Arbeidsloopbanen van uitstromers uit de bijstand naar werk in regio Rijnmond

Een onderzoek op basis van sequentieanalyse

Onderzoek en Business Intelligence

Gemeente Rotterdam

Colofon

Gemeente Rotterdam, afdeling Onderzoek en Business Intelligence (OBI)

Datum: 28 januari 2019
Auteur(s): Ludo van Dun, Marco Lips, mmv Cuneyt Ergun en Paul van der Aa

Vragen: onderzoek@rotterdam.nl

Samenvatting

Om inzicht te krijgen in factoren die van invloed zijn op patronen in de arbeidsloopbaan van uitstromers naar werk uit de bijstand, hebben we gebruik gemaakt van sequentieanalyse.

Sequentieanalyse is een techniek om (bijvoorbeeld) levenslopen te beschrijven en te analyseren; in ons geval gaat het over de arbeidsloopbaan nadat men uit de bijstand is gestroomd door het aanvaarden van werk. Van iedere persoon uit de onderzoeksgroep is gedurende 7 jaar maandelijks vastgesteld wat zijn arbeidsmarktpositie is: de hele rij van vastgestelde arbeidsmarktposities noemen we een *sequentie*. De sequentieanalyses worden uitgevoerd op de hele verzameling sequenties van alle onderzoekspersonen.

De onderzoeksgroep bestaat uit ruim 4.600 personen. Deze personen, tussen de 15 en 58 jaar en wonend in de arbeidsmarktregio Rijnmond, hebben in 2009 de bijstand verlaten wegens het aanvaarden van werk.

Arbeidsmarktposities
Om de loopbanen na uitstroom in kaart te brengen, hebben we een 9-tal arbeidsmarktposities onderscheiden:

1. Bijstand
2. Werk met aanvullende bijstand
3. Studerend
4. Studerend met werk
5. Zelfstandige
6. Werkende met laag inkomen
7. Werkende met middel inkomen
8. Werkende met hoog inkomen
9. Overig: pensioen, uitkeringen, overig actief

Clusteranalyse
De clusters zijn gevormd op basis van gelijkenis tussen sequenties. In een cluster zitten onderzoekspersonen bij elkaar die gedurende een (groot) deel van de onderzoeksperiode in dezelfde arbeidsmarktpositie hebben verkeerd.

De loopbanen na uitstroom hebben we teruggebracht tot vijf clusters:

- Cluster 1: werkenden met ontwikkeling, aandeel 23 procent
- Cluster 2: na periode werken terug in bijstand, aandeel 35 procent
- Cluster 3: werkenden zonder ontwikkeling, aandeel 18 procent
- Cluster 4: werkenden die afwisselend uitkeringen en bijstand ontvangen, aandeel 18 procent
- Cluster 5: zelfstandigen, aandeel 5 procent

Samenhang cluster met achtergrondkenmerken
Het ‘lidmaatschap’ van een cluster hebben we met achtergrondkenmerken sekse en leeftijd in verband gebracht.

Cluster 1: werkenden met ontwikkeling
Vrouwen komen vaker dan mannen en jongeren vaker dan ouderen in dit cluster.
Aan het eind van de onderzoeksperiode is bijna 80 procent uit dit cluster ook daadwerkelijk aan het werk.

Cluster 2: na periode werken terug in bijstand
Mannen keren vaker dan vrouwen en ouderen vaker dan jongeren na verloop van tijd terug in de bijstand.
Uit dit cluster zit op het einde van de onderzoeksperiode 86 procent daadwerkelijk in de bijstand.

Cluster 3: werkenden zonder ontwikkeling
Vrouwen komen vaker dan mannen in dit cluster terecht.
Leeftijd speelt hier een minder grote rol.
Van dit cluster is aan het eind van de 7-jarige onderzoeksperiode bijna 80 procent aan het werk.

Cluster 4: werkenden die afwisselend uitkeringen en bijstand ontvangen
Hier zien we geen verschillen tussen mannen en vrouwen. Wel komen ouderen eerder voor in dit cluster dan jongeren.
Uit dit cluster is op het einde van de onderzoeksperiode bijna een derde aan het werk (en daarvan werkt 12 procent met aanvullende bijstand), ruim een derde heeft pensioen of uitkering, en ruim een kwart zit in de bijstand.

Cluster 5: zelfstandigen
Het zijn vaker mannen dan vrouwen die voor zichzelf beginnen.
Uit het laatste cluster ten slotte is twee derde actief als zzp’er, en is daarnaast nog 15 procent aan het werk.

Arbeidsmarktpositie aan het einde van de onderzoeksperiode
Van de onderzoeksgroep van ruim 4.600 uitstromers uit de bijstand in de arbeidsmarktregio Rotterdam Rijnmond

in 2009, die we gedurende 7 jaar hebben gevolgd, is bijna de helft (47 procent of bijna 2.200 personen) aan het eind van de onderzoeksperiode aan het werk. De rest is niet actief: 36 procent is teruggekeerd in de bijstand, 17 procent zit niet in de bijstand maar is ook niet actief op de arbeidsmarkt (heeft bijvoorbeeld een pensioen).

Deskundigheidsbevordering

Het onderzoek is ook een traject geweest waarin een methodiek voor analyse van transities zou worden ontwikkeld die in vervolgonderzoeken voor andere subgroepen kan worden toegepast.

Het onderzoek heeft een procedure opgeleverd die het mogelijk maakt snel een sequentieanalyse uit te voeren op willekeurig welke groep werkzoekenden (onderscheiden naar sekse, leeftijd, opleiding), naar sector of regio.

Inhoudsopgave

1 Inleiding 8

1.1 Aanleiding 8

1.2 Onderzoeksvragen 8

1.3 Onderzoeksgroep 9

1.4 Aanpak dataverzameling en analyse 9

1.5 Leeswijzer 9

2 Analyseresultaten 10

2.1 Beschrijving sequentiedata 10

2.2 Clusteren 15

2.3 Multinomiale logistische regressieanalyse 19

3 Conclusies 22

Bijlage – uitleg sequentieanalyse 25

Introductie tot sequentieanalyse 25

Verkennen en clusteren van sequenties 27

Analyseren van clusters en sociale factoren 31

Gebruikte R-functies 34

Literatuur 35


samskip
DELFTRANS

1 Inleiding

1.1 Aanleiding

OBI heeft eind 2017 de tweede editie van de regionale arbeidsmarktanalyse Rijnmond gepubliceerd. Deze analyse beschrijft op basis van CBS-microgegevens en UWV-data-ontwikkelingen in vraag, aanbod en de verbinding tussen beide op de regionale arbeidsmarkt. Naast deze globale analyse worden specifieke thema's in aparte rapporten uitgediept. In de editie 2015 van de analyse zijn transities op de arbeidsmarkt onderzocht door SEOR onder de noemer 'dynamiek op de arbeidsmarkt'.

De begeleidingscommissie bij de arbeidsmarktanalyses die OBI uitvoert heeft verzocht om een verdiepingsanalyse van transitiepatronen op de arbeidsmarkt in de regio Rijnmond. Flexibilisering en economische herstructurering hebben tot gevolg dat steeds meer inwoners in de regio krijgen te maken met periodieke, deels noodgedwongen veranderingen in hun arbeidsmarktpositie. Tegelijkertijd verandert de kwalitatieve en kwantitatieve vraag van werkgevers in veel sectoren. Er is nog weinig inzicht in de vraag of en hoe transitiepatronen en veranderende vraag bijdragen aan nieuwe evenwichten op de regionale arbeidsmarkt en welke gevolgen dit heeft voor verschillende groepen werknemers, werkzoekenden en werkgevers.

Besloten is tot een verkennend onderzoek naar deze thematiek op basis van analyse van micro-data van CBS over arbeidsmarktposities door de tijd heen van delen van de regionale beroepsbevolking. Dit verkennend onderzoek richt zich op twee specifieke groepen: uitstromers naar werk uit de bijstand en werknemers die de financiële sector hebben verlaten. In het onderzoek is een methodiek voor analyse van transities ontwikkeld die in vervolgonderzoeken voor andere subgroepen kan worden toegepast.

Inzicht in het verloop van loopbanen van uitstromers uit de bijstand is van belang om beter te kunnen beoordelen welke mogelijkheden (bepaalde) groepen in de bijstand hebben om structureel op de arbeidsmarkt te re-integreren. Beschikbare analyses van het bijstandsbestand bieden hier geen zicht op. Over deze groep wordt in dit rapport verslag gedaan.

Over het onderzoek naar de financiële sector zijn separaat twee rapporten verschenen[1]. De kwantitatieve analyse van transities in de financiële sector wordt gecomplementeerd met een kwalitatief onderzoek over aanpassingen in deze sector onder werkgevers, beleidsadviseurs en onderwijsinstellingen.

Dit rapport is het verslag van het verdiepingsonderzoek naar werkzoekenden die in 2009 uit de bijstand naar werk zijn gestroomd. Met gebruikmaking van *sequentieanalyse* wordt een beeld geschetst van de arbeidsmarktposities die werkzoekenden, die in 2009 vanuit de bijstand naar werk zijn uitgestroomd, na hun uitstroom hebben ingenomen. Met behulp van sequentieanalyse worden de meest voorkomende patronen visueel in kaart gebracht, en wordt gekeken naar samenhang tussen patronen en enkele persoonskenmerken. Het rapport is mede gebaseerd op de afstudeerscriptie die Marco Lips voor het afronden van zijn studie Toegepaste Wiskunde aan de Haagse Hogeschool, locatie Delft, heeft geschreven.[2] Bovendien bevat de bijlage een uitgebreide beschrijving van het gebruik van sequentieanalyse, eveneens afkomstig uit de afstudeerscriptie.

1.2 Onderzoeksvragen

1. *Welke sequenties van arbeidsmarktposities doorlopen de onderzoekspersonen tussen het begin en einde van de onderzoeksperiode?*

[1] *De veranderende vraag naar arbeid in de bancaire sector in Rotterdam. Een kwalitatief onderzoek*, Gemeente Rotterdam, afdeling Onderzoek en Business Intelligence (OBI), augustus 2018. *Arbeidsloopbanen van uitstromers uit de financiële sector in regio Rijnmond. Een onderzoek op basis van sequentieanalyse,* Gemeente Rotterdam, afdeling Onderzoek en Business Intelligence (OBI), december 2018.

[2] *Arbeidsontwikkelingen na een bijstandsuitkering. Een sequentieanalyse van bijstandsgerechtigden in regio Rijnmond*, september 2018.

2. *In hoeverre zijn er clusters van vergelijkbare sequenties van arbeidsmarktposities te onderscheiden?*
3. *In hoeverre hangen gevonden clusters samen met persoonskenmerken?*
4. *In hoeverre hangen gevonden clusters samen met de arbeidsmarktpositie aan het einde van de onderzoeksperiode?*

1.3 Onderzoeksgroep

De onderzoeksgroep wordt gevormd door inwoners van de arbeidsmarktregio Rijnmond die:

- in de loop van het jaar 2009 vanuit de bijstand naar werk zijn gestroomd, en
- bij uitstroom tussen 15 en 58 jaar oud waren

Dat levert een onderzoeksgroep op van ruim 4.600 personen, die we gedurende 7 jaar (of 84 maanden) hebben gevolgd, waarbij we per maand hun arbeidsmarktpositie hebben vastgesteld.

1.4 Aanpak dataverzameling en analyse

Het verdiepingsonderzoek is uitgevoerd op microdatabestanden van het CBS met gegevens van alle Nederlanders en banen (sociaal statistisch bestand, SSB). Voor het koppelen van bestanden en het analyseren van de sequentiegegevens, hebben we gebruikgemaakt van het open source programma *R (en in het bijzonder het package TraMineR)*.

Op basis van de hierboven onder 1.3 genoemde criteria hebben we de relevante microdatabestanden – banen, sector van banen, standplaats van banen, kenmerken personen (geslacht, leeftijd, woonplaats, huishoudenspositie, opleidingsniveau bij uitstroom) – gekoppeld en in de voor sequentieanalyse benodigde vorm georganiseerd.

Voor de sequentieanalyse hebben we van alle onderzoekspersonen over een periode van 84 maanden na hun uitstroom uit de bijstand, bepaald wat op de eerste van elke maand hun arbeidsmarktpositie is. We zijn tot de volgende arbeidsmarktposities gekomen:

1. Bijstand (BY)
2. Werk met aanvullende bijstand (WBY)
3. Studerend (ST)
4. Studerend met werk (STW)
5. Zelfstandige (ZE)
6. Werkende met laag inkomen (WL)
7. Werkende met middel inkomen (WM)
8. Werkende met hoog inkomen (WH)
9. Overig: pensioen, uitkeringen, overig actief (O)

De arbeidsmarktposities ‘werkende met laag/middel/hoog inkomen’ (WL, WM en WH), zijn geconstrueerd door op elk peilmoment het basisloon in een van drie klassen te verdelen. De tertiaire klasse waarin deze valt is berekend op basis van het maximum en minimum van het basisloon van alle werkenden op dat peilmoment.

Niet van alle onderzoekspersonen is op elk peilmoment (eerste van de maand) zijn/haar arbeidsmarktpositie bekend. Het kan voorkomen dat iemand niet in het GBA staat ingeschreven en er dus niets bekend is over zijn/haar arbeidsmarktpositie (bijvoorbeeld wegens overlijden of emigratie). Wij hebben de vuistregel gehanteerd, dat van elke onderzoekspersoon maximaal een derde van het aantal peilmomenten onbekend mag zijn.

Naast de arbeidsmarktpositie, zijn ook geslacht, leeftijd (op moment van uitstroom) en huishoudenstype bekend. Opleiding is voor een gedeelte (vooral van jongeren) van de onderzoeksgroep bekend (van bijna twee derde).

1.5 Leeswijzer

Het rapport is als volgt opgebouwd: na het inleidende hoofdstuk, worden in hoofdstuk 2 de analyseresultaten besproken. In hoofdstuk 3 zetten we de antwoorden op de onderzoeksvragen op een rijtje en presenteren we conclusies. In de bijlage meer informatie over het gebruik van sequentieanalyse, de gebruikte *R-functies* en een overzicht van geraadpleegde literatuur.

2 Analyseresultaten

In dit hoofdstuk gaan we in op de resultaten van de sequentieanalyse van de arbeidsmarktposities van uitstromers uit de bijstand. In een kader geven we beknopt uitleg over de gehanteerde onderzoeksmethode.

2.1 Beschrijving sequentiedata

We beginnen met het beschrijven van het geheel van alle sequenties (dus van alle uitstromers). Dat doen we aan de hand van vier *plots*:

- Een distributieplot: hoe is de verdeling van de arbeidsmarktposities op elk meetmoment? Welke arbeidsmarktpositie komt op elk moment het vaakst voor? Een distributieplot (of verdelingsplot) laat per *state* de fractie zien die die state op ieder meetmoment uitmaakt van alle states. Zo is in één oogopslag duidelijk welke state domineert op elk meetmoment
- Een indexplot waarin de sequenties van alle onderzoekspersonen, wel of niet geordend, worden getoond (niet opgenomen vanwege onthullingsgevaar). Een indexplot stapelt alle sequenties op elkaar waardoor patronen in de sequenties goed te zien zijn
- Een frequentieplot: welke sequenties komen het vaakst voor?
- Gemiddelde tijd in een arbeidsmarktpositie (*state*)

Uit het distributieplot (figuur 2) valt op te maken dat ruim 45 procent aan de slag is gegaan in een baan met een relatief middel inkomen (groen), en een kwart in een baan met een relatief laag inkomen (oranje). Kleine aandelen zijn in een baan met een relatief hoog inkomen (roze, 14 procent) terecht gekomen, zijn gaan studeren naast een baan (8 procent) of zijn als zzp'er aan de slag gegaan (3 procent). Ook valt op dat er na pakweg 3 jaar niet veel meer verandert in de verdeling van de arbeidsmarktposities.

In het frequentieplot (figuur 3) is te zien welke sequenties het vaakst voor komen. Van de ruim 4.600 sequenties zijn er ruim 4.300 uniek, die komen dus maar één keer voor. De meest voorkomende sequentie is *WH, 84* met 51 keer (1,1 procent van alle sequenties). De sequentie *WH, 84*

Wat is sequentieanalyse?

Veel gebruikt in de moleculaire biologie (DNA-analyse), is sequentieanalyse in de sociale wetenschappen geïntroduceerd door Andrew Abbott in de jaren tachtig van de vorige eeuw. Uitgangspunt is een reeks (*sequence*) in de tijd geordende waarden (*states*) van een bepaalde variabele, in ons geval een reeks opeenvolgende arbeidsmarktposities van een uitstromer uit de financiele sector. De eenheid van analyse is een hele sequentie, niet één element uit de sequentie.

Doel van sequentieanalyse:

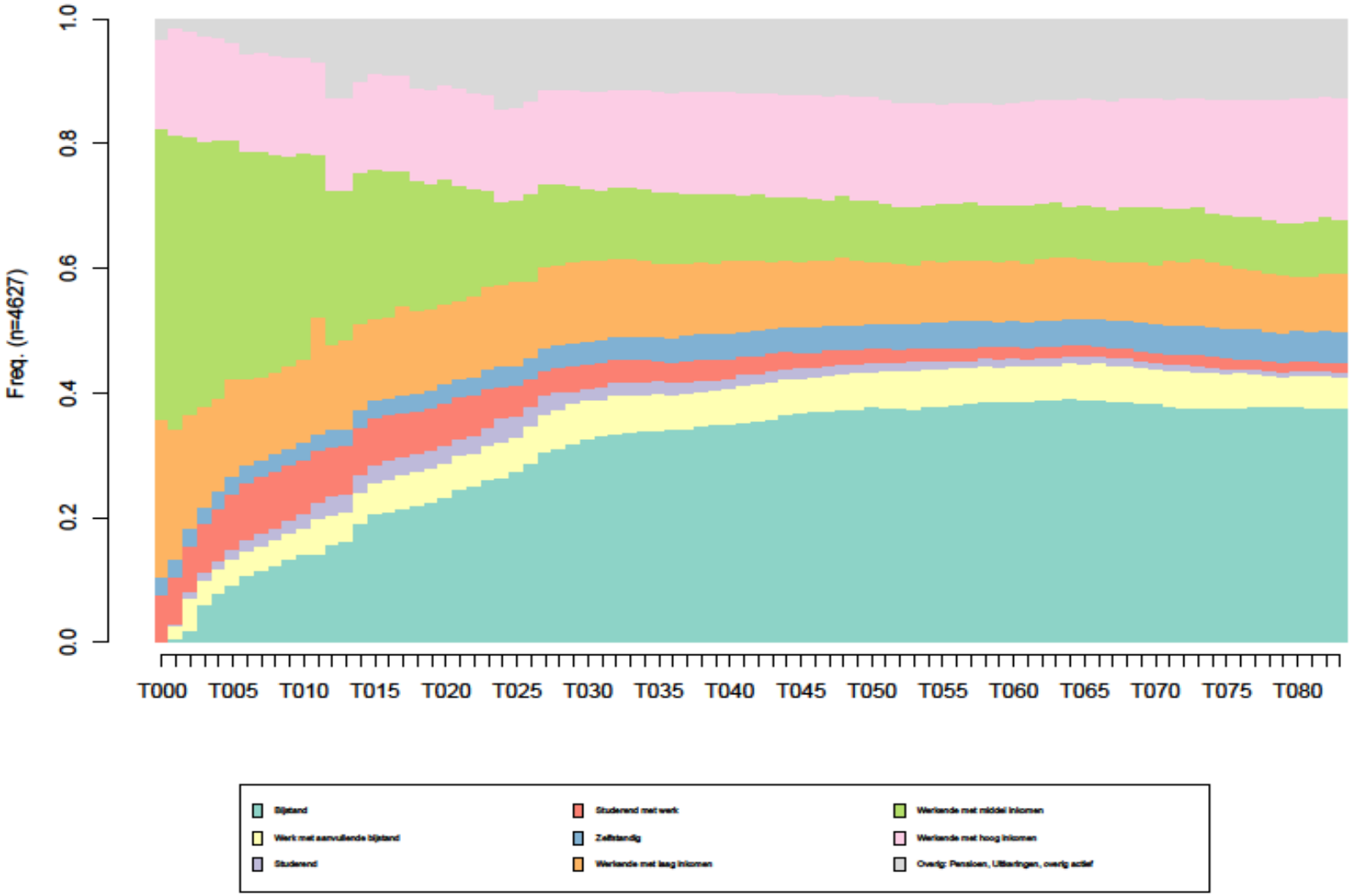
- Hele groep sequenties beschrijven
- Typische sequentiepatronen identificeren
- Samenhangen tussen patronen en individuele kenmerken opsporen

Centraal bij sequentieanalyse is het bepalen van de *afstandenmatrix*. Dat is een matrix waarin de afstand tussen sequenties is bepaald: hoe verschillend zijn sequenties? De meest gebruikte methode om de afstand tussen twee sequenties te bepalen is *Optimal Matching*. Wat zijn de minimale kosten om de ene sequentie te transformeren in de andere?

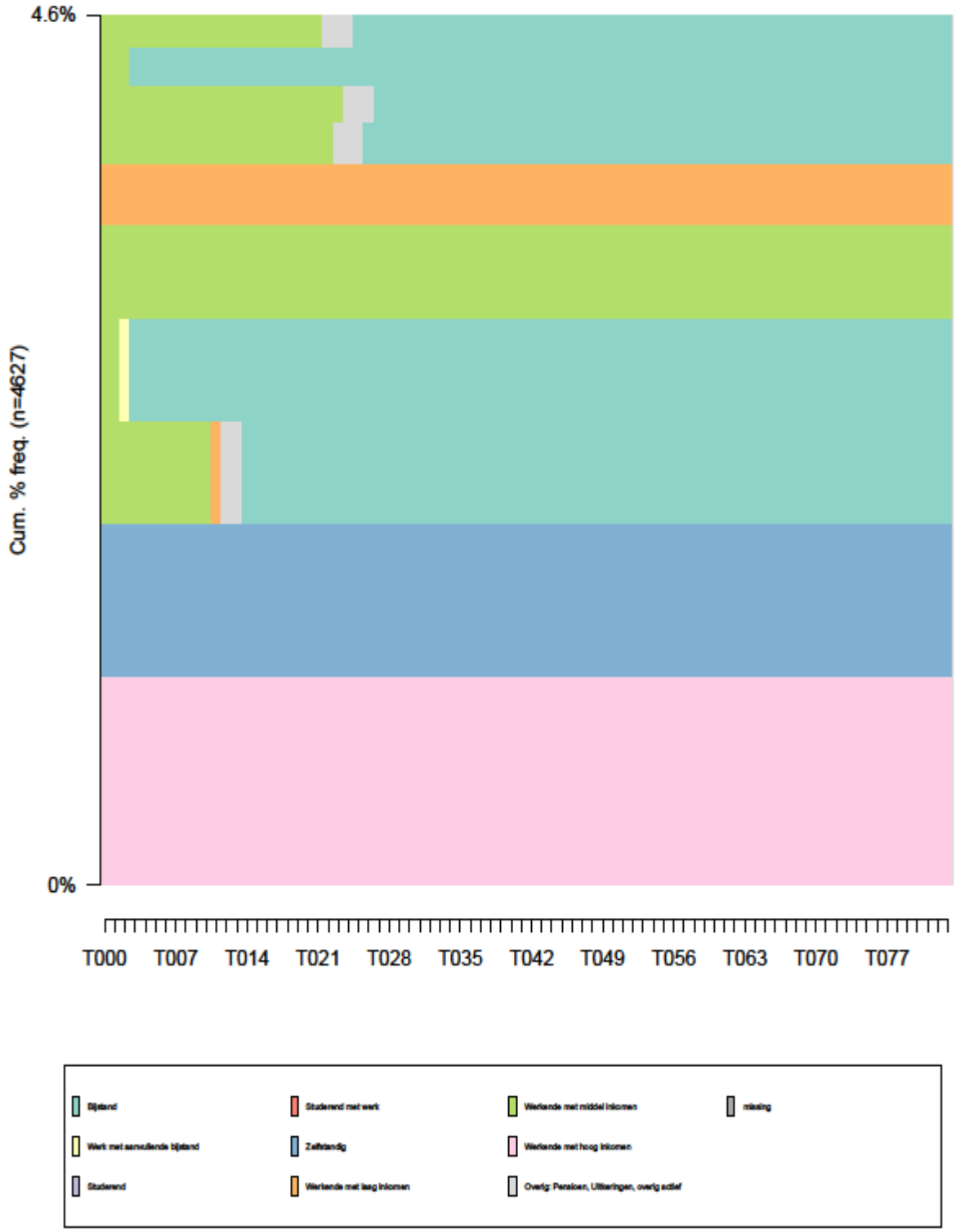
De afstandenmatrix wordt gebruikt om patronen te identificeren, en vormt de input van veel andere analyses binnen de sequentieanalyse (zoals *discrepancy analysis* om *regression trees* te maken).

staat voor iemand die na uitstroom uit de bijstand in een baan met een relatief hoog inkomen aan de slag is gegaan en dat is blijven doen gedurende de hele onderzoeksperiode van 84 maanden. De op een na meest voorkomende sequentie (37 keer of 0,8 procent) is *ZE, 84*. Dat zijn personen die gedurende alle 84 maanden als zelfstandige hebben gewerkt.

Figuur 2 Distributieplot



Figuur 3 Frequentieplot



In tabel 1 staan de tien meest voorkomende sequenties.
Figuur 3 is de visuele weergave van tabel 1.

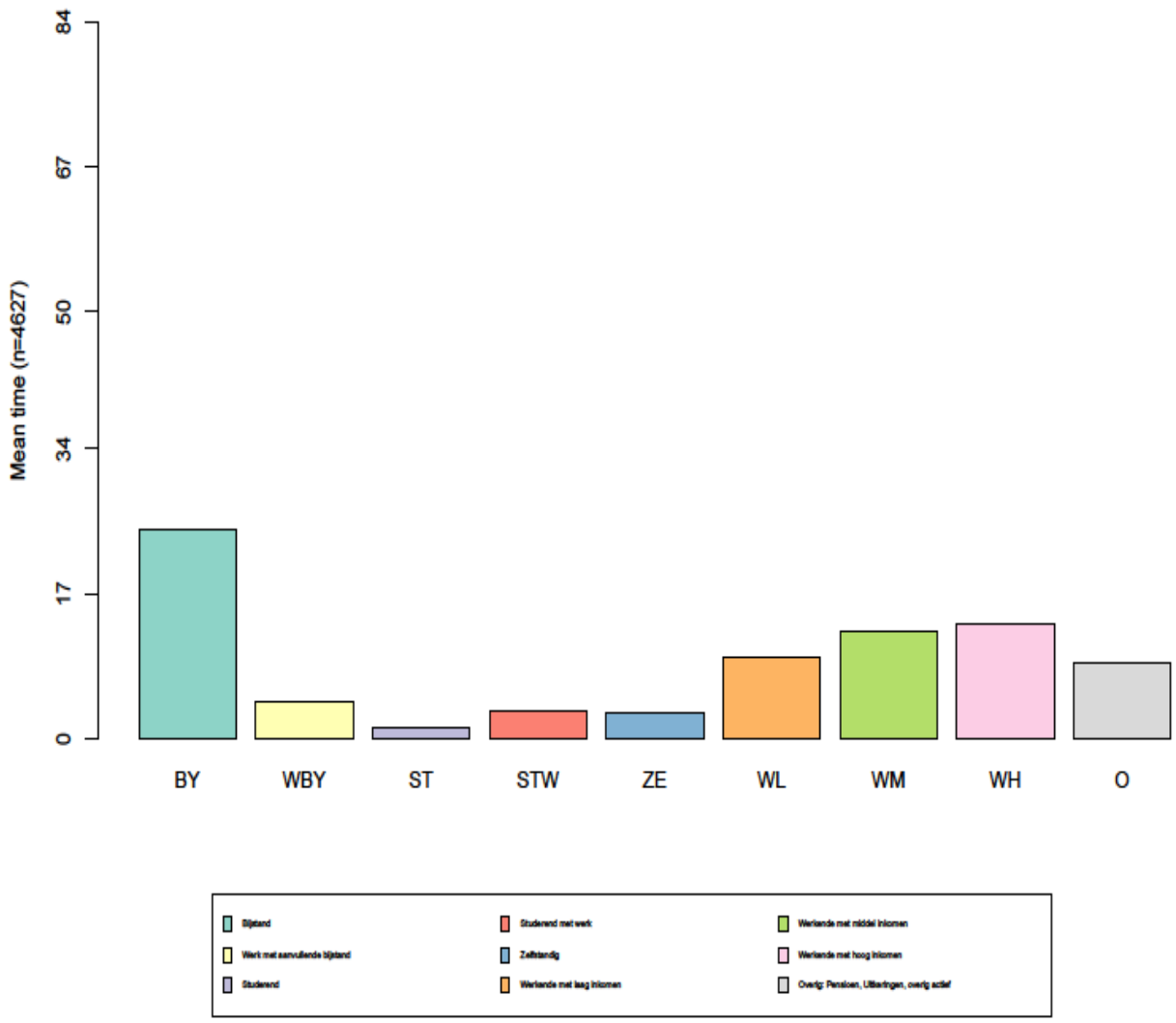
Tabel 1 Tien meest voorkomende sequenties

	Frequentie	%
WH/84	51	1,1
ZE/84	37	0,8
WM/11-WL/1-O/2-BY/70	25	0,5
WM/2-WBY/1-BY/81	25	0,5
WM/84	23	0,5
WL/84	15	0,3
WM/23-O/3-BY/58	10	0,2
WM/24-O/3-BY/57	9	0,2
WM/3-BY/81	9	0,2
WM/22-O/3-BY/59	8	0,2

Toelichting: WH=werkende met hoog inkomen; ZE=zelfstandige; WM=werkende met middel inkomen; WL=werkende met laag inkomen; O=overig; BY=bijstand; WBY=werk met aanvullende bijstand.

De laatste plot die we laten zien (figuur 4) toont wat de gemiddelde tijd is (in maanden) die in een arbeidsmarktpositie is doorgebracht. Het langst (24 maanden) heeft men gemiddeld in de positie ‘bijstand’ (BY) doorgebracht, gevolgd door ‘werkende met hoog inkomen’ (WH, 12 maanden).

Figuur 4 Gemiddelde tijd in een arbeidsmarktpositie (state) doorgebracht

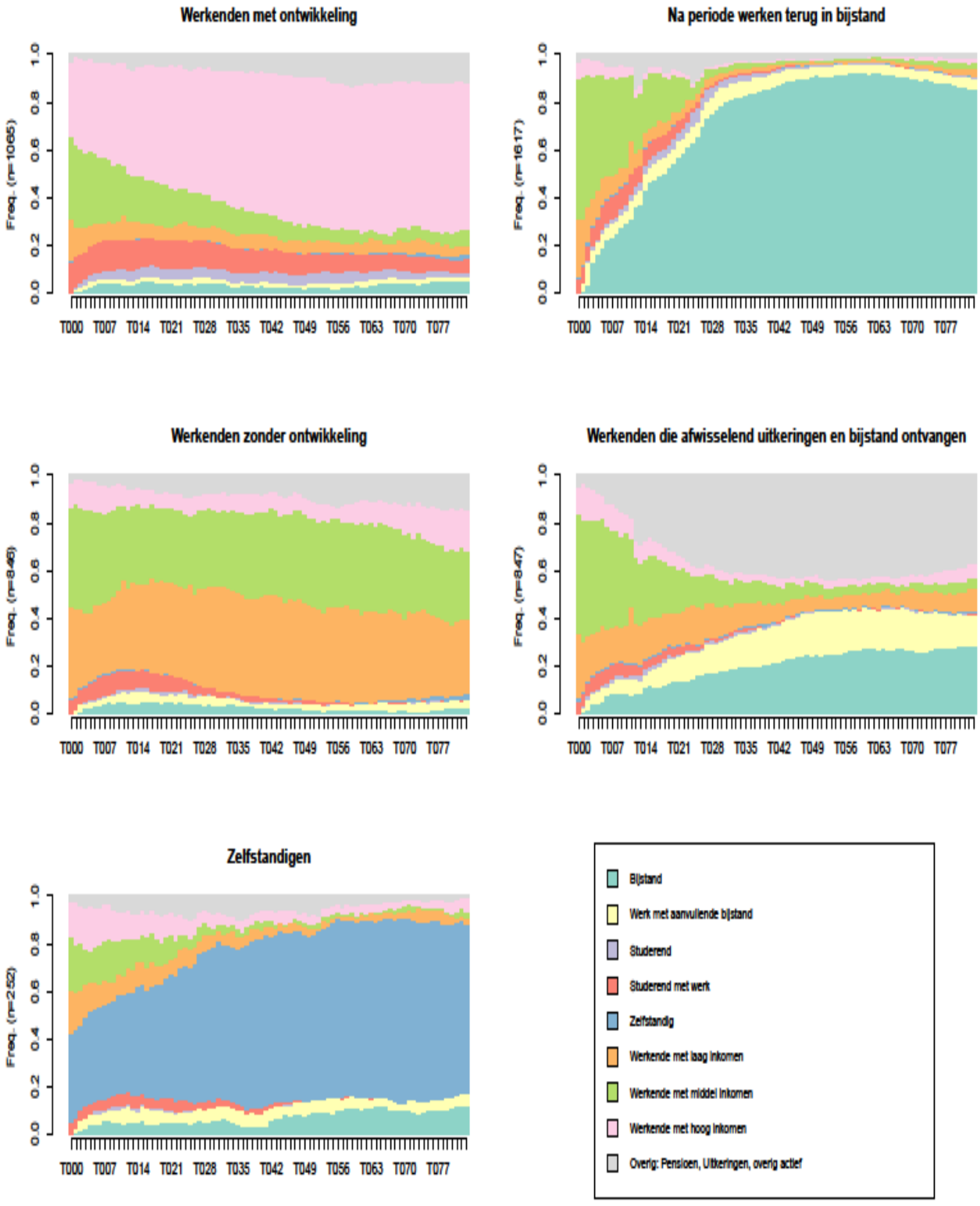


De verschillende visuele plots die we hierboven gebruikt hebben om inzicht te geven in de sequentiedata, zijn een indicatie voor wat we in de volgende paragraaf gaan zien: het proberen te vangen van de sequenties in samenhangende groepen of clusters.

2.2 Clusteren

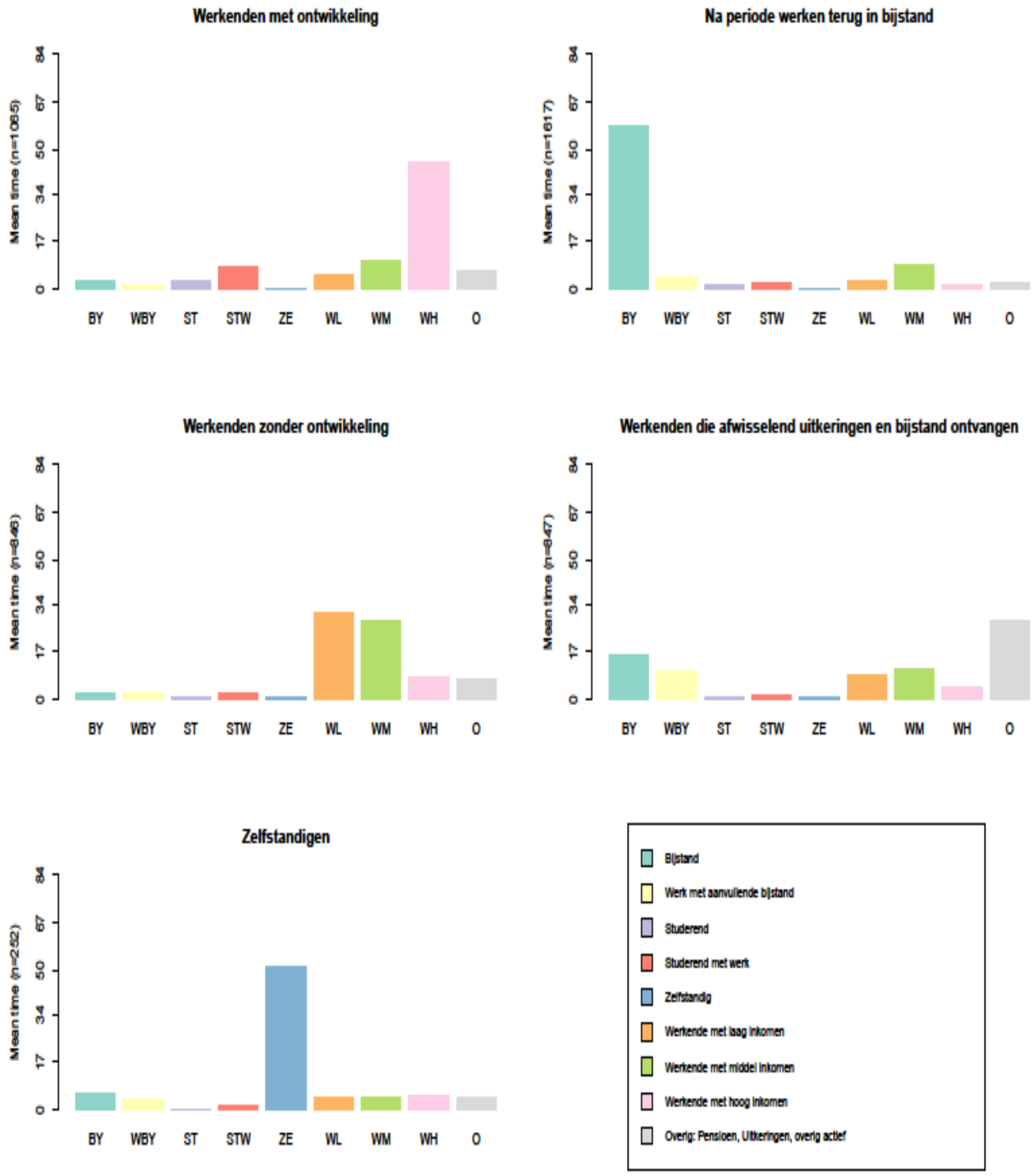
De volgende stap in de sequentieanalyse is het identificeren van (mogelijke) patronen met behulp van clusteranalyse. Bij clusteranalyse wordt onderzocht welke verzamelingen er gevormd kunnen worden in een dataset. Een veel gebruikte manier is *hiërarchisch clusteren*, en dit houdt in dat alle losse sequenties gezien worden als een cluster op zich en er stap voor stap clusters samengevoegd worden die het dichtst bij elkaar liggen. Dit gaat zo door totdat er nog maar één alles overkoepelend cluster over is. Daarna moet bepaald worden hoeveel clusters er nu precies geschikt en inhoudelijk te duiden zijn.

Figuur 5 Distributieplot per cluster



Wij zijn op een 5-clusteroplossing uitgekomen. Voor de interpretatie en naamgeving van de clusters, kijken we naar het distributieplot per cluster, aangevuld met de gemiddelde tijd dat personen in een bepaald cluster in elke *state* (arbeidsmarktpositie) hebben doorgebracht. Zie daarvoor de figuren 5 en 6.

Figuur 6 Gemiddelde tijd in een state doorgebracht, per cluster



Deze overwegingen leiden ertoe dat we tot de volgende (namen van de) clusters zijn gekomen:

- *Cluster 1: werkenden met ontwikkeling*. In dit cluster ontwikkelt men snel een stabiele positie op de arbeidsmarkt, waarbij men vaak na enkele jaren van een relatief laag/middel inkomen doorstroomt naar een baan met een relatief hoog inkomen.
- *Cluster 2: na periode werken terug in bijstand*. In dit cluster komt men vaak binnen 2 jaar weer in de bijstand terecht, om daar niet meer uit te komen.
- *Cluster 3: werkenden zonder ontwikkeling*. Personen in dit cluster vinden geen stabiele arbeidspositie, zij blijven wisselen tussen banen met een relatief laag en een relatief middel inkomen.

- *Cluster 4: werkenden die afwisselend uitkeringen en bijstand ontvangen*. Dit cluster is het lastigst te interpreteren: meest voorkomende state is 'overig', gevolgd door 'bijstand', maar men is ook af en toe aan het werk. Men wisselt dus tussen banen, uitkeringen (waaronder ook pensioen) en bijstand.
- *Cluster 5: zelfstandigen*. De meeste personen in dit cluster zijn overwegend actief als zelfstandige.

In tabel 2 zien we het aandeel van elk cluster in de totale onderzoeksgroep van ruim 4.600. Daarnaast geven we aan hoeveel transities (overgang van de ene naar de andere arbeidsmarktpositie) er gemiddeld per cluster hebben plaatsgevonden.

Het grootste cluster wordt gevormd door uitstromers die na een periode aan het werk te zijn geweest, terugkeren in de bijstand (35 procent). Ook valt op dat na pakweg 3 tot 4 jaar bijna iedereen in dit cluster is teruggekeerd in de bijstand. Bijna een kwart (23 procent) maakt na het aanvaarden van werk een ontwikkeling door, bijna een vijfde (18 procent) maakt geen ontwikkeling door. Ruim de helft (cluster 1 en 4 samen) is, na aanvankelijk aan het werk te zijn gegaan, aan het eind van de onderzoeksperiode overwegend niet aan het werk.

Opvallend is het gemiddelde aantal transities dat de totale onderzoeksgroep doormaakt: 10. Dat betekent over de hele onderzoeksperiode van 7 jaar dus gemiddeld meer dan 1 keer per jaar. Vooral in het cluster 'werkenden zonder ontwikkeling' maakt men gemiddeld veel transities door (bijna 16, dus gemiddeld meer dan 2 keer per jaar).

Tabel 2 Grootte cluster, gemiddeld aantal transities

	Abs	%	Gemiddeld aantal transities
Cluster 1: werkenden met ontwikkeling	1.065	23%	10,1
Cluster 2: na periode werken terug in bijstand	1.617	35%	6,6
Cluster 3: werkenden zonder ontwikkeling	846	18%	15,6
Cluster 4: werkenden die afwisselend bijstand en uitkeringen ontvangen	847	18%	10,7
Cluster 5: zelfstandigen	252	5%	6,6
Totaal	4.627	100%	9,8

We gebruiken de aanduiding 'overwegend' omdat personen in een bepaald cluster de meeste tijd hebben doorgebracht in één (of meer) posities. Zo betekent 'na periode werken terug in de bijstand' dat iedere persoon in dat cluster gemiddeld de meeste tijd heeft vertoefd in de arbeidsmarktpositie 'bijstand' (gemiddeld bijna 60 maanden); het betekent dus niet dat deze persoon de hele tijd in deze dominante positie heeft verkeerd. Het kan dus voorkomen dat hij of zij aan het eind van de onderzoeksperiode niet in de bijstand zit.

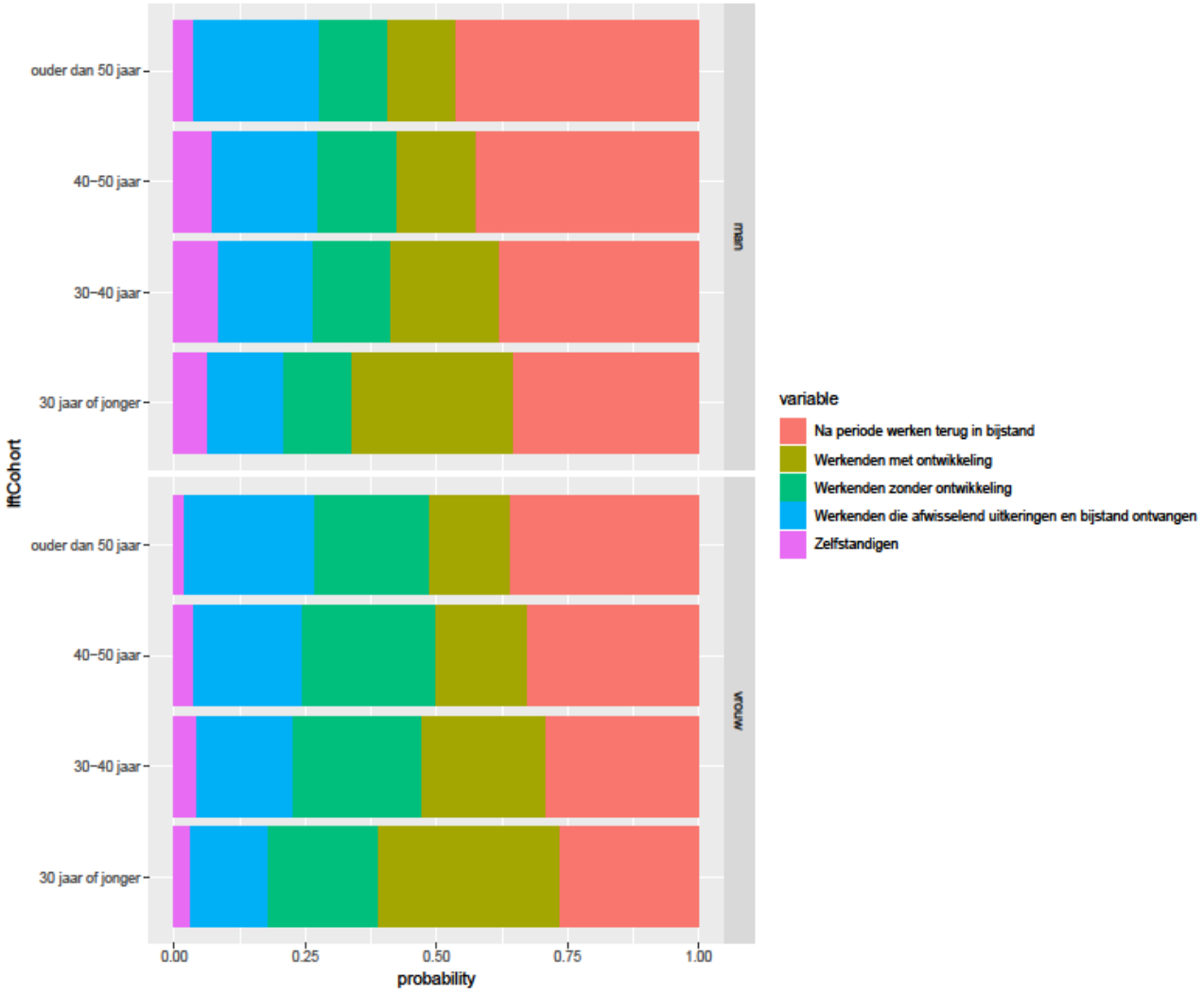
Van de onderzoeksgroep van ruim 4.600 uitstromers uit de bijstand in de arbeidsmarktregio Rotterdam Rijnmond in 2009, die we gedurende 7 jaar hebben gevolgd, is bijna de helft (47 procent of bijna 2.200 personen) aan het eind van de onderzoeksperiode aan het werk. De rest is niet actief: 36 procent is teruggekeerd in de bijstand, 17 procent zit niet in de bijstand maar is ook niet actief op de arbeidsmarkt aan het eind van de onderzoeksperiode (heeft bijvoorbeeld een pensioen). Was het aandeel werkende met laag inkomen aan het begin van de onderzoeksperiode nog 25 procent, aan het eind is dat 9 procent. Voor het aandeel werkende met middel inkomen geldt deze afname

in nog sterkere mate: van 46 naar 8 procent. Het aandeel werkende met hoog inkomen laat een stijging zien, van 14 naar 19 procent.

2.3 Multinomiale logistische regressieanalyse

Met behulp van multinomiale logistische regressieanalyse, zijn we nagegaan of er een relatie is tussen enkele achtergrondkenmerken en het clusterlidmaatschap. We hebben gekeken hoe geslacht en leeftijd van invloed zijn op het behoren tot een bepaald cluster.

Figuur 7 Waarschijnlijkheid om in een bepaald cluster te komen, naar sekse en leeftijd



In de figuur is te zien hoe waarschijnlijk het is om, afhankelijk van sekse en leeftijd, terecht te komen in een van de vijf onderscheiden clusters.

Kijken we naar het cluster *'werkenden met ontwikkeling'*, dan valt op dat vrouwen vaker dan mannen en jongeren vaker dan ouderen in dit cluster terechtkomen. Gemiddeld is de kans 21 procent om in dit cluster terecht te komen; de grootste kans om in dit cluster terecht te komen hebben vrouwen tot 30 jaar (35 procent), gevolgd door mannen tot 30 jaar (31 procent).

Bij het cluster *'na periode werken terug in bijstand'* zien we het tegenovergestelde als bij het eerste cluster: mannen vaker dan vrouwen en ouderen vaker dan jongeren keren na verloop van tijd terug in de bijstand. De gemiddelde kans om in dit cluster te komen is 36 procent. Bijna de helft

van de mannen ouder dan 50 jaar keert terug in de bijstand.

Voor het cluster *'werkenden zonder ontwikkeling'* geldt dat vrouwen vaker dan mannen in dit cluster terechtkomen; leeftijd speelt hier een kleine rol. De gemiddelde kans om in dit cluster te komen is 19 procent. Vrouwen tussen de 30 en 50 jaar oud (gemiddeld 25 procent) komen het vaakst in dit cluster terecht.

Het cluster *'werkenden die afwisselend uitkeringen en bijstand ontvangen'* laat geen verschillen zien tussen mannen en vrouwen. Wel komen ouderen eerder voor in dit cluster dan jongeren. De gemiddelde kans in dit cluster te komen is 19 procent. Die kans is het grootst bij mannen en vrouwen boven de 50 (gemiddeld 24 procent).

In het cluster *'zelfstandigen'* zijn het vaker mannen dan vrouwen die voor zichzelf beginnen. De gemiddelde kans in dit cluster terecht te komen is 5 procent. Bij mannen tussen de 30 en 50 jaar oud, is die kans gemiddeld 8 procent.

Samengevat:

- De kans om *terug te keren in de bijstand* is gemiddeld voor de hele groep 36 procent. Voor vrouwen tot 40 jaar is die kans gemiddeld 28 procent, voor mannen boven de 50 is die kans 46 procent
- Mannen zijn structureel vaker dan vrouwen te vinden in het cluster *'zelfstandigen'*
- Jongeren tot 30 jaar hebben de grootste kans om in het cluster *'werkenden met ontwikkeling'* te komen
- Kans om tussen baan, uitkering en bijstand te pendelen, is het grootst voor mannen en vrouwen boven de 50, met 24 procent

3 Conclusies

Wat kunnen we concluderen uit de hiervoor gepresenteerde resultaten? Dat doen we aan de hand van de onderzoeksvragen.

1. *Welke sequenties van arbeidsmarktposities doorlopen de onderzoekspersonen tussen het begin en einde van de onderzoeksperiode?*

We hebben in het onderzoek een 9-tal arbeidsmarktposities onderscheiden:

1. Bijstand
2. Werk met aanvullende bijstand
3. Studerend
4. Studerend met werk
5. Zelfstandige
6. Werkende met laag inkomen
7. Werkende met middel inkomen
8. Werkende met hoog inkomen
9. Overig: pensioen, uitkeringen, overig actief

2. *In hoeverre zijn er clusters van vergelijkbare sequenties van arbeidsmarktposities te onderscheiden?*

We zijn tot de volgende clusters gekomen:

- Cluster 1: werkenden met ontwikkeling
- Cluster 2: na periode werken terug in bijstand
- Cluster 3: werkenden zonder ontwikkeling
- Cluster 4: werkenden die afwisselend uitkeringen en bijstand ontvangen
- Cluster 5: zelfstandigen

Cluster 1, 2, 3 en 5 zijn duidelijke en sterke clusters, cluster 4 is minder sterk omdat er geen dominante arbeidsmarktpositie in dat cluster is.

3. *In hoeverre hangen gevonden clusters samen met persoonskenmerken?*

Cluster 1: werkenden met ontwikkeling
Vrouwen komen vaker dan mannen en jongeren vaker dan ouderen in dit cluster.

Cluster 2: na periode werken terug in bijstand
Mannen keren vaker dan vrouwen en ouderen vaker dan jongeren na verloop van tijd terug in de bijstand.

Cluster 3: werkenden zonder ontwikkeling
Vrouwen komen vaker dan mannen in dit cluster terecht. Leeftijd speelt hier een minder grote rol.

Cluster 4: werkenden die afwisselend uitkeringen en bijstand ontvangen
Hier zien we geen verschillen tussen mannen en vrouwen. Wel komen ouderen eerder voor in dit cluster dan jongeren.

Cluster 5: zelfstandigen
Het zijn vaker mannen dan vrouwen die voor zichzelf beginnen.

4. *In hoeverre hangen gevonden clusters samen met de arbeidsmarktpositie aan het einde van de onderzoeksperiode?*

De clusters zijn gevormd op basis van gelijkenis tussen sequenties. In een cluster zitten onderzoekspersonen bij elkaar die gedurende een (groot) deel van de onderzoeksperiode in dezelfde arbeidsmarktpositie hebben verkeerd.

Aan het eind van de onderzoeksperiode is bijna 80 procent uit het cluster ‘werkenden met ontwikkeling’ ook daadwerkelijk aan het werk.

Uit het cluster ‘na periode werken terug in bijstand’ zit op het einde van de onderzoeksperiode 86 procent daadwerkelijk in de bijstand.

Van het cluster ‘werkenden zonder ontwikkeling’ is aan het eind van de 7-jarige onderzoeksperiode bijna 80 procent aan het werk.

Uit het cluster ‘werkenden die afwisselend uitkeringen en bijstand ontvangen’ is op het einde van de onderzoeksperiode bijna een derde aan het werk (en daarvan werkt 12 procent met aanvullende bijstand), ruim een derde heeft pensioen of uitkering, en ruim een kwart zit in de bijstand.

Uit het laatste cluster ten slotte ‘zelfstandigen’ is twee derde actief als zzp’er, en is daarnaast nog 15 procent aan het werk.

Vervolgonderzoeken

Naast het beantwoorden van de onderzoeksvragen, is het onderzoek ook een traject geweest waarin een methodiek voor analyse van transities zou worden ontwikkeld die in vervolgonderzoeken voor andere subgroepen kan worden toegepast.

Het onderzoek heeft een procedure opgeleverd die het mogelijk maakt snel een sequentieanalyse uit te voeren op willekeurig welke groep werkzoekenden (onderscheiden naar sekse, leeftijd, opleiding), naar sector of regio.

Bijlage – uitleg sequentieanalyse

Introductie tot sequentieanalyse

In de sociale wetenschappen is sequentieanalyse een steeds populairdere methode geworden om levenslopen te beschrijven en te analyseren. Het proces van sequentieanalyse is in grote lijnen veelal hetzelfde. Eerst wordt er verkennend onderzoek gedaan naar de dataset van sequenties, dan wordt er een keuze gemaakt voor de manier waarop een kostenmatrix samengesteld wordt, waaruit vervolgens een afstandenmatrix van de sequenties geconstrueerd wordt. Aan de hand van deze afstandenmatrix worden de sequenties geclusterd. Vervolgens worden de clusters geïnterpreteerd en eventueel verklaard aan de hand van achtergrondkenmerken. Ook door middel van discrepantieanalyse worden statistische relaties tussen achtergrondkenmerken en sequenties onderzocht. In dit hoofdstuk wordt een introductie gegeven tot sequentieanalyse. Hierin wordt duidelijk gemaakt wat sequenties nu precies zijn, op welke manier er gewerkt wordt met sequenties, welke stappen er gemaakt moeten worden tot er daadwerkelijk analyseresultaten te verkrijgen zijn en op welke momenten er afwegingen door de onderzoeker gemaakt moeten worden.

Wat is een sequentie en wat zijn verschillende formats voor sequenties

Een arbeidsloopbaan van een persoon kan vertaald worden als een sequentie. Een verzameling van deze sequenties kan geanalyseerd worden door middel van sequentieanalyse. Sequentieanalyse bestaat uit twee analysevormen: clusteranalyse en discrepantieanalyse. Bij clusteranalyse wordt onderzocht welke verzamelingen er gevormd kunnen worden in een dataset. Deze verzamelingen kunnen vervolgens verklaard worden door de statistische relatie te onderzoeken met sociale kenmerken (covariabelen) als leeftijd, geslacht of opleidingsniveau. Bij discrepantieanalyse wordt er gekeken of deze covariabelen op zichzelf invloed uitoefenen op de sequentie van een persoon. Discrepantieanalyse geeft antwoord op vragen als bijvoorbeeld: zijn er duidelijke verschillen in structuur tussen sequenties van vrouwen en mannen?

Formats voor sequenties

Een sequentie wordt gedefinieerd door een serie toestanden van een bepaalde variabele. Als er bijvoorbeeld gesproken wordt over de arbeidsloopbaan van een persoon kan die onder andere beschreven worden met de toestanden:

- Middelbaar onderwijs (MO)
- Hoger onderwijs (HO)
- Wetenschappelijk onderwijs (WO)
- Traineeship (TR)
- Werkloos (WL)
- Werkende (WE)

Dit wordt een alfabet genoemd.

Een sequentie met deze toestanden kan dan beschreven worden in verschillende formats. Zo wordt er in de literatuur (Gabadinho, Studer, Ritschard, & Müller, 2010) gesproken over twee formats om een gehele sequentie aan te duiden: State Sequence (STS) en State Permanence Sequence (SPS). Een STS-format is de meest intuïtieve representatie van een sequentie, bijvoorbeeld:

MO-MO-MO-MO-MO-HO-HO-HO-HO-WO-TR-WE-WE

Het State Permanence Sequence format is een soort samengevatte versie van het STS-format zoals:

(MO,5) – (HO,4) – (WO,1) – (TR,1) – (WE,2)

Hierbij geven de getallen de hoeveelheid tijdsintervallen aan die de persoon doormaakt in de bijbehorende toestand. Als het van belang is om naar de verschillende states te kijken en de volgorde hiervan belangrijk is, maar niet hoeveel tijd iemand in de toestand doormaakt, wordt er gekeken naar een Distinct Successive State (DSS) format:

MO-HO-WO-TR-WE

Het proces van een sequentieanalyse

Sequentieanalyse bestaat uit twee analysevormen. Voordat deze twee analyses uitgevoerd kunnen worden zijn er nog verschillende stappen die doorlopen moeten worden. In

deze sectie wordt uiteengezet welke stappen er zijn en welke afwegingen bij elke stap gemaakt moeten worden.

De variëteit tussen sequenties in kaart brengen met een afstandenmatrix

Om beter inzicht te krijgen in eigenschappen en kenmerken van een verzameling van sequenties is het van belang om het verschil tussen sequenties in kaart te brengen. Het doel is om voor een sequentie een aantal waarden te genereren die vastleggen hoeveel deze sequentie lijkt op andere sequenties. Deze waarden, voor alle sequenties, worden vastgelegd in een afstandenmatrix. Hierin geldt hoe hoger de afstand tussen sequentie 1 op rij *x* en sequentie 2 op kolom *y*, hoe minder de twee sequenties op elkaar lijken. Deze afstandenmatrix is het begin van zowel de clusteranalyse als de discrepantieanalyse. Voor het berekenen van een afstandenmatrix moet echter wel bepaald worden hoe deze afstanden berekend worden. De meest gebruikte methode in de literatuur is Optimal Matching (OM) maar Longest Common Subsequence (LCS), Hamming (HAM) en Dynamic Hamming (DHD) zijn ook methoden in de R-package *TraMineR*. Deze methoden zijn echter allemaal gebaseerd op het principe van OM.

Om de afstand tussen twee sequenties te kunnen berekenen is het van belang te weten wat ervoor zorgt dat twee sequenties ver van elkaar liggen. Hierbij wordt gekeken naar de states die op ieder tijdstip van twee sequenties verschillend zijn. Omdat de afstand tussen twee states *niet altijd* voor alle states hetzelfde is (denk bijvoorbeeld aan 'werkloos zijn' is meer verschillend van 'werkend zijn' dan de state 'studeren en werken') wordt een kostenmatrix opgesteld. Hierin wordt bijgehouden hoeveel er bij de afstand tussen sequenties opgeteld moet worden als twee states van elkaar verschillen op een bepaald tijdstip. De kostenmatrix wordt vaak op verschillende wijzen opgesteld. Hierover wordt verderop meer verteld. De totale afstand tussen twee sequenties is dan ook de som van alle kosten tussen twee states op ieder tijdstip van de sequenties.

Clusteranalyse

Nadat vastgesteld is hoeveel sequenties van elkaar verschillen kunnen sequenties geclusterd worden. Hierbij worden sequenties die veel op elkaar lijken ingedeeld in een groep. Dit clusteren kan op 2 manieren gedaan worden: middels hiërarchisch clusteren en *k*-means clusteren (deze twee methoden worden verderop uitgelegd. Door sequenties te clusteren krijgt iedere sequentie een cluster toegekend. Met dit clusterlidmaatschap kan er gemakkelijk gezien worden welke sequenties veel op elkaar lijken en waardoor deze groep sequenties gekenmerkt wordt. Zo is het niet vreemd als een cluster gedomineerd wordt door één state (bijvoorbeeld werkloos). Om meer inzicht te krijgen in de personen die ingedeeld zijn in zo'n cluster kan gekeken worden naar sociale factoren van deze personen. Met verschillende methoden kan onderzocht worden of er een statistische relatie is tussen sociale factoren en het clusterlidmaatschap. Hebben vrouwen een significant grotere kans om in een bepaald cluster terecht te komen? Of speelt leeftijd een rol bij het hebben van een baan met een hoger inkomen? In dit rapport is gebruik gemaakt van multinomiale regressieanalyse.

Discrepantieanalyse

Covariabelen kunnen op zichzelf al flink wat invloed uitoefenen op een sequentie. Door discrepantieanalyse uit te voeren kan onderzocht worden of bijvoorbeeld leeftijd of sekse een significante invloed hebben op de sequentie van een persoon. Dit wordt onderzocht door de dataset op te delen in groepen gebaseerd op de waarden van de covariabelen. Alle mannen worden dus gegroepeerd en vervolgens alle vrouwen. Vervolgens worden statistieken berekend om de mate van ongelijkheid binnen groepen en tussen groepen te berekenen. Deze statistieken worden berekend met behulp van Sum of Squares. Deze Sum of Squares is afhankelijk van de afstanden tussen individuele sequenties. Ook discrepantieanalyse maakt dus gebruik van de afstandenmatrix. Verderop wordt meer uitleg gegeven over hoe discrepantieanalyse in zijn werk gaat.

Verkennen en clusteren van sequenties

De eerste stappen van een sequentieanalyse zijn het voorbereidend werk. Dit houdt in het verkennen van sequenties, het opstellen van een afstandenmatrix en het clusteren van sequenties. Ook kunnen er nog representatieve sequenties geselecteerd worden om een visueel beeld te krijgen van sequenties die zich in een cluster bevinden. Hier wordt het voorbereidend werk besproken dat gedaan moet worden voor het clusteren van sequenties. Daarbij wordt gekeken naar verschillende clustermethoden en technieken om de meest representatieve sequenties per cluster te ontdekken. Verderop wordt besproken hoe verschillende clusters en sequenties verklaard kunnen worden aan de hand van achtergrondkenmerken.

Verkennend onderzoek en het opstellen van een afstandenmatrix

Om de sequentiedata goed te kunnen analyseren is het belangrijk om al wat meer informatie over de data te verzamelen. Een indexplot is een goed begin om de variatie van de verschillende sequenties in beeld te brengen (zie Figuur 2). Een indexplot geeft alle *states* een kleur en stapelt alle sequenties vervolgens op elkaar. Het indexplot van Figuur 2 is gesorteerd op de afstand tot de meest voorkomende sequenties. Dit betekent dat sequenties die niet of nauwelijks groene states bevatten bovenaan komen te staan omdat er in deze dataset kennelijk veel sequenties zijn die een volledig groene sequentie hebben. Tegen de 35 sequenties hebben een volledig groene sequentie.

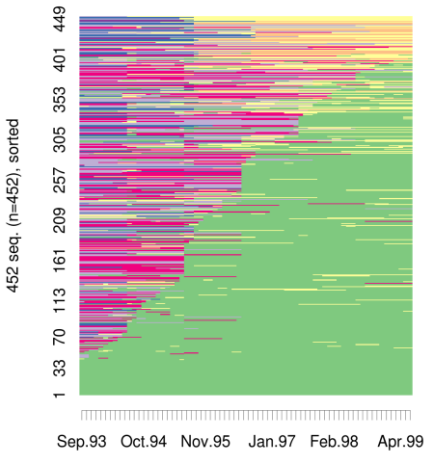
Opstellen van de kostenmatrix

Zoals eerder al is aangegeven, is er een afstandenmatrix nodig om clusters van sequenties te vormen.

Een in de literatuur veel voorkomende methode voor het opstellen van een afstandenmatrix is Optimal Matching[3] (Anyadike-Danes & Michael, 2000). Hierbij worden sequenties met elkaar vergeleken en wordt gekeken hoeveel toestanden er in een sequentie veranderd (gesubstitueerd) of toegevoegd/verwijderd (in de literatuur INDEL-operatie genoemd) moeten worden om twee sequenties gelijk aan elkaar te maken.

Het bepalen van een afstandenmatrix is een discussieonderwerp in het onderzoeksveld van de sequentieanalyse (Aisenbrey & Fasang, 2010). Zo zijn er onderzoekers die aankaarten dat het gebruik van verschillende methoden weinig effect heeft op uiteindelijke resultaten in een onderzoek. Aan de andere kant staan wetenschappers die vanuit een theoretisch oogpunt beargumenteren dat het correct bepalen van een afstandenmatrix een van de fundamenten is van sequentieanalyse.

Aan het substitueren of een INDEL-operatie van toestanden zitten echter wel bepaalde kosten die uiteindelijk de totale afstand tussen de twee sequenties bepalen. Om te bepalen hoeveel het kost om de ene toestand te substitueren met de andere toestand wordt een kostenmatrix opgesteld. In deze matrix wordt precies gedocumenteerd hoeveel het kost om bijvoorbeeld toestand WL uit het voorbeeld in de inleiding te substitueren met WE. Er is geen vaste methode voor het bepalen van de kosten in een kostenmatrix. Dit is een van de grootste kritiekpunten op sequentieanalyse. De



Figuur 1. Voorbeeld van een Indexplot. Bron: (Gabadinho, Studer, Ritschard, & Müller, 2010)

kosten worden namelijk in de ene situatie bepaald vanuit kennis die vergaard is uit de data, terwijl in de andere situatie het logischer is om alle kosten gelijk te stellen aan 2.[4] Om deze keuze te maken moet er nagegaan worden of toestanden op elkaar lijken of niet. Zo kan er beargumenteerd worden dat werkloos (WL) verder van middelbaar onderwijs (MO) ligt dan de state hoger onderwijs (HO). Omdat zowel MO als HO beide gerelateerd zijn aan onderwijs en WL hier niets mee te maken heeft. Dit betekent dat het

[3] Op het idee van OM zijn verschillende andere methoden gebaseerd zoals: LCS (Longest Common Subsequence), HAM (Hamming) en DHD (Dynamic Hamming). Zie documentatie bij de *TraMineR* functie seqdist().

[4] Eén verwijdering van de verkeerde toestand en dan één toevoeging van de juiste toestand, oftewel 2 INDEL-operaties.

substitueren van WL met MO niet dezelfde kosten zou moeten hebben als het substitueren van HO met MO.

Om deze kosten goed te kunnen bepalen kan de data zelf een goede uitkomst bieden. Een kostenmatrix kan bijvoorbeeld gebaseerd worden op een overgangsmatrix (zie Figuur 3). Een overgangsmatrix geeft in percentages aan hoe vaak personen in de data overstappen van state *x* naar state *y* in de gehele dataset (hierbij wordt het aantal overstappen tussen twee states geteld en gedeeld door het totaal aantal overstappen). Als een overstap dan veel voorkomt kan er gekozen worden om deze overstap lagere kosten te geven en andersom hogere kosten voor een overstap die nauwelijks gemaakt wordt.

	[-> EM]	[-> FE]	[-> HE]	[-> JL]	[-> SC]	[-> TR]
[EM ->]	0.99	0.00	0.00	0.01	0.00	0.00
[FE ->]	0.03	0.95	0.01	0.01	0.00	0.00
[HE ->]	0.01	0.00	0.99	0.00	0.00	0.00
[JL ->]	0.04	0.01	0.00	0.94	0.00	0.01
[SC ->]	0.01	0.01	0.02	0.01	0.95	0.00
[TR ->]	0.04	0.00	0.00	0.01	0.00	0.94

Figuur 2. Voorbeeld van een overgangsmatrix. Bron: (Gabadinho, Studer, Ritschard, & Müller, 2010)

In (Anyadike-Danes & Michael, 2000) worden substitutiekosten anders bepaald. Hier worden de toestanden ingedeeld in: geschoold, niet geschoold maar vakbekwaam, niet geschoold maar semi-vakbekwaam en niet vakbekwaam. De toestanden binnen deze groepen kunnen met lage kosten met elkaar gesubstitueerd worden terwijl toestanden tussen groepen met hoge kosten gesubstitueerd kunnen worden. Op deze manier worden kosten bepaald door middel van kennis over de verschillende toestanden. Dit blijft echter altijd subjectief.

Welke manier er ook gebruikt wordt voor het bepalen van de kostenmatrix, het is altijd belangrijk om bijpassende kosten te definiëren voor het toevoegen/verwijderen (INDEL) van states. In (Gauthier, Widmer, Bucher, & Notredame, 2009) worden INDEL kosten bepaald door het gemiddelde van de kostenmatrix te nemen (hierbij wordt de diagonaal van de matrix niet meegenomen in deze berekening omdat deze overal 0 is).

Verschillende clustertechnieken

In de *R-package* Cluster zijn er twee soorten technieken om data te clusteren. Er kan hiërarchisch geclusterd worden of geclusterd worden op basis van de gemiddelde afstand binnen de clusters (*k*-means).

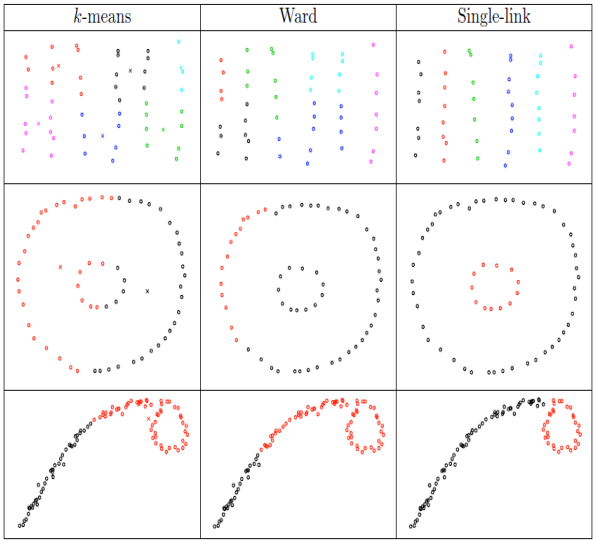
Een populaire manier van clusteren is hiërarchisch, dit betekent dat alle losse sequenties gezien worden als een cluster op zich en er stap voor stap clusters samengevoegd worden die het dichtst bij elkaar liggen. Dit gaat zo door totdat er nog maar één alles overkoepelend cluster over is. Daarna zal er bepaald moeten worden hoeveel clusters er nu precies geschikt en van belang zijn.

Bij het clusteren op basis van de gemiddelde afstand binnen een cluster (*k*-means clusteren) wordt van tevoren bepaald hoeveel clusters er gemaakt zullen worden. Dan worden er een *k* aantal (vooraf gespecificeerd) willekeurige punten gedefinieerd waarbij ieder punt een cluster representeert en worden sequenties vervolgens aan de punten van deze clusters toegewezen. Hierna worden de punten verschoven om de gemiddelde afstand van de punten tot de sequenties uit die clusters zo laag mogelijk te maken. Vervolgens worden deze twee stappen herhaald tot een bepaald criterium.

Hiërarchisch clusteren

Bij hiërarchisch clusteren zijn er verschillende methoden om te bepalen wanneer twee clusters samengevoegd moeten worden. In de package Cluster in R zijn hier 7 methoden voor: gemiddelde, enkele vergelijking, complete vergelijking, Wards methode, gewogen gemiddelde, flexibel en flexibel gemiddelde.

We bespreken de eerste 5 methoden. Bij de methode 'gemiddelde' worden de twee clusters met de kleinste gemiddelde afstand tot elkaar samengevoegd. Het gewogen gemiddelde spreekt vrijwel voor zich, bij het gewogen gemid-



Figuur 3. Voorbeelden van k-means, Ward en enkele vergelijking. Bron: (Tibshirani, 2009)

delde zijn de gemiddelde afstanden gewogen door opgegeven gewichten. Bij een enkele vergelijking wordt slechts gekeken naar het element uit een cluster met de kleinste afstand tot een element uit een ander cluster. Bij een complete vergelijking wordt juist gekeken naar de elementen met de grootste afstand tot elkaar uit de verschillende clusters.

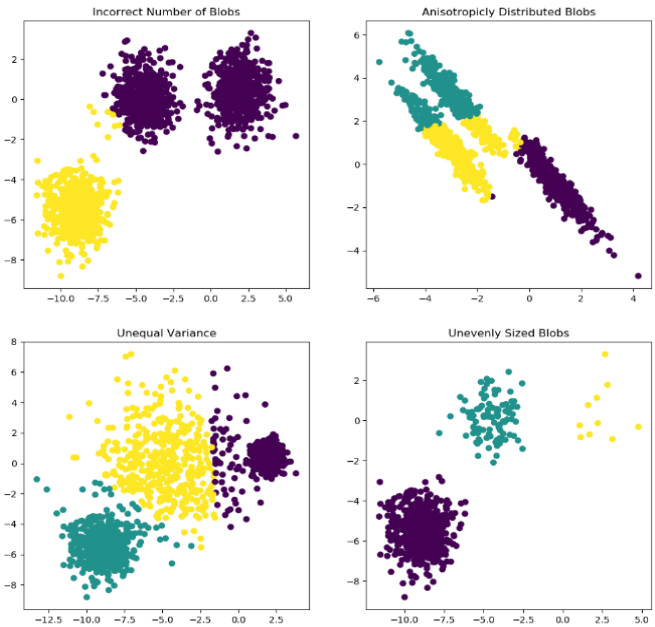
Wards methode is een populaire methode om sequenties te clusteren. Het kijkt naar de toename van de gekwadrateerde gemiddelde afstand binnen clusters als twee clusters samengevoegd zouden worden. Ondanks de populariteit van deze methode hoeft het niet altijd de beste clusteroplossing te bieden. Voorbeelden hiervan zijn te zien in Figuur 4. In Figuur 4 lijkt Single-link (enkele vergelijking) de beste methode, dit hoeft natuurlijk niet altijd zo te zijn. Per dataset zit er een groot verschil in de kwaliteit van de resultaten die een methode oplevert.

K-means clusteren

Het *k-means* clusteren is een algoritme waarbij er een vast aantal clusters bepaald wordt en de data zo goed mogelijk ingedeeld wordt per cluster. Dit houdt in dat de clusters een punt in de ruimte krijgen, vervolgens de data wordt toegewezen aan de cluster waarvan het punt het dichtstbij ligt, waarna het punt van de cluster verplaatst wordt zodat het de kleinste gemiddelde afstand heeft tot de data in dat cluster. Aan *k-means* clusteren zitten een aantal nadelen. Deze zijn gevisualiseerd in Figuur 5. Het eerste plot kan gezien worden als een van de grootste nadelen van de *k-means* clustertechniek. Omdat het lastig is vooraf te bepalen hoeveel clusters er gevormd moeten worden kan eerst hiërarchisch geclusterd worden. Hieruit kan afgeleid worden hoeveel clusters met deze methode het best gebruikt kan worden. Vervolgens kan dit aantal gebruikt worden met het *k*-means clusteren om te kijken of de clusters vrijwel hetzelfde zijn.

Verschillen tussen de twee cluster-methoden

Het grote verschil tussen de voorgaande twee clustermethoden is het moment van specificeren hoeveel clusters er gemaakt gaan worden. Bij hiërarchisch clusteren wordt dit namelijk pas op het laatst gedaan terwijl bij *k*-means dit aan het begin vastgelegd wordt. Aan hiërarchisch clusteren zitten voor- en nadelen. Zo blijft het altijd subjectief hoeveel clusters er uiteindelijk gedefinieerd worden en kunnen clusters niet meer veranderen als zij eenmaal samengevoegd zijn. Daartegenover staat dat er bij hiërarchisch clusteren wel een geschiedenis gedocumenteerd wordt. Zo is gemakkelijk te zien welke clusters samengevoegd worden en hoe deze zich verder ontwikkelen. Dit kan meer informatie opleveren dan er met *k*-means te krijgen is. Want ook *k*-means is niet vrij van nadelen. Zo geeft de *k*-means methode geen geschiedenis van de ontwikkeling van de clusters en is het resultaat afhankelijk van de eerste willekeurige punten die gekozen worden. *k*-means is daarentegen wel snel.



Figuur 4. Een aantal voorbeelden van nadelen van de k-means cluster techniek. Bron: (Gabow, 2007)

Multi-Channel Sequence Analysis

In de sociale wetenschappen is het niet altijd interessant genoeg om te kijken naar slechts één longitudinale variabele. Daarom bedachten (Gauthier, Widmer, Bucher, & Notredame, 2010) in 2010 een methode om meerdere longitudinale variabelen (kanalen of *channels*) te analyseren. De moeilijkheid van dit probleem zit in het bepalen van een afstandenmatrix. Zo kan er een gemiddelde afstandenmatrix berekend worden van de afzonderlijke afstandenmatrices van de verschillende kanalen maar dan wordt ervan uit gegaan dat alle kanalen dezelfde invloed hebben op bijvoorbeeld iemands arbeidsloopbaan. Dit hoeft niet altijd het geval te zijn. Een andere mogelijkheid is het opstellen van een alfabet met een permutatie van alle toestanden in de verschillende kanalen. Dit kan echter erg onoverzichtelijk en lastig worden met het bepalen van een kostenmatrix aangezien er dan states ontstaan als bijvoorbeeld: (werkend en getrouwd), (werkend en gescheiden). Het aantal states dat dan ontstaat is het aantal states van alfabet 1 keer het aantal states van alfabet 2. Daarom gebruiken

(Gauthier, Widmer, Bucher, & Notredame, 2010) een methode die de hiervoor genoemde methodes samenvoegt.

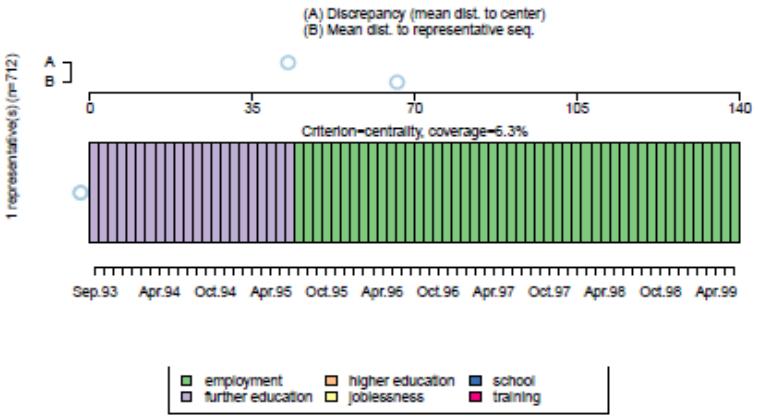
Bij het berekenen van de afstandenmatrix bij multi-channel sequence analysis wordt er naar de afstand van beide kanalen afzonderlijk gekeken. Als twee sequenties met twee kanalen met elkaar vergeleken worden, liggen de sequenties waarvan kanaal 1 wel overeenkomt, maar kanaal 2 niet, minder ver van elkaar dan twee sequenties waarbij beide kanalen verschillen. Hoe ver deze twee sequenties dan van elkaar verwijderd zijn wordt bepaald door de som van de substitutiekosten die voor de states (uit de kanalen die niet overeenkomen) gelden. Op deze manier worden beide longitudinale sequenties meegenomen in de analyse. Maar het is hierbij wel van belang om goede kosten te definiëren aangezien hiermee de relevantie van een kanaal gedefinieerd wordt. Deze methode van multi-channel sequence analysis is echter (nog) niet geïmplementeerd in de *R-package* TraMineR.

Opstellen van representatieve sequenties

Omdat het lastig is een goed beeld te krijgen van de karaktereigenschappen die sequenties binnen een cluster hebben, kan er een of meerdere sequenties per cluster geselecteerd worden. Deze sequenties geven een representatief beeld van de rest van de leden uit dat cluster.

Het proces van het vinden van representatieve sequenties bestaat uit twee stappen (Gabadinho, Ritschard, Studer, & Müller, 2010). Eerst wordt er voor ieder apart te onderscheiden sequentie een zogenoemde “representative score” berekend volgens een gekozen representatiecriterium. Verschillende criteria zijn: nabijheidsdichtheid, centraliteit, frequentie en de waarschijnlijkheid. Hiervan worden de eerste drie bepaald vanuit een afstandenmatrix van de sequenties en wordt de vierde berekend volgens een statistisch model. De nabijheidsdichtheid wordt berekend door het aantal sequenties dat zich binnen een vastgestelde nabijheidsradius bevindt. De nabijheidsradius kan gekozen worden als een fractie van de maximale afstand die twee sequenties kunnen hebben. De sequentie met de hoogste centraliteit wordt gedefinieerd als de sequentie met de kleinste som van de afstanden tot alle andere sequenties. Centraliteit is dus de kleinste som van alle afstanden tot de andere sequenties. Op deze manier wordt er een centraliteit score berekend. Frequentie wordt berekend door te tellen hoe vaak een sequentie voorkomt. De waarschijnlijkheid van een sequentie is lastiger te berekenen. Hierbij wordt er gekeken hoe groot de kans is dat een state uit een sequentie voorkomt op dat tijdstip.

Nadat er een representatiecriterium gekozen is kan de representatiescore berekend worden en kunnen alle sequenties gesorteerd worden op basis van deze score.



Figuur 5. Voorbeeld van een representatieve sequentie met zijn dekkingsgraad op basis van het centraliteitscriterium. Bron: (Gabadinho, Studer, Ritschard, & Müller, 2010)

De volgende stap is het verwijderen van overtollige sequenties. Er wordt begonnen met het selecteren van de eerste kandidaat uit de lijst met sequenties. Vervolgens wordt er per sequentie vanaf bovenaan de lijst gekeken of de sequentie buiten een vooraf bepaalde afstand tot de al gekozen sequentie(s) valt. Als de sequentie buiten de vooraf bepaalde afstand valt wordt de sequentie toegevoegd aan de verzameling representatieve sequenties. Anders wordt de volgende sequentie in de lijst bekeken. Iedere keer dat er een sequentie wordt toegevoegd aan de verzameling representatieve sequenties wordt er een dekkingsgraad berekend van de verzameling. Dit is het percentage sequenties dat een representatieve sequentie in zijn buurt heeft liggen. Deze buurt wordt bepaald door de vooraf bepaalde afstand. Het gehele proces stopt als er een vooraf gedefinieerde grens voor de dekkingsgraad bereikt is. In Figuur 6 is een voorbeeld te zien van een geplotte representatieve sequentie met zijn dekkingsgraad op basis van het centraliteitscriterium.

Analyseren van clusters en sociale factoren

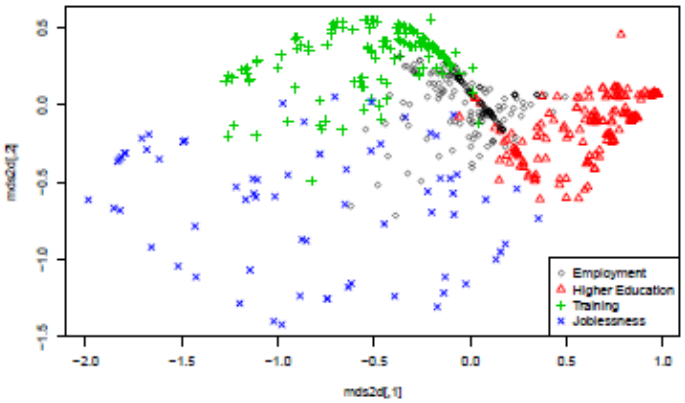
Het verkennen en clusteren van sequenties kan al veel informatie verschaffen maar het zegt nog niet veel over de achtergrond van deze clusters. Wat kenmerkt de sequenties uit een cluster? In sommige gevallen zijn er zelfs duidelijke verbanden te vinden tussen achtergrondvariabelen en het cluster waar sequenties toe behoren. Een bekend voorbeeld uit de literatuur is van jongeren in Ierland waarbij de loopbaan van deze jongeren een sterke relatie heeft met de arbeidsstatus van de vader (Anyadike-Danes & Michael, 2000). In deze sectie wordt besproken hoe clusters gekenmerkt kunnen worden en hoe statistische verbanden met achtergrondvariabelen onderzocht kunnen worden.

Het analyseren van clusters

Voordat clusters geïnterpreteerd kunnen worden is het belangrijk de kwaliteit van een cluster te onderzoeken. Dit kan gedaan worden door naar een verdelingsplot te kijken (zie figuur 2 in het rapport). Een verdelingsplot geeft weer hoeveel procent een state voorkomt op een meetmoment van alle sequenties vergeleken met alle states op een meetmoment. Als er in een verdelingsplot van een cluster een lange periode is waarbij een state een hoog percentage heeft, geeft dit cluster sequenties weer die gedomineerd worden door deze state. Op het moment dat een verdelingsplot geen eenduidig beeld geeft over de sequenties die zich in dat cluster bevinden kan er geconcludeerd worden dat de huidige clusteroplossing niet volstaat en er gekeken moet worden naar een andere clusteroplossing.

Verkennen door middel van multidimensionaal schalen

Als clusters bepaald worden op basis van twee variabelen kunnen deze gemakkelijk geplot worden. De twee variabelen vormen immers de twee dimensies en met kleuren kan het cluster van ieder punt aangeduid worden. Met het visualiseren van clusters uit sequenties is dat een stuk moeilijker omdat iedere sequentie in theorie zijn eigen dimensie is in de afstandenmatrix. Daarom kan multidimensionaal schalen een oplossing bieden. Multidimensionaal schalen brengt het aantal dimensies terug naar een vooraf gespecificeerd aantal. In Figuur 7 is een voorbeeld van multidimensionaal geschaalde sequenties te zien waarbij kleuren de clusterlidmaatschappen van de sequenties aanduiden. Met plots van multidimensionaal geschaalde sequenties, gekleurd op basis van hun clusterlidmaatschap, kan er snel visueel bekeken worden of clusters een logische verdeling hebben en niet door het hele plot verdeeld zijn. Hoewel er in Figuur 7 wel al namen aan de clusters gegeven zijn hoeft dit niet altijd de eerste stap te zijn. De clusters kunnen ook verkend worden zonder deze eerst te interpreteren.



Figuur 6. Voorbeeld van gevisualiseerde, multidimensionaal geschaalde sequenties. Bron: (Gabadinho, Studer, Ritschard, & Müller, 2010)

Het interpreteren van clusters

Om clusters te interpreteren kan er naar verschillende eigenschappen gekeken worden. Een cluster kan gedomineerd worden door sequenties die het grootste gedeelte van de tijd in de state ‘werkend’ zijn. Dit cluster kan dan geïnterpreteerd worden als het ‘werkend’ cluster.

Clusters kunnen echter een stuk uiteenlopender zijn. Een cluster kan bijvoorbeeld gedomineerd worden door sequenties waarbij de persoon eerst een aantal jaar gestudeerd heeft en daarna pas is gaan werken. Om een cluster goed te interpreteren zijn er in de literatuur verschillende methoden gevonden. Zo kan er een verdelingsplot van een cluster gemaakt worden om af te lezen welke state(s) op welk moment veel voorkomen in dit cluster. Een andere optie is het gebruiken van representatieve sequenties (zie hiervoor) per cluster. Deze representatieve sequenties dekken echter niet altijd alle sequenties uit een cluster, waardoor clusters dus niet volledig onderzocht kunnen worden voor het interpreteren van een cluster.

In (Anyadike-Danes & Michael, 2000) wordt er gekeken naar het gemiddeld aantal maanden dat de sequenties uit een cluster zich in een state bevinden. Sequenties uit een cluster kunnen gemiddeld erg lang in de state ‘werkend’ zitten, om deze reden kan dit cluster geïnterpreteerd worden als ‘werkenden’. Zeker als deze gemiddelden vergeleken worden met gemiddelden van de gehele dataset. Welke methode gebruikt wordt om clusters te interpreteren is altijd afhankelijk van de keuze van de onderzoeker en de kwaliteit van zijn beargumentering.

Causaliteit van sociale factoren testen

Nadat clusters geïnterpreteerd zijn kan verklaard worden waar deze verschillende clusters vandaan komen. Er wordt namelijk gedacht dat sociale factoren zoals opleidingsniveau een belangrijke rol spelen bij de ontwikkeling van een arbeidsloopbaan. In deze sectie worden methoden gepresenteerd hoe je zo'n relatie kunt onderzoeken.

(Multinomiale) logistische regressiemodellen

Er kan door middel van (multinomiale) logistische regressiemodellen bepaald worden wat de relatie is tussen covariabelen en clusterlidmaatschap. Met (multinomiale) logistische regressiemodellen kan bepaald worden hoe groot de kans is dat iemand met bepaalde sociale factoren behoort tot een bepaald cluster.

De belangrijkste operatie die plaatsvindt bij een logistisch regressiemodel is het vormen ("fitten") van een logistische kansverdeling op een manier waarbij de data zo goed mogelijk voorspeld wordt door de curve.

Bij het analyseren van de relatie tussen achtergrondkenmerken en clusterlidmaatschap, wordt het clusterlidmaatschap gezien als de responsvariabele. Een selectie van sociale factoren wordt gebruikt om het clusterlidmaatschap te voorspellen. Nadat er een goed model opgesteld is met alleen de causaal relevante variabelen die de data zo goed mogelijk voorspellen kan er nog onderzoek gedaan worden naar de specifieke waarden van achtergrondvariabelen en hun invloed op het clusterlidmaatschap. Heeft een cluster (dat bijvoorbeeld gedomineerd wordt door periodes in de bijstand) een sterkere positieve relatie met een lagere leeftijd? Dit kan onderzocht worden door te kijken naar de grootte van de coëfficiënten in een logistisch regressiemodel en vooral naar de e-macht van de coëfficiënten, ook wel de Odds Ratio genoemd.

Bij de covariabelen wordt er onderscheid gemaakt tussen binaire, categorische en continue variabelen. Het logistisch regressiemodel werkt standaard met binaire variabelen en continue variabelen. De categorische variabelen worden omgezet in meerdere dummyvariabelen waarin per categorie wordt bijgehouden of een persoon op die categorie scoort. Het interpreteren van continue variabelen is het gemakkelijkst. Bij continue variabelen stelt de Odds Ratio de toe/afname in kans per eenheid verschil voor (bijvoorbeeld per jaar dat iemand ouder wordt, wordt de kans dat iemand zich in cluster 1 bevindt met 1.1 (de Odds Ratio) vermenigvuldigd). Dat wil zeggen dat de kans op succes in de responsvariabele evenveel toeneemt bij de overgang van bijvoorbeeld leeftijd 15 naar 16 als van 84 naar 85. Dit geldt alleen als de variabelen voldoen aan de aanname dat deze een lineaire relatie hebben met de responsvariabele (Vach, 2012). Dit lineaire verband kan getest worden door de continue variabele in te delen in categorieën en te kijken of de Odds Ratio van de categorieën een lineair verband vertoont.

Bij binaire of categorische variabelen is het interpreteren lastiger. Hierbij wordt namelijk een referentiecategorie bepaald. De referentiecategorie wordt in veel statistische software aangegeven met de intercept. Het significantieniveau van de dummyvariabelen bij een categorische covariabele laat zien of een bepaalde categorie een significant hogere of lagere kans op succes heeft vergeleken met de intercept (referentiecategorie). Om te testen of een categorische variabele een significante impact heeft in een logistisch regressiemodel moet echter meer onderzoek gedaan worden. Er kan een "likelihood ratio test" uitgevoerd worden. Hierbij wordt een model met en zonder de categorische variabele getest op het verschil met de dataset. Als het model met de categorische variabele een significant lagere deviatie heeft, heeft de variabele een significant effect.

Sequentiekarakteristieken ontdekken door discrepantieanalyse

Optimal Matching hoeft niet alleen voor een clusteranalyse gebruikt te worden. Een afstandenmatrix geeft namelijk op zichzelf al veel informatie over het contrast tussen sequenties. In (Studer, Ritschard, Gabadinho, & Müller, 2011) worden een aantal statistieken gepresenteerd om het contrast tussen sequenties te analyseren. Deze methode wordt ook wel discrepantieanalyse genoemd. Bij discrepantieanalyse wordt er gekeken naar het discriminerend vermogen van een covariabele. Dat wil zeggen dat er gekeken wordt of de ene waarde van een covariabele een duidelijk andere sequentievorm heeft dan de andere waarde.

Om te beginnen wordt de Sum of Squares geïntroduceerd. Dit wordt ook wel gezien als de mate van discrepantie (ongelijkheid) binnen een verzameling (in dit geval sequenties). Sequenties kunnen gegroepeerd worden op basis van de achtergrondvariabelen zoals het geslacht van een persoon. Met groepen worden vanaf nu de verschillende waarden van een covariabele bedoeld. De totale Sum of Squares kan dan beschreven worden als de Sum of Squares tussen groepen plus de Sum of Squares binnen de groepen (de Huygens stelling (Studer, Ritschard, Gabadinho, & Müller, 2011)).

Vanwege deze stelling kunnen ANOVA-methoden toegepast worden op sequentieobjecten. Met deze statistieken kan de discrepantie tussen sequenties verklaard worden door de waarde van de achtergrondvariabelen. Omdat niet beargumenteerd kan worden dat sequenties normaal verdeeld zijn wordt dit gedaan door middel van een permutatie test van de F-statistiek. De waarde van de achtergrondvariabele wordt voor iedere sequentie veranderd naar een willekeurige waarde. Door dit zo vaak mogelijk te doen kan er een F-perm verdeling opgesteld worden. Als deze vergeleken wordt met de F-waarde van de dataset kan de achtergrondvariabele waarbij de F-waarde significant hoger ligt dan de F-perm verdeling gezien worden als een kenmerk dat verstoring veroorzaakt tussen sequenties. 5.000 permutaties zijn gebruikelijk bij een significantieniveau van 1% en 1.000 permutaties voor een significantieniveau van 5%.

De hierboven benoemde methode voor het blootleggen van statistische verbanden tussen covariabelen en sequenties, kijkt slechts naar het effect van één covariabele om de discrepantie tussen sequenties mee te verklaren. Als echter alle covariabelen tegelijk meegenomen worden in een model voor het verklaren van discrepantie tussen sequenties kan dit voor een completer beeld zorgen. Hiermee kunnen ook onderlinge relaties tussen covariabelen ontdekt worden. Voor dit probleem is in (Studer, Ritschard, Gabadinho, & Müller, 2011) multi-factor discrepantieanalyse geïntroduceerd.

Gebruikte R-functies

In deze bijlage worden alle R-functies benoemd die per onderdeel van de sequentieanalyse gebruikt zijn.

Maken van sequenties

seqdef(informat = “STS”)

Deze functie maakt van het onderzoeksbestand een sequentiebestand in het format “STS”.

Maken afstandenmatrix

seqdist()

Een functie voor het opstellen van de afstandenmatrix. Deze functie maakt gebruik van een sequentiebestand en vraagt om een methode, indelkosten en een kostenmatrix of een methode voor het berekenen van een kostenmatrix.

Verkennende plots

seqlplot()

Functie voor het maken van een indexplot.

seqdplot()

Functie voor het maken van een distributieplot.

seqmtplot()

Functie voor het maken van een plot met gemiddeld aantal states.

seqrplot()

Functie voor het vinden en plotten van representatieve sequenties.

cmdscale()

Functie voor het creëren van een multidimensionaal geschaalde matrix van de sequenties. Deze functie maakt gebruik van een afstandenmatrix. De multidimensionaal geschaalde matrix kan geplot worden met de functie plot().

Vinden van clusters

agnes()

Deze functie maakt gebruik van een afstandenmatrix om een dendogram op te stellen. Deze functie vereist een methode. Het object dat uit deze functie komt kan geplot worden met plot().

cutree()

Deze functie maakt een vector met voor iedere sequentie een clusterlidmaatschapnummer.

Logistische regressiemodellen

glm()

Deze functie maakt een logistisch regressiemodel mits: een responsvariabele en een verklarende variabele in de vorm Respons ~ verklarende is opgegeven en de parameter family = binomial(link =logit) is opgegeven.

summary()

Functie om de modelresultaten van de logistische modellen te bekijken.

Discrepantieanalyse

Dissassoc()

Deze functie voert een enkelvoudige discrepantieanalyse uit. Het object uit deze functie kan geprint worden.

dissmfacw()

Deze functie voert een meervoudige discrepantieanalyse uit.

Literatuur

Aisenbrey, S., & Fasang, A. E. (2010). New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. Sociological Methods & Research, 420-462.

Anyadike-Danes, & Michael, D. M. (2000). Predicting Successful and Unsuccessful Transitions from School to Work Using Sequence Methods. Belfast: Northern Ireland Economic Research Centre.

Babyak, M. A. (2004). What You See May Not Be What You Get: A Brief, Nontechnical Introduction. Psychosomatic Medicine, 411-421.

Birchfield, S. T. (2005). Microphone Array Position Calibration by basis-point Classical Multidimensional Scaling. IEEE, 1025-1034.

Gabadinho, A., Ritschard, G., Studer, M., & Müller, N. S. (2010). Extracting and Rendering Representative Sequences. Geneva: Department of Econometrics and Laboratory of Demography, University of Geneva.

Gabadinho, A., Studer, M., Ritschard, G., & Müller, N. S. (2010). Sequence analysis for social scientists. Part III - Describing and visualising sequence data sets (p. 54). Bristol: Department of Econometrics, University of Geneva.

Gabow, H. (2007). Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics Philadelphia.

Gauthier, J.-A., Widmer, E. D., Bucher, P., & Notredame, C. (2009). How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data. Sociological Methods & Research, 197-231.

Gauthier, J.-A., Widmer, E. D., Bucher, P., & Notredame, C. (2010). MULTICHANNEL SEQUENCE ANALYSIS APPLIED TO SOCIAL SCIENCE DATA. American Sociological Association, 34.

Kennis voor een sterk Rotterdams beleid

Onderzoek en Business Intelligence is een afdeling binnen de gemeente Rotterdam. De afdeling verzamelt informatie en doet onderzoek voor het maken en uitvoeren van beleid door de gemeente Rotterdam. Het onderzoek gaat over onderwerpen als gezondheid, zorg, onderwijs, re-integratie, demografie, ruimtelijke ordening en veiligheid. Soms is de gemeentelijke organisatie het onderwerp, vaker gaat het over de stad en haar bewoners. Het doel is steeds om met deze verzamelde kennis het beleid en de bedrijfsvoering van de gemeente te verbeteren.

Auteurs: Ludo van Dun, Marco Lips, mmv Cuneyt Ergun en Paul van der Aa

Gemeente Rotterdam