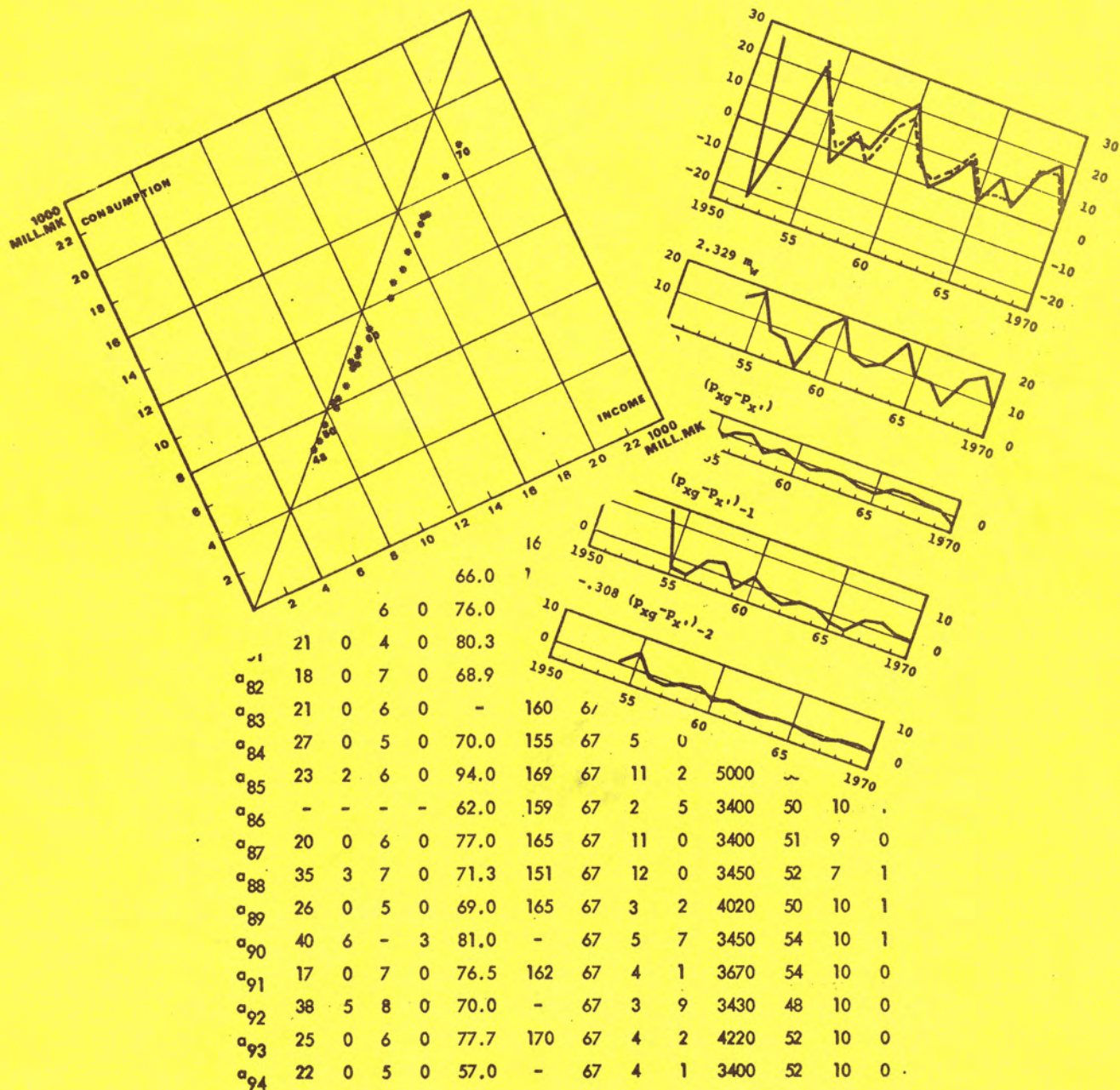


HELSINGIN YLIOPISTON TILASTOTIETEEN LAITOKSEN
OPETUSMONISTEITA N:O 3

TILASTOTIETEEN SOVELLUTUKSISTA
(toim. Yrjö O. Vartia)



Helsinki

marraskuu 1980

ISBN 951-45-2481-0

Toinen korjattu painos 1983

Toimittajan esipuhe

Helsingin yliopiston valtiotieteellisessä tiedekunnassa on viime vuosina luennoitu Prof. Mustosen aloitteesta kurssia tilastotieteen sovellutuksista. Kurssi on koostunut eräiden tärkeiden tilastotieteen sovellutusalueiden asiantuntijoiden pitämistä 4-6 luentotunnin jaksoista. Opetusohjelmassa kurssi seuraa välittömästi tilastotieteen johdantokurssia (entistä peruskurssia), joka kattaa vain tilastotieteen peruskäsitteiden, todennäköisyyslaskennan, parametrien estimoinnin ja hypoteesien testaamisen alustavan esittelyn.

Tilastotieteen sovellutuksia käsittelevällä kurssilla on syvennetty ja havainnollistettu havaintoaineiston hankintaa ja analysointia otantatutkimusten ja kokeellisen tutkimuksen yhteydessä sekä esitelty erityisesti Tilastokeskuksen toimintaa yhteiskuntatilastojen tuottajana. Johdantokurssiin nähden kokonaan uusia aihepiirejä kurssilla ovat edustaneet varianssi- ja regressioanalyysia, pääkomponentti- ja faktorianalyysia, aikasarjojen dekomponointia ja ennustamista, indeksilukujen konstruointia ja tilastollista tietojenkäsittelyä käsittelevät luentojaksot.

Allekirjoittaneen tehtävänä on ollut luentosarjan yleissuunnittelu, johon on sisältynyt käsiteltävien aihepiirien valinta ja rajaus. Luentomoniste perustuu vuoden vaihteessa 1979-80 pidettyihin luentoihin, joista esitel-

möijät ovat muokanneet oheiset esitykset. Luentoja suunniteltaessa on pyritty noudattamaan kaaviota

ONGELMA → HAVAINTOAINEISTO → MENETELMÄT → VASTAUS

tilastotieteen "keittokirjoissa" usein sovelletun kaavion

MENETELMÄ → HAVAINTOAINEISTO → VASTAUS (Mihin ongelmaan?)

sijaan.

Luentomonistetta ei voida vielä pitää kaikissa suhteissa tyydyttävänä, vaan esitystä on tarkoitus edelleen kehittää tarkoituksenmukaisempaan ja lukijaystävällisempään suuntaan. Tämän vuoksi toivon lukijalta kommentteja, jotka voidaan toimittaa allekirjoittaneelle.

Todettakoon lopuksi, ettei opetusmoniste ole tarkoitettu itseopiskelijoille, vaan luentojen oheismateriaaliksi, jota luennoilla edelleen täydennetään ja havainnollistetaan.

Helsingissä 5.11.1980

Yrjö O. Vartia

Yrjö O. Vartia

Opetusmonisteen toimittaja

Tilastotieteen apulaisprofessori

Toiseen korjattuun painokseen ovat luennoitsijat tehneet melko huomattavia korjauksia. Esitysjärjestystä on myös muutettu, eikä toista painosta voi enää suositella käytettävän rinnan ensimmäisen painoksen kanssa.

Helsingissä kesällä 1983

TILASTOTIETEEN SOVELLUTUKSISTA
(toim. Yrjö O. Vartia)

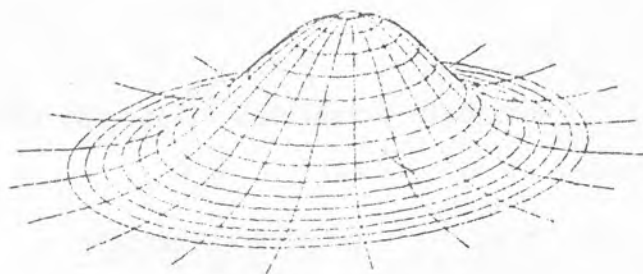
<u>Professori Seppo Mustonen (Tilastotieteen laitos):</u> <i>Tilastollisesta tietojenkäsittelystä</i>	4
<u>Apulaisprofessori Anders Ekholm (Tilastotieteen laitos):</u> <i>Varianssi- ja regressioanalyysistä</i>	28
<u>Assistentti Osmo Soininvaara (Tilastotieteen laitos):</u> <i>Faktorianalyysistä</i>	64
<u>Lehtori Markku Rahiala (Tilastotieteen laitos):</u> <i>Aikasarjojen vaihtelukomponenteista ja ennustamisesta</i>	74
<u>Apulaisprofessori Yrjö O. Vartia (Tilastotieteen laitos):</u> <i>Aikasarjojen konstruointi ja indeksiluvut</i>	100
<u>Lehtori Matti Liedes (Tilastotieteen laitos):</u> <i>Otanta tutkimuksesta</i>	140
<u>Lehtori Pyy-Matti Vasama (Tilastotieteen laitos):</u> <i>Kokeellisesta tutkimusmenetelmästä</i>	163
<u>VTL Aarno Laihonon (Tilastokeskus):</u> <i>Valtion tilastotoimi yhteiskuntatilaston tuottajana</i>	175

Seppo Mustonen

Tilastollisesta tietojenkäsittelystä

A SAMPLE OF THE INFORMATION AND A GRAPHICAL REPRESENTATION OF THE SAMPLE HAS BEEN OBTAINED AT THE END OF THE FIRST COURSE. ALSO STATISTICAL METHODS FOR THE ANALYSIS OF DATA ARE GIVEN IN THE COURSE.

BIVARIATE: $X - Y$



IN THE PREVIOUS COURSE, THE STUDENT HAS BEEN GIVEN AN IDEAL OF THE DATA. THE STUDENT HAS BEEN GIVEN AN IDEAL OF THE DATA. THE STUDENT HAS BEEN GIVEN AN IDEAL OF THE DATA.

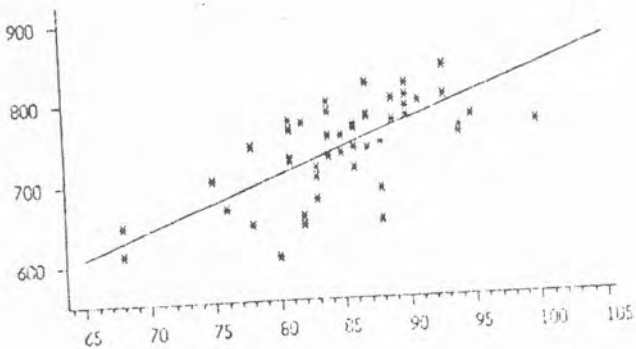
DECA: RESIDUAL (LOG) - CUM. FREQ (PROBIT)

A CORRELATION DIAGRAM OF THE DATA, ALSO A REGRESSION LINE (COMPUTED BY %)

DECA: WEIGHT - SHL

OPERATIONS FROM A POISSON DISTRIBUTION: FREQUENCY DISTRIBUTION OF THIS DATA

POISSON: $X - N(X)$



Automaattisen tietojenkäsittelyn kehitys on viime vuosikymmeninä vaikuttanut suuresti myös tilastotieteen tutkimusmahdollisuuksiin.

Tilastollisessa tutkimustyössä eivätkin riitä yksinomaan periaatteelliset ja teoreettiset tarkastelut vaan tarvitaan myös selviä laskennallisia tuloksia.

Tilastollisessa tietojenkäsittelyssä on ennen kaikkea kysymys tilastoaineistojen muokkauksesta ja analyysistä, mutta enenevässä määrässä on tilastollisiin ohjelmistoihin ryhdytty sisällyttämään myös muunlaisia toimintoja, jotka tukevat tilastollisten tutkimusten ja selvitysten läpivientiä.

Samoin tilastotieteen opetus edellyttää tilastollisilta ohjelmistoilta laadullisesti enemmän kuin mihin tavanomaisilla erillisillä ohjelmilla ja ohjelmapakkauksilla päästään.

Vaikka tilastollisissa tutkimuksissa käytetään usein mutkikkaitakin matemaattisia malleja, jolloin numeeristen menetelmien hallinta on tärkeää, tilastollinen tietojenkäsittely ei ole puhtaasti matemaattista luonteeltaan, vaan edellyttää ATK:n kaikkien osa-alueiden tuntemusta.

Helsingin yliopiston tilastotieteen laitoksella olemme viime vuosina tutkineet pientietokoneiden käyttömahdollisuuksia. 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 Wang 2200 uudentyypin "tilastollisen käyttöjärjestelmän" SURVO 76.

1.1 Tilastollisen tietojenkäsittelyn osa-alueet

Usein tilastollisen tietojenkäsittelyn alue rajoitetaan pelkästään tilastollisen data-analyysin laskentatehtävien toteuttamiseen ATK-laitteilla. Tässä yhteydessä tilastollista tietojenkäsittelyä tarkastellaan kuitenkin erityisesti tilastotieteen kannalta ja sen vuoksi käsite nähdään laajempaan.

Tilastollisen tietojenkäsittelyn katsotaan sisältävän ainakin seuraavia asioita ATK-laitteilla tehtynä:

1. Tilastolliset tietojärjestelmät:

tilastollisten aineistojen kerääminen, esittäminen, tarkastaminen ja ylläpito tietorekistereinä, tilastojen laadinta, tietojen hakujärjestelmät, otanta tietorekistereistä, tietojen siirrot, suurten tilastoaineistojen erityisongelmat,

2. Tilastolliset kuvaukset:

tilastoaineistojen, havaintoaineistojen, tilastojen, tilastollisten yhteenvedojen ja analyysitulosten numeerinen ja graafinen esittäminen, deskriptiiviset tilastolliset menetelmät,

3. Tilastollinen data-analyysi:

tilastollisten analyysimenetelmien edellyttämien laskentatehtävien toteuttaminen,

4. Simulointi:
tilastollisten ja stokastisten ilmiöiden simulointi (jäljittely)
ja seuranta,
5. Metodologinen tutkimus:
edellämainittujen kohtien hyväksikäyttö tilastollisten menetel-
mien tutkimuksessa, kokeilussa ja vertailussa,
6. Tilastotieteen opetus:
tilastollisen tietojenkäsittelyn hyväksikäyttö tilastotieteen ja
tilastollisten menetelmien opettamisessa,
7. Tilastolliset ohjelmistot ja käyttöjärjestelmät:
ohjelmistojen ja käyttöjärjestelmien tutkimus, suunnittelu,
kehittäminen ja vertailu.

Tämä jaottelu lienee jossain määrin epäjohdonmukainen (se sisältää eritasoisia ja osittain päällekkäin meneviä osia), mutta antaa kuitenkin käsityksen tilastollisen tietojenkäsittelyn alueen luonteesta ja laajuudesta.

1.2 ATK-sanastoa:

Kaikkia tietojenkäsittelytoimintoja voidaan tarkastella seuraavan pelkistetyn kaavion puitteissa:

INPUT --> PROCESSING --> OUTPUT
(syöttö) (käsittely) (tulostus)

Esim. Toteutettaessa yksinkertaista tilastollista tutkimusta ATK-laitteilla kaavion osilla voi olla seuraavat sisällöt:

INPUT: Havaintoaineiston kerääminen, lävistys reikäkortteille ja syöttö tietokoneeseen tai vaihtoehtoisesti suora syöttö tietokoneeseen elektronisten tietojentallennuslaitteiden avulla.

PROCESSING: Tietokoneelle syötetyn havaintoaineiston tarkastaminen ja muokkaaminen sekä tilastollisten analyysien laskentavaiheiden toteuttaminen.

OUTPUT: Saatujen tulosten siirto paperille kirjoittimilla ja/tai tallettaminen mahdollista jatkokäsittelyä varten.

Tässä tietojenkäsittelyprosessissa on tietenkin keskeinen asema laitteilla, joilla tapahtuu varsinainen laskeminen ja muu tiedon vaativampi muokkaus ja käsittely. Tietojenkäsittelylaitteita ovat laskimet, mikro- ja minitietokoneet sekä varsinaiset tietokoneet. Näitä kaikkia (myös ohjelmoitavia laskimia) pidetään seuraavassa "tietokoneina".

Tilastollisen tietojenkäsittelyn alueella kaikilla vaihtoehtoisilla laitteilla pienimmistä suurimpiin on arvokasta käyttöä. On erittäin harhaanjohtavaa katsoa, että vain 'suurilla' laitteistoilla voidaan tehdä 'suurta' tutkimusta. Valitettavasti ihmisten turhamaisuus ja vaillinaisen perehtyneisyys tietojenkäsittelyn mahdollisuuksiin on johtanut siihen, että tilastollisiakin tietojenkäsittelytehtäviä hoidetaan usein kalliilla ja epäkäytännöllisellä tavalla.

Vaikka automaattiseen tietojenkäsittelyyn tarkoitettut laitteet saattavat suuresti erota toisistaan tehokkuuden, nopeuden, kapasiteetin ja muiden ominaisuuksien suhteen, niille kaikille on ominaista mm. seuraavat piirteet:

- keskusyksikkö (prosessori), joka voi käsitellä tietoa käyttäjän määrittelemän tai valitseman ohjelman mukaisesti,
- keskusmuisti, joka tallettaa elektronisessa muodossa keskusyksikön tarvitsemat tiedot ja käsittelyyn tarvittavat ohjelmat. Yleensä tietokoneeseen liittyy vielä kapasiteetiltaan suurempia, mutta toiminnaltaan hitaampia magneettilevy- ja -nauhamuisteja (massamuisteja),
- ympäryslaitteet, joilla tapahtuu käsiteltävien tietojen syöttö ja tulostus samoin kuin käytettävien ohjelmien määrittely, valinta ja ulkoinen hallinta.

"Tietokoneelle" edellä esitetyssä mielessä on ennenkaikkea ominaista, että sen keskusyksikkö ei ole rajoittunut toimimaan vain yhden tai useamman ennalta määrätyn ohjelman mukaisesti (kuten esim. automaattinen pesukone), vaan sitä voidaan jatkuvasti "ohjelmoida" uudelleen uusia tehtäviä varten.

Tyypillistä on, että puhuttaessa tietokoneista tietojenkäsittelyjärjestelminä niissä erotetaan:

- hardware (kova), laitteen fyysinen toteutus keskusyksikköineen, muisteineen ja ympäryslaitteineen,
- software (pehmo), laitteen käyttöjärjestelmä ja varusohjelmisto, joita voidaan jatkuvasti kehittää muuttamatta itse laitetta ja joiden avulla tehdään yleensä kaikki laitteelle tarkoitettut ohjelmat ja sovellutusohjelmistot.

Tietokoneella teknisenä toteutuksena ilman hyvää käyttöjärjestelmää ja varusohjelmistoa ei ole paljon arvoa. Tietojenkäsittelykustannuksissa ohjelmistoilla on yhä enenevä merkitys.

Käyttötavan suhteen voidaan erottaa seuraavia toimintamuotoja:

- suora käyttö, koneen välitön käyttö, tyypillistä laskimille, mikrotietokoneille ja osittain myös minikoneille.
- osituskäyttö, tietokoneen käyttö kirjoituskone- tai näyttöpäätteen välityksellä siten, että keskuslaitteisto palvelee samanaikaisesti useita päätteitä ja käyttäjiä. Käyttö tapahtuu "keskustelemalla" keskuskoneen kanssa käyttöjärjestelmän välityksellä. Tulokset saadaan välittömästi päätteelle. (tyypillinen käyttötapa pienissä tehtävissä)
- (etäis)eräkäyttö, laskentatehtävän määrittely tapahtuu tavallisesti päätteeltä, mutta keskuskone ei välttämättä suorita tehtävää välittömästi, vaan työ joutuu odottamaan jonossa. (tyypillinen käyttötapa suurissa tehtävissä)

Puhuttaessa "tiedosta" tässä yhteydessä on syytä korostaa, että on luonnollista lukea tiedoksi yhtäläillä käsiteltävät tiedot eli datat kuin ohjelmat, joilla datojen käsittely tapahtuu. Tietojenkäsittelyprosessin aikana tiedot talletetaan joko väliaikaisesti tai pysyvästi käyttöjärjestelmän avulla ylläpidettyihin tiedostoihin massamuistilaitteissa.

Ohjelmoinnilla tarkoitetaan niiden ohjeitten laatimista, joilla tietokone saadaan toteuttamaan haluttu tehtävä. Ohjelmointi tapahtuu kirjoittamalla ja syöttämällä koneeseen ohjelma, joka on kirjoitettu jollakin ohjelmointikielellä.

Tavallisesti ohjelmat laaditaan kielellä, jota kone ei välittömästi "ymmärrä" ja ohjelma joudutaan (usein monessa vaiheessa) kääntämään koneen omalle "konekielelle". Tätä käännöstehtävää varten koneiden varusohjelmistoon kuuluu kääntäjiä, joilla ennen ohjelman toteutusta eli ajoa saadaan vaadittu käännös aikaan.

Toisena vaihtoehtona ovat ns. tulkkiohjelmat, jotka käsittelevät käyttäjän laatimaa ohjelmaa sen alkuperäisessä muodossa ja välittävät suoraan vaihe vaiheelta tarvittavat toimenpiteet prosessorin toteuttaviksi.

Ohjelmiksi voidaan laajassa mielessä lukea koko pehmo, siis koneen systeemi-ohjelmat, käyttöjärjestelmä, käännös- ja tulkki-ohjelmat, varusohjelmistot, sovellutusohjelmistot ja käyttäjän omat ohjelmat.

Jokainen ohjelma koostuu käskyistä tai lauseista, jotka toteutetaan yleensä esiintymisjärjestyksessä, ellei esim. ns. hyppykäskyillä tai ohjelmakierroskäskyillä toisin osoiteta. Monien ohjelmointikielten käskyt ja lauseet ovat lähes selväkielisiä, jolloin ohjelmaa on suhteellisen helppo lukea ja tarkastaa.

Seuraavassa esitetään pelkistetty esimerkki tilastollisesta ohjelmasta, joka on laadittu yksinkertaisella BASIC-ohjelmointikielellä. Mainittakoon, että koko SURVO 76-ohjelmisto on laadittu BASIC-kielen erästä laajennusta käyttäen.

Olkoon tehtävänä laatia ohjelma, joka laskee keskiarvon havainnoista $X_1, X_2, \dots, X_n$.

Oletamme, että havaintojen lukumäärä n tunnetaan ja se voidaan antaa havaintoarvojen alussa.

Seuraava BASIC-ohjelma toteuttaa vaaditun tehtävän:

```
10 INPUT N
20 S=0
30 FOR J=1 TO N
40 INPUT X
50 S=S+X
60 NEXT J
70 PRINT "KESKIARVO=";S/N
80 END
```

Ohjelman toiminta on tässä seuraava: Ensimmäinen käsky INPUT N pyytää käyttäjältä havaintojen lukumäärän N . Toinen käsky $S=0$ tyhjentää havaintoarvojen summaukseen tarkoitetun S -muuttujan; summa S on siis aluksi nolla. Kolmannesta käskystä riviltä 30 alkaa ohjelmakierros, joka toistetaan N kertaa J :n arvoilla $1, 2, \dots, N$. Joka kierroksella pyydetään uutta havaintoarvoa X (INPUT X rivillä 40), joka lisätään muuttujan S aikaisempaan arvoon käskyllä $S=S+X$. Käsky NEXT J osoittaa vain, mihin asti rivillä 30 alkanut ohjelmakierros yltää. Kun siis ohjelmakierros on tässä toistettu N kertaa, tullaan riville 70, jolla tulostetaan lainausmerkeissä oleva teksti KESKIARVO= ja tämän jälkeen numeerinen keskiarvo S/N eli summa jaettuna havaintojen lukumäärällä.

Tämän ohjelman laatiminen ja käyttö voi käytännössä tapahtua esim. seuraavasti:

Käyttäjä kirjoittaa päätteeltä ylläolevan ohjelman rivi riviltä. Kun koko ohjelma on kirjoitettu, kirjoitetaan vielä komentosana RUN, joka käynnistää ohjelman ajon. Tässä tapauksessa ohjelma pyytää ensiksi havaintojen lukumäärän kirjoittamalla kysymysmerkin ? ja jättämällä odottamaan käyttäjän vastausta. Kun käyttäjä on vastannut kirjoittamalla havaintojen lukumäärän, ajo jatkuu ja ohjelma pyytää ensimmäistä havaintoa jne. Kun kaikki havainnot on saatu käyttäjältä, ohjelma tulostaa keskiarvon. Toiminta tämän ohjelman kanssa voisi siis näyttää esim. seuraavalta:

(ohjelma kirjoitettu)

RUN

? 4

? 12.1

? 15

? 16.8

? 10.3

KESKIARVO= 13.55

Ylläoleva ohjelma oli vain voimakkaasti pelkistetty esimerkki tilastollisesta ohjelmasta. Vaikka se kelpaa tilapäistarkoituksiin, se on todellisen käytön kannalta puutteellinen mm. seuraavista syistä:

Joka kerran tätä ohjelmaa käytettäessä tulee antaa erikseen kaikki havaintoarvot eikä näitä arvoja talleteta mahdollisia lisälaskentoja varten. Toinen heikkous on se, ettei ohjelmaa tuntematon käyttäjä voi ajon alettua tietää, mitä on tapahtumassa eikä mitä kulloinkin kysytään. Kolmantena puutteena voi pitää sitä, ettei ohjelmaan ole sisällytetty minkäänlaisia varotoimenpiteitä mahdollisten järjettömien vastausten tai havaintoarvojen osalta.

Vaikka kaikki nämä puutteet voidaan korjata tässä yksittäisessä ohjelmassa, parempi ratkaisu on kuitenkin saavutettavissa siten, että synnytetään ensin yhtenäinen ja yleinen esitys- ja käsittelytapa kaikille tilastollisille aineistoille ja rakennetaan tämän varaan ohjelmakokonaisuus, joka kattaa mahdollisimman monet tilastollisen tietojenkäsittelyn tarpeet.

1.3 Tilastolliset tietojenkäsittelyjärjestelmät

Tilastolliset tietojenkäsittelyjärjestelmät ovat yleisohjelmistoja, jotka tekevät mahdolliseksi tilastolliseen tietojenkäsittelyyn liittyvät toiminnot yleensä suoraan ilman, että käyttäjän tarvitsisi laatia omia ohjelmia tai edes hallita ohjelmointia.

Useimmat yleisessä käytössä olevista tilastollisista ohjelmistoista keskittyvät erityisesti data-analyysin tehtäviin. Näistä tunnetuimpia ovat SAS, SPSS ja BMDP, kotimaisista HYLPS ja SURVO/71.

Tilastollisten tietojenkäsittelyjärjestelmien etuna on erillisiin ohjelmiin verrattuna mm. seuraavat piirteet:

- standardoidut tietorakenteet, jotka sopivat tilastolliseen laskentaan. Tällä tarkoitetaan mm. sitä, että havaintoaineistot talletetaan ennen varsinaisen laskennan alkua kokonaisina standardimuodossa havaintotiedostoihin, joista ne on saatavissa suoraan eri ohjelmille. Myös yleisesti esiintyville välituloksille (kuten frekvenssijakaumat ja -taulukot, korrelaatiomatriisit) on monissa ohjelmistoissa vakioesitysmuodot.

- operaatioiden ketjuttaminen, jolla tarkoitetaan mahdollisuutta toteuttaa suoraan monivaiheisia tilastollisia analyyseja siten, että edellisen vaiheen tulokset ovat seuraavien vaiheiden lähtötietoina. Ketjuttamisessa on suuresti apua standardoiduista tietorakenteista.
- ohjauskieli, jolla kuvataan tarvittavat tilastollisen tietojenkäsittelytoiminnot.

Esim. SURVO 66-järjestelmässä, joka on SURVO-nimisten ohjelmistojen kantamuoto, kuvattiin korrelaatioiden laskeminen muuttujista X1-X20 ja sellaisista havainnoista, joissa X24=1 (,joka tarkoittakoon tässä sitä, että kysymyksessä on tyyppiä MIES oleva havainto) seuraavasti:

```
EQUAL@ MIES X24 1
CORREL@ X1-X20 IF:MIES
```

Vastaavanlainen ohjauskieli on käytössä myös esim. Helsingin yliopiston laskentakeskuksen ylläpitämässä HYLPS-ohjelmistossa.

Ytimekkään ohjauskielen käyttö tilastollisen ohjelmiston yhteydessä on luonnollinen ratkaisu erityisesti etäisieräkäyttömuodossa. Sen sijaan, mikäli on mahdollista toimia vuorovaikutteisesti siten, että tarvittavat toimenpiteet saadaan toteutetuiksi välittömästi, ohjauskieli voidaan korvata normaalikielen käyttöön perustuvalla keskustelulla. Tämä on mahdollista osituskäytössä silloin, kun koneen vastausajat ovat kohtuullisen lyhyitä (ei siis liikaa kuormitusta keskustusneella) ja käytettävissä on riittävän nopea kuvaputkinäyttölaitte.

Helppimmin kuitenkin vielä tällä hetkellä on aito vuorovaikutteinen eli interaktiivinen työskentelytapa toteutettavissa mikro- ja minitietokoneilla. Tällainen ratkaisu on SURVO 76-ohjelmisto, jota seuraavassa kuvataan ja käytetään esimerkkijärjestelmänä.

2. SURVO 76-järjestelmä

SURVO 76 on keskustelevala järjestelmä eikä sitä käytettäessä tarvita mitään erityistä ohjaus- tai työmäärityskieltä, vaan järjestelmä toteuttaa tehtävät tekemiinsä suomen- (tai englannin) kielisiin kysymyksiin saamiensa vastausten perusteella. Järjestelmän peruskäyttötavat eivät edellytä myöskään minkään ohjelmointikielen taitoa.

Tilastollisen tietojenkäsittelyn yleisperiaatteiden tunteminen on kuitenkin tarpeen ja aivan välttämätöntä on tarvittavien tilastollisten menetelmien soveltamistaito. Käyttäjän on joka vaiheessa tiedettävä, mitä hän haluaa tehdä ja ilmaistava halunsa tekemällä aloitteita. Tämä tapahtuu nykyisessä WANG 2200-laitteistolle laaditussa toteutuksessa painamalla koneen ohjauspöydässä olevia ns. funktionappeja, joiden toimintaa voidaan säädellä ohjelmakohtaisesti. SURVO 76:ssa funktionappeja käytetään aloituskohtina kulloinkin käytössä olevan tilastollisen menetelmän (SURVO 76-modulin) ja työskentelyä valvovan SURVO 76-systeemiohjelman eri toimintoihin.

Jokainen tilastollinen menetelmä on SURVO 76:ssa paloiteltu suoritusjaksoihin, joita eri tavoin yhdistelemällä käyttäjä saa toteutetuksi menetelmän edellyttämät laskentatoimenpiteet haluamallaan tavalla.

Käyttäjälle itselleen jää tällöin vastuu valitseman suoritussyjärjestyksen mielekkyydestä. Olisi tietenkin helppo kytkeä osat yhteen johonkin vakiojärjestykseen, mutta se merkitsisi samalla käyttäjän joutumista systeemin armoille tavalla, joka on tyypillistä monille tilastollisille "pakkauksille". Näin menetettäisiin myös se kasvattava ominaisuus, joka järjestelmällä on velvoittaessaan käyttäjän omaehtoiisiin valintoihin.

Tavoitteena on siis, että järjestelmää pystytään käyttämään luovalla tavalla eikä pelkästään totuttuja latuja, tavanomaisia analyysiketjuja toistaen.

SURVO 76:n rakennetta hahmoteltaessa ja toteutettaessa on pyritty pitämään huolta myös vaativien soveltajien tarpeista. Järjestelmä tarjoaa tukea uusien ohjelmien teolle, sillä siihen sisältyy valmiita aliohjelmiä, joilla esim. hoidetaan havaintoaineiston siirrot ja esikäsittely. Tilastollisen ohjelman laatija voi keskittyä oman ohjelmansa erityispiirteiden selvittämiseen ja jättää rutiinitoimitukset järjestelmän huoleksi. Uudet ohjelmat voidaan myös helposti liittää järjestelmän varsinaisiksi osiksi.

Vaatiivaan käyttöön kuuluu mahdollisuus muuttaa ja laajentaa valmiita järjestelmän moduleja. Se on nykyisessä toteutuksessa melko helppoa, mutta edellyttää luonnollisesti kyseisen modulin toiminnan tuntemista ja ohjelmointitaitoa.

Pienempien muutosten teko kuuluu jopa normaalityöskentelyyn. Esim. muuttujatransformaatiot tai epälineaarisen mallin muodon määrittely regressioanalyysissä hoidetaan siten, että käyttäjä lisää tarvittavat käskyt ohjelmaan järjestelmän neuvomalla tavalla.

SURVO 76 on laadittu kokonaisuudessaan laajennettulla BASIC-kielellä, joka on WANG 2200:n ainoa ohjelmointikieli. Vastoin yleistä käsitystä tämä laajennettu BASIC on osoittautunut erittäin sopivaksi näinkin laajan ohjelmiston laadinnassa ja ylläpidossa. SURVO 76 on muodollisesti yksi ainoa BASIC-ohjelma ja käsittää tällä hetkellä hiukan alle miljoona tavua.

Seuraavaksi kuvataan lyhyesti eräitä tyypillisiä SURVO 76-toimintamuotoja.

2.1 Data-analyysi

SURVO 76-järjestelmän rungon muodostavat luonnollisesti tilastolliseen data-analyysiin tarkoitetut ohjelmamodulit. Suunnittelun ja toteutuksen alkuvaiheissa pääpaino oli standardimenetelmissä, mutta nykyisin kehitystä pyritään suuntaamaan vaativampiin erikoismenetelmiin.

Järjestelmä sisältää ohjelmia mm. seuraaviin data-analyysin muotoihin:

- perusjakaumat, taulukointi, perustunnusluvut
- havaintojen lajittelu, järjestystunnusluvut
- tilastolliset testit, otantajakaumat, testien kriittiset arvot
- korrelaatioanalyysi
- lineaarinen ja epälineaarinen regressioanalyysi
- pääkomponentti- ja faktorianalyysi
- erottelu-, luokittelu- ja ryhmittelyanalyysi
- aikasarja-analyysi

SURVO 76-ohjelmistossa on varauduttu myös puuttuvien havaintoarvojen käsittelyyn, se sisältää robusteja estimointitekniikkoja ja poikkeavien havaintojen tunnistuskeinoja.

Analyysekeinojen käytön tehokkuutta lisää olennaisesti se, että eri menetelmiä ja analyysien suorituvaiheita voi vapaasti kytkeä toisiinsa. Käyttö on aidosti vuorovaikutteista ja jokainen suorituvaihe tapahtuu käyttäjän ohjaamana välittömästi.

Erityistä huomiota on kiinnitetty havaintoaineistojen tallennukseen, editointiin, muunnoksiin ja muihin esikäsittelytoimintoihin. Koko aineistojen hallinta perustuu havaintomatriisimuotoiseen, standardoituun esitystapaan (havaintotiedostot), mikä takaa sen, että kaikki SURVO 76:n analyysi- ja muut aineiston käsittelytoimenpiteet voidaan kohdistaa mihin tahansa havaintotiedostoon. Myös tärkeimmille välituloksille (esim. korrelaatiomatriisit) on omat standardinsa, mikä yksinkertaistaa eri analyysohjelmien yhteistyötä.

Koska työskentely on vuorovaikutteista, on tärkeätä, että käyttäjä voi koko ajan seurata tapahtumia. Tässä seurannassa ja tulosten tarkastelussa auttaa se, että sekä muuttujat että havainnot voidaan SURVO 76-järjestelmän havaintotiedostoissa varustaa selväkielisin nimin.

Jokaisen SURVO 76-modulin oletetaan jatkuvasti kertovan näyttölaitteen välityksellä analyysin edistymisestä. Esim. aineistoa käsiteltäessä on tavallista, että kuvaputkella vilisevät käsittelyn kohteena olevien havaintojen nimet ja tärkeimmät välitulostiedot. Tarkoitus ei ole, että käyttäjä ehtisi lukea kaiken, mitä kuvaputkella näkyy, vaan tavallisesti riittää summittainen vaikutelma. Kuitenkin silloin, kun jokin tuntuu menevän odottamattomaan suuntaan, on hyvä väliaikaisesti keskeyttää toiminta ja pysähtyneestä "tajunnanvirrasta" havaita välittömästi, missä ollaan menossa. Tästä ominaisuudesta on erityisesti hyötyä uusiin ohjelmamoduleihin tutustuttaessa ja menetelmien ominaisuuksia kokeiltaessa. Tarvittaessa voi kuvaputkitulostuksen vauhdin alentaa normaalin lukemisnopeuden tasolle.

2.2 Graafiset esitykset

Kuvalliset keinot ovat varsin hyödyllisiä tilastollisessa tietojenkäsittelyssä. Juuri näiden osalta monet tilastolliset vakio-ohjelmistot ovat puutteellisia. SURVO 76:ssa voidaan piirtää esim. korrelaatiodiagrammeja, histogrammeja, aikasarjoja, analyyttisiä käyriä ja pintoja (perspektiivikuvina) sekä eräitä erikoisempia tilastollisia graafisia esityksiä. Tämä voi tapahtua suoraan tavallisella näyttölaitteella tai kirjoittimella, mutta parhaiten piirturin tai graafisen näyttölaitteen avulla.

Graafiset ohjelmamodulit on SURVO 76:ssa pyritty laatimaan niin, ettei käyttäjän tarvitse huolehtia toisarvoisista asioista. Niinpä esim. piirroksien skaalaus ja koordinaattiakseleille tulevat merkin-
nät voidaan tarvittaessa valita ohjelman antamien suositusten mukai-
sesti.

Myös mikä tahansa epälineaarinen asteikko on saatavissa käyttöön määrittämällä asteikkoa vastaava epälineaarinen muunnos aliohjelman-
na ja näin käyttäjä voi helposti piirtää esim. erityyppisille
logaritmi- ja todennäköisyyspapereille. Eri piirtämisrutiineja voi
yleensä vapaasti yhdistellä, jolloin on mahdollista synnyttää vai-
heittain hyvinkin mutkikkaita graafisia esityksiä ja jopa (graafisel-
la näyttölaitteella) korjailla niitä.

Tyypillistä SURVO 76-työskentelyä on esim. seuraava: Tutkittaessa
kahden muuttujan riippuvuutta piirretään aluksi niiden yhteisjakaumaa
kuvaava korrelaatiodiagramma. Tarkasteltaessa riippuvuuden luonnetta
päädytään sitten esim. johonkin lineaariseen tai epälineaariseen reg-
ressiomalliin. Lasketaan mallin parametrit ja piirretään lopuksi reg-
ressiokäyrä ja luottamusrajat alkuperäiseen piirrookseen.

SURVO 76:n modulaariselle rakenteelle on myös ominaista, ettei
esim. regressiokäyrien piirtämistä ohjata regressioohjelmasta, vaan
kaikki käyrät tehdään omalla tähän tarkoitukseen laaditulla erikois-
ohjelmalla. Tämä ratkaisutapa takaa huomattavasti monipuolisemmat
piirtämismahdollisuudet ja on parempi myös oppimismielessä, koska
käyttäjä tietää aina, mihin turvautua, kun esim. käyränpiirtämisestä
on kysymys.

2.3 Matriisilaskenta

Pientietokoneilla on ylimääräisten laskutoimitusten toteuttaminen
erilaisten ohjelmallisten toimintojen lomassa varsin tavallista ja
tämä koskee rajoitetusti myös matriisioperaatioita. Tilastollisessa
tietojenkäsittelystä esiintyy kuitenkin tehtäviä, joissa nämä suorat
aritmeettisten ja matriisilausekkeiden laskemiset eivät riitä.

Syntyy tilanteita, joissa käyttäjä haluaa välittömästi jalostaa
edelleen standardiohjelmilla saatuja tuloksia. SURVO 76 sisältää näi-
tä tarkoituksia varten laajan osajärjestelmän nimeltään MATRI, jonka
avulla voidaan toteuttaa normaalit tilastollisen tietojenkäsittelyn
yhteydessä esiintyvät matriisitoiminnot pöytälaskukonemaisesti siten,
että jokaista matriisioperaatiota vastaa oma "nappulansa".

Käsiteltävät matriisit voidaan antaa manuaalisesti (täytetään kuva-
putkella oleva taulukko matriisin alkioilla) tai matriiseja saatetaan
paimia käyttöön myös suoraan erilaisista SURVO 76-tiedostoista. Tulok-
set voidaan vastaavasti tallettaa erityisiin matriisitiedostoihin
jatkokäsittelyä varten.

MATRI huolehtii paitsi laskemisesta myös siitä, että esim. tulos-
matriisit varustetaan normaaleja matriisimerkintöjä vastaavilla ni-
millä. Jos siis esim. matriisit A ja B+C kerrotaan keskenään, tulosta
merkitään nimellä A(B+C). Käyttäjä voi vaiheittain synnyttää mutkik-
kaitakin matriisilausekkeitä ja aste asteelta seurata kuvaputkelta
työnsä edistymistä sekä formaalisella että numeerisella tasolla. Myös
matriisien rivit ja sarakkeet voi nimetä selväkielisesti ja MATRI
liikuttaa niitä operaatioiden luonteen mukaisesti siten, että tu-
losmatriisien otsikot ovat järkeviä.

MATRI pitää huolta siitä, että laajat matriisit tulostuvat auto-
maattisesti lohottuna tulostuslaitteen leveyttä vastaaviin osiin ja
että lohkoilla on täydelliset rivi- ja sarakeotsikot.

Jokainen matriisioperaatioiden muodostama ketju voidaan myös haluttaessa toistaa uusilla lähtömatriiseilla ja tallettaa levyllä käyttöohjeineen. Näin voidaan rakentaa yksinkertaisia matriisiohjelmia esim. linearisten mallien ja monimuuttujamenetelmien yhteydessä tarvittavia lisälaskentoja varten.

2.4 Simulointi

SURVO 76-järjestelmä ei ole tarkoitettu varsinaisiin simulointitehtäviin sanan laajassa mielessä, mutta se tarjoaa eräitä simulointiin liittyviä toimintamahdollisuuksia, joista on hyötyä tilastollisen tietojenkäsittelyn muiden toimintojen yhteydessä. Varsinkin tilastollisten menetelmien opetuksessa ja tutkimuksessa on havaittu hyödylliseksi käyttää keinotekoisia aineistoja erilaisten tilastollisten menetelmien toiminnan havainnollistamiseen ja vertailuun sekä yksinkertaisesti myös stokastisten ilmiöiden käyttäytymisen seurantaan.

SURVO 76:n eräät ohjelmamodulit (tärkein on SATTUMA) auttavat tällaisten kokeiden suunnittelussa. Halutessaan synnyttää esim. määrätyn regressiomallin mukaista aineistoa käyttäjä määrittelee mallin rakenteen SATTUMA:n antamien ohjeiden mukaisesti. Apuna on saatavilla eri perusjakaumia simuloivia aliohjelmia.

Simuloidut aineistot talletetaan levyllä havaintotiedostoiksi ja niitä on mahdollista analysoida kuten reaalisiaakin havaintoaineistoja.

Myös erilaisten perusjakaumien ja käyttäjän määrittelemien erikoisjakaumien käyttäytymistä on helppo havainnollistaa kuvaputkella. Käyttäjä valitsee tai määrittelee jakauman ja SATTUMA alkaa muodostaa havaintoja, jotka synnyttävät kuvaputkelle alati kasvavan, empiiristä jakaumaa kuvaavan histogramman.

2.5 Tekstinkäsittely

Vaikka SURVO 76-ohjelmistoa kehitettäessä on kiinnitetty paljon huomiota siihen, että analyysien antamat tulokset saadaan paperille helposti siistissä muodossa, vakio-ohjelmilla on kuitenkin mahdollonta saavuttaa tulosten esittämisessä jokaisen käyttäjän kannalta optimaalista esitysmuotoa ja tutkijat joutuvat usein tutkimusselostusta laatatiessaan muokkaamaan tulostaulukoita ja luonnollisesti myös kirjoittamaan selostuksen sanallisen osan.

Jonkin aikaa on kuitenkin jo harkittu SURVO 76:n "toimialan" laajennusta sanojenkäsittelyn suuntaan siten, että normaalin SURVO 76-keskustelun yhteydessä olisi mahdollista muotoilla tulostuksia ja kirjoittaa tekstiä laskelmien edetessä.

Tätä varten on nyt kehitetty SURVO 76:n yhteyteen yleinen tekstinkäsittelyohjelma, joka on hyödyllinen tutkimusraporttien laadinnassa. Käyttäjän on mahdollista vaihteittain kirjoittaa kuvaputkelle vapaa-muotoista tekstiä, yhdistää siihen välittömästi SURVO 76-tulostusta ja muokata raporttiaan erilaisilla editointioperaatioilla. Teksti voidaan tallettaa datalevylle ja halutut kohdat tulostaa paperille.

Tekstin ja tulostaulukkojen editointi tapahtuu editointikentässä, jonka laajuuden käyttäjä voi itse valita ja josta kuvaputkella näkyy vain osa. Käyttäjä voi siirtää helposti kuvaputkella näkyvää editointikentän osaa kaikkiin suuntiin.

Editointikenttään voi koska tahansa kirjoittaa uutta tekstiä ja kirjoitettua tekstiä voi editoida vapaasti käyttäen hyväksi funktio-nappeja samaan tapaan kuin Wangin BASIC-ohjelmia editoitaessa.

Lisäksi on mahdollista kirjoittaa editointikentän kaikille riveille editointioperaatioita, joilla hoidetaan vaativimmat toimenpiteet. Tällaisia ovat mm.

- kirjoitetun tekstin taittaminen haluttuun leveyteen,
- numeeristen taulukkojen muotoilu ja muokkaus,
- numeeristen taulukkojen talletus SURVO 76-tiedostoiksi,
- tulostaulukkojen haku editointikenttään SURVO 76-tiedostoista,
- editointikentän rivien numeerinen ja aakkoslaajittelu,
- editointikentän valittujen osien tulostus paperilla.

(Esim. nyt näkyvillä oleva teksti on kirjoitettu ja työstetty tällä ohjelmalla. Lopullinen taitto ja tulostus paperille on tapahtunut yhdellä ainoalla PRINT-operaatiolla.)

Kaikki tarvittavat editointioperaatiot kirjoitetaan myös editointikenttään ja niitä voidaan käsitellä samaan tapaan kuin tulostaulukkoita ja vapaata tekstiä. Samoin koko kenttä voidaan tallettaa operaatioineen levyille ja ottaa myöhemmin käyttöön uusia toimenpiteitä varten.

Tekstinkäsittelytoimintojen yhdistäminen tilastolliseen tietojenkäsittelyyn lisää monin tavoin varsinkin henkilökohtaisen ATK:n hyödyntämismahdollisuuksia.

Sen lisäksi, että se parantaa tutkimusselostusten ja muiden raporttien laatimista, myös tilastollisen tietojenkäsittelyn normaaleja toimintoja voidaan ottaa käyttöön uudella tavalla. Varsinkin useat yksinkertaiset tilastolliset operaatiot voidaan toteuttaa suoraan editointikentän muodostamassa tilassa editointioperaatioiden kaltaisesti.

2.6 Dokumentointi

Tietokoneohjelma tai -ohjelmisto vailla riittäviä käyttöohjeita on arvoton yleisessä käytössä. Interaktiivisissa ohjelmistoissa itse ohjelmateksti sisältää tavallisesti siinä määrin tarvittavaan keskusteluun viittaavia ilmaisuja, että ohjelman listaus saattaa jo kertoa varsin paljon ohjelmien toiminnasta.

Niinpä SURVO 76-ohjelmistoakin voidaan pitää varsin pitkälle itse-dokumentoivana. Varsinkin järjestelmän rakenneperiaatteet tunteva ja ohjelmointitaitoinen henkilö voi listaamalla ohjelmaa kuvaputkelle tai paperille etsiä tarvitsemansa tiedot.

SURVO 76-ohjelmistolle on lisäksi ominaista, että ainakin poikkeuksellisista toiminnoista on itse työskentelyn alkaessa saatavissa kuvaputkella riittävät ohjeet. Toisaalta laajempia aihekokonaisuuksia varten (esim. järjestelmän peruskäyttötavat ja matriisilaskenta) on omat opetusohjelmansa.

Muuten luotetaan siihen, että käyttäjä epäselvissä tapauksissa pyrkii itse hakemaan oikean menettelytavan käyttäen hyväksi aikaisempia kokemuksiaan ja "konetta pelkäämättä" kokeilee vaihtoehtoisilla tavoilla.

SURVO 76 ei ole "helppo" järjestelmä siinä mielessä, että se tekisi automaattisesti kaikki asiat käyttäjän puolesta, vaan päinvastoin se vaatii käyttäjää ottamaan koko ajan kantaa vaihtoehtoihin menettelytapoihin ja tekemään aloitteita. Toisaalta se tarjoaa tavanomaista runsaammin päätöksentekoa helpottavaa tietoa ja ehdotuksia.

SURVO 76-ohjelmiston toimintamahdollisuuksista on melko vaikea saada tyhjentävä selvitys kirjallisten dokumenttien avulla. Tämä johtuu järjestelmän dynaamisesta luonteesta, jota ei ole helppo havainnollistaa paperilla.

Sen vuoksi on katsottu hyväksi dokumentointitavaksi laatia erillisiä automaattisia esittelyohjelmia, jotka sisältävät valmiita kuvitellun käyttäjän ja SURVO 76-järjestelmän välisiä keskusteluja konkreettisista tilastollisista aiheista selityksin varustettuna.

Käyttäjä voi seurata näitä SURVO 76-keskusteluja kuten televisio-ohjelmaa, mutta hän voi myös halutessaan puuttua itse keskusteluun ja jatkaa sitä omalla tyyliällään.

SURVO 76 on nykyisellään jo varsin laaja tilastollinen järjestelmä. Siihen kuuluu noin 60 erilaista tilastollista ohjelmaa ja osakokonaisuutta, joista eräät (kuten esim. MATRI) muodostuvat vuorostaan monesta erillisestä ohjelmamodulista. Järjestelmän laajuutta kuvaa sekin, että tietolevyjä (ä 262 kilotavua) käytettäessä tarvitaan 7 levyä itse ohjelmistoa ja 5 levyä esittelyohjelmia ja muuta dokumentointia varten.

Järjestelmän kehitystyö tapahtuu nykyisin englanninkielellä, mutta siitä on jatkuvasti käytettävissä myös rinnakkainen suomenkielinen perusversio.

4. SURVO 76-kuvauksia

Kuten jo edellä kohdassa 2.2 todettiin, kuvioiden käyttö tilastollisen tutkimuksen yhteydessä on varsin hyödyllistä. Yksi kuva saattaa kertoa jopa enemmän kuin mutkikkaat laskennalliset analyysit, joihin tutkijat helposti ajautuvat sopivien graafisten keinojen puuttuessa.

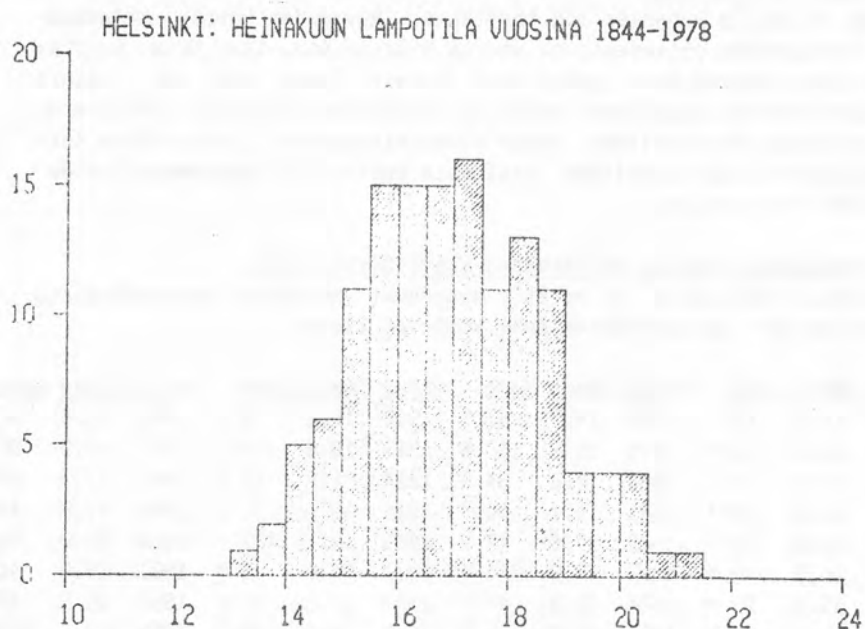
Seuraavassa tarkastellaan paria esimerkinomaista tilastollista tutkimustilannetta ja esitellään erilaisia kuvien piirtämismahdollisuuksia SURVO 76:n avulla.

4.1 Helsingin säätila heinäkuussa vuosina 1844-1978

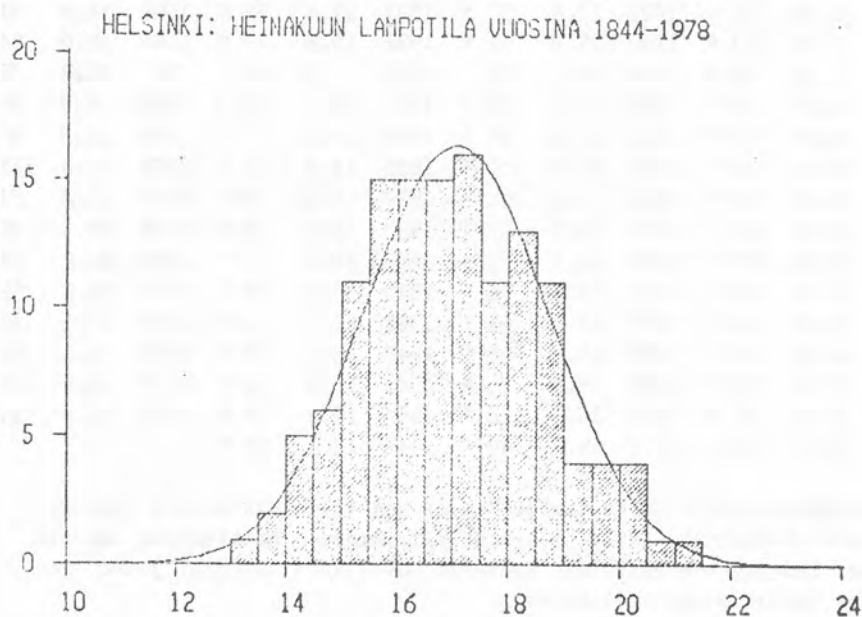
Helsingin säätilasta on kerätty seuraavat heinäkuun keskimääräistä lämpötilaa (C°) ja sademäärää (mm) koskevat tiedot:

vuosi	lämpö	sade	vuosi	lämpö	sade	vuosi	lämpö	sade	vuosi	lämpö	sade
1844	14.3	65 *	1878	14.3	57 *	1912	18.1	8 *	1946	18.8	35
1845	16.5	30 *	1879	15.5	148 *	1913	18.9	36 *	1947	17.9	95
1846	17.2	32 *	1880	16.2	84 *	1914	21.5	12 *	1948	17.9	25
1847	15.5	32 *	1881	15.1	70 *	1915	17.6	76 *	1949	17.3	49
1848	14.5	44 *	1882	17.0	44 *	1916	19.0	32 *	1950	15.8	48
1849	14.6	86 *	1883	15.9	204 *	1917	16.8	8 *	1951	15.6	40
1850	16.9	27 *	1884	16.9	49 *	1918	17.9	74 *	1952	16.3	45
1851	16.0	48 *	1885	18.6	65 *	1919	19.0	38 *	1953	17.1	121
1852	16.5	6 *	1886	17.1	92 *	1920	18.3	61 *	1954	17.7	78
1853	17.2	16 *	1887	16.4	26 *	1921	14.7	61 *	1955	18.3	16
1854	18.5	35 *	1888	15.6	75 *	1922	16.7	66 *	1956	16.0	104
1855	20.0	2 *	1889	16.0	83 *	1923	17.0	36 *	1957	18.2	77
1856	15.1	42 *	1890	15.4	78 *	1924	17.2	28 *	1958	15.5	73
1857	15.3	70 *	1891	17.7	27 *	1925	20.3	87 *	1959	18.7	55
1858	19.7	42 *	1892	14.8	122 *	1926	17.8	26 *	1960	18.2	110
1859	16.1	92 *	1893	16.1	51 *	1927	21.0	25 *	1961	16.5	65
1860	17.5	107 *	1894	17.0	68 *	1928	13.7	51 *	1962	15.3	53
1861	19.4	38 *	1895	16.1	118 *	1929	15.6	70 *	1963	16.7	31
1862	13.1	128 *	1896	19.1	33 *	1930	18.6	84 *	1964	17.2	36
1863	14.3	70 *	1897	17.5	67 *	1931	17.4	53 *	1965	14.9	61
1864	17.6	37 *	1898	16.3	71 *	1932	19.9	47 *	1966	18.0	54
1865	17.9	55 *	1899	19.4	27 *	1933	17.9	61 *	1967	16.8	25
1866	15.0	80 *	1900	15.6	88 *	1934	18.5	64 *	1968	15.7	59
1867	15.4	103 *	1901	20.0	34 *	1935	16.8	77 *	1969	16.9	67
1868	18.4	6 *	1902	13.9	103 *	1936	18.9	71 *	1970	16.3	111
1869	16.0	45 *	1903	16.2	91 *	1937	18.3	58 *	1971	17.2	21
1870	17.3	100 *	1904	14.2	37 *	1938	18.9	29 *	1972	20.4	80
1871	17.3	122 *	1905	16.6	78 *	1939	18.1	77 *	1973	20.2	24
1872	17.5	29 *	1906	18.2	56 *	1940	18.6	29 *	1974	16.2	52
1873	18.2	46 *	1907	17.0	61 *	1941	20.5	12 *	1975	18.1	13
1874	16.3	41 *	1908	15.5	46 *	1942	15.9	75 *	1976	16.1	50
1875	17.2	51 *	1909	15.8	75 *	1943	16.9	92 *	1977	14.9	106
1876	17.3	48 *	1910	16.7	113 *	1944	18.7	59 *	1978	15.9	97
1877	15.4	80 *	1911	15.6	40 *	1945	19.5	98 *			

Tällaisesta numeerisesta taulukosta on voi tietenkin poimia vaivatta erillisiä tietoja, mutta yleisten piirteiden hahmottaminen ei ole helppoa. Seuraavassa esitetään muutamia kuvallisia keinoja, jotka parantavat tarkastelumahdollisuuksia.

Kuva 1 Lämpötilajakauma histogrammana

Tässä tapauksessa on luonnollista odottaa, että jakauma olisi lähellä normaalijakaumaa. SURVO 76-ohjelmassa HISTO, jolla kuva on piirretty, voidaan tutkia erilaisten tilastollisten mallien hyvyttä myös graafisesti piirtämällä tutkittavan jakauman histogramman päälle haluttuja teoreettisia jakaumia. Seuraavassa kuvassa on kuvioon sovitettu normaalijakauma, jonka keskiarvo=17.05 ja keskihajonta=1.65 vastaavat havaittua lämpötilajakaumaa.

Kuva 2 Normaalijakauma sovitettuna lämpötilajakaumaan

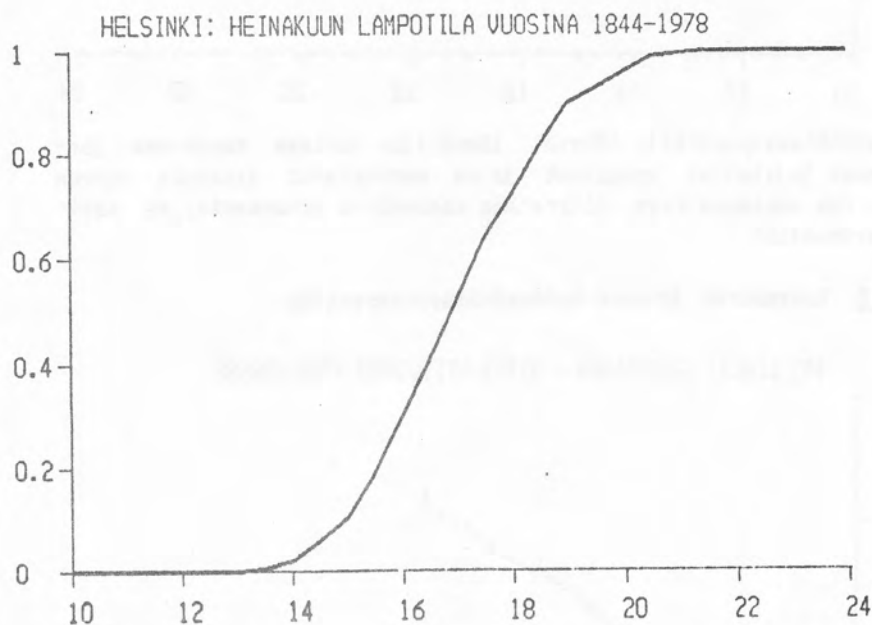
Yhteensopivuus näyttää jo silmämäärin hyvältä. Havaintoa vahvistaa myös χ^2 -testi, jossa luokitetun empiirisen jakauman frekvenssejä voidaan verrata teoreettisen jakauman vastaaviin frekvensseihin:

lämpö-tila	havaittu frekvenssi	teoreettinen frekvenssi
alle 14.6	8	8.2
14.6-15.0	6	6.1
15.1-15.5	11	9.0
15.6-16.0	15	11.9
16.1-16.5	15	14.4
16.6-17.0	15	15.9
17.1-17.5	16	16.1
17.6-18.0	11	14.8
18.1-18.5	13	12.5
18.6-19.0	11	9.6
19.1-19.5	4	6.7
yli 19.5	10	9.2

Testisuureen arvo on tässä tapauksessa 3.71 (vapausasteita 9), mikä merkitsee erittäin hyvää yhteensopivuutta, sillä todennäköisyys sille, että testisuure saisi yhteensopivuuden kannalta vielä paremman (pienemmän) arvon, on 0.07.

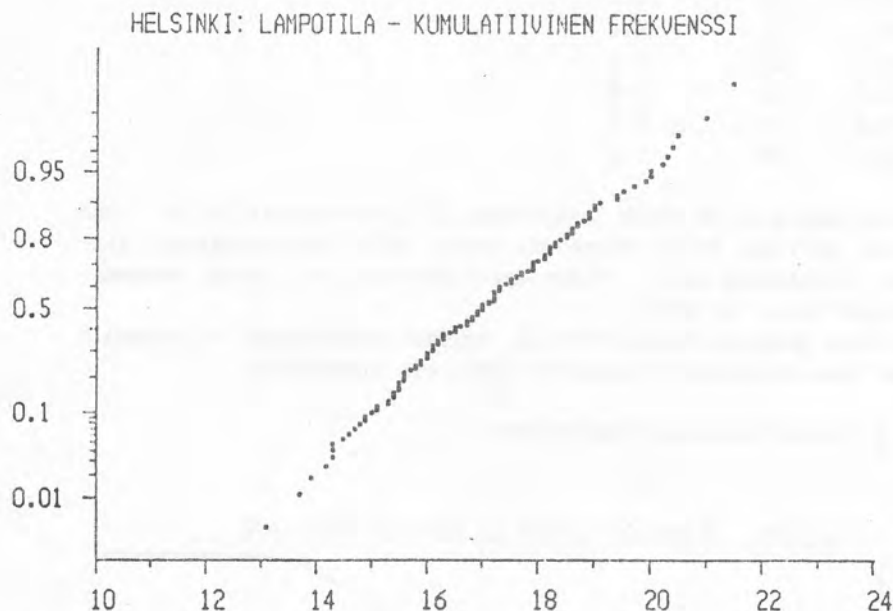
Empiirisen jakauman käyttäytymistä voidaan tutkia myös piirtämällä jakauman kumulatiivinen frekvenssifunktio eli summakäyrä:

Kuva 3 Lämpötilajakauman summakäyrä



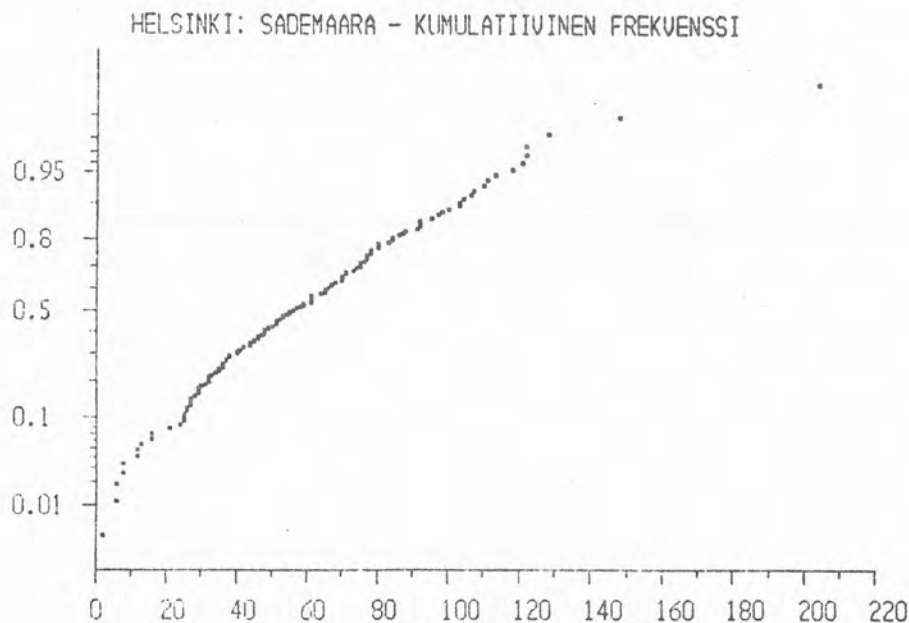
Summakäyrästä voidaan helposti lukea niiden tapausten suhteellinen osuus, jotka alittavat tietyn arvon. Esim. tässä tapauksessa nähdään, että heinäkuun keskilämpötila on alle 16° Helsingissä noin todennäköisyydellä 0.3. Summakäyrästä on kuitenkin verraten hankala tutkia jakauman muotoa ja esim. sen normaalisuutta. Tätä varten summakäyrää voidaan muuntaa siten, että se tarkan normaalijakauman tapauksessa oikenee suoraksi viivaksi. On olemassa valmiiksi viivoitettuja erikoispapereita (todennäköisyyspaperit), joissa pystysuorassa suunnassa on ns. probit-asteikko. Kuvassa 4 keskilämpötilan jakauma on piirretty todennäköisyyspaperiesitystä vastaavaan muotoon SURVO 76-ohjelmalla DIAGRAM.

Kuva 4 Keskilämpötilan jakauma todennäköisyyspaperilla



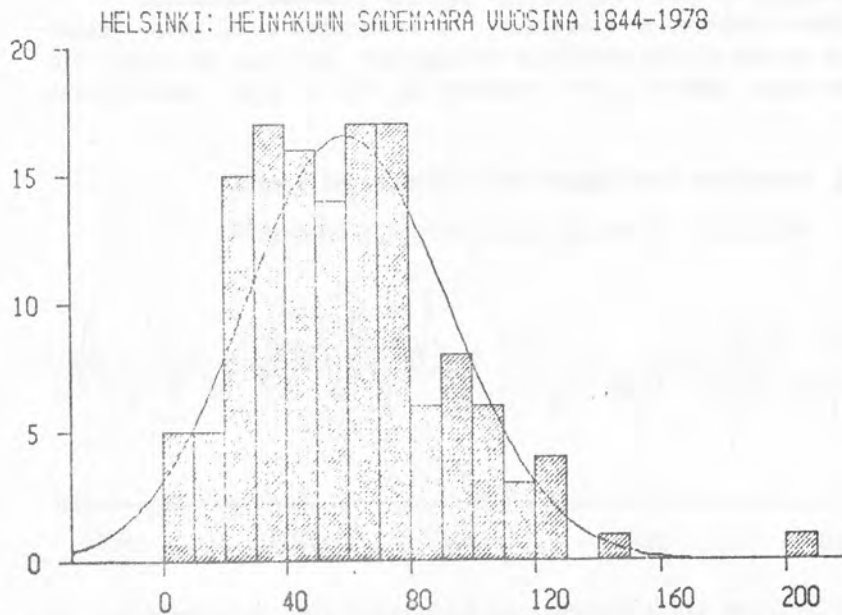
Todennäköisyyspaperilla näkyvät lämpötilan suhteen nousevaan järjestykseen lajitellut havainnot ja ne muodostavat likimain suoran viivan. Jos vastaava kuva piirretään sademäärän jakaumasta, se näyttää seuraavalta:

Kuva 5 Sademäärän jakauma todennäköisyyspaperilla



Havaintopisteet eivät ole yhtä kauniisti samalla suoralla, mikä osoittaa, ettei sademäärää voida pitää yhtä hyvin normaalisti jakautuneena. Itse asiassa tämä johtuu siitä, että jakauma on oikealle vino, mikä ilmenee selvästi myös histogrammasta:

Kuva 6 Sademäärän jakauma histogrammana

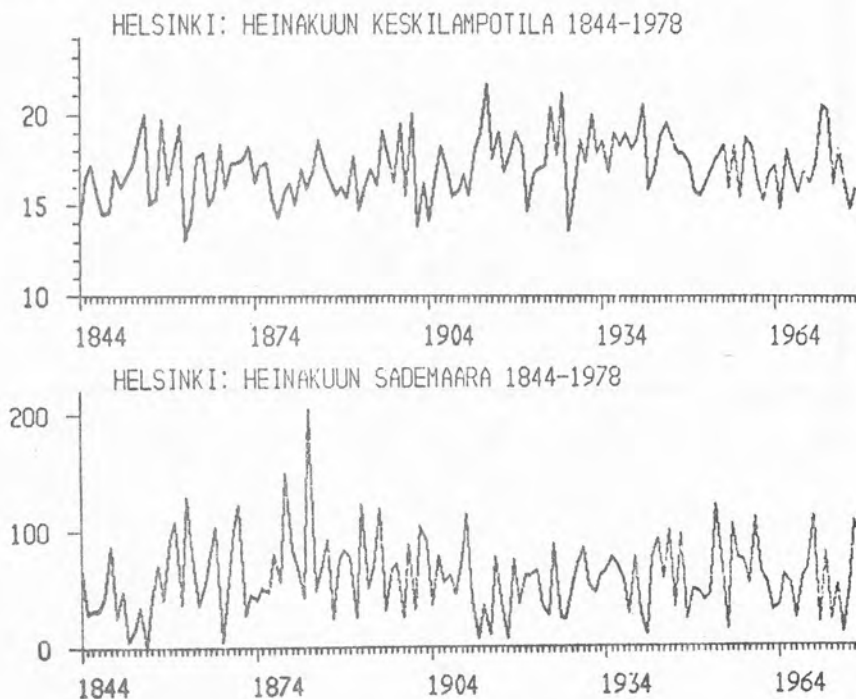


Kuvaan 6 on sovitettu myös normaali-jakauma empiirisen jakauman keskiarvoa 59.5 ja keskihajontaa 32.6 vastaavilla parametrin arvoilla.

χ^2 -testi saa nyt arvon 10.7 (vapausasteita 9), mikä ei viittaa hyvään yhteensopivuuteen. Todennäköisyys, että testisuure tulee havaittua arvoa suuremmaksi on 0.3. Normaalisuushypoteesia ei siis kuitenkaan voisi tällä perusteella hylätä.

Havainnot voidaan kuvata myös aikasarjoina ja ne näyttävät SURVO 76-ohjelmalla DIAGRAM piirrettyinä seuraavilta:

Kuva 7 Heinäkuun keskilämpötila ja sademäärä Helsingissä



Sademäärä näyttää vaihtelevan tällä aikavälillä täysin satunnaisluonteisesti. Sen sijaan keskilämpötilan kohdalla voidaan erottaa jo silmämääräisesti aivan lievää kasvua. Tämä havainto saa vahvistusta, kun aikasarjaan sovitetaan pienimmän neliösumman keinolla lineaarinen trendi. Trendin yhtälö on

$$\text{keskilämpötila} = 0.00876 \times \text{havaintovuosi} + 0.305.$$

Koska kertoimen 0.00876 suhde keskivirheeseensä (ns. t-arvo) on 2.45, on melko varmaa, että kerroin todella poikkeaa nolasta.

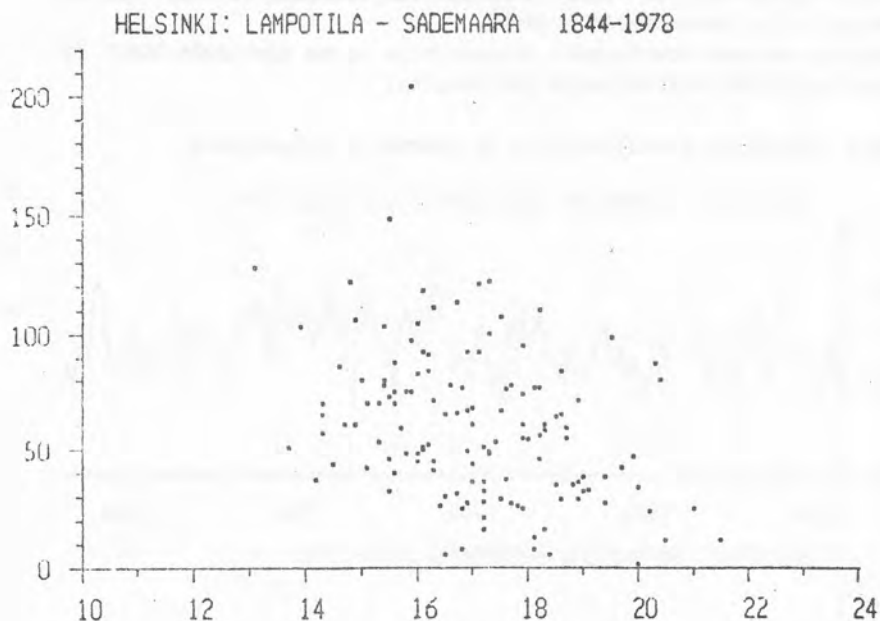
Jos tähän tulokseen on uskomista, se merkitsisi sitä, että viimeisen sadan vuoden aikana Helsingin heinäkuinen lämpötila on ollut lievässä kasvussa, joka on 0.009° vuodessa ja noin 1° koko tarkasteluvälillä.

Kuva 8 Heinäkuun keskilämpötilan "lineaarinen trendi"



Lopuksi voidaan vielä tarkastella keskilämpötilan ja sademäärän yhteisvaihtelua ns. korrelaatiodiagramman (pistediagramman) muodossa, jolloin ajallinen ulottuvuus jää jälleen pois tarkastelusta.

Kuva 9 Keskilämpötilan ja sademäärän yhteisjakauma

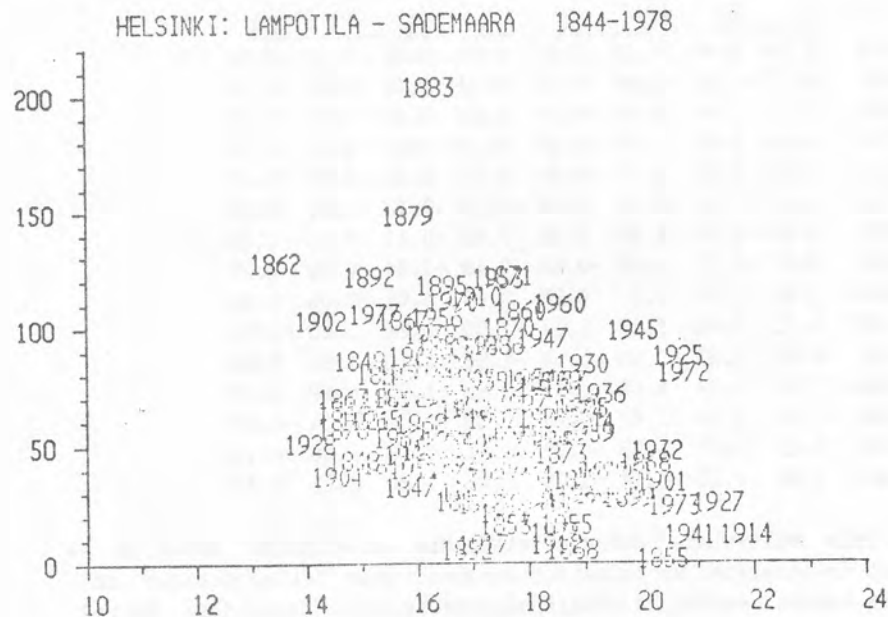


Tällaisesta pistediagrammasta on yleensä helppo havaita yhteisjakauman systemaattiset piirteet. Muuttujien välillä näyttää olevan lievää riippuvuutta (korrelaatiokerroin on itse asiassa -0.36) odotetun kaltaisella tavalla: Jos sataa paljon, ei voi olla kovin lämminä.

Pistediagrammasta näkyvät myös aineiston heikkoudet. Varsinkin, jos joukkoon on pujahtanut täysin virheellisiä havaintoarvoja tai aineiston yleiseen käyttäytymiseen nähden poikkeuksellisia havaintoja, ne huomataan. Tässä tapauksessa poikkeavin havainto lienee vuosi 1883, jonka kohdalla sademäärä näyttää olevan yli 200 mm. Tällaiset poikkeavuudet on syytä tutkia huolella, mutta niitä ei missään tapauksessa saa mennä aineistosta karsimaan ilman perusteellista selvitystä havainnon virheellisyydestä.

Poikkeavien havaintojen tarkastelua helpottaa, jos diagramma piirretään siten, että havaintopisteiden paikalle kirjoitetaan niiden tunnuksot, siis tässä tapauksessa asianomaiset vuosiluvut, mikä sekin saadaan tehdyksi suoraan DIAGRAM-modulilla.

Kuva 10 Havaintotunnuksilla merkitty pistediagramma



4.2 Moniulotteisen aineiston kuvaustapoja

Kun tutkittavana on yhden tai kahden muuttujan asemasta samanaikaisesti useampia muuttujia, kuvallisten keinojen käyttö vaikeutuu huomattavasti, sillä aidosti moniulotteista tilannetta on mahdoton liittää tasoon menettämättä olennaista informaatiota muuttujien ja havaintojen keskinäisestä suhteesta.

Viime aikoina on kuitenkin esitetty joitakin uusia keinoja, joilla saatetaan hieman helpottaa näitä kuvausongelmia.

Tarkastelemme esimerkkinä pientä jo valmiiksi tiivistettyä havaintoaineistoa antropologian alalta. Alkuperäinen aineisto sisältää pääkallon ja hampaiston rakennetta kuvaavia mittauksia ihmisistä, apinoista ja fossiileista. Mittauksia on ollut 8 ja ne on tehty useista havainnoista 15 eri ryhmässä.

Tässä esitettävä aineisto on tiivistetty siten, että erään monimuuttujamenetelmän, erotteluanalyysin avulla on määrätty 8 muunnettua muuttujaa $F_1, F_2, \dots, F_8$, jotka kuvaavat vertailtavien 15 ryhmän eroja mahdollisimman hyvin. Seuraavassa taulukossa esitetään vain näiden erottelumuuttujien ryhmäkohtaiset keskiarvot:

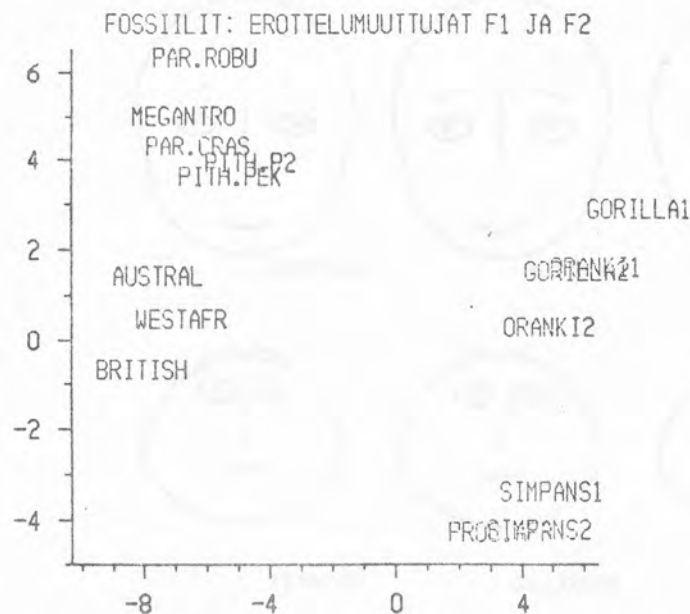
	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
WESTAFR	-8.09	0.49	0.18	0.75	-0.06	-0.04	0.04	0.03
BRITISH	-9.37	-0.68	-0.44	-0.37	0.37	0.02	-0.01	0.05
AUSTRAL	-8.87	1.44	0.36	-0.34	-0.29	-0.02	-0.01	-0.05
GORILLA1	6.28	2.89	0.43	-0.03	0.10	-0.14	0.07	0.08
GORILLA2	4.28	1.52	0.71	-0.06	0.25	0.15	-0.07	-0.10
ORANKI1	5.11	1.61	-0.72	0.04	-0.17	0.13	0.03	0.05
ORANKI2	3.60	0.28	-1.05	0.01	-0.03	-0.11	-0.11	-0.08
SIMPANS1	3.46	-3.37	0.33	-0.32	-0.19	-0.04	0.09	0.09
SIMPANS2	3.05	-4.21	0.17	0.28	0.04	0.02	-0.06	-0.06
PITH.PEK	-6.73	3.63	1.14	2.11	-1.90	0.24	1.23	-0.55
PITH.P2	-5.90	3.95	0.89	1.58	-1.56	1.10	1.53	0.58
PAR.ROBU	-7.56	6.34	1.66	0.10	-2.23	-1.01	0.68	-0.23
PAR.CRAS	-7.79	4.33	1.42	0.01	-1.80	-0.25	0.04	-0.87
MEGANTRO	-8.23	5.03	1.13	-0.02	-1.41	-0.13	-0.28	-0.13
PROC.AFR	1.86	-4.28	-2.14	-1.73	2.06	1.80	2.61	2.48

Käytetyin esitystapa moniulotteistenkin aineistojen kohdalla on edelleen tavanomainen korrelaatiodiagramma, joka perusmuodossaan sallii vain kahden muuttujan yhteisjakauman havainnollistamisen kerrallaan. Piirtämällä näitä diagrammoja eri muuttujapareille saadaan muodostetuksi jonkinlainen kokonaisnäkemys aineiston käyttäytymisestä.

Varsinkin pienten aineistojen tapauksessa on etua siitä, että pelkän havaintopisteen asemasta merkitään kuvioon havainnon nimi (kuten tehtiin jo edellä kuvassa 10) tai symboli, joka suuruutensa tai muotonsa puolesta antaa joitain havaintoa koskevia lisätietoja (kuten kaupunkien suuruudet kartalla).

Kuvassa 11 on fossiiliaineiston kahden ensimmäisen ja tärkeimmän erottelumuuttujan F_1 ja F_2 korrelaatiodiagramma ja se osoittaa aineiston jakautuvan melko selvästi useaan ryhmään (nykyihmiset, eräät ihmistyyppiset fossiilit, apinat). Tässä esityksessä ns. puuttuvaksi renkaaksi ihmisten ja apinoiden välillä joskus epäilty "Proconsul africanus" näyttää leiriytyneen simpanssien seuraan.

Kuva 11 Erottelumuuttujien F1 ja F2 ryhmäkohtaiset keskiarvot



Aitojen moniulotteisten graafisten tekniikkojen kehitys on vasta alullaan. Esimerkkeinä uusista mahdollisuuksista esitellään seuraavassa lyhyesti tunnetun amerikkalaisen tilastotieteilijän H.Chernoffin naamatekniikka (1973), jossa moniulotteinen informaatio puetaan ihmisen visuaalisen hahmotuskyvyn kannalta tuttuun muotoon ja D.F.Andrewsin käyrätekniikka (1972), jossa korrelaatiodiagrammien pisteet korvataan käyrillä.

Kuvassa 12 on fossiiliaineisto kuvattuna SURVO 76-modulin FACES piirtämänä Chernoffin naamatekniikalla. Jokainen tarkastelluista muuttujista F1, F2, ..., F8 liitetään tässä johonkin kasvonpiirteeseen (eräät jopa useampaan sellaiseen) käyttäjän harkitsemalla tavalla. Näitä piirteitä on Chernoffin alkuperäisessä koodauksessa käytettävissä 18 (esim. suun leveys, kaarevuus, silmien koko ja muoto jne.)

Tässä tapauksessa koodaus on valittu siten, että ihmisistä tulisi jossain määrin ihmisen kaltaisia ja apinoista apinoita. Merkitseekö tämä sitten sitä, että esim. Proconsul africanus tulee näyttämään todella omalta itseltään, on jo toinen asia; tuskin näin on.

On selvää, ettei Chernoffin naamatekniikka monessakaan suhteessa täytä objektiivisen kuvaustavan ehtoja ja on sellaisenaan pelkkä kursorisiteetti, mutta se tuo esiin mahdollisuuksia, joita kannattaa tutkia.

kuva 12 Ihmiset, apinat ja fossiilit Chernoffin naamoina



WESTAFR



BRITISH



AUSTRAL



GORILLA1



GORILLA2



ORANKI1



ORANKI2



SIMPANS1



SIMPANS2



PITH.PEK



PITH.P2



PAR.ROBU



PAR.CRAS



MEGANTRO



PROC.AFR

Andrewsin käyräesityksessä jokaista havaintoa muuttujista $F_1, F_2, \dots, F_8$ vastaa käyrä

$$f(t) = F_1/1.4142 + F_2 \sin(t) + F_3 \cos(t) + F_4 \sin(2t) + F_5 \cos(2t) + \dots,$$

joka on siis eräänlainen havainnon Fourier-kehitelmä. Funktio $f(t)$ on jaksollinen ja riittää piirtää se yhden jakson osalta. On selvää, että mitä enemmän havainnot muistuttavat toisiaan, sitä lähempänä toisiaan kulkevat niiden Andrews-käyrät.

On jopa mahdollista osoittaa, että itse asiassa käyrien väliin jäävän alueen pinta-ala vastaa havaintopisteiden euklidista etäisyyttä alkuperäisessä 8-ulotteisessa avaruudessa. Kuva 13 esittää aineistoamme Andrews-käyrinä piirrettynä SURVO 76-modulilla SCURVE ja tässä "panoraamakuvassa" (jokainen kiinteä t merkitsee määrättyä yksiulotteista tarkastelukulmaa) esim. Proconsul africanuksen (havainto 15) erikoinen luonne tulee jälleen selvästi esiin.

Kuva 13 Ihmisten, apinoiden ja fossiilien Andrews-käyrät

