

Tunnuste grupeerimine

Väärtuste summeerimine ja keskvaertuse arvutamine.
Väärtuste loendamine.
Klasteranalüüs.

Mitmemõtteline statistika Kairi Osula 2015/kevad

Tunnuste grupeerimine

- Oleme mõõtnud/uurinud teatavaid objektide omadusi ning soovime saadud infot kokku panna (**grupeerida**)
- El – oleme küsinud vastajate lemmikvärvi ja soovime kokku panna külmad ja soojad toonid
- El – meil on vastajate vanused ja soovime teha vanusegrupid
- JAH – võtame õpilaste erinevate ainete hinded ja soovime neid kirjeldada ühe, kokkuvõtva tunnuse, kaudu.
- JAH – teame treeningul osalejate kehakaalu ja pikkust, võime selle info põhjal arvutada KMI
- JAH – sotsiaalne staatus on mõõdetud erinevate tunnuste kaudu ning me soovime uurimiseks tekitada väiksemat arvu staatust kirjeldavaid tunnuseid
- JAH – väärtushoiakud, hinnangud jne.

Mitmemõtteline statistika Kairi Osula 2015/kevad

Tunnuste grupeerimine

Erinevad eesmärgid/olukorrad, miks tunnuseid koondada:

- Testide korral** (üksiküsimus, alamosa, kogu test):
 - õigete/valede vastuste arvu leidmine
 - punktide summeerimine
 - Ankeetide korral** (sisuliselt haakuvate küsimuste plokk):
 - teatud tüüpi vastuste arvu leidmine (nt "jah", vastamata,...)
 - teatud küsimusteploki põhjal koondtunnuse arvutamine (nt keskmine hinnang, summaarne kulu, ...)
 - latentsete e peidetud muutujate/tunnuste leidmine (nt hoiakud fenomeni eri aspektide suhtes, psühhosomaatilised seisundid, ...)
- Saan objektiivsema ja teistega võrreldava tulemuse, kui uurime nähtust läbi indikaatorite.

K.Niglas (loenguslaid)

Mitmemõtteline statistika Kairi Osula 2015/kevad

Loengu DVD 6

- Sissejuhatus
- Näide 1
 - eksam_inglise_keel.sav
 - ühe alamosa tulemuste koodamine e väärtuste summeerimine
 - väärtuste loendamine (kui paljud said 0 punkti)
- Näide 2
 - tagasiside.sav
 - koondhinnangu arvutamine (keskväärtus või summa)
- Näide 3
 - keeleshoiakud.sav
 - klasteranalüüs

Mitmemõtteline statistika Kairi Osula 2015/kevad

Väärtuste summeerimine

- Väärtuste kokkuliitmine funktsiooni SUM abil

	k1	k2	k3	summa_sum
1	2	1	2	5.00
2	2	1	.	3.00
3	2	9	.	2.00
4	0	1	.	1.00
5	0	9	.	.0

Numeric Expression:
sum(k1,k2,k3)

- Väärtuste kokkuliitmine kasutades liitmistehet

	k1	k2	k3	summa_p
1	2	1	2	5.00
2	2	1	.	.
3	2	9	.	.
4	0	1	.	.
5	0	9	.	.

Numeric Expression:
k1+k2+k3

Mitmemõtteline statistika Kairi Osula 2015/kevad

Väärtuste kokkuloendamine

- Transform/Count Values Within Cases ...

System missing
System or user missing

- Replace Missing Values...

k1	k2	k3	count_systemmissing	count_system_and_user_missing
2	1	2	0	0
2	1	.	1	1
2	9	.	1	2
0	1	.	1	1

Mitmemõtteline statistika Kairi Osula 2015/kevad

Näide 1



matemaatika2014.sav

- Mitu punkti kogusid õpilased 1.osa eest?
- Kirjeldage saadud koondtulemusi.
- Milliseid olulisusteste saaks rakendada?
- Mitu 0 punkti ül. oli igal õpilasel?
- Millele tuleb tähelepanu pöörata 2.osa ül. juures?

	Name	Type	Width	Decimals	Label
1	kood	Numeric	8	0	Õpilase kood
2	kool	String	40	0	Vastavastava Hariduskollegium
3	klass	Numeric	11	0	Klass
4	v1	Numeric	11	0	Õppekeele kood
5	v1_1	Numeric	11	0	1.osa (max=2p.)
6	v1_2	Numeric	11	0	1.osa (max=2p.)
7	v1_3	Numeric	11	0	1.osa (max=2p.)
8	v1_4	Numeric	11	0	1.osa (max=2p.)
9	v1_5	Numeric	11	0	1.osa (max=2p.)
10	v1_6	Numeric	11	0	1.osa (max=2p.)
11	v1_7	Numeric	11	0	1.osa (max=2p.)
12	v1_8	Numeric	11	0	1.osa (max=2p.)
13	v1_9	Numeric	11	0	1.osa (max=2p.)
14	v1_10	Numeric	11	0	1.osa (max=2p.)
15	v2_1	Numeric	11	0	2.osa (max=7p.)
16	v2_2	Numeric	11	0	2.osa (max=7p.)
17	v2_3	Numeric	11	0	2.osa (max=7p.)
18	v2_4	Numeric	11	0	2.osa (max=7p.)

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Näide 2



tagasiside.sav

- Ettevalmistus analüüsiks e andmetöötlus**
 - Andke ülevaade puuduvatest väärtustest.
 - Mida tuleks ette võtta nende vastajatega, kes ei andnud ühtegi hinnangut? Mida nendega, kellel on mõned hinnangud puudu? Kas asendada puuduvad väärtused?
 - Pöörake kõik skaalad samasuunalisteks.
 - Kas koondhinnangute saamiseks tuleks arvutada hinnangte summa või keskvärtus (kuidas tulemused erineksid?) Arvutage koondhinnang.
- Andmete analüüs**
 - Kirjeldage koondhinnangute tulemusi.
 - Testige koondhinnangute tulemusi vabalt valitud gruppide lõikes.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Näide 3



keeelhoiakud.sav

- 1 Kas Te loete ingliskeelseid filme vaadates alt eestikeelseid subtiitreid?
- 2 Kas Te jälgite ingliskeelset popmuusikat kuulates laulude sõnu?
- 3 Kas Te räägite oma sõpradega omavahel inglise keelt?
- 4 Kas Teil on tundeid (üllatust, armastust, viha jm) kergem vältendada inglise keeles?
- 5 Kas te kasutate inglise keelt internetis (jututuaas, e-mailides)?
- 6 Kas te loete meelelahuseks ingliskeelset kirjandust (raamatuid, ajakirju, koomikseid)?
- 7 Kas Te mõtlete inglise keeles?
- 8 Kas Te suhtlete kohalike muulastega eesti keeles?

Tunnused ei mõõda igaüks eraldi fenomeni vaid ühe fenomeni erinevaid aspekte

- Tunnuste grupeerimine**

JAH – millised tunnused võiksid kirjeldada ühe fenomeni eri aspekte?

Millised tunnused võiksid koos töötada?

EI – millised on vastajate tüüpilised grupid?

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Tunnuste grupeerimine



- Kui panen kaks tunnust kokku – need tunnused peegeldavad ühte ja sama fenomeni e neil on ühine kirjeldusosa
 - Kas kirjeldusosa on (võiks olla) 100%?
 - Igal tunnusel on
 - väike eripära, mis kirjeldab selle küsimuse olemust
 - ühine osa, mis peegeldab peidetud laiemat fenomeni.
 - Tunnuste kokkupanemisel:
 - usaldusväärsus kasvab
 - variatiivsus kasvab e tendents peaks selgemalt ilmema
 - Variatiivsus kaob – kui oleme kokku pannud juhuslikud tunnused, mis ei ole omavahel seotud
- Eeldame, et seos on 0,6-0,7 (PS),
sotsioloogias pannakse kokku ka nõrgemalt seotud tunnuseid.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Klasteranalüüs



- Avastab andmetes olevaid struktuure**, ilma neile selgitust või interpretatsiooni andmata
- Tõlgenduse lisab uurija/kasutaja
- Jagada gruppidesse
 - Tunnuseid
 - Objekte
- Gruppi kuuluvad objektid/tunnused oleksid omavahel **maksimaalselt** tugevalt seotud, grupid omavahel aga erineksid omavahel **maksimaalselt**

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Klasteranalüüs



- Tunnused – intervall tüüpi või binaarsed (Pearsoni kordaja)
- Hierarhilise klasteranalüüsi idee** seisneb selles, et samm-sammult hakatakse kokku panema neid tunnuseid, mis on omavahel kõige sarnasemad st millel on omavahel kõige suurem seos.
- Kauguse/sarnasuse mõõduks *Measure/Interval*: Pearson correlation

K.Niglas (loenguslaid)

Alguses on iga tunnus/objekt omaette grupp/klaster

Lõpuks on kõik tunnused/objektid koondatud üheks klasteriks

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Näide 3 (jätkub)

- **Tunnused:** k9-k40
- **Enne analüüsi**
 - Puuduvate väärtuste uuring
 - Korrelatsioonimaatriks
- **Problem**
 - Andmestik koosneb andmegruppidest (eesti, vene), mis ei sobi sisuliselt kokku. Tuleks valida analüüsi jaoks vaid eesti keelt kõnelevad vastajad.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

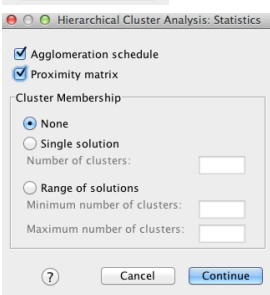
Klasteranalüüs SPSS-s

- Analyze/Classify/Hierarchical Cluster...
- **Variable(s):** uuritavad tunnused
- **Cluster:** ☒ Variables (grupeerime tunnuseid)
- **Display:** ☒ Statistics ☒ Plots



Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Statistics...



☒ Agglomeration schedule
☒ Proximity matrix

Cluster Membership
☒ None
☐ Single solution
 Number of clusters:
☐ Range of solutions
 Minimum number of clusters:
 Maximum number of clusters:

Cancel Continue

☒ Agglomeration schedule

Tunnuste kokkupanemise ülevaade sammude kaupa

☒ Proximity matrix

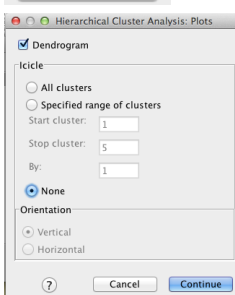
Sarnasus/kaugusmõõdude maatriks

Proximity matrix = korrelatsioonimaatriks, kui puuduvaid väärtuseid ei ole.

Kui on puuduvad väärtused, siis klasteranalüüsist jäetakse need objektid väljale, on kasvõi üks puuduv väärtus, kuid seosemaatriksis on need sees, seal arvestatakse väljajätmist paari-kaupa.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Plots...



☒ Dendrogram

IceTicle
☐ All clusters
☐ Specified range of clusters
 Start cluster:
 Stop cluster:
 By:
☒ None

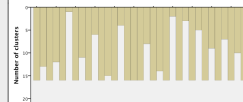
Orientation
☒ Vertical
☐ Horizontal

Cancel Continue

☒ Dendrogram

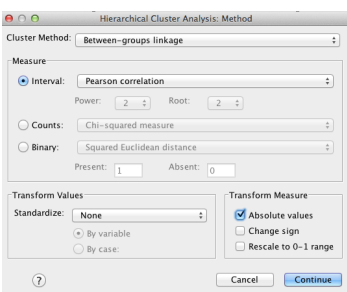
Tunnuste kokkupanemise graafiline ülevaade sammude kaupa

IceTicle



Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Method...



Cluster Method:

Measure
☒ Interval:
 Power: Root:
☐ Counts:
☐ Binary:
 Present: Absent:

Transform Values
 Standardize:
☐ By variable
☐ By case

Transform Measure
☒ Absolute values
☐ Change sign
☐ Rescale to 0-1 range

Cancel Continue

Cluster Method
Between-groups linkage
 klasterite vahelise sarnasus arvutatakse üksiktunnuste seosekordajate keskvaartustena

☒ Interval**Pearson correlation**

Tunnused on moodetud intervalli-skaalal ning kokkupanemise aluseks on Pearsoni korrelatsioonikordajate väärtused

☒ Absolute values

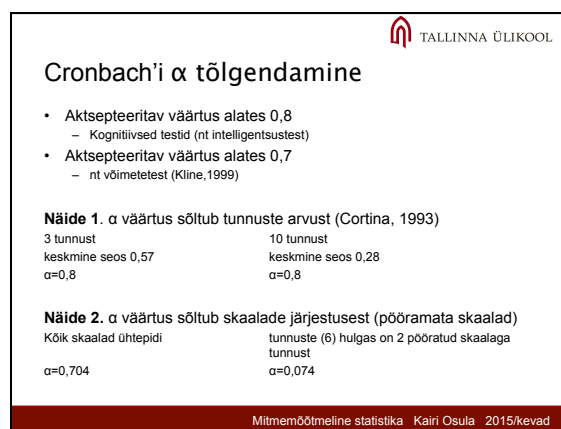
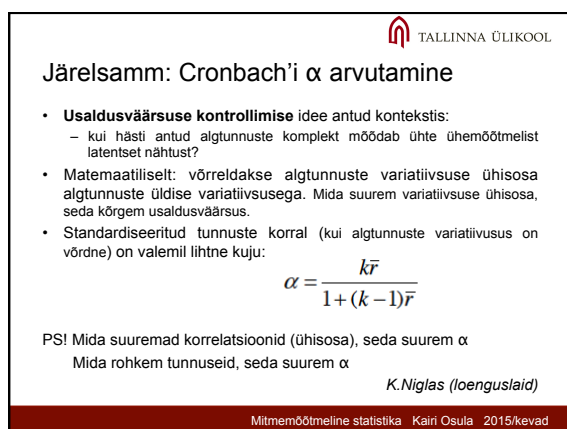
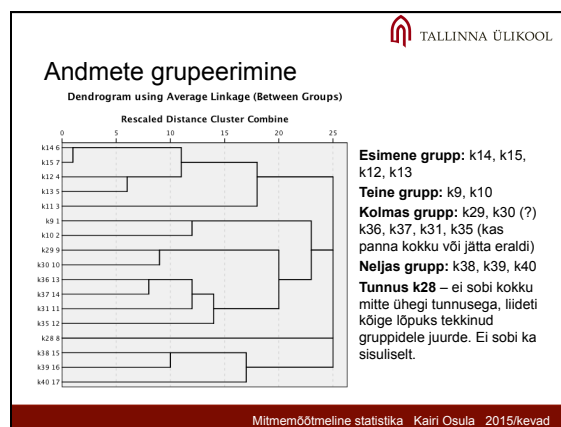
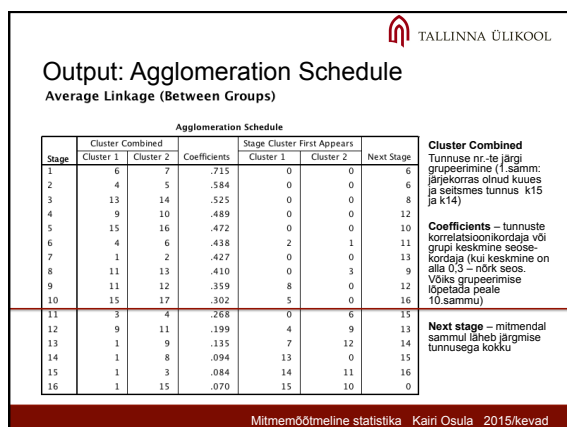
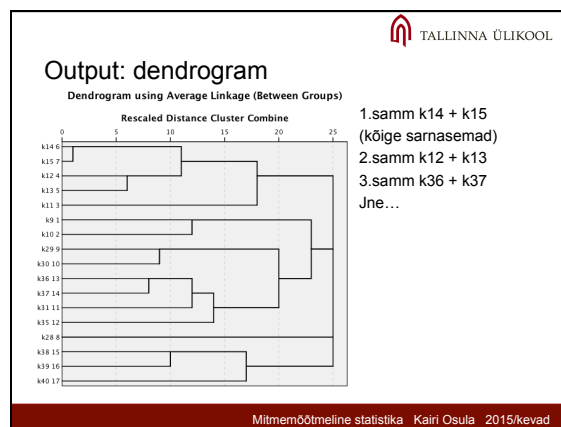
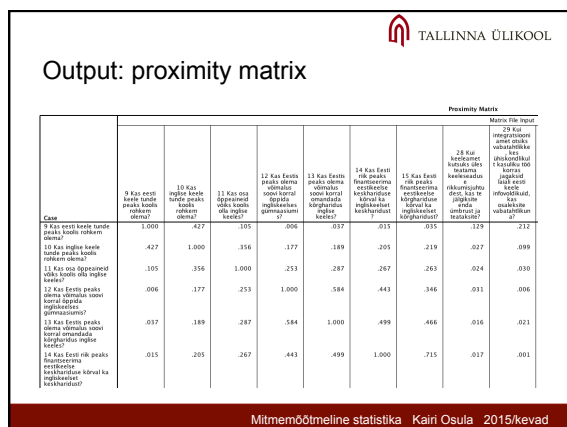
Ei erista gruppidesse koondamisel negatiivseid ja positiivseid seoseid

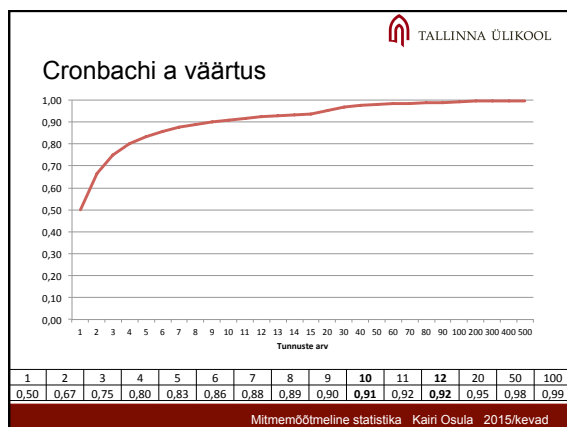
Mitmemõõtmeline statistika Kairi Osula 2015/kevad

Klasterite ühendamise meetodid

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

Mitmemõõtmeline statistika Kairi Osula 2015/kevad





TALLINNA ÜLICOOL

Cronbach'i α arvutamine SPSS-s

- Analyze/Scale/Reliability Analysis...

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLICOOL

Statistics...

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLICOOL

Output

Reliability Statistics

Cronbach's Alpha	N of Items
.776	5

N of items – 5 tunnust
Cronbach'i α väärtus on 0,776
 Kui α on üle 0,7 loetakse väärtust heaks. Parems oleks üle 0,8.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLICOOL

Output

Item-Total Statistics

Item	Total	Cronbach's Alpha if Item Deleted
11 Kas osa õppeaineid võiks koolis olla inglise keeles?	10.809	.809
12 Kas Eestis peaks olema võimalus soovi korral õppida ingliskeelses gümnaasiumis?	9.739	.739
13 Kas Eestis peaks olema võimalus soovi korral omandada kõrgharidus inglise keeles?	9.712	.712
14 Kas Eesti riik peaks finantseerima eestikeelse keskhariduse kõrval ka ingliskeelset keskharidust?	9.694	.694
15 Kas Eesti riik peaks finantseerima eestikeelse kõrghariduse kõrval ka ingliskeelset kõrgharidust?	9.712	.712

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLICOOL

Klasteranalüüs. DVD 8

Objektide grupeerimine

- Näide 1
 - Teooria (hierarhiline klasteranalüüs)
 - klaster_maad-väike.sav
- Näide 2
 - K-keskmiste klasterdamine
- Näide 3
 - Faktor- ja klasteranalüüs koos.

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

Vastajate/koolide tüüpilised grupid

- Teisena lõime koolide tüpoloogiat. Võtsime aluseks koolikliimat iseloomustavad näitajad kooliti Eristus (taas aluseks klasteranalüüs) kolm koolitüüpi:
 - edukaid-sooje-elujõulisi kooli** oli 65 seas 15 (23%). Neid iseloomustas teistest parem õppeedukus, õpilaste parem psühholoogiline ja kehaline heaolu ning õpilaste konstruktiivsed toimetulekustrategiad, ...;
 - keskmiselt edukaid-külmi-elujääjaid kooli** oli 65-st 34 (52%). Õppeedukuse poolest hõivasid need keskmise koha, kuid teistega võrreldes puudusid nende koolides õpilased põhjuseta kõige sagedamini ja jätsid oma kodutööd tegemata. ...;
 - väheedukaid-sooje-heitunud kooli** oli 16 (ligikaudu 25%). Neis koolides oli kõige madalam õppeedukus; õpilastele oli iseloomulik kõige väiksem tulevikuusk, ja seda just töö leidmise suhtes pärast kooli; ...

<http://www.rigikogu.ee/rito/index.php?id=13560&op=archive2>

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

Objektide grupeerimine

- Eesmärk (ehk miks objekte koondada):
 - Omavahel **sarnaste objektide leidmine** (sarnased = valitud tunnuste lõikes liigilähedasel samu vastuseid/väärtusi omavad vastajad/objektid)
 - Tüpoloogiate loomine**
- Võimalikud meetodid:
 - Vähe objekte:
 - Hierarhiline klasteranalüüs
 - k-keskmiste klasteranalüüs
 - Palju objekte:
 - k-keskmiste klasteranalüüs
 - Kategoriaalsed tunnused:
 - Tunnuste poolt defineeritud alamgruppidesse jagamine (abiks mitmemõõtmelised sagedustabelid)

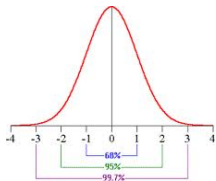
K.Niglas (loenguslaid)

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

Väärtuste standardiseerimine

- Vajalik, sest soovime ühikust lahti saada.
- Vaadati iga objekti jaoks erinevus keskmisest ja jagati see kaugus standardhälvega
- Jaotuse keskmine 0 ja st.hälve 1



Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

Klastrite ühendamise meetod:

- Between-groups linkage**
 - klastrite vaheline kaugus arvutatakse üksiktunnuste kauguste/sarnasuste keskvaertusena
- Nearest neighbor**
 - klastrite vaheline kaugus arvutatakse erinevates klastrites olevate omavahel kõige sarnasemate objektide vahelise kaugusena
- Centroid clustering**
 - klastrite vaheline kaugus arvutatakse klatri raskuskeskmete vahel
- Ward's method**
 - teistest erinev loogika (ANOVA'le sarnane) minimeeritakse klustrite sisest hajuvust $E = \text{summaarne objektide kaugus klatri keskpunktist}$ alul kõik eraldi ja $E=0$ igal sammul pannakse kokku need objektid, mille puhul E kasvab kõige vähem

K.Niglas (loenguslaid)

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

K-keskmiste klasterdamine

- K-keskmiste klasterdamise algoritm on samuti lihtne:
 - määrata klustrite arv
 - jagada objektid esialgsesse klustritesse
 - arvutada välja klustrite keskpunktid
 - võrrelda iga objekti kõigi klustrite keskpunktidega:
 - kui osutub, et objekti kaugus mõne muu klatri keskpunktist on väiksem kui selle klatri keskpunktist, milles ta parasjagu asub, siis tuleb objekt teise klustrisse ümber tõsta
 - peale objekti ümbertõstmist tuleb klustrite keskpunktid uuesti arvutada ja jätkata protsessi niikaua kui kõik objektid on klastri, mille keskpunktile nad kõige lähemal asuvad.

K.Niglas (loenguslaid)

Mitmemõõtmeline statistika Kairi Osula 2015/kevad

TALLINNA ÜLIKOOL

Klasteranalüüs

- Pole ühte ainuõiget lahendit => tõlgendus + stat.näitajad
- Kui andmetes ei ole sisulist grupeeritust, siis võib jagamine olla "meelevaldne" ja sõltuda (suurel määral) klasterdamise meetodist
- Tulemus võib sõltuda objektide sorteerimise järjekorrast
- Ekstreemsed väärtused => ühe-kahe objektiga klasterid
- Erinevad skaalad => standardiseerida
- Eksploratiivne meetod => F-testi e ANOVA't ei tohi klasteranalüüsi juures tõlgendada kui olulisustesti
- Kaalu, kas klasterdamise aluseks olevad tunnused on valitud "mõistlikult" (sobivad sisuliselt kokku, ükski tähtis tunnus ei puudu mudelist, jne)

K.Niglas (loenguslaid)

Mitmemõõtmeline statistika Kairi Osula 2015/kevad