KM 43 (1993)
pg. 119-149

TRANSITIETABELANALYSE EN ML-SCHATTINGEN VOOR PARTIEEL GECLASSIFICEERDE VERKIEZINGSDATA VIA LOGLINEAIRE MODELLEN

Marc Swyngedouw [1]

In dit artikel wordt ingegaan op de problematiek van partiële non-respons. Vertrokken wordt van een algemene situering van de verschillende technieken voor de aanpassing van steekproefgegevens vertekend door non-respons. Vervolgens worden de verschillende machanismen die in non-respons kunnen gevonden worden besproken. Hierbij wordt een onderscheid gemaakt tussen 'ignorable' (Missing Completely Ad Random en Missing Ad Random) en 'non-ignorable' non-responsmechanismen. Aan de hand van een eenvoudig voorbeeld wordt uitgelegd hoe in geval van transitietabellen, gebruik makende van het al dan niet 'genest' patroon van de non-respons toch ML-schattingen voor de te schatten celfrequenties kan verkregen worden d.m.v. een EM-algorithme. Terwijl tevens een test mogelijk wordt voor het veronderstelde loglineair model. Een analyse voor de verschuivingen tussen twee opeenvolgende verkiezingen (1981 - 1985) voor de (Belgische) provincie Limburg wordt uitgevoerd. Een 'movers - stayers' model gegeven de partiële non-respons wordt getest.

---

[1] dr. Marc Swyngedouw is werkzaam als hoofddocent aan de K.U.Brussel (02-412.43.59) en is onderzoekdirecteur van het Interuniversitair Steunpunt Politiek Opinie-onderzoek (ISPO), K.U.Leuven, Van Evenstraat 2c, 3000 Leuven, 016-28.31.59.

Bij de analyse van steekproefgegevens naar verschuivingen tussen twee opeenvolgende verkiezingen bemerken we regelmatig - al dan niet na post-stratificatie aan de hand van leeftijd, geslacht en sociaal niveau - door middel van een chi-kwadraat toets, dat niet mag aangenomen worden dat de steekproef-marginale verdeling voor de verkiezingen afkomstig zijn uit de populatieverdeling zoals we die kennen vanuit de verkiezingsresultaten. Dit kan verklaard worden door zeer verschillende factoren: bv. kiesintenties in plaats van kiesgedrag, non-respons, item non-respons o.i.v. vraagverwoording (Billiet, 1990) enz.. Door middel van een 'Iterative Proportional Fitting'-procedure (Bishop, Y, Fienberg, S, Holland, P, 1975) kunnen de hoofdeffecten van de transitietabel dan aangepast worden aan de verkiezingsuitslagen met behoud van de associatiestructuur van de geobserveerde transitietabel. Een andere meer voorkomende manier van weging is post-stratificatie op basis van de verkiezingsuitslagen. Pannekoek (1984) geeft aan onder welke voorwaarden beide methodes tot hetzelfde resultaat leiden.

Deze twee manieren van 'population weighting' hebben, zoals Kalton en Kasprzyk het in hun overzicht van de behandeling van 'missing survey data' (1986 : p. 1-16) opmerken, de bedoeling om zowel te compenseren voor steekproeffluctuaties en 'coverage errors' als voor non-respons. De laatst genoemde categorie kan dan slaan op de zogenaamde 'weigeraars' - deze respondenten die niet deelnemen aan het onderzoek - alsook op *die* respondenten die slechts partieel informatie hebben verstrekt. *Wel* worden binnen deze wegingsmethode de respondenten met een partieel responsgedrag voor de te analyseren variabelen weggelaten. Het grote voordeel van de hier gebruikte methoden is dat ze geen expliciete informatie veronderstelt met betrekking tot de non-respondenten.

Een totaal andere wijze van gegevensaanpassing betreft de zogenaamde 'imputation methodes'. Deze methodes behandelen de zogenaamde 'item-non-response' of partiële non-respons. De bedoeling van deze methodes (zie ook Kalton en Kasprzyk, 1986) is te komen tot 'volledige datarecords' voor alle respondenten die niet geweigerd hebben mee te werken aan het onderzoek, maar die om een of andere reden niet in staat waren of weigerden informatie te verstrekken betreffende bepaalde items. Zeer uiteenlopende methodes worden hiertoe gebruikt, gaande van 'mean imputation', 'random imputation' - beiden al dan niet 'within classes' - over 'hot-deck

imputation' - waaronder een duplicatieproces verstaan wordt waarbij de informatie uit de betreffende steekproef wordt gehaald - naar ingewikkelde 'regression imputation' en 'distance function matching'.

Zowel de wegings als de 'imputation'methode hebben in bepaalde omstandigheden, waarvan de bespreking buiten het bestek van dit artikel valt, het nadeel dat zij de meting van de relatie tussen variabelen kunnen verstoren. Opvallend is echter dat 'imputation' ook het nadeel heeft, de schattingen van de standaardfout te verstoren. De standaardfouten berekend op de traditionele wijze zullen bij data met 'imputation' in het algemeen onderschat zijn vanwege de creatie van de ingevulde data. In een poging dit te verhelpen, stelt Rubin (1986) voor om te werken met 'multiple imputation'. Hoewel dit in zekere zin een vorm van 'imputation' is, wordt deze methode in de literatuur veelal afzonderlijk behandeld. Bij 'multiple imputation' wordt de constructie van de volledige verzameling van gegevens verschillende malen uitgevoerd door middel van 'imputation' waarbij gebruik wordt gemaakt van *één* 'imputation' methode (die een bepaalde vorm van non-respons-mechanisme veronderstelt). De te schatten populatieparameter wordt dan voor elk van de gecreëerde groepen van gegevens berekend, alsmede het gemiddelde ervan over de verschillende gecreëerde 'volledige' gegevensgroepen. De variantie van de schatter wordt vervolgens bepaald aan de hand van twee componenten: de gemiddelde 'within imputation' variantie geassocieerd met de schatting *en* de 'between imputation' variantie van de schatting. Door vervolgens dezelfde procedure toe te passen met een andere 'imputation' methode, kan tevens de gevoeligheid van de schatter worden nagegaan voor de verschillende vormen van 'imputation'.

In theorie zou de 'imputation' methode ook bruikbaar kunnen zijn bij non-respons die ontstaat door 'weigeringen'. Dit veronderstelt dan wel dat over deze groep van non-respondenten voldoende apriorische informatie aanwezig is. Iets dat in de meeste gevallen niet het geval is. De toepassingen dienen dan ook voornamelijk in de sfeer van de item-non-respons gezien te worden.

Een derde methode om de analyse van gegevens met ontbrekende waarden te benaderen, maakt gebruik van statistische interferentie, die gebaseerd is op waarschijnlijkheden ('likelihood') afgeleid uit formele statistische modellen voor de gegevens en het

'missing data'-mechanisme. Bij een bepaald 'missing data'-mechanisme worden de meest aannemelijke schattingen voor de compleet geobserveerde gegevens en de partieel geobserveerde gegevens berekend vanuit een bepaald model. Ook hier geldt dezelfde opmerking als bij de 'imputation' methode, namelijk dat in theorie toepassingen bedacht kunnen worden voor de zogenaamde weigeraars. In de praktijk dienen de toepassingen beperkt te worden tot item-non-respons. Deze wijze van non-respons-benadering wordt ook wel beschreven als een methode die tracht de voldoende statistieken ('sufficient statistics') die men nodig heeft bij een bepaalde analysemethode, te schatten. Voldoende statistieken zijn deze gegevens zonder dewelke men bepaalde schattingsprocedures niet kan toepassen (zie Hagenaars, 1985: 140). Bij regressieanalyse of factoranalyse betreft het hier de schatting van de correlatie of covariantiematrix, bij loglineaire modellen schat men de celfrequenties.

Hieruit volgt, dat men de voorgestelde IPF-methode ook onder dit type van non-respons 'aanpassing' kan klasseren. Als verondersteld wordt dat de vastgestelde 'bias' op de marginale verdeling slechts het gevolg is van de non-respons en dat we de werkelijke marginaal kennen, zou een IPF-herschatting in functie van de gekende marginale verdeling, die meest aannemelijke schattingen voor de celfrequenties oplevert - en dus de voldoende statistieken - een toepassing zijn van dit type van non-respons-aanpassing.

Voor een uitgebreide bespreking, de stand van zaken, case-studies en theorieën betreffende de verschillende methodes die, al dan niet ad hoc, aangewend worden bij 'incomplete data' kan verwezen worden naar Madow e.a. (1983 a,b,c) en Little & Rubin (1987). Met het oog op de door ons behandelde gegevens, die van nominaal niveau zijn, behandelen we in het voorliggende artikel één specifieke inschattingsmethode voor partiële non-respons voor kruistabellen. Hierbij verloopt de modelspecificatie via de loglineaire modellen. Verder beperken we ons tot de tweedimensionale transitietabel voor twee opeenvolgende verkiezingen. We veronderstellen hierbij dat het 'missing data'-mechanisme 'Missing At Random' (MAR) is. Het is onze bedoeling een verkenning te maken van de mogelijkheden en moeilijkheden die deze vorm van statistische analyse met ontbrekende gegevens biedt.

Twee ons inziens belangrijke motieven liggen ten grondslag aan

deze oefening (zie Chen en Fienberg, 1974 : 637-638). Wanneer in de dagelijkse praktijk van de statistische analyse de respondenten met partiële informatie weggelaten worden uit de analyse, wordt er vanuit gegaan dat deze 'non-respondenten' een toevallige steekproef vormen uit de bestudeerde populatie die op de te onderzoeken karakteristieken en hun multivariate samenhang niet verschilt van de respondenten. Indien deze veronderstelling correct is, volgt hieruit dat de schattingen verkregen aan de hand van de respondenten, zuiver zijn. Dat wil zeggen, dat wanneer de steekproef N naar oneindig tendeert, de schattingen naar de ware waarden convergeren. Verderop zullen we zien dat hier een 'Missing Completely At Random' (MCAR-)mechanisme wordt verondersteld. Indien aan deze voorwaarde niet voldaan is zullen de verkregen schattingen niet zuiver zijn. Als daarentegen deze veronderstelling toch correct is, zullen de verkregen schattingen *wel* consistent zijn. Maar ze zullen minder efficiënt zijn, in *die* zin, dat hun variantie groter zal zijn, dan de schattingen die verkregen werden door gebruikmaking van de aanwezige partiële non-respons. Dit alles bij hetzelfde formele statistische model en 'missing data'-mechanisme. Hagenaars verwoordt dit als volgt (1987 : 57): aangezien de cellen van multivariate frequentietabellen voldoende waarnemingen moeten bevatten om betrouwbare schattingen van de parameters te verkrijgen en om de gepostuleerde modellen op hun empirische adequaatheid te kunnen toetsen, is het van belang alle gegevens zo optimaal mogelijk te benutten en niet te veel onderzoekselementen van de analyse uit te sluiten vanwege het ontbreken van scores op slechts een enkele variabele.

In de eerste paragraaf schetsen we de vorm waarin de door ons te analyseren gegevens zich aanbieden en presenteren we de gebruikte gegevens. In de volgende paragraaf gaan we kort in op de verschillende mogelijke 'missing data'-mechanismen en bespreken we de schattingsprocedure waarmee de vereiste ML-schattingen voor ons probleem verkregen kunnen worden. Tenslotte bekijken we de resultaten die verkregen werden in een door ons ondernomen toepassing.

1 Het patroon van de gegevens

We bekijken de situatie waarbij we de transitietabel tussen twee opeenvolgende verkiezingen (t, t+1) wensen te analyseren. We

hebben dan een tweewegs kruistabel waarbij m_{jk} het aantal volledig geclassificeerde respondenten voor cel (i,j) voorstelt. Een aantal respondenten kan of wil echter niet te kennen geven voor welke partij hij/zij gestemd heeft op tijdstip t en/of op tijdstip t+1. We verkrijgen dan drie supplementaire (partiële) tabellen M, R en Q. Het aantal respondenten die partieel geclassificeerd zijn op j^{de} rij wordt genoteerd met $r_{j.}$. Dat wil zeggen, we kennen hun stemgedrag op tijdstip t, maar zijn in het ongewisse betreffende hun stemgedrag op tijdstip t+1. Het aantal respondenten die partieel geclassificeerd zijn op de k^{de} kolom noemen we $q_{.k}$. Met andere woorden, we kennen hun stemgedrag op tijdstip t+1, maar niet op tijdstip t. Tenslotte hebben we nog de respondenten waarvan noch op tijdstip t noch op tijdstip t+1 bekend is waarvoor ze gestemd hebben. Deze laatste groep kan bestaan uit weigeraars of uit respondenten die de twee betreffende variabelen niet hebben beantwoord. Zij vormen een één-cel tabel z.

De totale steekproefomvang n bestaat dus uit

$$n = \Sigma_j\Sigma_k\, m_{jk} + \Sigma_j\, r_{j.} + \Sigma_k\, q_{.k} + z \qquad (1)$$

In de traditionele gegevensanalyse zouden we slechts gebruik maken van $\Sigma_j\Sigma_k\, m_{jk}$. Wij wensen echter ook de respondenten die over *een* van beide variabelen informatie hebben verstrekt bij de analyse te betrekken. Uitsluitend de groep waarvoor geen enkele informatie beschikbaar is (z), wordt buiten de analyse gelaten.

Het Studiecentrum W. Claes (1985) organiseerde op 13.10.'85 een 'exit-poll' bij 888 respondenten-kiezers die representatief voor de globale (Belgische) Limburgse kiezerspopulatie. Van hen hadden 788 deelgenomen aan zowel de verkiezingen van 1981 als aan deze van 1985. Het betreft hier een dwarsdoorsnede met retrospectieve bevraging over het stemgedrag van 1981.

Tabel 1 Volledig en partieel geclassificeerde geobserveerde gegevens met betrekking tot verschuivingen 1981 - 1985 voor de provincie Limburg. (Pseudo-Bayes-schattingen met uniforme apriori-probabiliteitsmatrix voor de volledig geclassificeerde gegevens [1]).

1985 1981	CVP	PVV	SP	VU	AGA	AND	BL/O	MISS '81	TOT
CVP	170,67	10,3	13,99	11,02	4,08	4,08	1,10	6,00	220,96
PVV	3,09	80,43	9,03	3,09	0,11	0,11	0,11	1,00	96,97
SP	8,04	3,09	116,70	3,09	4,08	4,08	1,10	7,00	197,17
VU	7,05	5,07	11,02	75,47	1,10	1,10	0,11	1,00	101,92
AGA	2,09	0,11	3,09	0,11	13,00	0,11	0,11	1,00	19,62
AND	0,11	1,10	2,09	4,08	4,08	3,09	1,10	0,00	15,64
BL/O	2,09	0,11	0,11	0,11	0,11	0,11	3,09	0,00	5,73
MISS'85	11,00	8,00	14,00	8,00	6,00	2,00	5,00		
TOT	204,14	107,93	220,03	104,96	32,55	14,67	11,72		712,00

Van alle respondenten gaven er 76 of 9,64% geen antwoord op de beide vragen naar hun stemgedrag in 1981 en in 1985. Daarenboven gaven 16 respondenten (2,03%) antwoord met betrekking tot hun stemgedrag in 1985 maar gaven geen antwoord met betrekking tot hun stemgedrag in 1981. Wel voor 1981 maar niet voor 1985 kregen we informatie van 54 respondenten of 6,85%. Indien we slechts met de volledig geclassificeerde respondenten zouden werken, zouden we een steekproefgrootte van 642 eenheden hebben. Betrekken we de partieel geclassificeerde respondenten mee in de analyse, dan hebben we 712 eenheden als steekproefomvang of een procentuele toename van 10,9%. Tabel 1 geeft de volledige en partiële gegevens weer.

2 Non-respons-mechanismen

"If units are selected by probability sampling, then the mechanism (leading to 'missing' survey variables for units that are not selected) is under control of the sampler (provided the design is succesfully implemented) and may be called 'ignorable'. If the

sampling frame is deficient or some units do not respond, however, then the mechanism leading to the observed data is not so well understood. Any analysis of the data is then dependent on assumptions about the missing-data mechanism, which should be made explicit." (Little & Rubin, 1987 : 9). Ook indien men partiële respons wil betrekken in de analyse van de gegevens, is het dus noodzakelijk dat men aangeeft op welke wijze de partiële non-respons tot stand is gekomen. Twee essentieel verschillende mechanismen worden hierbij onderkend (zie o.a. Rubin, 1976; Hagenaars, 1985,1988; Little, 1982; Rubin en Little, 1987): 'ignorable' en 'nonignorable' non-respons-mechanisme.

Bij het 'ignorable' non-respons-mechanisme wordt er van uitgegaan dat de non-respons tot stand komt op basis van toeval. Wanneer men er van uitgaat dat respondenten en non-respondenten twee aselecte toevalssteekproeven uit dezelfde populatie vormen, heeft men te maken met de meest vergaande vorm van het 'ignorable' non-respons-mechanisme. Het mechanisme wordt dan 'Missing completely At Random' (MCAR) genoemd. De (multivariate) frequentieverdelingen op de onderzochte variabele(n) worden verondersteld dezelfde te zijn voor respondenten en non-respondenten (indien deze laatsten geobserveerd zouden zijn). Deze vorm van non-respons-mechanisme wordt impliciet verondersteld in dat statistisch onderzoek waarin men de analyse beperkt tot de respondenten die op alle te onderzoeken variabelen hebben geantwoord.

Het is duidelijk, dat de veronderstelling van een MCAR-mechanisme zeer vergaand is en in de werkelijkheid zelden zal voorkomen. Een minder vergaand 'ignorable' non-respons-mechanisme stelt, dat de non-respons toevallig tot stand komt, niet bij de gehele populatie, maar binnen subgroepen. Er wordt dan aangenomen dat de populatie kan worden opgedeeld in subgroepen die elk een verschillende kans op respons hebben. Binnen deze subgroepen zal non-respons echter volledig op toeval berusten. Het non-respons-mechanisme wordt dan 'Missing At Random' (MAR) genoemd. Weegprocedures zoals post-stratificatie en celweging (door middel van de inversie van de respons-kans) veronderstellen deze afgezwakte vorm van 'ignorable' non-respons-mechanisme (Little en Rubin, 1987: 55 - 60).

Wanneer we te maken hebben met partiële non-respons, impliceert de MAR-veronderstelling dat bij bepaalde waarden van de geobser-

veerde variabelen, de 'vermiste' waarden op de andere variabelen toevallig zijn. Hagenaars (1985 : 139) beschrijft deze veronderstelling met al zijn consequenties als volgt: "Duiden we de variabelen waarvan de scores voor respondenten en non-respondenten bekend zijn, aan met A (A_1, A_2, ...) en die waarvan de scores voor de non-respondenten ontbreken, met B (B1, B2, ...) dan impliceert het MAR-mechanisme voor de variabelen die bij de correctieprocedure betrokken worden, dat de multivariate frequentieverdeling van de A-variabelen evenals die van de B-variabelen voor respondenten en non-respondenten verschillend is, maar dat de relatie tussen de A-variabelen en de B-variabelen voor respondenten en non-respondenten gelijk is en dat de verschillen wat betreft de B-variabelen verdwijnen met constanthouding van de A-variabelen. Binnen subgroepen, gevormd door de categorieën van A, is de multivariate frequentieverdeling van B voor respondenten en non-respondenten gelijk".

Fundamenteel verschillend met de voorgaande non-respons-mechanismen is de 'nonignorable' non-respons. De non-respons is dan niet toevallig, ook niet binnen subgroepen, maar hangt samen met de onderzochte variabelen. Hierbij dient een onderscheid te worden gemaakt tussen de situatie waarin men het 'nonignorable' non-respons-mechanisme kent en deze waarin dit mechanisme niet gekend is. Een voorbeeld van een gekend 'nonignorable' non-respons-mechanisme is de situatie waarbij een 'censurering' optreedt in de steekproef: vanaf een bepaalde waarde/tijdstip/punt c bijvoorbeeld zijn de gegevens 'missing'.

De correctiemethodes die ontwikkeld zijn om de partiële non-respons te betrekken in de data-analyse, gaan er meestal van uit dat het 'missing data'-mechanisme 'ignorable' is. De reden hiervan is, dat wanneer men aanneemt dat het mechanisme 'nonignorable' is, men precies dient aan te geven hoe het non-respons-mechanisme dan wel verloopt. In de meeste gevallen is men hiertoe niet in staat. Gebruikt men een foutief gespecificeerd 'nonignorable' non-respons-mechanisme in de modellering, dan kan dit ertoe leiden dat de vertekening slechts groter wordt.

3 ML-schattingen voor partieel geclassificeerde gegevens via loglineaire modellen bij een 'ignorable' MAR-non-respons-mechanisme

Dat de 'missing' observaties MAR zijn, wil dus zeggen dat de observaties van de volledige tabel en de observaties van de partiële tabellen die dezelfde karakteristieken hebben op de geobserveerde variabelen niet systematisch verschillen op andere variabelen. Mathematisch uitgedrukt geeft dit het volgende in bijvoorbeeld een analyse van een drie-dimensionale tabel {ABC}: als z een (partiële) observatie is die scoort op de variabelen A en B, met A = i, en B = j, dan is de kans dat z, de score k heeft op C als volgt bepaald:

$$Pr\,(C=k|A=i,B=j) = Pr(A=i,B=j,C=k)/Pr(A=i,B=j) \quad (2)$$

(Fuchs, 1982: 271).

Tabel 2 is een twee-variabelen voorbeeld van Little & Rubin (1987: 183-184) om dit duidelijker te maken. Wanneer we de ontbrekende gegevens op variabele Y_1 buiten beschouwing laten, hebben we een zogenaamd 'genest' patroon van volledig (op Y_1 en op Y_2) en partieel (alleen op Y_1) geclassificeerde data. We gaan uit van een 2 x 2 tabel waarbij Y_1 (j=1,2) de rijvariabele vormt en Y_2 (k=1,2) de kolomvariabele. Volledig geclassificeerd op de beide variabelen zijn 300 respondenten. Partieel geclassificeerd op Y_1 zijn 90 respondenten.

Tabel 2 Voorbeeld van partiële gegevens met algemeen patroon.

Y_2 / Y_1	1	2	TOT		Y_1	
1	100	50	150		1	30[a]
2	75	75	150		2	60[b]
TOT	175	125	300			
Y_2	1	2				
	28[c]	60[d]				

We noemen de volledig geobserveerde gegevens uit tabel 2, m_{jk}, de rijtotalen van de volledig geclassificeerde gegevens m_{j+}, de

equivalente kolomtotalen m_{+k}, de partieel geclassificeerde gegevens op Y_1, $r_{j.}$ en de som over alle volledig en partiële gegevens n. Little & Rubin geven aan dat de ML-schattingen voor de proporties in cel(jk) ($\hat{\pi}_{jk}$) bij partiële respons op Y_1 als volgt worden verkregen, bij een verzadigd model en de MAR-veronderstelling:

model:

$$\hat{\pi}_{jk} = \hat{\pi}_{j+} \cdot \hat{\pi}_{k|j}$$

met volgende ML-schatters

$$\hat{\pi}_{j+} = (m_{j+} + r_{j.})/n \qquad \hat{\pi}_{k|j} = m_{jk}/m_{j+}$$

We schatten dus de marginale proporties voor Y_1 en verdelen deze, binnen elke categorie van Y_1, in functie van de geobserveerde proportie van elke cel ten aanzien van de rijmarginaal zoals dit voor de volledig geclassificeerde eenheden wordt waargenomen. Anders gesteld: een proportie m_{jk}/m_{j+} van de partieel geclassificeerde observaties r_j wordt toegekend aan elke cel(jk).

In het voorbeeld wordt dit:
- voor de geschatte rijproporties:

$\hat{\pi}_{1+} = 180/390$ $\qquad$ $\hat{\pi}_{2+} = 210/390$

- voor de conditionele probabiliteiten bij m_{j+}:

$\hat{\pi}_{1|1} = 100/150$ $\qquad$ $\hat{\pi}_{2|1} = 50/150$

$\hat{\pi}_{1|2} = 75/150$ $\qquad$ $\hat{\pi}_{2|2} = 75/150$

In tabel 3 worden beide voorgaande gecombineerd tot ML-schattingen voor $\hat{\pi}_{jk}$ (genest patroon).

Tabel 3: Meest aannemelijkheidsschattingen $\hat{\pi}_{jk}$ voor de Little en Rubin gegevens volgens verschillende partiële data patronen en modellen (proporties) (1)

genest patroon Alg.patroon: verzadigd Alg.patroon: onafhankl.	Y_2 1	2	Tot.
Y_1	0,3077	0,1538	0,4615
	0,2795	0,1740	0,4535
1	0,2415	0,2201	0,4615
	0,2692	0,2692	0,5384
	0,2387	0,3078	0,5465
2	0,2817	0,2567	0,5385
	0,5769	0,4230	1,0000
	0,5182	0,4818	1,0000
Tot.	0,5232	0,4768	1,0000

(1): relevante populatieschattingen voor de verwachte volledige tabel,algemeen patroon, kunnen verkregen worden door de respectievelijke proporties te vermenigvuldigen met de totale steekproefomvang (volledig en partieel) van 478 eenheden. Voor genest patroon steekproefomvang vermenigvuldigen met 390 eenheden.

Deze directe en niet-iteratieve wijze om ML-schattingen voor de complete en partiële data te verkrijgen is mogelijk, omdat enerzijds de volledige en ontbrekende gegevens een 'genest' patroon vertonen en omdat anderzijds impliciet een verzadigd loglineair model werd gespecificeerd. Een 'genest' patroon is gedefinieerd als volgt (Hagenaars, 1986 : 70): "wanneer de variabelen zo kunnen worden geordend dat voor elke respondent geldt dat het hebben van een score op een bepaalde variabele impliceert dat ook voor alle voorafgaande variabelen de scores aanwezig zijn".

Wanneer we niet te doen hebben met een verzadigd model of een 'genest' gegevenspatroon, vereisen de ML-correctiemethodes (gebaseerd op loglinaire modellen) dat er gebruik wordt gemaakt van

een iteratieve procedure.

Fuchs (1982) heeft duidelijk gemaakt hoe ML-schattingen gevonden kunnen worden voor de 'missing sufficient statistics', die het mogelijk maken om de parameters van de gespecificeerde loglineaire modellen te schatten, als verondersteld wordt, dat aan de MAR voorwaarde voldaan is. Tevens geeft hij een toetsingsprocedure aan voor de aldus gepaste loglineaire modellen. Hiertoe gebruikt hij een variant van het zogenaamde EM-algoritme.

Little & Rubin (1987 : 129) omschrijven dit algoritme, in zijn algemeenheid beschouwd, als volgt: "The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) reestimate the missing values assuming the new parameter estimates are correct, (4) reestimate parameters, and so forth, iterating until convergence."

Het EM-algoritme vindt zijn naam in de twee stappen die het veronderstelt: een 'Expectation'- en een 'Maximalisation'-stap. De M-stap - de tweede stap in het algoritme - is eenvoudig te omschrijven en algemeen bekend van loglineaire modellen voor complete gegevens. Bij een bepaald gespecificeerd model dient de waarschijnlijkheid gemaximaliseerd te worden. Hiertoe kan het bestaande IPF- of Newton-Raphson-algoritme worden gehanteerd binnen de bestaande computersoftware.

De voorafgaande stap, de E-stap zorgt ervoor dat we de 'missing sufficient statistics' schatten. De toepassing hiervan naar loglinaire modellen toe houdt een uitbreiding in ten aanzien van de volledig gegevensanalyse situatie. Dit gebeurt als volgt: de te passen marginale verdelingen zoals gespecificeerd door het loglineair model dienen nu geschat te worden aan de hand van de volledig geclassificeerde gegevens plus de partieel geclassificeerde gegevens. Met andere woorden, in tegenstelling tot de situatie waar de gegevens volledig zijn en waarbij de te passen marginale verdelingen geobserveerd zijn, zullen we in de situatie waarbij partiële informatie meespeelt, moeten zien te werken met 'geschatte geobserveerde marginalen'. Een uitdrukking die aangeeft dat we de geobserveerde marginalen schatten, alsof er geen partiële non-respons zou geweest zijn, en de M(C)AR-assumptie geldt.

Als voorbeeld van de hier aangegeven werkwijze nemen we wederom een verzadigd model, maar nu voor tabel 2, waarbij we nu beide

partieel geobserveerde supplementaire tabellen in rekening brengen. Het voorbeeld is afkomstig van Little (1982) en wordt tevens besproken in Little & Rubin (1987). De meest aannemelijke schatters die gebruik maken van alle beschikbare gegevens kunnen verkregen worden door middel van het EM-algoritme. In de M-stap zijn er echter geen iteraties nodig, omdat we te doen hebben met een verzadigd model. We veronderstellen dus een verzadigd model met een MCAR mechanisme. Na convergentie van het EM-algorithme zullen meest waarschijnlijheids schatters verkregen hebben.

De toekenning van de non-respons (E-stap) aan de variabele waarop niet gescoord werd, heeft plaats op identieke wijze als bij het voorgaande voorbeeld. Dit betekent dat voor de eerste schatting gebruik wordt gemaakt van de conditionele verhoudingen in de volledig geclassificeerde tabel.

Van de 30 respondenten die scoren op Y_1 categorie 1, maar niet op Y_2 gaan er 30(100/150)= 20 naar Y_{11} en 30(50/150) = 10 naar Y_{12}. Van de 60 respondenten die scoren op Y_1 categorie 2, maar niet op Y_2 worden 60(75/150) = 30, toegewezen aan Y_{21} en Y_{22}. Van de 28 respondenten die scoren op Y_2, categorie 1, maar niet op Y_1 worden 28(100/175) = 16 toegewezen aan Y_{11} en 28(75/175) = 12 aan Y_{21}. Tenslotte worden van de 60 respondenten die scoren op Y_2, categorie 2, maar niet op Y_1, er 60(50/125) = 24 toegewezen aan Y_{12} en worden 60(75/125) = 36 toegewezen aan Y_{22}.

Gebruik makende van de in het geneste voorbeeld ontwikkelde notatie, maar die uitbreidende tot de twee partieel geclassificeerde supplementaire tabellen, verkrijgen we volgend model met meest aannemelijkheid schatters:

model:

$$\hat{\pi}_{jk} = m_{jk}/n + (r_{j.}/n \cdot m_{jk}/m_{j+}) + (q_{.k}/n \cdot m_{jk}/m_{+k})$$

waarbij n staat voor de optelsom van de volledig en partieel geclassificeerde gegevens, $r_{j.}$ voor de partieel geclassificeerde gegevens op Y_1 en $q_{.k}$ voor de partieel geclassificeerde gegevens op Y_2.

Dit geeft de volgende toewijzing aan elke cel, waarbij de letter in superscript aangeeft van welke partiële tabel en categorie het respectievelijke deel afkomstig is.

Stap 1:

		Y_2	
		1	2	totaal
Y_1	1	$100+20^a+16^c=136$	$50+10^a+24^d=84$	220
	2	$75+30^b+12^c=117$	$75+30^b+36^d=141$	258
totaal		253	225	478

Aangezien we met een verzadigd model te doen hebben is aan de M-stap eveneens voldaan en hebben we de verwachte waarden verkregen onder de assumptie van MAR. Omdat het patroon van de partiële data genest is, zijn deze schattingen maximale waarschijnlijkheids-schattingen van de celfrequenties. Deze niet-iteratieve wijze om maximale waarschijnlijkheidsschattingen te verkrijgen is mogelijk, gezien het genest zijn van de data, het verzadigd model en de MAR-assumptie, omdat de waarschijnlijkheidsfunctie kan opgedeeld worden en gemaximaliseerd worden, voor de afzonderlijke waar-schijnlijkheidsprodukten voor de volledig en partiëel geobserveer-de gegevens (Marini, Olsen, Rubin, 1980).

Als we echter MCAR veronderstellen kan op basis van de berekende verwachte waarden volgens stap 1 verder gegaan worden. De 30 res-pondenten die scoren op Y_1, categorie 1, maar niet op Y_2, worden nu als volgt toegekend aan respectievelijk Y_{11} en Y_{12}: 30(136/220) = 18,55 en 30(84/220) = 11,45. De andere partieel geclassificeerde gegevens worden op analoge wijze toegekend aan de respectievelijke cellen.

Vervolgens worden de in deze stap verkregen schattingen (aan de M-stap is ook voldaan) weer gebruikt om in een volgende stap een herschatting te maken. Afhankelijk van het convergentiecriterium - dit is de toegestane maximale afwijking in geschatte waarden tussen twee opeenvolgende iteraties - verkrijgt men convergentie na x-iteraties. Vanaf de vierde stap treden in dit voorbeeld geen noemenswaardige wijzigingen meer op. Indien we als convergentie-criterium 0,0001 stellen als maximale wijziging die mag optreden in de opeenvolgende schattingen voor de toewijzingen, verkrijgen we na 12 iteraties de populatieschattingen voor de verwachte volledige tabel, zoals die aan de hand van tabel 3 (algemeen pa-troon: verzadigd) kan berekend worden.

Fuchs (1982), Hagenaars (1985), Little & Rubin (1987) geven aan hoe deze werkwijze uitgebreid kan worden naar meerdere variabelen en niet-verzadigde loglineaire modellen. Algemeen gesteld komt dit op het volgende neer: "ML-estimation of (niet-verzadigde, ms) loglinear models involves distributing the partially classified counts into the full tabel using estimated conditional probabilities and the estimating of the classification probabilities from the filled-in table. The only difference (met het verzadigd model, ms) is that all probabilities are estimated subject to the constraints imposed by the loglinear model" (Little & Rubin, 1987 : 187). Dat wil zeggen, dat de relevante subtabellen in functie van het gespecificeerde loglineaire model in overweging moeten worden genomen bij de berekening van de conditionele probabiliteiten. De M-stap vereist in de meeste gevallen ook iteraties zoals dit het geval is bij een volledig geclassificeerde tabelanalyse.

Een eerste probleem met dit EM-algoritme doet zich voor, wanneer een lege cel voorkomt in de volledig geclassificeerde tabel of in een marginale tabel die correspondeert met een gespecificeerde term in een loglineair model, dat dienst doet als startwaarde voor het EM-algoritme. Zelfs indien de corresponderende cel een positieve frequentie heeft in een partiële, supplementaire tabel, zal het EM-algoritme geen allocatie toestaan aan de lege cel. Het gebruik van nulwaarden als startwaarden in bepaalde cellen, komt dan ook overeen met het plaatsen van structurele nullen op deze cellen. Een tweede probleem, nauw verwant met het voorgaande, doet zich voor wanneer in de volledig geclassificeerde tabel één of meerdere marginalen gelijk aan nul zijn, maar de corresponderende partiële, supplementaire tabellen niet leeg zijn. De verdeling van deze partiële non-respons over de lege cellen die aanleiding geven tot de lege marginale cellen van de volledige tabel, zal afhankelijk zijn van de startwaarden die gekozen worden. Er wordt dan geen unieke oplossing verkregen.

Daemen en Swyngedouw (1986), schreven het programma 'MISQUA' dat de allocatie van partiële non-respons mogelijk maakt voor R x K tabellen, waarbij verzadigde, onafhankelijkheids- en quasi-onafhankelijkheidsmodellen gepast kunnen worden en een M(C)AR-mechanisme verondersteld wordt.

We beschouwen de nulfrequenties die we vaststellen in de transitietabel als steekproefnullen (Swyngedouw, 1987). Vandaar ook dat

geopteerd werd voor de toepassing van een pseudo-Bayes-aanpassing met een uniforme apriori-probabiliteitsmatrix op de geobserveerde transitietabel (Bishop e.a. 1975; Swyngedouw 1989), zoals reeds kon worden gezien in de tabel met betrekking tot de verschuivingen 1981 - 1985 voor Limburg.

Als we de partieel geobserveerde gegevens mee modelleren via een bepaald loglineair model, dient ook de adequaatheid van het model getest te worden. Ook hiervoor ontwierp Fuchs (1982) een test. Als de verwachte waarden eenmaal berekend zijn met behulp van het model en het veronderstelde M(C)AR-mechanisme kunnen twee wegen worden bewandeld om de 'fit' van het model te bepalen. Beiden maken gebruik van de 'log-likelihood-ratio-chi-kwadraat'. De eerste testprocedure bestaat erin de partieel geclassificeerde gegevens te modelleren volgens een bepaald model en deze toe te voegen aan de geobserveerde volledig geclassificeerde gegevens. Vervolgens kan ditzelfde model gepast worden op de aldus verkregen tabel, alsof deze gegevens een volledig geobserveerde tabel zouden zijn. Deze methode werd voorgesteld door Hocking en Oxspring (1974). Ze wordt door Fuchs (1982 : 274) en Hagenaars (1986 : 66 & 72) afgewezen met als argument dat de toegewezen (partiële) gegevens de neiging zullen hebben het gespecificeerde model te bevestigen en dus de 'power' van de test doen afnemen. Verder is niet voldaan aan de voorwaarde dat voor alle variabelen onafhankelijke waarnemingen zijn verkregen (vanwege het bijschatten) en dat dus de chi-kwadraat verdeling onzeker wordt.

De tweede en betere manier om te toetsen gaat als volgt: men modelleert de volledig en partieel geclassificeerde gegevens volgens een bepaald gespecificeerd model en het veronderstelde M(C)AR-mechanisme. Dan worden de parameters $\hat{\pi}_{jk}$ geschat aan de hand van de geschatte volledige tabel. Door middel van deze verwachte waarden berekenen we de geschatte kansen (onder de conditie dat het model waar is), zowel voor de volledige tabel

$$\hat{\pi}_{jk} \cdot \Sigma_{jk} m_{jk}$$

als voor de marginale verdelingen

$$\hat{\pi}_{j+} \cdot \Sigma_j r_{j.} \quad (r_{j.} = \text{alleen } Y_1 \text{ bekend})$$

$$\hat{\pi}_{+k} \cdot \Sigma_k q_{.k} \quad (q_{.k} = \text{alleen } Y_2 \text{ bekend}).$$

Gegeven de geobserveerde frequenties en de verwachte frequenties kan men voor elke (sub)groep L^2 (minus twee maal de log-aannemelijkheid) berekenen en sommeren over alle (sub)groepen.

We werken het voorbeeld van Little en Rubin verder uit. De geschatte kansen onder het verzadigde loglineaire model binnen de volledige tabel, na modellering van de volledige en partiële gegevens worden weergegeven in tabel 3 (alg. patroon: verzadigd).

Merk op dat in dit MCAR model gesteld wordt dat $\Sigma_k \hat{\pi}_{jk}$ voor de verwachte volledige tabel gelijk is aan $\hat{\pi}_{j.}$ voor de partiële tabel. Hetzelfde geldt voor de andere marginale kansen.

Daar de steekproefgrootte 300 eenheden omvat in de volledig geclassificeerde geobserveerde tabel, verwachten we de waarden zoals aangegeven in tabel 4 onder de hoofding 'Exp. in sample, verzadigd'. Hiermee bedoelen we in navolging van Little en Rubin (1987) dat de nulhypothese wordt aangenomen dat het hier gespecificeerde model met een associatieterm tussen Y_1 en Y_2 de gegevens past.

Tabel 4: Verwachte waarden voor de volledige en de partiële tabellen onder de verschillende loglineaire modelspecificatie.

			Exp. in sample	Exp. in sample
		Obs.	verzadigd	onafhankelijk
Y_1	Y_2			
1	1	100	84	72
1	2	50	52	66
2	1	75	72	85
2	2	75	92	77
1	-	30	41	42
2	-	60	49	48
-	1	28	46	46
-	2	60	42	42

Zoals we reeds aangaven, heeft de allocatie van de partieel geclassificeerde gegevens hierbij plaats via de conditionele probabiliteiten $\hat{\pi}_{k|j}$ en $\hat{\pi}_{j|k}$ die geschat worden onder het gespecificeerde loglineaire model. Aangezien de groep die partieel geclassificeerd is op Y_1, uit 90 respondenten bestaat en de partieel geclassificeerde groep op Y_2 uit 88 eenheden, en de marginale geschatte kansen voor de categorieën van Y_1 en Y_2 bekend zijn (zie tabel 3), krijgen we hiervoor de eveneens in tabel 4 weergegeven verwachte waarden.

Bijvoorbeeld: de omvang van de partiële non-respons op variabele Y_1 is 90. Uit de marginale probabiliteiten van de geschatte volledige tabel blijkt dat 45,35% op categorie 1 zou moeten scoren, dus 90 x 0,4535 = 40,82 of afgerond 41.

De L^2-test voor het MCAR model waarbij een associatie tussen de beide variabelen verondersteld wordt, levert een waarde van 26,40.

Over de berekeningswijze - niet het aantal - van de vrijheidsgraden (df) dat deze toets kent, bestaat enige discussie. In tegenstelling tot Fuchs (1982) geven Hagenaars (1986 : 72) en Little & Rubin (1987 : 192) hetzelfde aan, maar op een andere wijze geformuleerd. Hagenaars formuleert het als volgt: "De vrijheidsgraden voor deze 'overall'-toets is gelijk aan: het aantal cellen in alle subgroepen min het aantal subgroepen min het aantal onafhankelijk te schatten parameters". In het voorbeeld betekent dit 8(cellen) - 3(subgroepen) - 3(parameters) = 2 df. Een andere berekeningswijze maar met hetzelfde resultaat: de volledige tabel heeft 3 vrijheidsgraden, de beide partiële tabellen elk 1; dit levert 5 vrijheidsgraden op en er moeten 3 parameters onafhankelijk geschat worden: d.i. 5 - 3 = 2 vrijheidsgraden. Dit levert een waarschijnlijkheid op die kleiner is dan 0,05. De test op het MCAR model is een test op de veronderstelling dat alle subtabellen uit dezelfde populatie afkomstig zijn, en dat het MCAR model binnen elke subtabel opgaat. Iets dat hier verworpen moet worden.

Een zelfde procédé kan nu gevolgd worden voor elk mogelijke meer restrictief model. Bij wijze van voorbeeld deden we dit nu ook voor het model dat veronderstelt det Y_1 en Y_2 onafhankelijk zijn. We gebruiken nogmaals de gegevens van Little & Rubin.

De meest aannemelijke schatting onder het onafhankelijkheidsmodel, berekend met zowel de partiële als de volledig geclassificeerde gegevens, en bij convergentie worden weergegeven in tabel 3 (alg.patroon: onafhankl.)

De verwachte waarden voor de volledige tabel en de partiële tabellen, bij gebruik van het onafhankelijkheidsmodel, zijn weergegeven in tabel 4.

De toets voor het niet-verzadigd model kan nu verkregen worden door 'min 2 maal de log-aannemelijkheids ratio' en het aantal vrijheidsgraden voor het niet-verzadigde model te berekenen en deze af te trekken van de verkregen resultaten voor het verzadigde model. Zo'n conditionele toets voor een onverzadigd model heeft hetzelfde aantal vrijheidsgraden als de toets voor hetzelfde niet-verzadigde model in een volledige gegevensanalyse. Hierover zijn alle auteurs het eens, en hieruit blijkt dan ook dat de formulering van Hagenaars en Little en Rubin, met betrekking tot de berekening van het aantal vrijheidsgraden, correct is.

Volgens Fuchs heeft deze werkwijze verschillende voordelen. Een

voordeel van deze conditionele test is, dat de 'power' van de test groter is dan bij dezelfde niet-conditionele test voor het gespecificeerde model. Eveneens is bij een conditionele test de benadering van de theoretische chi-kwadraat verdeling van L^2 beter. Een slechte theoretische chi-kwadraat benadering kan zich voordoen in de berekeningen van L^2 voor het verzadigde en niet-verzadigde model, vanwege de soms zeer kleine aantallen in de (sub)tabellen. Verder kan op deze wijze, dat wil zeggen door middel van conditionele toetsen, de significantie worden nagegaan van bepaalde effecten, als de twee geteste modellen 'genest' zijn (Fuchs, 1982 : 274).

Het aantal vrijheidsgraden voor het onafhankelijkheidsmodel bedraagt 3 en L^2 heeft een waarde van 35,95. De test op de adequaatheid van het model levert dus een L^2 op van 35,95 - 26,40 = 9,55 en heeft 3 - 2 = 1 vrijheidsgraad, wat een waarschijnlijkheid of overschrijdingskans heeft van 0,0001. Onafhankelijkheid van Y_1 en Y_2 moet dus verworpen worden.

De toets op het onafhankelijkheidsmodel, df = 3, L^2 = 35,95, is een test die aangeeft dat alle subtabellen uit dezelfde populatie komen *en* dat het onafhankelijkheidsmodel opgaat binnen deze populatie. Hieruit zou men kunnen concluderen dat de op basis van deze berekeningen uitgevoerde conditionele test op het model (verschil df en L^2 voor onafhankelijkheidsmodel en verzadigd model) een MCAR-mechanisme veronderstelt. Little en Rubin wijzen er echter op dat deze conditionele test geldig blijft, zelfs indien het non-responsmechanisme MAR is. Hagenaars is het hier mee eens en in een persoonlijke mededeling formuleert hij dit als volgt: "Op zich zijn de toetsen die hij (Fuchs) voorstelt MCAR, maar Little en Rubin wijzen erop dat de aannemelijkheidsfunctie te splitsen is in een deel dat refereert aan de eigenlijke modelassumptie en een deel dat te maken heeft met de aard van de non-respons (op voorwaarde dat die MAR is). Daardoor maakt het niet meer uit welk sterk of zwak 'missing at random'-mechanisme men veronderstelt, wanneer men een bepaald beperkt model vergelijkt met het verzadigde model: het non-respons deel van welke aard dan ook (mits 'missing at random') valt weg en wat je over houdt is een conditionele toets op de houdbaarheid van een bepaald spaarzaam model onder de assumptie dat het respons-mechanisme 'missing at random' is."

In de volgende paragraaf zullen we de resultaten van de analyse van de tabel met betrekking tot de verschuivingen 1981 - 1985 voor Limburg rapporteren.

4 Een beperkte toepassing met betrekking tot de verschuivingen 1981 -1985 voor de provincie Limburg

Een eerste vraag die we beantwoorden is of de non-respons als MCAR mag worden beschouwd. Zoals reeds werd uiteengezet, zou een positief antwoord hierop tot gevolg hebben, dat de analyses die werden uitgevoerd op de volledig geclassificeerde gegevens geldig, maar minder efficiënt zouden zijn. We passen hiertoe het volledig verzadigde model op de gegevens van de tabel, inclusief de partieel geclassificeerde gegevens. Ter illustratie geven we de verkregen 'geschatte volledige tabel' (die eveneens de 'verwachte waarden voor de volledige tabel is, gegeven het gepaste model') en de 'verwachte waarden voor de volledige en partiële tabellen'.

Tabel 5 'Geschatte volledige tabel' Limburg 1981 - 1985 onder verzadigd model

1985 1981	CVP	PVV	SP	VU	Agalev	Andere	Blan- co/ong.	TOT
CVP	185,105	11,124	15,328	12,253	5,121	4,847	1,933	235,771
PVV	3,289	87,693	9,727	3,371	0,136	0,128	0,194	104,538
SP	8,796	3,452	184,149	3,460	5,170	4,891	2,020	211,938
VU	7,514	5,523	11,857	82,433	1,355	1,284	0,193	110,159
Agalev	2,315	0,125	3,447	0,125	16,711	0,134	0,206	23,063
Andere	0,116	1,188	2,231	4,408	4,962	3,561	1,905	18,371
Blanco/ong.	2,209	0,119	0,117	0,119	0,134	0,127	5,337	8,162
TOT	209,344	109,224	226,856	106,169	33,589	14,972	11,848	712,002

Tabel 6 'Verwachte waarden voor de volledige en partiële tabellen' Limburg 81 - 85 onder verzadigd model

a) Volledige tabel

1985 1981	CVP	PVV	SP	VU	Agalev	Andere	Blanco/ong.	TOT
CVP	166,906	10,030	13,821	11,048	4,618	4,370	1,797	212,591
PVV	2,966	79,073	8,771	3,040	0,123	0,115	0,175	94,260
SP	7,931	3,113	166,044	3,120	4,662	4,410	1,821	191,101
VU	6,775	4,980	10,691	74,329	1,222	1,158	0,174	99,329
Agalev	2,087	0,113	3,108	0,113	15,068	0,121	0,186	20,796
Andere	0,105	1,071	2,012	3,975	4,474	3,211	1,718	16,565
Blanco/ong.	1,992	0,107	0,105	0,107	0,121	0,115	4,812	7,360
TOT	188,762	98,486	204,552	95,731	30,287	13,500	10,683	742,001

b) Partiële tabel rijvariabele (dat wil zeggen geclassificeerd op rijvariabele)

1981

CVP	PVV	SP	VU	Agalev	Andere	BL/O	TOT
5,298	2,349	4,763	2,475	0,518	0,413	0,183	16,000

c) Partiële tabel kolomvariabele (dat wil zeggen geclassificeerd op kolomvariabele)

1985

CVP	PVV	SP	VU	Agalev	Andere	BL/O	TOT
15,877	8,284	17,205	8,052	2,547	1,136	0,899	54,000

Het aantal vrijheidsgraden voor deze test kan als volgt bepaald worden. Naast de 63 cellen over alle (sub)tabellen heen, zijn er 3 (partiële) tabellen en schatten we 12 hoofdeffecten en 36 associatie-effecten. Dat wil zeggen, er resten 63 - 3 - 12 - 36 = 12 vrijheidsgraden. De 'log-likelihood-ratio-chi-kwadraat' bedraagt 22,307, wat een waarschijnlijkheid oplevert van ongeveer 4%. Gegeven de vele kleine waarden in de tabel kan men zich de vraag stellen of de asympthotische chi-kwadraat verdeling hier wel opgaat. Nochtans kwamen L^2 en de Pearson's chi-kwadraat relatief aardig overeen. Hoewel dit geen bewijs vormt, hebben we niets beter voor handen. We moeten dan ook besluiten dat de MCAR-hypo-

these verworpen dient te worden. Dit kan ons tot twee conclusies leiden. Enerzijds zou dit kunnen aangeven dat de partiële nonrespons andere marginalen heeft dan de volledige tabel en dat de nonrespons dus selectief is. Anderzijds, hoewel misschien ver gezocht, zou dit er op kunnen wijzen dat de de nonrespons wel MCAR is, maar heterogeen: de marginalen van de partiële nonrespons zijn gelijk aan deze van een subgroep uit de volledige tabel, bv. de 'movers' uit een 'movers -stayers' model. Hoe dan ook, analyses die uitsluitend gebaseerd zijn op de volledig geclassificeerde geobserveerde gegevens, zullen dan ook vertekend zijn.

In de volgende stap van de analyse schatten we de (quasi-) onafhankelijkheidsmodellen. Bij het EM-algoritme wordt de partiële non-respons dan toegekend, berekend op de proporties, zoals die verkregen worden onder de respectievelijke (quasi-)onafhankelijkheidsmodellen.

Bij de quasi-onafhankelijkheidsmodellen doen zich echter twee mogelijkheden voor betreffende de te schatten verwachte waarden voor de volledige en de partiële tabellen. Allereerst het gewone quasi-onafhankelijkheidsmodel of het zogenaamde 'movers-stayers'-model uit 1955 van Blumen, Kogan en McCarthy. (Upton 1978; Swyngedouw 1987). Dit is een latente klasse model, waarbij twee type kiezers onderscheiden worden. Eén type is de 'stayer'. Hij blijft op de twee tijdstippen in dezelfde responscategorie. Het andere type is de 'mover'. Hij verandert van responscategorie, van één punt in de tijd naar het andere en dit onafhankelijk van zijn vroegere responscategorie. Om dit model via loglineaire modellen te formuleren, plaatsen we structurele nullen op de diagonaal en passen een (quasi-)onafhankelijkheidsmodel. In het geval we uitsluitend te doen hebben met volledige gevens komt dit model volledig overeen met het 'movers-stayers'-model dat gepast kan worden via latente klasse analyse (zie o.a. Hagenaars 1985). Wanneer we in de M-stap van het EM-algoritme het quasi-onafhankelijkheidsmodel hebben gepast, kunnen we op basis van de verkregen loglineaire parameters van dit model, waarden gaan schatten voor de diagonaal. De geschatte waarden voor de off-diagonale cellen en de geschatte waarden voor de diagonale cellen worden in de E-stap aangewend om de toekenning van de partiële non-respons door te voeren bij het quasi-onafhankelijkheidsmodel. Het gevolg hiervan is, dat de nonrespons niet uitsluitend op de off-diagonale cellen terechtkomt,

maar dat ook een gedeelte wordt toegekend aan de diagonale cellen. Als convergentie in de berekeningen is opgetreden, bepalen we de 'fit' van het model zoals dit gebeurt bij het gewone quasi-onafhankelijkheidsmodel. Dit wil zeggen dat enkel de geschatte waarden voor de off-diagonale cellen en voor de partiële tabellen worden vergeleken door middel van de L^2 en dat er op geen enkele wijze rekening wordt gehouden met de uitgesloten diagonaal.

Een tweede wijze bestaat erin dat voor het modelleren van de non-respons hetzelfde quasi-onafhankelijkheidsmodel wordt gehanteerd, maar dat bij het toekennen van de non-respons, ook de diagonaal wordt uitgesloten. De non-respons wordt dan uitsluitend toegekend aan de off-diagonale cellen. Hiermee wordt afgeweken van het klassieke 'movers-stayers'-model.

Beide opties kunnen ook op inhoudelijke gronden verdedigd worden. Er wordt dan van uitgegaan dat de respondenten die partiële informatie verstrekken, uitsluitend behoren tot de groep van kiezers die behoren tot het type 'mover'. Het idee is dan dat de respondenten die hun stemgedrag niet veranderen, zich beter zullen kunnen herinneren voor welke partij ze in het verleden gestemd hebben. Voor de 'movers' daarentegen is het vaak moeilijker zich te herinneren voor welke partij ze in het verleden hebben gestemd en zij zullen dus sneller partiële informatie geven. Tevens kan bij de 'movers' ook de 'sociale wenselijkheid' een rol spelen. Veranderen van stemgedrag en de melding daarvan, betekent in zekere zin toegeven dat men zich in het verleden 'vergist' heeft. Deze melding zou dan als 'sociale onwenselijkheid' door de respondent-'mover' ervaren worden, daar hij niet graag zijn 'vergissing' te kennen geeft. Deze sociale wenselijkheidsvertekening zou aanleiding geven tot partiële non-respons.

In het eerste geval, waarbij de verdeling van de partiële non-respons ook over de diagonaal plaatsvindt, betekent dit, dat we het hierboven beschreven gedrag toekennen aan het type kiezers met de trek van 'latente movers'. In het tweede geval kennen we dit gedrag toe aan de kiezers die we de 'manifeste movers' noemen.

Het toekennen van de partiële non-respons aan enkel de 'manifeste movers' stelt ons echter voor het probleem dat het verondersteld 'ignorable' non-respons-mechanisme niet meer van toepassing is. Aangezien een gedeelte van de tabel (met name de diagonaal) expliciet uitgesloten is van toekenning van partiële

non-respons, bevinden we ons in de klasse van het 'nonignorable' non-respons-mechanisme. Het toekennen van de partiele non-respons aan de off-diagonale cellen zou nog volgens de MAR-assumptie kunnen verlopen, maar de kans dat de partiële non-respons aan de diagonaal wordt toegekend, wordt a priori op nul gezet. De toets van het model, zoals hierboven werd beschreven, gaat echter niet meer op, aangezien deze MAR veronderstelt. We zullen dit model dan ook niet kunnen passen.

Een overzicht van de resultaten die dit geeft, krijgen we in de tabellen 7 en 8. Deze resultaten werden verkregen door gebruik te maken van de computerprogramma's 'MISQUA' (Daemen en Swyngedouw, 1986) en 'PARTTAB' (Swyngedouw, 1987). De bepaling van de overschreidingskansen, gezien de L^2 en de respectievelijke vrijheidsgraden vond plaats aan de hand van de 'probchi: chi square distribution function' van SAS-IML (1985 : 147).

Tabel 7 'Goodness-of-fit'-statistieken voor de niet-verzadigde modellen, alsmede de test dat de volledig geobserveerde gegevens en de partieel-geobserveerde gegevens van dezelfde populatie afkomstig zijn

model	df	L^2	Prob.
verzadigd model	12	22,307	0,034
onafhankelijkheids-model	48	1045,175	0,000
Quasi-onafh. diagonaal (latente movers)	41	50,027	0,158

Het verschil tussen de betreffende modellen en het verzadigde model, geeft een test op de geldigheid van het model. Tabel 8 geeft deze resultaten

Tabel 8 'Goodness-of-fit'-statistieken voor niet-verzadigde modellen

model	df	L^2	Prob	AIC
onafhankelijkheids-model	36	1022,86	0,00	950,87
Quasi-onafh. diagonaal (latente movers)	29	27,720	0,53	-30,28

Een eerste conclusie die zich uit deze tabel opdringt, is dat het onafhankelijkheidsmodel niet opgaat. Dit was zelfs op het eerste gezicht reeds duidelijk. Een tweede conclusie bestaat erin dat het quasi-onafhankelijkheidsmodel (latente 'movers') een voldoende aanvaardbare 'fit' vertoont. Voor de analyse van de verschuivingsgegevens 1981-1985 voor Limburg, luidt de conclusie, dat de analyse van deze gegevens, met weglating van de partiële non-respons, een vertekende analyse moet zijn. Vervolgens kan er geconcludeerd worden, dat het quasi-onafhankelijkheidsmodel of het 'movers-stayers'-model, waarbij voor wat betreft de non-respons de assumptie van MAR wordt aangenomen, bij deze data past. De verwachte waarden voor de geschatte volledige tabel worden weergegeven in tabel 9.

Tabel 9 Verwachte waarden voor de geschatte volledige tabel onder quasi-onafhankelijkheid, 'movers-stayers'-model

1985 1981	CVP	PVV	SP	VU	Agalev	Andere	Blanco/ong.
CVP	17,2*	11,332	24,186	13,045	7,599	4,708	3,291
PVV	5,247	3,5*	7,360	3,970	2,313	1,433	1,002
SP	11,347	7,458	15,9*	8,585	5,001	3,099	2,166
VU	8,690	5,712	12,190	6,6*	3,830	2,373	1,659
Agalev	1,977	1,299	2,773	1,496	0,9*	0,540	0,377
Andere	3,708	2,437	5,202	2,806	1,634	1,0*	0,708
Blanco/ong.	0,749	0,492	1,050	0,567	0,330	0,204	0,1*

*: waarden verkregen aan de hand van de geschatte loglineaire parameters van het quasi-onafhankelijkheidsmodel; buiten de 'fit' van het model gelaten.

5 Besluit

We denken met deze studie van statistische analyse op nominaal niveau met partieel ontbrekende gegevens, aangeduid te hebben dat de praktijk waarin slechts gewerkt wordt met volledig-geobserveerde eenheden, tot vertekende resultaten kan leiden en dat zwakkere, meer aanvaardbare assumpties in verband met de non-respons gemaakt kunnen worden, met name MAR in plaats van MCAR, die statistisch handelbaar zijn. Het weinig of niet toegankelijk zijn van de noodzakelijke 'software' blijkt stilaan verholpen te worden. Zo kan verwezen worden naar het door Hagenaars ontwikkelde programma LCAG, dat in eerste instantie ontwikkeld werd voor het schatten van latente klassenmodellen via loglineaire modellen, maar dat ook geschikt is voor loglineaire modellen met ontbrekende gegevens, al dan niet binnen een latent klassenmodel (Hagenaars, 1985).

Gezien de MAR-assumptie is de uitbreiding van de bivariate situatie naar de multivariate geen probleem. De hiertoe benodigde algoritmen zijn een recht-toe-recht-aan uitbreiding van de hier beschrevene en kunnen, zoals reeds gezegd, gevonden worden bij Fuchs en Hagenaars. Problemen doen zich echter wel voor wanneer men uit de klasse van 'ignorable non-response' stapt en te maken krijgt met de 'nonignorable non-response'. Little (1982) gebruikt in een bivariate situatie een systeem waarbij aan de verschillende non-responscategorieën apriori-probabiliteiten worden toegekend. Het grote probleem hierbij is de basis waarop deze apriori-probabiliteiten worden bepaald. Little en Rubin (1987) behandelen de situatie van een bivariate tabel met één partiële tabel, Hagenaars (1988) behandelt een 'causal modelling'-systeem van Fay met betrekking tot een trivariate situatie. In deze gevallen maakt men gebruik van indicatorvariabelen voor het al dan niet aanwezig zijn van een observatie op een variabele. In beide gevallen blijkt de identificatie van verschillende modellen met een 'nonignorable' non-responsmechanisme een probleem.

Tot slot volgt een te overwegen stelling, onafgezien van de vooruitgang die geboekt wordt in verband met non-respons-correcties: "Finally, we should note that all methods of handling missing survey data must depend upon untestable assumptions. If the assumptions are seriously in error, the analyses may give mislea-

ding conclusions. The only secure safeguard against serious nonresponse bias in survey estimates is to keep the amount of missing data small (Kalton en Kasprzyk, 1986 : 14)."

Noten

1. Omdat we gebruik zullen maken van de loglineaire IPF-passingsmethode, is het noodzakelijk dat de geobserveerde tabel m_{jk} geen lege cellen bevat. Pseudo-Bayes-schattingsmethode is een soort van 'smoothing'-methode die de geobserveerde steekproefgrootte respecteert, doch aan de lege cellen een kleine waarde toekent. Zij is duidelijk te prefereren boven het ophogen van alle (of enkel de lege) cellen met een arbitraire kleine waarde zoals bv. 0.5. Zie Bishop, Fienberg en Holland, 1975, hfst. 12; Swyngedouw, 1989, hfst. 3.

ontvangen 1-11-1991
geaccepteerd 28-12-1992

Bibliografie

BILLIET, J., **Methoden van sociaal-wetenschappelijk onderzoek: ontwerpen en dataverzameling**, Acco, Leuven, 1990, 320 p.

BISHOP, Y.M.M, FIENBERG, S.E., HOLLAND, P.W., **Discrete Miltivariate Analysis. Theory and Practice**, MIT Press, Massachussets 1975, sixth printing 1980, 557 p.

CHEN, T., FIENBERG, S.E., Two-dimensional contingency tables with both completely and partially cross-classified data, in **Biometrics**, 1974, vol. 30, p. 629-642.

DAEMEN, L., SWYNGEDOUW, M., **Misqua**, A program to model partially classified data, K.U.Leuven, 1986.

FUCHS, Maximum likelihood estimation and model selection in contingency tables with missing data, in **Journal of the American Statistical Association**, 1982, vol. 77, no. 378, p. 270-278.

HAGENAARS, J.A.P., **Loglineaire analyse van herhaalde surveys. Panel-, trend-, en cohortonderzoek**. Katholieke Hogeschaal Tilburg, 1985, 414p.

HAGENAARS, J.A.P., Ontbrekende gegevens bij loglineaire analyse met en zonder latente variabelen. in Segers, J.H.G. en Bijnen, E.J., **Onderzoeken: reflecteren en meten**, Tilburg Univeristy Press, 1987, p. 57-77.

HAGENAARS, J.A.P., **Log-linear analysis with missing latent variables and missing data**, Paper presented at the International Conference on Social Science Methodology, Dubrovnik, 1988, 22 p.

HOCKING, R.R., OXSPRING, H.H., The analysis of partially categorized contingency data, in **Biometrics**, 1974, p. 469-483.

KALTON, G., KASPRZYK, D., The treatment of missing data, in **Survey Methodology**, 1986, vol. 12, no. 1, p. 1-16

LITTLE, R.J.A., Models for nonresponse in sample surveys, in **The Journal of the American Statistical Association**, 1982, vol. 77, no. 378, p. 237-250.

LITTLE, R.J.A., RUBIN, D.B., **Statistical analysis with missing data**, John Wiley and sons, New York, 1986, 278 p.

MADOW, W.G., NISSELSON, H., OLKIN, I. (red.), **Incomplete data in sample surveys, Volume 1, Report and case studies**, Academic Press, New York, 1983a.

MADOW, W.G., OLKIN, I. (red.) **Incomplete data in sample surveys, Volume 3, Proceeddings of the symposium**, Academic Press, New York, 1983b

MADOW, W.G., OLKIN, I. RUBIN, D.B. (red.) **Incomplete data in sample surveys, Volume 2, Theory and bibliographies** , Academic Press, New York, 1983c

MARINI, M., OLSEN, A., RUBIN, D., Maximum-likelihood estimation in panel studies with missing data, in **Sociological Methodology 1980**, Jossey-Bass, California, 1979, p. 314 - 357.

PANNEKOEK, J. **Applications of multiplicative models to problems in post-stratification**, Centraal Bureau voor de Statistiek, Voorburg, BPA no.: 13456-84-M1, 1984, 15 p.

RUBIN, D.B., **Inference and missing data**, in Biometrica, vol. 63, 1976, p. 581-592.

RUBIN, D.B., Basic Ideas of Multiple Imputation for Nonresponse, **Survey Methodology**, 1986, vol. 12, no. 1, p. 37 - 48.

S.A.S., **IML User's Guide**, Version 6 Edition, SAS Institute Inc. Cary, North Carolina, 1986.

STUDIECENTRUM WILLY CLAES, **Hoe stemden de Limburgers op 13.10.1985**, Hasselt, 1985, 70 p.

SWYNGEDOUW, M., **Pseudo**, A program for pseudo-Bayes estimations for R x N contingency tables., K.U.Leuven, 1985.

SWYNGEDOUW, M., **Parttab**, A program tot test models with M(C)AR classified data, Erasmus University, Rotterdam, 1987.

SWYNGEDOUW, M., A pilot study of Portuguese electorale shifts: 1976-1982, in **Quality and Quantity**, 21: 153-175, 1987

SWYNGEDOUW, M., A pilot study of Portuguese electoral shift 1976 - 1982, in **Quality and Quantity**, 21, 1987, p. 153 - 175.

SWYNGEDOUW, M., **De keuze van de kiezer. Naar een verbetering van de schattingen van verschuivingen en partijvoorkeur bij opeenvolgende verkiezingen en peilingen**, S.O.I./B.M.G., Leuven/Rotterdam, 1989, 333 p.

UPTON, G., **The analysis of cross-tabulated data**, Wiley and Sons, New York, 1978, 148 p.