KM 33(1990)
pag 109-116

RELATIONELE DATABASE-TOEPASSINGEN EN SPSS

Max Kommer[1]

Samenvatting

In dit artikel wordt ingegaan op de mogelijkheden om met behulp van de file-handling faciliteiten van SPSS een aantal aspecten van het relationele gegevensmodel te implementeren. De conclusie luidt dat dit inderdaad mogelijk is, maar dat het wenselijk zou zijn wanneer SPSS met enkele daarop speciaal toegesneden faciliteiten zou worden uitgerust

Inleiding

In hun zeer instructieve artikel over het gebruik van relationele databases in sociaal-wetenschappelijk onderzoek (KM 28, 1988) bepleiten Aerts en Duffhues dat statistische pakketten worden aangepast aan het gebruik van relationele databases. Dat zou onder andere moeten inhouden dat ze direct relationele databases kunnen benaderen, en dat ze selectie en inlezen van gegevens via een relationele vraagtaal mogelijk maken. Hoewel ik deze opvatting geenszins wil bestrijden lijkt het me goed om er op te wijzen dat veel van waar het hen werkelijk om gaat al sinds jaren mogelijk is met het voor de meeste sociaal-wetenschappers ongetwijfeld vertrouwde pakket SPSS. Dat die sociaal-wetenschappers volgens Aerts en Duffhues een eenkennigheid verweten mag worden is dan ook wellicht niet zozeer een gevolg van het feit dat ze nooit een blik geworpen hebben op "het erf van de economische wetenschap of op dat van de administratieve systemen" (A&D, p. 48), maar veeleer van het feit dat ze te weinig in hun manual gekeken hebben. Daarvoor zouden dan overigens wel twee excuses aangevoerd kunnen worden: die manual is nu niet direct het toppunt van toegankelijkheid, en voorzover in de opleiding aandacht geschonken wordt aan het gebruik van (pakketten als) SPSS gaat het vrijwel altijd om de statistische technieken, en zelden om gegevensstructuur en -manipulatie.

Alleen dat laatste al lijkt me reden genoeg om enige aandacht te besteden aan de mogelijkheden van SPSS voor toepassingen zoals beschreven door

[1] De auteur is als onderzoeker werkzaam bij het Wetenschappelijk Onderzoek- en Documentatiecentrum van het Ministerie van Justitie, Postbus 20301, 2500 EH 's-Gravenhage (tel. 070 - 706568).

Aerts en Duffhues. Ik zal mij daarbij overigens in een aantal opzichten beperken. In de eerste plaats zal ik niet ingaan op de vraag in hoeverre het concept van de relationele database management systeem (RDBMS) überhaupt al (volledig) gerealiseerd is (cf. Pascal, 1989). Voor het soort toepasing dat Aerts en Duffhues beschrijven is dat niet relevant: daarbij wordt in feite alleen gebruik gemaakt van de gegevensopslag- en benaderingsfaciliteiten die al door een pakket als dBase III geboden worden. Ik zal mij ook wat SPSS zelf betreft beperken, en wel tot de beide "IBM-dialecten": SPSS^X voor het mainframe, en SPSS/PC$^+$ voor DOS-machines. Een consequentie daarvan is dat ik niet inga op interfaces die in andere dialecten door SPSS of sommige bestaande database-systemen geboden worden, en bijvoorbeeld als user procedure te gebruiken zijn. Voor een aantal andere door mij niet te behandelen aspecten, zoals de efficiëntie, verwijs ik graag naar Johnson en Thomason, 1988.

SPSS en het relationele model

Eén van de eerste zaken waarmee de gebruiker van SPSS rekening dient te houden is dat het pakket gebaseerd is op wat de conventionele methode van gegevensopslag genoemd kan worden: de rechthoekige datamatrix. De gegevens worden beschreven met behulp van een data list, waarin de variabelen worden gedefinieerd die (in principe) voor elke case aanwezig zullen zijn, en die ook in elke case dezelfde betekenis hebben. Dit lijkt volstrekt strijdig met het relationele model, tenzij men zich realiseert dat een relationele database niets anders is dan een aantal rechthoekige datamatrices, die op een slimme wijze aan elkaar gekoppeld worden. Het grote verschil[2] tussen (relationele) database-pakketten en SPSS zit hem dan ook in de positie die de koppeling van bestanden in het concept inneemt: in de eerste is die koppeling zo ongeveer de basis waarop toepassingen worden gebouwd, in SPSS is koppeling van bestanden (afgezien dan natuurlijk van de koppeling die plaatsvindt bij het gebruik van VAR en VALUE LABELS) een bijna exotische mogelijkheid die in één hoofdstukje van de manual aan de orde komt. Daarbij is het zo dat de wijze waarop de koppeling technisch gerealiseerd wordt nogal verschilt, en dat heeft zowel consequenties voor de opbouw van de te koppelen bestanden als voor de wijze waarop er mee gewerkt wordt. Om dit duidelijk te maken is het nodig heel kort iets te zeggen over de wijze waarop de

[2] Uiteraard zijn er meer, en zeer belangrijke verschillen. Zo is de rol van de Data Dictionary in een relationele database-toepassing wel iets meer dan die van de mogelijkheid om in SPSS variable en value labels toe te kennen.

bestanden door de programmatuur benaderd worden.
Database-pakketten zijn ten principale record-geöriënteerd: ze zijn er op gericht om uit een bestand één record te selecteren en ter bewerking of beoordeling aan te bieden. Dit kan zowel interactief als in batch-verwerking. Om de selectie zo efficiënt mogelijk te laten verlopen zal over het algemeen gebruik gemaakt worden van geïndexeerde bestanden, dat wil zeggen dat elk record direct toegankelijk is via één of meer sleutel(s). SPSS daarentegen is bestand-georiënteerd: een bewerking (de relevante bewerkingen zijn hier datatransformaties en bestandsmanipulaties, niet de statistische procedures) wordt uitgevoerd voor elk record in het bestand. Tot en met versie 2 van SPSS^X worden de SPSS-commando's altijd als batch aangeboden[3]; in versie 3 is ook interactief werken mogelijk. Deze gang van zaken maakt het mogelijk een invoer-bestand eenvoudigweg record voor record van voren af aan te lezen (sequentiële benadering). In de meeste gevallen worden de records weer weggeschreven naar een hulpbestand; dit hulpbestand wordt het invoerbestand voor een volgende procedure. Het is uiteraard wel mogelijk om via een IF-statement een datatransformatie op slechts één record uit te voeren, maar dan worden toch alle records gelezen teneinde de voorwaarde te kunnen evalueren.
Het ligt voor de hand dat dit verschil in benaderingswijze gevolgen heeft voor de wijze waarop de koppeling technisch gerealiseerd kan worden. In het geval van een database-pakket kan bij elk willekeurig record uit het ene bestand direct het bijbehorende record uit één of meer andere bestanden gevonden worden. De consequentie is dat er geen enkele eis gesteld wordt aan de volgorde van de records. Bij SPSS ligt dat geheel anders: om een koppeling tot stand te kunnen brengen is het noodzakelijk dat alle invoerbestanden gesorteerd zijn op oplopende waarde van de sleutel(s). De koppeling vindt namelijk plaats doordat uit het eerste bestand het eerste record wordt gelezen, vervolgens wordt het eerste record uit het tweede bestand gelezen, enzovoort tot het laatste bestand. Zijn de waarden van de sleutels gelijk dan vindt koppeling plaats, en herhaalt de procedure zich. Zijn ze ongelijk dan wordt aan het hulpbestand een record toegevoegd waarin de variabelen uit het invoerbestand met de laagste waarde voor de sleutel worden overgenomen,

---

[3] Daarbij geldt dan ook nog dat commando's die geen procedure-status hebben pas worden uitgevoerd op het moment dat in de batch een procedure-commando wordt aangetroffen (dat laatste geldt overigens ook voor het interactieve SPSS-PC, en naar ik aanneem eveneens voor versie 3 van SPSS^X).

en die uit de andere bestanden een missing value krijgen. Daarna wordt het volgende record uit het invoerbestand met de laagste sleutelwaarde gelezen, en worden de sleutels opnieuw vergeleken.
Uit deze beschrijving moge duidelijk zijn dat de wijze waarop de koppeling plaatsvindt weliswaar verschilt tussen SPSS en echte database-pakketten, maar dat in principe dezelfde resultaten verkregen kunnen worden. Database-pakketten en SPSS maken het beide mogelijk om zowel een één-op-één koppeling uit te voeren (bij koppeling van twee bestanden betekent dit dat in beide bestanden evenveel records voorkomen, met elk een unieke waarde op de voor de koppeling gebruikte sleutel) als een één-op-veel koppeling (in het ene bestand komen verschillende records voor met dezelfde waarde op de voor de koppeling gebruikte sleutel, in het andere bestand komt elke waarde van die sleutel slechts één maal voor). Een één-op-veel koppeling kan overigens op twee verschillende manieren gerealiseerd worden: als output kan een bestand ontstaan waarin de gegevens uit het "één-bestand" worden gespreid over de corresponderende records van het "veel-bestand", of een waarin de gegevens uit het "veel-bestand" worden opgeslagen in een reeks variabelen van het "één-bestand" (zie voor een voorbeeld van het laatste Banks, 1988, p. 6). In principe is ook een veel-op-veel koppeling (bv. mensen aan verenigingen) mogelijk, doch dit is een bijzondere situatie waarbij het gebruik van de gekoppelde gegevens bepaalt op welke wijze de koppeling het best gerealiseerd kan worden.

Koppeling van SPSS-bestanden in de praktijk

Het commando waarmee binnen SPSS twee of meer bestanden gekoppeld kunnen worden is het MATCH FILES commando. In de eenvoudigste situatie, waarin twee parallelle files (bestanden met dezelfde cases, in dezelfde volgorde) gekoppeld moeten worden, hoeft dit commando alleen gespecificeerd te worden met de namen van de beide files. Nu is dit, zeker in het verband van dit artikel, geen interessante situatie. Waar het om gaat is de koppeling van bestanden via een sleutel. Ook dit is mogelijk. Stel dat de variabele idnr de sleutel is (hij moet dan in beide bestanden voorkomen, en beide bestanden moeten er op gesorteerd zijn). Een één- op-één koppeling vindt dan plaats met het volgende commando:

```
MATCH FILES    FILE = file1 / FILE = file2 / BY = idnr
```

De één-op-één koppeling houdt overigens niet in dat alle cases in beide

bestanden moeten voorkomen (zoals bij parallelle files), maar wel dat elke waarde voor de sleutel in elk bestand slechts één maal voorkomt. De variabelen van een record dat in één van de invoerbestanden niet voorkomt krijgen in het resultaatbestand een missing value.
Een één-op-veel koppeling vindt plaats door het bestand waarin elke waarde voor de sleutelvariabele slechts éénmaal voorkomt aan te duiden als TABLE inplaats van als FILE. De tweede file functioneert in dit geval inderdaad als "tabel": bij elk record uit de eerste file wordt aan de hand van de sleutel de tabel-file geraadpleegd, teneinde de waarden van de daarin voorkomende variabelen op te nemen in het resulterende bestand. Over het algemeen zal de tabel-file minder cases bevatten dan de andere file, maar het kan ook omgekeerd. In dat laatste geval bevat het resulterende bestand overigens maar evenveel cases als de oorspronkelijke; de cases uit de tabel-file waarvoor geen corresponderende case in de oorspronkelijke file aanwezig is worden niet (aangevuld met missing values) naar het resulterende bestand geschreven. Er zijn nog een aantal extra specificaties mogelijk bij het MATCH FILES commando. De belangrijkste daarvan is de mogelijkheid om verschillende logische variabelen te creëren: één om aan te geven of een case al dan niet in een bepaald bestand voorkwam (IN), en twee om aan te geven dat de betrokken case de eerste resp. laatste was met dezelfde sleutelwaarde (FIRST en LAST). Na koppeling zal veelal op een van deze variabelen geselecteerd moeten worden om een relationele bewerking te completeren.

Een praktische toepassing

In hun artikel geven Aerts en Duffhues een onderzoekvoorbeeld waarvoor het relationele model geschapen lijkt: een onderzoek naar de katholieke beweging in Arnhem. Voor dit onderzoek wordt gebruik gemaakt van netwerkanalyse, waarmee de relatiestrucuur tussen twee verzamelingen onderzoekseenheden (personen en organisaties) wordt blootgelegd. Dergelijke analyses kunnen worden uitgevoerd met behulp van speciale programmatuur (zoals GRADAP), maar Aerts en Duffhues laten zien dat door gebruikmaking van het relationele model voor de gegevensopslag ook met conventionele statistische programmatuur een deel van de onderzoekvragen te beantwoorden is.
Zij schetsen daartoe eerst de bestandsopbouw, waarvan het belangrijkste kenmerk is dat er twee entiteiten-bestanden zijn (één met gegevens over personenen en één met gegevens over organisaties), alsmede drie relatie-

bestanden (één dat gegevens bevat over de functies die personen in de organisaties bekleedden, één met gegevens over hun financiële bijdragen aan die organisaties, en één met familierelaties).

De analyse van het op bovenbeschreven wijze opgeslagen materiaal vindt plaats door het koppelen (op basis van de sleutels) van die records uit twee of meer van de bestanden, die aan een bepaald selectiecriterium voldoen. De meeste door Aerts en Duffhues besproken gevallen zijn ook in SPSS vrij eenvoudig te realiseren met de faciliteiten die het MATCH FILES commando biedt. Eén geval is echter beduidend lastiger, namelijk de echte netwerkanalyse.
Het probleem is de uitvoering van een inductie-operatie: het genereren van een bestand waarin relaties zijn vastgelegd tussen personen, bijvoorbeeld het feit dat ze in een bepaald peiljaar een functie vervulden in dezelfde organisatie. Het zal direct duidelijk zijn waarom SPSS hiermee een probleem heeft: omdat de relaties van de afzonderlijke personen met de organisatie(s) waarin ze een functie bekleedden zijn vastgelegd in een bestand met één record per relatie, komt het leggen van een relatie tussen personen neer op een "across-cases"-operatie, en dan nog één waarbij n functies n/ ((n-2)* 2) relaties opleveren. Aangenomen mag worden dat voor de meeste organisaties n > 2, en dat betekent dat uit een invoerbestand een uitvoerbestand opgebouwd worden met (beduidend) meer cases - niet direct een standaardbewerking in SPSS. Er zijn twee mogelijkheden om dit probleem op te lossen. De eerste is de gegevens uit het PEILJAARSELECTIE-bestand[4] in ruwe-data vorm weg te schrijven, en vervolgens met een speciaal geschreven INPUT PROGRAM weer zodanig in te lezen dat de koppeling over de cases tot stand komt. Dit is uiteraard alleen met SPSS^X mogelijk; SPSS/PC kent geen INPUT PROGRAM. De tweede is om op het PEILJAARSELECTIE-bestand een LAG-operatie uit te voeren, het resultaat als file weg te schrijven en daarop dan weer dezelfde operatie los te laten. Dit moet dan even vaak herhaald worden als het grootste aantal functies per organisatie. Het complete relatiebestand wordt daarna verkregen door alle uit deze stappen verkregen relatiebestanden samen te voegen.
Het is duidelijk dat in een geval als dit het gebruik van SPSS een veel omslachtiger manier van werken met zich meebrengt dan door Aerts en

---

[4] Hieronder wordt verstaan een selectie voor een bepaald peiljaar uit het organisatie/personen-relatiebestand.

Duffhues kon worden beschreven. In feite wordt namelijk van het feit dat de gegevens volgens het relationele model zijn opgeslagen alleen in die zin gebruik gemaakt dat de bewerking wordt uitgevoerd op een relatiebestand. Voor het overige gaat het om een relatief simpele maar tot vervelens toe herhaalde bewerking op dat éne bestand.

Conclusie

Allereerst moet vastgesteld worden dat SPSS er, in vergelijking met de door Aerts en Duffhues gebruikte vraagtaal, op het eerste gezicht niet gunstig afkomt als het gaat om het formuleren van compacte en inzichtelijke queries. Vooral de bovengeschetste inductie-operatie maakt dit duidelijk. Hierbij moet echter wel de kanttekening gemaakt worden dat, welk pakket ook gebruikt wordt, het exact uitdenken welke koppelingen en selecties tot stand gebracht moeten worden lastiger is dan het op een rijtje zetten van de commando's waarmee dat gebeurt. Een goede vraagtaal en de aanwezigheid van het hele arsenaal aan relationele operatoren kunnen daarbij een zeer krachtig hulpmiddel zijn, maar dat lost op zich slechts het technische en niet het logische probleem op.
In het verlengde hiervan kan vastgesteld worden dat de sequentiële verwerking waarop SPSS is gebaseerd moeilijk verenigbaar is met het relationele model. Wat dat betreft zou het al een stap voorwaarts zijn wanneer SPSS het gebruik van geïndexeerde bestanden zou ondersteunen; het telkens sorteren zou dan kunnen vervallen. Een veel ingrijpender stap zou natuurlijk zijn de volledige implementatie van het relationele model, inclusief een goede vraagtaal.
Zou de door Aerts en Duffhues bepleitte tussenweg, nl. het opnemen van een interface naar bestaande implementaties van het relationele model, een bruikbaar alternatief zijn? Het antwoord is m.i. ja en nee. Ja, omdat het waarschijnlijk de eenvoudigste oplossing[5] is (in sommige versies van SPSS bestaat reeds de mogelijkheid om direct relationele databases te benaderen; zie Keywords no. 2, 1988), en omdat er nu eenmaal op vrijwel alle terreinen waar SPSS-gebruikers zich mee bezighouden al veel databases zijn met uiterst relevante gegevens. Nee, omdat wie alleen

---

[5] Maar dan moet het wel een echte interface zijn, die het gebruik ondersteunt van de vraagtaal van het RDBMS. Een interface zoals SPSS/PC$^+$ nu reeds biedt naar dBASE III$^+$ is volstrekt onvoldoende: het komt er op neer dat men een (één) dBASE-bestand direkt kan inlezen, dus zonder het eerst als ASCII-bestand weg te schrijven. Het pakket dBASE STATS (een door ASHTON-TATE op de markt gebracht produkt van SPSS Inc.), dat het mogelijk maakt een aantal statistische procedures uit SPSS aan te roepen vanuit een dBASE applicatie, is dan al wat interessanter. Het heeft echter als bezwaar dat de toekenning van VAR en VALUE labels omslachtig is, en dat er veel tijd verloren gaat met het -toch weer- omzetten van het dBASE-bestand in een SPSS-workfile.

gebruik maakt van "nieuwe" gegevens en dus zelf de hele database nog moet bouwen en vullen, in dat geval toch weer met twee softwarepakketten moet werken, en omdat dan niet ten volle gebruik gemaakt kan worden van de uitstekende faciliteiten die SPSS biedt voor gegevensdocumentatie (LABELS, MISSING VALUES).

Op zich maakt het echter weinig uit op welke wijze relationele structuren geïmplementeerd worden in een pakket als SPSS. Wezenlijker is dat eenieder die zich met gegevensanalyse bezighoudt zich ervan bewust is dat het even zinvol en minstens even efficiënt kan zijn om aandacht te besteden aan de structurering van de gegevensbestanden als om te zoeken naar de meest geschikte statistische techniek. Wat dat betreft kan ik volledig met Aerts en Duffhues instemmen.

Referenties

Aerts, E. en T. Duffhues — Relationele databases in sociaal-wetenschappelijk onderzoek: een pleidooi
In: Kwantitatieve Methoden, 1988, no. 28

Banks, R. — SIR vs SPSS - the case of the General Household Survey
Contributed paper to the 1st user group conference of the SPSS European User's Group, Amsterdam, 1988

Johnson, M. & J.C. Thomason — Database Management, an Alternative Use for SPSS?
Contributed paper to the 1st user group conference of the SPSS European User's Group, Amsterdam, 1988

Pascal, F. — A Brave New World?
In: Byte, sept. 1989, pp. 247 - 256

SPSS, Inc. — Keywords, 1988, no. 2

Ontvangen: 01-1989
Geaccepteerd: 26-10-1989