

G Luku 1 Yksinkertainen korrespondenssianalyysi

Jussi Hirvonen

versio 1.5.9, tulostettu 2020-10-20

Sisällys

1 Data	5
1.1 Luvun 1 tavoitteet	5
1.2 Perhe ja muuttuvat sukupuolirootit - ISSP:n kyselytutkimuksen data 2012	6
1.3 Substanssimuuttujat, taustamuuttujat, muut	8
1.4 Aineiston rajaaminen	9
1.5 Datan valinnan vaiheet ja puuttuvat tiedot	22
1.6 Perusmuunnokset ISSP2012 - datalle	24
2 Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko	63
2.1 Äiti työssä	66
2.2 Korrespondenssianalyysin käsitteet	91
3 Tulkinnan perusteita	91
4 Yksinkertaisen korrespondenssianalyysin laajennuksia 1	104
4.1 Täydentävät muuttujat (supplementary points)	107
4.2 Korrespondenssianalyysin laajennuksia: vuorovaikutusmuuttuja ja osajoukon CA	121
4.3 Kvaliteetti ja stabiilius	130
4.4 Kartan rajaaminen	147
4.5 Subset CA	152
5 Yksinkertaisen korrespondenssianalyysin laajennuksia 2	175
5.1 Matriisiien yhdistäminen (stacked and concatenated matices)	175
5.2 MCA - multiple correspondence analysis	184
5.3 MCA	203

Versio 1.5.9 - lisätty MCA-analyysejä, niitä ennen yksi pinotti taulu - analyysi.

Versiot - vanha Galku - 5.6.2019 versio 1.5.1 Uusi Galku - 2.2.2020 versio 1.5.5, 4.2.2020 versio 1.5.6, 24.2.2020 versio 1.5.7, 6.9.2020 ver-

sio 1.5.8, 16.9.2020 versio 1.5.9.

Siivotaan datan käsitelyn koodilohkot, kopioadaan mahdollisesti hyödylliset koodipätkät tiedostoon siivous1.R (30.1.2020).

Uudet datan luku- ja muunnosskriptit (treeni2-projektista), korjaillaan virheitä ja editoidaan koodia.(31.1.2020)

(2.2.20) Toimii johdattelevaan esimerkkiin asti, myös PDF-tulostus. Kuvien otsikot vähän mitä sattuu, ja 'profiilikuviin asti maa-muuttujan järjestys "väärä", ts. eri kuin vanhemmissa versoissa. Korjattu, lisättiin johdattelevan esimerkin dataan myös maakoodi jossa Saksan ja Belgian jako (V3).

(4.2.20) Versio 1.5.6 - Galku toimii loppuun asti, tarkistettava ja editoitava. Poistetaan tarpeetonta tekstiä, vanha koodi voi jäädä selvästi merkittynä.

(24.2.20) Versio 1.5.7. Pieniä ja isompiakin korjailuja, koodin siistimistä jne.

(27.3.20) Muutetaan hieman karttojen koodilohkoja, html-tulosteessa kuvasuhde 1 mutta pdf-tulosteessa ei. Ero on pieni. (8.4.20)

(9.9.20) Pilkotaan liian pitkiä Rmd - tiedostoja pienemmiksi.

(7.10.20) Vanhat tekstit koodilohkoon "piiloon". Tarkistetaan, että jokaisella

taulukolla ja kartalla on otsikkotiedot, tavalla tai toisella. (10.10.20) Koodilohkon fig.cap ja plot-komennon main; jos main jätetään pois tulostuuko fig.cap? Ei tulostu html-versiossa!

HISTORIAA - koodilohkoon piiloon 7.10.20

```
# Vanhaa tekstiä dokumentista - pilottetaan koodilohkoon 7.10.2020
##6.8.2018 versio 1.0**
#
###Siistitään -> 12.8.2018 versio 1.05**
#
###Kommentit ja korjaukset -> 4.9.2018 versio 1.1**
#
#puuttuva riviprofiili kuva, siistimät interaktiomuuttujien koodaukset, ensimmäinen
#"pinottu taulu" - analyysi -> 19.9.2018 versio 1.2
#
###25.9.2018 siistitään datan käsitellyä; 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 #taulkoita. 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. Ensinnäkin maata, sitten 27 (Espanja pois). Valitaan myös muuttujat, jotta käsiteltävän datan listaukset ovat järkevämpiä. Aineistossa esim. Espanjan ja muutaman Unkarin poikkeavien vastausvaihtoehtojen listaukset ovat omia muuttujina, ja niiden arvo muille havainnoille on NAP (Not applicable). Samoin paljon maakohtaisia muuttujia, esim. koulutustaso. Mukaan otetaan vain #ku-vertailuihin kelpaavat muuttujat, muutama sellainen on myös aineistoon rakennettu. Jätetään pois kaikki perhesuhteisiin liittyvät kysymykset (esim. kotitöiden jakaminen) ja taustatiedot (esim. rahankäytö, 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 alkun 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.
#**5.112.18** Puuttuvat tiedot #ovat puuttuvia, ei voi mitään. Jos vähän ja selviää virheitä (ikä, sukupuoli), voidaan pudottaa havainnot. Muuten mukaan, periaatteessa.
#
#Data-jaksosta siirretään aineiston laajentamisen yhteyteen laajemman muuttujajoukon deskriptiiviset tarkastelu. Taulukko muuttujakuvauksesta jää data-lukuun. **5.12.18** Puuttuneisuuden taulukointia on, mutta siisti NA-taulukko puuttuu.
#
#**11.10.2018 Versio 1.4**
#
#- paperitulosteessa v1.3 kommentteja karttoihin ja ca:n numeerisiin tuloksiin, samoin muuttujalistauksien.
#- paperitulosteessa v1.4 samoin, ja puuttuneisuuden taulukointeja
#
#**11.10.2018 aloitetaan versio 1.5**
#- pieniä muutoksia ja kommentteja, aloitetaan uusi versio 1.51 5.12.2018
#
#**6.12.2018 1.5.1**
#- as_factor - funktio käyttöön; testaillaan miten toimii kun #(a) user_na - arvoja ei lueta ja (b) puuttuvat ovat mukana.
#

```

```

#
#***Muistilista:**
#
#1. Taulukot ja kuvat luvusta 2. alkaen eivät ole "bookdown-muodossa".
#CA-tulokset on tulostettu siisteinä taulukoina Bookdown-demo - dokumentissa.
#Voi tulostaa myös ca-outputin. Ominaisarvojen taulukko keskeneräinen, samoin
#"scree plot" kuvana 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-etaisyyden keskiarvorivistä/sarakkeesta?
#
#4. Kuvissa vielä hiottavaa, pdf-kuvia lisäiltyn img-hakemistoon.
#
#5. Data-tiedostojen nimeäminen (27.12.18)
#
#***ISSP2012*.data** - täysi aineisto
#
#***ISSP2012*jh1.data** - valikoitu aineisto (maat, muuttujat)
#
#***ISSP2012*esimi1.dat** - muuttujien muunnoksia ja uusia muuttujia; analyyseissä
#käytettävä data, tarkenne dat.
#
#6. kasitteet1.rmd - taulukko käsitteistä ja tärkeimmistä ISSP-dokumenteista
#
#***Historiaa ** (11.10.18)
#
#Vanhox kommetteja
#
## kirjastot/paketit ladataan jokaisessa Rmd-dokumentissa
## bib-formaan viitetietokantaa tullaan kokeilemaan
## kuvasuhde (aspect ratio) edelleen epäselvä juttu! Mutta näyttää PDF-tulosteessa
#olevan ok.
#
## Datan käsitteily 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 välitää
# + pidetään muuttujien ja tiedostojen nimeäminen selkeänä, tarkistetaan aika ajoin
#

```

```

## Taulukot: lisättiin riviprosentti- ja sarakeprosenttitaulut (25.4.18), kuva
## riviprofiileista 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
## kokeiluja 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 käsitetäulukko (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 30.1.20 Siivotaan, luodaan faktori-muuttujat heti alussa koko datalle (G1_1_data_fct1.Rmd).

Historiaa

edit tässä luvussa on paljon siistittävää, mutta data on ok. (13.5.2018). **edit** capaper - dokumentissa parempi uusi jäsentely (4.9.2018) **edit** ISSP-dataan perustietoa dokumentissa ISSP_data1.docx (4.9.2018) **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 - uusi versio (24.9.18) 10.10.2018 Maat ja muuttujat valittu.

edit 12.8.2020 käsitteet määriteltävä ennen käyttöä.

CA on eksploratiivinen ja graafinen(visuaalinen) menetelmä, peruskäsitteet esitellään rajatulla aineistolla. Se on kuitenkin oikea tutkimusaineisto kaikkine ongelmineen.

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".

1. Aineiston esittely, laajan kyselytutkimusaineiston tyypilliset ominaisuudet
2. Laadukkaan ja hyvin dokumentoidun aineiston edut
3. CA on käytetty monenlaisten aineistojen analyysiin (esim. ekologia ja biologia, arkeologia, kielen tutkimus). Kyselytutkmukset ovat yksi suosittu sovelluskohde. Muuttujat ovat usein kvalitatiivisia, ja aineistojen peruspulma on puuttuvissa havainnoissa.

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

luvun pitäisi olla mahdollisimman lyhyt (5.12.18)

Hieman historiaa datasta, sosiaalisesti määrätyneet sukupuolirootit (gender) tutkimusaiheena neljässä ISSP:n kyselytutkimuksessa. Avainsana "cross-cultural comparative studies", "cross country" ehkä myös? Tämä on varsinainen teollisuuden ala! Maiden väliset vertailut ovat aina hyvin hankalia.

#V Ajankohtainen tutkimusaihe, ei ISSP-dataa mutta samoja kysymyksiä. (<https://www.economist.com/finance-and-economics/2020/10/03/why-east-and-west-german-women-still-work-vastly-different-hours>) Tutkimus: (https://www.diw.de/documents/publikationen/73/diw_01.c.799295.de/dwr-20-38.pdf)

Tärkeät linkit - ISSP 2012 data

Toimivat html-tulosteessa, PDFtiedostoissa saa toimimaan (vaati tarkat formatoinnit Rmd-koodissa).

www.issp.org, tutkimushankkeen historiaa. Löytyy myös bibliografia tutkimuksesta, joissa aineistojen käytetty.

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

data ja dokumentaatio (selattavissa): zacad.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

Alla myös suora linkki

Linkitys dokumentteihin on hankalaa

- monta portaalia, joista pääsee monien organisaationimien taakse
- tärkeimmät linkit ISSP-tutkimuksen "kotisivu" ja selkeät **muuttujavaukset ja muut tiedot**

- käytännössä linkittäminen “syvälle” johonkin sivustoon tai www-palveluun ei ole järkevää, parempi antaa selkeät viitetiedot ja tiedot organisaatioista. Ne säilyvät, tai jäljille pääsee.

Edit Refworksissa on kerätty viitteitä, tässä päijätään kolmen saitin osoitteilla. Alla linkkejä jotka eivät näy PDF-tulosteessa, lisätty tekstinä. Suomessa tutkimusta koordinoi [yhteiskuntatieteellinen tietoarkisto](<http://urn.fi/urn:nbn:fi:fsd:T-FSD2820>). Kuvaus datasta ja muuta hyödyllistä suomeksi, esim. lomake (ZA5900_q_fi-fi.pdf).

Aineistot <https://dbk.gesis.org/dbksearch/sdesc2.asp?no=5900&db=e2012> (19.10.20 tiedostoluettelo).

[Muuttujakuvaukset, data ja dokumentaatio] (<http://zacad.gesis.org/webview/index.jsp?object=http://zacad.gesis.org/obj/fStudy/ZA5900>) **Täällä on kaikki.** Tietoarkisto, Leibniz Institute for the Social Sciences.

Dokumentointi on kattava, tiedot löytyvät haastattelumenetelmista (parerilomake, tietokoneavusteinen haastattelu, jne), maakohtaisten taustamuuttujien harmonisoinnista maittain, otantamenetelmistä jne. Esittelen vain aineiston tärkeimmät rajaaukset. Monitorointiraportti kertoo puuttuneisuuden määrään, otantamenetelmät jne maittain. ”Code book” kertoo muuttujien määritelmät sekä yhteisille että maakohtaisille muuttujille. Kaikista muuttujista on taulukko maittain. Lisäksi raportti kyselylomakkeen laadinnasta.

```
issp_docname <- c("Variable Report", "Study Monitoring Report", "Basic Questionnaire",
                  "Contents of ISSP 2012 module", "Questionnaire Development")
issp_docdesc <- c("Perusdokumentti, muuttujien kuvaukset ja taulukot",
                  "tiedokerun toteutus eri maissa",
                  "Maittain sovellettava kyselylomake", "substanssikysymykset taulukkona",
                  "kyselylomakkeen laatiminen")
issp_docfile <- c("ZA5900_cdb.pdf", "ZA5900_mr.pdf", "ZA5900_bq.pdf", "ZA5900_overview.pdf",
                  "ssoar-2014-scholz_et_al-ISSP_2012_Family_and_Changing.pdf")

col_isspdocs <- c("dokumentti", "sisältö", "tiedosto")
# colnames(ISSPdocsT.df) <- col_isspdocs
# Vanha df-koodi
# ISSPdocsT.df <- data_frame(issp_docname, isspp_docdesc, isspp_docfile)
# knitr::kable(ISSPdocsT.df, booktab=TRUE)

# varoitukset data_framen käytöstä, toimisiko tibble()?
ISSPdocsT.tbl <- tibble(issp_docname, isspp_docdesc, isspp_docfile)
colnames(ISSPdocsT.tbl) <- col_isspdocs
knitr::kable(ISSPdocsT.tbl, booktab = TRUE,
            caption = ' ISSP 2012: tärkeimmät dokumentit')
```

Taulukko 1: ISSP 2012: tärkeimmät dokumentit

dokumentti	sisältö	tiedosto
Variable Report	Perusdokumentti, muuttujien kuvaukset ja taulukot	ZA5900_cdb.pdf
Study Monitoring Report	tiedokeruun toteutus eri maissa	ZA5900_mr.pdf
Basic Questionnaire	Maittain sovellettava kyselylomake	ZA5900_bq.pdf
Contents of ISSP 2012 module	substanssikysymykset taulukkona	ZA5900_overview.pdf
Questionnaire Development	kyselylomakkeen laatiminen	ssoar-2014-scholz_et_al-ISSP_2012_Family_and_Changing.pdf

1.3 Substanssimuuttujat, taustamuuttujat, muut

zxy capaper - dokumentissa uusi jäsentely (4.9.2018)

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 editoimissa ja tiedonkeruun suunnittelussa on nähty paljon vaivaa vertailukelpoisuuden vuoksi. Tästä esimerkkejä, esim. "mitä puoluetta äänestit".

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

Parempi muotoilu: Varsinaiset substanssimuuttujat eli kyselylomakkeet on kottetti hioa mahdollisimman yhdenmukaisiksi. Silti pieniä eroja löytyy, ja isoakin (Espanja on pudottanut neutraalin "en samaa enkä eri mieltä" - vaihtoehdon pois, ja Unkarissa on muutamat vastausvaihtoehdot valittu omalla tyyllillä). Taustamuuttujissa on pyritty samaan, ja aineistoon on myös rakennettu kansainvälisesti vertailukelpoisia muuttuja kansallisesti kerätyistä tiedoista. Näitä ovat erityisesti tuloihin liittyvät tiedot, ja mone muutkin. Muuttujat jakautuvat substanssi- ja taustamuuttujaan, ja taustamuuttujista monet tiedot on kerätty kansallisiin ainiestossa maan kirjantunnisteella alkaviin muuttujaan.

zxy HUOM! Dataa ei ole kerätty vain kansainvälisiin vertailuhiihin! 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.)

1.4 Aineiston rajaaminen

1. Eurooppa ja samankaltaiset maat (25)

(24.2.20) Aineistosta valittiin ensin joukko suhteellisen samankaltaisia kehityneitä teollisuusmaita. Sitten valittiin osa kysymyksistä, ja vielä suppeampi valikoima kiinnostavia taustamuuttuja. Muutama maa pudotettiin pois tämän valinnan jälkeen.

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

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, Iso-Britannia, USA pois -> **25 maata (11.10.18)**

Espanja jätettiin pois, koska siellä kysymyksissä jätettiin pois neutraali vaihtoehto ("en puolesta enkä vastaan / en osaa sanoa"). USA ja GB pois koska kiinnostava TOPBOT-muuttuja puuttuu (puuttui 11.10.18, sittemmin USA:n ainestoa on täydennetty).

3. Datan hallinta - reproducible research- periaate

Helposti toistettava analyysi, ei "haurasta" datan muokkauksen koodia.

edit 24.2.20 Vanhoja perusideoita

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

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

Kommnetti 3.10.18

R-spesifit: R-koodissa tarkemmin, kaikki yksityiskohdat.

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. Siksi muuttujien tyypimuunnokset (yleensä faktorointi) tallennetaan uusiksi muuttujiksi, metatieto säilyy vanhassa muuttujassa.

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

Puuttuva tieto voidaan koodata monella tavalla (ei halua vastata jne), ja SPSS (datan jakelutiedosto) sallii kolme koodia puuttuville tiedoille. Ne voi lukea R-dataan, mutta puuttuneisuutta ei tässä työssä tutkita sen tarkemmin. Detaljitet R-koodissa (haven-paketin read_spss-funktion user_na -optio, ei käytetä tässä).

Tiedostonimistä (10.10.18, 30.1.20, 11.2.20, 22.9.20)

edit 22.9.20 Tarkista kun valmista.

ISSP2012.data - täysi aineisto, luetaan SPSS-tiedostosta ISSP2012jh1.data - valittu osa aineistosta (maat, muuttujat) ISSP2012*.jh1.dat - valittu osa aineistosta, luotu uusia muuttujia ja muunnettua muuttuja. Alkuperäiset muuttujat säilytetään, voi aina tarkistaa ja verrata. ISSP2012esim1, 2 jne, tarkenne .dat rajattuja aineistoja joissa uusia muuttujia ja muuttujien nimiä. Näitä luodaan analyysin eri vaiheissa.

Datan perusmuokkauksen vaiheet

1. Data r-dataaksi

```
ISSP2012jh.data <- read_spss("data/ZA5900_v4-0-0.sav")
```

2. valitaan maat(25)

```
ISSP2012jh1a.data <- filter(ISSP2012jh.data, V4 %in% incl_countries25)
```

3. valitaan muuttujat

```
ISSP2012jh1b.data <- select(ISSP2012jh1a.data, all_of(jhvars1))
```

Poistetaan havainnot joilla tieto sukupuolesta tai ikä puuttuu.

```
ISSP2012jh1c.data <- filter(ISSP2012jh1b.data, (!is.na(SEX) & !is.na(AGE)))
```

Perusmuunnokset (G1_1_data_fct1.Rmd)

```
ISSP2012jh1d.data <- ISSP2012jh1c.data
```

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

DATA RAJAAMISTA - maat(5.10.2018)

```
# Aineiston rajaamisen kolme vaihetta (10.2018)
#
# TIEDOSTOJEN NIMEÄMINEN
#
# R-datatiedostot .data - tarkenteella ovat osajoukkoja koko ISSP-dataasta ISSP2012.data
# R-datatiedostot .dat - tarkenteella: mukana alkuperäisten muuttujien muunnoksia
# (yleensä as_factor), alkuperäisissä muuttujissa mukana SPSS-tiedoston metadata.
#
# Luokittelumuuttujan tyyppi on datan lukemisen jälkeen yleensä merkkijono (char)
# ja haven_labelled.
#
# Muutetaan R-datassa ordinaali- tai nominaaliasteikon muuttujat haven-paketin
# as_factor - funktiolla faktoreiksi. R:n faktorityyppin muuttujille voidaan tarvittaessa
# määritellä järjestys, toistaiseksi niin ei tehdä (25.9.2018).
```

```

#
# Muunnetun muuttujan rinnalla säilytetään SPSS-tiedostosta luettu muuttaja, metatiedot säilytetään alkuperäisessä.
#
# R-datatiedostot joiden nimen loppuosa on muotoa *esim1.dat: käytetään analyyseissä
#
# 1. VALITAAN MAAT (25) -> ISSP2012jh1a.data. Muuttujat koodilohkossa dataset_vars1
#
# kolme maa-muuttujaan datassa. V3 erottlee joidenkin maiden alueita, V4 on koko maan koodi 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 ainestoa täydennettiin, koska siinä oli puutteita. Jako ei siis ole oleellinen
# mutta muut ovat. Tähdellä merkityt maat valitaan johdattelevaan esimerkkiin.
#
# Maat (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 muutujia 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
#



ISSP2012jh.data <- read_spss("data/ZA5900_v4-0-0.sav") #luetaan alkuperäinen data R- dataksi

#str(ISSP2012jh.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(ISSP2012jh.data)
#str(ISSP2012jh.data) #61754 obs. of 420 variables - kaikki

ISSP2012jh1a.data <- filter(ISSP2012jh.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() # muuttujen nimet
# Maakohtaiset muuttujat (kun on poikettu ISSP2012 - vastausvaihtoehdosta tms.)
# on aineistossa eroteltu maatunnus-etulittueellä (esimerkiksi ES_V7).
# Demografisissa ja muissa taustamuuttujissa suuri osa tiedoista on kerätty maa-
# kohtaisilla lomakkeilla. Vertailukelpoiset muuttujat on konstruoitu niistä.
# Muuttuja on 420, vain osa yhteisiä kaikille maille.

```

DATAN RAJAAMISTA - MUUTTUJAT (5.10.2018)

SPSS-tiedostosta saadaan luettua haven-paketin read_spss-funktiolla paljon metatietoja.

```

# 2. VALITAAN MUUTTUJAT -> ISSP2012jh1b.data. Maat valittu koodilohkossa dataset_country1

# METADATA

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

# MAA - maakoodit ja maan kahden merkin tunnus

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

# SUBSTANSSIMUUTTUJAT - Attitudes towards family and gender roles (9)
#
# Yhdeksän kysymystä (lyhennetyt versiot, englanniksi), vastausvaihtoehdot Q1-Q2
#
# 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
#
# Q3a Should women work: Child under school age
# Q3b Should women work: Youngest kid at school
# 1= kokopäivätö, 2 = osa-aikatyö, 3 = pysyä kotona, 8 = en osaa sanoa (can't choose), 9 =

```

```

#
# Kysymysten Q3a ja Q3b eos-vastaus ei ole sama kuin "en samaa enkä eri mieltä" (ns. neutraali)
# vaihtoehto), mutta kieltyymisiä jne. (koodi 9) on aika vähän. Kolmessa
# maassa ne on yhdistetyt:
# (8 Can't choose, CA:can't choose+no answer, KR:don't know+refused, NL:don't know).
# Kun SPSS-tiedostosta ei ole tuotu puuttuvan tiedon tarkempaa luokittelua,
# erottelua ei voi tehdä.
#
#
#
# substvars1 <- c("V5","V6","V7","V8","V9","V10","V11","V12","V13") # 9 muuttujaan

# Nämä yhteiset muuttujat pois (maaspesifien muuttujien lisäksi) :
#
# "V14", "V15", "V16", "V17", "V18", "HU_V18", "V19", "HU_V19", "V20", "HU_V20", "V21",
# "V28", "V29", "V30", "V31", "V32", "V33", # "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"
#
#
# DEMOGRAFiset JA MUUT TAUSTAMUUTTUJAT (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 enter university)
# other ISCED level 3 programs - designed to prepare students for direct entry into the labour market
# 4 Post secondary, non-tertiary (other upper secondary programs toward labour market or tertiary)
# 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 kategoriaa 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
#
# Voisi koodata dummymuuttujaksi lapsia (1) - ei lapsia (0).
# Ranskan datassa on erittäin iso osa puuttuvia tietoja ("99", n. 20 %), myös
# Australialla aika paljon. Sama tilanne myös muissa perheen kokoon liittyvissä
# kysymyksissä.
#
# 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
#
# URBURRAL - 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. Vaihotehdon 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, all_of(jhvars1))

# laaja aineisto - mukana havainnot joissa puuttuvia tietoja
# hauska detalji URBURRAL - muuttujan metatiedoissa viite jonkun työaseman hakemistoon
# str(ISSP2012jh1b.data) #32969 obs. of 23 variables
#
# SUBSTANSSIMUUTTUJAT
#
# $ V5      : 'haven_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 mother"
# ..- attr(*, "labels")= Named num  0 1 2 3 4 5 8 9
#
# ISSP2012jh1b.data$V5 näyttää tarkemmin rakenteen
#
# glimpse(ISSP2012jh1b.data)
# str(ISSP2012jh1b.data) # 32969 obs. of 23 variables

# Poistetaan havainnot, joissa ikä (AGE) tai sukupuolitieto puuttuu (5.7.2019)

```

```

ISSP2012jh1c.data <- filter(ISSP2012jh1b.data, (!is.na(SEX) & !is.na(AGE)))

# str(ISSP2012jh1c.data) # 32823 obs. of 23 variables, 32969-32823 = 146
# TARKISTUS 8.6.20 dplyr 1.0.0 havaintojen ja muuttujien määrä ok.

# ISSP2012jh1c.data %>% summary() %>% kable()

```

Metatietojen (3) ja maa-muuttujien (3) lisäksi aineistossa on seitsemäntoista muuttuja. Yhdeksän muuttujaa ovat ns. substanssikysymysten vastauksia, joilla luodataan asenteita sukupuolirooleihin ja perhearvoihin. Taustamuuttuja on kahdeksan.

Yhdeksän kysymystä (lyhennetty versiot, englanniksi), vastausvaihtoehdot

Vastausvaihtoehdot:

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

edit 14.8.20 Kysymyksissä Q1a ja Q2a vastausten järjestys on tulkinnan (moderni vs.perinteinen tai liberaali vs. konsertavi) erilainen.

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

Vastausvaihtoehdot: "Work full-time" "Work part-time" "Stay at home", "Can't choose" 1 = W, 2 = w, 3 = H, NA = 6,8,9 ei tässä eriteltyä. 6 on Taiwanin oma vastausvaihtoehdo, 8 = en osaa sanoa ja 9 = no answer.

Muuttuja taulukkona - karkea tapa

```

tabVarNames <- c(substvars1,bgvars1) # muuttujanimet muuttujille

# Kysymysten lyhyet versiot englanniksi
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 compar
    "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
    "Legal partnership status: married, civil partnership...",
    "Place of living: urban - rural"
)
#tabVarDesc

# Taulukko

# luodaan df - varoitus: data_frame() is deprecated, use tibble" (4.2.20),
# vaihdetaan tibbleen (21.2.20)

# jhVarTable1.df <- data_frame(tabVarnames,tabVarDesc) OLD
jhVarTable1.tbl <- tibble(tabVarnames,tabVarDesc)
cols_jhVarTable1 <- c("muuttuja","kysymyksen tunnus, lyhennetty kysymys")
colnames(jhVarTable1.tbl) <- cols_jhVarTable1
str(jhVarTable1.tbl)

## tibble [17 x 2] (S3: tbl_df/tbl/data.frame)
## $ muuttuja : chr [1:17] "V5" "V6" "V7" "V8" ...
## $ kysymyksen tunnus, lyhennetty kysymys: chr [1:17] "Q1a Working mother can have warm re
# Suomalaiset pitkät kysymykset
vastf1 <- c("Q1a Työssäkäyvä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän
          ja turvallisen suhteen kuin äiti, joka ei käy työssä")

vastf2 <- c("Q1b Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työss
vastf3 <- c("Q1c Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätyö.")
vastf4 <- c("Q1d On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat
          ensisijaisesti kodin ja lapsia.")
vastf5 <- c("Q1e Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyon tekeminen.")
vastf6 <- c("Q2a Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen")
vastf7 <- c("Q2b Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja per
vastf8 <- c("Q3a Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa
          Kun perheessä on alle kouluikäinen lapsi")
vastf9 <- c("Q3b 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) OLD
jhVarTable1_fi.tbl <- tibble(tabVarnames_subst,tabVarDesc_fi)
cols_jhVarTable1 <- c("muuttuja","Kysymyksen tunnus, suomenkielisen lomakkeen kysymys")
colnames(jhVarTable1_fi.tbl) <- cols_jhVarTable1

# TAULUKODEN TULOSTUS

# kable(booktab = T) # booktab = T gives us a pretty APA-ish table
# Lyhyet kysymykset englanniksi

knitr::kable(jhVarTable1.tbl, booktab = TRUE,
             fig.cap = "ISSP2012: Työelämä ja perhearvot - valitut muuttujat")

```

muuttuja kysymyksen tunnus, 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	Maintain status: work, unemployed, in education...
TOPBOT	Top-Bottom self-placement (10 pt scale)
HHCHILD	How many children in household: children between [school age] and 17 years of age
MARITAL	Legal partnership status: married, civil partnership...
URBRUR	Place of living: urban - rural

```
# Suomen lomakkeen kysymykset (löytyy myös kuva lomakkeen sivustasta)
```

```
knitr::kable(jhVarTable1_fi.tbl, booktab = TRUE,
             fig.cap = "ISSP2012: suomenkielisen lomakkeen kysymykset")
```

muuttuja kysymyksen tunnus, suomenkielisen lomakkeen kysymys

V5	Q1a Työssäkävä äiti pystyy luomaan lapsiinsa aivan yhtä lämpimän
----	--

muuttu **Kysymyksen tunnus, suomenkielisen lomakkeen kysymys**

- ja
tur-
valli-
sen
suh-
teen
kuin
äiti,
joka
ei
käy
työssä
- V6 Q1b Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä.
- V7 Q1c Kaiken kaikkiaan perhe-elämä kärsii, kun naisella on kokopäivätö.
- V8 Q1d On hyvä käydä töissä mutta tosiasiassa useimmat naiset haluavat ensisijaisesti
ko-
din
ja
lapsia.
- V9 Q1e Kotirouvana oleminen on aivan yhtä antoisaa kuin ansiotyön tekeminen.
- V10 Q2a Sekä miehen että naisen tulee osallistua perheen toimeentulon hankkimiseen.
- V11 Q2b Miehen tehtävä on ansaita rahaa; naisen tehtävä on huolehtia kodista ja perheestä.
- V12 Q3a Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?
- Kun
per-
hees-
sä
on
alle
kou-
lui-
käi-
nen
lapsi
- V13 Q3b Millä tavoin naisten pitäisi mielestäsi käydä työssä seuraavissa tilanteissa?

muuttu **Kysymyksen tunnus, suomenkielisen lomakkeen kysymys**

Kun
nuo-
rin
lapsi
on
aloit-
ta-
nut
koulunkäynnin

Taulukot voivat olla hankalia eristyisesti PDF-tulostuksessa, jos ne ovat
monimutkaisia tai solujen "koot" (merkkiä/solu) vaihtelevat paljon.

Tarkemmat kuvaukset lähes tuhatsivuisessa koodikirjassa ZA5900_cdb.pdf (**refworks-viite pitäisi löytyä**, ja ISSP dokumentit kerrotaan luvun alussa).

Bookdown-versiossa taulukot omiksi koodilohkoiksi, ja fig.caption - optiolla taulukon otsikko.

Kysymyslomakkeen kuva, vai kuva liitteisiin? **Liitteisiin**.

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

Seuraavaksi perheeseen, työhön ja kotitöihin liittyviä kysymyksiä.						
23. Mitä mieltä olet seuraavista väittämistä? Rengasta jokaiselle... riviltä vain yksi valitsesto						
	Taydin samaan mieltä	Samaa mieltä	En samaa eri mieltä	Eri mieltä	Taydin eri mieltä	En osaa
a) Työssäkäyvä illi pystyy luomaan lapsineen avan yhtä lämpimän ja turvallisen suhteen kuin illi, joka ei käy työssä.....	1	2	3	4	5	8
b) Alle koulukäinä lapsi todennäköisesti karsii, jos hänen arinna käy työssä.....	1	2	3	4	5	8
c) Kaiken kaikkiaan perhe-elämä karsii, kun he eivät ole mukana työpäivässä.....	1	2	3	4	5	8
d) On hyvä käydä töissä mutta toisaalta se useimmitaasek haluvat ensteijäiseksi kodin ja lapsia.....	1	2	3	4	5	8
e) Kotirovuna oleminen on aivan yhtä antoisaa kuin ansioityn tekeminen	1	2	3	4	5	8
24. Mitä mieltä olet seuraavista väittämistä? Rengasta kummattakin riviltä vain yksi valitsesto.						
	Taydin samaan mieltä	Samaa mieltä	En samaa eri mieltä	Eri mieltä	Taydin eri mieltä	En osaa sanoo
a) Sekä miehen että naisen tulisi osallistua perheen toimintatulon hoitaminnoon	1	2	3	4	5	8
b) Miehen tehtävät on ainaa rahaa; naisen tehtävät on huolehtia kodista ja perheestä	1	2	3	4	5	8
25. Millä tavoin naisten pitäisi mielstävä käydä työssä seuraavissa tilanteissa? Rengasta kummattakin riviltä vain yksi valitsesto.						
Naisen tulisi...	käydä kokopäiväyksessä	käydä osa-alkiyössä	pysyä kotona	En osaa		
a) Kun perheessä on alle koulukäinä lapsi	1	2	3	8		
b) Kun nuorin lapsi on aloittanut koulutuksensa	1	2	3	8		

Kuva 1: Suomen lomake

edit 10.10.20 “En osaa sanoa” on mielipide, mutta tässä tutkielmanissa se tulkitaan puuttuvaksi vastaukseksi. Puuttuvia vastauksia voisi SPSS-datasta analysoida tarkemminkin (kolme SPSS-koodia). Kysymyksessä 25 on jätetty

neutraali vaihtoehto pois, antaa mahdollisuuden arvioida neutraalien vastausten ja ”en osaa sanoa”- vastausten tai puuttuvat vastauksen yhteyksiä.

edit 10.10.20 Kysymysten suunta vaihtelee. Vaihtoehto ”täysin samaa mieltä (1)” on vahvasti liberaali/moderni kysymyksissä 23a ja 24a ja vahvasti konservatiivinen/perinteinen kysymyksissä 23b ja 24b. Kolme Kysymystä (23c - 23e) ovat hieman monimerkityksisiä (“double barreld”), miten ”samaa mieltä” - vastaukset pitääsi tulkita? Samaa mieltä voi olla myös siksi, että ”näin nämä asiat nyt ovat”, omasta mielipiteestä riippumatta, realistisena arviona.

1.5 Datan valinnan vaiheet ja puuttuvat tiedot

edit 24.2.20 Toistoa

1. Vastauskato on kyselytutkimusten suurin ongelma.

Johdattelevassa esimerkissä on kolme muuttujaa, ei ongelma, aika vähän puuttuvia.

Isomman 25 aineiston osalta tarkistetaan, mitä ”listwise deletion” saa aikaan. Aineisto pienenee nopeasti, ja vaikeasti hahmotettavalla tavalla. Tämä erävastauskato ei ole tutkielman ydinauhe, mutta laajemman aineiston käytössä se täytyy ottaa huomioon. Yksikkövastauskato ei käsitellä, tutkimuksen toteutukseen raporteissa on kerrottu tarkemmin miten kyselyn toteuttajat ovat tämän huomioineet. Yksikkövastauskato eli otokseen poimitut joita ei ole tavoitettu ollekaan on kansallisen tason ongelma, joka on ratkaistu vaihelevin tavoin. Tiedot löytyvät aineiston dokumentatiosta. Aineistossa on myös mukana maakohtaiset painomuuttujat, mutta ei painoja maiden vertailuun. Vastausprosentit (response rate) vaihelevat maittain, kts. monitoring report. (**edit** toistoa! 24.2.20)

CA:n eräs etu on se, että muuttujien otetaan olevan luokittelasteikon (nominaliasteikon) muuttuja, ja puuttuva havainto on yksi luokka lisää. Puuttuvat havainnot otetaan mukaan laajemmassa aineistossa myös siksi, että CA ja MCA edellyttää yleensä useamman muuttujan analyyseissä sitä. Jokaisen kahden muuttujan parittaisen ristiintaulukoihin reunajakaumien pitää olla samoja.

ks Perusasiat havaintojen puuttellisuudesta kyselytutkimuksissa. Yksikkövastauskato (unit non-response), erävastauskato (item non-response). Mitä on raportoitava, kun käytetään valmista aineistoa? Erävastauskatoa analysoidaan, kun käytetään kaikkia valittuja muuttujia.

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

Datakatalogi-dokumentista näkee vastausten jakauman jokaisen kysymyksen osalta, myös puuttuvien tietojen tarkemman koodauksen.

1. Valitaan 25 maata ja muuttujat

2. Johdattelevissa esimerkeissä valitaan kuusi maata ja kolme muuttuja. Jätetään pois kaikki havainnot (vastaukset) joissa on puuttuvia tietoja (“listwise deletion”)
3. Kun laajempi aineisto otetaan käyttöön, joudutaan pohtimaan miten puuttuvia havaintoja käsitellään. Jos kyse on selvistä virheistä (esim. haastateltavan ikä puuttu) havainnot jätetään pois, muuten mietitään.

Miten puuttuvia tietoja (erävastuskato, havainnossa puuttu joku tieto) käsitellään?

1. Miksi tieto puuttuu, mitä “puuttuva tieto” tarkoittaa? Lavea kysymys!

Joissain kysymyksissä (V12, V13) puuttuvaksi tiedoksi kirjautuu vastaus (“en osaa sanoa”) “ei vastausta” - vaihtoehdon lisäksi. Nämä mukaan.

Ikä ja sukupuoli: ilmeinen virhe, joten jätetään havainnot pois (näitä ei ole paljon).

“Listwise delete” on raaka ratkaisu, kun muuttujia on paljon. Imputointi, mutta CA:lla voi analysoida puuttuvaa arvoa yhtenä luokittelumuuttujan “modaliteettina”.

2. Puuttuvien tietojen jakauma?

edit 24.2.20) Kun laajempi aineisto ja puuttuvat arvot otetaan mukaan analysiin loppuluvuissa, vilkaistaan pikaisesti erävastauskodon rakennetta.

3. Onko puuttuvia tietoja tasaisesti eri maissa, vai vaihteleeko niiden suhteellinen osuus?

Vaihtelee, ja jo tästäkin syystä puuttuvien käsitteily on oleellinen asia. Ensimmäisenä ratkaisuna ne voi pitää mukana ca/mca - analyyseissä, reunajakaumat pysyvät samoina.

4. Onko joissain tai jossain maassa huomattava määrä puuttuvia tietoja?

Joissain muuttujissa on kohtalaisen paljon puuttuvia tietoja joissain maissa.

Tarkemmin puuttuneisuutta ei analysoida. Esimerkkejä löytyy (MG, CAiP ja ”vihreä kirja”). Kaksi R-pakettia, joilla pikaisesti vilkaistaan dataa, ei vielä mukana tässä (24.2.20). **edit** Viite!

Koko aineistossa (valitut 25 maata) kysymyksen Q1b (muuttuja V6) vastauksista puuttuvia tietoja on 3,5 prosenttia (1219/34271). **Huom:** kun pudotetaan havainnot joilta SEX tai AGE puuttuu, N = 32823. On oikea määrä (5.7.2019, kts. treeni2- projekti, Data_iso1.R).

edit kaksi vanhaa koodilohkoo, olkoon toistaiseksi mukana (11.2.20)

Puuttuvien tietojen tarkempi koodaus ISSP-datassa:

0: Not applicable (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."

1.6 Perusmuunnokset ISSP2012 - datalle

Datatiedosto on muunnosten jälkeen **ISSP2012jh1d.data**, luokittelumuuttujat muunnetaan R:n factor- muuttujaksi.

Jokaisesta muuttujasta on kaksi versiota, toisessa puuttuvat tiedot ovat R:n "NA"- arvoja ja toisessa "NA"-arvo on eksplisiittinen muuttuja ("missing").

Substanssimuuttujien luokkien tunnuiset ("faktorilabelit") muutetaan graafisiin analyyseihin sopivan lyhyiksi. Taustamuuttujien luokittelua ja luokkien tunnuksia pohditaan, kun ne otetaan käyttöön.

Factor: määritelmä

"Very short : levels are the input, labels are the output in the factor() function. A factor has only a level attribute, which is set by the labels argument in the factor() function. This is different from the concept of labels in statistical packages like SPSS, and can be confusing in the beginning." (<https://stackoverflow.com/questions/5869539/confusion-between-factor-levels-and-factor-labels>)

Haven-paketin labelled_spss-luokka (<https://github.com/tidyverse/haven/issues/172>), kaksi toisiaan täydentävää käyttötapaa. Se on yksi "välimuoto" kun dataa luetaan SPSS/SAS/Stata formaatista R-formaatteihin. Toisaalta labelled-paketin avulla luokan olioita voi monipuolisesti muokata monipuolisesti ja jakaa tuloksia takaisin muiden ohjelmistojen tiedostoformaatteihin. Tässä ensimmäinen vaihtoehto käytössä.

#V Tärkein lähde McNamara&Horton(2017) "Wranglin with categorical data in R".

R-maailmassa on monta tapaa tehdä asioita. Tässä käytetäänforcats-paketin funktioita, ei dplyr-paketin kuten em. artikkeliissa.

Muunnokset: mutate (<https://suzan.rbind.io/2018/02/dplyr-tutorial-2/#changing-column-names-after-mutation>)

Faktorit - recode

dplyr

"You can use recode() directly with factors; it will preserve the existing order of levels while changing the values. Alternatively, you can use recode_factor(),

which will change the order of levels to match the order of replacements. See the `forcats` package for more tools for working with factors and their levels.”

”This is a vectorised version of `switch()`: you can replace numeric values based on their position or their name, and character or factor values only by their name. This is an S3 generic: `dplyr` provides methods for numeric, character, and factors. For logical vectors, use `if_else()`. For more complicated criteria, use `case_when()`.” (<https://dplyr.tidyverse.org/reference/recode.html>)

`forcats::fct_recode` (<https://r4ds.had.co.nz/factors.html>)

1.6.1 Vaihe 1 - muuttujat joissa ei ole puuttuvia tietoja

Aineistosta on jätetty pois ne havainnot, joissa ikä (AGE) tai sukupuoli (SEX) on puuttuva tieto. Aika paljon tarkistuksia, kolme maa-muuttujaan järjestetään C_ALPHAN - muuttujan järjestykseen. Ikä-muuttuja säilyy numeerisena. Ensimmäiseen faktori-tyypin muuttujaan jää tyhjänä luokkana puuttuva tieto, luokka poistetaan.

```
# VAIHE 1 - muuttujat joissa ei ole puuttuvia tietoja

# vaihe 1.1 haven_labelled ja chr -> as_factor

ISSP2012jh1d.dat <- ISSP2012jh1c.data %>%
  mutate(maa = as_factor(C_ALPHAN), # ei puuttuvia, ei tyhjiä leveleitä
         maa3 = as_factor(V3), # maakoodi, jossa aluejako joillan mailla
         sp1 = as_factor(SEX), # ei puuttuvia, tyhjä level "no answer" 999
         )

# C_ALPHAN - maa - maa3 tarkistuksia

# V3
# "Pulma" on järjestys. C_ALPHAN ("chr") on aakkosjärjestysessä, kun luodaan
# maa = as_factor(C_ALPHAN) järjestys muuttuu (esiintymisjärjestys datassa?)
# maa3 muunnetaan maakoodista ('haven_labelled' num), jonka

str(ISSP2012jh1d.dat$maa) #Country Prefix ISO 3166 Code - alphanumeric

## Factor w/ 25 levels "AU","AT","BG",...: 1 1 1 1 1 1 1 1 1 1 ...
## - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
# attributes(ISSP2012jh1d.dat$maa) # ei tyhiä levels-arvoja, 25 levels
# ISSP2012jh1d.dat$maa %>% fct_unique()
# ISSP2012jh1d.dat$maa %>% fct_count() # summary kertoo samat tiedot (20.2.20)
# sum(is.na(ISSP2012jh1d.dat$maa)) # ei puuttuvia tietoja
ISSP2012jh1d.dat$maa %>% summary() # mukana vain valitut 25 maata
```

```

##   AU   AT   BG   CA   HR   CZ   DK   FI   FR   HU   IS   IE   LV   LT   NL   NO
## 1557 1182 1003  953  997 1804 1403 1171 2409 1012 1172 1166 1000 1187 1315 1444
##   PL   RU   SK   SI   SE   CH   BE   DE   PT
## 1115 1525 1128 1034 1059 1237 2192 1761  997
# str(ISSP2012jh1d.dat$maa3)  # "Country/ Sample ISO 3166 Code
#(see V4 for codes for whole nation states)"
# 29 levels
# str(ISSP2012jh1d.dat$V3)

# attributes(ISSP2012jh1d.dat$maa3) # ei tyhiä levels-arvoja, 29 levels
# sum(is.na(ISSP2012jh1d.dat$maa3)) # nolla ei ole puuttuva tieto! (3.2.20)
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
# ISSP2012jh1d.dat$maa3 %>% fct_count()
# Vain näissä on jaettu maan havainnot (3.2.20)
#
# [38] BE-FLA-Belgium/ Flanders
# [39] BE-WAL-Belgium/ Wallonia
# [40] BE-BRU-Belgium/ Brussels
# [41] DE-W-Germany-West
# [42] DE-E-Germany-East
# [43] PT-Portugal 2012: first fieldwork round (main sample)
# [44] PT-Portugal 2012: second fieldwork round (complementary sample)

# ISSP2012jh1d.dat$maa3 %>% fct_count() #miksi ei tulosta mitään? (3.2.2020)

# ISSP2012jh1d.dat$maa3 %>% summary()
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
# maa3: 25 maata, havaintojen määrä. Poisjätetyissä havaintoja 0.
# glimpse(ISSP2012jh1d.dat$maa3)
# head(ISSP2012jh1d.dat$maa3)
# length(levels(ISSP2012jh1d.dat$maa3))

# C_ALPHAN alkuperäinen järjestys, maa aakkosjärjestyssä (2.2.20)
#
# Huom1: Myös merkkijonомуuttujaa C_ALPHAN tarvitaan jatkossa.
#
# Huom2: kun dataa rajataan, on tarkistettava ja tarvittaessa poistettava
# "tyhjät" R-factor - muuttujan "maa" luokat (3.2.2020)

# vaihe 1.2 tyhjät luokat (levels) pois faktoreista

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
  mutate(sp = fct_drop(sp1),
        maa3 = fct_drop(maa3))

```

```

# Poistetaan maa3-muuttujan tyhjät luokat (3.2.20)

# maa3 - tarkistuksia

# str(ISSP2012jh1d.dat$maa3) # 29 levels

attributes(ISSP2012jh1d.dat$maa3) #

## $levels
## [1] "AU-Australia"
## [2] "AT-Austria"
## [3] "BG-Bulgaria"
## [4] "CA-Canada"
## [5] "HR-Croatia"
## [6] "CZ-Czech Republic"
## [7] "DK-Denmark"
## [8] "FI-Finland"
## [9] "FR-France"
## [10] "HU-Hungary"
## [11] "IS-Iceland"
## [12] "IE-Ireland"
## [13] "LV-Latvia"
## [14] "LT-Lithuania"
## [15] "NL-Netherlands"
## [16] "NO-Norway"
## [17] "PL-Poland"
## [18] "RU-Russia"
## [19] "SK-Slovakia"
## [20] "SI-Slovenia"
## [21] "SE-Sweden"
## [22] "CH-Switzerland"
## [23] "BE-FLA-Belgium/ Flanders"
## [24] "BE-WAL-Belgium/ Wallonia"
## [25] "BE-BRU-Belgium/ Brussels"
## [26] "DE-W-Germany-West"
## [27] "DE-E-Germany-East"
## [28] "PT-Portugal 2012: first fieldwork round (main sample)"
## [29] "PT-Portugal 2012: second fieldwork round (complementary sample)"
##
## $class
## [1] "factor"
##
## $label
## [1] "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation states)"

```

```

#sum(is.na(ISSP2012jh1d.dat$maa3)) # nolla ei ole puuttuva tieto! (3.2.20)
# ISSP2012jh1d.dat$maa3 %>% summary()
# ISSP2012jh1d.dat$maa3 %>% fct_unique()
ISSP2012jh1d.dat$maa3 %>% fct_count() # miksi ei tulosta? Tulostaa komentoriviltä!

```

f	n
AU-Australia	1557
AT-Austria	1182
BG-Bulgaria	1003
CA-Canada	953
HR-Croatia	997
CZ-Czech Republic	1804
DK-Denmark	1403
FI-Finland	1171
FR-France	2409
HU-Hungary	1012
IS-Iceland	1172
IE-Ireland	1166
LV-Latvia	1000
LT-Lithuania	1187
NL-Netherlands	1315
NO-Norway	1444
PL-Poland	1115
RU-Russia	1525
SK-Slovakia	1128
SI-Slovenia	1034
SE-Sweden	1059
CH-Switzerland	1237
BE-FLA-Belgium/ Flanders	1090
BE-WAL-Belgium/ Wallonia	543
BE-BRU-Belgium/ Brussels	559
DE-W-Germany-West	1205
DE-E-Germany-East	556
PT-Portugal 2012: first fieldwork round (main sample)	894
PT-Portugal 2012: second fieldwork round (complementary sample)	103

```

str(ISSP2012jh1d.dat$C_ALPHAN)

## chr [1:32823] "AU" ...
## - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## - attr(*, "format.spss")= chr "A20"
## - attr(*, "display_width")= int 22

```

```

attributes(ISSP2012jh1d.dat$C_ALPHAN)

## $label
## [1] "Country Prefix ISO 3166 Code - alphanumeric"
##
## $format.spss
## [1] "A20"
##
## $display_width
## [1] 22

ISSP2012jh1d.dat %>% tableX(C_ALPHAN, maa)

```

C_ALPHAN/BG	CA	HR	CZ	DK	FI	FR	HU	IS	IE	LV	LT	NL	NO	PL	RU	SK	SI	SE	CH	BE	DE	PT
AT	0	11820	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AU	15570	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21920	0
BG	0	0	10030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CA	0	0	0	9530	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12370	0	0
CZ	0	0	0	0	0	18040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17610
DK	0	0	0	0	0	14030	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FI	0	0	0	0	0	0	11710	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FR	0	0	0	0	0	0	0	24090	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HR	0	0	0	0	9970	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HU	0	0	0	0	0	0	0	0	10120	0	0	0	0	0	0	0	0	0	0	0	0	0
IE	0	0	0	0	0	0	0	0	0	11660	0	0	0	0	0	0	0	0	0	0	0	0
IS	0	0	0	0	0	0	0	0	11720	0	0	0	0	0	0	0	0	0	0	0	0	0
LT	0	0	0	0	0	0	0	0	0	0	11870	0	0	0	0	0	0	0	0	0	0	0
LV	0	0	0	0	0	0	0	0	0	0	10000	0	0	0	0	0	0	0	0	0	0	0
NL	0	0	0	0	0	0	0	0	0	0	0	13150	0	0	0	0	0	0	0	0	0	0
NO	0	0	0	0	0	0	0	0	0	0	0	14440	0	0	0	0	0	0	0	0	0	0
PL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99
RU	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SE	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10590	0	0
SI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10340	0	0
SK	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11280	0	0
Total	155711821003953997180414031171240910121172116610001187131514441115152511281034105912372192176199																					

```

ISSP2012jh1d.dat %>% tableX(C_ALPHAN, maa3)

```

```
ISSP2012jh1d.dat %>% tableX(maa, maa3)
```

	AU	AT	BG	CA	HR	Cze	EKF	I	F	R	H	U	S	I	E	L	V	N	O	P	R	S	U	S	W	D	Y	W	E	East	sample
AU155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
AT0	1182	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
BG0	0	1003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
CA0	0	0	9530	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
HR0	0	0	0	9970	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
CZ0	0	0	0	0	1804	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
DK0	0	0	0	0	1403	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
FI0	0	0	0	0	0	1170	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
FR0	0	0	0	0	0	0	2409	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
HU0	0	0	0	0	0	0	0	1012	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
IS0	0	0	0	0	0	0	0	0	1172	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
IE0	0	0	0	0	0	0	0	0	0	1166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
LV0	0	0	0	0	0	0	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
LT0	0	0	0	0	0	0	0	0	0	0	0	1187	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
NL0	0	0	0	0	0	0	0	0	0	0	0	0	1315	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
NO0	0	0	0	0	0	0	0	0	0	0	0	0	0	1444	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
PL0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1115	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
RU0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1525	0	0	0	0	0	0	0	0	0	0	0	0	0		
SK0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1128	0	0	0	0	0	0	0	0	0	0	0			
SI0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1034	0	0	0	0	0	0	0	0	0	0			
SE0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1059	0	0	0	0	0	0	0	0	0	0		
CH0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1237	0	0	0	0	0	0	0	0	0		
BE0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1090543	559	0	0	0	0	0	0	0		
DE0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1205560	0	0	0			
PT0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	894		
Total	557182003539971804403172409012172166000187315444115525128034059237090543	559	120556894	103																											

ISSP2012jh1d.dat %>% **tableX**(V3, maa3)

sp, sp1, SEX - tarkistuksia

```
ISSP2012jh1d.dat$sp %>% fct_count()
```

	f	n
Male	14789	
Female	18034	

```
ISSP2012jh1d.dat$sp %>% fct_count()
```

	f	n
Male	14789	
Female	18034	

```
ISSP2012jh1d.dat %>% tableX(SEX,sp1)
```

SEX/sp1	Male	Female	No answer	Total
1	14789	0	0	14789
2	0	18034	0	18034
Total	14789	18034	0	32823

```
ISSP2012jh1d.dat %>% tableX(SEX,sp)
```

SEX/sp	Male	Female	Total
1	14789	0	14789
2	0	18034	18034
Total	14789	18034	32823

```
ISSP2012jh1d.dat %>% tableX(sp1,sp)
```

sp1/sp	Male	Female	Total
Male	14789	0	14789
Female	0	18034	18034
No answer	0	0	0
Total	14789	18034	32823

```
# vaihe 1.3 uudet "faktorilabelit"  
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%  
  mutate(sp =
```

```

fct_recode(sp,
           "m" = "Male",
           "f" = "Female")
)

# Tarkistuksia

ISSP2012jh1d.dat$sp %>% fct_unique()

## [1] m f
## Levels: m f

ISSP2012jh1d.dat$sp %>% fct_count()



|   | f     | n |
|---|-------|---|
| m | 14789 |   |
| f | 18034 |   |



ISSP2012jh1d.dat$sp %>% summary()

##      m      f
## 14789 18034
# AGE -> ika
# AGE----
ISSP2012jh1d.dat$ika <- ISSP2012jh1d.dat$AGE

# Tarkistuksia
attributes(ISSP2012jh1d.dat$ika) # tyhjä level "No answer"

## $label
## [1] "Age of respondent"
##
## $format.spss
## [1] "F3.0"
##
## $labels
## 15 years 16 years 17 years 18 years 102 years No answer
##          15        16        17        18       102       999
##
## $class
## [1] "haven_labelled" "vctrs_vctr"     "double"
# str(ISSP2012jh1d.dat$ika)
ISSP2012jh1d.dat$ika %>% summary()

```

		[1]	[2]
[1]AGE	[1]AGE	1.00	
[2]ika	[2]ika	1.00	1.00

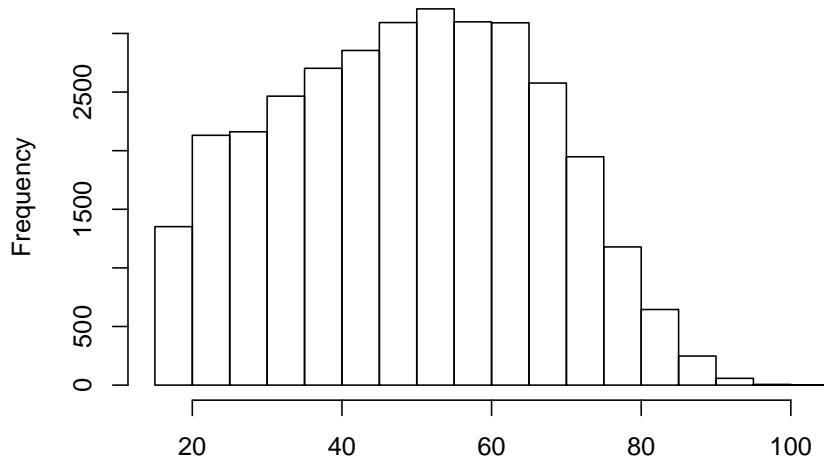
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15	36	50	49.51607	63	102

```
ISSP2012jh1d.dat %>%
  tableC(AGE, ika, cor_type = "pearson", na.rm = FALSE, rounding = 5,
         output = "text", booktabs = TRUE, caption = NULL, align = NULL,
         float = "htb") %>% kable()

## N = 32823
## Note: pearson correlation (p-value).

ISSP2012jh1d.dat$ika %>% hist(main = "ISSP 2012: vastaajan ikä")
```

ISSP 2012: vastaajan ikä



```
# str(ISSP2012jh1d.dat) - tarkistus
```

1.6.2 Vaihe 2

Vaihessa 2 luodaan samalla samalla periaatteella substanssi- ja taustamuuttujille kaksi R-factor- tyypin muuttuja. Toisessa (esim. Q1a) puuttuva tieto on R-ohjelmiston sisäinen NA-arvo. Toisessa (Q1am) puuttuva tieto on yksi luokittelumuuttujan arvo(“missing”).

1.6.3 Vaihe 2.1

```
# Substanssi- ja taustamuuttujat R-faktoreiksi
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
  mutate(Q1a1 = as_factor(V5), #labels
         Q1b1 = as_factor(V6),
         Q1c1 = as_factor(V7),
         Q1d1 = as_factor(V8),
         Q1e1 = as_factor(V9),
         Q2a1 = as_factor(V10),
         Q2b1 = as_factor(V11),
         Q3a1 = as_factor(V12), #labels = vastQ3_labels (W,w,H)
         Q3b1 = as_factor(V13), #labels = vastQ3_labels
         edu1 = as_factor(DEGREE),
         msta1 = as_factor(MAINSTAT),
         sosta1 = as_factor(TOPBOT),
         nchild1 = as_factor(HHCHILD),
         lifsta1 = as_factor(MARITAL),
         urbru1 = as_factor(URBRURAL)

      )

# Muuttujat Q1a1...urbru1 ovat apumuuttuja, joissa on periaatteessa kaikki SPSS-
# tiedostosta siirtyvä metatieto. Poikkeus on SPSS:n kolme tarkentavaa koodia
# puuttuvalle tiedolle, ne saisi mukaan read_spss - parametrin avulla (user_na=TRUE)
#
# Tarkistuskset
# ISSP2012jh1d.dat %>% summary()

ISSP2012jh1d.dat %>%
  select(Q1a1, Q1b1, Q1c1, Q1d1, Q1e1, Q2a1, Q2b1, Q3a1, Q3b1) %>%
  summary()
```

Q1a1	Q1b1	Q1c1	Q1d1	Q1e1	Q2a1	Q2b1	Q3a1	Q3b1
Agree :12352	Disagree :9003	Disagree :8706	Disagree :7863	Agree :8342	Agree :13464	Disagree :9210	Work full-time : 5373	Work full-time :13722

Q1a1	Q1b1	Q1c1	Q1d1	Q1e1	Q2a1	Q2b1	Q3a1	Q3b1
Strongly Agree :8389 agree :11116	Agree :8263	Agree :7672	Neither agree nor	Strongly agree :11305 disagree:7841	Strongly disagree :8917	Work part-time :15655	Work part-time :13817	
Disagree :4074	Neither agree :nor nor	Neither agree :nor nor	Neither agree :nor di-	Disagree :7267 neither agree :7403	Neither agree :sagree:	Stay at home :8367 disagree:6109	Stay at home :1762	
Neither agree :5547 nor disagree :5960	Strongly di- :5960	Strongly di- :5016	Strongly di- :3462	Disagree :1929 sagree :2704	Agree :5164 sagree :2704	TW: women should decide :0	TW: women should decide :0	
Strongly disagree :2747 : 1051	Strongly agree :2838	Strongly agree :2818	Strongly agree :3357	Strongly agree :403	Strongly agree :403	Can't choose, CA:+NA, KR:DK,ref.KR:DK,ref., NL:DK: 0	Can't choose, CA:+NA, KR:DK,ref.KR:DK,ref., NL:DK: 0	
(Other) : 0	(Other) : 0	(Other) : 0	(Other) : 0	(Other) : 0	(Other) : 0	No answer : 0	No answer : 0	
NA's : 848	NA's :1188	NA's :1056	NA's :2051	NA's :2554	NA's : 683	NA's : 719	NA's : 3428	NA's : 3522

ISSP2012jh1d.dat %>%

```
select(edu1,msta1, sosta1, nchchild1, lifsta1, urbru1) %>%
summary()
```

edu1	msta1	sosta1	hchild1	lifsta1	urbru1
Lower secondary (secondary completed does not allow entry to university: obligatory school) :7811	In paid work :17967	06 :6889 children :75102	No children :75102	Married or a small city :9203	A town or a small city :9203
Upper secondary (programs that allows entry to university :7115	Retired : 7999	05 :6798 child :4378	One married/ never in a civil partnership, single : 7535	Never married/ never in a village :8646	A country village :8646

edu1	msta1	sosta1	child1	lifsta1	urbru1
Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formation):5658	Unemployed and looking for a job, HR: incl never had a job: 1769	07 :5778 2 : 2643	Divorced from spouse/ legally separated from civil partner: 2997	A big city :8442	
Lower level tertiary, first stage (also technical schools at a tertiary level) :5147	In education : 1763	08 :3477598	Widowed/ civil partner died : 2763	The suburbs or outs- skirts of a big city:4386	
Upper level tertiary (Master, Dr.) :4762	Domestic work : 1180	04 :3346117	Civil partnership : 1035	A farm or home in the country :1902	
(Other) :2022	(Other) : 1775	(Other) :1753	(Other) : 486	(Other) : 0	
NA's : 308	NA's : 370	NA's : NA's : :1777:	NA's : 434	NA's : 244	
		940			

```
# Substanssimuuttujat - ristiintaulukoinnit riittävät (6.2.20)
```

```
# ISSP2012jh1d.dat$Q1a1 %>% fct_count()
# ISSP2012jh1d.dat$Q1b1 %>% fct_count()
# ISSP2012jh1d.dat$Q1c1 %>% fct_count()
# ISSP2012jh1d.dat$Q1d1 %>% fct_count()
# ISSP2012jh1d.dat$Q1e1 %>% fct_count()
# ISSP2012jh1d.dat$Q2a1 %>% fct_count()
# ISSP2012jh1d.dat$Q2b1 %>% fct_count()
# ISSP2012jh1d.dat$Q3a1 %>% fct_count()
#ISSP2012jh1d.dat$Q3b1 %>% fct_count()
```

```
# Taustamuuttujat - ristiintaulukoinnit riittävät (6.2.20)
```

```
# ISSP2012jh1d.dat$edu1 %>% fct_count()
# ISSP2012jh1d.dat$msta1 %>% fct_count()
# ISSP2012jh1d.dat$sosta1 %>% fct_count()
# ISSP2012jh1d.dat$ncchild1 %>% fct_count()
# ISSP2012jh1d.dat$lifsta1 %>% fct_count()
```

```
# ISSP2012jh1d.dat$urbru1 %>% fct_count()
```

Taustamuuttujien luokitteluja (esim. luokkien yhdistäminen) pohditaan tarkemmin, kun muuttujat otetaan käyttöön.

1.6.4 Vaihe 2.2

Poistetaan muuuttujista luokittelumuuttujien arvot, joissa ei ole havaintoja. Näitä tyhjiä luokkia siirryy SPSS-tiedostosta haven_labelled -luokan tietoihin.

```
# Poistetaan tyhjät luokat muuttujista
```

```
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
  mutate(Q1a = fct_drop(Q1a1),
         Q1b = fct_drop(Q1b1),
         Q1c = fct_drop(Q1c1),
         Q1d = fct_drop(Q1d1),
         Q1e = fct_drop(Q1e1),
         Q2a = fct_drop(Q2a1),
         Q2b = fct_drop(Q2b1),
         Q3a = fct_drop(Q3a1),
         Q3b = fct_drop(Q3b1),
         edu = fct_drop(edu1),
         msta = fct_drop(msta1),
         sosta = fct_drop(sosta1),
         nchild = fct_drop(nchild1),
         lifsta = fct_drop(lifsta1),
         urbru = fct_drop(urbru1)

  )
# Tarkistuksia 1

ISSP2012jh1d.dat %>% summary()
```



```
ISSP2012jh1d.dat %>%
  select(Q1a, Q1b, Q1c, Q1d, Q1e, Q2a, Q2b, Q3a, Q3b) %>%
  str()

## # tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
## # $ Q1a: Factor w/ 5 levels "Strongly agree",...: 5 1 2 2 1 NA 2 4 2 2 ...
## # ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not work"
## # $ Q1b: Factor w/ 5 levels "Strongly agree",...: 1 5 4 4 4 NA 4 3 4 3 ...
## # ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## # $ Q1c: Factor w/ 5 levels "Strongly agree",...: 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"
## # $ Q1d: Factor w/ 5 levels "Strongly agree",...: 3 5 5 2 4 NA 4 5 4 5 ...
## # ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## # $ Q1e: Factor w/ 5 levels "Strongly agree",...: 3 1 2 3 4 NA 2 4 4 1 ...
## # ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working outside home"
## # $ Q2a: Factor w/ 5 levels "Strongly agree",...: 1 3 4 2 2 NA 2 5 2 1 ...
## # ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## # $ Q2b: Factor w/ 5 levels "Strongly agree",...: 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"
## # $ Q3a: Factor w/ 3 levels "Work full-time",...: 3 NA NA 2 2 NA 2 NA 2 2 ...
## # ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
## # $ Q3b: Factor w/ 3 levels "Work full-time",...: 2 NA 2 1 2 NA 2 NA 2 2 ...
## # ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa"
```

```

ISSP2012jh1d.dat %>%
  select(Q1a1, Q1b1, Q1c1, Q1d1, Q1e1,Q2a1,Q2b1,Q3a1, Q3b1) %>%
  str()

## # tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
## $ Q1a1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 6 2 3 3 2 NA 3 5 3 3 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
## $ Q1b1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 2 6 5 5 5 NA 5 4 5 4 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ Q1c1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 4 6 3 5 5 NA 5 3 5 3 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
## $ Q1d1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 4 6 6 3 5 NA 5 6 5 6 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## $ Q1e1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 4 2 3 4 5 NA 3 5 5 2 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working outside the home"
## $ Q2a1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 2 4 5 3 3 NA 3 6 3 2 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## $ Q2b1: Factor w/ 8 levels "NAP: ES","Strongly agree",...: 4 6 5 5 5 NA 3 6 5 2 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
## $ Q3a1: Factor w/ 6 levels "Work full-time",...: 3 NA NA 2 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
## $ Q3b1: Factor w/ 6 levels "Work full-time",...: 2 NA 2 1 2 NA 2 NA 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa

ISSP2012jh1d.dat %>%
  select(edu, msta, sosta, nchild,lifsta, urbru) %>%
  str()

## # tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
## $ edu : Factor w/ 7 levels "No formal education",...: 3 6 6 4 3 NA NA 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for international comparison"
## $ msta : Factor w/ 9 levels "In paid work",...: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
## $ sosta : Factor w/ 10 levels "Lowest, Bottom, 01",...: 3 7 8 NA 7 2 7 NA 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
## $ nchild: Factor w/ 11 levels "No children",...: NA NA 4 2 1 NA 1 1 2 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age] and [age of child]"
## $ lifsta: Factor w/ 6 levels "Married","Civil partnership",...: 6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
## $ urbru : Factor w/ 5 levels "A big city","The suburbs or outskirts of a big city",...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa

ISSP2012jh1d.dat %>%
  select(edu1, msta1, sosta1, nchild1,lifsta1, urbru1) %>%
  str()

```

```

## tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
## $ edu1 : Factor w/ 8 levels "No formal education",...: 3 6 6 4 3 NA NA 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for intern...
## $ msta1 : Factor w/ 10 levels "In paid work",...: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
## $ sosta1 : Factor w/ 14 levels "Not available: GB,US",...: 4 8 9 NA 8 3 8 NA 11 7 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
## $ nchild1: Factor w/ 14 levels "No children",...: NA NA 4 2 1 NA 1 1 2 NA ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age ...
## $ lifsta1: Factor w/ 9 levels "Married","Civil partnership",...: 6 1 1 6 1 6 1 1 1 NA ...
##   ..- attr(*, "label")= chr "Legal partnership status"
## $ urbru1 : Factor w/ 7 levels "A big city","The suburbs or outskirts of a big city",...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa...
# Tarkistuksia 2 - ristiintaulukointi Q1a/Q1am riittää (6.2.20)

# Substanssimuuttujat

# ISSP2012jh1d.dat %>% tableX(Q1a,Q1a1)
# ISSP2012jh1d.dat %>% tableX(Q1b,Q1b1)
# ISSP2012jh1d.dat %>% tableX(Q1c,Q1c1)
# ISSP2012jh1d.dat %>% tableX(Q1d,Q1d1)
# ISSP2012jh1d.dat %>% tableX(Q1e,Q1e1)
# ISSP2012jh1d.dat %>% tableX(Q2a,Q2a1)
# ISSP2012jh1d.dat %>% tableX(Q2b,Q2b1)
# ISSP2012jh1d.dat %>% tableX(Q3a,Q3a1)
# ISSP2012jh1d.dat %>% tableX(Q3b,Q3b1)

# Taustamuuttujat

# ISSP2012jh1d.dat %>% tableX(edu,edu1)
# ISSP2012jh1d.dat %>% tableX(msta,msta1)
# ISSP2012jh1d.dat %>% tableX(sosta,sosta1)
# ISSP2012jh1d.dat %>% tableX(nchild,nchild1)
# ISSP2012jh1d.dat %>% tableX(lifsta,lifsta1)
# ISSP2012jh1d.dat %>% tableX(urbru,urbru1)

```

1.6.5 Vaihe 2.3

Luodaan uusi muuttuja, jossa puuttuva tieto (NA) on mukana luokittelumuuttujan uutena arvona ("missing").

```
# Uusi muuttuja, jossa NA-arvot ovat mukana muuttujan uutena luokkana. Muuttujat
# nimetään Q1a -> Q1am.
```

```
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
```

```

    mutate(Q1am = fct_explicit_na(Q1a, na_level = "missing"),
           Q1bm = fct_explicit_na(Q1b, na_level = "missing"),
           Q1cm = fct_explicit_na(Q1c, na_level = "missing"),
           Q1dm = fct_explicit_na(Q1d, na_level = "missing"),
           Q1em = fct_explicit_na(Q1e, na_level = "missing"),
           Q2am = fct_explicit_na(Q2a, na_level = "missing"),
           Q2bm = fct_explicit_na(Q2b, na_level = "missing"),
           Q3am = fct_explicit_na(Q3a, na_level = "missing"),
           Q3bm = fct_explicit_na(Q3b, na_level = "missing"),
           edum = fct_explicit_na(edu, na_level = "missing"),
           mstam = fct_explicit_na(msta, na_level = "missing"),
           sostam = fct_explicit_na(sosta, na_level = "missing"),
           nchilddm = fct_explicit_na(ncchild, na_level = "missing"),
           lifstam = fct_explicit_na(lifsta, na_level = "missing"),
           urbrum = fct_explicit_na(urbru, na_level = "missing"),
           )
# Tarkistuksia 3

ISSP2012jh1d.dat %>%
  select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
  summary()

```

Q1am	Q1bm	Q1cm	Q1dm	Q1em	Q2am	Q2bm	Q3am	Q3bm
Strongly agree :11116	Strongly agree :2747	Strongly agree :2838	Strongly agree :2818	Strongly agree :3357	Strongly agree :11305	Strongly agree :2704	Work full-time: 13722	Work full-time: 5373
Agree :12352	Agree :8389	Agree :8263	Agree :7672	Agree :8342	Agree :13464	Agree :5164	Work part-time: 15655	Work part-time: 13817
Neither agree nor disagree: 3382	Neither agree nor disagree: 5919	Neither agree nor disagree: 6019	Neither agree nor disagree: 7403	Neither agree nor disagree: 7841	Neither agree nor disagree: 5039	Stay at home: 8367	Stay at home: 1762	me : 8367
Disagree : 4074	Disagree : 9003	Disagree : 8706	Disagree : 7863	Disagree : 7267	Disagree : 1929	Disagree : 9210	missing : 3428	missing : 3522
Strongly disagree : 1051	Strongly disagree : 5547	Strongly disagree : 5960	Strongly disagree : 5016	Strongly disagree : 3462	Strongly disagree : 403	Strongly disagree : 8917	NA : 719	NA : 719
missing : 848	missing : 1188	missing : 1056	missing : 2051	missing : 2554	missing : 683	missing : 719	NA : 719	NA : 719

ISSP2012jh1d.dat %>%

```
select(edum,mstam, sostam,nchilf, lilstam, urbrum) %>%
summary()
```

edum	mstam	sostam	nchilf	lilstam	urbrum
Lower secondary (secondary completed does not allow entry to university: obligatory school) :7811	In paid work :17967	06 :6889	No children:24102	Married :17573	A big city :8442
Upper secondary (programs that allows entry to university :7115	Retired :7999	05 :6798	One child :1035	Civil partnership :4378	The suburbs or outskirts of a big city:4386
Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formation):5658	Unemployed and looking for a job, HR: incl never had a job: 1769	07 :5778	Separated from spouse/ civil partner (still legally married/ still legally in a civil partnership):486	Divorced from spouse/ legally separated from civil partner :2997	A town or a small city :9203
Lower level tertiary, first stage (also technical schools at a tertiary level) :5147	In education :1763	08 :3477	missing	Widowed/ civil partner died : 2763	A country village :8646
Upper level tertiary (Master, Dr.) :4762	Domestic work : 1180	04 :3346	3 : 598	missing	A farm or home in the country :1902
Primary school (elementary school) :1531 (Other) : 799	Permanently sick or disabled : 1093 (Other) : 1052	03 : 2221	4 : 117	Never married/ never in a civil partnership, single : 7535 (Other) : 434 missing : 434	missing : 244 NA

```

ISSP2012jh1d.dat %>%
  select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
  str()

## tibble [32,823 x 9] (S3: tbl_df/tbl/data.frame)
## $ Q1am: Factor w/ 6 levels "Strongly agree",...: 5 1 2 2 1 6 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
## $ Q1bm: Factor w/ 6 levels "Strongly agree",...: 1 5 4 4 4 6 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ Q1cm: Factor w/ 6 levels "Strongly agree",...: 3 5 2 4 4 6 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
## $ Q1dm: Factor w/ 6 levels "Strongly agree",...: 3 5 5 2 4 6 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## $ Q1em: Factor w/ 6 levels "Strongly agree",...: 3 1 2 3 4 6 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working outside the home"
## $ Q2am: Factor w/ 6 levels "Strongly agree",...: 1 3 4 2 2 6 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## $ Q2bm: Factor w/ 6 levels "Strongly agree",...: 3 5 4 4 4 6 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
## $ Q3am: Factor w/ 4 levels "Work full-time",...: 3 4 4 2 2 4 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q3a Should women work: Child under school age"
## $ Q3bm: Factor w/ 4 levels "Work full-time",...: 2 4 2 1 2 4 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q3b Should women work: Youngest kid at school"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa"

ISSP2012jh1d.dat %>%
  select(edum,mstam, sostam,nchilidm,lifstam, urbrum) %>%
  str()

## tibble [32,823 x 6] (S3: tbl_df/tbl/data.frame)
## $ edum : Factor w/ 8 levels "No formal education",...: 3 6 6 4 3 8 8 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for international comparison"
## $ mstam : Factor w/ 10 levels "In paid work",...: 6 6 3 1 6 5 6 2 1 5 ...
##   ..- attr(*, "label")= chr "Main status"
## $ sostam : Factor w/ 11 levels "Lowest, Bottom, 01",...: 3 7 8 11 7 2 7 11 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
## $ nchilidm: Factor w/ 12 levels "No children",...: 12 12 4 2 1 12 1 1 2 12 ...
##   ..- attr(*, "label")= chr "How many children in household: children between [school age] and [age of child]"
## $ lifstam: Factor w/ 7 levels "Married","Civil partnership",...: 6 1 1 6 1 6 1 1 1 7 ...
##   ..- attr(*, "label")= chr "Legal partnership status"
## $ urbrum : Factor w/ 6 levels "A big city","The suburbs or outskirts of a big city",...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa

# Taustamuuttuja, puuttuva tieto mukana - ristintaulkointi riittää (6.2.20)

# ISSP2012jh1d.dat$edum %>% fct_count()

```

```

# ISSP2012jh1d.dat$mstam %>% fct_count()
# ISSP2012jh1d.dat$sostam %>% fct_count()
# ISSP2012jh1d.dat$nchildm %>% fct_count()
# ISSP2012jh1d.dat$lifstam %>% fct_count()
# ISSP2012jh1d.dat$urbrum %>% fct_count()

# Substanssimuuttujat, puuttuva tieto mukana - ristiintaulkointi riittää (6.2.20)

# ISSP2012jh1d.dat$Q1am %>% fct_count()
# ISSP2012jh1d.dat$Q1bm %>% fct_count()
# ISSP2012jh1d.dat$Q1cm %>% fct_count()
# ISSP2012jh1d.dat$Q1dm %>% fct_count()
# ISSP2012jh1d.dat$Q1em %>% fct_count()
# ISSP2012jh1d.dat$Q2am %>% fct_count()
# ISSP2012jh1d.dat$Q2bm %>% fct_count()
# ISSP2012jh1d.dat$Q3am %>% fct_count()
# ISSP2012jh1d.dat$Q3bm %>% fct_count()

```

1.6.6 Vaihe 2.4

Lopuksi luodaan uudet “faktorilabelit” substanssimuuttujille. Näkyvät komennolla levels(). Graafisessa analyysissä kuvia on saatava mukaan kaikki oleellinen, mutta ei mitään sen lisäksi. Näitä muuttujan arvojen tunnuksia muokataan tarvittaessa.

Taustamuuttujien “faktorilabeleita” säädetään kun ne otetaan käyttöön.

```
# Vaihe 2.4.1
```

```

# Q1a - Q1e, Q2a, Q2b Viisi vastausvaihtoehtoa - ei eksplisiittistä NA-tietoa("missing")
# Q3a - Q3b kolme vastausvaihtoehtoa

ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
  mutate(Q1a = fct_recode(Q1a,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
  Q1b = fct_recode(Q1b,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
  Q1c = fct_recode(Q1c,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"))

```

```

    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
Q1d = fct_recode(Q1d,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
Q1e = fct_recode(Q1e,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
Q2a = fct_recode(Q2a,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree" ),
Q2b = fct_recode(Q2b,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree"),
Q3a = fct_recode(Q3a,
    "W" = "Work full-time",
    "w" = "Work part-time",
    "H" = "Stay at home" ),
Q3b = fct_recode(Q3b,
    "W" = "Work full-time",
    "w" = "Work part-time",
    "H" = "Stay at home" )
)

# Tarkistuksia 1
ISSP2012jh1d.dat %>%
  select(Q1a, Q1b, Q1c, Q1d, Q1e, Q2a, Q2b, Q3a, Q3b) %>%
  summary()

```

Q1a	Q1b	Q1c	Q1d	Q1e	Q2a	Q2b	Q3a	Q3b
S :11116	S :2747	S :2838	S :2818	S :3357	S :11305	S :2704	W : 5373	W : 11116
s :12352	s :8389	s :8263	s :7672	s :8342	s :13464	s :5164	w :15655	w :12352
? : 3382	? :5949	? :6000	? :7403	? :7841	? : 5039	? :6109	H : 8367	H : 3382
e : 4074	e :9003	e :8706	e :7863	e :7267	e : 1929	e :9210	NA's: 3428	NA : 4074
E : 1051	E :5547	E :5960	E :5016	E :3462	E : 403	E :8917	NA	NA : 1051
NA's: 848	NA's:1188	NA's:1056	NA's:2051	NA's:2554	NA's: 683	NA's: 719	NA	NA : 848

```
# Vaihe 2.4.2 - muuttujassa eksplisiittinen NA-tieto
ISSP2012jh1d.dat <- ISSP2012jh1d.dat %>%
  mutate(Q1am = fct_recode(Q1am,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree",
    "P" = "missing"),
  Q1bm = fct_recode(Q1bm,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree",
    "P" = "missing"),
  Q1cm = fct_recode(Q1cm,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree",
    "P" = "missing"),
  Q1dm = fct_recode(Q1dm,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree",
    "P" = "missing"),
  Q1em = fct_recode(Q1em,
    "S" = "Strongly agree",
    "s" = "Agree",
    "?" = "Neither agree nor disagree",
    "e" = "Disagree",
    "E" = "Strongly disagree",
```

```

        "P" = "missing"),
Q2am = fct_recode(Q2am,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
Q2bm = fct_recode(Q2bm,
                    "S" = "Strongly agree",
                    "s" = "Agree",
                    "?" = "Neither agree nor disagree",
                    "e" = "Disagree",
                    "E" = "Strongly disagree",
                    "P" = "missing"),
Q3am = fct_recode(Q3am,
                    "W" = "Work full-time",
                    "w" = "Work part-time",
                    "H" = "Stay at home",
                    "P" = "missing"),
Q3bm = fct_recode(Q3bm,
                    "W" = "Work full-time",
                    "w" = "Work part-time",
                    "H" = "Stay at home",
                    "P" = "missing")
    )
# Tarkistuksia 4

ISSP2012jh1d.dat %>%
  select(Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm) %>%
  summary()

```

Q1am	Q1bm	Q1cm	Q1dm	Q1em	Q2am	Q2bm	Q3am	Q3bm
S:11116	S:2747	S:2838	S:2818	S:3357	S:11305	S:2704	W: 5373	W:13722
s:12352	s:8389	s:8263	s:7672	s:8342	s:13464	s:5164	w:15655	w:13817
?: 3382	?:5949	?:6000	?:7403	?:7841	?: 5039	?:6109	H: 8367	H: 1762
e: 4074	e:9003	e:8706	e:7863	e:7267	e: 1929	e:9210	P: 3428	P: 3522
E: 1051	E:5547	E:5960	E:5016	E:3462	E: 403	E:8917	NA	NA
P: 848	P:1188	P:1056	P:2051	P:2554	P: 683	P: 719	NA	NA

```
# Tarkistuksia 5
```

```
# Substanssimuuttuja
```

ISSP2012jh1d.dat %>%

tableX(Q1a,Q1am)

Q1a/Q1am	S	s	?	e	E	P	Total
1	11116	0	0	0	0	0	11116
2	0	12352	0	0	0	0	12352
3	0	0	3382	0	0	0	3382
4	0	0	0	4074	0	0	4074
5	0	0	0	0	1051	0	1051
Missing	0	0	0	0	0	848	848
Total	11116	12352	3382	4074	1051	848	32823

ISSP2012jh1d.dat %>%

tableX(Q1b,Q1bm)

Q1b/Q1bm	S	s	?	e	E	P	Total
1	2747	0	0	0	0	0	2747
2	0	8389	0	0	0	0	8389
3	0	0	5949	0	0	0	5949
4	0	0	0	9003	0	0	9003
5	0	0	0	0	5547	0	5547
Missing	0	0	0	0	0	1188	1188
Total	2747	8389	5949	9003	5547	1188	32823

ISSP2012jh1d.dat %>%

tableX(Q1c,Q1cm)

Q1c/Q1cm	S	s	?	e	E	P	Total
1	2838	0	0	0	0	0	2838
2	0	8263	0	0	0	0	8263
3	0	0	6000	0	0	0	6000
4	0	0	0	8706	0	0	8706
5	0	0	0	0	5960	0	5960
Missing	0	0	0	0	0	1056	1056
Total	2838	8263	6000	8706	5960	1056	32823

ISSP2012jh1d.dat %>%

tableX(Q1d,Q1dm)

Q1d/Q1dm	S	s	?	e	E	P	Total
1	2818	0	0	0	0	0	2818
2	0	7672	0	0	0	0	7672
3	0	0	7403	0	0	0	7403
4	0	0	0	7863	0	0	7863
5	0	0	0	0	5016	0	5016
Missing	0	0	0	0	0	2051	2051
Total	2818	7672	7403	7863	5016	2051	32823

ISSP2012jh1d.dat %>%
 tableX(Q1e,Q1em)

Q1e/Q1em	S	s	?	e	E	P	Total
1	3357	0	0	0	0	0	3357
2	0	8342	0	0	0	0	8342
3	0	0	7841	0	0	0	7841
4	0	0	0	7267	0	0	7267
5	0	0	0	0	3462	0	3462
Missing	0	0	0	0	0	2554	2554
Total	3357	8342	7841	7267	3462	2554	32823

ISSP2012jh1d.dat %>%
 tableX(Q2a,Q2am)

Q2a/Q2am	S	s	?	e	E	P	Total
1	11305	0	0	0	0	0	11305
2	0	13464	0	0	0	0	13464
3	0	0	5039	0	0	0	5039
4	0	0	0	1929	0	0	1929
5	0	0	0	0	403	0	403
Missing	0	0	0	0	0	683	683
Total	11305	13464	5039	1929	403	683	32823

ISSP2012jh1d.dat %>%
 tableX(Q2b,Q2bm)

Q2b/Q2bm	S	s	?	e	E	P	Total
1	2704	0	0	0	0	0	2704
2	0	5164	0	0	0	0	5164
3	0	0	6109	0	0	0	6109
4	0	0	0	9210	0	0	9210

Q2b/Q2bm	S	s	?	e	E	P	Total
5	0	0	0	0	8917	0	8917
Missing	0	0	0	0	0	719	719
Total	2704	5164	6109	9210	8917	719	32823

```
ISSP2012jh1d.dat %>%
  tableX(Q3a,Q3am)
```

Q3a/Q3am	W	w	H	P	Total
1	5373	0	0	0	5373
2	0	15655	0	0	15655
3	0	0	8367	0	8367
Missing	0	0	0	3428	3428
Total	5373	15655	8367	3428	32823

```
ISSP2012jh1d.dat %>%
  tableX(Q3b,Q3bm)
```

Q3b/Q3bm	W	w	H	P	Total
1	13722	0	0	0	13722
2	0	13817	0	0	13817
3	0	0	1762	0	1762
Missing	0	0	0	3522	3522
Total	13722	13817	1762	3522	32823

```
ISSP2012jh1d.dat %>% # tableX muotoilee taulukkoa!
  tableX(Q3am,Q3a)
```

Q3am/Q3a	1	2	3	Missing	Total
W	5373	0	0	0	5373
w	0	15655	0	0	15655
H	0	0	8367	0	8367
P	0	0	0	3428	3428
Total	5373	15655	8367	3428	32823

```
ISSP2012jh1d.dat$Q3a %>% levels()
## [1] "W"  "w"  "H"
```

```
ISSP2012jh1d.dat$Q3am %>% levels()
```

```
## [1] "W" "w" "H" "P"
# Taustamuuttujat
```

```
ISSP2012jh1d.dat %>%
  tableX(edu, edum)
```

	Lower secondary (secondary completed Primary does not school allow entry No (ele- for- men- university: mal tary edu (education school)	Upper secon- dary (pro- grams that allows entry to university	Post secondary, non-tertiary (other upper secondary programs toward labour market or technical formation)	Lower tertiary, first stage (also labour market or technical formation)	Upper tertiary (Mas- ter, Dr.)	miss	Total	
1	491	0	0	0	0	0	491	
2	0	1531	0	0	0	0	1531	
3	0	0	7811	0	0	0	7811	
4	0	0	0	7115	0	0	7115	
5	0	0	0	0	5658	0	5658	
6	0	0	0	0	0	5147	0	5147
7	0	0	0	0	0	4762	0	4762
Missing	0	0	0	0	0	0	308308	
Total	491	1531	7811	7115	5658	5147	4762	30832823

```
ISSP2012jh1d.dat %>%
```

```
tableX(msta, mstam)
```

	Unemployed and looking In for a job, HR: paid incl never msta/mstam had a job	In compulsory military	Permanently service or Domestic community work service	In
1	179670	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
2	0 1769	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
3	0 0 1763	0 0 0 0 0	0 0 0 0 0	0 0 0 0 0
4	0 0 0 189	0 0 0 0 0	0 0 0 0 0	0 0 0 0 189
5	0 0 0 1093	0 0 0 0 0	0 0 0 0 0	0 0 0 0 1093

msta/ workdm	In paid incl never had a job	Unemployed and looking for a job, HR:			Apprenti- ce or education			Permanently sick or disabled			Domest- ic work			In compulsory military service or community service			Other miss- ing			Total
		In workdm	In workdm	In workdm	or education	appren- tice	educa- tion	sick or disabled	Perma- nently disabled	Retired	Domest- ic work	Community service	Domest- ic work	Community service	Retired	Domest- ic work	Community service	Retired	Domest- ic work	Community service
6	0	0			0	0	0		79990	0				0	0	0	0	0	0	7999
7	0	0			0	0	0		0	1180	0			0	0	0	0	0	0	1180
8	0	0			0	0	0		0	0	9			0	0	0	0	0	0	9
9	0	0			0	0	0		0	0	0			484	0	0	0	0	0	484
Missing	0				0	0	0		0	0	0			0	370	370	0	0	0	370
Total	17967	1769			1763	189	1093		79991	1180	9			484	370	370	484	370	32823	

ISSP2012jh1d.dat %>%

tableX(sosta, sostam)

sosta/sostam	Lowest, Bottom,									Highest, Top, 10				missing	Total
	02	03	04	05	06	07	08	09	Top, 10	missing	Total				
1	562	0	0	0	0	0	0	0	0	0	0	0	0	562	
10	0	0	0	0	0	0	0	0	442	0	442	0	0	0	442
2	0	866	0	0	0	0	0	0	0	0	0	0	0	866	
3	0	0	2221	0	0	0	0	0	0	0	0	0	0	2221	
4	0	0	0	3346	0	0	0	0	0	0	0	0	0	3346	
5	0	0	0	0	6798	0	0	0	0	0	0	0	0	6798	
6	0	0	0	0	0	6889	0	0	0	0	0	0	0	6889	
7	0	0	0	0	0	0	5778	0	0	0	0	0	0	5778	
8	0	0	0	0	0	0	0	3477	0	0	0	0	0	3477	
9	0	0	0	0	0	0	0	0	667	0	0	0	0	667	
Missing	0	0	0	0	0	0	0	0	0	0	0	1777	1777	1777	1777
Total	562	866	2221	3346	6798	6889	5778	3477	667	442	442	1777	1777	32823	

ISSP2012jh1d.dat %>%

tableX(nchild, nchildm)

nchild/nchildm	No children	One child	2 children	3	4	5	6	7	8	18	21 children	missing	Total
	nchild/nchildm	nchildm	nchildm	nchildm	nchildm	nchildm	nchildm	nchildm	nchildm	nchildm	nchildm		
1	24102	0	0	0	0	0	0	0	0	0	0	0	24102
10	0	0	0	0	0	0	0	0	0	1	0	0	1
11	0	0	0	0	0	0	0	0	0	0	1	0	1
2	0	4378	0	0	0	0	0	0	0	0	0	0	4378
3	0	0	2643	0	0	0	0	0	0	0	0	0	2643
4	0	0	0	598	0	0	0	0	0	0	0	0	598

nchild/nchil	No children	One child	2 children	3	4	5	6	7	8	18	21 children	missing	Total
5	0	0	0	0	117	0	0	0	0	0	0	0	117
6	0	0	0	0	0	20	0	0	0	0	0	0	20
7	0	0	0	0	0	0	13	0	0	0	0	0	13
8	0	0	0	0	0	0	0	7	0	0	0	0	7
9	0	0	0	0	0	0	0	0	3	0	0	0	3
Missing	0	0	0	0	0	0	0	0	0	0	940	940	
Total	24102	4378	2643	598	117	20	13	7	3	1	1	940	32823

ISSP2012jh1d.dat %>%

tableX(lifsta, lifstam)

lifsta/Mitglied/partnership	Civil still legally in a civil partnership)	Divorced from spouse/ legally separated from civil partner		Widowed/civil never in a part-ner		single	missing	Total
		Separated from spouse/ civil partner (still legally married/	Civil still legally in a civil	partner	died			
1	17570	0		0	0	0	0	17573
2	0	1035	0	0	0	0	0	1035
3	0	0	486	0	0	0	0	486
4	0	0	0	2997	0	0	0	2997
5	0	0	0	0	2763	0	0	2763
6	0	0	0	0	0	7535	0	7535
Missing	0	0	0	0	0	0	434	434
Total	17570	3035	486	2997	2763	7535	434	32823

ISSP2012jh1d.dat %>%

tableX(urbru, urbrum)

urbru/urbrum	A big city	The suburbs or outskirts of a big city	A town or a small city	A country village	A farm or home in the country	missing	Total
1	8442	0	0	0	0	0	8442
2	0	4386	0	0	0	0	4386
3	0	0	9203	0	0	0	9203
4	0	0	0	8646	0	0	8646
5	0	0	0	0	1902	0	1902
Missing	0	0	0	0	0	244	244
Total	8442	4386	9203	8646	1902	244	32823

Muunnosten testaus, varmistetaan että muuttujat säilyvät samanlaisina. Muunnosten jälkeen on koodilohkossa passiivisina riveinä taulukointeja ja muita testauksia. (16.9.2020)

```
# (16.9.2020) Testaus uusille muuttujille
# Koodilohkoissa on jo testattu taulukoimalla muuttujia. Tässä varmistetaan, että
# muuttujat pysyvät sellaisina millaisiksi ne on luotu.

# ika - onpas hankala testata !
# Min. 1st Qu. Median Mean 3rd Qu. Max.
# 15.00 36.00 50.00 49.52 63.00 102.00
# ikatest <- ISSP2012jh1d.dat$ika %>% summary()
# ikatest <- ikatest[2,]
#validate_that(are_equal(ikatest, c(15, 36, 50, 49.5, 63, 102)))
#str(ISSP2012jh1d.dat)
#ISSP2012jh1d.dat %>%

# substanssimuuttujat 1
# Q1a, Q1b, Q1c, Q1d, Q1e, Q2a, Q2b, Q3a, Q3b (r. 423->)

validate_that(length(levels(ISSP2012jh1d.dat$Q1a)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1a),
c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1b)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1b),
c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1c)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1c),
c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1d)) == 5)

## [1] TRUE
```

```

validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1d),
                      c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1e)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1e),
                      c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q2a)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q2a),
                      c("S", "s", "?", "e", "E")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q2b)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q2b),
                      c("S", "s", "?", "e", "E")))

## [1] TRUE
# substanssimuuttujat 2

validate_that(length(levels(ISSP2012jh1d.dat$Q3a)) == 3)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q3a),
                      c("W", "w", "H")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q3b)) == 3)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q3b),
                      c("W", "w", "H")))

## [1] TRUE
# substanssimuuttujat, puuttuva tieto muuttujan arvona
# Q1am, Q1bm, Q1cm, Q1dm, Q1em, Q2am, Q2bm, Q3am, Q3bm

```

```

validate_that(length(levels(ISSP2012jh1d.dat$Q1am)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1am),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1bm)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1bm),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1cm)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1cm),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1dm)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1dm),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q1em)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q1em),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q2am)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q2am),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE

```

```

validate_that(length(levels(ISSP2012jh1d.dat$Q2bm)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q2bm),
c("S", "s", "?", "e", "E", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q3am)) == 4)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q3am),
c("W", "w", "H", "P")))

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$Q3bm)) == 4)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012jh1d.dat$Q3bm),
c("W", "w", "H", "P")))

## [1] TRUE
# taustamuuttujat puuttuvilla tiedoilla ja ilman
# testataan vain tasojen määrä, ei labeleita jotka ovat
# alkuperäisestä datasta.

# edu, edum
validate_that(length(levels(ISSP2012jh1d.dat$edu)) == 7)

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$edum)) == 8)

## [1] TRUE
# msta, mstam
validate_that(length(levels(ISSP2012jh1d.dat$msta)) == 9)

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$mstam)) == 10)

## [1] TRUE
# sosta, sostam
validate_that(length(levels(ISSP2012jh1d.dat$sosta)) == 10)

## [1] TRUE

```

```

validate_that(length(levels(ISSP2012jh1d.dat$sostam)) == 11)

## [1] TRUE
# nchild, ncildm
validate_that(length(levels(ISSP2012jh1d.dat$nchild)) == 11)

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$ncildm)) == 12)

## [1] TRUE
# lifsta, lifstam
validate_that(length(levels(ISSP2012jh1d.dat$lifsta)) == 6)

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$lifstam)) == 7)

## [1] TRUE
# urbru, urbrum
validate_that(length(levels(ISSP2012jh1d.dat$urbru)) == 5)

## [1] TRUE
validate_that(length(levels(ISSP2012jh1d.dat$urbrum)) == 6)

## [1] TRUE

```

2 Yksinkertainen korrespondenssianalyysi - kahden luokittelumuuttujan taulukko

Vanhaa jäsenystä

Yksinkertainen esimerkki, yksi kysymys (V6/Q1b) ja kuusi maata ristiintaulukituna. Johdatteluna aiheeseen esitellään ca-käsitteet profili, massa ja reunajakauma. Havainnollistetaan rivi- ja sarakeprofilien vertailua vastaaviin keskiarvo-profileihin.

Taulukoita tarkastella ensin rivien ja sitten sarakkeiden suhteen. Miten ne poikkeavat keskiarvostaan, miten toisistaan saman kategorian profilista. Usein taulukoissa muuttujilla on selvästi eri rooli, kuten tässä. Koitan hahmottaa maiden (=aggregoituja yksilöitä) eroja ja yhtäläisyysjä. Sarakkeiden vertailussa taas näemmme, miten muuttujien profilit poikkeavat keskiarvostaan. Monia riippuvuuksia ja poikkeamia näyttäisi olevan. Klassinen ongelma, Pearson ja Fisher. Luokittelumuuttujien yhteys ("korrelaatio") on hankala juttu.

Riippumattomuushypoteesi ja χ^2 - riippumattomuustesti (pieni huomautus - on monta tapaa testata taulukon riippuvuuksia). Riippumattomuushypoteesi eh-dollisena todennäköisytenä reunajakauman suhteen. Riippumattomuustulkinta ei aina päde, jos aggregoidut havainnot/rivi-tai sarakeprofiilit/”samples” MG:n terminologiassa eivät ole riippumattomia. Esimerkki Barentsin merenpohjan lajiston havainnot (lukumäärit, “abundance”) öljylauttojen liepeiltä (havainnot ryväksiä).

zxy Tämä puuttuu kaavoista!

Käsitteitä

1. Taulukko

“Ranskaisella terminologia”: käsitellään yksilöiden tai havaintoyksiköiden pilveä ja muuttujien pilveä . Taulukot saadaan yksinkertaisen CA:n tapauksessa aggregoimalla “cloud of individuals”. MG:n termi “sample”.

#V MOOC, LeReoux

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

Dataa valitaan, aggregoidaan, ryhmitellään. Aktiivisia valintoja Blasius emt. “data ei löydy kadulta”, taulukot vielä vähemmän.

ISSP-datan etu: hyvin dokumentoitut ja editoitut, laadukas aineisto. Iso (ja kallis) työvaihe on jo tehty. Aineistoa on myös käytetty useissa tutkimuksissa, mm. MG:n oppikirjoissa.

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

3. CA:ssa vaikea juttu on (Blasius, “vizualisation - verkkokirja”) rivien ja sarakkeiden **tekkinen symmetria**. No ei se nyt niin hämäävä ehkä ole, oleellinen juttu (21.2.20). Kts. myös MG:n didaktiset esittelyt, skaalataan ”hajontamittarilla” ja painotetaan massoilla. **edit 6.9.20** Tätä havainnoslistataan teorian esittelyn yhteydessä johdattelevan aineiston datalla. Pienellä taulukolla helpompa.

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

Muutama versio tiiviiksi kuvaukseksi - toistoa on (10.4.20)

Dimensioiden vähentäminen tärkein asia (“the essence”), pienessä taulossa ei ihan ilmeinen. Esimerkin pienissä taulukoissa on toisaalta helppo katsoa datasta, mistä on kyse. Toinen tavoite on visualointi, yleensä kaksilotteisena kuvana (karttana). Kartta on metaforana hieman hankala. Kartalla esitetään kahden pistejoukon (“pilven”) projektiot, jotka säilyttävät maksimimääärän alkuperäisen n-ulotteisen pistejoukon hajonnasta (inertiasta). Projektiossa lähekkäin olevat saman pilven pistet voivat kuitenkin olla n-ulotteisessa pilvessä hyvinkin kauhana toisistaan. Tulkinnassa tärkeitä ovat “ääripääät”, ja numeeriset tulokset kertovat

kuinka hyvin piste on tasossa esitetty. Pisteiden väiset etäisyydet suhteellisia, ja eri pistejoukkojen välisillä etäisyyksillä ei ole suoraan mitään tulkintaa. Tämä ei oikein vastaa mielikuvaan kartasta, josta helposti näkee kuinka kaukana on Uudenmaan raja.

Yksinkertainen korrespondenssianalyysi on kahden luokitteluaesteikon 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 kaksilottaisena 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 ylhällä 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 - tehty jo

1. Ensimmäinen taulukko: profilit, massat, keskiarvoprofilit, 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. Käsittteistä oma Rmd (ja Excel jos osoittautuu kätevämmäksi), kaavaliite Dispo-repossa ja myös Rmd-muodossa.

edit(10.4.20): kaavaliitteen lisäksi voi tekstiin upottaa muutaman r-koodi-esimerkin

Ensimmäinen symmetrinen kartta

Tulkinnat ja yksinkertaisimmat perussäänöt. Dimensiot ja kuinka paljon alkuperäisen taulukon inertiaa saadaan esitettyä kartalla. Sitten asian ydin, akseleiden

tulkinta (“mitä on oikealla ja vasemmalla”). Jos pisteet ovat alkuperäisessä “pil vessä” kaukana toisistaan, ne ovat sitä myös projektiossa. Kartta, mutta etäisyksillä ei suoraa tulkintaa paitsi eteisyksinllä origoon. Rivipisteiden suhteelliset etäisydet, samoin sarakepisteidet. Mitä tarkoittavat prosentit akselleilla?

Varoitus virhetulkinnasta: ryhmien tunnistaminen rivi- ja sarakepisteiden läheisyyden avulla, myös pelkästään rivi- tai 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 sarekepistettä. MG: “loose ends” - paperi, symmetrinen kuva eräs suurin sekaannuksen lähde. Tätä koitetaan selventää myös MG:n JASA-artikkelissa.

zxy(teoria/historia-jaksoon,104.20).Termi korrespondenssi: “neglected multivariate method” - paperissa käännetty näin englanniksi ransk. termi (Benzecri) rivien ja sarakkeiden “correspondence” eli yhteys, “riippuvuus”, vastaavuus tms. **edit 4.7.20** Kts. myös Funmooc-muistiinpanot, opk! Mitä kartta esittää? Kaikki edellä kuvattu esitetään suhteellisina eroina koko aineiston keskiarvosta, riippumattomuushypoteesi.

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.

Aineisto muuttujat Q1a-Q1e (arvot 1-5, täysin samaa mieltä - täysin eri mieltä) ovat vastauksia ensimmäiseen kysymyspatteriin (kts. lomake).

edit 10.4.20 Muuttujien “suunta” samaksi, jos monta. Laajemman aineiston käsittelyyn tästä huomautus.

(V6/Q1b) Alle kouluikäinen lapsi todennäköisesti kärsii, jos hänen äitinsä käy työssä. V6 muunnetaan uudeksi luokittelumuuttujaksi (R:ssä factor) Q1b. Tämä ei vielä tee kuvista ahtaita kun sarakkeita ja rivejä on vähän. Pudotetaan tarvittaessa turha Q-kirjain pois. Alkuperäisessä muuttujassa metatieto säilyy varmemmin, ja tarkistuksia on helpompi tehdä.

Valitaan esimerkin data edellisessä luvussa luodusta R-datasta ISSP2012jh1d.dat). Ihan yhtä hyvin voisi aina lukea suoraan alkuperäisestä spss-tiedostosta, mutta pidemmässä raportissa tämä on siistimpi tapa (23.3.2019). Kun havaintoja ja maita jätetään pois, uuteen dataan jää tyhjiä luokittelumuuttujien luokkia, ne poistetaan.

```
# UUSI DATA 30.1.20
#
# LUETAAN DATA G1_1_data2.Rmd - tiedostossa, luodaan faktorimuuttujat
```

```

# G1_1_data_fct1.Rmd-tiedostossa -> ISSP2012jh1d.dat (df)
# 23 muuttuja (9 substanssimuuttuja, 8 taustamuuttuja, 3 maa-muuttuja, 3 metadatamuuttuja)
# 25 maata.
# Poistettu 146 havaintoa, joilla SEX tai AGE puuttuu
# Johdattelevassa esimerkissä kuusi maata, kaksi taustamuuttuja ja yksi kysymys
# (V6/Q1b)

# Kuusi maata

countries_esim1 <- c(56, 100, 208, 246, 276, 348) #BE,BG,DK,FI,DE,HU
ISSP2012esim3.dat <- filter(ISSP2012jh1d.dat, V4 %in% countries_esim1)
# str(ISSP2012esim3.dat) - pitkä listaus pois (24.2.20)

#neljä maamuuttuja, kysymys Q1b, ikä ja sukupuoli

vars_esim1 <- c("C_ALPHAN", "V3", "maa", "maa3", "Q1b", "sp", "ika")
ISSP2012esim2.dat <- select(ISSP2012esim3.dat, all_of(vars_esim1))

str(ISSP2012esim2.dat) # 8542 obs. of 7 variables, ja sama 8.6.2020

## tibble [8,542 x 7] (S3:tbl_df/tbl/data.frame)
## $ C_ALPHAN: chr [1:8542] "BG" "BG" "BG" "BG" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
## $ V3      : dbl+lbl [1:8542] 100, 100, 100, 100, 100, 100, 100, 100, 100, 100...
##   ..@ label      : chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation)"
##   ..@ format.spss: chr "F5.0"
##   ..@ labels     : Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bul...
## $ maa      : Factor w/ 25 levels "AU","AT","BG",...: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## $ maa3     : Factor w/ 29 levels "AU-Australia",...: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na...
## $ Q1b      : Factor w/ 5 levels "S","s","?","e",...: 3 2 3 4 3 3 4 3 2 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ sp       : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##   ..- attr(*, "label")= chr "Sex of Respondent"
## $ ika      : dbl+lbl [1:8542] 64, 43, 63, 31, 52, 46, 51, 40, 57, 64, 41, 60, 21, 4...
##   ..@ label      : chr "Age of respondent"
##   ..@ format.spss: chr "F3.0"
##   ..@ labels     : Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
##   - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa...

```

```

# C_ALPHAN: chr, maa: Factor w/ 25

# Poistetaan havainnot, joilla Q1b - muuttujassa puuttuva tieto 'NA'
# sum(is.na(ISSP2012esim2.dat$Q1b)) = 399

ISSP2012esim1.dat <- filter(ISSP2012esim2.dat, !is.na(Q1b))

str(ISSP2012esim1.dat) # 8143 obs. of 6 variables

## tibble [8,143 x 7] (S3:tbl_df/tbl/data.frame)
## $ C_ALPHAN: chr [1:8143] "BG" "BG" "BG" "BG" ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##   ..- attr(*, "format.spss")= chr "A20"
##   ..- attr(*, "display_width")= int 22
## $ V3      : dbl+lbl [1:8143] 100, 100, 100, 100, 100, 100, 100, 100, 100, 100...
##   ..@ label      : chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation"
##   ..@ format.spss: chr "F5.0"
##   ..@ labels     : Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##   .. ..- attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bul...
## $ maa      : Factor w/ 25 levels "AU","AT","BG",...: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## $ maa3     : Factor w/ 29 levels "AU-Australia",...: 3 3 3 3 3 3 3 3 3 3 ...
##   ..- attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na...
## $ Q1b      : Factor w/ 5 levels "S","s","?","e",...: 3 2 3 4 3 3 4 3 2 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ sp       : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##   ..- attr(*, "label")= chr "Sex of Respondent"
## $ ika      : dbl+lbl [1:8143] 64, 43, 63, 31, 52, 46, 51, 40, 57, 64, 41, 60, 21, 4...
##   ..@ label      : chr "Age of respondent"
##   ..@ format.spss: chr "F3.0"
##   ..@ labels     : Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
##   - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa...

# Tarkistuksia - miksi nämä eivät tulosta mitään? (3.2.20)

fct_count(ISSP2012esim1.dat$sp)


```

f	n
m	3799
f	4344

```
fct_count(ISSP2012esim1.dat$Q1b)
```

f	n
S	810
s	1935
?	1367
e	2125
E	1906

```
fct_count(ISSP2012esim1.dat$maa)
```

f	n
AU	0
AT	0
BG	921
CA	0
HR	0
CZ	0
DK	1388
FI	1110
FR	0
HU	997
IS	0
IE	0
LV	0
LT	0
NL	0
NO	0
PL	0
RU	0
SK	0
SI	0
SE	0
CH	0
BE	2013
DE	1714
PT	0

```
fct_count(ISSP2012esim1.dat$maa3)
```

f	n
AU-Australia	0
AT-Austria	0
BG-Bulgaria	921

f	n
CA-Canada	0
HR-Croatia	0
CZ-Czech Republic	0
DK-Denmark	1388
FI-Finland	1110
FR-France	0
HU-Hungary	997
IS-Iceland	0
IE-Ireland	0
LV-Latvia	0
LT-Lithuania	0
NL-Netherlands	0
NO-Norway	0
PL-Poland	0
RU-Russia	0
SK-Slovakia	0
SI-Slovenia	0
SE-Sweden	0
CH-Switzerland	0
BE-FLA-Belgium/ Flanders	1012
BE-WAL-Belgium/ Wallonia	490
BE-BRU-Belgium/ Brussels	511
DE-W-Germany-West	1167
DE-E-Germany-East	547
PT-Portugal 2012: first fieldwork round (main sample)	0
PT-Portugal 2012: second fieldwork round (complementary sample)	0

```
# Toimivat tarkistukset (3.2.20)
summary(ISSP2012esim1.dat$sp)

##      m      f
## 3799 4344
#sp: 3799 + 4344 = 8143
summary(ISSP2012esim1.dat$Q1b)

##      S      s      ?      e      E
## 810 1935 1367 2125 1906
#   S      s      ?      e      E
# 810 + 1935 + 1367 + 2125 + 1906 = 8143

# EDELLINEN DATA - havaintojen määrit mat kuin uudella datalla (31.1.20)
#
```

```

# 8557 obs. ennen kuin sexagemissing poistettiin, nyt 8542, 8557-8542 = 15
#
# Poistetaan havainnot joissa puuttuva tieto muuttujassa V6 (Q1b) n = 399
# 8542-399 = 8143

# Tyhjät "faktorilabelit" on poistettava

ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
  mutate(maa = fct_drop(maa),
         maa3 = fct_drop(maa3)
  )

summary(ISSP2012esim1.dat$maa)

##   BG    DK    FI    HU    BE    DE
##  921  1388 1110  997 2013 1714
summary(ISSP2012esim1.dat$maa3)

##          BG-Bulgaria           DK-Denmark           FI-Finland
##             921                  1388                  1110
##          HU-Hungary BE-FLA-Belgium/ Flanders BE-WAL-Belgium/ Wallonia
##             997                  1012                  490
##          BE-BRU-Belgium/ Brussels        DE-W-Germany-West        DE-E-Germany-East
##             511                  1167                  547
# str(ISSP2012esim1.dat$maa)
# attributes(ISSP2012esim1.dat$maa)

# str(ISSP2012esim1.dat$maa3)
# attributes(ISSP2012esim1.dat$maa3)

ISSP2012esim1.dat %>% tableX(maa, Q1b, type = "count")

```

maa/Q1b	S	s	?	e	E	Total
BG	118	395	205	190	13	921
DK	70	238	152	232	696	1388
FI	47	188	149	423	303	1110
HU	219	288	225	190	75	997
BE	191	451	438	552	381	2013
DE	165	375	198	538	438	1714
Total	810	1935	1367	2125	1906	8143

```
fct_count(ISSP2012esim1.dat$Q1b)
```

	f	n
S	810	
s	1935	
?	1367	
e	2125	
E	1906	

```
# fct_count(ISSP2012esim1.dat$sp)
# fct_unique(ISSP2012esim1.dat$maa)
# fct_count(ISSP2012esim1.dat$maa)
ISSP2012esim1.dat %>% tableX(maa, C_ALPHAN, type = "count")
```

maa/C_ALPHAN	BE	BG	DE	DK	FI	HU	Total
BG	0	921	0	0	0	0	921
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
BE	2013	0	0	0	0	0	2013
DE	0	0	1714	0	0	0	1714
Total	2013	921	1714	1388	1110	997	8143

```
# maa3 - siistitäään "faktorilabelit" kaksikirjaimisiksi
#
# 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
# Tähän pitäisi päästä
# levels = c("100", "208", "246", "348", "5601", "5602", "5603", "27601", "27602"),
# labels = c("BG", "DK", "FI", "HU", "bF", "bW", "bB", "dW", "dE"))
levels(ISSP2012esim1.dat$maa3)

## [1] "BG-Bulgaria"                  "DK-Denmark"
## [3] "FI-Finland"                   "HU-Hungary"
## [5] "BE-FLA-Belgium/ Flanders"    "BE-WAL-Belgium/ Wallonia"
## [7] "BE-BRU-Belgium/ Brussels"     "DE-W-Germany-West"
## [9] "DE-E-Germany-East"
```

```

ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
  mutate(maa3 =
    fct_recode(maa3,
      "BG" = "BG-Bulgaria",
      "DK" = "DK-Denmark",
      "FI" = "FI-Finland",
      "HU" = "HU-Hungary",
      "bF" = "BE-FLA-Belgium/ Flanders",
      "bW" = "BE-WAL-Belgium/ Wallonia",
      "bB" = "BE-BRU-Belgium/ Brussels",
      "dW" = "DE-W-Germany-West",
      "dE" = "DE-E-Germany-East")
  )
# tarkistuksia
levels(ISSP2012esim1.dat$maa3)

## [1] "BG" "DK" "FI" "HU" "bF" "bW" "bB" "dW" "dE"
# str(ISSP2012esim1.dat$maa3) # 9 levels
summary(ISSP2012esim1.dat$maa3)

##   BG   DK   FI   HU   bF   bW   bB   dW   dE
## 921 1388 1110  997 1012  490  511 1167  547

# TÄSSÄ TOISTOA! (4.2.20)

# Muutetaan muuttujien "maa" ja "maa3" arvojen (levels) järjestys samaksi kuin alkuperäisen
# muuttujan C_ALPHAN. Helpomi verrata aikaisempia tuloksia.

# maa samaan järjestukseen kuin C_ALPHAN - olisiko aakkosjärjestys?
# tämä vain siksi, että muuten esimerkin ca-kartta "kääntyy"
# "vanha" maa-muuttuja talteen - ei ehkä tarpeen? (4.2.20)

ISSP2012esim1.dat$maa2 <- ISSP2012esim1.dat$maa # "alkuperäinen" maa talteen

ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
  mutate(maa =
    fct_relevel(maa,
      "BE",
      "BG",
      "DE",
      "DK",
      "FI",
      "HU"))
ISSP2012esim1.dat <- ISSP2012esim1.dat %>%
  mutate(maa3 =

```

```
fct_relevel(maa3,
  "bF",
  "bW",
  "bB",
  "BG",
  "dW",
  "dE",
  "DK",
  "FI",
  "HU"))
```

maa2/maa	BE	BG	DE	DK	FI	HU	Total
BG	0	921	0	0	0	0	921
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
BE	2013	0	0	0	0	0	2013
DE	0	0	1714	0	0	0	1714
Total	2013	921	1714	1388	1110	997	8143

```
ISSP2012esim1.dat %>% tableX(maa,C_ALPHAN, type = "count")
```

maa/C_ALPHAN	BE	BG	DE	DK	FI	HU	Total
BE	2013	0	0	0	0	0	2013
BG	0	921	0	0	0	0	921
DE	0	0	1714	0	0	0	1714
DK	0	0	0	1388	0	0	1388
FI	0	0	0	0	1110	0	1110
HU	0	0	0	0	0	997	997
Total	2013	921	1714	1388	1110	997	8143

```
str(ISSP2012esim1.dat)
```

```
## # tibble [8,143 x 8] (S3: tbl_df/tbl/data.frame)
## # $ C_ALPHAN: chr [1:8143] "BG" "BG" "BG" "BG" ...
## #   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## #   ..- attr(*, "format.spss")= chr "A20"
## #   ..- attr(*, "display_width")= int 22
## # $ V3      : dbl+lbl [1:8143] 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100, 100...
```

```

##     ..@ label      : chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole nation"
##     ..@ format.spss: chr "F5.0"
##     ..@ labels     : Named num [1:45] 32 36 40 100 124 152 156 158 191 203 ...
##     ... - attr(*, "names")= chr [1:45] "AR-Argentina" "AU-Australia" "AT-Austria" "BG-Bul
##     $ maa       : Factor w/ 6 levels "BE","BG","DE",...: 2 2 2 2 2 2 ...
##     .. - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##     $ maa3      : Factor w/ 9 levels "bF","bW","bB",...: 4 4 4 4 4 4 4 4 4 ...
##     .. - attr(*, "label")= chr "Country/ Sample ISO 3166 Code (see V4 for codes for whole na
##     $ Q1b       : Factor w/ 5 levels "S","s","?","e",...: 3 2 3 4 3 3 4 3 2 3 ...
##     .. - attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
##     $ sp        : Factor w/ 2 levels "m","f": 2 2 1 2 2 2 1 1 2 1 ...
##     .. - attr(*, "label")= chr "Sex of Respondent"
##     $ ika       : dbl+lbl [1:8143] 64, 43, 63, 31, 52, 46, 51, 40, 57, 64, 41, 60, 21, 4...
##     ..@ label     : chr "Age of respondent"
##     ..@ format.spss: chr "F3.0"
##     ..@ labels     : Named num [1:6] 15 16 17 18 102 999
##     ... - attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
##     $ maa2      : Factor w/ 6 levels "BG","DK","FI",...: 1 1 1 1 1 1 ...
##     .. - attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
##     - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa

```

TODO (1) Taulukot erotettava omiksi koodilohkoiksi bookdowniin. (2) Ikä - maa - taulukko vain tarkistuksiin, ihan liian pitkä.

```

# Taulukoita (31.1.2020) ja tarkistuksia

# 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

# assert_that ehkä tarpeeton - expect_equivale testaa levelien
# järjestyksen ja määrään (20.2.20)

validate_that(length(levels(ISSP2012esim1.dat$sp)) == 2)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012esim1.dat$sp),
c("m", "f")))

## [1] TRUE

```

```

validate_that(length(levels(ISSP2012esim1.dat$maa)) == 6)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012esim1.dat$maa),
c("BE", "BG", "DE", "DK", "FI", "HU")))

## [1] TRUE
validate_that(length(levels(ISSP2012esim1.dat$maa3)) == 9)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012esim1.dat$maa3),
c("bF", "bW", "bB", "BG", "dW", "dE", "DK", "FI", "HU")))

## [1] TRUE
validate_that(length(levels(ISSP2012esim1.dat$Q1b)) == 5)

## [1] TRUE
validate_that(are_equal(levels(ISSP2012esim1.dat$Q1b),
c("S", "s", "?", "e", "E")))

## [1] TRUE
# testthat - paketti - pois käytöstä 16.9.20
# expect_ ei anna ok-ilmoitusta, ainoastaan virheilmoituksen? (11.4.20)

# expect_equivalent(levels(ISSP2012esim1.dat$maa),
#                   c("BE", "BG", "DE", "DK", "FI", "HU"))

# expect_equivalent(levels(ISSP2012esim1.dat$maa3),
#                   c("bF", "bW", "bB", "BG", "dW", "dE", "DK", "FI", "HU"))

# expect_equivalent(levels(ISSP2012esim1.dat$sp), c("m", "f"))

# expect_equivalent(levels(ISSP2012esim1.dat$Q1b),
#                   c("S", "s", "?", "e", "E"))

ISSP2012esim1.dat %>% tableX(maa, ika, type = "row_perc")

```

maa\ika	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53							
BE	0.00	0.00	0.79	0.29	0.24	0.19	0.51	0.49	0.34	0.41	0.19	0.69	0.34	0.39	0.61	0.59	0.04	0.72	0.04	0.89	0.59	0.74	0.49	0.54	0.89	0.81	0.69	0.69	0.54	0.69	0.62	0.68	0.79	0.51	0.89	0.79	0.92								
BC	0.00	0.00	0.41	0.09	0.98	0.98	0.76	0.19	0.76	0.19	0.40	0.98	0.98	0.09	0.50	0.98	0.30	0.50	0.98	0.74	0.74	0.30	0.95	0.19	0.06	0.72	0.62	0.50	0.40	0.76	0.87	0.52	0.30	0.41	0.70	0.87	0.17								
DE	0.00	0.00	0.01	0.11	0.69	0.23	0.58	0.40	0.98	0.46	0.58	0.17	0.40	0.23	0.46	0.28	0.52	0.40	0.99	0.34	0.28	0.34	0.11	0.23	0.23	0.87	0.52	0.16	0.81	0.98	0.73	0.80	0.02	0.80	0.51	0.80	0.59	0.31	0.80	0.23	0.37	0.10	0.02	0.66	0.87
DK	0.00	0.00	0.78	0.30	0.30	0.23	0.52	0.74	0.95	0.15	0.08	0.73	0.37	0.41	0.02	0.09	0.41	0.42	0.59	0.15	0.73	0.78	0.37	0.80	0.02	0.80	0.51	0.80	0.59	0.31	0.80	0.23	0.37	0.10	0.02	0.66	0.87								
FI	0.72	0.80	0.17	0.62	0.08	0.35	0.17	0.68	0.26	0.53	0.35	0.41	0.26	0.17	0.61	0.17	0.62	0.08	0.80	0.35	0.41	0.53	0.98	0.60	0.90	0.71	0.35	0.26	0.08	0.80	0.62	0.89	0.70	0.61	0.98	0.71	0.53	0.0	0.0	0.0					

maa5/16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53				
HU	0.00	0.00	90.20	0.00	80.91	91.10	50.00	40.30	91.60	81.50	41.50	72.92	01.60	61.50	40.10	01.40	32.22	31.91	50.30	50.20	1.00	10.26	16.22	31.20	38.51	78.40	39.20	46.25	61.36	62.52	60.71	54.73	62.50	52.85	63.73	90.65	77.93	37.83	81.85	61.82	0.00
All	0.10	26.16	22.31	20.38	51.78	40.39	20.46	25.61	36.62	52.60	71.54	73.62	50.52	85.63	73.90	65.77	93.37	83.83	81.85	61.82	0.00																				

```
# Riviprofiilit
```

```
# ISSP2012esim1.dat %>% tableX(maa,ika,type = "row_perc")
ISSP2012esim1.dat %>% tableX(maa,sp ,type = "row_perc")
```

maa/sp	m	f	Total
BE	47.44	52.56	100.00
BG	40.72	59.28	100.00
DE	48.66	51.34	100.00
DK	49.42	50.58	100.00
FI	42.88	57.12	100.00
HU	47.44	52.56	100.00
All	46.65	53.35	100.00

```
# Kysymyksen Q1b vastaukset
```

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

maa/Q1b	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

```
# Kuuluu ehkä vasta seuraavaan jaksoon ? (20.2.20)
```

```
ISSP2012esim1.dat %>% tableX(maa3,Q1b,type = "row_perc")
```

maa3/Q1b	S	s	?	e	E	Total
bF	5.04	23.81	25.89	30.83	14.43	100.00
bW	10.82	21.02	18.57	24.08	25.51	100.00
bB	17.03	20.94	16.63	23.87	21.53	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00

maa3/Q1b	S	s	?	e	E	Total
dW	11.40	26.82	11.83	32.13	17.82	100.00
dE	5.85	11.33	10.97	29.80	42.05	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

```
# str(ISSP2012esim1.dat) # 8143 obs. of 7 variable,
# sama kuin vanhassa Galku-koodissa.
```

Taulukot ja kuvat omina koodilohkoina bookdown-versioon

Frekvenssitaulukko

```
# Esimerkki - siisti taulukko (20.2.20)
```

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

Taulukko 52: 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 53: 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

	S	s	?	e	E	Total
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 54: 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”) keskimääriäiseen riviin, sarakemuuttujien eli vastausvaihtoehtojen reunajakaumaan. Vastavasti sarakkeita voi verrata rivi-muuttujien reunajakaumasarakkeeseen (“Total2”). Eniten vastaajia on Belgiasta (25 %) ja Saksasta (21 %), vähiten Unkarista (12 %). **edit 24.2.20** Lisätty karttoihin versio, jossa maiden painot skaalattu yhtä suuraksi. Esimerkkilaskelma CAcalc_1.R.

```
# CA tässä, jotta saadaan rivi- ja sarakeprofiilikuvat

simpleCA1 <- ca(~maa + Q1b,ISSP2012esim1.dat)

# Maiden järjestys kääntää kuvan (1.2.20) - esimerkki on
# vähän kuriositeetti. Kartta voi tietysti "flipata" koordintaattien suhteen ainakin
# neljällä tavalla (? 180 astetta molempien akseleiden ympäri molempien suuntiin?)
# (18.2.20). Tämän maa2-muuttujaan käyttävän kuvan voi jättää pois (8.4.20)
```

```

# simpleCA2 <- ca(~maa2 + Q1b, ISSP2012esim1.dat)

# Oikeastaan maiden vertailussa pitäisi niiden massat skaalata yhtä suuriksi, tässä
# pikainen kokeilu (20.2.20)
# Riviprosentit taulukoksi, nimet sarakkeille ja riveille (ei kovin robustia...)

johdesim1_rowproc.tab <- simpleCA1$N / rowSums(simpleCA1$N)
colnames(johdesim1_rowproc.tab) <- c("S", "s", "?", "e", "E")
rownames(johdesim1_rowproc.tab) <- c("BE", "BG", "DE", "DK", "FI", "HU")

# Miten tibbleenä? Ei toimi, ei maa-muuttuja ollenkaan
# johdesim1_rowproc.tbl <- as_tibble(johdesim1_rowproc.tab)
# str(johdesim1_rowproc.tab)

# TARKISTUKSIA (20.2.20)
# johdesim1_rowproc.tab
# rowSums(johdesim1_rowproc.tab)
# str(johdesim1_rowproc.tab)

simpleCA3 <- ca(johdesim1_rowproc.tab)

# Kartta piirretään koodilohkossa simpleCAmap1, r. 773 noin.

# Riviprosentit tarkistusta varten
#      S   s   ?   e   E
#BE 9.49  22.40  21.76  27.42  18.93
#BG 12.81  42.89  22.26  20.63  1.41
#DE 9.63  21.88  11.55  31.39  25.55
#DK 5.04  17.15  10.95  16.71  50.14
#FI 4.23  16.94  13.42  38.11  27.30
#HU 21.97  28.89  22.57  19.06  7.52
#
# Ja datan saa leikepöydän kautta, jos on tarve pikatarkistuksiin
# read <- read.table("clipboard")

# 9.4.2020 CAcalc_1.R - laskentoa ca-funktion tuloksilla (16 objektiin lista)

```

TODO 2.2.20

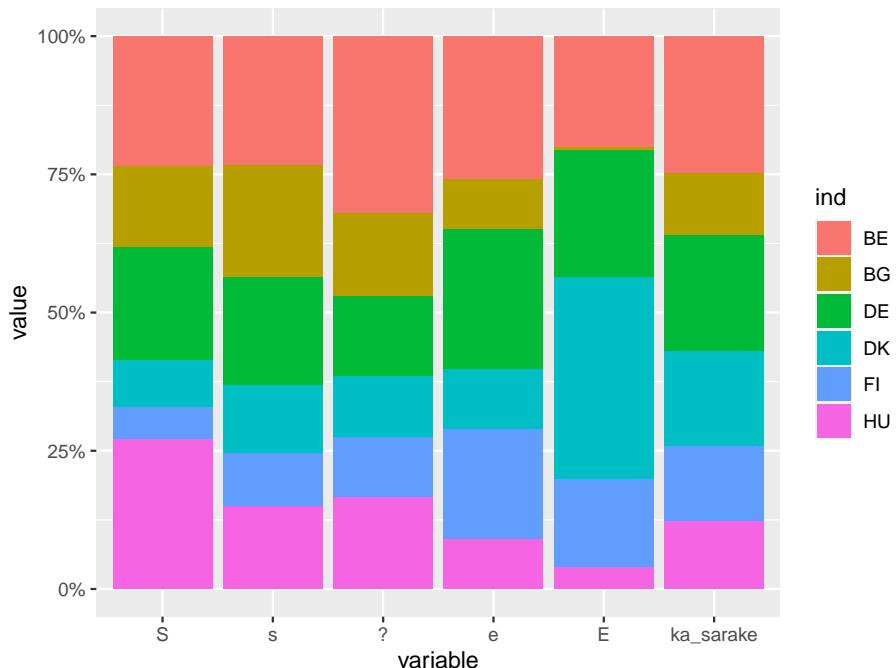
Onko tämä kuva tallennettava kuvatiedostoksi, vai onnistuuko sen tuottaminen Bookdownissa. Ei taida onnistua? (4.9.18)

Sarakeprofilit, oikea järjestys maa-muuttujan tasoilla. Faktoreiden järjestys voi

tuottaa yllätyksiä, kun dataa muokataan ggplot - grafiikaksi.

```
#mutkikas kuvan piirto - sarakeprofiliit vertailussa
#ggplot vaatii df-rakenteen ja 'long data' - muotoon
##https://stackoverflow.com/questions/9563368/create-stacked-barplot-where-each-stack-is-sc
#
# käytetään ca - tuloksia
apu1 <- (simpleCA1$N)
colnames(apu1) <- c("S", "s", "?", "e", "E")
rownames(apu1) <- c("BE", "BG", "DE", "DK", "FI", "HU")
apu1_df <- as.data.frame(apu1)
#lasketaan rivien reunajakauma
apu1_df$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())
```



Kuva 2: Q1b:Sarakeprofiliit ja keskiarvoprofilii

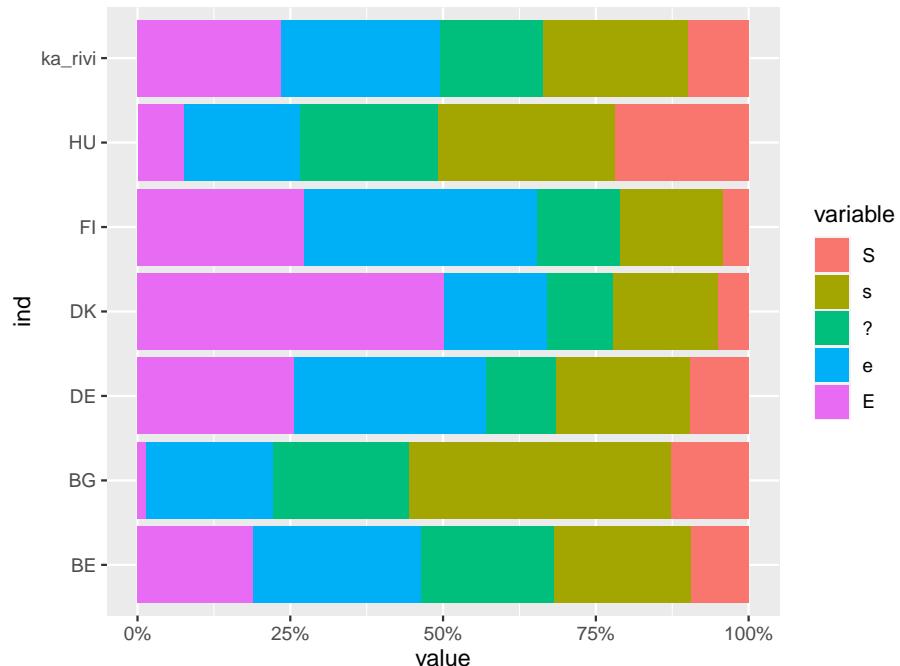
```
# apu1_df  
# apu1b_df
```

Edit 28.5.20 Idea: ca havainnollistaa rivien ja sarakkaeiden riippuvuuksia eroina keskiarvosta, yksinkertainen khii2-tulkinta riippumattomuushypoteesina. Rivi-ja sarakeprofilit standardoidaan (tulkintaa!) PCA-tyyliin (poikkeama keskiarvosta jaetaan hajonnalla, lukumäärädata ja POisson-jakaumassa hajonta=odotusarvo). Rivi- ja sarakeratkaisuiden symmetria ja sidos esitetään tässä.

Sarakekuvassa näkyy E-sarakkeenselvästi ero (DK), muut erot eivät niinkään. S-sarakkeessa HU erottuu, samoin eos(?)-sarakkeessa BE.

TODO 2.2.20 Voisi harkita taulukoiden (rivi- ja sarakeprosentit) sijoittamista kuvien viereen?

```
# riviprofiilit ja keskiarvorivit - 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'))  
# str(apu2b_df)  
# glimpse(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 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())
```



edit 28.5.20 Tanska ja Unkari erottuvat ka-rivistä E-vaihtoehdon (“modaliteetin”) osuuksissa selvimmin. Bulgarialla S+ on hieman suurempi kuin Unkarilla, mutta s-osuuus on suurempi. “Ääripäitä” S- ja E-osuuksissa edustavat HU - FI, DK ja BG-HU - DK.

Graafinen analyysi ja R

Käytänön neuvoja data-analyysiin, kuulunee tekstiin vai meneekö “ohjelmisto-ympäristö” -liitteeseen? Tärkeää juttu!

Kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (`X11()`) ja sitten plot. Voi tallentaa pdf-muodossa grafiikkaikkunasta, ja ladata outputtiin 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. RStudiossa voi avata komentokehointeessa grafiikkaikkunan. Siitä käsin tallennettu pdf-kuva on ladattu alla Rmarkdonin omalla komennolla, kohdistus keskelle. Parhaiten näyttäisi toimivan knitrin funktio, mutta oletuskuvakolla saa ca-kuvasta näköjään aika lähelle oikeanlaisen ilman mitään temppuja.

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

CA-ratkaisun (algoritmin) lähtötieto: suhteelliset frekvenssit (korrespondenssi-

matriisi P) (30.3.20)

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

Taulukko 55: Kysymyksen Q1b 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

Massat ja skaalaus Tätä ensimmäistä kuvaaa on muistiinpanoissa kommentoitu (löytyy printattuna) Kolme karttaa. Maiden vertailussa on järkeväksi vakioida niiden massat (kolmas kartta). Massan käsite on CA:n ydinasioita, siksi maiden massat ovat jatkossa mukana. Kartta määräytyy maiden otoskokojen suuruisilla painoilla, mutta ero ei ole kovin suuri.

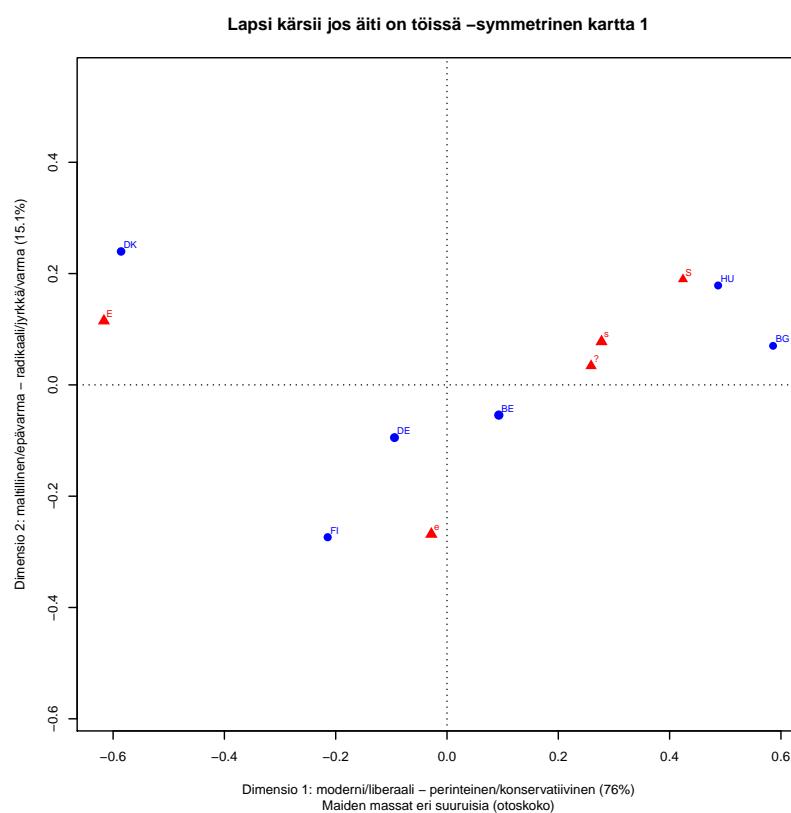
```
#simpleCA1 <- ca(~maa + V6,ISSP2012esim1.dat) suoritetaan ennen värikuvaa, tuloksia tarvitaan siinä.
```

```
# TODO(11.4.20) fig.cap koodilohkossa tekee kuvasta "kelluvan", ja kuvat numeroidaan.
# Miten plot-komennon kuvaotsikot vaikuttavat?
# Pitäiskö (a) jokaiselle kuvalle oma koodilohko (b) esittää nämä kaksi yhdessä vierekkäin
# Pohditaan kun koodataan capaper-projektia.
```

```
# Symmetrinen kartta
# Akselien tekstit "käsityönä" - esimerkki (3.5.2020)
```

```
par(cex = 0.6)
plot(simpleCA1, map = "symmetric", mass = c(TRUE,TRUE),
      main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta 1",
      xlab = "Dimensio 1: moderni/liberaali - perinteinen/konservatiivinen (76%)",
      ylab = "Dimensio 2: maltillinen/epävarma - radikaali/jyrkkä/varma (15.1%)",
      sub = "Maiden massat eri suuruisia (otoskoko)")
```

```
# plot(simpleCA2, map = "symmetric", mass = c(TRUE,TRUE),
#       main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta ",
#       sub = "maa-muuttuja maa2,järjestys as_factor(C_ALPHAN)")
# Kartta käännyt ympäri - esimerkki faktoroinnin arvaamattomista
```



Kuva 3: Q1b: lapsi kärsii jos äiti on töissä

```
# seurauksista (30.3.20)
```

```
# 13.5.2018
```

```
# kuvasuhteen saa oikeaksi, kun avaa g-ikkunan (X11()) ja sitten plot. Voi tallentaa pdf-mu-
```

```
# grafiikkaikkunasta, ja ladata outputiin knitr-vaiheessa. Parempi tulostaa kuvatdsto pdf-a-
```

```
# jos lopulliseen versioon joutuu
```

```
# näin tekemään.
```

edit 2.5.2020 Riviprofilitaulukossa rivimassat ovat vakioita (=1), mutta ca-ratkaisussa skaalautuvat eri arvoksi (vakio). Oleellinen vaikutus karttaan on pienien massan pisteiden (BG, FI, HU) siirtymien kohti origoa. Ei kyllä ole kovin selvä, ja näkyy selvimmin "ääripäissä" (? 5.9.20) Entäs sarakepisteet, niiden massat eivät muutu? **TODO 5.9.20** Alempana taulukko khii2-etäisyyksistä, tässä kaksoi projisoitu samaan kuvaan.

```
# Sama kartta - maiden massat vakiotu
```

```
# CA:n lähtötietona riviprofiilit
```

```
par(cex = 0.6)
```

```
plot(simpleCA3, map = "symmetric", mass = c(TRUE,TRUE),
main = "Lapsi kärsii jos äiti on töissä -symmetrinen kartta 2",
sub = "Maidet massat vakioitu (riviprofiilidata)")
```

Näitä karttoja vertaillaan seuraavassa luvussa tarkemmin.

Toinen tapa - kuvatiedoston lataaminen include_graphics - funktiolla. Pitää miettiä mikä on järkevää, dataa tutkaillessa piirretään useita kuvia. PDF-muodossa ne ovat skaalautuvia, kommentteja voi lisätä jne.

Rivien (1) ja sarakkeiden (2) khii2-etäisyydet keskiarvosta.

Rivimassat skaalautuvat ca-ratkaisussa vakioksi 1/6 (0.167), kun lähtötietona on riviprofilien taulukko jossa rivimassat ovat 1. Sarakemassat skaalautuvat uudelleen. Massojen summa on 1. **TODO 16.9.20** knittr::kabble nimeää taulukot koodilohkon nimen mukaan, siis vain yksi taulukko? Taulukot omiksi koodilohkoiksi.

```
# khii2 - etäisyyksien taulukko
```

```
#str(simpleCA1)
```

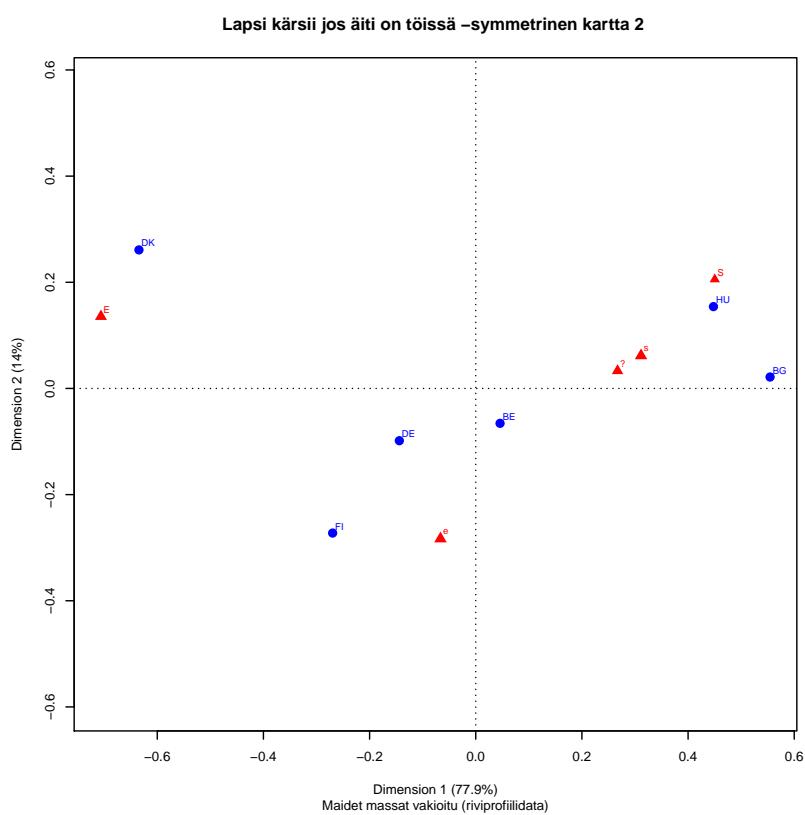
```
#simpleCA1$rowdist
```

```
#str(simpleCA1$rowdist)
```

```
# verrataan "tavallisen" ca:n ja riviprofiili-ca:n khii2-etäisyyksiä origosta
# 5.9.2020
```

```
# khii2 - etäisyydet origosta
```

```
# simpleCA1$rownames
```



Kuva 4: Q1b: lapsi kärsii jos äiti on töissä

```

# simpleCA1$rowdist
# simpleCA3$rowdist

# simpleCA3$colnames
# simpleCA1$coldist
# simpleCA3$coldist

# massat - huom! Riviprofiilien ca: rivimassojen summa on 1 !(5.9.20)
# sum(simpleCA1$rowmass)
# sum(simpleCA3$rowmass)
# rivit

# simpleCA1$rownames
# simpleCA1$rowmass
simpleCA3$rowmass

## [1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667

# sum(simpleCA1$rowmass)
# sum(simpleCA3$rowmass)

# sarakkeet

# simpleCA3$colnames
# simpleCA1$colmass
# simpleCA3$colmass
# sum(simpleCA1$colmass)
# sum(simpleCA3$colmass)

# rivien khii2 - etäisyydet, massat ja vakiodut massat
rowdist.tbl <- as_tibble(rbind(simpleCA1$rowdist, simpleCA3$rowdist), .name_repair = c("unique"))

## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
## * ...

str(rowdist.tbl)

## tibble [2 x 6] (S3: tbl_df/tbl/data.frame)
## $ ...1: num [1:2] 0.158 0.147
## $ ...2: num [1:2] 0.631 0.59
## $ ...3: num [1:2] 0.175 0.206
## $ ...4: num [1:2] 0.634 0.689

```

```

## $ ...5: num [1:2] 0.348 0.384
## $ ...6: num [1:2] 0.55 0.508
colnames(rowdist.tbl) <- simpleCA1$rownames

knitr::kable(rowdist.tbl,digits = 3,
             caption = "Rivietäisyydet keskiarvosta (khii2) - toisella rivilla rivimassat yh",
             booktabs = TRUE)

```

Taulukko 56: Rivietäisyydet keskiarvosta (khii2) - toisella rivilla rivimassat yhtä suuret

BE	BG	DE	DK	FI	HU
0.158	0.631	0.175	0.634	0.348	0.550
0.147	0.590	0.206	0.689	0.384	0.508

```

# rowdist.tbl ok 15.9.20

# sarakkeiden khii2 - etäisyydet, massat ja vakioidut massat

coldist.tbl <- as_tibble(rbind(simpleCA1$coldist, simpleCA3$coldist), .name_repair = c("unique"))

## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5

# coldist.tbl
colnames(coldist.tbl) <- simpleCA1$colnames
rowid_to_column(coldist.tbl)

```

rowid	S	s	?	e	E
1	0.5246525	0.3248840	0.3078230	0.2721699	0.6271108
2	0.5587368	0.3567818	0.3025459	0.2944703	0.7190317

```

# print(coldist.tbl)

knitr::kable(coldist.tbl,digits = 3, booktabs = TRUE,
              caption = "Sarake-etaisyydet keskiarvosta (khii2) - toisella rivilla rivimassat yh"

```

Taulukko 58: Sarake-etäisyydet keskiarvosta (khii2) - toisella rivilla rivimassat yhtä suuret

	S	s	?	e	E
	0.525	0.325	0.308	0.272	0.627
	0.559	0.357	0.303	0.294	0.719

```
colmass.tbl <- as_tibble(rbind(simpleCA1$coldist, simpleCA3$coldist), .name_repair = c("unique"))
## New names:
## * `` -> ...1
## * `` -> ...2
## * `` -> ...3
## * `` -> ...4
## * `` -> ...5
colnames(colmass.tbl) <- simpleCA1$colnames
rowid_to_column(colmass.tbl)
```

rowid	S	s	?	e	E
1	0.5246525	0.3248840	0.3078230	0.2721699	0.6271108
2	0.5587368	0.3567818	0.3025459	0.2944703	0.7190317

```
knitr::kable(colmass.tbl, digits = 3, booktabs = TRUE,
             caption = "Sarakkeiden massat - toisella rivilla rivimassat yhtä suuret")
```

Taulukko 60: Sarakkeiden massat - toisella rivilla rivimassat yhtä suuret

	S	s	?	e	E
	0.525	0.325	0.308	0.272	0.627
	0.559	0.357	0.303	0.294	0.719

Rivi- ja sarake-etäisyydet (khii2) keskiarvosta rivien massoilla ja vakioiduilla massoilla. Massoiltaan (havaintojen lukumäärä tai joitain siihen verrannollista) pienet rivit HU ja BG ovat vakiomassojen ratkaisussa lähempänä keskiarvoa. Niin on myös Belgia. Saksa, Tanska ja Suomi ovat kauempana origosta. **Erot kuitenkin pieniä (17.9.20)**

Grafiikan hienosäätiö on hieman haastavaa: analyysivaiheessa kannattaa tallentaa kuvia RStudion grafiikkakunasta pdf-muodossa talteen, graafisessa data-analyysissä niitä tieteenkin syntyy aika paljon. HTML- ja pdf- formaatin kuvat viimeistellään bookdown-ympäristössä.

2.2 Korresponduenssianalyysin käsitteet

edit 15.9.20 Näitä taulukoitu edellä, vertailtu "normaaleilla massoilla" ja vakioiduilla rivimassilla (riviprofileilla) laskettuja ca-ratkaisuja, khii2-etäisyyksiä ja sarakemassojen skaalautumista kun rivimassat vakioidaan.

Triplet

1. Profilit
2. Massat
3. Profilien etäisyydet (khii2): (a) saman pistejoukon pisteen (b) eri pistejoukkon pisteen

Tätä "triplettiä" täydentää neljä siitä johdettua käsitettä, viite muistiinpanoissa.
#V Tässäkin CAIP ja MG2017HY-luentokalvot.

3 Tulkinnan perusteita

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

Tärkein asia CA:ssa kaikki on suhteellista

Tärkeä asia 1 Symmetrinen kuva, kaksi pistepilveä samassa koordinaatistossa. **Tärkeä asia 2** Rivi- ja sarakeratkaisun duaalisuus (vai käsitelläänkö jo johdattelevassa esimerkissä?)

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

Asymmetrinen kartta, jossa riviprofilit ovat päärakenteen koordinaateissa ja sarakeprofilit standardkoordinaateissa.

- (1) Sarakkeet ideaalipisteinä, edustavat kuvittellisia maita joissa kaikki ovat vastanneet vain yhdellä tavalla. Sarakepisteet ovat barysentrisen koordinaatiston akselita.
- (2) Sarakepisteet kaukana origosta, koska skaalattu ja
- (3) Rivipisteet kasautuneet keskiarvopisteen ympärille. Symmetrinen kuva on usein hyvä oletus tästä syystä.
- (4) Rivi- ja sarakepisteiden suhteelliset sijannit samat kuin symmetrisessä kuvassa
- (5) Tässäkin kuussa pisteen koko kuva sen massaa. Sarakkeista "täysin samaa mieltä" (ts) ja "ei samaa eikä eri mieltä" ovat massoiltaan pienimmät.

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

(6) 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 lähettilä 45 asteen kulmaa, sarake määrittää yhtä paljon molempia ulottuvuuksia. **#V lähde?**

Standardikooridaateissa esitetyt sarakepisteet ovat fiktiivisiä "maapisteitä", joissa kaikki vastaukset ovat yhdessä luokittelumuuttujan arvossa. Alkuperäisessä täydessä avaruudessa ne ovat simpleksin kärkipisteet, simpleksin sisällä ovat riviprofilit.

```
# asymmetrinen kartta - rivit pc ja sarakkeet sc
# sarakkeet vektorikuvina
# HUOM! simpleCA1 luodaan G1_2_johdesim.Rmd - tiedostossa
#
#
#
# Kuva tiedostoon - ennen plot-komentoa avataan tiedosto
# pdf("img/sCA1asymm1.pdf")
par(cex = 0.6)
plot(simpleCA1, map = "rowprincipal",
      arrows = c(FALSE, TRUE),
      # main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta 1
      sub = "asymmetrinen kartta")

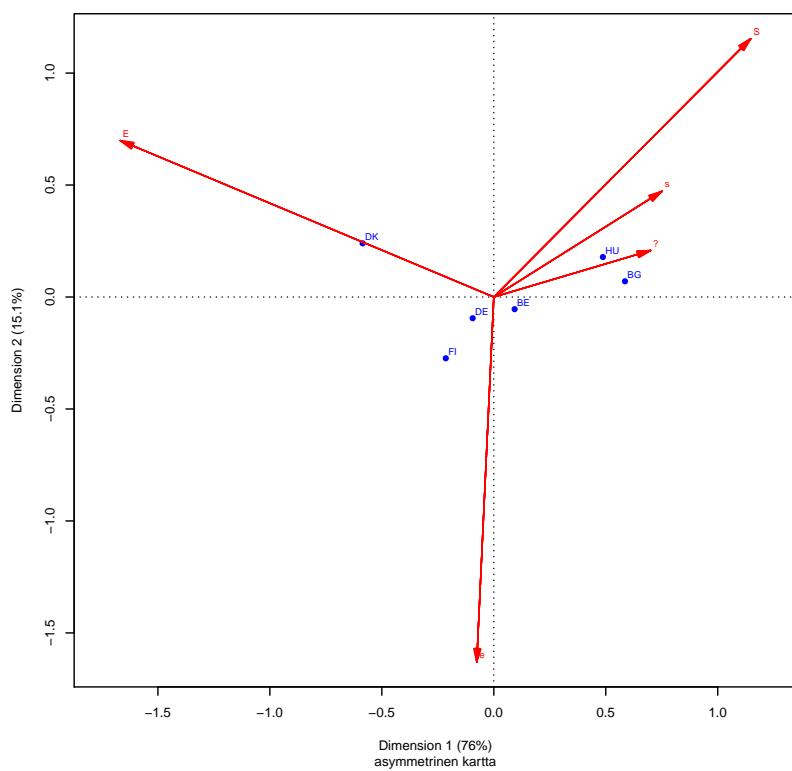
# Kuva tiedostoon - suljetaan
# dev.off()
```

Sarakkeen "Eri mieltä" (e) määrittää toisen ulottuvuuden, jonka voisi tulkita erottelevan "maltilliset" mielipiteen tiukemmista. Sarake "täysin samaa mieltä" (S) määrittää toista ulottuvuutta lähes yhtä paljon kuin ensimmäistä, mutta "täysin eri mieltä" (E) on vasemmalla ja kolme vastausvaihtoehtoa oikealla. Kovin terävästi dimensiot eivät eroa toisistaan?

Asymmetrinen kartta - rivipisteet (profilit) sarakepisteiden standardikoordinaatien keskiarvopisteinä (ns. barysentrisen keskiarvo).

```
# Barysentrisen keskiarvon "viivakuviota" kehitetty CA_calc1.R - skriptissä
# simpleCA1-objektista saa std-koordinaatit, muunnoksella rivien pääkoordinaatit
# rpc.

# Jos plot-komennotoon "MapObj1 <- ", saadaan pisteen koordinaatit
# plot-funktiolla ensin "raamit" ja pisteen talteen, sitten pisteen Suomen
# pistestä lines(x,y) sarakevektoreihin? (29.5.20)
```



Kuva 5: Q1b: lapsi kärsii jos äiti on töissä

```

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

# Kuva tiedostoon - ennen plot-komentoa avataan tiedosto
# pdf("img/sCA1asymm1.pdf") ja lopuksi suljetaan tiedosto

# Piirretään Suomen riviprofiilista janat sarakepisteisiin - barysentrisen keskiarvo
# Rivipisteet pääkoordinaatteina (principal coordinates)

simpleCA1.rpc <- simpleCA1$rowcoord %*% diag(simpleCA1$sv)

# X11()
par(cex = 0.6)
plot(simpleCA1, map = "rowprincipal",
      arrows = c(FALSE, FALSE),
      # main = "Lapsi kärsii jos äiti on töissä -asymmetrinen kartta 2",
      sub = "Suomen profiili sarakkeiden barysentrisenä keskiarvona")
segments(simpleCA1.rpc[5,1],simpleCA1.rpc[5,2],simpleCA1$colcoord[, 1],
         simpleCA1$colcoord[, 2], col = "pink")

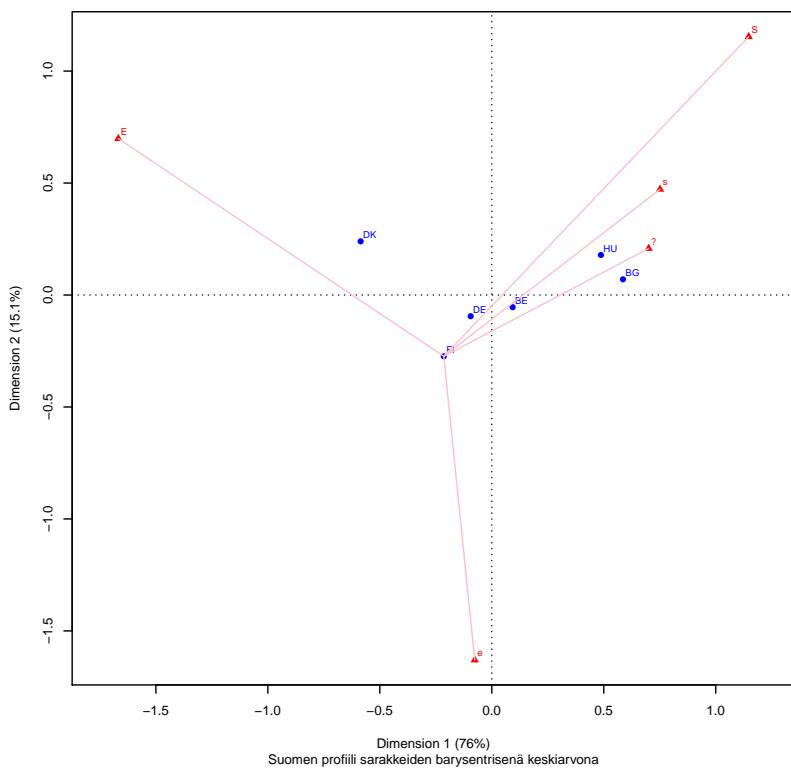
# Kuva tiedostoon - suljetaan
# dev.off()

```

Edit 3.5.20 Selvää: sarakevektroit ovat standardikoordinaateissa, ideaalipisteitä (“maa jossa kaikki samaa mieltä”). Miksi ne ovat kartalla “reilusti” ykköstä suurempia? Vastaus: ideaalipisteet esitetään rivipisteiden koordinaatistossa - > skaalaus.

Edit 11.6.20 - tulkinta ja data?TARKISTA Origo on koko aineiston barysentrisen keskipiste. Janan pituus on kääntäen verrannollinen sarakkeen (“ideaalipisteen”) suhteelliseen osuuteen. Maapiste (profilipiste) on saravektoreiden barysentrisen keskiarvo, ja etäisyydet kertovat kyseisen sarakkeen suhteellisen osuuden maaprofilisissa.(? 11.6.20). " In an asymmetric map where the rows, for example, are in principal co-ordinates (i.e. the row analysis), distances between displayed row points are approximate khii2-distances between row profiles; and distances from the row profile points to a column vertex point are, as a general rule, inversely related to the row profile elements for that column." CAiP, s. 72., tarkemmin s.62-. Pisteiden väliset etäisyydet voidaan optimaalisessa tilanteessa (symm. kuva sama pistejoukko, asymm. kuva myös sarake- ja rivipisteet) tulkita vain approksimaatioina.

Verifying the profile-vertex interpretation (emt., s 68): Each row profile point (staff group) is at a weighted average position of the column vertex points (smoking categories), where the weights are the elements of the respective row



Kuva 6: Q1b: lapsi kärsii jos äiti on töissä

profile. As a general rule, assuming that the display is of good quality, which is true in this case, the closer a profile is to that vertex, the higher its profie value is for that category.

Verifointi verteksi kerrallaan, kaikki on suhteellista! Suomi on kaikkein lähimänä e-verteksiä. Niinpä Suomen profiilissa e-vastausten suhteellinen osuus on suurempi kuin muilla mailla. Tanska vastaanasti lähipänä E-verteksiä. Projisoidaan rivipisteet sarakevektoreille -> järjestys. **todo 10.10.20** Tarkista tämä tulkinta!

Perusidea: kartta antaa yleiskuvan riippuvuudesta, approksimaation tarkkuuden ja laadun rajoissa. Yksityiskohtien etsikely ei ole oleellista, väärrien johtopäätösten välttäminen on. Erityisesti symmetrisessä kartassa ei voi tulkita mitenkään (tiukasti ottaen) eri pistejoukkojen etäisyysjä. Ei voi tunnistaa klustereita!

edit 14.8.20 Barysentrinen koordinaatisto on ideaalipisteiden simpleksin kärkipisteiden koordinaatisto.

```
# X11() komentoriville ja plot-komento -> grafiikkaikkuna
par(cex = 0.6)
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 2a (rowgreen)",
      sub = "sarakevektorin ja rivipisteiden värin tummuus = kontribuutio(absolute)")
```

Greenacre (2006, “loose ends -artikkeli”) ehdotti asymmetrisessä kuvassa standardikoordinaattien skaalaamista niin, että ne kerrotaan massan neliöjuurella. Tämä skaalaus toimii hyvin pienien ja suuren inertian tapauksessa. Kartoissa pätee sama sääntö kuin muussakin graafisessa data-analyyisissä, kuvien on esitettävä oleelliset yhteydet, mutta mielellään vain ne. **todo 10.10.20** Inertian maksimi on ratkaisun dimensioiden suurin lukumäärä. **teoria-jaksoon**.

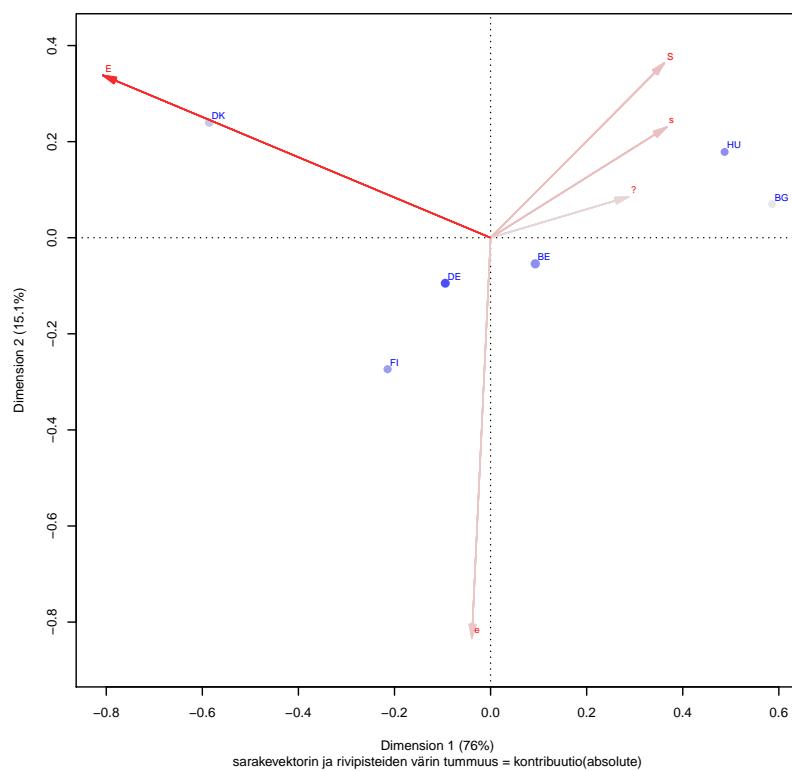
Sama kuva, kontribuutiot “relative”. **edit 24.2.20** Ero selittäävä!

```
#X11() komentoriville ja plot-komento
par(cex = 0.6)
plot(simpleCA1, map = "rowgreen",
      contrib = c("relative", "relative"),
      mass = c(TRUE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Lapsi kärsii jos äiti on töissä - asymmetrinen kartta 2b (rowgreen)",
      sub = "sarakevektorin ja pisteen värin tummuus = kontribuutio(relative)")
```

Asymmetrisessä kartassa 2 pisteiden koko on suhteessa niiden massaan, ja värisävy absoluuttiseen tai suhteelliseen kontribuutioon.

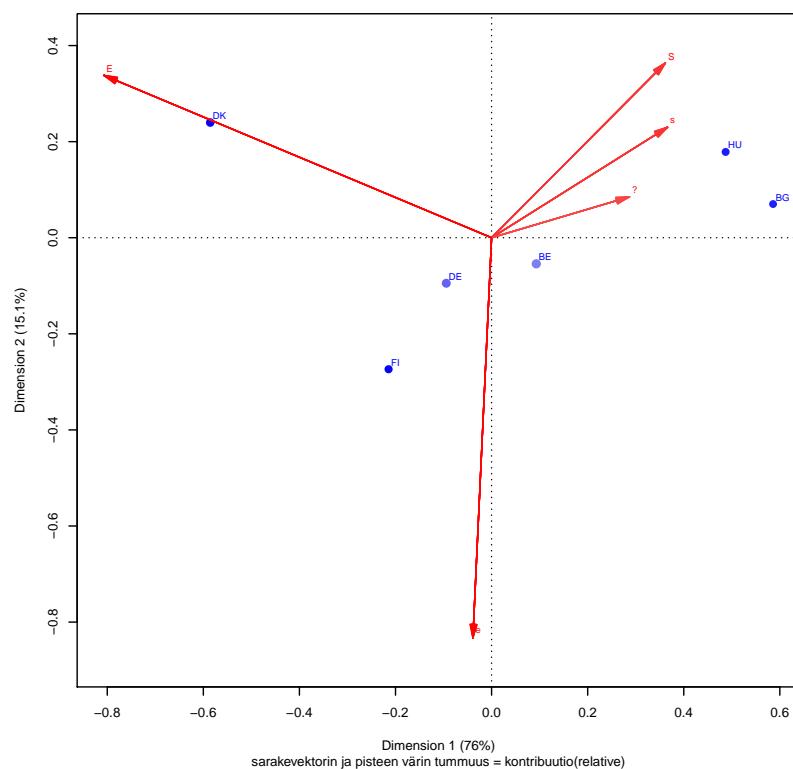
Sarakevektorin kulma akseliin - mitä pienempi sitä enemmän määrittää. Jos

Lapsi kärsii jos äiti on töissä – asymmetrinen kartta 2a (rowgreen)



Kuva 7: Q1b: lapsi kärsii jos äiti on töissä

Lapsi kärsii jos äiti on töissä – asymmetrinen kartta 2b (rowgreen)



Kuva 8: Q1b: lapsi kärsii jos äiti on töissä

lähellä 45 asteen lävistääjää, kontribuutiota on molempien akseleihin

Tulkinta: rivipisteiden ortogonaalinen projektio “sarakevektorille”

Rivipisteet voidaan projisoida ortogonaalisesti sarakevektorille ja sen pisteillä merkitylle jatkeelle. Järjestys on sama kuin sarakkeen modaliteetin suhteellinen osuuus rivipisteen profilissa. Ensimmäisessä kartassa on käsivaralla piirretty suurinpiirtein kohtisuorat projektiot lievempää erimielisyyttä edustavalle s-ideaalipisteeseen (0,1,0,0,0) origosta piirrettylle suoralle.

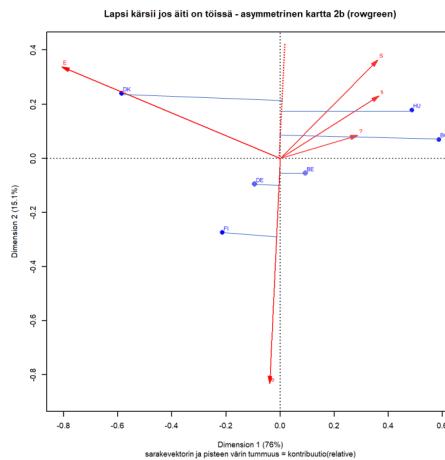
Onko näin? (17.9.20) Toisessa kuvassa sama E-vektorille. Maiden suhteellinen järjestys on oikea, mutta tarkasti etäisyydet eivät (välimatka-asteikolla) ole vertailukelpoisia. Kartta on approksimaatio. Asymmetrisen kartan kuitenkin kaksoiskuva (biplot).

Vertailun vuoksi riviprofilien sarakkeita vastavat arvot suuruusjärjestyksessä:

e: FI 38, DE 31, BE 27, BG 21, HU 19, DK 17 (%-yksikköä)

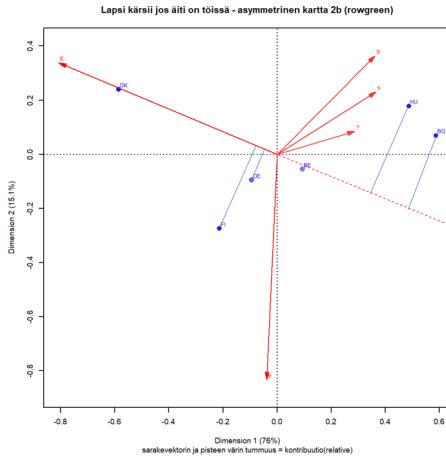
todo 15.9.20 Voiko tämän tehdä myös skaalatulle (rowgreen) asymmetriselle kuvalle?

```
knitr::include_graphics('img/simpleCAasymmTulk1.png')
```



E: DK 50, FI 27, DE 26, BE 19, HU 8, BG 1 (%-yksikköä)

```
knitr::include_graphics('img/simpleCAasymmTulk2.png')
```



“kaikki on suhteellista” Esimerkiksi Tanskan riviprofilissa E-vastauksien suhteellinen osuus poikkeaa eniten keskiarvopisteestä (origossa). Asymmetrisessä kuvassa voimme projisoida rivipisteet sarakepisteen vektorille.

Korresponduenssianalyysin ratkaisun yksityiskohdat: ca-funktion tuloste ja “tulosobjekti”. **Tulosobjekti “teoriajaksoon”, vai viitataako siellä tähän kohtaan?**

Numeeriset tulokset, jälkimmäisessä rivimassat vakioitu yhtä suuriksi (“riviprofilaulukon ca”).

```
# CA:n numeeriset tulokset
# (11.4.20) yhdistää koodilohkoon khii2dist1 (G1_2_johdesim.Rmd, r. 665)
# CA:n numeeristen tulosten käsitteilyä myös CAcalc_1.R -skriptissä.

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 |

# vertailun vuoksi numeeriset tulokset, kun maiden massat vakiot
summary(simpleCA3)

##
## Principal inertias (eigenvalues):
##
##   dim value % cum% scree plot
## 1 0.167678 77.9 77.9 *****
## 2 0.030095 14.0 91.9 ***
## 3 0.013206 6.1 98.0 **
## 4 0.004296 2.0 100.0
##   -----
## Total: 0.215275 100.0
##
##
## Rows:
##   name mass qlt inr k=1 cor ctr k=2 cor ctr
## 1 | BE | 167 295 17 | 46 97 2 | -66 199 24 |
## 2 | BG | 167 884 270 | 554 882 306 | 22 1 3 |
## 3 | DE | 167 718 33 | -144 489 21 | -98 229 54 |
## 4 | DK | 167 993 367 | -635 849 400 | 261 144 377 |
## 5 | FI | 167 999 114 | -270 494 72 | -272 505 411 |
## 6 | HU | 167 870 200 | 448 778 199 | 154 92 132 |
##
## Columns:
##   name mass qlt inr k=1 cor ctr k=2 cor ctr
## 1 | S | 105 785 153 | 450 649 127 | 206 135 148 |
## 2 | s | 250 792 148 | 311 762 145 | 62 30 32 |
## 3 | | 171 792 73 | 267 780 73 | 33 12 6 |
## 4 | e | 256 976 103 | -66 51 7 | -283 925 681 |
## 5 | E | 218 1000 524 | -706 964 649 | 136 36 133 |

```

```

# Rivi- ja sarake-etäisyydet (keskiarvosta/sentroidista)
# HUOM! Edellisessä jaksossa taulukko rivi- ja sarake-etäisyyksistä. Tuskin
# kannattaa tässä toistaa. Muuta analyysiä numeerisista tuloksista. (10.4.20)

# simpleCA1$rownames
# simpleCA1$rowdist
# simpleCA3$rowdist

# simpleCA1$colnames
# simpleCA1$coldist
# simpleCA3$coldist

# Hieman laskentaa

# Massojen summat 1 - onko ero ca-tulosten massoissa pyöristysvirhe? (17.10.20)
#sum(simpleCA1$colmass)
#sum(simpleCA3$colmass)
#testMassDiff <- simpleCA3$colmass - simpleCA1$colmass
#testMassDiff
# 0.005812145 0.012608847 0.002977493 -0.005426916 -0.015971569

```

editEdellisessä jaksossa esimerkki siistimmästä taulukosta,samoin bookdown-testiprojektissa.

Tulkinta: - ensin ratkaisu alkuperäisillä massoilla (otoskoot), sitten vertaillaan ratkaisuun vakiomassilla -> massojen vaikutus.

1. Pääakselien inertiat - ratkaisun yleinen laatu (ne prosentit kuvissa!)
 - maksimi-inertia (teoreettinen) on 4, yleisesti $\min(J-1, K-1)$, J on rivien ja K sarakkeiden lukumäärä
 - alkuperäisten “pistepilvien” inertia jaetaan ca-ratkaisussa pääakseleille suurimmasta pienimpään (esim. Rows:inr - sarakkeen summa on 1)
 - kokonaisinertia 0.18 (vertailu teoreettiseen maksimiin ei kovin kiinnostava?)
 - ensimmäinen akseli ei täysin dominoi, mutta kaksi ensimmäistä kuvaavat jo 91 prosenttia aineiston kokonaishajonnasta (käytän termejä hajonta ja inertia vaihtelevästi)
 - 9 prosenttia jää kuitenkin 3. ja 4. dimensiolle
2. Kartta tulkitaan katseella - numeerisista tuloksista tarkistetaan ja varmistetaan

Mitä on vasemmalla ja oikealla (1. dimensio 76 % inertiesta), ylhällä ja alhaalla (15 % inertiesta) -> sarakkeet ja niiden avulla akselien tulkinta (etäisyydet appr. khii2).

Pisteet ja niiden (suhteellinen sijanti) origoon ja toisiinsa (samassa pilvessä) appr. kii2 jos kvalitetti kelvollinen. Muuten “lähellä voi olla kaukana”, kaukana on kuitenkin aika kaukana.

Numeerista tuloksista katsotaan:

Sarakkeet ja akseleiden tulkinnan tarkistus (ensin):

1. Kvalitetti kaikilla ok (>785): sarakkeet ja niiden etäisyydet hyvin esitetty
2. Sarakkeiden osuudet kokonaisinertiasta: E-sarake yli puolet (524), eos-sarake pienin (73).
3. Akselien tulkinnan varmistaminen (pääakselit ja koordinaatit), koordinaatien suunta (+ vai -)

cor: relative contributions (out of 1000) of each dimension to the inertia of individual points. These are also interpreted as squared correlations (1000)

CA-ratkaisun ulottuvuuksien (yleensä 2) suhteellinen kontribuution pisteen inertialle; kuinka ”kaukana” tai ”lähellä” piste on akseleiden määräಮäästä tasosta? CA-ratkaisu minimoi massoilla painotetut poikkeamat.

cor1 + cor2 = qlt

ctr: contributions (out of 1000) of each point to the principal inertia of a dimension - pisteen suhteellinen kontribuutio akselin (pää) inertialle. Huom! suhteellinen, akselin osuus pilven inertiasta huomioitava.

ensimmäinen akseli (76% kokonaisinertiasta); E “tasossa kiinni” (cor=966) ja osuus akselin kontribuutiosta yli puolet (651), koordinaatin etumerkistä näkyy suunta (-). - oikealle (+) S ja s yht neljännes (265), eos(“?”) hieman ja e mitään (2) - ensimmäinen akseli ”selittää” kaikkien muiden pisteen inertiasta suurimman osan, e poikkeus (cor 11)

- hyvin selkeää, visuaalinen tulkinta on oikea E vasemmalla - kontrastina oikealla S ja s, eos-kategoria hieman.

toinen akseli (15% kokonaisinertiasta): e “tasossa kiinni” (cor 971) ja osuus akselin inertiasta n. 70% (mutta akselin osuus kokonaisinertiasta 15%), suunta (-). eos ei vaikuta akseliin(ctr 7) eikä akseli eos-pisteen inertiaan(cor.

Positiiviseen suuntaan akselin inertiaan vaikuttavat tiukat mielipiteet E (ctr 114) ja S (ctr 132), s jonkin verran (53)

Mallillinen ”eri mieltä” määritää yksin toisen dimension, hieman yllättäen neutraalilla (eos) vaihtoehdolla ei ole yhteyttä dimensioon. Sen ainoa vaikutus ratkaisuun (kun inertia näin dekomponoidaan) on pieni S-suuntaan ensimmäisellä aksellilla.

ctr/cor - epäsymmetrinen suhde (teoriajaksoon).

Rivit

1. kvaliteeetti (qlt): onko joku piste huonosti edustettu ratkaisussa (projektiossa)?
 - Belgia qlt kehnoin, Saksan toiseksi kehnoin. Belgialla suurin massa, Saksalla toiseksi suurin, ja silti niiden vaikutus karttaan (ctr molemmille akseleille) on heikko
2. inr: part of total inertia of the point in the full space (rows or columns)
 - pisteen suhteellinen osuus (1000) koko alkuperäisen pilven inertiesta
 - DK suurin (381), BG ja HU lähes puolet (yht. 457), kolme muuta yhteensä 16 %
 - taas hämmentää Saksan ja Belgian pieni osuus kokonaisinertiaasta
 - vai onko seuraus siitä, että massa molemmilla suuri? Lähellä aineiston keskiarvopistettä? Tämä nähdään kun verrataan CA-ratkaisuun jossa mailla sama paino!

Vertailu vakioitujen rivimassojen ca-ratkaisuun

1. Tätä voisi käyttää esimerkinä numeeristen tulosten vertailussa?
2. Kokonaisinertia kasvaa ($0,18 \rightarrow 0,22$), koordinaatisto muuttuu mutta ei kovin radikaalisti.
3. Kvaliteetti, kontribuutiot? Miten vertailla oleellisiaasioita?

Belgian laatu putoavaa, ja kontribuutiot pienenevät entisestään.

Saksan laatu paranee aika paljon, ja kontribuutiotkin jonkin verran. Aika ou-toa, että suurimman massan maiden (DE,BE lähes puolet datasta) kontribuutiot ovat niin pieniä (24.2.20).

Saksa siirtyy x-akselilla vasemmalle (x-koordinaatti $-94 \rightarrow -144$). Belgia siiryy x-akselilla läheemmäs origoa ($93 \rightarrow 46$) ja y-akselilla vähän kauemmas. Eihän tässä näin pitänyt käydä!

Kuvasta näkee muutoksen vähäisyyden. Aineisto on pieni, mutta kuva on selvästi tehokkaampi tapa kuvata taulukon rivien ja sarakkeiden yhteyksiä.

4. **Sarakemassojen muutos?** Kaavaliitteessä eksplisiittisesti rivi- ja sarakeratkaisujen yhteys, massat mukana. Sarake- ja rivimassojen summat ovat
 1. Sarekemassat ja inr muuttuvat hieman. Onko pyöristysvirhettä?

4 Yksinkertaisen korrespondenssianalyysin laajennuksia 1

Korrespondenssianalyysi sallii rivien tai sarakkeiden yhdistelyn tai "jakamisen". Tämä onnistuu esimerkkiaineistossa lisäämällä rivejä eli jakamalla eri maiden vastausksia useampaan ryhmään. Kartalle voidaan myös lisätä apumuuttuja tai

täydentäviä muuttuja (supplementary points). Ne eivät vaikuta ca:n tuloksiin, vaan esittävät lisätietoa.

edit 9.9.20 - vähän vanhentunutta Sen avulla voi myös tarkastella ja vertailua erilaisia ryhmien välisiä tai ryhmien sisäisiä (within groups - between groups) eroja hieman. Teknisesti yksinkertaista korrespondenssianalyysiä sovelletaan muodukkauun matriisiin. Datamatriisi rakennetaan useammasta alimatriisista, joko "pinoamalla" osamatriiseja (stacked matrices) tai muodostamalla symmetrisen lohkomatriisi (ABBA).

edit 9.9.20 Uusi johdanto 1. Saksan ja Belgian jako

- jakamaekvivalenssi(?) (distributional equivalence), voidaan jakaa ja yhdistää rivejä
- Belgia leviää pystysuuntaan, Saksa vaakasuuntaan

```
# HUOM! Tässä ei vielä supp.points mukana!
par(cex = 0.6)
suppointCA1 <- ca(~maa3 + Q1b, ISSP2012esim1.dat)
plot(suppointCA1, main = "Belgian ja Saksan ositteet",
      sub = "symmetrinen kartta")
```

2. Täydentävät pisteet (supplementary points)

- eivät vaikuta karttaan (koordinaatistoon, rivi- tai sarakepisteiden sijaintiin)
- useita tapoja käyttää, tässä Saksan ja Belgian maapisteet
- barysentriinen keskiarvo
- laajempi esittely capaper-projektissa

3. Lisämuuttujat - yksinkertainen ca useamman muuttujan analyysissä

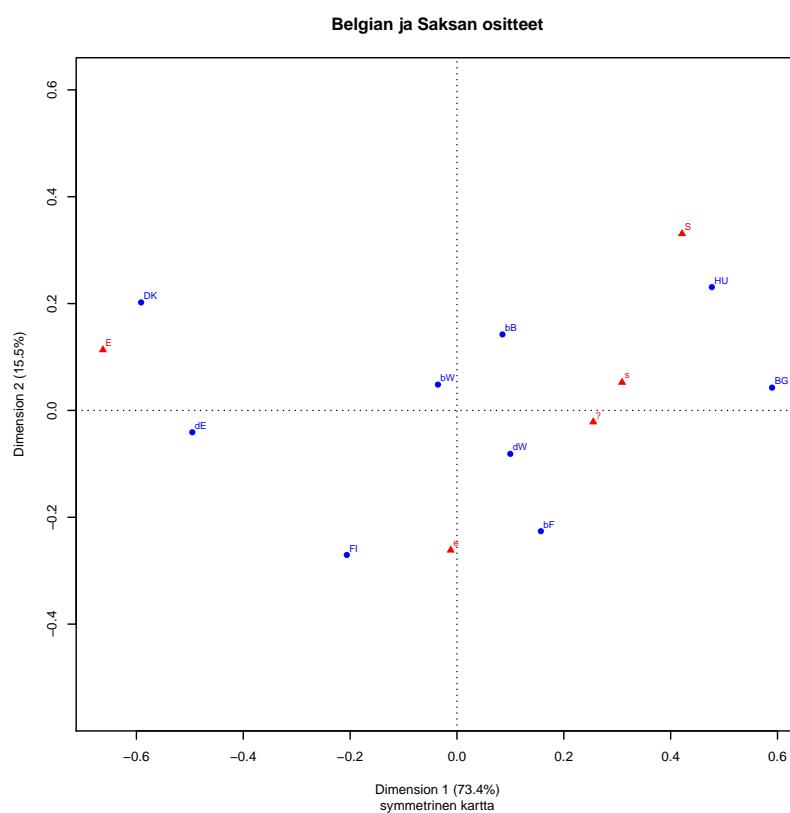
3.1 Yksinkeraisin malli: vuorovaikutusmuuttuja (interactive coding) - yksi rivi (sample) jaetaan useammaksi, sama idea kuin edellisessä esimerkissä - käytännön raja yleensä kolmen muuttujan vuorovaikutusmuuttuja, esimerkinä maan, jän ja sukupuolen yhdistetty muuttuja

3.2 Kartta on data-analyysin väline - miten sitä voi selkeyttää?

Hyvässä kartassa on vain tarpeellinen määrä informaatiota, epäselvä tukkoinen kuva ei toimi. Yksinkertainen keino analyysivaiheessa on "leikata ja liimata" kuvia ja tallentaa ne pdf-muodossa. Voi tehdä muistiinpanoja, lisätä kommentteja jne.

Tässä esitellään kaksi muuta: kuva-alan rajaaminen BaseR-grafiikan plot-funktion parametreilla. Tässä esitellään yleisempi osajoukon korrespondenssianalyysi (subset ca). Osajoukko CA:ssa reunajakaumat (massat) ja valitun datan profilit eivät muutu. Idea on tavallaan tehdä ensin CA-kalkyylit ja valita sitten tarkasteltava joukko.

Paljon parempi idea kuin suoravivainen aineiston rajaaminen, jossa "kaikki muuttuisi" jokaisella aineiston rajaucksella. Yhteys suhteellisten osuu-



Kuva 9: Belgian ja Saksan aluejako

sien/koostumusdatan (compositional DA) analyysiin: ca ei ole “subcompositionally coherent”.

Täyden aineiston inertia voidaan dekompoida l. jakaa osajoukoille, eräs CA:n perusideoita yleisemminkin.

Käytetään tässäkin johdattelevan esimerkin dataa, johon muunnokset on jo alustavasti tehty.

(14.9.20) G1_4_CAlaaj1.Rmd jaettu kolmeksi tiedostoksi, uusia G1_4_CAlaaj2.Rmd ja G1_4_CAlaaj3.Rmd

Vanhaa koodia kolme koodilohkoa

4.1 Täydentävät muuttujat (supplementary points)

zxy Piste sinne piirretään, mutta muuttujassa on se tieto. “Täydentävä piste” kuulostaa huonolta. Lisämuuttujat, havainnot, lisäpisteet?

Viite:CAip ss 89, HY2017_MCA.

Aineistossa on havaintoja (rivejä) tai muuttuja (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 järkevästi painottaa (massa) otoskoolla, vaan vakioidaan (jotenkin) sama (suhteellinen) massa kaikille. Samalla oikaistaan myös naisten yliedustus aineistossa.

Käsitteitä: (a) Active point, aktiivinen piste (aktiivinen havainto tai muuttuja). Pistettä käytetään CA-laskennassa.

- (b) Supplementary point, täydentävä piste (täydentävä havainto), täydentävä rivi- tai sarakeprofiili. Pistettä ei käytetä CA-ratkaisun laskennassa, mutta sillä lasketaan koordinaatit ratkaisuun.

Täydentävien muuttujien kolme käyttötapaa:

1. Sisällöllisesti tutkimusongelman kannalta poikkeava tai erilainen rivi tai sarake
2. Outlayerit, poikkeava havainto jolla pieni massa (esimerkissä uusi sarake-muuttuja, jossa kovin vähän havaintoja)
3. osaryhmät **EDIT** capaper- jäsentelyssä ja bookdown-dokumentissa selitetty täydentävät/lisäpisteet tarkemmin (18.9.2018).

17.10.20 Muutin hieman koodia, aluejako-taulukon rivien järjestys muuttui.

#

Belgian ja Sakasan jako, maapisteet lisäpisteinä 24.5.2018

```

# CA-tulos: suppointCA2

# Be ja De jako maa3-muuttujassa
# str(ISSP2012esim1.dat$maa3)
# attributes(ISSP2012esim1.dat$maa3)

suppoint1_df1 <- select(ISSP2012esim1.dat, maa3,Q1b)

# tarkistuksiin jos koodi suoritetaan rivi kerrallaan
# str(suppoint1_tab1)

suppoint1_tab1 <- table(suppoint1_df1$maa3, suppoint1_df1$Q1b)
suppoint1_tab1

```

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

```

#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)

glimpse(suppoint2_df)

## Rows: 3,727
## Columns: 2
## $ maa <fct> BE, ...
## $ Q1b <fct> e, E, S, e, s, s, s, S, S, s, s, s, s, ?, e, ?, ?, S, ?, ...
suppoint2_tab1 <- table(suppoint2_df$maa, suppoint2_df$Q1b)
suppoint2_tab1 # tarkistus 1

```

/	S	s	?	e	E
BE	191	451	438	552	381
BG	0	0	0	0	0
DE	165	375	198	538	438
DK	0	0	0	0	0
FI	0	0	0	0	0
HU	0	0	0	0	0

```
# Huom! tämä komento vain kerran - tai koko Rmd-tiedosto uudestaan (17.10.20)
suppoint2_tab1 <- suppoint2_tab1[-c(2,4:6) ,]
suppoint2_tab1 # tarkistus 2
```

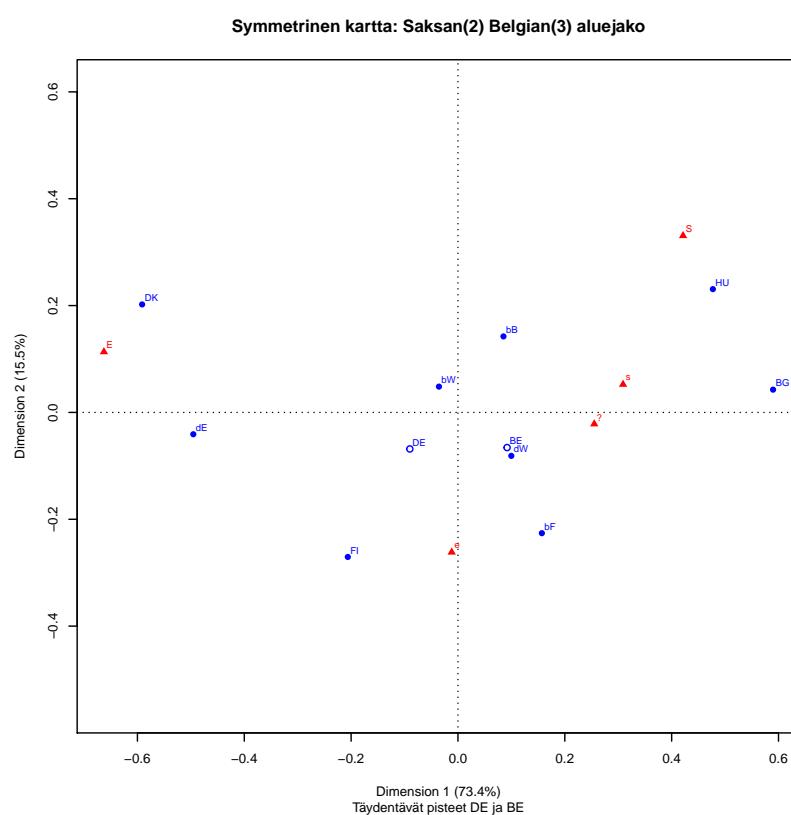
/	S	s	?	e	E
BE	191	451	438	552	381
DE	165	375	198	538	438

```
# suppoint2_tab1 # Belgian ja Saksan summat yli ositteiden
# suppoint2_tab1
#lisätään rivit maa3-muuttujan taulukkoon

suppoint1_tab1 <- rbind(suppoint1_tab1, suppoint2_tab1)
suppoint1_tab1
```

	S	s	?	e	E
bF	51	241	262	312	146
bW	53	103	91	118	125
bB	87	107	85	122	110
BG	118	395	205	190	13
dW	133	313	138	375	208
dE	32	62	60	163	230
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75
BE	191	451	438	552	381
DE	165	375	198	538	438

```
suppointCA2 <- ca(suppoint1_tab1[,1:5], suprow = 10:11)
par(cex = 0.6)
plot(suppointCA2, main = "Symmetrinen kartta: Saksan(2) Belgian(3) aluejako",
     sub = "Täydentävät pistet DE ja BE" )
```



Kuva 10: Belgian ja Saksan aluejako

Saksan ja Belgian summarivit ovat ositteiden painotettuja (barysentrisiä) keskiarvoja (sentoideja), esim.läntisen ja itäisen Saksan rivipisteiden välisellä janalla on koko maan summapiste DE. Entisen länsi-Saksan piste on keskiarvopistettä lähempänä, etäisyys on käänteisessä suhteessa massaan.

18.9.20 Hajonta kasvaa, ja kartan oikealla puolella sarakkeiden etäisyydet kasvavat. Erityisest S-piste nousee toisen dimension suuntaan ylöspäin. Muiden maiden pisteet eivät juuri muuta suhteellisia sijaintejaan, mutta "Belgia leviää pystysuuntaan ja Saksa vaakasuuntaan." Bryssel (bB) ja Flanders (bF) ovat konservatiivisempiä kuin lievästi liberaalimpi Wallonia (bW). Flandersin ja Brysselin ero on mielipiteen varmuudessa tai tiukkuudussa, ja Bryssel saattaa olla polarisoitunut (lievä Guttman-efekti).

```
# riviprofiilitaulukko joteskin (18.9.20)
ISSP2012esim1.dat %>% tableX(maa3, Q1b, type = "row_perc")
```

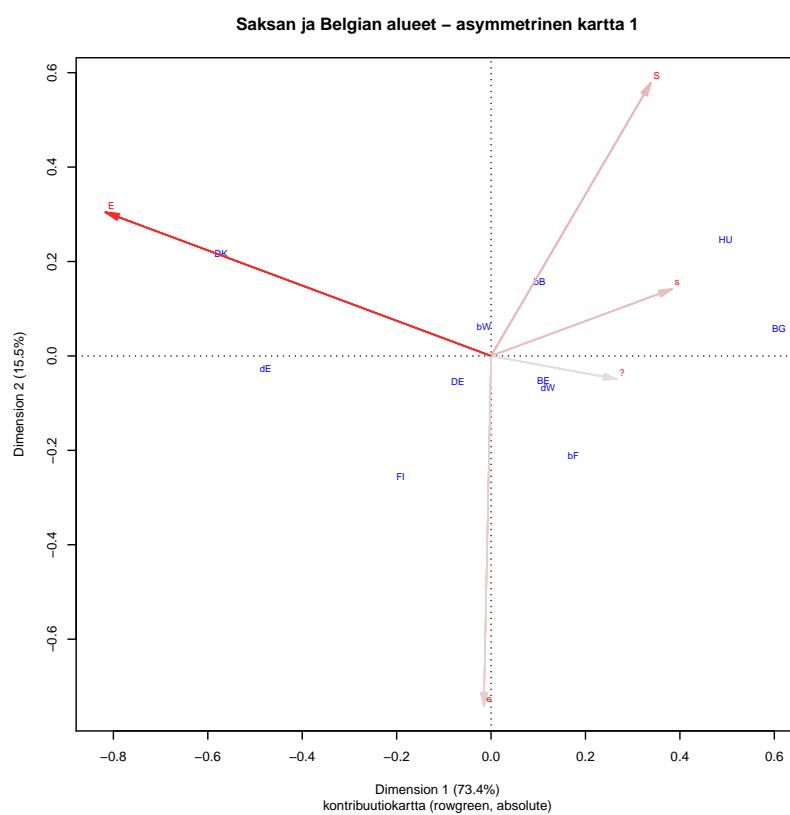
maa3/Q1b	S	s	?	e	E	Total
bF	5.04	23.81	25.89	30.83	14.43	100.00
bW	10.82	21.02	18.57	24.08	25.51	100.00
bB	17.03	20.94	16.63	23.87	21.53	100.00
BG	12.81	42.89	22.26	20.63	1.41	100.00
dW	11.40	26.82	11.83	32.13	17.82	100.00
dE	5.85	11.33	10.97	29.80	42.05	100.00
DK	5.04	17.15	10.95	16.71	50.14	100.00
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

Piirretään vertailun vuoksi vielä asymmetrisen kartta ("kontribuutio-kartta, kontribuutio-kaksoiskuva"). **edit 3.5.20** Minne katoavat pisteet, outoa?

```
par(cex = 0.6)
plot(suppointCA2, map = "rowgreen",
      contrib = c("absolute", "absolute"),
      mass = c(TRUE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Saksan ja Belgian alueet - asymmetrisen kartta 1",
      sub = "kontribuutiokartta (rowgreen, absolute)")
```

Maapisteet täydentävinä pisteinä

```
# Sama kuva, maasummat lisäpisteinä (4.2.20)
par(cex = 0.6)
plot(suppointCA2, map = "rowgreen",
      contrib = c("relative", "relative"),
      mass = c(TRUE, TRUE),
```

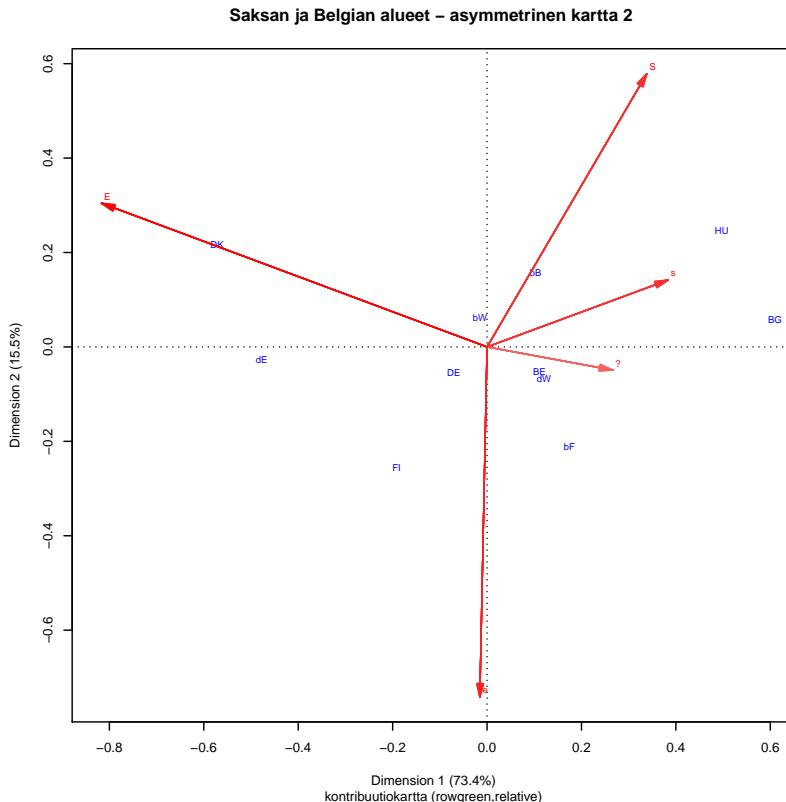


Kuva 11: Belgian ja Saksan aluejako

```

arrows = c(FALSE, TRUE),
main = "Saksan ja Belgian alueet - asymmetrisen kartta 2",
sub = "kontribuutiokartta (rowgreen,relative)")

```



Kuva 12: Belgian ja Saksan aluejako

Kaksi asymmetristä kontribuutio-karttaa (MG:n keksintö) osoittavat, että tulkinnan hankaluksista huolimatta symmetrinen kartta on usein selkeämpi. molemissa ideaalipisteet sijatsevat kaukana, vaikka ne on skaalattu hieman läheemmäs origoa. Maapisteiden hajontaa on aika vaikeaa nähdä. Belgian täydentävä maa-piste (BE) peittyy läntisen Saksan (dW) alle. “rowgreen-kartoista” puuttuvat jostain syystä pistet.

(17.10.20) Kontribuutiokartoista riittää toinen, absolute tai relative.

Tulostetaan numeeriset taulukot.

```
# CA – numeeriset tulokset
```

```
summary(suppointCA2)
```

```

## 
## 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 | bF | 124  650  69 | 157 212  20 | -226 438 195 |
## 2 | bW |  60  388   3 | -36 137   0 |  48 252   4 |
## 3 | bB |  63  481  17 |  85 127   3 | 142 354  39 |
## 4 | BG | 113  878 215 | 590 874 255 |  43  5   6 |
## 5 | dW | 143  345  33 | 100 208   9 | -81 138  29 |
## 6 | dE |  67  966  82 | -495 960 107 | -41  7   3 |
## 7 | DK | 170  971 327 | -591 869 387 | 202 102 214 |
## 8 | FI | 136  957  79 | -206 352  38 | -271 605 307 |
## 9 | HU | 122  927 177 | 477 751 181 | 231 176 201 |
## 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 |

```

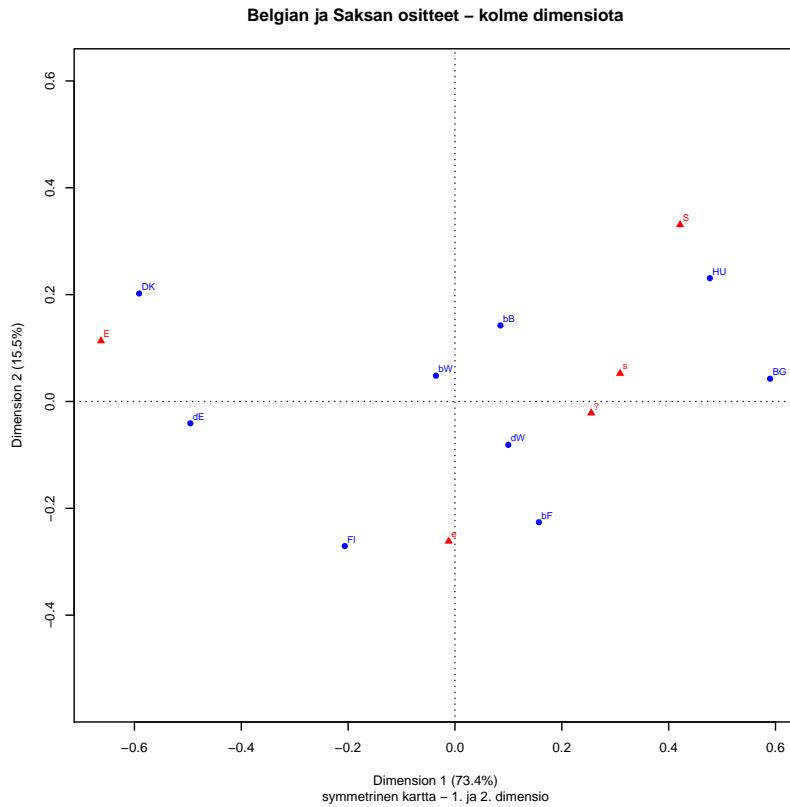
Kolmiulotteisesta kuvasta voi tulostaa molempien akseleiden ja uuden kolmannen akselin kartat. R-ohjelmistossa voi tulostaa näytölle kolmiulotteisen kuvan, siitä voisi ehkä ottaa kuvakaappauksena esimerkin raporttiin?

edit Kommentti 3d-ratkaisusta: tuo esiin Belgian ("belgioiden") erilaisuuden. "Belgiat ovat 2-3 - dimension kartassa diagonaalilla, ja 1-3 kartassa hieman samoin kuin 2-d - ratkaisun kartassa. Tarvittaessa voi liittää myös 3-d - kuvia, pitäisi saada myös dynaamisen pdf-tiedoston? (8.6.2020). Mutta 2d-aproksimaatio on aika hyvä, 89 % kokonaisinertiasta. Miten pitäisi jatkaa? Analysoida maiden sisäisiä eroja? Siinä erilaiset aluejaot ovat aika herkästi korvikemuuttuja joillekin muille vaikuttaville tekijöille. Entäs kaupungit - isot ja pienet - ja maaseutu?

Elinkeinorakenne, tulot jne...

```
# Näkyisikö Belgian aluejako kolmannessa dimensiossa? (19.2.20). Näkyy, ja  
# kaksiulotteisen ratkaisun numeerisista tuloksista näkee myös pistet joiden  
# kvalitetti on heikko.
```

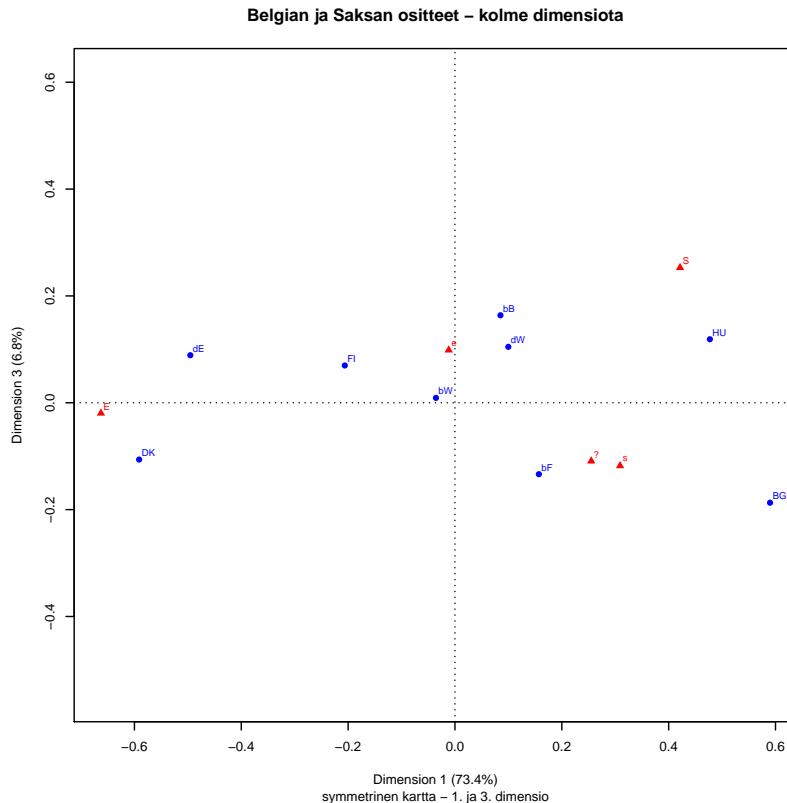
```
suppointCA3 <- ca(~maa3 + Q1b, ISSP2012esim1.dat, nd = 3)  
# (24.2.20)  
# Tulostetaan kolme karttaa - ensimmäinen ja toinen akseli uuden kolmannen kera  
par(cex = 0.6)  
plot(suppointCA3, dim = c(1,2),  
      main = "Belgian ja Saksan ositteet - kolme dimensiota",  
      sub = "symmetrinen kartta - 1. ja 2. dimensio")
```



Kuva 13: Belgian ja Saksan aluejako - 3D

```
par(cex = 0.6)  
plot(suppointCA3, dim = c(1,3),
```

```
main = "Belgian ja Saksan ositteet - kolme dimensiota",
sub = "symmetrinen kartta - 1. ja 3. dimensio")
```

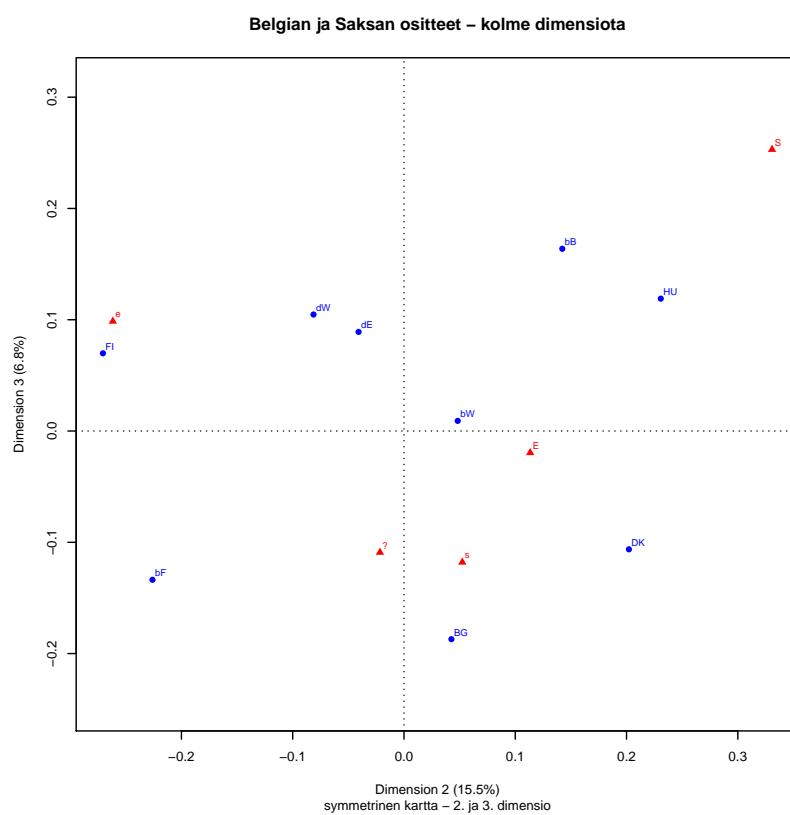


Kuva 14: Belgian ja Saksan aluejako - 3D

```
par(cex = 0.6)
plot(suppointCA3, dim = c(2,3),
      main = "Belgian ja Saksan ositteet - kolme dimensiota",
      sub = "symmetrinen kartta - 2. ja 3. dimensio")

# KUMMALLINEN JUTTU, summary() EI TOIMI KUN 3d-ratkaisu!# KUMMALLINEN JUTTU,
# summary() EI TOIMI KUN 3d-ratkaisu!
# summary(suppointCA3)
# Error in rsc %*% diag(sv) : non-conformable arguments

# Virheilmoitus "Error in rsc %*% diag(sv) : non-conformable arguments" ?!
# onko vika täydentävässä pisteissä? Ei ole, eivät ole mukana
# ISSP2012esim1.dat %>% tableX(maa3, Q1b)
```



Kuva 15: Belgian ja Saksan aluejako - 3D

```

# suppointCA3

# Virheilmoituksen selvittelyä (24.2.20)
# str(suppointCA3)
# tämä matriisikertolasku ei onnistu - 3d-ratkaisussa on vain kolme koordinaattia!
# suppointCA3$rowcoord
# diag(suppointCA3$sv)
# suppointCA1$rowcoord %*% diag(suppointCA1$sv)

#Tämä toimii
#
# suppointCA1$rowcoord
# suppointCA1$sv
# suppointCA1$rowcoord %*% diag(suppointCA1$sv)
# summary(suppointCA1)
#
# Kolmiulottein kuva grafiikkaikkunaan
# X11()
# plot3d(suppointCA3, c(1,2,3))

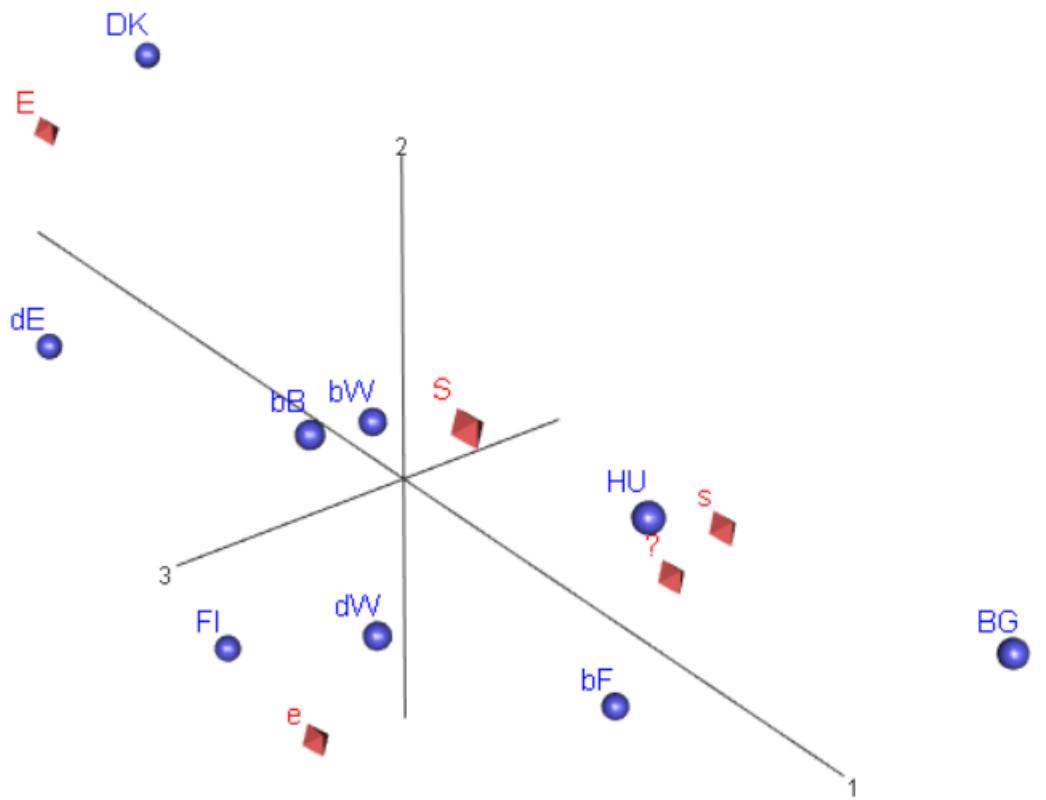
# Hyödyllinen, mutta aika vaikea

```

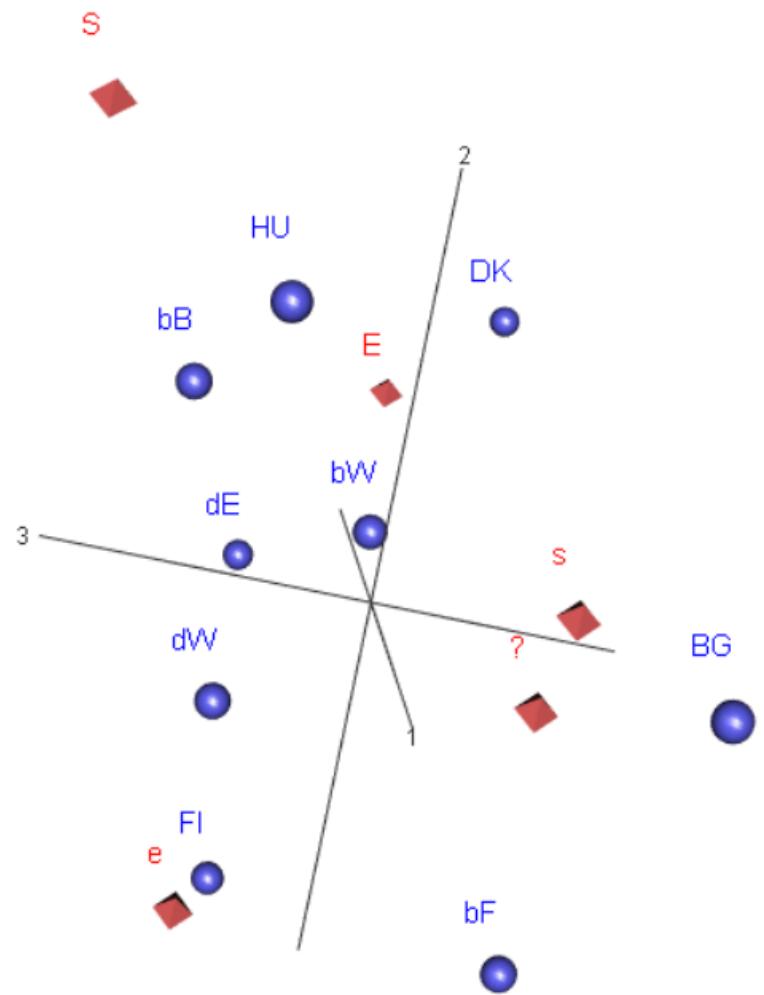
Kolmiulotteisesta kuvasta Belgian ja Saksan alueiden hajonta tasosta näkyy hieman selvemmin jos kuva olisi dynaaminen. On mahdollista, mutta tätä ei nyt tehdä (17.9.20). Kolmannen dimension osuus on noin seitsemän prosenttia kokonaisinertiastä, kaksiulotteisten karttojen tulkinta ei ole ihan helppoa.

TODO 3d-numeeriset tulokset, summary() ei toimi?

```
knitr::include_graphics('img/3dSymMap_1.PNG')
```



```
knitr::include_graphics('img/3dSymMap_2.PNG')
```



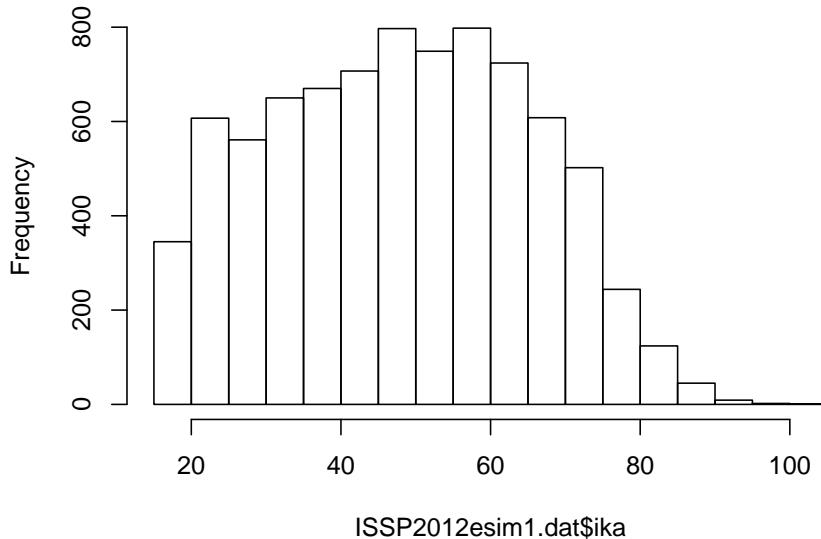
4.2 Korrespondenssianalyysin laajennuksia: vuorovaikutusmuuttuja ja osajoukon CA

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 interaktiomuuttuja 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 vuorovaikutusmuuttuja 1
#
# Uusi R-data: ISSP2012esim1b.dat - MIKSI, TARVITAANKO? *esim1.dat kelpaisi?(4.2.20)
# Vaihdetaan tiedoston nimi (ISSP2012esim2 -> ISSP2012esim1b), ensimmäistä käytetään jo ai
#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$AGE)
hist(ISSP2012esim1.dat$iika)
```

Histogram of ISSP2012esim1.dat\$ika



```
ISSP2012esim1b.dat <- mutate(ISSP2012esim1.dat, age_cat = ifelse(ika %in% 15:25, "1",
  ifelse(ika %in% 26:35, "2",
  ifelse(ika %in% 36:45, "3",
  ifelse(ika %in% 46:55, "4",
  ifelse(ika %in% 56:65, "5", "6")))))

ISSP2012esim1b.dat <- ISSP2012esim1b.dat %>%    # uusi (4.2.20)
  mutate(age_cat = as_factor(age_cat)) # järjestys omituinen! (4.2.20)
# Tarkistuksia

# str(ISSP2012esim2.dat$age_cat)
# levels(ISSP2012esim2.dat$age_cat)
# ISSP2012esim2.dat$age_cat %>% summary()

# Järjestetään ikäluokat uudelleen

ISSP2012esim1b.dat <- ISSP2012esim1b.dat %>%
  mutate(age_cat =
    fct_relevel(age_cat,
      "1",
      "2",
      "3",
```

```

        "4",
        "5",
        "6")
)

# Tarkistuksia

# Iso taulukko, voi tarkistaa että muunno ok.
# test6 %>% tableX(AGE, age_cat, type = "count")
# taulu42 <- ISSP2012esim2.dat %>% tableX(maa,age_cat,type = "count")
# kable(taulu42,digits = 2, caption = "Ikäluokka age_cat")
#
#
# UUdet taulukot (4.2.20)

ISSP2012esim1b.dat %>%
  tableX(maa,age_cat,type = "count") %>%
  kable(digits = 2, caption = "Ikäluokka age_cat")

```

Taulukko 66: 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

```

ISSP2012esim1b.dat %>%
  tableX(maa,age_cat,type = "row_perc") %>%
  kable(digits = 2, caption = "age_cat: suhteelliset frekvenssit")

```

Taulukko 67: age_cat: suhteelliset frekvenssit

	1	2	3	4	5	6	Total
BE	10.33	16.54	16.69	18.63	18.28	19.52	100.00
BG	8.36	12.49	17.26	16.07	21.50	24.32	100.00
DE	11.96	13.01	15.99	20.89	16.80	21.35	100.00
DK	14.91	15.35	17.65	19.52	16.86	15.71	100.00
FI	13.69	14.95	14.86	20.09	21.44	14.95	100.00
HU	10.33	16.15	19.86	17.15	19.66	16.85	100.00

	1	2	3	4	5	6	Total
All	11.69	14.87	16.91	18.99	18.69	18.85	100.00

Ikäjäkauma painottuu kaikissa maissa jonkin verran vanhempia 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).

```
# ga - ikäluokka ja sukupuoli
# Uusi tiedostonimi ISSP2012esim2.dat -> ISSP2012esim1b.dat (10.10.20)

# case_when: ikä ja sukupuoli
ISSP2012esim1b.dat <- mutate(ISSP2012esim1b.dat, ga = case_when((age_cat == "1")&(sp == "m")
  (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(ISSP2012esim1b.dat$ga) # chr-muuttuja, mutta toimii (4.2.20)

## chr [1:8143] "f5" "f3" "m5" "f2" "f4" "f4" "m4" "m3" "f5" "m5" "m3" "f5" ...
# str(ISSP2012esim2.dat)
# str(ISSP2012esim1.dat$ga2)
# ga on merkkijono, samoin ga2, pitäisikö muuttaa faktoriksi?
# str(ISSP2012esim1.dat)

#Tulostetaan taulukkoina ga2 - muuttuja.

ISSP2012esim1b.dat %>% tableX(maa,ga,type = "count") %>%
kable(digits = 2, caption = "Ikäluokka ja sukupuoli ga")
```

Taulukko 68: Ikäluokka ja sukupuoli ga

	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
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

```
ISSP2012esim1b.dat %>% tableX(maa,ga,type = "row_perc") %>%
kable(digits = 2, caption = "ga: suhteelliset frekvenssit")
```

Taulukko 69: ga: suhteelliset frekvenssit

	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
BE	5.76	9.84	8.64	9.89	9.24	9.19	4.57	6.71	8.05	8.74	9.04	10.33	100.00
BG	4.34	6.95	10.21	9.23	12.38	16.18	4.02	5.54	7.06	6.84	9.12	8.14	100.00
DE	5.95	7.00	8.87	10.85	7.88	10.79	6.01	6.01	7.12	10.04	8.93	10.56	100.00
DK	5.98	7.93	9.80	10.52	9.22	7.13	8.93	7.42	7.85	9.01	7.64	8.57	100.00
FI	8.47	8.56	8.47	10.63	12.79	8.20	5.23	6.40	6.40	9.46	8.65	6.76	100.00
HU	5.42	8.63	9.53	9.13	9.43	10.43	4.91	7.52	10.33	8.02	10.23	6.42	100.00
All	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.

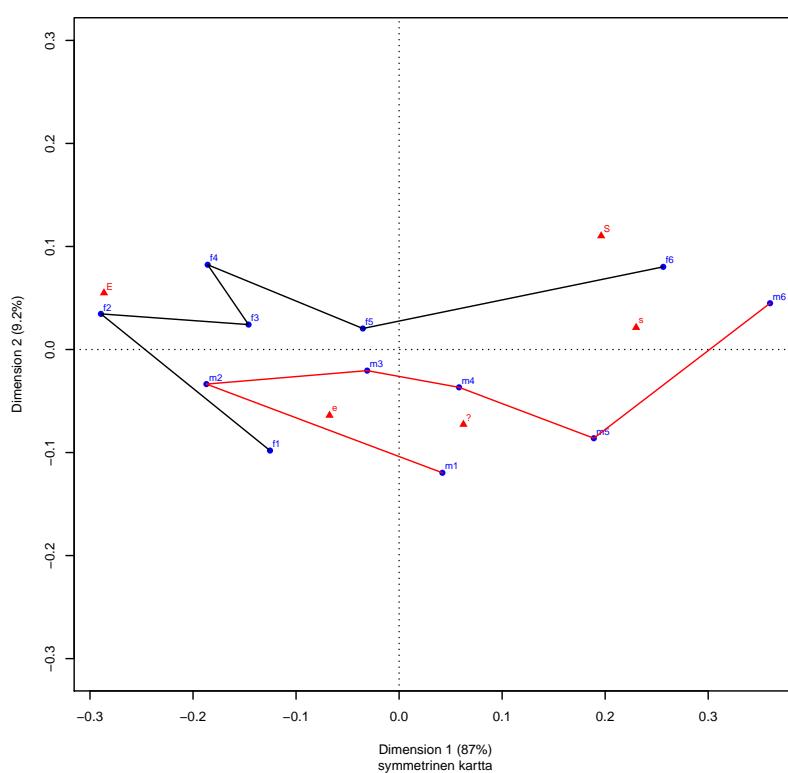
```
gaTestCA1 <- ca(~ga + Q1b,ISSP2012esim1b.dat)

# maapisteiden pääkoordinaatit janojen piirtämiseen

gaTestCA1.rpc <- gaTestCA1$rowcoord %*% diag(gaTestCA1$sv)

par(cex = 0.6)
plot(gaTestCA1, main = "Äiti töissä: ikäluokka ja sukupuoli",
      sub = "symmetrinen kartta")
# naiset
lines(gaTestCA1.rpc[1:6,1],gaTestCA1.rpc[1:6,2])
#miehet
lines(gaTestCA1.rpc[7:12,1],gaTestCA1.rpc[7:12,2], col = "red")
```

Äiti töissä: ikäluokka ja sukupuoli



Kuva 16: Iän ja sukupuolen yhdistetty muuttuja

```

#segments(gaTestCA1.rpc[1:6,1],gaTestCA1.rpc[1:6,2])
#
#      ,
#      gaTestCA1.rpc[4,1],gaTestCA1.rpc[4,2]
#      )
#segments(gaTestCA1.rpc[4,1],gaTestCA1.rpc[4,2],
#      gaTestCA1.rpc[3,1],gaTestCA1.rpc[3,2]
#      )

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  qlt  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 Aika yksiulotteinen (87 prosenttia ensimmäisellä dimensiolla!). Data on “as

it is”, ei ole vakioitu ga-luokkien kokoja (massat max(f4 101), min (m1 57)).

zxy miten pitäisi tulkita “oikealle kaatunut U - muoto” miehillä ja naisilla? Järjestys ei toimi, S s-sarakkeen vasemmalla puolella. Miehet konservatiivisempia, mutta maltillisempia? Nuorin ikäluokka on poikkeava. Epävarmoja tai maltillisesti e, sitten loikka vasemmalle ja sieltä konservatiiviseen suuntaa oikealle. Naisilla poikkeama f3 - f4. VAnhimmat ikäluokat tiukemmin konservatiivisia (f6, m6). Jos vertaa sukupuolten eroja samassa ikäluokassa, on aika samanlainen (miehet konservatiivisia, naiset liberaaleja). Naisista vain vanhin ikäluokka oikealla, miehistä nuorin ja kolme vanhinta.

zxy Tulkinnassa muistettava, että ikäluokat yli maiden. 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.

edit 14.8.20 Viite?

edit 14.8.20 Numeeriset tulokset: nuorimpien miesten (qlt 659) ja erityisesti keski-ikäisten miestén m3 (qlt 457) pistet huonosti esitetty kartalla. Tulkitaan myös cor ja ctr, riveille ja sarakkeille.

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

```
ISSP2012esim1b.dat <- mutate(ISSP2012esim1b.dat, maaga = paste(maa, ga, sep = ""))

# tarkistus, muunnos ok
# ISSP2012esim1b.dat %>% tableX(maa, maaga)
# head(ISSP2012esim2.dat)
# str(ISSP2012esim2.dat)
```

Maa - ikäluokka - sukupuoli - interaktiomuuttuja maaga

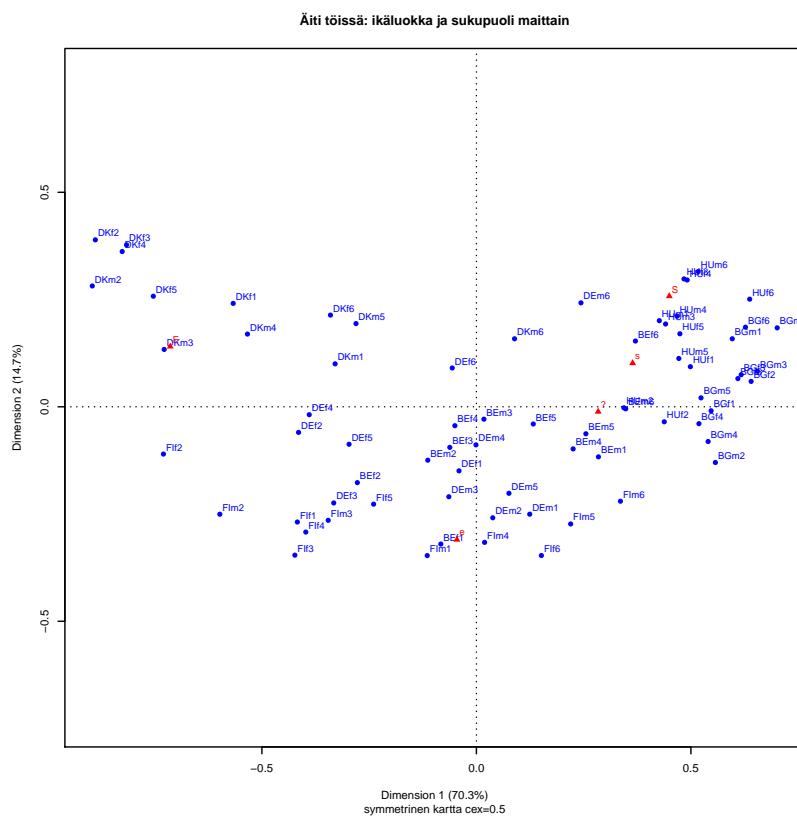
Tehty jo 26.9.2018!

```
maagaCA1 <- ca(~maaga + Q1b, ISSP2012esim1b.dat)
```

```
# maapisteiden pääkoordinaatit janojen piirtämiseen
# HUOM! maagaCA1.rpc on matriisi
```

```
maagaCA1.rpc <- maagaCA1$rowcoord %*% diag(maagaCA1$sv) #Missä käytetään? (10.10.20)

par(cex = 0.5)
plot(maagaCA1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
      sub = "symmetrinen kartta cex=0.5")
```



Kuva 17: Ikä, sukupuoli ja maa

```
#Kuvatiedoston koko säädettävä tarkemmin (30.3.20)

# pdf("img/maagaCA1_symm1.pdf")
# par(cex = 0.5)
# plot(maagaCA1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
#       sub = "symmetrinen kartta cex=0.5")
# dev.off()
```

Kartta on aika tukkoinen, yhdistetään muuttujat janoilla.

4.3 Kvaliteetti ja stabiilius

maaga-talukossa on paljon pieniä frekvenssejä (alle 5). Periaatteessa pienien frekvenssin rivejä tai sarakkeita voi yhdistää (distr. equivivalece), ja näin kannattaa tehdä jotta kartta ei mene tukkoon. **9.9.20** Pienet solufrekvenssin vs. pienet reunajakauman frekvenssin, miten edellisen kanssa pitäisi toimia?

Kartan herkyyttä joillekin pienien massan rivipisteille on tutkittu. Ei ilmeistä syytä huoleen, mutta (a) joidenkin pisteen huono kvaliteetti ja (b) pienet solufrekvenssit ovat huono juttu. Jälkimmäisen voisi korjata yhdistelemättä luokkia, hyöty olisi kuvan selkeytyminen ja haitta kiinnostavien piirteiden peittymine. Erityisesti nuorimman ja toiseksi nuorimman ikäluokan ero.

Vertailu voi tehdä

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

2. Maiden välillä

- a. miten ikä-sukupuoliluokat sijaitsevat suhteessa maiden keskiarvopisteisiin
- b. mikä on niiden järjestys

Ratkaisun numeerisia tuloksia voi katsoa, löytyykö profileja joilla on pieni massa mutta suuri vaikutus akseleihin.

```
# (24.2.20) Miten voisi kätevästi tarkistaa, että mikään pieni massa piste ei
# vaikuta (kontribuutiot) liikaa karttaan?
#str(maagaTestCA1)
```

ISSP2012esim1b.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

maaga/Q1b	S	s	?	e	E	Total
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
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

maaga/Q1b	S	s	?	e	E	Total
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
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

```

maagaCA1num <- summary(maagaCA1)
# maagaCA1num
# str(maagaCA1num) numeriset tulokset tibbleksi - rivit
maagaCAnum2 <- as_tibble(maagaCA1num$rows, .name_repair = c("unique"))

## New names:
## * cor -> cor...6
## * ctr -> ctr...7
## * cor -> cor...9
## * ctr -> ctr...10

# maagaCAnum2
# str(maagaCAnum2)
summary(maagaCAnum2)

```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEf1	Min.	Min.	Min.	Min. :-	Min.	Min.	Min. :-	Min.	Min.
: 1	: 5.00	:108.0	: 1.00	895.00	: 0.0	: 0.00	347.000	: 0.00	: 0.00

	name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEf2	1st	1st	1st	1st	1st	1st	1st	1st	1st	1st
: 1	Qu.:10.00	Qu.:704.0	Qu.:6.75	Qu.:330.00	Qu.:351.0	Qu.:4.50	Qu.:156.000	Qu.:42.75	Qu.:2.75	
BEf3	Median	Median	Median	Median	Median	Median	Median	Median	Median	Median
: 1	:13.00	:838.0	:11.00	:	:667.5	:11.00	: -	:141.50	:10.00	
					82.50			13.500		
BEf4	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean	Mean
: 1	:13.97	:772.6	:13.88	:	:573.3	:13.92	:	:199.28	:13.86	
					46.49			1.653		
BEf5	3rd	3rd	3rd	3rd	3rd	3rd	3rd	3rd	3rd	3rd
: 1	Qu.:17.00	Qu.:953.0	Qu.:15.00	Qu.:472.50	Qu.:830.0	Qu.:17.00	Qu.:265.0	Qu.:21.25		
BEf6	Max.	Max.	Max.	Max.	Max.	Max.	Max.	Max.	Max.	Max.
: 1	:26.00	:999.0	:57.00	701.00	:982.0	:66.00	389.000	:834.00	:61.00	
(Other)	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

```
colnames(maagaCAnum2) # välilyötejä nimen alussa
```

```
## [1] "name"      "mass"       " qlt"       " inr"       " k=1"       "cor...6"
## [7] "ctr...7"    " k=2"       "cor...9"     "ctr...10"
names(maagaCAnum2)[3] <- "qlt"
# maagaCAnum2 %>% rename( qlt, qlt)
```

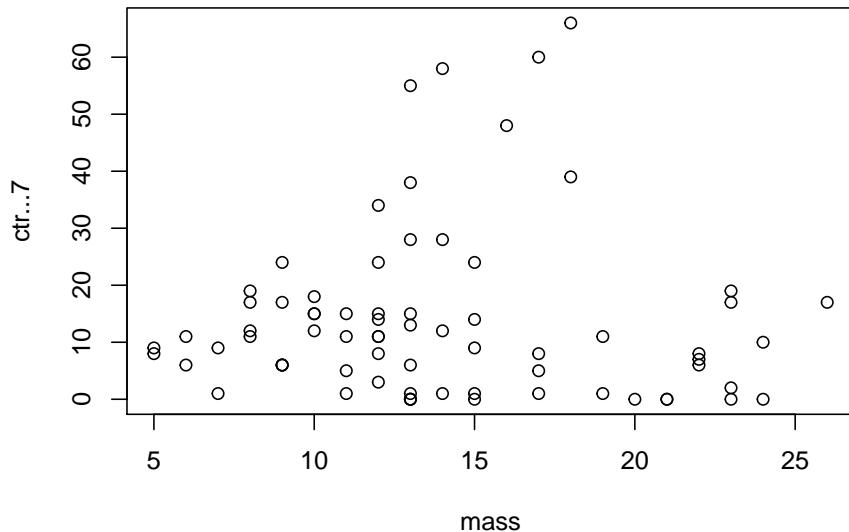
```
arrange(maagaCAnum2 , mass)
```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BGf1	5	531	11	547	531	8	-9	0	0
BGm1	5	940	7	596	878	9	159	62	3
BGm2	6	830	9	557	788	11	-130	43	3
HUm1	6	935	5	426	766	6	201	170	6
FIm1	7	787	5	-115	78	1	-347	710	22
HUf1	7	723	9	499	698	9	93	25	1
BGf2	8	860	14	640	853	17	59	7	1
BGm3	8	709	19	655	698	19	83	11	1
BGm4	8	771	11	540	754	12	-81	17	1
HUm6	8	726	15	517	529	11	315	197	20
BGm6	9	692	27	701	647	24	184	45	8
FIm2	9	977	14	-598	832	17	-250	146	14
FIm3	9	998	6	-345	629	6	-265	369	16
FIm6	9	911	6	336	637	6	-220	274	12
HUm2	9	381	11	344	381	6	-2	0	0

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BGf4	10	932	12	519	927	15	-39	5	0
BGm5	10	979	11	524	977	15	21	2	0
DKf1	10	991	15	-567	839	18	241	152	15
HUm4	10	999	10	468	830	12	211	169	11
BEm1	11	429	9	284	367	5	-117	62	4
FIf6	11	835	7	151	134	1	-347	701	35
HUf2	11	689	11	438	685	11	-35	4	0
HUf4	11	768	18	491	564	15	296	204	25
BGf3	12	815	21	617	804	24	75	12	2
DKf6	12	808	9	-340	579	8	214	229	14
FIf1	12	980	11	-417	693	11	-269	287	21
FIf2	12	927	26	-730	907	34	-110	21	4
FIf3	12	984	13	-423	590	11	-346	394	36
FIm5	12	734	7	220	289	3	-273	446	23
HUf3	12	808	18	484	586	15	298	222	27
HUf5	12	850	13	474	753	14	170	97	9
DEF1	13	425	3	-41	29	0	-149	395	7
DEM1	13	912	4	124	180	1	-250	732	20
DEM2	13	766	4	38	16	0	-259	749	22
DKm2	13	989	43	-895	900	55	282	89	26
DKm3	13	982	28	-728	950	38	134	32	6
DKm5	13	643	9	-281	435	6	194	208	13
FIm4	13	837	6	19	3	0	-316	834	33
HUf6	13	671	34	637	581	28	251	90	21
HUm3	13	957	12	441	803	13	193	154	12
HUm5	13	942	12	472	891	15	113	51	4
BEf1	14	678	9	-83	43	1	-320	635	38
BGf5	14	880	23	609	870	28	66	10	2
DKf2	14	991	49	-888	831	58	389	160	53
FIf4	14	991	14	-398	644	12	-292	347	32
DEF2	15	938	10	-415	919	14	-60	19	1
DEM3	15	737	4	-64	63	0	-210	674	17
DKm1	15	981	7	-329	898	9	100	83	4
DKm4	15	941	19	-534	855	24	170	86	11
DKm6	15	355	5	89	85	1	158	270	9
DKf5	16	998	38	-753	894	48	258	105	27
BEm2	17	372	5	-113	169	1	-125	203	7
DEF5	17	839	7	-297	772	8	-87	67	3
DKf3	17	963	53	-816	793	60	377	170	61
FIf5	17	952	8	-240	502	5	-227	450	23
BGf6	18	921	32	627	846	39	186	74	16
DKf4	18	977	57	-826	820	66	362	157	61
DEF3	19	846	13	-333	582	11	-224	264	24
DEM5	19	603	5	76	75	1	-202	529	20

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEm3	20	108	1	17	29	0	-29	79	0
BEf3	21	320	3	-62	96	0	-95	224	5
DEM4	21	137	5	-1	0	0	-89	137	4
BEm4	22	966	5	225	812	6	-98	154	5
BEm5	22	728	8	255	686	8	-63	42	2
DEM6	22	849	12	244	427	7	242	422	34
BEf5	23	332	5	133	304	2	-40	28	1
BEf6	23	832	17	371	710	17	153	121	14
DEF4	23	985	13	-390	982	19	-18	2	0
DEF6	23	116	8	-56	32	0	90	84	5
BEf2	24	914	11	-278	650	10	-177	264	20
BEf4	24	164	3	-50	92	0	-44	71	1
BEm6	26	788	15	348	788	17	-5	0	0

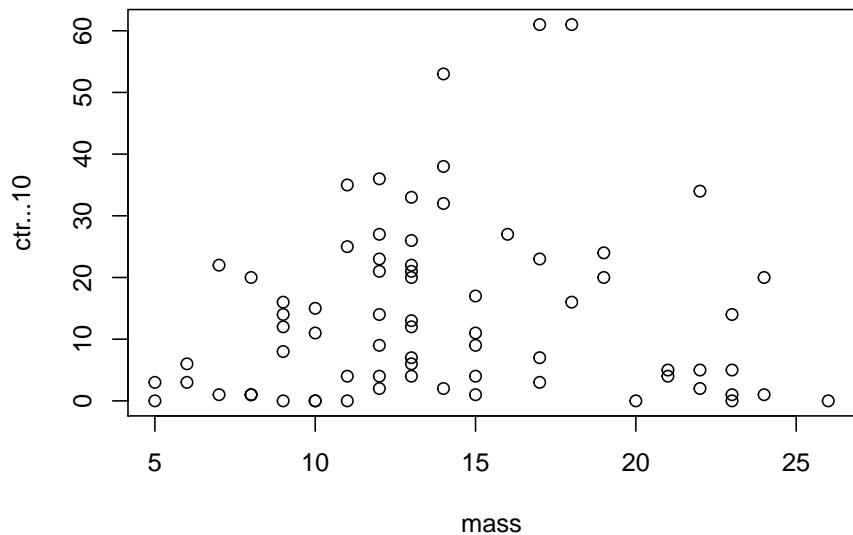
```
# maagaCAnum2
# plot(maagaCAnum2, x = c("mass"), y = c("ctr...7"), xlim = c(0,30), ylim = c(0, 1000))
with(maagaCAnum2, plot(mass, ctr...7))
```



```
tail(arrange(maagaCAnum2 ,ctr...7))
```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BGf6	18	921	32	627	846	39	186	74	16
DKf5	16	998	38	-753	894	48	258	105	27
DKm2	13	989	43	-895	900	55	282	89	26
DKf2	14	991	49	-888	831	58	389	160	53
DKf3	17	963	53	-816	793	60	377	170	61
DKf4	18	977	57	-826	820	66	362	157	61

```
with(maagaCAnum2, plot(mass, ctr...10))
```



```
tail(arrange(maagaCAnum2 ,ctr...10))
```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
FIf6	11	835	7	151	134	1	-347	701	35
FIf3	12	984	13	-423	590	11	-346	394	36
BEf1	14	678	9	-83	43	1	-320	635	38
DKf2	14	991	49	-888	831	58	389	160	53
DKf3	17	963	53	-816	793	60	377	170	61
DKf4	18	977	57	-826	820	66	362	157	61

```

str(maagaCAnum2)

## # tibble [72 x 10] (S3: tbl_df/tbl/data.frame)
## $ name     : Factor w/ 72 levels "BEf1","BEf2",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ mass      : num [1:72] 14 24 21 24 23 23 11 17 20 22 ...
## $ qlt       : num [1:72] 678 914 320 164 332 832 429 372 108 966 ...
## $ inr       : num [1:72] 9 11 3 3 5 17 9 5 1 5 ...
## $ k=1       : num [1:72] -83 -278 -62 -50 133 371 284 -113 17 225 ...
## $ cor...6   : num [1:72] 43 650 96 92 304 710 367 169 29 812 ...
## $ ctr...7   : num [1:72] 1 10 0 0 2 17 5 1 0 6 ...
## $ k=2       : num [1:72] -320 -177 -95 -44 -40 153 -117 -125 -29 -98 ...
## $ cor...9   : num [1:72] 635 264 224 71 28 121 62 203 79 154 ...
## $ ctr...10  : num [1:72] 38 20 5 1 1 14 4 7 0 5 ...

arrange(maagaCAnum2, qlt)

```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEm3	20	108	1	17	29	0	-29	79	0
DEF6	23	116	8	-56	32	0	90	84	5
DEM4	21	137	5	-1	0	0	-89	137	4
BEf4	24	164	3	-50	92	0	-44	71	1
BEf3	21	320	3	-62	96	0	-95	224	5
BEf5	23	332	5	133	304	2	-40	28	1
DKm6	15	355	5	89	85	1	158	270	9
BEm2	17	372	5	-113	169	1	-125	203	7
HUm2	9	381	11	344	381	6	-2	0	0
DEF1	13	425	3	-41	29	0	-149	395	7
BEm1	11	429	9	284	367	5	-117	62	4
BGf1	5	531	11	547	531	8	-9	0	0
DEM5	19	603	5	76	75	1	-202	529	20
DKm5	13	643	9	-281	435	6	194	208	13
HUf6	13	671	34	637	581	28	251	90	21
BEf1	14	678	9	-83	43	1	-320	635	38
HUf2	11	689	11	438	685	11	-35	4	0
BGm6	9	692	27	701	647	24	184	45	8
BGm3	8	709	19	655	698	19	83	11	1
HUf1	7	723	9	499	698	9	93	25	1
HUm6	8	726	15	517	529	11	315	197	20
BEm5	22	728	8	255	686	8	-63	42	2
FIm5	12	734	7	220	289	3	-273	446	23
DEM3	15	737	4	-64	63	0	-210	674	17
DEM2	13	766	4	38	16	0	-259	749	22
HUf4	11	768	18	491	564	15	296	204	25
BGm4	8	771	11	540	754	12	-81	17	1
FIm1	7	787	5	-115	78	1	-347	710	22

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEm6	26	788	15	348	788	17	-5	0	0
DKf6	12	808	9	-340	579	8	214	229	14
HUf3	12	808	18	484	586	15	298	222	27
BGf3	12	815	21	617	804	24	75	12	2
BGm2	6	830	9	557	788	11	-130	43	3
BEf6	23	832	17	371	710	17	153	121	14
FIf6	11	835	7	151	134	1	-347	701	35
FIm4	13	837	6	19	3	0	-316	834	33
DEF5	17	839	7	-297	772	8	-87	67	3
DEF3	19	846	13	-333	582	11	-224	264	24
DEM6	22	849	12	244	427	7	242	422	34
HUf5	12	850	13	474	753	14	170	97	9
BGf2	8	860	14	640	853	17	59	7	1
BGf5	14	880	23	609	870	28	66	10	2
FIm6	9	911	6	336	637	6	-220	274	12
DEM1	13	912	4	124	180	1	-250	732	20
BEf2	24	914	11	-278	650	10	-177	264	20
BGf6	18	921	32	627	846	39	186	74	16
FIf2	12	927	26	-730	907	34	-110	21	4
BGf4	10	932	12	519	927	15	-39	5	0
HUm1	6	935	5	426	766	6	201	170	6
DEF2	15	938	10	-415	919	14	-60	19	1
BGm1	5	940	7	596	878	9	159	62	3
DKm4	15	941	19	-534	855	24	170	86	11
HUm5	13	942	12	472	891	15	113	51	4
FIf5	17	952	8	-240	502	5	-227	450	23
HUm3	13	957	12	441	803	13	193	154	12
DKf3	17	963	53	-816	793	60	377	170	61
BEm4	22	966	5	225	812	6	-98	154	5
DKf4	18	977	57	-826	820	66	362	157	61
FIm2	9	977	14	-598	832	17	-250	146	14
BGm5	10	979	11	524	977	15	21	2	0
FIf1	12	980	11	-417	693	11	-269	287	21
DKm1	15	981	7	-329	898	9	100	83	4
DKm3	13	982	28	-728	950	38	134	32	6
FIf3	12	984	13	-423	590	11	-346	394	36
DEF4	23	985	13	-390	982	19	-18	2	0
DKm2	13	989	43	-895	900	55	282	89	26
DKf1	10	991	15	-567	839	18	241	152	15
DKf2	14	991	49	-888	831	58	389	160	53
FIf4	14	991	14	-398	644	12	-292	347	32
DKf5	16	998	38	-753	894	48	258	105	27
FIm3	9	998	6	-345	629	6	-265	369	16

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
HUm4	10	999	10	468	830	12	211	169	11

```
head(arrange(maagaCA1num2, qlt))
```

name	mass	qlt	inr	k=1	cor...6	ctr...7	k=2	cor...9	ctr...10
BEm3	20	108	1	17	29	0	-29	79	0
DEF6	23	116	8	-56	32	0	90	84	5
DEM4	21	137	5	-1	0	0	-89	137	4
BEF4	24	164	3	-50	92	0	-44	71	1
BEF3	21	320	3	-62	96	0	-95	224	5
BEF5	23	332	5	133	304	2	-40	28	1

```
# Hieman hankalaan kätevästi järjestää numeerisia tuloksia massan mukaan

#str(maagaCA1num)
#maagaCA1num$rows
#maagaRows.df <- maagaCA1num$rows
# sarakenimet eivät yksikäsitteisiä
#maagaRows.df
#str(maagaRows.df)
#names(maagaRows.df)
#str(maagaRows.df$mass)
# ei toimi AscmaagaRows.df <- maagaRows.df[order(mass),]
```

Massa ja kontribuutiot akselleille 1 ja 2: epäilyttäviä havaintoja joilla pieni massa ja suuri kontribuutio ei näytä olevan.

Huonosti esitetettyjä pisteitä on erityisesti Belgiasta, myös Saksan (DEF6,DEF1 ja DEM4), Tanskan vanhat miehet (DKm6) ja Unkarin nuorehkot miehet (HUm2) kuuluvat tähän joukkoon.

Maapisteet täydentäviksi pisteiksi - tarkistuksia.

```
# Miten maa-rivit täydentäviksi riveiksi - alla siisti ratkaisu
# Miten labelit hieman lähemmäkis 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 keskiarvona

# HUOM! maagaTab1 integer matriisi, dimnames-attribuutilla kaksi arvoa
#ikäluokka - sukupuoli ja maa - maaga-muuttuja
maagaTab1 <- table(ISSP2012esim1b.dat$maaga, ISSP2012esim1b.dat$Q1b)
```

```

#dim(testTab1) #72 riviä, 5 saraketta

# maa-rivit
maagaTab_sr <- table(ISSP2012esim1b.dat$maa, ISSP2012esim1b.dat$Q1b)
#maagaTab_sr

maagaTab1 <- rbind(maagaTab1,maagaTab_sr)
# str(maagaTab1)
# maagaTab1
# dim(maagaTab1) #78 riviä, 5 saraketta, 1-72 data ja 73-78 täydentävät rivit

spCAmaaga1 <- ca(maagaTab1[,1:5], suprow = 73:78)
# X11()

# Plot toimii (4.2.20), ja par() (4.5.20)
par(cex = 0.5)
plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
      sub = "symmetrinen kartta, maat täydentävinä pisteinä, cex=0.5"
            )

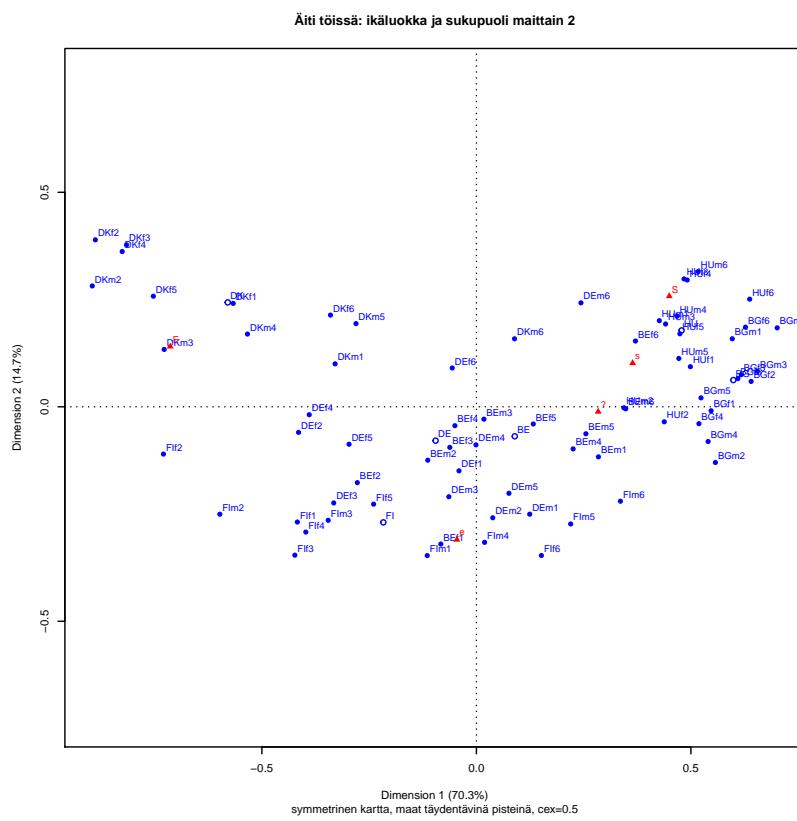
#asymmetrinen kartta
#X11()
par(cex = 0.5)
plot(spCAmaaga1, map = "rowgreen",
      contrib = c("absolute", "absolute"),
      mass = c(TRUE,TRUE),
      arrows = c(FALSE,TRUE),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 3",
      sub ="absoluuttiset kontribuutiot ('rowgreen'),cex=0.5",
            )

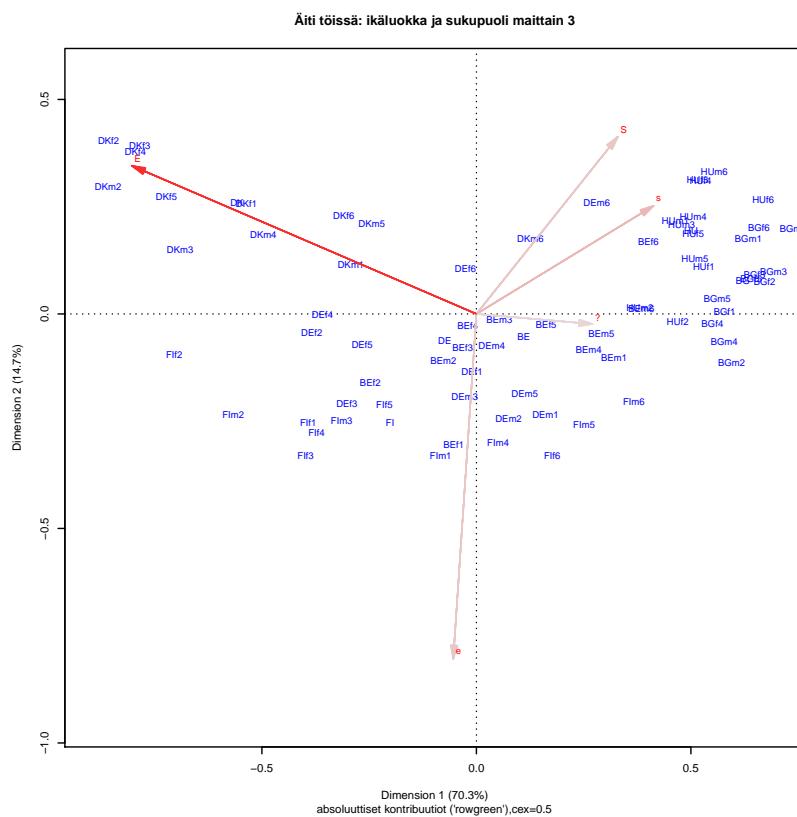
#asymmetrinen kartta (14.8.20)

par(cex = 0.5)
plot(spCAmaaga1, map = "rowgreen",
      contrib = c("absolute", "absolute"),
      mass = c(TRUE,TRUE),
      arrows = c(FALSE,TRUE),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 3",
      sub = "suhteelliset kontribuutiot ('rowgreen') ,cex=0.5"
            )

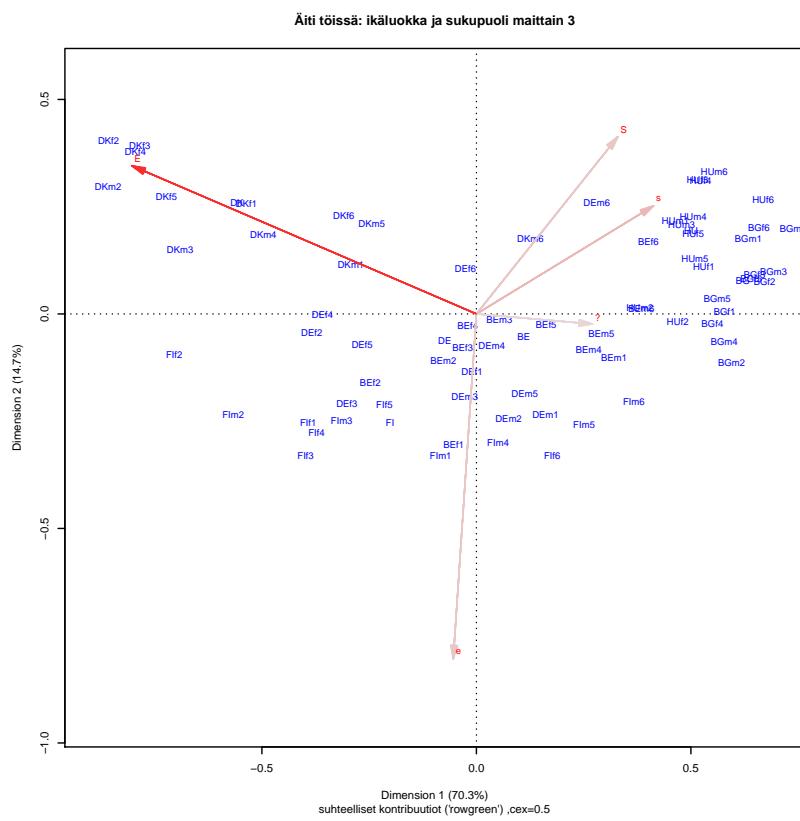
```

Asymmetrinen kontribuutiokartta on hyödyllinen sarakkeiden ja niiden avulla akseleiden tulkinnassa. Rivipisteet pakkautuvat kuitenkin tiiviimmin kartan keskelle kuin symmetrisessä kuvassa.





Kuva 19: Ikä-sukupuoli-maa



Kuva 20: Ikä-sukupuoli-maa

```

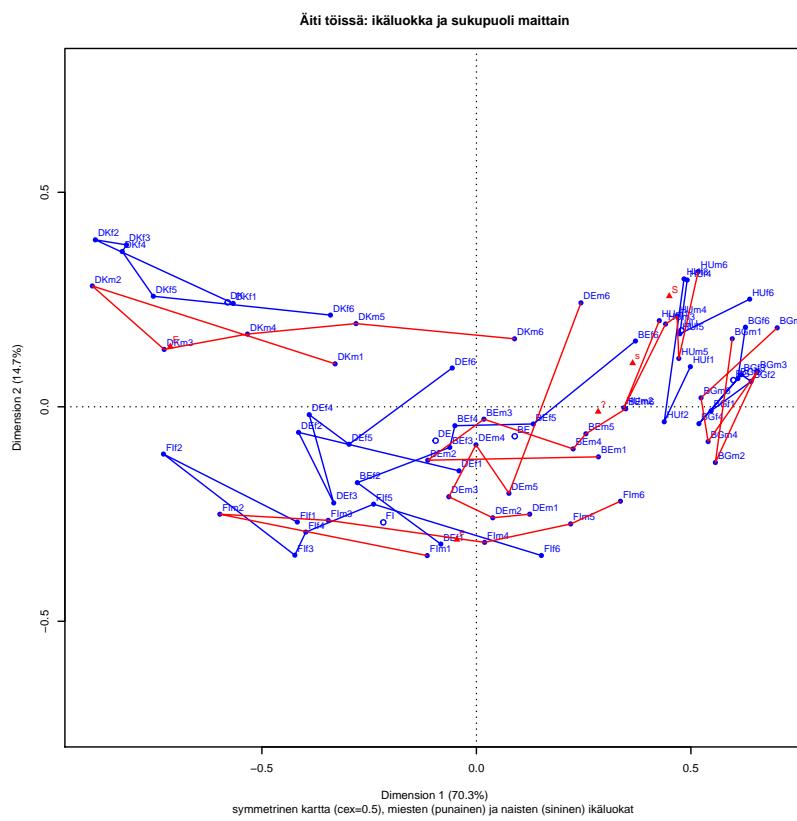
# pisteiden koordinaatit (symmetrinen kartta, päärakennus)
maagaLines1 <- cacoord(spCAmaaga1, type = "symmetric")

# vain kaksi ensimmäistä dimensioita
maagaLines1 <- maagaLines1$rows[, 1:2]

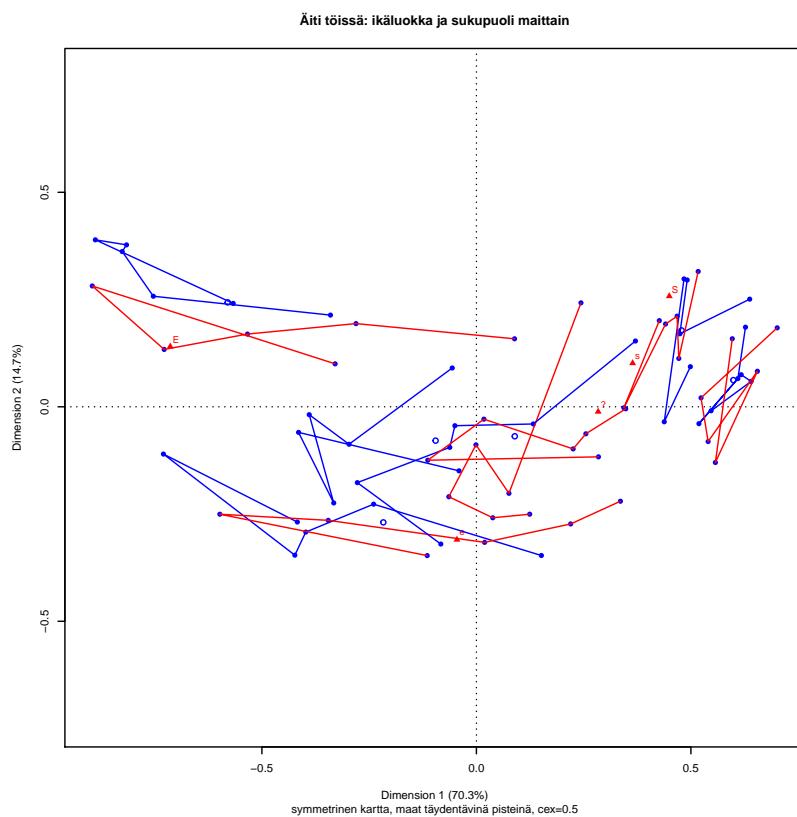
# Tarkistuksia
#maagaLines1
#str(maagaLines1)
#class(maagaLines1)
# Onko yhtään vähemmän tukkoinen? Eipä juuri (17.9.20)
par(cex = 0.5)
plot(spCAmaaga1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
     sub = "symmetrinen kartta (cex=0.5), miesten (punainen) ja naisten (sininen")
)
lines(maagaLines1[49:54,1],maagaLines1[49:54,2], col="blue") #FIf
lines(maagaLines1[55:60,1],maagaLines1[55:60,2], col="red") #FIm
lines(maagaLines1[25:30,1],maagaLines1[25:30,2], col="blue") #DEF
lines(maagaLines1[31:36,1],maagaLines1[31:36,2], col="red") #DEM
lines(maagaLines1[37:42,1],maagaLines1[37:42,2], col="blue") #DKf
lines(maagaLines1[43:48,1],maagaLines1[43:48,2], col="red") #DKm
lines(maagaLines1[1:6,1],maagaLines1[1:6,2], col="blue") #BEf
lines(maagaLines1[7:12,1],maagaLines1[7:12,2], col="red") #BEm
lines(maagaLines1[13:18,1],maagaLines1[13:18,2], col="blue") #BGf
lines(maagaLines1[19:24,1],maagaLines1[19:24,2], col="red") #BGm
lines(maagaLines1[61:66,1],maagaLines1[61:66,2], col="blue") #HUF
lines(maagaLines1[67:72,1],maagaLines1[67:72,2], col="red") #HUM

# Onko yhtään vähemmän tukkoinen? Eipä juuri (17.9.20)
par(cex = 0.5)
plot(spCAmaaga1, labels = c(0,2) , main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
     sub = "symmetrinen kartta, maat täydentävinä pisteinä, cex=0.5")
)
lines(maagaLines1[49:54,1],maagaLines1[49:54,2], col="blue") #FIf
lines(maagaLines1[55:60,1],maagaLines1[55:60,2], col="red") #FIm
lines(maagaLines1[25:30,1],maagaLines1[25:30,2], col="blue") #DEF
lines(maagaLines1[31:36,1],maagaLines1[31:36,2], col="red") #DEM
lines(maagaLines1[37:42,1],maagaLines1[37:42,2], col="blue") #DKf
lines(maagaLines1[43:48,1],maagaLines1[43:48,2], col="red") #DKm
lines(maagaLines1[1:6,1],maagaLines1[1:6,2], col="blue") #BEf
lines(maagaLines1[7:12,1],maagaLines1[7:12,2], col="red") #BEm
lines(maagaLines1[13:18,1],maagaLines1[13:18,2], col="blue") #BGf
lines(maagaLines1[19:24,1],maagaLines1[19:24,2], col="red") #BGm
lines(maagaLines1[61:66,1],maagaLines1[61:66,2], col="blue") #HUF
lines(maagaLines1[67:72,1],maagaLines1[67:72,2], col="red") #HUM

```



Kuva 21: Ikä-sukupuoli-maa



Kuva 22: Ikä-sukupuoli-maa

Rivipisteiden tunnisteiden (label) poistaminen tuo pisteparvien ominaisuuden hieman selvemmin esiin. Suomi alimpana ja Tanska ylimpänä vasemmalla ovat melko samanlaisia. Vaihtelu on lähes täysin ensimäisen dimension suuntaan, Molemmilla mailla miesten pisteet ovat oikealla ja alempaan. Keskellä Saksan ja Belgian pistejoukoissa vaihtelu toisen dimension suunnassa on jo suurempaa ja miesten pisteet ovat selvästi oikealla verrattuna naisten pisteisiin. Bulgarian ja Unkarin pistet sijaitsevat kauimpana oikealla, ja vaihtelu on suurimmalta osin toisen dimensio suuntaan, ja pisteet ovat tiukemmin lähellä toisiaan. Kaikissa maissa vanhemmat ikäluokat ovat nuoria konservatiivimempia, ja kaikissa nuorin ikäluokka on melko konservatiivinen. Keskimäiset ikäluokat ovat liberaaleimpia. Unkari ja Bulgaria poikkeavat muista selvästi.

4.4 Kartan rajaaminen

0. Analyysin voi tehdä vain osalle dataa - vähän kehno ratkaisu, ei yhteyttää koko aineiston antamaan yleiskuvaan muuttujien yhteyksistä.
1. Kartasta voi suurentaa osan käsitönä (“leikkaa ja liimaa”). Toimii, yksinkertainen ja ehkä siksi paras? Mutta ei voi koodata! Sopii eksploratiiviseen vaiheeseen, pdf-kuviin voi liittää kommentteja ja tekstiä jne.
2. BaseR plot-funktiolla voi rakentaa kartan pala kerrallaan ja rajata kuvan johonkin kartan osaan. Työlästä, hankalaa? Yllättäen ei, kun asettaa koodilohkon option Rmarkdowissa oikein! (9.9.20). Jos käytetään yhtä ca-funktion tulosobjektiä, pelkkää plot-komentoja.
3. Osajoukon korrespondenssianalyysi (subset ca)

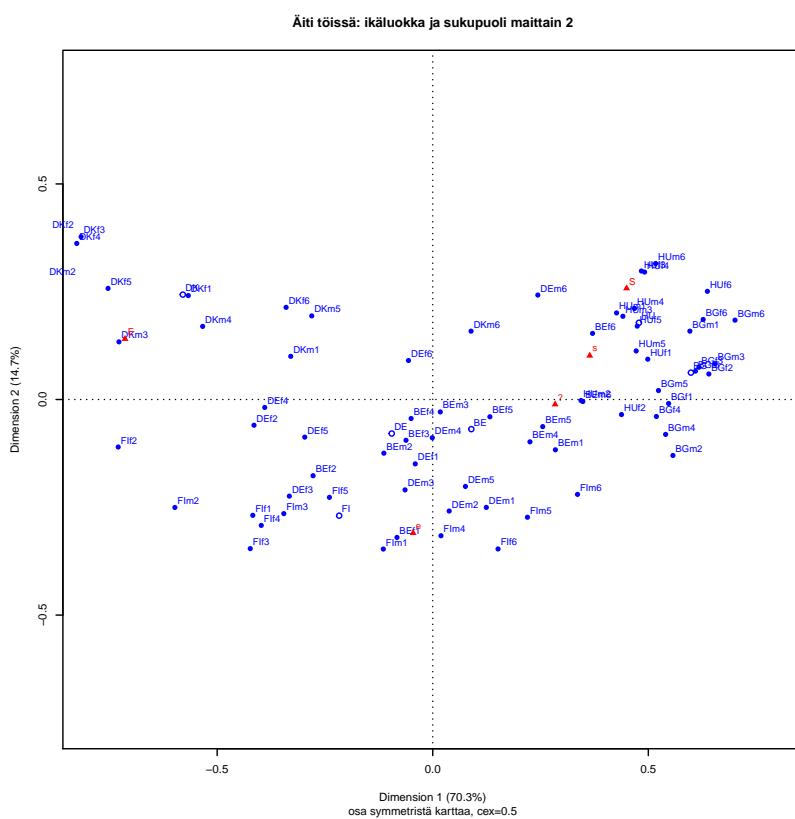
*zxy** Tässä vain lyhyt kuvaus subset ca:n perusideasta

4.4.1 Kuva-alueen rajaaminen - ei oikein toimi (16.10.20)

```
# Zoomaus - esimerkki (24.2.20) xlim=c(-0.5,0.5), ylim=c(-0.6,0.4)
# EI TOIMII - PISTEET piirrettävä, ei voi käyttää ca.plot - funktiota? (4.5.20).
# MG2017 laskarit - day3: mca:n tuloslistaa käytetty luottamusellipseissä ja toimii?
# Piirtää pisteytä koko "plot-alueelle"?
#X11()

# Rivi- ja sarakekoordinaatit (principal coordinates) talteen
maagaCA1.rpc <- spCAMAAGA1$rowcoord %*% diag(simpleCA1$sv)
maagaCA1.cpc <- spCAMAAGA1$colcoord %*% diag(simpleCA1$sv)
par(cex = 0.5)

plot(spCAMAAGA1, xlim = c(-0.75,0.75), ylim = c(-0.75,0.75),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
      sub = "osa symmetristä karttaa, cex=0.5"
```



Kuva 23: Ikä-sukupuoli-maa

```
# TÄTÄ KEHITELLÄÄ CAcalc_1.R - skriptissä (15.6.20)
# ei toimi ihan toivotulla tavalla - tarkoitettu komentoriviltä
# grafiikkaikkunaan tulostukseen ?
# Vai pitääkö ensin piirtää kuvan "kehys" ilman pisteitä xlim- ja ylim- parametreilla
# ja sitten vasta pisteet?
# Kuvasuhde oikein, kun xlim = ylim, miten turhat pisteet pois? (29.5.20). Menevät
# näköjään kuva-alueen ja ulomman marginaalin väliin.

# Voisiko valita objektista vain osan pisteistä? Tai plot ilman tulostusta, tulos
# objektiin ja sieltä pisteet kuvaan? (29.5.20)
# str(spCAmaaqaa)
```

Ei toimi ihan odotetulla tavalla, mutta kuvasuhde näyttäisi olevan molemmissa oikea. Hieman hankala menetelmä.

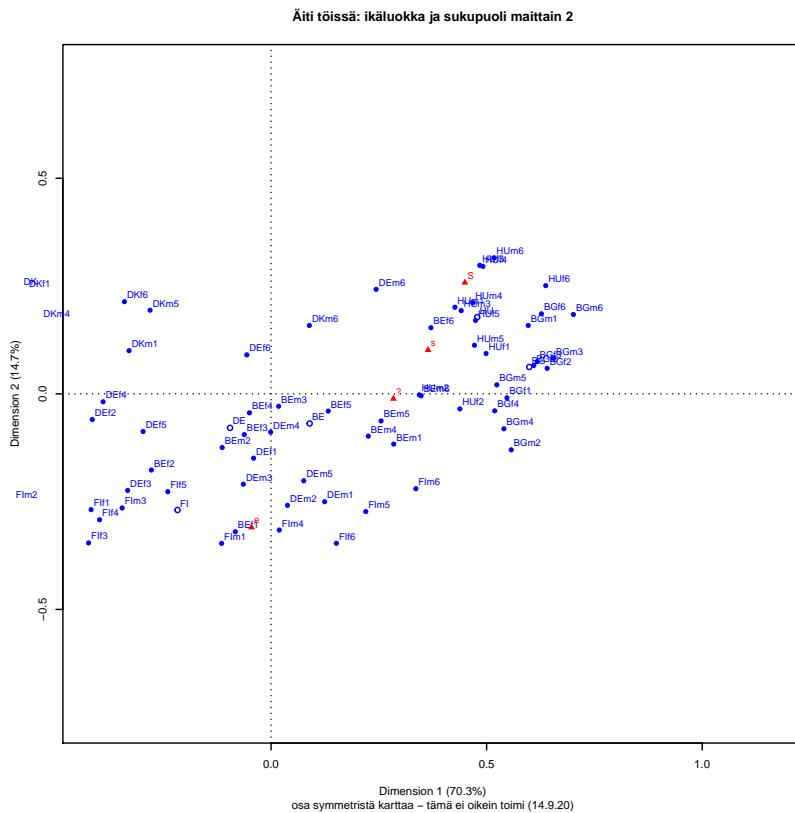
```

# Zoomaus - esimerkki (24.2.20) xlim=c(-0.5,0.5), ylim=c(-0.6,0.4)
# EI TOIMI - PISTEET piirrettävä, ei voi käyttää ca.plot - funktiota? (4.5.20).
# MG2017 laskarit - day3: mca:n tuloslistaa käytetty luottamusellipseissä ja toimii?
# Piirtää pisteitä koko "plot-alueelle"?


# Rivi- ja sarakekoordinaatit (principal coordinates) talteen
maagaCA1.rpc <- spCAmaaga1$rowcoord %*% diag(simpleCA1$sv)
maagaCA1.cpc <- spCAmaaga1$colcoord %*% diag(simpleCA1$sv)
par(cex = 0.5)

plot(spCAmaaga1, xlim = c(0,0.75), ylim = c(-0.75,0.75),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
      sub = "osa symmetristä karttaa - tämä ei oikein toimi (14.9.20)"
      )

```



Kuva 24: Ikä-sukupuoli-maa

```

# TÄTÄ KEHITELLÄÄ CAcalc_1.R - skriptissä (15.6.20)
# ei toimi ihan toivotulla tavalla - tarkoitettu komentoriviltä
# grafiikkaikkunaan tulostukseen ?
# Vai pitääkö ensin piirtää kuvan "kehys" ilman pisteitä xlim- ja ylim- parametreilla
# ja sitten vasta pisteet?
# Kuvasuhde oikein, kun xlim = ylim, miten turhat pisteet pois? (29.5.20). Menevät
# näköjään kuva-alueen ja ulomman marginaalin väliin.

# Voisiko valita objektista vain osan pisteistä? Tai plot ilman tulostusta, tulos
# objektiin ja sieltä pisteet kuvaan? (29.5.20)
# str(spCAMAAGA1)

Rivipisteiden inertia on suurempi kuin sarakkeiden. Asymmetrisen kartta näyttää
sarakeet selvemmin.

# Zoomaus - esimerkki (24.2.20) xlim=c(-0.5,0.5), ylim=c(-0.6,0.4)
# EI TOIMI - PISTEET piirrettävä, ei voi käyttää ca.plot - funktiota? (4.5.20).
# MG2017 laskarit - day3: mca:n tuloslistaa käytetty luottamusellipseissä ja toimii?
# Piirtää pisteitä koko "plot-alueelle"?

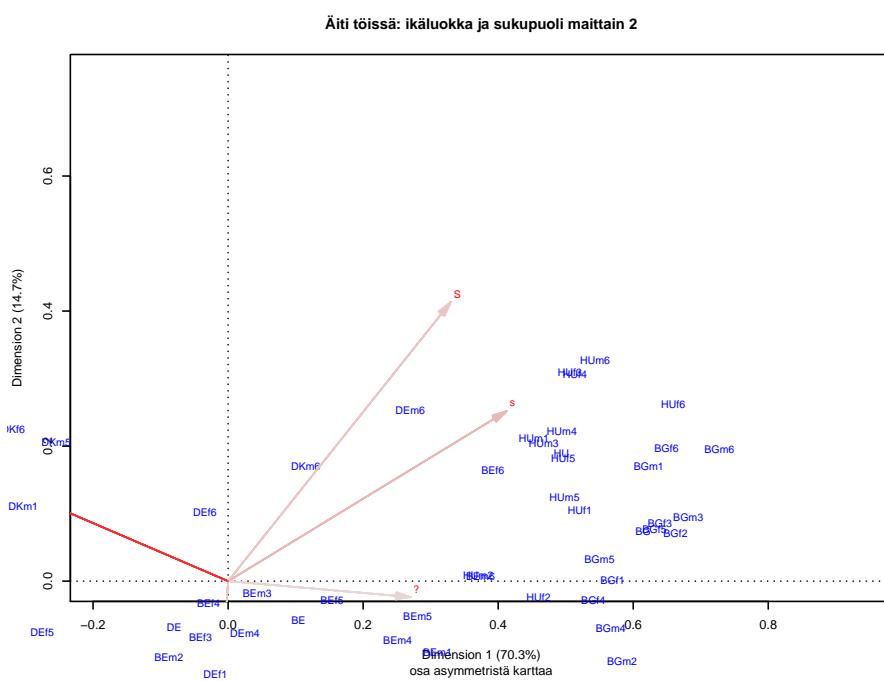
# Rivi- ja sarakekoordinaatit (principal coordinates) talteen
# maagaCA1.rpc <- spCAMAAGA1$rowcoord %*% diag(simpleCA1$sv)
# maagaCA1.cpc <- spCAMAAGA1$colcoord %*% diag(simpleCA1$sv)

par(cex = 0.5)
plot(spCAMAAGA1, map = "rowgreen",
      contrib = c("absolute", "absolute"),
      mass = c(TRUE, TRUE),
      arrows = c(FALSE, TRUE),
      xlim = c(0,0.75), ylim = c(0,0.75),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
      sub = "osa asymmetristä karttaa"
    )

# TÄTÄ KEHITELLÄÄ CAcalc_1.R - skriptissä (15.6.20)
# ei toimi ihan toivotulla tavalla - tarkoitettu komentoriviltä
# grafiikkaikkunaan tulostukseen ?
# Vai pitääkö ensin piirtää kuvan "kehys" ilman pisteitä xlim- ja ylim- parametreilla
# ja sitten vasta pisteet?
# Kuvasuhde oikein, kun xlim = ylim, miten turhat pisteet pois? (29.5.20). Menevät
# näköjään kuva-alueen ja ulomman marginaalin väliin.

# Voisiko valita objektista vain osan pisteistä? Tai plot ilman tulostusta, tulos
# objektiin ja sieltä pisteet kuvaan? (29.5.20)
# str(spCAMAAGA1)

```



Kuva 25: Ikä-sukupuoli-maa

Edellinen kartta on vähän epäilyttäävä, asettuvatkohan skaalatut sarakevektorin oikein? (9.9.20) ***

4.5 Subset CA

Teoria esitetään myöhemmin, käytännön hyödyllisyys osoitetaan tässä. CA-kartoissa on usein aivan liian paljon pisteitä, vain karkeat yleispiirteet näkyvät. Ovatko kiinnostavat asia piilossa? "Poikkeavat havainnot ovat ainoina todella kiinnostavia havaintoja".

Data Ensimmäinen osajoukko-ca käyttää datan perustiedostoa (ISSP2012esim2.dat). Maa-rivit ovat mukana kokonaislukumatriisissa maagaTab1.

Dataassa maa-sukupuoli-ikäluokkarivit ovat näillä riveillä: BE 1-12, BG 13-24, DE 25-36, DK 37-48, FI 49-60, HU 61-72 . Maa-profilit ovat taulukon(matriisin) maagaTab1 riveillä 73-78 samassa järjestyksessä. **edit 9.9.20** Tarkistettava, miten ca-funktio laskee täydentävien rivien koordinaatit, kun riveistä otetaan analyysiin vain osa.

```
#31.8.20 Testataan subset ca
#ISSP2012esim2.dat %>% tableX(maaga, Q1b)
maagaCA2subset1BEBG <- ca(~maaga + Q1b, ISSP2012esim1b.dat, subsetrow = 1:24)
par(cex = 0.5)
plot(maagaCA2subset1BEBG,
      sub = "symmetrinen kartta:Belgia ja Bulgaria,osajoukon ca (subset ca)"
    )

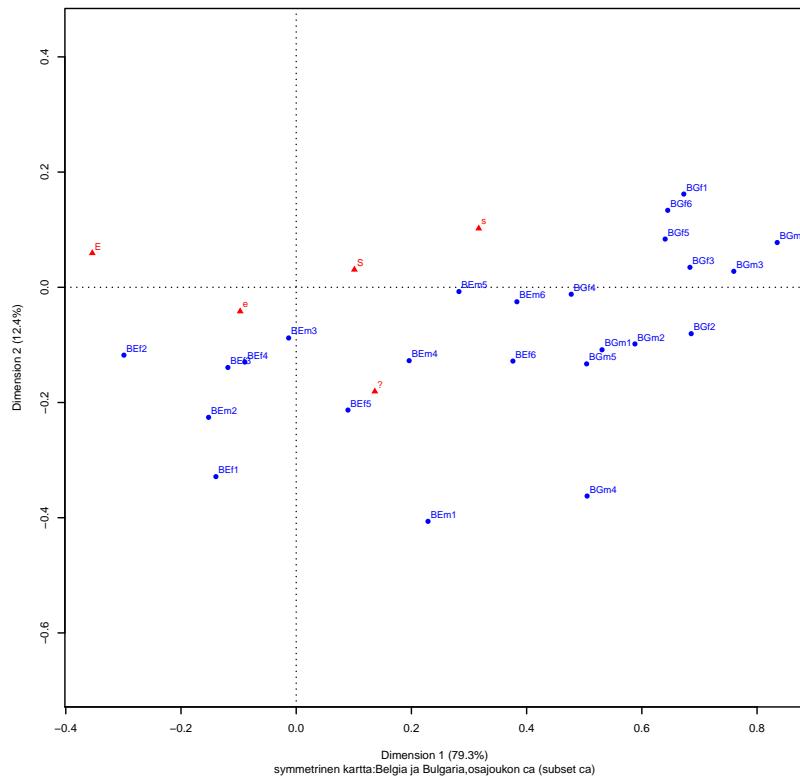
# main = "Äiti töissä: ikäluokka ja sukupuoli maittain" mikä on plot-funktion
# title-asetusten ja koodilohkon asetusten suhde? Ainakin plot-funkiton
# main = "Äiti töissä: ikäluokka ja sukupuoli maittain" korvaa koodilohkossa
# määritellyn (9.9.20). R-markdownissa kuvaan pääotsikko putoaa kokonaan pois,
# mitenköhän tulosteissa?
```

Yhdistetään pisteitä.

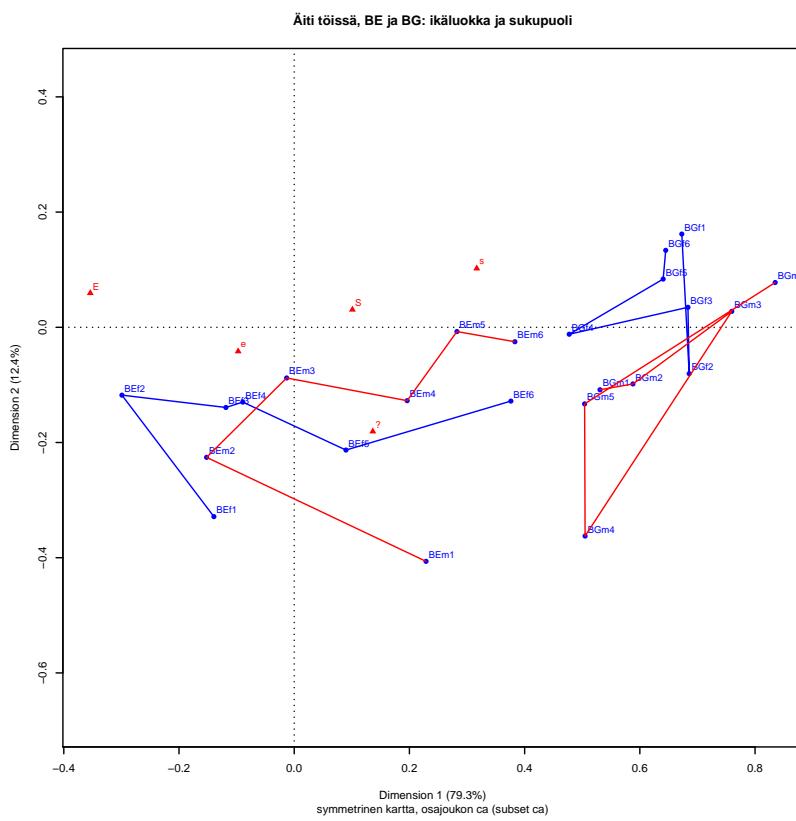
```
# CA- tulosoijekti maagaCA2subset1

maagaLinesBEBG <- cacoord(maagaCA2subset1BEBG, type = "symmetric")
maagaLinesBEBG <- maagaLinesBEBG$rows[, 1:2]
par(cex = 0.5)
plot(maagaCA2subset1BEBG,main = "Äiti töissä, BE ja BG: ikäluokka ja sukupuoli",
      sub = "symmetrinen kartta, osajoukon ca (subset ca)"
    )
lines(maagaLinesBEBG[1:6,1],maagaLinesBEBG[1:6,2], col="blue") #BEf
lines(maagaLinesBEBG[7:12,1],maagaLinesBEBG[7:12,2], col="red") #BEm
lines(maagaLinesBEBG[13:18,1],maagaLinesBEBG[13:18,2], col="blue") #BBGf
lines(maagaLinesBEBG[19:24,1],maagaLinesBEBG[19:24,2], col="red") #BEm
```

Kartan tulkintaa, kokonaisinertia voidaan esittää maa kerrallaan. (9.9.20)



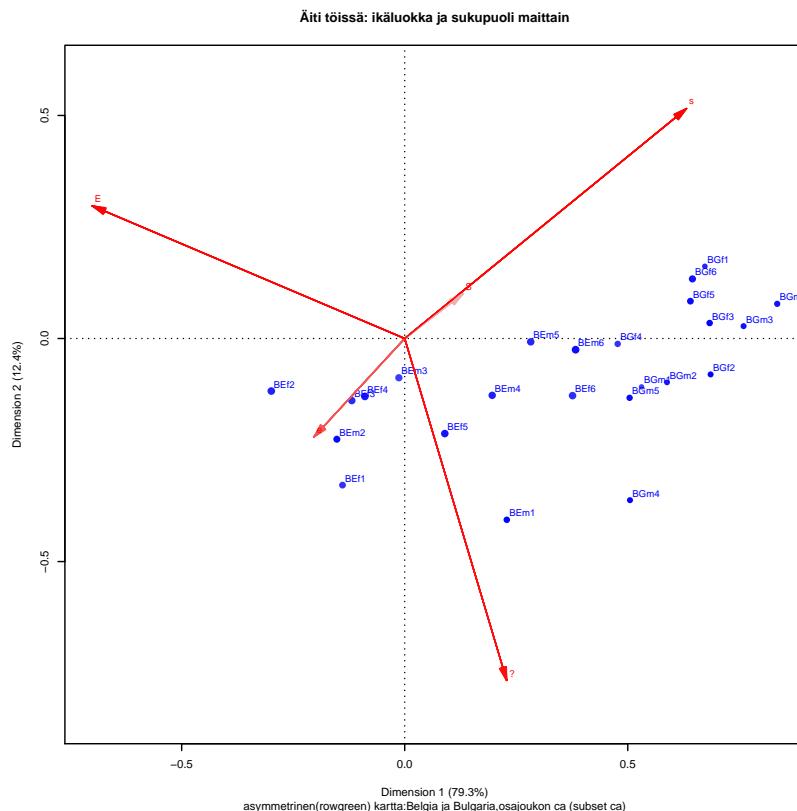
Kuva 26: Ikä, sukupuoli ja maa



Kuva 27: Ikä, sukupuoli ja maa - Belgia

```
# Datana maagaTab1 - viimeisinä riveinä maarivit

maagaCA2sub1 <- ca(maagaTab1[,1:5], subsetrow = 1:24)
par(cex = 0.5)
plot(maagaCA2sub1, map = "rowgreen",
      contrib = c("relative", "relative"),
      mass = c(TRUE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
      sub = "asymmetrinen(rowgreen) kartta:Belgia ja Bulgaria,osajoukon ca (subset ca"
      )
```



Lisätään kuvaan täydentäväänä (passiivisena) pisteenä Belgian maapiste, tarkistettava tuleeko se “oikeaan” kohtaan. **edit 10.10.20** Ei toimi, maapisteiden koordinaatit pitäisi laskea uudelleen. Syy: ca rivien osajoukolle. Maapisteiden osajoukon keskiarvopiste ei ole origo, mutta sarakepisteiden on.

```
# Datana maagaTab1 - viimeisinä riveinä maarivit
maagaTab1
```

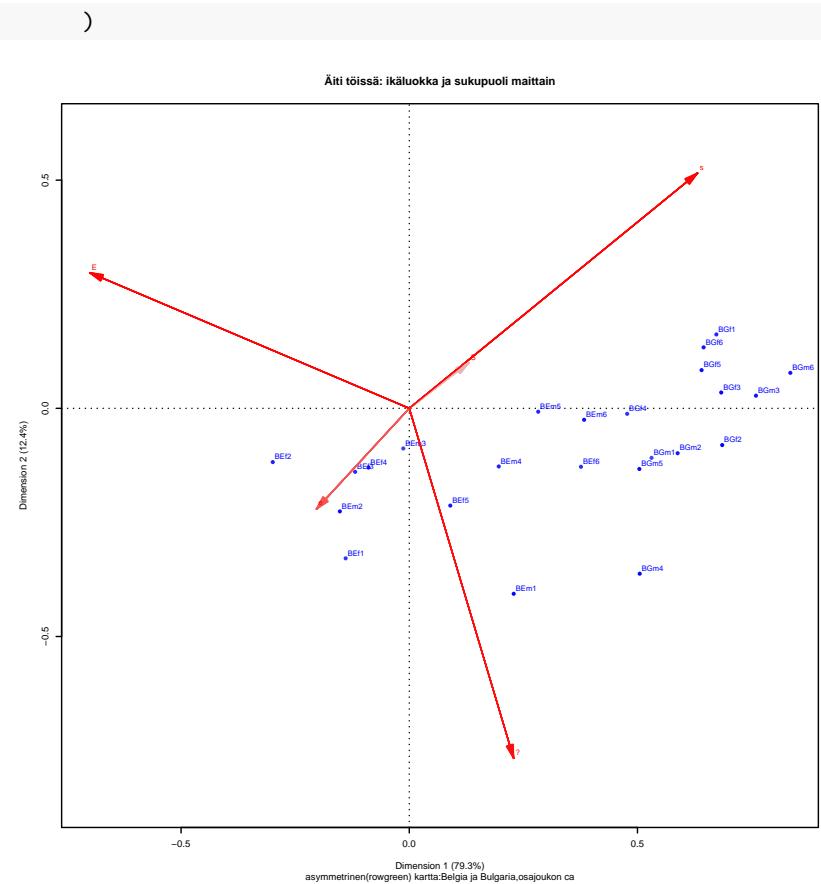
	S	s	?	e	E
BEf1	5	15	28	43	25
BEf2	10	26	34	66	62
BEf3	19	27	33	53	42
BEf4	21	34	40	55	49
BEf5	21	38	46	48	33
BEf6	25	58	50	30	22
BEm1	9	19	30	24	10
BEm2	10	19	31	40	35
BEm3	18	33	31	44	36
BEm4	19	46	37	51	23
BEm5	15	61	34	49	23
BEm6	19	75	44	49	21
BGf1	2	21	7	9	1
BGf2	7	28	17	12	0
BGf3	10	44	21	18	1
BGf4	14	30	15	24	2
BGf5	16	51	21	25	1
BGf6	27	66	26	27	3
BGm1	8	12	9	7	1
BGm2	4	21	12	14	0
BGm3	5	33	16	11	0
BGm4	7	19	21	15	1
BGm5	12	29	21	19	3
BGm6	6	41	19	9	0
DEF1	5	28	13	33	23
DEF2	9	14	14	37	46
DEF3	10	22	12	59	49
DEF4	11	31	20	53	71
DEF5	8	27	12	43	45
DEF6	31	40	15	50	49
DEM1	6	26	20	36	15
DEM2	7	26	13	39	18
DEM3	11	24	15	45	27
DEM4	22	39	17	57	37
DEM5	11	43	19	54	26
DEM6	34	55	28	32	32
DKf1	7	11	9	15	41
DKf2	4	15	7	13	71
DKf3	3	20	15	14	84
DKf4	5	24	8	19	90
DKf5	6	16	11	22	73
DKf6	5	26	11	17	40
DKm1	10	21	18	28	47
DKm2	2	11	9	16	65

	S	s	?	e	E
DKm3	2	13	12	23	59
DKm4	4	24	14	24	59
DKm5	11	14	23	18	40
DKm6	11	43	15	23	27
FIf1	3	9	13	36	33
FIf2	5	6	3	34	47
FIf3	2	8	13	39	32
FIf4	3	15	13	47	40
FIf5	6	26	17	52	41
FIf6	3	22	21	34	11
FIm1	1	9	13	22	13
FIm2	2	5	6	28	30
FIm3	2	10	9	27	23
FIm4	8	23	13	43	18
FIm5	5	31	15	35	10
FIm6	7	24	13	26	5
HUF1	11	13	16	11	3
HUF2	15	19	25	22	5
HUF3	22	26	26	12	9
HUF4	24	25	20	14	8
HUF5	21	28	19	19	7
HUF6	33	30	18	21	2
HUM1	9	15	12	8	5
HUM2	18	13	15	22	7
HUM3	15	38	24	16	10
HUM4	14	29	17	13	7
HUM5	19	31	24	21	7
HUM6	18	21	9	11	5
BE	191	451	438	552	381
BG	118	395	205	190	13
DE	165	375	198	538	438
DK	70	238	152	232	696
FI	47	188	149	423	303
HU	219	288	225	190	75

```

maagaCA2sub2 <- ca(maagaTab1[,1:5], subsetrow = 1:24)
par(cex = 0.4)
plot(maagaCA2sub2, map = "rowgreen",
      contrib = c("relative", "relative"),
      mass = c(FALSE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
      sub = "asymmetrinen(rowgreen) kartta:Belgia ja Bulgaria,osajoukon ca"

```



Belgian kolme aluetta täydentäväksi pisteiksi? **Ei tehdä, aika iso duuni.**(9.9.20)

```

# ISSP2012esim2.dat
# spCAmaaga1    maaga-ca-objekti (täydentävillä maa-pisteillä)
# maagaTab1      taulukko jossa maaga-rivit ja maat täydentävinä pisteinä
#
# Ongelma 1: miten saa maarivit kätevästi? Tässä tapauksessa näin
# maagaTab1
# Taulukon viimeisillä riveillä maa-profiilit frekvensseinä
# maaga-rivit ovat samassa järjestyksessä, kuusi naisten ja kuusi miesten
# ikäryhmää
# ISSP2012esim2.dat %>% tableX(maaga, Q1b)
#
# BE   191 451 438 552 381
# BG   118 395 205 190  13
# DE   165 375 198 538 438

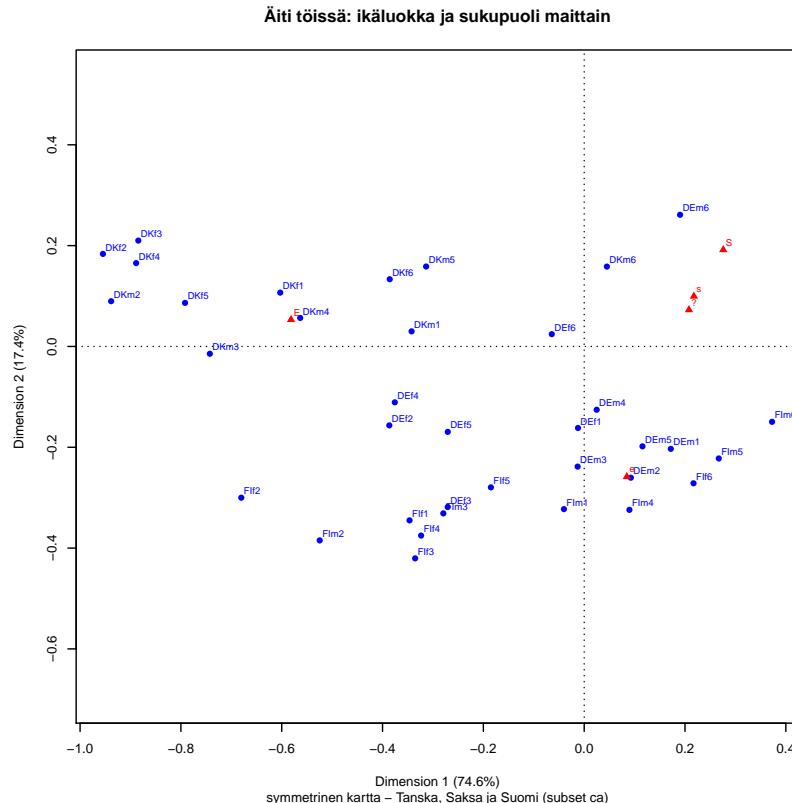
```

```

# DK      70 238 152 232 696
# FI      47 188 149 423 303
# HU      219 288 225 190  75
#
# BE 1-12, BG 13-24, DE 25-36, DK 37-48, FI 49-60, HU 61-72, maarivi
# Hoitaako ca-paketti automaattisesti täydentävien pisteen "skaalauksen
# subsetCA:ssa? Sarakepisteiden keskiarvo on origossa, mutta rivien osajoukon
# keskiarvo ei ole ja tämä pitäisi korjata.

maagaCA2sub2 <- ca(~maaga + Q1b, ISSP2012esim1b.dat, subsetrow = 25:60)
par(cex = 0.6)
plot(maagaCA2sub2, main = "Äiti töissä: ikäluokka ja sukupuoli maittain",
     sub = "symmetrinen kartta - Tanska, Saksa ja Suomi (subset ca)")

```



Sama kuva - yhdistetään pisteitä janoilla.

```

# ca-tulosobjekti maagaCA2sub2, DK DE FI

maagaLinesDKDEFI <- cacoord(maagaCA2sub2, type = "symmetric")

```

```
maagaLinesDKDEFI <- maagaLinesDKDEFI$`rows[ , 1:2]
maagaLinesDKDEFI
```

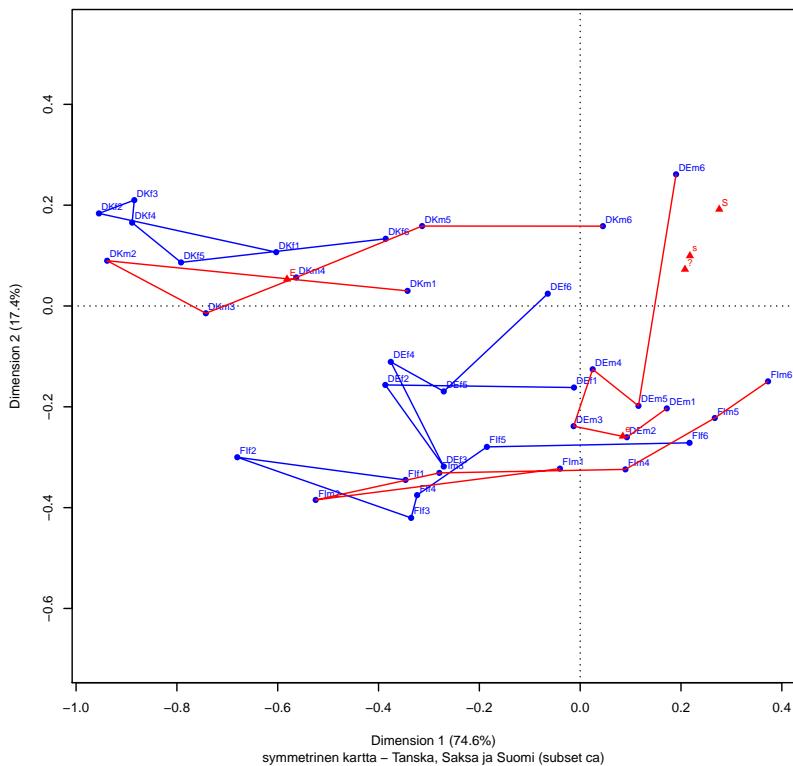
	Dim1	Dim2
DEF1	-0.0122683	-0.1617268
DEF2	-0.3865882	-0.1566004
DEF3	-0.2706794	-0.3184381
DEF4	-0.3756920	-0.1108647
DEF5	-0.2706035	-0.1694576
DEF6	-0.0642786	0.0244026
DEm1	0.1718294	-0.2031323
DEm2	0.0926227	-0.2603888
DEm3	-0.0130310	-0.2383625
DEm4	0.0247644	-0.1255945
DEm5	0.1156083	-0.1981363
DEm6	0.1901933	0.2611931
DKf1	-0.6029847	0.1067835
DKf2	-0.9546637	0.1836315
DKf3	-0.8845039	0.2100040
DKf4	-0.8887806	0.1653375
DKf5	-0.7917943	0.0864100
DKf6	-0.3858083	0.1333505
DKm1	-0.3423812	0.0299842
DKm2	-0.9383647	0.0897735
DKm3	-0.7426186	-0.0144571
DKm4	-0.5633286	0.0565446
DKm5	-0.3135585	0.1585389
DKm6	0.0450350	0.1583916
FIf1	-0.3464501	-0.3449977
FIf2	-0.6802179	-0.3000565
FIf3	-0.3353403	-0.4203573
FIf4	-0.3234953	-0.3751587
FIf5	-0.1850220	-0.2795225
FIf6	0.2168754	-0.2714628
FIm1	-0.0401893	-0.3226569
FIm2	-0.5246685	-0.3847647
FIm3	-0.2794497	-0.3311589
FIm4	0.0897328	-0.3241366
FIm5	0.2669923	-0.2222380
FIm6	0.3726818	-0.1495112

```
par(cex = 0.6)
plot(maagaCA2sub2,
```

```

sub = "symmetrinen kartta - Tanska, Saksa ja Suomi (subset ca)")
lines(maagaLinesDKDEFI[1:6,1],maagaLinesDKDEFI[1:6,2], col="blue") #DEf
lines(maagaLinesDKDEFI[7:12,1],maagaLinesDKDEFI[7:12,2], col="red") #DEM
lines(maagaLinesDKDEFI[13:18,1],maagaLinesDKDEFI[13:18,2], col="blue") #DKf
lines(maagaLinesDKDEFI[19:24,1],maagaLinesDKDEFI[19:24,2], col="red") #DKm
lines(maagaLinesDKDEFI[25:30,1],maagaLinesDKDEFI[25:30,2], col="blue") #FIf
lines(maagaLinesDKDEFI[31:36,1],maagaLinesDKDEFI[31:36,2], col="red") #FIm

```

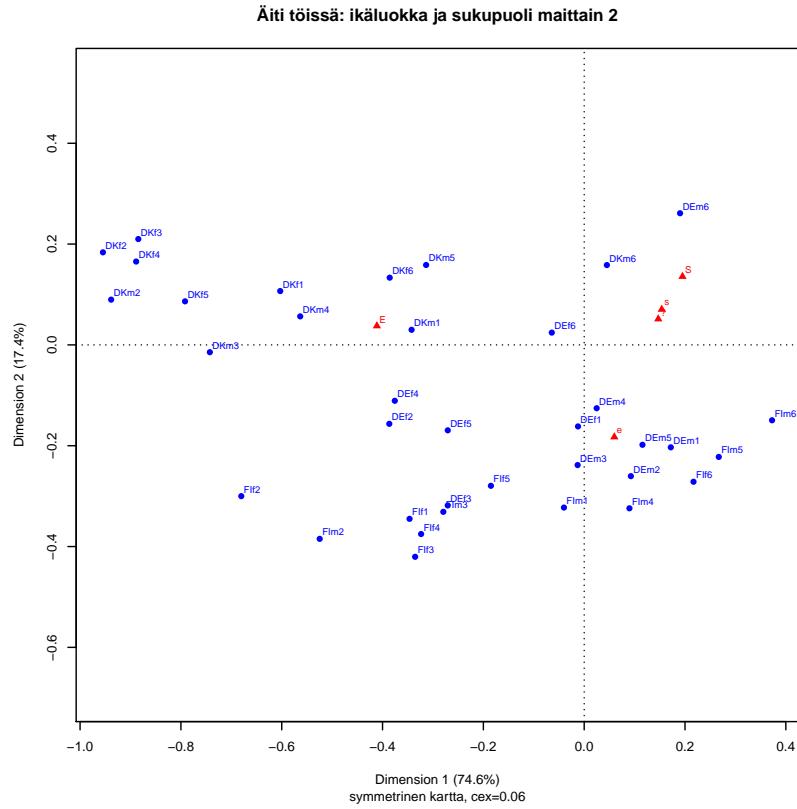


Kuva 28: Ikä, sukupuoli ja maa:Tanska-Saksa-Suomi

```

spCAmaagasub1 <- ca(maagaTab1[,1:5], subsetrow = 25:60 )
par(cex = 0.6)
plot(spCAmaagasub1, main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
     sub = "symmetrinen kartta, cex=0.06"
)

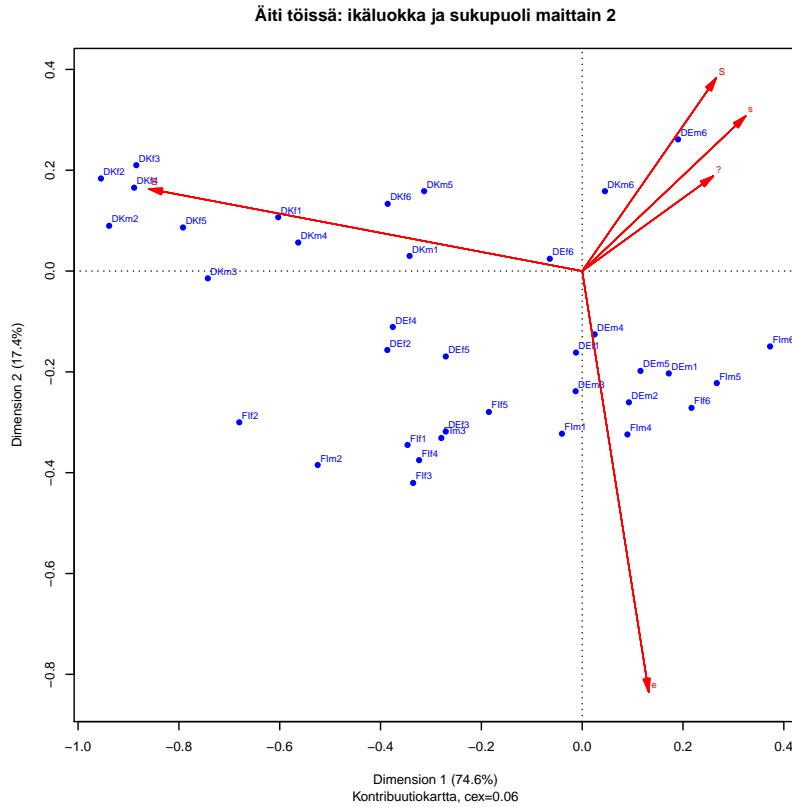
```



Sama kartta kuin edellinen, mutta kontribuutiokarttana.

```
# Osajoukko-ca ja täydentävät maapisteet

par(cex = 0.6)
plot(spCAMAAGASUB1, map = "rowgreen",
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä: ikäluokka ja sukupuoli maittain 2",
      sub = "Kontribuutiokartta, cex=0.06"
)
```



17.10.20 Kontribuutiokartta näyttää paremmalta.

Tutkitaan CA-ratkaisun tuloksia.

```
summary(spCAMAAGASUB1)
```

```
##  
## Principal inertias (eigenvalues):  
##  
##   dim      value      %    cum%    scree plot  
##   1      0.053545  74.6  74.6  ****  
##   2      0.012492  17.4  92.0  ***  
##   3      0.003297   4.6  96.6  *  
##   4      0.002441   3.4 100.0  *  
##  
##   -----  
##   Total: 0.071776 100.0  
##  
##  
## Rows:  
##       name     mass   qlt   inr     k=1 cor ctr     k=2 cor ctr
```

```

## 1 | DEf1 |   6 467   5 | -12  3  0 | -162 464 13 |
## 2 | DEf2 |   7 930 19 | -387 799 21 | -157 131 14 |
## 3 | DEf3 |   9 919 25 | -271 385 13 | -318 533 76 |
## 4 | DEf4 |  11 993 25 | -376 913 30 | -111 80 11 |
## 5 | DEf5 |   8 893 13 | -271 641 11 | -169 252 19 |
## 6 | DEf6 |  11   48 15 | -64  42  1 |  24   6  1 |
## 7 | DEm1 |   6 827   8 | 172 345  3 | -203 482 21 |
## 8 | DEm2 |   6 855   8 |   93  96  1 | -260 759 34 |
## 9 | DEm3 |   7 874   7 |  -13   3  0 | -238 871 34 |
## 10 | DEm4 |  11 285   8 |   25  11  0 | -126 274 13 |
## 11 | DEm5 |   9 684 10 | 116 174  2 | -198 510 30 |
## 12 | DEm6 |  11 750 22 | 190 260  8 | 261 490 61 |
## 13 | DKf1 |   5 979 27 | -603 949 35 | 107 30  5 |
## 14 | DKf2 |   7 996 89 | -955 960 115 | 184 36 18 |
## 15 | DKf3 |   8 985 98 | -885 933 122 | 210 53 29 |
## 16 | DKf4 |   9 983 104 | -889 950 132 | 165 33 20 |
## 17 | DKf5 |   8 1000 69 | -792 988 92 |  86 12  5 |
## 18 | DKf6 |   6 834 17 | -386 745 17 | 133 89  9 |
## 19 | DKm1 |   8 978 13 | -342 971 17 |  30  7  1 |
## 20 | DKm2 |   6 997 79 | -938 988 104 |  90  9  4 |
## 21 | DKm3 |   7 989 52 | -743 989 69 | -14  0  0 |
## 22 | DKm4 |   8 962 36 | -563 952 45 |  57 10  2 |
## 23 | DKm5 |   7 682 16 | -314 543 12 | 159 139 13 |
## 24 | DKm6 |   7 291   9 |   45  22  0 | 158 269 15 |
## 25 | FIIf1 |   6 951 20 | -346 478 13 | -345 474 55 |
## 26 | FIIf2 |   6 941 48 | -680 788 50 | -300 153 42 |
## 27 | FIIf3 |   6 952 24 | -335 370 12 | -420 582 82 |
## 28 | FIIf4 |   7 999 25 | -323 426 14 | -375 573 82 |
## 29 | FIIf5 |   9 982 14 | -185 299  6 | -280 683 55 |
## 30 | FIIf6 |   6 704 13 |  217 274  5 | -271 430 33 |
## 31 | FIm1 |   4 624   8 |  -40  10  0 | -323 614 30 |
## 32 | FIm2 |   4 984 26 | -525 640 22 | -385 344 52 |
## 33 | FIm3 |   4 990 12 | -279 412  6 | -331 578 38 |
## 34 | FIm4 |   6 944 11 |   90  67  1 | -324 877 54 |
## 35 | FIm5 |   6 722 14 |  267 426  8 | -222 295 23 |
## 36 | FIm6 |   5 911 11 |  373 785 12 | -150 126  8 |

##
## Columns:
##      name mass qlt inr k=1 cor ctr k=2 cor ctr
## 1 |   S   99 731 107 | 195 493 71 | 136 238 147 |
## 2 |   s   238 832 114 | 154 688 105 | 70 144 94 |
## 3 |   | 168 647 88 | 147 576 68 | 51 70 35 |
## 4 |   e   261 992 135 | 60 96 17 | -183 896 697 |
## 5 |   E   234 1000 556 | -411 992 739 | 38 8 27 |

```

Kolmen maan osajoukon ratkaisussa 2. dimensiolla (maltillinen liberaali?) on

inertiasta 17 prosenttia, edellä ollut paljon yksiuotteinempia ratkaisuja. Huono kvaliteetti on DEF1 (467) ja DEF6 (48!), DEM4 (285). Tanskan havainnoista vanhimmat miehet (DKm6,291) ovat kaikkein huonoimmin esitettyjä ratkaisussa, ja hieman nuoremmatkin (DKm5, 682). Suomen aineistossa vain nuoret miehet (FIM1, 624) on esitetty kartalla huonosti. Kaksi dimensiot selittävät osajoukon kokonaishajonnasta 92%, mutta muutaman ryhmän hajonta on muissa dimensioissa. Saksan naisten iäkkääin ikäluokka (DEF6) ja keski-ikäisen miehet (DEM4) vain näyttävät olevan lähekkääin origon tuntumassa, samoin muutama muu huonosti tasoon sijoitettu piste.

Huonosti kuvatuista pisteistä ei kuva ei siis kerro oikeastaan mitään muuta.

Sarakkeet on esitetty kohtalaisen hyvin, ja symmetrisessä kartassa tärkeimmälle dimensioille projisodut sarakepisteet ovat odotetussa järjestyksessä.

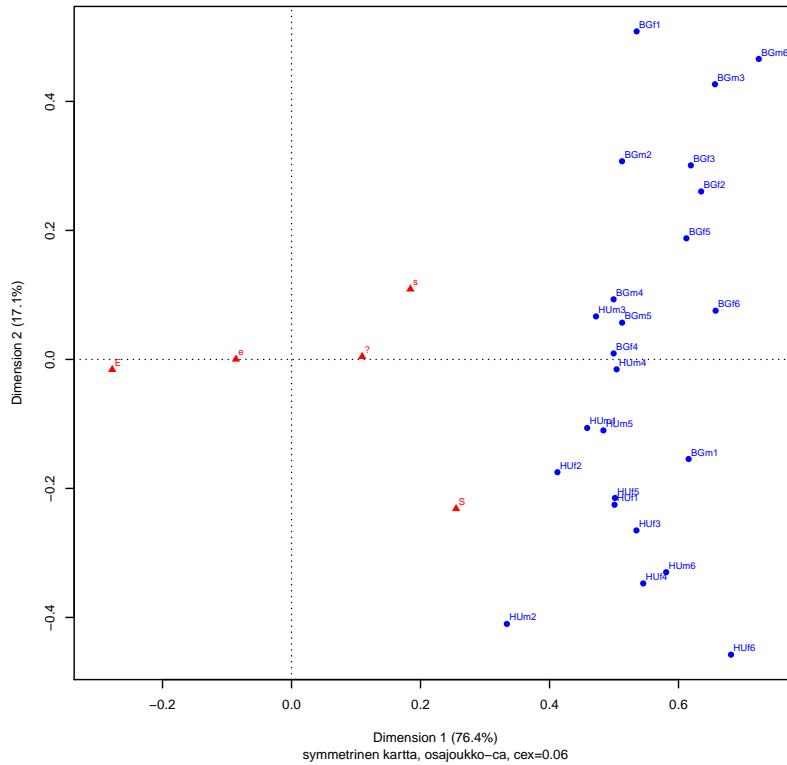
Kontribuutiokartasta nähdään, että tärkein kontrasti on tiukan erimielisyiden (E) ja kaikkien muiden vastausvaihtoehtojen välillä. Epävarmojen tai maltillissten (e) kontrasti hallitsee toista dimensiota, erityisesti S- ja s- kategorioiden kanssa. Samalla kuvasta näkee (ja numeerisist tuloksista voi vahvistaa), että S-piste on lähempänä (kulma on pienempi) pystyakselia. Kontribuutio on suurempi (147 vs. 71 x-akselille). Toisaalta x-akseli selittää selvästi suurimman osan kaikkien muiden sarakepisteiden inertiesta, ja y-akseli taas lähes täysin e-pisteen inertian.

Osajoukko-ca neljälle ryhmälle Belgia on vähän rajatapaus, kokeillaan osajoukko-ca:ta neljällä maaryhmällä. BG-HU, BE-DE-DK-FI, DEDKFI ja BEBGHSubset

```
# Neljä maaryhmää
BGHSubset <- c(13:24, 61:72)
BEDEDKFISubset <- c(1:12, 25:36, 37:48, 49:60)
DEDKFISubset <- c(25:36, 37:48, 49:60)
BEBGHSubset <- c(1:12, 13:24, 61:72)

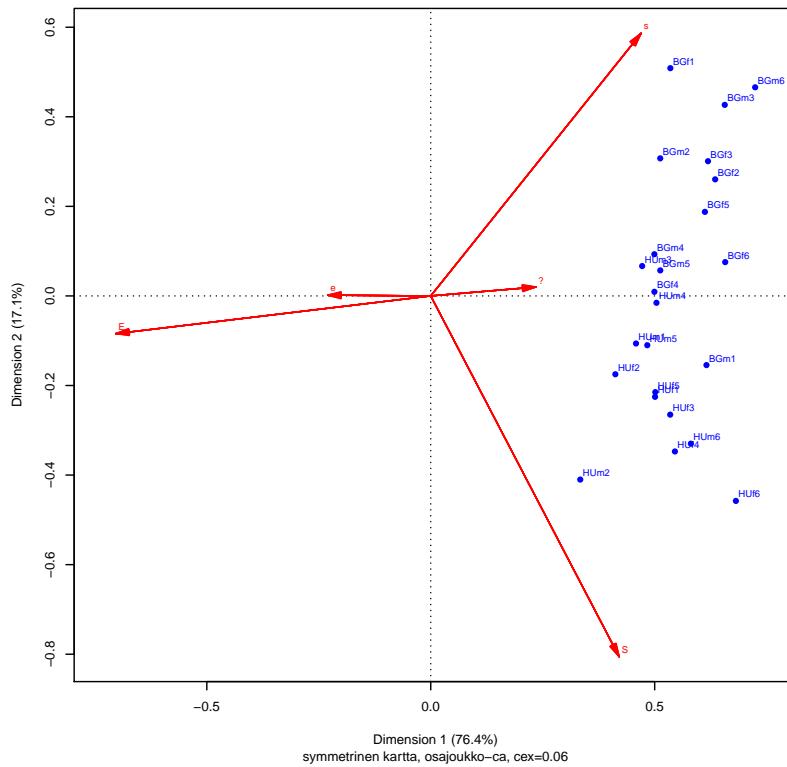
spCAmaagasub3 <- ca(maagaTab1[,1:5], subsetrow = BGHSubset)
par(cex = 0.6)
plot(spCAmaagasub3, main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
     sub = "symmetrinen kartta, osajoukko-ca, cex=0.06"
)
```

Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



```
spCAMAAGASUB3 <- ca(maagaTab1[,1:5], subsetrow = BGHUsubset)
par(cex = 0.6)
plot(spCAMAAGASUB3, map = "rowgreen",
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "symmetrinen kartta, osajoukko-ca, cex=0.06"
)
```

Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



```
# xlim=c(-0.1,0.8), ylim=c(-0.,0.4) kuvaa voisi säättää, mutta tärkein E-sarake
# jäisi pois
summary(spCAMAAGASUB3)

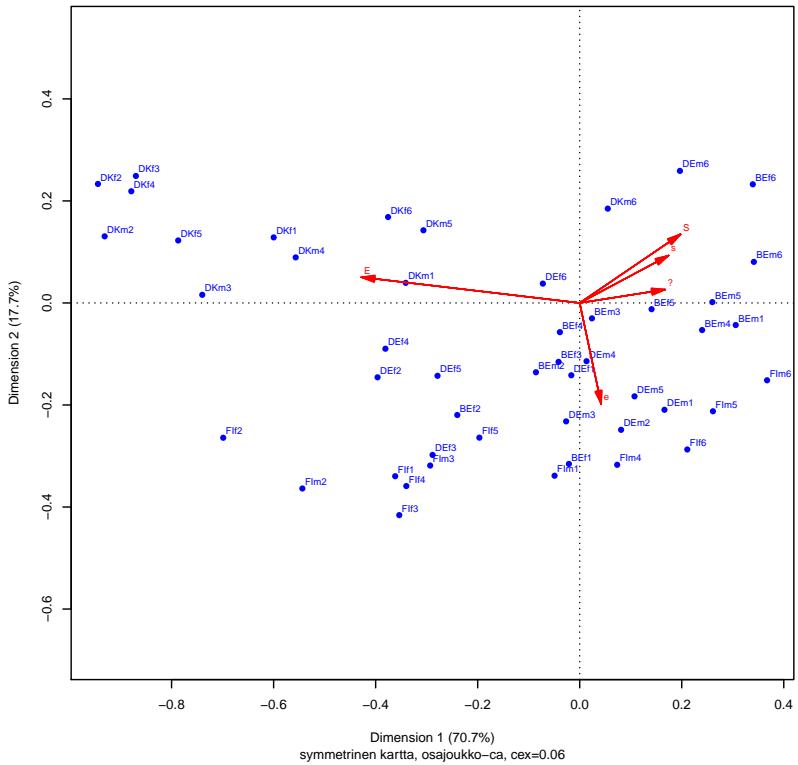
##
## Principal inertias (eigenvalues):
##
##   dim      value      %   cum%   scree plot
## 1    0.036585 76.4 76.4 ****
## 2    0.008208 17.1 93.5 ****
## 3    0.002065  4.3 97.8 *
## 4    0.001054  2.2 100.0 *
##   -----
##   Total: 0.047912 100.0
##
##
## Rows:
##       name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | BGf1 |    2 966  29 | 535 507 19 | 509 459 77 |
```

```

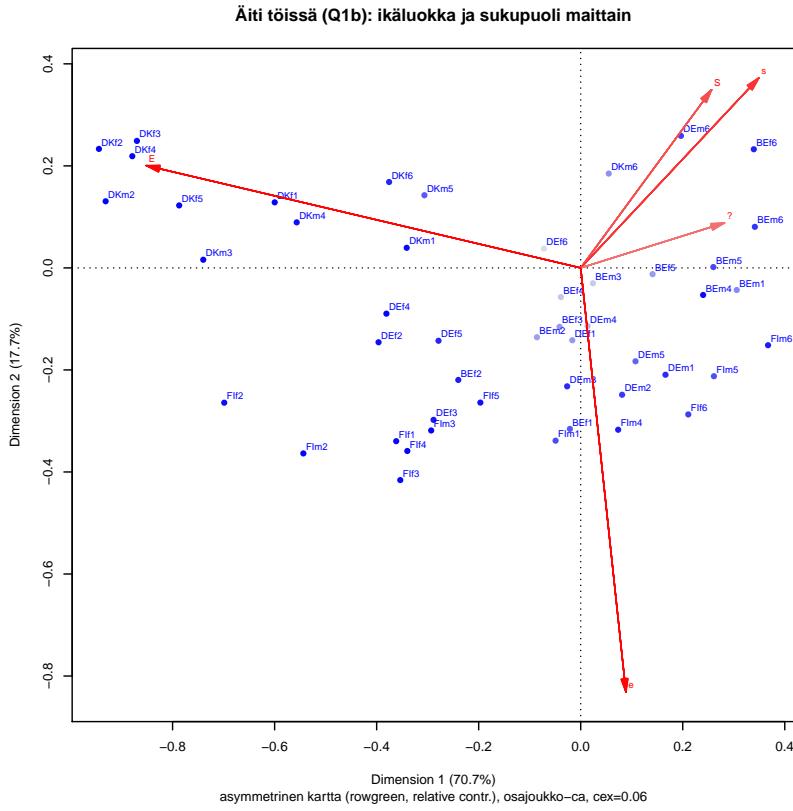
## 2 | BGf2 | 4 980 39 | 635 839 43 | 260 141 32 |
## 3 | BGf3 | 6 999 57 | 619 808 60 | 301 191 64 |
## 4 | BGf4 | 5 859 32 | 499 859 36 | 9 0 0 |
## 5 | BGf5 | 7 960 62 | 612 878 72 | 188 83 30 |
## 6 | BGf6 | 9 942 89 | 657 930 108 | 75 12 6 |
## 7 | BGm1 | 2 994 19 | 616 935 24 | -154 59 7 |
## 8 | BGm2 | 3 906 26 | 512 667 22 | 307 240 36 |
## 9 | BGm3 | 4 997 51 | 656 701 47 | 427 296 89 |
## 10 | BGm4 | 4 667 31 | 499 644 26 | 93 22 4 |
## 11 | BGm5 | 5 948 30 | 513 936 37 | 57 12 2 |
## 12 | BGm6 | 5 977 73 | 724 691 66 | 466 286 122 |
## 13 | HUf1 | 3 847 25 | 501 704 23 | -225 143 21 |
## 14 | HUf2 | 5 717 31 | 412 607 25 | -175 109 20 |
## 15 | HUf3 | 6 890 49 | 535 715 46 | -265 176 50 |
## 16 | HUf4 | 6 976 50 | 545 694 45 | -347 282 82 |
## 17 | HUf5 | 6 996 36 | 501 842 40 | -215 154 32 |
## 18 | HUf6 | 6 964 93 | 681 664 81 | -458 300 163 |
## 19 | HUm1 | 3 932 15 | 458 885 17 | -106 48 4 |
## 20 | HUm2 | 5 900 30 | 334 359 14 | -410 541 94 |
## 21 | HUm3 | 6 939 32 | 472 920 39 | 67 18 3 |
## 22 | HUm4 | 5 964 27 | 504 963 34 | -15 1 0 |
## 23 | HUm5 | 6 984 33 | 484 936 40 | -110 48 9 |
## 24 | HUm6 | 4 883 41 | 581 668 36 | -330 215 52 |
##
## Columns:
##      name   mass   qlt   inr    k=1 cor ctr    k=2 cor ctr
## 1 | S | 99 987 249 | 255 542 177 | -231 446 649 |
## 2 | s | 238 965 235 | 184 715 221 | 109 250 344 |
## 3 | | 168 568 74 | 110 567 55 | 4 1 0 |
## 4 | e | 261 727 56 | -86 727 53 | 0 0 0 |
## 5 | E | 234 983 385 | -278 980 494 | -16 3 7 |
spCAmaagasub4 <- ca(maagaTab1[,1:5], subsetrow = BEDEDKFIsubset)
par(cex = 0.6)
plot(spCAmaagasub4,
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "symmetrinen kartta, osajoukko-ca, cex=0.06"
)

```

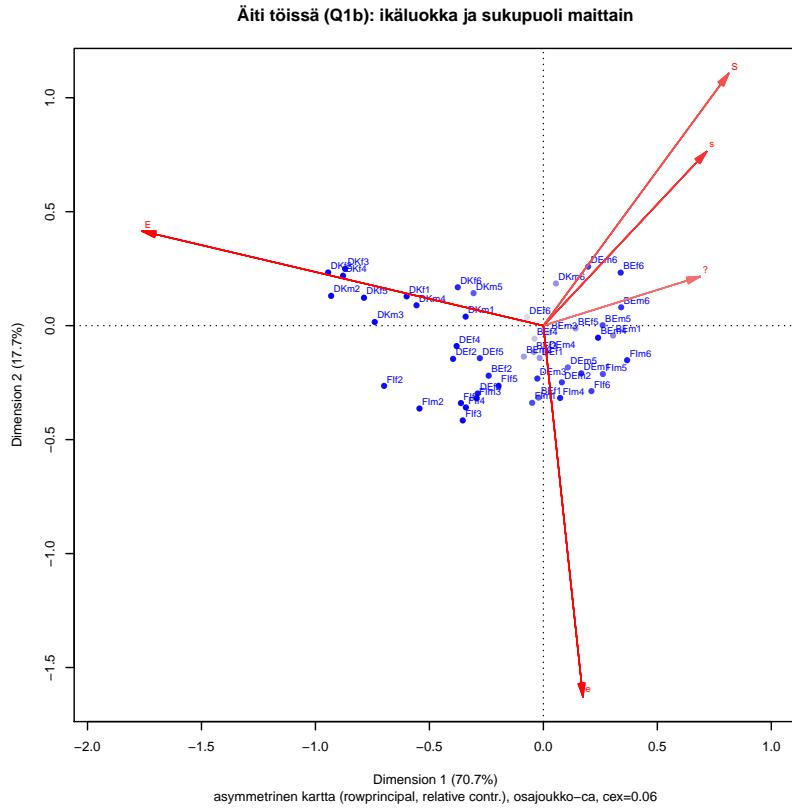
Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



```
# asyymmetrinen kartta
par(cex = 0.6)
plot(spCAMAAGASUB4, map = "rowgreen",
      contrib = c("relative", "relative"),
      mass = c(FALSE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "asyymmetrinen kartta (rowgreen, relative contr.), osajoukko-ca, cex=0.06")
)
```



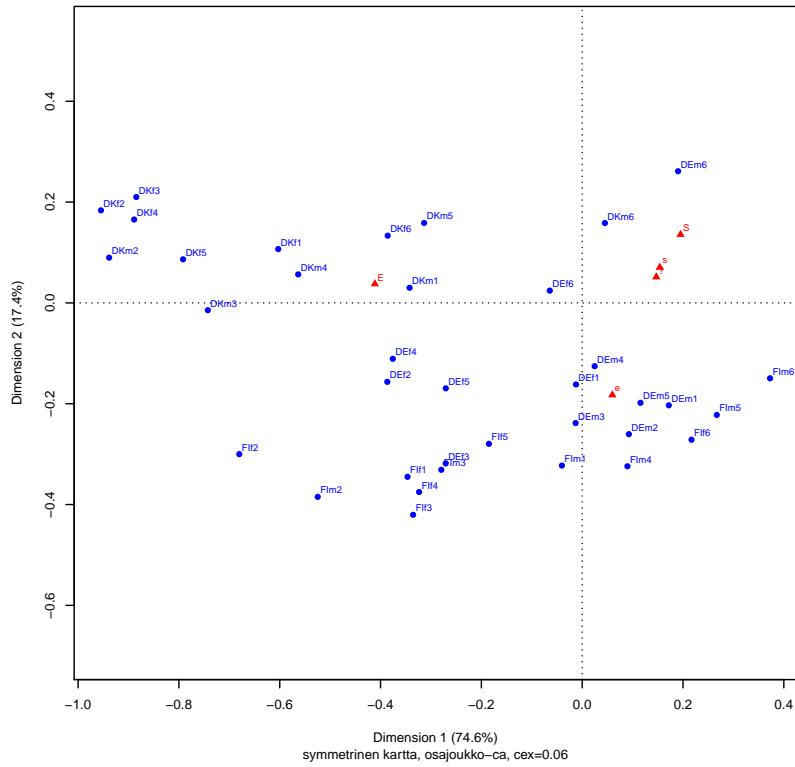
```
par(cex = 0.6)
plot(spCAMAAGASUB4, map = "rowprincipal",
      contrib = c("relative", "relative"),
      mass = c(FALSE, TRUE),
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "asymmetrisen kartta (rowprincipal, relative contr.), osajoukko-ca, cex=0.06"
)
```



17.10.20 Kontribuutiokartta on paras. Asymmetrinen on liian tukkoinen.

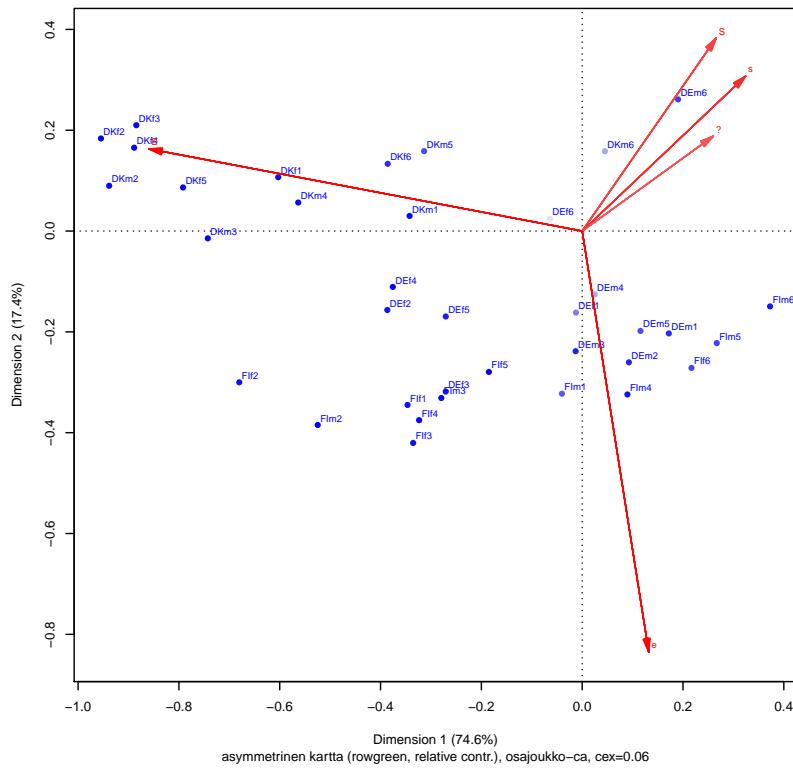
```
spCAmaagasub5 <- ca(maagaTab1[,1:5], subsetrow = DEDKFIs subset)
par(cex = 0.6)
plot(spCAmaagasub5, main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "symmetrinen kartta, osajoukko-ca, cex=0.06"
)
```

Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



```
# asymmetrinen kartta
par(cex = 0.6)
plot(spCAMAAGASUB5, map = "rowgreen",
     contrib = c("relative", "relative"),
     mass = c(FALSE, TRUE),
     arrows = c(FALSE, TRUE),
     main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
     sub = "asymmetrinen kartta (rowgreen, relative contr.), osajoukko-ca, cex=0.06"
)
```

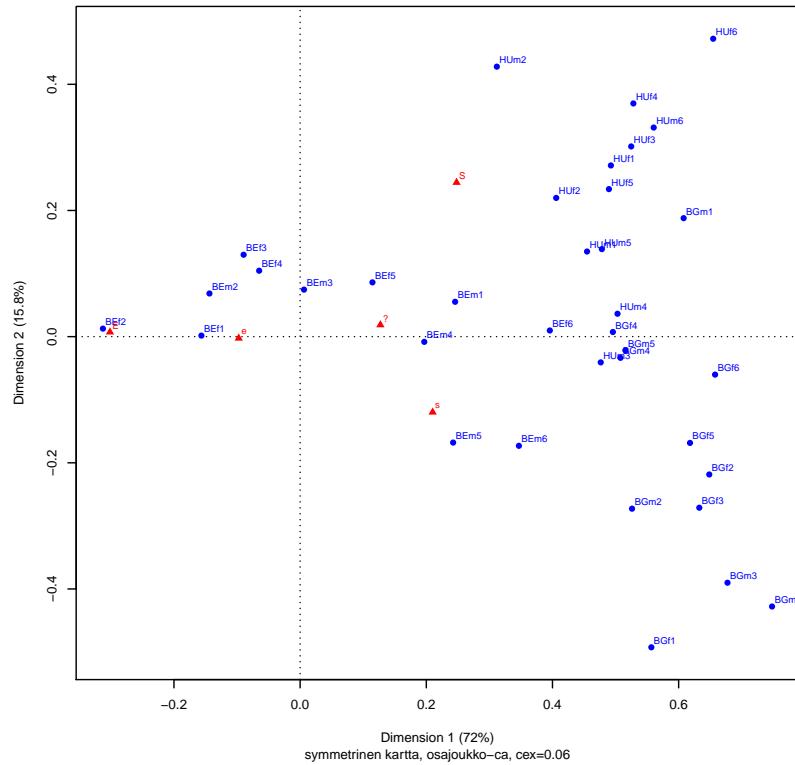
Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



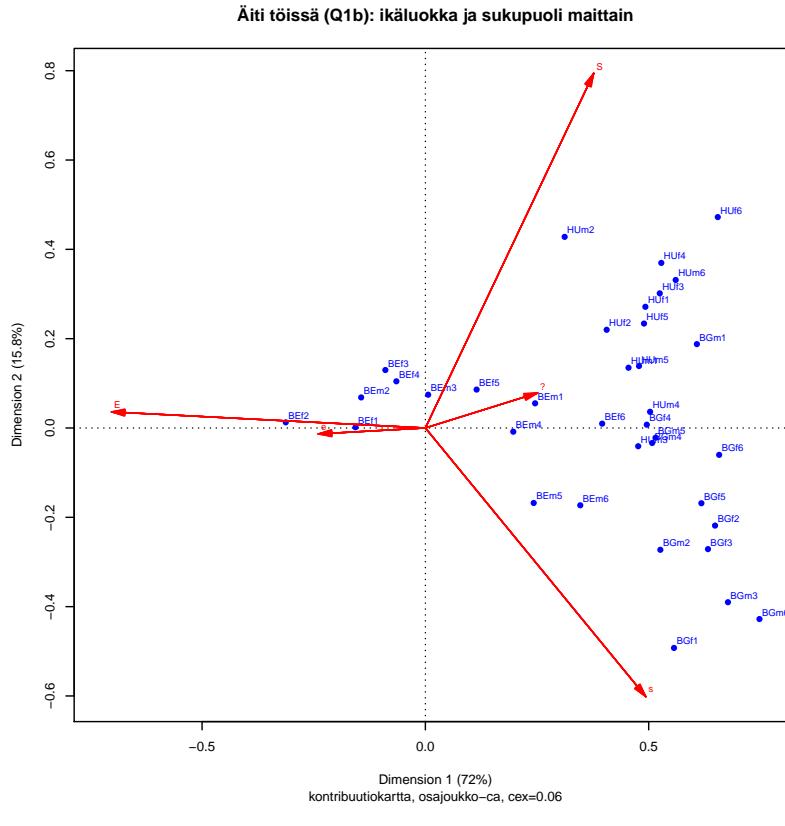
asymmetrinen kartta (rowgreen, relative contr.), osajoukko-ca, cex=0.06

```
spCAMAAGASUB6 <- ca(maagaTab1[,1:5], subsetrow = BEBGHUSUBSET)
par(cex = 0.6)
plot(spCAMAAGASUB6, main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
     sub = "symmetrinen kartta, osajoukko-ca, cex=0.06"
)
```

Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain



```
plot(spCAMAAGASUB6, map = "rowgreen",
      arrows = c(FALSE, TRUE),
      main = "Äiti töissä (Q1b): ikäluokka ja sukupuoli maittain",
      sub = "kontribuutiokartta, osajoukko-ca, cex=0.06"
)
```



5 Yksinkertaisen korrespondenssianalyysin laajennuksia 2

Kaksi ensimmäistä lukua ehkä omaksi kokonaisuudeksi, ohitetaan aika kevyesti.

5.1 Matriisien yhdistäminen (stacked and concatenated matrices)

Pinotuista tauluista yksinkertainen esimerkki rajatulla aineistolla, selitetään periaate. Ei laajenneta aineistoa, sillä puuttuvat tiedot aiheuttavat pulmia joihin sopii parhaiten MCA. MCA on muuttujien välisten suhteiden analyysiä, näin puuttuvista tiedosta saadan otetta.

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

Concatenated tables (yhdistetyt taulut tai matriisit): (a) kaksi luokittelumuuttuja (b) useita muuttujia stacked (“pinotaan”).

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

```
# Data
ISSP2012Concat1jh.dat <- select(ISSP2012esim1b.dat, Q1b, maa,sp, age_cat)

# mjca-funktioita -> Burt-matriisi
Concat1jh.Burt <- mjca(ISSP2012Concat1jh.dat, ps=""")$Burt

# Burt-matriisi symmetrinen
#dim(Concat1jh.Burt)
# 19 x 19
#rownames(Concat1jh.Burt)
#[1] "Q1bS"      "Q1bs"      "Q1b?"      "Q1be"      "Q1bE"      "maaBE"      "maaBG"      "maaDE"
#[10] "maaFI"     "maaHU"     "spm"       "spf"       "age_cat1"   "age_cat2"   "age_cat3"   "age_cat4"
#[19] "age_cat6"

# maat - vastaukset
ISSP2012Concat2jh.dat <- Concat1jh.Burt[6:11, 1:5]
# ISSP2012Concat2jh.dat
# sukupuoli ja vastaukset
ISSP2012Concat2jh.dat <- rbind(ISSP2012Concat2jh.dat, Concat1jh.Burt[12:13 ,1:5])
# ISSP2012Concat2jh.dat
# ikäluokka ja vastaukset
ISSP2012Concat2jh.dat <- rbind(ISSP2012Concat2jh.dat, Concat1jh.Burt[14:19 ,1:5])
# ISSP2012Concat2jh.dat

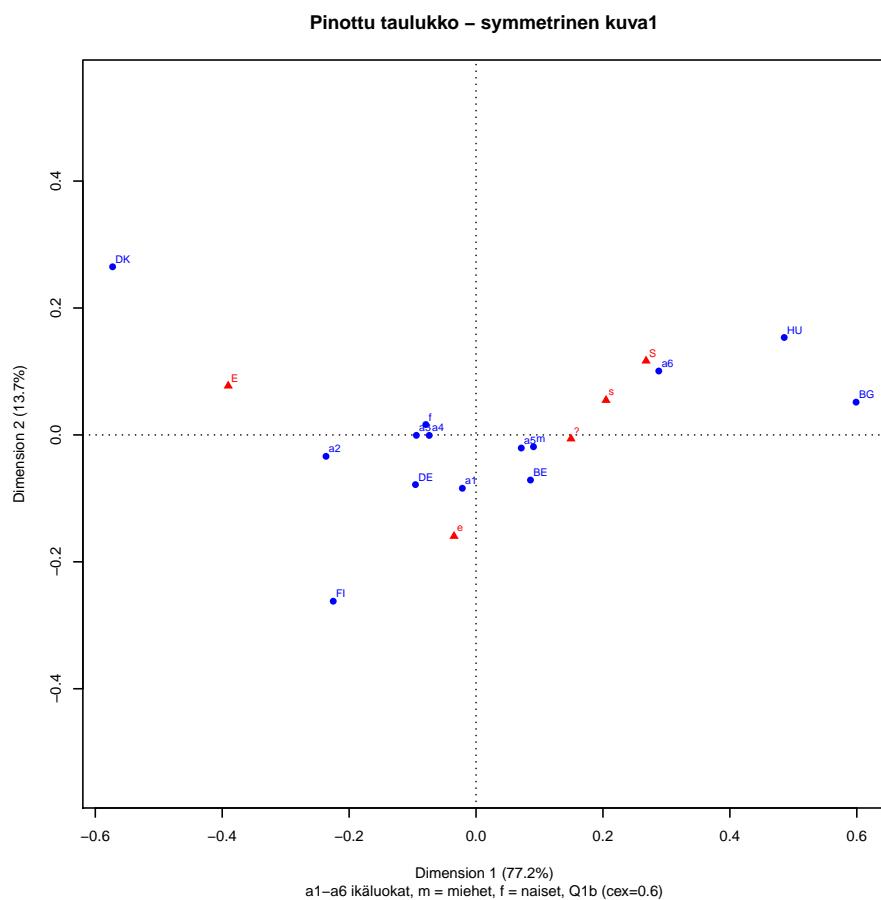
Concat1jh.CA1 <- ca(ISSP2012Concat2jh.dat)
# plot(Concat1jh.CA1)

# Siistitään muuttujien nimet

Concat1jh.CA1$colnames <- c("S", "s", "?", "e", "E")
Concat1jh.CA1$rownames <- c("BE", "BG", "DE", "DK", "FI", "HU", "m", "f",
                           "a1", "a2", "a3", "a4", "a5", "a6")
# Käänetään kuva x-akselin ympäri
Concat1jh.CA1$rowcoord[, 2] <- -Concat1jh.CA1$rowcoord[, 2]
Concat1jh.CA1$colcoord[, 2] <- -Concat1jh.CA1$colcoord[, 2]
```

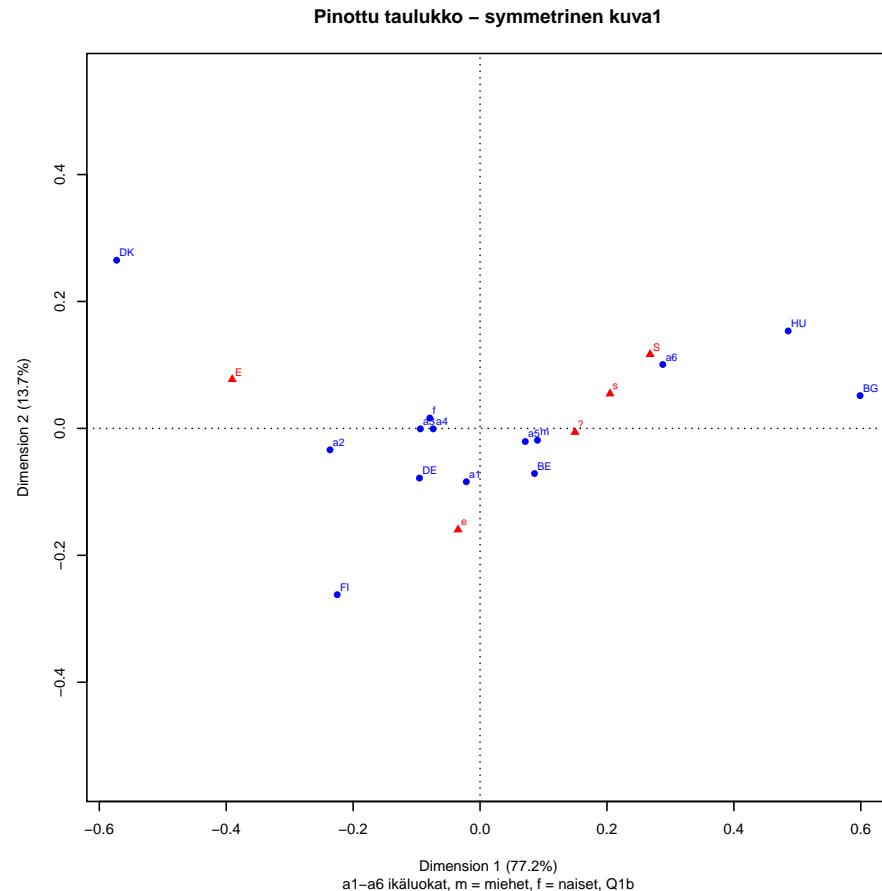
Piirretään karttoja.

```
par(cex = 0.6)
plot(Concat1jh.CA1,
      main = "Pinottu taulukko - symmetrinen kuva1",
      sub = "a1-a6 ikäluokat, m = miehet, f = naiset, Q1b (cex=0.6)"
      )
```



Kuva 29: Pinottu matriisi - kartta 1

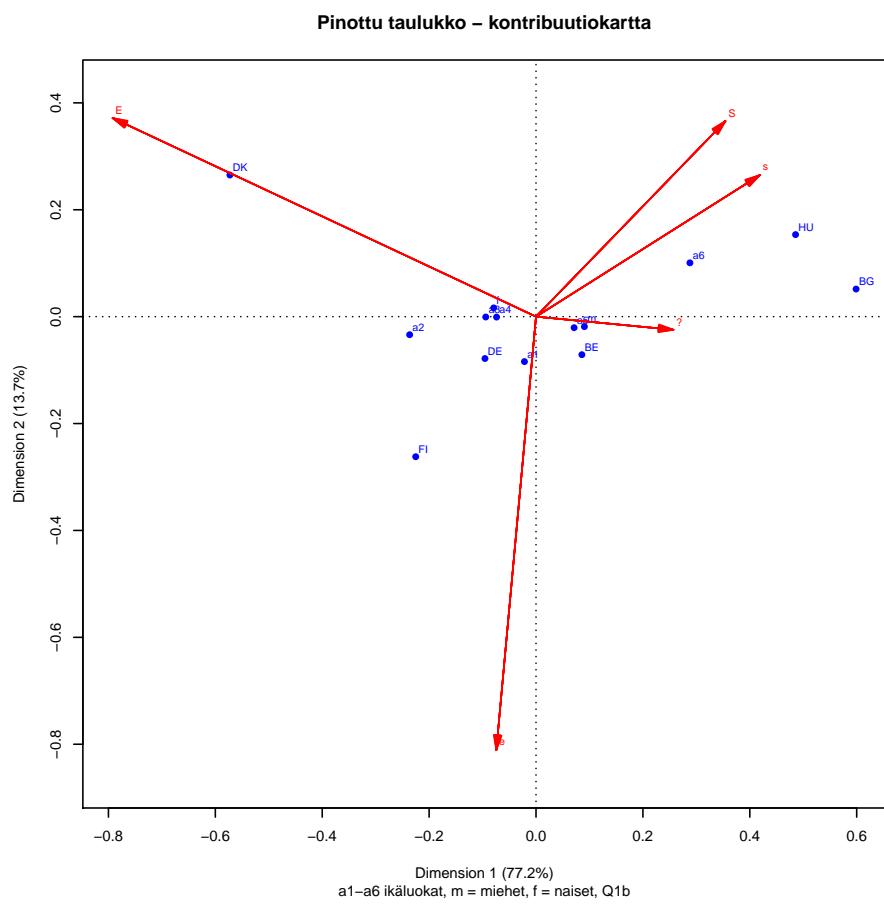
```
plot(Concat1jh.CA1,
      main = "Pinottu taulukko - symmetrinen kuva1",
      sub = "a1-a6 ikäluokat, m = miehet, f = naiset, Q1b"
    )
```



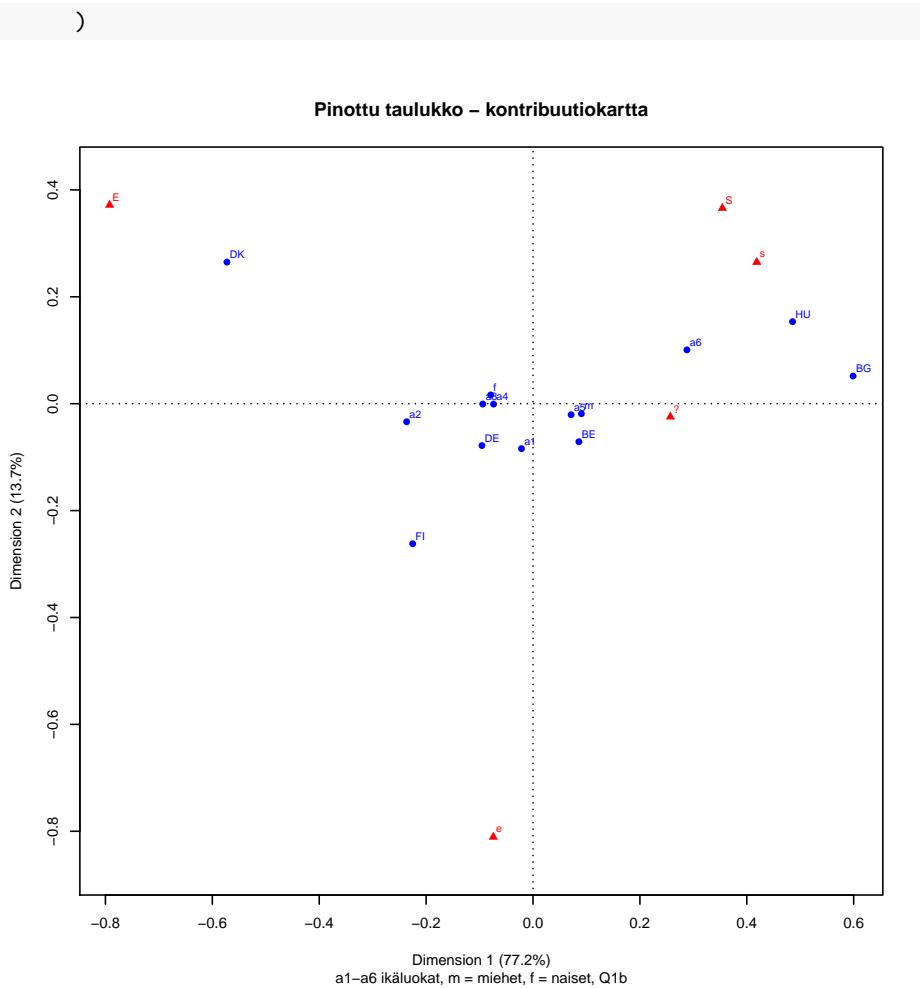
Kuva 30: Pinottu matriisi - kartta 1

```
plot(Concat1jh.CA1, map = "rowgreen",
      arrows = c(FALSE, TRUE),
      main = "Pinottu taulukko - kontribuutiokartta",
      sub = "a1-a6 ikäluokat, m = miehet, f = naiset, Q1b"
    )

plot(Concat1jh.CA1, map = "rowgreen",
      main = "Pinottu taulukko - kontribuutiokartta",
      sub = "a1-a6 ikäluokat, m = miehet, f = naiset, Q1b"
```



Kuva 31: Pinottu matriisi - kartta 1

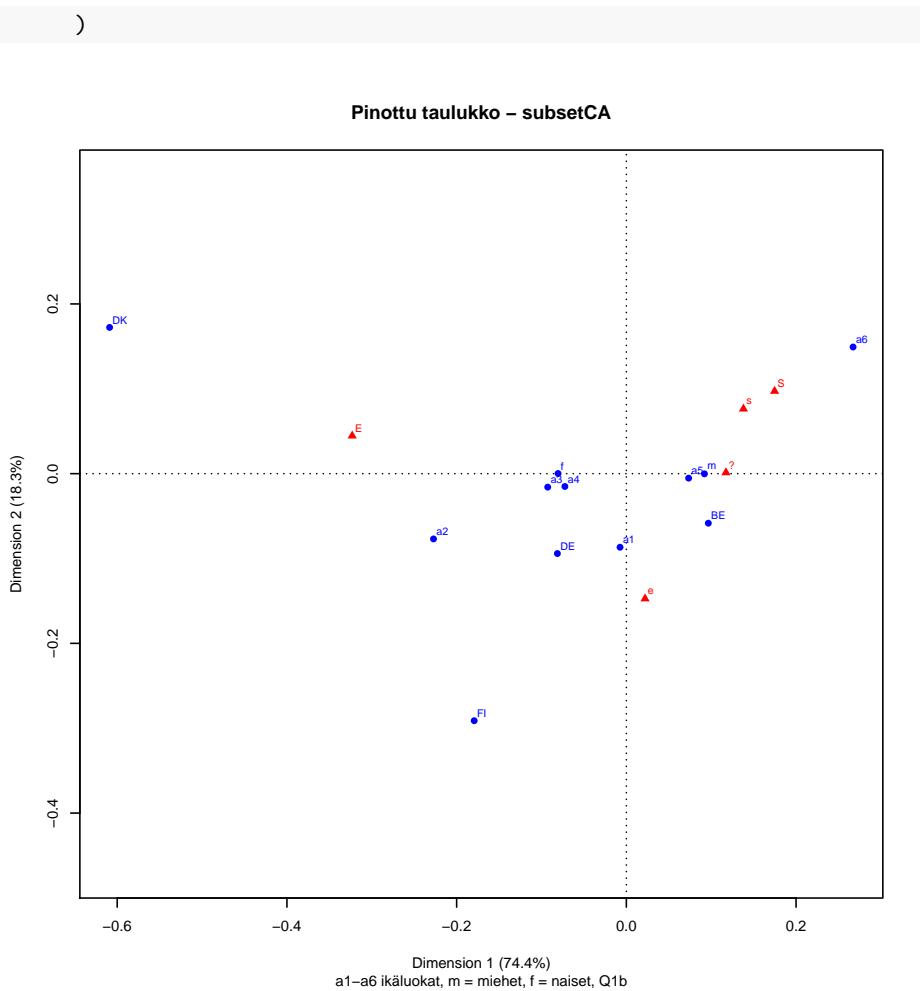


Kuva 32: Pinottu matriisi - kartta 1

```
# Tukkoinen kuva - rajataan pois HU ja BG (rivit 2, 6)
# dim(ISSP2012Concat2jh.dat)

subConcat1jh.CA1 <- ca(ISSP2012Concat2jh.dat[,1:5], subsetrow = c(1:1, 3:5, 7:14))
subConcat1jh.CA1$colnames <- c("S", "s", "?", "e", "E")
subConcat1jh.CA1$rownames <- c("BE", "DE", "DK", "FI", "m", "f",
                                "a1", "a2", "a3", "a4", "a5", "a6")

plot(subConcat1jh.CA1,
      main = "Pinottu taulukko - subsetCA",
      sub = "a1-a6 ikäluokat, m = miehet, f = naiset, Q1b"
```

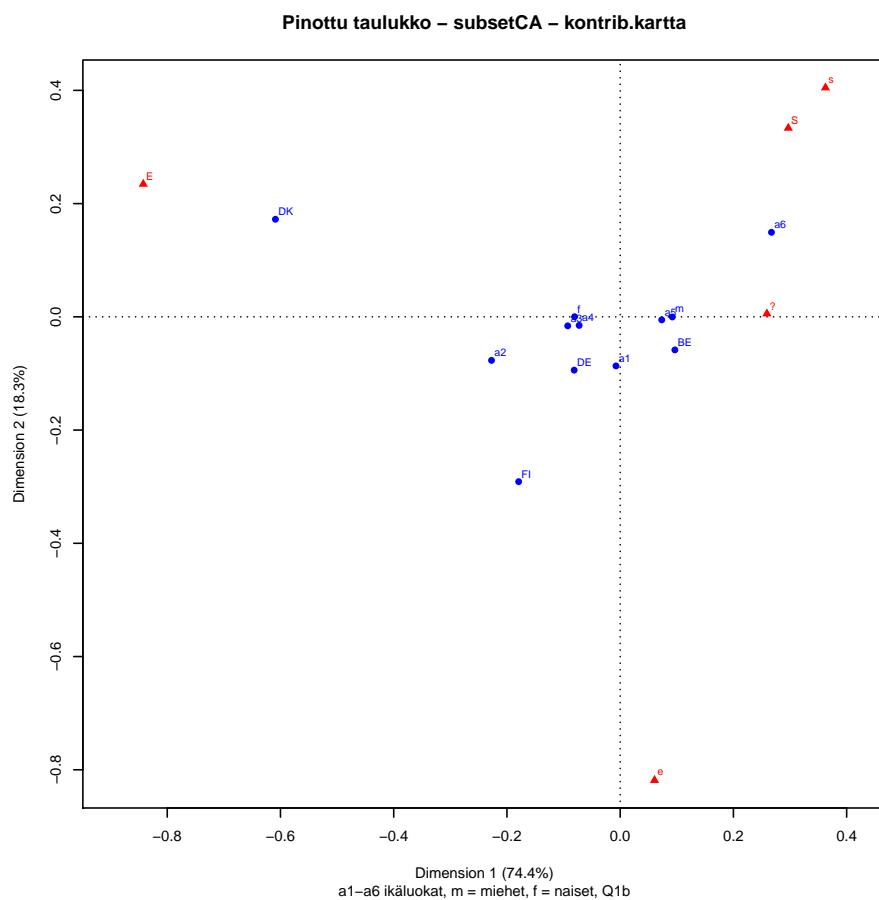


Kuva 33: Pinottu matriisi - kartta 1

```
# Kokeillaan kontribuutiokuvaaa
plot(subConcat1jh.CA1, map = "rowgreen",
      main = "Pinottu taulukko – subsetCA – kontrib.kartta",
      sub = "a1–a6 ikäluokat, m = miehet, f = naiset, Q1b"
    )

summary(Concat1jh.CA1)

## 
## Principal inertias (eigenvalues):
## 
##   dim     value      %   cum%   scree plot
```



Kuva 34: Pinottu matriisi - kartta 1

```

## 1      0.056877  77.2  77.2 ****
## 2      0.010116  13.7  91.0 ***
## 3      0.003923   5.3  96.3 *
## 4      0.002711   3.7 100.0 *
##
## -----
## Total: 0.073628 100.0
##
##
## Rows:
##       name  mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | BE   | 82   498   28 | 86 295 11 | -71 203 41 |
## 2 | BG   | 38   907   204 | 599 901 238 | 52 7 10 |
## 3 | DE   | 70   498   29 | -95 298 11 | -78 200 43 |
## 4 | DK   | 57   990   310 | -573 816 328 | 265 175 394 |
## 5 | FI   | 45   987   75 | -225 419 40 | -262 568 309 |
## 6 | HU   | 41   856   168 | 486 778 169 | 153 78 95 |
## 7 | m    | 156  910   20 | 91 873 22 | -19 37 5 |
## 8 | f    | 178  910   17 | -79 873 20 | 16 37 5 |
## 9 | a1   | 39   501   8 | -22 31 0 | -84 470 27 |
## 10 | a2   | 50   958   40 | -236 939 49 | -34 19 6 |
## 11 | a3   | 56   958   7 | -94 958 9 | -1 0 0 |
## 12 | a4   | 63   841   6 | -74 841 6 | -1 0 0 |
## 13 | a5   | 62   868   5 | 71 801 6 | -21 67 3 |
## 14 | a6   | 63   957   83 | 288 852 92 | 101 104 63 |
##
## Columns:
##       name  mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
## 1 | S    | 99   786  147 | 268 661 126 | 117 125 134 |
## 2 | s    | 238  843  172 | 205 787 175 | 55 56 70 |
## 3 |      | 168  640   80 | 150 639 66 | -6 1 1 |
## 4 | e    | 261  970   97 | -35 44 6 | -160 926 657 |
## 5 | E    | 234 1000  504 | -390 962 628 | 77 38 138 |

```

Kartan tulkinta; miten eroaa yhteisvaikutusmuuttujan analyysistä?

Mikä on maapisteiden ja kahden selittävän (eksogeisen) muuttujan pisteiden yhteyks sarakepisteisiin?

Ikäluokkapisteet ovat koko aineiston keskiarvopisteitä, niiden sijantia voi tulkittaa pistejoukko kerrallaan kuten maapisteidenkin. Mitään yhteisvaikutuksia ei analysoida eksplisiittisesti. Karttaa voi verrata sukupuoli-ikäluokka yhteisvaikutusmuuttujan analyysin aiemmin. Naispiste on tiukassa nipussa ikäluokkien a3 ja a4 kanssa aivan origon vasemmalla puolella. Miesten keskiarvopiste on hieman origosta oikealle, yhdessä ikäluokan a5 kanssa.

Koko aineiston kartassa ikäluokkapisteet ja sukupuolipisteet ovat pakkautuneet maapisteitä tiiviimmin origon ympärille. Lisäpisteet on hyvin esitetty, niiden etäisyksiä voi luotettavasti arvioida kuvasta. Poikkeus on nuorin ikäluokka

(a1, qlt = 501). Inertian osuudet (inr) ovat yhtä vaativimattomia kuin Belgian (28) ja Saksan (29), (m = 20, f = 17, a2 = 40, a6 = 83), samoin kontribuutiot akseleiden inertiaan. 1. dimension kontribuutio (cor) on suuri (>800) kaikilla paitsi nuorimmalla ikäryhmällä (a1) jolla 2. dimension selittää lähes puolet sen inertiaasta (470).

Kun Bulgaria ja Unkari jätetään pois, origon lähelle pakkautuneet pistet erottuvat hieman paremmin (osajoukon CA).

edit 17.10.20 Barysentrisen ominaisuudet?

Inertian dekomponointi alimatriiseille.

Tärkeä oletus: reunajakaumat tauluissa samat, ei puuttuvia tietoja muuttujissa.

5.1.1 Matched matrices

Huom! (16.10.20) Jos ja kun ei tehdä analyysiä, ei tarvitse omaa jaksoa. Kannattaa mainita, ehkä vain teoriajaksossa? Idea: matriisien yhdistämisellä saadaan ote monenlaiseen tutkimusongelmaan. Benzecri: data-analyysissä on vain löydettävä oikea matriisi joka diagonalisoidaan-

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 arvointi. Alkuperäinen taulukko jaetaan kahdeksi tauluksi sukupuolen mukaan. Matriisien yhdistäminen (concatenation) riveittäin tai sarakkeittain ei näytä optimaalisesti mm - matriisien eroa.

Ryhmiä välisen ja ryhmien sisäisen inertian erottaminen, **ABBA** on yksi ratkaisu (ABBA matrix, teknisesti block circular matrix).

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

Esimerkinä "Attitudes to women working in 2012".

5.2 MCA - multiple correspondence analysis

MCA on samantyyppisten luokittelusteknikoiden muuttujien välisen yhteyksien analyysiä.

Data Substanssi muuttujien (kysymysten) ja taustamuuttujien (demografiset, koulutus, asuinpaikka) analysissä ydin on substanssimuuttujien välissä suhteissa.

Subsanssimuuttujista valitaan seitsemän kysymystä (naisten rooli työmarkkinoilla) joissa vastausvaihtoehtoja on viisi. Tämä on suositus tai yleinen käytäntö, joka yksinkertaistaa analyysiä.

Taustamuuttujista valitaan kolme: koulutustaso edu ja asuinpaikka urbru ovat (jos tiedokerun erot unohdetaan) taustatietoja. Kolmas muuttujasosta/“Top-Bottom self-placement”) on kysymys mutta ei pohdita tätä enempää.

Lisämuuttujina ovat sukupuoli (sp), maa ja ikä. Ikä luokitellaan kuuteen ryhmään ja luodaan ikaluokan ja sukupuolen yhteisvaikutusmuuttuja ga.

```
# str(ISSP2012jh1d.dat) - luotu skripteissä G1_1_data2.Rmd ja G1_1_data_fct1.Rmd

#Valitaan muuttujat joissa puuttuva tieto on koodattu muuttujan arvoksi

MCAvars1 <- c("Q1am", "Q1bm", "Q1cm", "Q1dm", "Q1em", "Q2am", "Q2bm", "edum",
           "sostam", "urbrum", "maa", "ika", "sp" )

MCAdataljh.dat <- ISSP2012jh1d.dat %>% select(all_of(MCAvars1))
dim(MCAdataljh.dat)

## [1] 32823     13
names(MCAdataljh.dat)

##  [1] "Q1am"    "Q1bm"    "Q1cm"    "Q1dm"    "Q1em"    "Q2am"    "Q2bm"    "edum"
##  [9] "sostam"   "urbrum"   "maa"      "ika"      "sp"
# luodaan ikaluokka-muuttuja ja ikaluokka-sukupuoli - muuttuja
#age_cat
#ikä 1=15-25, 2 =26-35, 3=36-45, 4=46-55, 5=56-65, 6= 66 and older
MCAdataljh.dat <- mutate(MCAdataljh.dat, age_cat = ifelse(ika %in% 15:25, "1",
                                                               ifelse(ika %in% 26:35, "2",
                                                               ifelse(ika %in% 36:45, "3",
                                                               ifelse(ika %in% 46:55, "4",
                                                               ifelse(ika %in% 56:65, "5", "6")))))

# str(MCAdataljh.dat$age_cat)
# uusi (4.2.20)
MCAdataljh.dat <- MCAdataljh.dat %>%
  mutate(age_cat = as_factor(age_cat))
#tarkastuksia - outo järjestys
#levels(MCAdataljh.dat$age_cat)
# str(MCAdataljh.dat$age_cat)
MCAdataljh.dat<- MCAdataljh.dat %>%
  mutate(age_cat = fct_relevel(age_cat,
                                "1",
                                "2",
```

```

    "3",
    "4",
    "5",
    "6"))

```

Tarkistuksia ehkä? (16.10.20)

```

MCAdataljh.dat %>%
  tableX(maa,age_cat,type = "count") #%%>%

```

maa/age_cat	1	2	3	4	5	6	Total
AU	109	134	265	332	353	364	1557
AT	115	228	206	228	182	223	1182
BG	89	123	169	157	210	255	1003
CA	112	32	72	156	260	321	953
HR	136	201	177	198	166	119	997
CZ	195	278	377	361	279	314	1804
DK	210	214	252	273	235	219	1403
FI	160	171	171	230	254	185	1171
FR	149	357	426	451	466	560	2409
HU	106	163	201	172	197	173	1012
IS	224	187	188	201	194	178	1172
IE	26	161	255	266	234	224	1166
LV	168	159	187	216	203	67	1000
LT	151	169	186	220	203	258	1187
NL	60	129	191	250	283	402	1315
NO	165	205	256	309	264	245	1444
PL	158	173	185	178	219	202	1115
RU	222	232	254	253	276	288	1525
SK	75	116	189	254	264	230	1128
SI	97	157	161	178	202	239	1034
SE	80	134	164	211	196	274	1059
CH	137	184	217	265	185	249	1237
BE	216	346	349	400	399	482	2192
DE	208	229	275	370	300	379	1761
PT	115	144	185	174	166	213	997
Total	3483	4626	5558	6303	6190	6663	32823

```

#kable(digits = 2, caption = "Ikäluokka age_cat")

MCAdataljh.dat %>%
  tableX(maa,age_cat,type = "row_perc") #%%>%

```

maa/age_cat	1	2	3	4	5	6	Total
AU	7.00	8.61	17.02	21.32	22.67	23.38	100.00
AT	9.73	19.29	17.43	19.29	15.40	18.87	100.00
BG	8.87	12.26	16.85	15.65	20.94	25.42	100.00
CA	11.75	3.36	7.56	16.37	27.28	33.68	100.00
HR	13.64	20.16	17.75	19.86	16.65	11.94	100.00
CZ	10.81	15.41	20.90	20.01	15.47	17.41	100.00
DK	14.97	15.25	17.96	19.46	16.75	15.61	100.00
FI	13.66	14.60	14.60	19.64	21.69	15.80	100.00
FR	6.19	14.82	17.68	18.72	19.34	23.25	100.00
HU	10.47	16.11	19.86	17.00	19.47	17.09	100.00
IS	19.11	15.96	16.04	17.15	16.55	15.19	100.00
IE	2.23	13.81	21.87	22.81	20.07	19.21	100.00
LV	16.80	15.90	18.70	21.60	20.30	6.70	100.00
LT	12.72	14.24	15.67	18.53	17.10	21.74	100.00
NL	4.56	9.81	14.52	19.01	21.52	30.57	100.00
NO	11.43	14.20	17.73	21.40	18.28	16.97	100.00
PL	14.17	15.52	16.59	15.96	19.64	18.12	100.00
RU	14.56	15.21	16.66	16.59	18.10	18.89	100.00
SK	6.65	10.28	16.76	22.52	23.40	20.39	100.00
SI	9.38	15.18	15.57	17.21	19.54	23.11	100.00
SE	7.55	12.65	15.49	19.92	18.51	25.87	100.00
CH	11.08	14.87	17.54	21.42	14.96	20.13	100.00
BE	9.85	15.78	15.92	18.25	18.20	21.99	100.00
DE	11.81	13.00	15.62	21.01	17.04	21.52	100.00
PT	11.53	14.44	18.56	17.45	16.65	21.36	100.00
All	10.61	14.09	16.93	19.20	18.86	20.30	100.00

```
#kable(digits = 2, caption = "age_cat: suhteelliset frekvenssit")

# Ikäluokka-sukupuoli - muuttuja
MCAdataljh.dat <- mutate(MCAdataljh.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"
    ))
}

#Sosiaalinen status: oma arvio "Top-Bottom self-placement"
str(ISSP2012jh1d.dat$sosta)

## Factor w/ 10 levels "Lowest, Bottom, 01",...: 3 7 8 NA 7 2 7 NA 10 6 ...
## - attr(*, "label")= chr "Top-Bottom self-placement"
#Koulutustaso
str(ISSP2012jh1d.dat$edu)

## Factor w/ 7 levels "No formal education",...: 3 6 6 4 3 NA NA 7 6 7 ...
## - attr(*, "label")= chr "Highest completed degree of education: Categories for internati
#Asuipaikka
str(ISSP2012jh1d.dat$urbru)

## Factor w/ 5 levels "A big city","The suburbs or outskirts of a big city",...: 1 1 1 NA 1
## - attr(*, "label")= chr "Place of living: urban - rural"
# Muunnetaan faktorimuuttujia, mahdollisimman lyhyet tunnisteet kategorioille
MCAdataljh.dat <- MCAdataljh.dat %>%
mutate(E = fct_recode(edum,
    "1" = "No formal education",
    "2" = "Primary school (elementary school)",
    "3" = "Lower secondary (secondary completed does not allow entry to university: obli
    "4" = "Upper secondary (programs that allows entry to university",
    "5" = "Post secondary, non-tertiary (other upper secondary programs toward labour ma
    "6" = "Lower level tertiary, first stage (also technical schools at a tertiary level
    "7" = "Upper level tertiary (Master, Dr.)",
    "P" = "missing"),
    S = fct_recode(sostam,
    "1" = "Lowest, Bottom, 01",
    "2" = "02",
    "3" = "03",
    "4" = "04",
    "5" = "05",
    "6" = "06",
    "7" = "07",
    "8" = "08",
    "9" = "09",
    "10"= "Highest, Top, 10",
    "P" = "missing"),

```

```

U = fct_recode(urbrum,
  "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",
  "P" = "missing")
)

names(MCADATA1jh.dat)

## [1] "Q1am"      "Q1bm"      "Q1cm"      "Q1dm"      "Q1em"      "Q2am"      "Q2bm"
## [8] "edum"       "sostam"    "urbrum"    "maa"        "ika"        "sp"        "age_cat"
## [15] "ga"         "E"          "S"          "U"

dim(MCADATA1jh.dat)

## [1] 32823     18

MCADATA1jh.dat$E %>% levels()

## [1] "1" "2" "3" "4" "5" "6" "7" "P"
MCADATA1jh.dat$S %>% levels()

## [1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "P"
MCADATA1jh.dat$U %>% levels()

## [1] "1" "2" "3" "4" "5" "P"
MCADATA1jh.dat$age_cat %>% levels()

## [1] "1" "2" "3" "4" "5" "6"
str(MCADATA1jh.dat$ga) # toimiikohan - chr-muuttuja? (16.10.20)

## chr [1:32823] "m5" "f5" "f3" "f1" "f6" "m6" "f5" "m5" "f3" "f6" "m1" "m3" ...
MCADATA1jh.dat <- MCADATA1jh.dat %>%
  mutate(gaf = as_factor(ga))

str(MCADATA1jh.dat$gaf)

## Factor w/ 12 levels "m5","f5","f3",...: 1 2 3 4 5 6 2 1 3 5 ...
levels(MCADATA1jh.dat$gaf) # järjestyksellä ei liene väliä? (16.10.20)

## [1] "m5" "f5" "f3" "f1" "f6" "m6" "m1" "m3" "f2" "f4" "m4" "m2"
# gaf ja ga: sama järjestys

```

```

MCAdataljh.dat <- MCAdataljh.dat %>%
  mutate(gaf = fct_relevel(gaf,
    "f1",
    "f2",
    "f3",
    "f4",
    "f5",
    "f6",
    "m1",
    "m2",
    "m3",
    "m4",
    "m5",
    "m6"))

# Lopuksi substanssimuuttutien nimet lyhyiksi

MCAdataljh.dat <- MCAdataljh.dat %>% mutate(a1 = Q1am,
  b1 = Q1bm,
  c1 = Q1cm,
  d1 = Q1dm,
  e1 = Q1em,
  a2 = Q2am,
  b2 = Q2bm)

#Tarkistus

# MCAdataljh.dat %>% tableX (a1, Q1am)
# MCAdataljh.dat %>% tableX (b1, Q1bm)
# MCAdataljh.dat %>% tableX (c1, Q1cm)
# MCAdataljh.dat %>% tableX (d1, Q1dm)
# MCAdataljh.dat %>% tableX (e1, Q1em)
# MCAdataljh.dat %>% tableX (a2, Q2am)
# MCAdataljh.dat %>% tableX (b2, Q2bm)
# MCAdataljh.dat %>% tableX(gaf, ga)

# Perustietoja

MCAdataljh.dat %>% tableX (maa,a1, type = "row_perc")

```

maa/a1	S	s	?	e	E	P	Total
AU	22.16	44.32	10.40	16.89	3.79	2.44	100.00
AT	36.46	34.60	9.39	12.69	3.98	2.88	100.00
BG	13.96	42.37	15.65	20.54	3.59	3.89	100.00
CA	28.75	40.82	9.44	13.96	5.98	1.05	100.00

maa/a1	S	s	?	e	E	P	Total
HR	29.59	41.32	8.22	15.15	5.12	0.60	100.00
CZ	33.09	27.83	17.52	11.97	6.10	3.49	100.00
DK	60.51	26.51	3.42	5.77	3.14	0.64	100.00
FI	39.03	35.87	8.37	10.42	2.13	4.18	100.00
FR	51.39	28.89	6.64	8.14	3.07	1.87	100.00
HU	29.35	31.92	19.17	12.25	5.53	1.78	100.00
IS	41.98	44.62	6.14	6.31	0.77	0.17	100.00
IE	29.50	41.60	9.09	15.18	2.83	1.80	100.00
LV	31.70	34.50	11.10	16.70	5.50	0.50	100.00
LT	8.42	44.48	21.57	19.55	2.11	3.88	100.00
NL	13.54	45.40	14.68	16.43	4.79	5.17	100.00
NO	23.61	47.09	9.56	14.34	1.80	3.60	100.00
PL	17.76	44.04	9.24	22.69	4.57	1.70	100.00
RU	27.02	37.44	15.28	14.10	2.03	4.13	100.00
SK	54.43	24.20	9.04	7.45	2.66	2.22	100.00
SI	41.39	42.17	6.87	6.09	0.87	2.61	100.00
SE	36.45	39.66	11.52	7.55	2.08	2.74	100.00
CH	30.32	47.78	7.68	12.29	1.54	0.40	100.00
BE	33.21	35.90	11.18	10.17	2.83	6.71	100.00
DE	58.83	27.31	2.56	8.01	2.10	1.19	100.00
PT	24.47	50.75	7.22	14.84	2.01	0.70	100.00
All	33.87	37.63	10.30	12.41	3.20	2.58	100.00

```
MCAdata1jh.dat %>% tableX (maa,b1, type = "row_perc")
```

maa/b1	S	s	?	e	E	P	Total
AU	5.01	25.18	17.85	35.45	13.23	3.28	100.00
AT	18.44	37.82	14.47	17.34	8.29	3.64	100.00
BG	11.76	39.38	20.44	18.94	1.30	8.18	100.00
CA	5.35	22.14	18.78	32.21	20.15	1.36	100.00
HR	7.52	26.48	19.06	32.60	13.34	1.00	100.00
CZ	9.65	21.73	22.34	23.00	19.68	3.60	100.00
DK	4.99	16.96	10.83	16.54	49.61	1.07	100.00
FI	4.01	16.05	12.72	36.12	25.88	5.21	100.00
FR	10.63	22.87	17.60	19.47	25.90	3.53	100.00
HU	21.64	28.46	22.23	18.77	7.41	1.48	100.00
IS	1.11	11.77	15.87	47.10	23.12	1.02	100.00
IE	4.37	20.58	16.38	39.71	16.21	2.74	100.00
LV	18.80	39.50	15.60	20.90	3.80	1.40	100.00
LT	4.21	36.90	33.36	18.53	1.85	5.14	100.00
NL	4.49	22.51	18.40	33.84	14.90	5.86	100.00
NO	1.59	12.88	15.65	40.10	25.28	4.50	100.00
PL	9.87	35.43	13.90	32.74	5.74	2.33	100.00

maa/b1	S	s	?	e	E	P	Total
RU	16.00	35.54	23.61	16.66	2.75	5.44	100.00
SK	10.37	21.81	20.30	26.42	17.55	3.55	100.00
SI	3.77	26.31	19.34	35.30	12.67	2.61	100.00
SE	2.74	11.71	20.68	26.06	33.33	5.48	100.00
CH	7.19	34.84	17.95	29.51	9.05	1.46	100.00
BE	8.71	20.57	19.98	25.18	17.38	8.17	100.00
DE	9.37	21.29	11.24	30.55	24.87	2.67	100.00
PT	7.32	49.55	15.55	21.46	5.22	0.90	100.00
All	8.37	25.56	18.12	27.43	16.90	3.62	100.00

```
MCAdata1jh.dat %>% tableX (maa,c1, type = "row_perc")
```

maa/c1	S	s	?	e	E	P	Total
AU	6.55	28.45	16.76	29.54	15.74	2.95	100.00
AT	17.26	36.38	15.57	18.02	10.07	2.71	100.00
BG	8.37	30.21	24.73	26.32	5.28	5.08	100.00
CA	4.83	21.09	15.01	34.00	23.40	1.68	100.00
HR	8.32	24.97	18.05	33.30	14.34	1.00	100.00
CZ	9.37	21.01	25.83	21.29	18.90	3.60	100.00
DK	5.49	12.54	8.48	14.68	58.16	0.64	100.00
FI	2.65	10.25	11.61	33.30	37.06	5.12	100.00
FR	11.58	21.59	17.39	21.00	25.16	3.28	100.00
HU	17.59	27.57	25.89	18.68	8.79	1.48	100.00
IS	2.22	14.33	17.15	40.78	25.09	0.43	100.00
IE	6.69	26.67	13.04	32.68	17.92	3.00	100.00
LV	18.50	33.00	19.50	22.50	5.20	1.30	100.00
LT	3.37	34.04	32.35	23.08	2.70	4.47	100.00
NL	5.55	27.45	19.01	28.37	14.75	4.87	100.00
NO	2.15	17.38	18.35	36.29	20.84	4.99	100.00
PL	7.17	28.43	13.81	39.28	8.43	2.87	100.00
RU	17.84	36.39	22.16	16.92	3.02	3.67	100.00
SK	12.85	24.91	22.25	24.02	14.27	1.68	100.00
SI	5.03	32.88	21.18	27.95	10.06	2.90	100.00
SE	2.83	13.22	16.62	27.20	35.79	4.34	100.00
CH	10.35	36.62	16.65	27.08	8.08	1.21	100.00
BE	9.08	24.04	17.66	23.68	17.47	8.07	100.00
DE	10.39	20.73	12.10	28.00	26.12	2.67	100.00
PT	6.32	36.11	19.16	29.29	8.22	0.90	100.00
All	8.65	25.17	18.28	26.52	18.16	3.22	100.00

```
MCAdataljh.dat %>% tableX (maa,d1, type = "row_perc")
```

maa/d1	S	s	?	e	E	P	Total
AU	5.59	20.81	23.96	30.44	14.96	4.24	100.00
AT	7.87	21.24	20.73	21.74	19.37	9.05	100.00
BG	8.18	39.28	28.32	14.06	1.69	8.47	100.00
CA	3.57	14.80	27.39	30.64	20.25	3.36	100.00
HR	11.13	28.49	24.67	24.47	9.43	1.81	100.00
CZ	16.41	29.43	30.27	12.36	5.76	5.76	100.00
DK	4.92	11.33	14.68	21.74	42.34	4.99	100.00
FI	4.95	22.29	20.67	24.94	14.35	12.81	100.00
FR	11.50	22.58	20.22	18.35	21.75	5.60	100.00
HU	20.95	32.81	30.34	9.49	3.85	2.57	100.00
IS	2.82	20.65	21.50	32.17	20.73	2.13	100.00
IE	4.97	20.24	20.41	30.53	19.55	4.29	100.00
LV	16.10	32.40	24.80	19.50	3.60	3.60	100.00
LT	3.62	23.76	32.52	22.83	3.03	14.24	100.00
NL	1.29	13.38	17.11	34.07	26.24	7.91	100.00
NO	2.15	12.47	18.63	35.04	22.99	8.73	100.00
PL	9.51	29.69	18.30	31.75	4.75	6.01	100.00
RU	14.30	29.84	26.95	18.82	3.87	6.23	100.00
SK	23.32	36.97	27.30	7.09	1.42	3.90	100.00
SI	6.67	31.24	22.53	24.27	9.48	5.80	100.00
SE	3.49	14.16	23.80	20.02	27.29	11.24	100.00
CH	6.95	27.41	23.44	32.26	8.33	1.62	100.00
BE	9.35	18.11	20.99	23.81	17.61	10.13	100.00
DE	5.91	14.88	12.32	32.65	28.79	5.45	100.00
PT	6.82	33.80	21.36	26.48	9.03	2.51	100.00
All	8.59	23.37	22.55	23.96	15.28	6.25	100.00

```
MCAdataljh.dat %>% tableX (maa,e1, type = "row_perc")
```

maa/e1	S	s	?	e	E	P	Total
AU	12.46	35.32	25.11	16.96	5.33	4.82	100.00
AT	12.61	20.39	21.83	18.10	16.07	11.00	100.00
BG	10.67	36.89	25.62	16.95	2.39	7.48	100.00
CA	11.54	30.95	26.23	17.63	8.29	5.35	100.00
HR	9.73	24.57	18.15	27.38	16.55	3.61	100.00
CZ	9.20	18.63	31.43	19.90	12.69	8.15	100.00
DK	11.69	17.03	19.67	19.53	22.31	9.76	100.00
FI	10.25	22.80	20.67	23.06	10.08	13.15	100.00
FR	10.75	17.10	23.41	22.62	18.76	7.35	100.00
HU	15.51	25.99	29.94	19.86	6.13	2.57	100.00

maa/e1	S	s	?	e	E	P	Total
IS	8.36	32.08	30.97	21.76	4.35	2.47	100.00
IE	12.78	31.56	22.13	22.04	7.03	4.46	100.00
LV	13.90	30.60	24.60	21.30	4.50	5.10	100.00
LT	4.55	22.07	34.71	18.96	2.78	16.93	100.00
NL	3.50	21.67	25.48	26.16	13.84	9.35	100.00
NO	3.53	17.24	23.89	31.09	13.64	10.60	100.00
PL	10.13	34.08	20.18	27.62	4.13	3.86	100.00
RU	13.57	30.89	26.82	16.39	3.41	8.92	100.00
SK	16.67	23.32	25.09	21.28	8.78	4.88	100.00
SI	5.51	29.69	19.83	29.21	9.19	6.58	100.00
SE	5.67	15.20	32.58	18.51	11.14	16.90	100.00
CH	10.75	46.00	16.98	20.21	3.88	2.18	100.00
BE	12.59	23.31	22.35	19.57	10.58	11.59	100.00
DE	10.85	19.42	13.52	27.88	20.39	7.95	100.00
PT	7.22	27.48	18.86	32.10	10.83	3.51	100.00
All	10.23	25.42	23.89	22.14	10.55	7.78	100.00

MCAdata1jh.dat %>% **tableX** (maa,a2)

maa/a2	S	s	?	e	E	P	Total
AU	216	574	487	205	31	44	1557
AT	322	492	195	107	22	44	1182
BG	431	491	50	11	7	13	1003
CA	150	376	261	122	27	17	953
HR	409	451	93	34	6	4	997
CZ	1012	547	165	36	18	26	1804
DK	827	291	195	34	48	8	1403
FI	388	507	169	71	10	26	1171
FR	1218	711	321	76	34	49	2409
HU	340	398	206	47	10	11	1012
IS	356	599	159	47	8	3	1172
IE	287	391	276	157	19	36	1166
LV	344	429	153	59	8	7	1000
LT	162	703	233	54	4	31	1187
NL	218	556	333	115	35	58	1315
NO	422	775	184	31	6	26	1444
PL	249	593	135	117	13	8	1115
RU	392	706	275	86	9	57	1525
SK	530	383	165	32	9	9	1128
SI	432	506	69	11	3	13	1034
SE	495	408	112	21	5	18	1059
CH	202	614	231	172	8	10	1237
BE	812	758	353	134	30	105	2192

maa/a2	S	s	?	e	E	P	Total
DE	631	729	177	134	31	59	1761
PT	460	476	42	16	2	1	997
Total	11305	13464	5039	1929	403	683	32823

MCAdat1jh.dat %>% **tableX** (maa,b2)

maa/b2	S	s	?	e	E	P	Total
AU	47	202	301	564	399	44	1557
AT	118	267	268	248	248	33	1182
BG	129	253	279	245	68	29	1003
CA	21	95	154	354	319	10	953
HR	56	142	186	389	215	9	997
CZ	332	506	432	322	178	34	1804
DK	33	63	124	162	1017	4	1403
FI	25	78	170	437	416	45	1171
FR	99	196	326	477	1267	44	2409
HU	174	266	325	170	65	12	1012
IS	10	66	110	505	479	2	1172
IE	42	100	159	479	357	29	1166
LV	233	285	236	196	33	17	1000
LT	113	281	498	226	24	45	1187
NL	39	127	242	462	385	60	1315
NO	21	52	160	565	609	37	1444
PL	176	321	177	353	78	10	1115
RU	335	469	404	231	26	60	1525
SK	258	348	300	153	62	7	1128
SI	28	185	187	372	245	17	1034
SE	18	42	123	271	575	30	1059
CH	64	242	196	461	268	6	1237
BE	152	235	367	597	743	98	2192
DE	122	167	194	611	630	37	1761
PT	59	176	191	360	211	0	997
Total	2704	5164	6109	9210	8917	719	32823

MCAdat1jh.dat %>% **tableX**(gaf, ga)

gaf/ga	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
f1	1867	0	0	0	0	0	0	0	0	0	0	0	1867
f2	0	2652	0	0	0	0	0	0	0	0	0	0	2652
f3	0	0	3180	0	0	0	0	0	0	0	0	0	3180
f4	0	0	0	3508	0	0	0	0	0	0	0	0	3508

gaf/ga	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
f5	0	0	0	0	3270	0	0	0	0	0	0	0	3270
f6	0	0	0	0	0	3557	0	0	0	0	0	0	3557
m1	0	0	0	0	0	0	1616	0	0	0	0	0	1616
m2	0	0	0	0	0	0	0	1974	0	0	0	0	1974
m3	0	0	0	0	0	0	0	0	2378	0	0	0	2378
m4	0	0	0	0	0	0	0	0	0	2795	0	0	2795
m5	0	0	0	0	0	0	0	0	0	0	2920	0	2920
m6	0	0	0	0	0	0	0	0	0	0	0	3106	3106
Total	1867	2652	3180	3508	3270	3557	1616	1974	2378	2795	2920	3106	32823

MCAdata1jh.dat %>% **tableX**(maa, age_cat)

maa/age_cat	1	2	3	4	5	6	Total
AU	109	134	265	332	353	364	1557
AT	115	228	206	228	182	223	1182
BG	89	123	169	157	210	255	1003
CA	112	32	72	156	260	321	953
HR	136	201	177	198	166	119	997
CZ	195	278	377	361	279	314	1804
DK	210	214	252	273	235	219	1403
FI	160	171	171	230	254	185	1171
FR	149	357	426	451	466	560	2409
HU	106	163	201	172	197	173	1012
IS	224	187	188	201	194	178	1172
IE	26	161	255	266	234	224	1166
LV	168	159	187	216	203	67	1000
LT	151	169	186	220	203	258	1187
NL	60	129	191	250	283	402	1315
NO	165	205	256	309	264	245	1444
PL	158	173	185	178	219	202	1115
RU	222	232	254	253	276	288	1525
SK	75	116	189	254	264	230	1128
SI	97	157	161	178	202	239	1034
SE	80	134	164	211	196	274	1059
CH	137	184	217	265	185	249	1237
BE	216	346	349	400	399	482	2192
DE	208	229	275	370	300	379	1761
PT	115	144	185	174	166	213	997
Total	3483	4626	5558	6303	6190	6663	32823

MCAdata1jh.dat %>% `tableX`(maa, gaf)

maa/gaf	f1	f2	f3	f4	f5	f6	m1	m2	m3	m4	m5	m6	Total
AU	70	87	163	192	173	181	39	47	102	140	180	183	1557
AT	74	130	104	124	99	114	41	98	102	104	83	109	1182
BG	46	68	99	88	117	163	43	55	70	69	93	92	1003
CA	74	17	34	76	85	93	38	15	38	80	175	228	953
HR	63	102	88	111	95	75	73	99	89	87	71	44	997
CZ	111	163	212	190	146	175	84	115	165	171	133	139	1804
DK	83	110	140	148	129	100	127	104	112	125	106	119	1403
FI	94	97	96	121	147	102	66	74	75	109	107	83	1171
FR	94	242	299	296	294	333	55	115	127	155	172	227	2409
HU	55	86	97	91	94	106	51	77	104	81	103	67	1012
IS	100	103	105	93	87	78	124	84	83	108	107	100	1172
IE	18	118	191	191	119	117	8	43	64	75	115	107	1166
LV	85	98	109	123	124	45	83	61	78	93	79	22	1000
LT	78	99	110	128	108	171	73	70	76	92	95	87	1187
NL	41	77	112	157	135	183	19	52	79	93	148	219	1315
NO	107	104	140	156	135	112	58	101	116	153	129	133	1444
PL	63	93	102	98	124	122	95	80	83	80	95	80	1115
RU	121	140	169	159	183	206	101	92	85	94	93	82	1525
SK	42	69	94	127	144	129	33	47	95	127	120	101	1128
SI	46	84	84	94	114	136	51	73	77	84	88	103	1034
SE	52	82	88	117	98	137	28	52	76	94	98	137	1059
CH	74	84	114	129	89	127	63	100	103	136	96	122	1237
BE	118	205	181	207	196	230	98	141	168	193	203	252	2192
DE	103	123	152	191	142	193	105	106	123	179	158	186	1761
PT	55	71	97	101	93	129	60	73	88	73	73	84	997
Total	1867	2652	3180	3508	3270	3557	1616	1974	2378	2795	2920	3106	32823

MCAdata1jh.dat %>% `tableX`(maa, S)

maa/S	1	2	3	4	5	6	7	8	9	10	P	Total
AU	21	19	35	62	263	304	336	265	52	43	157	1557
AT	4	7	31	81	328	333	219	117	27	35	0	1182
BG	50	94	237	219	260	93	31	12	4	1	2	1003
CA	12	6	23	36	105	168	222	196	41	34	110	953
HR	15	26	77	103	344	184	130	64	11	7	36	997
CZ	22	54	162	294	530	277	222	125	24	7	87	1804
DK	8	7	38	52	208	295	379	259	42	37	78	1403
FI	13	17	36	78	159	241	315	226	40	16	30	1171
FR	44	52	225	293	577	463	310	121	23	16	285	2409
HU	35	110	195	228	213	114	67	38	5	1	6	1012

maa/S	1	2	3	4	5	6	7	8	9	10	P	Total
IS	10	15	28	62	245	261	225	116	13	14	183	1172
IE	21	14	37	48	113	299	239	188	70	55	82	1166
LV	23	32	116	187	265	189	119	40	9	2	18	1000
LT	17	59	128	195	258	215	175	96	15	4	25	1187
NL	25	22	59	114	172	259	359	185	47	18	55	1315
NO	17	18	36	82	279	377	330	194	41	15	55	1444
PL	13	37	81	131	302	289	145	85	16	16	0	1115
RU	90	117	234	246	272	393	100	50	13	8	2	1525
SK	9	30	92	193	297	256	165	78	4	4	0	1128
SI	6	11	46	102	339	238	143	67	17	12	53	1034
SE	10	6	36	57	213	277	254	118	9	25	54	1059
CH	4	11	41	100	255	246	261	225	36	15	43	1237
BE	71	39	78	123	343	449	520	279	37	31	222	2192
DE	8	21	53	103	188	529	441	308	55	17	38	1761
PT	14	42	97	157	270	140	71	25	16	9	156	997
Total	562	866	2221	3346	6798	6889	5778	3477	667	442	1777	32823

MCAdata1jh.dat %>% **tableX**(maa, U)

maa/U	1	2	3	4	5	P	Total
AU	411	498	277	141	185	45	1557
AT	421	88	316	324	33	0	1182
BG	463	30	150	359	1	0	1003
CA	299	195	347	49	57	6	953
HR	256	150	325	266	0	0	997
CZ	645	82	639	434	3	1	1804
DK	387	314	391	203	102	6	1403
FI	97	406	289	219	140	20	1171
FR	405	382	750	722	136	14	2409
HU	351	28	313	320	0	0	1012
IS	365	372	264	70	55	46	1172
IE	157	275	336	164	230	4	1166
LV	417	61	284	193	45	0	1000
LT	434	5	407	336	5	0	1187
NL	242	87	428	502	36	20	1315
NO	342	181	371	304	240	6	1444
PL	291	61	346	411	5	1	1115
RU	750	18	372	385	0	0	1525
SK	99	30	417	571	11	0	1128
SI	147	101	199	240	344	3	1034
SE	259	191	295	198	108	8	1059
CH	106	121	300	658	51	1	1237
BE	505	284	468	792	84	59	2192

maa/U	1	2	3	4	5	P	Total
DE	374	186	595	578	28	0	1761
PT	219	240	324	207	3	4	997
Total	8442	4386	9203	8646	1902	244	32823

MCAdat1jh.dat %>% **tableX**(maa, E)

maa/E	1	2	3	4	5	6	7	P	Total
AU	5	42	348	237	133	474	239	79	1557
AT	0	0	824	92	104	0	162	0	1182
BG	14	58	193	236	264	43	195	0	1003
CA	3	23	66	136	228	365	128	4	953
HR	32	9	275	481	73	123	0	4	997
CZ	5	0	689	876	26	31	150	27	1804
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
HU	8	25	476	281	55	120	46	1	1012
IS	10	23	323	109	232	262	161	52	1172
IE	8	8	207	252	246	177	267	1	1166
LV	3	7	157	270	323	0	240	0	1000
LT	5	37	283	184	440	204	28	6	1187
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
RU	60	0	141	258	682	384	0	0	1525
SK	4	10	501	418	24	23	148	0	1128
SI	14	49	392	317	70	176	15	1	1034
SE	7	99	281	208	0	159	279	26	1059
CH	1	26	224	55	590	187	152	2	1237
BE	76	187	430	545	153	418	348	35	2192
DE	0	18	174	74	1015	152	325	3	1761
PT	44	367	195	233	10	74	73	1	997
Total	491	1531	7811	7115	5658	5147	4762	308	32823

MCAdat1jh.dat %>% **tableX**(maa, S, type = "row_perc")

maa/S	1	2	3	4	5	6	7	8	9	10	P	Total
AU	1.35	1.22	2.25	3.98	16.89	19.52	21.58	17.02	3.34	2.76	10.08	100.00
AT	0.34	0.59	2.62	6.85	27.75	28.17	18.53	9.90	2.28	2.96	0.00	100.00
BG	4.99	9.37	23.63	21.83	25.92	9.27	3.09	1.20	0.40	0.10	0.20	100.00
CA	1.26	0.63	2.41	3.78	11.02	17.63	23.29	20.57	4.30	3.57	11.54	100.00

maa/S	1	2	3	4	5	6	7	8	9	10	P	Total
HR	1.50	2.61	7.72	10.33	34.50	18.46	13.04	6.42	1.10	0.70	3.61	100.00
CZ	1.22	2.99	8.98	16.30	29.38	15.35	12.31	6.93	1.33	0.39	4.82	100.00
DK	0.57	0.50	2.71	3.71	14.83	21.03	27.01	18.46	2.99	2.64	5.56	100.00
FI	1.11	1.45	3.07	6.66	13.58	20.58	26.90	19.30	3.42	1.37	2.56	100.00
FR	1.83	2.16	9.34	12.16	23.95	19.22	12.87	5.02	0.95	0.66	11.83	100.00
HU	3.46	10.87	19.27	22.53	21.05	11.26	6.62	3.75	0.49	0.10	0.59	100.00
IS	0.85	1.28	2.39	5.29	20.90	22.27	19.20	9.90	1.11	1.19	15.61	100.00
IE	1.80	1.20	3.17	4.12	9.69	25.64	20.50	16.12	6.00	4.72	7.03	100.00
LV	2.30	3.20	11.60	18.70	26.50	18.90	11.90	4.00	0.90	0.20	1.80	100.00
LT	1.43	4.97	10.78	16.43	21.74	18.11	14.74	8.09	1.26	0.34	2.11	100.00
NL	1.90	1.67	4.49	8.67	13.08	19.70	27.30	14.07	3.57	1.37	4.18	100.00
NO	1.18	1.25	2.49	5.68	19.32	26.11	22.85	13.43	2.84	1.04	3.81	100.00
PL	1.17	3.32	7.26	11.75	27.09	25.92	13.00	7.62	1.43	1.43	0.00	100.00
RU	5.90	7.67	15.34	16.13	17.84	25.77	6.56	3.28	0.85	0.52	0.13	100.00
SK	0.80	2.66	8.16	17.11	26.33	22.70	14.63	6.91	0.35	0.35	0.00	100.00
SI	0.58	1.06	4.45	9.86	32.79	23.02	13.83	6.48	1.64	1.16	5.13	100.00
SE	0.94	0.57	3.40	5.38	20.11	26.16	23.98	11.14	0.85	2.36	5.10	100.00
CH	0.32	0.89	3.31	8.08	20.61	19.89	21.10	18.19	2.91	1.21	3.48	100.00
BE	3.24	1.78	3.56	5.61	15.65	20.48	23.72	12.73	1.69	1.41	10.13	100.00
DE	0.45	1.19	3.01	5.85	10.68	30.04	25.04	17.49	3.12	0.97	2.16	100.00
PT	1.40	4.21	9.73	15.75	27.08	14.04	7.12	2.51	1.60	0.90	15.65	100.00
All	1.71	2.64	6.77	10.19	20.71	20.99	17.60	10.59	2.03	1.35	5.41	100.00

```
MCAdata1jh.dat %>% tableX(maa, U, type = "row_perc")
```

maa/U	1	2	3	4	5	P	Total
AU	26.40	31.98	17.79	9.06	11.88	2.89	100.00
AT	35.62	7.45	26.73	27.41	2.79	0.00	100.00
BG	46.16	2.99	14.96	35.79	0.10	0.00	100.00
CA	31.37	20.46	36.41	5.14	5.98	0.63	100.00
HR	25.68	15.05	32.60	26.68	0.00	0.00	100.00
CZ	35.75	4.55	35.42	24.06	0.17	0.06	100.00
DK	27.58	22.38	27.87	14.47	7.27	0.43	100.00
FI	8.28	34.67	24.68	18.70	11.96	1.71	100.00
FR	16.81	15.86	31.13	29.97	5.65	0.58	100.00
HU	34.68	2.77	30.93	31.62	0.00	0.00	100.00
IS	31.14	31.74	22.53	5.97	4.69	3.92	100.00
IE	13.46	23.58	28.82	14.07	19.73	0.34	100.00
LV	41.70	6.10	28.40	19.30	4.50	0.00	100.00
LT	36.56	0.42	34.29	28.31	0.42	0.00	100.00
NL	18.40	6.62	32.55	38.17	2.74	1.52	100.00
NO	23.68	12.53	25.69	21.05	16.62	0.42	100.00
PL	26.10	5.47	31.03	36.86	0.45	0.09	100.00

maa/U	1	2	3	4	5	P	Total
RU	49.18	1.18	24.39	25.25	0.00	0.00	100.00
SK	8.78	2.66	36.97	50.62	0.98	0.00	100.00
SI	14.22	9.77	19.25	23.21	33.27	0.29	100.00
SE	24.46	18.04	27.86	18.70	10.20	0.76	100.00
CH	8.57	9.78	24.25	53.19	4.12	0.08	100.00
BE	23.04	12.96	21.35	36.13	3.83	2.69	100.00
DE	21.24	10.56	33.79	32.82	1.59	0.00	100.00
PT	21.97	24.07	32.50	20.76	0.30	0.40	100.00
All	25.72	13.36	28.04	26.34	5.79	0.74	100.00

```
MCAdata1jh.dat %>% tableX(maa, E, type = "row_perc")
```

maa/E	1	2	3	4	5	6	7	P	Total
AU	0.32	2.70	22.35	15.22	8.54	30.44	15.35	5.07	100.00
AT	0.00	0.00	69.71	7.78	8.80	0.00	13.71	0.00	100.00
BG	1.40	5.78	19.24	23.53	26.32	4.29	19.44	0.00	100.00
CA	0.31	2.41	6.93	14.27	23.92	38.30	13.43	0.42	100.00
HR	3.21	0.90	27.58	48.24	7.32	12.34	0.00	0.40	100.00
CZ	0.28	0.00	38.19	48.56	1.44	1.72	8.31	1.50	100.00
DK	2.85	1.14	5.27	6.27	26.30	40.06	18.10	0.00	100.00
FI	0.00	9.99	6.83	31.43	22.72	16.48	11.96	0.60	100.00
FR	4.77	9.71	30.47	14.78	0.00	20.71	18.22	1.33	100.00
HU	0.79	2.47	47.04	27.77	5.43	11.86	4.55	0.10	100.00
IS	0.85	1.96	27.56	9.30	19.80	22.35	13.74	4.44	100.00
IE	0.69	0.69	17.75	21.61	21.10	15.18	22.90	0.09	100.00
LV	0.30	0.70	15.70	27.00	32.30	0.00	24.00	0.00	100.00
LT	0.42	3.12	23.84	15.50	37.07	17.19	2.36	0.51	100.00
NL	0.84	2.28	25.63	11.41	18.63	21.52	18.25	1.44	100.00
NO	0.97	0.00	23.27	19.18	3.88	12.81	39.34	0.55	100.00
PL	1.08	13.09	6.37	55.07	4.84	4.75	14.80	0.00	100.00
RU	3.93	0.00	9.25	16.92	44.72	25.18	0.00	0.00	100.00
SK	0.35	0.89	44.41	37.06	2.13	2.04	13.12	0.00	100.00
SI	1.35	4.74	37.91	30.66	6.77	17.02	1.45	0.10	100.00
SE	0.66	9.35	26.53	19.64	0.00	15.01	26.35	2.46	100.00
CH	0.08	2.10	18.11	4.45	47.70	15.12	12.29	0.16	100.00
BE	3.47	8.53	19.62	24.86	6.98	19.07	15.88	1.60	100.00
DE	0.00	1.02	9.88	4.20	57.64	8.63	18.46	0.17	100.00
PT	4.41	36.81	19.56	23.37	1.00	7.42	7.32	0.10	100.00
All	1.50	4.66	23.80	21.68	17.24	15.68	14.51	0.94	100.00

```
MCAdataljh.dat %>% tableX(maa, age_cat, type = "row_perc")
```

maa/age_cat	1	2	3	4	5	6	Total
AU	7.00	8.61	17.02	21.32	22.67	23.38	100.00
AT	9.73	19.29	17.43	19.29	15.40	18.87	100.00
BG	8.87	12.26	16.85	15.65	20.94	25.42	100.00
CA	11.75	3.36	7.56	16.37	27.28	33.68	100.00
HR	13.64	20.16	17.75	19.86	16.65	11.94	100.00
CZ	10.81	15.41	20.90	20.01	15.47	17.41	100.00
DK	14.97	15.25	17.96	19.46	16.75	15.61	100.00
FI	13.66	14.60	14.60	19.64	21.69	15.80	100.00
FR	6.19	14.82	17.68	18.72	19.34	23.25	100.00
HU	10.47	16.11	19.86	17.00	19.47	17.09	100.00
IS	19.11	15.96	16.04	17.15	16.55	15.19	100.00
IE	2.23	13.81	21.87	22.81	20.07	19.21	100.00
LV	16.80	15.90	18.70	21.60	20.30	6.70	100.00
LT	12.72	14.24	15.67	18.53	17.10	21.74	100.00
NL	4.56	9.81	14.52	19.01	21.52	30.57	100.00
NO	11.43	14.20	17.73	21.40	18.28	16.97	100.00
PL	14.17	15.52	16.59	15.96	19.64	18.12	100.00
RU	14.56	15.21	16.66	16.59	18.10	18.89	100.00
SK	6.65	10.28	16.76	22.52	23.40	20.39	100.00
SI	9.38	15.18	15.57	17.21	19.54	23.11	100.00
SE	7.55	12.65	15.49	19.92	18.51	25.87	100.00
CH	11.08	14.87	17.54	21.42	14.96	20.13	100.00
BE	9.85	15.78	15.92	18.25	18.20	21.99	100.00
DE	11.81	13.00	15.62	21.01	17.04	21.52	100.00
PT	11.53	14.44	18.56	17.45	16.65	21.36	100.00
All	10.61	14.09	16.93	19.20	18.86	20.30	100.00

Taustamuuttujien taulukoissa on yllättävä isoja eroja, jotkut taulukoiden luvat ovat nollia tai hyvin vähän havaintoja. Luokkia pitäisi ehkä yhdistellä, jo pelkästään “kuvaroskan” takia. Ei tehdä.

Puuttuneisuuden yleiskuva

```
# Puuttuvien tietojen yleiskuva

# Puuttuvat tiedot aineistossa - viite datan dokumentointiin jossa taulukot.
# Vaihtelee maittain ja muuttujittain, paljon.

# Koko data (G1_1_data2.Rmd - skriptissä valitut muuttujat ja 25 maata)
#
#sum(!complete.cases(ISSP2012jh1d.dat)) = 9455
#dim(ISSP2012jh1d.dat) = 32823
```

```

#9455/32823 = 0.2880602

# Puuttuvat tiedot valitussa MCA-aineistossa

#missingMCAvars1 <- c("Q1a", "Q1b", "Q1c", "Q1d", "Q1e", "Q2a", "Q2b", "edu",
#                         "sosta", "urbru", "maa", "ika", "sp" )
#missingTestMCA1.dat <- ISSP2012jh1d.dat %>% select(all_of(missingMCAvars1))

#sum(!complete.cases(missingTestMCA1.dat)) = 6101
#dim(missingTestMCA1.dat) = 32823
#6101/32823 = 0.1858758 Puuttellisten havaintojen osuus.

```

Koko tähän tutkimukseen valitussa aineistossa (25 maata ja muuttujat, poistettu havainnot joissa ikä tai sukupuoli puuttuu) 71% havainnoista on kaikki tiedot.

MCA-analyyseihin valitun $7 + 3 = 10$ muuttujan aineiston havainnoista 81% on vailla puuttuvia tietoja. Jos puuttuvat tiedot poistetaan (ns. “listwise delete” poistetaan jos yksi tai useampi tieto puuttuu) viidesosa datasta jää pois. Kannattaako puuttuvia tietoja hieman analysoida?

Datassa edellisen luvun ikäluokka-sukupuoli - muuttuja.

edit Tässä keskityttävä data-analyysin **tutkimusongelmiin**, johdantoa MCA-lukuun.

5.3 MCA

```

# Ensimmäiset MCA-kartat - viiden vastausvaihtoehdon kysymykset ja puuttuvat tiedot

# glimpse((MCAdataljh.dat))
mcaDat11jh.dat <- MCAdataljh.dat %>% select(a1,b1,c1, d1, e1,a2,b2)
glimpse(mcaDat11jh.dat)

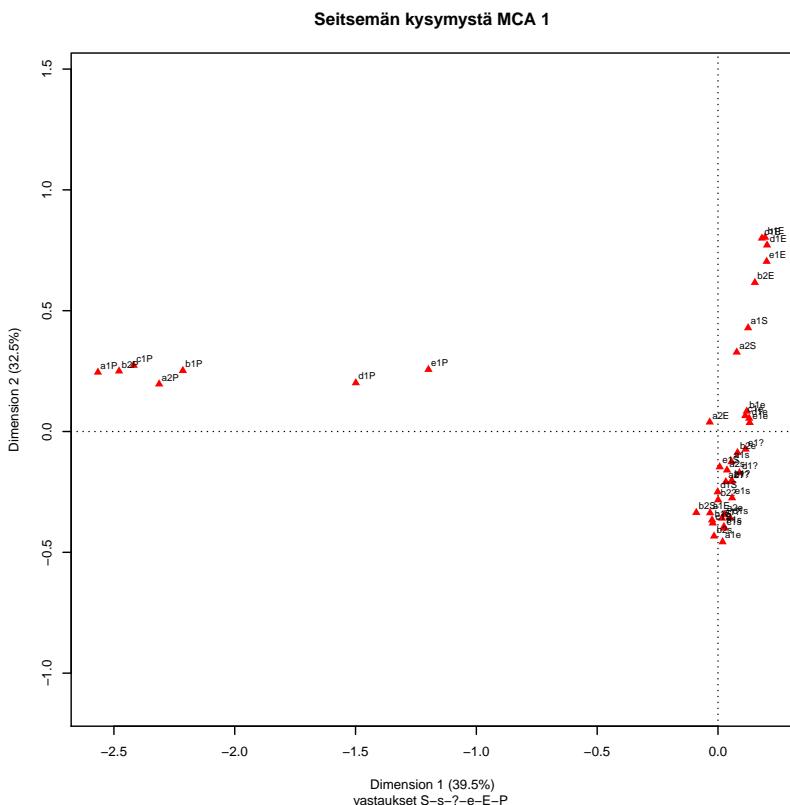
## Rows: 32,823
## Columns: 7
## $ a1 <fct> E, S, s, S, P, s, e, s, s, e, s, s, s, S, P, s, ?, s, s, s, S...
## $ b1 <fct> S, E, e, e, P, e, ?, e, ?, ?, E, e, e, ?, ?, e, ?, P, e, e, e...
## $ c1 <fct> ?, E, s, e, e, P, e, s, e, s, ?, e, e, ?, s, e, s, ?, e, e, ?
## $ d1 <fct> ?, E, E, s, e, P, e, E, e, E, ?, e, e, E, ?, s, ?, e, s, P, e, ?
## $ e1 <fct> ?, S, s, ?, e, P, s, e, e, S, e, s, s, e, s, s, S, s, s, s, e, s...
## $ a2 <fct> S, ?, e, s, s, P, s, E, s, S, s, s, S, s, s, ?, s, s, s, S...
## $ b2 <fct> ?, E, e, e, P, s, E, e, E, e, E, e, ?, E, s, e, e, ?

Qmuuttujat1.mca <- mjca(mcaDat11jh.dat, ps="")

# ps="" muuttujan ja sen kategorian eroitinmerkki
par(cex=0.6)

```

```
plot.mjca(Qmuuttujat1.mca,
           main = "Seitsemän kysymystä MCA 1",
           sub = "vastaukset S-s-?-e-E-P ")
```



Kuva 35: MCA-kartta: viiden vastausvaihtoehdon kysymykset

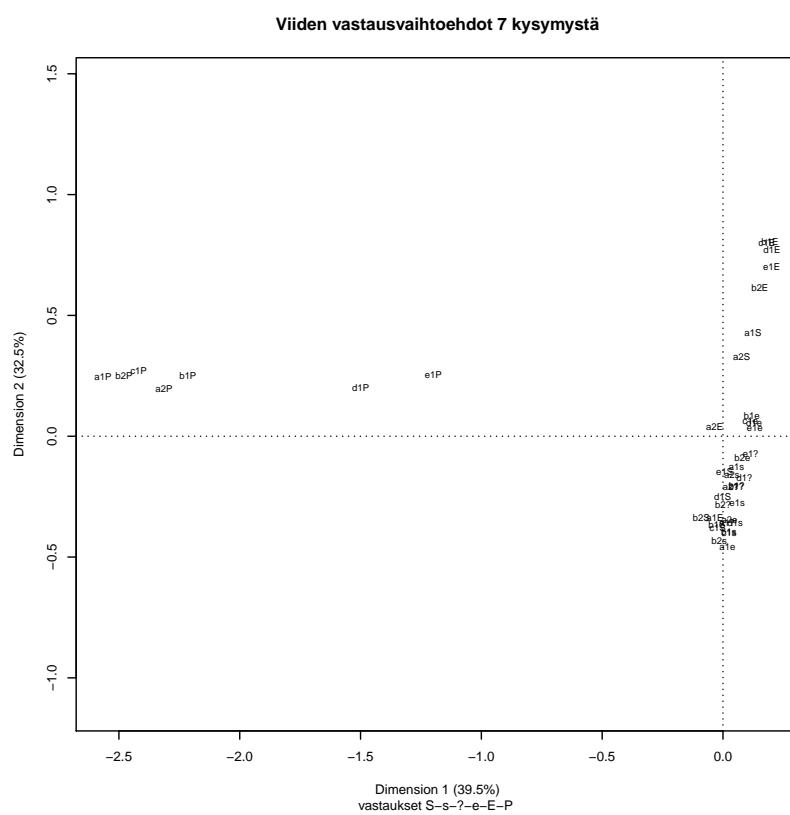
```

plot.mjca(Qmuuttujat1.mca, labels = c(2,1),
           main = "Viiden vastausvaihtoehdot 7 kysymystä",
           sub = "vastaukset S-s-?-e-E-P ")

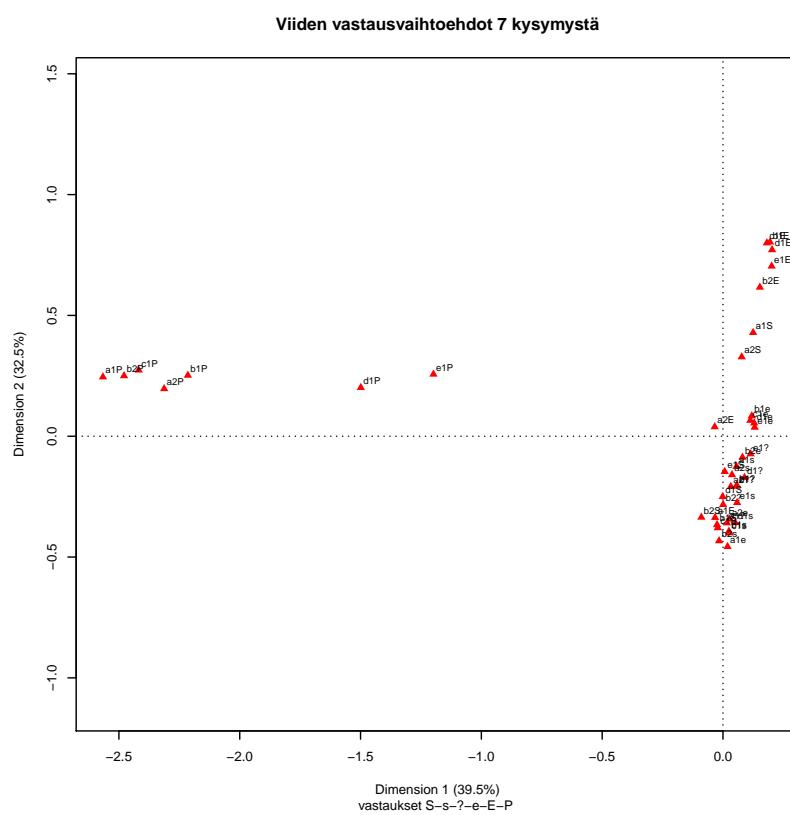
plot.mjca(Qmuuttujat1.mca, labels = c(1,2),
           main = "Viiden vastausvaihtoehdot 7 kysymystä",
           sub = "vastaukset S-s-?-e-E-P "
         )

# EI TOIMI pch = c(19, 1, 17,24) (pisteen symboli) 16.10.20
#
# pch Vector of length 4 giving the type of points to be used for row active and
# supplementary, column active and supplementary points

```



Kuva 36: MCA-kartta: viiden vastausvaihtoehdon kysymykset

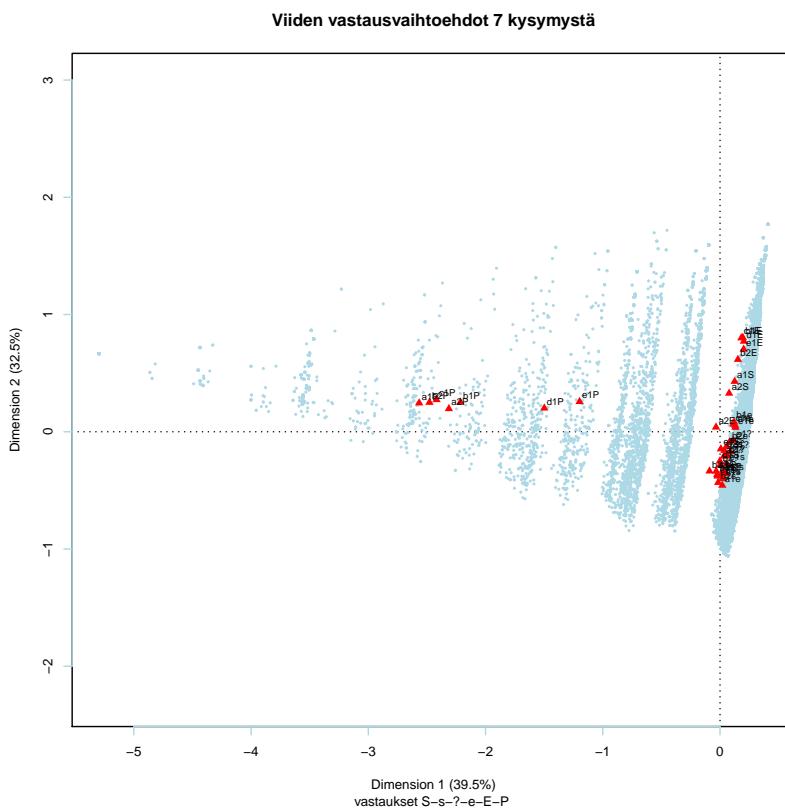


Kuva 37: MCA-kartta: viiden vastausvaihtoehdon kysymykset

```

par(cex=0.6)
plot.mjca(Qmuuttujat1.mca, what = c("all","all"), labels = c(0,2),
           col = c("lightblue", "red"),
           main = "Viiden vastausvaihtoehdot 7 kysymystä",
           sub = "vastaukset S-s-?-e-E-P "
         )

```



Kuva 38: MCA-kartta: viiden vastausvaihtoehdon kysymykset

```

summary(Qmuuttujat1.mca)

##
## Principal inertias (eigenvalues):
##
##   dim   value     %   cum%   scree plot
## 1   0.156867 39.5 39.5 *****
## 2   0.129043 32.5 71.9 *****
## 3   0.064154 16.1 88.0 ****

```

```

## 4      0.014818  3.7 91.8 *
## 5      0.008877  2.2 94.0 *
## 6      0.000538  0.1 94.1
## 7      0.000171  0.0 94.2
## 8      7.3e-050  0.0 94.2
## 9      2e-06000 0.0 94.2
##
##      -----
## Total: 0.397574
##
##
## Columns:
##   name   mass   qlt   inr      k=1 cor ctr      k=2 cor ctr
## 1 | a1S    48   916   21 | 125 71 5 | 429 845 69 |
## 2 | a1s    54   267   15 | 56 44 1 | -125 223 7 |
## 3 | a1?    15   633   20 | 17 1 0 | -359 632 15 |
## 4 | a1e    18   709   21 | 19 1 0 | -457 708 29 |
## 5 | a1E    5    112   23 | -33 1 0 | -337 111 4 |
## 6 | a1P    4    961   40 | -2566 952 155 | 245 9 2 |
## 7 | b1S    12   133   28 | -25 1 0 | -366 132 12 |
## 8 | b1s    37   696   21 | 25 3 0 | -394 693 44 |
## 9 | b1?    26   305   19 | 57 23 1 | -202 283 8 |
## 10 | b1e   39   138   19 | 119 92 4 | 84 46 2 |
## 11 | b1E   24   859   32 | 195 48 6 | 804 811 121 |
## 12 | b1P   5    945   41 | -2215 932 162 | 252 12 3 |
## 13 | c1S    12   140   28 | -22 0 0 | -379 140 14 |
## 14 | c1s    36   699   21 | 26 3 0 | -400 696 45 |
## 15 | c1?    26   318   19 | 56 22 1 | -207 296 9 |
## 16 | c1e    38   105   19 | 113 79 3 | 65 26 1 |
## 17 | c1E    26   866   32 | 182 42 5 | 800 823 129 |
## 18 | c1P    5    943   42 | -2418 931 171 | 273 12 3 |
## 19 | d1S    12   81    25 | -1 0 0 | -250 81 6 |
## 20 | d1s    33   755   20 | 50 15 1 | -358 741 33 |
## 21 | d1?    32   439   17 | 89 93 2 | -171 345 7 |
## 22 | d1e    34   164   18 | 129 139 4 | 55 25 1 |
## 23 | d1E    22   933   28 | 203 60 6 | 771 873 101 |
## 24 | d1P    9    969   35 | -1499 952 128 | 201 17 3 |
## 25 | e1S    15   46    23 | 7 0 0 | -146 46 2 |
## 26 | e1s    36   775   17 | 58 34 1 | -275 741 21 |
## 27 | e1?    34   316   16 | 115 224 3 | -73 92 1 |
## 28 | e1e    32   239   17 | 131 221 3 | 37 17 0 |
## 29 | e1E    15   978   24 | 201 74 4 | 704 904 58 |
## 30 | e1P    11   984   32 | -1198 940 102 | 257 43 6 |
## 31 | a2S    49   871   18 | 78 46 2 | 328 825 41 |
## 32 | a2s    59   466   14 | 37 24 1 | -160 442 12 |
## 33 | a2?    22   595   18 | 32 14 0 | -208 581 7 |
## 34 | a2e    8    571   20 | 27 3 0 | -343 567 8 |

```

```

## 35 | a2E | 2 3 20 | -34 1 0 | 39 2 0 |
## 36 | a2P | 3 971 32 | -2313 964 101 | 196 7 1 |
## 37 | b2S | 12 175 24 | -90 12 1 | -335 163 10 |
## 38 | b2s | 22 737 21 | -16 1 0 | -433 736 33 |
## 39 | b2? | 27 580 19 | 1 0 0 | -283 580 17 |
## 40 | b2e | 40 122 18 | 81 56 2 | -87 66 2 |
## 41 | b2E | 39 933 28 | 153 54 6 | 616 879 114 |
## 42 | b2P | 3 964 35 | -2479 954 123 | 250 10 2 |

#X11()
# subsetcat = (6,12,18,24,30,36,42) - väärä formaatti
# subsetcat=(1:42)[-c(1:5,7:11,13:17,19:23,24:29,31:35, 37:41)]) väärä formaatti

Qmuuttujat2.mca <- mjca(mcaDat11jh.dat, ps="", subsetcat=(1:42)[-c(6,12,18,24,30,36,42)])

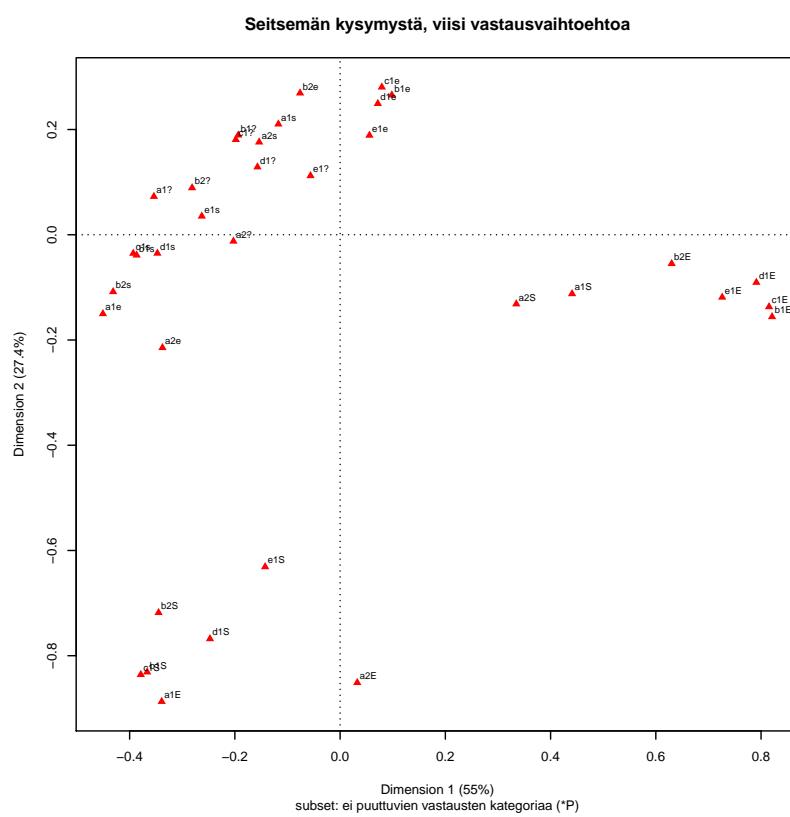
plot.mjca(Qmuuttujat2.mca,
            main="Seitsemän kysymystä, viisi vastausvaihtoehtoa",
            sub = "subset: ei puuttuvien vastausten kategoriaa (*P)")

plot.mjca(Qmuuttujat2.mca,what = c("all","all"),labels = c(0,2),
            col = c("lightblue", "red"),
            main="Seitsemän kysymystä, viisi vastausvaihtoehtoa",
            sub = "subset: ei puuttuvien vastausten kategoriaa (*P)")

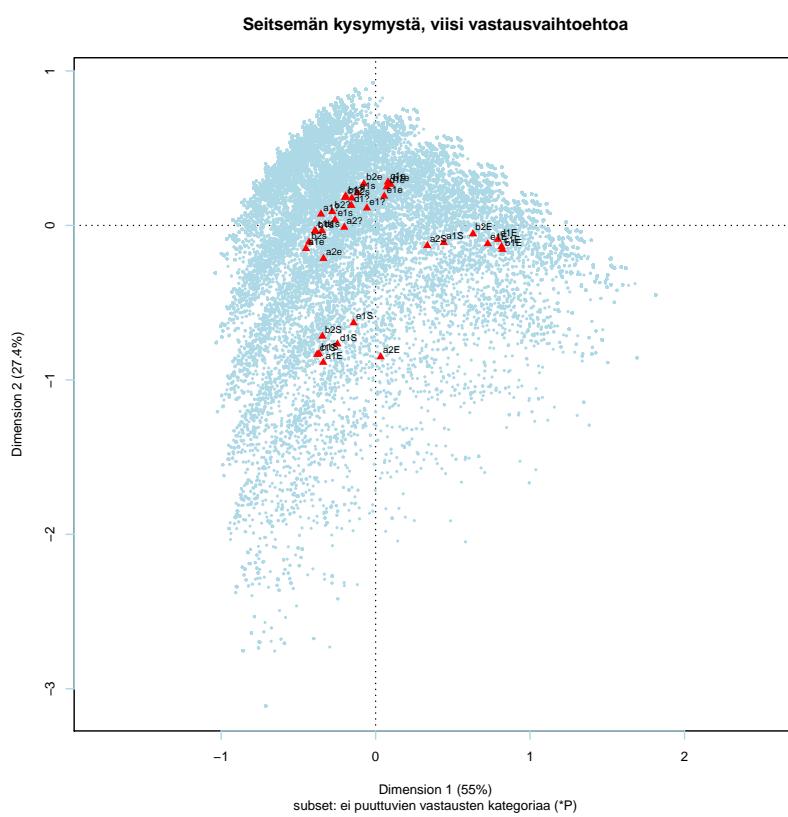
summary(Qmuuttujat2.mca)

##
## Principal inertias (eigenvalues):
##
##   dim   value    %  cum%  scree plot
##   1     0.129087 55.0 55.0 ****
##   2     0.064296 27.4 82.4 ****
##   3     0.014807  6.3 88.7 **
##   4     0.008871  3.8 92.5 *
##   5     0.000538  0.2 92.7
##   6     0.000221  0.1 92.8
##   7     0.000156  0.1 92.9
##   8     6.9e-050  0.0 92.9
##   9     1e-06000 0.0 92.9
##   -----
##   Total: 0.234728
##
##
## Columns:
##   name   mass  qlt  inr   k=1 cor ctr   k=2 cor ctr
## 1 | a1S | 48 975 29 | 441 916 73 | -112 59 9 |

```



Kuva 39: MCA-kartta: viiden vastausvaihtoehdon kysymykset



Kuva 40: MCA-kartta: viiden vastausvaihtoehdon kysymykset

```

## 2 | a1s | 54 869 20 | -117 206 6 | 210 663 37 |
## 3 | a1? | 15 653 27 | -354 626 14 | 73 26 1 |
## 4 | a1e | 18 788 29 | -451 709 28 | -150 79 6 |
## 5 | a1E | 5 884 30 | -339 113 4 | -887 771 56 |
## 6 | b1S | 12 818 37 | -367 133 12 | -831 685 128 |
## 7 | b1s | 37 696 28 | -387 689 42 | -39 7 1 |
## 8 | b1? | 26 527 25 | -194 269 8 | 189 258 14 |
## 9 | b1e | 39 555 25 | 98 67 3 | 266 488 43 |
## 10 | b1E | 24 889 43 | 821 858 126 | -156 31 9 |
## 11 | c1S | 12 826 38 | -379 141 14 | -836 685 134 |
## 12 | c1s | 36 698 29 | -393 692 43 | -35 6 1 |
## 13 | c1? | 26 517 25 | -198 282 8 | 181 235 13 |
## 14 | c1e | 38 550 26 | 79 41 2 | 281 509 46 |
## 15 | c1E | 26 890 43 | 815 865 134 | -137 24 8 |
## 16 | d1S | 12 851 34 | -247 80 6 | -768 771 112 |
## 17 | d1s | 33 751 26 | -347 744 31 | -35 8 1 |
## 18 | d1? | 32 561 23 | -157 336 6 | 129 226 8 |
## 19 | d1e | 34 620 24 | 72 47 1 | 249 572 33 |
## 20 | d1E | 22 944 38 | 791 932 106 | -91 12 3 |
## 21 | e1S | 15 910 31 | -142 44 2 | -631 866 90 |
## 22 | e1s | 36 752 23 | -263 738 19 | 35 13 1 |
## 23 | e1? | 34 350 22 | -56 70 1 | 112 279 7 |
## 24 | e1e | 32 597 23 | 56 48 1 | 189 549 18 |
## 25 | e1E | 15 1012 32 | 726 985 62 | -119 26 3 |
## 26 | a2S | 49 1005 24 | 335 870 43 | -132 134 13 |
## 27 | a2s | 59 989 19 | -154 428 11 | 176 561 28 |
## 28 | a2? | 22 584 24 | -203 582 7 | -12 2 0 |
## 29 | a2e | 8 794 26 | -338 565 7 | -215 228 6 |
## 30 | a2E | 2 870 27 | 33 1 0 | -851 869 20 |
## 31 | b2S | 12 922 33 | -345 173 11 | -718 750 94 |
## 32 | b2s | 22 783 28 | -431 736 32 | -108 46 4 |
## 33 | b2? | 27 640 25 | -281 582 16 | 89 58 3 |
## 34 | b2e | 40 711 24 | -76 53 2 | 269 658 45 |
## 35 | b2E | 39 939 37 | 630 932 119 | -55 7 2 |

```

```
# mutta ei hyväksy viimeisiä sarakkeita poistettaviksi (37:41)
```

```
#Qmuuttujat3.mca <- mjca(mcaDataJH2.dat, ps="", subsetcat=(1:42)[-c(1:5, 7:11, 13:17, 19:23])
```

```
#plot.mjca(mca1Qmuuttujat3)
#summary(mca1Qmuuttujat3)
```

```
#mca1Qmuuttujat3 <- mjca(mcaDataJH2.dat, ps="", subsetcat=(1:42)[-c(1:5, 7:11, 13:17, 19:23)])
```

```
#pchlist()
```

5.3.1 subset MCA ja täydentävät sarakkeet (19.10.20)

```
# Täydentävät sarakkeet - ei toimi! (19.20.20)
# dim(mcaDat11jh.dat) 7 kysymystä
dim(MCADATA1jh.dat)

## [1] 32823    26
str(MCADATA1jh.dat)

## tibble [32,823 x 26] (S3: tbl_df/tbl/data.frame)
## $ Q1am   : Factor w/ 6 levels "S","s","?","e",...: 5 1 2 2 1 6 2 4 2 2 ...
##   ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
## $ Q1bm   : Factor w/ 6 levels "S","s","?","e",...: 1 5 4 4 4 6 4 3 4 3 ...
##   ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ Q1cm   : Factor w/ 6 levels "S","s","?","e",...: 3 5 2 4 4 6 4 2 4 2 ...
##   ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
## $ Q1dm   : Factor w/ 6 levels "S","s","?","e",...: 3 5 5 2 4 6 4 5 4 5 ...
##   ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## $ Q1em   : Factor w/ 6 levels "S","s","?","e",...: 3 1 2 3 4 6 2 4 4 1 ...
##   ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working outside the home"
## $ Q2am   : Factor w/ 6 levels "S","s","?","e",...: 1 3 4 2 2 6 2 5 2 1 ...
##   ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## $ Q2bm   : Factor w/ 6 levels "S","s","?","e",...: 3 5 4 4 4 6 2 5 4 1 ...
##   ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
## $ edum   : Factor w/ 8 levels "No formal education",...: 3 6 6 4 3 8 8 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for international comparisons"
## $ sostam : Factor w/ 11 levels "Lowest, Bottom, 01",...: 3 7 8 11 7 2 7 11 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
## $ urbrum : Factor w/ 6 levels "A big city","The suburbs or outskirts of a big city",...
##   ..- attr(*, "label")= chr "Place of living: urban - rural"
## $ maa    : Factor w/ 25 levels "AU","AT","BG",...: 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## $ ika    : dbl+lbl [1:32823] 58, 59, 40, 20, 72, 68, 64, 57, 45, 71, 19, 41, 68, ...
##   ..@ label     : chr "Age of respondent"
##   ..@ format.spss: chr "F3.0"
##   ..@ labels    : Named num [1:6] 15 16 17 18 102 999
##   .. ..- attr(*, "names")= chr [1:6] "15 years" "16 years" "17 years" "18 years" ...
## $ sp     : Factor w/ 2 levels "m","f": 1 2 2 2 2 1 2 1 2 2 ...
##   ..- attr(*, "label")= chr "Sex of Respondent"
## $ age_cat: Factor w/ 6 levels "1","2","3","4",...: 5 5 3 1 6 6 5 5 3 6 ...
## $ ga     : chr [1:32823] "m5" "f5" "f3" "f1" ...
## $ E      : Factor w/ 8 levels "1","2","3","4",...: 3 6 6 4 3 8 8 7 6 7 ...
##   ..- attr(*, "label")= chr "Highest completed degree of education: Categories for international comparisons"
## $ S      : Factor w/ 11 levels "1","2","3","4",...: 3 7 8 11 7 2 7 11 10 6 ...
##   ..- attr(*, "label")= chr "Top-Bottom self-placement"
```

```

## $ U      : Factor w/ 6 levels "1","2","3","4",...: 1 1 1 6 1 2 6 2 2 6 ...
## ..- attr(*, "label")= chr "Place of living: urban - rural"
## $ gaf    : Factor w/ 12 levels "f1","f2","f3",...: 11 5 3 1 6 12 5 11 3 6 ...
## $ a1     : Factor w/ 6 levels "S","s","?","e",...: 5 1 2 2 1 6 2 4 2 2 ...
## ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
## $ b1     : Factor w/ 6 levels "S","s","?","e",...: 1 5 4 4 4 6 4 3 4 3 ...
## ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ c1     : Factor w/ 6 levels "S","s","?","e",...: 3 5 2 4 4 6 4 2 4 2 ...
## ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
## $ d1     : Factor w/ 6 levels "S","s","?","e",...: 3 5 5 2 4 6 4 5 4 5 ...
## ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## $ e1     : Factor w/ 6 levels "S","s","?","e",...: 3 1 2 3 4 6 2 4 4 1 ...
## ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working mom"
## $ a2     : Factor w/ 6 levels "S","s","?","e",...: 1 3 4 2 2 6 2 5 2 1 ...
## ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## $ b2     : Factor w/ 6 levels "S","s","?","e",...: 3 5 4 4 4 6 2 5 4 1 ...
## ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sample.RData"
# Data: 7 kysymystä, täydentävät muuttujat maa, sp, S (status), U (asuinpaiikka),
# E (koulutustaso)

mcaDat21jh.dat <- MCAdat1jh.dat %>% select(a1,b1,c1, d1, e1,a2,b2,E,S,U,maa,gaf)
str(mcaDat21jh.dat)

## # tibble [32,823 x 12] (S3:tbl_df/tbl/data.frame)
## $ a1 : Factor w/ 6 levels "S","s","?","e",...: 5 1 2 2 1 6 2 4 2 2 ...
## ..- attr(*, "label")= chr "Q1a Working mom: warm relationship with children as a not working mom"
## $ b1 : Factor w/ 6 levels "S","s","?","e",...: 1 5 4 4 4 6 4 3 4 3 ...
## ..- attr(*, "label")= chr "Q1b Working mom: Preschool child is likely to suffer"
## $ c1 : Factor w/ 6 levels "S","s","?","e",...: 3 5 2 4 4 6 4 2 4 2 ...
## ..- attr(*, "label")= chr "Q1c Working woman: Family life suffers when woman has full-time job"
## $ d1 : Factor w/ 6 levels "S","s","?","e",...: 3 5 5 2 4 6 4 5 4 5 ...
## ..- attr(*, "label")= chr "Q1d Working woman: What women really want is home and kids"
## $ e1 : Factor w/ 6 levels "S","s","?","e",...: 3 1 2 3 4 6 2 4 4 1 ...
## ..- attr(*, "label")= chr "Q1e Working woman: Being housewife is as fulfilling as working mom"
## $ a2 : Factor w/ 6 levels "S","s","?","e",...: 1 3 4 2 2 6 2 5 2 1 ...
## ..- attr(*, "label")= chr "Q2a Both should contribute to household income"
## $ b2 : Factor w/ 6 levels "S","s","?","e",...: 3 5 4 4 4 6 2 5 4 1 ...
## ..- attr(*, "label")= chr "Q2b Men's job earn money, women's job look after home"
## $ E : Factor w/ 8 levels "1","2","3","4",...: 3 6 6 4 3 8 8 7 6 7 ...
## ..- attr(*, "label")= chr "Highest completed degree of education: Categories for interview"
## $ S : Factor w/ 11 levels "1","2","3","4",...: 3 7 8 11 7 2 7 11 10 6 ...
## ..- attr(*, "label")= chr "Top-Bottom self-placement"
## $ U : Factor w/ 6 levels "1","2","3","4",...: 1 1 1 6 1 2 6 2 2 6 ...
## ..- attr(*, "label")= chr "Place of living: urban - rural"

```

```

## $ maa: Factor w/ 25 levels "AU","AT","BG",...: 1 1 1 1 1 1 1 1 1 1 ...
## ..- attr(*, "label")= chr "Country Prefix ISO 3166 Code - alphanumeric"
## $ gaf: Factor w/ 12 levels "f1","f2","f3",...: 11 5 3 1 6 12 5 11 3 6 ...
## - attr(*, "notes")= chr [1:45] "document Plan File: /Users/marcic/Desktop/old/GPS2011 sa...
dim(mcaDat21jh.dat)

## [1] 32823     12
glimpse(mcaDat21jh.dat)

## Rows: 32,823
## Columns: 12
## $ a1 <fct> E, S, s, S, P, s, e, s, s, e, s, s, S, P, s, ?, s, s, s, ...
## $ b1 <fct> S, E, e, e, P, e, ?, e, ?, ?, E, e, e, ?, ?, e, ?, P, e, e, ...
## $ c1 <fct> ?, E, s, e, e, P, e, s, e, s, ?, e, e, ?, s, e, s, ?, e, e, ...
## $ d1 <fct> ?, E, E, s, e, P, e, E, e, E, ?, e, e, E, ?, s, ?, e, s, P, e, ...
## $ e1 <fct> ?, S, s, ?, e, P, s, e, e, S, e, s, s, e, s, s, S, s, s, s, e, ...
## $ a2 <fct> S, ?, e, s, s, P, s, E, s, S, s, s, s, S, s, s, ?, s, s, s, ...
## $ b2 <fct> ?, E, e, e, e, P, s, E, e, S, e, e, e, E, e, ?, E, s, e, e, ...
## $ E <fct> 3, 6, 6, 4, 3, P, P, 7, 6, 7, 4, 6, 3, 6, 6, 2, 2, 7, 6, 6, 3, ...
## $ S <fct> 3, 7, 8, P, 7, 2, 7, P, 10, 6, 4, 5, P, 7, 9, 3, P, 8, 7, 6, 5, ...
## $ U <fct> 1, 1, 1, P, 1, 2, P, 2, 2, P, 2, 2, 3, 2, 2, 5, 5, 1, 2, 5, 4, ...
## $ maa <fct> AU, ...
## $ gaf <fct> m5, f5, f3, f1, f6, m6, f5, f3, f6, m1, m3, f6, f2, m1, m6, ...
# kysymysten puuttuvat pois subsetcat=(1:42)[-c(6,12,18,24,30,36,42)])
# maa ja gaf - ei puuttuvia tietoja
#
mcaDat21jh.dat$E %>% fct_count() # P-kategoria 50

```

f	n
1	491
2	1531
3	7811
4	7115
5	5658
6	5147
7	4762
P	308

```
mcaDat21jh.dat$S %>% fct_count() # P-kategoria 61
```

f	n
1	562
2	866

f	n
3	2221
4	3346
5	6798
6	6889
7	5778
8	3477
9	667
10	442
P	1777

```
mcaDat21jh.dat$U %>% fct_count() # P-kategoria 67
```

f	n
1	8442
2	4386
3	9203
4	8646
5	1902
P	244

```
# tämä ei toimi colsup = (1:50)[c(43:50)], supcol
# mcaDat21jh.dat[1:10,1:8]
# test1QjaMuut1.mca <- mjca(mcaDat21jh.dat[,1:8], ps = "", 
#                               what = c("none", "passive"),
#                               subsetcat=(1:50)[-c(6,12,18,24,30,36,42,50,61)],
#                               supcol = 8:9
# )
# QjaMuut1.mca
# summary(QjaMuut1.mca)
# plot(test1QjaMuut1.mca)
#X11()
```