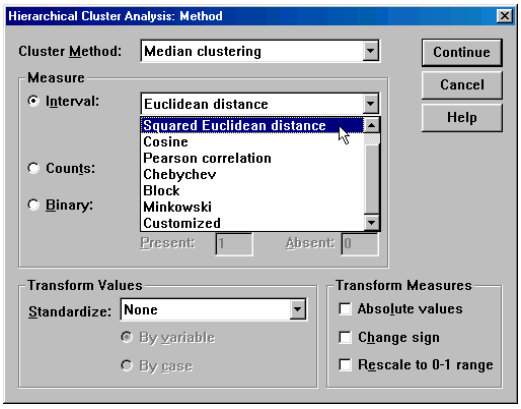
[1] Sarnasust tuleb mõõta matemaatiliste meetoditega, millest teen lühikese ülevaate allpool.

- Kui mingi tunnus sisaldab objektide identifikaatoreid e nimetusi, siis paiguta see väljale *Label Cases by:*
  NB! selleks saab kasutada ainult tunnuseid tüübiga *string*
- Objektide klassifitseerimise korral jäta kastis *Cluster* aktiivseks väli *Cases*

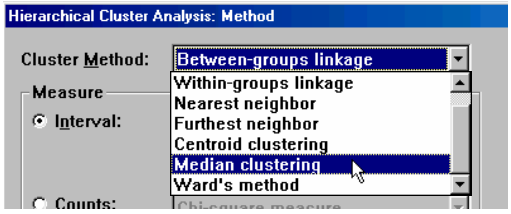


- **Klasterdamise matemaatilise meetodi valikuks** vajuta nuppu *Method...*

1. Vali sobiv **sarnasuse** e **kauguse mõõt** kastis *Measure*:
   arvtunnuste puhul väljalt *Intreval:*
   binaarsete e kahe väärtusega tunnuste puhul väljalt *Binary:*



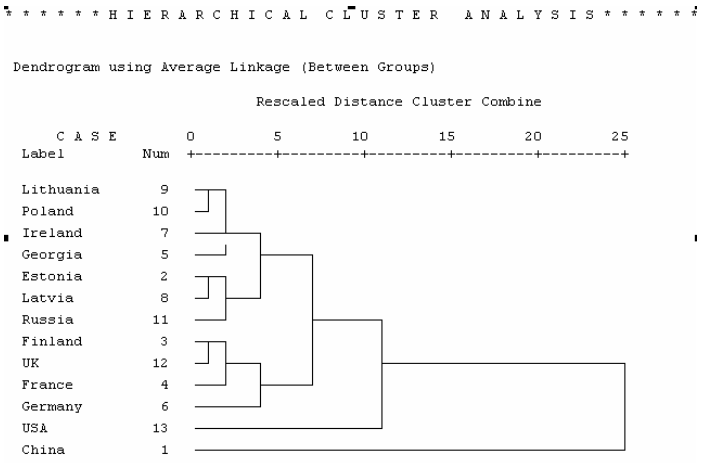
2. Vali sobiv klasterdamise meetod väljal *Cluster Method:*



3. Kui klasterdamise aluseks olevad tunnused on mõõdetud erinevatel skaaladel, siis vali tunnuste standardiseerimiseks[2] kastis *Transform Values* väljal *Standardize: Z scores*

[2] See on vajalik selleks, et suuremate väärtuste ja hajuvusega (seega ka absoluutväärtuselt suuremaid erinevusi andval) skaalal mõõdetud tunnused ei mõjutaks klasterdamise tulemust enam kui väikese hajuvusega tunnused.

- Liitmise protsessi ja tulemust kirjeldava diagrammi e **dendrogrammi** tellimiseks vajuta nuppu *Plots...* ning märgista väli *Dendrogram*; kastis *Icicle* märgista väli *None*.



Diagrammi ülaservas olev skaala väljendab standardiseeritud kaugusmõõtu. Jooniselt on näha, et meie näites osutusid väga sarnasteks:
I. Leedu ja Poola, millele liitusid ka Iiri ja Gruusia
II. Eesti ja Läti, millele liitus ka Venemaa
III.Soome ja Suurbritannia, millele liitus Prantsusmaa ja ka Saksamaa
Hiina ja eriti Ameerika olid teistest riikidest niivõrd erinevad, et nad liideti eelviimasel ja viimasel sammul kui kõik teised riigid olid juba omavahel ühte klastrisse liidetud.
Seega võib antud näite puhul rääkida kolmest tekkinud klastrist ja kahest teistest erinevast objektist. Põhimõtteliselt võib nüüd uurija võtta vastu otsuse jätta need kaks objekti mudelist välja ja korrata analüüsi.

- Põhiaknas olev nupp *Statistics...* võimaldab tellida:

  Agglomeration schedule liitmise protsessi kirjeldava tabeli

Liidetavad objektid (nr tuleneb rea numbrist andmestikus). Klastri tähiseks saab esimese liidetava objekti number!

Klastrite vaheline kaugus liitmisel

Samm, millel antud objekt esimest korda teisega liideti. (Kui väärtus on 0, siis pole enne liidetud)

Liitmise samm

Järgmine samm, millel tekkinud klaster mõne teisega liidetakse

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	9	10	.475	0	0	6
2	2	8	.811	0	0	5
3	3	12	.999	0	0	4
4	3	4	2.464	3	0	8
5	2	11	2.501	2	0	9
6	7	9	2.797	0	1	7
7	5	7	3.979	0	6	9
8	3	6	6.270	4	0	10
9	2	5	7.586	5	7	10
10	2	3	11.852	9	8	11
11	2	13	19.268	10	0	12
12	1	2	47.050	0	11	0

Tabel kirjeldab täpselt sedasama protsessi, mida väljendas ka dendrogramm, ainult natuke täpsemal kujul. Näeme, et viimasel ja eelviimasel sammul on liidetavate klastrite kaugus väga suur ning, et esimene kauguse "hüppeline" suurenemine on kaheksandal sammul – enne seda on meil 6 klastrit (täpsemalt 3 klastrit ja 3 üksikobjekti).

Proximity matrix üksikobjektide vaheliste kaugusmõõtude maatriksi

Proximity Matrix

Case	Squared Euclidean Distance												
	1:China	2:Estonia	3:Finland	4:France	5:Georgia	6:Germany	7:Ireland	8:Latvia	9:Lithuania	10:Poland	11:Russia	12:UK	13:USA
1:China		46.468	53.371	57.050	25.980	70.558	41.773	45.012	33.902	38.549	39.220	59.414	53.300
2:Estonia	46.468		8.104	14.728	8.241	16.132	9.147	.811	2.847	2.931	3.080	7.781	22.997
3:Finland	53.371	8.104		2.192	11.825	5.425	2.642	12.331	5.059	2.810	11.558	.999	16.605
4:France	57.050	14.728	2.192		15.173	8.068	3.927	20.184	9.366	6.923	18.970	2.736	10.032
5:Georgia	25.980	8.241	11.825	15.173		29.361	5.199	9.843	2.447	4.291	11.460	14.076	16.956
6:Germany	70.558	16.132	5.425	8.068	29.361		13.928	20.593	16.687	12.063	16.179	5.317	28.951
7:Ireland	41.773	9.147	2.642	3.927	5.199	13.928		13.398	3.302	2.292	13.572	4.069	11.665
8:Latvia	45.012	.811	12.331	20.184	9.843	20.593	13.398		4.355	4.990	1.922	12.860	27.393
9:Lithuania	33.902	2.847	5.059	9.366	2.447	16.687	3.302	4.355		.475	5.215	6.829	16.347
10:Poland	38.549	2.931	2.810	6.923	4.291	12.063	2.292	4.990	.475		5.036	4.372	16.029
11:Russia	39.220	3.080	11.558	18.970	11.460	16.179	13.572	1.922	5.215	5.036		13.339	28.871
12:UK	59.414	7.781	.999	2.736	14.076	5.317	4.069	12.860	6.829	4.372	13.339		16.099
13:USA	53.300	22.997	16.605	10.032	16.956	28.951	11.665	27.393	16.347	16.029	28.871	16.099	

This is a dissimilarity matrix

Cluster Mebership -

tabeli, kus on iga objekti jaoks kirjas, millisesse klastrisse ta kuulub (klastrite arv tuleb ise määrata)

Cluster Membership

Case	5 Clusters
1:China	1
2:Estonia	2
3:Finland	3
4:France	3
5:Georgia	4
6:Germany	3
7:Ireland	4
8:Latvia	2
9:Lithuania	4
10:Poland	4
11:Russia	2
12:UK	3
13:USA	5

- Kui sobiv klastrite arv on leitud, siis saab nupu *Save...* abil tellida andmetabelisse tunnuse(d), kus iga objekti puhul on väärtuseks klastri number, millesse ta kuulub.

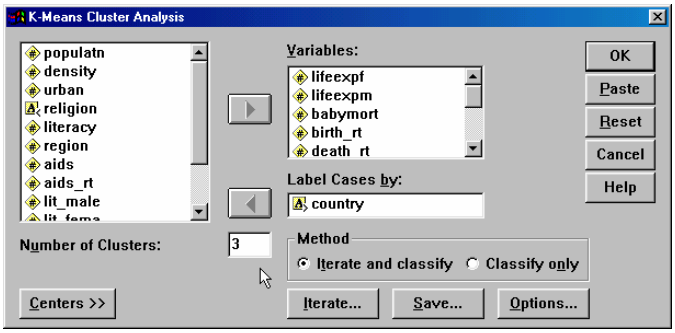
k-keskmiste klasterdamine

... 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. **k-keskmiste klasterdamise algoritm on samuti lihtne**:

1. kõigepealt tuleb määrata klastrite arv, siis
2. jagada objektid esialgsetesse klastritesse, edasi
3. arvutada välja klastrite keskpunktid ning
4. 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.
5. Peale objekti ümbertõstmist tuleb pöörduda uuesti sammu 3 juurde ja jätkata protsessi niikaua kui kõik objektid on klastris, mille keskpunktile nad kõige lähemal asuvad.

k-keskmiste klasteranalüüsi tellimine SPSS'is ja **metoodilised valikud**:

- Vali *Analyze/Classify/K-Means Cluster...*
- Paiguta klassifitseerimise aluseks olevad tunnused väljale *Variable(s):*
- Kui mingi tunnus sisaldab objektide identifikaatoreid e nimetusi, siis paiguta see väljale *Label Cases by:*
  NB! selleks saab kasutada ainult tunnuseid tüübiga *string*
- Määra klastrite arv väljal *Number of Clusters:*



- Vajutades nuppu *Iterate...* avaneb võimalus valida, kas

  A. klastrite keskpunktid arvutatakse ümber peale iga üksiku objekti ümbertõstmist (selleks märgista väli *Use running means*) või

  B. peale seda kui kõik objektid on antud keskpunktide suhtes võrreldud ja vajadusel ümber tõstetud (selleks jäta väli *Use running means* märgistamata)

  PS! Suurte andmestike korral võib esimese variandi järgi arvutamine suhteliselt kaua aega võtta, kuna ümberarvutamisi tuleb arvutil sooritada väga palju!

  Samas saad suurendada ka lubatud ümberpaigutamiste arvu (*Maximum Iterations:*)

k-keskmiste klasteranalüüsi põhitulem:

Final Cluster Centers

	Cluster		
	1	2	3
Naiste keskmine eluiga	75	79	77
Meeste keskmine eluiga	66	73	68
Surnud imikuid 1000 elava kohta	23.8	6.6	16.6
Sünde 1000 elaniku kohta	14.3	12.8	14.3
Surnuid 1000 elaniku kohta	11	10	11
Sündivuse ja suremuse suhe	1.38	1.26	1.36
Keskmine laste arv	2.0	1.8	2.0

Selles tabelis on toodud tekkinud klastrite keskpunktid kõigi klasterdamise aluseks olevate tunnuste lõikes –

näeme, et esimene klaster on kõige madalama eluea ja kõige suurema imikute surevusega, teine klaster aga kõige kõrgema eluea ja kõige madalama imikute surevusega. Teiste tunnuste osas on klastrite erinevused väiksemad

PS! tabelis antav komakohtade arv sõltub iga tunnuse defineeritud komakohtade arvust!

Distances between Final Cluster Centers

Cluster	1	2	3
1		19.276	7.781
2	19.276		11.593
3	7.781	11.593	

Selles tabelis on toodud tekkinud klastrite omavahelised kaugused –

näeme, et esimene ja teine klaster erinevad omavahel kõige enam, kusjuures kõige sarnasemad on omavahel esimene ja kolmas klaster

Number of Cases in each Cluster

Cluster	1	3.000
	2	4.000
	3	3.000
Valid		10.000
Missing		.000

Selles tabelis on toodud igassse klastrisse kuuluvate objektide arv –

näeme, et esimesse klastrisse kuulub 3 riiki, teise 4 ja kolmandasse 3 riiki. Kokku oli 10 anlüüsitavat riiki, millest ühteigi ei jäetud puuduvate väärtuste pärast analüüsist välja.

- Nupu *Options...* abil on lisaks võimalik tellida:

  Initial cluster centers esialgsete klastrite keskpunktid

  ANOVA table dispersioonanalüüsi tabel klastrite võrdlemiseks klasterdamise aluseks olevate tunnuste lõikes

ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Naiste keskmine eluiga	16.092	2	.774	7	20.795	.001
Meeste keskmine eluiga	49.717	2	2.952	7	16.840	.002
Surnud imikuid 1000 elava kohta	261.737	2	4.661	7	56.155	.000
Sünde 1000 elaniku kohta	3.008	2	1.440	7	2.088	.194
Surnuid 1000 elaniku kohta	.208	2	1.440	7	.145	.868
Sündivuse ja suremuse suhe	1.382E-02	2	6.630E-02	7	.208	.817
Keskmine laste arv	5.805E-02	2	2.951E-02	7	1.967	.210

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

Selle tabeli eesmärgiks on anda ülevaade sellest, millistel tunnustel on klasterdamise juures kõige suurem mõju ehk milliste tunnuste osas on tekkinud klastrid kõige erinevamad. Seda peegeldab statistiku F suurus.
PS! Antud olukorras ei tohi ANOVA tulemusi kasutada statistilise olulisustestina, sest võrreldavad grupid ehk klastrid on saadud põhimõttel, et erinevused oleksid maksimaalsed.
-
Näeme, et kõige enam erinevad tekkinud klastrid imikute suremuse osas kuid ka keskmise oodatava eluea osas. Kõige sarnasemad on klastrid suhtelise suremuse ning sündivuse ja suremuse suhte osas.

Cluster information for each case tabel, kus on iga objekti jaoks kirjas, millisesse klastrisse ta kuulub ning tema kaugus klastri keskpunktist

Cluster Membership

Case Number	COUNTRY	Cluster	Distance
1	Estonia	3	3.021
2	Finland	2	1.838
3	Georgia	1	4.307
4	Germany	2	1.964
5	Ireland	2	2.337
6	Latvia	1	3.187
7	Lithuania	3	1.087
8	Poland	3	3.084
9	Russia	1	3.970
10	UK	2	1.599

Näeme, et esimesse klastrisse kuuluvad Gruusia, Läti ja Venemaa; teise Soome, Saksa, Iirimaa ning Suurbritannia; kolmandasse Eesti, Leedu ja Poola.
Kolmandas klastris on kõige tüüpilisemaks e kõige klastri keskpunktile lähemaks riigiks Leedu, esimeses klastris aga kõige erinevamaks aga Gruusia.

- Kui sobiv klastrite arv on leitud, siis saab nupu *Save...* abil tellida andmetabelisse tunnuse(d), kus iga objekti puhul on väärtuseks klastri number, millesse ta kuulub.

Tunnuste grupeerimise ...

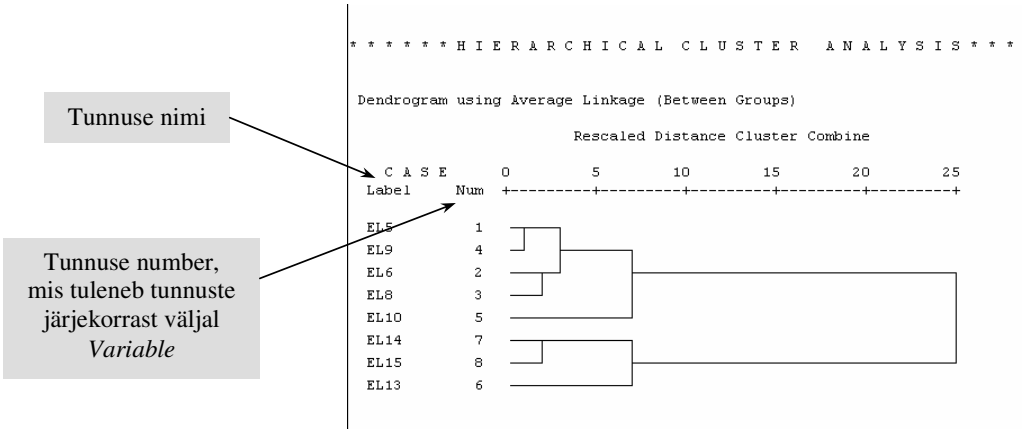
... korral on eesmärgiks leida sarnased, ehk omavahel kõige enam seotud tunnused ja moodustada selle põhjal ligilähedast fenomeni peegeldavate tunnuste grupid. Klasteranalüüsi kõrval saab siin kasutada ka faktoranalüüsi, mis on matemaatiliselt keerukam ja täpsem kuid seetõttu ka andmete suhtes suuremaid eeldusi tegev statistiline meetod.

Tunnuste grupeerimiseks tuleb kasutada hierarhilist klasteranalüüsi, millest oli juba juttu eespool seoses objektide grupeerimisega. Kasutatav algoritm on siin sama; erinevus seisneb vaid tunnuste jaoks sobiva sarnasus- ehk kaugusmõõdu valikus, milleks on Pearsoni korrelatsioonikordaja.

- Vali *Analyze/Classify/Hierarchical Cluster...*
- Paiguta grupeeritavad tunnused väljale *Variable(s):*
- Tunnuste klassifitseerimise korral märgista kastis *Cluster* väli *Variables*
- Klasterdamise matemaatilise meetodi valikuks vajuta nuppu *Method...*
  1. Vali kauguse mõõduks kastis *Measure Interval: Pearson correlation*
  2. Vali sobiv klasterdamise meetod väljal *Cluster Method:*

Between-groups linkage	klastrite vaheline sarnasus arvutatakse üksiktunnuste seosekordajate keskväärtusena
Nearest neighbor	klastrite vaheline sarnasus arvutatakse erinevates klastrites olevate omavahel kõige enam seotud tunnuste vahelise seosekordajana. See meetod annab tulemuseks nn **suurima korrelatsiooni tee**

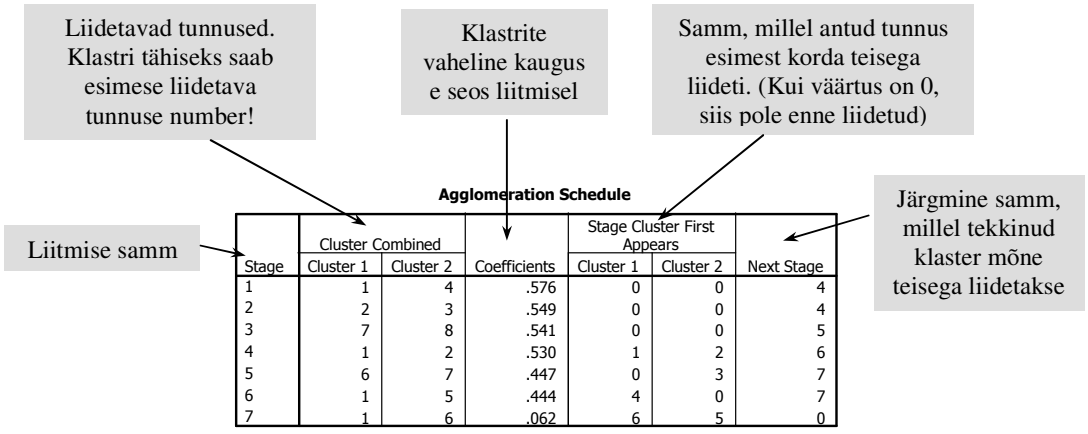
  3. Kui soovid, et sarnaseks peetaks ka tunnuseid, kus on (tugev) negatiivne seos, siis märgista kastis *Transform Measures:* väli *Absolute values*
- Liitmise protsessi ja tulemust kirjeldava diagrammi e **dendrogrammi** tellimiseks vajuta nuppu *Plots...* ning märgista väli *Dendrogram*; kastis *Icicle* märgista väli *None*.



Näeme, et esimesel sammul liideti esimene ja neljas tunnus ning et tekib kaks selgelt eristuvat klastrit.
NB! vaata tunnuste kirjeldusi tabelist Proximity Matrix järgmisel lehel!

- Põhiaknas olev nupp *Statistics...* võimaldab tellida:

  Agglomeration schedule liitmise protsessi kirjeldava tabeli



Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	4	.576	0	0	4
2	2	3	.549	0	0	4
3	7	8	.541	0	0	5
4	1	2	.530	1	2	6
5	6	7	.447	0	3	7
6	1	5	.444	4	0	7
7	1	6	.062	6	5	0

Näeme, et esimesel sammul liideti esimene ja neljas tunnus ning et viimasel sammul seos klastrite vahel sisuliselt puudub, mistõttu on mõistlik valida mudel, kus on kaks klastrit.

Proximity matrix üksiktunnuste vaheliste kaugusmõõtude maatriksi (korrelatsioonimaatriks antud juhul)

Proximity Matrix

Case	Matrix File Input							
	inimeste heaolu	demokraatia	inimõigused	sotsiaalsed garantiid	julgeolek	liikmesriikide vaheline ebavõrdsus	rahvusliku identiteedi kadumine	liigne sekkumine liikmesriikide sisepoliitikasse
1 inimeste heaolu		.569	.499	.576	.425	.060	.082	.082
2 demokraatia	.569		.549	.495	.429	.019	.116	.122
3 inimõigused	.499	.549		.558	.438	.066	.059	.039
4 sotsiaalsed garantiid	.576	.495	.558		.483	.031	.079	.039
5 julgeolek	.425	.429	.438	.483		.059	.041	.037
6 liikmesriikide vaheline ebavõrdsus	.060	.019	.066	.031	.059		.445	.449
7 rahvusliku identiteedi kadumine	.082	.116	.059	.079	.041	.445		.541
8 liigne sekkumine liikmesriikide sisepoliitikasse	.082	.122	.039	.039	.037	.449	.541	

Cluster Mebership tabeli, kus on iga tunnuse jaoks kirjas, millisesse klastrisse ta kuulub (klastrite arv tuleb ise määrata)

Cluster Membership

Case	2 Clusters
inimeste heaolu	1
demokraatia	1
inimõigused	1
sotsiaalsed garantiid	1
julgeolek	1
liikmesriikide vaheline ebavõrdsus	2
rahvusliku identiteedi kadumine	2
liigne sekkumine liikmesriikide sisepoliitikasse	2

N Toodud näites kasutatud tunnused olid kõik mõõdetud sama tüüpi skaalal: vastajal oli palutud viiepalli skaala hinnata kuivõrd tugevalt iseloomustavad toodud märksõnad tema arvates EL'i. Näeme, et esimene tekkinud klastritest koosneb positiivsetest omadustest, teine aga negatiivsetest omadustest. See, et need klastrid omavahel seotud pole, viitab sellele, et negatiivsete aspektide nägemine ei välista nõustumist positiivse poolega ja vastupidi.