

Programmatuur-sectie

TCINFO: een automatische programmadetector

Dik Bakker/Ulbe Brouwer/Pierre Debets *

1. Inleiding

TCINFO is een aan het Technisch Centrum van de Universiteit van Amsterdam ontwikkeld programma voor het automatisch zoeken van het juiste programma bij een gevraagde (statistische) methode of techniek. Aan de hand van, door het programma interactief gestelde, vragen kan de gebruiker uiteindelijk één of meer programma's vinden, die op de SARA Cyber computer voor het betreffende doel beschikbaar zijn. Aanvullende informatie over de gevonden programma's wordt op verzoek door TCINFO geleverd.

TCINFO is ontworpen om de veel voorkomende vraag: "Bestaat er een programma voor(volgt een of andere statistische techniek)" middels het computersysteem zélf te beantwoorden.

Langs interactieve weg zal de gebruiker zeer gemakkelijk zo uitvoerig mogelijk informatie kunnen krijgen over voor hem/haar van toepassing zijnde computerprogramma's. Optioneel is het mogelijk alle verwijzingen naar dokumentatie, fouten en beperkingen van programma's naar een printfile te laten schrijven, voor latere consultatie.

Regelmatig zullen de meest recente ontwikkelingen m.b.t. nieuw aangeschafte programmatuur, onlangs ontdekte fouten enz., op de informatie file worden bijgeschreven.

In het onderstaande zal iets nader worden ingegaan op de volgende punten:

*) de auteurs zijn werkzaam bij het Technisch Centrum F.S.W. van de Universiteit van Amsterdam.

- indeling van het probleemgebied
- globale indeling van het programma
- enige technische aspecten
- een voorbeeld van het gebruik

2. Systematische indeling in statistische deelgebieden

Regelmatig zijn door diverse instanties pogingen ondernomen om het toegankelijke statistische terrein voor de sociale wetenschappers op zinvolle wijze in te delen.

Deze pogingen zijn vaak slechts zeer ten dele geslaagd. De onduidelijkheid, ambiguïteit, mate van overlap in dergelijke gekonstrueerde schema's zijn over het algemeen groot.

Een blik op de vele systematische indelingen voor statistiek in bibliotheken geeft dit duidelijk weer.

Ook bij het ontwerpen van TCINFO dienden dergelijke problemen zich aan. Desalniettemin is uiteindelijk een opsplitsing gemaakt van het statistisch terrein in tien deelgebieden, te weten:

1. beschrijvende statistiek (eenvoudige grootheden en associatiematen)
2. hypothese toetsende enkelvoudige technieken
3. regressie technieken
4. variantie-analytische technieken
5. multidimensional scalings technieken incl. 1 dim.
6. tijdreeksen
7. multivariate technieken
8. graphen analytische technieken
9. testmodellen
10. loglineaire modellen

Op deze indeling is direkt kritiek te leveren, immers, zij

suggereert dat bijvoorbeeld multivariate regressie en variantieanalyse niet thuis horen onder de noemer multivariate technieken; hetgeen vanzelfsprekend niet het geval is. Het heeft niet in de bedoeling gelegen een exclusieve, disjunkte opsomming van deelgebieden te geven. Een sterke mate van overlap is aanwezig, echter, het leek zinvol erg veel gebruikte statistische bewerkingstechnieken zoals regressie en variantie analyse apart te kategoriseren, onder meer omdat voor deze bewerkingen een breed scala van mogelijke programmatuur bestaat.

Het deelgebied multivariate technieken is hier gereserveerd voor factoranalytisch georiënteerde technieken, discriminant analyse, canonische correlaties en clustertechnieken. Zelfs door deze inperking van het terrein der multivariate technieken worden in deze module nog meer dan 30 programma's genoemd.

Ook de benoeming van de verschillende deelgebieden kan verwarring opleveren. Zinnvolle suggesties van gebruikerskant zullen in de toekomst gaarne worden ingewacht.

Wat dient men bijvoorbeeld te verstaan onder testmodellen; deze vraag zal, gesteld aan drie methodologen, drie antwoorden kunnen opleveren, afhankelijk van de oriëntatie: psychologisch, sociologisch, politicologisch. In TCINFO wordt hierbij gedacht aan o.a. onderwerpen uit de psychometrika zoals itemanalyse, logistische testmodellen, betrouwbaarheidscoëfficiënten, enz., kortom statistische aanpakken die nauw verbonden zijn aan de psychologische testleer (deze technieken zijn trouwens ook eenvoudig toe te passen op andere inhoudelijke vakgebieden). Ook min of meer verwante technieken als latente class analyse (Iazarsfeld) worden in deze module ondergebracht, hoewel de overeenkomst

met factoranalyse ook ~~er~~ ot is.

- Het gebruik van TCINFO zal uitmaken of de hier gehanteerde indeling zinvol en doorzichtig is; aanpassing blijft te allen tijde mogelijk.

3. Indeling van het programma

TCINFO bestaat uit drie onderdelen, die elk als een afzonderlijk geheel kunnen worden beschouwd.

Deel 1. Hierin kan de gebruiker zijn te analyseren data set aan het programma aanbieden. Er zijn twee niveaus te onderscheiden:

- a. de gebruiker heeft geen idee over de te hanteren statistische bewerkingstechniek. In principe moet het nu mogelijk zijn om via "goed gekozen" vragen de gebruiker te sturen naar de meest voor de hand liggende oplossing. Dit onderdeel zal uiterst ingewikkeld zijn: het ligt niet in de bedoeling aan de verdere uitwerking ervan hoge prioriteit te geven. De vraag is zelfs, of een dergelijke edukatieve zoekprocedure wel zinvol op het SARA-systeem is te programmeren (mogelijk wél op het PLATO-systeem!)
- b. de gebruiker kent (wellicht vaag) de weg waarlangs hij wil gaan om tot een afgeronde statistische bewerking te komen. In deze situatie is het uiterst zinvol vooronderstellingen m.b.t. de uit te voeren analyse vooraf te toetsen, bijv.:

- homoskedasticiteit
- zijn de variabelen normaal verdeeld ?
- is een correlatie matrix positief definitief?
- bestaat er lineaire afhankelijkheid tussen variabelen?

- zijn de scores afkomstig uit een bepaalde, bekende verdeling?
- etc.

In dit eerste deel van TCINFO zal gegeven de data getoetst worden of aan de voorwaarden voor het uitvoeren van enige analyse voldaan is.

Een aantal voorwaarden zal overigens lastig te toetsen zijn, laat staan te programmeren. De verwachting is, dat nog veel studie gepaard zal gaan aan de verdere uitwerking van dit onderdeel.

Deel 2. Hierin kan de gebruiker zoeken naar programmatuur bij de gekozen techniek.

Normaliter weet de gebruiker van statistische programmatuur welke statistische methode en/of techniek hij voor de analyse van zijn data nodig heeft. Deel 2 levert hem nu de programma's. De door de computer gestelde vragen zijn van dien aard, dat altijd aan het eind van een sessie een of meer programma's worden gevonden.

Alhoewel in veel opzichten een beperking (daarover in punt 4 iets meer), levert dit een snelle zoekprocedure op: teleurstelling bij de gebruiker om na een lange, intelligente computersessie de mededeling: "geen programma beschikbaar" te krijgen, is hiermede uitgesloten.

E.e.a. houdt echter wel in, dat frustratie kan ontstaan door de nogal grote rigiditeit in het aantal mogelijke antwoorden op een vraag.

Bij aanschaf van nieuwe programmatuur zullen nieuwe vragen

worden geïntroduceerd c.q. meer alternatieven op een vraag worden geboden.

M.a.w. hoe meer programma's beschikbaar zijn op een deelgebied, hoe gedifferentieerder benadering via vragen mogelijk is.

Zo bevat momenteel de module graphen analyse slechts een beperkt aantal vragen, terwijl de module voor hypothese toetsende technieken zeer veel mogelijkheden kent (meer dan 25 programma's).

Het ligt in de bedoeling, dit onderdeel regelmatig aan te passen en uit te breiden.

Zo zijn recentelijk, voorafgaande aan de vragen m.b.t. statistische methoden, enige vragen ingevoegd m.b.t. programma-tuur voor data cleaning en niet-numerieke data (i.c. teksten). Naast het aflopen van een boomstructuur met vragen aangaande (statistische) methoden, op zoek naar bijbehorende programma-tuur, is het ook mogelijk, direkt de informatie over de aanwezige programma's op te vragen; de gebruiker typt eenvoudigweg de naam van het bewuste programma in.

Deel 3. Hierin kan verdere informatie en ook (onderlinge) evaluatie worden gevraagd van de gekozen programma's.

Als, aan het eind van deel 2, de programma's bekend zijn, kan de gebruiker op verzoek nadere informatie verkrijgen over de gevonden programma's. Dit betreft een korte omschrijving van de methode, geheugenbeslag, ntl parameter, dokumentatie e.d.

Via deel 3 kan verdere aanvullende informatie worden verstrekt, zoals:

- bekende fouten

- onvolkomenheden in de methode
- rekenbetrouwbaarheid
- mogelijke opties enz. enz.

Indien mogelijk wordt een vergelijking gemaakt tussen programma's voor dezelfde bewerking en aldus verwezen naar het "beste" programma.

Vergelijkingskriteria zullen in de output worden opgesomd, zodat de gebruiker ook zelf een keuze kan maken, afhankelijk van het gewicht dat hij aan de criteria hecht.

4. Technische aspecten

Van het, in Fortran en voor een klein deel in Compass geschreven, programma zijn momenteel het middelste deel - het zoeken naar programmatuur bij een computer analyse probleem - en de eerste helft van het derde deel - het verschaffen van een korte programma omschrijving plus gegevens over geheugenbeslag, vindplaats van het programma en de dokumentatie - operationeel.

De programma informatie bevindt zich op een random file, welke "onzichtbaar" voor de gebruiker wordt geladen en weer aan het systeem teruggegeven. Er is een programma ontwikkeld, waarmee deze file ge'update kan worden; naar het zich laat aanzien, zullen binnen enkele maanden alle ± 175 nu aan TCINFO bekende programma's op uniforme wijze op genoemde file beschreven staan.

Op verzoek van de gebruiker wordt, van de gevonden programma's, in elf regels (één scherm vol op de 711 terminal) gepresenteerd :

- korte inhoudelijke omschrijving

- wijze van opslag in de Cyber (vindplaats)
- enige bijzonderheden, zoals geheugenbeslag, libraries etc.
- aanduiding van de beschikbare dokumentatie en literatuur

Wederom optioneel kan deze informatie bewaard worden op de file output, welke na afloop van de sessie naar de printer ge'route kan worden.

De vragen worden op twee wijzen gesteld (zie het voorbeeld onder punt 5.):

1. er verschijnt een zin op de terminal die is:
 - a. een ja/nee vraag; antwoord ja of nee (j of n is genoeg)
 - b. een opsomming van twee of drie alternatieven: de eerste, onderscheidende letter(s) van het gekozen alternatief zijn als antwoord voldoende.
2. er verschijnt een lijst van (maximaal 10) alternatieven, voorafgegaan door een cijfer (1 t/m n). Het bij de keuze behorende cijfer dient te worden getypt.

Bij een 'onmogelijk' antwoord geeft het programma in het kort aan welke antwoorden werden verwacht.

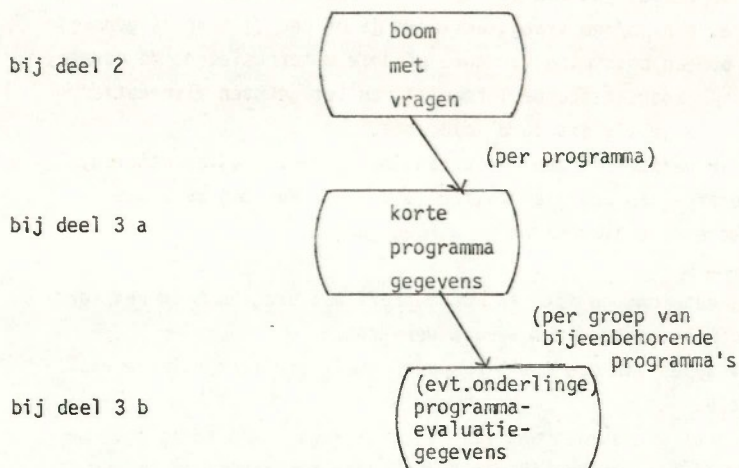
Het typen van een ? als antwoord geeft een iets ruimere explicatie.

Door te antwoorden met een asterisk * gaat men terug naar de voorgaande vraag; dit met oog op door typefouten of andere vergissingen ontstane ongewenste keuzen.

Ten gevolge van de beperkte geheugenruimte voor interactieve programma's (60.000 oktaal op het SARA systeem) is er een, nu reeds dringende, beperking aan de omvang van de af te zoeken boomstructuur, en de hoeveelheid informatie (en daarmee aan

de duidelijkheid) in de door het programma gestelde vragen.

Overwogen wordt, om ook de, nu nog in TCINFO geprogrammeerde, af te zoeken boom met vragen, naar schijf over te hevelen. Hiermede verkrijgt men de volgende, hierarchische structuur voor de informatie file :



Uiteraard zal het afzoeken van een boom op schijf een vertraging betekenen tijdens de zoekfase van deel 2. Echter, het programma zou wel een (optimale) deelboom 'binnen kunnen halen'. Bovendien biedt de vrijgekomen geheugenruimte de mogelijkheid om zoekprocedures te versnellen via methoden (cf. binary search op gealfabetiseerde programmalijst) voor

de programmering waarvan nu geen ruimte beschikbaar is. Verder liggen, naast het nog niet gerealiseerde deel 1, en de opbouw van de evaluatie file voor deel 3 b (waarvoor diepgaande research m.b.t. de aanwezige programmatuur een absolute voorwaarde is), afhankelijk van de beschikbare geheugenruimte, o.m. de volgende uitbreidingen in het verschiet :

- de mogelijkheid voor gebruikers om, min of meer onafhankelijk van de nogal dwingende vraagstructuur, door het zelf verschaffen van key words (cf. "factor analyse", "cleaning", "schaal", "gemiddelde en standaarddeviatie", etc.) programmatuur op te sporen.

Hiervoor is een (bescheiden) tekstverwerkende komponent noodzakelijk, en een routine, die dergelijke sleutelbegrippen kan match'en tegen de ergens in de informatie file aanwezige begrippen; hierop zou het programma zonodig dan weer door moeten kunnen vragen.

Een eerste element is inmiddels ingebouwd in de routine, welke naar een door de gebruiker opgegeven programma zoekt. Is de bewuste programma-naam onbekend, dan wordt, op basis van de vooronderstelling dat het om een type of spellingsfout gaat, een overzicht gegeven van de meest op de ingetypte naam gelijkende wel bestaande programma namen.

Zo levert "FACT" (géén programma) op:

FACTOR

FACAN

FACAN SCORES

JFACTOR

en geeft "ACT" :

FACTOR

ECTA

- geeft de genoemde spellingsoptie alle programma's waarvan

de naam (maar niet noodzakelijkerwijs de functie) min of meer overeenstemt, men zou ook op basis van bepaalde typologieën (gebruikte methode, al of niet interactief, edukatief, utility, in een bepaald pakket aanwezig, fixed of free field data format, al of niet key word versie etc.) naar (een verzameling van) programma's willen kunnen zoeken. Hiertoe biedt de programma-naam noch de plaats in de boomstructuur in het algemeen voldoende soelaas.

Aan alle programma's in de informatie file is een retrieval-kode gekoppeld, welke echter slechts in beperkte mate is geactiveerd. Uitbreiding van de mogelijkheden in deze liggen in het verschiet.

- het ligt in de bedoeling de mogelijkheid te creëren voor de gebruiker om vragen, opmerkingen e.d., ontstaan n.a.v. het gebruik van TCINFO, via het programma zelf aan het Technisch Centrum door te spelen; door periodieke verwerking van e.e.a. kan wellicht nog meer dan in het verleden rekening worden gehouden met de suggesties, wensen en problemen van gebruikers, die de weg naar de User's Service niet weten te vinden.

5. Voorbeeld van het gebruik

Onderstaand een weergave van een (korte) sessie met TCINFO. In dit geval is gekozen voor het bij een gewenste methode zoeken van bijbehorende programmatuur, liever dan directe informatie over reeds gekozen programma's.
(zie volgende pagina's)

TOINFO/T.C./USE IS FREE OF CHARGE

12/09/79

11.55.55.

WILT U INSTRUKTIE, TYPE DAN JA, ANDERS NEE
nee

WILT U INFORMATIE OVER EEN OF MEER SPECIFIEKE PROGRAMMA'S ?
nee

HEEFT U NUMERIEKE OF NIET-NUMERIEKE DATA ?
n

ANTWOORD MOET BEGINNEN MET NU NI
num

WILT U INFORMATIE OVER CLEANING, MANIPULATIE OF STATISTISCHE BEWERKING ?
st

WELKE METHODE/TECHNIEK KIEST U ?

1. BESCHRIJVENDE STATISTIEK (EENVOUDIGE GROOTHEDEN EN ASS.MATEN)
2. HYPOTHESE TOETSENDE ENKELVOUDIGE TECHNIEKEN
3. REGRESSIE TECHNIEKEN
4. VARIANTIE-ANALYTISCHE TECHNIEKEN
5. MULTIDIMENS.SCHAALTECHNIEKEN (INKL. 1 DIM.)
6. TIJDREEKSEN
7. MULTIVARIATE TECHNIEKEN
8. GRAPHEN ANALYSE
9. TESTMODELLEN
10. LOGLINEAIRE MODELLEN

44

EEN WAARDE TUSSEN 1 EN 10 A.U.B.
4

VARIANTIE-ANALYTISCHE TECHNIEKEN

UNIVARIAAT OF MULTIVARIAAT ?
u

BAYESIAANSE OF KLASSIEKE AANPAK ?
?

ANTWOORD MOET BEGINNEN MET B K

DOORGAAN MET ETX - VRAGEN MET ? - CORREKTIE MET * - STOPPEN MET ?STOP

BAYESIAANSE OF KLASSIEKE AANPAK ?
k1

PARAMETRISCH OF NONPARAMETRISCH ?⁹⁶⁻
p

EEN OF MEER FAKTOREN ?
meer

WILT U UW KEUZE CORRIGEREN ?
nee

U HEEFT DE KEUZE UIT DE VOLGENDE PROGRAMMA'S :

- 1 VARANU
- 2 ANOVA
- 3 BMDP2V
- 4 BMDP3V

WILT U EXTRA INFORMATIE ?
ja

OVER WELK PROGRAMMA ?
anova

EEN WAARDE TUSSEN 1 EN 4 A.U.B.

2

ANOVA

OMSCHRIJ- VARIANTEANALYSE MET 1 AFHANKELYKE EN 1-5 ONAFH. VARIABELEN.
VING FACTORIEEL DESIGN. BY ONGELYKE GELFREKVENTIES KEUZE UIT:
KLASSIEKE, HIERARCHIESE OF REGRESSIEBENADERING. FIXED MODEL.

VINDPLAATS SPSS
BIJZONDER- PUBLIC FILE
HEDEN

DOKUMEN- SPSS-MANUAL
TATIE

WILT U DEZE INFORMATIE BEWAREN ?
j

U HEEFT DE KEUZE UIT DE VOLGENDE PROGRAMMA'S :

- 1 VARANU
- 2 ANOVA * OUTPUT
- 3 BMDP2V
- 4 BMDP3V

WILT U EXTRA INFORMATIE ?

J

OVER WELK PROGRAMMA ?

3

BMDP2V

OMSCHRIJ- (CO)VARIANTIE-ANALYSE MET 1 AFH.VAR. EN MEERDERE FACTOREN.
VING FIXED EFFECTS MODELS EN REPEATED MEASURES MODELS VOOR
BALANCED DESIGNS MET GELIJKE EN ONGELIJKE GELFREKVENTIES.
MEERDER GROEPS- EN TRIALFACTOREN MOGELYK, ALSOOK COVARIATEN.
VINDPLAATS TCLIB.
BIJZONDER- LIBRARY: TCLIB.
HEDEN

DOKUMEN- BMDP-MANUAL.
TATIE

WILT U DEZE INFORMATIE BEWAREN ?

J

U HEEFT DE KEUZE UIT DE VOLGENDE PROGRAMMA'S :

1	VARANU		
2	ANOVA	*	OUTPUT
3	BMDP2V	*	OUTPUT
4	BMDP3V		

WILT U EXTRA INFORMATIE ?

n

WILT U DOORGAAN OF STOPPEN ?
stop

N.B. INFORMATIE M.B.T. 2 PROGRAMMA'S OP FILE OUTPUT
STOP

.951 CP SECONDS EXECUTION TIME

COMMAND-