

SUOMEN TILASTOSEURA 70 VUOTTA

STATISTISKA SAMFUNDET I FINLAND 70 ÅR

THE FINNISH STATISTICAL SOCIETY 70 YEARS

**Toimittaja
Redaktör
Editor**

Ilkka Mellin

HELSINKI 1991

Seppo Mustonen

Tilastotiede ja tietojenkäsittely

Kun niin moni asia maailmassa on tällä hetkellä muuttumassa, saatetaan myös kysyä, mitä on tapahtumassa **tilastotieteelle**. Onko tilastotieteen itsenäisyys uhattuna? Eikö se ole hajoamassa ympärillä olevien vahvemman identiteetin omaavien tieteenalojen, ennen kaikkea **matematiikan** ja **tietojenkäsittelyn** jalkoihin?

Vaikka tilastotieteellä on omat kunniakkaat perinteensä, sille kerääntyä jatkuvasti lisää vaikeuksia tulla toimeen näiden voimakkaampien kumppaniensa varjossa. Teoriapuolella matematiikka on näkyvämpi; tekniikassa ja tilastotiedon tuotannossa taas tietojenkäsittely valtaa yhä alaa. Monilla sovellusaloilla tilastotieteen itsenäinen osuus on jo kauan sitten pystytty peittämään.

Tilastotiedettä ja sen merkitystä eivät eri alojen tutkijat, saati sitten tavalliset valveutuneetkaan ihmiset, huomaa näiden paremmin tiedostettujen ja tunnustettujen alojen takaa.

Onneksi on joitakin muiden tieteiden erikoisaloja, kuten **biometria** ja **ekonometria**, jotka tunnustavat tilastollisen alkuperänsä. On kuitenkin aika havaita, ettei tilastotiede ole tässä maailmassa yksin ja että sen on kestävä erilaiset liittoutumat pysyäkseen hengissä.

Tilastotieteen suhde tietojenkäsittelyyn

Tilastotieteen suhde tietojenkäsittelyyn on erityisen pulmallinen mutta samalla kertaa myös mielenkiintoinen. Useimmat tutkijat kohtaavat tilastolliset asiat lähinnä vain ATK:n avulla ylläpidettyjen tietokantojen ja tilastollisten ohjelmien ja tietojärjestelmien kautta. On näennäisen helppoa harrastaa laajamittaista data-analyysia ja käyttää "syvällisiä" tilastollisia menetelmiä ymmärtämättä paljoakaan esim. otannasta ja tilastollisen ajattelun pelisäännöistä. Tässä ongelmana on se, ettei koko ongelmaa välttämättä tiedosteta. ATK-asiantuntijan apu ja jonkin ohjelmiston pikakurssi koetaan hyödyllisemmiksi kuin tilastollisten perusmenetelmien kunnolli-

nen opettelu.

Usein ei edes tajuta, että ollaan tekemisissä tilastollisten ongelmien kanssa. Nykyisin esim. tilastolliset piirrokset kulkevat "business-grafiikan" nimellä ja monet tilastolliset tiivistykset pannaan taulukkolaskennan ja tietokantojen hallinnan tiliin. Mutkikkaampia tilastolliseen ajatteluun perustuvia ratkaisuja saatetaan kutsua "liikkeenjohdon tukijärjestelmiksi" jne.

Tilastotieteilijöiden ongelmat

Tilastotieteen omat tutkijat ja asiantuntijat kohtaavat tietojenkäsittelyssä toisenlaisia pulmia, joita ei niitäkään aina hyvin tiedosteta. Osittain ongelmat ovat samoja kuin muillakin aloilla ja koskevat tietojenkäsittelyn roolia ja sen merkittävyyttä tutkijan työssä.

On esim. jotain hassunkurista siinä, että melkein jokainen matemaattisten alojen tutkija kokee nykyisin velvollisuudekseen oppia käyttämään jotain matemaattisen tekstinkäsittelyn ohjelmaa saadakseen hengentuotteen-
sa painokuntoon. Vaikka jotkut ovatkin taitavia tässäkin työssä ja se ilo heille suotakoon, ei voi olla kokonaisuuden kannalta järkevää, että niin monet tutkijat itse toimivat eräänlaisina kirjapainon latojina ja faktoreina. Yleensäkin tutkijoiden työssä ja keskinäisessä ajatustenvaihdossa tuntuu ATK-knoppologia nousevan tarpeettoman paljon esiin.

Jatkuvasti edistyvän ja uusia muotoja hamuavan tietotekniikan pinnallisten rituaalien omaksumiseen uhrataan liikaa aikaa tärkeämpien taitojen kustannuksella. Ollaan kaiketi ns. yleisten markkinavoimien armoilla.

Resurssien jako

Kun alituisesti valitetaan resurssien puutetta, voi myös kysyä, miten paljon toistaiseksi kannattaa uhrata esim. elektronisen postin ylläpitoon paikallisissa olosuhteissa. Kansainvälisessä ja pitkien etäisyyksien päästä tapahtuvassa tietojen vaihdossa tällainen toiminta on luonnollisesti paikallaan ja usein jopa välttämätöntä.

Samoin yhä edelleen jopa tekstinkäsittelyä ja muuta pienimuotoista tietojenkäsittelyä saatetaan harrastaa erittäin epätaloudellisesti keskustietokoneilla, kun "se ei tutkijalle itselleen maksa mitään!"

Ennustamisen vaikeus

Tietotekniikan kehityksen riipeys ja hallitsemattomuus sekä tähän liittyvä kaupallisuuden lietsoma toiveajattelu saavat viisaatkin ihmiset lyhytnäköisten ratkaisujen tielle. Jonkinlainen muodinmukaisuus ohjaa valintoja enemmän kuin kokemusperäinen tieto.

Ennustukset alan kehitysnäkymistä menevät usein pieleen. Kun noin 15 vuotta sitten, hieman ennen mikrojen tuloa, lähdin rakentamaan ensimmäistä vuorovaikutteista Survon versiota, eräät asiantuntijat esittivät voimakkaita epäilyjä. Jotkut kysyivät, milloin taas palaan tekemään kunnon tutkimusta ja lopetan leikkimisen koneilla. Toiset taas katsoivat vuorovaikutteisen käyttötavan olevan ylellisyyttä ja resurssien tuhlausta ja ettei sitä lainkaan tarvita tilastollisessa analyysissä.

Ohjelmointi vaikeaa

Myönnän itse erehtyneeni siinä, kun uskoin aikoinaan tietotekniikan omaksumisen olevan helppoa tutkijoille. Ilmeisesti kuitenkin vain harvat tilastotieteilijät kykenevät toimimaan myös tietojenkäsittelyn puolella luovina yksilöinä.

Kysymys ei liene pelkästään motivaation puutteesta, vaan esim. ohjelmointi on ihmisille yleensä yllättävän vaikeaa. Vaikka osattaisiinkin tehdä ohjelmia omaan käyttöön, yleisten, myös muille kelpaavien ratkaisujen kehittäminen ei välttämättä onnistu. Ei kyetä asettumaan "tavallisen" käyttäjän rooliin. Olenkin sitä mieltä, että meillä on aina olleet paremmat tietokoneet kuin ansaitsisimme.

Se, etteivät tietojärjestelmät ole tarpeeksi hyviä ja käyttäjien kannalta ystävällisiä, johtuu eniten meistä itsestämme.

Tilastollinen tietojenkäsittely

Selvintä yhteyttä ainakin nimellisesti tilastotieteen ja tietojenkäsittelyn välillä edustaa tilastollinen tietojenkäsittely (engl. statistical computing), joka on tälle alueelle omistautunut tilastotieteen (tai vaihtoehtoisesti tietojenkäsittelyn) erikoisuuntaus. Sen näkyvinä tunnusmerkkeinä ovat ISI:n alaisuudessa toimiva IASC-järjestö (**I**nternational **A**ssociation for **S**tatisti-

cal Computing) ja tämän järjestön kahden vuoden välein Euroopassa pitämät COMPSTAT-symposiot.

Tilastollisen tietojenkäsittelyn alalle on syntynyt myös lehtiä, joista **Statistical Computing Quarterly** on parhaillaan sulautumassa toiseen lehteen **Statistical Computing and Data Analysis**.

Toimintaa näillä nimikkeillä on ollut jo kymmenen vuoden ajan, mutta yhä vielä alue vaikuttaa selkiintymättömältä.

Tutkimusalueet

Näkyvimpinä tutkimuksen ja keskustelun aiheina ovat olleet mm.

- tilastolliset algoritmit, ohjelmat ja ohjelmistot
- numeerisesti hankalat tilastolliset menetelmät
- teoreettisesti vaikeiden ongelmien käsittely simuloimalla
- tilastollisiin tietokantoihin liittyvät kysymykset

sekä varsinkin viime vuosina lisääntyvässä määrin

- asiantuntijajärjestelmät ja
- tilastolliset käyttöympäristöt.

Nämä kaikki ovat mielekkäitä aiheita ja sopivat tilastollisen tietojenkäsittelyn piiriin. Liian usein varsinkin lehdissä on kuitenkin esiintynyt tarkasteluja, joissa tilastollisella tietojenkäsittelyllä ei ole mitään mainittavaa osuutta. Tyypillinen esimerkki tällaisesta on jonkin teoreettisen ongelman pohdinta, jota on jatkettu muodollisella simulointikokeella.

Arvostusongelmat

Alan selkiintymättömyys ilmenee parhaiten siinä, että sen arvomaailma on juuttunut perinteisen tieteenarvostuksen tasolle, vaikka tässä on alue, jolla uusien tuulien soisi puhaltavan. Useiden mielestä esim. tietojärjestelmän suunnittelu ja ohjelmointi yms. "likainen työ" ei ole tutkimusta, mutta valmiiden ohjelmien kaavamainenkin vertailu on. Arvostuksen puuttuminen haittaa myös kyvykkäiden ihmisten hakeutumista alalle.

Toisaalta innostutaan helposti esittelemään huimia tulevaisuuden suunnitelmia ja näköaloja, joista tuskin on lainkaan näyttöä. Toimivia ratkaisuu-

ja tuodaan harvemmin esiin. Tällaiset piirteet leimaavat varsinkin asiantuntijajärjestelmien ympärillä käytävää keskustelua.

Uudet mahdollisuudet

Tietotekniikka asettaa useanlaisia haasteita tilastotieteen tutkimukselle.

Mielenkiintoisia uusia mahdollisuuksia avautuu sekä symbolisen laskennan yhdistämisessä numeeriseen menetelmiin että rinnakkaislaskennan vaatimuksissa tilastollisten algoritmien suhteen.

Asiantuntijajärjestelmiin kohdistuva mielenkiinto on monessa suhteessa paikallaan ja sekin tulee tarjoamaan monenlaisia virikkeitä perinteiselle tilastotieteen tutkimukselle.

On myös alettu palauttaa kunniaan sellaisia käsilaskennan aikakaudella kehitettyjä menettelyjä, jotka tähän asti vaikeasti automatisoitavina ovat jääneet vaille huomiota.

Erityisesti tilastollisen aineiston graafinen esittäminen elää uutta kuokoistuskautta.

Koulutustarve

Tulevaisuuden visioista puhuttaessa ei kuitenkaan saisi unohtaa sitä tosi-seikkaa, että kehittyneet (tai kuten jotkut haluavat sanoa "syvälliset") tilastotieteen menetelmät ovat harvojen herkkua. Kun ajatellaan tilastotieteen käyttömahdollisuuksia laajemmin, tavallisten keskiarvojen laskeminen ja oikea tulkinta on tärkeämpää kuin yleistettyjen lineaaristen mallien ymmärtäminen ja yksinkertaisten otosten järkevä vertailu olennaisempaa kuin konfirmatorinen faktorianalyysi.

Jotta tilastotiede ei myöskään hukkuisi valmiiksi pureksittujen, automaattisten asiantuntijaratkaisujen alle, pitäisi havahtua näkemään se ilmeinen koulutus- ja valistustarve, joka tulee vastaan kouluopetuksesta alkaen.

Tilastotiede kouluissa

Nuoriso ei tällä hetkellä ole kovin kiinnostunut matemaattisista aineista. Lukiossa, jo opettajien asenteista johtuen (kun tilastotiede ei ole mikään luonnontiede), tilastotieteen opetus jää hyvin vähäiseksi. Tästä tiedon ja motivaation puutteesta on seurauksena, että yliopistollisiin tilastotieteen

opintoihin hakeudutaan melko sattumanvaraisesti.

Nämä tosiasiat ja niiden aiheuttama huoli tilastotieteen tulevaisuudesta ovat synnyttäneet kiinnostusta parantaa tilastotieteen imagoa koululaitoksen piirissä. Mielestäni eräs parhaita keinoja on lyöttäytyä yhteen sekä koulujen matematiikan että tietotekniikan opetuksen kanssa ja tarjota mahdollisuuksia näyttää ja opettaa tilastotiedettä aidosti tietotekniikkaa hyväksikäyttäen.

SURVOS-kokeilu

Jo kohta kolmen vuoden ajan on ollut käynnissä yhteistyössä Kouluhallituksen kanssa kokeilu, jossa lähes kymmenen lukion matematiikan opettajat ovat voineet perehtyä SURVO 84C -järjestelmään. Kokeilu on nyt johtanut siihen, että Survosta on kehitetty rajoitettu, erityisesti kouluopetuksen tarpeisiin suunnattu **SURVOS-versio**. Tähän koulusurvoon on rakennettu Survon omalla makrokielellä laaja opetusohjelmasarja, joka käsittelee paitsi Survon oppimista myös matematiikan ja tilastotieteen asioita. Sama suomenkielinen opetussarja kuuluu myös laajan Survon uusimpaan laitokseen.

Jopa suppea SURVOS-versio sisältää opetusohjelmien ja yleisesti Survon makrojen eli ns. **sukrojen** tekoon tarvittavat työvälineet. Tavoitteena on saada opettajat laatimaan lisää omia, enemmän yksityiskohtiin meneviä opetusohjelmia. Mallina on käytettävissä nykyinen opetussarja.

Muut opetuskokeilut

Tällä hetkellä on jo olemassa muitakin Survoon liittyviä ja tilastotieteen opetusta tukevia opetusohjelmia.

Esim. **Jouko Manninen** Helsingin kauppakorkeakoulusta on tehnyt tilastotieteen peruskurssin harjoittelussa sovellettavan **TILJAT**-nimisen opetussarjan, joka ei suoraan edellytä Survon hallintaa.

Myös Helsingin yliopiston psykologian laitoksella ja Tampereen yliopiston tilastotieteen laitoksella Survo on ollut peruskurssitasoisessa opetuskäytössä jo parin vuoden ajan.

Survo opetusjärjestelmänä

Survo nykymuodossaan tarjoaa ymmärtääkseni moniin muihin tilastollisiin järjestelmiin verrattuna hyvän pohjan opetusohjelmien ja käyttäjäkohtaisesti räätälöityjen sovellusten teolle mm. seuraavista syistä.

Survossa tilastolliset menetelmät eivät ole yksin vaan järjestelmä tukee yhdenmukaisesti myös muita toimintoja, joita opetuksessa ja ammattisovelluksissa tarvitaan. Näitä toimintoja ovat mm. tekstinkäsittely, erilaiset vapaamuotoiset laskentatavat, grafiikka, tietokantojen hallinta ja julkaisujen laadinta. Eri toimintojen yhdistämiseen opetus- ja asiantuntijasovelluksiksi on lisäksi käytettävissä omia ohjelmointitapoja kuten em. sukrokieli ja matriisitulkki.

Näillä keinoilla on opetusohjelmissa mahdollista tarvittaessa peittää Survon oma käyttöliittymä ja siirtyä esim. täysin valikko-ohjattuun työtapaan. Näin on tehty mm. edellä mainituissa opetusohjelmissa ja Tiedekeskus Heurekaan laaditussa tilastotieteen perusteita koskevassa esittelyssä. Survoa käytetään tällöin siis valmiita perusrakenteita ja -ratkaisuja sisältävänä käyttöympäristönä ja sovelluskehittimenä.

Tietojenkäsittelystä ajatusten käsittelyyn

Samalla periaatteella edistyneet Survon käyttäjät kehittelevät omia vähemmän tiedäville tarkoitettuja asiantuntijasovelluksiaan. Esim. monivaiheiset ja teknisesti vaativat tilastollisen analyysin muodot voidaan Survossa muuntaa sukrotekniikalla hyvin helpokäyttöisiksi tarvitsematta tehdä töitä esim. C-kielellä, jolla itse Survo on koodattu.

Kaikessa tässä tähdätään työtapaan, jossa pienin askelin ja suhteellisen yksinkertaisin menetelmin keskustellaan tutkittavien aineistojen kanssa. Kysymys on ennen kaikkea **ajatusten käsittelystä**, jota harrastetaan tietojenkäsittelyn tuella. Siinä tilastollisella analyysillä saattaa olla oma tärkeä osuutensa, mutta ei aina.

Kohti yleistä käyttöympäristöä

Suurin osa ihmisistä, joilla mikrotietokone on jokapäiväisenä työvälineenään, tarvitsee kuitenkin muutakin kuin pelkkää tekstinkäsittelyä. Kun he huomaavat, että rutiinitoimenpiteisiin on mahdollista liittää myös las-

kutoimituksia ja aineiston tilastollista käsittelyä, kiinnostus tilastotiedettä kohtaan saattaa lisääntyä aidolla tavalla.

Omassa työssäni pidän sen vuoksi tärkeimpänä tavoitteena luoda edellytykset aidolle asiantuntijan käyttöympäristölle, jossa tilastolliset analyysi- ja laskentamenettelyt ovat välittömästi saatavilla työskentelyn luonnollisina osina. Vaikka erikoisohjelmia ja -järjestelmiä yhä tarvitaan, kehityksen painopiste tulee olemaan yleisten työympäristöjen parantamisessa.

S.Mustonen: Statistics and Computer Science

Relations between Statistics and Computer Science are studied. Especially the role of statisticians in statistical computing and related areas is considered. Some new trends in statistical data processing are described.