

De eerste S van SAS
enkele kritische kanttekeningen

G.C.Bos *)
J.Mulder **)
H.v.d.Knaap ***)
H.Doppenberg ****)

*) Shell Nederland Informatieverwerking, Rijswijk
**) M.S.A., K.U.Nijmegen
***) Unilever Research Laboratorium, Vlaardingen
****) Mars, Veghel

Abstract.

In this paper the results of the "Kwaliteitstoetsing SAS" project of the SSP (Statistical Software Packages) subgroup SAS (Statistical Analysis System) are presented.

During 1983 and 1984 the documentation and output of many statistical procedures provided by SAS were evaluated with respect to their usefulness, completeness and correctness.

In general the statistical correctness was satisfactory. Most of our comments are with respect to the documentation, support and output facilities.

We feel that the attention of SAS Institute is shifting from the statistical area to other tasks, such as developing graphical facilities and a user-friendly way of datahandling.

Samenvatting

Een verslag van de resultaten van de werkgroep 'Kwaliteitstoetsing' van de SSP (Sectie Statistische Programmatuur), subgroup SAS (Statistical Analysis System).

In de afgelopen twee jaar zijn vele statistische SAS procedures beoordeeld op hun bruikbaarheid, volledigheid, juistheid en enkele documentatieaspecten.

In het algemeen worden deze kwaliteitscriteria positief beoordeeld. Met name zijn er nauwelijks echte fouten gevonden in de statistische procedures.

Echter plaatsen wij enkele kritische kanttekeningen bij documentatie, gebruikersondersteuning en uitvoermogelijkheden.

Wij hebben de indruk dat de aandacht van SAS Institute verschuift van statistiek naar andere toepassingen, zoals het ontwikkelen van grafische hulpmiddelen en een gebruikersvriendelijke manier van gegevensverwerking.

1 Inleiding

1.1 Wat is SAS?

SAS (Statistical Analysis System) is een software-systeem dat gereedschappen biedt voor data-analyse : informatie-opslag en retrieval, data modificatie, rapportering, statistische analyse en file-handling.

De SAS taal heeft een Engelsachtige syntax. Een SAS programma is opgebouwd uit stappen: DATA stappen waarin datamanipulatie en informatie opslag gebeurt, en PROC stappen waarin analyses en rapportering worden uitgevoerd: de zogenaamde "procedures".

SAS biedt ca. 50 procedures voor statistische analyses. Daarnaast biedt SAS grafische procedures in het SAS/GRAPH product, spreadsheet mogelijkheden in SAS/FSP, operations research technieken in SAS/OR en toepassingen op econometrisch gebied plus tijdreeks analyses in SAS/ETS. Gebruikers kunnen eigen procedures aan SAS toevoegen. Als deze aan bepaalde kwaliteitscriteria voldoen, worden zij opgenomen in een Supplemental Users Library, die wordt verspreid onder alle SAS gebruikers. Veel officiële SAS procedures vinden hier hun oorsprong.

Het SAS pakket wordt op lease basis ter beschikking gesteld door SAS Institute, een bedrijf dat gevestigd is in Cary (North Carolina) en dat een aantal vertegenwoordigingen elders in de wereld heeft. SAS Heidelberg is het contact voor Nederland van SAS Institute.

Bij SAS Institute bestaat een uitgebreid systeem van gebruikersondersteuning en communicatie. Enkele belangrijke services zijn:

SASWARE BALLOTS: Alle wensen en verlangens van gebruikers, die bij SAS binnenkomen, worden verzameld. Wensen die niet op korte termijn gerealiseerd kunnen worden, zet SAS op de SASware ballot. Een keer per jaar krijgen de gebruikers deze ballot, waarop ze kunnen aangeven wat zij belangrijk vinden. De prioriteitenlijst die hieruit gemaakt wordt, gebruikt SAS bij het verbeteren van het pakket. Voor de Europese gebruikers geldt, dat het niet duidelijk is wat er met hun wensen gebeurt; het ballot formulier komt namelijk op het moment dat het al een paar weken ingeleverd had moeten zijn.

SAS USAGE NOTES: Als gebruikers een fout vinden, kunnen ze contact opnemen met SAS Institute. Deze fouten worden gedocumenteerd in de Usage Notes. In deze Usage Notes staat tevens in welk release de fout eventueel verholpen zal zijn en een oplossing om het probleem te omzeilen, als dat mogelijk is. De coördinator van een rekencentrum kan na aanvraag deze notes tot zijn beschikking krijgen op zijn computersysteem.

SUGI: De Internationale SAS Gebruikers groep organiseert jaarlijks een groot congres in de USA. In andere delen van de wereld worden soortgelijke congressen gehouden, zoals SEUGI in Europa. SAS Institute stimuleert deze congressen om ook op deze manier contact te hebben met gebruikers.

1.2 De subgroep SAS.

In Nederland zijn de gebruikers van SAS verenigd in de, op 25 juni 1982 opgerichte, subgroep "SAS" van de Sectie Statistische Programmatuur, die deel uitmaakt van de VVS. Deze subgroep stelt zich ten doel een vinger aan de pols te houden van het SAS produkt voor wat betreft fouten en nieuwe ontwikkelingen, en ervaringen uit te wisselen. Dit doel omvat zoveel facetten, dat er drie werkgroepen geformeerd zijn, die elk een deel bekijken. De drie werkgroepen zijn:

1. Kwaliteitstoetsing
2. Opleidingen
3. Grafische Toepassingen

Om de verkregen resultaten kenbaar te maken aan mensen buiten de "SAS" subgroep, heeft de werkgroep Kwaliteitstoetsing haar bevindingen in het hierna volgende verslag vastgelegd.

2 Begrenzing werkkerrein

2.1 Eigen werk

De aanpak van de werkgroep kwaliteitstoetsing is in essentie toegespitst op het gebruik van SAS als *statistisch* pakket bij het eigen werk.

Dit lijkt misschien een wat grove beperking. Een dergelijke selectieve aanpak heeft echter grote voordelen:

- het beoordelen betekent aldus een minimale extra belasting bij het dagelijks werk: een voorwaarde voor het goed functioneren van een werkgroep.
- de beoordelingen worden uitgevoerd met realistische gegevens, niet met gefingeerde data.
- er wordt geen kruit verschoten aan het moeizaam opstellen van objectieve kwaliteitscriteria: de beoordelingen mogen dan subjectief zijn, maar zijn voor gebruikers, die veel met dit pakket werken, zeker relevant.

De beoordelingen zijn gedaan op versie 82.4 en eerdere. Als er een statistische procedure van het pakket wezenlijk veranderd was in een nieuwere versie, dan werd dat onderdeel opnieuw beoordeeld.

2.2 Criteria

Het bekende voorbeeld waarin statistische programmatuur geevalueerd wordt, is het werk "A comparative review of statistical software" van Francis(1979). Deze geeft een indrukwekkende lijst van 61 criteria. De criteria van Francis blijken voor ons niet goed hanteerbaar, aangezien deze bedoeld zijn voor het vergelijken van statistische pakketten, terwijl wij ons bezighouden met de kwaliteit van een pakket.

Wij hebben drie hoofdcriteria onderscheiden:

1. overzichtelijkheid
2. juistheid
3. volledigheid

Deze criteria zijn zowel toegepast op de documentatie als op de output van SAS. Ter toelichting van wat wij onder deze criteria verstaan, het volgende.

1 Overzichtelijkheid

- a) Is snel duidelijk hoe een procedure aangestuurd moet worden na lezing van de *documentatie*? Is hierin terug te vinden wat in de output staat?
- b) Is direkt duidelijk wat in de *output* staat?

2 Juistheid

- a) *Documentatie* juistheid kan worden beoordeeld met de vraag: wat in de documentatie staat, kan ik dat in de output vinden?
- b) De uitkomsten op de *output* werden gecontroleerd op hun juistheid en, wanneer dat belangrijk geacht werd, op de rekennauwkeurigheid.

3 Volledigheid

De vraag die hier centraal staat is: Kan ik met SAS mijn probleem op een voor mijzelf bevredigende wijze oplossen?

- a) Is alles wat in de output staat ook in de *documentatie* terug te vinden?
- b) Krijg ik in de *output* alle relevante informatie?

3 Resultaten

3.1 Resultaten per procedure

De resultaten van ons onderzoek zijn in een tabel in de appendix samengevat. In deze tabel staan per procedure vermeld waar wij commentaar of wensen hebben en op welk van de genoemde criteria. Tevens is er een toelichting gegeven op het commentaar of de wens. Voor een aantal gevallen is deze korte toelichting niet voldoende. In het volgende hoofdstuk zullen we o.a. deze gevallen nader bekijken.

3.2 Samenvatting resultaten per criterium

1a *Overzichtelijkheid* van output

Alle procedures werden positief beoordeeld op overzichtelijkheid.

1b *Overzichtelijkheid* documentatie

Bij 2 van de 24 procedures werd de documentatie niet overzichtelijk gevonden.

2a *Juistheid* van de output

Bij de procedures VARCOMP, SURVTEST, RSREG en ANOVA/GLM zijn onjuistheden geconstateerd.

Bij de procedure VARCOMP worden foutieve resultaten verkregen als in een model zowel fixed als random factoren zitten.

De procedure SURVTEST komt in een oneindige loop terecht als er een toets voor een exponentieel model gevraagd wordt en er geen failures zijn. Deze procedure is door een gebruiker geschreven en wordt door SAS meegeleverd.

In de procedure RSREG worden missende waarden voor de onafhankelijke variabelen niet goed verwerkt bij een lack-of-fit toets.

Bij de procedures ANOVA en GLM wordt in een aantal gevallen na een variantie analyse de Student-Newman-Keuls voortzetting niet goed uitgevoerd. Ook zijn in enkele niet-statistische procedures (DATASETS en CONTENTS) fouten ontdekt.

2b *Juistheid* van de documentatie

Bij de procedures PCTL, CANDISC en FORMAT werd een discrepantie geconstateerd tussen output en documentatie.

3a *Volledigheid* output

Bij 12 van de 24 procedures werd een wens tot uitbreiding van de mogelijkheden geformuleerd. Hierbij werd voor 8 procedures de compleetheid toch wel positief beoordeeld. De wensen tot uitbreiding liggen in het algemeen op het vlak van:

verbetering aansturen preciese rapportformaat.

uitbreiding mogelijkheden om statistische tussenresultaten op datasets weg te schrijven.

uitbreiding van het aantal toetsen of parameter schattingen binnen een procedure.

3b *Volledigheid* van documentatie

Bij 7 van de 24 procedures was er commentaar op de volledigheid van de documentatie. In het algemeen hebben we geconstateerd, dat er vaak een duidelijke literatuurreferentie ontbreekt. Referenties staan als een lijstje aan het einde van elke procedurebeschrijving, maar in de tekst zelf wordt er niet naar verwezen. Raadpleging van deze literatuur is soms nodig om duidelijkheid te verkrijgen over een specifieke (soms default) SAS optie.

4 Beschouwing

4.1 SAS produkt

Beschouwen we de kwantitatieve resultaten (par 3.2) van onze beoordelingen, dan vinden wij hierin inderdaad de algemene indruk weerspiegeld die wij van SAS hebben.

SAS wordt door de werkgroep beschouwd als een bruikbaar en betrouwbaar statistisch pakket.

Toch geeft het samenvatten van beoordelingen van 24 losse procedures geen compleet, samenhangend beeld van de kwaliteit van SAS.

Door het fragmentarisch beoordelen (d.w.z. per procedure) zijn twee punten uit het oog geraakt: de overzichtelijkheid van het geheel en de taaleigenschappen van SAS.

-Overzichtelijkheid geheel.

Men hoort in de wandelgangen van SAS gebruikers de klacht dat men door de bomen (een omvangrijke stapel SAS Guides) het bos niet kan zien.

Een voorbeeld:

Van een variabele wil men de gestandaardiseerde waarden gebruiken. Dit lijkt logisch gezien thuis te horen in de DATA stap (voor berekeningen) en men zoekt in het betreffende manual de bijbehorende functie. Resultaat : niet te vinden, noch bij de functies noch in de index. Toch kan dit wel, maar met de procedure STANDARD, die gedocumenteerd is in een ander manual.

De onoverzichtelijkheid ontstaat op het moment dat de gebruiker moet gaan kiezen uit verschillende procedures.

-Taaleigenschappen.

Omdat de procedures aparte programma's zijn die binnen SAS worden aangeroepen, ligt de syntax hiervoor niet vast zoals voor de statements in een DATA stap. Elke procedure kent dus zijn eigen syntax, die voortkomt uit zijn statistische aard. Bij nieuwe procedures wordt niet goed gelet op wat reeds bestaat. Dit verschijnsel wordt nog versterkt doordat gebruikers eigen geschreven procedures kunnen toevoegen, en daarbij hun eigen ideeën over syntax kunnen botvieren.

Dit heeft tot gevolg dat er homoniemen en synoniemen ontstaan. Een voorbeeld van een homoniem is het woord OUTPUT. Dit woord wordt in sommige procedures als optie opgegeven en betekent dat een output dataset moet worden gemaakt. Bij andere procedures wordt OUTPUT gebruikt als een statement waarachter men wensen kan opgeven over een output dataset. In een datastap betekent OUTPUT, dat op dat punt in het programma, de data in het geheugen in een SAS-dataset moet worden geplaatst. Het heeft wel in elk van de gevallen met SAS-datasets te maken, maar het is niet netjes. In het geval van wensen over een output dataset had bijvoorbeeld beter het woord OUTSPEC gekozen kunnen worden.

Een voorbeeld van een synoniem is het volgende. Bij twee veel gebruikte procedures, FREQ en TABULATE, kunnen tabellen gevraagd worden. Bij de procedure FREQ gebeurt dat met het TABLES statement en bij TABULATE met TABLE. Als elk van de mogelijkheden gewenst is, waarom dan niet beide mogelijkheden toegestaan bij elk van de twee procedures?

Voor wie geïnteresseerd is in een beschouwing over de taaleigenschappen van SAS, verwijzen wij naar het artikel van Haux(1983).

4.2 SAS Institute

De assistentie van SAS Institute is meestal zeer adequaat. Het SAS Usage Notes en SASware Ballot systeem dragen daaraan bij. In die gevallen waar echter echte fouten op statistisch gebied in de programmatuur worden gevonden, is de feedback van SAS Institute log te noemen. De fouten zoals door ons geregistreerd in de procedures VARCOMP, CONTENTS en DATASETS zijn inmiddels al 8 tot 18 maanden bekend maar er is nog geen andere actie genomen dan een rapportage van SAS Heidelberg naar SAS Cary, NC-USA.

Uit de resultaten van de SASware Ballot(1984) blijkt dat de wensen op statistisch terrein internationaal niet hoog scoren: in de top 100 van wensen verschijnt de eerste statistische wens pas op de 60-ste plaats. Hieruit blijkt dat SAS van een echt statistisch pakket is uitgegroeid tot een breder toepasbaar (zoals graphics, datamanagement, operations research, etc.) stuk gereedschap. We verwachten daarom van SAS Institute op statistisch gebied niet veel vernieuwingen. Door deze accentverschuivingen vrezen we dat de eerste S van SAS een steeds minder grote rol gaat spelen.

5 Literatuur.

- 1) Francis I.(1979), A comparative review of statistical software, North Holland Publishing Company.
- 2) Haux R.(1983), Die Syntax der Kommandosprache von SAS -Überblick und Kritik, Statistical Software Newsletter, series B, number 5, 28-37.
- 3) SAS User's Guide(1982), de delen Basics en Statistics, SAS Institute.

Appendix

In de tabel zijn de volgende afkortingen gebruikt:

d = documentatie
o = output
k = kommentaar
w = wens

proc	over- zicht	juist- heid	vol- ledig	toelichting
	d o	d o	d o	
ANOVA		k k k	w	optie voor keuze MS of SS. geen aanduiding voor automatisch poolen sources. P<.0001 wordt afgedrukt als P=.0001 . SNK-voortzetting is fout.
CANDISC			w	meer documentatie over afgedrukte matrices.
CHART		k k	k	aantal mogelijkheden beperkt. indeling klassebreedten niet altijd logisch.
CONTENTS		k	w	bij historie ontbreken eerste programmaregels. in een run overzicht variabelen en # observaties.
COXREGR			k k w	voor de hand liggende toets ontbreekt. max aantal covariaten = 20. combinatietoetsen van parameters.
DATASETS		k		rename-statement samen met change werkt niet.
DISCRIM			w	eigenwaarden afdrukken.
GLM		k k	w w	zie toelichting ANOVA. bij random optie geldt toelichting VARCOMP. parameters naar een output dataset.

proc	over-	juist	vol-	toelichting
	zicht	heid	ledig	
	d o	d o	d o	
KSLTEST			w w k	uitgebreidere toetsen voor normaliteit. verdelingstoetsen met zelf opgegeven parameters. biedt niet veel meer dan proc UNIVARIATE.
MATRIX			k	documentatie is didactisch slecht.
NLIN			k k k k	uiterst zinvolle uitbreiding regressiearsenaal. bij method=dud beter meer startwaarden opgeven.
NPAR1WAY			w w	uitbreiding naar universeler nonparam-methoden.
PCTL		k		slechte en soms foutieve documentatie.
PHGLM			w	combinatietoetsen van parameters.
RANK	w		k k	aanvulling op NPAR1WAY voor nonparam-statistiek. kracht PROC laten zien met betere voorbeelden.
REG			k	grote overlap met proc GLM.
RSQUARE			w	naast R**2 ook residuele standaarddeviatie.
RSREG		k		in l-o-f test foute behandeling missing values.
STEPPDISC				
STEPWISE			w	predicted en residual waarden naar output dataset.
SURVTEST		k	k	onvoldoende documentatie. komt soms in oneindige loop.
TABULATE	k			procedure is niet helder beschreven. w kolomtitels rechts aanschuiven. w verkorte notatie in table-statement.
UNIVARIATE			k k w w w	zeer groot arsenaal beschrijvende kengetallen. meer percentielen afdrukken. keuze schaling plot. testen verdelingen met zelf opgegeven parameters.
VARCOMP		k		alleen juist bij volledig random modellen.