

G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

versio 1.5 dev, tulostettu 2018-10-13

Sisältö

1	Data	3
1.1	Luvun 1 tavoitteet	3
1.2	Perhe ja muuttuvat sukupuoliroolit - ISSP:n kyselytutkimuksen data 2012	3
1.3	Aineiston rajaaminen	4
1.4	Puuttuvat tiedot	16
1.5	Substanssimuuttujat, taustamuuttujat, muut	29
2	Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko	30
2.1	Äiti työssä	31
2.2	Korrespondenssianalyysin käsitteet	40
3	Tulkinnan perusteita	40
4	Yksinkertaisen korrespondenssianalyysin laajennuksia 1	44
4.1	Täydentävät muuttujat (supplementary points)	46
4.2	Lisämuuttujat: ikäluokka ja sukupuoli	52
5	Yksinkertaisen korrespondenssianalyysin laajennuksia 2	63
5.1	Päällekkäiset matriisit (stacked matrices)	63
5.2	Matched matrices	63

Versiot

6.8.2018 versio 1.0

Siistitään -> 12.8.2018 versio 1.05

Kommentit ja korjaukset -> 4.9.2018 versio 1.1

puuttuva riviprofilikuva, siistimmät interaktiivimuuttujien koodaukset, ensimmäinen “pinottu taulu” - analyysi -> 19.9.2018 versio 1.2

25.9.2018 siistitään datan käsittelyä; ei huomioida puuttuvan tiedon tarkempaa koodausta (read_spss - funktion user_na = TRUE asetus)

1.10.2018 Versio 1.3

Muutokset tarkemmin Readme.md - tiedostossa.

Uusi jakso yksinkertaisen CA:n laajennuksille, joissa otetaan analyysiin useampia muuttujia “pinoamalla” ja/tai yhdistämällä taulukoita. Tässä jaksossa otetaan myös käyttöön isompi aineisto (enemmän maita ja muuttujia). Siisti koodipätkä täydentävien muuttujien lisäämiseen.

3.10.2018 Versio 1.4

Siistitään pois turhat datan listaukset. Aineiston rajaaminen selkeäksi. Ensin kuusi maata, sitten 27 (Espanja pois). Valitaan myös muuttujat, jotta käsiteltävän datan listaukset ovat järkevämpiä. Aineistossa esim. Espanjan ja muutaman Unkarin poikkeavien vastausvaihtoehtojen vastaukset ovat omina muuttujina, ja niiden arvo muille havainnoille on NAP (Not applicable). Samoin paljon maakohtaisia muuttujia, esim. koulutustaso. Mukaan otetaan vain kv-vertailuihin kelpaavat muuttujat, muutama sellainen on myös aineistoon rakennettu. Jätetään pois kaikki perhesuhteisiin liittyvät kysymykset (esim. kotitöiden jakaminen) ja taustatiedot (esim.

rahankäyttö, puolison eri tiedot jne.), koska muuten jouduttaisiin miettimään miten näiden osalta käsitellään perheettömiä. Muutamia muuttujia otetaan mukaan (lasten lkm jne.).

8.10.2018

Datan valinta. Data-jaksossa aluksi, voi miettiä siirtääkö esimerkki-lukuun ja “pinotut taululut” - luvun alkuun kuvailut. Tavallaan siistiä, jos alussa lyhyesti.

10.10.2018

Maiden ja muuttujien valinta. TOPBOT halutaan mukaan, joten USA ja GB on jätettävä pois. Muuttuja on kuitenkin hankala, usealla maalla puuttuva tieto yli 10 prosentissa, ja muutamalla nolla tai ihan muutamia. Pohditaan aikanaan. Data-jaksosta siirretään aineiston laajentamisen yhteyteen laajemman muuttujajoukon deskriptiiviset tarkastelu. Taulukko muuttujakuvauksesta jää data-lukuun.

11.10.2018 Versio 1.4

- paperitulosteessa v1.3 kommentteja karttoihin ja ca:n numeerisiin tuloksiin, samoin muuttujalistauksiin.

11.10.2018 aloitetaan versio 1.5

Muistilista:

1. Taulukot ja kuvat luvusta 2. alkaen eivät ole “bookdown-muodossa”. CA-tulokset on tulostettu Bookdown-demo - dokumentissa. Ominaisarvojen taulukko keskeneräinen, samoin “scree plot” kuva puuttuu.
2. Osa kuvista (esim. profiilikuva) pitää varmaan tulostaa pdf-muodossa ja ottaa capaper-dokkariin `include_graphics` - funktiolla.
3. Puuttuvia tai mahdollisesti lisättäviä taulukoita (nämä saa ca-funktion tuloksista suoraan)
 - khii2 - etäisyydet riveille ja sarakkeille - on tulostettu ilman muotoiluja (11.10.18)
 - massoilla painotetut khii2-etäisyyden keskiarvorivistä/sarakkeesta?
4. Kuvissa vielä hiottavaa, pdf-kuvia lisäilty `img`-hakemistoon.

V MG & Blasius, “vihreä kirja”: contributions to inertia

Historiaa (11.10.18)

Vanhoja kommentteja

- kirjastot/paketit ladataan jokaisessa Rmd-dokumentissa
- bib-formaatin viitetietokantaa tullaan kokeilemaan
- kuvasuhde (aspect ratio) edelleen epäselvä juttu! Mutta näyttää PDF-tulosteessa olevan ok.
- Datan käsittely ja hallinta +SPSS:n sallima kolme puuttuvan tiedon koodia saadaan mukaan `read_spss`-funktion (haven) parametrilla `USER_NA = TRUE` (mutta tarkistettava!) (25.4.18)
 - faktoreita ei ainakaan toistaiseksi muuteta ordinaaliasteikolle, CA ei tästä välitä
 - pidetään muuttujien ja tiedostojen nimeäminen selkeänä, tarkistetaan aika ajoin
- Taulukot: lisättiin riviprosentti- ja sarakeprosenttitaulut (25.4.18), kuva riviprofileista puuttu vielä (15.5.2018)
- Datan esittelyssä on turhaa välitulostusta, ja samoin vähän muuallakin. Html on helpompi lukea, kun koodi on oletuksena piilossa
- PDF-tulosteessa koodi pääsääntöisesti näkyy toistaiseksi
- kokeilu CA-karttojen tulostamiseen (a) suoraan koodilla ja (b) r-grafiikkaikkunasta tallennetun pdf-kuvan avulla. Paras toistaiseksi (a), jätin kokeilu näkyviin. Analyysit R:n grafiikkaikkunassa, jotta `asp=1`, ja tulkintaa varten voi tallentaa PDF-muodossa.
- rakenteeseen muutoksia (näkyvät sisällysluettelossa), ei erillistä teorialiitettä vaan sopivina annoksina. Lukuun 3 perusasiat, kaavat, määritelmät
- tehdään käsitetaulukko (kirjoittamista varten)

- 20.5.2018 (a) tulkita-osuuteen karttakuvia ja ca-tulokset (b) siistimpi taulukoiden tulostus löytyi (c) kaavaliite laajeni (dispo-haarassa)
 - 23.5.2018 lisätään dataan toinen maa-muuttuja maa2, ikäluokkamuuttuja age_cat ja iän ja sukupuolen vuorovaikutusmuuttuja ga.
 - 24.5.2018 lisättiin ca-kartta, jossa Saksan ja Belgian ositteet ja summarivit täydentävinä (passiivisina)
-

1 Data

edit tässä luvussa on paljon siistittävää, mutta data on ok. (13.5.2018). **edit** capaper - dokumentissa parempi uusi jäsentely (4.9.2018) **edit** ISSP-datan perustietoa dokumentissa ISSP_data1.docx (4.9.2018) **edit** koodilohkoja ei vielä siistitä, eikä nimetä capaper-vaatimusten mukaan.

edit 24.9.18 Poistettiin turhaa, uusi versio tiedostosta (G1_1_data1.Rmd -> G1_1_data2.Rmd).

1.1 Luvun 1 tavoitteet

Datan esittely ja kuvailut - tämä luku täysin uusiksi (24.9.18)

10.10.2018 maat ja muuttajat valittu.

TODO Miten tämä dokkari siistitään? Vanha teksti omiksi tэгätyiksi pätkiksi?

2012 data, muuttujaluokat (subst, demog.). Lisäksi maakohtaisia juttuja.

1. Eksploratiivinen ja graafinen menetelmä tarvitseen aineiston, hankalaa esitellä jollain synteettisellä esimerkkiaineistolla. **edit** Eksp&graaf menetelmät määriteltävä johdantoluvussa. Esimerkkiaineistoja (synteettisiä kuten smoke, myös muita) on mm. ca - paketissa.
2. CA (ja MCA) sopivat isojen moniulotteisten ja mutkikkaiden aineistojen analyysiin, siksi iso aineisto. Samalla analyysiä voi laajentaa moneen suuntaan. **V** Benzecri: "kun data menee miljoonaan suuntaan".
3. Aineiston esittely, laajan kyselytutkimusaineiston tyypilliset ominaisuudet
4. Laadukkaan ja hyvin dokumentoidun aineiston edut
5. Tärkeä raja: CA sopii ja sitä on käytetty myös hyvin toisen tyyppisiin aineistoihin (ekologia ja biologia, arkeologia, kielen tutkimus)

1.2 Perhe ja muuttuvat sukupuoliroolit - ISSP:n kyselytutkimuksen data 2012

Hieman historiaa datasta, sosiaalisesti määräytyneen sukupuoliroolit (gender) tutkimusaiheena neljässä kansainvälisessä kyselytutkimuksessa.

Tärkeitä linkit

www.issp.org, tutkimushankkeen historiaa. Löytyy myös bibliografia tutkimuksista, joissa aineistoja on käytetty.

www.gesis.org - tutkimuksen "sihteeristö", dokumentaatio ja datat.

data ja dokumentaatio (selattavissa): zacat.gesis.org

edit tässä järkevä viite ISSP - dataan ISSP Research Group (2016): International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012. GESIS Data Archive, Cologne. ZA5900 Data file Version 4.0.0, doi:10.4232/1.12661 **tämä doi-linkki ei toimi**

Linkitys on hankalaa

- monta portaalia, joista pääsee monien organisaationimien taakse
- tästä lyhyt selostus
- tärkeimmät linkit ISSP-tutkimuksen “kotisivu” ja selkeä **muuttujakuvaukset ja muut tiedot**
- käytännössä linkittäminen “syvälle” johonkin sivustoon tai www-palveluun ei ole järkevää, parempi antaa selkeät viitetiedot ja tiedot organisaatioista. Ne kyllä säilyvät, tai jäljille pääsee.

Aineistot 2012 **toimii - ja viitetieto tuossa edellä! V**

Muuttujakuvaukset ja muut tiedot **OK - täältä löytyy oikeastaan kaikki!**

Data ja dokumentit **vie vain aineiston dokumentoinnin etusivulle**

Suomenkielinen lomake (ZA5900_q-fi-fi.pdf) **vie vain aineiston dokumentoinnin etusivulle**

Käyttöehdot: **GESIS-palvelun datan yleiset käyttöehdot, viittauskäytännöt**

Tiedonkeruumenetelmä ja otoskoko: ** Viimeisin Portugali 29.06.2014 - 31.01.2015, ensimmäinen Bulgaria 16.08.2011 - 20.09.2011. Suurin osa muista 2012-13, kuten Suomi (21.09.2012 - 07.12.2012).

Vie tutkimushankkeen “kotisivulle” ZA5900: International Social Survey Programme: Family and Changing Gender Roles IV - ISSP 2012

Havaintojen lukumäärät voi tarkistaa täältä . **Vie aineiston dokumentointisivustoon, jossa helppo navigoida** zacat.gesis.org.

edit: aineiston kuvailua voi ja kannattaakin jatkossa tarkentaa, ja laittaa se liitteeksi(?- tuskinpa). Dokumentointi on hyvin tarkka, tiedot löytyvät haastattelumenetelmistä (parerilomake, tietokoneavusteinen haastattelu, jne), maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaukset.

1.3 Aineiston rajaaminen

zxy Aineiston kuvailu omana osanaan (7.8.2018). **zxy** capaper - dokumentissa uusi jäsentely (4.9.2018)

Aineistossa (jatkossa ISSP2012) on kyselytutkimukseen tulokset 41 maasta. Lisäksi aineistossa on runsaasti demografisia ja muita taustatietoja. R-koodista selviää käytetty versio (SPSS-tiedoston nimi) ja rajauksessa käytetyt muuttujat.

Rajaukset

zxy Aineiston luonne: maakohtaisesti eri tavoin kerätty data, jossa pyritään yhtenäisiin käytäntöihin ja tietosisältöihin. Silti myös substanssikysymyksissä eroja, isoja ja pienempiä. Näin vain on, en pohdi miksi. Ei ole mitenkään ainutlaatuista. Aineiston editoinnissa ja tiedonkeruun suunnittelussa on nähty paljon vaivaa vertailukelpoisuuden vuoksi. Tästä esimerkkejä, esim. “mitä puoluetta äänestit”.

1. Eurooppa ja samankaltaiset maat (28)

Bulgaria, Czech Republic, Denmark, Finland, France, Germany, Great Britain, Ireland, Latvia, Lithuania, Norway, Poland, Sweden, Slovakia, Slovenia, Spain, Switzerland, Australia, Austria, Canada, Croatia, Iceland, Russia, United States, Belgium, Hungary, Netherlands, Portugal (28) - **Espanja, Great Britain, USA pois -> 25 (11.10.18)**

Pois 13: Argentiina, Turkki, Venezuela, Etelä-Afrikka, Korea, Intia, Kiina, Taiwan, Filippiinit, Meksiko, Israel, Japani, Chile.

2. Maat joissa varsinaisissa tutkimuskysymyksissä on käytetty poikkeavia luokitituksia tms.

Espanja ja Unkari (V18 - V20) ovat muuttaneet ISSP-kysymyksiä. Lähes kaikilla mailla on monissa hieman erilaisia muotoiluja.

Lisäksi iso joukko tietoja on kerätty kansallisilla kysymysversioilla.

Espanja pois, USA ja GB pois koska TOPBOT-muuttuja puuttuu (11.10.18)

25 maata, johdattelevassa esimerkissä 6 maata (11.10.18)

3. kaikki havainnot, joissa on puuttuvia tietoja.

Puuttuvia tietoja on yllättävän vähän. Jätetään silti kaikki ne havainnot pois, joissa joku tieto puuttuu. Rajataan kuitenkin tätä vaikutusta niin, että sitä sovelletaan vain käytettäviin muuttujiin.

Johdattelevassa esimerkissä on kolme muuttujaa, ei ongelma.

Isomman 25 aineiston osalta tarkistetaan, mitä “listwise deletion” saa aikaan. Yksinkertaisinta olisi pudottaa kaikki havainnot, joissa on puuttuvia tietoja joissain valituissa (isommassa joukossa) muuttujia (10.10.18).

Tämä rajausta on kyselytutkimuksessa ankara, tai oikeastaan kelvoton. Oikea menettely olisi imputoida jollain menetelmällä puuttuvat tiedot, mutta rajausta otantatutkimuksen menetelmät tutkielman ulkopuolelle (aiheesta löytyy artikkeleita...). Yksittäisten vastausten puuttuminen eli erävastauskato ohitetaan aluksi, mutta siihen palataan. Korrespondenssianalyysiin on helppo ottaa mukaan myös puuttuvat tiedot, sillä data on luokitteluasteikon dataa. Yksikkövastauskato eli otokseen poimitut joita ei ole tavoitettu ollenkaan on kansallisen tason ongelma, joka on ratkaistu vaihtelevin tavoin. Tiedot löytyvät aineiston dokumentaatiosta. Aineistossa on myös mukana painomuuttujat, mutta ne soveltuvat vain jokaisen maan omaan aineistoon.

zxy Tärkein rajausta esimerkianalyysissä, ja voidaan esitellä CA:n käyttö puuttuvien vastausten analysoinnissa (Likert-asteikkolla).

4. Datan hallinta liittyy reproducible research- periaatteeseen

Aineistoa käsitellään ja muokataan niin, että jokaisen analyysin voi mahdollisman yksinkertaisesti toistaa suoraan alkuperäisestä datasta.

Aineiston muokkauksen (muuttujien ja havaintojen valikointi, muunnokset ja uusien muuttujien luonti jne.) dokumentoidaan r-koodiin.

zxy 3.10.18

Kun SPSS-tiedosto luetaan R:n data frame - tiedostoksi, mukana tulee myös metadata. Uusien muuttujien luonnissa tai data-formaatin vaihtuessa (esim. matriisiksi, taulukoksi jne) metadata katoaa.

Helposti toistettava tutkimus: polku alkuperäisestä datasta analyysien dataan selkeä (ja lyhyt jos mahdollista).

Muuttujien tyyppimuunnokset (yleensä faktorointi) tallennetaan uusiksi muuttujiksi, metatieto säilyy vanhassa muuttujassa.

Tiedostonimistä 10.10.18

ISSP2012.data (df) jossa alkuperäinen SPSS-data ISSP2012jh1.data osajoukko edellisestä ISSP2012jh1a.data - valitaan maat jne. Kerrottu alempana. **huom** tiedostonimessä ei kerrota R-tiedostotyyppiä (?), tulisi liikaa lyhenteitä.

ISSP2012esim1.dat edellisen osajoukkoja, joissa uusia muuttujia ja tyyppimuunnoksia. Nämä vaihtuvat analyysin vaihtuessa mukaan, jotta polku olisi lyhyt. Tästä tulee toistoa R-koodiin, mutta ei liikaa ja sopii myös Markdown-työskentelyyn. Jaksot erillisiä Rmd-tiedostoja, jokaisen alussa ladataan r-paketit ja data. Tallennetaan datan lukukoodi omaksi tiedostoksi, näin on jo tehty paketeille (paketit.R)

zxy Datan rajaaminen (maat. muuttujat) heti alussa, pieni ristiriita eksploraatiivisen data-analyysin perusidena kanssa? No oikeastaan ei, sillä 420 muuttujan aineisto on hieman työläs vaikkapa listauksissa. Muuttujien nimiä joutuu kaivelemaan pitkistä listauksista.

zxy R-koodiin jätetään myös tarkistuksia yms. joita ei raportoida tässä, samoin niiden tuloksia. Voiko R-koodi olla fingelskaa?

DATA RAJAAMISTA - maat(5.10.2018)

```
# Aineiston rajaamisen kolme vaihetta (10.1018)
#
# TIEDOSTOJEN NIMEÄMINEN
```

```

#
# R-datatiedostot .data - tarkenteella ovat osajoukkoja koko ISSP-datasta ISSP2012.data
# R-datatiedostot .dat - tarkenteella: mukana alkuperäisten muuttujien muunnoksia
# (yleensä as_factor), alkuperäisissä muuttujissa mukana SPSS-tiedoston metadata.
# Muutetaan R-datatiedossa alunperin ordinaali- tai nominaaliasteikon muuttuja haven-paketin
# as_factor - funktiolla faktoreiksi. R:n faktorityypin muuttujille voidaan tarvittaessa
# määritellä järjestys, toistaiseksi niin ei tehdä (25.9.2018)
#
# R-datatiedostot joiden nimen loppuosa on muotoa *esim1.dat: käytetään analyyseissä
#
# 1. VALITAAN MAAT (25) -> ISSP2012jh1a.data. Muuttujat koodilohkossa datasel_vars1
#
# kolme maa-muuttujaa datassa. V3 erottelee joidenkin maiden alueita, V4 on koko maan
# ja C_ALPHAN on maan kaksimerkkinen tunnus.
#
# V3 - Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)
# V3 erot valituissa maissa
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
# 62001 PT-Portugal 2012: first fieldwork round (main sample)
# 62002 PT-Portugal 2012: second fieldwork round (complementary sample)
# Myös tämä on erikoinen, näyttää olevan vakio kun V4 = 826:
# 82601 GB-GBN-Great Britain
# Portugalissa aineistoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellinen,
# mutta muut ovat. Tähdellä merkityt maat valitaan johdattelevaan esimerkkiin.
#
# Maat (27, ei Espanjaa).Myös US ja Great Britan jätettiin lopulta pois eli maita jää 25
#
# 36 AU-Australia
# 40 AT-Austria
# 56 BE-Belgium*
# 100 BG-Bulgaria*
# 124 CA-Canada
# 191 HR-Croatia
# 203 CZ-Czech Republic
# 208 DK-Denmark*
# 246 FI-Finland*
# 250 FR-France
# 276 DE-Germany*
# 348 HU-Hungary*
# 352 IS-Iceland
# 372 IE-Ireland
# 428 LV-Latvia
# 440 LT-Lithuania
# 528 NL-Netherlands
# 578 NO-Norway
# 616 PL-Poland
# 620 PT-Portugal
# 643 RU-Russia
# 703 SK-Slovakia

```

```

# 705 SI-Slovenia
# 752 SE-Sweden
# 756 CH-Switzerland
# 826 GB-Great Britain and/or United Kingdom - jätetään pois jotta saadaan TOPBOT
# -muuttuja mukaan (top-bottom self-placement) .(9.10.18)
# 840 US-United States - jätetään pois, jotta saadaan TOPBOT-muuttuja mukaan.(10.10.18)
#
# Belgian ja Saksan alueet:
# V3
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
#
# Unkari (348) toistaiseksi mukana, mutta joissain kysymyksissä myös Unkarilla on
# poikkeavia vastausvaihtoehtoja(HU_V18, HU_V19,HU_V20). Jos näitä muuttujia käytetään,
# Unkari on parempi jättää pois.
#
#
# (25.4.2018) user_na
# haven-paketin read_spss - funktiolla voi r-tiedostoon lukea myös SPSS:n sallimat kolme
# (yleensä 7, 8, 9) tarkempaa koodia puuttuvalle tiedolle.
# "If TRUE variables with user defined missing will be read into labelled_spss objects.
# If FALSE, the default, user-defined missings will be converted to NA"
# https://www.rdocumentation.org/packages/haven/versions/1.1.0/topics/read_spss
#
# (25.9.2018) jätetään pois. Tietoa ei käytetä, koodauksissa on myös eroja maiden ja eri
# kysymysten välillä. Kaikki puuttuvat tiedot saavat R-tiedostossa arvon NA.
# R-ohjelmiston "implisiittinen konversio" muuntaa monet muuttujat (esim. Likert-asteikon
# vastaukset 1,...,5) merkkijonomuuttujiksi.
#

ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav") #luetaan alkuperäinen data R- dataksi (df).

#str(ISSP2012.data)

incl_countries25 <- c(36, 40, 56,100, 124, 191, 203, 208, 246, 250, 276, 348, 352,
372, 428, 440, 528, 578, 616, 620, 643, 703, 705, 752, 756)

#str(ISSP2012.data)
#str(ISSP2012.data) #61754 obs. of 420 variables - kaikki

ISSP2012jh1a.data <- filter(ISSP2012.data, V4 %in% incl_countries25)

#head(ISSP2012jh1a.data)
#str(ISSP2012jh1a.data) #34271 obs. of 420 variables, Espanja ja Iso-Britannia
# pois (9.10.2018)
#str(ISSP2012jh1a.data) # 32969 obs. of 420 variable, Espanja Iso-Britannia,
# USA pois (10.10.2018)
#
#names(ISSP2012jh1.data) # muuttujen nimet
# Maakohtaiset muuttujat (kun on poikettu ISSP2012 - vastausvaihtoehtoista tms.)

```

```
# on aineistossa eroteltu maatunnus-etuliitteellä (esimerkiksi ES_V7).
# Demografisissa ja muissa taustamuuttujissa suuri osa tiedoista on kerätty maa-
# kohtaisilla lomakkeilla. Vertailukelpoiset muuttujat on konstruoitu niistä.
# Muuttujia on 420, vain osa yhteisiä kaikille maille.
```

Data-tiedoston tutkailua

Koodilohkossa voi tarkastaa, mitä valittujen maiden koko datassa on.

```
# Datan selailua (9.10.2018)
#
# datatiedoston ominaisuuksia 1
# str(ISSP2012jh1.data$V5)

# tarpeetonta testailua 24.9.2018 - puuttuvien havaintojen kaivelua
# TÄMÄ VASTA KUN MUUTTUJAT ON VALITTU
#
# test1 <- is.na(ISSP2012jh1.data$V5)
# missV5 <- ISSP2012jh1.data[test1,]
# head(missV5)
# str(missV5)
# hist(missV5$V5)
# summary(missV5$V5)
# puuttuvan tiedon tarkempi koodaus - jos olisi user_na=TRUE read_spss-komenossa
# attr(ISSP2012jh1.data$V5, 'labels')
# attr(ISSP2012jh1.data$V5, 'na_values')

# Tästä näkyvät maakohtaiset substanssimuuttujat helposti!
# attr(ISSP2012jh1.data, 'names')

# hist(ISSP2012jh1.data$V5) #kätevä mutta ei lopulliseen käyttöön
# hist(ISSP2012jh1.data$DEGREE)
#
# PERUSKOMENNOT DATATIEDOSTOSON TUTKIMISEEN
typeof(ISSP2012jh1a.data)

## [1] "list"

class(ISSP2012jh1a.data)

## [1] "tbl_df"      "tbl"        "data.frame"

storage.mode(ISSP2012jh1a.data)

## [1] "list"

#attributes(ISSP2012jh1.data)
#
# tiedoston metadata : muuttujat V1, V2 ja DOI
#
ISSP2012jh1a.data[1:3,1:3]

## # A tibble: 3 x 3
##   V1      V2      DOI
##   <dbl+lbl> <chr>   <chr>
## 1 5900    4.0.0 (2016-11-23) doi:10.4232/1.12661
## 2 5900    4.0.0 (2016-11-23) doi:10.4232/1.12661
```


3 5900 4.0.0 (2016-11-23) doi:10.4232/1.12661

Kolme ensimmäistä muuttujaa ovat datan metatietoja. Muuttujissa V12 ja V13 vastausvaihtoehdot ovat erilaiset, neutraali "ei samaa eikä eri mieltä" puuttuu. "En osaa sanoa" - vaihtoehto kasvattaa puuttuvien havaintojen määrää. Näissä EOS merkitykseltään erilainen. Labels (W, w, H ja jotain eos-vaihtoehdolle).

Ikä (AGE) on kokonaisluku (**tarkista R-tyyppi 13.10.18**), taustamuuttujissa on monenlaisia koodauksia. Yhdistellään ja luokitellaan uudelleen tarvittaessa.

DATAN RAJAAMISTA - MUUTTUJAT (5.10.2018)

```
# 2. VALITAAN MUUTTUJAT -> ISSP2012jh1b.data. Maat valittu koodilohkossa dataset_country1
#
#
# Muuttujat on luokiteltu dokumentissa ZA5900_overview.pdf
# https://zcat.gesis.org/webview/index.jsp?object=http://zcat.gesis.org/obj/fStudy/ZA5900
# Study Description -> Other Study Description -> Related Materials
#
#

# METADATA

metavars1 <- c("V1", "V2", "DOI")

#MAA - maakoodit ja maan kahden merkin tunnus

countryvars1 <- c("V3","V4","C_ALPHAN")

#temp <- select(ISSP2012jh1.data, metavars1)
#str(temp)

# SUBSTANSSIMUUTTUJAT - Attitudes towards family and gender roles (7)
#
# Seitsemän kysymystä (lyhennetyt versiot, englanniksi), vastausvaihtoehdot
#
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri mieltä,
# 4 = eri mieltä, 5 = täysin eri mieltä
#
# Q1a Working mother can have warm relation with child
# Q1b Pre-school child suffers through working mother
# Q1c Family life suffers through working mother
# Q1d Women's preference: home and children
# Q1e Being housewife is satisfying
#
# Q2a Both should contribute to household income
# Q2b Men's job is earn money, women's job household
#
#
# Kysymysten Q3a ja Q3b (pitäisikö naisen olla töissä jos (a) alle kouluikäinen lapsi (b) nuorin lapsi
# vastauksissa on paljon "en osaa sanoa" - valintoja. Muut vastaustvaihtoehdot ovat 1= kokopäivätyö, 2
# 3 = pysyä kotona. Eos-vastaus ei ole sama kuin "en samaa enkä eri mieltä" (ns. neutraali vaihtoehto),
# muista puuttuvan tiedon koodeista. Jätetään siis pois(ks. koodilohko dataprop1).

substvars1 <- c("V5", "V6", "V7", "V8", "V9", "V10", "V11", "V12", "V13")

# Nämä jäävät pois:
```

```

#
# "V14", "V15", "V16", "V17", "V18", "HU_V18", "V19", "HU_V19", "V20", "HU_V20", "V21", "V28", "V29", "V30", "V31",
# "V34", "V35", "V36", "V37", "V38", "V39", "V40", "V41", "V42", "V43", "V44", "V45",
# "V46", "V47", "V48", "V49", "V50", "V51", "V52", "V53", "V54", "V55", "V56", "V57", "V58", "V59",
# "V60", "V61", "V62", "V63", "V64", "V65", "V65a", "V66", "V67"
#
# DEMOGRAFISSET JA MUUT TAUSTAMUUTTUAJAT (8)
#
# AGE, SEX
#
# DEGREE - Highest completed degree of education: Categories for international comparison.
# Slightly re-arranged subset of ISCED-97
#
# 0 No formal education
# 1 Primary school (elementary school)
# 2 Lower secondary (secondary completed does not allow entry to university: obligatory school)
# 3 Upper secondary (programs that allow entry to university or programs that allow to entry other ISCED
# prepare students for direct entry into the labour market)
# 4 Post secondary, non-tertiary (other upper secondary programs toward labour market or technical form
# 5 Lower level tertiary, first stage (also technical schools at a tertiary level)
# 6 Upper level tertiary (Master, Dr.)
# 9 No answer, CH: don't know
# Yhdistelyt?
#
# MAINSTAT - main status: Which of the following best describes your current situation?
#
# 1 In paid work
# 2 Unemployed and looking for a job, HR: incl never had a job
# 3 In education
# 4 Apprentice or trainee
# 5 Permanently sick or disabled
# 6 Retired
# 7 Domestic work
# 8 In compulsory military service or community service
# 9 Other
# 99 No answer
# Armeijassa tai yhdyskuntapalvelussa muutamia, muutamissa maissa. Kategoriassa 9
# on hieman väkeä. Yhdistetään 8 ja 9. Huom! Esim Puolassa ei yhtään eläkeläistä
# eikä kategoriassa 9, Saksassa ei ketään kategoriassa 9.
#
# TOPBOT - Top-Bottom self-placement (10 pt scale)
#
# "In our society, there are groups which tend to be towards the top and groups
# which tend to be towards the bottom. Below is a scale that runs
# from the top to the bottom. Where would you put yourself on this scale?"
# Eri maissa hieman erilaisia kysymyksiä.
#
# HHCHILDR - How many children in household: children between [school age] and
# 17 years of age
#
# 0 No children
# 1 One child

```

```

# 2 2 children
# 21 21 children
# 96 NAP (Code 0 in HOMPOP)
# 97 Refused
# 99 No answer
# koodataan dummymuuttujaksi lapsia (1) - ei lapsia (0) ?
#
# MARITAL - Legal partnership status
#
# What is your current legal marital status?
# The aim of this variable is to measure the current 'legal' marital status '.
# PARTLIV - muuttujassa on 'de facto' - tilanteen tieto parisuhteesta
#
# 1 Married
# 2 Civil partnership
# 3 Separated from spouse/ civil partner (still legally married/ still legally
#   in a civil partnership)
# 4 Divorced from spouse/ legally separated from civil partner
# 5 Widowed/ civil partner died
# 6 Never married/ never in a civil partnership, single
# 7 Refused
# 8 Don't know
# 9 No answer
#
# URBRURAL - Place of living: urban - rural
#
# 1 A big city
# 2 The suburbs or outskirts of a big city
# 3 A town or a small city
# 4 A country village
# 5 A farm or home in the country
# 7 Other answer
# 9 No answer
# 1 ja 2 vaihtelevat aika paljon maittain, parempi laskea yhteen. Unkarista puuttuu
# jostain syystä kokonaan vaihtoehto 5. Vaihtotehdon 7 on valinnut vain 4 vastaajaa Ranskasta.
# Yhdistetään 1 ja 2 = city, 3 = town, rural= 4, 5, 7
#

bgvars1 <- c( "SEX", "AGE", "DEGREE", "MAINSTAT", "TOPBOT", "HHCHILDR", "MARITAL", "URBRURAL")

#Valitaan muuttujat

jhvars1 <- c(metavars1, countryvars1, substvars1, bgvars1)

#jhvars1
ISSP2012jh1b.data <- select(ISSP2012jh1a.data, jhvars1)
str(ISSP2012jh1b.data) #32969 obs. of 23 variables

## Classes 'tbl_df', 'tbl' and 'data.frame':   32969 obs. of  23 variables:
## $ V1      : 'labelled' num  5900 5900 5900 5900 5900 5900 5900 5900 5900 5900 ...
## .. attr(*, "label")= chr "GESIS Data Archive Study Number"
## .. attr(*, "format.spss")= chr "F4.0"
## .. attr(*, "labels")= Named num 5900
## .. ..- attr(*, "names")= chr "GESIS Data Archive Study Number ZA5900"

```

```

## $ V2      : chr "4.0.0 (2016-11-23)" "4.0.0 (2016-11-23)" "4.0.0 (2016-11-23)" "4.0.0 (2016-11-23)"
##   ..- attr(*, "label")= chr "GESIS Archive Version"
##   ..- attr(*, "format.spss")= chr "A25"
##   ..- attr(*, "display_width")= int 26
## $ DOI      : chr "doi:10.4232/1.12661" "doi:10.4232/1.12661" "doi:10.4232/1.12661" "doi:10.4232/1.12661"
##   ..- attr(*, "label")= chr "Digital Object Identifier"
##   ..- attr(*, "format.spss")= chr "A50"
##   ..- attr(*, "display_width")= int 26
## $ V3      : 'labelled' num 36 36 36 36 36 36 36 36 36 36 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)"
##   ..- attr(*, "format.spss")= chr "F5.0"
##   ..- attr(*, "labels")= Named num 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulgaria" ...
## $ V4      : 'labelled' num 36 36 36 36 36 36 36 36 36 36 ...
##   ..- attr(*, "label")= chr "Country ISO 3166 Code (see V3 for codes for the sample)"
##   ..- attr(*, "format.spss")= chr "F3.0"
##   ..- attr(*, "labels")= Named num 32 36 40 56 100 124 152 156 158 191 ...
##   .. ..- attr(*, "names")= chr "AR-Argentina" "AU-Australia" "AT-Austria" "BE-Belgium" ...
## $ C_ALPHAN: chr "AU" "AU" "AU" "AU" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
## $ V5      : 'labelled' num 5 1 2 2 1 NA 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V6      : 'labelled' num 1 5 4 4 4 NA 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V7      : 'labelled' num 3 5 2 4 4 NA 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V8      : 'labelled' num 3 5 5 2 4 NA 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V9      : 'labelled' num 3 1 2 3 4 NA 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working for pay"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V10     : 'labelled' num 1 3 4 2 2 NA 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num 0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V11     : 'labelled' num 3 5 4 4 4 NA 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"

```

```

##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  0 1 2 3 4 5 8 9
##   .. ..- attr(*, "names")= chr   "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ V12      : 'labelled' num  3 NA NA 2 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 3 6 8 9
##   .. ..- attr(*, "names")= chr   "Work full-time" "Work part-time" "Stay at home" "TW: women should d
## $ V13      : 'labelled' num  2 NA 2 1 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 3 6 8 9
##   .. ..- attr(*, "names")= chr   "Work full-time" "Work part-time" "Stay at home" "TW: women should d
## $ SEX      : 'labelled' num  1 2 2 2 2 1 2 1 2 2 ...
##   ..- attr(*, "label")= chr "Sex of Respondent"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 9
##   .. ..- attr(*, "names")= chr   "Male" "Female" "No answer"
## $ AGE      : 'labelled' num  58 59 40 20 72 68 64 57 45 71 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "format.spss")= chr "F3.0"
##   ..- attr(*, "labels")= Named num  15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr   "15 years" "16 years" "17 years" "18 years" ...
## $ DEGREE   : 'labelled' num  2 5 5 3 2 NA NA 6 5 6 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for international comp
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  0 1 2 3 4 5 6 9
##   .. ..- attr(*, "names")= chr   "No formal education" "Primary school (elementary school)" "Lower se
## $ MAINSTAT: 'labelled' num  6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num  1 2 3 4 5 6 7 8 9 99
##   .. ..- attr(*, "names")= chr   "In paid work" "Unemployed and looking for a job, HR: incl never had
## $ TOPBOT   : 'labelled' num  3 7 8 NA 7 2 7 NA 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num  0 1 2 3 4 5 6 7 8 9 ...
##   .. ..- attr(*, "names")= chr   "Not available: GB,US" "Lowest, Bottom, 01" "02" "03" ...
## $ HHCHILDR: 'labelled' num  NA NA 3 1 0 NA 0 0 1 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age] and 17 ye
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num  0 1 2 21 96 97 99
##   .. ..- attr(*, "names")= chr   "No children" "One child" "2 children" "21 children" ...
## $ MARITAL  : 'labelled' num  6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 3 4 5 6 7 8 9
##   .. ..- attr(*, "names")= chr   "Married" "Civil partnership" "Separated from spouse/ civil partner
## $ URBURAL  : 'labelled' num  1 1 1 NA 1 2 NA 2 2 NA ...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 3 4 5 7 9
##   .. ..- attr(*, "names")= chr   "A big city" "The suburbs or outskirts of a big city" "A town or a s
## - attr(*, "notes")= chr "document Plan File: /Users/marcic/Desktop/old/GPS2011 sampling/ISSP2013.s

```

```
#summary(ISSP2012jh1b.data$C_ALPHAN)
```

Metatietojen ja maa-muuttujien lisäksi aineistossa on viisitoista muuttujaa. Seitsemä muuttujaa ovat ns. substanssikysymysten vastauksia, joilla luodataan asenteita sukupuolirooleihin ja perhearvoihin.

Yhdeksän kysymystä (lyhennetyt versiot, englanniksi), vastausvaihtoehdot

Vastausvaihtoehdot:

1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri mieltä, 4 = eri mieltä, 5 = täysin eri mieltä

Q1a Working mother can have warm relation with child Q1b Pre-school child suffers through working mother
Q1c Family life suffers through working mother Q1d Women's preference: home and children Q1e Being housewife is satisfying
Q2a Both should contribute to household income Q2b Men's job is earn money, women's job household

Vastausvaihtoehdot: "Work full-time" "Work part-time" "Stay at home" 1 = W, 2 = w, 3 = H, NA = 6,8,9 ei tässä eriteltynä

Q3a Should women work: Child under school age Q3b Should women work: Youngest kid at school

Kahdeksan demografista ja muuta taustamuuttujaa on kuvattu koodilohkon kommentteissa.

```
# Muuttuja taulukkona
#head(ISSP2012jh1b.data)
#str(ISSP2012jh1b.data)
#
#attributes(ISSP2012jh1b.data)
#temp1 <- attr(ISSP2012jh1b.data$V5, 'label') #labelin saa luettua
#temp1
#attr(ISSP2012jh1b.data[,7:21], 'labels') # tuloksena NULL
#temp1[1]
#str(temp1$names)
#ISSP2012jh1b.data$V6
#lapply(ISSP2012jh1b.data, class) # muuttujat joilla labeleita
#labels(ISSP2012jh1b.data, which = c("V6","V7")) hämähä
#str(substvars1)
# Karkea tapa
tabVarNames <- c(substvars1,bgvars1) # muuttujanimet muuttujille, ei maa- tai metamuuttujia
#tabVarNames
tabVarDesc <- c("Q1a Working mother can have warm relation with child",
  "Q1b Pre-school child suffers through working mother",
  "Q1c Family life suffers through working mother",
  "Q1d Women's preference: home and children",
  "Q1e Being housewife is satisfying",
  "Q2a Both should contribute to household income",
  "Q2b Men's job is earn money, women's job household",
  "Q3a Should women work: Child under school age",
  "Q3b Should women work: Youngest kid at school",
  "Respondents age ",
  "Respondents gender",
  "Highest completed degree of education: Categories for international comparison",
  "Main status: work, unemployed, in education...",
  "Top-Bottom self-placement (10 pt scale)",
  "How many children in household: children between [school age] and 17 years of age",
  "Legal partnership status: married, civil partnership...",
  "Place of living: urban - rural"
)
```

```

#tabVarDesc

# Taulukko
# luodaan df
jhVarTable1.df <- data_frame(tabVarNames,tabVarDesc)
cols_jhVarTable1 <- c("muuttuja","lyhennetty kysymys")
colnames(jhVarTable1.df) <- cols_jhVarTable1
#jhVarTable1.df

# Suomalaiset pitkät kysymykset
vastf1 <- c("Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän
           ja turvallisen suhteen kuin äiti, joka ei käy työssä")

vastf2 <- c("Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä.")
vastf3 <- c("Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.")
vastf4 <- c("On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat
           ensisijaisesti kodin ja lapsia.")

vastf5 <- c("Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.")
vastf6 <- c("Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen.")
vastf7 <- c("Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä.")
vastf8 <- c("Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
           Kun perheessä on alle kouluikäinen lapsi")
vastf9 <- c("Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
           Kun nuorin lapsi on aloittanut koulunkäynnin")

tabVarDesc_fi <- c(vastf1,vastf2,vastf3,vastf4,vastf5,vastf6,vastf7, vastf8,vastf9)
#tabVarDesc_fi
tabVarNames_subst <- c(substvars1)
jhVarTable1_fi.df <- data_frame(tabVarNames_subst,tabVarDesc_fi)
cols_jhVarTable1 <- c("muuttuja","suomenkielisen lomakkeen kysymys")
colnames(jhVarTable1_fi.df) <- cols_jhVarTable1

# kable(booktab = T) # booktab = T gives us a pretty APA-ish table
# Lyhyet kysymykset englanniksi
knitr::kable(jhVarTable1.df, booktab=TRUE)

```

muuttuja	lyhennetty kysymys
V5	Q1a Working mother can have warm relation with child
V6	Q1b Pre-school child suffers through working mother
V7	Q1c Family life suffers through working mother
V8	Q1d Women's preference: home and children
V9	Q1e Being housewife is satisfying
V10	Q2a Both should contribute to household income
V11	Q2b Men's job is earn money, women's job household
V12	Q3a Should women work: Child under school age
V13	Q3b Should women work: Youngest kid at school
SEX	Respondents age
AGE	Respondents gender
DEGREE	Highest completed degree of education: Categories for international comparison
MAINSTAT	Main status: work, unemployed, in education. . .
TOPBOT	Top-Bottom self-placement (10 pt scale)

muuttuja	lyhennetty kysymys
HHCHILDR	How many children in household: children between [school age] and 17 years of age
MARITAL	Legal partnership status: married, civil partnership. . .
URBRURAL	Place of living: urban - rural

```
# Suomen lomakkeen kysymykset (löytyy myös kuva lomakkeen sivusta)
knitr::kable(jhVarTable1_fi.df, booktab=TRUE)
```

muuttuja	suomenkielisen lomakkeen kysymys
V5	Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä
V6	Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä.
V7	Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.
V8	On hyvä käydä töissä mutta tosiasiallisesti useimmat naiset haluavat ensisijaisesti kodin ja lapsia.
V9	Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.
V10	Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen.
V11	Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä.
V12	Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa? Kun perheessä on alle kouluikäinen lapsi
V13	Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa? Kun nuorin lapsi on aloittanut koulunkäynnin

Tarkemmat kysymysten muotoilut poikkeavat tietysti hieman eri maiden välillä. Suomen lomakkeet täydelliset kysymykset voi tarkista tiedostosta ZA5900_q_fi-fi.pdf, löytyy zcat-sivustolta. Tarkemmat kuvaukset lähes tuhatsivuisessa koodikirjassa ZA5900_cdb.pdf ([refworks-viite?](#)).

Kysymyslomakkeesta voisi ehkä näyttää myös kuvan?

```
knitr::include_graphics('img/substvar_fi_Q1Q2.png')
```

““

TODO 9.10.18 Tarkista tiedostojen nimeämiset

1.4 Puuttuvat tiedot

zxy Kun muuttujat on valittu, voi lyhyesti vilkaista puuttuneisuutta. Muuten asia käsitellään aina, kun muuttujia otetaan mukaan analyysiin. Tai voi sen tässäkin ehkä esitellä, kerralla? Muuttujat voisi luetella laskevassa järjestyksessä puuttuneisuuden mukaan?

zxy Perusasiat havaintojen puuttellisuudesta kyselytutkimuksissa. Yksikkövastauskato (unit non-response), eräsvastauskato (item non-response). Mitä on raportoitava, kun käytetään valmista aineistoa? Eräsvastauskato on silti ongelma, vaihtelee kysymyksittäin, vaikka se ei kovin suuri olekaan.

Yksikkövastauskato on otettu vaihtelevasti huomioon, kun kyselyn toteuttaja on editoinut ja tarkastanut datan. Eri maiden datassa on (mutta ei aina!) mukana painot mm. vastauskadon oikaiksemiseen **Viittet - tekninen raportti**. Myös selaimella voi zcat-sivustolla tutkailla kysymyksittäin.

Aineistossa on tarkempi kolmen luokan koodaus puuttuvalle tiedolle, ja sen saa halutessaan luettua R-dataan.

zxy Miten puuttuneisuus kannattaa kuvailla? Löytyy dokumentaatiosta!

Puuttuneisuus muuttujassa V6 (kysymys Q1b) (esimerkki): Ehkä tarpeetonta.

Seuraavaksi perheeseen, työhön ja kotitöihin liittyviä kysymyksiä.

23. Mitä mieltä olet seuraavista väittämistä?
Rengasta jokaiseltä riviltä vain yksi vaihtoehto.

	Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a) Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä	1	2	3	4	5	8
b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä	1	2	3	4	5	8
c) Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö	1	2	3	4	5	8
d) On hyvä käydä töissä mutta tosiasiasa useimmat naiset haluavat ensisijaisesti kodin ja lapsia	1	2	3	4	5	8
e) Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen	1	2	3	4	5	8

24. Mitä mieltä olet seuraavista väittämistä?
Rengasta kummaltakin riviltä vain yksi vaihtoehto.

	Täysin samaa mieltä	Samaa mieltä	En samaa enkä eri mieltä	Eri mieltä	Täysin eri mieltä	En osaa sanoa
a) Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen	1	2	3	4	5	8
b) Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä	1	2	3	4	5	8

25. Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
Rengasta kummaltakin riviltä vain yksi vaihtoehto.

Naisen tulisi...	käydä kokopäivätyössä	käydä osaaikatyössä	pysyä kotona	En osaa sanoa
a) Kun perheessä on alle kouluikäinen lapsi	1	2	3	8
b) Kun nuorin lapsi on aloittanut koulunkäynnin	1	2	3	8

Kuva 1: Suomen lomake

```
# Miksi teksti on sinistä, ja oikean yläkulman painikkeet puuttuvat? (13.10.18)
#Vähän vanhentunutta puuttuneisuuden (erävastauskato) kuvailua
temp <- ISSP2012jh1b.data
temp$maa <- as_factor(temp$C_ALPHAN)
temp$Q1b <- as_factor(temp$V6 )

#summary(temp)
#str(taulu1)
taulu1 <- temp %>% tableX(V6,maa,type = "count")
taulu1a <- taulu1[ ,1:13]
knitr::kable(taulu1a,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen Q1b vastaukset maittain")
```

Taulukko 3: Kysymyksen Q1b vastaukset maittain

	AU	AT	BG	CA	HR	CZ	DK	FI	FR	HU	IS	IE	LV
1	82	218	118	51	75	174	70	47	256	219	13	56	188
2	405	447	395	215	265	392	238	188	551	288	138	250	395
3	285	171	205	181	190	403	152	149	424	225	186	197	156
4	568	205	190	317	327	415	232	423	469	190	552	478	209
5	215	98	13	194	133	355	696	303	624	75	271	197	38
Missing	57	43	82	14	10	65	15	61	85	15	12	37	14
Total	1612	1182	1003	972	1000	1804	1403	1171	2409	1012	1172	1215	1000

```
taulu1b <- taulu1[ ,14:26]
knitr::kable(taulu1b,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen Q1b vastaukset maittain")
```

Taulukko 4: Kysymyksen Q1b vastaukset maittain

	LT	NL	NO	PL	RU	SK	SI	SE	CH	BE	DE	PT	Total
1	50	59	23	110	244	117	39	29	89	193	165	73	2758
2	438	206	186	205	542	246	272	124	421	454	276	405	8422

```

#Faktoreiksi substanssi- ja taustamuuttujat TÄHÄN KELPO TIEDOSTONIMI
#temp$maa <- as_factor(temp$C_ALPHAN)
#temp$Q1a <- as_factor(temp$V5) #labels ainakin näihin
#temp$Q1b <- as_factor(temp$V6)
#temp$Q1c <- as_factor(temp$V7)
#temp$Q1d <- as_factor(temp$V7)
#temp$Q1e <- as_factor(temp$V7)
#temp$Q2a <- as_factor(temp$V7)
#temp$Q2b <- as_factor(temp$V7)
#temp$sp <- as_factor(temp$SEX) # tähän levels, labels
#temp$ika <- temp$AGE
#temp$edu <- as_factor(temp$DEGREE)
#temp$socstat<- as_factor(temp$MAINSTAT)
#temp$class <- as_factor(temp$TOPBOT)
#temp$nchild<- temp$HHCHILDR
#temp$legstat <- as_factor(temp$MARITAL)
#temp$urb<- as_factor(temp$URBRURAL)

#test <-summary(temp)
#str(test)
#head(test)
#test
#temp5 <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V6, type = "count")
#str(temp5)
#temp5
#maat ja havaintojen lukumäärät
#temp6 <- temp5[,7]
#temp6

```

Puuttuvat tiedot ja “listwise deletion” - pientä pohdintaa...

```

# Pohditaan hieman ovatko kaikki muuttujat käyttökelpoisia eli puuttuuko liikaa vastauksia
# Nämä taulukoinnit kuuluvat jaksoon, jossa lisämuuttujat otetaan käyttöön

# ISSP2012jh1b.data %>% tableX(C_ALPHAN, TOPBOT, type = "row_perc")
# puuttuvia tietoja yhteensä 3110/34271 9 prosenttia!
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V5, type = "count")

```

##		V5						
##	C_ALPHAN	1	2	3	4	5	Missing	Total
##	AT	431	409	111	150	47	34	1182
##	AU	358	715	167	270	60	42	1612
##	BE	730	789	247	225	63	148	2202
##	BG	140	425	157	206	36	39	1003
##	CA	278	400	91	136	57	10	972
##	CH	375	591	95	152	19	5	1237
##	CZ	597	502	316	216	110	63	1804
##	DE	1041	481	45	141	37	21	1766
##	DK	849	372	48	81	44	9	1403
##	FI	457	420	98	122	25	49	1171
##	FR	1238	696	160	196	74	45	2409
##	HR	295	413	82	153	51	6	1000
##	HU	297	323	194	124	56	18	1012
##	IE	357	500	109	189	34	26	1215

##	IS	492	523	72	74	9	2	1172
##	LT	100	528	256	232	25	46	1187
##	LV	317	345	111	167	55	5	1000
##	NL	178	597	193	216	63	68	1315
##	NO	341	680	138	207	26	52	1444
##	PL	198	491	103	253	51	19	1115
##	PT	244	508	73	149	20	7	1001
##	RU	412	571	233	215	31	63	1525
##	SE	387	420	122	80	22	29	1060
##	SI	428	436	71	63	9	27	1034
##	SK	614	273	102	84	30	25	1128
##	Total	11154	12408	3394	4101	1054	858	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V6, type = "count") #on jo ylemmänä
```

##		V6						
##	C_ALPHAN	1	2	3	4	5	Missing	Total
##	AT	218	447	171	205	98	43	1182
##	AU	82	405	285	568	215	57	1612
##	BE	193	454	440	554	381	180	2202
##	BG	118	395	205	190	13	82	1003
##	CA	51	215	181	317	194	14	972
##	CH	89	431	222	365	112	18	1237
##	CZ	174	392	403	415	355	65	1804
##	DE	165	376	199	538	441	47	1766
##	DK	70	238	152	232	696	15	1403
##	FI	47	188	149	423	303	61	1171
##	FR	256	551	424	469	624	85	2409
##	HR	75	265	190	327	133	10	1000
##	HU	219	288	225	190	75	15	1012
##	IE	56	250	197	478	197	37	1215
##	IS	13	138	186	552	271	12	1172
##	LT	50	438	396	220	22	61	1187
##	LV	188	395	156	209	38	14	1000
##	NL	59	296	242	445	196	77	1315
##	NO	23	186	226	579	365	65	1444
##	PL	110	395	155	365	64	26	1115
##	PT	73	495	157	215	52	9	1001
##	RU	244	542	360	254	42	83	1525
##	SE	29	124	219	276	354	58	1060
##	SI	39	272	200	365	131	27	1034
##	SK	117	246	229	298	198	40	1128
##	Total	2758	8422	5969	9049	5570	1201	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V7, type = "count")
```

##		V7						
##	C_ALPHAN	1	2	3	4	5	Missing	Total
##	AT	204	430	184	213	119	32	1182
##	AU	105	459	267	477	252	52	1612
##	BE	201	531	387	521	384	178	2202
##	BG	84	303	248	264	53	51	1003
##	CA	46	207	146	329	227	17	972
##	CH	128	453	206	335	100	15	1237
##	CZ	169	379	466	384	341	65	1804

##	DE	183	365	215	493	463	47	1766
##	DK	77	176	119	206	816	9	1403
##	FI	31	120	136	390	434	60	1171
##	FR	279	520	419	506	606	79	2409
##	HR	83	250	180	334	143	10	1000
##	HU	178	279	262	189	89	15	1012
##	IE	80	327	159	394	214	41	1215
##	IS	26	168	201	478	294	5	1172
##	LT	40	404	384	274	32	53	1187
##	LV	185	330	195	225	52	13	1000
##	NL	73	361	250	373	194	64	1315
##	NO	31	251	265	524	301	72	1444
##	PL	80	317	154	438	94	32	1115
##	PT	64	360	192	294	82	9	1001
##	RU	272	555	338	258	46	56	1525
##	SE	30	140	176	288	380	46	1060
##	SI	52	340	219	289	104	30	1034
##	SK	145	281	251	271	161	19	1128
##	Total	2846	8306	6019	8747	5981	1070	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V8, type = "count")
```

##		V8						
##	C_ALPHAN	1	2	3	4	5	Missing	Total
##	AT	93	251	245	257	229	107	1182
##	AU	91	337	384	488	242	70	1612
##	BE	206	399	461	523	388	225	2202
##	BG	82	394	284	141	17	85	1003
##	CA	34	143	268	295	199	33	972
##	CH	86	339	290	399	103	20	1237
##	CZ	296	531	546	223	104	104	1804
##	DE	104	262	217	577	510	96	1766
##	DK	69	159	206	305	594	70	1403
##	FI	58	261	242	292	168	150	1171
##	FR	277	544	487	442	524	135	2409
##	HR	111	284	248	245	94	18	1000
##	HU	212	332	307	96	39	26	1012
##	IE	61	248	243	372	235	56	1215
##	IS	33	242	252	377	243	25	1172
##	LT	43	282	386	271	36	169	1187
##	LV	161	324	248	195	36	36	1000
##	NL	17	176	225	448	345	104	1315
##	NO	31	180	269	506	332	126	1444
##	PL	106	331	204	354	53	67	1115
##	PT	68	338	213	266	90	26	1001
##	RU	218	455	411	287	59	95	1525
##	SE	37	150	252	213	289	119	1060
##	SI	69	323	233	251	98	60	1034
##	SK	263	417	308	80	16	44	1128
##	Total	2826	7702	7429	7903	5043	2066	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V9, type = "count")
```

##		V9					
##	C_ALPHAN	1	2	3	4	5	Missing Total

##	AT	149	241	258	214	190	130	1182
##	AU	198	568	402	277	86	81	1612
##	BE	278	514	492	430	233	255	2202
##	BG	107	370	257	170	24	75	1003
##	CA	112	304	252	171	82	51	972
##	CH	133	569	210	250	48	27	1237
##	CZ	166	336	567	359	229	147	1804
##	DE	193	342	238	494	359	140	1766
##	DK	164	239	276	274	313	137	1403
##	FI	120	267	242	270	118	154	1171
##	FR	259	412	564	545	452	177	2409
##	HR	97	245	182	274	166	36	1000
##	HU	157	263	303	201	62	26	1012
##	IE	155	380	263	274	87	56	1215
##	IS	98	376	363	255	51	29	1172
##	LT	54	262	412	225	33	201	1187
##	LV	139	306	246	213	45	51	1000
##	NL	46	285	335	344	182	123	1315
##	NO	51	249	345	449	197	153	1444
##	PL	113	380	225	308	46	43	1115
##	PT	72	274	189	323	108	35	1001
##	RU	207	471	409	250	52	136	1525
##	SE	60	161	346	196	118	179	1060
##	SI	57	307	205	302	95	68	1034
##	SK	188	263	283	240	99	55	1128
##	Total	3373	8384	7864	7308	3475	2565	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V10, type = "count")
```

##		V10						
##	C_ALPHAN	1	2	3	4	5	Missing	Total
##	AT	322	492	195	107	22	44	1182
##	AU	223	598	505	206	32	48	1612
##	BE	817	760	355	134	30	106	2202
##	BG	431	491	50	11	7	13	1003
##	CA	156	382	265	124	27	18	972
##	CH	202	614	231	172	8	10	1237
##	CZ	1012	547	165	36	18	26	1804
##	DE	633	731	178	134	31	59	1766
##	DK	827	291	195	34	48	8	1403
##	FI	388	507	169	71	10	26	1171
##	FR	1218	711	321	76	34	49	2409
##	HR	410	453	93	34	6	4	1000
##	HU	340	398	206	47	10	11	1012
##	IE	296	409	284	168	19	39	1215
##	IS	356	599	159	47	8	3	1172
##	LT	162	703	233	54	4	31	1187
##	LV	344	429	153	59	8	7	1000
##	NL	218	556	333	115	35	58	1315
##	NO	422	775	184	31	6	26	1444
##	PL	249	593	135	117	13	8	1115
##	PT	461	479	42	16	2	1	1001
##	RU	392	706	275	86	9	57	1525
##	SE	495	408	113	21	5	18	1060
##	SI	432	506	69	11	3	13	1034

```
##      SK      530      383      165      32      9      9      1128
##      Total 11336 13521 5073 1943 404 692      32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V11, type = "count")
```

```
##           V11
## C_ALPHAN 1    2    3    4    5    Missing Total
## AT      118   267   268   248   248   33      1182
## AU      49   208   310   581   416   48      1612
## BE     153   239   369   598   744   99      2202
## BG     129   253   279   245   68    29      1003
## CA      22    96   157   361   326   10      972
## CH      64   242   196   461   268    6      1237
## CZ     332   506   432   322   178   34      1804
## DE     122   168   194   612   633   37      1766
## DK      33    63   124   162  1017    4      1403
## FI      25    78   170   437   416   45      1171
## FR      99   196   326   477  1267   44      2409
## HR      56   142   186   392   215    9      1000
## HU     174   266   325   170   65    12      1012
## IE      46   107   163   499   367   33      1215
## IS      10    66   110   505   479    2      1172
## LT     113   281   498   226   24    45      1187
## LV     233   285   236   196   33    17      1000
## NL      39   127   242   462   385   60      1315
## NO      21    52   160   565   609   37      1444
## PL     176   321   177   353   78    10      1115
## PT      59   176   192   362   212    0      1001
## RU     335   469   404   231   26    60      1525
## SE      18    42   123   271   576   30      1060
## SI      28   185   187   372   245   17      1034
## SK     258   348   300   153   62    7      1128
##      Total 2712 5183 6128 9261 8957 728      32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V12, type = "count")
```

```
##           V12
## C_ALPHAN 1    2    3    Missing Total
## AT      21   483   563   115      1182
## AU      45   695   588   284      1612
## BE     375  1064   461   302      2202
## BG     213   365   332   93       1003
## CA     134   373   271   194       972
## CH      44   849   292   52       1237
## CZ     259   739   666   140      1804
## DE     214   991   296   265      1766
## DK     481   708    72   142      1403
## FI     318   503   165   185      1171
## FR     291  1270   556   292      2409
## HR     307   436   227   30       1000
## HU     166   481   342   23       1012
## IE     126   558   309   222      1215
## IS     346   608    75   143      1172
## LT     128   613   339   107      1187
## LV     101   456   421   22       1000
```

##	NL	70	751	335	159	1315
##	NO	421	711	161	151	1444
##	PL	137	288	611	79	1115
##	PT	237	550	186	28	1001
##	RU	176	767	466	116	1525
##	SE	231	571	108	150	1060
##	SI	321	468	179	66	1034
##	SK	222	431	376	99	1128
##	Total	5384	15729	8397	3459	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, V13, type = "count")
```

##		V13				
##	C_ALPHAN	1	2	3	Missing	Total
##	AT	254	649	135	144	1182
##	AU	345	964	28	275	1612
##	BE	856	941	61	344	2202
##	BG	430	342	142	89	1003
##	CA	403	338	36	195	972
##	CH	136	957	93	51	1237
##	CZ	682	771	217	134	1804
##	DE	454	955	82	275	1766
##	DK	904	353	5	141	1403
##	FI	641	355	26	149	1171
##	FR	1060	967	41	341	2409
##	HR	577	297	94	32	1000
##	HU	409	468	111	24	1012
##	IE	362	592	29	232	1215
##	IS	599	420	21	132	1172
##	LT	458	550	63	116	1187
##	LV	439	450	71	40	1000
##	NL	238	866	43	168	1315
##	NO	853	412	20	159	1444
##	PL	573	328	146	68	1115
##	PT	708	219	44	30	1001
##	RU	513	770	115	127	1525
##	SE	492	407	10	151	1060
##	SI	767	184	30	53	1034
##	SK	606	336	102	84	1128
##	Total	13759	13891	1765	3554	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, SEX, type = "count")
```

##		SEX			
##	C_ALPHAN	1	2	Missing	Total
##	AT	537	645	0	1182
##	AU	699	876	37	1612
##	BE	1055	1138	9	2202
##	BG	422	581	0	1003
##	CA	578	382	12	972
##	CH	620	617	0	1237
##	CZ	807	997	0	1804
##	DE	857	909	0	1766
##	DK	693	710	0	1403
##	FI	514	657	0	1171

```
##      FR      851    1558  0      2409
##      HR      464    536  0      1000
##      HU      483    529  0      1012
##      IE      432    774  9      1215
##      IS      606    566  0      1172
##      LT      493    694  0      1187
##      LV      416    584  0      1000
##      NL      610    705  0      1315
##      NO      690    754  0      1444
##      PL      513    602  0      1115
##      PT      453    548  0      1001
##      RU      547    978  0      1525
##      SE      485    574  1      1060
##      SI      476    558  0      1034
##      SK      523    605  0      1128
##      Total 14824 18077 68      32969
```

```
missAGE <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, AGE, type = "count")
missAGE[,86:87]
```

```
##      AGE
## C_ALPHAN Missing Total
##      AT      0      1182
##      AU      53      1612
##      BE      7      2202
##      BG      0      1003
##      CA      15      972
##      CH      0      1237
##      CZ      0      1804
##      DE      5      1766
##      DK      0      1403
##      FI      0      1171
##      FR      0      2409
##      HR      3      1000
##      HU      0      1012
##      IE      47      1215
##      IS      0      1172
##      LT      0      1187
##      LV      0      1000
##      NL      0      1315
##      NO      0      1444
##      PL      0      1115
##      PT      4      1001
##      RU      0      1525
##      SE      0      1060
##      SI      0      1034
##      SK      0      1128
##      Total 134      32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, DEGREE, type = "count")
```

```
##      DEGREE
## C_ALPHAN 0 1 2 3 4 5 6 Missing Total
##      AT  0 0 824 92 104 0 162 0 1182
##      AU  5 44 357 242 135 486 246 97 1612
```


##	BE	76	188	432	546	153	418	349	40	2202
##	BG	14	58	193	236	264	43	195	0	1003
##	CA	3	25	68	139	228	369	129	11	972
##	CH	1	26	224	55	590	187	152	2	1237
##	CZ	5	0	689	876	26	31	150	27	1804
##	DE	0	18	175	74	1019	152	325	3	1766
##	DK	40	16	74	88	369	562	254	0	1403
##	FI	0	117	80	368	266	193	140	7	1171
##	FR	115	234	734	356	0	499	439	32	2409
##	HR	32	9	276	483	73	123	0	4	1000
##	HU	8	25	476	281	55	120	46	1	1012
##	IE	8	9	217	257	253	187	276	8	1215
##	IS	10	23	323	109	232	262	161	52	1172
##	LT	5	37	283	184	440	204	28	6	1187
##	LV	3	7	157	270	323	0	240	0	1000
##	NL	11	30	337	150	245	283	240	19	1315
##	NO	14	0	336	277	56	185	568	8	1444
##	PL	12	146	71	614	54	53	165	0	1115
##	PT	45	368	195	234	10	74	73	2	1001
##	RU	60	0	141	258	682	384	0	0	1525
##	SE	7	99	281	208	0	159	280	26	1060
##	SI	14	49	392	317	70	176	15	1	1034
##	SK	4	10	501	418	24	23	148	0	1128
##	Total	492	1538	7836	7132	5671	5173	4781	346	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, MAINSTAT, type = "count")
```

##	MAINSTAT											
##	C_ALPHAN	1	2	3	4	5	6	7	8	9	Missing	Total
##	AT	708	52	38	0	0	324	51	0	9	0	1182
##	AU	916	36	35	3	31	389	88	0	41	73	1612
##	BE	1112	150	116	12	61	576	120	0	32	23	2202
##	BG	454	82	34	0	31	353	26	0	23	0	1003
##	CA	412	24	89	2	18	367	22	0	4	34	972
##	CH	771	31	46	18	20	256	74	4	14	3	1237
##	CZ	1056	54	130	3	44	403	5	0	73	36	1804
##	DE	990	98	94	32	22	437	87	0	0	6	1766
##	DK	861	40	131	31	55	241	11	0	33	0	1403
##	FI	656	47	132	12	11	268	17	2	24	2	1171
##	FR	1243	105	64	23	37	766	87	1	83	0	2409
##	HR	494	225	13	5	7	218	30	0	5	3	1000
##	HU	497	98	51	5	0	301	44	0	6	10	1012
##	IE	650	74	29	4	33	266	111	0	19	29	1215
##	IS	772	25	114	10	40	128	18	0	12	53	1172
##	LT	606	90	92	1	35	321	38	0	4	0	1187
##	LV	603	93	73	3	29	138	50	0	11	0	1000
##	NL	678	45	35	2	50	389	77	0	0	39	1315
##	NO	936	17	93	13	84	213	24	2	37	25	1444
##	PL	583	58	72	2	336	0	64	0	0	0	1115
##	PT	521	112	42	5	12	258	44	0	6	1	1001
##	RU	829	61	73	1	57	464	37	0	3	0	1525
##	SE	616	26	54	2	35	250	5	0	32	40	1060
##	SI	487	55	84	1	13	351	33	0	7	3	1034
##	SK	569	74	36	1	36	367	27	0	7	11	1128
##	Total	18020	1772	1770	191	1097	8044	1190	9	485	391	32969

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, TOPBOT, type = "count")
```

```
##          TOPBOT
## C_ALPHAN 1  10  2  3  4  5  6  7  8  9  Missing Total
##   AT      4  35  7  31  81 328 333 219 117 27  0      1182
##   AU      24  44 19  35  65 271 314 344 270 56 170      1612
##   BE      71  31 40  78 124 345 451 521 279 38 224      2202
##   BG      50  1  94 237 219 260 93  31  12  4  2      1003
##   CA      13  36  7  23  36 106 172 223 198 43 115      972
##   CH       4  15 11  41 100 255 246 261 225 36 43      1237
##   CZ      22  7  54 162 294 530 277 222 125 24 87      1804
##   DE       8  17 21  53 103 188 531 443 309 55 38      1766
##   DK       8  37  7  38  52 208 295 379 259 42 78      1403
##   FI      13  16 17  36  78 159 241 315 226 40 30      1171
##   FR      44  16 52 225 293 577 463 310 121 23 285      2409
##   HR      15  7  26 77 103 344 185 131 64  11 37      1000
##   HU      35  1 110 195 228 213 114 67  38  5  6      1012
##   IE      22  60 15  37  52 119 307 244 197 72 90      1215
##   IS      10  14 15  28  62 245 261 225 116 13 183      1172
##   LT      17  4  59 128 195 258 215 175 96  15 25      1187
##   LV      23  2  32 116 187 265 189 119 40  9  18      1000
##   NL      25  18 22  59 114 172 259 359 185 47 55      1315
##   NO      17  15 18  36  82 279 377 330 194 41 55      1444
##   PL      13  16 37  81 131 302 289 145 85  16  0      1115
##   PT      14  9  42 97 157 272 140 71  25  16 158      1001
##   RU      90  8 117 234 246 272 393 100 50  13  2      1525
##   SE      10  25  6  36  57 213 277 254 119 9  54      1060
##   SI       6  12 11  46 102 339 238 143 67  17 53      1034
##   SK       9  4  30 92 193 297 256 165 78  4  0      1128
##   Total 567 450 869 2221 3354 6817 6916 5796 3495 676 1808 32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, HHCHILDR, type = "count")
```

```
##          HHCHILDR
## C_ALPHAN 0    1    18 2    21 3    4    5    6    7 8 Missing Total
##   AT     993 102  0  70  0 15  2  0  0  0 0 0      1182
##   AU    1077 147  1 146  1 50 13  0  0  0 0 177      1612
##   BE    1646 263  0 177  0 31 11  3  1  1 1 68      2202
##   BG     785 148  0  60  0  6  3  1  0  0 0 0      1003
##   CA     740 97  0  51  0 11  2  1  0  0 1 69      972
##   CH     972 147  0  95  0 20  2  1  0  0 0 0      1237
##   CZ    1376 264  0 137  0  8  1  1  1  0 0 16      1804
##   DE    1406 199  0 129  0 18  5  0  0  0 0 9      1766
##   DK     933 189  0 204  0 59 16  1  1  0 0 0      1403
##   FI     874 160  0 101  0 27  4  0  2  2 1 0      1171
##   FR    1283 293  0 225  0 55  6  0  1  0 0 546      2409
##   HR     751 154  0  74  0 18  1  0  0  0 0 2      1000
##   HU     777 129  0  73  0 29  3  0  1  0 0 0      1012
##   IE     892 138  0 109  0 60 11  4  0  0 0 1      1215
##   IS     701 226  0 147  0 45  9  2  0  0 0 42      1172
##   LT     921 186  0  72  0  6  2  0  0  0 0 0      1187
##   LV     741 185  0  66  0  8  0  0  0  0 0 0      1000
##   NL    1092 107  0  89  0 22  3  2  0  0 0 0      1315
##   NO    1004 211  0 140  0 44 10  1  1  1 0 32      1444
```

```
##      PL      759    212    0   105    0   26    7    1    3    2    0    0      1115
##      PT      790    152    0    53    0    6    0    0    0    0    0    0      1001
##      RU     1201    257    0    58    0    6    1    0    2    0    0    0      1525
##      SE      833    128    0    81    0   13    4    1    0    0    0    0      1060
##      SI      810    143    0    72    0    8    1    0    0    0    0    0      1034
##      SK      844    152    0   114    0   13    2    2    0    1    0    0      1128
##      Total 24201 4389    1  2648    1   604  119  21  13  7  3  962      32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, MARITAL, type = "count")
```

```
##           MARITAL
## C_ALPHAN 1      2      3      4      5      6      Missing Total
## AT      719    0      0    154    82    227    0          1182
## AU      974    0     39    141    79    332    47          1612
## BE     1150   159    52    213   126   482    20          2202
## BG      536    86    17    70    166   127    1          1003
## CA      595    85    15    66    53    148   10          972
## CH      705    13    23    88    50    347   11          1237
## CZ     1017    2     9   237   155   366   18          1804
## DE      975    3     23   148   129   488    0          1766
## DK      719    0     20   141    62   461    0          1403
## FI      619    3     10    88    28   400   23          1171
## FR     1235   118    61   249   216   496   34          2409
## HR      571    42     4    75    75   226    7          1000
## HU      427    0     18   165   128   274    0          1012
## IE      750    23    41    45    64   255   37          1215
## IS      532   174    12    57    46   208   143          1172
## LT      604    0     12   132   200   233    6          1187
## LV      473    0     28   151    86   262    0          1000
## NL      726    83     5   107   103   271   20          1315
## NO      795    37    16   121    42   408   25          1444
## PL      655    0     10    55   134   261    0          1115
## PT      485    0     22    76   132   278    8          1001
## RU      654    0     23   187   298   350   13          1525
## SE      521    40    12   108    62   299   18          1060
## SI      525   143    10    49   105   199    3          1034
## SK      686    25     9    82   154   163    9          1128
##      Total 17648 1036 491 3005 2775 7561 453          32969
```

```
ISSP2012jh1b.data %>% tableX(C_ALPHAN, URBURURAL, type = "count")
```

```
##           URBURURAL
## C_ALPHAN 1      2      3      4      5      Missing Total
## AT      421    88   316   324   33      0          1182
## AU      427   515   283   144   188   55          1612
## BE      506   286   471   795   84    60          2202
## BG      463    30   150   359    1     0          1003
## CA      307   197   356    49   57     6          972
## CH      106   121   300   658   51     1          1237
## CZ      645    82   639   434    3     1          1804
## DE      375   186   598   579   28     0          1766
## DK      387   314   391   203   102    6          1403
## FI       97   406   289   219   140   20          1171
## FR      405   382   750   722   136   14          2409
## HR      256   152   325   267    0     0          1000
```

##	HU	351	28	313	320	0	0	1012
##	IE	161	288	348	172	241	5	1215
##	IS	365	372	264	70	55	46	1172
##	LT	434	5	407	336	5	0	1187
##	LV	417	61	284	193	45	0	1000
##	NL	242	87	428	502	36	20	1315
##	NO	342	181	371	304	240	6	1444
##	PL	291	61	346	411	5	1	1115
##	PT	219	241	327	207	3	4	1001
##	RU	750	18	372	385	0	0	1525
##	SE	259	191	296	198	108	8	1060
##	SI	147	101	199	240	344	3	1034
##	SK	99	30	417	571	11	0	1128
##	Total	8472	4423	9240	8662	1916	256	32969

#Taulukko puuttuvista tiedosista 11.10.2018

```
missQ1a <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V5, type = "count")
missQ1b <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V6, type = "count")
missQ1c <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V7, type = "count")
missQ1d <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V8, type = "count")
missQ1e <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V9, type = "count")
missQ2a <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V10, type = "count")
missQ2b <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, V11, type = "count")
missSP <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, SEX, type = "count")
missAGE <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, AGE, type = "count")
```

#missAGE[,86:87]

```
missDEGREE <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, DEGREE, type = "count")
missMAINSTAT <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, MAINSTAT, type = "count")
missTOPBOT <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, TOPBOT, type = "count")
missHHCHILDR <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, HHCHILDR, type = "count")
missMARITAL <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, MARITAL, type = "count")
missURBRURAL <- ISSP2012jh1b.data %>% tableX(C_ALPHAN, URBRURAL, type = "count")
```

USA:n datassa ei ole muuttujaa TOPBOT, ja puuttuvien tietojen osuus on yli kymmenen prosenttia. USA:n voisi ehkä jättää pois, ja muuttaa tässä MISSING-arvon numeerikseksi 99.? Olisiko ok? Tulee silti pulmia, joissain maissa puuttuvia tietoja on nolla tai ihan muutaman. Muiden 1-10 - asteikon muuttujia voi yhdistellä, mutta puuttuvaa tietoa on hankala yhdistää mihinkään. Ehkä tämä muuttuja jätetään pois? Jätetään USA pois(10.10.18)

ainoa ratkaisu taitaa olla se, että (a) katsotaan paljonko putoaa havaintoja jos “listwise deletion” ja (b) koodataan puuttuva tieto omaksi kategoriaksi

Puuttuvien tietojen tarkempaa koodausta ei enään mietitä (3.10.2018)

Puuttuvien tietojen tarkempi koodaus ISSP-datassa:

(zxy 24.9.2018 Tällaisia eroja löytyy, ohitetaan mutta mainitaan. Pitäisikö (a) pudottaa kaikki joilla puuttuvia joissain muuttujissa vai (b) yhdistää NA ja eos?)

Esimerkiksi Ruotsin puuttuviksi tiedoiksi koodatuista 29 havainnosta 19 valitsi “can’t choose”(8) ja 10 kieltäytyi vastaamasta (9) tms. Dokumentti, s.12.

Tarkastellaan aineiston puuttuvia havaintoja hieman tarkemmin. Puuttuvat tiedot on koodattu aineistoon näin: 0: Not applicapble (NAP), Not available (NAV) 7: (97,997, 9997,...): Refused 8: (98, 998, 9998,...): Don’t know 9: (99, 999, 9999,...): No answer

NAP ja NAV määritellään

"GESIS adds ‘Not applicable’(NAP) codes for questions that have filters. NAP indicates that only a subsample

and not all of respondents were asked. Also in the case of country specific variables, all the other countries are coded NAP.

GESIS adds 'Not available' for variables, which in single countries may not have been conducted for whatever reason."

** (3.10.2018) Puuttuvien tietojen rajaava vaikutus raportoidaan, kun tietoja rajataan. Ei pohdita tämän enempää**

1.5 Substanssimuuttujat, taustamuuttujat, muut

zxy capaper - lukuun.

zxy muuttujien kuvaukset.

zxy tässä myös maakohtaisen poikkeavat kysymykset, joita riittää aika lailla.

zxy HUOM! Dataa ei ole kerätty vain kansainvälisiin vertailuihin! Sitä voi ja ehkä pitäisikin analysoida maa kerrallaan, ja vertailla näitä tuloksia. (#V Blasiuksen artikkeli, jossa arvioidaan yhden ISSP-tutkimuksen vertailukelpoisuutta. Kysymykset eivät kovin hyvin näytä toimivan samalla tavalla eri maissa.)

zxy Myös maakohtaiset erot, ja niiden vaikutus aineiston rajaamiseen

zxy yksi kappale: Aineitoa on harmonisoitu, kysymyksiä hiottu, vertailukelpoisuuteen on pontevasti pyritty. Silti eroja löytyy, osa ymmärrettäviä (lisäkysymykset jne) ja osa ei (Espanja!). Tällaista on kansainvälisen kyselytutkimuksen data.

edit: nämä merkinnät ovat muistiinpanoja, kun tarkemmin luin muuttujadokumenttia ensimmäistä kertaa. TÄMÄN PÄTKÄN VOI POISTAA - TURHAA

Kysymyksissä on vaihtelua, ja tavallaan niin pitääkin olla kansainvälisessä kyselytutkimuksessa. Vastaaajien on ymmärrettävä kysymyksen suurinpiirtein samalla tavalla. Kaikki on tarkasti dokumentoitu.

edit: täsmennettävä, periaatteessa vastaukset on harmonisoitu. Joistain maista joku tieto puuttuu, jos sitä ei ole kysytty. Joissain tapauksissa kysymysten vaihtoehdot poikkeavat standardista.

Aineistossa on ns. substanssimuuttujia 63 (V5 - V67). Suurin osa on kerätty jollain haastattelumenetelmällä, ja yleisin vastausvaihtoehto on viiden arvon Likert-skaala (1 = täysin samaa mieltä, samaa mieltä, en samaa enkä eri mieltä, eri mieltä, täysin eri mieltä =5). Eri maiden lomakkeissa on vaihtelua puuttuviksi tiedoiksi koodattujen muiden vastausten välillä.

Esimerkiksi Suomen lomakkeessa on kuudes vaihtoehto "en osaa sanoa", ja lisäksi on koodattu vastaamisesta kieltäytyminen tai muuten puuttuva tieto. Ensimmäisessä aineiston rajauksessa nämä kaikki jätetään pois, käytetään "yleistä" puuttuvan tiedon määritelmää (eli joku noista em.).

Espanjan lisäksi Unkarin osatutkimuksessa kysymyksen V18 V19 V20 vastausvaihtoehdot ovat poikkeavat siten, että keskimäinen neutraali vaihtoehto on jätetty pois (em.dok, s. 48).

Islannissa kysymykseen V28 (Consider a couple who both work full-time and now have a new born child. One of them stops working for some time to care for their child. Do you think there should be paid leave available and, if so, for how long?) on tarjolla oma vastausvaihtoehto ((97) "Yes, but don't know how many months"). Kysymyseen "V29 - Q9 Paid leave: Who should pay ja V30(Paid leave: How to divide between parents) Bulgarian kysely on poikkeava (0 NAP (code 0,98 in V28), s. 91).

Hollannin vastausvaihtoehdoissa kysymykseen V35 (Elderly people: Provider of domestic help) on oma variantti "5 Employers", jonka kuitenkin on valinnut vain 6 vastajaa (0,5 %).

V39, V40, V41, V42, V43, V44, V45, V46, V47, V48, V50, V51, V52, V53, V54: paljon poikkeamia, aika vaikeaselkoisia kysymyksiä. Näitä ehkä pitää tutkailla... V55 (Life in general: How happy on the whole) ok.

V56-57 poikkeamia, V58 (Health status) ok V59 "ketjutettu kysymys", samoin V60-V64. s. 174 - puolison koulutus...

Muuttujat, kysymykset: miten viitata?

SPSS-datassa muuttujat on nimetty V1,...,V67. Metadatatassa taas kerrotaan kysymys, esim. V6 on vastaus kysymykseen Q1b. Suomenkielisessä lomakkeessa ensimmäinen kysymyspatteri Q1 on kysymys 23. 23b:Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä. Miten kysymyksiin kannattaa viitata?

2 Yksinkertainen korrespondenssianalyysi - kahden luokittelu-muuttujan taulukko

jäsennystä

Tässä esitellään yksinkertainen esimerkki, yksi kysymys (esim. V6) ja muutamia maita ristiintaulukoituna. Johdatteluna aiheeseen esitellään ca-käsitteet profiili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofiilien vertailua vastaaviin keskiarvoprofiileihin.

Taulukoita kannattaa tarkastella ensin rivien (kuva puuttuu) ja sitten sarakkeiden suhteen. Miten ne poikkeavat keskiarvostaan, miten toisistaan saman kategorian profiilista. Usein taulukoissa muuttujilla on selvästi eri rooli, kuten tässä. Koitamme hahmottaa maiden (=agregoituja yksilöitä) eroja ja yhtäläisyyksiä. Sarakkeiden vertailussa taas näemme, miten muuttujien profiilit poikkeavat keskiarvostaan. Monia riippuvuusia ja poikkeamia näyttäisi olevan. Klassinen ongelma, Pearson ja Fisher (ehkä turhaa tässä?).

Riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi ehdollisena todennäköisyytenä reunajakauman suhteen. **zxy** Tämä puuttuu kaavoista!

zxy

Tarvitaanko käsitteellistä täsmentämistä, tai selkiinnyttämistä?

1. Taulukon käsite

Erityisesti CA, jossa “ranskalaisella terminologialla” käsitellään yksilöiden tai havaintoyksiköiden pilveä ja muuttujien pilvelä (nominaaliasteikko). Taulukot saadaan yksinkertaisen CA:n tapauksessa aggregoimalla “cloud of individuals”. **#V** MOOC, LeReoux

2. Kontingenssitaulu (kts. viite, jossa ohje “yhteys aina riviä pitkin”), frekvenssitaulu, ristiintaulukointi

- dataa valitaan, aggregoidaan, ryhmitellään. Aktiivisia valintoja. Blasius emt. “data ei löydy kadulta”, ja vaikka siitä ei ole epäilystäkään ISSP-datan tapauksessa, niin siitäkin jatketaan eteenpäin.

3. Peruskäsitteiden yksinkertaisessa esityksessä tärkein lähde MG:n CAiP **#V** Siellä tästäkin on sananen: substanssiero usein on.

4. CA:ssa hämäävä juttu (Blasius, “vizualisation - verkkokirja”) rivien ja sarakkeiden **tekni-**nen symmetria.

χ^2 - etäisyys, yhteys hajontaan eli inertiaan ca-terminologiassa.

Dimensioiden vähentämisen idea (“the essence”), joka ei pienessä taulossa ole ihan ilmeinen. Toinen tavoite on visualisointi, yleensä kaksiulotteisena kuvana (karttana).

Yksinkertainen korrespondenssianalyysi on kahden luokitteluasteikon muuttujan riippuvuuksien geometrista analyysiä. Lähtökohta on kahden muuttujan ristiintaulukointi, alkuperäinen data voi olla muillakin asteikoilla mitattua. Menetelmän ydin on tarkastella molempien muuttujien – taulukon rivien ja sarakkeiden – riippuvuuksia kaksiulotteisena kuvana. Kuvaa kutsutaan myös kartaksi, ja tulkinnan ensimmäinen askel on kartan “koordinaatiston” tulkinta. Kaikki etäisyydet kuvassa ovat suhteellisia, vain rivi- ja sarakepisteiden etäisyydet kuvan origosta voidaan tulkita tarkasti. Koordinaatiston tulkinta aloitetaan “katsomalla mitä on oikealla ja vasemmalla, ja mitä on ylhäällä ja alhaalla” (viite LeRoux et.al, Bezecri-sitaatti). Vaikka pisteiden etäisyyksiä

edes rivi- ja sarakepisteiden välillä ei voi tarkkaan tulkita (approksimaatioita), projektiossa kaukana toisistaan olevat pisteet ovat kaukana toisistaan myös alkuperäisessä “pistepilvessä”.

Akseleiden tulkinta “ääripäiden” kautta (“kontrasti”?). Huom “ääripää” ei välttämättä Likert-asteikolla tarkoita “äärimielipidettä”, vaan se voi tarkoittaa myös selvää tai varmaa mielipidettä.(3.10.18).

Vanha lista:

1. Ensimmäinen taulukko: profiilit, massat, keskiarvoprofiilit, khii2 - riippumattomuustesti ja etäisyysmitta
2. Hyvin tiivis esitys CA:n perusideasta, mutta ilman aivan simppeleitä kolmiulotteisia kuvia (niitä on jo)
3. Ensimmäinen symmetrinen kartta, perustulkinta (mitä kuvasta voidaan sanoa, mitä ei)
4. Lyhyt viittaus graafisen esityksen tulkintapulmiin, jotka eivät ole kovin pahoja. CA-kartta kaksoiskuvana (ts. informaatio voidaan palauttaa, skalaaritulo)?
5. Tulkinnan syventäminen - CA-käsitteiden tarkempi esittely

Haaste: käsitteet ja niiden suhteet ovat abstraktien matemaattisten rakenteiden tuloksia (barycentric, sentroidi), ja ne pitää jotenkin johdonmukaisesti pala kerrallaan tuoda esimerkkien kautta tekstiin. Käsitteistä oma Rmd (ja Excel jos osoittautuu kätevämmäksi), kaavaliite Dispo-repossa ja myös Rmd-muodossa. **edit** Kaavaliitteessä pieniä eroja, ja tekstiä on LaTeX-versiossa enemmän.

Ensimmäinen symmetrinen kartta

Tulkinnat ja yksinkertaisimmat perussäännöt. Dimensiot ja kuinka paljon alkuperäisen taulukon inertiaa saadaan esitettyä kartalla. Sitten asian ydin, akseleiden tulkinta (“mitä on oikealla ja vasemmalla”). Jos pisteet ovat alkuperäisessä “pilvessä” kaukana toisistaan, ne ovat sitä myös projektiossa. Kartta, mutta etäisyyksillä ei suoraa tulkintaa paitsi etäisyyksillä origoon. Rivipisteiden suhteelliset etäisyydet, samoin sarakepisteidet. Mitä tarkoittavat prosentit akseleilla?

Varoitus virhetulkinnasta: ryhmien tunnistaminen rivi, jopa rivi- ja sarakepisteistä koostuvien ryhmien. **zxy** Ja silti tavallaan voi. Sarake- ja rivipisteiden etäisyyksille ei ole suoraa tulkintaa, mutta on “vetovoima” (attraktio) ja “työntövoima” (repulsio). Jos profiilissa sarakemuuttujan osuus on suuri (siis suurempi kuin keskiarvopisteessä, suhteellinen ero), se “ajautuu” lähelle sarakepistettä. MG: “loose ends” - paperi, symmetrinen kuva eräs suurin sekaannuksen lähde. Tätä koitetaan selventää myös MG:n JASA-artikkelissa.

zxy termi korrespondenssi: “neglected multivariate method” - paperissa käännetty näin englanniksi ransk. termi, tätä itsekkin nykyään käyttävät), rivien ja sarakkeiden “correspondence” eli yhteys/“riippuvuus”/vastaavuus tms.

zxy . Tarina: valitaan edellisessä luvussa esitetyn pohjalta osa muuttujista, perustellaan miksi työmarkkia-asenteen ovat kiinnostavia, valitaan esimerkkianalyysiin **yksi** muuttuja ja kuusi maata.

2.1 Äiti työssä

zxy Perustellaan aineiston valinnan vaiheet. Esimerkiksi otetaan yksi kysymys.

zxy Suhde data-lukuun, siellä pitäisi esitellä aineisto sisällöllisesti. Tässä vain valitan esimerkkiä varten yksi kysymys ja kuusi maata.

zxy Muuttujien nimeäminen vaikuttaa (a) muuttujien faktorointiin ja (b) kuviin ja taulukoihin.

Aineisto muuttujat V5-V9 ovat vastauksia ensimmäiseen kysymyspatteriin (Q1) (1-5 Likert, täysin samaa mieltä - täysin eri mieltä) seuraaviin kysymyksiin (suomenkielinen lomake, kysymys 23): **Käytänkö muuttujanimenä Q1b vai V6? Jälkimmäinen lyhyempi, ja tätä muuttujaa käytetään. Toisaalta kun faktoroidaan, pitää tehdä uusi muuttuja, muuten metadata häviää** (3.10.2018)

- (a) Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän ja turvallisen suhteen kuin äiti, joka ei käy työssä

- (b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä
- (c) Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö
- (d) On hyvä käydä töissä mutta tosiasiaa useimmat naiset haluavat ensisijaisesti kodin ja lapsia
- (e) Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen

zxy Tässä koodilohkossa esimerkki helposti toistettavasta tutkimuksesta: alkuperäisestä datasta liikkeelle. `user_na = true` - jutun voi mainita, ja viitata koodiliitteeseen. Alaviitteeksi, sillä ei analysoida tarkemmin?

zxy koodilohkon loppu ehkä tarpeeton, tarkistettava!

```
# Alkuperäinen data (ei käytetä user_na =TRUE 25.9.2018)
ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav")
#
# str(ISSP2012.data)
#61754 obs. of  420 variables ja 61754 obs. of  420 variables 25.4.18
#

# Yksi kysymys ja kuusi maata
incl_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU)

ISSP2012esim1.dat <- filter(ISSP2012.data, V4 %in% incl_esim1)
#str(ISSP2012esim1.dat) #8557 obs. of  420 variables
#
# mukaan muuttujat, V3 jos halutaan jakaa Saksa ja Belgia
# SEX 1=male, 2=female AGE haastateltava ikä haastatteluhetkellä
#
ISSP2012esim1.dat <- select(ISSP2012esim1.dat, C_ALPHAN, V3,V4, V6, SEX, AGE)

#str(ISSP2012esim1.dat) #8557 obs. of  6 variables
#
```

zxy Tehdään aineistoon muutama muutos (eli faktoreiksi, mutta ei järjestystä), jotta sen käsittely on helpompaa.

zxy taulukot erotettava omiksi koodilohkoiksi bookdowniin.

```
# muutetaan muuttujia faktoreiksi
#
# Luokittelumuuttujien tasoille labelit
#
# sp (sukupuoli) m = 1, f = 2
sp_labels <- c("m","f")
# S = täysin samaa mieltä, s = samaa mieltä, ? = ei samaa eikä eri, e = eri mieltä, E = täysin eri miel
vastaus_labels <- c("S","s","?","e","E")

# Faktoreiksi
ISSP2012esim1.dat$maa <- factor(ISSP2012esim1.dat$C_ALPHAN)
ISSP2012esim1.dat$sp <- factor(ISSP2012esim1.dat$SEX, labels = sp_labels)
ISSP2012esim1.dat$Q1b <- factor(ISSP2012esim1.dat$V6, labels = vastaus_labels)
str(ISSP2012esim1.dat$Q1b)

## Factor w/ 5 levels "S","s","?","e",...: 3 2 3 4 3 3 4 3 2 3 ...

#
# toinen maa-muuttuja, jossa Saksan ja Belgian jako
# V3
```



```

# 5601      BE-FLA-Belgium/ Flanders
# 5602      BE-WAL-Belgium/ Wallonia
# 5603      BE-BRU-Belgium/ Brussels
# 27601     DE-W-Germany-West
# 27602     DE-E-Germany-East
#
# Tarkastuksia
#
#ISSP2012esim1.dat %>% tableX(maa,V6,type = "count") # 400 missing
#ISSP2012esim1.dat %>% tableX(maa,AGE,type = "count") # 12 missing (7 BE, 5 DE)
#ISSP2012esim1.dat %>% tableX(maa,SEX ,type= "count") # 9 missing (BE)
#
# Jos yhdelläkään havainnolla ei puutu tietoja useammasta muuttujasta:
# 400 + 12 + 9 = 421
# 8557- 421 = 8136
#summary(ISSP2012esim1.dat$sp)
ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "cell_perc")

```

maa/Q1b	1	2	3	4	5	Missing	Total
BE	2.26	5.31	5.14	6.47	4.45	2.10	25.73
BG	1.38	4.62	2.40	2.22	0.15	0.96	11.72
DE	1.93	4.39	2.33	6.29	5.15	0.55	20.64
DK	0.82	2.78	1.78	2.71	8.13	0.18	16.40
FI	0.55	2.20	1.74	4.94	3.54	0.71	13.68
HU	2.56	3.37	2.63	2.22	0.88	0.18	11.83
Total	9.49	22.66	16.01	24.86	22.31	4.67	100.00

```

#Apuvälineitä - lisätietoa muuttujista
# kun faktoroidaan V6, niin metadata katoaa?
str(ISSP2012esim1.dat)

```

```

## Classes 'tbl_df', 'tbl' and 'data.frame':   8557 obs. of  9 variables:
## $ C_ALPHAN: chr  "BG" "BG" "BG" "BG" ...
## .. attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## .. attr(*, "format.spss")= chr "A20"
## .. attr(*, "display_width")= int 22
## $ V3 : 'labelled' num  100 100 100 100 100 100 100 100 100 100 ...
## .. attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states"
## .. attr(*, "format.spss")= chr "F5.0"
## .. attr(*, "labels")= Named num  32 36 40 100 124 152 156 158 191 203 ...
## .. ..- attr(*, "names")= chr  "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bulgaria" ...
## $ V4 : 'labelled' num  100 100 100 100 100 100 100 100 100 100 ...
## .. attr(*, "label")= chr "Country ISO 3166 Code (see V3 for codes for the sample)"
## .. attr(*, "format.spss")= chr "F3.0"
## .. attr(*, "labels")= Named num  32 36 40 56 100 124 152 156 158 191 ...
## .. ..- attr(*, "names")= chr  "AR-Argentina" "AU-Australia" "AT-Austria" "BE-Belgium" ...
## $ V6 : 'labelled' num   3 2 3 4 3 3 4 3 2 3 ...
## .. attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## .. attr(*, "format.spss")= chr "F1.0"
## .. attr(*, "labels")= Named num   0 1 2 3 4 5 8 9
## .. ..- attr(*, "names")= chr  "NAP: ES" "Strongly agree" "Agree" "Neither agree nor disagree" ...
## $ SEX : 'labelled' num   2 2 1 2 2 2 1 1 2 1 ...

```

```
##   ..- attr(*, "label")= chr "Sex of Respondent"
##   ..- attr(*, "format.spss")= chr "F1.0"
##   ..- attr(*, "labels")= Named num  1 2 9
##   .. ..- attr(*, "names")= chr  "Male" "Female" "No answer"
## $ AGE      : 'labelled' num  64 43 63 31 52 46 51 40 57 64 ...
##   ..- attr(*, "label")= chr "Age of respondent"
##   ..- attr(*, "format.spss")= chr "F3.0"
##   ..- attr(*, "labels")= Named num  15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr  "15 years" "16 years" "17 years" "18 years" ...
## $ maa      : Factor w/ 6 levels "BE","BG","DE",...: 2 2 2 2 2 2 2 2 2 ...
## $ sp       : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
## $ Q1b      : Factor w/ 5 levels "S","s","?","e",...: 3 2 3 4 3 3 4 3 2 3 ...
## - attr(*, "notes")= chr  "document Plan File: /Users/marcic/Desktop/old/GPS2011 sampling/ISSP2013.s

# typeof(ISSP2012esim1.dat$V6) # what is it?
# class(ISSP2012esim1.dat$V6) # what is it? (sorry)
# storage.mode(ISSP2012esim1.dat$V6) # what is it? (very sorry)
# length(ISSP2012esim1.dat$V6) # how long is it? What about two dimensional objects?
# attributes(ISSP2012esim1.dat$V6) # does it have any metadata?
# str(ISSP2012esim1.dat) #8143 obs. of 8 variables
```

Poistetaan havainnot, joissa puuttuvia tietoja

```
#poistetaan havainnot, joissa puuttuvia tietoja
ISSP2012esim1.dat <- filter(ISSP2012esim1.dat, (!is.na(V6) & !is.na(SEX) & !is.na(AGE)))
#str(ISSP2012esim1.dat)
# 8143 obs. of 6 variables
# muutamalla havainnolla on useampi puuttuva tieto kolmessa muuttujassa (8143-8136 = 7)
```

Taulukot ja kuvat omina koodilohkoina

Frekvenssitaulukko

```
taulu2 <- ISSP2012esim1.dat %>% tableX(maa, Q1b, type = "count")
knitr::kable(taulu2,digits = 2, booktabs = TRUE,
              caption = "Kysymyksen Q1b vastaukset maittain")
```

Taulukko 6: Kysymyksen Q1b vastaukset maittain

	S	s	?	e	E	Total
BE	191	451	438	552	381	2013
BG	118	395	205	190	13	921
DE	165	375	198	538	438	1714
DK	70	238	152	232	696	1388
FI	47	188	149	423	303	1110
HU	219	288	225	190	75	997
Total	810	1935	1367	2125	1906	8143

Riviprosentit

```
taulu3 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "row_perc")

knitr::kable(taulu3,digits = 2, booktabs = TRUE,
              caption = "Kysymyksen Q1b vastaukset, riviprosentit")
```

Taulukko 7: Kysymyksen Q1b vastaukset, riviprosentit

	S	s	?	e	E	Total
BE	9.49	22.40	21.76	27.42	18.93	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
DE	9.63	21.88	11.55	31.39	25.55	100.00
DK	5.04	17.15	10.95	16.71	50.14	100.00
FI	4.23	16.94	13.42	38.11	27.30	100.00
HU	21.97	28.89	22.57	19.06	7.52	100.00
All	9.95	23.76	16.79	26.10	23.41	100.00

Sarakeprosentit

```
taulu4 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "col_perc")

knitr::kable(taulu4,digits = 2, booktabs = TRUE,
  caption = "Kysymyksen Q1b vastaukset, sarakeprosentit")
```

Taulukko 8: Kysymyksen Q1b vastaukset, sarakeprosentit

	S	s	?	e	E	All
BE	23.58	23.31	32.04	25.98	19.99	24.72
BG	14.57	20.41	15.00	8.94	0.68	11.31
DE	20.37	19.38	14.48	25.32	22.98	21.05
DK	8.64	12.30	11.12	10.92	36.52	17.05
FI	5.80	9.72	10.90	19.91	15.90	13.63
HU	27.04	14.88	16.46	8.94	3.93	12.24
Total	100.00	100.00	100.00	100.00	100.00	100.00

Taulukoissa on kuuden maan vastausten jakauma kysymykseen “Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä”. Taulukko on pieni, mutta havaintoja 8143. Alemman suhteellisten frekvenssien taulukon rivejä voi verrata toisiinsa ja alimpaan (“Total”) keskimääräiseen riviin, sarakemuuttujien eli vastausvaihtoehtojen reunajakaumaan. Vastavasti sarakkeita voi verrata rivimuuttujien reunajakaumasarakkeeseen (“Total2”). Eniten vastaajia on Belgiasta (25 %) ja Saksasta (21 %), vähiten Unkarista (12 %).

EDIT: Pienenkin taulukon pyörittely johdattelee hyvin, mihin korrespondenssianalyysiä tarvitaan. Näistä riippuvuuden rakenteet näkee ilmankin, jos on tarpeeksi nokkela. Muiden pitää käyttää CA:ta.

```
simpleCA1 <- ca(~maa + Q1b,ISSP2012esim1.dat)
#tämä ajetaan jotta saadaan hieno kuva piirrettyä
```

edit: Riviprofiileista tarvitaan myös kuva, mutta hiotaan myöhemmin (13.5.2018) **zxy** Onko tämä kuva tallennettava kuvatiedostoksi, vai onnistuuko sen tuottaminen Bookdownissa. Ei taida onnistua? (4.9.18)

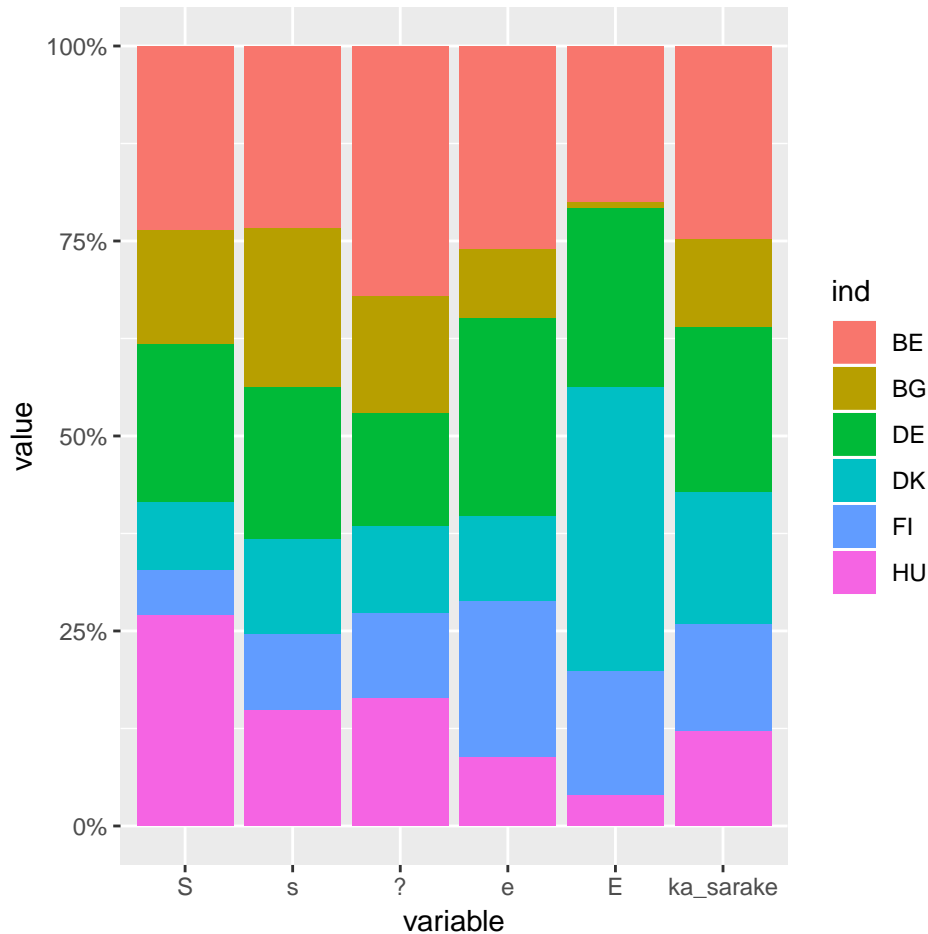
```
#mutkikas kuvan piirto - sarakeprofiilit vertailussa
#ggplot vaatii df-rakenteen ja 'long data' - muotoon
##https://stackoverflow.com/questions/9563368/create-stacked-barplot-where-each-stack-is-scaled-to-sum-
#
# käytetään ca - tuloksia
apu1 <- (simpleCA1$N)
colnames(apu1) <- c("S", "s", "?", "e", "E")
rownames(apu1) <- c("BE", "BG", "DE", "DK", "FI", "HU")
```

```

apu1_df <- as.data.frame(apu1)
#lasketan rivien reunaajakauma
apu1_df$ka_sarake <- rowSums(apu1_df)
#muokataan 'long data' - muotoon
apu1b_df <- melt(cbind(apu1_df, ind = rownames(apu1_df)), id.vars = c('ind'))

ggplot(apu1b_df, aes(x = variable, y = value, fill = ind)) +
  geom_bar(position = "fill", stat = "identity") +
  scale_y_continuous(labels = percent_format())

```



```
#apu1b_df
```

zxy Massat saa mukaan vaikka viittaamalla frekvenssitauluun (4.9.2018)

Riviprofilikuva toimii, mutta vaatii vielä viilausta (18.9.2018)

```

# riviprofiilit ja keskiarvorivi - 18.9.2018
apu2_df <- as.data.frame(apu1)
apu2_df <- rbind(apu2_df, ka_rivi = colSums(apu2_df))

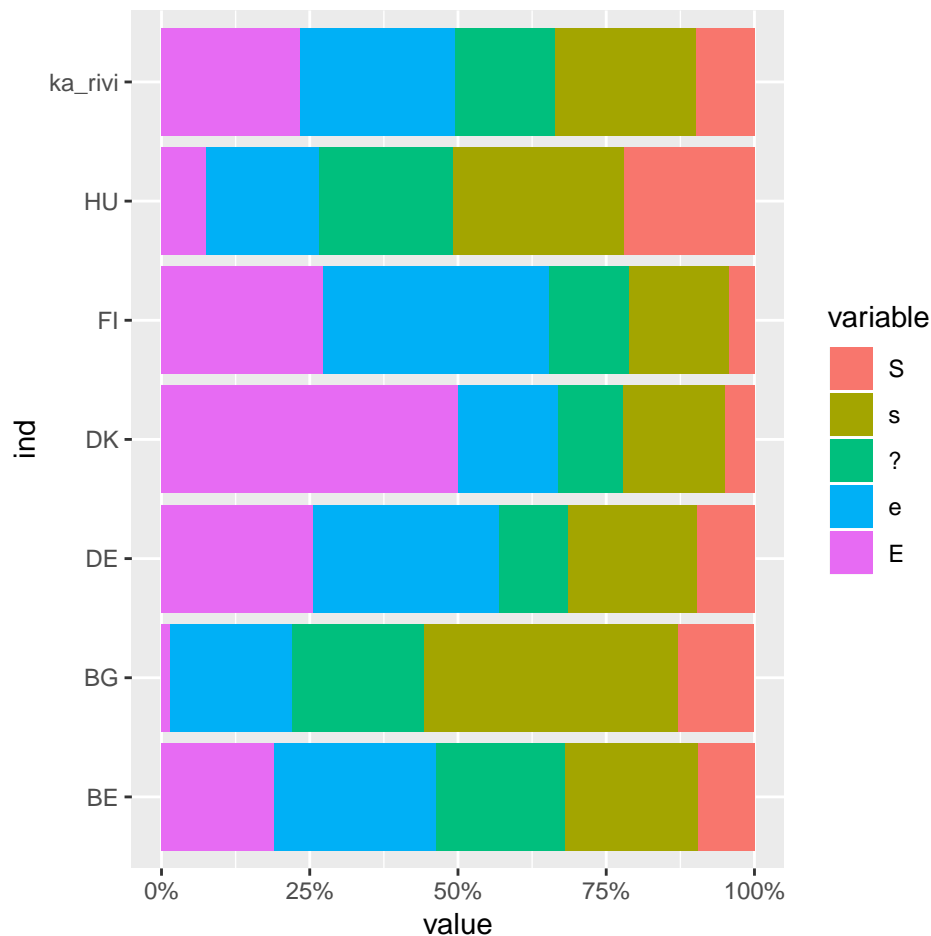
#apu2_df

#str(apu2_df)
## typeof(apu2_df) # what is it?

```

```
## class(apu2_df) # what is it? (sorry)
## storage.mode(apu2_df) # what is it? (very sorry)
## length(apu2_df) # how long is it? What about two dimensional
## objects?
# attributes(apu2_df)

# temp1 <- cbind(apu2_df, ind = rownames(apu2_df))
# temp1
##muokataan 'long data' - muotoon
apu2b_df <- melt(cbind(apu2_df, ind = rownames(apu2_df)), id.vars = c('ind'))
#apu2b_df
#
#
#ggplot(apu2b_df, aes(x = value, y = ind, fill = variable)) +
#  geom_bar(position = "fill", stat = "identity") +
#  #coord_flip() +
#  scale_x_continuous(labels = percent_format())
#versio2 # perkele, tämä toimii! 18.9.2018
ggplot(apu2b_df, aes(x = ind, y = value, fill = variable)) +
  geom_bar(position = "fill", stat = "identity") +
  coord_flip() +
  scale_y_continuous(labels = percent_format())
```



Graafinen analyysi ja R

Käytännön neuvoja data-analyysiin, kuulunee tekstiin, vai meneekö “ohjelmistoympäristö” -liitteeseen? Tärkeä juttu!

Kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-muodossa grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-ajurilla, jos lopulliseen versioon joutuu näin tekemään (13.5.2018). Tämä voi olla järkevä tapa analyysivaiheessa? Teksti kopsattu alla olevasta koodilohkosta.

Ensimmäinen korrespondenssianalyysi - kokeiluja kuvasuhteen säätämiseksi output-dokumentissa. RStudiassa voi avata komentokehoitteessa grafiikka-ikkunan. Siitä käsin tallennettu pdf-kuva on ladattu alla Rmarkdownin omalla komennolla, kohdistus keskelle. Parhaiten näyttäisi toimivan knitrin funktio, mutta oletuskuvakoolla saa ca-kuvasta näköjään aika lähelle oikeanlaisen ilman mitään temppuja.

zxy Selventäisikö vielä khii2-etäisyyksien taulukko, tai ehkä seuraavassa luvussa? **#V** MG&Blasius, “vihreän kirja”, johdanto.

```
# khii2 - etäisyyksien taulukko
#str(simpleCA1)
#simpleCA1$rowdist
#str(simpleCA1$rowdist)
#tablRowDist <- simpleCA1$rowdist
#rownames(tablRowDist) <- simpleCA1$rownames
simpleCA1$rowdist
```

```
## [1] 0.1579735 0.6309909 0.1750128 0.6340627 0.3477331 0.5504040
```

```
simpleCA1$coldist
```

```
## [1] 0.5246525 0.3248840 0.3078230 0.2721699 0.6271108
```

Rivien ja sarakkeiden khii2 - etäisyydet, siistimpi taulukko jos tarpeen (11.10.18)

Lähtökohta: suhteelliset frekvenssit (korrespondenssimatriisi P)

```
taulu5 <- ISSP2012esim1.dat %>% tableX(maa,Q1b,type = "cell_perc")
knitr::kable(taulu5,digits = 2, booktabs = TRUE,
             caption = "Kysymyksen V6 vastaukset maittain (%)")
```

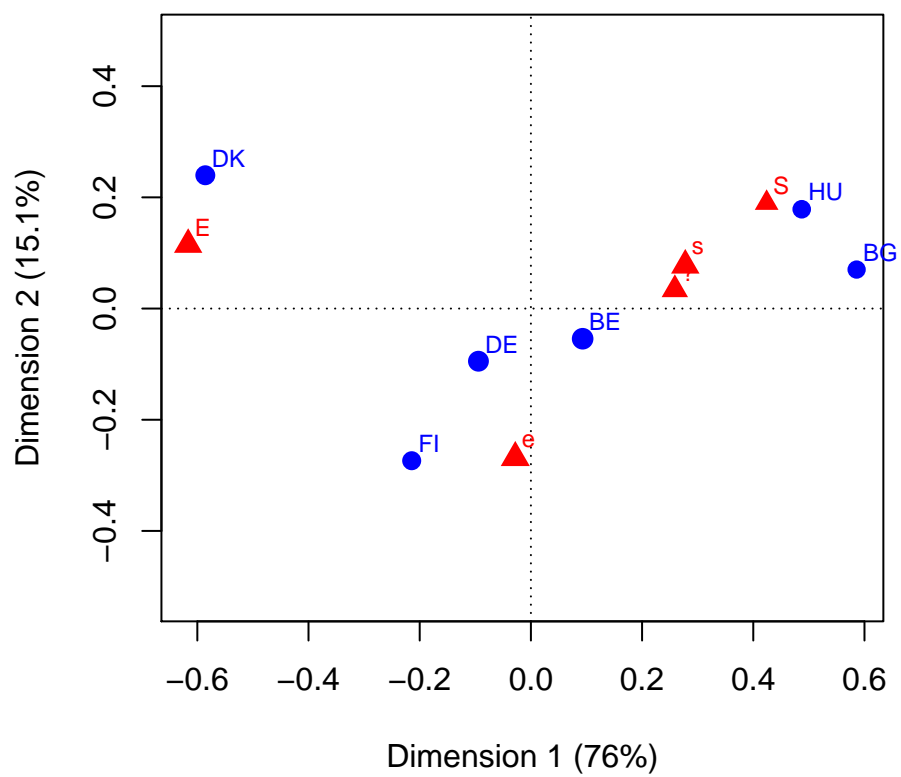
Taulukko 9: Kysymyksen V6 vastaukset maittain (%)

	S	s	?	e	E	Total
BE	2.35	5.54	5.38	6.78	4.68	24.72
BG	1.45	4.85	2.52	2.33	0.16	11.31
DE	2.03	4.61	2.43	6.61	5.38	21.05
DK	0.86	2.92	1.87	2.85	8.55	17.05
FI	0.58	2.31	1.83	5.19	3.72	13.63
HU	2.69	3.54	2.76	2.33	0.92	12.24
Total	9.95	23.76	16.79	26.10	23.41	100.00

zxy Tätä ensimmäistä kuvaa on muistiinpanoissa kommentoitu (löytyy printattuna)

```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuva, tuloksia tarvitaan #sinä.
#symmetrinen kartta
```

```
plot(simpleCA1, map = "symmetric", mass = c(TRUE,TRUE))
```



Kuva 2: V6: lapsi kärsii jos äiti on töissä

```
#str(simpleCA1)
# 13.5.2018
# kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-muodossa
# grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-ajurilla, jos
# näin tekemään.
# näitä kokeiltiin chunk-optioissa mutta ei toimineet (out.width = "6", out.height = "6") (13.5.2018), vaan
# pandoc failed with error 43
#
```

Ja toinen tapa - kuvatiedoston lataaminen include_graphics - funktiolla. Ei esitetä tässä. Nämä toiminevat vain pdf-tulostuksessa?

2.2 Korrespondenssianalyysin käsitteet

1. Profilit
2. Massat
3. Profilien etäisyydet (khii2)

zxy Ja tätä “triplettii” täydentää neljä siitä johdettua käsitettä, viite muistiinpanoissa. **#V** Tässäkin CAiP ja MG2017HY-luentokalvot.

3 Tulinnan perusteita

Luvussa syvennetään esimerkin tulinnan perusteita. Miksi symmetrinen kartta on yleensä paras vaihtoehto, siksi se oletusarvoisesti esitetäänkin. Milloin voi käyttää vaihtoehtoisia esitystapoja? **Ydinluku.**

Esimerkkiaineistossa tulee jo pohdittavaa, Guttman (arc, horseshoe) - efekti, ratkaisun dimensiot jne.

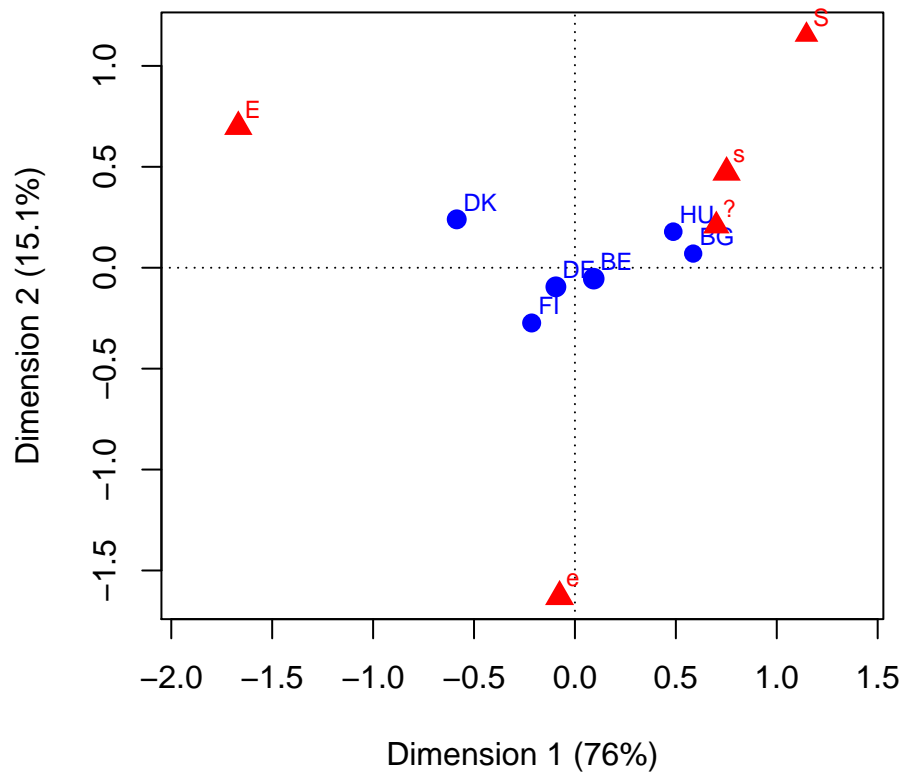
Asymmetrinen kartta, jossa riviprofiilit ovat pääkomponentti-koordinaateissa ja sarakeprofiilit standardikoordinaateissa.

- (1) Sarakkeet ideaalipisteinä, edustavat kuvittellisia maita joissa kaikki ovat vastanneet vain yhdellä tavalla.
- (2) Sarakepisteet kaukana origosta, koska skaalattu
- (3) Rivipisteet kasautuneet keskiarvopisteen ympärille
- (4) Rivi- ja sarakepisteiden suhteelliset sijannit samat kuin symmetrisessä kuvassa
- (5) Tässäkin kuvassa pisteen koko kuvaa sen massaa. Sarakkeista “täysin samaa mieltä” (ts) ja “ei samaa eikä eri mieltä” ovat massoiltaan pienimmät.
- (6) Pisteiden koko kuvaa rivin tai sarakkeen massaa.

```
# asymmetrinen kartta - rivit pc ja sarakkeet sc
# HUOM! simpleCA1 luodaan G1_2_johdesim.Rmd - tiedostossa

plot(simpleCA1, map = "rowprincipal",
     mass = c(TRUE,TRUE),
     main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta" )
```


Lapsi kärsii jos äiti on töissä –asymmetrinen kartta



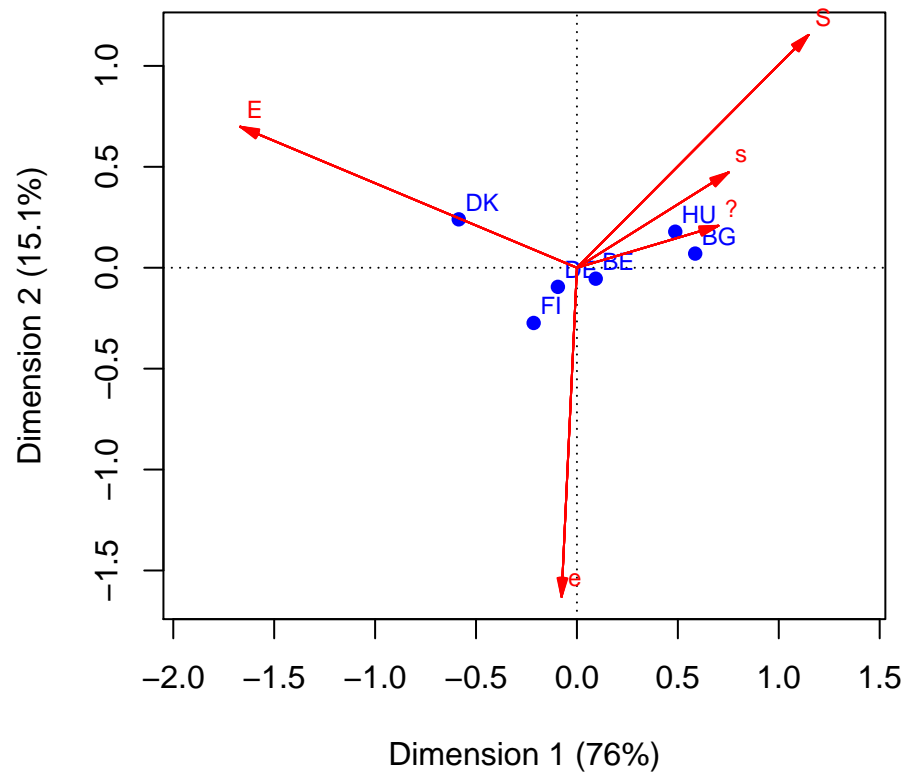
Tarinaa voi tarvittaessa jatkaa, tämä on CA:n hankalin asia. Kaksi koordinaatistoa, ja niiden yhteys.

(7) Asymmetrinen kuva ja akseleiden / dimensioiden tulkinta

Piirretään sama asymmetrinen kartta uudelleen, mutta yhdistetään sarakepisteet keskiarvopisteeseen (sentroidiin) suorilla. Mitä terävämpi on sarakesuoran (vektorin?) ja akselin kulma, sitä enemmän sarake määrittää tätä ulottuvuutta. Jos vektori on lähettä 45 asteen kulmaa, sarake määrittää yhtä paljon molempia ulottuvuuksia.

```
# asymmetrinen kartta - rivit pc ja sarakkeet sc
# sarakkeet vektorikuvina
# HUOM! simpleCA1 luodaan G1_2_johdesim.Rmd - tiedostossa
plot(simpleCA1, map = "rowprincipal",
     arrows = c(FALSE, TRUE),
     main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta 1" )
```

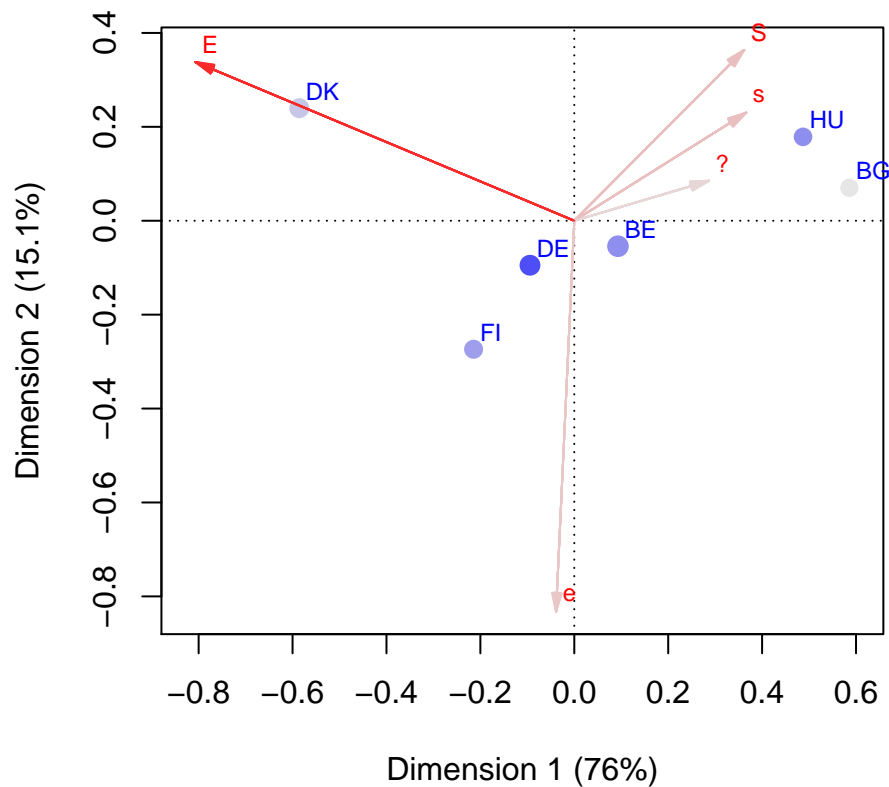
Lapsi kärsii jos äiti on töissä –asymmetrinen kartta



Tärkein havainto on sarakkeen “Eri mieltä” (e) ja toisen ulottuvuuden yhteys. Myös sarake “täysin samaa mieltä” (ts) määrittää toista ulottuvuutta lähes yhtä paljon kuin ensimmäistä.

```
#X11() komentoriville ja plot-komento
plot(simpleCA1, map = "rowgreen",
     contrib= c("absolute", "absolute"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
     main = "Lapsi kärsii jos äiti on töissä - asymmetrinen kartta 2" )
```

Lapsi kärsii jos äiti on töissä – asymmetrinen kartta



Greenacre (2006, “loose ends -artikkeli”) ehdotti asymmetrisessä kuvassa standardikoordinaattien skaalaamista niin, että ne kerrotaan massan neliöjuurella. Tämä skaalaus toimii hyvin pienen ja suuren inertian tapauksessa. Kartoissa pätee sama sääntö kuin muussakin graafisessa data-analyysissä, kuvien on esitettävä oleelliset yhteydet, mutta mielellään vain ne.

Tulkinta: rivipisteiden ortogonaalinen projektio “sarakevektorille”

Asymmetrisessä kartassa 2 pisteiden koko on suhteessa niiden massaan, ja värisävy absoluuttiseen kontribuutioon (voi olla myös suhteellinen kontribuutio).

```
# CA:n numeeriset tulokset
summary(simpleCA1)
```

```
##
## Principal inertias (eigenvalues):
##
## dim   value      %   cum%   scree plot
## 1     0.136619  76.0  76.0   *****
## 2     0.027089  15.1  91.1   ****
## 3     0.010054   5.6  96.7   *
## 4     0.005988   3.3 100.0   *
## -----
## Total: 0.179751 100.0
##
##
```

```
## Rows:
##      name  mass  qlt  inr      k=1 cor ctr      k=2 cor ctr
## 1 |   BE |   247  465   34 |   93 347  16 |  -54 118  27 |
## 2 |   BG |   113  874  251 |  586 862 284 |   70  12  21 |
## 3 |   DE |   210  584   36 |  -94 291  14 |  -95 293  70 |
## 4 |   DK |   170  996  381 | -586 853 428 |  240 143 362 |
## 5 |   FI |   136 1000   92 | -214 380  46 | -274 620 377 |
## 6 |   HU |   122  889  206 |  487 783 213 |  179 105 144 |
##
## Columns:
##      name  mass  qlt  inr      k=1 cor ctr      k=2 cor ctr
## 1 |    S |    99  784  152 |  424 653 131 |  190 131 132 |
## 2 |    s |   238  788  140 |  278 731 134 |   78  57  53 |
## 3 |      |   168  720   88 |  259 707  82 |   34  12   7 |
## 4 |    e |   261  982  108 |  -28  11   2 | -268 971 693 |
## 5 |    E |   234 1000  512 | -616 966 651 |  115  34 114 |
```

zxy Taulukon käsitteiden läpikäynti ja pureskelu kuulunee seuraavaan lukuun.

4 Yksinkertaisen korrespondenssianalyysin laajennuksia 1

Korrespondenssianalyysi sallii rivien tai sarakkeiden yhdistelyn tai “jakamisen”. Tämä onnistuu esimerkkiaineistossa lisäämällä rivejä eli jakamalla eri maiden vastauksia useampaan ryhmään.

Sen avulla voi myös tarkastella ja vertailla erilaisia ryhmien välisiä tai ryhmien sisäisiä (within groups - between groups) eroja hieman. Teknisesti yksinkertaista korrespondenssianalyysiä sovelletaan muokattuun matriisiin. Datamatriisi rakennetaan useammasta alimatriisista, joko “pinoamalla” osamatriiseja (stacked matrices) tai muodostamalla symmetrinen lohkomatriisi (ABBA).

Lisätään esimerkkitietoihin uusia muuttujia, vastaajan luokitelut ikä ja sukupuoli.

**** EDIT: **** Koitetaan aina pitää alkuperäinen data mahdollisimman “lähellä”, luodaan siis kaikki uudestaan. Tarketeena .data jos koko aineisto ja .dat jos rajattu. Aineisto laajennetaan myöhemmin?

Toinen pulma: milloin laajennetaan dataa useampaan maahan?

```
# Saksan ja Belgian aluejako - täydentävät pisteet

ISSP2012esim1.data <- read_spss("data/ZA5900_v4-0-0.sav") # Alkuperäinen data, ( user_na = TRUE pois 25

#str(ISSP2012esim1.data)
#61754 obs. of  420 variables
#
# KUUSI MAATA

incl_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU)

ISSP2012esim1.dat <- filter(ISSP2012esim1.data, V4 %in% incl_esim1)

#str(ISSP2012esim1.dat) #8557 obs. of  420 variables
#
# mukaan muuttujat, V3 jos halutaan jakaa Saksa ja Belgia
# SEX 1=male, 2=female AGE haastateltava ikä haastatteluhetkellä
```

```

# MUUTTUJAT

ISSP2012esim1.dat <- select(ISSP2012esim1.dat, C_ALPHAN, V3,V4, V6, SEX, AGE)

#str(ISSP2012esim1.dat) #8557 obs. of 6 variables
#
# Poistetaan havainnot, joissa puuttuvia tietoja

ISSP2012esim1.dat <- filter(ISSP2012esim1.dat, (!is.na(V6) & !is.na(SEX) & !is.na(AGE)))

#str(ISSP2012esim1.dat) #8143 havaintoa, 6 muuttujaa
#8557-8143 = 414 havaintoa vähemmän

# sp (sukupuoli) m = 1, f = 2
sp_labels <- c("m","f")
#
# vastausvaihtoehdot
#
# 1 = täysin samaa mieltä, 2 = samaa mieltä, 3 = ei samaa eikä eri, 4 = eri mieltä, 5 = täysin eri miel
vastaus_labels <- c("S","s","?", "e","E")

# Faktoreiksi - onko ihan oikein? On(26.9.18) - faktoroitu uudeksi muuttujaksi, vanhassa säilyvät metat

ISSP2012esim1.dat$maa <- factor(ISSP2012esim1.dat$C_ALPHAN)
ISSP2012esim1.dat$sp <- factor(ISSP2012esim1.dat$SEX, labels = sp_labels) #pitäisikö lisätä levels?
ISSP2012esim1.dat$Q1b <- factor(ISSP2012esim1.dat$V6, labels = vastaus_labels) #pitäisikö lisätä levels
#str(ISSP2012esim1.dat)
#str(ISSP2012esim1.dat$sp)
#summary(ISSP2012esim1.dat)
#ISSP2012esim1.dat %>% tableX(sp, V6, type = "row_perc")

```

EDIT: Uudet muuttujat omassa koodilohkossa pätkänä

```

# 23.5.2018 maa2 - muuttuja
# ISO 3166 Code kansallisvaltiolle muuttujassa V4
#
# ISO 3166 Code V3 - maiden jaot
# 5601 BE-FLA-Belgium/ Flanders
# 5602 BE-WAL-Belgium/ Wallonia
# 5603 BE-BRU-Belgium/ Brussels
# 27601 DE-W-Germany-West
# 27602 DE-E-Germany-East
ISSP2012esim1.dat$maa2 <- factor(ISSP2012esim1.dat$V3,
                                levels = c("100","208","246","348","5601","5602","5603","27601","27602"),
                                labels = c("BG","DK","FI","HU","bF","bW","bB","dW","dE"))

#head(ISSP2012esim1.dat)
#str(ISSP2012esim1.dat$maa2)
#taulu41 <- ISSP2012esim1.dat %>% tableX(maa,maa2,type = "count") # Tarkistus maa2-muuttujalle
#kable(taulu41,digits = 2, caption = "Uusi muuttuja maa2: Belgian ja Saksan ositus")

```

zxy Edellä pelkkä tarkistus, tuloksen voi kopsata koodilohkoon kun homma on hoidettu.

4.1 Täydentävät muuttujat (supplementary points)

zxy Piste sinne piirretään, mutta muuttujassa on se tieto. "Täydentävät piste" kuulostaa huonolta. Lisämuuttujat, havainnot?

Ref:CAip ss 89, HY2017_MCA

Aineistossa on havaintoja (rivejä) tai muuttujia (sarakkeita), joista voi olla hyötyä tulosten tulkinnassa. Nämä lisäpisteet voidaan sijoittaa kartalle, jos niitä voidaan jotenkin järkevästi vertailla kartan luomisessa käytettyihin profileihin (riveihin ja sarakkeisiin).

EDIT Lisätään Belgian ja Saksan aluejako täydentäviksi riveiksi. Sopii tarinaan, dimensioiden tulkinta ei ollut esimerkissä kovin kirkas. Viite CAip:n lukuun, jossa vain todetaan että maita ei ole järkevää painottaa (massa) otoskoolla, vaan vakioidaan (jotenkin) sama (suhteellinen) massa kaikille. Samalla oikaistaan myös naisten yliedustus aineistossa.

Active point, aktiivinen piste (aktiivinen havainto tai muuttuja).

Täydentävä piste (täydentävä havainto).

Täydentävien muuttujien kolme käyttötapaa:

- sisällöllisesti tutkimusongelman kannalta poikkeava tai erilainen rivi tai sarake
- outlierit, poikkeava havainto jolla pieni massa (esimerkissä uusi sarakemuuttuja, jossa kovin vähän havaintoja)
- osaryhmät **EDIT** capaper- jäsentelyssä ja bookdown-dokumentissa selitetetty täydentävät/lisäpisteet tarkemmin (18.9.2018).

```
# Kömpelöä koodia, harjoitellaan taulukoiden yhdistelyä (CAtest1.Rmd)
# Belgian ja Saksan jako lisäpisteinä 24.5.2018
#head(ISSP2012esim1.dat)
```

```
# HUOM! Tässä ei vielä supp.points mukana!
suppointCA1 <- ca(~maa2 + Q1b,ISSP2012esim1.dat)
#plot(suppointCA1, main = "Belgian ja Saksan ositteet")
#kuva kääntyy ympäri, kerrotaan koordinaattivektorit luvulla -1
#summary(suppointCA1)
#print(suppointCA1)
#str(suppointCA1)
#
# Käännetään kuva
```

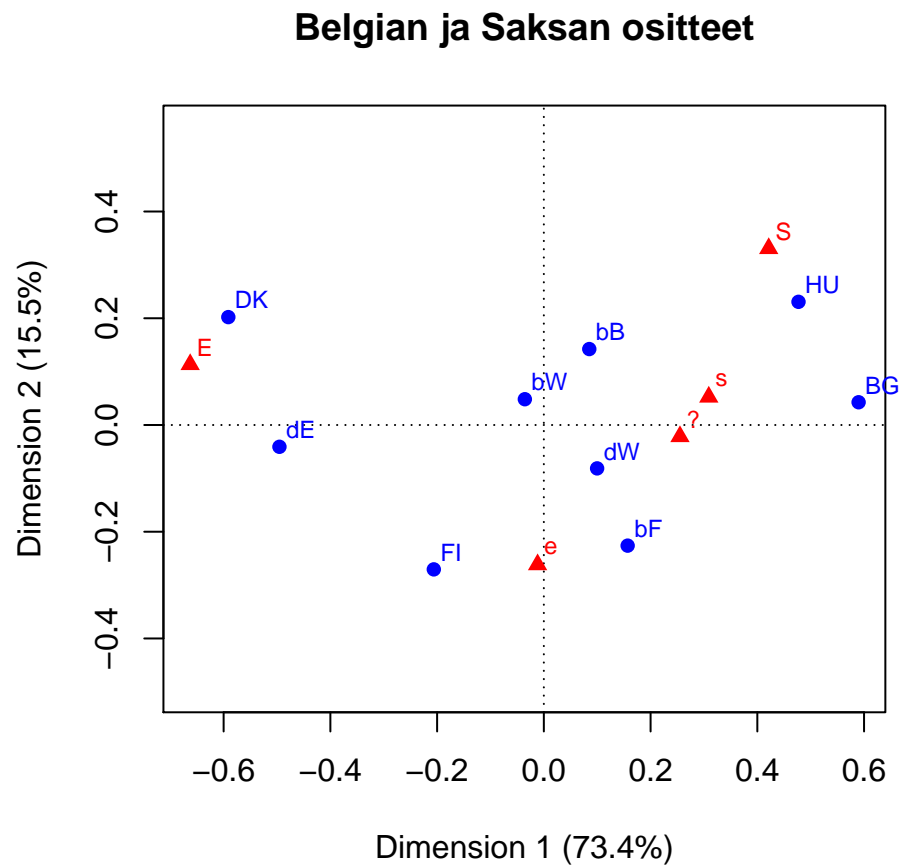
```
suppointCA1b <- suppointCA1
suppointCA1b$rowcoord <- suppointCA1b$rowcoord[,] * (-1)
suppointCA1b$colcoord <- suppointCA1b$colcoord[,] * (-1)
suppointCA1b$rowcoord
```

	Dim1	Dim2	Dim3	Dim4
BG	1.5024575	0.2364976	-1.5646535	1.2274009
DK	-1.5060223	1.1214678	-0.8891868	0.1996764
FI	-0.5252216	-1.5009862	0.5841156	0.1935193
HU	1.2154623	1.2803425	0.9947716	-0.9386679
bF	0.4000647	-1.2540425	-1.1182121	-1.6025782
bW	-0.0906315	0.2679979	0.0761877	-0.7901000
bB	0.2169124	0.7893585	1.3697862	-0.5617393
dW	0.2543232	-0.4511235	0.8757353	1.5124903
dE	-1.2620072	-0.2265947	0.7448562	-0.2844804

```
suppointCA1b$colcoord
```

	Dim1	Dim2	Dim3	Dim4
S	1.0733103	1.8351327	2.1160478	-0.2360525
s	0.7872571	0.2909285	-0.9861563	1.2374779
?	0.6497888	-0.1199336	-0.9123790	-1.9203632
e	-0.0298593	-1.4515479	0.8247769	0.2094281
E	-1.6881081	0.6291103	-0.1632819	-0.0121801

```
plot(suppointCA1b, main = "Belgian ja Saksan ositteet")
```



```
# Miten lisärivit? (24.5.2018)
# Luetaan data tauluksi - ei toimi, char-table
# yritetään uudestaan table-funktiolla
# data maa2-muuttujalla
suppoint1_df1 <- select(ISSP2012esim1.dat, maa2,Q1b)
#str(suppoint1_df1)
#head(suppoint1_df1)
suppoint1_tab1 <- table(suppoint1_df1$maa2, suppoint1_df1$Q1b)
suppoint1_tab1
```

/	S	s	?	e	E
BG	118	395	205	190	13
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75
bF	51	241	262	312	146
bW	53	103	91	118	125
bB	87	107	85	122	110
dW	133	313	138	375	208
dE	32	62	60	163	230

```

#plot(ca(~maa2 + V6, suppoint1_df1)) #toimii
#
# Saksan ja Belgian summarivit
#
suppoint2_df <- filter(ISSP2012esim1.dat, (maa == "BE" | maa == "DE"))
suppoint2_df <- select(suppoint2_df, maa, Q1b)
#head(suppoint2_df)
#tail(suppoint2_df)
#str(suppoint2_df)
#suppoint2_df
suppoint2_tab1 <- table(suppoint2_df$maa, suppoint2_df$Q1b)
#suppoint2_tab1
suppoint2_tab1 <- suppoint2_tab1[-2,]
# kömpelösti kolme kertaa
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]
suppoint2_tab1 <- suppoint2_tab1[-3,]
#suppoint2_tab1

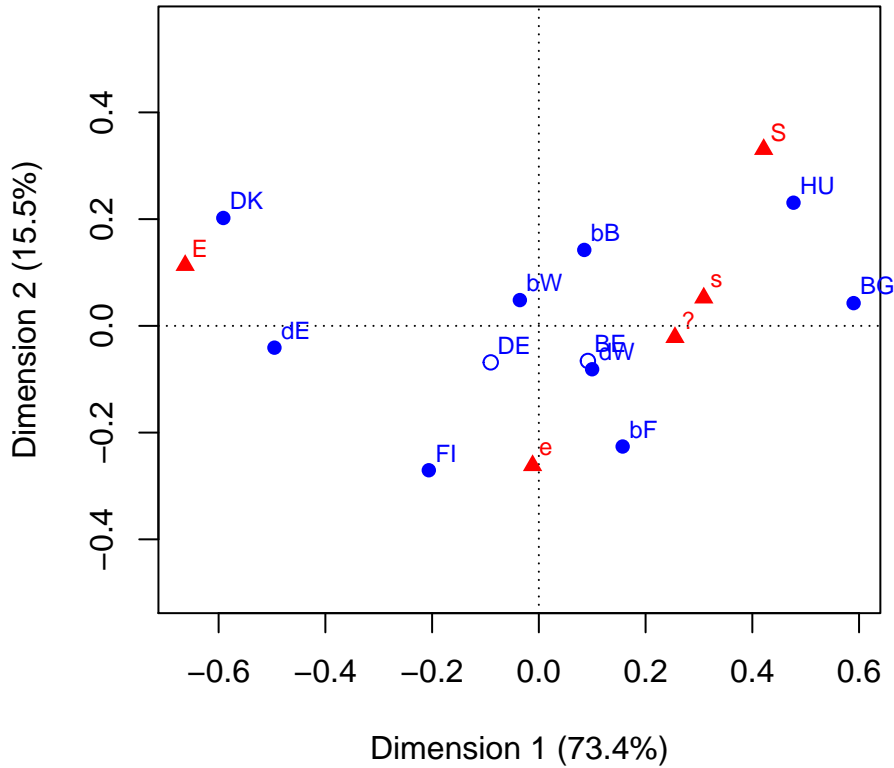
#lisätään rivit maa2-muuttujan taulukkoon

suppoint1_tab1 <- rbind(suppoint1_tab1, suppoint2_tab1)
#suppoint1_tab1
suppointCA2 <- ca(suppoint1_tab1[,1:5], suprow = 10:11)
#käännetään kuva
suppointCA2b <- suppointCA2
suppointCA2b$rowcoord <- suppointCA2b$rowcoord[,] * (-1)
suppointCA2b$colcoord <- suppointCA2b$colcoord[,] * (-1)

plot(suppointCA2b, main = "Passiiviset pisteet DE ja BE" )

```


Passiiviset pisteet DE ja BE



```
# ca- output
#names(suppoincA2b)
#str(suppoincA2b)
#str(suppoincA2b$rowcoord)
#suppoincA2b
#suppoincA2b$rowcoord
#apply(suppoincA2b$rowcoord, 2, sum)
#suppoincA2b$rowdist
#suppoincA2b$colldist
summary(suppoincA2b)
```

```
##
## Principal inertias (eigenvalues):
##
## dim      value      %      cum%      scree plot
## 1         0.154101   73.4   73.4   *****
## 2         0.032489   15.5   88.9   ****
## 3         0.014294    6.8   95.7   **
## 4         0.008944    4.3  100.0   *
##
## -----
## Total: 0.209828 100.0
##
##
```

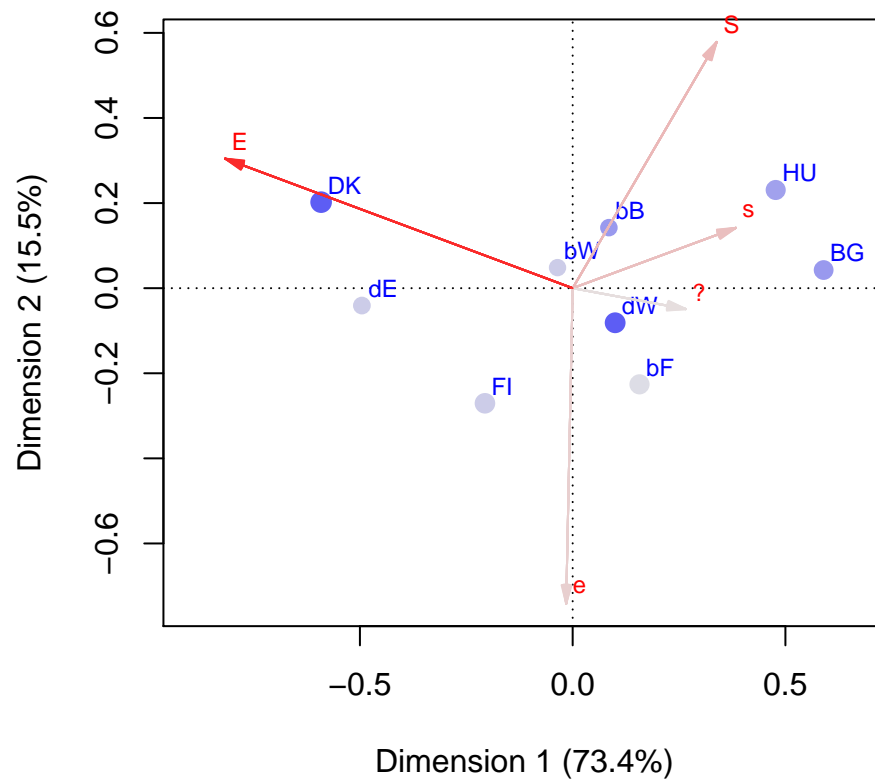
```
## Rows:
##      name  mass  qlt  inr    k=1 cor  ctr    k=2 cor  ctr
## 1 |   BG |  113  878  215 |  590 874  255 |   43  5   6 |
## 2 |   DK |  170  971  327 | -591 869  387 |  202 102  214 |
## 3 |   FI |  136  957   79 | -206 352   38 | -271 605  307 |
## 4 |   HU |  122  927  177 |  477 751  181 |  231 176  201 |
## 5 |   bF |  124  650   69 |  157 212   20 | -226 438  195 |
## 6 |   bW |   60  388    3 |  -36 137    0 |   48 252    4 |
## 7 |   bB |   63  481   17 |   85 127    3 |  142 354   39 |
## 8 |   dW |  143  345   33 |  100 208    9 |  -81 138   29 |
## 9 |   dE |   67  966   82 | -495 960  107 |  -41  7    3 |
## 10 | (*)BE | <NA>  512 <NA> |   92 338 <NA> | -66 173 <NA> |
## 11 | (*)DE | <NA>  418 <NA> |  -90 265 <NA> | -68 153 <NA> |
##
## Columns:
##      name  mass  qlt  inr    k=1 cor  ctr    k=2 cor  ctr
## 1 |   S |   99  816  167 |  421 505 115 |  331 311 335 |
## 2 |   s |  238  781  143 |  309 759 147 |   52  22  20 |
## 3 |   |  168  594   88 |  255 589  71 |  -22   4   2 |
## 4 |   e |  261  871   98 |  -12   2   0 | -262 870 550 |
## 5 |   E |  234  999  505 | -663 971 667 |  113  28  93 |
```

Saksan ja Belgian summarivit ovat ositteiden painotettuja keskiarvoja (sentroideja), läntisen ja itäisen Saksan rivipisteiden välisellä janalla on koko maan summapiste DE.

Piirretään vertailun vuoksi vielä asymmettrinen kartta (“kontribuutio-kartta, kontribuutio-kaksoiskuva”).

```
#X11()
plot(suppointCA1b, map = "rowgreen",
     contrib= c("absolute", "absolute"),
     mass = c(TRUE,TRUE),
     arrows = c(FALSE, TRUE),
     main = "Saksan ja Belgian alueet - asymmettrinen kartta 1" )
```

Saksan ja Belgian alueet – asymmetrinen kartta 1



Tulostetaan numeeriset taulukot.

```
summary(suppointCA1b)
```

```
##
## Principal inertias (eigenvalues):
##
## dim    value      %   cum%   scree plot
## 1      0.154101  73.4  73.4   *****
## 2      0.032489  15.5  88.9   ****
## 3      0.014294   6.8  95.7   **
## 4      0.008944   4.3 100.0   *
## -----
## Total: 0.209828 100.0
##
##
## Rows:
##   name  mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
## 1 |  BG |  113  878  215 |  590 874 255 |  43  5  6 |
## 2 |  DK |  170  971  327 | -591 869 387 | 202 102 214 |
## 3 |  FI |  136  957   79 | -206 352  38 | -271 605 307 |
## 4 |  HU |  122  927  177 |  477 751 181 | 231 176 201 |
## 5 | bF |  124  650   69 |  157 212  20 | -226 438 195 |
## 6 | bW |   60  388    3 |  -36 137   0 |  48 252   4 |
```

```
## 7 |   bB |   63 481 17 |   85 127 3 | 142 354 39 |
## 8 |   dW |  143 345 33 |  100 208 9 | -81 138 29 |
## 9 |   dE |   67 966 82 | -495 960 107 | -41 7 3 |
##
## Columns:
##      name  mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 |   S |   99 816 167 |  421 505 115 | 331 311 335 |
## 2 |   s |  238 781 143 |  309 759 147 |  52 22 20 |
## 3 |   |  168 594 88 |  255 589 71 | -22 4 2 |
## 4 |   e |  261 871 98 |  -12 2 0 | -262 870 550 |
## 5 |   E |  234 999 505 | -663 971 667 | 113 28 93 |
```

4.2 Lisämuuttujat: ikäluokka ja sukupuoli

zxy Otsikkoa pitää harkita, CAip - kirjassa tämä on ensimmäinen esimerkki yksinkertaisen CA:n laajennuksesta. Otsikkona on “multiway tables”, ja tästä yhteisvaikutusmuuttujan (interactive coding) luominen on ensimmäinen esimerkki. Menetelmää taivutetaan sen jälkeen moneen suuntaan.

Luodaan luokiteltu ikämuuttua `age_cat`, ja sen avulla iän ja sukupuolen interaktiivimuuttuja `ga`. Maiden välillä on hieman eroja siinä, kuinka nuoria vastaajia on otettu tutkimuksen kohteeksi. Suomessa alaikäraja on 15 vuotta, monessa maassa se on hieman korkeampi. Ikäluokat ovat (1=15-25, 2=26-35, 3=36-45, 4=46-55, 5=56-65, 6=66 tai vanhempi). Vuorovaikutusmuuttuja `ga` koodataan `f1, ..., f6` ja `m1, ..., m6`. Muuttujien nimet kannattaa pitää mahdollisimman lyhyinä.

```
# Iän ja sukupuolen vuorovaikutusmuuttujia 1
#
# Uusi R-data: ISSP2012esim2.dat
#
#age_cat
#AGE 1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 and older
#
#summary(ISSP2012esim1.dat$AGE)
#hist(ISSP2012esim1.dat$AGE)
ISSP2012esim2.dat <- mutate(ISSP2012esim1.dat, age_cat = ifelse(AGE %in% 15:25, "1",
  ifelse(AGE %in% 26:35, "2",
    ifelse(AGE %in% 36:45, "3",
      ifelse(AGE %in% 46:55, "4",
        ifelse(AGE %in% 56:65, "5", "6"))))))
ISSP2012esim2.dat$age_cat <- factor(ISSP2012esim2.dat$age_cat)

#test6 %>% tableX(AGE, age_cat, type = "count") aika iso taulukko, voi tarkistaa että muunnos ok.
taulu42 <- ISSP2012esim2.dat %>% tableX(maa,age_cat,type = "count")
kable(taulu42,digits = 2, caption = "Ikäluokka age_cat")
```

Taulukko 13: Ikäluokka `age_cat`

	1	2	3	4	5	6	Total
BE	208	333	336	375	368	393	2013
BG	77	115	159	148	198	224	921
DE	205	223	274	358	288	366	1714
DK	207	213	245	271	234	218	1388
FI	152	166	165	223	238	166	1110
HU	103	161	198	171	196	168	997
Total	952	1211	1377	1546	1522	1535	8143

1	2	3	4	5	6	Total
---	---	---	---	---	---	-------

```
taulu43 <- ISSP2012esim2.dat %>% tableX(maa,age_cat,type = "cell_perc")
kable(taulu43,digits = 2, caption = "age_cat: suhteelliset frekvenssit")
```

Taulukko 14: age_cat: suhteelliset frekvenssit

	1	2	3	4	5	6	Total
BE	2.55	4.09	4.13	4.61	4.52	4.83	24.72
BG	0.95	1.41	1.95	1.82	2.43	2.75	11.31
DE	2.52	2.74	3.36	4.40	3.54	4.49	21.05
DK	2.54	2.62	3.01	3.33	2.87	2.68	17.05
FI	1.87	2.04	2.03	2.74	2.92	2.04	13.63
HU	1.26	1.98	2.43	2.10	2.41	2.06	12.24
Total	11.69	14.87	16.91	18.99	18.69	18.85	100.00

Ikäjäkauma painottuu kaikissa maissa jonkin verran vanhempiin ikäluokkiin. Nuorempien ikäluokkien osuus on (alle 26-vuotiaan ja alle 26-35 - vuotiaat) varsinkin Bulgariassa (BG) ja Unkarissa (HU) pieni.

zxy Siistimmät versioit muuttujien luonnista (case_when - rakenne) (19.9.2018).

```
# case_when: ikä ja sukupuoli
ISSP2012esim2.dat <- mutate(ISSP2012esim2.dat, ga = case_when((age_cat == "1") & (sp == "m") ~ "m1",
  (age_cat == "2") & (sp == "m") ~ "m2",
  (age_cat == "3") & (sp == "m") ~ "m3",
  (age_cat == "4") & (sp == "m") ~ "m4",
  (age_cat == "5") & (sp == "m") ~ "m5",
  (age_cat == "6") & (sp == "m") ~ "m6",
  (age_cat == "1") & (sp == "f") ~ "f1",
  (age_cat == "2") & (sp == "f") ~ "f2",
  (age_cat == "3") & (sp == "f") ~ "f3",
  (age_cat == "4") & (sp == "f") ~ "f4",
  (age_cat == "4") & (sp == "f") ~ "f4",
  (age_cat == "5") & (sp == "f") ~ "f5",
  (age_cat == "6") & (sp == "f") ~ "f6",
  TRUE ~ "missing"
))

#ISSP2012esim1.dat %>% tableX(ga,ga2) # tarkistus uudelle muuttujan luontikoodille
# muuttujien tarkistuksia 19.9.2018
#str(ISSP2012esim1.dat$ga)
#str(ISSP2012esim1.dat$ga2)
# ga on merkkijono, samoin ga2, pitäisikö muuttaa faktoriksi?
#str(ISSP2012esim1.dat)

#Tulostetaan taulukkoina ga2 - muuttuja.
taulu46 <- ISSP2012esim2.dat %>% tableX(maa,ga,type = "count")
kable(taulu46,digits = 2, caption = "Ikäluokka ja sukupuoli ga2")
```

Taulukko 15: Ikäluokka ja sukupuoli ga2

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	116	198	174	199	186	185	92	135	162	176	182	208	2013

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BG	40	64	94	85	114	149	37	51	65	63	84	75	921
DE	102	120	152	186	135	185	103	103	122	172	153	181	1714
DK	83	110	136	146	128	99	124	103	109	125	106	119	1388
FI	94	95	94	118	142	91	58	71	71	105	96	75	1110
HU	54	86	95	91	94	104	49	75	103	80	102	64	997
Total	489	673	745	825	799	813	463	538	632	721	723	722	8143

```
taulu47 <- ISSP2012esim2.dat %>% tableX(maa,ga,type = "cell_perc")
kable(taulu47,digits = 2, caption = "ga2: suhteelliset frekvenssit")
```

Taulukko 16: ga2: suhteelliset frekvenssit

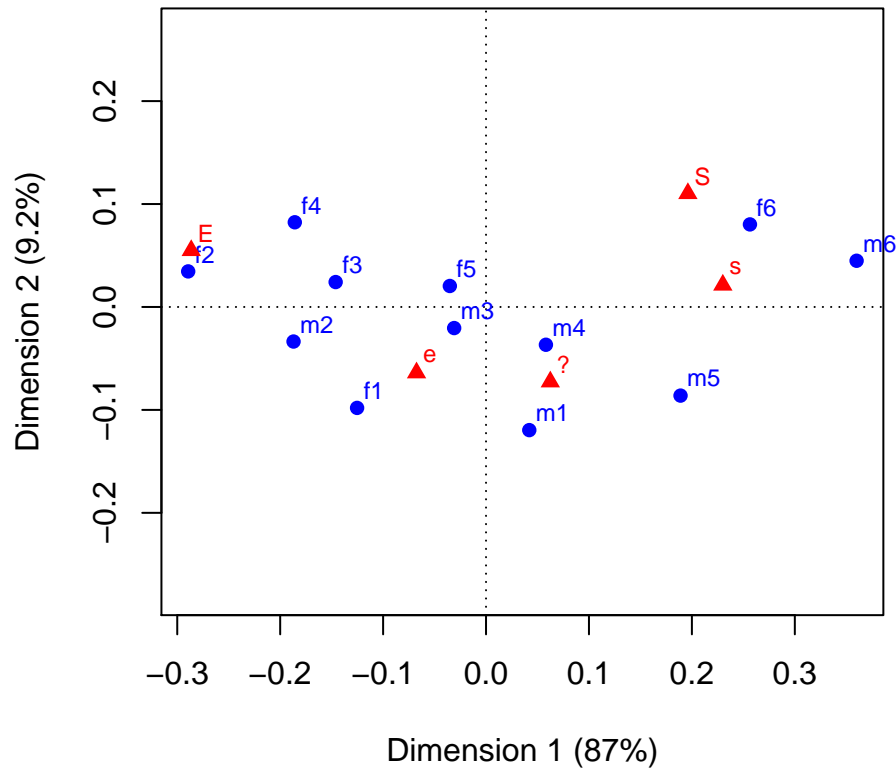
	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	1.42	2.43	2.14	2.44	2.28	2.27	1.13	1.66	1.99	2.16	2.24	2.55	24.72
BG	0.49	0.79	1.15	1.04	1.40	1.83	0.45	0.63	0.80	0.77	1.03	0.92	11.31
DE	1.25	1.47	1.87	2.28	1.66	2.27	1.26	1.26	1.50	2.11	1.88	2.22	21.05
DK	1.02	1.35	1.67	1.79	1.57	1.22	1.52	1.26	1.34	1.54	1.30	1.46	17.05
FI	1.15	1.17	1.15	1.45	1.74	1.12	0.71	0.87	0.87	1.29	1.18	0.92	13.63
HU	0.66	1.06	1.17	1.12	1.15	1.28	0.60	0.92	1.26	0.98	1.25	0.79	12.24
Total	6.01	8.26	9.15	10.13	9.81	9.98	5.69	6.61	7.76	8.85	8.88	8.87	100.00

edit Vain tarkistuksiin, toisen voi poistaa (19.9.2018)!

CAiP, ch16, täällä myös maa- ja sukupuoli- uudelleenpainotus.

```
gaTestCA1 <- ca(~ga + Q1b,ISSP2012esim2.dat)
plot(gaTestCA1, main = "Äiti töissä: ikäluokka ja sukupuoli")
```

Äiti töissä: ikäluokka ja sukupuoli



```
summary(gaTestCA1)
```

```
##
## Principal inertias (eigenvalues):
##
## dim    value      %   cum%   scree plot
## 1      0.037448  87.0  87.0   *****
## 2      0.003977   9.2  96.2   **
## 3      0.001041   2.4  98.6   *
## 4      0.000590   1.4 100.0
## -----
## Total: 0.043055 100.0
##
## Rows:
## name    mass    q1t    inr    k=1 cor ctr    k=2 cor ctr
## 1 | f1 | 60 | 990 | 36 | -125 614 25 | -98 376 145 |
## 2 | f2 | 83 | 997 | 163 | -289 983 185 | 35 14 25 |
## 3 | f3 | 91 | 984 | 47 | -146 958 52 | 24 26 13 |
## 4 | f4 | 101 | 1000 | 97 | -186 836 93 | 82 164 172 |
## 5 | f5 | 98 | 879 | 4 | -35 658 3 | 20 221 10 |
## 6 | f6 | 100 | 951 | 176 | 256 866 175 | 80 85 162 |
## 7 | m1 | 57 | 659 | 32 | 42 72 3 | -120 587 205 |
```

```
## 8 | m2 | 66 977 57 | -187 946 62 | -34 30 19 |
## 9 | m3 | 78 457 5 | -31 318 2 | -20 139 8 |
## 10 | m4 | 89 674 14 | 58 482 8 | -37 192 30 |
## 11 | m5 | 89 988 90 | 189 818 85 | -86 170 166 |
## 12 | m6 | 89 978 277 | 360 963 307 | 45 15 45 |
##
## Columns:
##      name    mass  qlt  inr      k=1 cor ctr      k=2 cor ctr
## 1 | S | 99 915 128 | 196 695 102 | 110 220 304 |
## 2 | s | 238 969 304 | 230 961 336 | 21 8 27 |
## 3 | | 168 777 46 | 62 330 17 | -73 447 223 |
## 4 | e | 261 897 58 | -68 473 32 | -64 424 268 |
## 5 | E | 234 997 464 | -286 962 513 | 55 35 177 |
```

zxy Ei kovin kiinnostava, mutta voi verrata sekä edellisiin maa-vertailuihin että maan, ikäluokan ja sukupuolen yhteisvaikutusmuuttujan tuloksiin. MG tutkailee eri kysymyksellä tätä samaa asiaa, ja havaitsee että (a) maiden erot suuria ja sukupuolten pieniä (b) naiset liberaalimpia kuin miehet.

zxy miten pitäisi tulkita “oikealle kaatunut U - muoto” miehillä ja naisilla? Järjestys ei toimi, jotain muuta pelissä?

zxy On kiinnostava, mutta aika yksiulotteinen (87 prosenttia ensimmäisellä dimensiolla). **pisteet voisi yhdistää? (29.9.18)**

```
# Luodaan aineistoon kolmen muuttujan yhdysvaikutusmuuttuja maaga, maa, ikäluokka ja sukupuoli.
# Yleensä ei yhdysvaikutuksissa mennä yli kolmen luokittelumuuttujan, ja tässäkin vain maiden pieni lukum
# tekee tarkastelun aika helpoksi.
```

```
ISSP2012esim2.dat <- mutate(ISSP2012esim2.dat, maaga = paste(maa, ga, sep = ""))

#ISSP2012esim2.dat %>% tableX(maa, maaga) # tarkistus, muunnos ok

#head(ISSP2012esim2.dat)
#str(ISSP2012esim2.dat)
```

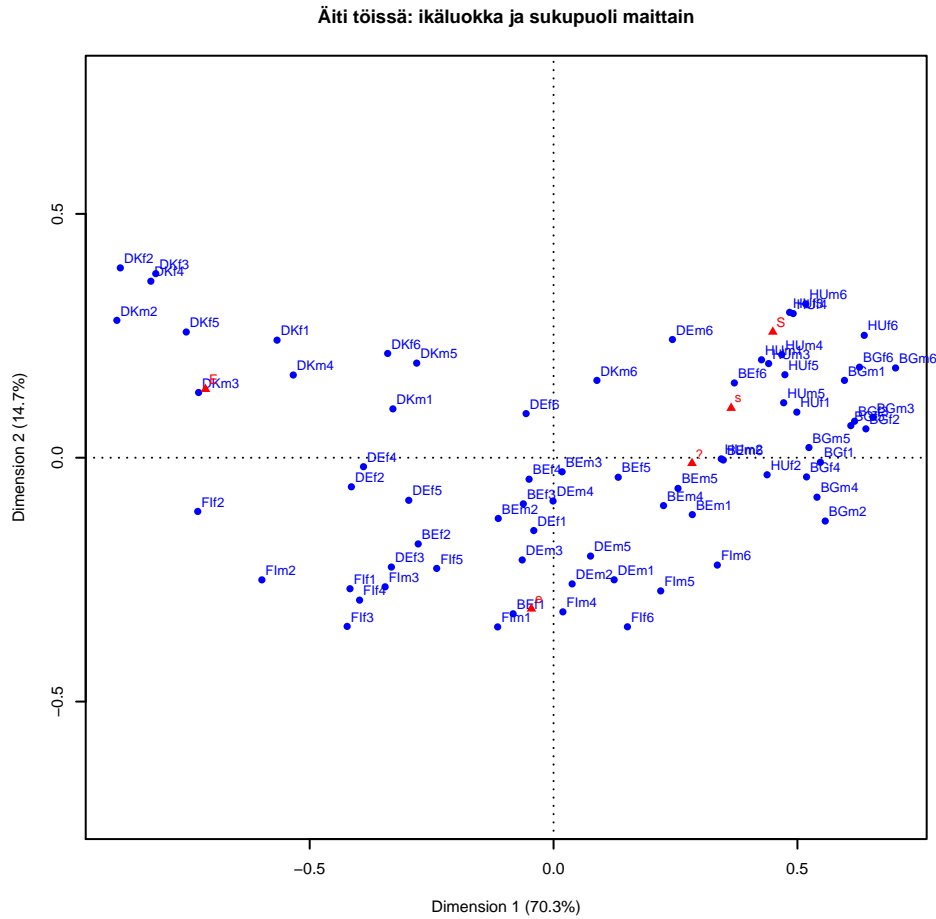
TARKISTA - ja maat voisi lisätä täydentävinä pisteinä (26.9.2018)

```
maagaTestCA1 <- ca(~maaga + Q1b,ISSP2012esim2.dat)
# par("cex"= 0.5, "offset" = 0.5) ei toimi
par("cex"= 0.5)
plot(maagaTestCA1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain", "offset" = 0.5)
```

```
## Warning in plot.window(...): "offset" is not a graphical parameter
```

```
## Warning in plot.xy(xy, type, ...): "offset" is not a graphical parameter
```

```
## Warning in title(...): "offset" is not a graphical parameter
```

```
#str(maagaTestCA1)
# lisätään maapisteet frekvenssitaulukkoon maagaTestCA1$N (26.9.18)? Aika hankalaa...
# maagaTestCA1$N
#maagaTestCA1$rownames
ISSP2012esim2.dat %>% tableX(maaga, Q1b) # aika pieniä frekvenssejä soluissa!
```

maaga/Q1b	S	s	?	e	E	Total
BEf1	5	15	28	43	25	116
BEf2	10	26	34	66	62	198
BEf3	19	27	33	53	42	174
BEf4	21	34	40	55	49	199
BEf5	21	38	46	48	33	186
BEf6	25	58	50	30	22	185
BEm1	9	19	30	24	10	92
BEm2	10	19	31	40	35	135
BEm3	18	33	31	44	36	162
BEm4	19	46	37	51	23	176
BEm5	15	61	34	49	23	182
BEm6	19	75	44	49	21	208
BGf1	2	21	7	9	1	40
BGf2	7	28	17	12	0	64
BGf3	10	44	21	18	1	94
BGf4	14	30	15	24	2	85

maaga/Q1b	S	s	?	e	E	Total
BGf5	16	51	21	25	1	114
BGf6	27	66	26	27	3	149
BGm1	8	12	9	7	1	37
BGm2	4	21	12	14	0	51
BGm3	5	33	16	11	0	65
BGm4	7	19	21	15	1	63
BGm5	12	29	21	19	3	84
BGm6	6	41	19	9	0	75
DEf1	5	28	13	33	23	102
DEf2	9	14	14	37	46	120
DEf3	10	22	12	59	49	152
DEf4	11	31	20	53	71	186
DEf5	8	27	12	43	45	135
DEf6	31	40	15	50	49	185
DEm1	6	26	20	36	15	103
DEm2	7	26	13	39	18	103
DEm3	11	24	15	45	27	122
DEm4	22	39	17	57	37	172
DEm5	11	43	19	54	26	153
DEm6	34	55	28	32	32	181
DKf1	7	11	9	15	41	83
DKf2	4	15	7	13	71	110
DKf3	3	20	15	14	84	136
DKf4	5	24	8	19	90	146
DKf5	6	16	11	22	73	128
DKf6	5	26	11	17	40	99
DKm1	10	21	18	28	47	124
DKm2	2	11	9	16	65	103
DKm3	2	13	12	23	59	109
DKm4	4	24	14	24	59	125
DKm5	11	14	23	18	40	106
DKm6	11	43	15	23	27	119
FIf1	3	9	13	36	33	94
FIf2	5	6	3	34	47	95
FIf3	2	8	13	39	32	94
FIf4	3	15	13	47	40	118
FIf5	6	26	17	52	41	142
FIf6	3	22	21	34	11	91
FIm1	1	9	13	22	13	58
FIm2	2	5	6	28	30	71
FIm3	2	10	9	27	23	71
FIm4	8	23	13	43	18	105
FIm5	5	31	15	35	10	96
FIm6	7	24	13	26	5	75
HUf1	11	13	16	11	3	54
HUf2	15	19	25	22	5	86
HUf3	22	26	26	12	9	95
HUf4	24	25	20	14	8	91
HUf5	21	28	19	19	7	94
HUf6	33	30	18	21	2	104
HUm1	9	15	12	8	5	49
HUm2	18	13	15	22	7	75

maaga/Q1b	S	s	?	e	E	Total
HUm3	15	38	24	16	10	103
HUm4	14	29	17	13	7	80
HUm5	19	31	24	21	7	102
HUm6	18	21	9	11	5	64
Total	810	1935	1367	2125	1906	8143

```

# Miten maa-rivit täydentäviksi riveiksi - alla siisti ratkaisu
# Miten labelit hieman lähemmäksi pistettä? offset-jotenkin toimii...

# rakennetaan taulukko, jossa alimpina riveinä "maa-rivit"
# otetaan karttaan mukaan täydentävinä pisteinä
# karttaa on helpompi tulkita, kun nähdään miten ikä-sukupuoli-ryhmät sijatsevat keskiarvonsa ympärillä

#ikäluokka - sukupuoli ja maa - maaga-muuttuja
testTab1 <- table(ISSP2012esim2.dat$maaga, ISSP2012esim2.dat$Q1b)
#dim(testTab1) #72 riviä, 5 saraketta

# maa-rivit
testTab_sr <- table(ISSP2012esim2.dat$maa, ISSP2012esim2.dat$Q1b)
#testTab_sr

testTab1 <- rbind(testTab1,testTab_sr)
#dim(testTab1)
#dim(testTab1) #78 riviä, 5 saraketta, 1-72 data ja 73-78 täydentävät rivit

spCAmaaga1 <- ca(testTab1[,1:5], suprow = 73:78)
#X11()
par("cex"= 0.75, "asp" = 1, "offset" = 0.5)

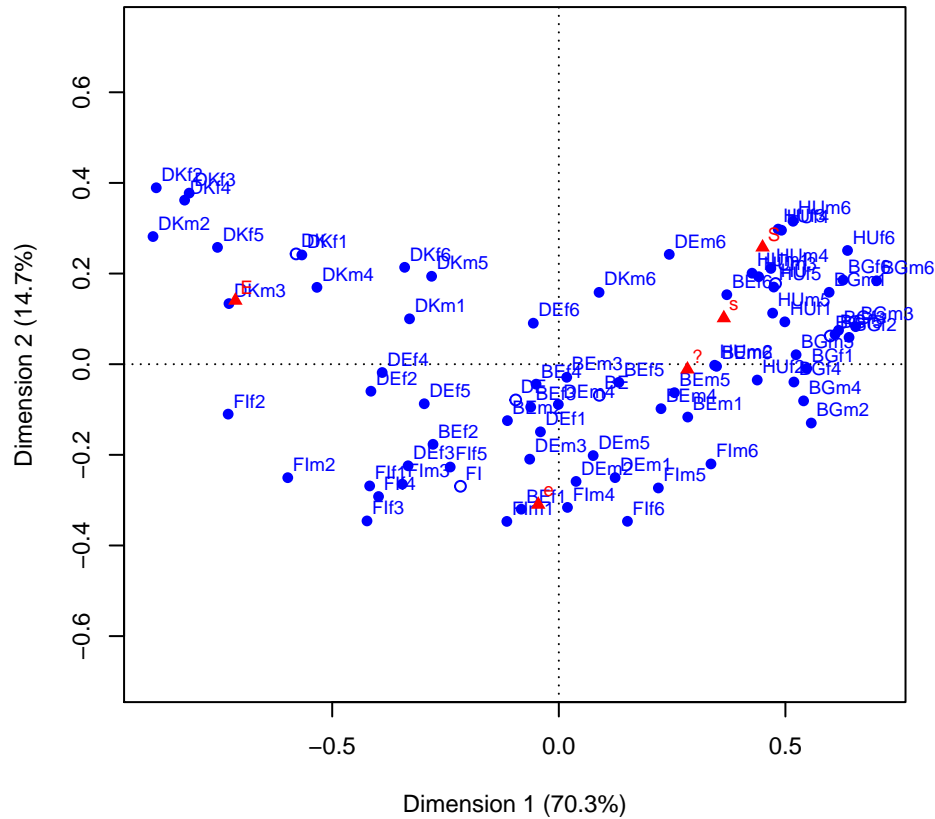
## Warning in par(cex = 0.75, asp = 1, offset = 0.5): "asp" is not a graphical
## parameter

## Warning in par(cex = 0.75, asp = 1, offset = 0.5): "offset" is not a
## graphical parameter

plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2 - maat täydentävinä pisteinä"
      )

```

Äiti töissä: ikäluokka ja sukupuoli maittain 2 – maat täydentävinä pist



```
#par()

#asymmetrinen kartta
#X11()
#par("cex"= 0.75, "asp" = 1, "offset" = 0.5)
#plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 3 (kontributiot) - liian tukkois",
#      map = "rowgreen",
#      contrib= c("absolute", "absolute"),
#      mass = c(TRUE,TRUE),
#      arrows = c(FALSE,TRUE)
#    )
#numeeriset tulokset
summary(spCAmaaga1)
```

```
##
## Principal inertias (eigenvalues):
##
## dim    value      %   cum%   scree plot
## 1      0.184895  70.3  70.3  *****
## 2      0.038751  14.7  85.0  ****
## 3      0.024006   9.1  94.1  **
## 4      0.015502   5.9 100.0  *
## -----
```

Total: 0.263154 100.0

##

##

Rows:

##		name	mass	qlt	inr	k=1	cor	ctr	k=2	cor	ctr				
## 1		BEf1		14	678	9		-83	43	1		-320	635	38	
## 2		BEf2		24	914	11		-278	650	10		-177	264	20	
## 3		BEf3		21	320	3		-62	96	0		-95	224	5	
## 4		BEf4		24	164	3		-50	92	0		-44	71	1	
## 5		BEf5		23	332	5		133	304	2		-40	28	1	
## 6		BEf6		23	832	17		371	710	17		153	121	14	
## 7		BEm1		11	429	9		284	367	5		-117	62	4	
## 8		BEm2		17	372	5		-113	169	1		-125	203	7	
## 9		BEm3		20	108	1		17	29	0		-29	79	0	
## 10		BEm4		22	966	5		225	812	6		-98	154	5	
## 11		BEm5		22	728	8		255	686	8		-63	42	2	
## 12		BEm6		26	788	15		348	788	17		-5	0	0	
## 13		BGf1		5	531	11		547	531	8		-9	0	0	
## 14		BGf2		8	860	14		640	853	17		59	7	1	
## 15		BGf3		12	815	21		617	804	24		75	12	2	
## 16		BGf4		10	932	12		519	927	15		-39	5	0	
## 17		BGf5		14	880	23		609	870	28		66	10	2	
## 18		BGf6		18	921	32		627	846	39		186	74	16	
## 19		BGm1		5	940	7		596	878	9		159	62	3	
## 20		BGm2		6	830	9		557	788	11		-130	43	3	
## 21		BGm3		8	709	19		655	698	19		83	11	1	
## 22		BGm4		8	771	11		540	754	12		-81	17	1	
## 23		BGm5		10	979	11		524	977	15		21	2	0	
## 24		BGm6		9	692	27		701	647	24		184	45	8	
## 25		DEf1		13	425	3		-41	29	0		-149	395	7	
## 26		DEf2		15	938	10		-415	919	14		-60	19	1	
## 27		DEf3		19	846	13		-333	582	11		-224	264	24	
## 28		DEf4		23	985	13		-390	982	19		-18	2	0	
## 29		DEf5		17	839	7		-297	772	8		-87	67	3	
## 30		DEf6		23	116	8		-56	32	0		90	84	5	
## 31		DEm1		13	912	4		124	180	1		-250	732	20	
## 32		DEm2		13	766	4		38	16	0		-259	749	22	
## 33		DEm3		15	737	4		-64	63	0		-210	674	17	
## 34		DEm4		21	137	5		-1	0	0		-89	137	4	
## 35		DEm5		19	603	5		76	75	1		-202	529	20	
## 36		DEm6		22	849	12		244	427	7		242	422	34	
## 37		DKf1		10	991	15		-567	839	18		241	152	15	
## 38		DKf2		14	991	49		-888	831	58		389	160	53	
## 39		DKf3		17	963	53		-816	793	60		377	170	61	
## 40		DKf4		18	977	57		-826	820	66		362	157	61	
## 41		DKf5		16	998	38		-753	894	48		258	105	27	
## 42		DKf6		12	808	9		-340	579	8		214	229	14	
## 43		DKm1		15	981	7		-329	898	9		100	83	4	
## 44		DKm2		13	989	43		-895	900	55		282	89	26	
## 45		DKm3		13	982	28		-728	950	38		134	32	6	
## 46		DKm4		15	941	19		-534	855	24		170	86	11	
## 47		DKm5		13	643	9		-281	435	6		194	208	13	
## 48		DKm6		15	355	5		89	85	1		158	270	9	
## 49		FIf1		12	980	11		-417	693	11		-269	287	21	

```

## 50 | FIif2 | 12 927 26 | -730 907 34 | -110 21 4 |
## 51 | FIif3 | 12 984 13 | -423 590 11 | -346 394 36 |
## 52 | FIif4 | 14 991 14 | -398 644 12 | -292 347 32 |
## 53 | FIif5 | 17 952 8 | -240 502 5 | -227 450 23 |
## 54 | FIif6 | 11 835 7 | 151 134 1 | -347 701 35 |
## 55 | FIim1 | 7 787 5 | -115 78 1 | -347 710 22 |
## 56 | FIim2 | 9 977 14 | -598 832 17 | -250 146 14 |
## 57 | FIim3 | 9 998 6 | -345 629 6 | -265 369 16 |
## 58 | FIim4 | 13 837 6 | 19 3 0 | -316 834 33 |
## 59 | FIim5 | 12 734 7 | 220 289 3 | -273 446 23 |
## 60 | FIim6 | 9 911 6 | 336 637 6 | -220 274 12 |
## 61 | HUf1 | 7 723 9 | 499 698 9 | 93 25 1 |
## 62 | HUf2 | 11 689 11 | 438 685 11 | -35 4 0 |
## 63 | HUf3 | 12 808 18 | 484 586 15 | 298 222 27 |
## 64 | HUf4 | 11 768 18 | 491 564 15 | 296 204 25 |
## 65 | HUf5 | 12 850 13 | 474 753 14 | 170 97 9 |
## 66 | HUf6 | 13 671 34 | 637 581 28 | 251 90 21 |
## 67 | HUum1 | 6 935 5 | 426 766 6 | 201 170 6 |
## 68 | HUum2 | 9 381 11 | 344 381 6 | -2 0 0 |
## 69 | HUum3 | 13 957 12 | 441 803 13 | 193 154 12 |
## 70 | HUum4 | 10 999 10 | 468 830 12 | 211 169 11 |
## 71 | HUum5 | 13 942 12 | 472 891 15 | 113 51 4 |
## 72 | HUum6 | 8 726 15 | 517 529 11 | 315 197 20 |
## 73 | (*)BE | <NA> 510 <NA> | 89 321 <NA> | -69 189 <NA> |
## 74 | (*)BG | <NA> 911 <NA> | 599 901 <NA> | 62 10 <NA> |
## 75 | (*)DE | <NA> 498 <NA> | -95 295 <NA> | -79 203 <NA> |
## 76 | (*)DK | <NA> 983 <NA> | -580 836 <NA> | 243 147 <NA> |
## 77 | (*)FI | <NA> 990 <NA> | -217 389 <NA> | -269 600 <NA> |
## 78 | (*)HU | <NA> 860 <NA> | 478 755 <NA> | 178 105 <NA> |
##
## Columns:
## name mass qlt inr k=1 cor ctr k=2 cor ctr
## 1 | S | 99 653 155 | 450 492 109 | 258 162 171 |
## 2 | s | 238 741 174 | 364 687 170 | 102 54 63 |
## 3 | | 168 535 96 | 284 534 73 | -11 1 1 |
## 4 | e | 261 941 103 | -45 20 3 | -310 921 646 |
## 5 | E | 234 1000 471 | -714 962 645 | 141 37 119 |

```

Kuvissa on aika ahdasta. Kuvan voisi rajata johonkin alueeseen erityisesti oikea yläosa on täynnä pisteitä. Maiden täydentävät pisteet ovat ikäluokka-sukupuoli - luokkien keskiarvopisteitä. Maiden väliset erot dominoivat, mutta maiden välillä on isoja eroja.

Kartan herkkyyttä joillekin pienen massan rivipisteille pitää tutkia tarkemmin.

Vertailu voi tehdä

1.Maiden sisällä, ikä-sukupuoli - luokkien välillä. Ovatko naiset kaikissa ikäluokissa mies-ikäluokkien oikealla vai vasemmalla puolella?

2.Maiden välillä

- miten ikä-sukupuoliluokat sijaitsivat suhteessa maiden keskiarvopisteisiin
- mikä on niiden järjestys

5 Yksinkertaisen korrespondenssianalyysin laajennuksia 2

ZXY Tässä laajennetaan data isommaksi aineistoksi, lisää maita. **TODO 10.10.18** Data-jaksosta koodia tänne!

```
#valittavien maiden kolminumeroinen ISO 3166 - koodi vektoriin - TÄSSÄ KAIKKI MAAT (27, ei Espanjaa)
#incl_countriesALL <- c(36, 40, 56, 100, 124, 191, 203, 208, 246, 250, 276, 348, 352, 372, 428, 440,
#                        528, 578, 616, 620, 643, 703, 705, 752, 756, 826, 840)
#ISSP2012.data <- read_spss("data/ZA5900_v4-0-0.sav") # (user_na = TRUE pois 27.9.18)
#
#str(ISSP2012.data) #61754 obs. of 420 variables
#ISSP2012jh1.data <- filter(ISSP2012.data, V4 %in% incl_countriesALL)
#
```

5.1 Pällekkäiset matriisit (stacked matrices)

Ref:CAip, CA_Week2.pdf (kalvot MCA-kurssilta 2017)

Concatenated tables (yhdistetyt taulut tai matriisit): (a) kaksi luokittelumuuttujaa (b) useita muuttujia stacked ("pinotaan").

MCA 2017 laskareissa ja kalvoissa esitetään, miten nämä saadaan kätevästi CA-paketin MJCA-funktion BURT-optiolla.

5.2 Matched matrices

Ref:CAip ss. 177, HY2017_MCA, Greenacre JAS 2013 (sovellus ISSP 1989, 4 kysymystä 'pitäisikö äidin olla kotona', 8 maata), tässä artikkelissa "SVD-based methods", joista yksi CA (muut biplots, PCA, compositional data/log ratios).

Edellisen menetelmän variantti, jossa ryhmien väliset ja sisäiset erot saadaan esiin. Inertian jakaminen. Samanlaisten rivien ja sarakkeiden kaksi samankokoista taulua, esimerkiksi sukupuolivaikutusten arviointi. Alkuperäinen taulukko jaetaan kahdeksi tauluksi sukupuolen mukaan. Matriisien yhdistäminen (concatenation) riveittäin tai sarakkeittain ei näytä optimaalisesti mm - matriisien eroja.

Ryhmien välisen ja ryhmien sisäinen inertian erottaminen, **ABBA** on yksi ratkaisu (ABBA matrix, teknisesti block circulanMat matrix).

Luokittelu voi olla myös kahden indikaattorimuuttujan avulla jako neljään taulukkoon (esim. miehet vs. naiset länsieuroopassa verrattuna samaan asetelmaan itä-Euroopassa). Samaa ideaa laajennetaan.

Esimerkkinä "Attitudes to women working in 2012".