

Tilastollisesta tietojenkäsittelystä¹⁾

Seppo Mustonen
Tilastotieteen laitos
Helsingin yliopisto

Automaattisen tietojenkäsittelyn valtava tekninen kehitys on viime vuosikymmeninä vaikuttanut suuresti eri tieteenalojen tutkimusmahdollisuuksiin. Myös tilastotieteessä tämä vaikutus ilmenee voimakkaana. Tilastollisessa tutkimustyössä eivät riitä yksinomaan periaatteelliset ja teoreettiset tarkastelut vaan tarvitaan myös selviä laskennallisia tuloksia.

Ennen tietokoneiden käyttöönottoa laskentavälineiden puutteellisuus rajoitti tuntuvasti tilastollisten menetelmien kehittämistä ja käyttöä sekä tilastotieteen omassa piirissä että tilastotiedettä soveltavilla aloilla. Silti jo pelkän käsinlaskennankin avulla ahkerat tutkijat suoriutuivat hankalista tehtävistä, vaikka aikaa kului paljon. Esim. muutaman muuttujan regressiomallin laskeminen saattoi viedä kokonaisen työpäivän, mutta sitä tarkemmin laskelmoitiin jo etukäteen, mitä kannatti laskeskella.

Saattaa uskoa, että tuollainen vaikkapa 1950-luvulla toiminut tilastollisia menetelmiä soveltanut tutkija olisi suunnattoman innostunut niistä keinoista, jotka nykyisin ovat tarjolla eri laskentatehtäviin. Kuitenkaan en ole varma, että hän olisi kovinkaan ilahtunut havaitessaan, kuinka tehottomasti näitä loistavia mahdollisuuksia käytetään hyväksi. Kielämättä nykyisin tehdään tilastollisia analyyseja valtavasti paljon, mutta onko tutkimusten taso siitä noussut?

Vaikuttaa siltä, että useat ATK:n apuun luottavat tutkijat ovat tietokoneiden ja niille laadittujen ohjelmistojen armoilla vailla kunnon käsitystä omista vaikutusmahdollisuuksistaan. He kamppailevat näiden uusien laskentavälineiden kanssa suorastaan sääliä herättävällä tavalla. He ovat vieraantuneet tärkeästä työkohteestaan, itse laskentaprosessista, jonka seuraaminen ja ohjaaminen oli itsestään selvä asia käsin tai pöytälaskukoneilla toimittaessa. Nykyisin saattaa syntyä empiirisiä tutkimuksia, joiden tekijä ei ole missään vaiheessa kunnolla tutustunut aineistoonsa eikä pahimmassa tapauksessa ole edes nähnyt havaintoarvoaan, koska tutkimusapulaiset ovat koonneet aineiston ja »menetelmäasiantuntijat» tai »tietokonemiehet» ovat toteuttaneet analyysit ja selittäneet tulokset.

¹⁾Virkaanastujaisesityelmä 16.11.1977



Tilannetta ei paranna se, että tietokoneiden usein laadultaan kyseenalaiset vakio-ohjelmistot houkuttelevat tutkijoita valitsemaan kaavamaisia ratkaisutapoja, joiden yksityiskohtiin heillä itsellään ei ole mitään mahdollisuutta puuttua. Toisaalta tutkijat voivat muodollisesti soveltaa kirjasto-ohjelmien avulla menetelmiä, joiden perusteista ja merkityksestä heillä ei ole riittävää käsitystä.

Tällaisen valmiiksi pakatun laskenta-avun entistä helpompi saatavuus on siis johtanut arveluttaviin käyttötapoihin, ja virhesovellutuksiin kohdistettu täysin oikeutettu arvostelu on osaltaan heittänyt varjonsa myös väärinkäytettyihin tilastollisiin menetelmiin.

Tällä asiantilalla on oma osuutensa siihen, että vielä 1960-luvulla varsin suositut monimuuttujamenetelmät, ennen muuta faktorianalyysi, ovat nykyisin miltei pannassa eräillä tutkimusaloilla. Faktorianalyysillä oli aikoinaan erittäin keskeinen asema suomalaisessa käyttäytymis- ja yhteiskuntatieteellisessä tutkimuksessa. Siitä tuli laskentamahdollisuuksien parantuessa muotimenetelmä, jota saatettiin soveltaa valmiiden ohjelmien suosiolla avustuksella ja onnistuneita tutkimusasetelmia matkien mitä kummallisimmissa tilanteissa eikä juuri tavattu soveltajaa, joka olisi ollut omiin tuloksiinsa tyytymätön. Faktorianalyysi on näet mekaanisesti käytettynä verrattavissa metsään, joka vastaa niin kuin sille huutaa. Faktorianalyysi on saanut sittemmin samoista syistä osakseen kohtuutonta innostusta myös muissa reaalitieteissä. Yhteiskuntatieteissä se johtui kuitenkin 1960-luvun lopulla ankaran arvostelun kohteeksi, mikä johtui ennen muuta sovellutusten heikosta tasosta ja samalla se oli helppo leimata epäilyttävien »positivististen» tutkimusmenetelmien joukkoon kuuluvaksi.

Nyt ei tietenkään ole tarkoitus moittia tietokoneiden suomia laskentamahdollisuuksia siitä, että faktorianalyysia tai muita tutkimusmenetelmiä saatetaan käyttää väärin. Jokaisen tutkimusedellytysten parannus voi edesauttaa myös huonojen tutkimusten tekoa, eikä uudistuksia näillä syillä tule torjua.

Jotain on kuitenkin mennyt vikaan. ATK:n tarjoama apu on nähdäkseni nykyisellään painottunut väärin eikä vastaa kaikin osin tilastollisen tutkimustyön tarpeita. Tällä väärisyyteisyydellä on oma »historiallinen» taustansa.

Kun tilastollisia ohjelmia lähdettiin alunperin kehittämään, huomion kohteeksi joutuivat etenkin laskentatyöltään raskaat menetelmät, joissa tietokoneen hyötysuhde on suurimmillaan. Yleisiä ohjelmia ei näyttänyt kannattavan laatia yksinkertaisia tilastollisia kuvaustapoja, laskentoja ja päättelyjä varten, vaikka nämä ovat käytännössä yleisimpiä ja sikäli myös tärkeimpiä. Uskottiin, että tutkijat hoitavat nuo helpommat tehtävät entiseen tapaan. Näin uskotaan kai edelleenkin, sillä vaikka entistä laajempia tilastollisia ohjelmistoja on otettu käyttöön, vielä tällä hetkellä tilanne on se, että tutkijan on vaivattomampaa laskettaa tietokoneilla esim. faktorianalyysia kuin toteuttaa oman aiheensa kannalta tärkeitä yksilöllisiä alkeistilastollisia laskentatehtäviä.

Niinpä syntyy houkutus käydä suoraan pitkälle automatisoitujen, laskennallisesti vaativien menetelmien kimppuun ja jättää alustavanluonteiset yksinkertaiset tarkastukset ja yhteenvedot suorittamatta. Tällaisista olosuhteiden yllyttämistä laiminlyönnestä seuraa, että tutkijalta jäävät havaitsematta monet aineistonsa erityispiirteet ja jopa koodaus- ja

lävistysvirheet saattavat kohtalokkaasti vääristää hienojen analyysien tuloksia. Mitään ylivoimaisia esteitä ei toki ole järkevien tarkistusten suorittamiselle ja tunnollinen tutkija hoitaa ne, vaikka se tuottaakin ylimääraistä vaivaa.

Tämänkaltaiset tilastollisen tietojenkäsittelyn epäkohdat eivät yksinomaan johdu liian kaavamaisista tilastollisista ohjelmista eikä tutkijoiden puutteellisesta metodologisesta koulutuksesta, vaan hyvin paljon myös yleisesti käytössä olevista ATK-laitteistoista ja niiden käyttömuodoista, jotka eivät aina sovellu joustavaan, ihmisläheiseen työskentelyyn.

Tilastolliseen tietojenkäsittelyyn tarkoitettut järjestelmät on jouduttu toteuttamaan ATK:n ehdoilla eikä tutkimuksenteon omia erityispiirteitä ole voitu ottaa huomioon riittävästi. Näitä ongelmia kärjistää käsitykseni mukaan se, että kaikki toiminta perustuu liikaa keskitettyihin, suuriin ATK-laitteistoihin ja niiden byrokraattisiin, laskentakeskusvaltaisiin käyttömuotoihin.

Keskitettyihin laitteistoihin tukeutuvat ratkaisut näyttivät vielä tämän vuosikymmenen alussa houkuttelevilta ja taloudellisesti ainoilta mahdollisilta. Tuntui edulliselta siirtää kaikki ATK-toiminnot harvoille, nimellisesti tehokkaille jättiläiskoneille. Uskottiin niiden moniajokäyttöjärjestelmien pystyvän palvelemaan kaikkia. Monessa suhteessa nämä toiveet ovat toteutuneet, mutta usein vasta kalliiden lisäinvestointien avulla.

Myös tilastollisen tietojenkäsittelyn osalta tähän kehitykseen liittyy paljon myönteistä. Esim. laajat simulointia tai valtakunnallisia tietorekistereitä tarvitsevat tutkimukset eivät olisi mahdollisia ilman näitä suuria ATK-järjestelmiä. On kuitenkin harhaanjohtavaa kuvitella, että tavanomaiset tilastolliset tutkimukset vaatisivat suuria laitteistoja ja byrokraattisia järjestelmiä, vaikka miltei yksinomaan sellaisia nykyisin käytetään. Useimmat tilastollisen tietojenkäsittelyn alaan kuuluvista lukuisista korkeakoulujen tutkimus- ja opetussovelluksista ovat varsin pieniä suurten laitteistojen mahdollisuuksia ajatellen. Tällaisissa pienissä tehtävissä suurten keskitettyjen järjestelmien käyttö on kankeaa ja epätaloudellista. Käyttäjät turhautuvat ja jopa luopuvat yrityksistään.

Tällaista taustaa vasten tuntuu luonnolliselta, että ns. pientietokoneet ja hajautetut tietojenkäsittelymenetelmät yleensäkin tarjoavat uuden varteenotettavan tuen tilastollisen tietojenkäsittelyn tehtäviin.

Helsingin yliopiston tilastotieteen laitoksella on nyt parin vuoden ajan tutkittu pientietokoneiden käyttömahdollisuuksia tilastollisen tietojenkäsittelyn alaan kuuluvissa tehtävissä. Päähuomio on kiinnitetty tilastotieteen omien tutkimus- ja opetusmahdollisuuksien kehittämiseen ATK:n avulla, mutta monessa suhteessa tällainen tutkimustyö antaa viitteitä myös siitä, millaista tilastollinen tietojenkäsittely voisi olla laajassa mittakaavassa lähitulevaisuudessa.

Olemme suunnitelleet laitoksen omalle pientietokoneelle uudentyypin tilastollisen käyttöjärjestelmän ja sitä laajennetaan jatkuvasti.

Kokemustemme mukaan on selvää, että sopivalla tilastollisella järjestelmällä varustetulla pientietokoneella saavutetaan tilastollisen tietojenkäsittelyn tehtävissä palvelutaso, joka laadultaan olennaisesti ylittää nykyisten yleisessä käytössä olevien järjestelmien mahdollisuudet. Tämä perustuu siihen, että pientietokoneella päästään välittömään, monipuoliseen

ja keskustelunomaiseen työskentelyyn ilman arvaamattomien ulkopuolisten tekijöiden aiheuttamia viivytyksiä. Tutkija voi itse vaivattomasti pitää yhteyttä sekä ohjelmistoon että käyttämiinsä aineistoihin. Hän voi haluamassaan järjestyksessä vaihteittain soveltaa yhtälailla yksinkertaisia tilastollisia kuvaustapoja kuin vaativia analyysimenetelmiä ja koko ajan seurata, mitä hänen aineistolleen tapahtuu.

Pientietokoneiden avulla voidaan myös palauttaa suosioon monia ihmisläheisiä työtapoja, jotka on syyttä suotta unohdettu tällä ns. tietokoneaikakaudella. Samalla työskentelyä on mahdollista suunnata muotoon, jossa tutkijalta jälleen edellytetään tarvitsemiensa menetelmien hallinta ja jossa hän voi monipuolisesti käyttää hyväksi taitojaan.

Esim. omassa järjestelmässämme koneen nopea kuvaputkinäyttölaitte mahdollistaa eräitä sellaisia osittain näköhavaintoon perustuvia tilastollisia tarkastelutapoja, jotka vaikeasti automatisoituvina ovat syrjäytyneet viime vuosikymmeninä. Faktorianalyysin ns. rotaatiovaihe on tästä sopiva osoitus. Sen alkuperäinen suoritustapa oli visuaalinen ja perustui tiettyihin peräkkäisiin graafisiin konstruktioihin. Kun rotaatio-ongelma ohjelmoidaan tietokoneella, jolla ei ole joustavaa keskustelunomaista ja kuvallista käyttötapaa, ratkaisun kriteerit joudutaan muokkaamaan sokeaan analyttiseen muotoon. Eräät tällaiset analyttiset rotaatiomenetelmät antavat muodollisesti hyviä tuloksia, mutta niiden varjopuolena on täysi tunteettomuus tutkijan mahdollisia erityistoivomuksia kohtaan. Monet ovat tosin sitä mieltä, että tämä onkin etu, koska analyttinen kriteeri estää liian subjektiiviset ratkaisut.

Tällaisissa kiistanalaisissa tilanteissa on hyödyllistä antaa tutkijalle tilaisuus vaihtoehtoihin menettelytapoihin ja niiden vertailuun.

Omassa järjestelmässämme faktorianalyysin rotaatiovaihe toteutetaan kuvaputken avulla siten, että päämenettelynä tutkija soveltaa graafista menetelmää, mutta hänellä on koko ajan samalla saatavilla neuvonantajinaan tietyt analyttiset rotaatiokriteerit.

Lienee selvää, että tällä tapaa työskentelevä tutkija voi olennaisesti paremmin ymmärtää ja hyväksikäyttää aineistoaan ja menetelmiä kuin heittäessään tehtävänsä täysin automaattiselle ohjelmalle, jonka toimintaa ei pääse lainkaan seuraamaan. Samoin menetelmien opetuksessa tällaiset havainnollistamiskeinot ovat korvaamattomia.

Monipuolinen pientietokonejärjestelmä on siten tutkijan ja opettajan käytössä todellinen data-analyttinen laboratorio, jonka tuottama hyöty esim. metodologian tutkimuksessa ja opetuksessa saattaa olla paljon enemmän kuin se sinänsä merkittävä laskenta-apu, jota ATK on tähän asti tarjonnut.

Vaikka toistaiseksi olisi syytä karttaa esittämästä kovin yltiöpäisiä ja yleistettyjä käsityksiä hajautetun tietojenkäsittelyn merkityksestä tutkimukselle ja opetukselle, uskon vakaasti, että yliopistojen ja korkeakoulujen ATK-resursseja kannattaisi viipymättä kohdistaa entistä enemmän pientietokoneisiin sekä niiden käyttöä koskevaan tutkimukseen ja suunnitteluun. Keskitettyjä laitteistoja ja laskentakeskusorganisaatiota toki tarvitaan edelleenkin, mutta niiden luonne ja tehtävät olisi ainakin tilastollisen tietojenkäsittelyn osalta arvioitava uudelleen.

ATK:ta käyttävien tutkijoiden ja opettajien tulisi itse pohtia, millaisia laitteistoja ja pal-

veluja he työssään tarvitsevat ja osallistua aktiivisesti oman sovellutusalan ohjelmistojen kehittämiseen.

Nykyisin on liikaa mukauduttu siihen, mitä sattuu olemaan tarjolla ja tietty muuttomattomuus uhkaa lyödä leimansa tietojenkäsittelytoimintoihin. Vaikka jäykkiä akateemisia traditioita yhä ylläpidetään, eivät ne ainakaan tilastollisessa tietojenkäsittelyssä ole vielä paikallaan.