Miksi valtava datamäärä tuottaa niin vähän tietoa?

Olen saanut 1970-luvun oloissa parhaan mahdollisen tilastotieteellisen koulutuksen professorien Leo Törnqvist ja Seppo Mustonen oppilaana. Tilastollinen tutkimus erilaisten ilmiöiden keskinäisistä riippuvuuksista oli silloin paljolti käsityötä. Dataa oli käytettävissä niukasti ja siksi vähistä tiedoista oli otettava kaikki irti. Professori Törnqvist opetti aloittamaan tutkimuksen aina lyijykynän ja millimetripaperin kanssa. Jokainen havainto merkittiin paperille, jotta ymmärrettäisiin, millaista dataa käsitellään. Seppo Mustosen kehittämä Survo-ohjelma lisäsi tässä tuottavuutta monikymmenkertaisesti, kun saman pystyi tekemään tietokoneen näytöllä. Dataa piti usein korjata, koska tiedoissa oli virheitä tai johonkin havaintoon liittyi jokin poikkeuksellinen seikka. Jos esimerkiksi halusi selvittää junalippujen hinnan vaikutusta junalla matkustamisen suosioon, ei ollut järkevää ottaa aineistoon havaintoa ajalta, jolloin junat seisoivat lakon takia.

Professori Törnqvist korosti käytettävän mallin järkevyyttä. Ei ollut mieltä käyttää lineaarista mallia tilanteessa, jossa riippuvuus ei voinut olla lineaarista. Pahinta mitä saattoi tehdä, oli ”dimensiovirhe”, jossa tulos muuttuisi toiseksi, jos esimerkiksi pituutta mitattaisiin metrien sijasta jalkoina.

Tiesimme, että uusi aika tehokkaine tietokoneineen ja suurine datamäärineen oli tulossa. Tämä tulisi merkitsemään jättiharppausta yhteiskunnallisten ilmiöiden ja niiden keskinäisten riippuvuuksien ymmärtämisessä. Edessä piti olla yhteiskuntatieteiden kukoistuskausi. Professori Yrjö Ahmavaara hahmotteli ehkä vähän orwelmaiselta haiskahtavaa kyberneettistä yhteiskuntapolitiikkaa, jossa päätöksentekijät voisivat optimoida hyvinkin monimutkaista hyötyfunktiota, kun toimenpiteiden monimutkaiset vaikutukset olisivat tiedossa.

Laskentakapasiteetin ja ennen kaikkea käytössä olevan datamäärän kasvu on ylittänyt kaiken sen, mitä saatoimme kuvitella, mutta sitä suurta yhteiskuntatieteiden kukoistuskautta yhä odotetaan. Miksi datan tulva ja valtavasti kasvanut laskentakapasiteetti ei ole tuottanut tiedollista vallankumousta?

Valtavaan datamäärään sisältyy aina virheellisiä havaintoja. Me jouduimme tavallisesti poistamaan aineistosta noin kymmenennen osan havainnoista tai korjaamaan niitä. Tätä ei voi tehdä ainakaan käsin, jos havaintoja on miljoonia. Pieni määrä oikeita havaintoja antaa tarkemmat estimaatit kuin valtava määrä virheellisten havaintojen saastuttamaa aineistoa. Tilastolliset menetelmät perustuvat tavallisesti pienimmän neliösumman menetelmään, jolloin muusta aineistosta poikkeavat havainnot saavat hyvin suuren painoarvon.

Tutkijan pitää myös tuntea käyttämänsä menetelmät ja mittarit. Niin yksinkertainen asia kuin korrelaatiokertoimen tulkintakin voi mennä aivan metsään. Esimerkiksi muuttujien x ja y välinen riippuvuus voi olla hyvinkin voimakasta, vaikka havaintoaineistossa niiden välinen korrelaatio on vähäinen. Korrelaatio mittaa lineaarista riippuvuutta ja vain sitä. Korrelaatiokerroin voi olla matala vaikka muuttuja y olisi suorastaan muuttujan x funktio, jos riippuvuus on epälineaarinen.

1970-luvun lopulla faktorianalyysi teki tuloaan yhteiskuntatieteisiin. Tämä menetelmä on tehokas, mutta hyvä se on vain sellaisen tutkijan käsissä, joka ymmärtää, miten analyysi toimii. Aivan järkyttäviä nollatutkimuksia julkaistiin tuolloin jopa väitöskirjoina, kun tutkija tulkitsi faktorilatauksia kuin Delfoin oraakkeli. Nyt näkee tutkimuksia, jotka analysoitu jollain minulle tuntemattomalla tavalla. Kun tutkijalta kysyy, miten se on analysoitu, saa vastaukseksi kaupallisen ohjelmiston nimen. Siihen, mitä tuo ohjelmisto tarkkaan ottaen tekee, ei vastausta tule – ohjelman toimintaperiaate saattaa olla jopa liikesalaisuus! Miten sellaista voi käyttää tutkimuksessa?

Yhteiskunnallinen tutkimus ja erityisesti epidemiologinen tutkimus kärsii myös väärin ymmärretystä tietosuojasta. Kaupalliset tahot saavat tallettaa ihmisistä tietovarantoihinsa asioita, joista akateemiset tutkijat eivät voi kuin haaveilla. Kukaan ei antaisi ikinä tutkijalle lupaa kerätä yksityiskohtaisia tietoja tavallisten kansalaisten päivittäisistä ostoksista tai oikeutta analysoida

ihmisten toisilleen lähettämiä sähköpostiviestejä ja tarkkailla heidän tietohakujaan, mutta kaupan keskusliikkeet, Facebook ja Google saavat näin tehdä tai ainakin tekevät. Tähän verrattuna aika vaatimaton hanke oli HSL:n yritys tallettaa matkakorttia käyttävien matkat, jotta reitit ja aikataulut osattaisi suunnitella paremmin. Se kiellettiin tietosuojaa loukkaavana.

Kuvitelkaa, mitä kaikkea voisi tehdä, jos epidemiologinen tutkimus voisi käyttää samanlaisia tietovarastoja, joita nuo edellä mainitut yksityiset yritykset käyttävät! Tutkijoiden tulisi nousta barrikadeille puolustamaan sitä, että olemassa olevia ja laajenevia valtavia tietovarastoja saisi käyttää myös yhteiskunnalliseen ja lääketieteelliseen tutkimukseen eikä vain yksityisten yritysten tarpeisiin.

Jokainen tutkija ei voi kouluttautua tilastotieteen huippuosaajaksi, mutta jokaisen tutkimusryhmän käytettävissä pitäisi sellainen olla. Tilastollisia osaajia pitäisi kouluttaa selvästi enemmän. Eikä tilastotieteen koulutus saisi olla vain matemaattisten teorioiden pänttäämistä – nekin on kyllä hyvä osata – sillä hyvä tilastotieteilijä on ennen kaikkea käsityöläinen.

Havaintoaineistoista pitäisi aina piirtää kuvia koska kovista ihminen hahmottaa asioita paljon paremmin kuin luvuista. Enää ei onneksi tarvitse käyttää millimetripaperia ja lyijykynää, vaan tasokkaita kuvia saa tietokoneen näyttää murto-osasekunnissa. Aineistoa on katseltava eri puolilta. Usein järkevän kuvan muuttujien y ja x riippuvuudesta saa vasta, kun on poistanut molemmista kolmannen muuttujan z vaikutuksen. Kun tuntee aineistonsa hyvin, ei yleensä tarvitse kovin monimutkaisia tutkimusmenetelmiä. Tutkimusmenetelmä ei oikeastaan koskaan saisi olla niin kehittynyt, ettei tutkija ymmärrä, miten se toimii.

On myös poistettava aineistoa häiritsevät virheelliset havainnot tai korjattava niitä. Jos niitä ei pysty poistamaan käsin, ne voi poistaa tähän tarkoitukseen suunnitellulla algoritmilla. Jos havainto poikkeaa muista liikaa – useita standardipoikkeamia – melkein varmasti virheellinen. Tämä voi johtaa myös oikeiden havaintojen hylkäämiseen, mutta tämä riski on paljon pienempi kuin riski virheellisiin havaintoihin perustuvasta väärästä johtopäätöksestä.

Osmo Soininvaara
Valt.lis
Eduskunta, kansanedustaja 19.4.2015 saakka Siitä alkaen luennoija