Sotsiaalse analüüsi meetodid:
kvantitatiivne lähenemine

Klasteranalüüs

Indrek Soidla

Andmeanalüüsi meetodid

Otsime seaduspärasid tunnuste vahel

- Analüüs: tihti suunatud *tunnustevaheliste seoste* otsimisele
- Nt soovime teada, mis on seotud kliendi otsusega mingit toodet osta
- Samas oleks vajalik eelnevalt teada, millised segmendid meie klientuuris esinevad
- Koondame omavahel sarnased indiviidid rühmadesse, mis üksteisest erinevad
- Seejärel saab juba nt uurida, kuidas grupikuuluvus on seotud ostukäitumisega

Otsime mustreid indiviidide seas

Klasterdamine ja liigitamine

- Liigitamine *(classification)*
  - rühmakuuluvuse kriteeriumid on teada, indiviidid liigitatakse nende alusel
  - Juhendatud õpe *(supervised learning)*
  - Nt spämmifiltrid (eeldusel, et teada, millised meilid on spämm)
- Klasterdamine, klasteranalüüs *(clustering)*
  - teatud tunnuste alusel sarnastest indiviididest moodustatakse rühmad, kriteeriumid (tunnuste väärtuste kombinatsioonid) pole eelnevalt teada
  - Juhendamata õpe *(unsupervised learning)*
  - Nt liikluskindlustuses kõrge riskikoefitsiendiga klientide hulgas eristuvate gruppide leidmiseks

Klasteranalüüs

- Põhimõtteliselt indiviidide rühmitamine valitud tunnuste *(variables, features)* alusel
- Põhilised eesmärgid:
  - Andmetes struktuuri/mustrite leidmine
    - nt avastuslikus analüüsis uute uurimisküsimuste / hüpoteeside formuleerimiseks
  - Homogeensete gruppide moodustamine
    - nt andmete komplekssuse vähendamiseks
    - nt teatud tunnuste põhjal moodustunud klastrite uurimiseks teiste tunnuste lõikes
- Suhteliselt nõrk statistilises tõestuses
- Seevastu võimaldab andmeid paremini mõista ja tõlgendada
- Tulemused võimaldavad täpsemini uurida tunnustevahelisi seoseid

Klasteranalüüsi käik

1. Indiviidide ehk objektide (valimi) valik
2. Tunnuste valik
3. Tunnuste standardiseerimine
4. Kauguse/läheduse kriteeriumi valik
5. Klasterdusmeetodi valik
6. Klastrite arvu valik
7. Tõlgendamine, testimine, replikeerimine, valideerimine

Milligan (1996)

(1) Indiviidide ehk objektide valik

- Oluline esinduslikkus
  - Populatsiooni suhtes
  - Klasterstruktuuri suhtes, mille olemasolu eeldame
- Praktikas valik tihti ette antud kasutada olevate andmetega
- Valiku teema relevantsem juhul, kui
  - võimalik kasutada mitmeid erinevaid andmestikke
  - kasutada pole populatsiooni ega esindusliku valimi andmeid
  - uurime mingit konkreetset gruppi, mitte kogu valimit

(2) Tunnuste valik

- Tunnused peaksid sisaldama vajalikku informatsiooni indiviidide klastritesse jagamiseks
- Kõrgelt korreleeritud tunnused
  - ei pruugi lisada olulist informatsiooni
  - võib lisada ebavajalikku komplekssust
- Teisalt, kaasatud tunnused annavad klastritele tähenduse
  - kui tunnus teoreetilises raamistikus oluline, lisab klasterdusse olulist infot
- Võimalik tunnuste valikut automatiseerida, aga
- Oluline silmas pidada klasterdamise eesmärki ja teoreetilist/sisulist tausta

(3) Tunnuste standardiseerimine

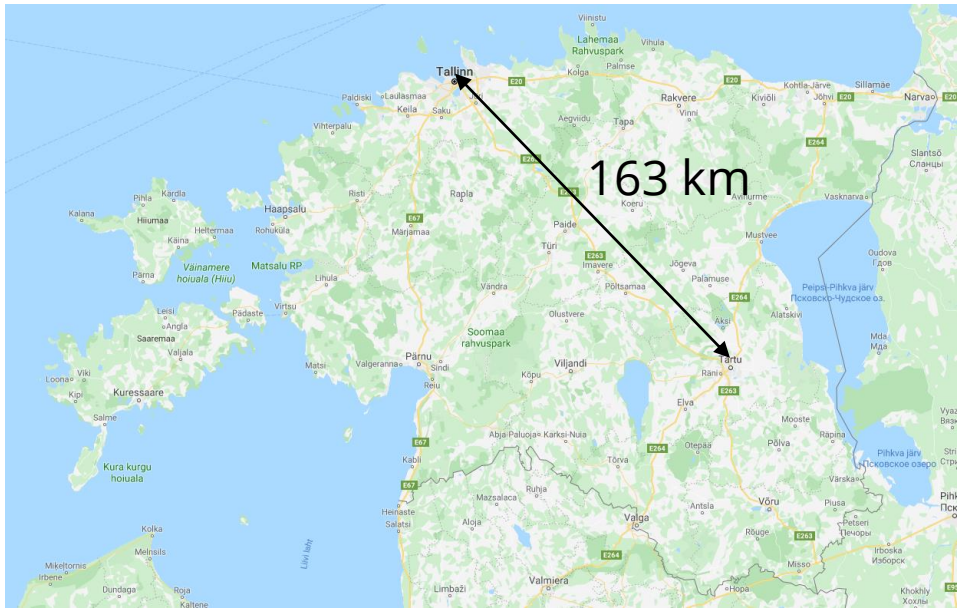
- Vaatame enne kauguse kriteeriumi valikut, et paremini mõista
  - klasteranalüüsi sisu ja
  - tunnuste standardiseerimise mõtet

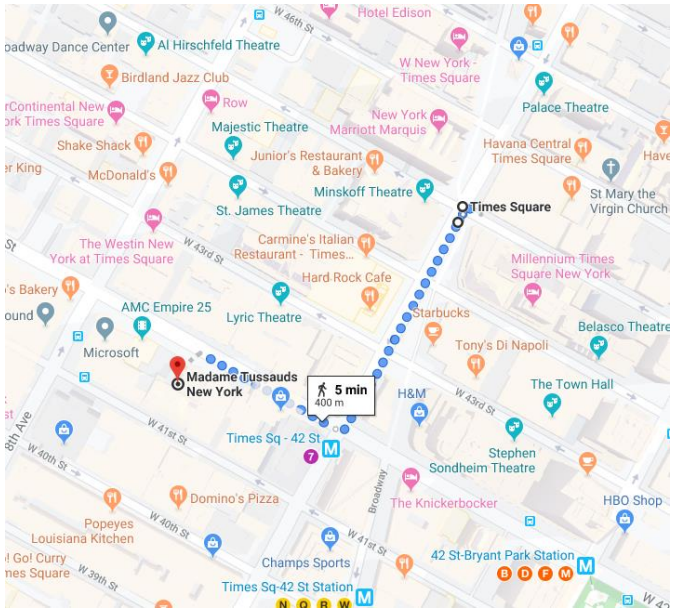
(4) Kauguse / läheduse kriteeriumi valik

- Kaugus / lähedus: *distance / proximity*
- Vrd erinevus / sarnasus: *dissimilarity / similarity*
- Mitmeid erinevaid kriteeriume ehk indiviidide vaheliste kauguste arvutamise meetodeid
- Pole ühte „õiget“ kauguse arvutamise meetodit
- Sõltub,
  - kuidas sarnasust / erinevust uurijana defineerime
  - milline on klasterdamise eesmärk

Kuidas leiame sarnased indiviidid?

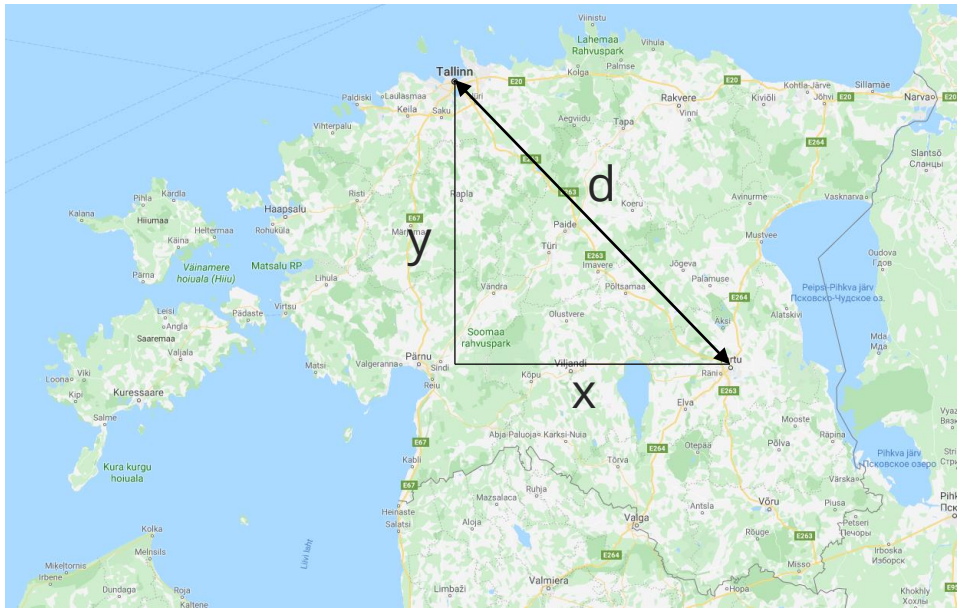
- Tunnustes esinevate väärtuste alusel
- Kuidas leida mitme tunnuse alusel kokkuvõtlik sarnasuse/läheduse näitaja?
- Võrdlus geograafiast:





Kuidas leiame sarnased indiviidid?

- Geograafiline kaugus



Eukleidiline kaugus (kaugus linnulennul): $d = \sqrt{x^2 + y^2}$

Linnakaugus *(Manhattan distance)*: $d = x + y$

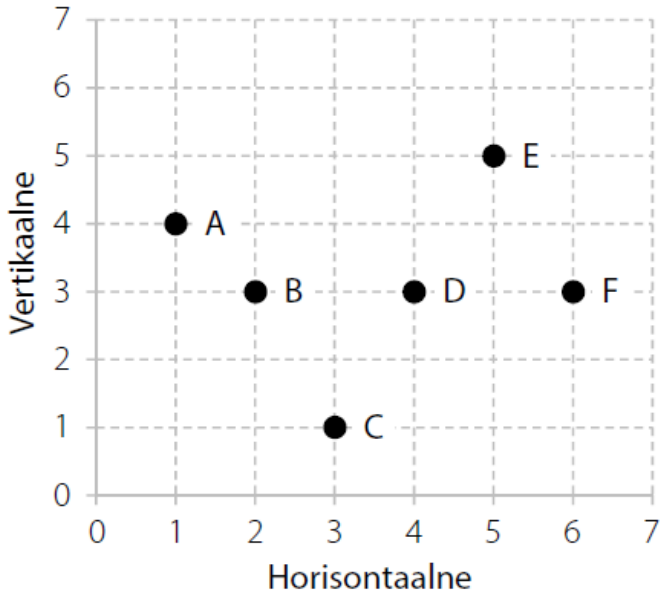
- Sarnaselt võimalik leida indiviidide kaugus, lähtudes nende väärtustest tunnustes

Kuidas leiame sarnased indiviidid?

- Indiviidide o_1 ja o_2 vaheline eukleidiline kaugus kahe tunnuse lõikes

$$d(o_1, o_2) = \sqrt{(x_1^1 - x_2^1)^2 - (x_1^2 - x_2^2)^2}$$

$$d(A, B) = \sqrt{(2-1)^2 + (4-3)^2} = \sqrt{2} = 1{,}41$$



- Eelnev näide on kauguse arvutamine kahe tunnuse põhjal
- Rohkemate tunnuste korral arvutatakse ruutkaugused kõigi tunnuste lõikes
- Ruutkaugused summeeritakse, summast võetakse ruutjuur
- Indiviididevaheline eukleidiline kaugus *M* arvu tunnuste lõikes ($M = i_{max}$):

$$d(o_1, o_2) = \sqrt{\sum_{i=1}^{M} (x_1^i - x_2^i)^2}$$

- Linnakaugus:

$$d(o_1, o_2) = \sum_{i=1}^{M} \left| x_1^i - x_2^i \right|$$

Tooding (2015)

Kuidas leiame sarnased indiviidid: dihhotoomsed tunnused

- Kahe indiviidi vaheline kaugus = kokkulangevate väärtuste osakaal tunnustes

Teine indiviid / Esimene indiviid	Jah	Ei	Kokku
Jah	a	b	$a + b$
Ei	c	d	$c + d$
Kokku	$a + c$	$b + d$	$a + b + c + d$

- Eukleidiline kaugus: $d(o_1, o_2) = \sqrt{b + c}$
  - Kui kõik väärtused ühtivad, siis $d(o_1, o_2) = 0$
  - Ei olene sellest, kui palju on ühtivusi puudumise alusel

Tooding (2015)

Kuidas leiame sarnased indiviidid: dihhotoomsed tunnused

Esimene indiviid \ Teine indiviid	Jah	Ei	Kokku
Jah	a	b	$a+b$
Ei	c	d	$c+d$
Kokku	$a+c$	$b+d$	$a+b+c+d$

- Ühtivuskaugus: $d(o_1, o_2) = \frac{b+c}{a+b+c+d} = 1 - \frac{a+d}{a+b+c+d}$ Läheduskordaja: $s(o_1, o_2) = \frac{a+d}{a+b+c+d}$
  - *d* ja *s* varieeruvad nullist üheni
  - Olenevad mh sellest, kui palju on ühtivusi puudumise alusel
- Jaccardi kaugus: $d(o_1, o_2) = \frac{b+c}{a+b+c}$ Lance'i-Williamsi kaugus: $d(o_1, o_2) = \frac{b+c}{2a+b+c}$
  - *d* varieerub nullist üheni
  - Ühtivustest võetakse arvesse ainult ühtivusi olemasolu (kokkulangevad jah-väärtused) alusel
- Kas võtta arvesse ühtivusi ainult olemasolu alusel?
  - Oleneb kategooriate tähendusest ja uurimiseesmärgist
  - Välja töötatud kümneid kauguse arvutamise variatsioone eri juhtumite jaoks

Tooding (2015)

Kuidas leiame sarnased indiviidid: järjestustunnused

- Skaala pole arvuline, ainult järjestatav
- Tunnuses arvulised koodid, aga tõlgenduslik skaalapunktide kaugus ei pruugi vastata arvulistele väärtustele =>
- Indiviididevaheliste kauguste arvutamine keeruline
- Pmst saaks teisendada dihhotoomseteks, aga osa infot läheks kaotsi
- Vahel kasutatakse arvulisi koode, aga vt pt 2
  - Põhineb eeldusel, et skaala arvuliste väärtuste vahed = skaalapunktide tõlgenduslikud kaugused
- Võimalused:
  - Omistada tunnuse väärtusteks astakud vm astakutel põhinevad ümberarvutused
  - Kasutada järjestustunnuseid võimaldavaid analüüsimeetodeid (nt *latent class clustering*)
- Lihtsuse ja ajalise piiratuse tõttu antud kursuses klasteranalüüs arvuliste tunnustega

Hennig (2015)

(3) Tunnuste standardiseerimine

- Kuidas mõjutab klasteranalüüsi see, kui tunnused on mõõdetud erinevatel skaaladel?
- Indiviidide vahelised erinevused pikema skaalaga tunnuse lõikes määravamad
- Võimalik standardiseerida
  - [0, 1]-skaalale
    - Pole hea, kui tunnuses on erindid (suruvad ülejäänud väärtused väikesele alale)
- Kuidas mõjutab klasteranalüüsi see, kui (sama skaalaga) tunnuste hajuvus on erinev?
- Suurem hajuvusega (variatiivsusega) tunnus klasterduses määravam
- Võimalik standardiseerida, jagades tunnuse väärtused läbi
  - standardhälbega või arvutada z-skoorid
  - kvartiilhaarde (IQR) või absoluutse mediaanhälbega (MAD)

Hennig (2015)

(3) Tunnuste standardiseerimine

- Oluline võib olla ka tunnuste teisendamine
- Sisulised kaalutlused määravad
- Näide säästude mõõtmisel:
  - Kas erinevus säästude vahel summas 300 ja 3000 on sisulises plaanis 10x vähem oluline, kui erinevus 30 000 ja 300 000 vahel?
  - Pigem sama oluline?
  - Siis võiks kasutada logaritm-teisendust

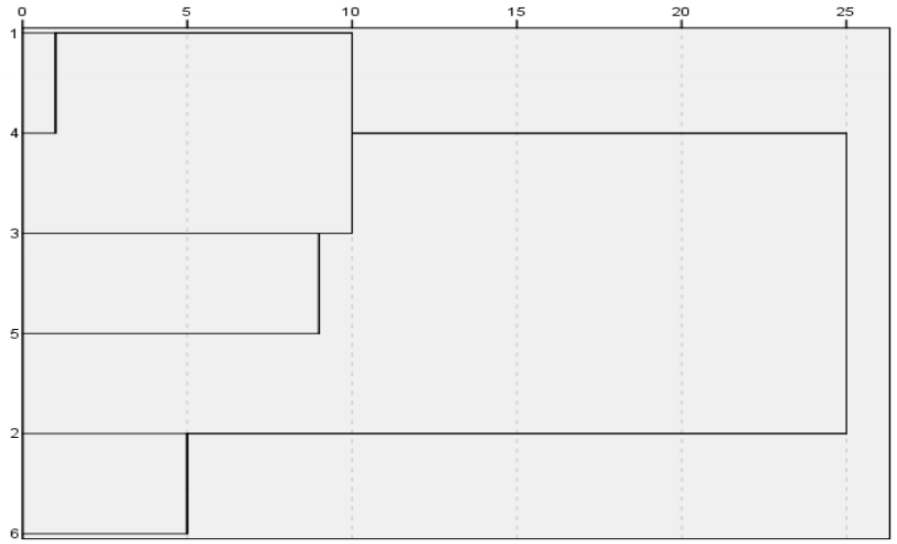
Hennig (2015)

(5) Klasterdusmeetodi valik

- Klasteranalüüsis keskne, aga kauguse arvutamise meetodi valik võib olla isegi olulisem!
- Hierarhilised klasterdusmeetodid
  - Muudatused samm-sammult, igal sammul üks muudatus klasterstruktuuris
  - Võimalik jälgida klastrite samm-sammulist moodustumist
  - Tulemusi võimalik esitada klasterduspuuna
  - Klastrite arvu ei pea ette andma, sobiv arv võib nähtuda klasterduspuult
  - Sobivad paremini väiksema kogumi klasterdamiseks

(5) Klasterdusmeetodi valik

- Hierarhilised klasterdusmeetodid
  - Näide: ETF grantide tulemuslikkuse analüüs (Ainsaar, Soidla, Roots 2019)
  - Humanitaarteadlaste hulgas on rahulolematus seniste teaduse taseme mõõtmisega
  - Peavad bibliomeetrial põhinevaid mõõdikuid endi suhtes ebaõiglasteks
  - Kuivõrd humanitaarteadlased erinevad teistest selle poolest, milliseid teaduse hindamise mõõdikuid peetakse headeks?



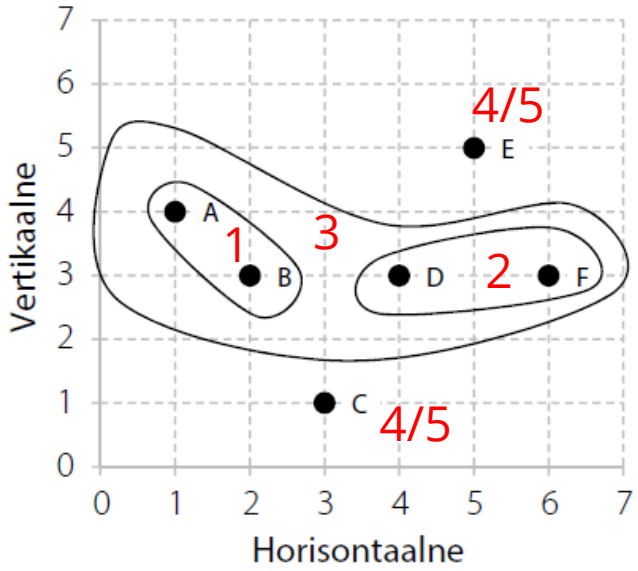
Joonis 13. Eri teadusvaldkondades töötavate inimeste lähedus teineteisele indikaatorite eelistusel teadusprojektide mõjude hindamisel (1= arstit., 4 = loodust.; 3 = inseneri- ja tehnikat; 5 = põllumajandust., 2 = humanitaart.; 6 = sotsiaalt.)

(5) Klasterdusmeetodi valik

- Hierarhilised klasterdusmeetodid
  - Liigendavad meetodid *(divisive clustering)*
    - Alustatakse ühest tervikust, samm-sammult eraldatakse kaugemaid indiviide
  - Ühendavad meetodid *(agglomerative clustering)*
    - Kõik indiviidid alguses eraldi, samm-sammult ühendatakse lähemad
- Millest lähtutakse indiviidide eraldamisel / ühendamisel?
  - Loomulikult kaugustest
  - Oluline siiski ka eraldamise / ühendamise järjekord ja reeglid =>
  - Palju erinevaid klasterdusmeetodeid

Ühe seose ehk lähima naabri meetod

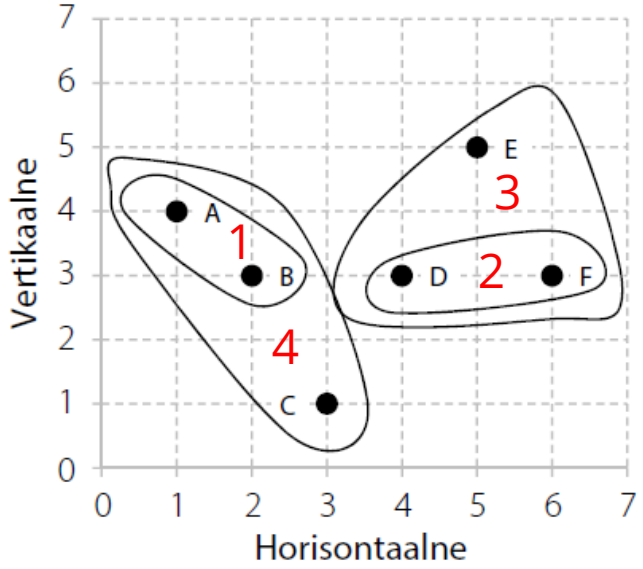
- *Single linkage / nearest neighbour clustering*
- Indiviidide rühmitamisel lähtutakse kahe klastri *lähimate* elementide vahelisest kaugusest
- Ühendatakse indiviidid, mis on teineteisele kõige lähemal
- Klastrite puhul ühendatakse need, mille lähimad indiviidid on kõige lähemal
- Ahelaefekt



Tooding (2015)

Täieliku seose ehk kaugeima naabri meetod

- *Complete linkage / farthest neighbour clustering*
- Indiviidide rühmitamisel lähtutakse kahe klastri *kaugeimate* elementide vahelisest kaugusest
- Ühendatakse indiviidid, mis on teineteisele kõige lähemal
- Klastrite puhul ühendatakse need, mille kaugeimad indiviidid on teineteisele kõige lähemal
- Keskendub suurima klastrisisese kauguse võimalikult väikesena hoidmisele
- Väldib ahelaefekti
- Moodustuvad klastrid enam-vähem võrdse diameetriga
- Võib olla erindite suhtes tundlik



Kabacoff (2015), Tooding (2015)

Klastritevahelise keskmise kauguse meetod

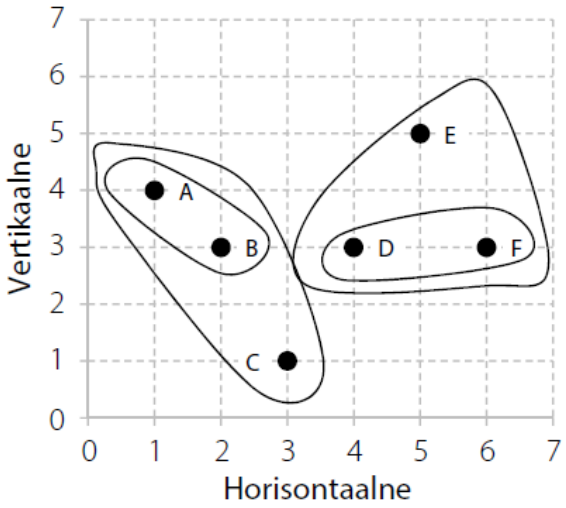
- Hennig (2015): nii ühe kui täieliku seose algoritmid võivad anda liiga äärmusliku tulemuse
- *Average linkage*
- Kahe klastri vaheline kaugus: keskmine klastrite kõikvõimalike indiviidipaaride vahelistest kaugustest
- Vähem ahelaefekti ja samal ajal vähem tundlik erindite suhtes

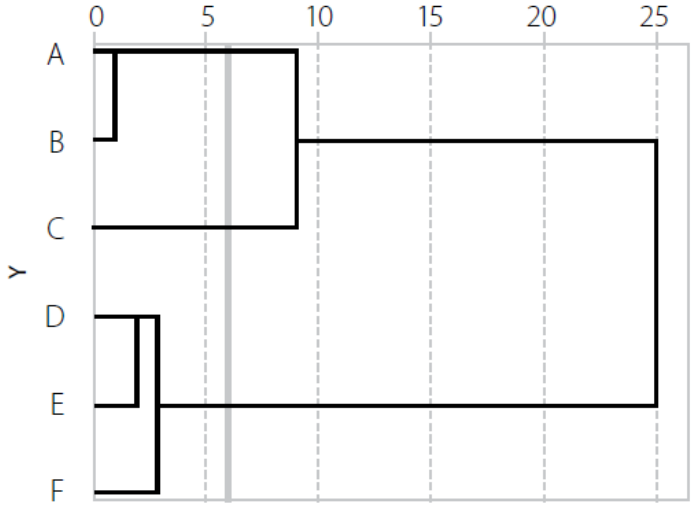
Wardi meetod

- Igal sammul teostatakse selline klastrite/indiviidide ühendamine, mille tagajärjel klasterstruktuuri summaarne ruuthälve iga klastri keskpunktist on väikseim

$$\sum_{i,j,k}(x_{ijk} - m_{ik})^2 \rightarrow \min$$

- x_{ijk} – k-nda tunnuse väärtus i-nda klastri j-ndal indiviidil
- m_{ik} – k-nda tunnuse keskmine i-ndas klastris
- Moodustuvad klastrid enam-vähem võrdse indiviidide arvuga





Tooding (2015)

k-keskmiste / k-mediaanide meetod

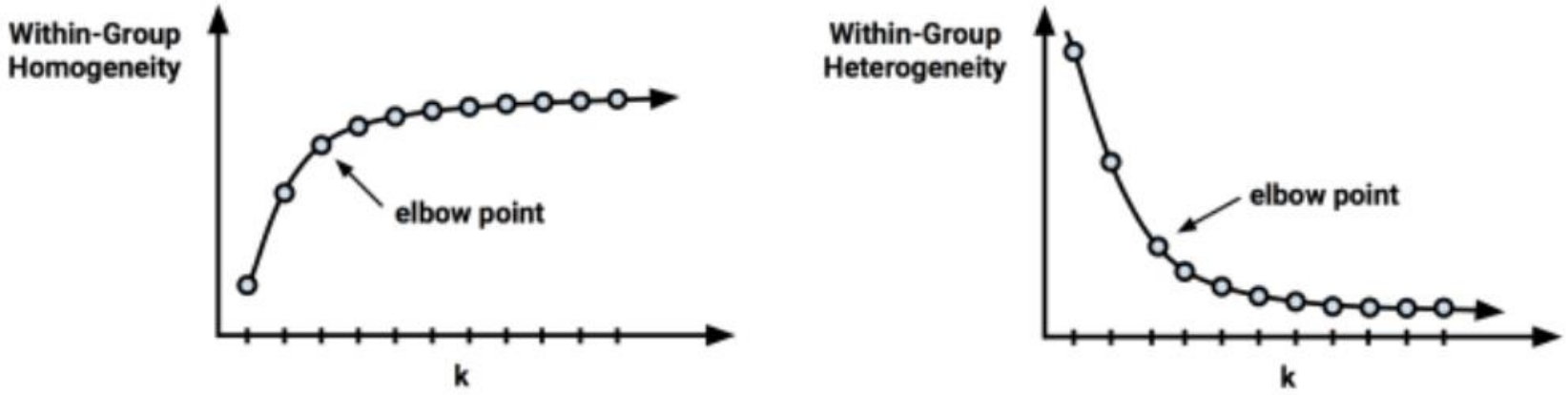
- Sobib ka suure indiviidide arvu korral
- Klastrite arv tuleb eelnevalt defineerida
- Määratakse iga klastri tsentroid (keskpunkt klasterdustunnuste alusel)
  - Kas reaalsed indiviidid või juhuslikud punktid
- Arvutatakse iga indiviidi kaugus tsentroididest ja vähima kaugusega indiviid ja tsentroid rühmitatakse =>
- Arvutatakse uued keskmised ja saadakse tsentroidide uued asukohad
- Iteratiivne protsess, mis lõpeb, kui tsentroidide asukohad enam ei muutu

Klasterdusmeetodi valik

- Indiviidide klasterdamise/liigitamise aluseks võetakse mingi kauguse arvutamise viis
  - Analüüsiprogrammis klasterdamise väljund indiviididevahelisi kaugusi ei pruugi esitada
  - Võib olla siiski oluline kaugusi kontrollida (erindite probleem)

(6) Klastrite arvu valik

- Mitmeid erinevaid tehnilisi kriteeriume
- Tehnilisi kriteeriumid käsitletakse vahel „objektiivsetena“
- Valik nende vahel ikkagi subjektiivne
- Üks võimalus nt nn *elbow method*



- Igasuguste tehniliste kriteeriumide kõrval olulisem siiski klastrite sisuline tähendus
- Võib läbi töötada erineva klastrite arvuga klasterdusi ja teha otsus empiiriliselt

Lantz (2015)

(7) Tõlgendamine, testimine, replikeerimine, valideerimine

- Valideerimine ülalolevatest mõistetest kõige laiem, pmst kaasab teisi
- (kuigi vahel mõistetakse valideerimise all ka ainult klastrite arvu analüüsi)
- Sisemine valideerimine
  - Mitmeid erinevaid parameetreid, mis võimaldavad hinnata klastrisisest homogeensust või klastritevahelist erinevust
  - Oht väikestele erinevustele parameetrites suure kaalu omistamisel
  - Olulisem võib olla väline valideerimine
  - Sissejuhatuseks vt Hennig (2015)
- Väline valideerimine
  - Mitteformaalne – ekspertteadmise põhjal klasterjaotuse hindamine
    - *Does it make sense?*
  - Formaalne – klasterjaotuse seosed teiste tunnustega
    - Kui eeldame/näeme klasterstruktuuris teatud iseärasusi, mis eelneva teadmise kohaselt peaks olema seotud teiste tunnustega, saame nende seoste olemasolu kontrollida
  - Tavaliselt ei saa ühte teisest lahutada, vaja mõlemat

Hennig (2015)