MIXTURE HIDDEN MARKOV MODELLEN IN SOCIAAL-WETENSCHAPPELIJK ONDERZOEK

DOOR | **STEF BOUWHUIS, DIMITRIS PAVLOPOULOS, MAURICIO GARNIER-VILLARREAL & LAURA EBERLEIN**
Faculteit Sociale Wetenschappen, Vrije Universiteit Amsterdam

1. INLEIDING

De laatste decennia is steeds meer longitudinale data van hoge kwaliteit beschikbaar gekomen voor sociaal-wetenschappelijk onderzoek. Dit heeft ertoe geleid dat procesmatige manieren van het analyseren van data aan populariteit gewonnen hebben (Abbott, 1983). In plaats van 'slechts' naar enkele transities te kijken, kunnen we met een procesmatige methode een hele periode en de volgorde waarin gebeurtenissen plaatsvinden, onderzoeken. Deze manieren van data-analyse hebben dus als voordeel dat zij een omvattend en gedetailleerd beeld kunnen schetsen van ontwikkelingen in de levensloop van mensen, bijvoorbeeld op het gebied van carrières (Fasang & Liao, 2014; Studer & Ritschard, 2016). Een veelgebruikte procesmatige analysemethode in de sociale wetenschappen is sequentie analyse (SA). In SA worden sequenties van een bepaald fenomeen geconstrueerd voor elk individu in de studie, bijvoorbeeld sequenties van arbeidsmarktstatussen. Vervolgens worden deze sequenties meestal geclusterd op basis van overeenkomstigheid met behulp van clusteringalgoritmes, voornamelijk met de Ward clustering. Een uitbreiding op deze methode is *multi-channel* SA, waarbij van twee of meer fenomenen gecombineerd sequenties vervaardigd worden. *Multi-channel* SA is geschikt om processen met verschillende dimensies te bestuderen. Een voorbeeld is het onderzoek van Mattijssen et al. (2023), waarin Nederlandse registerdata gebruikt is om sequentietypes van schoolverlaters te maken op basis van contracttype en (gecategoriseerd) inkomen. Door gebruik te maken van *multi-channel* SA, kon dit onderzoek de rijke variatie aan type loopbanen van werknemers die op de arbeidsmarkt instromen, beschrijven.

SA is een deterministische methode, in de zin dat sequenties aan clusters toegewezen worden waarna geanalyseerd kan worden tot welke clusters individuen behoren. Een ander type modellen, waaronder Markov modellen, is probabilistisch van aard (Liao et al., 2022; Vermunt, 2010a). In dit type modellen wordt modelmatig geschat hoe groot de kans is dat een situatie zich op moment t voordoet, gegeven de situatie op *t-1*. Een voorbeeld is het weer: hoe groot is de kans dat het op maandag zonnig is gegeven dat het op zondag zonnig was? Het feit dat dit modelmatig is gedaan betekent dat het model voor meetfouten corrigeert. Het weer van vandaag is medebepaald maar niet perfect bepaald door het weer van gisteren. Een uitbreiding op dit model is het hidden Markov model (HMM). *Hidden* refereert hier aan het latente karakter van het betreffende (sociale) fenomeen. Of het buiten zonnig is, is direct te observeren. Veel sociale fenomenen zijn echter niet zo eenvoudig direct te observeren en zijn dus latent. Dit type variabelen wordt geobserveerd door middel van indicatoren die wel direct te observeren zijn. HMMs zijn geschikt voor multi-dimensionele

concepten, zoals opvattingen over democratie. Verschillende dimensies in dit soort onderzoek zijn bijvoorbeeld: een liberale dimensie die meet in hoeverre verkiezingen en bescherming van minderheden als onderdeel van democratie gezien worden en een redistributieve dimensie, die meet in hoeverre redistributie van inkomen als onderdeel van democratie gezien wordt. Individuen die vergelijkbaar scoren op beide dimensies worden gegroepeerd in clusters, in HMM staten genoemd. In HMM worden verschillende indicatoren als het ware samengevoegd tot één latente categoriale variabele en onderzoeken we de bewegingen tussen de categorieën van die variabele (de staten). Dit heeft als voordeel ten opzichte van multi-channel SA, waarin transities voor elk van de variabelen afzonderlijk onderzocht worden, dat de uitkomsten van het model overzichtelijker zijn. Een ander voordeel is dat een HMM als uitkomst een transitiekansen matrix geeft, waarmee inzicht verkregen kan worden in hoe groot de kans is dat een individu van een bepaalde staat naar een andere staat beweegt. Multi-channel SA geeft geen vergelijkbaar overzicht om de transities tussen staten te kunnen interpreteren.
Een hidden Markov model kan uitgebreid worden naar een Mixture hidden Markov model (MHMM), dat enigszins lijkt op SA. De uitbreiding houdt in dat mixtures (trajecten) gemodelleerd worden door individuen met vergelijkbare transitiepatronen tussen de staten van de latente variabele te clusteren. De veronderstelling is dat de trajecten waartoe mensen behoren ook een latente variabele is, met de initiële staat, de staat waarin een individu zich bevindt op t=0, en de transitiekansen als indicatoren. Het model schat de kans om tot elk traject te behoren. Een belangrijk voordeel van MHMM ten opzichte van multi-channel SA is dat predictoren van de clustering in staten en trajecten aan het model toegevoegd kunnen worden. Daardoor zijn we in staat om verschillende soorten causale vragen te beantwoorden met MHMMs.

Het doel van dit artikel is om de lezer bekend te maken met MHMMs. In de methoden sectie zullen we ingaan op de stappen die je als onderzoeker dient te zetten om een MHMM uit te voeren, met de nadruk op de te maken keuzes. In de resultatensectie bespreken we een voorbeeld van de toepassing van een MHMM om de te zetten stappen te illustreren en om een indruk te geven van hoe de resultaten van dergelijke modellen gepresenteerd en gevisualiseerd kunnen worden. We sluiten af met een conclusie waarin we kort reflecteren op het gebruik van MHMMs in sociaal-wetenschappelijk onderzoek.

2. METHODEN

MHMMs behoren tot de familie van latente variabelenmodellen. Daarom beginnen we met een korte uiteenzetting over latente klasse analyse (LKA), waarna we ingaan op hoe MHMM daarop voortborduurt.
LKA is een analysemethode waarmee op basis van cross-sectionele data onderzoek gedaan kan worden naar discrete latente variabelen (Masyn, 2013; Nylund-Gibson et al., 2019; Nylund-Gibson & Choi, 2018). Individuen met vergelijkbare scores op de indicatoren van de latente variabele worden in een staat gegroepeerd. In deze zin is LKA een persoons-georiënteerde methode in tegenstelling tot factoranalyse, wat een variabele-georiënteerde methode is (Collins & Lanza, 2010). Een belangrijke keuze die gemaakt moet worden in LKA is het optimale aantal staten. Dit wordt in twee stappen gedaan: (1) modellen met een oplopend aantal staten worden geschat;

en (2) een onderzoeker kiest het meest optimale model. Dit wordt gedaan zowel op basis van statistische parameters (Bayesian Information Criterion, Aikake Information Criterion) als op basis van de interpreteerbaarheid van de staten. Voor de tweede stap zijn latente klasse heterogeniteit en onderscheidingsvermogen (ook wel separatie of entropie genoemd) van belang. Dat houdt in dat de voorkeur uitgaat naar een model met homogene staten en waarin de staten voldoende van elkaar verschillen. De staten in het gekozen model vertegenwoordigen feitelijk categorieën van de latente variabele. Elk van de categorieën dient een label te krijgen op basis van de scores van de individuen in de betreffende categorie op de indicatoren van de latente variabele. Vervolgens wordt voor elk individu voor elke staat geschat wat de kans is dat dit individu zich in de betreffende staat bevindt. Het individu wordt toegewezen aan de staat waarvoor de kans het hoogst is dat hij of zij zich daarin bevindt. Het LKA-model bestaat uit twee delen: (1) *measurement* deel; en (2) *structural* deel. Het eerste deel heeft betrekking op hoe de indicatoren gerelateerd zijn aan de latente variabele en het tweede deel op de distributie van individuen over de categorieën van de latente variabele.

Hidden Markov Models (HMMs) zijn een longitudinale extensie van LKA (Nylund-Gibson et al., 2023; van der Nest et al., 2020; Vermunt, 2010b). Dat betekent dat de staten op basis van longitudinale data geïdentificeerd worden en dat het model bewegingen (transities) tussen de staten modelleert. Dit laatste wordt, zoals gezegd, gedaan op basis van een Markov transitie structuur: de staat waarin een individu zich op moment *t* bevindt wordt uitsluitend voorspeld door de staat waarin hij of zij zich op *t-1* bevond. De aanname is dat de waardes van de geobserveerde indicatoren op verschillende tijdspunten onafhankelijk zijn, conditioneel op de latente (*hidden*) staten. Dat betekent dat de afhankelijkheid tussen tijdspunten verklaard moet worden door de autocorrelatie structuur van de latente klassen. Net als LKA bestaan HMMs uit een *measurement* deel en een *structural* deel. Ook in HMM refereert het *meausurement* deel aan de relatie tussen de indicatoren en de latente variabele in de vorm van *item-response probabilities*: de kans op een score op een indicator, gegeven de staat waarin een individu gegroepeerd is. Het *structural* deel refereert in HMM aan de initiële staat, dat wil zeggen, de staat waarin een individu geclassificeerd is op *t=0* en aan de transitiekansen (*transition probabilities*): dit zijn de kansen dat een individu van een staat naar een andere staat beweegt tussen tijdspunt *t-1* en *t*. Het aantrekkelijke aan HMM (en LKA) is dat het mogelijk is om predictoren aan zowel dit *measurement* deel en aan het *structural* deel toe te voegen. Het is dus bijvoorbeeld mogelijk om te onderzoeken of leeftijd van invloed is op de initiële staat van een individu en/of de transitiekansen. In het standaardmodel wordt ervan uitgegaan dat deze transitiekansen op elk tijdspunt gelijk zijn. Stel je voor dat je 40 tijdspunten hebt, dan is de assumptie dat de transitiekansen van tijdspunt 1 naar tijdspunt 2 identiek zijn aan de transitiekansen van tijdspunt 39 naar tijdspunt 40. Dat is niet in alle gevallen waarschijnlijk. Het is dan ook mogelijk om een tijdselement aan het model toe te voegen, zodat voor elk tijdspunt aparte transitiekansen berekend worden.

Een verdere extensie zijn de MHMMs. Bovenop de structuur van HMMs worden aan deze modellen *mixtures*, oftewel trajecten, toegevoegd (Vermunt, 2010b). Deze trajecten kunnen ook gezien worden als categorieën van een tijdsconstante latente variabele. In dit geval bestaat de latente variabele uit de patronen van ontwikkeling die individuen door de tijd meemaken, met andere woorden, de transities tussen staten van de eerdere latente variabele. Individuen worden gegroepeerd in trajecten op basis van

hun initiële staat en transitiekansen. Ook hier speelt de onderzoeker een belangrijke rol in het kiezen van het optimale aantal trajecten. Eerst wordt een aantal modellen geschat met een oplopend aantal trajecten en vervolgens bepaalt de onderzoeker, op basis van de statistische criteria die hierboven genoemd zijn en de interpreteerbaarheid van de trajecten, welk aantal trajecten optimaal is. Hoe groter het aantal trajecten hoe meer tijdrovend het schatten van het model is. Om de schattingsduur te verminderen kan de twee-stappen procedure van Bakk en Kuha (2018) toegepast worden. In deze procedure worden de geschatte parameters van het *measurement* deel (uit een model met maar één mixture) gefixeerd bij het modelleren van de trajecten. De uitkomst van de schatting is dus ook hier een discrete (categoriale) variabele. De categorieën worden wederom gelabeld aan de hand van de scores op de indicatoren van de latente trajecten, namelijk de initiële staat en transitiekansen. Ook aan trajecten kunnen predictoren toegevoegd worden. Het is aan de onderzoeker om te bepalen waar in het model (initiële staat, transitiekansen en/of trajecten) predictoren toegevoegd worden. Bij de keuze voor de optimale plek voor de predictoren kunnen zowel statistische overwegingen (*model fit*) en theoretische overwegingen een rol spelen.

3. EEN VOORBEELD VAN GEBRUIK VAN MHMMS

In deze sectie gaan we het gebruik van MHMMs voor het analyseren van longitudinale data toelichten aan de hand van een onderzoek naar de carrières van jonge werknemers (Eberlein et al., 2024). Informatie over de data en studiepopulatie in dit onderzoek is te vinden in de tekst box.

Data en studiepopulatie

Voor dit onderzoek is registerdata van het Centraal Bureau voor de Statistiek (CBS) gebruikt[1]. Deze data bevat maandelijkse gegevens over onder andere contractvorm, werkuren en inkomen. Individuen die tussen de periode 2009-2013 hun opleiding hebben verlaten (met of zonder diploma) zijn in dit onderzoek voor een periode van 72 maanden gevolgd. Alleen mensen die een opleiding verlaten hebben (met of zonder diploma) die bedoeld zijn om een arbeidsmarktkwalificatie te verkrijgen, zijn geïncludeerd (MBO, HBO of WO, ISCED niveaus 353, 354, 645, 655, 747 en 757). Alleen schoolverlaters met werk in minimaal één van deze maanden zijn geïncludeerd en de follow-up periode begon in de eerste maand dat een schoolverlater werk had. In totaal voldeden 672.757 individuen aan de inclusiecriteria. Vanwege de rekenkracht die nodig is voor MHMMs is uit deze groep een willekeurige steekproef van 12.000 individuen getrokken.

Het doel van dit onderzoek was om inzicht te krijgen in hoe *employment quality* zich ontwikkelt in de eerste jaren van de loopbanen van Nederlandse schoolverlaters: mensen die hun opleiding verlaten hebben (met of zonder diploma) en de arbeidsmarkt betreden hebben. De latente variabele in het onderzoek is dus *employment quality*, een

(1) Bakker, B. F., Van Rooijen, J., & Van Toor, L. (2014). The system of social statistical datasets of Statistics Netherlands: An integral approach to the production of register-based social statistics. Statistical Journal of the IAOS, 30(4), 411-424.

multi-dimensioneel concept met twee dimensies: (1) werkzekerheid; en (2) hoogte van het inkomen. Een eerste vraag die beantwoord moet worden, is: welke indicatoren gebruiken we om deze latente variabele te meten? In dit onderzoek is ervoor gekozen om deze variabele te meten aan de hand van de volgende indicatoren: (1) werkstatus (in loondienst, zelfstandig, niet werkzaam); (2) contracttype (vast, tijdelijk, oproep/ uitzendkracht, geen contract); (3) werkuren; en (4) maandelijks inkomen. Het is mogelijk dat er overlap bestaat tussen de indicatoren werkstatus en contracttype. Er is toch voor gekozen om deze als aparte indicatoren in het model op te nemen, omdat het bijvoorbeeld mogelijk is dat een individu zowel werkt als zelfstandige als in loondienst is.

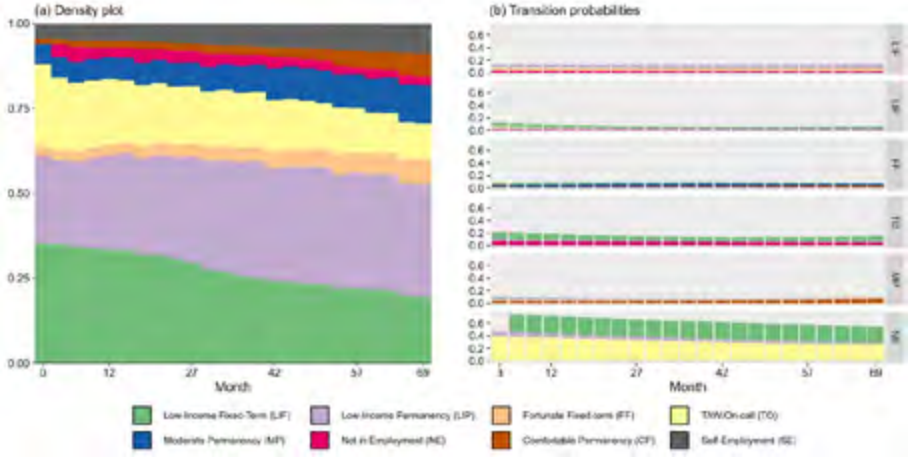
De eerste stap in het uitvoeren van een MHMM is het kiezen van het optimale aantal staten (of categorieën) van de latente variabele, in dit geval dus *employment quality*. In het hier beschreven onderzoek is gekozen voor het model met acht staten, wat betekent dat onze latente variabele acht categorieën met verschillende niveaus of vormen van *employment quality* heeft. Deze categorieën zijn vervolgens gelabeled op basis van hoe de individuen in de staten scoren op de hierboven genoemde indicatoren (zie Tabel 1). 'Laag inkomen, tijdelijk contract' bijvoorbeeld, is zo genoemd omdat 99% van de mensen in deze staat een tijdelijk contract hebben en zij een relatief laag inkomen hebben. De volgende acht staten hebben we onderscheiden (we presenteren eerst de Nederlandse namen en daarna de oorspronkelijke Engelse namen): (1) laag inkomen, tijdelijk contract (*low income, fixed-term*); (2) hoog inkomen, tijdelijk contract (*fortunate fixed term*); (3) uitzend- en oproepkrachten (*TAW/on-call*); (4) zelfstandigen (*self-employed*); (5) laag inkomen, vast (*low income permanency*); (6) matig inkomen, vast (*moderate permanency*); (7) comfortabel vast (*comfortable permanency*); en (8) niet in werk (*not in employment*).

TABEL 1: LATENTE STATEN VAN EMPLOYMENT QUALITY (PERCENTAGES)

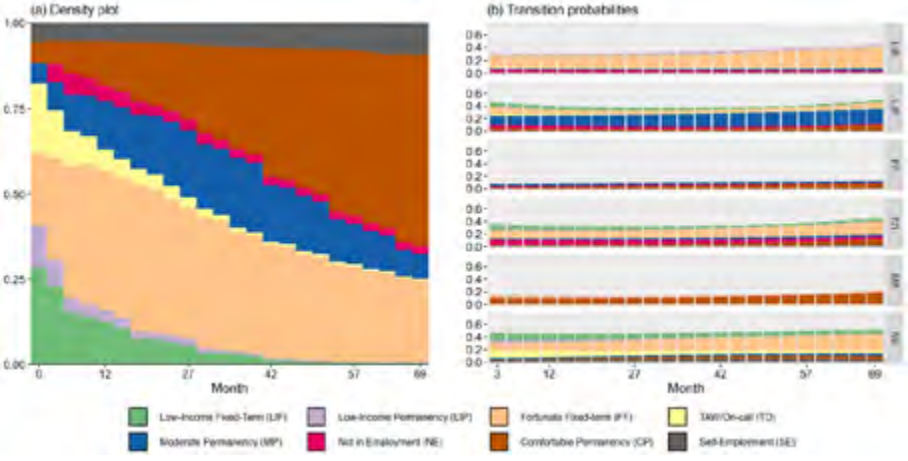
	Acht staten van de latente variabele employment quality							
Indicatoren	**Laag inkomen, tijdelijk contract**	**Hoog inkomen, tijdelijk contract**	**Uitzend- en oproep-krachten**	**Zelf-standigen**	**Laag inkomen, vast**	**Matig inkomen, vast**	**Comfortabel vast**	**Niet in werk**
Omvang staat	21	13	12	6	18	12	9	10
Contract Type								
Permanent	0	0	1	0	100	100	100	0
Tijdelijk	99	100	2	0	0	0	0	0
Uitzendwerk	0	0	53	0	0	0	0	0
Oproepcontract	0	0	44	0	0	0	0	0
Geen contract	0	0	0	100	0	0	0	100
Maandelijks inkomen								
Geen inkomen	0	0	0	5	0	0	0	48
€ 1-€ 750	16	0	19	20	13	0	0	12
€ 751-€ 1500	32	0	32	20	26	1	0	31
€ 1501-€ 2250	51	3	32	14	61	5	0	7
€ 2251-€ 3000	1	70	13	11	0	95	0	2
€ 3001-€ 3750	0	19	3	10	0	0	66	0
> € 3750	0	8	1	19	0	0	34	0
Arbeidsmarktstatus								
Loondienst	100	100	100	0	100	100	100	0
Zelfstandig	0	0	0	99	0	0	0	0
Niet-werkend	0	0	0	1	0	0	0	100
Werkuren per week	30	39	23	–	30	38	40	0

Naast het *measurement* deel, dat inzicht geeft in de relatie tussen de indicatoren en de latente variabele, wordt in deze stap ook inzicht verkregen in het *structural* deel: de categorie van *employment quality* waarin individuen de arbeidsmarkt betreden (initiële staat) en de manier waarop individuen bewegen tussen de acht onderscheiden categorieën. Zoals gezegd wordt in eerste instantie één transitiematrix geschat, met alle mogelijke transitiekansen, berekend voor de hele studieperiode. Door een tijdsvariabele toe te voegen aan het structurele deel van het model, kan onderzocht worden of deze transitiekkansen gedurende de periode veranderen.
Nadat de eerste stap uitgevoerd is kan de tweede stap gezet worden. In deze stap wordt het optimale aantal trajecten, of *mixtures*, gekozen. Dit is gedaan door verschillende modellen met een oplopend aantal mixtures te schatten. Zoals gezegd worden in mixtures individuen met vergelijkbare initiële staten en transitiekansen gegroepeerd. In dit onderzoek zijn vier van zulke trajecten onderscheiden. Wederom worden de trajecten gelabeld op basis van hoe de individuen in de trajecten scoren op de initiële staten en vooral de transitiekansen. Het eerste traject in dit onderzoek is 'stabiel in de val' (*stable entrapment*) genoemd, omdat individuen voor langere perioden in de 'laag inkomen, tijdelijk contract' en 'laag inkomen, vast contract' staten verkeren. Figuur 1 laat de *density plot* en de transitiekansen van traject 1 zien. In de *density plot* is voor elk tijdsmoment te zien hoe de individuen in de studie verdeeld zijn over de acht staten die geïdentificeerd zijn. Er is te zien dat het aantal mensen 'laag inkomen, tijdelijk contract' en 'uitzend- en oproepkrachten' afneemt en het aantal mensen in 'laag inkomen, vast' en 'matig inkomen, vast' toeneemt. De *density plot* laat echter niet zien hoe individuen tussen de staten bewegen. Daar geven de transitiekansen (ook Figuur 1) meer inzicht in. De afkortingen van de staten helemaal rechts in de figuur geven de staat op *t-1* weer, en de balken in de figuur de hoogte van de kans om naar één van de andere staten te bewegen. Zo is te zien dat mensen uit 'laag inkomen, tijdelijk contract' (LIF) vooral bewegen naar 'laag inkomen, vast' (LIP), 'niet in werk' (NE) en 'hoog inkomen, tijdelijk contract' (FF). Mensen uit 'uitzend- en oproepkrachten' (TO) bewegen vooral naar 'laag inkomen, tijdelijk contract' term in het begin van de studieperiode en naar 'niet in werk' gedurende de hele periode. Het tweede traject is 'opwaartse mobiliteit' (*upward mobility*) genoemd: veel individuen beginnen in een staat met veel tijdelijke contracten, maar maken gedurende de studieperiode een transitie naar 'comfortabel vast' mee. Het derde traject is 'transities naar vast' (*moving to permanency*) genoemd. Ook hier beginnen veel individuen in een staat met veel tijdelijke contracten. Veel van deze individuen maken een transitie naar 'matig inkomen vast' mee: een staat met veel permanente contracten, maar met een lager maandinkomen dan 'comfortabel vast'. Het vierde en laatste traject is 'transities uit werk' (*moving out of employment*) genoemd. Zoals de naam suggereert maken veel individuen in dit traject een transitie naar de 'niet in werk' staat mee.

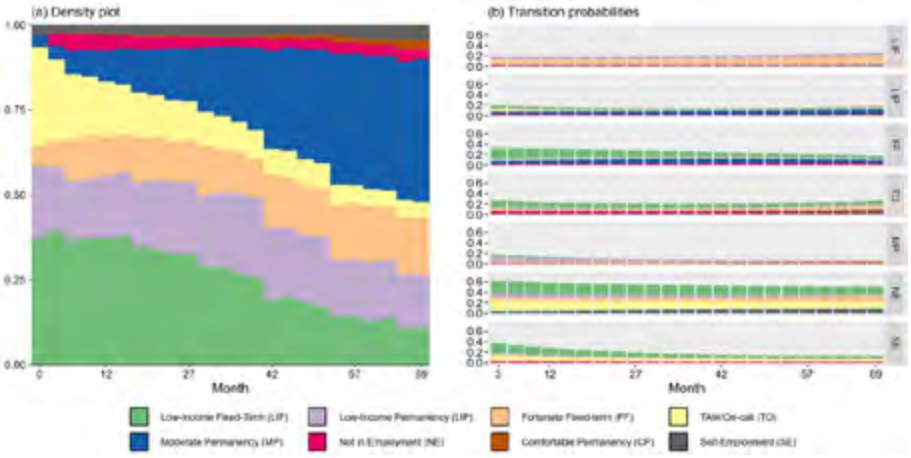
FIGUUR 1: VISUELE WEERGAVE VAN TRAJECT 1: 'STABIEL IN DE VAL'



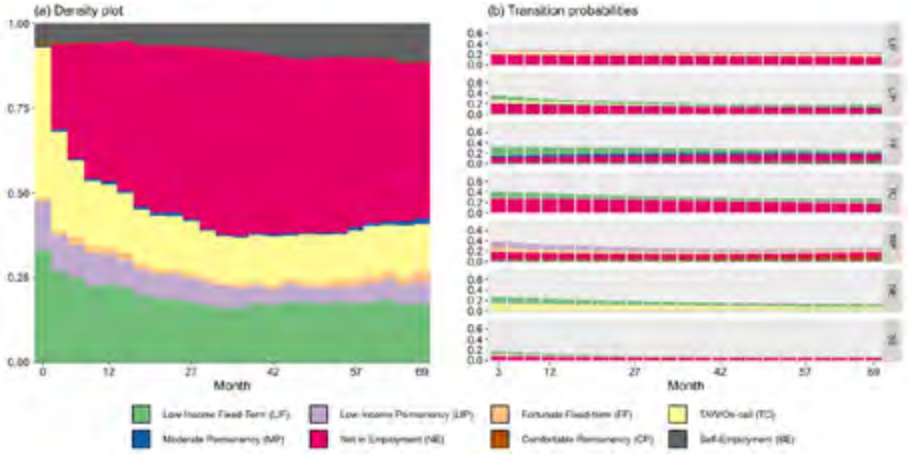
FIGUUR 2: VISUELE WEERGAVE VAN TRAJECT 2: 'OPWAARTSE MOBILITEIT'



FIGUUR 3: VISUELE WEERGAVE VAN TRAJECT 3: 'TRANSITIES NAAR VAST'



FIGUUR 4: VISUELE WEERGAVE VAN TRAJECT 4: 'TRANSITIES UIT WERK'



Een belangrijke vraag is hoe de resultaten van een MHMM het best gevisualiseerd kunnen worden. Het is belangrijk om nogmaals op te merken dat dit een probabilistisch model is: het model berekent de kans dat een individu zich in één van de categorieën van de latente variabele bevindt en de kans dat een individu zich in één van de trajecten bevindt. Desondanks zijn de visualisaties van de trajecten met *density plots* hierboven in essentie deterministisch: we gebruiken de voorspelde staat van een persoon op elk tijdstip. Deze deterministische visualisatie van de resultaten is weinig problematisch wanneer de entropie (onderscheidingsvermogen) van de staten en de trajecten hoog is. In dit model is de entropie gelijk aan 0,70. Dit is nogal laag maar

nog een steeds acceptabele waarde. Het is belangrijk te melden dat dit niet geldt voor de visualisatie van transitiekansen: deze zijn een directe output van het model en zijn dus niet gebaseerd op de voorspelde status.

Een laatste mogelijkheid van MHMMs die we hier bespreken is het toevoegen van predictoren. Zoals gezegd kunnen predictoren in verschillende delen van een MHMM toegevoegd worden: het *measurement* deel, het *structural* deel en/of het niveau van trajecten. In dit onderzoek is op basis van theoretische en statistische overwegingen gekozen voor het toevoegen van predictoren op het niveau van trajecten (opleidingsniveau en of een schoolverlater een diploma behaald had toen hij of zij de opleiding verliet) en initiële staten (duur tot het vinden van de eerste baan en leeftijd waarop de arbeidsmarkt betreden werd). In Tabel 2 en 3 staan de coëfficiënten van de predictoren, geschat door middel van een multinomiale regressieanalyse. Een positieve coëfficiënt wijst op een positief verband, een negatieve coëfficiënt op een negatief verband).[2] In Tabel 2 staan de coëfficiënten van de predictoren van de initiële staat. De positieve coëfficiënt voor 'matig inkomen, vast' (0,2329) in Tabel 2 laat zien dat naarmate het langer duurt om een baan te vinden, de kans groter is dat een individu zijn of haar loopbaan start in deze staat. Daarnaast is te zien dat naarmate iemand ouder is wanneer hij of zij zijn of haar eerste baan vindt, de kans groter is om in 'comfortabel, vast' te starten (0,3664).

TABEL 2: LOGIT COËFFICIËNTEN IN EFFECTS CODING, PREDICTOREN VAN INITIËLE STAAT

		Duur tot vinden baan	Leeftijd vinden baan
Initiële staat	Laag inkomen, tijdelijk contract	0,0361	-0,1847
	Hoog inkomen, tijdelijk contract	-0,0216	-0,1143
	Uitzend- en oproepkrachten	-0,0008	0,0704
	Zelfstandigen	0,0689	-0,1762
	Laag inkomen, vast	-0,3147	0,1936
	Matig inkomen, vast	0,2329	-0,1259
	Comfortabel vast	-0,0899	0,3664
	Niet in werk	0,0891	-0,0292

In Tabel 3 staan de coëfficiënten van de predictoren van de trajecten. In deze tabel is onder andere te zien dat studenten die de universiteit verlaten hebben een grotere kans hebben op 'opwaartse mobiliteit' (1,5604) dan studenten die een opleiding op een ander niveau verlaten hebben. Studenten die een diploma verkregen hebben,

(2) We presenteren de effect coëfficiënten. Deze tellen op tot 0. Een andere mogelijkheid is om een referentiecategorie te kiezen. De coëfficiënten laten dan zien hoe groot de kans is om in een bepaalde initiële staat te starten, ten opzichte van de referentie initiële staat. Dit kan ook voor trajecten gedaan worden.

hebben ook een grotere kans op 'opwaartse mobiliteit' (1,3799) dan studenten die hun opleiding verlaten hebben zonder diploma.

TABEL 3: LOGIT COËFFICIËNTEN IN EFFECTS CODING, PREDICTOREN VAN TRAJECTEN

		Opleidingsniveau			**Diploma verkregen**
		Universiteit	***Hoger beroepsonderwijs***	***Middelbaar beroepsonderwijs***	
Traject	Stabiel in de val	-1,141	-0,0588	1,1998	-0,5133
	Opwaartse mobiliteit	1,5604	0,1737	-1,7341	1,3799
	Transities naar vast	-0,0498	0,2986	-0,2488	0,6131
	Transities uit werk	-0,3696	-0,4136	0,7831	-1,4797

4. CONCLUSIE

Het doel van dit artikel was om de lezer bekend te maken met MHMM. Dat hebben we gedaan door het onderscheid met andere procesmatige analysemethoden, zoals SA, te bespreken, uit een te zetten welke stappen gezet moeten worden bij het schatten van een MHMM, welke keuzes onderzoekers daarin moeten maken, en het gebruik van MHMM te illustreren aan de hand van een voorbeeld.

De keuze tussen SA en MHMM is vooral bepaald door onze perceptie over het sociaal fenomeen dat we onderzoeken; moeten wij dit als probabilistisch bestuderen of is het genoeg als wij dit als deterministisch beschouwen? Aangezien SA een minder complexe en tijdrovende methode is, lijkt die meer geschikt te zijn om eendimensionale fenomenen te onderzoeken. Echter, bij multidimensionele fenomenen worden de resultaten van multichannel SA heel complex en dus moeilijk te evalueren en te interpreteren. MHMM, in tegenstelling tot SA, modelleert de latente structuur van het fenomeen en identificeert dus de meest relevante kenmerken van de latente staten en de trajecten. Dit is direct te zien wanneer we de resultaten van multichannel SA en MHMM vergelijken. MHMM is in staat om trajecten makkelijker te onderscheiden met hele heterogene data. De artikelen van Mattijssen et al. (2023) en Eberlein et al. (2024) hebben vergelijkbare registerdata op een vergelijkbare populatie gebruikt. Mattijssen et al komen op 14 trajecten van inkomensstatus en contractstatus met behulp van SA, terwijl Eberlein et al slechts 4 trajecten vinden door een MHMM te implementeren. Dit maakt de resultaten van onderzoek eenvoudiger te interpreteren en te communiceren.

MHMM is tenslotte meer geschikt in onderzoek naar causale verbanden. Aangezien MHMM een model is, kunnen causale relaties onderzocht worden door aan verschillende delen van MHMMs predictoren toe te voegen.

LITERATUUR

Abbott, A. (1983). Sequences of Social Events: Concepts and Methods for the Analysis of Order in Social Processes. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *16*(4), 129-147. https://doi.org/10.1080/01615440.1983.10594107

Bakk, Z., & Kuha, J. (2018). Two-Step Estimation of Models Between Latent Classes and External Variables. *Psychometrika*, *83*(4), 871-892. https://doi.org/10.1007/s11336-017-9592-7

Collins, L. M., & Lanza, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Science*. John Wiley & Sons, Inc.

Eberlein, L., Pavlopoulos, D., & Garnier-Villarreal, M. (2024). Starting flexible, always flexible? The relation of early temporary employment and young workers employment trajectories in the Netherlands. *Research in Social Stratification and Mobility*, *89*, 100861. https://doi.org/10.1016/j.rssm.2023.100861

Fasang, A. E., & Liao, T. F. (2014). Visualizing Sequences in the Social Sciences. *Sociological Methods & Research*, *43*(4), 643–676. https://doi.org/10.1177/0049124113506563

Liao, T. F., Bolano, D., Brzinsky-Fay, C., Cornwell, B., Fasang, A. E., Helske, S., Piccarreta, R., Raab, M., Ritschard, G., Struffolino, E., & Studer, M. (2022). Sequence analysis: Its past, present, and future. *Social Science Research*, 102772. https://doi.org/10.1016/j.ssresearch.2022.102772

Masyn, K. E. (2013). Latent Class Analysis and Finite Mixture Modeling. In P. E. Nathan & T. D. Little (Eds.), *The Oxford Handbook of Quantitative Methods: Vol. Volume 2: Statistical Analysis* (p. 63). Oxford University Press.

Mattijssen, L., Pavlopoulos, D., & Smits, W. (2023). Does it pay off to specialize? The interplay between educational specificity, level and cyclical sensitivity. *Social Science Research*, *109*, 102782. https://doi.org/10.1016/j.ssresearch.2022.102782

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440–461. https://doi.org/10.1037/tps0000176

Nylund-Gibson, K., Garber, A. C., Carter, D. B., Chan, M., Arch, D. A. N., Simon, O., Whaling, K., Tartt, E., & Lawrie, S. I. (2023). Ten frequently asked questions about latent transition analysis. *Psychological Methods*, *28*(2), 284–300. https://doi.org/10.1037/met0000486

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from Latent Classes: A Demonstration of Different Approaches to Include Distal Outcomes in Mixture Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967-985. https://doi.org/10.1080/10705511.2019.1590146

Studer, M., & Ritschard, G. (2016). What matters in differences between life trajectories: A comparative review of sequence dissimilarity measures. *Journal of the Royal Statistical Society. Series A: Statistics in Society*, *179*(2), 481-511. https://doi.org/10.1111/rssa.12125

van der Nest, G., Lima Passos, V., Candel, M. J. J. M., & van Breukelen, G. J. P. (2020). An overview of mixture modelling for latent evolutions in longitudinal data: Modelling approaches, fit statistics and software. *Advances in Life Course Research*, *43*, 100323. https://doi.org/10.1016/j.alcr.2019.100323

Vermunt, J. K. (2010a). Longitudinal Research Using Mixture Models. In K. Montfort, J. H. Oud, & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 119–152). Springer. https://doi.org/10.1007/978-3-642-11760-2_4

Vermunt, J. K. (2010b). Longitudinal Research Using Mixture Models. In K. Montfort, J. H. Oud, & A. Satorra (Eds.), *Longitudinal Research with Latent Variables* (pp. 119–152). Springer. https://doi.org/10.1007/978-3-642-11760-2_4

INHOUDSTAFEL

MIXTURE HIDDEN MARKOV MODELLEN IN SOCIAAL-WETENSCHAPPELIJK ONDERZOEK