WESP, Waarlijk Eenvoudig Statistisch Pakket *)

L.Th. van der Weele, D.M. van der Sluis, T. Wierstra, T. van der Meer, H.F. Vogt.

REKENCENTRUM DER RIJKSUNIVERSITEIT GRONINGEN

Inleiding

Hoe waarlijk eenvoudig WESP ook is, het is ondoenlijk in een klein half uur te vertellen hoe het pakket precies in elkaar zit. Ik zal me daarom beperken tot
- de filosofie achter het pakket
- heel in het algemeen de opzet
- en een aantal punten waarop pakketten beoordeeld kunnen worden.

WESP is een algemeen statistisch pakket met een batch versie en een interactieve versie; het eenvoudige zit hem meer in de eenvoud in het gebruik dan in de eenvoud in het aantal mogelijkheden dat het biedt. Naar onze mening is WESP door deze eenvoud zeer geschikt als eerste kennismaking met statistische programmatuur. Het wordt wel eens een "teaching pakket" genoemd, maar is tot een zekere diepgang toch ook zeer bruikbaar als "research pakket". Globaal past WESP in de uitspraak van G. Bernard gedaan op de afgelopen COMPSTAT: "Easy-to-use packages, with consequently reduced possibilities, are thus still useful, along with more sophisticated packages"

Eerst iets over de geschiedenis. WESP is ontstaan in 1971 toen Groningen overging op een andere computer installatie. In een situatie met een aantal losse programma's die niet of nauwelijks op elkaar waren afgestemd, werd besloten in overleg met de gebruikers van statistische programmatuur over te gaan op een geïntegreerd systeem van statistische programmatuur. Geschikte pakketten waren op dat moment niet beschikbaar: derhalve besloten we zelf maar een pakket te ontwikkelen. De eerste versie werd geschreven in ALGOL 60 om een aantal redenen waar ik nu niet op in wil gaan. Het pakket werd geleidelijk uitgebreid. In 1974 kwam een interactieve versie beschikbaar. Later werd WESP omgezet in FORTRAN met een klein aantal assembler routines. De redenen voor deze omzetting waren
- we hadden een versie nodig voor een DEC system-10
- we wilden het geheugenbeslag en de executietijd verminderen.

Als nevenproduct werd de portabiliteit verhoogd.

Op de CYBER 74-18 heeft WESP aan 50.000 octale woorden genoeg om een redelijk grote job te draaien. De executiesnelheid is tamelijk hoog, wat geïllustreerd wordt door de 3.7 sec. CPU tijd die nodig is om 13 modulen te laten werken op een bestand van 500 individuen en 13 variabelen.

Filosofie

De filosofie achter het pakket is samen te vatten met de uitspraken: een statistische berekening moet een middel zijn en geen doel, of anders gezegd: een statistische berekening is niet meer dan een stap in het analyse proces en moet geen eindproduct zijn. Verder gaat de ambitie niet verder dan voor, zeg, 90% van de gebruikers, zeg 90% van hun problemen op te lossen. De overige problemen worden verwezen naar speciale statistische programmatuur. U ziet geheel conform Bernard, al hebben wij het bedacht voor dat hij de desbetreffende uitspraak deed.

Deze filosofie gaf aanleiding tot de volgende voorwaarden
1. Het pakket moet gemakkelijk te leren zijn, hetgeen een eenvoudige stuurtaal betekent.
2. Een beperking tot algemene statistische methoden, omdat sophisticated methoden in handen van minder geoefenden slechts tot misbruik kunnen leiden.
3. Compacte invoer, overzichtelijke uitvoer.
4. Een minimum aan opties, niet alleen omdat het onthouden van de betekenis van opties lastig is, maar ook omdat het maken van een keus nogal problematisch is voor minder geoefenden.
5. Uitgebreide controles op het correct zijn van de invoer; met name controles op het bereik van variabelen en op het kompleet zijn van gevallen wanneer de gegevens voor een individu in meer dan één kaart zijn geponst.

*) Tekst van een voordracht gehouden door L.Th. van der Weele op de jaarvergadering van de Contactgroep Statistische Programmatuur, Utrecht 12 september 1978

6. Duidelijke foutmeldingen; dit lijkt vanzelfsprekend, maar helaas ontbreekt hier nog al eens wat aan, zelfs bij befaamde pakketten.
7. Een minimum aan "operating system" handelingen.

Structuur

De structuur van een WESP programma is strikt modulair; dat betekent dat een moduul wordt aangeroepen op de plaats waar de actie nodig is. Zodoende kan de gebruiker op elk moment beslissen wat hij in de volgende stap gaat doen. Er zijn geen voorbereidende acties nodig, zoals het maken van permanente omcoderingen in SPSS. Evenmin is er een voorschrift betreffende de volgorde van modulen. Dit is van groot belang, vooral bij de interactieve versie. Als een moduul zijn werking heeft voltooid keert het programma terug op een punt waar een willekeurig moduul gespecificeerd kan worden. Een WESP programma is in feite niets anders dan een opeenvolging van moduulaanroepen met bijbehorende specificaties.

Binnen een moduul is de structuur als volgt.

1. Het keyword om het gewenste moduul te selecteren.
2. Een lijst van parameters om de benodigde waarden, "opties" en "statistics" aan te geven. Wanneer slechts defaults worden gebruikt vervalt deze kaart. In de interactieve versie wordt dit aangegeven met NONE. Voor "opties" en "statistics" worden woorden gebruikt en geen getallen zoals b.v. SPSS doet. De volgorde van de parameters is volstrekt willekeurig; ze worden met komma's van elkaar gescheiden.
3. De nummers van de te gebruiken variabelen. Van deze lijsten zijn diverse vormen gedefinieerd, afhankelijk van het type moduul. Bij sommige modulen moet hier een instructie worden gespecificeerd, zoals een omcoderingsvoorschrift.
4. In een enkel geval moeten hier aanvullende gegevens worden gegeven; b.v. een correlatie-matrix als deze slechts in kaarten beschikbaar is. Dit wordt met een parameter aangegeven.
5. Het woord END om de moduul-instructie af te sluiten. De input is in vrij-formaat met de komma of de punt als scheidingsteken.

Een voorbeeld van een moduul-instructie is:

CORRELATE	voor product-moment correlatie coëfficiënten
MISDAT,GROUP,COR=OUTFILE	met de parameter MISDAT wordt aangegeven dat missing data paarsgewijze weggelaten moeten worden. Lijstsgewijs weglaten is default. GROUP betekent dat alle berekeningen moeten worden uitgevoerd voor subgroepen. De default betekent negeren van een groepstructuur ook al zou die wel aanwezig zijn. COR=OUTFILE betekent dat de correlatie matrix naar een file wordt geschreven voor verdere analyses.
1,3-8,10	Voor selectie van de variabalen 1, 3 t/m 8 en 10
END	Om de instructie voor het moduul te beëindigen.

Invoer

Zoals gebruikelijk in dit soort pakketten is de rechthoekige scorematrix de basis voor de berekeningen. De scores kunnen formaat gebonden als wel in vrij formaat worden ingevoerd. Uiteraard kan ook met een WESP system file gewerkt worden. Binnenkort wordt er een procedure opgenomen om ook een SPSS system file in te kunnen lezen. Ook kan dan een SPSS system file gecreëerd worden. Bij relevante modulen kan ook een correlatiematrix of een factormatrix worden ingelezen, in WESP formaat en binnenkort ook in SPSS formaat.

Data manipulatie

De gebruikelijke variabelentransformaties kunnen worden uitgevoerd, zoals het optellen en aftrekken van variabelen, het vermenigvuldigen, log transformatie, omcoderingen, enzovoorts. Van belang is op te merken dat als een opdracht ongedefinieerd is, zoals delen door nul, het resultaat automatisch een missing data is. Voorbeelden van transformaties zijn:

```
LOG 8,8.
RECODE 14=iF3=1A4=1THEN1 ELSE
          iF3=2A4=2THEN2 ELSE 3.
```

De scorematrix kan ook in subgroepen onderverdeeld worden, b.v. als volgt:

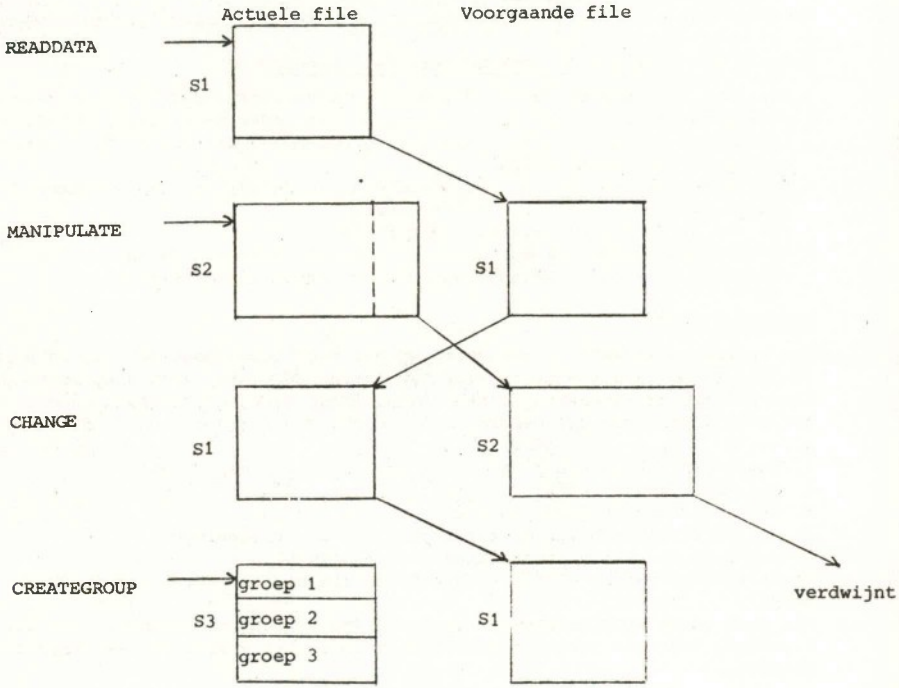
```
CREATEGROUP
utrecht
GROUP1  (4=1A5=1), GROUP2 (4=2)
END
```

Eveneens kan de scorematrix getransponeerd worden.

```
TRANSPOSE
csp
END
```

Tenslotte kunnen individuen op grond van een kriterium geselecteerd worden en kan een aselecte steekproef, met of zonder teruglegging, getrokken worden.

Bij het veranderen van de scorematrix wordt geen onderscheid gemaakt tussen permanente en tijdelijke omcoderingen. Dit wordt gerealiseerd met behulp van twee files: een actuele, waarop de modulen werken, en een voorgaande waarop de voorlaatste scorematrix is verzameld. Door de status van actuele- en voorgaande file te wisselen kunnen de laatste transformaties ongedaan worden gemaakt. Op deze wijze hoeft nooit vooraf beslist te worden of een transformatie tijdelijk dan wel permanent is. Dit is een groot voordeel bij interactief werken met het pakket.



Missing data

Voor de missing data is een vaste code gekozen, en wel de spatie: een spatie betekent altijd een missing data en omgekeerd wordt een missing data altijd voorgesteld door een spatie. In geen geval is het mogelijk berekeningen te maken waarbij missing data als getallen worden behandeld. Dit is, naar ons gevoel, de enige veilige methode om missing data te definieren. Met een apart moduul kunnen andere coderingen tot missing data (spaties dus) getransformeerd worden. Missing data kunnen paarsgewijs als wel lijstgewijs worden weggelaten. Vanwege veiligheid t.a.v. de analyse is hierbij het lijstgewijs weglaten default.

Batch versus interactief gebruik

Wij geloven dat, afgezien van dat naar onze mening WESP gemakkelijker is te gebruiken dan elk ander pakket dat wij kennen, het grote voordeel van WESP is dat het te gebruiken is in de batch als wel interactief met precies dezelfde instructie, afgezien van de procedures typisch voor interactieve handelingen. Deze procedures behelzen.

1. Het programma vraagt expliciet om de invoer van keywords, parameterlijsten, variabelenlijsten en andere gegevens.
2. Na uitvoering van een moduul keert het programma terug op een punt waarop een nieuw keyword kan worden geselecteerd.
3. Fouten kunnen binnen een moduul worden opgevangen en hersteld.
4. Berekeningen kunnen binnen een moduul herhaald worden voor andere variabelenlijsten.
5. De output van een moduul kan "gestuurd" worden
   - de hele output kan op de terminal getoond worden
   - de hele output kan naar een regeldrukker of een andere terminal gestuurd worden
   - een regel of een aantal regels kan bekeken worden om na te gaan of de resultaten correct zijn, waarna opnieuw beslist kan worden wat er met de output gedaan dient te worden.

Eveneens kan een zoekprocedure voor een "character string" gebruikt worden.

Gebruik van WESP bij de Rijksuniversiteit Groningen

- Per jaar worden meer dan 30.000 WESP jobs verwerkt, dat betekent meer dan 8% van alle jobs. Behalve WESP worden nog gebruikt onder meer: SPSS, MULTIVARIANCE, CLUSTAN en een groot aantal speciale programma's die veelal zijn verzameld in onze programmatheek LISTOR.
- Een WESP-programma bevat gemiddeld 6 à 7 modulen; er zijn gevallen dat een programma veel meer dan 100 modulen bevat.
- 62% van de WESP jobs bevatten geen fouten.
- De gebruikers zijn afkomstig uit alle disciplines van de universiteit, waarbij uiteraard de Sociale Wetenschappen de grootste afnemers zijn.

Onderwijs WESP

Het onderwijs in WESP is gebaseerd op het z.g. Cursusboek dat aan de hand van gegeven data het analyse proces met behulp van WESP stap voor stap introduceert. Daarnaast wordt uiteraard de WESP Handleiding gebruikt, evenals de data waarop het practicum wordt uitgevoerd. Zo'n cursus duurt 3 à 4 dagen, afhankelijk van de pretentie van de cursusgever en de "snelheid" van de studenten. De cursus omvat:

1. Het leren van de stuurtaal en de mogelijkheden van WESP.
2. Het schrijven van WESP programma's.
3. Het ponsen van kaarten en het verwerken van jobs.
4. Het interpreteren van de resultaten van de berekeningen.
5. Het oefenen met de interactieve versie TTWESP.

Na voltooiing van de cursus kunnen de meeste studenten zelfstandig werken met WESP.
Per jaar wordt WESP onderwezen aan zo'n 450 studenten, afkomstig van allerlei disciplines. Voor een aantal studierichtingen is het volgen van een WESP cursus verplicht.

De inhoud

Het pakket bevat de voor dit soort pakketten gebruikelijke faciliteiten met een nadruk op algemene statistische methoden. Met name:

- uitgebreide invoer en data management procedures
- frequentie tabellen, kruistabellen en regeldrukker-plots
- distributie parameters
- chi^2 toets, product-moment en rangcorrelatie coëfficiënten, partiële correlatie, verscheidene methoden voor factoranalyse, multiple regressie, twee-steekproeven toetsen(wel en niet parametrisch), toetsen voor gepaarde waarnemingen en eenvoudige variantie-analyse.

Documentatie

Er bestaat een komplete gebruikershandleiding geschreven in het Nederlands. Daarentegen zijn de teksten in het programma geheel in het Engels. Bovendien is er een Cursusboek om met WESP te leren werken.

Literatuur

G. Bernard.	A Comparison of three Statistical Packages: GENSTAT, BMDP, and SPSS, In: COMPSTAT 1978, Proceedings in computational statistics, Wien 1978.
N.H. Nie et al.	SPSS, Statistical Package for the Social Sciences, New York 1975
D.M. van der Sluis	WESP, de informatica aspecten. Intern RC-Rapport, Rekencentrum Rijksuniversiteit Groningen, 1976.
L.Th. van der Weele	WESP, an easy to use statistical package. In: ECODU-17 Proceedings, Davos 1974.
L.Th. van der Weele	Handleiding WESP, R.C.-Publikatie nr. 8, Rekencentrum Rijksuniversiteit Groningen, 1977.
L.Th. van der Weele	De rol van de Projectgroep Statistische Applicaties (PSA) van het Rekencentrum bij de bewerking van (medische) onderzoekgegevens. In: Lezingencyclus Medische Informatica, Groningen 1977.
T. Wierstra	TTWESP, an interactive statistical package. In: ECODU-21 Proceedings, Geilo 1976.
W. Zeelenberg	Cursusboek WESP, Overdinkel 1978.