Muste: Survomainen editoriaalinen käyttöympäristö R:lle

Reijo Sund

Tilastollinen tietojenkäsittely ei ole mahdollista ilman siihen soveltuvia työvälineitä. Käytännössä se tarkoittaa, että pelkkä matematiikan taitaminen ei riitä, vaan jokainen aineistoja analysoiva tilastotieteilijä tarvitsee työnsä tueksi ohjelmiston, jonka avulla hän saa tietokoneen valjastettua tekemään aineistolle haluamiaan toimenpiteitä. Moiseen puuhaan eivät rajallisen valikoiman ”käyttäjäystävällisiä” valikoita tarjoavat ohjelmat riitä, vaan kunnon ohjelmiston on tarjottava aineistojen analysointiin joustava ympäristö ja ominaisuudet, joita voi helposti ja vapaasti laajentaa tarpeen niin vaatiessa. Viime vuosina suursuosioon noussut R-ohjelmisto on ”tilastollisena ohjelmointikielenä” hyvä esimerkki tällaisesta joustavuudesta. Toinen erinomainen esimerkki on professori Seppo Mustosen elämäntyönään kehittämä Survo-ohjelmisto, jota voisi luonnehtia aikaansa edellä olevia innovaatioita täynnä olevaksi ”käyttöjärjestelmäksi tilastolliseen tietojenkäsittelyyn”.

Mustosesta ja Survosta

Professori Seppo Mustosen rooli tilastotieteen opetuksen uudistajana erityisesti 1960- ja 1970-lukujen vaihteen tuntumaan sijoittuneiden Helsingin yliopiston ensimmäisten tilastotieteen cum laude -kurssien toimeenpanijana, joiden yhteydessä muuten myös tilastotieteilijöiden ainejärjestö Moodi sai alkunsa, on kiistatta erittäin merkittävä. Lisäksi Mustosella, jota on joskus leikkimielisesti luonnehdittu ”Suomen ensimmäiseksi nörtiksi”, oli näppinsä pelissä suomalaisen tietojenkäsittelytieteen syntymisessä. Mustosen aloitteiden ja muistioiden pohjalta nimittäin perustettiin Pohjoismaiden ensimmäinen tietojenkäsittelyn professuuri

Yhteiskunnalliseen korkeakouluun (nykyään Tampereen yliopisto). Mustonen on myös uranuurtaja empiiristen aineistojen tilastollisiin analyyseihin soveltuvien eli tutkijoiden ajatusprosesseja tukevien tietojenkäsittelyohjelmistojen kehittämistyön saralla. Tästä näkyvin merkki on Mustosen elämäntyönään luoma Survo-järjestelmä, joka lienee yksi vanhimmista suomalaisista edelleen aktiivisessa kehityksessä olevista ohjelmistoprojekteista.

Mustonen oli kehittänyt vuodesta 1960 alkaen tilastollisten operaatioiden ohjelmakirjastoa Elliott 803 -koneelle Kaapelitehtaan elektroniikkaosastolla ja sen myötä syntyivät myös ensimmäiset konkreettiset ajatukset Survosta. Vuonna 1962 Mustonen oli Münchenissä IFIPin (International Federation for Information Processing) kokouksessa. Hän istui kuumana loppukesän päivänä ulkona kahvilassa yhdessä Martti Tienarin (myöhemmin Helsingin yliopiston ensimmäinen tietojenkäsittelytieteen professori), joka työskenteli Kaapelitehtaan elektroniikkaosastolla eli ”Salmisaaren yliopistolla” kuten Mustonenkin, ja Teknillisen korkeakoulun professori Olli Lokin kanssa. Mustonen ja Tienari kertoivat Kaapelitehtaalla syntyneistä ajatuksista. Lokki suhtautui niihin hyvin innostuneesti ja hänen kannustuksensa pohjalta ajatus muotoutui lopulta tilastolliseksi ohjelmointikieleksi, jonka toteutus jäi Mustosen harteille. Mustonen kehitteli ideoita edelleen myös yhdessä Timo Alangon kanssa tilastolliseksi ohjelmointijärjestelmäksi SURVO 64 vuosina 1963-1964. Ensimmäinen kokonaisuudessaan toteutettu Survo-järjestelmä oli SURVO 66 Elliott 803 -koneella. Jo siinä oli mahdollistettu käyttäjän aktiivinen rooli siten, että käyttäjä pystyi itse määrittelemään suoritettavat analyysit.

Survon seuraava sukupolvi (SURVO 76) toimi Wang 2200 -pientietokoneella. Siinä vuorovaikutteisuutta oli merkittävästi lisätty ja sen käyttäminen vastasi keskustelua tietokoneen kanssa. Huomattavin seuraavista edistysaskeleista oli niin sanotun

editoriaalisen käyttöliittymän sisällyttäminen Survoon vuonna 1979. Idea oli kytenyt Mustosella jo pitkään, sillä hän oli koulupojasta alkaen ollut innostunut ruutupaperin ja kynän kanssa työskentelystä ja kaipasi jotain vastaavaa myös tietokoneella työskentelyyn. 1980-luvun kynnyksellä koneet olivat myös kehittyneet sen verran, että alkoi olla realistista ajatella editoriaalisen käyttöliittymän toteuttamista. Hiukan yllättävästi varsinaisena alkusykähdyksenä editoriaalisen käyttöliittymän toteuttamiselle oli kuitenkin Mustosen (silloin 12-vuotiaan) Olli-pojan tarve nuottien puhtaaksipiirtämiselle, jota varten Mustonen kehitti (teksti)editorin, jonka avulla pystyi vapaasti kirjoittamaan nuottien tuottamiseen tarvittavia ohjauskoodeja ja edelleen piirtämään niiden mukaista nuottikirjoitusta joko ruudulle tai rumpupiirturille. Mustonen havaitsi nopeasti, että vastaava editoriaalinen käyttötapa soveltuisi erittäin hyvin myös tilastolliseen tietojenkäsittelyyn. Erityisesti yhtenä ensimmäisistä editoriaalisen käyttöliittymän komennoista toteutettu SORT vakuutti Mustosen ja näin ollen hän päätti hylätä (monista nykyisistä Windows-ohjelmista tutun) valikkopohjaisen käyttöliittymän ensisijaisena vaihtoehtona, koska se osoittautui rajoittavaksi ja ”vanhanaikaiseksi” jo 1980-luvun alussa. Siitä lähtien Survon sydämenä on ollut editoriaalinen toimintatapa toimituskenttineen, jossa rivien alut tähdittävä kontrollisarake on ”näköisjäänne” suomalaisen Hans-Peter Sehmin Wangille ohjelmoiman tekstieditorin ulkoasusta. Jo Wang-ajan Survolla sai myös piirrettyä kätevästi niin hyvännäköistä grafiikkaa, että Wangin graafisten näyttölaitteiden suhteellinen menekki oli silloin Suomessa koko maailman suurinta ja sitä ihmeteltiin valmistajan taholtakin.

PC-koneille siirtymisen jälkeen Survo (84C) oli jo laajentunut interaktiivisesta tilastollisesta ohjelmasta monialaiseksi yleiskäyttöympäristöksi. Survo 98:n myötä siirryttiin 32-bittiseen maailmaan ja Windowsin mahdollistama ikkunointi tuli käyttöön Survo MM -versiossa vuonna 2000. Tätä kirjoitettaessa Survo MM on

versiossaan 3.28 ja kehitys jatkuu (emeritus)professori Mustosen johdolla aktiivisesti edelleen. Lisätietoja Survosta löytyy osoitteesta http://www.survo.fi. Kannattaa katsastaa sieltä myös Survoon liittyvät julkaisut, jotka valottavat mielenkiintoisella tavalla sen pitkää kehityshistoriaa.

Survosta Musteeseen

Survon mahdollisuuksia on hyödynnetty monilla tieteenaloilla ja sillä on vakiintunut käyttäjäkunta. Survo on tällä hetkellä kuitenkin suljetun lähdekoodin ohjelmisto, joka toimii (vain) Windows-ympäristössä. Nykytrendinä monialustaisuus, eli käytännössä tarve siirtyä Windowsista muiden käyttöjärjestelmien (Linux, Mac OS X) käyttäjäksi, näyttäisi kuitenkin olevan kasvussa. Vuosien varrella tehtyjen erinäisten kokeilujen ja keskustelujen jälkeen ehdotinkin Mustosen vetämässä tilastollisen tietojenkäsittelyn seminaarissa vuoden 2009 helmikuussa, että Survon keskeiset ominaisuudet toteutettaisiin avoimen lähdekoodin monialustaisena projektina. Keskeinen ajatukseni oli integroida Survo osaksi valtaisan suosittua avoimen lähdekoodin R-ohjelmistoa, koska siten Survon kaikki tärkeimmät ominaisuudet tulisivat aidosti R:n osaksi täydentäen sitä muun muassa käyttöliittymän ja tiedostojen käsittelyn osalta merkittävällä tavalla. Projekti sai nimekseen Muste. Ehkäpä nimi onnistuu heijastelemaan erinäisiä ideoita liittyen Mustosen kehittämään ruutupaperin ja kynän korvaavaan editoriaaliseen käyttöliittymään, jonka huomaa siihen totuttuaan ”välttämättömäksi” sujuvan työskentelyn kannalta.

Kävi varsin nopeasti selväksi, että R:n perusjakelun mukana tuleva ensisijaisesti ikkunoiden hallitsemiseen tarkoitettu Tcl/Tk-rajapinta mahdollisti helposti Survon editorin perustoiminnan jäljittelemisen ja R:n avustamana joiltain osin jopa aidon Survon ohjelmamodulien suorittamisen. Kokeilut olivat niin lupaavia, että professori Mustosen suotuisalla myötävaikutuksella päätettiin kokeilla pystyttäisiinkö Survon C-kielistä lähdekoodia hyödyntämään moisessa Survolle uudessa ympäristössä.

Varsin nopeasti osoittautui, että hyödyntäminen onnistuu kohtuullisen helposti R:n C-ohjelmoinnille tarjoaman rajapinnan puitteissa ja näin ollen valtaosa Survon toiminnoista pystyttäisiin suoraan kopioimaan lähdekooditasolta lähtien eli ilman varsinaista uudelleenohjelmointia. Käytännössä vain järjestelmäriippuvaiset osat, kuten ruudulle piirtäminen ja näppäimistön luku ovat vaatineet uutta ohjelmakoodia ja nekin on Survossa alusta alkaen eristetty omiksi helposti hallittaviksi muutaman ydinfunktion kokonaisuuksiksi.

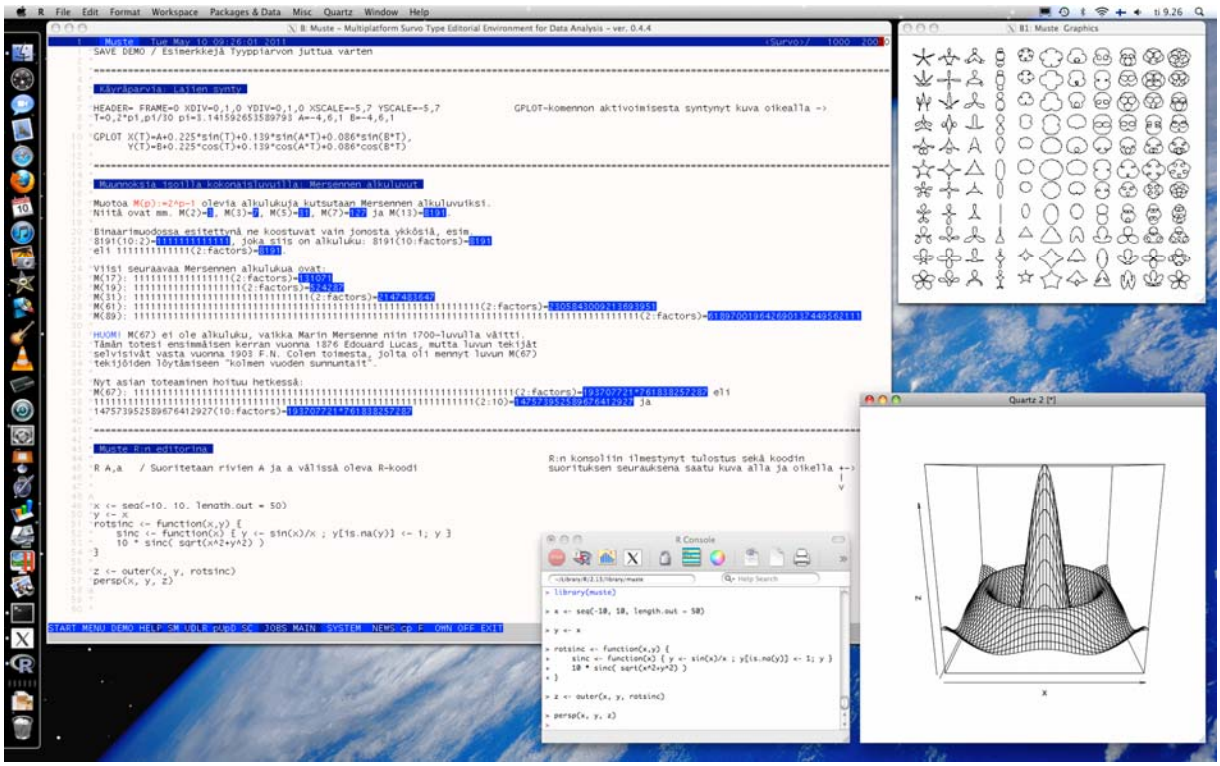
Seppo Mustonen on tarjonnut Muste-projektin käyttöön kaiken tarvittavan Survon C-kielisen lähdekoodin ja olen askarrellut hankkeen parissa oman toimen ohella aina silloin, kun siihen on löytynyt aikaa. Kyseessä on ollut mielenkiintoinen seikkailu ohjelmoinnin maailmassa, jossa ”puhutaan” sekaisin C:tä, R:ää ja Tcl/Tk:ta. Enemmän tai vähemmän yllättäviä haasteita on ollut paljon, mutta tähän mennessä niistä on selvitty kohtuullisen hyvin. Suunnitelmissa on ollut ensin saada Survon keskeiset ominaisuudet toimimaan Musteessa ja sen jälkeen aikomuksena olisi lisätä Muste-pakettiin toiminnallisuutta, joka helpottaa R-ohjelmiston muiden ominaisuuksien käyttöä Muste-paketin tarjoaman editoriaalisen käyttöympäristön kanssa. Tätä kirjoitettaessa tuorein Musteen versio on 0.4.4 ja se tarjoaa R-paketin muodossa Survon editorin miltei kokonaisuudessaan (mainittavimpana puutteena kosketuslaskenta), tiedostojen käsittelyyn tarvittavat komennot sekä erinäisiä muita ominaisuuksia. Myös kuvaruutugrafiikan toteuttaminen on lähtenyt hyvin käyntiin, vaikka se onkin ollut eniten järjestelmäriippuvainen osa Survoa. Miltei kaikkien muiden Survon toimintojen lisäämisen pitäisi olla kohtuullisen suoraviivaista, joskaan ei ihan automaattista. Uudet R-kytköksiä tukevat ominaisuudet vaatinevat sitten enemmän panostusta ja pohdintaa parhaista toiminta- ja toteuttamistavoista.

Esimerkkejä Musteen käytöstä

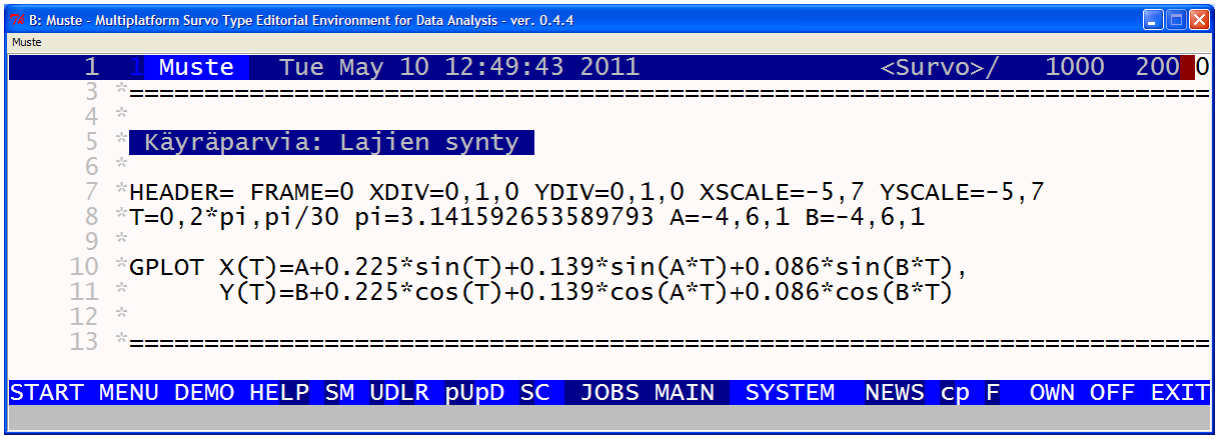
Kuvassa 1 on yleisnäkymä Musteella työskentelystä Mac OS X -käyttöjärjestelmässä. Suurimmassa ikkunassa näkyvään Musteen toimituskenttään on koottu muutamia yksinkertaisia esimerkkejä. Ensimmäinen niistä on lainattu suoraan Survon suomenkielisestä opetussarjasta ja siinä piirretään ”lajien synty” -käyräparvi, joka näkyy kuvan 1 oikeassa yläkulmassa. Näkymä samaan esimerkkiin toimituskentässä Windows XP -järjestelmässä on kuvassa 2. Kuva saadaan siis aikaan aktioimalla rivin 10 GPLOT-komento.

Toisessa esimerkissä laskeskellaan Mersennen alkulukuja ja hiukan pienempään ikkunaan typistetty esitys samasta aiheesta on luettavissa kuvasta 3. Esimerkissä on käytetty editoriaalisen aritmetiikan mahdollisuuksia mm. määrittelemällä rivillä 17 väliaikainen funktio, jolla voidaan laskea Mersennen lukuja sekä tekemällä (isojen) kokonaislukujen muunnoksia kannasta toiseen ja jakamalla kokonaislukuja alkutekijöihinsä.

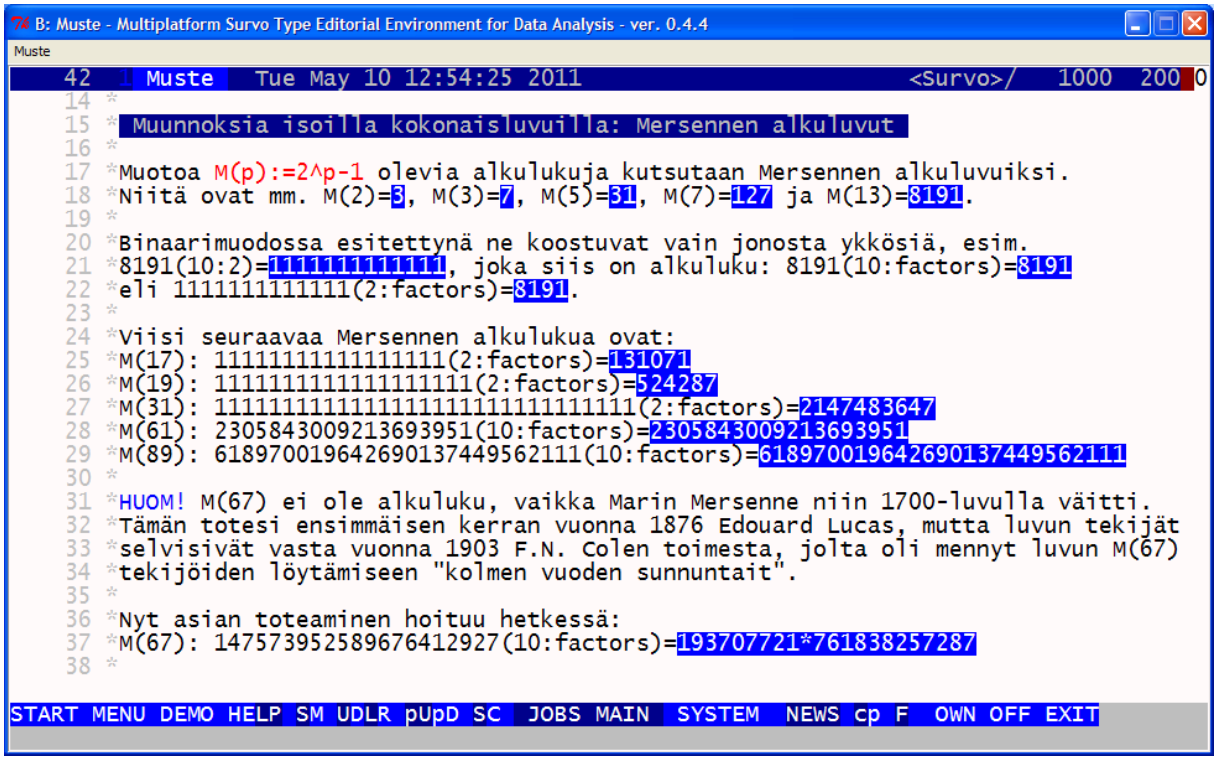
Kolmannessa esimerkissä on yksinkertaisesti kirjoitettu R-koodia Musteen toimituskenttään, josta se voidaan suorittaa aktivoimalla rivillä 44 näkyvä R-komento (kuva 4).



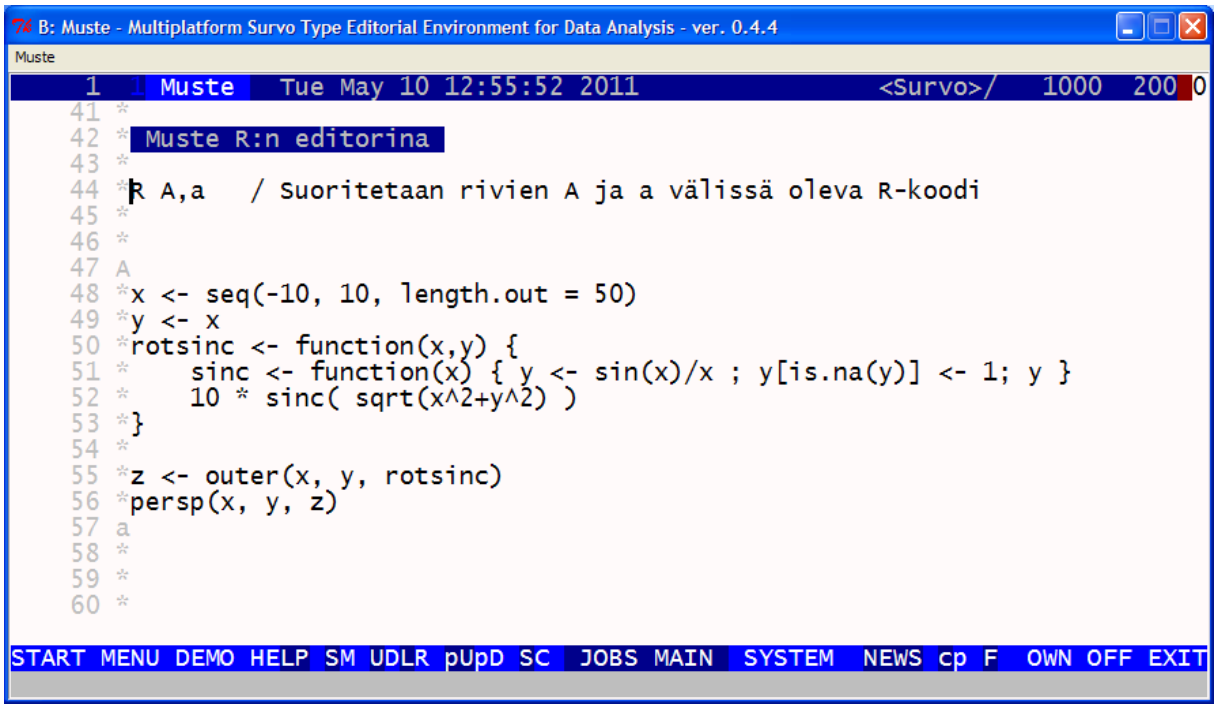
Kuva 1. Yleisnäkymä Musteen käyttöön Mac OS X -käyttöjärjestelmässä



Kuva 2. Esimerkki Musteen käytöstä: Käyräparven piirto



Kuva 3. Esimerkki Musteen käytöstä: Mersennen alkulukujen laskemista



Kuva 4. Esimerkki Musteen käytöstä: Muste R:n editorina

Musteen saatavuus

Toistaiseksi Mustetta ei ole laitettu yleiseen jakeluun, mutta sen binaarisen R-paketin saa testikäyttöön lähettämällä pyynnön ja tiedon kohdekäyttöjärjestelmästä sähköpostilla osoitteeseen reijo.sund@helsinki.fi. Kehitystilannetta voi myös seurata Musteen kotisivuilta osoitteesta http://www.survo.fi/muste, josta löytyy lisäksi muuta aiheeseen liittyvää materiaalia.

Reijo Sund, VTT, soveltavan tilastotieteen dosentti, tutkimuspäällikkö, Terveyden ja hyvinvoinnin laitos