KM 43 (1993)
pg. 175-183

Het statistisch pakket ARIEL+PLUS

Marien C.J. Lina[1]

Distributiegegevens

Auteur/Producent:	Sistemas Integralas José Miguel de la Barra 412 Casilla 13168 Santiago de Chile Chile
Distributie adres:	Sistemas Integralas 80 Rue du Fabourg Saint-Denis 75010 Paris Tel (33 1) 48 00 06 92
Besturingssysteem:	MS-DOS
Tevens beschikbaar voor:	IBM-mainframes, Digital, NCR, Wang en Nixdorf onder UNIX, DOS, OS en MVS
Omvang:	2 Mbytes (MS-dos harde schijf)
Prijs:	fl. 10.586,-

Inleiding

ARIEL+plus wordt ontwikkeld voor IBM-PC's en compatibles door Sistemas Integrales in Chili, voor de statistische verwerking van survey-research data. Het pakket wordt toegepast in 'health-service' projecten in derde wereld landen. Het vond onder anderen aftrek bij de Wereldbank. Door lid te worden van de 'ARIEL-club' abonneert men zich automatisch op het bulletin 'ARIEL News'. Het bulletin geeft een indruk van de nieuwste ontwikkelingen van ARIEL+plus.

Installatie

Bij het pakket wordt een handleiding meegeleverd, waarin de installatieprocedure duidelijk wordt omschreven. Tijdens de installatie kan men een keuze maken uit meerdere conversatie-talen, te weten: Engels, Duits, Frans, Spaans en Portugees. Tijdens de installatie wordt de interface en de herkenning van commando's aangepast voor de gekozen taal. Voor deze bespreking is de engelstalige versie getest. Voor de installatie van de DOS-versie van ARIEL+plus op een harde schijf is het nodig om ca. 2 Megabyte opslagcapaciteit te reserveren. De installatie van de testversie verliep probleemloos. Binnen een ARIEL setup kan de taal worden ingesteld met het RESET commando.

[1] CBS, Hoofdafdeling Automatisering, afd. Statistische Informatica

Handleiding en tutorial

Bij het pakket wordt gebruikersdocumentatie in de gewenste taal meegeleverd (zie installatie). In de Command-Summary van de handleiding worden alle opdrachten beschreven die aan het programma kunnen worden opgegeven. ARIEL+PLUS is commando-gestuurd. Het merendeel van de beschikbare keywords zijn herkenbare woorden, zoals bijvoorbeeld REGRESSION, MODIFY, LOGIT, MODEL-SAVING en SCATTER. Een alfabetische index is opgenomen in de COMMAND SUMMARY.

De eenvoudige procedures zoals de diverse mogelijkheden om tabellen te produceren worden in de command-summary duidelijk aangegeven. Voor de meer geavanceerde statistische procedures zoals factoranalyse en correspondentie analyse is de informatie enigszins beknopt.

Daarnaast wordt een tutorial meegeleverd. De survey-data die voor dit voorbeeld worden meegeleverd zijn onttrokken aan een 'National household Survey', uitgevoerd in Chili. De inrichting van de tutorial is geschikt voor beginnende gebruikers van het pakket, waarbij inhoudelijk statistische kennis bekend verondersteld wordt. De tutorial verstrekt de benodigde informatie in een praktische en logische volgorde. Na een begrippendefinitie volgt een beschrijving hoe een datastructuur kan worden opgezet. Daarna wordt uitleg gegeven over de inhoud van de aanwezige procedures. In de voorbeelden worden enkele eenvoudige procedures uitgevoerd op de beschikbare data. Uitvoervoorbeelden daarvan zijn in de tutorial opgenomen.

Aansturing van het programma

Na het opstarten van een introductiescherm, kan men in een geïntegreerde editor de batchfiles creëren en editen. Vanuit het introductiescherm kan men ook een run opstarten van de geselecteerde batchfile. De gebruiker kan interactief een naam opgeven voor de benodigde files met de data en de commando's en de uitvoerfile waar de resultaten naar worden geschreven. Daarna kan de uitvoering via het indrukken van een toets geactiveerd worden. Na verwerking keert men automatisch terug in de editor, waarin de gegenereerde uitvoerfile te raadplegen en aan te passen is.

Binnen ARIEL+PLUS worden gereserveerde woorden gehanteerd voor het aanroepen van de procedures. De naamgeving van data matrices, procedures en opties is soms enigszins cryptisch, en kan dan onoverzichtelijk zijn. Voor de aanwijzing van variabelen en data matrices worden geen namen maar nummers gehanteerd. Sommige keywords bestaan uit 1 letter of 1 cijfer, waarbij er niet altijd een logisch verband bestaat tussen de code en de betekenis daarvan (vergelijk de keuze van options bij SPSS). Voor mensen die het pakket goed kennen kan de bondigheid van de keywords een voordeel zijn. Anderen zullen terug moeten grijpen naar de handleiding wanneer zij bijv. de commandoregel lezen: V 3 SX 3 12 14.

Voor sommige zaken wordt een specificatie van opties verwacht, die zeker voor een beginnend gebruiker van het pakket verwarrend kunnen zijn. Zo wordt voor het opstarten van een factoranalyse de opgave van een categoriale ('qualitative') variabele verwacht, die door het pakket intern gebruikt wordt. Dit gaf bij het testen enige verwarring over het begrip

'qualitative variable', en de optie is overbodig wanneer het pakket de creatie van deze variabele intern af zou handelen.

De begrippen worden binnen het pakket consistent gebruikt. De specifieke aansturing van het programma komt verder aan de orde bij de behandeling van data invoer en afhandeling van procedures.

Data-organisatie

De ruwe data input wordt door ARIEL+plus ondergebracht in een ARIEL systemfile, die wordt aangegeven als 'MAIN TOME'. Een tome (bundel) is een verzameling datasets, die in één file worden opgeslagen. De tome kan een zeer grote datafile worden. Als eerste stap van de data-organisatie wordt door de gebruiker een tome gecreëerd, waarbij de (maximale) afmeting moet worden opgegeven, die daarna op schijf wordt gereserveerd. De omvang van een tome kan overigens achteraf uitgebreid worden. Het is zinvol om meerdere datasets in 1 tome onder te brengen als aggregatie van die datasets beoogd wordt. In andere gevallen kan de tome onnodig groot worden.

Op de main tome kunnen bewerkingen worden uitgevoerd voor reorganisatie van data. Een tome kan in gewijzigde of gereduceerde vorm worden gekopieerd naar een secundary tome. De gewijzigde data-organisatie kan dan voor de analyse gebruikt worden.

Een dataset of datamatrix wordt in ARIEL+PLUS aangeduid als een 'THEME'. In een tome kunnen meerdere theme's worden ondergebracht. Deze datasets kunnen verschillen van organisatie. Elk theme heeft een nummer. Voor elke statistische analyse is het noodzakelijk dat er een theme geselekteerd is. Dit is vergelijkbaar met de aanroep van een systemfile in SPSS. Een verwerking van data uit verschillende themes is mogelijk door herhaling van de procedures voor verschillende themes.

Een tome is dus anders gezegd een verzameling van van elkaar afwijkende themes (data matrices), die bewaard worden in 1 file (tome). Dit kan ten goede komen aan de overzichtelijkheid van data directories. Een nadeel hiervan ligt in de mogelijkheid van het ontstaan van onwerkbaar grote datafiles. De gehanteerde datastructuur maakt het niet noodzakelijk dat verschillende datasets in één file worden ondergebracht. Een situatie waarbij een tome gevormd wordt door een gereserveerde directory met themes lijkt meer voor de hand te liggen.

De themes worden aangeboden als ruwe ASCII-textfiles in fixed format. Bij een analyse wordt telkens verwezen naar een tome. Binnen die tome wordt een theme geselecteerd. De theme's worden geïdentificeerd met een nummer. Wanneer er binnen een run data uit meerdere theme's geanalyseerd worden vereist dit enige administratie, en zou het handig zijn om de variabelen uit verschillende theme's met een zelfde naam op te kunnen roepen. Een naam is gemakkelijker te onthouden dan een cijfer. Een zelfde naam zou bovendien in meerdere theme's naar verschillende variabele nummers kunnen verwijzen.

Bij het programma wordt een oefenfile meegeleverd. Dit betreft een dataset van een huishoudensenquête die gehouden is in Chili. De aanwezigheid van dit voorbeeld verhoogt de gebruikswaarde van de handleiding voor situaties waarin men beperkt ervaring heeft met het

organiseren van survey research en de statistische verwerking daarvan.

Tabel: Lijst van mogelijke opdrachten in Ariel+plus voor het beheer en voor het structureren van data

AGGREGATION	Vertaalt data naar een ander hiërarchisch nivo, waarbij selecties en gewichten toegepast kunnen worden.
DESTINATION	Definieert de bestemming van de onder aggregate vertaalde data
BACKUP	Maakt een Backup van een tome
CATALOG	Geeft de inhoud van een tome weer
COPY	Kopieert een dataset of de structuur daarvan
DESTROY	Verwijdert een theme
ELIMINATE	Verwijdert een variabele uit een theme
FORMATS	Geeft een Overzicht van formats van ruwe data files
INITIATE	Creëert een tome file
INVALID	Geeft een overzicht van invalid data in de theme
INPUT	Feature voor tape invoer
LISTING	Print een data list voor gespecificeerde variabelen
MENDING	Voor het uitvoeren van structurele correcties in de data
MODELSAVING	Bewaart resultaten van procedures als variabelen
PASSAGE	Transporteert data naar een externe file
RESTORE	Restore een backup van een tome
THEME	Selecteert een TEME
TITLE	Definieert een titel voor de uitvoer
SUBTITLE	Definieert een subtitel voor de uitvoer
USER	Definieert een gebruikers melding voor de uitvoer
VARIABLES	Geeft een overzicht van variabelen in een theme

Zoals reeds aangegeven wordt een datamatrix als invoer in fixed format aangeboden. Het inlezen van data in free format wordt niet ondersteund. Bij het definiëren van variabelen wordt derhalve een specificatie van het dataformat verwacht.

Al met al zijn er in vergelijking tot de gangbare pakketten bij het inlezen van nieuwe data sets nogal wat stappen nodig voordat er met de eigenlijke analyse kan worden begonnen.

Specificatie van variabelen en procedures

Voor de beginnende gebruiker wordt uitgelegd hoe een kodeboek kan worden samengesteld, en hoe men variabele specificaties kan opgeven. De specificaties worden ingebracht met behulp van een geintegreerde editor, en worden bewaard in een 'demanding file', een batchfile met ARIEL commando's. Bij het creëren van deze batchfile is er geen help rubriek of interne controle binnen het programma. Een dergelijke batchfile met data definities ziet er als volgt uit:

Voorbeeld van een setup voor data-definitie

```
THEME  1
TITLE    "NATIONAL HOUSEHOLD SURVEY: HOUSEHOLD LEVEL"
SUBTITLE "PRACTICE SURVEY FOR ARIEL+PLUS TUTORIAL"
USER     "ENGLISH TUTORIAL"
V  1 "HOUSEHOLD NUMBER"          1  2  7   ORDERED
V  2 "PLACE OF RESIDENCE"        1  9  1
C  1  URBAN                 2
C  2  RURAL
V  3 "TYPE OF HOUSING"           1 10  1
C  1 "HOUSE OR APT"         3
C  2  TRADITIONAL
C  3  OTHER
V  4 "TENANCY STATUS"            1 11  1
C  1  OWNED                  4
C  2  RENTED
C  3  BORROWED
C  4  OTHER
V  5 "NUMBER OF BEDROOMS"        1 12  1
V  6 "NUMBER OF BEDS"            1 13  1
V  7 "DISTANCE POTABLE WATER"    1 14  1
C  1 "IN DWELLING"        7
C  2 "LESS     THAN 25M"  7
C  3 "MORE     THAN 25M"  7
V  8 "SOURCE POTABLE WATER"      1 15  1
C  1 "PUBLIC SYSTEM"        8
C  2  WELL
C  3  SPRING
C  4  RIVER
C  5  OTHER
V  9 "EXCRETA DISPOSAL"          1 16  1
C  1 "SEWAGE SYSTEM"        9
C  2 "SEPTIC TANK"
C  3 "PIT    LATRINE"
C  4  NONE
V 10 "GARBAGE DISPOSAL"          1 17  1
C  1 "PUBLIC COLLECT"     10
C  2  INCINERATION
C  3  BURIED
C  4  OTHER
V 11  ELECTRICITY                1 18  1
C  1  AVAILABLE           11
C  2  UNAVAILABLE
VARIABLES
CATALOG
```

De batchfile ('demanding file') wordt opgestart vanuit het introductie scherm.

Hiërarchische dataverwerking

Ten opzichte van de meest gebruikte statistische pakketten heeft ARIEL+plus een optie die aparte vemelding verdient, de hiërarchische databewerking. Dit betreft de mogelijkheid om op redelijk eenvoudige wijze data van een bepaald observatie niveau (bijvoorbeeld het individu) te aggregeren naar een dataset op een hoger observatieniveau (bijvoorbeeld het huishouden, de woonplaats, de provincie of nationaal). Andersom is het mogelijk om waarden van een hoger observatieniveau toe te kennen aan bijv. individuen. De procedure genereert dan een datafile op het geaggregeerde niveau, en de oorspronkelijke dataset blijft

aanwezig in de tome. Bij deze procedure kunnen gewichten en selecties worden toegepast voor variabelen in de oorspronkelijke dataset.

Data uit twee of meer themes kunnen gecombineerd worden in 1 nieuw theme. Per theme kan de volgorde van de te kopiëren variabelen worden opgegeven. Combinatie van data uit verschillend georganiseerde datasets is mogelijk.

Opties voor het manipuleren van variabelen

Het programma kent diverse opties voor het manipuleren met variabelen. Nieuwe variabelen kunnen worden berekend uit de waarden van bestaande varabelen. Er kan een variabele als gewicht worden gehanteerd tijdens een verwerking.

Tabel: Enkele commando's voor variabele specificatie en manipulatie

V	Definieert een variabele
C	Definieert een categorie voor een variabele
CREATE	Creatie en wijziging van een variabelen
EXCLUDE	Verwijdert een categorie uit een categoriele variabele.
GROUPING	Categoriseren van kwantitatieve variabelen.
SEGMENTATION	Sorteren van observaties in homogene groepen
SL	Selectie statements beginnen met dit keyword, zowel te gebruiken voor data manipulaties als voor selectieve analyses.
TR	Statement voor specificaties in TRANSFORM en GROUPING
TRANSFORM	Creatie van nieuwe variabele uit waarden van een bestaande variabele (groepering)
TRIAL	Instelling (beperkt) aantal records voor een test run
UNLOCK	verwijdert de protectie status in de dataset, waarna een variabele verwijderd kan worden uit de dataset.
WEIGHT	kent een gewicht (waarde van een variabele) toe aan individuele cases.
WEIGHTHALT	Beeindigt de toepassing van gewichten

Per analyse kan na het invoeren van de regel met een procedure in de volgende regel van de batchfile een selectie opdracht worden gespecificeerd. Het select statement dient (desgewenst) na elke procedure opnieuw te worden gespecificeerd en is na voltooiing van de procedure niet meer van toepassing.

Met het MENDING keyword kunnen de waarden van bestaande variabelen worden gewijzigd op diverse manieren, bijvoorbeeld, met een constante worden vermenigvuldigd, of in het geval dat er sprake is van missing data met een constante worden gevuld. De waarden van variabelen kunnen met 1 of meer opdrachten worden overgebracht naar een set bestaande of nieuwe variabelen. Deze opties zijn vergelijkbaar met de RECODE en COMPUTE procedures in SPSS. De toepassing van die procedures is in vergelijking met SPSS enigszins omslachtig.

Bijvoorbeeld, een samengestelde teller en een samengestelde noemer moeten in ARIEL eerst in aparte statements in hulpvariabelen worden ondergebracht, voordat ze op elkaar kunnen worden gedeeld. In ARIEL zijn meer dan 5 statements nodig voor het uitrekenen van de formule: *V30 = 200 * (V14-V17) / (V14+V17)* (één statement in SPSS).

Tabel: Statistische procedures

ANOCOR	Correspondentie Analyse
BARYCENTERS	Centroid diagram voor 2 kwantitatieve variabelen en een of meer categoriele variabelen.
CLUSTER	Cluster analyse
COMPONENTS	Principale componenten analyse
CORRELATION	Correlatie
CROSSTAB	produceert 2 dimensionele kruistabellen
DISTRIBUTION	Cumulatieve frequentiecurve van een kwantitatieve (niet categoriele) variabele.
GINI	GINI concentratie curve en statistische grootheden voor een kwantitatieve variabele.
HISTOGRAM	Histogram kwantitatieve variabelen, ev. per klasse van een categoriele variabele.
LINEAR	Regressie voor 1 afhankelijke en n onafhankelijke variabelen (n<=100).
LOGIT	Logit Maximum Likelyhood analyse
MCLAS	Multipele classificatie analyse
MCORRESPOND	Multipele correspondentie analyse
PERCENTILES	Percentielen van variabelen.
PIECHART	Taartdiagram.
PREVIEW	Frequentie-tabellen
PROBIT	Probit Maximum Likelyhood Analyse.
QUANTICROSS	Kruistabel met gemiddelde en som.
RECORD	Structureert de ruwe datafile in een theme.
REGRESSION	Voert een regresie analyse uit.
SCATTER	Scatterplot van 2 variabelen.
SHOWDOWN	Cumulatieve frequentie tabel.
STATISTICS	Statistische opties voor kruistabellen
TABLEAU	Een breakdown voor gemiddelde, som en standaard afwijking in multivariate tabellen.

Ondersteunde statistische procedures

ARIEL+plus kent uiteenlopende statistische procedures. Voor de analytische statistische procedures is het aanbod van procedures nogal grillig. Zo worden wel logit-analyse, correspondentie-analyse en een factoranalyse (principale componenten analyse) ondersteund, maar procedures voor een T-test of voor een variantieanalyse zijn niet beschikbaar. Regressie en correlatie zijn wel beschikbaar. Rechte tellingen kunnen worden geproduceerd met het PREVIEW-commando. Voor verdere beschrijvende statistische procedures (kruistabellen en scatterplots) zijn voldoende mogelijkheden aanwezig. De procedure QUANTICROSS maakt het mogelijk om gewogen kruistabellen te produceren. In de tabel wordt het gemiddelde en de som van de gewogen frequenties afgedrukt.

Het aanbod van de procedures en opties is per saldo beperkter dan in de meest gangbare statistische pakketten.

De uitvoer

De uitvoer van de analyses is over het geheel genomen overzichtelijk te noemen. Beschrijvende procedures zoals frequenties en kruistabellen leveren standaard uitvoer met kolom- en rijpercentages. Via het STATISTICS commando kan men vooraf specificeren welke percentages worden afgedrukt. Bij de kruistabellen kunnen geen associatiematen of toetsingsgrootheden worden berekend.

Er worden door het programma uitvoerfiles geproduceerd die eenvoudig zijn te bewerken met elke eigentijdse textverwerker die een standaard ASCII-textfile kan inlezen. Een rechtstreekse afdruk leverde het probleem op dat de instelling van de regelbreedte uitgaat van een 'brede wagen' (132 karakters breed). Dit is binnen het pakket niet anders in te stellen.

Machine-onafhankelijkheid

ARIEL+plus is een BATCH georienteerd programma, dat oorspronkelijk op een mainframe geïmplementeerd is. Het programma is op diverse systemen beschikbaar. Deze beschrijving heeft betrekking op de PC versie. Het programma is geïnstalleerd op diverse personal computers 286/386), en bleek op al deze IBM compatible PC's probleemloos te functioneren. Er wordt binnen het programma, voor zover gebleken geen gebruik gemaakt van grafische hardware. Een mathematische coprocessor is niet vereist.

Gebruiksgemak

Voor het uitvoeren van de testprocedure koste het enige moeite om een datafile voor ARIEL+plus aan te maken, en er statistische procedures op te laten uitvoeren. De user interface voegt weinig toe aan de batchverwerking en is duidelijk minder geavanceerd dan de meeste gangbare statistische pakketten op personal computers. Hoewel de toe te passen procedures in de handleiding beschreven staan, zal het in veel gevallen enige tijd en inspanning vergen om het pakket goed te doorgronden. Daarna blijft het een qua gebruiksgemak minder geavanceerd pakket.

Resumé

ARIEL+plus bleek een pakket, dat zich zonder problemen liet installeren. Het organiseren van de data en de aansturing van het programma via een batchfile verliep aanvankelijk nogal stroef, mede doordat het programma een complexe data organisatie kent. Daarnaast werd het als een stap terug in de tijd ervaren, dat voor de selectie van variabelen en definitie van opties geen namen maar nummers moeten worden opgegeven. Deze ongemakken bezwaren zal men voor lief moeten nemen wil men het programma gebruiken.

Na enkele tests uitgevoerd te hebben bleek dat ARIEL+plus een handzaam programma, waarbij redelijk snel statistische standaardprocedures kunnen worden verwerkt. Tijdens het testen is een bestand van ca. 3000 records met 13 variabelen probleemloos en met redelijke snelheid verwerkt.

De mogelijkheid om zonder kunstgrepen data te reorganiseren op verschillende hiërarchische niveaus van observatie is een duidelijk pluspunt ten opzichte van de gangbare statistische pakketten.

Het arsenaal aan statistische procedures en opties is voldoende voor een verwerking van een survey sample. Het aanbod van geavanceerdere statistische procedures en opties is in vergelijking met gangbare pakketten zoals SPSS of SAS enigszins beperkt. Wat verder vragen oproept is het afwijkende dataformaat van de systemfiles en de minder gangbare namen voor enkele procedures.

De aanschaf van het pakket zou overwogen kunnen worden wanneer men geen data uitwisselt met andere pakketten of instanties. Het pakket kan interessant zijn door de analysemogelijkheden voor hiërarchische data. Een nieuwe versie van het pakket is in ontwikkeling. Bij de relatief hoge aanschafprijs van de 'first copy' (zie de kop van dit artikel) zij opgemerkt dat 'subsequent copies' circa fl. 1600,- kosten. Gezien het aanbod van ondersteunde procedures, en de batch-georiënteerde aansturing, is het vooralsnog de vraag of ARIEL+plus zich een plaats op de statistische software-markt in Nederland kan verwerven.