De Arbeidsmarkt In Beweging

Het vinden van clusters op de Nederlandse arbeidsmarkt

Christian de Groot

Begeleider: Dr. A.J. Koning
Meelezer: Dr. D. Fok

17 juni 2013

Abstract

De arbeidsmarkt, waarschijnlijk de meest bruisende en energieke markt die er is. Iedereen wil graag carrière maken en heeft daar zijn eigen redenen voor. Het kan zijn om je gezin goed te onderhouden of bijvoorbeeld het nastreven van een jeugddroom. Hoe het ook zij, dit zorgt voor een hoop verschuivingen en bewegingen op de arbeidsmarkt. In deze paper is gekeken of er bepaalde clusters van sectoren te vinden zijn waarbij de sectoren binnen zo'n cluster vergelijkbare bewegingen vertonen. Gebruik is gemaakt van een multinomiaal logit model, een clusterstructuur en de schattingsmethode Maximum Likelihood om hier een duidelijk antwoord op te vinden. De resultaten van het onderzoek laten zien dat deze clusters er wel degelijk zijn en dat de sectoren detailhandel en horeca en contractcatering redelijk tot goed vergelijkbaar zijn.

***Key words**: Arbeidsmarkt, Clusters, Maximum Likelihood Estimation, Markov keten*

Inhoudsopgave

1 Introductie

300.000, 507.000, 6,4, zijn op het eerste gezicht zomaar een paar willekeurige getallen. In werkelijkheid weerspiegelen deze getallen de werkloosheidscijfers van de afgelopen jaren in Nederland. Het eerste getal is de werkloze beroepsbevolking in het jaar 2008, het tweede getal is weer de werkloze beroepsbevolking, alleen dan vier jaar later oftewel 2012. Dit is wellicht al een moment waar de schrik vele mensen om het hart slaat. Met name voor zo'n klein land als Nederland is een toename van maar liefst 207.000 werklozen bepaald niet bevorderlijk voor de economische gesteldheid van het land. Het laatste getal is het werkloosheidspercentage in het jaar 2012 . Dit is, ook met het oog op voorgaande jaren (2011=5,4% 2008=3,8%), een forse stijging. Zeker als je het bekijkt in aantallen mensen.

Deze zorgelijke getallen zijn veelal het gevolg van de financiële crisis, ookwel de credit crunch genoemd, die wereldwijd om zich heen begon te grijpen vanaf het jaar 2008. Tot op de dag van vandaag zijn de gevolgen van deze crisis, helaas, nog steeds duidelijk voelbaar. Het consumentenvertrouwen is dan ook vooralsnog bijzonder laag wat betekent dat ook de huishoudens de economische situatie voor nu en de komende tijd, nog bepaald niet rooskleurig inzien.

Naast de groeiende werkloosheid is er voor Nederland nog een ander groot probleem dat opdoemt aan de horizon, namelijk de vergrijzing. Mensen worden steeds ouder, dit met name als gevolg van de medische vooruitgang. De levensverwachting voor vrouwen staat momenteel op 82,6 jaar en voor mannen iets lager, namelijk 78,5 jaar. Dit verschil is hoogstwaarschijnlijk te verklaren uit het feit dat mannen over het algemeen te maken hebben met een hoger stressniveau. Verder zorgt ook het babyboom effect vlak na De Tweede Wereldoorlog, ervoor dat de ratio tussen het aantal ouderen en jongeren opzienbarend hoog is. Dit gecombineerd met het feit dat veel mensen ontkerkelijkt zijn, wat over het algemeen inhoudt dat er minder kinderen geboren worden, heeft tot gevolg dat de grijze druk erg hoog is en naar prognose van het Centraal Bureau voor de Statistiek (CBS), alleen nog maar zal stijgen. Onder grijze druk verstaan we het aantal mensen met de leeftijd van 65 jaar en ouder ten opzichte van het aantal mensen in de leeftijdscategorie 20-64. De verwachting is dat in het jaar 2040 de ouderenpopulatie (65 en ouder) zal zijn gegroeid tot 4,6 miljoen mensen. Dit is, vergeleken met de 2,6 miljoen ouderen van vandaag, ongeveer een verdubbeling. Een gebied waar de vergrijzing het meeste toeslaat is toch wel het onderwijs. Volgens het SBQ (SectorBestuur OnderwijsArbeidsmarkt) is de instroom van mensen die een lerarenopleiding willen volgen de laatste jaren behoorlijk afgenomen wat uiteindelijk zal leiden tot een fors tekort aan leraren. Om enigszins grip op dit probleem te krijgen is een beter inzicht vereist in de stromingen op de arbeidsmarkt. De reden dat de vergrijzing in Nederland zo'n groot probleem is, is omdat hier gebruik gemaakt wordt van het omslagstelsel. Dit stelsel houdt in dat de werkenden van vandaag de pensioenuitkeringen betalen van de inactieven. Echter, door de scheve verhouding tussen het aantal werkenden en het

aantal inactieven wordt het onmogelijk voor de huidige werkende beroepsbevolking om al deze uitkeringen te financieren. De Nederlandse overheid heeft reeds de maatregel getroffen van het verhogen van de AOW-leeftijd (Algemene Ouderdoms Wet) van 65 jaar naar 67 jaar. Of dit het probleem echter oplost, zal nog moeten blijken. Vooralsnog blijft het iets dat hoog op de agenda staat.

Gezien de eerder beschreven problemen lijkt het erg nuttig om de bewegingen op de arbeidsmarkt eens goed onder de loep te nemen. Uiteraard is de arbeidsmarkt zeer divers en is deze op te delen in verschillende sectoren zoals de bouw, detailhandel, de zorg etc. De bewegingen die tussen deze sectoren plaatsvinden kunnen van waarde zijn. Ook veranderingen zoals mensen die van werkend naar werkloos gaan of juist precies andersom, zijn zaken die absoluut het onderzoeken waard zijn. Daarnaast is het interessant om te kijken of bepaalde sectoren geclusterd zouden kunnen worden aan de hand van hun statistische eigenschappen. Met andere woorden welke banen enigszins vergelijkbaar zijn.

De data die in deze paper zullen worden gebruikt zijn afkomstig van het SSB (Sociaal Statistisch Bestand). Het SSB volgt alle mensen op de arbeidsmarkt en aggregeert deze informatie tot zogenaamde overgangsmatrices. Deze overgangsmatrices tonen de overgangskansen en behoudkansen van de arbeidssectoren in een bepaald jaar. Oftewel de kans dat een persoon in bepaald jaar van sector a naar sector b gaat. Deze overgangsmatrices zijn erg informatief aangezien zij de stromingen tussen de verschillende sectoren goed weergeven.

Met behulp van een aantal modellen die gebruik maken van de eerder genoemde overgangsmatrices zal ik onderzoeken hoe de bewegingen op de arbeidsmarkt zijn te verklaren. Hiermee zijn we dan ook aangekomen bij de onderzoeksvraag van deze paper:
Zijn er clusters van sectoren te vinden zodanig dat sectoren binnen een bepaalde cluster vergelijkbare bewegingen op de arbeidsmarkt laten zien?

Deze paper is als volgt opgebouwd: in de hierop volgende sectie zal de gebruikte data worden besproken. Vervolgens de methodiek, waar duidelijk wordt gemaakt welke methoden voor het onderzoek zullen worden gebruikt. Dit begint met de beschrijving van een clusteranalyse en vervolgens de modelfase. De sectie resultaten toont de resultaten van het onderzoek en in de laatste sectie, de conclusie, zullen de belangrijkste bevindingen worden besproken.

2 Data

Zoals eerder gezegd is de data die in deze paper gebruikt zal worden, afkomstig van het SSB. Het SSB volgt alle mensen op de arbeidsmarkt. Elk ontslag dat er valt, elke baanwisseling, wordt door hun genoteerd. Vervolgens aggregeert het SSB deze informatie tot de de eerder genoemde overgangsmatrices.Deze matrices tonen de overgangskansen en de behoudkansen van de arbeidsectoren in een

bepaald jaar. De verschuivingen op de Nederlandse arbeidsmarkt kunnen worden onderzocht aan de hand van deze verschillende sectoren of arbeidsituaties. Een soortgelijk onderzoek, namelijk Maria (2013), maakt gebruik van dezelfde dataset. Hier zal later nog op worden teruggekomen.

Categorie	Leeftijd	Categorie	Leeftijd
1	15-20	7	45-50
2	20-25	8	50-55
3	25-30	9	55-60
4	30-35	10	60-65
5	35-40	11	65-70
6	40-45		

Tabel 1: Leeftijdscategoriën

De dataset van het SSB bestaat uit elf overgangsmatrices per jaar, waarbij elk van deze elf matrices correspondeert met een bepaalde leeftijdcategorie. Tabel 1 toont een overzicht van deze leeftijdscategoriën. De dataset bevat de overgangsmatrices uit de periode 1999-2000 tot en met 2007-2008. Elk van deze overgangsmatrices bevat 30 verschillende states, deze weerspiegelen de verschillende arbeidsectoren plus arbeidsituaties. Tabel 2 geeft hier een overzicht van. Om een indruk te krijgen van hoe zo'n overgangsmatrix er precies uitziet volgt hieronder een voorbeeld voor de leeftijdscategorie 15-20 in het jaar 1999-2000.

$$\begin{array}{l|cccc} \text{1999-2000 categorie 1} & sector1 & sector2 & \dots & sector30 \\ sector1 & 0.5501 & 0.0009 & \dots & 0.0130 \\ sector2 & 0.1966 & 0.4692 & \dots & 0.0078 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ sector30 & 0.0662 & 0.0004 & \dots & 0.6444 \end{array}$$

Als we dan bijvoorbeeld het getal 0.5501 bekijken dan is dit de kans dat een persoon in het jaar 1999-2000 niet van sector verandert. Wat meteen al opvalt is dat de kans dat een persoon in zijn eigen sector blijft in alle overgangsmatrices de grootste waarde heeft. Hier zullen we later in de paper op terugkomen.

In deze paper zal alleen gekeken wordt naar de leeftijdscategorie 25-30 aangezien dit hoogstwaarschijnlijk een van de meest dynamische categoriën is. Daarnaast richt dit onderzoek zich ook op een select aantal sectoren namelijk: sector 1 Inactief sector 2 WW sector 3 Detailhandel sector 4 Horeca en contractcatering sector 5 Banken en Verzekeringswezen. De keuze is gevallen op deze sectoren omdat naar verwachting in de detailhandel, horeca en contractcatering alsmede bank en verzekeringswezen veel mensen werkzaam zijn. Anderzijds vallen gezien de vergrijzing en de economische recessie steeds meer mensen in de sectoren

inactief en WW. Het jaar waarvan de verschuivingen onderzocht zullen worden betreft de overgang 2007-2008. In de volgende sectie worden de methoden uiteengezet die gebruikt zullen worden om de data te analyseren.

Sector	Arbeidsmarktpositie
1	Inactief
2	Werkend SBI onbekend
3	WWB (Wet Werk en Bijstand)
4	AO
5	WW
6	Landbouw, bosbouw en jacht, visserij
7	winning delfstoffen
8	Traditionele industrie
9	Aardolie-/splijt- en kweekstoffenindustrie/chemie, rubber
10	metaalnijverheid
11	elektriciteit, aardgas, water (nutsbedrijven)
12	Bouwnijverheid
13	Bouwinstallatie
14	Vervoermiddelen (handel, reparatie, benzine)
15	Groothandel
16	Detailhandel
17	Horeca en contractcatering
18	Transport, opslag, post, telecommunicatie
19	Banken en verzekeringswezen
20	Informatie en communicatietechnologie
21	Schoonmaak
22	Verhuur en handel in roerend en onroerend goed
23	R&D/ Advocaten, accountants, economisch adviesbureaus / ing en architecten
24	Uitzendwezen en bemiddeling, reclame en rest zakelijke diensten, beveiliging
25	Openbaar bestuur
26	Onderwijs
27	Zorg
28	Welzijn
29	Cultuur, sport en recreatie
30	Milieudienstverlening, ideële organisaties en overige dienstverlening

Tabel 2: Sectoromschrijving

3 Methode

In deze sectie zal de methodiek van het onderzoek worden beschreven. We beginnen met een uitleg die duidelijk maakt op basis waarvan potentiële clusters gevormd zouden kunnen worden, de clusteranalyse. Vervolgens zal worden ingegaan op wat voor model zal worden opgesteld met betrekking tot het schatten van de overgangskansen.

3.1 Cluster Model

Het doel van het clusteren is om de relatief grote groep objecten, op te delen in een aantal deelverzamelingen. De objecten binnen zo'n deelverzameling zijn dan relatief gezien gelijkwaardig. Er bestaan veel verschillende soorten clustermethoden of mooier gezegd classificatiemethoden. Je kan namelijk op monothetische wijze groeperen. De clusters worden dan samengesteld op basis van één kenmerk. Een andere manier is polythetisch, dan worden er meerdere kenmerken gebruikt voor de classificatie. Daarnaast hebben we ook nog de methode genaamd iteratieve partitionering waarbij, zoals ook Fraley and Raftery (2002) vermelden, observaties steeds verplaatst worden van de ene groep naar de andere totdat er geen verbetering meer te behalen is in een bepaald criterium. Wanneer de observaties worden samengevoegd in steeds grotere groepen dan gaat het om een agglomeratieve methode ookwel bekend als bottom-up methode. Als de clusteranalyse als resultaat een dendrogram oplevert (een diagram met een boomstructuur) dan noemt men de classificatie hiërarchisch. Te denken valt bijvoorbeeld aan een stamboom.

In deze paper zal worden geclusterd op basis van iteratieve partitionering met als criterium de loglikelihood. Om een idee te krijgen wat precies met loglikelihood bedoeld wordt, volgt nu eerst een korte omschrijving van de schattingsmethode Maximale Aannemelijkheid. Deze methode, in het Engels beter bekend als Maximum Likelihood Estimation (MLE), is een manier om de parameters van een bepaald model te schatten. Laten we als voorbeeld nemen een model met de parameters gemiddelde en standaarddeviatie. MLE zorgt er nu voor dat, bij de gegeven dataset, er een set van waarden gevonden wordt waarbij de modelparameters gemiddelde en standaarddeviatie de zogenaamde Likelihood functie maximaliseren. Deze Likelihood functie wordt door de onderzoeker zelf opgesteld en moet zo goed mogelijk bij de dataset passen. In het algemeen neemt men de logaritme van de Likelihood functie omdat deze makkelijker is om mee te werken, dit levert dan als eindresultaat de eerder genoemde loglikelihood op.

De loglikelihood speelt bij onze methode een sleutelrol aangezien we aan de hand hiervan de clusters zullen bepalen. Hoe de precieze Likelihood functie gedefinieerd is voor dit onderzoek en waarom, wordt uitgelegd in de hierop volgende secties. Nu eerst een overzicht van de clustermethode.

Stap 1 Allereerst kiezen we zelf een bepaald clusteraantal, in dit onderzoek vijf aangezien we vijf sectoren hebben.

Stap 2 Start met het plaatsen van alle 5 sectoren in één cluster en bepaal vervolgens de loglikelihood.

Stap 3 Je stelt de mogelijkheid voor 2 clusters en gaat vervolgens alle mogelijke combinaties af om 5 sectoren in twee clusters op te delen. Kies die combinatie die de hoogste loglikelihood oplevert. Voer vervolgens dezelfde stap uit voor 3,4

en 5 clusters.

Stap 4 De optimale combinaties die je krijgt voor de verschillende cluster-aantallen vergelijk je aan de hand van het Akaike informatie criterium (AIC). Kies het clustermodel waarbij de AIC het laagst is.

Ter verantwoording: het kiezen van een optimale combinatie bij een gegeven clusteraantal, mag gedaan worden op basis van de loglikelihood. Echter, het vergelijken van combinaties van verschillende clusteraantallen is niet toegestaan omdat dit leidt tot verkeerde conclusies. Als je namelijk modelllen vergelijkt met een verschillend clusteraantal, is het aantal parameters ook verschillend. Met andere woorden de vergelijking die je maakt is niet helemaal eerlijk meer, vandaar AIC. De AIC is als volgt gedefinieerd:

$$AIC = 2k - 2log(L) \tag{1}$$

k stelt hier het aantal parameters voor en L de gemaximaliseerde waarde van de Likelihood functie. Overigens moet vermeld worden dat bij het selecteren van een optimale combinatie we de loglikelihood eerst met -2 vermenigvuldigen, hiermee verkrijgen we dan de zogeheten deviance. Op basis hiervan kiezen we. Dit doen we omdat programmeertechnisch gezien, dit handiger is. Met andere woorden in plaats van de loglikelihood te maximaliseren, minimaliseren we de deviance. Uiterraard komt dit op hetzelfde neer.

Het genereren van alle mogelijke combinaties om sectoren in clusters op te delen is gedaan naar hoe vermeld staat in de paper van Orlov (2002). Hierin worden een aantal bewijzen en pseudocodes gegeven met betrekking tot het bepalen van partities. Orlov's methode is als volgt: hij stelt eerst dat er een bijectie is tussen alle partities van een set $\{x_1, ..., x_n\}$ en een set $\{k_1, ..., k_n\}$. Hierbij is de set k als volgt gedefinieerd:

$$\{\{k_1, ...k_n\} : \forall 2 \leq i \leq n : k_1 \leq k_i \leq max_{1 \leq j < i}\, k_j + 1, k_i \in Z\} voor\ alle\ n \in N en\ k_1 \in Z$$

Vervolgens gaat Orlov van de set x de deelverzamelingen bepalen en dit doet hij op basis van de laagste index. Bijvoorbeeld:

$$Partitie\ 1 = \{\{x_5, x_3\}, \{x_1, x_2\}, \{x_7, x_6, x_4\}$$

$$=> \{\{x_1, x_2\}^1, \{x_3, x5\}^2, \{x_4, x_6, x_7\}^3$$

$$k_1 = \{1, 1, 2, 3, 2, 3, 3\}$$

Elk element in de k-vector geeft aan in welke cluster, in dit geval 1,2 of 3, een x met die index (van de k-vector) zich bevindt. Dit is slechts een voorbeeld en in dit onderzoek gebruiken we het algoritme dat Orlov geeft om alle mogelijke partities te bepalen.

3.2 Markov Keten

Voordat de modelfase van dit onderzoek zal worden besproken, eerst een korte, noodzakelijke omschrijving van Markov ketens. Deze spelen namelijk bij de gegeven data wel degelijk op de achtergrond een rol.

Een stochastisch proces $\{X_t, t \geq 0\}$ met een eindige toestandsruimte wordt een discrete tijd Markov keten genoemd. De grond aanname voor Markov ketens is dat de toestand, ookwel state, alleen afhangt van het heden en niet van de states in het verleden, de geheugenloosheid eigenschap. Met andere woorden:

$$Pr(X_{t+1} = j | X_t = i, X_{t-1}, ..., X_0) = Pr(X_{t+1} = j | X_t) \tag{2}$$

Hierbij is X_t de state op tijdstip t. Een Markov keten wordt weergegeven door een overgangsmatrix met hierin de overgangskansen. Een overgangskans is de kans dat een individu van een bepaalde state verandert naar een andere op een bepaald tijdstip. Als we nu voor dit onderzoek een statistisch model willen opstellen voor de Markov keten dan relateren we de overgangskansen aan een nog onbekende parameter θ. Dit ziet er dan als volgt uit:

$$Pr(X_{t+1} = j | X_t) = \pi_t^{ij}(\theta) \tag{3}$$

De theoretische overgangsmatrix ziet er dan als volgt uit:

$$\Pi_t(\theta) = \begin{bmatrix} \pi_t^{1,1} & \pi_t^{1,2} & \dots & \pi_t^{1,5} \\ \pi_t^{2,1} & \pi_t^{2,2} & \dots & \pi_t^{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_t^{5,1} & \pi_t^{5,2} & \dots & \pi_t^{5,5} \end{bmatrix}.$$

π_i moet hier als functie van de vector θ_i gezien worden. Verder volgt uit de Markoveigenschappen dat alle overgangskansen zich in het bereik [0,1] bevinden en dat $\sum_{j=1}^{J} \pi_{ij} = 1$.

3.3 Empirische data

Om het uiteindelijke model en de bijbehorende parameters goed te kunnen specificeren bekijken we eerst een voorbeeld op basis van empirische data. Hiertoe definiëeren we nu eerst de N-matrix, een empirische aantallen matrix. Elk getal in deze matrix weerspiegelt een aantal mensen dat van de ene state naar de andere state verandert.

$$N_t = \begin{bmatrix} n_t^{1,1} & n_t^{1,2} & \dots & n_t^{1,5} \\ n_t^{2,1} & n_t^{2,2} & \dots & n_t^{2,5} \\ \vdots & \vdots & \ddots & \vdots \\ n_t^{5,1} & n_t^{5,2} & \dots & n_t^{5,5} \end{bmatrix}.$$

Met betrekking tot deze N-matrix alvast een opmerking: als we een rij selecteren uit deze N-matrix en deze voor het moment even los zien van de context en we bekijken deze rij getallen als een gebeurtenis dan volgt de kans dat deze gebeurtenis optreedt een mulitnomiale kansverdeling.

Vervolgens de D_t matrix, een diagonaalmatrix met op de diagonaal het totaal aantal mensen in een een bepaalde state i op tijdstip t.

$$D_t = \begin{bmatrix} d_t^{1,1} & 0 & \dots & 0 \\ 0 & d_t^{i,i} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_t^{5,5} \end{bmatrix}.$$

Merk op dat een element $d_t^{i,i}$ de som is van alle elementen van één rij van N_t.

Als we nu het matrixproduct van N_t en D_t nemen, verkrijgen we de empirische overgangsmatrix P, namelijk P $= D^{-1}N$. Elk element van de overgangsmatrix P stelt de kans voor dat een individu op een tijdstip t van een state i naar een state j gaat. We zien nu de gelijkenis tussen $\Pi_t(\theta)$ en P. De matrix P voldoet aan eerder genoemde markoveigenschappen namelijk: alle kansen bevinden zich op het bereik [0,1] en $\sum_{j=1}^{J} p_{ij} = 1$. Gezien deze informatie kiezen we als model het multinomiaal logit model, deze lijkt, ook zoals Maria (2013) aangeeft, het best op de data te passen en neemt ook de markoveigenschappen mee.

3.4 Schattingsmodel

Nu volgt een overzicht van een algemeen model dat als het ware als basis dient voor het uiteindelijke model waarmee de overgangskansen zullen worden geschat. Dit zal, zoals al eerder gezegd, gebaseerd zijn op Maximum Likelihood. Voor dit onderzoek zal gebruik gemaakt worden van een likelihood functie die gebaseerd is op het multinomiale logit model. Hoe dit model gespecificeerd is, is ook terug te vind in Paap (2013) en Fok (2012). De kans dat een persoon dan van een bepaalde state i verandert naar state j ziet er dan als volgt uit:

$$\pi_{ij}(\theta_i) = Pr(X_{t+1} = j | X_t = i) = \frac{exp(\theta_{ij})}{\sum_{l=1}^{J} exp(\theta_{il})} \; voor \; j = 1, ..., J \qquad (4)$$

θ_{ij} stelt hier het j^e element voor van de vector θ_i. Het probleem dat we nu echter hebben, is dat de parameters niet geïdentificeerd zijn. Dit houdt eigenlijk in dat er geen start is om van te beginnen, geen referentiepunt. Vandaar dat we deze referentie zelf instellen door te zeggen b=basis. Het model ziet er dan als volgt uit:

$$\pi_{ij}(\theta_i) = \frac{exp(\theta_{ij})}{1 + \sum_{l \neq b}^{J} exp(\theta_{il})} \; voor \; j = 1, ..., J voor j = 1, ..., J \qquad (5)$$

en

$$\pi_{ib}(\theta_i) = \frac{1}{1 + \sum_{l \neq b}^{J} exp(\theta_{il})} \; voor \; j = 1, ..., J \tag{6}$$

We gebruiken hier de kans voor een persoon om in zijn eigen sector te blijven als referentiecategorie. Dit houdt in dat alle θ_{ii} op nul gezet zullen worden. De model parameters zullen geschat worden door middel van Maximum Likelihood, dit aan de hand van de volgende Likelihood functie:

$$L(\theta) = \binom{(\sum_{j=1}^{J} n_{ij})}{n_{i1} \; ... \; n_{iJ}} \prod_{j=1}^{J} \prod_{i=1}^{I} (\pi_{ij}(\theta_i))^{n_{ij}}) \tag{7}$$

Zoals eerder vermeld neemt men in het algemeen de logaritme van de Likelihood functie omdat deze makkelijker is om mee te werken. Deze wordt dan gegeven door:

$$ll(\theta) = log(\binom{(\sum_{j=1}^{J} n_{ij})}{n_{i1} \; ... \; n_{iJ}}) + \sum_{j=1}^{J} \sum_{i=1}^{I} (n_{ij} log(\pi_{ij}(\theta_i))) \tag{8}$$

Om de aannemelijkheidschatter θ te verkrijgen zal de loglikelihood worden gemaximaliseerd. Wel moet hierbij vermeld worden dat bij het maximaliseren de eerste term in de Likelihood functie niet hoeft te worden meegenomen. Dit omdat de eerste term een constante is. Het resultaat is een JxJ matrix $\widehat{\Theta}$ met alle parameterschattingen. De geschatte overgangsmatrix die hierbij hoort is $\widehat{\Pi}(\theta)$ met de overgangskansen $\pi_{ij}(\widehat{\theta})$

3.5 De Clusterstructuur

Dan nu de beschrijving van het daadwerkelijke model waarin rekening mee zal worden gehouden met clusterstructuren. Als we weer de theoretische overgangsmatrix $\Pi(\theta)$ in ogenschouw nemen, willen we nu door middel van clusters een bepaalde structuur aanbrengen op deze matrix. Een belangrijk voordeel dat hiermee behaald kan worden is, ook zoals Maria (2013) aangeeft, dat het aantal te schatten paramters, dat nu op JxJ staat, behoorlijk kan worden teruggebracht. De clusters geven bepaalde indelingen van de sectoren in groepen. Een dergelijke indeling kunnen we nu toepassen op de matrix $\Pi(\theta)$. We bekijken hier een voorbeeld met ook 5 sectoren, $\Pi(\theta)$ komt er dan als volgt uit te zien:

$$\Pi_t(\theta) = \begin{bmatrix} \pi_t^{1,1} & \pi_t^{1,2} & \pi_t^{1,3} & \pi_t^{1,4} & \pi_t^{1,5} \\ \pi_t^{2,1} & \pi_t^{2,2} & \pi_t^{2,3} & \pi_t^{2,4} & \pi_t^{2,5} \\ \pi_t^{3,1} & \pi_t^{3,2} & \pi_t^{3,3} & \pi_t^{3,4} & \pi_t^{3,5} \\ \pi_t^{4,1} & \pi_t^{4,2} & \pi_t^{4,3} & \pi_t^{4,4} & \pi_t^{4,5} \\ \pi_t^{5,1} & \pi_t^{5,2} & \pi_t^{5,3} & \pi_t^{5,4} & \pi_t^{5,5} \end{bmatrix}.$$

Stel nu dat deze vijf sectoren zijn onderverdeeld in twee clusters C1 en C2.

C1 bevat sector 1,2 en 3 en C2 bevat 4 en 5. Om aan te geven waar we naartoe willen, de volgende visualisatie:

$$\Pi_t(\theta) = \left[\begin{array}{ccc|cc} \pi_t^{1,1} & \pi_t^{1,2} & \pi_t^{1,3} & \pi_t^{1,4} & \pi_t^{1,5} \\ \pi_t^{2,1} & \pi_t^{2,2} & \pi_t^{2,3} & \pi_t^{2,4} & \pi_t^{2,5} \\ \pi_t^{3,1} & \pi_t^{3,2} & \pi_t^{3,3} & \pi_t^{3,4} & \pi_t^{3,5} \\ \hline \pi_t^{4,1} & \pi_t^{4,2} & \pi_t^{4,3} & \pi_t^{4,4} & \pi_t^{4,5} \\ \pi_t^{5,1} & \pi_t^{5,2} & \pi_t^{5,3} & \pi_t^{5,4} & \pi_t^{5,5} \end{array}\right] \Pi_t(\theta) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right].$$

Blok A vormt nu de bewegingen binnen cluster 1 en blok D de bewegingen binnen cluster 2. Blokken B en C bevatten de kansen wat betreft uitwisselingen tussen de twee clusters. Wat je nu dankzij deze structuur krijgt is dat je eigenlijk alleen nog maar te maken hebt met de kansen om in een cluster te blijven en de kansen om tussen clusters te wisselen. Dit betekent dat we voor dit voorbeeld met vier onbekende parameters te maken krijgen en dat onze parameter matrix Θ er dan als volgt uitziet:

$$\Theta = \begin{bmatrix} \beta_{11} & \beta_{11} & \beta_{11} & \beta_{12} & \beta_{12} \\ \beta_{11} & \beta_{11} & \beta_{11} & \beta_{12} & \beta_{12} \\ \beta_{11} & \beta_{11} & \beta_{11} & \beta_{12} & \beta_{12} \\ \beta_{21} & \beta_{21} & \beta_{21} & \beta_{22} & \beta_{22} \\ \beta_{21} & \beta_{21} & \beta_{21} & \beta_{22} & \beta_{22} \end{bmatrix}.$$

- β_{11} is de overgangskans van cluster 1 naar cluster 1
- β_{12} is de overgangskans van cluster 1 naar cluster 2
- β_{21} is de overgangskans van cluster 2 naar cluster 1
- β_{22} is de overgangskans van cluster 2 naar cluster 2

Hieruit is ook direct af te leiden dat we in plaats van JxJ parameters nog maar KxK parameters hoeven te schatten, waarbij K het aantal clusters voorstelt. De parameter matrix Θ zal geschat worden met MLE schatter $\widehat{\Theta}$. Hiermee verkrijgen we dan de geschatte overgangsmatrix $\Pi(\widehat{\theta})$. β_{ij} is hier overigens weer het j^e van de vector β_i voor een bepaalde i en j.

Als we nogmaals de data bekijken, valt het meteen op dat de diagonaalelementen bij alle overgangsmatrices de grootste waarde hebben. Met andere woorden de kans dat een persoon in zijn eigen sector blijft is in alle gevallen het grootst. Dit is zeker iets waar tijdens het schatten rekening mee moet worden gehouden. De oplossing die Maria (2013) hiervoor biedt is om nog een extra parameter te introduceren, in ons geval γ, speciaal voor de diagonaalelementen. Deze extra parameter kan het best worden gezien als een soort bonus voor de kans om in je eigen sector te blijven. Dit betekent dat onze parameter matrix er nu als volgt uitziet:

$$\Theta = \begin{bmatrix} \beta_{11}+\gamma & \beta_{11} & \beta_{11} & \beta_{12} & \beta_{12} \\ \beta_{11} & \beta_{11}+\gamma & \beta_{11} & \beta_{12} & \beta_{12} \\ \beta_{11} & \beta_{11} & \beta_{11}+\gamma & \beta_{12} & \beta_{12} \\ \beta_{21} & \beta_{21} & \beta_{21} & \beta_{22}+\gamma & \beta_{22} \\ \beta_{21} & \beta_{21} & \beta_{21} & \beta_{22} & \beta_{22}+\gamma \end{bmatrix}.$$

Deze onbekende parameters β en γ zullen worden geschat aan de hand van π_{ij} deze stellen de overgangskansen voor om van sector i in cluster k naar sector j in cluster l te gaan, hier het algemene overzicht:

$$\pi_{ij}(\beta) = \frac{exp(\beta_{kl})}{\sum_{h=1}^{J} exp(\beta_{kh}) + exp(\beta_{kk}+\gamma)} \; voor \; h = 1, ..., J \tag{9}$$

en voor de diagonaalelementen π_{ii} :

$$\pi_{ii}(\beta) = \frac{exp(\beta_{kk}+\gamma)}{\sum_{h=1}^{J} exp(\beta_{kh}) + exp(\beta_{kk}+\gamma)} \; voor \; h = 1, ..., J \tag{10}$$

Uiteraard moet ook hier een identificatie restrictie worden toegepast. Het handigste is om de kans dat een persoon binnen zijn eigen cluster blijft, als referentiecategorie te nemen. Dus voor $i, j \in C_k \; \beta_{kk} = 0$, dit betekent dat het model er als volgt uit ziet:

$$\pi_{ij}(\beta) = \frac{exp(\beta_{kl})}{\sum_{h\neq k}^{J} exp(\beta_{kh}) + exp(\gamma)} \; voor \; h = 1, ..., J \tag{11}$$

En voor de diagonaalelementen:

$$\pi_{ii}(\beta) = \frac{exp(\gamma)}{1 + \sum_{h\neq k}^{J} exp(\beta_{kh}) + exp(\gamma)} \; voor \; h = 1, ..., J \tag{12}$$

En voor de referentiecategorie:

$$\pi_{ib}(\beta) = \frac{1}{1 + \sum_{h\neq k}^{J} exp(\beta_{kh}) + exp(\gamma)} \; voor \; h = 1, ..., J \tag{13}$$

We verkrijgen dan de volgende likelihood functie:

$$L(\beta) = \prod_{j=1}^{J} \prod_{i=1}^{I} (\pi_{ij}(\beta))^{n_{ij}} \tag{14}$$

4 Resultaten

In dit hoofdstuk zullen de resultaten van het onderzoek worden toegelicht. Voor elk clusteraantal is die indeling van sectoren gekozen die de minimale deviance opleverde. Hieronder nu een overzicht van deze resultaten.

Allereerst bekijken we, wat verkregen werd als we alle sectoren in één cluster stopten. De matrices β en $\Pi(\beta)$ zagen er als volgt uit:

$$\beta = \begin{bmatrix} 3,745 & 0,000 & 0,000 & 0,000 & 0,000 \\ 0,000 & 3,745 & 0,000 & 0,000 & 0,000 \\ 0,000 & 0,000 & 3,745 & 0,000 & 0,000 \\ 0,000 & 0,000 & 0,000 & 3,745 & 0,000 \\ 0,000 & 0,000 & 0,000 & 0,000 & 3,745 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,914 & 0,022 & 0,022 & 0,022 & 0,022 \\ 0,022 & 0,914 & 0,022 & 0,022 & 0,022 \\ 0,022 & 0,022 & 0,914 & 0,022 & 0,022 \\ 0,022 & 0,022 & 0,022 & 0,914 & 0,022 \\ 0,022 & 0,022 & 0,022 & 0,022 & 0,914 \end{bmatrix}.$$

Deviance= 152590 AIC= 152592 Optimale partitie=[1 1 1 1 1]

Even een uitleg wat betreft de partitie, het betekent hier sector 1 naar cluster 1, sector 2 naar cluster 1 etc. We zien dat in de matrix β slechts op de diagonaal de coëfficiënten ongelijk aan nul zijn. Dit is logische aangezien we de kans dat een persoon binnen zijn eigen cluster blijft op nul hadden gezet in verband met identificatie van de parameters. We hebben hier maar één cluster dus dat klopt. Om aan te tonen dat er wel degelijk wordt geoptimaliseerd tonen we bij de clusteraantallen 2,3 en 4 ook de een na beste schattingen. Uit het steeds relatief grote verschil in AIC tussen de beste en de een na beste schatting is af te leiden dat de beste duidelijk optimaal is. Dit kan niet gedaan worden voor de clusteraantallen 1 en 5 omdat er simpelweg bij beide maar één combinatie mogelijk is, of alle sectoren in 1 cluster of elke sector in een cluster.

Dan nu **2** clusters, de beste:

$$\beta = \begin{bmatrix} 3,635 & -1,252 & 0,000 & 0,000 & 0,000 \\ 4,495 & 3,635 & 4,495 & 4,495 & 4,495 \\ 0,000 & -1,252 & 3,635 & 0,000 & 0,000 \\ 0,000 & -1,252 & 0,000 & 3,635 & 0,000 \\ 0,000 & -1,252 & 0,000 & 0,000 & 3,635 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,920 & 0,007 & 0,024 & 0,024 & 0,024 \\ 0,226 & 0,096 & 0,226 & 0,226 & 0,226 \\ 0,024 & 0,007 & 0,920 & 0,024 & 0,024 \\ 0,024 & 0,007 & 0,024 & 0,920 & 0,024 \\ 0,024 & 0,007 & 0,024 & 0,024 & 0,920 \end{bmatrix}.$$

Deviance= 144310 AIC= 144315 Optimale partitie= [1 2 1 1 1]

De een na beste :

$$\beta = \begin{bmatrix} 4,511 & 0,684 & 0,684 & 0,684 & 0,684 \\ 1,805 & 4,511 & 0,000 & 0,000 & 0,000 \\ 1,805 & 0,000 & 4,511 & 0,000 & 0,000 \\ 1,805 & 0,000 & 0,000 & 4,511 & 0,000 \\ 1,805 & 0,000 & 0,000 & 0,000 & 4,511 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,920 & 0,020 & 0,020 & 0,020 & 0,020 \\ 0,061 & 0,909 & 0,010 & 0,010 & 0,010 \\ 0,061 & 0,010 & 0,909 & 0,010 & 0,010 \\ 0,061 & 0,010 & 0,010 & 0,909 & 0,010 \\ 0,061 & 0,010 & 0,010 & 0,010 & 0,909 \end{bmatrix}.$$

Deviance= 144910 AIC= 144915 Optimale partitie= [1 2 2 2 2]

Nu **3** clusters, de beste:

$$\beta = \begin{bmatrix} 3,332 & -1,627 & 0,000 & 0,000 & -1,064 \\ 4,740 & 3,332 & 4,740 & 4,740 & 1,984 \\ 0,000 & -1,627 & 3,332 & 0,000 & -1,064 \\ 0,000 & -1,627 & 0,000 & 3,332 & -1,064 \\ -0,717 & -0,999 & -0,717 & -0,717 & 3,332 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,917 & 0,006 & 0,033 & 0,033 & 0,011 \\ 0,302 & 0,074 & 0,302 & 0,302 & 0,019 \\ 0,033 & 0,006 & 0,917 & 0,033 & 0,011 \\ 0,033 & 0,006 & 0,033 & 0,917 & 0,011 \\ 0,016 & 0,012 & 0,016 & 0,016 & 0,939 \end{bmatrix}.$$

Deviance= 141610 AIC= 141620 Optimale partitie= [1 2 1 1 3]

De een na beste:

$$\beta = \begin{bmatrix} 4,051 & -1,133 & 0,442 & 0,442 & 0,442 \\ 4,546 & 4,051 & -2,180 & -2,180 & -2,180 \\ 1,082 & -0,987 & 4,051 & 0,000 & 0,000 \\ 1,082 & -0,987 & 0,000 & 4,051 & 0,000 \\ 1,082 & -0,987 & 0,000 & 0,000 & 4,051 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,920 & 0,005 & 0,025 & 0,025 & 0,025 \\ 0,620 & 0,378 & 0,001 & 0,001 & 0,001 \\ 0,047 & 0,006 & 0,915 & 0,016 & 0,016 \\ 0,047 & 0,006 & 0,016 & 0,915 & 0,016 \\ 0,047 & 0,006 & 0,016 & 0,016 & 0,915 \end{bmatrix}.$$

Deviance= 142490 AIC= 142500 Optimale partitie= [1 2 3 3 3]

Vervolgens **4** clusters, de beste:

$$\beta = \begin{bmatrix} 3,685 & -1,429 & 0,238 & 0,238 & -0,793 \\ 4,204 & 3,685 & 3,044 & 3,044 & -5,610 \\ 0,960 & -0,807 & 3,685 & 0,000 & -1,470 \\ 0,960 & -0,807 & 0,000 & 3,685 & -1,470 \\ 0,023 & -0,564 & -1,162 & -1,162 & 3,685 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,925 & 0,006 & 0,029 & 0,029 & 0,011 \\ 0,450 & 0,268 & 0,141 & 0,141 & 0,000 \\ 0,059 & 0,010 & 0,903 & 0,023 & 0,005 \\ 0,059 & 0,010 & 0,023 & 0,903 & 0,005 \\ 0,024 & 0,014 & 0,007 & 0,007 & 0,947 \end{bmatrix}.$$

Deviance= 139550 AIC= 139567 Optimale partitie= [1 2 3 3 4]

De een na beste:

$$\beta = \begin{bmatrix} 3,573 & -1,065 & 0,000 & -0,266 & -0,740 \\ 4,018 & 3,573 & 4,018 & -1,166 & -4,639 \\ 0,000 & -1,065 & 3,573 & -0,266 & -0,740 \\ 0,833 & -0,329 & 0,833 & 3,573 & 0,039 \\ 0,016 & 0,089 & 0,016 & -3,331 & 3,573 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,932 & 0,009 & 0,026 & 0,020 & 0,012 \\ 0,378 & 0,242 & 0,378 & 0,002 & 0,000 \\ 0,026 & 0,009 & 0,932 & 0,020 & 0,012 \\ 0,055 & 0,017 & 0,055 & 0,848 & 0,025 \\ 0,026 & 0,028 & 0,026 & 0,001 & 0,918 \end{bmatrix}.$$

Deviance= 142670 AIC= 142687 Optimale partitie= [1 2 1 3 4]

En als laatste **5** clusters:

$$\beta = \begin{bmatrix} 3,164 & -2,024 & -0,563 & -0,746 & -0,522 \\ 3,307 & 3,164 & 1,509 & -0,240 & -1,266 \\ -0,315 & -1,989 & 3,164 & -1,515 & -1,576 \\ 1,141 & -0,568 & 0,224 & 3,164 & -1,155 \\ -0,427 & -1,107 & -0,942 & 0,111 & 3,164 \end{bmatrix} \Pi(\beta) = \begin{bmatrix} 0,930 & 0,005 & 0,022 & 0,019 & 0,023 \\ 0,483 & 0,418 & 0,080 & 0,014 & 0,005 \\ 0,029 & 0,005 & 0,948 & 0,009 & 0,008 \\ 0,108 & 0,020 & 0,043 & 0,818 & 0,011 \\ 0,025 & 0,013 & 0,015 & 0,043 & 0,905 \end{bmatrix}.$$

Deviance= 140950 AIC= 140976 Optimale partitie= [1 2 3 4 5]

Om alles nog even goed op een rijtje te zetten de volgende tabel met alle beste schattingen:

Clusteraantal	Deviance	AIC	Optimale partitie
1	152590	152592	[1 1 1 1 1]
2	144310	144315	[1 2 1 1 1]
3	141610	141620	[1 2 1 1 3]
4	139550	139567	[1 2 3 3 4]
5	140950	140976	[1 2 3 4 5]

Wat we hier dus zien is dat het clusteraantal dat de kleinste defiance en daarmee de hoogste loglikelihood oplevert, het getal 4 is. Met een deviance van 139550 is dit duidelijk de meest logische keuze. De optimale indeling van sectoren die hierbij hoort is als volgt: in cluster 1 de sector inactief, in cluster 2 de sector WW, in cluster 3 de sectoren detailhandel en horeca en contractcatering en in cluster 4 de sector bank en verzekeringswezen. Het blijkt dus dat mensen in de leeftijdscategorie 25-30 jaar in de sectoren detailhandel en horeca en contractcatering zich op vergelijkbare wijze bewegen. We lichten de matrix β behorende bij dit clusteraantal en een matrix om de structuur van β aan te geven nog even uit:

$$\beta_{4clusters} = \begin{bmatrix} 3,685 & -1,429 & 0,238 & 0,238 & -0,793 \\ 4,204 & 3,685 & 3,044 & 3,044 & -5,610 \\ 0,960 & -0,807 & 3,685 & 0,000 & -1,470 \\ 0,960 & -0,807 & 0,000 & 3,685 & -1,470 \\ 0,023 & -0,564 & -1,162 & -1,162 & 3,685 \end{bmatrix}.$$

$$\beta_{4clusters} = \begin{bmatrix} \gamma & \beta_{12} & \beta_{13} & \beta_{13} & \beta_{14} \\ \beta_{21} & \gamma & \beta_{23} & \beta_{23} & \beta_{24} \\ \beta_{31} & \beta_{32} & \gamma & \beta_{33} & \beta_{34} \\ \beta_{31} & \beta_{32} & \beta_{33} & \gamma & \beta_{34} \\ \beta_{41} & \beta_{42} & \beta_{43} & \beta_{43} & \gamma \end{bmatrix}.$$

Met name de tweede rij van de matrix β toont in vergelijking tot de andere grote coëfficiënten, oftewel behoorlijke uitwisseling tussen die clusters. Het betreft hier de de uitwisseling tussen cluster 2 en 1,3 en 4, dat is tussen WW enerzijds en inactief, horeca en contractcatering, detailhandel en bank en verzekeringswezen anderzijds. Als we dan ook weer even terugkijken naar de $\Pi(\beta)$ die hierbij hoort, zie je dat ook hier de tweede rij de grootste kansen bevat in vergelijking tot de andere. Met andere woorden de kans om cluster 2 te verlaten voor 1,3 of 4 is over het algemeen het grootst.

5 Conclusie

We begonnen dit onderzoek met een dataset die aangaf hoe mensen wisselden tussen de sectoren op de arbeidsmarkt. Het doel van het onderzoek was om een beter inzicht te krijgen in deze verschuivingen. We hebben er hier voor gekozen om mensen in de leeftijdscategorie 25 tot 30 jaar te bekijken. Dit omdat deze leeftijdscategorie waarschijnlijk het meest dynamisch is. Mensen zijn nog relatief jong, misschien net klaar met studeren of al aan het werk, met andere woorden een interessante categorie. Verder hebben we deze mensen bekeken voor de sectoren inactief, WW, detailhandel, horeca en contractcatering en bank en verzekeringswezen voor de jaarovergang 2007-2008. De centrale onderzoeksvraag die we hierbij probeerden te beantwoorden luidde:
Zijn er clusters van sectoren te vinden zodanig dat sectoren binnen een bepaalde cluster vergelijkbare bewegingen op de arbeidsmarkt laten zien?
In deze paper hebben we een model gespecificeerd om een zo goed mogelijk antwoord op deze vraag te vinden. We hebben gebruik gemaakt van een multinomiaal logit model om zo de overgangsmatrices te schatten. Daarnaast hebben we een zogenaamde clusterstructuur aangebracht op deze overgangsmatrices. Op basis van de loglikelihood is vervolgens bepaald welke indeling van de sectoren in clusters het meest waarschijnlijk is.

Om dan nu de onderzoeksvraag te beantwoorden, hebben we dergelijke clusters kunnen vinden? Ja, het bleek dat het clusteraantal dat het meest waarschijnlijk is, het clusteraantal 4 is. De indeling die hierbij hoort is als volgt: cluster 1 de sector inactief, cluster 2 de sector WW, cluster 3 de sectoren detailhandel en horeca en contractcatering en in cluster 4 de sector bank en verzekeringswezen. Met andere woorden de sectoren detailhandel en horeca en contractcatering zijn op basis van deze analyse, de sectoren die vergelijkbare bewegingen tonen op de arbeidsmarkt. Dit is ook wel een enigszins logisch resultaat aangezien beide sectoren ongeveer hetzelfde opleidingsniveau vereisen en waarschijnlijk ook dezelfde hoeveelheid aan loon uitbetalen. Deze sectoren zijn ook erg toegankelijk voor mensen en lenen zich bijvoorbeeld ook goed voor bijbaantjes. Een sector zoals bank en verzekeringswezen is dan toch weer heel wat anders. Vandaar dat het ook niet vreemd is dat, ook als we de andere twee sectoren erbij betrekken, bank en verzekeringswezen in een aparte cluster is geplaatst. Het resultaat dat

de sectoren inactief en WW in aparte clusters zijn gezet is logisch te verklaren uit het feit dat mensen in de sector WW nog steeds actief op zoek zijn naar werk terwijl dat voor de inactieven over het algemeen niet het geval is.

We zagen dat mensen die zich in cluster 2, oftewel de WW, bevinden de grootste kans hebben om van cluster te wisselen. Dit heeft waarschijnlijk te maken met het feit dat de mensen die zich in deze cluster bevinden actief op zoek zijn naar werk en dit op ieder moment kunnen vinden. Desalniettemin blijft het het meest waarschijnlijk dat mensen in het algemeen binnen hun eigen cluster blijven. Iedereen wil toch blijkbaar een zekere standvastigheid en zekerheid hebben en is dan het meest geneigd om alles bij het oude te houden. Mensen zijn en blijven routine wezens.

6 Toekomstig Onderzoek

Voor een eventueel vervolgonderzoek zou gekeken kunnen worden naar alle 30 sectoren die het SSB beschikbaar heeft gesteld. Aangezien dit een veel grotere dataset betreft wordt het clusteren allicht een stuk complexer maar het zou wel tot interessante inzichten kunnen leiden.

Referenties

Dennis Fok(2012). Slides econometrie 2.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

Marchella Maria. Where is the break, 2013.

Michael Orlov. Efficient generation of set partitions. *Engineering and Computer Sciences, University of Ulm, Tech. Rep*, 2002.

Richard Paap(2013). Slides marketing models.