Kyselytutkimusaineistojen tiivistäminen ja visualisointi

Kimmo Vehkalahti

yliopistonlehtori, soveltavan tilastotieteen dosentti
Helsingin yliopisto, sosiaalitieteiden laitos
http://www.helsinki.fi/people/Kimmo.Vehkalahti/suomeksi.html

Sosiaalitutkimuksen tilastolliset menetelmät, jakso 4
7.–8.2.2012

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Jakso 4: tavoitteet ja sisältö

Jakson tavoitteena on oppia tiivistämään ja visualisoimaan **kyselytutkimusaineistoja.**

Tällaiset aineistot ovat tyypillisiä mm. yhteiskuntatieteellisessä tutkimuksessa, jossa mielenkiinto kohdistuu mm. asenteisiin, arvoihin ja mielipiteisiin. Näitä moniulotteisia ilmiöitä mitataan kysely- ja haastattelulomakkeilla (ks. myös aikaisemmat jaksot sekä tilastotieteen johdantokurssi).

1. **Mittaus kyselytutkimuksissa**
   - Mittausmalli
   - Mittauksen laatu
2. **Aineiston tiivistäminen**
   - Oletukset ja rajoitukset
   - Faktorianalyysi ja sen tulkinta
   - Faktoripiste- ja summamuuttujat
3. **Aineiston visualisointi**
   - Hierarkkinen ryhmittely
   - Korrespondenssianalyysi

1. Mittaus kyselytutkimuksissa

Tutkimuskysymyksistä johdettu **mittausmalli** ohjaa lomakkeen tekoa ja aineiston käsittelyä tiivistämisestä lähtien ja antaa hyvän pohjan analyyseihin ja visualisointeihin.

- **mittausmalli: mitä** mitataan, **millä** ja **miten**
- mittausinstrumentti: **millä** mitataan ja **miten**
- kysely- ja haastattelututkimuksen instrumentti: **lomake**
- **mittari:** kokoelma saman aihepiirin mittauksia

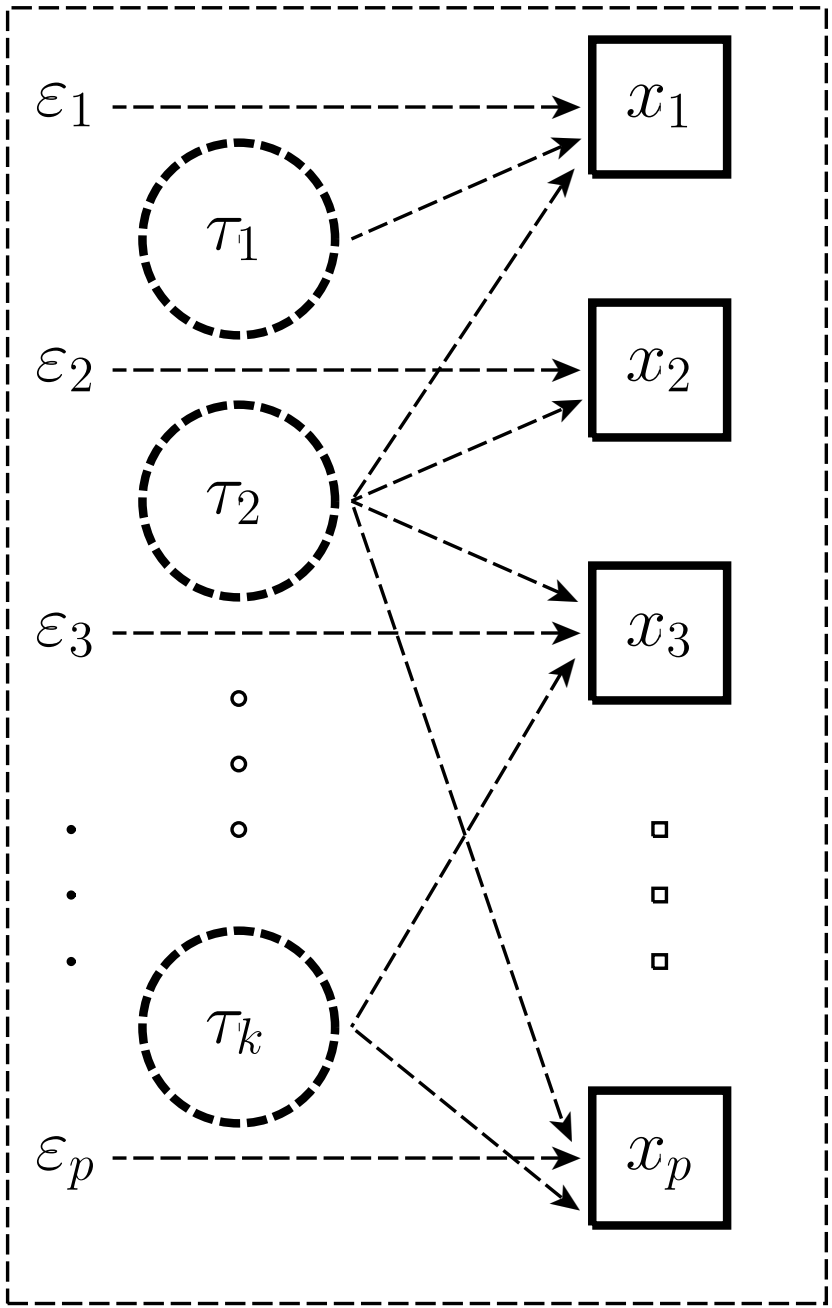
Mittauksen laatu:

1. **validiteetti:** mitataanko sitä mitä pitikin?
2. **reliabiliteetti:** mitataanko riittävän tarkasti?

Mittaustaso vaikuttaa menetelmävalikoimaan:

- luokittelu — järjestäminen — numeerinen mittaus
- useimmat menetelmistä edellyttävät numeerista mittausta
- eräissä menetelmissä luokittelukin riittää
- järjestystasoinen mittaus usein ongelmallisinta

Mittausmallin hahmotelma (vrt. johdantokurssi)

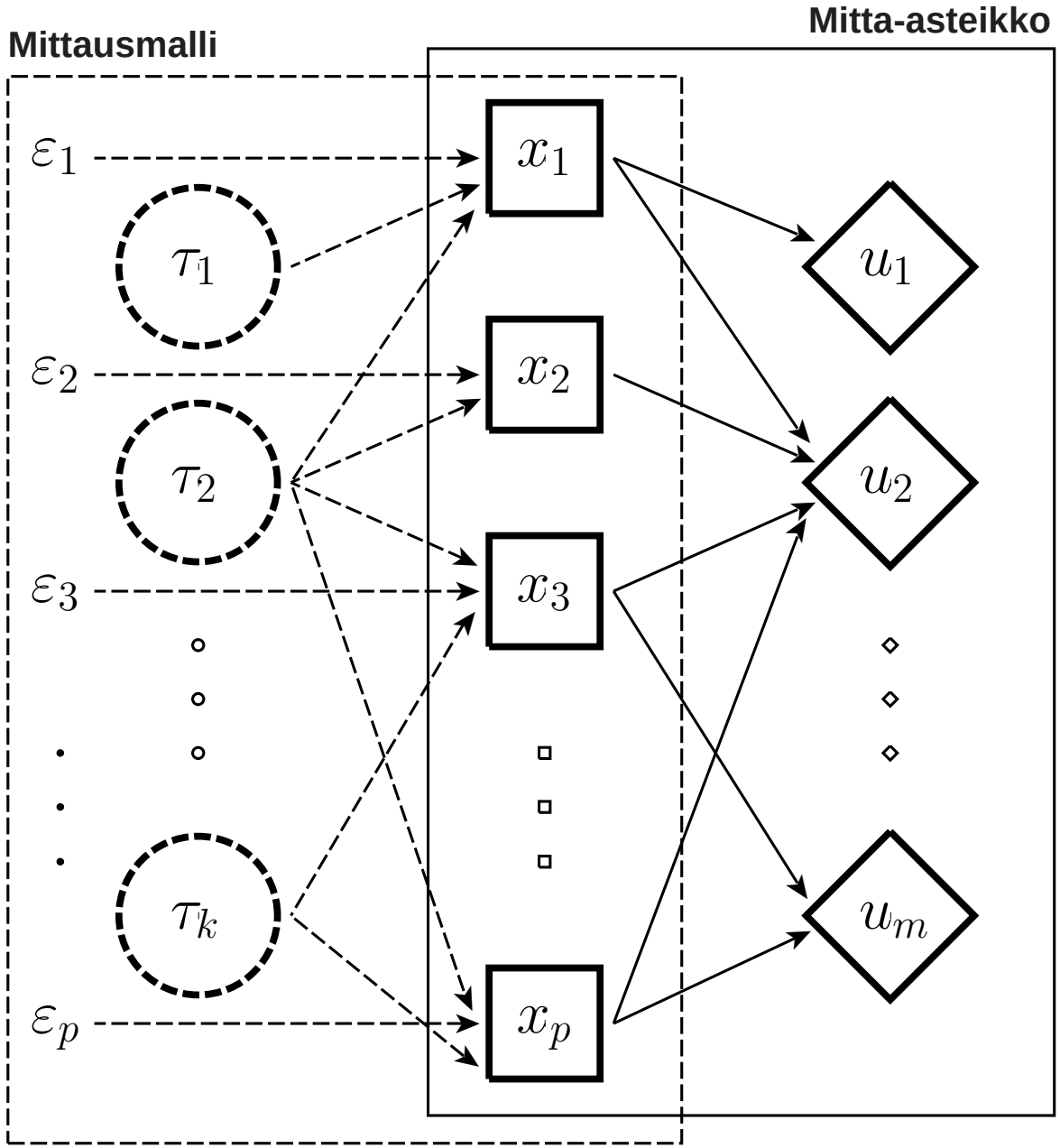


2. Aineiston tiivistäminen

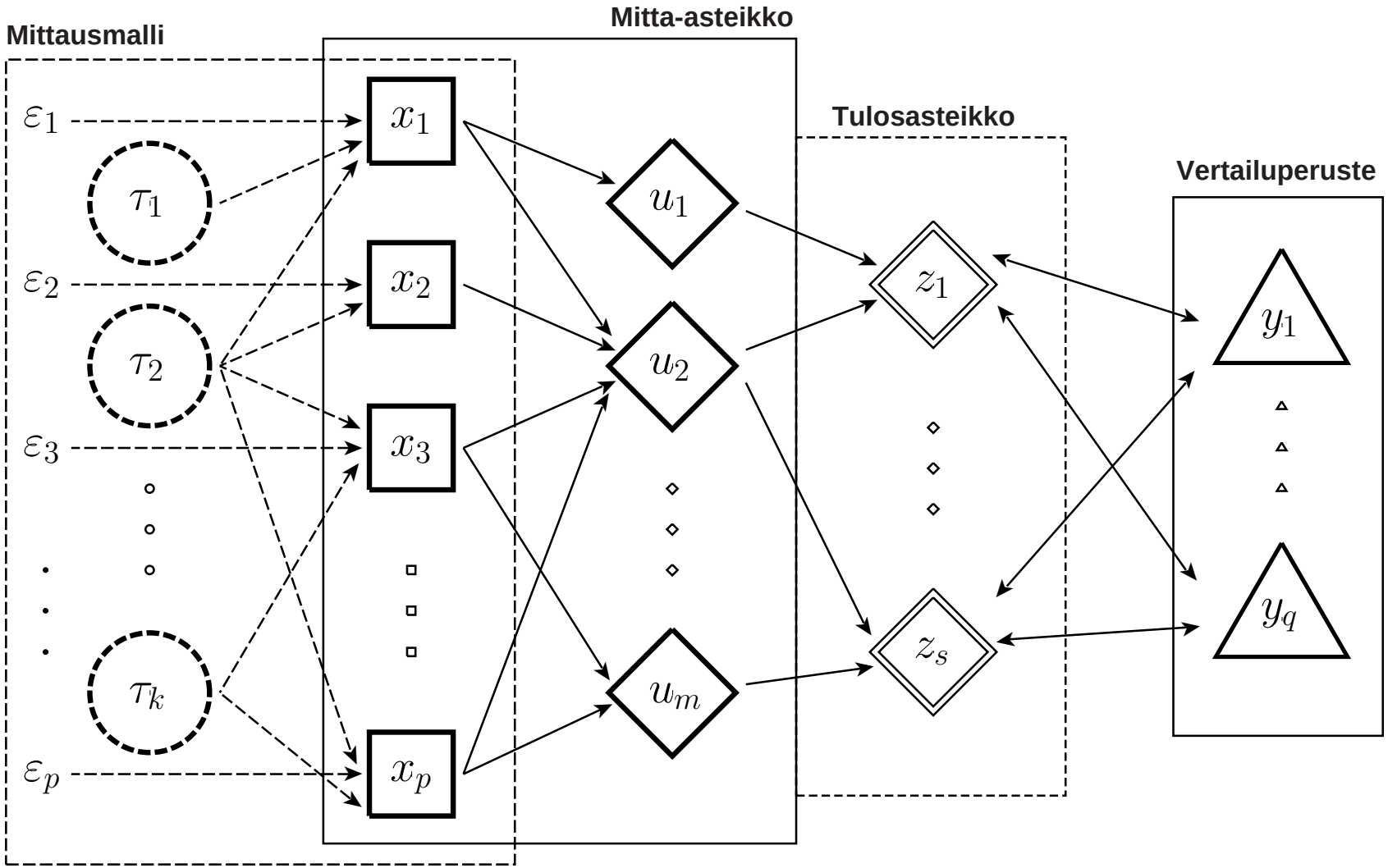
Tilastollisten menetelmien yleinen tavoite on **tiivistää** aineistoon sisältyvää informaatiota kuviksi, tunnusluvuiksi, taulukoiksi yms.

- tiivistäminen (ja muu analysointi) edellyttää ehdottomasti kunnollista aineistoon **tutustumista**
- tutustumisen ja tiivistämisen kannalta keskeistä: **jakaumien kuvaileminen** (graafisesti ja tunnusluvuin)
  - (empiirinen) **jakauma**: aineiston yhden muuttujan **kaikki** (mitatut ja koodatut) arvot
- tiivistäminen edellyttää kuitenkin yleensä enemmän:
  - kyselyaineistoissa on paljon (”liikaa”) muuttujia
  - muuttujien **yhdistely** tutkimusalan teorian perusteella
  - **faktorianalyysi**
  - muut monimuuttuja- ym. analyysimenetelmät

Mittausmallista aineiston tiivistämiseen (faktorianalyysi)



Mittauskehikko: ”suuntaviivat suunnittelusta analyyseihin”



- perusta mittauksen laadun johdonmukaiseen arviointiin
- **keskeinen tilastollinen menetelmä: faktorianalyysi**
- kokonaisuus kattaa myös useita muita menetelmiä

Mittauskehikon neljä osaa lyhyesti

Mittausmalli

- **Mitä** ilmiötä tutkitaan? **Montako** ulottuvuutta siinä on?
- **Millä** ilmiötä mitataan – mahdollisimman hyvin?
- **todentaminen käytännössä: faktorianalyysi**

Mitta-asteikko

- osioiden eli mitattujen muuttujien yhdistelmä
- esimerkkejä: faktoripisteet, summamuuttujat, indeksit jne.
- **aineiston tiivistäminen**

Nämä tässä yhteydessä lähinnä lisätietona:

Tulosasteikko

- syntyy regressio-, erottelu- ym. analyysien tuloksena
- kytkee toisiinsa mittaamisen ja monimuuttujamenetelmät

Vertailuperuste

- mittausmallin ulkopuolella määritelty kriteeri
- vastaajien vertailuun, järjestelyyn, ryhmittelyyn jne.

Faktorianalyysi ja sen tulkinta

Faktorianalyysi on monimuuttujamenetelmä, ts. siinä käsitellään yhtaikaa useita muuttujia. Muuttujien oletetaan mittaavan samaa, tyypillisesti moniulotteista ilmiötä (kuten asenteet, arvot jne.). Tämänkaltaiset ilmiöt ovat *latentteja*, eli niitä ei voida suoraan mitata, vaan joudutaan käyttämään epäsuoria keinoja, siis useita kysymyksiä tai väitteitä.

Faktorianalyysissa erottuu **kaksi eri vaihetta.** Ensin on tavoitteena hahmottaa aineiston avulla taustalla oleva **rakenne**, jota voidaan etukäteen kuvata mittausmallilla. Kirjallisuudessa erotetaan usein kaksi lähestymistapaa: *eksploratiivinen* eli aineistoperustainen ja *konfirmatorinen* eli malliperustainen faktorianalyysi. Käytännössä useimmat analyysit ovat jotain näiden kahden väliltä: kun ilmiöstä tiedetään ennalta enemmän, toiminta perustuu enemmän mittausmalliin – uusia ilmiöitä tutkiessa joudutaan vääjäämättä toimimaan enemmän aineiston varassa.

Faktorianalyysi ja sen tulkinta

Faktorianalyysin toinen vaihe on aineiston **tiivistäminen** löydetyn faktorirakenteen perusteella. Siinä "päästään eroon" suuresta määrästä muuttujia, joita on tarvittu ilmiön mittaamiseen, mutta joista on vaikea sellaisenaan saada kunnollista kokonaiskäsitystä.

Tiivistämisen onnistuminen edellyttää, että on löydetty oikea **faktorilukumäärä**, joka siis vastaa ilmiön ulottuvuuksien määrää. Tässä ennalta pohdittu mittausmalli on avainasemassa, sillä jos lukumäärä hahmotetaan vain aineiston pohjalta, jää turhan paljon epävarmuuksia. On tutkijan tehtävä päättää oikea lukumäärä.

Toinen edellytys on, että faktorit on pystytty **nimeämään** ja **tulkitsemaan** ymmärrettävästi. Myös tässä ilmiön tuntemus on ratkaisevan tärkeää. Ellei tulkinta onnistu, jää aineiston tiivistäminen keinotekoiseksi eikä siitä ole jatkossa vastaavaa hyötyä muiden analyysien ja visualisointien kannalta.

Seuraavassa perehdytään faktorianalyysiin ja sen tulkintaan pienen empiirisen esimerkin välityksellä.

Esimerkki: faktorianalyysi (ESS)

ESS (European Social Survey), round 5, 2010, Suomi (n=1878)

Tarkastellaan tässä vain seuraavia muuttujia:

- How interested in politics (1=very, ..., 4=not at all)

Seuraavissa asteikko 0=no trust at all, ..., 10=complete trust:

- Trust in country's parliament
- Trust in the legal system
- Trust in the police
- Trust in politicians
- Trust in political parties
- Trust in the European Parliament
- Trust in the United Nations

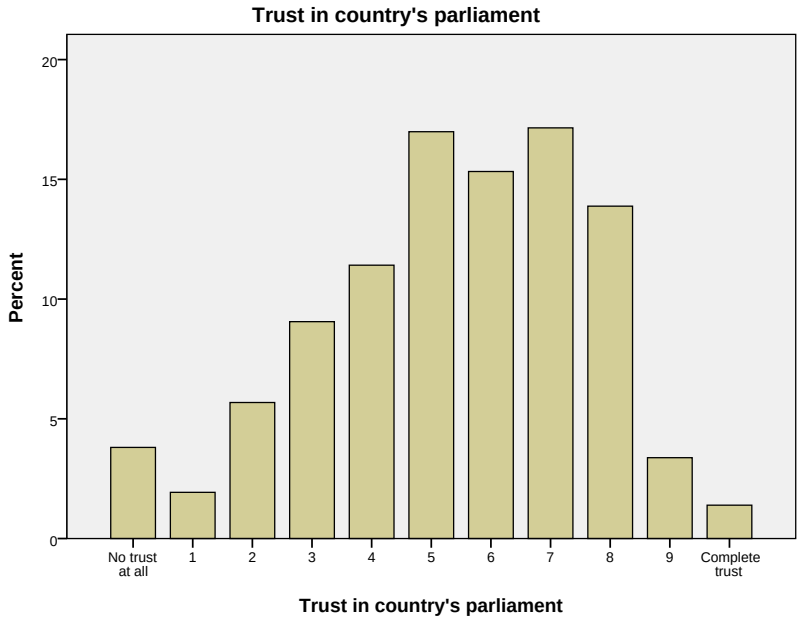
Seuraavissa asteikko 0–10 eri sanamuodoin (10=positiivisin):

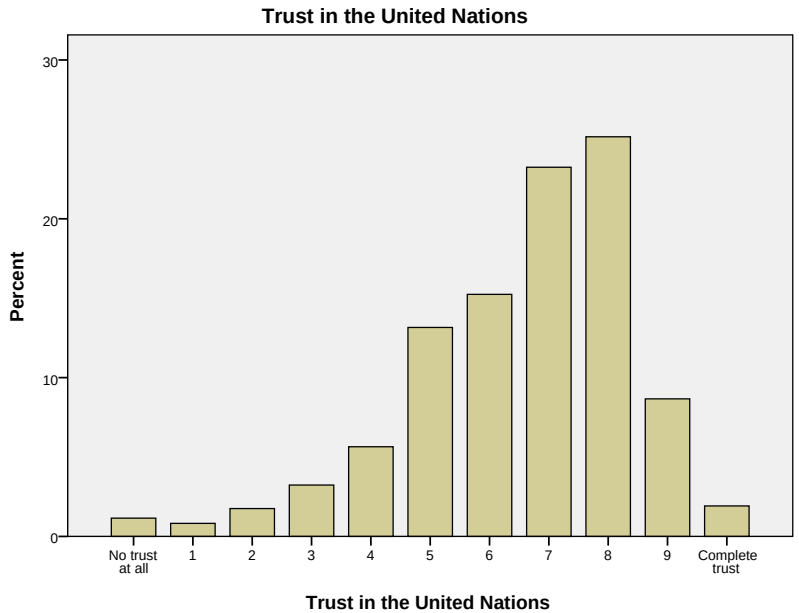
- Most people can be trusted or you can't be too careful
- Most people try to take advantage of you, or try to be fair
- Most of the time people are helpful or mostly looking out for themselves

(Kysymysten tarkemmat sanamuodot on tässä sivuutettu.)

Esimerkki: faktorianalyysi (ESS); oletuksista

Muuttujien jakaumat vaihtelevat, olennaista on mittaustaso:





Oletetaan nyt, että olisi hahmoteltu **2–3 faktorin** mittausmalli (ulottuvuuksina luottamus mm. poliittisiin toimijoihin ja ihmisiin).

Tehdään tältä pohjalta faktorianalyysi SPSS:llä *(luennolla).*

(Esimerkki on sisällöltään keinotekoinen, fokus on menetelmässä.)

Esimerkki: faktorianalyysi (ESS); tulosteista

Faktorianalyysin olennaisin tuloste on ns. rotatoitu faktorimatriisi:

Rotated Factor Matrix[a]

	Factor		
	1	2	3
Trust in politicians	,895	,249	,140
Trust in political parties	,876	,237	,135
Trust in country's parliament	,742	,205	,301
Trust in the European Parliament	,703	,187	,213
Trust in the United Nations	,452	,210	,368
How interested in politics	-,212	-,005	-,088
Most people can be trusted or you can't be too careful	,160	,720	,183
Most people try to take advantage of you, or try to be fair	,099	,686	,153
Most of the time people helpful or mostly looking out for themselves	,195	,582	,104
Trust in the legal system	,378	,214	,745
Trust in the police	,177	,206	,651

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

(järjestettynä faktorilatausten perusteella faktoreittain)

Esimerkki: faktorianalyysi (ESS); tulosteista

On syytä tarkastella myös muuttujien kommunaliteetteja:

Communalities

	Initial	Extraction
How interested in politics	,055	,053
Trust in country's parliament	,656	,683
Trust in the legal system	,538	,743
Trust in the police	,392	,498
Trust in politicians	,801	,883
Trust in political parties	,776	,841
Trust in the European Parliament	,619	,575
Trust in the United Nations	,459	,383
Most people can be trusted or you can't be too careful	,397	,577
Most people try to take advantage of you, or try to be fair	,345	,504
Most of the time people helpful or mostly looking out for themselves	,299	,387

Extraction Method: Maximum Likelihood.

(valitettavasti SPSS antaa ne eri järjestyksessä eri taulukkoon!)

Esimerkki: faktorianalyysi (ESS); tulkinnoista

(Taulukoita ja tulkintoja katsastellaan tarkemmin luennolla.)

Muista tulostaulukoista ilmenee, että kolme faktoria ”selittää” yhteensä 55.7 % kyseisten muuttujien välisestä vaihtelusta. Olennaisempia ovat muuttujien väliset riippuvuudet, joita tulkitaan faktorilatausten avulla *(vrt. mittausmallin nuolet)*.

Ensimmäinen faktori (”luottamus poliittisiin toimijoihin”?) on voimakkain. Toinen voisi olla nimeltään ”luottamus ihmisiin” ja kolmas ”luottamus auktoriteetteihin”. Mitä enemmän nojataan tutkimusalan teoriaan, sitä helpompi on nimetä faktorit. (Teoria saattaa myös tukea *korreloivia faktoreita*; tässä ne on oletettu korreloimattomiksi, mikä on ainakin alkuvaiheessa selkeämpää.)

Kiinnostus politiikkaan, jolla on negatiivinen lataus (huomaa asteikon suunta!), ei nouse juurikaan esiin millään faktorilla. Samaa kuvastaa myös sen kommunaliteetti, joka on käytännössä nolla.

Eräs esimerkki faktorianalyysin tulosten esittämisestä

Aineisto: suomalaisten, hollantilaisten ja englantilaisten kuluttajien ruoan terveellisyyttä ja makua koskevat asenteet

- **Lähde:** K. Roininen, H. Tuorila, E.H. Zandstra, C. de Graaf, K. Vehkalahti, K. Stubenitsky, and D.J. Mela (2001). Differences in Health and Taste Attitudes and Reported Behaviour among Finnish, Dutch and British Consumers: a Cross-National Validation of the Health and Taste Attitude Scales (HTAS). *Appetite*, **37**, 33–45.

Seuraavassa tarkastellaan vain Suomen aineistoa ($N = 467$), ja sen HTAS-mittaria, jonka **Health**-osassa on 3 ulottuvuutta:

1. General Health Interest
2. Light Product Interest
3. Natural Product Interest

Huom! Tämä on esimerkkinä siitä, miten faktorianalyysin tulokset kannattaa esittää tiivistetysti. (Sisältöön ei tässä juuri puututa.)

Eräs esimerkki faktorianalyysin tulosten esittämisestä

Factor structure of Health sub-scales (Finland)	F1	F2	F3	h^2
General Health Interest				
I am very particular about the healthiness of food.	**0.75**	0.16	0.13	0.61
I always follow a healthy and balanced diet.	**0.73**	0.13	-0.02	0.56
It is important for me that my diet is low in fat.	**0.65**	0.12	0.26	0.50
It is important for me that my daily diet contains a lot of vitamins and minerals.	**0.64**	*0.31*	0.10	0.52
(R) I eat what I like and I do not worry about healthiness of food.	**0.47**	*0.34*	*0.31*	0.43
(R) I do not avoid any foods, even if they may raise my cholesterol.	**0.46**	*0.40*	0.27	0.44
(R) The healthiness of food has little impact on my food choices.	**0.42**	*0.31*	*0.50*	0.53
(R) The healthiness of snacks makes no difference to me.	**0.32**	0.24	*0.50*	0.41
Light Product Interest				
(R) In my opinion, the use of light products does not improve ones health.	0.03	**0.77**	0.26	0.67
(R) I do not think that light products are healthier than conventional products.	-0.09	**0.74**	0.17	0.58
I believe that eating light products keeps one's cholesterol level under control.	0.29	**0.61**	-0.04	0.46
(R) In my opinion light products don't help to drop cholesterol levels.	-0.07	**0.61**	0.09	0.38
I believe that eating light products keeps one's body in good shape.	*0.37*	**0.54**	-0.10	0.43
In my opinion by eating light products one can eat more without getting too many calories.	0.23	**0.33**	-0.18	0.19
Natural Product Interest				
(R) I do not care about additives in my daily diet.	*0.49*	0.06	**0.52**	0.51
(R) In my opinion, organically grown foods are no better for my health than those grown conventionally.	0.11	0.20	**0.52**	0.32
(R) In my opinion, artificially flavored foods are not harmful for my health.	0.08	-0.09	**0.50**	0.27
I try to eat foods that do not contain additives.	*0.63*	-0.07	**0.37**	0.54
I would like to eat only organically grown vegetables.	*0.46*	-0.03	**0.34**	0.32
I do not eat processed foods, because I do not know what they contain.	*0.50*	-0.15	**0.23**	0.33
Sum of squares	4.05	2.94	2.01	9.00
Variance explained %	20.3	14.7	10.0	45.0

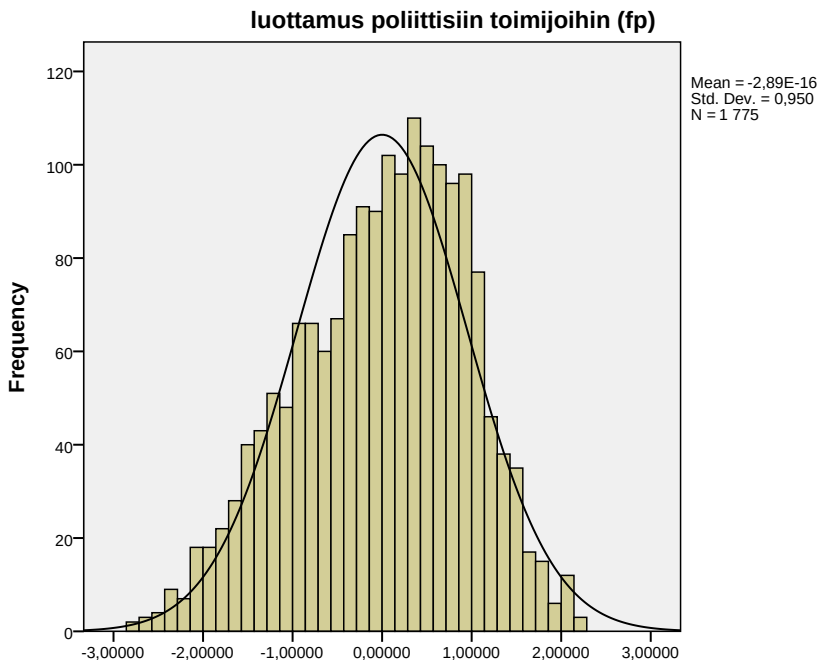
(R) = reversed (negative) statements, h^2 = communalities

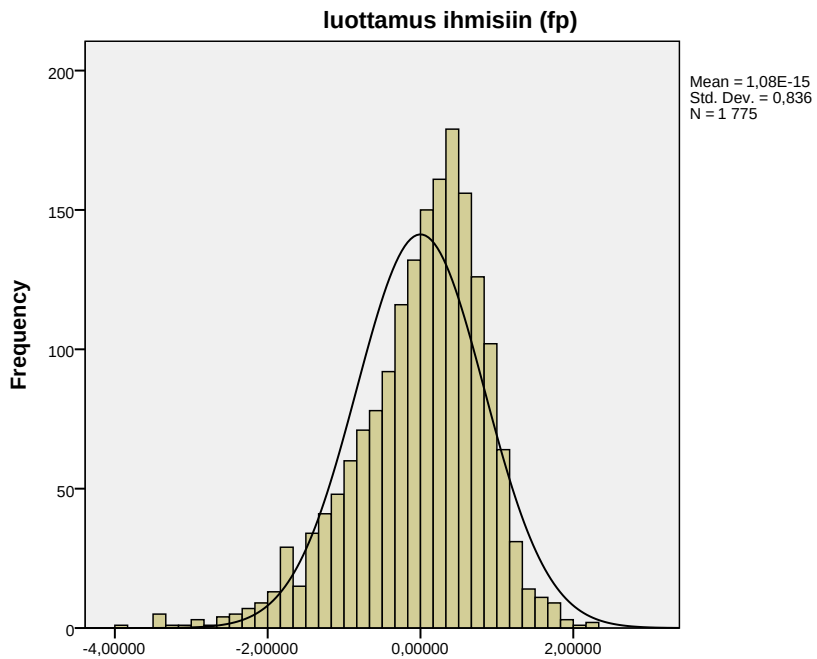
Esimerkki: faktorianalyysi (ESS); vaihe 2: tiivistys

Kun faktorit on tulkittu ja nimetty, on aika siirtyä takaisin aineiston havaintoyksikkötasolle. Muodostetaan uudet faktoripistemuuttujat, jotka kertovat, mihin kohtaan mitäkin ulottuvuutta kuvaavaa jatkumoa kukin vastaaja sijoittuu.

Vaihtoehtoinen tapa on tehdä **summamuuttujia**, mutta ne eivät hyödynnä faktorianalyysista saatua tietoa niin hyvin (esim. sitä, *miten hyvin* eri muuttujat mittaavat kutakin ulottuvuutta).

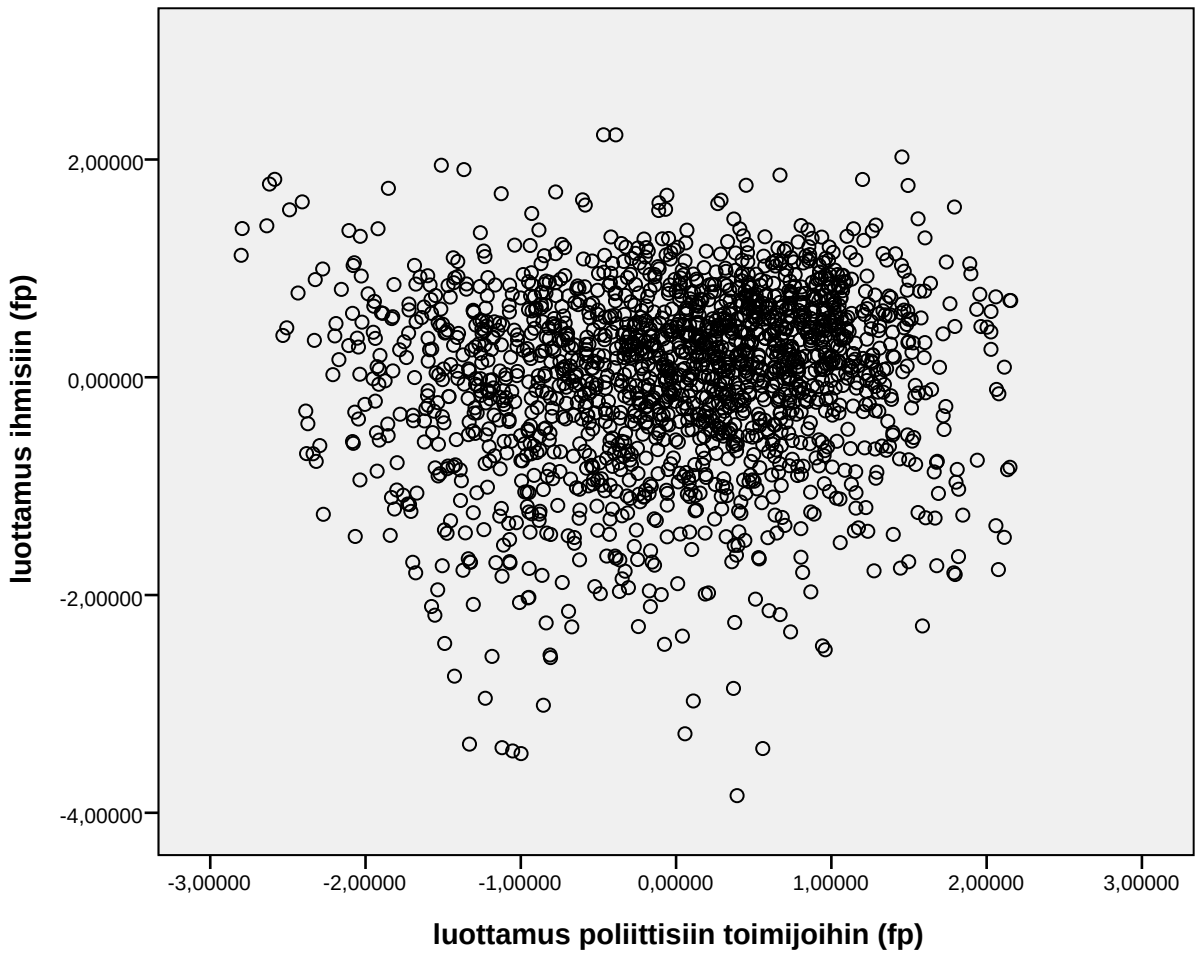
Visualisoidaan kahden faktoripistemuuttujan jakaumia:





Esimerkki: faktorianalyysi (ESS); faktoripisteistä

Ne ovat jatkuvampia kuin alkuperäiset (vinouttakin on toki yhä).
Visualisoidaan näiden samojen muuttujien yhteisjakaumaa:



- Hahmotatko "luottamus/epäluottamus"-nelikentän? Tulkitse!

Esimerkki: faktorianalyysi (ESS); faktoripisteistä

Faktoripisteet voidaan edelleen tiivistää esim. ala- ja yläkvartiilien avulla (<25 %, 25–75 %, >75 %) kolmeen luokkaan: 1="epäluottamus", 2="siltä väliltä", 3="luottamus".

Kahden tällaisen muuttujan välinen ristiintaulukointi:

luottamus poliittisiin toimijoihin (123) * luottamus ihmisiin (123) Crosstabulation

Count

		luottamus ihmisiin (123)			
		1 epäluottamus	2 siltä väliltä	3 luottamus	Total
luottamus poliittisiin toimijoihin (123)	1 epäluottamus	139	205	100	444
	2 siltä väliltä	218	461	208	887
	3 luottamus	88	220	136	444
Total		445	886	444	1775

Tämän tyyppiset tiivistykset ovat tyypillisiä yhteiskuntatieteissä. Jatkuvat muuttujat toimivat edellä vain välivaiheena. Informaatiota hukkuu (paljon), mutta kokonaisuus voi olla helpompi hahmottaa. (Monesti saatetaan tiivistää vain kahteen luokkaan!)

Faktoripiste- ja summamuuttujat

Edellä on mainittu summamuuttujat faktoripisteiden vaihtoehtona. Kummankin käyttöön on perusteensa. Vertaillaan vähän:

Faktoripisteet muodostetaan faktorianalyysin perusteella, jolloin ne pyrkivät vastaamaan mahdollisimman hyvin aikaansaatuja faktoreita. Paremmat muuttujat huomioidaan suuremmilla painoilla kuin huonommat. Muuttujat voivat olla erilaisilla asteikoilla mitattuja sekä eri suuntaisia. Jos faktorit oletetaan keskenään korreloimattomiksi, faktoripisteetkään eivät korreloi keskenään.

Summamuuttujat muodostetaan valitsemalla vain parhaita muuttujia, mutta painottamalla niitä keskenään samanarvoisesti. Tutkimusalan teoria voi "sanella" valittavat muuttujat, jolloin ei faktorianalyysia periaatteessa tarvita lainkaan. On kuitenkin hyvä tarkistaa, miltä tilanne aineistossa näyttää. Muuttujien on joka tapauksessa oltava vertailukelpoisilla asteikoilla mitattuja sekä samansuuntaisia. Summamuuttujat korreloivat yleensä keskenään.

Aineiston tiivistäminen: johtopäätöksiä

Aineiston tiivistäminen on tärkeä ja välttämätön vaihe, kun työskennellään kyselytutkimusaineiston parissa. Mittaus ei onnistu kovin vähillä muuttujilla, mutta kokonaiskäsitysten saamiseksi muuttujia on aluksi ”liikaa”. Tiivistämisen pohjalta pitäisi saada hyvät mahdollisuudet jatkaa eteenpäin. On paljon helpompi jatkaa, kun yhtaikaa käsiteltävien muuttujien määrä on pienempi. *Monenlaiset kiinnostavat analyysit voivat alkaa vasta tästä!*

Jatkon kannalta olennaisen tärkeää on faktorianalyysin toteutus huolellisesti: hahmotetaan faktoreiden oikea (sopiva) lukumäärä sekä nimetään ja tulkitaan ne sisällöllisesti mielekkäästi.
Nämä asiat ovat myös tämän jakson keskeisin sisältö.
Monia kohtia on selitetty ja käyty läpi tarkemmin luennoilla, mutta selostusta löytyy myös oheismateriaalista kurssin kotisivulta.

3. Aineiston visualisointi

Katsotaan lopuksi joitain keinoja aineiston **visualisoimiseen**, kun sitä on ensin tiivistetty faktorianalyysilla.

Sisällysluettelossa esiintyi kaksi monimuuttujamenetelmää, jotka kummatkin ovat luonteeltaan hyvin visuaalisia:

- Hierarkkinen ryhmittely
- Korrespondenssianalyysi

Tässä ei ole mahdollisuutta perusteelliseen läpikäyntiin, mutta lähestytään sen sijaan menetelmiä yleisluontoisemmin perehtyen niiden yleisiin tavoitteisiin ja periaatteisiin.

Menetelmien **päätavoite** on visualisoida aineistoa, samalla kaivautuen siihen syvemmälle ja jatkaen sen tiivistämistä. Ahkera tiivistys voi lopulta tuottaa hyödyllisiä kiteytymiä. Parhaimmillaan ne voivat olla näyttäviä visualisointeja, jotka auttavat *(myös tutkijaa itseään!)* ymmärtämään saatuja tuloksia.

Yleistä visualisointimenetelmistä

Hierarkkiset (ym.) **ryhmittelymenetelmät** tiivistävät yleensä aineistoa **havaintojen** (vastaajien) suhteen (muuttujien osalta tiivistettiin jo edellä). On usein selvempää lähteä liikkeelle faktoripisteistä, koska niitä on paljon vähemmän kuin alkuperäisiä muuttujia. Ryhmittelyt perustuvat heuristisiin kokeiluihin (ilman tilastollista mallia), jolloin tutkimusalan ja ilmiön tuntemus korostuu entisestään.

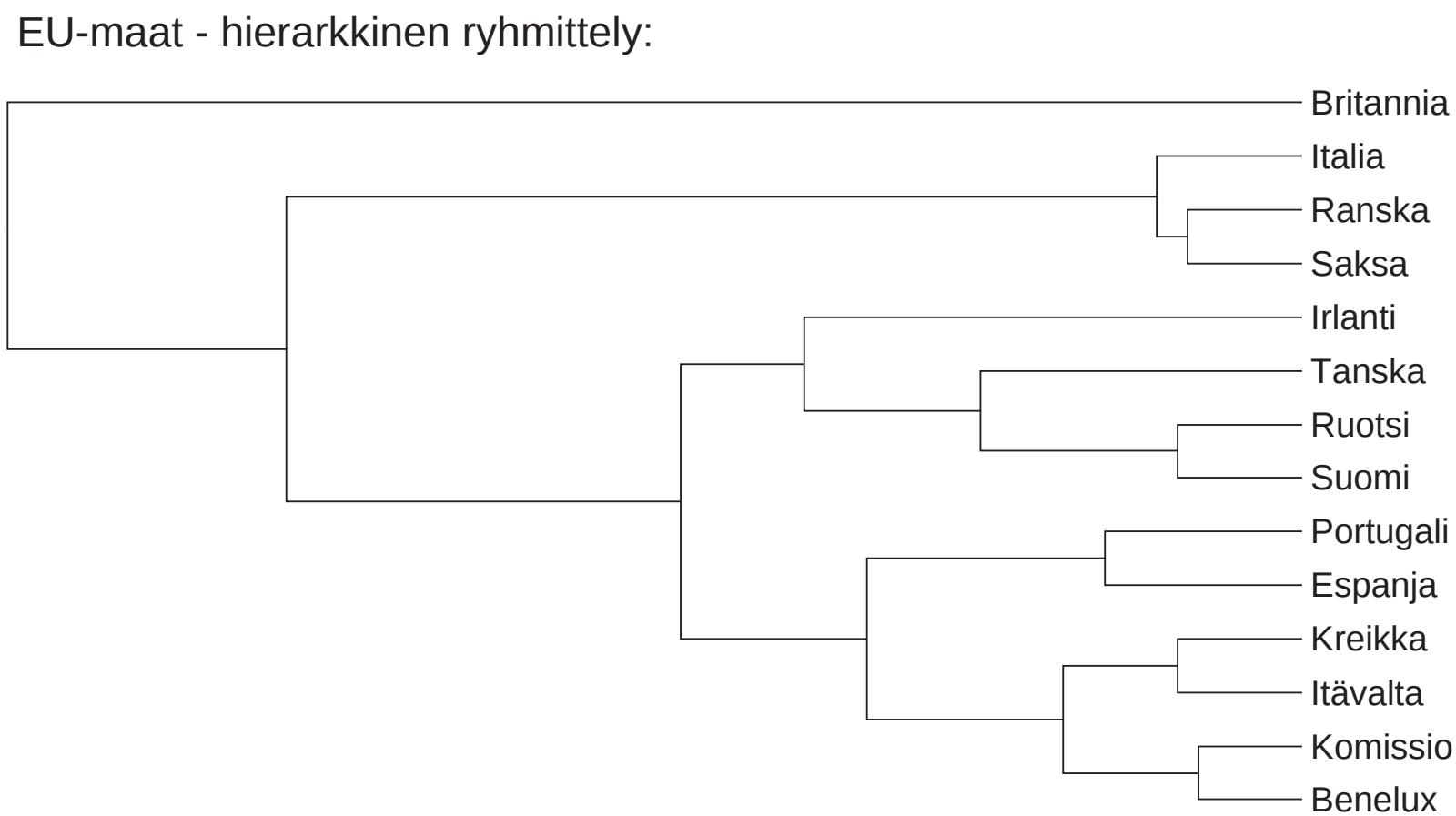
Korrespondenssianalyysi puolestaan lähtee jo pidemmälle tiivistetystä tilanteesta, koska se perustuu luokiteltujen muuttujien käyttöön. Tämä tuo myös joustavuutta, sillä samaan analyysiin voi tuoda monenlaisia muuttujia: (luokiteltuja) faktoripisteitä tai erilaisia (luokittelu- tai järjestystasoisia) taustamuuttujia.

Parhaimmillaan menetelmien tulokset (visualisoinnit) auttavat vastaamaan tutkimuskysymyksiin — tai löytämään uusia!

Esimerkki: hierarkkinen ryhmittely

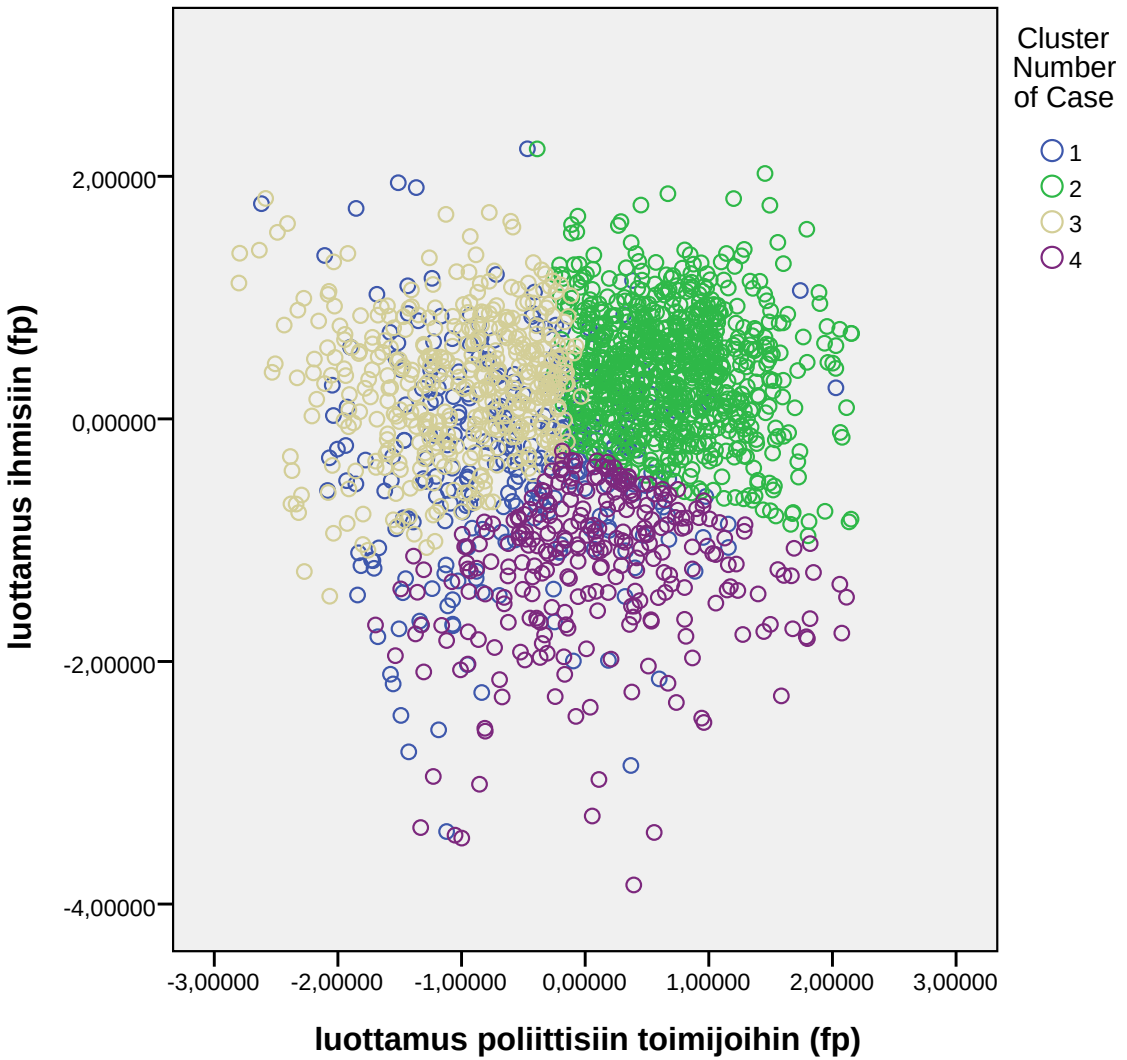
Tavoitteena on hakea profiileiltaan samankaltaisia havaintoja ja muodostaa niistä uusia ryhmiä joillakin sopivilla kriteereillä.

Hierarkkinen ryhmittely soveltuu pienemmille (osa-)aineistoille, joissa havainnoilla on selvät nimet, kuten ohessa:



Esimerkki: ryhmittelyanalyysi (ESS)

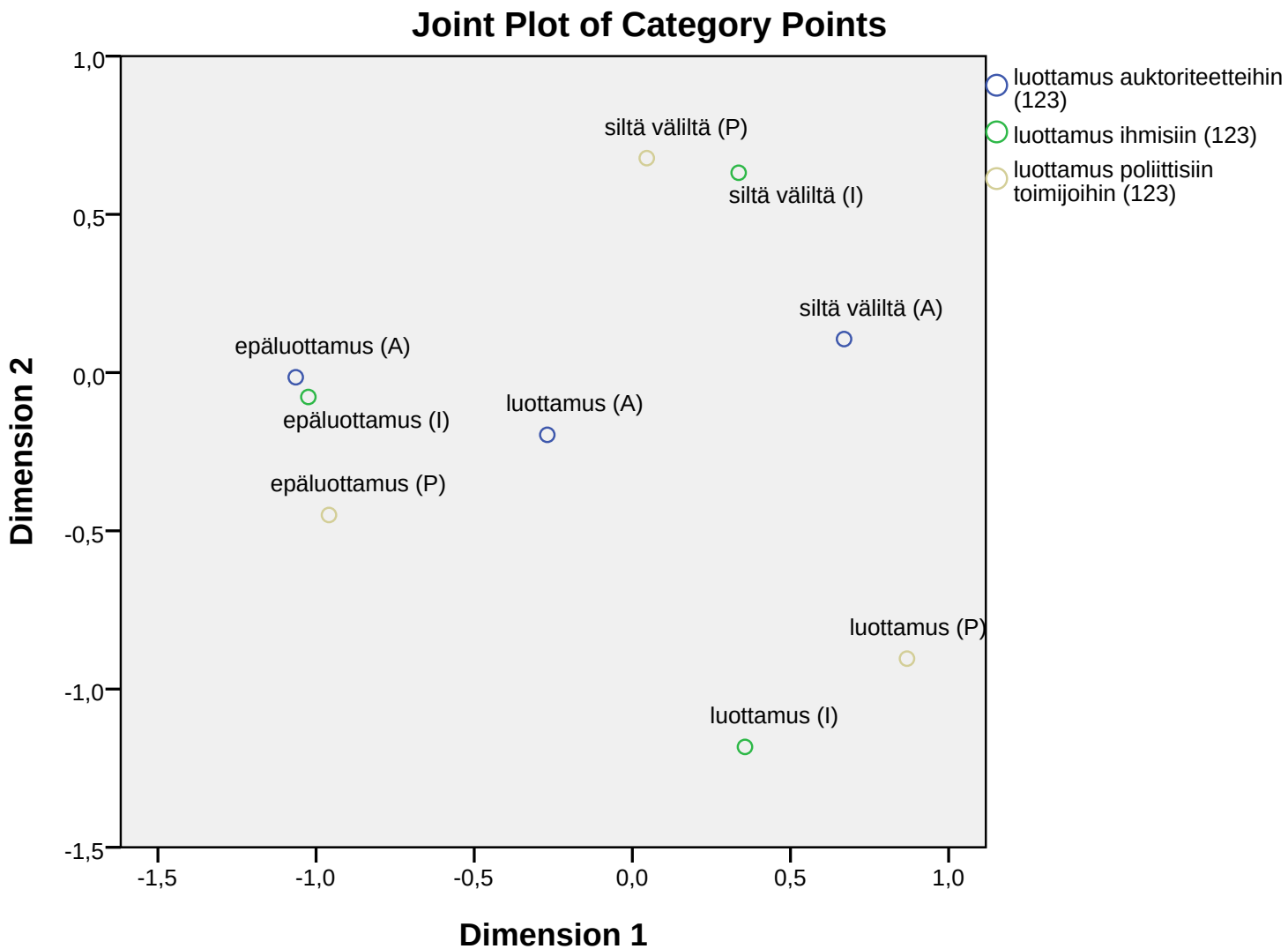
Esimerkki ns. *K-means*-ryhmittelystä ja sen visualisoinnista:



(ryhmät on merkitty aiempaan faktoripisteiden hajontakuvaan)

Esimerkki: korrespondenssianalyysi (ESS)

Kolme faktoripistemuuttujaa yhtaikaa, mutta nyt luokiteltuina:



Esimerkki: korrespondenssianalyysi (ESS); taustat

Lisäksi kaksi taustamuuttujaa (sukupuoli ja ikä):