Multi-niveau latente klasse analyse: met een toepassing bij het simultaan clusteren van landen en consumenten

Jeroen K. Vermunt, Universiteit van Tilburg, vakgroep Methodologie en Statistiek, Faculteit der Sociale Wetenschappen, Postbus 90153, 5000 LE Tilburg. Telefoon: 013 - 4662544; Fax: 013 - 4663002. E-mail adres: j.k.vermunt@uvt.nl

Tammo H. A. Bijmolt, Rijksuniversiteit Groningen, vakgroep Marketing, Faculteit der Economische Wetenschappen, Postbus 800, 9700 AV Groningen. Telefoon: 050-363 7065 / 3686; Fax: 050-363 3720. E-mail adres: t.h.a.bijmolt@eco.rug.nl

Leo J. Paas, Vrije Universiteit, vakgroep Marketing, Faculteit der Economische Wetenschappen en Bedrijfskunde, De Boelelaan 1105, 1081HV, Amsterdam. Telefoon: 020 - 598 6185; Fax: 020 - 598 6005. E-mail adres: lpaas@feweb.vu.nl (tevens de redacteur voor dit artikel)

SAMENVATTING: In dit artikel bespreken we latente klasse analyse, een model gebaseerde clustertechniek. Speciale aandacht wordt besteed aan een recent voorgestelde uitbreiding van latente klasse analyse die gebruikt kan worden voor de analyse van multi-niveau gegevens, zoals gegevens van respondenten die gegroepeerd kunnen worden in postcodes, van klanten gegroepeerd in filialen of van consumenten gegroepeerd in landen. Na een introductie waarin we latente klasse analyse bespreken en vergelijken met andere clusteranalyse methoden, behandelen we het multi-niveau latente klasse model. Vervolgens illustreren we dit multi-niveau model met een toepassing waarin consumenten worden geclusterd op basis van hun portfolio aan financiële producten en landen worden geclusterd op basis van de verdeling van consumenten over de verschillende segmenten. We bespreken zowel de marketingimplicaties van de bevindingen als de bredere toepassing van het multi-niveau latent klasse model voor marktonderzoek.

ABSTRACT: In this paper we describe latent class analysis, a model-based clustering technique. Special attention is given to a recently introduced extension of latent class analysis for the analysis of multilevel data, such as data in which respondents are nested within postal codes, clients nested within branch offices, or consumers nested within countries. After an introduction in which we describe latent class analysis and compare it with other clustering methods, we describe in more detail the multilevel latent class model. We illustrate the method with an application in which consumers are clustered on the basis of their financial product portfolios and countries are clustered based on the distribution of consumers over segments. We discuss the marketing implications of our findings, as well as the broader applicability of the multilevel latent class model for marketing research.

Jeroen K. Vermunt is als Hoogleraar verbonden aan het departement Methoden en Technieken van Onderzoek (Faculteit Sociale Wetenschappen) van de Universiteit van Tilburg. Zijn onderzoeksinteresses betreffen methodologische onderwerpen bij sociaal en gedragswetenschappelijk onderzoek. Hij ontwikkelde samen met Jay Magidson de Latent GOLD and Latent GOLD Choice programma's voor latent klasse analyse. Zijn werk verscheen in onder andere in *International Journal of Research in Marketing, Canadian Journal of Marketing Research, Quirk's Marketing Research Review, Sociological Methodology, Applied Psychological Meaurement* en *Psychometrika.* E-mail adres: j.k.vermunt@uvt.nl

Tammo H.A. Bijmolt is Hoogleraar Marketing-onderzoek onderzoek bij het Departement Marketing van de Economische Faculteit, Rijksuniversiteit Groningen. Hij promoveerde cum laude in 1996 bij dit departement en werkte tot 2004 bij de Universiteit van Tilburg. Tammo Bijmolt is voorzitter van de NIMA/MOA-examencommissie Marketing-onderzoek en Informatiemanagement (MIM). Hij heeft gedoceerd aan verschillende doelgroepen als doctoraal studenten, promovendi en post-doctoraal managers over onderwerpen als CRM, marktonderzoek en segmentatie. Hij heeft een groot aantal "research-based consultancy" projecten uitgevoerd voor onder andere: Bavaria, Brova, Carlson Marketing Group, GfK en Unilever. Onderzoeksmethodologie, CRM en loyaliteitsprogramma's vormen een belangrijk deel van zijn onderzoeksportefeuille. Tammo Bijmolt heeft over deze en vergelijkbare onderzoeksprojecten een groot aantal lezingen gegeven bij (inter-)nationale congressen en publiceerde hierover in internationale toptijdschriften zoals: *Journal of Marketing Research, Journal of Consumer Research, International Journal of Research in Marketing* en *Journal of Classification*; en daarnaast regelmatig in Nederlandse vaktijdschriften zoals het *Jaarboek van de MOA* en *Tijdschrift voor Marketing*. E-mail adres: t.h.a.bijmolt@eco.rug.nl

Leo J. Paas is als Universitair Docent verbonden aan het departement Marketing (Faculteit der Economische Wetenschappen en Bedrijfskunde) van de Vrije Universiteit in Amsterdam. Veel van zijn onderzoek gaat over de afname van producten en diensten door consumenten in de financiële sector. Andere onderzoeksinteresses betreffen consumentengedrag in de kunstsector, onderzoek en methodologie betreffende meetschalen in marketing en ontwikkelingen in de marketingliteratuur. Zijn werk is o.a. verschenen in *Journal of Economic Psychology*, *International Journal of Research in Marketing* en *The Service Industries Journal*. E-mail adres: lpaas@feweb.vu.nl

Multi-niveau latente klasse analyse: met een toepassing bij het simultaan clusteren van landen en consumenten

Jeroen K. Vermunt

Tammo H.A. Bijmolt

Leo J. Paas

1. Inleiding

Consumenten verschillen sterk in preferenties voor producten en diensten en in reacties op reclames en veel andere voor marketing relevante variabelen. Eén van de doelstellingen bij marktonderzoek is dan ook het beschrijven van de heterogeniteit in de markt op basis van waargenomen of uitgesproken preferenties van consumenten (revealed versus stated preferences). Bij waargenomen preferenties gaat het bijvoorbeeld om informatie over het bezit of de aankoop van bepaalde producten, terwijl het bij uitgesproken preferenties meestal gaat om attitudes of gedragsintenties.

Afhankelijk van de toepassing zal men de heterogeniteit in preferenties beschouwen als een continu of juist als een discreet verschijnsel. In het eerste geval zal men gebruik maken van, bijvoorbeeld, factoranalyse of random-effects (hierarchical Bayes) regressietechnieken. Opsporen van discrete categorieën consumenten met verschillende preferenties (marktsegmenten) geschiedt over het algemeen met behulp van clusteranalytische technieken. In de academische marketing wereld is een debat gaande over welke van deze twee benadering de beste is, waarbij de kwaliteit van de voorspelling van (toekomstig) individueel gedrag als voornaamste vergelijkingscriterium wordt gehanteerd. Sommige auteurs vinden dat modellen met continue heterogeniteit het beter doen, terwijl anderen aantonen dat modellen met discrete heterogeniteit even goed in staat zijn om individueel gedrag te voorspellen.

In dit artikel vinden we segmenten met behulp van clusteranalyse technieken (Wedel en Kamakura, 2000). De specifieke techniek die wij gebruiken is het latente klasse model, ook wel aangeduid als mixture model. In tegenstelling tot andere clusteranalyse methoden betreft het hier een op een statistisch model gebaseerde clustertechniek. Vandaar dat ook de term “model-based clustering” wordt gebruikt (McLachlan en Peel, 2000). Om precies te zijn: we veronderstellen dat de geobserveerde data tot stand zijn gekomen via een bepaald proces dat kan worden gemodelleerd. Het simpelste voorbeeld van een dergelijk proces is dat de populatie bestaat uit *S* segmenten met verschillende preferenties. De indeling in segmenten leidt tot verschillen in de kans om, bijvoorbeeld, een bepaald product te bezitten.

Doel van dit artikel is inzicht te geven in het gebruik van latente klasse analyse voor marktsegmentatie. Hierbij gaan we vooral in op een veelbelovende nieuwe variant, het multi-niveau latente klasse model. Na de inleiding vergelijken we methoden voor clusteranalyse en latente klasse analyse. Daarna bespreken we standaard latente klasse analyse en enkele voor marktonderzoek relevante varianten. Vervolgens besteden we aandacht aan latente klasse modellen voor multi-niveau data. We illustreren het multi-niveau latente klasse model met een toepassing waarin consumenten worden geclusterd op basis van hun portfolio aan financiële producten en landen worden geclusterd op basis van de verdeling van consumenten over de verschillende segmenten. De marketingimplicaties van de bevindingen van deze analyse worden vervolgens besproken en het artikel wordt afgesloten met een korte discussie over bredere toepassingen van het multi-niveau latent klasse model voor marktonderzoek.

2. Vergelijking tussen verschillende methoden voor clusteranalyse

Latente klasse analyse, clusteranalyse op basis van een statistisch model, verschilt sterk van andere clustermethoden, zoals hiërarchische en niet-hiërarchische clustertechnieken (Wedel en Kamakura, 2000). Bij hiërarchische clustertechnieken wordt eerst een tabel met paarsgewijze afstanden

(similariteiten) berekend tussen respondenten en vervolgens worden clusters gevormd door, via een bepaald algoritme, de respondenten (en clusters) die op elkaar lijken samen te voegen. Hoewel deze methoden een aantal ad hoc elementen bevatten (welke afstandsmaat, welke procedure van samenvoegen), werken ze over het algemeen goed bij kleine steekproeven.

Latente klasse analyse maakt geen gebruik van een dergelijke similariteitsmatrix en is evenals niet-hiërarchische clustertechnieken zoals K-means juist meer geschikt voor segmentatie op basis van grotere steekproeven. K-means clustering en latente klasse analyse zijn sterk verwante technieken. Hoewel de eerste niet is gebaseerd op een expliciet populatiemodel, is het wel degelijk mogelijk om de impliciete aannames te vertalen in een zeer restrictief model-based clustering model (Magidson en Vermunt, 2001). Een belangrijk verschil is dat K-means een harde partitionering geeft terwijl latent klasse analyse leidt tot een fuzzy partitioning. Dat wil zeggen, bij clusteranalyse moet er worden bepaald tot welk segment elk individu behoort, terwijl bij latente klasse analyse voor elk individu een segmentspecifieke lidmaatschapskans wordt bepaald. Ambroise en Govaert (2000) tonen echter aan dat het heel gemakkelijk is om een variant van latente klasse analyse te construeren waarin de mate van fuzzyness kan worden gemanipuleerd. Als de fuzzyness parameter op 0 wordt gezet, verkrijgt men een harde partitioning. De fuzzyness parameter kan echter niet alleen gebruikt worden ter verkrijging van minder fuzzy of zelfs een harde partitioning, maar ook van juist een meer fuzzy partitioning.

Ondanks de overeenkomsten tussen K-means clusteranalyse en latente klasse analyse, merken we op dat model gebaseerde clustering als belangrijke voordeel heeft dat men een statistisch model gemakkelijk kan aanpassen aan het type data dat men wil gebruiken. Zo kan men in plaats van een standaard latente klasse model een wat ingewikkelder model specificeren dat, bijvoorbeeld, geschikt is voor de analyse van data, die zijn verkregen via een keuze experiment (choice-based conjoint studies). Ook zijn allerlei verfijningen mogelijk zoals clustering in meerdere dimensies, modellen voor longitudinale data en de in dit artikel gepresenteerde modellen voor multi-niveau data. Bovendien kan er relatief gemakkelijk worden omgegaan met variabelen van verschillende meetniveaus. Nadeel van

een statistisch model is dat modelaannames noodzakelijk zijn, zoals, bijvoorbeeld, de aanname van locale onafhankelijkheid. Over het algemeen kunnen deze echter worden getoetst en deels worden losgelaten. Bovendien maken ook andere methoden aannames, al zijn die vaak impliciet en niet toetsbaar.

3. Latente klasse analyse

Er zijn veel varianten van latente klasse analyse en mixture modellen. Hieronder zullen we een aantal daarvan kort bespreken. We beginnen met het standaard ongerestricteerde model voor dichotome of polytome de responsvariabelen - de gemeten variabelen in de latent klasse analyse. Daarna laten we zien hoe het model wordt aangepast voor de analyse van antwoorden verkregen via een keuze experiment. Tot slot bespreken we kort modellen voor continue responsvariabelen.

Startpunt is dat we voor iedere persoon de waarden kennen op een set responsvariabelen. De respons van individu *i* op variabele *t* duiden we aan met y_{it} en het aantal variabelen als *T*. De volledige responsvector voor individu *i* wordt aangeduid als $\mathbf{y}_i$. Een bepaalde latente klasse (segment) zal worden aangeduid als *s* en het totaal aantal klassen als *S*, waarbij $1 \leq s \leq S$. Het basis idee van latente klasse modellen is dat de kansverdeling van $\mathbf{y}_i$ – $P(\mathbf{y}_i)$ – een gewogen soms is van *S* klasse specifieke kansverdelingen – $P(\mathbf{y}_i|s)$ (McLachlan and Peel, 2000; Vermunt en Magidson, 2002). Weging is gebaseerd op de proportie individuen in elke latente klasse $P(s)$. Dat wil zeggen dat:

$$P(\mathbf{y}_i) = \sum_{s=1}^{S} P(s)P(\mathbf{y}_i \mid s) \qquad (1)$$

In het standaard latent klasse model wordt dit basis idee gecombineerd met de zogenaamde aanname van locale onafhankelijkheid (Goodman, 1974, Magidson en Vermunt, 2004). The *T* geobserveerde variabelen worden verondersteld onafhankelijk van elkaar te zijn binnen klassen, hetgeen als volgt wordt weergegeven:

$$P(\mathbf{y}_i \mid s) = \prod_{t=1}^{T} P(y_{it} \mid s) \qquad (2)$$

De geschatte klasse-specifieke responskansen $P(y_{it}|s)$, voor de T indicatoren, geven aan hoe klassen van elkaar verschillen en kunnen worden gebruikt om de latente klassen te labelen. Wanneer we de twee basisvergelijkingen (1) en (2) combineren krijgen we het volgende model voor $P(\mathbf{y}_i)$:

$$P(\mathbf{y}_i) = \sum_{s=1}^{S} P(s) \prod_{t=1}^{T} P(y_{it} \mid s) \qquad (3)$$

De onbekende coëfficiënten in het model zijn de modelkansen $P(s)$ en $P(y_{it}|s)$. Deze worden over het algemeen geschat via de maximum likelihood methode. Bij modelselectie maakt men meestal gebruik van likelihood-ratio toetsen en informatiecriteria zoals AIC, BIC en CAIC. Zoals bij elke vorm van clusteranalyse betreft het belangrijkste modelselectie vraagstuk het aantal klassen of clusters dat men nodig heeft om de data in voldoende mate te beschrijven.

Wanneer latente klasse analyse gebruikt wordt als een clusteranalyse techniek zal men niet enkel de onbekende modelkansen willen schatten maar tevens personen willen toewijzen aan klassen. De kans dat persoon *i*, gegeven haar/zijn waardes op de responsvariabelen, behoort tot klasse *s*, $P(s|\mathbf{y}_i)$, verkrijgen we met behulp van de Bayes formule. De meest gebruikte classificatieregel is modale toewijzing. Dit is toewijzen aan het segment waarvoor $P(s|\mathbf{y}_i)$ het grootste is. Het zal duidelijk zijn dat in een goede clusteroplossing, een oplossing waarin men clusters goed van elkaar kan onderscheiden, de grootste kans voor de meeste respondenten dicht bij 1 ligt.

Latente klasse analyse is voor veel marketing doeleinden gebruikt. Bijvoorbeeld, het zogenaamde latente klasse keuzemodel wordt gebruikt voor de analyse van data verkregen via keuze experimenten. In keuze experimenten heeft elk van de responscategorieën (elk van de producten waar men uit kiest) bepaalde kenmerken (een bepaalde prijs, formaat, kwaliteit, kleur, etc.). Verondersteld wordt dat de kans dat een respondent een bepaald product kiest afhangt van de utiliteit van deze kenmerken: naarmate de utiliteit van een product in vergelijking tot die van de andere producten in de

keuzeset hoger is, is de kans dat men dat product kiest groter. Deze utiliteiten kunnen worden geschat met behulp van een zogenaamd discrete-keuzemodel of conditional logit model dat is ontwikkeld door Nobelprijswinnaar McFadden (1974). Typisch voor de latente klasse analyse variant van dit type model is dat verondersteld wordt dat de populatie heterogeen is wat betreft de utiliteiten van productkenmerken: iedere klasse of segment heeft haar eigen set utiliteiten (Kamakura, Wedel en Agrawal, 1994). Een latente klasse choice-based conjoint model wordt verkregen door restricties op te leggen aan de responskansen $P(y_{it}|s)$, met behulp van een conditional logit model:

$$P(y_{it} = m \mid s) = \frac{\exp(\sum_{j=1}^{J} \beta_{js} z_{mtj})}{\sum_{m'=1}^{M_t} \exp(\sum_{j=1}^{J} \beta_{js} z_{m'tj})}. \qquad (4)$$

Hier is *m* een bepaalde responscategorie, *j* een bepaald productkenmerk, z_{mtj} de waarde van product *m* in keuzeset *t* voor kenmerk *j* en β_{js} de utiliteit van kenmerk *j* voor latente klasse *s*.

Behalve met discrete responsvariabelen kan latente klasse analyse ook worden gebruikt met continue variabelen, en zelfs met combinaties van discrete en continue responsvariabelen. In het geval van een set continue variabelen zal de kans $P(\mathbf{y}_i \mid s)$ in vergelijking (1) de vorm hebben van een multivariaat normale verdeling (McLachlan en Peel, 2000; Vermunt en Magidson, 2002, 2004). In het meest algemene model zal dit een volledig ongerestricteerde verdeling betreffen, wat inhoudt dat gemiddelden, varianties en covarianties vrij geschat worden en ongelijk zijn tussen klassen. Een voorbeeld van een meer gerestricteerde variant is een model waarin de gemiddelden verschillen tussen klassen, maar de klassen dezelfde variantie-covariantie matrix hebben. Dit is een specificatie die ook wordt gebruikt in lineaire discriminant analyse. Een andere mogelijke restrictie is om, zoals in het standaard latente klasse model, te veronderstellen dat de variabelen onafhankelijk zijn binnen klassen, hetgeen bij continue indicatoren inhoudt dat hun covarianties op nul worden gezet. Een zeer gerestricteerde covariantie structuur die zeer lijkt op wat impliciet wordt verondersteld bij K-means

clusteranalyse wordt verkregen door de covarianties op nul te zetten en de varianties gelijk te maken tussen klassen en tussen indicatoren (Magidson en Vermunt, 2001, 2004).

Andere voorbeelden van relevante uitbreidingen model zijn modellen met meerdere latente variabelen en modellen waarin klasse lidmaatschap wordt voorspeld met behulp van covariaten (zie Magidson en Vermunt, 2004 voor een overzicht). Hieronder bespreken we een relatief recente uitbreiding van het latente klasse model, die veelbelovend is voor marketingtoepassingen, namelijk het multi-niveau latente klasse model.

4. Latent klasse modellen voor multi-niveau data

Multi-niveau latente klasse modellen worden gebruikt in situaties waarin het onjuist is om te veronderstellen dat waarnemingen onafhankelijk van elkaar zijn. Voorbeelden zijn data verkregen via clustersteekproeven, data verkregen via huishoudpanels, en allerlei vormen van geneste data, zoals van consumenten behorende tot bepaalde regio's, postcodes of filialen van een supermarktketen.

Er zijn op zijn minst drie manieren om dit type afhankelijke waarnemingen te behandelen binnen de context van latente klasse modellen. Dat zijn:

1. het gebruik van pseudo maximum likelihood schattingen met zogenaamde "survey design corrected" standaardfouten,
2. het gebruik van normaal verdeelde random effecten,
3. het gebruik van niet-parametrisch random effecten, hetgeen neerkomt op het clusteren van hogere niveau eenheden, ofwel de multi-niveau variant van het latente klasse model dat we hier zullen gebruiken.

De eerste wijze van aanpak werd voorgesteld door Wedel, Ter Hofstede en Steenkamp (1998) en Vermunt (2001). Afhankelijkheid tussen waarnemingen wordt daarin als een statistisch probleem gezien en dat probleem wordt opgelost door zogenaamde "survey corrected standaardfouten" voor de

modelparameters te berekenen. Deze zullen over het algemeen wat groter zijn dan de niet gecorrigeerde standaardfouten.

De tweede aanpak is sterk verwant aan hoe een normale multi-niveau regressieanalyse wordt uitgevoerd. Daarin wordt verondersteld dat de intercept (de constante) en mogelijk ook één of meerdere regressiecoëfficiënten variëren tussen groepen (Snijders en Bosker, 1999). Dit wordt gemodelleerd door aan te nemen dat deze parameters uit een multivariate normale verdeling komen. Vandaar dat in deze context ook wel de term random effect of random coëfficiënt wordt gebruikt. Vermunt (2003) stelde voor om hetzelfde principe toe te toepassen in de context van het latente klasse model. Dit kan door de omvang van de klassen te laten variëren tussen groepen. Daarbij wordt aangenomen dat (na een logistische transformatie) de klassenproporties normaal verdeeld zijn tussen groepen. In feite wordt een random-effects multinomiaal logistisch regressiemodel (Hedeker, 2003) geschat voor de klassenproporties.

In dit artikel kiezen we voor de derde benadering, die overigens sterk verwant is aan de tweede. Heterogeniteit in de klassenproporties tussen groepen wordt niet gemodelleerd als een continu verschijnsel, maar als een discreet fenomeen. Er wordt aangenomen dat groepen geclusterd kunnen worden in een beperkt aantal clusters (latente klassen), op basis van het vóórkomen van de verschillende latente klassen van respondenten in elke groep (Vermunt, 2003). Voordeel van deze aanpak is dat er inzicht wordt gegeven in de verbanden op verschillende niveaus. Als er, bijvoorbeeld, individuen in verschillende landen zijn geïnterviewd, dan levert het multi-niveau latente klasse model inzicht in de gelijkenis tussen individuen onderling en tussen de verschillende landen in de dataset. Dit soort informatie kan zeer relevant zijn voor marketing. Niet alleen wordt duidelijk welke klantsegmenten er per land zijn, maar ook worden landen op relevante kenmerken geclusterd. Stel dat de situatie in Figuur 1 opgaat met betrekking tot een segmentatie op productbezit. Er zijn vijf consumentsegmenten en in elk van de vijf segmenten vinden we consumenten met een bepaald productbezit patroon. Het voorkomen van deze segmenten, met bepaald productbezit patronen, kan per

land verschillen. In Nederland vinden we, bijvoorbeeld, dat slechts 5% van de consumenten het productbezit patroon heeft dat segment 1 typeert. Verder is in dit land 25% van de consumenten ingedeeld in segment 2 en maar liefst 55% in segment 3. In Duitsland is een soortgelijke segmentatiestructuur gevonden. In een multi-niveau latente klasse analyse worden Nederland en Duitsland daarom waarschijnlijk in hetzelfde landencluster ingedeeld. In Spanje en Portugal is er echter sprake van een sterk afwijkende verdeling over segmenten: er zijn veel consumenten met productbezit zoals in segment 4. Deze landen zullen daarom waarschijnlijk zijn in te delen in een ander landencluster dan Nederland en Duitsland, maar wel bij elkaar in hetzelfde landsegment. Een bedrijf kan op basis van deze clustering selecteren welke landen de juiste waarden hebben op relevante kenmerken voor, bijvoorbeeld, de introductie van een nieuw product.

[Voeg Figuur 1 hier in]

Hoewel de meer technische aspecten van dit model – zoals het schatten van model parameters via maximum likelihood – niet eenvoudig zijn, is het conceptueel wel een relatief eenvoudig model. De volledige responsvector voor case *i* uit groep *j* is nu aangeduid als $\mathbf{y}_{ij}$. Het enige dat verder verandert ten opzichte van het basismodel beschreven in vergelijking (1) is dat de klassengrootten – P(s) – nu afhangen van het groepssegment, *r*, waartoe groep *j* behoort. Voor de meer wiskundige lezer betekent dat:

$$P(\mathbf{y}_{ij} \mid r) = \sum_{s=1}^{S} P(s \mid r) P(\mathbf{y}_{ij} \mid s) \qquad (5)$$

Verdere details over hoe dit model kan worden geschat en uitgebreid zijn te vinden in Vermunt (2003). Het multi-niveau latente klasse model is geïmplementeerd in het softwarepakket Latent GOLD 4.0 (Vermunt en Magidson, 2005), een pakket dat is toegespitst op de toepassing van een breed scala aan latente klasse modellen en dat draait onder MS-Windows.

5. Een toepassing bij het simultaan segmenteren van landen en consumenten

De toepassing die we willen gebruiken om de multi-niveau extensie van het latente klasse model te illustreren betreft een onderzoek naar het bezit van financiële producten in 15 Europese landen. De dataset die we gebruiken betreft de Eurobarometer 56.0, verzamelt tussen 22 augustus en 27 september 2001 door een consortium van marktonderzoekbureaus in opdracht van de Europese Commissie, Directoraat-generaal Media en Communicatie, Opinie Enquêtes (Christensen, 2001). De Eurobarometer enquête omvat de populatie (15 jaar en ouder) van de 15 EU lidstaten in 2001. Er zijn 17 steekproef zones: Duitsland is verdeeld in Oost en West, en het Verenigd Koninkrijk in Groot-Brittannië en Noord Ierland. Met de term "land" duiden we deze steekproef zones aan.

De steekproefomvang is ongeveer 1000 per land, met uitzondering van Luxemburg (600) en Noord Ierland (300). De totale steekproefomvang is 16,200. We gebruiken gewichten uit de Eurobarometer database zodat elke nationale steekproef representatief is met betrekking tot demografische variabelen en er wordt gecorrigeerd voor de verschillende steekproef ratio's tussen landen (zie Tabel 1).

[Voeg Tabel 1 hier in]

De informatie over het bezit van financiële producten leiden we af uit vragen 25 tot met 28 van de Eurobarometer 56.0 survey (Christensen, 2001). Het betreft het bezit van een lopende rekening, een spaarrekening, een creditkaart, een andere bankpas, een chequeboek, rood staan, een hypotheek, en andere leningen. Het gaat om het bezit van individuele consumenten. Tabel 1 geeft beschrijvende informatie met trekking tot het bezit van deze 8 producten in de 17 landen.

De theoretische motivatie van de hier gepresenteerde analyse wordt gedetailleerd beschreven in Bijmolt, Paas en Vermunt (2004). Hoewel we het voorbeeld hier gebruiken om het multi-niveau latente klasse model te illustreren, is het van belang om de twee elementen van de theoretische

motivatie te benadrukken. Ten eerste past de studie in de traditie van onderzoek naar acquisitiepatronen (Kamakura, Ramaswami en Srivastava, 1991; Paas, 2003; Paas en Molenaar, 2005). De vraag is hierbij of producten in een bepaalde volgorde worden aangeschaft, ofwel of er consumentsegmenten bestaan die verschillen in productbezit en die kunnen worden beschouwd als personen in verschillende levensfasen. Het tweede element in deze studie is het segmenteren van landen (Steenkamp en Ter Hofstede, 2002). In plaats van landen te segmenteren op basis van macro gegevens op landniveau gebruiken we hier micro data, ofwel de informatie over het productbezit van consumenten binnen landen. Om precies te zijn, we gaan na of er landclusters te vinden zijn die vergelijkbaar zijn wat betreft de mate van vóórkomen van de gevonden consumentclusters.

De eerste stap in de analyse is het vaststellen van het aantal consument en landsegmenten dat noodzakelijk is om de gegevens in voldoende mate te beschrijven. Dat doen we met hulp van CAIC (Consistent Akaike Information Criterion). We hebben het model geschat met 1 t/m 15 consumentsegmenten en 1 t/m 8 landsegmenten. Het model met 14 clusters van consumenten en 7 van landen heeft de laagste CAIC waarde en is daarom het beste model.

[Voeg Tabel 2 hier in]

Tabel 2 laat a posteriori klasse lidmaatschapkansen zien voor de 17 landen in onze studie. De classificatie van landen in segmenten valt sterk samen met de kaart van Europa. Het eerste cluster bevat de lage landen (België and Nederland) en Duitsland. Het tweede cluster bevat de Scandinavische landen en Oostenrijk. Luxemburg blijkt zowel kenmerken van cluster 1 als cluster 2 te hebben. Groot-Brittannië, Ierland, en Noord-Ierland vormen cluster 3. In deze drie landen lijken de segmentstructuren sterk op elkaar. In tegenstelling tot de andere delen van Europa bestaat Zuid-Europa uit meerdere kleine segmenten: Italië en Portugal vormen samen cluster 4 en Spanje, Griekenland en Frankrijk vormen ieder een eigen cluster. In elk van deze landen komt een unieke segmentstructuur voor.

Het bovenste deel van Tabel 3 geeft informatie over het productbezit binnen elk van de 14 consumentsegmenten. Om de interpretatie te vereenvoudigen zijn de clusters zodanig geordend dat de geaggregeerde penetratie over alle producten oploopt. In segment 1 bezitten consumenten de minste producten en consumenten in segment 14 de meeste. Opvallend is de zeer lage penetratiegraad in de eerste drie segmenten van zowel een basis betaalmiddel (lopende rekening) als van meer geavanceerde betaalmiddelen (creditkaart of andere bankpas). Segment 1 heeft een lage kans voor alle producten, terwijl segment 2 enkel een vrij hoge penetratiegraad heeft voor spaarrekening en segment 3 voor spaarrekening en chequeboek. Consumenten in segmenten 4 tot en met 9 hebben penetratiegraden van bijna 1 voor lopende rekening en enkele andere betaalmiddelen, waarbij het type betaalmiddel verschilt tussen deze segmenten. Bijvoorbeeld, creditkaart bezit is heel hoog in segment 5 terwijl chequeboek bezit juist meer voorkomt in segmenten 8 and 9. Segmenten 10 tot en met 14 kenmerken zich door een hoge kans op het bezitten van minstens één kredietproduct (rood staan, hypotheek of andere lening). Segmenten 13 and 14 bevatten de meeste intensieve gebruikers met relatief hoge penetratiegraden voor alle acht financiële producten, waarbij vooral de hoge waarde voor hypotheken in segment 14 opvalt.

[Voeg Tabel 3 hier in]

De resultaten in het onderste gedeelte van Tabel 3 koppelen de landsegmenten aan de consumentsegmenten. Het betreft informatie over de relatieve omvang van de 14 consumentclusters binnen ieder van de 7 landclusters. Veertien consumentsegmenten lijkt veel, maar veel segmenten bevatten maar 1% of zelfs maar 0.1% van de consumenten in de meeste landsegmenten. Dit betekent dat het werkelijke aantal consumentsegmenten in een bepaald landsegment steeds beduidend kleiner is dan 14. Bijna alle consumentsegmenten, vooral degene met hoge productpenetraties (3 t/m 14), zijn groot (zeg groter dan 10%) in maar één of twee landsegmenten. Anderzijds zijn consumenten segmenten 1 and 2, met een lage penetratie voor alle producten, vrij groot in vier landsegmenten.

Voor elk landsegment zien we een paar veel voorkomende consumentsegmenten, een aantal met lage penetratiegraden (consumentsegmenten 1 t/m 3), een aantal met midden tot hoge graden maar lage graden voor kredietproducten (consumentsegmenten 4 t/m 9) en een aantal met hoge graden voor de meeste producten (segmenten 10 t/m 14). Landsegment 2 (Oostenrijk en de Scandinavische landen) bevat bijvoorbeeld consumenten in segmenten 2, 6, 7, 10 en 13, die allemaal een lage penetratiegraad hebben voor chequeboek. De zeer diverse set van consumentsegmenten die in landsegment 3 (Ierland, Noord-Ierland, Groot-Brittannië) veel voorkomen gaat van segmenten met zeer lage kansen voor alle producten (consumentsegment 1) tot en met segmenten met extreem hoge kansen voor alle producten (consumentsegment 14). Consumentsegment 2, met lage penetratiegraden voor alle producten behalve voor de spaarrekening, is extreem dominant in Griekenland (landsegment 6). Frankrijk (landsegment 7) bevat vrijwel alleen consumentsegmenten 2, 8, 11, and 12, allen met een hoge kans op bezit van een chequeboek.

6. Marketing implicaties van de resultaten

Onze empirische resultaten zijn zeer relevant voor internationaal opererende financiële instellingen. We zien dat, tegen verwachtingen in (Berger, Dai, Ongena en Smith, 2003; Ganesh, 1998), er in Europa nog steeds sprake blijkt te zijn van grote verschillen tussen landen en consumenten. Er zijn wel een aantal landen in hetzelfde landsegment in te delen, bijvoorbeeld, Duitsland, Nederland, Luxemburg en België, maar dit is wellicht een gevolg van culturele overeenkomsten van voor het EU-tijdperk. Voor financiële dienstverleners suggereren deze bevindingen dat Europa nog niet als een geheel kan worden benaderd. Verder is er ook duidelijk geworden welke groepen landen mogelijk hetzelfde kunnen worden benaderd, bijvoorbeeld, Duitsland, Nederland, Luxemburg en België, en welke specifieke landen niet, Frankrijk, Griekenland en Spanje.

Ook voor het targetten van een enkel product zijn onze bevindingen zeer relevant. Bijvoorbeeld, een creditcard bedrijf kan de bevindingen die zijn gerapporteerd in dit artikel gebruiken

om landsegmenten te selecteren voor de introductie van hun product. Vervolgens kan het bedrijf ook bepalen welke segmenten binnen elk land voor benadering in aanmerking komen. Als het bedrijf een nieuwe creditcard introduceert, waarmee men bestaande gebruikers van creditcards hoopt aan te spreken, dan kan de financiële dienstverlener consumentensegment 5 in landensegment 5 (Spanje) benaderen. Andere segmenten met veel bezitters van creditcards zijn consumentensegment 11, dat veel voorkomt in landensegment 7 (Frankrijk) en consumentensegment 14, dat veel is te vinden in landensegment 3 (Groot Brittanië, Ierland en Noord-Ierland).

7. Discussie

In dit artikel hebben we aandacht besteed aan een voor marktonderzoek zeer interessante uitbreiding van het latente klasse model. Deze uitbreiding maakt het mogelijk om een simultane clusteranalyse uit te voeren van eenheden op een lager en op een hoger niveau. Het model werd geïllustreerd met behulp van een voorbeeld waarin consumenten werden geclusterd op basis van het bezit van financiële producten en landen op basis van de omvang van de verschillende consumentsegmenten.

Inhoudelijk gezien is dit een goed gekozen voorbeeld omdat het gemakkelijk te volgen is en de marketingrelevantie evident is. Aan de andere kant zijn er wellicht nog betere voorbeelden denkbaar. In ons voorbeeld was het aantal hogere niveau eenheden (landen) vrij klein terwijl multi-niveau analyse nog meer voordelen biedt in situaties waarin het aantal groepen groot is (50 of meer). Ook is in ons voorbeeld het aantal personen per groep groot, meestal ongeveer 1000 consumenten per land, terwijl tussen de 5 a 30 personen per groep de meer standaard situatie is bij een multiniveau analyse. Er zijn waarschijnlijk verschillende marktonderzoektoepassingen die nog beter passen bij de meer standaard situatie waarin veel meer groepen onderzocht worden die ieder bestaan uit veel minder onderzochte consumenten.

Literatuur

Ambroise, C, en G. Govaert, 2000. 'Clustering by maximizing a fuzzy classification maximum likelihood criterion'. In: W. Jansen and J.G. Bethlehem (eds.), Proceedings in Computational Statistics 2000. Heidelberg: Physica-Verlag, pp. 187-192.

Berger, A.N., Q. Dai, S. Ongena, en D.C. Smith, 2003. To what extent will the banking industry be globalized? A study of bank nationality and reach in 20 European nations. Journal of Banking & Finance, 27, 383-415.

Bijmolt, T.H.A., Paas, L.J., en J.K. Vermunt, 2004. Country and consumer segmentation: Multi-level latent class analysis of financial product ownership. International Journal of Research in Marketing, 21, 323-340.

Christensen, T., 2001. Eurobarometer56.0: Information and communication technologies, financial services, and cultural activities, August-September 2001. Brussels: European Opinion Research Group.

Ganesh, J., 1998. Converging trends within the European Union: Insights from an analysis of diffusion patterns. Journal of International Marketing, 6, 32-48.

Goodman, L.A., 1974. The analysis of systems of qualitative variables when some of the variables are unobservable: Part I - A modified latent structure approach. American Journal of Sociology, 79, 1179-1259.

Hedeker, D., 2003. A mixed-effects multinomial logistic regression model. Statistics in Medicine, 22, 1433-1446.

Kamakura, W.A., S.N. Ramaswami en R.K. Srivastava, 1991. Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. International Journal of Research in Marketing, 8, 329-349.

Kamakura, W.A., Wedel, M. en J. Agrawal, 1994. Concomitant variable latent class models for the external analysis of choice data. International Journal of Research in Marketing, 11, 451-464.

Magidson, J. en J.K. Vermunt, 2002. Latent class modeling as a probabilistic extension of K-means clustering. Quirk's Marketing Research Review, March 2002, 20 & 77-80.

Magidson, J. en J.K. Vermunt, 2004. 'Latent class analysis'. In: D. Kaplan (ed.), The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oakes: Sage Publications, pp. 175-198.

McFadden, D., 1974. 'Conditional logit analysis of qualitative choice behavior'. In: I. Zarembka (Ed.), Frontiers in econometrics. New York: Academic Press, pp. 105-142.

McLachlan, G.J., en D. Peel, 2000. Finite Mixture Models. New York: John Wiley & Sons, Inc.

Paas, L.J. (2003). 'De toepassing van het Mokkenschaalmodel voor acquisitiepatroonanalyse in de financiële dienstverlening'. In: A.E. Bronner, P. Dekker, J.C. Hoekstra, E. de Leeuw, Th.B.C. Poiesz, K. de Ruyter en A. Smidts (Eds.), Jaarboek Marktonderzoeksassociatie, Haarlem: Vrieseborch, pp. 121-137.

Paas, L.J. en I.W. Molenaar, 2005. Analysis of acquisition patterns: A theoretical and empirical evaluation of alternative methods. International Journal of Research in Marketing, 22, 87-100.

Snijders, T.A.B. en R.J. Bosker, 1999. Multilevel Analysis. London: Sage Publications.

Steenkamp, J.-B.E.M., en F. Ter Hofstede, 2002. International market segmentation: Issues and perspectives. International Journal of Research in Marketing, 19, 185-213.

Vermunt, J.K., 2002. Comments on "Latent class analysis of complex sample survey data". Journal of the American Statistical Association, 97, 736-737.

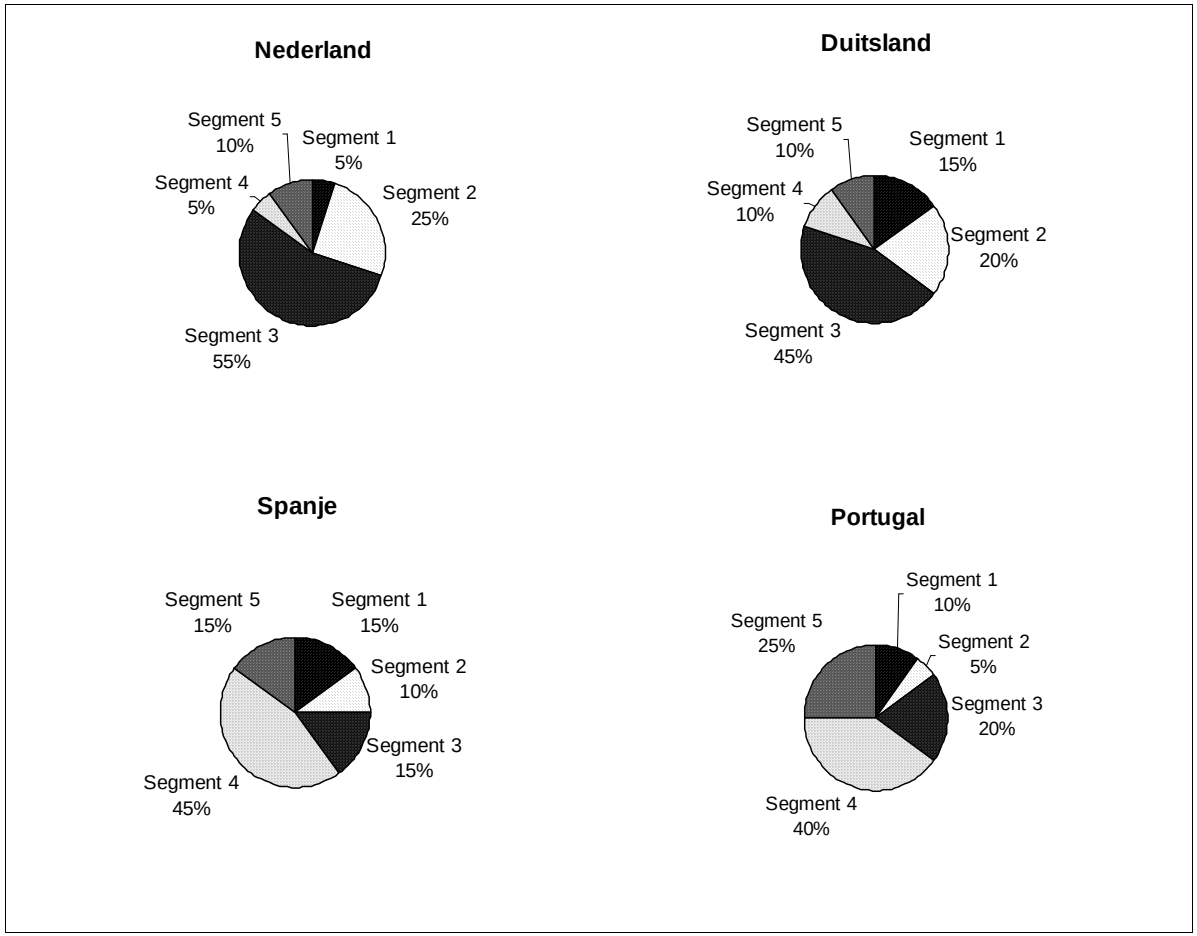
Vermunt, J.K., 2003. Multilevel latent class models. Sociological Methodology, 33, 213-239.

Vermunt, J.K. en J. Magidson, 2002. 'Latent class cluster analysis'. In: J. Hagenaars en A. McCutcheon (eds.), Applied latent class models. Cambridge: Cambridge University Press, pp. 89-106.

Vermunt, J.K. en J. Magidson, 2005. Latent GOLD User's Manual. Boston: Statistical Innovations Inc.

Wedel, M. en W.A. Kamakura, 2000. Market segmentation: Conceptual and methodological foundations (second edition). Dordrecht: Kluwer.

Wedel, M., F. Ter Hofstede en J.-B.E.M. Steenkamp, 1998. Mixture model analysis of complex samples. Journal of Classification, 15, 225-244.



Figuur 1: Fictief voorbeeld van segmentstructuren in verschillende landen

Tabel 1. Beschrijvende Statistieken voor de Internationale Steekproef

Land	Steekproef-omvang	Gemiddeld gewicht	Bezit van Financiële Producten (Steekproefproporties)							
			Lopende rekening	Spaar-rekening	Credit-kaart	Bankpas	Cheque-boek	Rood staan	Hypo-theek	Lening
Oostenrijk	1093	.34	71.5	82.3	33.7	61.0	21.6	41.4	17.7	21.8
België	1031	.43	85.2	85.9	39.2	74.0	34.2	33.3	25.9	21.1
Denemarken	1001	.22	78.5	63.2	48.2	60.8	33.9	55.2	51.8	36.3
Finland	1023	.21	87.5	50.2	31.5	84.7	0.7	16.0	22.0	26.5
Frankrijk	1002	2.42	87.8	69.8	57.7	31.0	87.9	50.6	18.6	26.6
Oost Duitsland	1024	.67	91.8	76.1	22.5	81.2	41.5	35.7	13.0	23.0
West Duitsland	1023	2.87	89.5	84.2	29.9	78.0	40.9	39.8	16.8	16.5
Groot-Brittannië	1041	2.37	75.2	77.1	52.3	58.7	76.4	29.3	37.1	20.1
Griekenland	1001	.45	11.0	79.7	18.7	25.9	6.3	3.4	14.6	11.6
Ierland	1002	.15	51.4	71.7	32.3	40.3	45.1	16.2	25.7	26.6
Italië	998	2.52	65.6	19.4	36.3	51.3	62.7	10.0	12.3	12.8
Luxemburg	609	.03	84.7	81.8	65.0	69.6	49.6	50.9	29.7	30.0
Nederland	1047	.65	89.5	82.5	37.2	94.3	26.6	63.6	33.6	14.9
Noord Ierland	305	.22	62.3	59.7	42.3	41.3	62.3	21.3	35.7	14.1
Portugal	1000	.42	70.0	44.2	33.0	33.0	60.8	2.5	13.0	8.0
Spanje	1000	1.70	61.6	67.2	52.1	33.2	17.4	8.2	19.4	17.2
Zweden	1000	.37	76.1	77.5	57.0	59.4	19.7	19.0	34.7	25.8

Tabel 2. Model Resultaten: Landsegmenten

Land-segment	Relatieve Omvang	A Posteriori Kansen voor Landsegment Lidmaatschap $\{P(r\|\mathbf{y}_j)\}$ *	
		Land	Kans
1	.256	België, Oost Duitsland, West Duitsland, Nederland	1.00
		Luxemburg	.81
2	.260	Oosterijk, Denemarken, Finland, Zweden	1.00
		Luxemburg	.19
3	.175	Groot-Brittanië, Ierland, Noord Ierland	1.00
4	.119	Italë, Portugal	1.00
5	.064	Spanje	1.00
6	.064	Griekenland	1.00
7	.064	Frankrijk	1.00

* Alle niet gerapporteerde a posteriori kansen zijn kleiner dan 0.01

Tabel 3. Model Resultaten: Consumentsegmenten

	Consumentsegmenten:													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Financiële producten:	Kansen op Productbezit $\{P(y_{ijt}=1\|s)\}$													
Lopende Rekening	.05	.05	.33	1.00	1.00	.92	.98	1.00	1.00	.88	1.00	.99	.98	1.00
Spaarrekening	.39	.85	.61	.14	.65	.61	.91	.91	.15	.71	.78	.73	.93	.86
Creditkaart	.00	.09	.00	.29	.84	.23	.16	.62	.76	.81	.96	.49	.51	.83
Bankpas	.00	.19	.00	.62	.32	.80	.85	.53	.90	.69	.00	.99	.95	.87
Chequeboek	.09	.01	.86	.79	.13	.02	.29	1.00	.95	.30	.98	1.00	.50	1.00
Rood staan	.00	.03	.31	.02	.04	.42	.19	.11	.22	.29	.65	.68	.68	.60
Hypotheek	.01	.11	.25	.01	.14	.06	.00	.03	.29	.60	.22	.06	.55	.89
Lening	.02	.08	.21	.01	.14	.31	.00	.02	.27	.47	.33	.33	.30	.39
Landsegmenten:	Relatieve Omvang van Consumentsegmenten $\{P(s\|r)\}$													
1	.07	.03	.01	.00	.00	.17	.39	.03	.00	.00	.00	.08	.22	.01
2	.05	.17	.04	.00	.06	.26	.11	.01	.00	.18	.00	.01	.11	.01
3	.15	.11	.09	.03	.01	.00	.04	.26	.00	.02	.00	.07	.00	.23
4	.37	.00	.06	.29	.02	.00	.00	.01	.26	.00	.00	.00	.00	.00
5	.16	.27	.01	.01	.34	.00	.02	.03	.01	.15	.01	.00	.00	.01
6	.13	.79	.01	.00	.00	.00	.01	.02	.00	.04	.00	.00	.00	.00
7	.06	.00	.11	.07	.00	.02	.02	.14	.00	.00	.41	.14	.00	.04