Auteursrechterlijke overeenkomst

Opdat de Universiteit Hasselt uw eindverhandeling wereldwijd kan reproduceren, vertalen en distribueren is uw akkoord voor deze overeenkomst noodzakelijk. Gelieve de tijd te nemen om deze overeenkomst door te nemen, de gevraagde informatie in te vullen (en de overeenkomst te ondertekenen en af te geven).

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling met

Titel: Definiëren van socio-demografische profielen op basis van activiteiten-gebaseerd verplaatsingsgedrag
Richting: 2de masterjaar in de verkeerskunde - verkeersveiligheid Jaar: 2009

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Ik ga akkoord,

BLEUX, Kristof

Datum: 14.12.2009

Definiëren van socio-demografische profielen op basis van activiteiten-gebaseerd verplaatsingsgedrag

Kristof Bleux

promotor :
Prof. dr. Davy JANSSENS

Eindverhandeling voorgedragen tot het bekomen van de graad
master in de verkeerskunde verkeersveiligheid

universiteit hasselt

Woord vooraf

Deze Masterproef vormt het einde van mijn Masteropleiding Verkeerskunde afstudeerrichting verkeersveiligheid aan de Universiteit Hasselt te Diepenbeek. Dankzij de hulp en steun van een aantal personen ben ik er in geslaagd om dit project tot een goed einde te brengen.

Graag zou ik mijn promotor, Davy Janssens, willen bedanken voor zijn deskundige begeleiding, waardevolle aanbevelingen en hulp bij het tot stand brengen van deze Masterproef.

Daarnaast wil ik ook mijn begeleider, Marlies Vanhulsel, bedanken voor haar begeleiding en aanbevelingen. Alsook wil ik haar nog bedanken voor de goede uitleg bij het toepassen van de verschillende softwareprogramma's.

Verder wens ik ook mijn ouders te bedanken voor de steun die ze mij hebben gegeven de voorbije jaren.

Kristof Bleux,
Mei 2009

Samenvatting

Het onderzoek van deze masterthesis situeert zich in het domein van transportmodellering. Het doel van deze masterproef is dan op basis van dagboeken, opgesteld door verschillende individuen, socio-demografische profielen aanmaken met een gelijkaardig activiteiten-verplaatsingspatroon. Dit zal gebeuren aan de hand van een analyse waarbij dan homogene groepen van individuen bepaald worden. Deze gegevens zullen dan gekoppeld worden aan de socio-demografische kenmerken van de individuen.

Om de vraag naar socio-demografische profielen te begrijpen zullen de verschillende verkeersmodellen in het eerste hoofdstuk, de literatuurstudie, worden uitgelegd. Hierbij gaan de belangrijkste verschillen tussen de modellen verduidelijkt worden. De belangrijkste conclusie die dan volgt uit dit hoofdstuk, is dat de vraag naar activiteitgebaseerde modellen wordt gestimuleerd door de wil om het menselijke gedrag beter te begrijpen en hierdoor tevens de beleidsmaatregelen beter te modelleren.

In het 2de hoofdstuk wordt dan de opzet van het onderzoek verder uitgelegd. Doordat de vraag naar de nieuwe activiteitgebaseerde modellen groot is, is er nood aan nieuwe basisdata of inputgegevens. De basisdata die nodig zijn voor het aanmaken van deze modellen zijn vooral dagboeken en demografische gegevens. Eveneens zijn ook de persoonlijke gegevens van de ondervraagde individuen in dit onderzoek van belang. Omdat deze gegevens in werkelijkheid zeer uitlopend zijn, gaan er meer geaggregeerde profielen opgesteld worden waardoor de modellen niet te complex worden.

In het derde hoofdstuk worden dan de onderzoeksvragen weergegeven. Deze zullen een leidraad vormen doorheen de masterproef.

In het 4de hoofdstuk wordt de methodologie die doorheen de masterproef wordt gebruikt opgesomd. Er zijn hierin 3 hoofdtaken te onderscheiden: toepassen van SAM-methode, het opstellen van een afstandmatrix en het aanmaken van clusters met een evenwichtige verhouding. Deze worden eerst kort toegelicht in het eerste deel van het hoofdstuk. In het tweede deel van dit hoofdstuk zijn de gebruikte methodes meer gedetailleerd beschreven in deelhoofdstukken. De belangrijkste methode is de SAM-methode. Hierin gaat er een 'pairwise alignment' gebeuren. Dit is een proces dat omschreven kan worden als het gelijkstellen van bronwaarde en doelwaarde door het gebruik van enkele

toegestane functies. Het doel is hierbij het minimaliseren van de verschillen of het maximaliseren van de gelijkheden tussen de 2 sequenties of activiteitenpatronen.

In hoofdstuk5 worden de algemene kenmerken van de data beschreven die gebruikt worden doorheen de hele masterproef. Deze dataset is opgebouwd uit een aantal persoonskenmerken en bevat eveneens de activiteiten die elke persoon heeft gerapporteerd gedurende een week. De data zijn afkomstig uit dagboeken en enquêtes ingevuld door 325 personen.

In het volgende hoofdstuk worden de analyses uitgevoerd en besproken. Hierin worden 2 verschillende analyses gedaan. De eerste analyse gebeurt op basis van een activiteitenpatroon dat geen rekening houdt met de duur van activiteiten. Bij de 2de analyse wordt de duur wel in rekening gebracht en wordt de dag verdeeld in 96 periodes van 15 minuten. In beide gevallen worden er 3 scenario's opgesteld waarbij aan de activiteiten en locaties in de activiteitenpatronen verschillende gewichten worden toegekend. Er wordt in beide analyses gekozen voor het 3de scenario verder uit te werken wegens de opbouw van de datastructuur. Enkel de modellen van de 2 andere scenario's zouden gebouwd worden om te kijken naar verschillen in opbouw van de modellen.

Dan volgt er een analyse van de gemaakte clusters op basis van de activiteiten en de persoonlijke kenmerken van de personen. Deze worden dan samen nog eens omschreven in een algemene conclusie van de beide analyses. In de laatste stap worden dan de modellen opgebouwd en de socio-demografische profielen bepaald. Het modelalgoritme dat in dit onderzoek wordt gebruikt zijn beslissingsbomen. De belangrijkste kenmerken die bijdrage voor het kiezen van dit modelalgoritme zijn: de eenvoud om het model te begrijpen, kan omgaan met 'missing data' en is goed in de verwerking van categorische variabelen.

De belangrijkste variabelen die gebruikt worden om de beslissingsboom op te stellen zijn het aantal kinderen in een huishouden, de leeftijd (categorisch), het beroep, het inkomen en of de personen een partner hebben. Deze komen ook terug in verschillende andere studies en zorgen dus voor een extra garantie dat de modellen een goede schatting zijn. De socio-demografische profielen zijn niet eenduidig per cluster en worden dus onderverdeeld in meerdere types. Er zijn echter wel enkele gelijkenissen terug te vinden tussen deze verschillende types binnen 1 cluster.

De belangrijkste verschillen die duidelijk worden bij het vergelijken van de uiteindelijke modellen van de 3 scenario's is dat de variabele weekdag minder van belang is als aan de activiteit minder gewicht wordt toegekend.

In hoofdstuk 7 wordt dan een pleidooi gegeven voor toekomstig onderzoek. Het gaat hier vooral om het zoeken naar standaardwaarden om de parameterinstellingen in te stellen.

In hoofdstuk 8 wordt dan in het kort nog een algemene conclusie gegeven. De belangrijkste kenmerken hierin zijn dat de modellen een goede voorspelling maken van de clusters en socio-demografische profielen en er nog een duidelijke nood aan een standaardaanpak is.

Inhoudsopgave

Lijst van figuren

Lijst van tabellen

Inleiding

Om het verplaatsingsgedrag te kunnen begrijpen moeten de verplaatsingen van individuen en de organisatie ervan in tijd en ruimte, niet als loshangende fragmenten worden beschouwd maar als een geheel van een dagelijks activiteitenpatroon. De verplaatsingen zijn immers een afgeleide van de vraag naar activiteiten van individuen. Doordat activiteiten echter beperkt zijn in tijd en ruimte is er een onderlinge relatie tussen de verschillende elementen van een activiteitenpatroon. (Miller, 2003; Rietveld, 1994)

De meeste verkeersmodellen zijn opgesteld op basis van meetpunten langs bepaalde strategische plaatsen. Echter momenteel vindt er een verschuiving plaats in deze verkeersmodellen waarbij er een overstap wordt gemaakt van 4-staps modellen naar activiteitgebaseerde modellen.

Een volgende stap zou kunnen zijn dat deze worden opgebouwd door socio-demografische gegevens. Hierdoor moet er eerst onderzocht worden welke types activiteitspatronen bij deze gegevens horen. Het onderzoek dat wordt behandeld in deze masterproef zal daarover gaan. Op basis van dagboeken, opgesteld door verschillende individuen, zal er een analyse worden gemaakt. In de analyse worden dan homogene groepen van individuen bepaald. Deze gegevens worden dan gekoppeld aan de socio-demografische kenmerken van de individuen om zo te komen tot profielen van individuen met een gelijkaardig activiteiten-verplaatsingspatroon.

1 Literatuurstudie

1.1 Evolutie van transportmodellen

Het onderzoek dat in deze masterthesis wordt behandeld, situeert zich in het domein van transportmodellering. Transportmodellering is daarentegen echter ontstaan uit een vraag vanuit het transportbeleid. Het transportbeleid had vroeger enkel betrekking tot het verminderen van de autotoename op de wegen of toch tenminste er een bepaalde controle over houden. Om dit te bestuderen moesten er dus voorspellingen of schattingen van verplaatsingen gemodelleerd worden. Deze modellen moeten ervoor zorgen dat de beleidsmaatregelen gesimuleerd kunnen worden voordat ze uitgevoerd worden. Eveneens vormen zulke modellen een ondersteuning voor de beleidsmakers met betrekking tot hun genomen beslissingen. (Janssens, 2008)

De aanvankelijke transportmodellen werden behandeld via een standaard methodologische aanpak, beter bekend in de transportwereld als het 4-staps model (Ruiter and Ben-Akiva, 1978). Dit model is zeer geliefd door zijn eenvoud in wiskundige berekeningen en de gemakkelijkheid waarbij de weginfrastructuren kunnen worden aangepast. Doorheen de jaren is er echter al veel kritiek gegeven op het 4-staps model maar door zijn eenvoud en leesbaarheid blijft deze methode nog altijd een geliefde methode bij beleidsmakers. (Wilson, 1967; Ortúzar and Willumsen, 2002)

Nochtans zijn er doorheen de jaren nog factoren bijgekomen die een rol spelen op het vlak van transportbeleid naast enkel het volume van het verkeer, zoals de uitstoot van het verkeer, de congestie, het landgebruik, … (Dijst, 1997) Beleidsmakers proberen dan ook hiervoor maatregelen op te stellen die deze negatieve invloed verminderen of ze tenminste controleren. Deze maatregelen staan beter bekend als 'Travel Demand Management'(TDM-) maatregelen. TDM-maatregelen hebben als doel het verplaatsingsgedrag te veranderden zonder al te veel infrastructurele wijzigingen, het beter benutten van de bestaande transportmodus en het controleren van de negatieve invloeden van het toenemende privé transportgebruik. (Krygsman, 2004) Dit kan op 2 manieren: technologische maatregelen en gedragsaanpassingen. Om deze maatregelen te analyseren en te modelleren is er een nieuwe manier van modelleren nodig. Dit komt door het feit dat het 4-staps model vooral geschikt is om infrastructurele uitbreidingen te simuleren. Hierdoor is er een grote vraag naar modellen die een realistische weergave

van het beslissingsproces van individuen weergeeft en die ook meerdere beleidsmaatregelen kan incorporeren. Daarom is nu de overstap bezig naar activiteitengebaseerde modellen.

1.2 Opbouw van de modellen

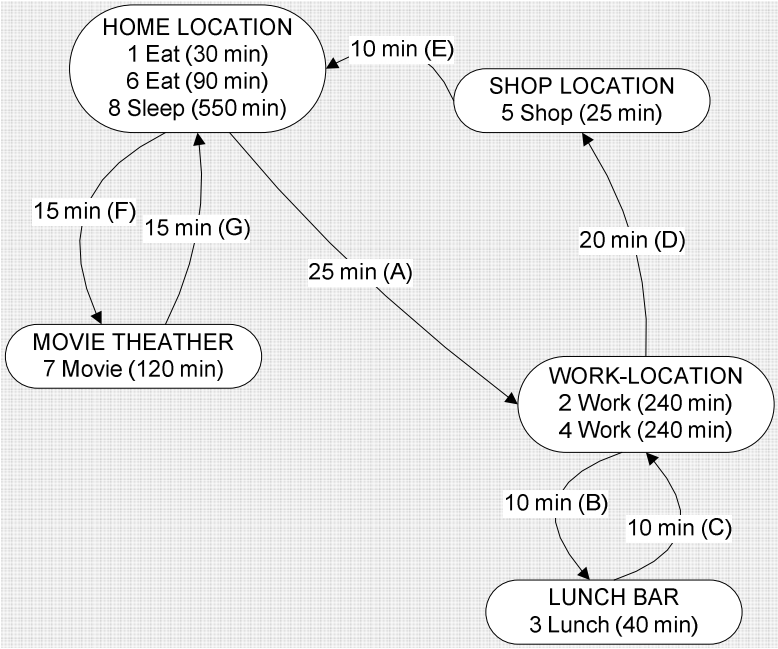
Er worden 3 manieren van modelleren besproken in dit hoofdstuk: tripgebaseerde(4-staps model), tourgebaseerde en activiteitengebaseerde modellen.
De modellen worden dus verder verduidelijkt zodat de verschillen tussen de modellen zichtbaar worden, ze beter te begrijpen zijn en geïnterpreteerd kunnen worden.

Het eerste model dat dan ook besproken wordt is het oudste model maar steeds nog toegepaste 4-staps model. (McNally, 2008).

Het 4-staps model bestaat uit 4 fases/stappen:

- Productie en attractie (trip generation)
- Distributie (trip distribution)
- Keuze vervoersmiddel (mode choice)
- Toedeling (trip assignment)

De routekeuze toedeling in dit model is wel logisch als het gaat om 1 enkele trip maar verplaatsingbeslissingen zoals hoe vaak, naar welke locatie, welk vervoersmiddel, … worden gemaakt voor meer dan 1 trip op een gegeven moment. Ook wanneer onderweg meerdere stops worden gemaakt, zullen deze afhankelijk zijn van de eerder genomen beslissingen. Dit zal in het kort uitgelegd worden via een kleine weergave.



Figuur 1: activiteitenschema

In de figuur is een activiteitenschema weergegeven van 1 dag. Op basis van een tripgebaseerde aanpak kunnen er 7 onafhankelijke trips onderscheiden waarvan 4 thuisgebaseerde trips. Niet-thuisgebaseerde trips worden echter zeer matig gemodelleerd in tripgebaseerde modellen doordat ze moeilijk kunnen gekoppeld worden aan bepaalde verblijfslocaties en de verschillende types van huishouden.

In de tourgebaseerde aanpak kunnen er 2 thuisgebaseerde tours en 1 werkgebaseerde tour onderscheiden worden. 2 tours zijn een simpele 'heen en weer' tour zonder tussenstop(F-G en B-C). Echter bij de 'thuis-werk' tour(A-D-E) vindt er wel een tussenstop plaats op weg naar huis namelijk bij de winkellocatie. De tourgebaseerde modellen zijn al een vooruitgang ten opzichte van tripgebaseerde modellen maar ze missen ook nog de basis van het gedrag tussen de verschillende trips. Bijvoorbeeld als de persoon gaat winkelen in de 'thuis-werk' tour zal deze activiteit minder waarschijnlijk zijn in de 'thuis-cinema' tour of in de 'werk-lunch' tour. Dus een verandering in de ene tour kan een gevolg hebben in een andere tour. (Miller E.J., Roorda M.J. and Carrasco J.A., 2005)

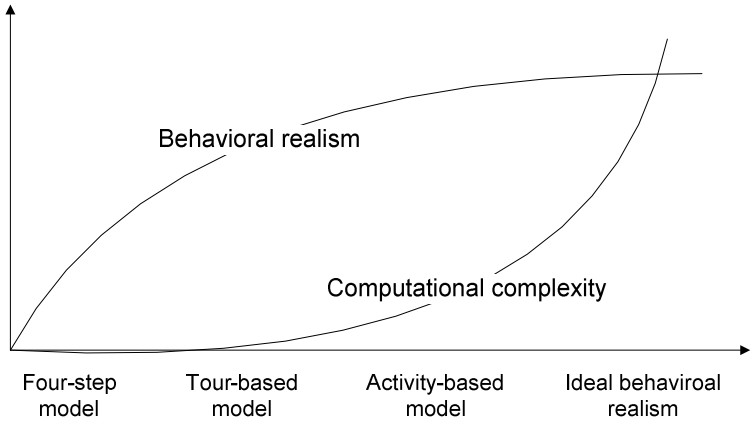
Het derde en laatste model is gebaseerd op een dagelijks activiteitenpatroon. Dit wil zeggen dat alle trips die gemaakt worden tijdens de dag aan elkaar gekoppeld zijn door hetzelfde beslissingproces. Naar dit model wordt meestal verwezen als een activiteitgebaseerde aanpak van modelleren door de gezamenlijke aanpak van de verschillende activiteiten gedurende de dag. (McNally, 2000)

Tabel 1: verschillende aanpakken en bijhorende beslissingen

Aanpak	Model beslissingen	Trips
Trip-based	1 home-based work (HBW) trip	A
	3 non-home-based work (NHBW) trips	B,C,D
	3 home-based other (HB0) trips	E,F,G
Tour-based	1 home-based work tour, with stop on the way home	A,D,E
	1 work-based tour, with no extra stops	B,C
	1 home-based other tour, with no extra stops	F,G
Activity-based	1 primary work tour, with work-based sub-tour, stop on the way home, and secondary non-work tour	A,B,C,D, E,F,G

De laatste jaren komen de activiteitgebaseerde verplaatsingsanalyses meer en meer in de belangstelling te staan ter vervanging van tripgebaseerde verplaatsingsanalyses. Deze vooruitgang komt er door de theoretische aanpak die rekening houdt met: de vraag naar activiteiten, de interrelatie tussen de verschillende trips en de interactie binnen het huishouden. Onderzoek naar activiteitgebaseerde modellen wordt gestimuleerd door de wil om het menselijke gedrag beter te begrijpen en daarbij de beleidsmaatregelen beter te modelleren.

De volgende figuur toont de evolutie van tripgebaseerde modellen(4-staps model) naar meer geavanceerde modellen waarbij er een toename is in complexiteit van het model samen met een betere voorstelling van de werkelijkheid. Op de figuur valt te zien dat naarmate de modellen verbeteren, de complexiteit van het modelleren exponentieel stijgt en het realisme stijgt maar afvlakt naar het einde.



Figuur 2: gedragrealisme en berekeningskracht van de modellen in relatie met de verplaatsingsmodellen

De manier waarop modellen gesimuleerd worden ondergaat dus een verandering waardoor ook een deel van de basisdata veranderen. De verandering van deze basisdata voor de transportmodellen zal dan ook het doel zijn van deze masterthesis.

De modellen worden realistischer en vragen dan ook meer gegevens. Om toch een overzicht te kunnen bewaren en de modellen niet te complex te maken bij de activiteitgebaseerde modellen worden er clusters of groepen ontwikkeld die gemeenschappelijke kenmerken of activiteiten bevatten. Dit komt ook de rekenkracht die het model nodig heeft ten goede. Om deze clusters te kunnen aanmaken zal er gebruik gemaakt worden van een 'string-alignment method' (SAM). Daarna zal er bepaald worden via persoonlijke kenmerken tot welke cluster een individu thuishoort. Mijn onderzoek zal dus een invoer kunnen vormen voor een activiteitgebaseerd model.

2 Opzet van het onderzoek

Door de veranderingen van de modellen in de transportgemeenschap is er ook nood aan nieuwe of aangepaste basisdata. Het oude 4-staps model bestaat slechts uit een trip-gebaseerde aanpak en hierbij komt ook nog eens het feit dat deze modellen meestal enkel piekperiodes beslaan. Zoals eerder vermeld worden vooral de 'thuis-werk' trips gemodelleerd. De informatie die nodig is voor deze modellen is dan ook zeer beperkt.

De meest recent ontwikkelde modellen zijn activiteitgebaseerde modellen. Deze modellen vragen meer gedetailleerde data. Het verzamelen van deze data zal dan ook op een andere manier worden gedaan. De data die nodig zijn voor het aanmaken van deze modellen zijn vooral dagboeken en demografische gegevens. Eveneens zijn ook de persoonlijke gegevens van de ondervraagde individuen in dit onderzoek van belang. Het doel van deze masterproef is op basis van deze dagboeken, opgesteld door verschillende individuen op basis van hun verplaatsingsgedrag, onder te brengen in verschillende types/categorieën. De categorieën bestaan uit de verzameling van verschillende individuen met een gelijkaardig activiteiten-verplaatsingspatroon. Het uiteindelijke doel is het opstellen van socio-demografische profielen op basis van deze activiteiten-verplaastingspatronen.

Hierdoor kan dus op basis van algemene persoonskenmerken een verplaatsingsprofiel gedefinieerd worden. Deze algemene persoonskenmerken zijn vaak al eenvoudig te vinden bij de lokale gemeentes of bij grote volkstellingen uitgevoerd door de overheid. Daardoor zullen de modellen gedetailleerder zijn dan hun voorgangers. Door enkele steekproeven in het studiegebied te doen zal er voldoende informatie aanwezig zijn het model zo waarheidsgetrouw te maken. In dit onderzoek zal dan ook een dergelijke steekproef worden uitgewerkt die later mogelijk in een activiteitgebaseerd model kan ingevoerd worden.

3 Onderzoeksvragen

Om een houvast te creëren waarnaar gestreefd wordt, zullen onderstaande onderzoeksvragen de leidraad vormen gedurende deze masterproef:

- Hoe is de dataset opgesteld en wat zijn de algemene kenmerken?
- Hoeveel significante groepen met een gelijkaardig verplaatsingsprofiel zijn er te vinden?
- Welke zijn de kenmerken van de activiteiten en verplaatsingspatronen van deze groepen?
- Speelt de afstand/locatie een belangrijke in deze patronen en/of per clusters?
- Welke socio-demografische variabelen zijn bepalend voor het toewijzen van individuen aan clusters?
- Kunnen de resultaten bijdrage tot andere onderzoeken?

4 Methodologie

Om de doelstellingen te kunnen bereiken, moet er ook gekeken worden naar de gebruikte methodes. De methodologie die gevolgd gaat worden zal dan ook verduidelijkt worden in dit hoofdstuk.

Vooraleer een methode aan te duiden moet eerst de data goed begrepen worden. Daarom zal eerst de term 'activiteitenpatroon' verduidelijkt worden. Een activiteitenpatroon bevat een sequentie van activiteiten en verplaatsingen. Mensen maken dagelijks beslissingen over hoe hun dag eruit zal zien: in welke volgorde de activiteiten worden uitgevoerd, waar deze activiteiten plaats vinden, welke vervoersmodus gebruikt wordt om de locaties te bereiken, hoe lang de activiteiten duren, … Deze beslissingen zijn gerelateerd. (Gärling T., Gillholm R., Romanus J., and Selart M., 1997)

Wanneer 2 sequenties met elkaar vergeleken worden zullen er wel enkele gelijkenissen aanwezig zijn zoals een gelijkaardige activiteit of locatie maar de positie in de sequentie kan bvb. verschillend zijn. De methodologie die gebruikt wordt in dit onderzoek om sequenties te vergelijken, is gebaseerd op 'string-alignment method'(SAM) (C-H. Joh, T. Arentze, and H. Timmermans, 2007).

In het huidige onderzoek bevat elk patroon echter meer dan één dimensie namelijk, een activiteit en een locatie. Daarom wordt er gebruik gemaakt van een multidimensionele methode om de patronen te vergelijken. Het doel is dus op zoek te gaan naar bepaalde groepen/clusters met een gelijkaardig basispatroon.

Daartoe zal er gebruik gemaakt worden van de volgende fases om tot de clusters of groepen te komen:

- Toepassen van SAM-methode
    a) Wat is SAM-methode
    b) Opstellen van sequenties aan de hand van gegevens verzamelt door dagboeken
    c) Bepalen van parameters
    d) Toekennen van gewichten aan activiteit en locatie

e) SAM-analyse uitvoeren

- Opstellen van afstandsmatrix
- Aanmaken van de gelijkaardige clusters
    a) Op welke methode
    b) Parameters bij het bepalen van het aantal clusters

Nadat deze stappen zijn doorlopen zullen in een volgende fase deze clusters dan worden geanalyseerd naar hun activiteitenstructuur en socio-demografische structuur. Ten laatste wordt er dan nog een model gemaakt om zo te bepalen welke variabele een belangrijke rol spelen zoals leeftijd, geslacht, huishoudsamenstelling,… Op basis hiervan zullen er dan profielen worden opgesteld voor elke cluster.

4.1 SAM-Methode

De oorsprong van deze methode is terug te vinden in de biologie (Sankoff and Kruskal, 1983). De methode meet de afstand tussen 2 strings/sequenties onderling. Om de gegevens van een activiteitenpatroon in een computerprogramma te kunnen verwerken worden deze omgezet naar symbolen. Het doel bij deze methode is het minimaliseren van de verschillen of het maximaliseren van de gelijkheden tussen de 2 sequenties. Dit wordt verwezenlijkt door toevoeging, verwijdering of vervanging van elementen/symbolen in de sequenties. De acties krijgen een bepaalde waarde toegekend en afhankelijk van de grootte ervan wordt dus de afstand groter tussen de 2 sequenties die onderzocht worden. (Wilson C., 2006)

4.1.1 Pairwise alignment

In dynamische programmering termen is 'pairwise alignment' een proces van het gelijkstellen van bronwaarde en doelwaarde door het gebruik van enkele toegestane functies. Het doel zoals eerder beschreven is dus het minimaliseren van de verschillen of het maximaliseren van de gelijkheden tussen de 2 sequenties. De mogelijke functies die toegestaan zijn om dit resultaat te bereiken zijn: toevoeging, verwijdering of vervanging van elementen/symbolen in de sequenties. Elk van deze functies heeft zijn specifieke kost bepaald door de onderzoeker. Exacte gelijkenissen tussen de 2 sequenties hebben

dan ook geen kost. Om dit duidelijker te maken wordt er een voorbeeld van 'pairwise alignment' gegeven aan de hand van de engelse woorden 'true' en 'untrue':

- - t r u e

u n t r u e

De letters 'u' en 'n' moeten verwijderd worden uit het woord 'untrue' om overeen te komen met het woord 'true'. (Wilson, C. 2008) Echter kan er ook gezegd worden dat de letters 'u' en 'n' kunnen worden toegevoegd aan 'true' om hetzelfde resultaat te verkrijgen. De 2 streepjes stellen de spaties voor die nodig zijn om het eerste woord overeen te laten komen met het tweede woord. Alignment algoritmes definiëren een afstand tussen 2 sequenties door het vinden van het pad met de minste kost in een vergelijkingstabel. De vergelijkingstabel is opgebouwd uit één sequentie die boven aan staat en de andere sequentie aan de zijkant, en de cellen van de tabel bevatten waarden afhankelijk van de kost die bij de uitgevoerde functies hoort zoals eerder werd vermeld. De optimale alignment is dan het pad van de cel linksbovenaan tot de cel rechtsonderaan in de tabel met de minste kost.

Tabel 2 geeft de berekening weer voor de alignment van de sequenties 'true' en 'untrue'. De kosten voor de functies (van toevoeging, verwijdering of vervanging van een letter) worden vastgelegd op een strafpunt van waarde 1. Tabel 2 geeft dan de 3 sommen weer die geminimaliseerd worden voor elke cel.

Tabel 2: pairwise alignment voorbeeld met bijhorende celwaardes

		Doelwaarde						
bronwaarde		0	u	n	t	r	u	e
	0	0	1	2	3	4	5	6
	t	1	2,1,2	2,2,3	3,2,4	3,4,5	4,5,6	5,6,7
	r	2	3,2,2	3,2,3	3,3,3	4,2,4	3,4,5	4,5,6
	u	3	4,2,3	3,3,3	4,3,4	4,4,3	4,2,4	3,4,5
	e	4	5,4,3	4,3,4	4,4,4	5,4,4	5,4,3	4,2,4

De waarde in de cel (i,j) wordt berekend volgens volgende methode:

- de celwaarde links (1) vermeerderd met 1 voor de kost van verwijdering van 'u' in de doelwaarde
- de celwaarde links-boven (0) vermeerderd met 1 voor de kost van niet-overeenstemmen van 'u' en 't'
- de celwaarde boven (1) vermeerderd met 1 voor de kost van invoeging van 'u' in de bronwaarde

Aan de hand van bovenstaande tabel wordt dan een 2de tabel aangemaakt met enkel de laagste kost voor elke cel, voorgesteld in Tabel 3. Het minimale pad wordt dan voorgesteld door de grijze kaders die een minimum kost bevatten van 2.

Tabel 3: pairwise alignment voorbeeld met minimum celwaarde

		Doelwaarde						
bronwaarde		0	u	n	t	r	u	e
	0	**0**	**1**	**2**	3	4	5	6
	t	1	1	2	**2**	3	4	5
	r	2	2	2	3	**2**	3	4
	u	3	2	3	3	3	**2**	3
	e	4	3	3	4	4	3	**2**

4.1.2 Toevoeging van locatiegegevens

Om de locatiegegevens te kunnen toevoegen moet er een aangepast algoritme worden opgesteld om de berekeningen te kunnen maken, een multidimensionele algoritme. (Wilson, C. 2008) In deze berekeningen worden de locatiegegevens omgevormd tot euclidische afstanden. In het aangepaste algoritme worden dan de afstandkosten/strafpunten die zijn toegekend door het niet overeenstemmen van de activiteiten (voorgesteld door letters) vervangen door de gewogen som van de afstandkosten, bepaald door een verschillende activiteit, en de locatiekosten, bepaald door een verschillende locatie, voorgesteld door de euclidische afstand tussen de locaties waar de activiteit plaats vind. Dit levert de volgende formule op,

$$q(a_i, b_j) = u * d(a_i, b_j) + v * e(a_i, b_j)$$

waarbij:

- u en v de gewichten zijn waarvan de som gelijk is aan 1,
- $d(a_i,b_j)$ de strafpunten zijn door het niet overeenstemmen van de activiteit,
- $e(a_i,b_j)$ de strafpunten zijn door het niet overeenstemmen van de locatie en,
- $q(a_i,b_j)$ de gewogen som van alle strafpunten.

Uit deze formule kan afgeleid worden dat wanneer de activiteiten identiek zijn en doorgaan op dezelfde locatie, de gewogen som gelijk is aan 0. Als echter alleen de activiteit of locatie identiek is dan wordt de gewogen som bepaald door de strafpunten van de factor die niet overeenstemt. Als de activiteit en locatie verschillend zijn dan

worden beide delen van de formule gebruikt. Een voorbeeld wordt weergegeven in onderstaande tabel.

Tabel 4: voorbeeld berekening van gewichten bij locatiegegevens

	Activiteit(75%) Locatie (25%)	Activiteit(50%) Locatie (50%)	Activiteit(25%) Locatie (75%)
Activiteit identiek Locatie identiek	0,75*0 +0,25*0=0	0,50*0 +0,50*0=0	0,25*0 +0,75*0=0
Activiteit identiek Locatie 4 strafpunten	0,75*0 +0,25*4=1	0,50*0 +0,50*4=2	0,25*0 +0,75*4=3
Activiteit 4 strafpunten Locatie identiek	0,75*4 +0,25*0=3	0,50*4 +0,50*0=2	0,25*4 +0,75*0=1
Activiteit 4 strafpunten Locatie 4 strafpunten	0,75*4 +0,25*4=4	0,50*4 +0,50*4=4	0,25*4 +0,75*4=4

Het locatiebestand bestaat uit x en y coördinaten van alle activiteiten van alle sequenties. Doordat de matrix $e(a_i,b_j)$ bepaald door alle afstanden tussen alle activiteiten van elke personen in een onderzoek ontzettend groot kan zijn is dit soms onoverzichtelijk. Een voorbeeld kan dit verduidelijken. Als een onderzoek 5 000 locaties zou bevatten betekent dit dat de matrix e(i,j) bijna 12 500 000 unieke elementen zou bevatten. Deze matrix zou dan voor elke vergelijking van iedere pairwise alignment gelezen moeten worden. Dit zou echter leiden tot een algoritme wat tergend traag zou werken. Dit is geen conceptueel probleem, maar om toch tot een vlotte uitvoering van de alignments te komen, is het gemakkelijker om te werken met locaties op discrete wijze dan op een continu vlak in een studiegebied. (Wilson, C. 2006)

Het software-programma dat in deze studie gebruikt wordt om de alignments uit te voeren heet 'ClustalTXY' (Wilson, C. 2006). Dit programma definieert een rechthoekig rooster dat het hele studiegebied omvat waarin de activiteiten plaatsvinden. Vervolgens berekent het programma een matrix 'E', dat de Euclidische afstand bevat van de zone waar de activiteit van sequentie 'a' gebeurt tot de zone waar de activiteit van sequentie 'b' gebeurt. Dit zorgt dan voor een kleinere afstandsmatrix die is opgebouwd uit de coördinaten van het middelpunt in elke zone in plaats van de individuele activiteitlocaties.

In het programma dat hier wordt gebruikt zal dit rooster bestaan uit een 5x5 matrix van zones, die de verschillende activiteitlocaties bevatten, en 1 extra zone die buiten het studiegebied valt. De grootte van de matrix 'E' is dan 26x26 voor dit programma en ons studiegebied. (Wilson, C. 2006)

4.1.3 Dagelijkse activiteiten omvormen tot sequenties

De activiteiten die genoteerd werden in de dagboeken, moeten eerst omgevormd worden tot een bruikbare code om de analyse uit te kunnen voeren. Hierbij moeten de activiteiten en hun locatie tot sequenties omgevormd worden die dan gelezen kunnen worden door het programma 'ClustalTXY'. Hierbij worden de activiteiten omgevormd tot een lettercode alsook de locaties. De specifieke codes voor de activiteiten worden weergegeven in bijlage 1. De lettercodes voor de locaties beginnen bij elke persoon terug van vooraf: a is locatie 1, b is locatie 2, c is locatie 3, … Dus locatie a voor persoon x is niet dezelfde locatie voor persoon y. De activiteiten worden eerst vermeld en daarna de bijhorende locatie. Eveneens is het de gewoonte dat de lettercode 'a' gebruikt wordt voor de thuislocatie. (Wilson C., 2007)

Wanneer de activiteiten en locaties zijn omgevormd tot de bijhorende codes worden de sequenties per persoon gevormd. Door middel van een kort voorbeeld zal het werkingsprincipe worden uitgelegd. In onderstaande tabel wordt een klein fragment uit de dataset gebruikt.

Tabel 5: informatie tot aanmaken van sequentie

ID	Activiteit	Start van Activiteit (min.)	Weekdag	Locatie
HH4101GL10027_2	H	390	2	39282
HH4101GL10027_2	W	470	2	39281
HH4101GL10027_2	B	720	2	39289

De sequentie die hier getoond wordt is de sequentie van persoon 'HH4101GL10027' op weekdag 2. De weekdag is dan ook het laatste cijfer van de ID-nummer. De sequentie zal er dan als volgt uitzien:
HaWbBc

Deze procedure zal nu voor iedere persoon uitgevoerd worden en voor elke dag.

4.1.4 Bepalen van parameters

Nadat de sequenties zijn opgesteld kunnen deze in het programma 'ClustalTXY' ingevoerd worden. Daarnaast worden ook de locatiegegevens geïmporteerd om een matrix van 5x5 op te stellen zoals eerder vermeld.

Echter moet er naast de gegevensinvoer ook nog parameters van de pairwise alignment bepaald worden. Een eerste parameter die bepaald moet worden is de 'gap opening'. De standaardwaarde vastgelegd in ClustalTXY is 1.

De tweede parameter is de 'gap extension'. Deze zorgt ervoor dat, hoe langer een opening blijft tussen 2 sequenties, hoe meer strafpunten aan deze opening wordt toegerekend. Deze is standaard bepaald op 0,1.

Om nu de juiste parameters toe te passen wordt de literatuur geraadpleegd. Hiervoor maak ik gebruik van 2 toegepaste bewijzen in de verkeerskundige context en een algemene toepassing van deze parameters.

Verkeerskundige context
In een eerste studie wordt aan de 'gap opening' een waarde van 3 toegekend en aan de 'gap extension' een waarde 1 bij het gebruiken van korte sequenties. Deze parameters worden verantwoord door het feit dat de tijdsduur bij deze sequenties niet wordt meegerekend. De invoeging van een opening ('gap') is dan meestal te wijten aan het feit dat de ene sequentie gedetailleerder is dan de andere sequentie. De openingsstrafpunten voor waar de duur wel in rekening mee gehouden is, zijn hoger. Dit komt doordat de tijd wel meetelt en de openingen die gemaakt worden scores ontwikkelen door het feit dat er een verschil is in activiteiten door de dag. Hierdoor worden in deze sequenties een 'gap opening' van 10 en een 'gap extension' van 5 vastgelegd (Wilson, 1998,a)

In een tweede studie werden aan de 'gap opening' en 'gap extension' respectievelijk een waarde van 5 en 3 toegekend. (Wilson, 1998,b).

Algemene context
In een studie toegepast in het domein van de biologie werden de waarden van de 'gap opening' vastgelegd op 0,4X groter dan de 'gap extension'. (Pons & Vogler, 2006)

Als conclusie heb ik dan besloten om te kiezen voor een 'gap opening' van 5 en 'gap extension' van 3.

4.1.5 Toekennen van gewichten aan locatie en activiteit

In ClustalTXY bestaat de mogelijkheid om gewichten toe te kennen aan de locaties en activiteiten. Hierdoor kan in functie van de belangrijkheid van een bepaald criterium, activiteit of locatie, de afstandsmatrix bepaald worden tussen alle sequenties. Hierdoor zal er onderzoek gedaan worden of er verschil bestaat tussen een evenwichtige vergelijking tussen de activiteit en locatie, en waar één van deze aspecten een meer belangrijkere rol speelt. Voor deze vergelijking te doen worden er 3 sets van gewichten toegepast:

- Scenario 1: activiteit telt voor 50% en locatie ook voor 50%
- Scenario 2: activiteit telt voor 25% en locatie voor 75%
- Scenario 3: activiteit telt voor 75% en locatie voor 25%

4.2 Opstellen van de afstandsmatrix

Nu alle kenmerken van het proces beschreven zijn worden de resultaten besproken. De sequenties zijn nu immers met elkaar vergeleken en hebben hun bijhorend resultaat bekomen. De output van een 'pairwise alignment' in 'ClustalTXY' ziet er als volgt uit.

```
Sequences (   1:   2) %id:   60   Distance:      15  HH121032GL134409 HH121311GL135185
Sequences (   1:   3) %id:    0   Distance:      23  HH121032GL134409 HH121478GL135588
Sequences (   1:   4) %id:    0   Distance:      17  HH121032GL134409 HH121656GL136058
Sequences (   1:   5) %id:    0   Distance:      16  HH121032GL134409 HH122109GL137164
Sequences (   1:   6) %id:  100   Distance:       6  HH121032GL134409 HH122496GL138185
Sequences (   1:   7) %id:    0   Distance:      16  HH121032GL134409 HH122521GL138235
Sequences (   1:   8) %id:  100   Distance:       6  HH121032GL134409 HH122524GL138252
Sequences (   1:   9) %id:   60   Distance:      10  HH121032GL134409 HH122571GL138404
Sequences (   1:  10) %id:   80   Distance:      10  HH121032GL134409 HH122948GL139343
```

Figuur 3: voorbeeld van output uit ClustalTXY

Deze tabel moet dan omgevormd worden tot een afstandsmatrix. Dit is immers nodig om de analyse te kunnen uitvoeren die gaat bepalen van hoeveel clusters er aanwezig zijn. De afstandsmatrix bevat dan ook de sequentie ID's in de rijen en kolommen waarbij de cellen in de matrix de afstanden tussen deze sequenties weergeven. Om dit te verwezenlijken, wordt er gebruik gemaakt van een code/formule in Microsoft Visual

Studio die dit automatisch berekent[1]. Een klein voorbeeld van dit resultaat wordt in de onderstaande tabel weergegeven.

Tabel 6: afstandsmatrix

	1	2	3	4	5	6
1	0	15	23	17	16	6
2	15	0	26	20	19	9
3	23	26	0	14	11	21
4	17	20	14	0	5	15
5	16	19	11	5	0	14
6	6	9	21	15	14	0

Hoe groter de waarde in de cel, hoe meer de sequenties van elkaar verschillen. Doordat elke sequentie gelijk is aan zichzelf, bevat de diagonaal de waarde nul.

4.3 Bepalen van het aantal clusters

Op basis van de afstandsmatrices, die bepaald werden door het gebruiken van het programma 'ClustalTXY', gaan er clusters gevormd. Er zijn een aantal manieren om het aantal clusters te bepalen. De methode die in dit onderzoek gebruikt gaat worden, noemt de fuzzy cluster methode. Deze methode kan toegepast worden in het programma 'R' en krijgt daar de naam 'fanny' (Kaufman L. and Rousseeuw PJ, 1990). De keuze voor 'R' komt doordat het gebruik van het programma mij reeds bekend is door eerder onderzoek en door de lessen gegeven in de 1ste en 2de master verkeerskunde van deze opleiding. Bovendien is deze gratis te verkrijgen en gemaakt voor onderzoek en statistische analyses.

De fuzzy cluster analyse houdt in dat elke observatie of sequentie wordt toegewezen aan een cluster op basis van hun lidmaatschapwaarden. De lidmaatschapswaarden tot een cluster kunnen niet negatief zijn en variëren tussen de waarde 0 en 1. De som van alle lidmaatschapswaarden van één observatie of sequentie moeten samen 1 vormen. Dus als voorbeeld waarbij er 3 clusters zijn, kan een observatie voor 30% toegewezen worden aan cluster 1, voor 20% aan cluster 2 en voor 50% aan cluster 3.

[1] De code is aangemaakt door mijn begeleider: Marlies Vanhulsel.

Er is echter 1 probleem: de methode berekent niet automatisch het optimaal aantal clusters. Hierdoor moeten er parameters onderzocht worden zoals de 'Dunns' Validity Index' en de 'Silhouette Validation Method'.

4.3.1 Dunns' Validity Index

De dunn's validiteit index(Dunn, 1974) is een eerste parameter om het optimale aantal clusters te bepalen. Het idee achter deze theorie bestaat uit het identificeren van clusters die homogeen zijn in hun eigen elementen en heterogeen in vergelijking met andere clusters. Dit betekent dat de intracluster afstanden klein moeten zijn, d.w.z. sequenties in eenzelfde cluster lijken zeer sterk op elkaar, en dat de intercluster afstanden groot moeten zijn, d.w.z. sequenties in verschillende clusters zijn significant verschillend.

Voor elke verdeling in clusters, waarbij c_i de i-de cluster voorstelt van zo'n patroon, kan de dunn's validiteit index berekend worden aan de hand van de volgende formule,

$$D = \min_{1 \le i \le n} \left\{ \min_{\substack{1 \le j \le n \\ i \neq j}} \left\{ \frac{d(c_i, c_j)}{\max_{1 \le k \le n} \{d'(c_k)\}} \right\} \right\}$$

waarbij:

- $d(c_i,c_j)$ de afstand is tussen clusters c_i en c_j (intercluster afstanden)
- $d'(c_k)$ de intracluster afstanden zijn van cluster c_k en
- n het aantal clusters.

Het uiteindelijke doel is dan het maximaliseren van de intercluster afstanden en minimaliseren van de intracluster afstanden. Hieruit volgt dat het aantal clusters dat D maximaliseert gekozen wordt als het optimale aantal clusters.

Om het effect van het aantal clusters dat er vooraf bepaald wordt te minimaliseren, wordt de genormaliseerde dunn's coëfficiënt berekend. Deze coëfficiënt wordt berekend aan de hand van de volgende formule:

$$FN_m = \frac{F_m - (1/m)}{1 - (1/m)} = \frac{mF_m - 1}{m - 1}$$

waarbij:

- FN_m de genormaliseerde dunn's coëfficiënt is,

- F_m de dunn's coëfficiënt is van het aantal cluster en,
- M het aantal clusters dat er gekozen is.

Deze waarden bevinden zich tussen 0 en 1 waarbij het aantal clusters dat er gekozen worden geen belang spelen. De waarde '0' stelt een totale verspreiding over de verschillende clusters voor en de waarde '1' een totale toewijzing van iedere sequentie tot een bepaalde cluster.

4.3.2 Silhouette Validation Method

De tweede parameter die kan gebruikt worden voor het kiezen van het optimale aantal clusters, is de Silhouette Validation Method (Rousseeuw, 1987). Deze techniek bestaat uit het berekenen van de silhouettebreedte voor iedere sequentie en de gemiddelde silhouettebreedte voor ieder cluster en voor de totale dataset. In dit geval zal enkel de gemiddelde silhouettebreedte voor de hele dataset worden gebruikt. De silhouettebreedte, S(i), wordt berekend door de volgende formule:

$$S(i) = \frac{(b(i) - a(i))}{\max\{a(i), b(i)\}}$$

waarbij:

- a(i) het gemiddelde verschil is van het i-de object (sequentie) tot alle andere objecten in dezelfde cluster en
- b(i) het gemiddelde minimum verschil is van het i-de object (sequentie) tot alle andere objecten in de andere cluster (de kortbij gelegen cluster).

De gemiddelde silhouettebreedte voor de hele dataset is het gemiddelde van alle S(i) objecten in de dataset. Het optimale aantal cluster wordt dan bepaald door de maximale gemiddelde silhouettebreedte voor de hele dataset.

4.3.3 Minimaal en maximaal aantal clusters

Doordat bij de functie 'fanny' in het programma 'R' het aantal clusters op voorhand moet bepaald zijn, worden in hier minimum en maximum waarden vastgelegd. Bij het onderzoeken naar het optimale aantal clusters per scenario is dan gekozen om minimaal 3 clusters te creëren en maximaal 10 clusters. Minimaal 3 om toch voldoende verschil

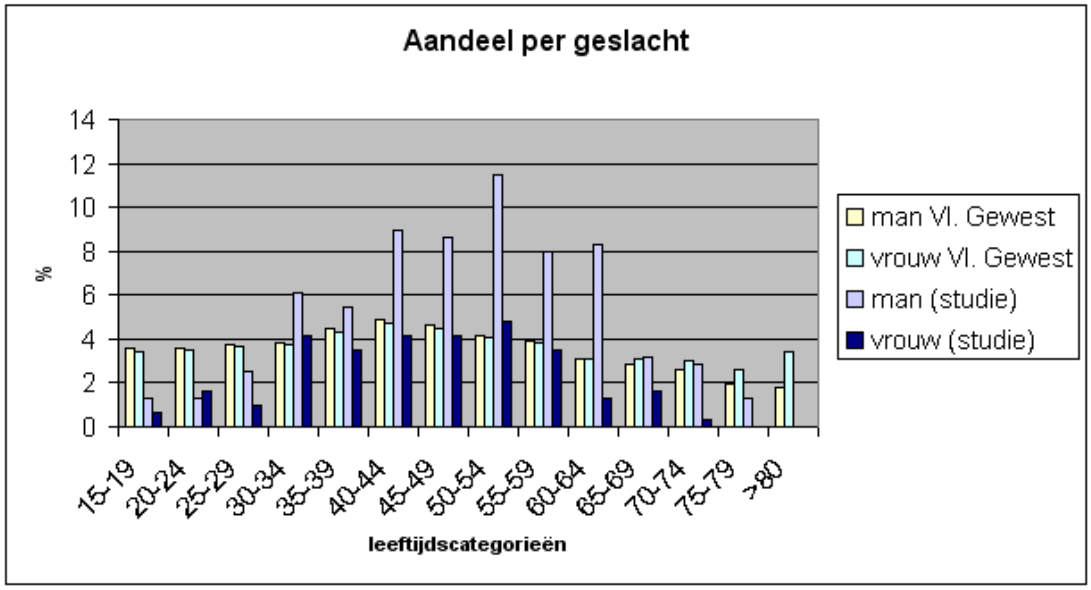
tussen de clusters te creëren. De keuze voor maximaal 10 clusters te analyseren komt door het feit dat de clusters geen waarden kunnen bevatten of slechts zeer kleine aantallen wanneer het aantal clusters stijgt.

5 Algemene data

In dit hoofdstuk worden de data beschreven waarmee in de masterproef zal gewerkt worden. De dataset die voorhanden is, is opgebouwd uit een aantal persoonskenmerken en bevat eveneens de activiteiten die elke persoon heeft gerapporteerd gedurende een week.

5.1 Persoonlijke kenmerken

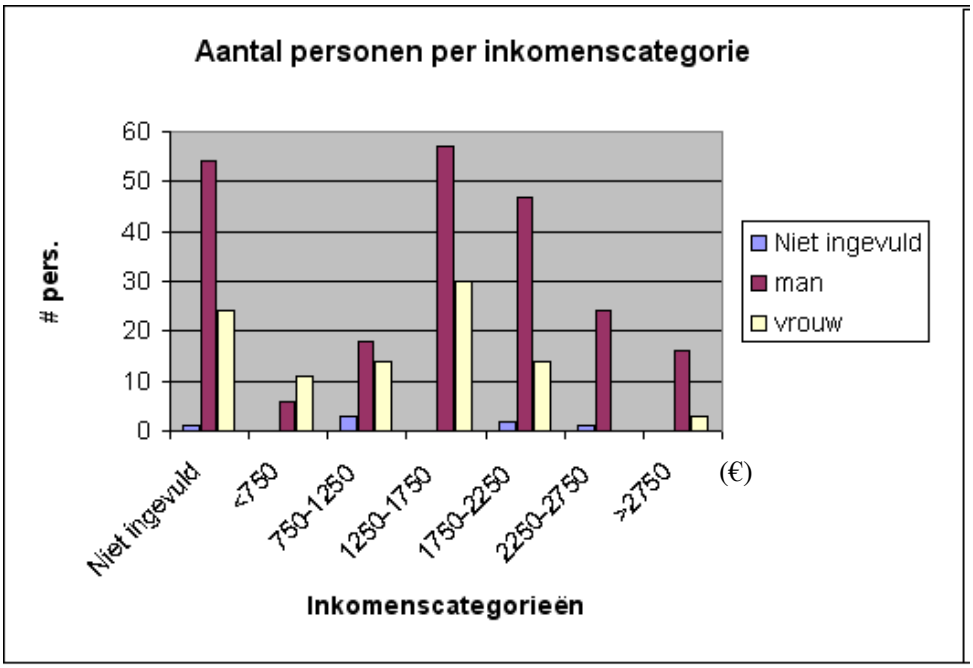
De volgende persoonskenmerken werden meegenomen in het onderzoek: het geboortejaar, het geslacht, de inkomenscategorie, aantal werkuren, aantal kinderen, partner, huishoudgrootte, rijbewijs, huwelijksstatus, diploma en beroep. Aan de hand van deze kenmerken kan er al een globaal beeld worden gegeven van hoe de dataset is opgesteld. De enquête werd door 325 personen ingevuld. Van deze 325 personen waren er 222 mannen en 96 vrouwen. 7 personen deelden echter hun geslacht niet mee. Verder hebben 316 van de ondervraagden een rijbewijs. In het totaal hebben 139 bevraagde personen hun aantal werkuren niet meegedeeld. Voor een aantal kan dit te verklaren zijn doordat ze niet gaan werken momenteel, thuis blijven om het gezin te onderhouden, nog naar school gaan, gepensioneerd zijn,… Als verondersteld wordt dat een persoon vanaf 35 werkuren per week een voltijds baan heeft dan werken er 49 deeltijds en 137 voltijds. Uit verdere gegevensanalyses zijn de volgende resultaten gekomen. De eerste figuur geeft het aandeel personen per geslacht en per leeftijdscategorie weer. De gegevens van het Vlaamse gewest zijn verkregen door raadpleging van het Nationaal Instituut voor de Statistiek.



Figuur 4: verhouding van de leeftijdscategorieën in Vlaanderen en enquête

De meeste personen van onze studie bevinden zich in de leeftijdscategorie van 50-54 jaar. De jongste en oudste leeftijdscategorieën zijn minder aanwezig in dit onderzoek. De beperkte aanwezigheid van respondenten in de leeftijdscategorie '–18 jaar' is te verklaren doordat deze leeftijdscategorie meestal niet meegenomen wordt in dergelijke onderzoeken. Wat duidelijk opvalt, is dat de mannen veel meer vertegenwoordigd zijn dan de vrouwen in onze studie. Echter het aandeel bij de vrouwen tussen de 30 en 60 jaar is gelijkaardig aan het aandeel van het Vlaamse gewest. Dit betekent dan vooral dat buitenste leeftijdscategorieën vervangen zijn door een teveel aan mannen in de leeftijdscategorieën tussen de 30 en 60 jaar.

De tweede figuur die gemaakt werd laat de verhoudingen zien van het aantal personen per geslacht en per inkomenscategorieën.



Figuur 5: aantal personen per inkomenscategorie enquête

Uit bovenstaande figuur blijkt dat 79 personen hun inkomenscategorie niet hebben ingevuld. De grootste groep van de ondervraagde (±1/4), 57 mannen en 30 vrouwen, vallen binnen de inkomenscategorie van €1250-€1750. De kleinste groep van de ondervraagde zijn terug te vinden in de kleinste en hoogste inkomenscategorieën met respectievelijk 17 en 19 personen.

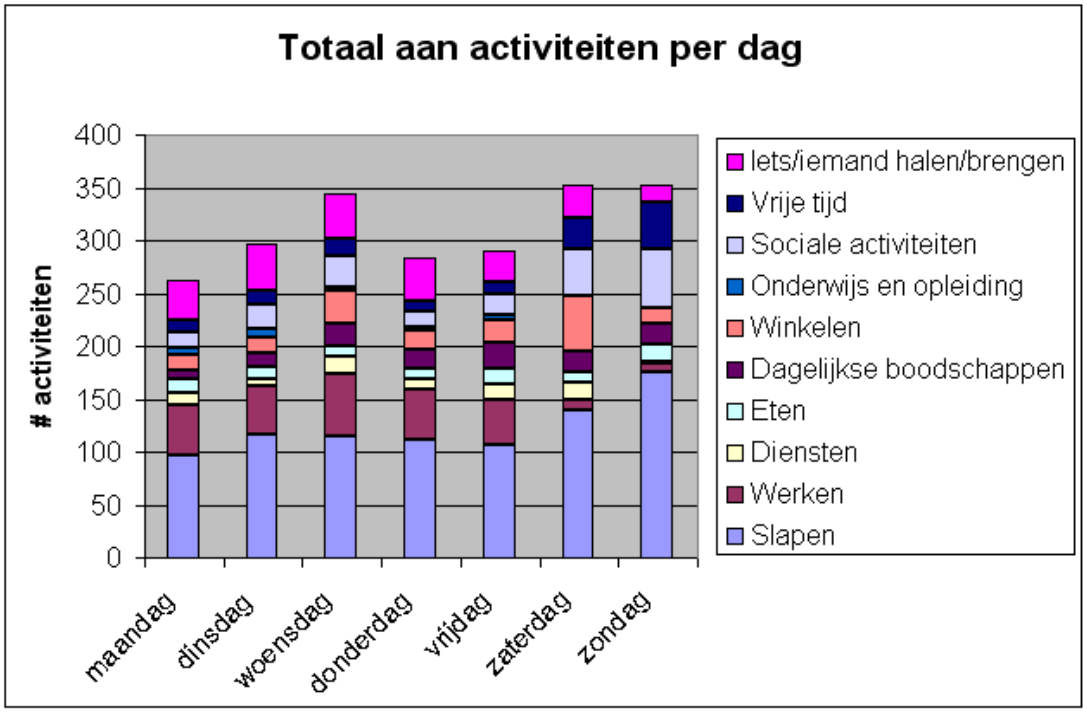
5.2 Dagboekdata kenmerken

In dit onderdeel worden de algemene kenmerken van de dagboeken van alle personen tezamen geanalyseerd. Om de grafiek overzichtelijker te maken worden de thuisactiviteiten afzonderlijk weergegeven in een tabel. De overige activiteiten worden weergegeven in de onderstaande figuur.

Tabel 7: aantal thuisactiviteiten per weekdag

	maandag	dinsdag	woensdag	donderdag	vrijdag	zaterdag	zondag
Thuisactiviteiten	510	492	575	521	500	626	683

De thuisactiviteiten zijn vooral aanwezig in het weekend. Ook zijn er meer thuisactiviteiten op woensdag in vergelijking met de andere weekdagen.
De slaapactiviteit is ongeveer gelijk aan de thuisactiviteit. Wat hiermee bedoeld wordt, is dat deze activiteit ook dezelfde distributie behoudt als de thuisactiviteiten. De slaapactiviteit wordt dan ook het meeste uitgevoerd in het weekend. De slaapactiviteiten komen ongeveer 109 keer voor per weekdag. De 3de hoofdactiviteit is de werkactiviteit die gemiddeld 48 keer voor komt tijdens de weekdagen. Eveneens komt deze activiteit ook voor tijdens het weekend maar zeer beperkt met respectievelijk 9 en 8 activiteitmeldingen. In het weekend vinden er dan weer meer winkelactiviteiten en activiteiten met een sociaal karakter plaats. De winkelactiviteiten zijn het populairste op een zaterdag gevolgd door vrijdag en woensdag. Ook nog kenmerkend is dat door de week meer activiteiten worden geregistreerd dan in het weekend.



Figuur 6: aantal activiteiten per weekdag

5.3 Locatiegegevens

In het onderzoek zijn 2072 locaties opgenomen. Op onderstaande kaart worden de meest uiterste punten van ons onderzoeksgebied weergegeven. Deze punten vormen dan ook de grenzen van het rooster dat wordt bepaald voor de analyse. Dit gebied zal dus in 25 cellen worden opgedeeld.

Figuur 7: uiterste punten in het onderzoeksgebied

De punten stellen volgende gemeenten voor:

- rood: De Panne
- geel: Hoogstraten
- blauw: Maaseik
- groen: Halle

6 Analyse van de dataset

In dit hoofdstuk wordt een onderscheid gemaakt tussen 2 types van sequenties. De eerste sequentie wordt beschreven als activiteitensequentie of korte sequentie. Deze sequentie bevat alle activiteiten die een persoon doet achtereenvolgens. De 2de sequentie is de dagsequentie of lange sequentie en omvat ook alle activiteiten van een persoon maar er is nu rekening gehouden met de duur van de activiteiten. De sequentie wordt onderverdeeld in periodes van 15 minuten. Een dagsequentie bestaat dus minimum uit 96 periodes. Mits er sommige korte verplaatsingen van minder dan 15 minuten in sequenties voorkomen kan het zijn dat deze sequenties meer dan 96 periodes bevatten. Deze verplaatsingen moeten toch worden opgenomen omdat ze noodzakelijk zijn om de correctheid van de sequentie te bewaren.

6.1 Activiteitensequentie (korte sequentie)

6.1.1 Parameters toegepast

Wanneer het aantal clusters bepaald wordt door de functie 'fanny' in het softwareprogramma 'R' komen enkele waarden naar voor die worden vergeleken met de eerder bepaalde parameters. Omdat deze waarden berekend werden volgens het toegewezen gewicht aan de activiteit en de locatie (3 scenario's) op de hele dataset zijn er 3 resultaten.

Scenario 1(50act.-50loc.)

Tabel 8: clusterparameters voor scenario 1 bij de activiteitensequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,3643102	0,0464653	0,3589693
4	0,2734954	0,0313272	0,3435179
5	0,2847499	0,1059374	0,3730723
6	0,2871762	0,1446114	0,3761072
7	0,2281312	0,0994863	0,3520182
8	0,1814612	0,0645271	0,2737149
9	0,3085036	0,2220666	0,4691736
10	0,2947385	0,2163761	0,4585543

De parameters die worden gebruikt of geanalyseerd zijn de genormaliseerde dunn's coëfficiënt en gemiddelde silhouettebreedte. Voor beide parameters geldt dat de hoogste waarde de optimale cluster aanduidt. Voor de genormaliseerde dunn's coëfficiënt is het optimale aantal clusters gelijk aan 9. De optimale keuze voor het aantal clusters bij de gemiddelde silhouettebreedte parameter is 9. De keuze is dus ook gevallen voor het aanmaken van 9 clusters omdat dit aantal de beste waarde van beide combineert.

Scenario 2(25act.-75loc.)

Tabel 9: clusterparameters voor scenario 2 bij de activiteitensequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,3843285	0,0764927	0,3667477
4	0,2904183	0,0538911	0,3464681
5	0,2850979	0,1063724	0,4333232
6	0,269609	0,1235308	0,3965194
7	0,2531949	0,1287274	0,3819216
8	0,2650638	0,1600729	0,3720034
9	0,2520491	0,1585553	0,3772783
10	0,2393404	0,1548227	0,3792509

Het optimale aantal clusters volgens de genormaliseerde dunn's coëfficiënt is gelijk aan 8. De beste keuze voor het aantal clusters bij de gemiddelde silhouettebreedte is 5. De lichte stijging van de waarden bij de gemiddelde silhouettebreedte bij hogere cluster aantallen ligt in het feit dat hoe kleiner de clusters zijn, hoe meer dat de gevallen per cluster op elkaar gelijken, en dus hoe hoger de score hier wordt. Hierdoor wordt er meer belang gehecht aan de Dunn waarden in dit werk dan aan de silhouettewaarden. De keuze is dus gevallen voor het aanmaken van 8 clusters.

Scenario 3(75act.-25loc.)

Tabel 10: clusterparameters voor scenario 3 bij de activiteitensequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,4070162	0,1105242	0,3521779
4	0,3257992	0,1010657	0,3066248
5	0,2818433	0,1023041	0,2758154
6	0,2503327	0,1003992	0,2697295
7	0,2259901	0,0969885	0,187798
8	0,2651093	0,1601249	0,2880958
9	0,248315	0,1543544	0,3372628
10	0,2252708	0,1391898	0,335847

Als het optimale aantal clusters zou gekozen worden op basis van de genormaliseerde dunn's coëfficiënt dan zouden er 8 clusters aanwezig moeten zijn. Het aantal clusters bij de gemiddelde silhouettebreedte dat optimaal geacht wordt, is 3. De keuze is echter gevallen voor het aanmaken van 8 clusters. Dit wordt verklaard doordat de genormaliseerde dunn's coëfficiënt bij 8 clusters het hoogste is en een goede gemiddelde silhouettebreedte heeft. Tevens is er ook in de vorige scenario's telkens gekozen voor de beste genormaliseerde dunn's coëfficiënt. Deze lijn wordt hier dan ook verder doorgetrokken.

Conclusie:

Na de bovenstaande scenario's te vergelijken lijkt het beste dat de analyse zal gebeuren op basis van scenario 3. Het argument voor het kiezen van dit scenario bevindt zich in de eerder bepaalde roosterstructuur. In onderstaande tabel wordt de matrix die op de kaart van Vlaanderen wordt toegepast getoond.

Tabel 11: lettercodes bij toepassing van een locatie matrix

E	**J**	**O**	**T**	**Y**
D	**I**	**N**	**S**	**X**
C	**H**	**M**	**R**	**W**
B	**G**	**L**	**Q**	**V**
A	**F**	**K**	**P**	**U**

Als deze matrix op Vlaanderen wordt gelegd dan blijven er grote gebieden over waarin de meeste mensen hun activiteiten blijven uitvoeren. De verplaatsing voor dezelfde activiteiten zal weinig uitmaken sinds de afstanden relatief gezien worden. Om dit verder te ondersteunen worden er een vergelijkende sequentie getoond die dit bevestigt. Er wordt slechts 1 vergelijking getoond maar er bevinden zich zo meerdere in de hele dataset.

Tabel 12: vergelijking van dezelfde activiteiten op een verschillende locatie

Sequentie	Activiteitenpatroon	Scenario 1	Scenario 2	Scenario 3
1	SrHrLrHrSr	/	/	/
251	SnHnLnHnSn	24	16	8

Uit de tabel blijkt dat de activiteiten in dezelfde volgorde worden uitgevoerd maar op een andere locatie in het rooster. De personen voeren dus dezelfde handeling uit op een dag maar doen de activiteiten op een andere locatie. Als voorbeeld kan dus beschouwd worden dat een persoon in West-Vlaanderen hetzelfde activiteitenprofiel heeft als een persoon in Limburg.

Eveneens geldt voor beide activiteitenpatronen dan ook dat de locatie binnen 1 cel van het rooster blijft. Dit valt te verklaren door de grootte van de roosterstructuur zoals eerder vermeld. De kans is dus ook zeer groot dat de activiteiten van één persoon binnen dezelfde cel van het rooster zullen afspelen. Daarom is het misschien beter om meer aandacht te geven aan de activiteiten zelf en minder aan de locatie waar deze activiteiten doorgaan.

Een diagonale kijk over de hele dataset lijkt de vorige veronderstelling te bevestigen waarbij de meeste activiteiten van iedere sequentie zich in dezelfde cel afspelen. Hierdoor heb ik dan besloten om verder te werken met scenario 3.

6.1.2 Analyses van de clusters op basis van activiteiten

De analyses van de clusters die besproken zullen worden gaan over scenario 3 waarbij een gewicht van 75% wordt toegekend aan de activiteit. De gegevens per cluster zijn berekend per persoon en staan ook per type activiteit geordend. Zo worden het aantal activiteiten per cluster getoond en ook hoeveel personen deze activiteit doen. Personen kunnen in meerdere clusters voorkomen doordat de activiteiten per dag zijn ingedeeld. Dit wil zeggen dat bijvoorbeeld maandag van persoon X in cluster2 zit en woensdag in cluster4. Procentueel wordt dan nog weergegeven hoeveel personen per cluster een bepaalde activiteit doen. Uiteindelijk worden dan ook nog enkele gemiddelden weergegeven als de activiteit door een persoon wordt uitgevoerd. Deze resultaten hebben betrekking tot: het gemiddelde aantal activiteiten, de gemiddelde duur van een activiteit, de gemiddelde totale duur van een activiteit en de gemiddelde reistijd die personen doen voor een activiteit uit te voeren.

Tabel 13: Cluster1 van de activiteitensequentie

Cluster1	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	728	107	98	6,80	211	1436	26
Slapen	172	82	75	2,10	468	982	5
Werken	53	33	30	1,61	418	671	22
Diensten	10	9	8	1,11	83	92	13
Eten	12	10	9	1,20	84	101	10
Dagelijkse boodschappen	25	20	18	1,25	35	44	7
Winkelen	28	25	23	1,12	44	49	9
Onderwijs en opleiding	1	1	1	1,00	15	15	0
Sociale activiteiten	29	23	21	1,26	158	199	11
Vrije tijd	30	22	20	1,36	163	222	12
Iets/iemand halen/brengen	15	12	11	1,25	15	19	11
Totaal	1103	109		20,06			

In deze cluster bevinden zich 109 personen van wie bijna 1/3 een werkactiviteit doet. Dagelijkse boodschappen, winkelen, sociale activiteiten en vrije tijd worden door 1/5 van de cluster gedaan. Het verschil in tijdsbesteding aan sociale activiteiten en vrije tijd is niet groot, slechts 5 minuten. Dit verschil is ook tussen de 2 overige vermelde activiteiten, dagelijkse boodschappen en winkelen, op te merken. Echter komen de dagelijkse boodschappen meer voor dan winkelen. Echter is de tijd die aan deze activiteiten wordt besteed minder.

Gemiddeld zijn er ongeveer 7 thuisactiviteiten per persoon. De totale duur van de thuisactiviteit bedraagt 24 uur of ±3,5 uur per uitgevoerde thuisactiviteit. De personen die in deze cluster gaan werken, spenderen er per activiteit ±7 uur aan. De langst durende verplaatsing in deze cluster wordt ook geleverd voor het uitvoeren van de werkactiviteit. De activiteit 'iets/iemand halen/brengen' wordt slechts door 11% van de cluster gedaan.

Tabel 14: Cluster2 van de activiteitensequentie

Cluster2	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	447	73	97	6,12	176	1078	24
Slapen	114	63	84	1,81	445	805	20
Werken	43	20	27	2,15	437	940	22
Diensten	12	9	12	1,33	85	113	15
Eten	16	13	17	1,23	116	143	14
Dagelijkse boodschappen	28	21	28	1,33	26	35	10
Winkelen	25	19	25	1,32	54	71	13
Onderwijs en opleiding	5	2	3	2,50	231	578	33
Sociale activiteiten	34	23	31	1,48	126	186	14
Vrije tijd	26	21	28	1,24	191	236	13
Iets/iemand halen/brengen	23	15	20	1,53	13	20	18
Totaal	773	75		22,05			

De cluster is opgebouwd door 75 personen en 97% van deze personen hebben een thuisactiviteit gedaan. De thuisactiviteit wordt gemiddeld 6 keer gedaan met een tijdsbesteding van 3 uur per uitgevoerde activiteit. Een verschil met de vorige cluster is dat de sociale activiteiten, vrije tijd en dagelijkse boodschappen in deze cluster meer vertegenwoordigd zijn. Hierbij voert 1 persoon op 3 een vorige vermelde activiteit uit. Het verschil in tijd tussen winkelen en dagelijkse boodschappen is groter geworden in relatie met de vorige cluster. Er is nu een verschil van ±30 minuten op te merken.

De werkactiviteit in deze cluster wordt door een gelijkaardig aantal personen uitgevoerd in verhouding met de totale clusterpopulatie. Echter is er een verschil op te merken in de totale duur van de activiteit en ook in de uitvoering ervan. De werkactiviteit wordt gemiddeld 2x uitgevoerd en duurt daardoor langer in het totaal. De gemiddelde tijd besteed aan een werkactiviteit is ongeveer gelijk in beide clusters.

Een andere opmerking is dat 1/5 van deze cluster de activiteit 'iets/iemand halen/brengen' doet. In de vorige cluster bedroeg dit maar 10%.

Tabel 15: Cluster3 van de activiteitensequentie

Cluster3	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	679	82	100	8,28	151	1250	31
Slapen	143	69	84	2,07	458	949	13
Werken	70	31	38	2,26	273	616	24
Diensten	42	24	29	1,75	43	75	14
Eten	36	18	22	2,00	62	124	13
Dagelijkse boodschappen	41	27	33	1,52	23	35	12
Winkelen	70	40	49	1,75	39	68	13
Onderwijs en opleiding	16	6	7	2,67	168	448	29
Sociale activiteiten	83	39	48	2,13	107	228	19
Vrije tijd	39	25	30	1,56	131	204	14
Iets/iemand halen/brengen	169	49	60	3,45	16	55	32
Totaal	1388	82		29,43			

De derde cluster bevat 82 personen. De thuisactiviteit in deze cluster wordt gemiddeld 8x uitgeoefend. Dit is duidelijk meer dan in de vorige clusters. De duur van deze activiteit is echter minder lang in deze cluster. Daardoor is de totale tijdsbesteding bij deze cluster minder dan in de eerste cluster maar meer dan in de tweede cluster.

De sociale activiteiten en winkelen wordt door de helft van cluster beoefend. Ze worden beide ongeveer een 2-tal keer gedaan. Dit is beduidend meer dan in de 2 vorige clusters. De tijdbesteding van sociale activiteiten is in deze cluster eveneens minder lang. Dit kan te verklaren zijn door het feit dat de totale activiteit in deze cluster meer zijn opgesplitst in meerdere activiteiten. Echter is de totale duur van de sociale activiteit in tegenstelling tot de thuisactiviteiten in deze cluster langer dan in de vorige clusters.

De werkactiviteit heeft in deze cluster eveneens een redelijk aandeel in de gedane activiteiten per persoon. Ze vindt meer dan 2x plaats maar duurt slechts 4,5 uur per activiteit. De totale werkduur komt dan meer overeen met de eerste cluster dan die van de tweede cluster.

De activiteiten ‘diensten’ en ‘eten’ komen voor het eerste duidelijk naar voren in een cluster. Ze worden dan ook door bijna ¼ van de cluster gedaan en worden meer uitgeoefend dan in de vorige clusters.

Een ander verschil ten opzichte van de vorige cluster is dat de activiteit ‘iets/iemand halen/brengen’ door 60% van de cluster wordt gedaan. In deze situatie is er eveneens

geen verschil op te merken in de gemiddelde duur per activiteit. Echter is dit wel het geval bij het aantal keer dat de activiteit wordt uitgeoefend.

Tabel 16: Cluster4 van de activiteitensequentie

Cluster4	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	521	81	100	6,43	203	1306	19
Slapen	123	69	85	1,78	452	806	5
Werken	28	14	17	2,00	445	890	21
Diensten	6	5	6	1,20	79	95	15
Eten	6	6	7	1,00	101	101	0
Dagelijkse boodschappen	13	13	16	1,00	50	50	10
Winkelen	19	16	20	1,19	73	87	19
Onderwijs en opleiding	1	1	1	1,00	165	165	5
Sociale activiteiten	25	21	26	1,19	120	143	16
Vrije tijd	15	11	14	1,36	191	260	19
Iets/iemand halen/brengen	8	7	9	1,14	46	53	20
Totaal	765	81		19,30			

Cluster4 bevat alle activiteiten die er zijn opgenomen in het onderzoek. De cluster bestaat uit 81 personen en ze nemen allen deel in thuisactiviteiten. De thuisactiviteiten van deze cluster komen dan ook het beste overeen met die van cluster1. Het percentage personen die een werkactiviteit doen, is de laagste tot nu toe gemeten in alle clusters, namelijk 17%. De opbouw van de werkactiviteit is wel gelijkaardig met die in cluster2.

Enkel winkelen en sociale activiteiten worden door meerdere personen uitgevoerd in vergelijking tot de werkactiviteit. Er is echter een verschil op te merken in verband met deze twee activiteiten als de vergelijking ten opzichte van de andere clusters wordt gedaan. De totale duur van de sociale activiteiten in deze cluster is veel lager dan in de totale duur in de 3 voorgaande clusters. Een omgekeerde situatie vindt plaats bij de activiteit ‘winkelen’. De totale duur van winkelen alsook de duur van 1 winkelactiviteit is hoger in deze cluster in vergelijking met de eerder vermelde clusters.

Tabel 17: Cluster5 van de activiteitensequentie

Cluster5	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	336	63	100	5,33	153	816	23
Slapen	72	49	78	1,47	468	688	10
Werken	64	24	38	2,67	350	933	18
Diensten	9	7	11	1,29	74	95	5
Eten	15	11	17	1,36	92	125	25
Dagelijkse boodschappen	20	17	27	1,18	14	16	8
Winkelen	20	14	22	1,43	59	84	13
Onderwijs en opleiding	3	2	3	1,50	176	264	10
Sociale activiteiten	33	25	40	1,32	162	214	13
Vrije tijd	25	17	27	1,47	184	271	12
Iets/iemand halen/brengen	25	17	27	1,47	13	19	16
Totaal	622	63		20,48			

Cluster5 is de laatste cluster waarbij alle activiteiten worden beoefend. De cluster is samengesteld door 63 personen die allemaal de thuisactiviteit uitvoeren. De thuisactiviteit is niet zo talrijk aanwezig als in de vorige cluster waardoor ook het aantal keer dat deze activiteit wordt uitgevoerd, daalt. Ze wordt namelijk een 5-tal keer beoefend door een persoon en duurt in het totaal maar ±13,5 uur.

De sociale activiteiten en werkactiviteiten worden door 2 op 5 personen uitgevoerd in deze cluster. De activiteit 'werken" heeft het hoogste gemiddelde aantal uitgevoerde werktaken overheen alle clusters, namelijk zo'n 2,67 keer. In het totaal besteedt een persoon hier 15,5 uur aan in het totaal.

Dagelijkse boodschappen, vrije tijd en iets/iemand halen/brengen worden door 1/5 van de cluster gedaan. De totale tijd die een persoon besteed aan vrije tijd is de hoogste in deze cluster van alle clusters. De duur van de dagelijkse boodschappen daarentegen zijn de laagste van alle clusters. Ze duren slechts 15 minuten. De meeste clusters besteden er echter een minimum van 30 minuten aan. Het patroon van de activiteit 'iets/iemand halen/brengen' is gelijkaardig in cluster1 en cluster2.

Tabel 18: Cluster6 van de activiteitensequentie

Cluster6	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	551	67	100	8,22	205	1686	21
Slapen	114	56	84	2,04	467	951	10
Totaal	665	67		10,26			

Dit is de eerste cluster waar er enkel thuisactiviteiten worden uitgevoerd naast slapen en is opgebouwd uit 67 personen. Ze worden dan ook door alle personen in de cluster gedaan. De gemiddelde duur besteed aan 1 thuisactiviteit is ±3,5 uur en wordt gemiddeld 8 keer uitgeoefend.

Tabel 19: Cluster7 van de activiteitensequentie

Cluster7	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	632	62	100	10,19	168	1713	36
Slapen	120	56	90	2,14	480	1029	0
Totaal	752	62		12,34			

Cluster7 is de 2de cluster waar er enkel thuisactiviteiten plaatsvinden. Het aantal personen in deze cluster bedraagt 62. Het verschil tussen deze cluster en cluster6 wordt dan verklaard door het aantal thuisactiviteiten en niet door de totale duur. Dit komt doordat de totale duur voor beide clusters ongeveer dezelfde is. Het aantal thuisactiviteiten is dan echter groter in deze cluster met 2 thuisactiviteiten meer.

Tabel 20: Cluster8 van de activiteitensequentie

Cluster8	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	13	3	100	4,33	245	1062	13
Slapen	2	2	67	1,00	560	560	0
Werken	2	1	33	2,00	203	406	5
Winkelen	2	1	33	2,00	85	170	30
Vrije tijd	1	1	33	1,00	100	100	10
Iets/iemand halen/brengen	1	1	33	1,00	20	20	0
Totaal	21	3		11,33			

De laatste cluster wordt maar opgebouwd door 3 personen. Dit is een zeer klein aantal. Echter is de verdeling in 8 clusters beter dan die in 7 clusters. Van deze cluster kan er dan ook niet zoveel van gezegd worden enkel dat het aantal thuisactiviteiten de minste is van alle clusters.

Enkele hypothesen over welk type personen er in de clusters zitten:

Als de werkactiviteit wordt bekeken dan kan er een eerste hypothese naar voor geschoven worden. In cluster2, cluster4 en cluster5 zijn personen opgenomen die een volledige dag werken. Terwijl de totale tijd in cluster1 en cluster3 veel minder bedraagt waarbij dus de veronderstelling kan gemaakt worden dat de personen hier maar tijdelijk werken of mensen zijn die deze dag maar een halve dag moeten werken. Wanneer deze veronderstellingen aan een variantie-analyse worden getoetst door middel van een ANOVA-tabel, kan er onderzocht worden of de verschillen tussen de cluster wel significant zijn. Hierbij wordt dus onderzocht of de verschillen in de cluster per toeval zijn ontstaan of er systematische verschillen zijn op te merken. Als nu de ANOVA-tabel wordt aangemaakt voor alle werkclusters, cluster1 tot en met cluster5, resulteert dit in geen significant verschil in de totale werkduur tussen de 5 clusters. Echter als de 2 groepen samengenomen worden en er een nieuwe ANOVA-tabel wordt opgesteld blijkt er wel een significant verschil te bestaan tussen beide groepen op een betrouwbaarheidsniveau van 90%. De verwachtingen dat cluster 2, 4 en 5 dan meer werken is dan ook correct gestaafd.

Tabel 21: ANOVA-tabel van de totale werkduur tussen de 2 werktypes

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	850930	850930	3,476	0,0647
Residuals	120	29375180	244793		

Wanneer cluster1 en cluster3 vergeleken worden valt er op dat in cluster3 veel meer activiteiten doorgaan dan in cluster1. Om dit te verifiëren is er opnieuw een ANOVA-tabel opgemaakt.

Tabel 22: ANOVA-tabel van het aantal activiteiten tussen cluster1 en cluster3

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	2168,7	2168,7	18,374	2,89E-05
Residuals	189	22307	118		

Uit de tabel kan er dus opgemaakt worden dat deze aanname significant is. Daaruit volgt dan dat cluster3 meer personen bevat die actiever bezig zijn met andere zaken dan alleen werken. Hierdoor kan er gesteld worden dat cluster3 meer is opgebouwd uit

personen die na hun werkactiviteit nog andere zaken doen. Een goed voorbeeld is de hoge activiteit 'iets/iemand halen/brengen'. Dit kan er op wijze dat cluster3 meer personen bevat die een gezin met kinderen hebben dan cluster1. Het hoge aantal thuisactiviteiten daaraan gekoppeld versterkt dit alleen maar. Cluster1 zou dan personen kunnen bevatten die gaan werken maar geen kinderen in hun gezin hebben. Verder gaat 1/3 in deze cluster werken waardoor er meer oudere personen in deze cluster zitten die gepensioneerd zijn, of tijdelijk niet kunnen werken of gewoon weg instaan voor het huishouden.

Cluster5 kan voor een groot stuk gelijkaardig worden bevonden aan cluster3. Echter is er voor cluster5 een verschil in het aantal activiteiten dat wordt uitgevoerd. Dit is veel minder. De test via de variantie-analyse toont dit echter ook weer aan als significant verschil.

Tabel 23: ANOVA-tabel van het aantal activiteiten tussen cluster3 en cluster5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	1772,7	1772,7	14,081	2,54E-04
Residuals	143	18002,5	125,9		

Een verklaring kan zijn dat de werkactiviteit, sinds er geacht wordt dat dit een cluster is met een volledige werkdag, een grote hoeveelheid tijd in beslag neemt waardoor er minder tijd overblijft voor andere activiteiten. Hierdoor gaan mensen slechts 1 of 2 activiteiten naast de werkactiviteit uitvoeren. Ook bevat deze cluster een groot aandeel van personen die enkel thuisactiviteiten verrichten en enkele bijkomende activiteiten.

Het verschil tussen cluster2 en cluster4 zit ook in het aantal activiteiten en de hoeveelheid personen die een activiteit uitvoeren. De werkactiviteit en thuisactiviteit zijn voor beide clusters gelijkaardig. Hierdoor zit het verschil in de overige activiteiten. Om dit te testen is er eerst een ANOVA-tabel aangemaakt van het totale aantal activiteiten per cluster. Deze toonde echter aan dat er geen significant verschil was op te merken tussen de 2 clusters. Echter nadat de werkactiviteiten en thuisactiviteiten waren verwijderd uit de analyse blijken de overige activiteiten wel significant te verschillen van elkaar. De resultaten zijn dan ook weergegeven in onderstaande tabel.

Tabel 24: ANOVA-tabel van het aantal activiteiten uitgezonderd van thuisactiviteiten en werkactiviteiten tussen cluster2 en cluster4

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	19,247	19,247	11,229	1,07E-03
Residuals	125	214,249	1,714		

Cluster2 bevat veel meer personen die meerdere activiteiten uitvoeren dan cluster4. Hierdoor kan er verondersteld worden dat in cluster2 meer personen zijn die niet werken of thuisblijven en dus meer activiteiten uitvoeren op een andere locatie. Het is dus een actievere bevolking meestal ook gekenmerkt door jongere personen. Dus de veronderstelling die hier gemaakt kan worden is dat cluster2 waarschijnlijk jongere personen zal bevatten t.o.v. cluster4

De hypothese bij cluster6 en cluster7 kunnen verschillende veronderstellingen aannemen. Dit kunnen personen zijn die niet werken zoals huismoeders/huisvaders, gepensioneerden en tijdelijk werklozen. Ook kunnen dit personen zijn die een dag thuis werken in de tuin, een dag verlof hebben genomen, … Eveneens kunnen dit dagen zijn in het weekend waarbij deze personen gewoon thuisgebleven zijn om te ontspannen. Er zijn dus verschillende mogelijkheden. Het verschil tussen de clusters zit enkel in het aantal activiteiten.

6.1.3 Analyses van de clusters op basis van persoonlijke kenmerken

In dit hoofdstuk zullen de clusters per persoonlijk kenmerk besproken worden. Voor de exacte gegevens per cluster wordt verwezen naar bijlage 2. De huwelijksstatus is niet meegenomen doordat de meeste personen in de dataset getrouwd zijn. Het gemiddelde aantal leden in een gezin is ook niet meegenomen omdat deze waarden zijn terug te vinden in de criteria: het aantal kinderen in een gezin en indien de personen een partner hadden, ja of nee. Cluster8 zal nauwelijks besproken worden doordat er maar 3 personen in de cluster zitten.

Man vs. Vrouw:

In alle clusters is er een groter aandeel mannen dan vrouwen. Het verschil tussen de clusters is dan ook niet echt te onderscheiden. De percentages van het aandeel mannen ligt tussen 65% en 73%.

Gemiddeld aantal werkuren:

Enkel bij cluster4 en cluster6 is een hoger aantal gemiddelde werkuren dan 2200 vastgesteld. Sinds cluster6 geen werkactiviteit bevat kan er al een conclusie worden genomen dat deze cluster weekenddagen bevat of dagen dat deze personen niet moeten werken. Voorts levert de ANOVA-tabel weer dat er geen significant verschil is tussen het aantal werkuren in de clusters.

Tabel 25: ANOVA-tabel van het aantal werkuren tussen alle clusters bij de activiteitensequentie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	6	2261206	376868	0,930	4,73E-01
Residuals	427	173074914	405328		

Partner:

Het criteria partner vindt zijn grootste aandeel in cluster4 met 81%. In cluster1 en cluster6 zijn het minste aandeel terug te vinden met respectievelijk 60% en 64%. Hierbij kan het hoogste aandeel gekenmerkt worden door samenwonende of getrouwd koppels terwijl cluster1 en cluster6 meer alleenstaande personen bevatten.

Gemiddeld aantal kinderen:

In cluster4, cluster6 en cluster7 is het gemiddelde lager dan 1. Dit betekent dat de clusters meerdere personen bevat waarvan kinderen geen deel uitmaken van het gezin. Dit kunnen dan oudere personen zijn waar de kinderen het huis al uit zijn, personen die momenteel meer belang hechten aan hun carrière, … Het hoogste gemiddelde is terug te vinden in cluster3. Dit kunnen dan vooral de jongere gezinnen voorstellen. Uit de tabel kan geconcludeerd worden dat de variabele 'aantal kinderen in een huishouden' heel significant is. Deze zou dan later ook terug moeten komen in het model.

Tabel 26: ANOVA-tabel van het aantal kinderen tussen alle clusters bij de activiteitensequentie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	6	65,24	10,87	8,457	6,24E-09
Residuals	842	1082,68	1,29		

Leeftijd:

Cluster1 bestaat uit 1/3 personen die zich in de leeftijdscategorie 35-49 bevinden. De meest voorkomende leeftijdscategorie was echter 50-54 met 15%. Daarentegen is cluster2 opgebouwd uit een oudere bevolkingsgroep. De cluster is vooral opgebouwd uit personen die een leeftijd hebben tussen de 50-59 jaar. Het aandeel binnen de cluster is voor deze groep gelijk aan 32%. De categorie 40-49 volgt echter op een kleine afstand met respectievelijk 29%.

In de derde cluster zijn de leeftijden tussen 35-44 het beste vertegenwoordigd met een aandeel van 35%. De tweede groep bestaat uit de personen die een leeftijd hebben tussen 45-59. Deze groep is eveneens goed voor 1/3 van de cluster. In cluster4 komt een leeftijdscategorie naar voor als belangrijkste die nog niet vermeld is in de vorige clusters. Het gaat om personen die een leeftijd hebben tussen de 60 en 64 jaar. Ze maken voor 16% deel uit van de cluster. Verder worden de leeftijden tussen 50-59 voor 29% vertegenwoordigd in de cluster.

In de 5de cluster is er een duidelijk onderscheid tussen 2 groepen. Langst de ene zijde is er de groep van personen die tussen de 35-49 jaar oud zijn en langste de andere zijde de personen met een leeftijd tussen de 55-64 jaar. De jongste groep heeft een aandeel van bijna 50% terwijl de oudste groep een aandeel heeft van 20%. Cluster6 bestaat vooral

uit de leeftijdscategorieën van 50-54 en 55-59. Ze zorgen samen voor een 40% aandeel in de cluster.

De 7de cluster is voor 40% opgebouwd uit personen die een leeftijd hebben tussen de 50 en 64 jaar.

Inkomen:
Het inkomen in iedere cluster is tussen de 20% en 25% niet vermeld. In cluster3 komen meer huishoudens voor met een inkomen < €750 in vergelijking met de andere clusters. In cluster7 komt deze categorie echter het minste voor.

De inkomenscategorie €750-€1250 is het meest vertegenwoordigd in cluster6 met 15%. De andere clusters hebben een gelijkaardige verdeling voor deze categorie ±11%. De categorie €1250-€1750 is dan het minste voorkomend in cluster6. Dit kan dus betekenen dat er voor de twee laatst vermelde inkomenscategorieën bij cluster6 een wisseling heeft voorgedaan. De overige cluster hebben een gemiddelde van 30% bij de categorie €1250-€1750.

De laagste toewijzing bij de categorie €1750-€2250 bedraagt 14% en is terug te vinden bij cluster5. De gemiddelde bijdrage voor deze categorie tot een cluster bedraagt 20%.

Van de twee laatste categorieën kan gezegd worden dat ze samen goed zijn voor ±12% in een cluster. Opmerking hierbij zijn dat in cluster7 meer hogere inkomens voorkomen van >€2750, met name een aandeel van 10%. Een andere opmerking is dat cluster5 beduidend meer hogere inkomenscategorieën bevat dan de overige clusters. Het aandeel van de 2 laatste categorieën voor cluster5 samen is 17%.

Diploma:
Het diploma dat het meeste voorkomt is het Hoger Niet-universitair diploma. Het heeft dan ook een aandeel per cluster tussen de 34% en 44%. Echter is er één uitzondering waarbij dit diploma nog meer voorkomt. Dit geldt voor cluster5 waarbij 54% van de cluster dit diploma heeft behaald.

Het verschil tussen het tweede en derde meest voorkomende diploma is minder te onderscheiden. Toch kan gezegd worden dat het diploma Hoger Middelbaar Onderwijs

Technisch/Beroeps ook veel voorkomt, vooral in de clusters 2, 3, 4 en 6. In de overige clusters komen ze ook voor maar minder frequent. Een ander diploma dat veel voorkomt is het Hoger Universitair diploma. Echter is dit niet geval in cluster1 en cluster6.

De diploma's Hoger Middelbaar Onderwijs Algemeen Vormend en Lager Middelbaar Onderwijs Technisch/Beroeps vinden hun plaats op de 4de rangorde.

Verklaringen van deze rangordes zijn te verklaren door het feit dat een groot deel van de bevolking met een Hoger Middelbaar Onderwijs Algemeen Vormend gaat doorstuderen en uitkomen met een Hoger Niet-universitair diploma of Hoger Universitair diploma. Het hoge aandeel in het Hoger Middelbaar Onderwijs Technisch/Beroeps diploma wordt dan verklaard door het feit dat deze personen meteen op de arbeidsmarkt terechtkomen zonder verder te studeren. De diploma's Hoger Middelbaar Onderwijs Algemeen Vormend en Lager Middelbaar Onderwijs Technisch/Beroeps kunnen van de oudere generatie zijn. Deze personen moesten vroeger gaan werken en daardoor konden zij hun schooltijd niet volledig doorlopen.

Beroep:

Gepensioneerde personen komen in elke cluster voor, maar toch zijn er verschillen op te merken. In cluster4 en cluster7 zijn 1 op 3 personen gepensioneerd en in cluster1 en cluster6 is dit 1 op 4 personen. In deze clusters hebben de gepensioneerden dus een groot aandeel. De gepensioneerde personen zijn in de overige clusters minder aanwezig met slechts 1 gepensioneerde op 7 personen.

Het beroep 'bediende geen kader' is het vaakst voorkomende beroep over alle clusters heen. In cluster2 en cluster3 hebben ze het grootste aandeel voor hun rekening met ±30%. Ook in cluster1 en cluster6 staan ze op de eerste plaats, echter is het verschil met de gepensioneerden, die op de tweede plaats staan, minder beduidend. Nochtans maken ze voor ¼ deel uit van de cluster.

Het beroep 'ambtenaar' wordt in elke cluster terug gevonden met een gemiddeld aandeel van 12%. Uitzondering hierop is cluster5 waarbij het beroep 'ambtenaar' het meest voorkomende beroep is met 21%. Echter is het verschil met de andere beroepen minder klein in deze cluster.

De beroepen 'bediende kader' en 'arbeider' vullen de 4de en 5de plaats in op de ranglijsten in de clusters. Ze maken elke een 10% deel uit per cluster.

Weekdagen en weekenddagen:
Eerst wordt er per cluster bekeken. Daarna wordt er over de dagen zelf nog gezien.

Cluster1 bevat vooral weekenddagen en vrijdagen ten opzichte van de andere dagen in de cluster. Cluster2 daarentegen heeft minder donderdagen. De weekenddagen zijn ook meer vertegenwoordigd in deze cluster. De derde cluster bevat minder zondagen maar meer woensdagen. Cluster4 is uit ¼ zondagen opgesteld. In cluster5 zijn er geen dagen die meer opvallen door hun aanwezigheid of afwezigheid. Cluster6 is dan vooral opgesteld door weekenddagen. Ze staan in voor bijna de helft van de dagen, namelijk 43%. Cluster7 valt dan weer op door de hoge aanwezigheid van maandagen en zondagen.
Nu worden de dagen elk afzonderlijk geanalyseerd te beginnen met maandag. Voor de maandag valt er op dat deze dag meer aanwezig is in cluster1 dan in de andere clusters. De dinsdag blijkt vervolgens ook het meeste aanwezig te zijn in cluster1 en het minste in cluster6. Als de resterende dagen geanalyseerd worden blijkt dat ongeveer ¼ van alle woensdagen plaats vinden in cluster3. Cluster3 blijkt ook de grootste hoeveelheid donderdagen te bevatten. Het minste aantal donderdagen komt dan weer voor in cluster2. De vrijdag blijkt vervolgens voor ¼ terug te komen in de eerste cluster.

Tenslotte resten dan nog de weekenddagen. Deze zijn het meest voorkomend in de hele dataset. Voor de zaterdag blijkt dat cluster1 ±¼ van alle zaterdagen bevat en cluster5 slechts 1/10. De zondagen zijn duidelijker onder te verdelen per cluster dan de andere dagen. Cluster1, cluster4 en cluster6 hebben elke een aandeel van ±20% en cluster3 en cluster5 elke een aandeel van 8%.

6.1.4 Conclusie van de analyses

Wanneer de hypotheses in de activiteitenanalyse getoetst worden met de demografische analyses blijken enkele veronderstelling te kloppen. Er zullen dan ook per cluster socio-demografische profielen worden opgesteld op basis van deze analyses en besproken worden in hoeverre ze afwijken van de eerdere veronderstellingen.

De eerste cluster bevat personen die tussen de 35 en 49 jaar zijn. Echter hebben de personen in deze cluster niet zoveel Hoger Universitaire diploma's dan de andere clusters. De beroepen die het meeste voorkomen zijn 'bediende geen kader' en gepensioneerd. Samen zijn ze goed voor de helft van de cluster. Cluster1 bevat vervolgens ook de vaakst voorkomende dagen in een week op uitzondering van woensdag en donderdag.

Cluster2 bestaat uit personen die zich bevinden in de leeftijdscategorieën 50-59 en 40-49. Deze staan in voor meer dan de helft van de cluster met elk een aandeel van ±30%. Voorts is deze cluster vooral opgebouwd uit personen die het beroep 'Bediende geen Kader' uitoefenen. Dit is zo voor 1 op 3 personen in deze cluster.

De volgende cluster, cluster3, bestaat uit personen met een leeftijd tussen 35-44 jaar en 45-59 jaar. Deze twee leeftijdsgroepen bevatten elk 1/3 van de cluster. Voorts is dit ook de cluster waar gezinnen met meerdere kinderen toebehoren. Verder hebben 1 op 3 personen in deze cluster een job met als functie 'Bediende geen Kader'. Ook wordt deze cluster gekenmerkt door de hoge aanwezigheid van de weekdagen: woensdag en donderdag.

Cluster4 is zowat het tegenovergestelde van cluster3. Ze wordt gekenmerkt door oudere leeftijdscategorieën gaande van 50-59 en 60-64. De oudste leeftijdscategorie bevat 1 op 6 personen en de 2^{de} leeftijdscategorie zorgt voor nog eens 2 op 6 personen. Samen zijn dus de beide leeftijdscategorieën goed voor de helft van de cluster. Tevens wordt deze cluster bijkomend gekenmerkt door een lager aantal kinderen per gezin ten opzichte van de andere clusters. Een logisch gevolg van de oude leeftijdscategorieën is dat de meeste personen in deze cluster dan ook het beroep 'gepensioneerd' uitoefenen. De dag van de week die in deze cluster het meeste aanwezig is, is de zondag.

De vijfde cluster wordt voor de helft ingevuld door de leeftijdscategorie 35-49. De 2 hoogste inkomenscategorieën zijn hier meer vertegenwoordigd dan in de andere clusters. Tevens onderscheidt deze cluster zich ook van de andere clusters doordat het beroep 'ambtenaar' het vaakst voorkomt. Dit beroep wordt dan ook door 1/5 van de cluster beoefend. Opmerkelijk bij deze cluster is ook dat meer dan de helft een Hoger Niet-universitair diploma heeft.

Cluster6 is de voorlaatste cluster die besproken zal worden. De leeftijden die het grootste deel uit maken van deze cluster zijn gelegen tussen de 50-59 jaar, met name 40%. Deze cluster wordt ook gekenmerkt door de hoge aanwezigheid van weekenddagen, in het bijzonder de zondag. Het aantal kinderen per gezin is hier ook minder dan in de andere clusters. Als het criterium beroep nader bekeken wordt blijkt dat de meeste personen gepensioneerd zijn of de functie 'bediende geen kader' uitoefenen. De diploma's die voor deze groep bepalend zijn: Technisch/Beroeps en Hoger Niet-universitair. Verder kan er nog vermeld worden dat deze cluster gekenmerkt wordt door de inkomensklasse €750-€1250. Dit is ook een cluster waar er enkel thuisactiviteiten worden uitgevoerd.

De laatste cluster, cluster7, heeft veel weg van cluster4. De leeftijdscategorie 50-64 maakt voor 40% deel uit van de cluster. Het is dan ook niet verwonderlijk dat de meeste personen in deze cluster gepensioneerd zijn. Door de oude leeftijden zijn er ook minder gezinnen aanwezig met kinderen in het huishouden. Tevens is deze cluster, alsook de vorige cluster, opgebouwd uit alleen maar thuisactiviteiten.

De hypotheses van de activiteitenanalyses zullen in het kort worden besproken of ze correct waren. De eerste hypothese was dat cluster2, cluster4 en cluster5 een volledige dag zouden werken en cluster1 en cluster3 halve dagen of deeltijds. Dit is echter moeilijk te controleren en kan dan ook niet bevestigd worden. De 2de hypothese dat cluster3 bestaat uit een actievere bevolking en gezinnen met kinderen dan cluster1, die dan hoofdzakelijk zal bestaan uit minder actievere bevolking zoals gepensioneerde personen, komt helemaal overeen met de sociodemografische profielen van de twee clusters. Dit was dan ook een terechte aanname. De 3de hypothese dat cluster2 meer uit jongere personen zou bestaan dan cluster4 blijkt ook te kloppen. De gemiddelde leeftijd is hetzelfde in beide clusters, maar de uiterste leeftijdscategorieën zijn duidelijk tegengesteld aan elkaar. De laatste hypothese dat cluster6 en cluster7 zou bestaan uit hoofdzakelijk weekenddagen, niet-actieve bevolking en actieve bevolking met een dag vrijaf komt ook terug in de opgestelde profielen. De meeste mensen in deze clusters zijn van de leeftijdscategorie ouder dan 50 jaar en dus ook gepensioneerd. Tevens is cluster6 vooral vertegenwoordigd door de weekenddagen.

6.1.5 Opstellen van een model

Nu dat de socio-demografische hypotheses in het vorige hoofdstuk zijn opgesteld moet er een model gebouwd worden dat personen gaat toewijzen aan de activiteitprofielen, of de zogenaamde clusters, om te komen tot de socio-demografische profielen. Hierbij gaat er dus gezocht worden naar verborgen patronen in de socio-demografische kenmerken van personen die de toewijzing zullen leveren. Het zoeken naar verborgen patronen in een dataset wordt in de literatuur beschreven als data-mining. Bij het opstellen van een model dat dit gaat verklaren, zijn er verschillende mogelijkheden. Dit kan door het maken van een regressieanalyse, het opstellen van beslissingsbomen, relationele leermodellen, ... (Fayyad U., Piatetsky-Shapiro G., and Smyth P., 1996)

Belangrijk hierbij te onthouden is dat er geen optimale data-mining algoritme of model bestaat. Elk algoritme heeft zijn voordelen en nadelen. Beslissingbomen blijken nuttig te zijn voor het vinden van structuur in hoge-dimensionele gegevens en om te werken met data die zowel uit categorische als numerische variabelen bestaan. Het is door de ervaring met deze modellen en het toepassingsgebied dat de juiste algoritmes gekozen worden. (Langley P. and Simon H. A., 1995; Hand, D. J. 1994)

Doordat onze data vooral bestaan uit een mix van categorische en numerische variabelen kiezen we voor het aanmaken van een beslissingsboom. Enkele voordelen die hierbij gepaard gaan worden kort even vermeld. Ten eerste is de boomstructuur gemakkelijk te interpreteren door de lezer, natuurlijk voorzien door middel van een voorbeeld. Ten tweede, beslissingbomen kunnen omgaan met waarden die niet zijn ingevuld in de dataset, de zogenaamde 'missing values'. De meeste andere algoritmes vragen datanormalisatie waarbij de lege velden verwijderd of ingevuld moeten worden. Een andere datanormalisatie moet hier ook niet uitgevoerd worden, de aanmaak van dummy variabelen. Ten derde kunnen grote datasets eenvoudig en snel worden berekend op een persoonlijke computer. (Witten I.H. and Frank E., 2000)

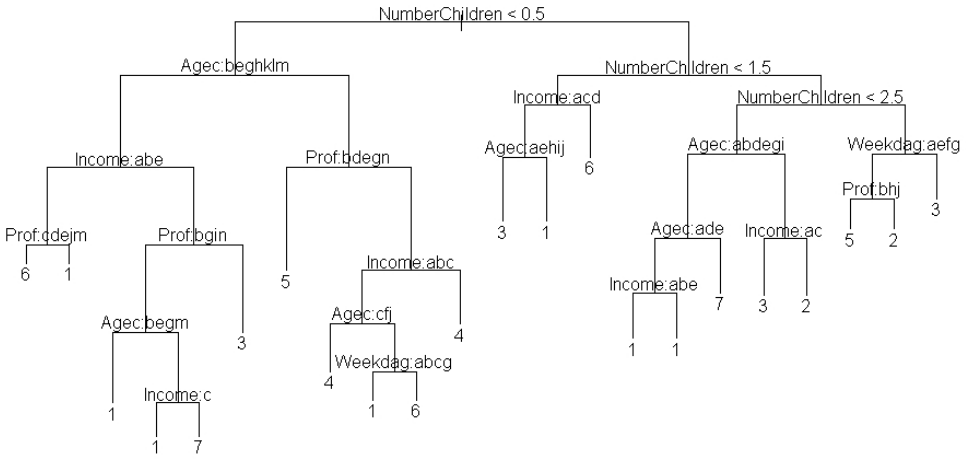
In de literatuur wordt er verder nog een onderscheid gemaakt tussen 2 type bomen: een regressieboom en een classificatieboom. Het verschil zit in het feit dat regressiebomen continue waarden uitkomt, en classificatiebomen discrete waarden. In ons onderzoek gaat er dus gewerkt worden met classificatiebomen. (Witten I.H. and Frank E., 2000)

Om het model te kunnen valideren wordt de dataset onderverdeeld in een trainingsset en een testset. De trainingsset zal bestaan uit 2/3 van de dataset en de testset uit de resterende data.

Tijdens de aanmaak van de beslissingsboom worden alle variabelen toegevoegd. Hieruit kon al snel geconcludeerde worden dat de volgende variabelen geen rol speelde: aantal leden van het huishouden, rijbewijs, werkuren en huwelijksstatus. Dit is te verklaren doordat het aantal leden in het huishouden ook terug te vinden is in de afzonderlijke variabelen: partner en aantal kinderen in het huishouden. Dit was dus een gecorreleerde variabele. De andere 3 variabelen zijn weggelaten doordat ze in alle clusters eigenlijk hetzelfde zijn en dus geen extra bijdrage kunnen leveren tot een onderscheid. Ook diploma en beroep zijn met elkaar gecorreleerd. Echter leverde de variabele 'beroep' beter resultaten dan de variabele 'diploma'. Tenslotte was er nog de variabele 'partner'. Deze heeft het uiteindelijke model niet behaald doordat deze geen extra bijkomende informatie leverde. Ze voegde als het ware niets meer toe tot verbetering van het model.

In Figuur 8 wordt het finale model weergegeven dat is aangemaakt via het gratis verkrijgbare software programma 'R'. De gebruikte functie hierbij is de functie 'tree'. De betekenissen van de lettercodes in de boom zijn terug te vinden in Bijlage 4. In Bijlage 5 worden de knopen van de beslissingsboom uitgeschreven.
Het uiteindelijke model bevat dan 5 variabelen. Hierbij blijkt het aantal kinderen in een gezin een belangrijke parameter te zijn. Deze variabele komt dan ook voor in de eerste takken. De leeftijdscategorieën, beroepen en inkomenscategorieën komen voor in de middelste regionen van de boom. De variabele weekdag blijkt vooral een finaal onderscheid te maken in bepaalde takken van de boom en zijn dan ook beneden in de boomstructuur te vinden.

Om de boom te verduidelijken wordt er een voorbeeld beschreven van hoe de boom nu moet gelezen worden. Als een persoon geen kinderen heeft volgt deze de tak naar links. Dan wordt er gevraagd naar de leeftijd. Is deze persoon bijvoorbeeld tussen de 20 en 24 jaar dan (categorie b) dan volgt men opnieuw de tak naar links. Is vervolgens het inkomen <€750 (categorie a) dan volgt men opnieuw de linkse tak. Waarbij de laatste beslissing wordt genomen. Heeft deze persoon bijvoorbeeld het beroep 'Scholier' (categorie k) dan volgt men deze keer de rechtse tak. Hieruit blijkt dan dat de persoon met deze kenmerken de meeste kans heeft om te toe te horen tot cluster1.



Figuur 8: beslissingsboom voor scenario3 bij activiteitensequentie

Als de trainingsdata voorspeld zou moeten worden aan de hand van de opgemaakte boom dan zouden er 209 van de 568 juist voorspeld worden of 37%. Als deze boom nu gebruikt wordt om de testdata te voorpellen blijkt dat er 77 van 284 juist worden voorspeld of 27%. Dit lijkt misschien weinig maar als er gewoon random zou voorspeld worden is er een kans van 12,5% dat de juiste cluster zou worden gekozen sinds er 8 clusters zijn in dit scenario. De beslissingsboom heeft dus een redelijke voorspellingswaarde.

De demografische profielen kunnen nu worden opgesteld per cluster volgens de beslissingsboom. Dit zal dan ook gebeuren in de volgende paragrafen. Elke cluster wordt weliswaar gekenmerkt door verschillende types personen.

Cluster1 bestaat uit 3 types personen:

- Een eerste type personen wordt gekenmerkt door personen die geen kinderen hebben en de 5 meeste voorkomende beroepen uitoefen zoals telkenmale vermeld in de hypothese in het vorige hoofdstuk. Voorts zijn ze opgebouwd uit 3 belangrijke leeftijdsgroepen: 35-39, 45-54 en 65-74.

- Een 2de type personen dat onderscheiden kan worden zijn: personen met 1 kind, een inkomen tussen €1250-€2250 en een leeftijd tussen de 20-34, 40-49 en ouder dan 65 jaar
- Een derde en laatste type personen die gekenmerkt wordt door deze cluster zijn personen met minstens 2 kinderen en een leeftijd tussen 30-39 jaar.

Cluster2 is opgebouwd door 2 types personen:

- Het eerste type kenmerkt zich door een huishouden met 2 kinderen, een leeftijd tussen 25-29 jaar of 40-44 jaar of 50-54 jaar of ouder dan 60 jaar en een inkomen tussen €750-€1250 of meer dan €1750.
- Het tweede type is samengesteld door personen met meer dan 2 kinderen in een huishouden en waarbij de volgende beroepen niet worden uitgeoefend: ambtenaar, bediendekader en huishouden. Voorts kan er nog gezegd dat dit type personen gekenmerkt wordt door activiteiten die in het weekend plaatsvinden of op maandag en vrijdag.

Cluster3 heeft zijn weerslag in 4 verschillende types personen:

- Type 1 is een persoon die geen kinderen heeft, een leeftijd heeft die zich bevindt in de volgende leeftijdscategorieën: 35-39, 45-54 en 65-74. Ook heeft deze persoon de volgende beroepenstatus niet: ambtenaar, bediende geen kader, gepensioneerd en zelfstandige.
- Type 2 personen bestaan uit personen die 1 kind te onderhouden hebben in het huishouden, een inkomen hebben tussen €1250-€2250 en een leeftijd hebben tussen 35-39 jaar of 50-64 jaar.
- Het derde type is opgesteld uit huishoudens met 2 kinderen, een leeftijd tussen 40-44 jaar of 50-54 jaar of ouder dan 60 jaar en een inkomen kleiner dan €750 of tussen de €1250-€1750.
- Het laatste type bevat personen met meer dan 2 kinderen in een huishouden en waarbij activiteiten worden uitgevoerd op een dinsdag, woensdag of donderdag.

Cluster4 is opgebouwd uit 2 types personen. Echter zijn er gemeenschappelijke kenmerken terug te vinden. De gemeenschappelijke kenmerken zijn het aantal kinderen in een huishouden, of anders gezegd in dit geval het gebrek aan kinderen in een huishouden, de leeftijden die gelegen zijn tussen 25-34 jaar of 40-44 jaar of 55-64 jaar en de beroepsstatus die wordt uitgeoefend: bediende kader en gepensioneerd. Het

verschil is dan terug te vinden in de variabele 'inkomen' waarbij de inkomens die hoger zijn dan €1750 een eerste type weergeven. Als het inkomen kleiner is dan €1750 komt er nog een bijkomende factor aan bod waarbij de leeftijd een rol speelt. Hebben de personen dan specifiek nog een leeftijd tussen 25-29 jaar of 40-44 jaar of 60-64 jaar dan behoort men tot de 2de categorie of type.

Cluster5 bestaat uit 2 verschillende types:

- Personen die geen kinderen hebben, een leeftijd tussen 25-34 jaar of 40-44 jaar of 55-64 jaar en de volgende beroepsstatus niet hebben: bediende kader of gepensioneerd.
- Het andere type wordt beschreven door personen met meer dan 2 kinderen en waarbij de personen een beroepsstatus hebben van ambtenaar, bediendekader of huishouden. De activiteiten die door deze personen gedaan worden vinden plaats in het weekend of op de weekdagen: maandag of vrijdag.

De socio-demografische profielen die door de beslissingsboom voor cluster6 worden opgesteld zijn:

- Personen die één van de 5 meest voorkomende beroepen uitoefenen, uitgezonderd van het beroep arbeider. Ze hebben ook een inkomen van minder dan <€1250 en een leeftijd die zich bevindt in de leeftijdscategorieën: 20-24, 35-39, 45-54 of >65.
- Een tweede groep wordt gekenmerkt worden door personen zonder kinderen, een leeftijd tussen de 14-19 jaar, 30-34 jaar of 55-59 jaar, een inkomen dat kleiner is dan €1750. De beroepen die dan het beste deze groep voorstellen zijn scholieren, bediende kaders, huishouden en gepensioneerden.
- De laatste groep in deze cluster zijn personen met 1 kind in het huishouden waarvan het inkomen ligt tussen €750-€1250 of hoger is dan €2250

De laatste groep of cluster7 bevat de volgende socio-demografische kenmerken:

- Enerzijds is deze cluster opgesteld door personen die geen kinderen hebben in het huishouden, een leeftijd tussen 50-54 jaar of 65-74 jaar en het volgende beroep uitoefenen: ambtenaar of bediende geen kader ofwel gepensioneerd zijn.
- Anderzijds is deze cluster ook opgebouwd uit personen met een huishouden van 2 kinderen en hebben een leeftijd tussen 20-24 jaar of 45-49 jaar.

Als de hypotheses uit het vorige hoofdstuk met de socio-demografische profielen, die werden opgesteld aan de hand van de beslissingsboom, vergeleken worden kan er een algemene conclusie genomen worden. De hypotheses waren meestal gegrond en vonden dan ook hun weerslag terug in het model. Echter zijn in de socio-demografische profielen nog onderscheiden te vinden binnen de clusters, de verschillende types. Met deze opvatting heb ik geen rekening gehouden bij het opstellen van de hypotheses. Daarbij kunnen de verschillende variabellen aangehaald in de hypotheses terug gevonden in bepaalde types van de algemene socio-demografische profielen maar niet allemaal.

6.1.6 Vergelijking tussen de modellen van de verschillende scenario's

In dit hoofdstuk worden de verschillende modellen van de 3 scenario's met elkaar vergeleken. De twee beslissingsbomen van scenario 1 en scenario 2 zijn terug te vinden in bijlage 6.

Tabel 27: vergelijking beslissingsbomen over de 3 scenario's

	Scenario 1 (act.50)	Scenario 2 (act.25)	Scenario 3 (act.75)
testdata juist voorspeld (284)	74 (26%)	91 (32%)	77 (27%)
trainingsdata juist voorspeld (568)	192 (33%)	249 (43%)	209 (37%)
# clusters	9	7	8
# variabelen	5	3	5
random juiste voorspelling	11%	14%	12,5%

Uit bovenstaande tabel blijkt dat de modellen een voorspellingswaarde tussen de 26% en 32% bevatten. Als dit vergeleken wordt met de random situatie is dit toch een beduidend betere methode. Scenario 2 heeft slechts aan 3 variabelen genoeg om een redelijke beslissingsboom op te stellen terwijl dit voor de andere twee scenario's er nog 2 variabelen bijkomen. De weggelaten variabelen in scenario 2 zijn: weekdag en inkomen. Hierbij wordt nog eens versterkt dat weekdag enkel een verdieping was bij de scenario3 en geen hoofdstructuurbijdrager. Wanneer de locatie dus meer invloed krijgt, zoals bij scenario 2, verliezen de variabelen weekdag en inkomen aan voorspellingswaarden. Dit zou er toe kunnen leiden dat deze meer invloed hebben op de activiteiten en minder op de locatie. Dit kan een nieuw inzicht brengen voor verder onderzoek.

6.2 Dagsequentie (lange sequentie)

De stappen uit hoofdstuk 6.1 worden hier herhaald voor de dag sequenties waarbij dus rekening gehouden wordt met de duur van de activiteiten.

6.1.1 Parameters toegepast

Het aantal clusters moet opnieuw bepaald worden doordat de sequenties anders zijn opgebouwd. Eveneens worden de 3 scenario's met hun bijhorende gewicht aan activiteit en locatie berekend.

Scenario 1(50act.-50loc.)

Tabel 28: clusterparameters voor scenario 1 bij de dagsequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,6190938	0,4286407	0,6629059
4	0,5813591	0,4418122	0,7012598
5	0,5021786	0,3777233	0,6318288
6	0,4776517	0,3731821	0,6308013
7	0,4356615	0,341605	0,5538989
8	0,4314138	0,3501871	0,5994232
9	0,3531849	0,272333	0,4879438
10	0,3259861	0,2510956	0,446388

De gemiddelde silhouettebreedte daalt in dit scenario naarmate het aantal clusters toeneemt. Dit zou kunnen leiden tot het kiezen van een beperkt aantal clusters. Bij het analyseren van de genormaliseerde dunn's coëfficiënt blijkt ook dat de kleinere clusters betere resultaten weergegeven. Daarom is er gekozen om bij dit scenario 5 clusters aan te maken. De beste waarden zijn echter niet gekozen omdat ik liever 1 cluster extra meeneem.

Scenario 2(25act.-75loc.)

Tabel 29: clusterparameters voor scenario 2 bij de dagsequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,6521624	0,4782436	0,663585
4	0,6140702	0,4854269	0,7481796
5	0,5490277	0,4362846	0,7442756
6	0,4903752	0,3884502	0,6697366
7	0,4494248	0,3576623	0,5602992
8	0,413048	0,3291977	0,4780338
9	0,3872081	0,3106091	0,4801859
10	0,3784331	0,3093701	0,4790342

De resultaten van dit scenario zijn gelijkaardig met die van scenario 1. Echter zijn de resultaten van dit scenario hoger dan die in het eerste scenario maar dit heeft met de toekenning van de gewichten te maken dan. Vervolgens worden er dan bij dit scenario ook 5 clusters aangemaakt in plaats van 4 clusters. Het verschil tussen 4 clusters en 5 clusters is dan ook veel kleiner dan bij scenario 1.

Scenario 3(75act.-25loc.)

Tabel 30: clusterparameters voor scenario 3 bij de dagsequentie

# clusters	Dunn coeff	Normalized	Avg. Silhouette width
3	0,4700868	0,2051303	0,4637173
4	0,4540418	0,2720558	0,5592621
5	0,4176814	0,2721018	0,5533134
6	0,3777307	0,2532768	0,5624531
7	0,3457267	0,2366811	0,5743627
8	0,3007177	0,2008203	0,5540504
9	0,2885936	0,1996678	0,5162408
10	0,2943856	0,215984	0,5331304

Als het optimale aantal clusters zou gekozen worden op basis van de genormaliseerde dunn's coëfficiënt dan zouden er 5 clusters aanwezig moeten zijn. Het aantal clusters bij de gemiddelde silhouettebreedte dat optimaal geacht wordt is 7. De keuze is dan gevallen voor het aanmaken van 5 clusters. Dit wordt verklaard doordat de genormaliseerde dunn's coëfficiënt bij 5 clusters het hoogste is en een redelijk gemiddelde silhouettebreedte. Bovendien is er in de vorige scenario's iedere keer gekozen voor 5 clusters. Deze lijn wordt hier dan ook verder aangenomen.

6.2.2 Analyses van de clusters op basis van activiteiten

De analyses van de clusters bij de dagsequenties zullen eveneens gaan over scenario 3 waarbij de activiteit 75% van de gewichten werd toegekend. Dit wegens dezelfde redenen als vermeld in hoofdstuk 6.1.1. bij de conclusie.

Tabel 31: Cluster1 van de dagsequentie

Cluster1	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	627	90	98	6,97	158	1101	29
Slapen	138	66	72	2,09	463	968	12
Werken	113	46	50	2,46	414	1017	28
Diensten	10	7	8	1,43	55	79	15
Eten	25	14	15	1,79	76	136	13
Dagelijkse boodschappen	22	14	15	1,57	24	38	12
Winkelen	35	23	25	1,52	52	79	17
Onderwijs en opleiding	4	3	3	1,33	188	251	17
Sociale activiteiten	42	24	26	1,75	187	327	24
Vrije tijd	43	29	32	1,48	177	262	13
Iets/iemand halen/brengen	48	25	27	1,92	18	35	17
Totaal	1107	92		24,31			

In deze cluster wordt de werkactiviteit door de helft van de cluster uitgeoefend. De gemiddelde duur van een werkactiviteit bedraagt ±7 uur. Ruw geschat zorgt deze cluster voor 2 tot 3 werkdagen in de week. De thuisactiviteiten worden een 7-tal keer uitgevoerd per week en duren per activiteit 2 uur en 30 minuten. Verder beoefent ook minstens 1 op 4 personen de volgende activiteiten uit: sociale activiteiten, iets/iemand halen/brengen, winkelen en vrijetijdsbestedingen. Deze activiteiten worden vervolgens 1 à 2 keer gedaan per persoon. De sociale activiteiten worden hier meer uitgeoefend dan de vrijtijdsbestedingen. Dit heeft dan ook lichtjes zijn gevolg in de totale tijdsduur.

Tabel 32: Cluster2 van de dagsequentie

Cluster2	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	962	75	100	12,83	188	2411	19
Slapen	198	67	89	2,96	471	1392	8
Werken	36	15	20	2,40	275	660	17
Diensten	22	14	19	1,57	65	102	14
Eten	12	1	1	12,00	96	1152	10
Dagelijkse boodschappen	32	21	28	1,52	28	43	8
Winkelen	29	20	27	1,45	42	61	13
Onderwijs en opleiding	2	2	3	1,00	105	105	20
Sociale activiteiten	44	29	39	1,52	121	184	13
Vrije tijd	20	17	23	1,18	127	149	13
Iets/iemand halen/brengen	49	16	21	3,06	16	49	30
Totaal	1406	75		41,48			

Wat in deze cluster meteen kenmerkend is, is het hoge aantal thuisactiviteiten. Dit wordt dan ook meteen vertaald in een hoge totale tijdsinname. De werkactiviteit in deze cluster wordt door 1/5 van de totale cluster gedaan. De tijdsbesteding per werkactiviteit is ook beduidend lager dan in de eerste cluster. Er zijn andere activiteiten die hier door meerdere personen worden gedaan en dus meer kenmerkend zullen zijn voor de cluster. Het gaat hier vooral om de sociale activiteiten, winkelen, dagelijkse boodschappen en vrije tijd. Echter is de totale tijd besteed aan sociale activiteiten en vrije tijd minder dan in cluster1. De activiteit 'iets/iemand halen/brengen' wordt hier een 3-tal keer uitgeoefend, wat 1x meer is dan bij de vorige cluster.

Tabel 33: Cluster3 van de dagsequentie

Cluster3	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	669	86	100	7,78	151	1175	41
Slapen	168	67	78	2,51	453	1136	10
Werken	101	30	35	3,37	382	1286	33
Diensten	22	13	15	1,69	36	61	12
Eten	32	14	16	2,29	94	215	14
Dagelijkse boodschappen	18	13	15	1,38	21	29	14
Winkelen	36	24	28	1,50	70	105	16
Onderwijs en opleiding	16	3	3	5,33	194	1035	57
Sociale activiteiten	52	30	35	1,73	139	241	13
Vrije tijd	42	25	29	1,68	195	328	18
Iets/iemand halen/brengen	94	24	28	3,92	16	63	39
Totaal	1250	86		33,18			

Cluster3 bevat weer meer werkactiviteiten dan cluster2. Er gaan nu gemiddeld 1 op 3 personen werken. Ze doen dit een 6,5 uur per dag. Het wordt ook een 3 keer uitgevoerd. Het is dus meer aanvullend samen met de ander clusters. Ook de thuisactiviteit is goed vertegenwoordigd met gemiddeld bijna 8 activiteiten. De activiteit 'iets/iemand halen/brengen' is eveneens hoog voor deze cluster met gemiddeld 4 activiteiten. Winkelen, sociale activiteiten en vrijtijdsbesteding worden ook door ±1/3 van de cluster gedaan. De vrijetijdsbestedingen nemen in deze cluster meer tijd in beslag dan de sociale activiteiten.

Tabel 34: Cluster4 van de dagsequentie

Cluster4	**act.**	**pers.**		**gem.**			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	573	61	100	9,39	233	2189	42
Slapen	129	50	82	2,58	479	1236	0
Werken	5	4	7	1,25	246	308	18
Diensten	5	5	8	1,00	77	77	18
Eten	8	6	10	1,33	98	131	18
Dagelijkse boodschappen	19	12	20	1,58	30	48	9
Winkelen	29	16	26	1,81	56	102	14
Onderwijs en opleiding	2	1	2	2,00	123	246	10
Sociale activiteiten	19	4	7	4,75	88	418	11
Vrije tijd	18	11	18	1,64	129	211	15
Iets/iemand halen/brengen	19	10	16	1,90	16	30	9
Totaal	826	61		29,24			

Deze cluster bevat al veel minder activiteiten dan de vorige clusters. Toch is er hier een groot aantal thuisactiviteiten nog aanwezig. De werkactiviteit is ook maar door 10% vertegenwoordigd in de cluster. De sociale activiteiten daarentegen worden een 4 à 5 keer uitgevoerd. Dit is opmerkelijk hoger dan in de andere clusters. De tijd die hier aan besteed wordt is dan ook vervolgens hoger. Echter is het aandeel van deze activiteit in de gehele cluster lager dan in de vorige clusters. De winkelactiviteit wordt door ¼ van de personen gedaan en duurt gemiddeld een uur per activiteit.

Tabel 35: Cluster5 van de dagsequentie

Cluster5	act.	pers.		gem.			
	#	#	%	#	duur	Tot. duur	reistijd
Thuisactiviteit	1076	86	100	12,51	203	2540	28
Slapen	227	77	90	2,95	476	1403	0
Werken	5	4	5	1,25	191	239	13
Diensten	20	8	9	2,50	67	168	16
Eten	8	8	9	1,00	76	76	15
Dagelijkse boodschappen	36	22	26	1,64	33	54	12
Winkelen	35	25	29	1,40	47	66	14
Onderwijs en opleiding	2	2	2	1,00	178	178	8
Sociale activiteiten	47	22	26	2,14	98	209	26
Vrije tijd	13	9	10	1,44	146	211	18
Iets/iemand halen/brengen	31	16	19	1,94	30	58	37
Totaal	1500	86		29,76			

De thuisactiviteiten in cluster5 zijn ook weer van hoge aard. Er worden dan ook gemiddeld een 12 à 13 keer uitgevoerd per persoon. De werkactiviteit is ook weer minder aanwezig met slechts 5%. Dagelijkse boodschappen, winkelen en sociale activiteiten worden door ¼ van de cluster vertegenwoordigd.

Enkele hypothesen over welk type personen er in de clusters zitten:

Cluster4 en Cluster5 worden vooral gekenmerkt door de lage werkactiviteiten. Deze clusters kunnen dan vooral zijn opgebouwd door personen die niet werken, bijvoorbeeld gepensioneerden, en weekenddagen wanneer er niet gewerkt moet worden. Doordat er in cluster4 meer sociale activiteiten worden uitgevoerd en er meer tijd wordt aan besteed kan dit erop wijzen dat deze cluster jongere personen bevat.

Cluster1, cluster2 en cluster3 stellen vervolgens dan de meer actievere bevolking weer of de werkende klasse. Dit is vooral het geval in cluster1 waar de helft een werkactiviteit uitvoerde. Cluster2 en cluster3 kunnen dan meer niet-werkende personen bevatten dan cluster1. Ook zijn cluster1 en cluster3 meer opgebouwd uit personen die een hele dag werken en cluster2 bevat dan meer halve werkdagen.

Tabel 36: ANOVA-tabel van de totale werkduur tussen de clusters 1&3 en cluster2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	1990267	1990267	4,408	0,03861
Residuals	89	40187479	451545		

Bijkomend heeft cluster2 ook beduidend meer thuisactiviteiten dan de twee andere clusters.

Tabel 37: ANOVA-tabel van het aantal thusiactiviteiten tussen de clusters 1&3 en cluster2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	1569,5	1569,5	16,772	5,70E-05
Residuals	249	23301,5	93,6		

Dit kan er op wijze dat hier meer niet-werkend personen of gepensioneerde personen inzitten.

Cluster1 heeft de activiteit 'iets/iemand halen/brengen' veel minder aanwezig dan de andere twee clusters. Dit kan er op wijzen dat cluster2 en cluster3 meer huishoudens bevat met kinderen. Zoals uit de tabel blijkt, is dit echter al niet significant op een betrouwbaarheidsniveau van 90% dus zal dit ook niet terug komen in de verdere analyses.

Tabel 38: ANOVA-tabel van het aantal activiteiten 'iets/iemand halen/brengen' tussen de clusters 2&3 en cluster1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	1	42,14	42,14	2,645	1,09E-01
Residuals	63	1003,61	15,93		

6.2.3 Analyses van de clusters op basis van persoonlijke kenmerken

In dit hoofdstuk zullen de clusters per persoonlijk kenmerk besproken worden. Dit is dus een herhaling van wat er in hoofdstuk 6.1.3 is gebeurd maar nu toegepast op de dag sequentie.

Man vs. Vrouw:
In alle clusters zijn de mannen vertegenwoordigd tussen de 61% en 73%. Een duidelijk verschil is er dus niet op te merken.

Gemiddeld aantal werkuren:
Het gemiddelde aantal werkuren is gelijk aan 2200 minuten. Hierbij zijn er 2 clusters die hier wat van afwijken. Cluster1 zit net onder dit gemiddelde en cluster5 zit ruim boven dit gemiddelde. Echter volgens de ANOVA-tabel is er geen significant verschil tussen de 5 clusters op te merken.

Tabel 39: ANOVA-tabel van het aantal werkuren tussen alle clusters bij de dagsequentie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	4	1252918	313230	0,810	5,19E-01
Residuals	429	165881063	386669		

Partner:
Dit criterium is duidelijk te onderscheiden tussen de clusters. Cluster3 en cluster5 zijn duidelijk de clusters waar er meer personen aanwezig zijn met een partner met name 8 op de 10 personen. In de overige clusters hebben 6 op 10 personen een partner.

Gemiddeld aantal kinderen:
Het hoogste aantal kinderen in een gezin komt voor in cluster1 met een gemiddelde van 1,3. Enkel in cluster5 is het aantal kinderen in gezin kleiner dan 1, wat er op wijst dat hier meer huishoudens aanwezig zijn met geen kinderen. Wordt de variantie-analyse uitgevoerd voor het aantal kinderen, dan resulteert dit in een significant verschil tussen de 5 clusters.

Tabel 40: ANOVA-tabel van het aantal kinderen tussen alle clusters bij de dagsequentie

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Cluster	4	42,83	10,46	7,985	2,55E-06
Residuals	847	1109,08	1,31		

Leeftijd:

De oudste leeftijden zijn terug te vinden in cluster5. De leeftijdscategorie die hier het hoogste scoort is dan ook die van 55-64. Cluster4 volgt hier kort op met de leeftijdscategorie 45-54 als meeste voorkomende klasse. Cluster3 wordt dan het meeste gevormd door personen die een leeftijd hebben tussen de 40-49 jaar.
Cluster1 en cluster2 zijn minder kenmerkend. Cluster1 bevat dan ook geen specifieke leeftijdscategorieën waarbij de leeftijden schommelen tussen de 35-59 jaar. Cluster2 onderscheidt zich dan toch op een manier van cluster1 doordat de leeftijden tussen 35-44 jaar als middencategorie naar voor geschoven worden en nog meer uiterste leeftijdscategorieën bevat zoals 30-34 en 65-69.

Inkomen:

De inkomenscategorie < €750 komt het meeste voor in de clusters 3 en 4. Cluster1 daarentegen bevat de hoogste inkomenscategorieën. Ze bevatten samen 1/5 van de cluster. Cluster2 onderscheidt zich van de andere clusters door het feit dat de inkomenscategorie €750- €1250 het hoogste is van alle clusters en de categorie €1250-€1750 de laagste.

Diploma:

In alle clusters komt het diploma 'Hoger Niet-universitair' het meeste voor. Het diploma 'Hoger universitair' komt niet voor in cluster4 als één van de belangrijkste diploma's. Nochtans is dit diploma bij cluster3 en cluster5 meer voorkomend. 1 op 5 personen heeft dit diploma dan ook in de vorig vermelde clusters. Het Hoger Middelbaar Onderwijs Technisch/Beroeps diploma is in elke cluster goed voor ±15%.

Beroep:

De beroepen die het meeste voorkomen in alle clusters zijn 'Bediende geen Kader' en 'gepensioneerd'. In cluster 1 zitten dan de meeste personen met het beroep van 'Bediende geen Kader' met name ongeveer 40%. In cluster5 is dit beroep echter minder aanwezig, 15%, maar staat het nog steeds op de 2de plaats qua voorkomen. De

gepensioneerden zijn het beste vertegenwoordigd in cluster5 met 29%. In de overige clusters komen ze telkens op de 2de plaats te staan.

Ambtenaar is het beroep dat zich tevreden moet stellen op een derde plaats in de rangschikking. In alle clusters is dit beroep aanwezig met 12%. Toch zijn er verhoudingsgewijs meer ambtenaren in cluster2 aanwezig dan in de andere clusters. Hier gaat de aanwezigheid zelfs naar de 1 op 5 personen toe. Bediende met een kader functie zijn eveneens in elke cluster terug te vinden telkens met een aandeel van ±10%.

Het aandeel van de arbeiders in de clusters schommelt tussen de 5% en 13%.

Weekdagen en weekenddagen:
Eerst wordt er weer gekeken doorheen de cluster, daarna per dag.

De verdeling van de dagen in de eerste cluster is gelijkmatig. Er is enkel vrijdag die een 3% afwijkt van het gemiddelde van de cluster. Bij de 2de cluster blijken er minder dinsdagen, woensdagen en donderdagen voor te komen dan de andere 4 dagen. Cluster3 bevat meer woensdagen en zondagen in vergelijking met de andere dagen van de week in de cluster. Cluster4 is duidelijk opgebouwd door zaterdagen en zondagen en is dus duidelijk een voorbeeld van de weekenddagen. Ook cluster5 is zo opgebouwd. Echter is hier het verschil tussen de andere dagen minder maar toch nog steeds goed herkenbaar.

De analyse overheen de dagen van de week leert ons dat cluster1 een minder aantal weekenddagen bevat en een hoger aantal donderdagen. Cluster2 onderscheidt zich dan van de rest door een hoog aantal maandagen en vrijdagen. Cluster3 maakt zich herkenbaar door meer woensdagen dan de andere clusters. Cluster4 bevat dan weer een minimum aantal weekdagen overheen de dagen. De weekenddagen zijn in deze cluster normaal verdeeld. Cluster5 is de cluster waar de weekenddagen het beste vertegenwoordigd zijn. Ook blijken de meeste dinsdagen in deze cluster voor te komen.

6.2.4 Conclusie van de analyses

Eveneens worden hier de hypotheses in de activiteitenanalyse getoetst worden met de demografische analyses. Er zullen dan ook weer per cluster socio-demografische hypotheses worden opgesteld en besproken worden in hoeverre ze afwijken van de eerdere activiteiten hypotheses.

Cluster1 is een groep die zich kenmerkt door geen specifieke leeftijdscategorie. De leeftijden schommelen tussen de 35-59 jaar. Veder oefenen de personen in deze cluster voor 40% het beroep 'Bediende geen Kader' uit. Ze vertegenwoordigen ook de hoogste inkomensklasse van alle clusters. De cluster bevat ook geen specifieke dagen van de week. Wat wel kenmerkend is voor deze cluster is het hoogste aantal kinderen per huishouden van alle clusters.

Cluster2 kan omschreven worden als de cluster waarbij de lage gemiddelde inkomens, €750-€1250, meer voorkomen. De hoge gemiddelde inkomens, €1250-€1750, komen dan ook minder voor in deze groep. De leeftijden in deze cluster zijn meer die van de middelbare leeftijd, 35-44 jaar, aangevuld met enkele uiterste leeftijdscategorieën zoals 30-34 en 65-69. De dagen in de week die belangrijk zijn in deze cluster zijn: maandagen en vrijdagen.

De 3de cluster wordt gekenmerkt door de hoge aanwezigheidsgraad van woensdagen. Ze is ook opgesteld door de personen die een leeftijd hebben tussen de 40 jaar en 49 jaar. Een grote groep van personen in deze cluster hebben ook een partner. Tevens wordt er nog een kenmerk aan de cluster toegekend namelijk het hoogste aantal personen die een lage inkomensklasse hebben.

De voorlaatste cluster bevat de 2de oudste leeftijdscategorie van alle clusters namelijk de leeftijdscategorie '45-54'. Deze cluster wordt ook gekenmerkt door de weekenddagen en de laagste inkomensklasse.

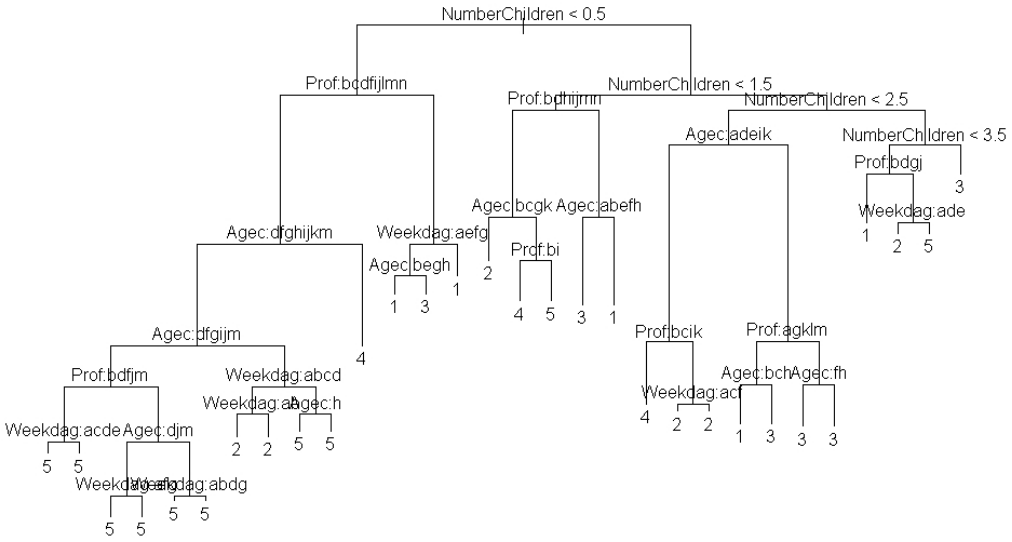
De 5de cluster wordt dan tenslotte bepaald door de oudste leeftijdscategorie. De leeftijden schommelen dan ook tussen de 55 en 64 jaar. Het is dus ook niet verwonderlijk dat het meest uitgeoefende beroep 'gepensioneerd zijn' is. Ook is het lage aantal kinderen in een

huishouden hier niet ongewoon te verwachten dan. Tevens wordt deze cluster ook nog opgebouwd door personen met een partner. Dit geldt namelijk voor 77% van de cluster.

De hypothesen in de activiteitenanalyse van de dag sequentie worden nu even gecontroleerd of ze correct waren. De eerste hypothese was dat cluster4 en cluster5 meer weekenddagen zou bevatten dan de andere clusters doordat er minder werkactiviteiten waren. Na de socio-demografische analyse blijkt deze veronderstelling te kloppen. Ook wordt cluster5 nog eens gekenmerkt door een hoge graad van gepensioneerde personen wat leidt tot de 2de hypothese dat cluster4 jonger zou zijn dan cluster5. Dit klopt echter ook maar het verwachte verschil was groter.
Een 3de hypothese was dat cluster1 meer werkende personen bevat dan cluster2 en cluster3. Dit klopt immers ook doordat de gepensioneerden meer vertegenwoordigd zijn in de twee laatst genoemde clusters en de 1ste cluster opgebouwd is uit 40% bediende zonder kader functie.
De aanname dat cluster2 meer halve werkdagen zou bevatten, kan ook deels correct worden beschouwd doordat de weekdagen, maandag en vrijdag, meer voorkomen dan de andere dagen in de week.
De laatste aanname dat cluster1 meer huishoudens met kinderen zou bevatten dan cluster2 en cluster3, is niet bekrachtigd door de analyse van socio-demografische kenmerken van de clusters. Dit zou dan ook een foute aanname kunnen zijn.

6.2.5 Opstellen van een model

Voor de keuze van het modelalgoritme wordt verwezen naar hoofdstuk 6.1.5. Het finale model voor scenario 3 bij de dagsequentie is opgebouwd uit 4 variabelen: weekdag, aantal kinderen in een huishouden, beroep en leeftijd. De boom wordt opnieuw meteen bepaald door het aantal kinderen in een huishouden gevolgd door het beroep dat een persoon uitoefent. De leeftijd wordt dan meer gebruikt om een onderscheid in het midden van de boom te maken. De variabele weekdag is dan ook meestal de laatste variabele om de toekenning tot een cluster te bepalen. De uitleg bij de takken en knopen is wederom terug te vinden in Bijlage 5.



Figuur 9: beslissingsboom voor scenario3 bij dagsequentie

Eveneens kunnen dan nu de socio-demografische profielen opgesteld worden aan de hand van de beslissingsboom.

De 1ste cluster heeft 5 verschillende types in zijn socio-demografische profiel:

- Het eerste type wordt gekenmerkt door geen kinderen in het huishouden. De beroepen die hier vooral niet worden beoefend zijn: ambtenaar, gepensioneerd zijn en zorgen voor het huishouden. Alsook vinden de activiteiten plaats op dinsdag, woensdag of donderdag
- Het 2de type is gelijkaardig aan type 1 maar de activiteiten worden nu op de andere dagen uitgevoerd. Alsook zijn de leeftijden van de personen terug te vinden in de leeftijdsklassen: 35-39 of 45-54.
- In het derde type van cluster1 is er wel sprake van kinderen in het huishouden namelijk 1 kind. Tevens hebben de personen een beroepsstatus die verschillend is van ambtenaar, gepensioneerd, bediende kader of huishouden. Alsook zijn de leeftijden van de personen terug te vinden in de leeftijdsklassen: 25-34, 45-49 of >55.
- In het voorlaatste type worden de personen omschreven als personen met een huishouden met 2 kinderen waarbij ze het beroep 'bediende geen kader' uitvoeren en ze een leeftijd hebben tussen 20 en 29 jaar.
- Het 5de en laatste type wordt bepaald door personen met 3 kinderen in het huishouden en hebben de volgende beroepsstatus: bediende geen kader, ambtenaar en huishouden.

Cluster2 bestaat uit 4 verschillende types ook hoofdzakelijk te onderscheiden door het aantal kinderen in het huishouden. De types zijn dan:

- Type1: personen met een huishouden waarin geen kinderen voorkomen en waarbij de personen het beroep 'ambtenaar' uitoefenen of gepensioneerd zijn of zorgen voor het huishouden. Hebben deze personen dan ook nog een leeftijd <30jaar of tussen de leeftijdklassen: 35-39, 50-54 of 65-74 en doen ze dit allemaal op weekdagen uitgezonderd van de vrijdag, dan behoren ze tot dit type.
- Type 2: deze is samengesteld uit personen met 1 kind in het huishouden en hebben een beroepsstatus gelijk aan ambtenaar of gepensioneerd. De leeftijden van de personen bevinden zich in de volgende leeftijdsklassen: 20-29, 45-49 of 65-69.
- Type 3: personen waarvan het huishouden uit 2 kinderen bestaat. De leeftijden die gelden voor dit type zijn tussen de 30 jaar en 39 jaar of maken deel uit van de leeftijdsklassen 55-59 of 65-69. De beroepen die hoofdzakelijk niet gedaan worden door dit type zijn: ambtenaar of gepensioneerd.

- Type 4: personen met 3 kinderen in het huishouden. Ze hebben vervolgens volgende beroepsstatus niet: bediende geen kader, ambtenaar en huishouden. Voorts worden de activiteiten gedaan op een maandag, donderdag, vrijdag of zondag.

Cluster3 heeft ook 4 hoofdtypes:

- Type 1 kan omschreven worden als personen die een huishouden hebben waarin geen kinderen voorkomen en waarbij de personen het beroep 'ambtenaar' uitoefenen of gepensioneerd zijn of zorgen voor het huishouden. De leeftijden van dit type personen zijn niet terug te vinden in de leeftijdsklassen: 35-39 of 45-54. De activiteiten vinden ook nog plaats op een dinsdag, woensdag of donderdag.
- Type 2 is opgesteld uit personen met een huishouden dat bestaat uit 1 kind. Tevens hebben de personen een beroepsstatus die verschillend is van ambtenaar, gepensioneerd, bediende kader of huishouden. Alsook zijn de leeftijden van de personen niet terug te vinden in de leeftijdsklassen: 25-34, 45-49 of >55.
- Type 3 zijn dan personen waarvan het huishouden uit 2 kinderen bestaat. Als vervolgens de personen niet thuishoren in de volgende leeftijdscategorieën, 20-29 of 50-54, dan behoren ze tot dit type.
- Type 4 is het eenvoudigste te omschrijven als het type met meer dan 3 kinderen in een huishouden.

De 4de cluster is opgesteld uit 3 types:

- Het eerste type bestaat uit personen met geen kinderen. Ze zijn vervolgens terug te vinden in de leeftijdsklasse 35-39 of 70-74 of zijn jonger dan 29 jaar. De beroepen die plaatsvinden bij dit type zijn dan ook hoofdzakelijk: ambtenaar, zelfstandige, huishouden en gepensioneerd.
- Type 2 is dan samengesteld door personen met 1 kind in het huishouden en hebben een beroepsstatus gelijk aan ambtenaar of gepensioneerd. De leeftijden van de personen bevinden zich in de volgende leeftijdsklassen: 30-44 of 50-64 of ouder dan 70 jaar.
- Type 3 bevat dan personen wiens huishouden uit 2 kinderen bestaat. De leeftijden die gelden voor dit type zijn tussen de 30 jaar en 39 jaar of maken deel uit van de leeftijdsklassen 55-59 of 65-69. De beroepen die hoofdzakelijk door dit type worden gedaan zijn: ambtenaar of gepensioneerd.

De 5de en laatste cluster bestaat uit 4 verschillende types personen:

- De eerste 2 groepen zijn personen met geen kinderen waarbij de personen het beroep 'ambtenaar' uitoefenen of gepensioneerd zijn of zorgen voor het huishouden. Hebben deze personen dan ook nog een leeftijd tussen de 30-34 of 40-49 of 55-64 of ouder zijn dan 75 jaar, dan behoort deze persoon tot het 1ste type. Echter als de personen niet in de vermelde leeftijdsklasse zitten dan is er nog steeds kans dat deze tot dezelfde cluster horen. De activiteiten moeten dan uitgevoerd worden op weekenddagen of op een vrijdag. Deze personen vormen dan ook het 2de type.
- Het derde type is opgesteld uit personen met 1 kind in het huishouden en hebben een leeftijd tussen de 30-44 of 50-64 of ouder dan 70 jaar. Tevens hebben deze personen nog de beroepsstatus: bediende kader, huishouden of zelfstandige.
- Het laatste type personen in deze cluster zijn personen met 3 kinderen in het huishouden en die de volgende beroepsstatus niet hebben: bediende geen kader, ambtenaar en huishouden. Voorts worden de activiteiten gedaan op een dinsdag, woensdag of zaterdag.

Eveneens kunnen de hypotheses in het vorige hoofdstuk met de socio-demografische profielen, die werden opgesteld aan de hand van de beslissingsboom, vergeleken worden voor de dagsequentie. De conclusie hierbij is hetzelfde als bij de activiteitensequentie waarbij de hypotheses meestal gegrond waren en vonden dan ook hun weerslag terug in het model. Echter zijn ook hier in de socio-demografische profielen nog verschillen te vinden binnen de clusters. Daarbij kunnen de verschillende variabelen die aangehaald worden in de hypotheses teruggevonden worden in bepaalde types van de algemene socio-demografische profielen, maar niet in allemaal.

6.2.6 Vergelijking tussen de modellen van de verschillende scenario's

In dit hoofdstuk worden de verschillende modellen van de 3 scenario's met elkaar vergeleken. De twee beslissingsbomen van scenario 1 en scenario 2 zijn terug te vinden in Bijlage 7.

Tabel 41: vergelijking beslissingsbomen over de 3 scenario's

	Scenario 1 (act.50)	Scenario 2 (act.25)	Scenario 3 (act.75)
testdata juist voorspeld (284)	126 (44%)	152 (53%)	98 (35%)
trainingsdata juist voorspeld (568)	331 (58%)	341 (60%)	296 (52%)
# clusters	5	5	5
# variabelen	4	4	4
random juiste voorspelling	20%	20%	20%

Ten eerste is er een verschil tussen de variabellen die niet wordt getoond in de tabel. In scenario 1 en scenario 2 is de variabele 'Weekdag' van scenario 3 vervangen door de variabele 'Partner'. Een aanleiding kan zijn dat naarmate de activiteiten meer doorwegen dat de dagen van de week ook een belangrijker rol worden toegedeeld. Een verklaring kan zijn dat er meer onderscheid komt tussen de weekdagen en weekenddagen sinds er aan andere activiteiten die dagen worden gedaan. Uit de tabel blijkt voorts dat alle modellen een beter voorspelling geven dan het kiezen van een random selectie. Scenario 3 komt er minder goed uit maar dit kan mogelijk verklaard worden door het hoge activiteitgewicht. Toch blijft het een goed voorspellingsmodel. Voorts zijn de trainingsdata voorspellingen meer gelijkend op elkaar dan de voorspellingen van de testdata.

6.3 Vergelijking van de verklarende variabele met de literatuur

De variabelen die telkens terug komen in mijn onderzoek zijn: weekdag, aantal kinderen in het huishouden, beroep, leeftijd, partner en inkomen. Wanneer er naar de literatuur gekeken wordt, blijken veel van deze variabelen echter terug te komen.

De variabelen: weekdag, beroep, aantal kinderen in het huishouden en leeftijd zijn ook terug te vinden als significant in de paper van 'Activity Patterns of Canadian Women' (Wilson C., 1998 b).

In een andere studie bleken eveneens de variabelen: leeftijd, aantal kinderen in het huishouden, weekdag, inkomen terug te komen als significante variabelen (C-H. Joh., 2004).

7 Verder onderzoek

Doorheen de masterproef zijn er heel wat parameters ingesteld: gewichten die toegekend worden aan de activiteit en locatie, strafpunten die toegekend zijn aan het maken van openingen en continuatie ervan in de sequenties, … Ook zijn er beslissingen genomen bij het aanmaken van de sequenties: moet er rekening gehouden worden met de duur of niet? Zo ja, welke duur moet er dan gebruikt worden? In mijn onderzoek heb ik gekozen voor een indeling in 15 minuten. Ook de dataset zelf heeft al een belangrijke bijdrage. Hierbij denk ik maar aan de locatie die overal bijna in dezelfde cel van een rooster plaatsvond. Ook werd bij het bepalen van de locatie gebruik gemaakt van relatieve afstanden. Een verdere verdieping in mijn onderzoek zou dan kunnen gebeuren door het toepassen van een specifieker gebied zoals een provincieniveau.

Doordat de parameters die ingesteld moeten worden geen duidelijke weergave hebben in de literatuur van dit domein en zelf hoofdzakelijk te bepalen zijn door de gebruiker, moet er gestreefd worden naar een standaard. Het uitvoeren van een sensiviteitsanalyse op deze parameters kunnen dan ook een grote bijdrage leveren in dit domein van de verkeersmodellering. Dit wordt zelf ook aangehaald door verschillende onderzoekers. (Wilson C., 1998 b)

8 Conclusie

Na het hele proces te doorlopen van het aanmaken van sequenties tot het opstellen van het model, is er gebleken dat er veel invloeden zijn die het resultaat kunnen bepalen door de verschillende parameters die worden ingesteld. Eén van de belangrijkste conclusie is dat in de masterproef vooral geconcentreerd geweest is op het 3de scenario of het scenario waarbij de activiteit het belangrijkste was. Dit was echter zo door de opmaak van de dataset.

Door het onderscheid te maken tussen de activiteitensequentie, waarbij de duur van de activiteit geen belang heeft, en dagsequentie, waarbij de duur van de activiteit wel meespeelt, zijn er enkele gelijkenissen en verschillen opgetreden. Persoonlijk vond ik dat de clusters in de activiteiten sequentie gemakkelijker te identificeren waren met socio-demografische profielen dan de dag sequentie. Het verschil tussen de weekdagen en weekenddagen blijkt dan weer beter te onderscheiden in de dag sequentie. Het zoeken naar een evenwicht tussen deze 2 is dan ook een deel van verder onderzoek zoals vermeld in het laatste hoofdstuk.

Bij het bouwen van het model bleken toch dezelfde significante variabelen terug te komen. De invloed van de opbouw in sequentie bleek hier echter dan geen verschil uit te maken. Het is dan enkel terug te vinden in het maken van de clusters dat de sequentie meespeelt. De modellen zijn enkel goed als ze hoofdzakelijk overeenkomen met andere studies en dus een gelijkenis vertonen. Deze gelijkenis is dan ook conform met de literatuur.

Het algemene besluit is dan ook dat de modellen hoofdzakelijke een goede voorspelling maken van de clusters. Echter is er nog een duidelijke nood aan een standaard aanpak die door alle onderzoekers in het vakgebied kan worden gebruikt.

Bibliografie

- Dijst, M. (1997). Spatial policy and passenger transportation. Netherlands Journal of Housing and the Built Environment, No. 12: 91-111.
- Dunn, J.C. (1974). Well separated clusters and optimal fuzzy partitions. Journal on Cybernetics, No.4: 95-104.
- Fayyad U., Piatetsky-Shapiro G., and Smyth P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, AI Magazine: Fall 1996: 37-54.
- Gärling T., Gillholm R., Romanus J., and Selart M.. (1997). Interdependent Activity and Travel Choices: Behavioral Principles of Integration of Choice Outcomes. In Activity-Based Approaches to Travel Analysis (D. F. Ettema and H. J. P. Timmermans, eds.), Pergamon, Oxford, United Kingdom, 1997, pp. 135–149.
- Hand, D. J. (1994). Deconstructing Statistical Questions. Journal of the Royal Statistical Society A. 157(3): 317–356.
- Janssens, D. (2008). Activity based models: course introduction.
- Joh, C-H. (2004). Measuring and predicting adaptation in multidimensional activity-travel patterns.
- Joh C-H., Arentze T. and Timmermans H. (2007). Identifying Skeletal Information of Activity Patterns by Multidimensional Sequence Alignment. Transportation Research Record: Journal of the Transportation Research Board, No. 2021: 81-88.
- Kaufman, L. and Rousseeuw, PJ. (1990). Finding Groups in Data: An Introduction to Cluster Analysis.
- Krygsman, S. (2004). Activity and Travel Choice in Multimodal Public Transport Systems. PhD. Thesis, Faculty of Geographical Sciences, Utrecht University, Utrecht.
- Langley, P. and Simon, H. A. (1995). Applications of Machine Learning and Rule Induction. Communications of the ACM 38:55–64.
- McNally, M. (1996). An Activity-Based Microsimulation Model For Travel Demand Forecasting. ITS Working Paper UCI-ITS-AS-WP-96-1, Irvine CA.
- McNally, M. (2000). The Activity-Based Approach. ITS Working Paper UCI-ITS-AS-WP-00-4, Irvine CA.
- McNally, M. (2008). The Four Step Model. ITS Working Paper UCI-ITS-AS-WP-07-2, Irvine CA.

- Miller E.J., Roorda M.J. and Carrasco J.A. (2005). A tour-based model of travel mode choice. Transportation, Vol. 32: 399-422.
- Miller, H.J. (2003). What about people in geo geographic information science. Computers, Environment and Urban Systems, Vol. 27: 447-453.
- Ortúzar and Willumsen. (2002). Modelling Transport, Third edition. John Wiley & Sons, LTD.
- Pons, J. and Vogler, A.P. (2006). Size, frequency, and phylogenetic signal of multiple-residue indels in sequence alignment of introns. Cladistics, Vol. 22, Issue 2: 144-156.
- Rietveld, P. (1994). Spatial economic impacts of transport infrastructure supply. Transportation Research Part A: Policy and Practice, Vol. 28: 329-341.
- Rousseeuw, PJ. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics, No. 20: 53-65.
- Ruiter, E.R. and Ben-Akiva, M.E. (1978). Disaggregate Travel Demand Models for the San Francisco Bay Area. Transportation Research Record, No. 673: 121-128.
- Sankoff, D. and Kruskal, J. (1983). Time warps, string edits and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading, MA.
- Wilson, C. (1998,a). Analysis of Travel Behavior Using Sequence Alignment Methods. Transportation Research Record, No. 1645: 52-59.
- Wilson, C. (1998,b). Activity Patterns of Canadian Women: Application of ClustalG Sequence Alignment Software. Transportation Research Record, No. 1777: 55-67.
- Wilson, C. (1999). ClustalG: Software for analysis of activities and sequential events.
- Wilson, C. (2006). Activity Patterns in Space and Time: Calculating Representative Hagerstrand Trajectories.
- Wilson, C. (2007). Notes on use of ClustalTXY (beta).
- Wilson, C. (2008). Activity patterns in space and time: calculating representative Hagerstrand trajectories. Transportation, Vol. 35: 485-499.
- Wilson, F.R. (1967). Journey to Work - Modal Split. Maclaren and Sons, LTD.
- Witten, I.H. and Frank, E. (2000). Data Mining: Practical Machine Learning Tools and Techniques with JAVA Implementations.

Bijlage

Bijlage 1: lettercodes van activiteiten

Tabel 42: activiteit omzetten in lettercodes

Activiteit	**Letter**
Thuisactiviteit	H
Slapen	S
Werken	W
Diensten	V
Eten	F
Dagelijkse boodschappen	G
Winkelen	P
Onderwijs en opleiding	E
Sociale activiteiten	C
Vrije tijd	L
Iets/iemand halen/brengen	B
Toeren	R
Andere	O
Niet bekend	*X*
Verplaatsing	*T*

Bijlage 2: socio-demografische kenmerken per cluster voor de activiteitensequentie

Tabel 43: socio-demografische kenmerken van cluster1 en cluster2 horende bij de activiteitensequentie

		Cluster1	Cluster2
# pers.		109	75
Man (%)		65	71
Gem. # werkuren		2191	2191
Partners (%)		64	76
Gem. # kinderen		1,1	1,2
5 meest voorkomende leeftijdscat. (#)		50-54 (16)	50-54 (12)
		40-44 (13)	55-59 (12)
		45-49 (13)	40-44 (12)
		35-39 (12)	45-49 (10)
		65-69 (12)	35-39 (9)
Inkomen	<750	7	4
	750-1250	13	8
	1250-1750	32	21
	1750-2250	20	17
	2250-2750	10	5
	>2750	3	3
4 meest voorkomende diploma's (#)		HogerNietUniversitair (46)	HogerNietUniversitair (33)
		HMOAlgemeenVormend (17)	HMOTechnischBeroeps (15)
		HMOTechnischBeroeps (14)	HogerUniversitair (13)
		HogerUniversitair (9)	LMOTechnischBeroeps (6)
5 meest voorkomende beroepen (#)		BediendeGeenKader (27)	BediendeGeenKader (23)
		Gepensioneerd (27)	Gepensioneerd (12)
		Ambtenaar (16)	Ambtenaar (10)
		BediendeKader (10)	Arbeider (8)
		Arbeider (7)	BediendeKader (6)
# Weekdagen	Maandag	23	12
	Dinsdag	21	19
	Woensdag	22	13
	Donderdag	19	7
	Vrijdag	27	16
# Weekend-dagen	Zaterdag	31	18
	Zondag	33	22

Tabel 44: socio-demografische kenmerken van cluster3 en cluster4 horende bij de activiteitensequentie

		Cluster3	Cluster4
# pers.		82	81
Man (%)		66	69
Gem. # werkuren		2181	2224
Partners (%)		74	81
Gem. # kinderen		1,4	0,8
5 meest voorkomende leeftijdscat. (#)		40-44 (15)	60-64 (13)
		35-39 (14)	50-54 (12)
		50-54 (11)	55-59 (11)
		45-49 (10)	40-44 (11)
		55-59 (10)	45-49 (10)
Inkomen	<750	8	4
	750-1250	7	8
	1250-1750	28	22
	1750-2250	13	18
	2250-2750	5	5
	>2750	4	4
4 meest voorkomende diploma's (#)		HogerNietUniversitair (37)	HogerNietUniversitair (29)
		HogerUniversitair (16)	HogerUniversitair (17)
		HMOTechnischBeroeps (14)	HMOTechnischBeroeps (16)
		LMOTechnischBeroeps (6)	LMOTechnischBeroeps (8)
5 meest voorkomende beroepen (#)		BediendeGeenKader (25)	Gepensioneerd (25)
		Ambtenaar (11)	BediendeGeenKader (10)
		Gepensioneerd (11)	Ambtenaar (9)
		BediendeKader (10)	BediendeKader (9)
		Arbeider (7)	Arbeider (6)
# Weekdagen	Maandag	16	15
	Dinsdag	16	14
	Woensdag	28	20
	Donderdag	24	12
	Vrijdag	17	14
# Weekend-dagen	Zaterdag	23	21
	Zondag	12	28

Tabel 45: socio-demografische kenmerken van cluster5 en cluster6 horende bij de activiteitensequentie

		Cluster5	Cluster6
# pers.		63	67
Man (%)		73	70
Gem. # werkuren		2172	2228
Partners (%)		68	60
Gem. # kinderen		1,2	0,9
5 meest voorkomende leeftijdscat. (#)		45-49 (12)	50-54 (14)
		35-39 (10)	55-59 (13)
		40-44 (8)	40-44 (8)
		55-59 (6)	30-34 (5)
		60-64 (6)	65-69 (5)
Inkomen	<750	4	5
	750-1250	7	10
	1250-1750	20	11
	1750-2250	9	14
	2250-2750	7	5
	>2750	5	4
4 meest voorkomende diploma's (#)		HogerNietUniversitair (34)	HogerNietUniversitair (23)
		HogerUniversitair (11)	HMOTechnischBeroeps (13)
		HMOTechnischBeroeps (9)	LMOTechnischBeroeps (10)
		HMOAlgemeenVormend (4)	HMOAlgemeenVormend (8)
5 meest voorkomende beroepen (#)		Ambtenaar (13)	BediendeGeenKader (18)
		BediendeGeenKader (11)	Gepensioneerd (16)
		Gepensioneerd (10)	Ambtenaar (6)
		BediendeKader (9)	Zelfstandige (6)
		Arbeider (6)	Arbeider (4)
# Weekdagen	Maandag	12	12
	Dinsdag	15	11
	Woensdag	12	14
	Donderdag	13	15
	Vrijdag	12	12
# Weekend-dagen	Zaterdag	11	20
	Zondag	13	28

Tabel 46: socio-demografische kenmerken van cluster7 en cluster8 horende bij de activiteitensequentie

		Cluster7	Cluster8
# pers.		62	3
Man (%)		66	33
Gem. # werkuren		2169	1680
Partners (%)		74	67
Gem. # kinderen		0,6	1,3
5 meest voorkomende leeftijdscat. (#)		60-64 (11)	50-54 (1)
		50-54 (7)	35-39 (1)
		55-59 (7)	14-19 (1)
		40-44 (7)	
		45-49 (5)	
Inkomen	<750	1	0
	750-1250	6	0
	1250-1750	18	2
	1750-2250	13	0
	2250-2750	2	0
	>2750	6	0
4 meest voorkomende diploma's (#)		HogerNietUniversitair (24)	HogerNietUniversitair (2)
		HogerUniversitair (13)	LagerOnderwijs (1)
		HMOTechnischBeroeps (7)	
		HMOAlgemeenVormend (6)	
5 meest voorkomende beroepen (#)		Gepensioneerd (19)	Ambtenaar (1)
		BediendeGeenKader (9)	BediendeGeenKader (1)
		BediendeKader (9)	Scholier (1)
		Ambtenaar (6)	
		Arbeider (3)	
# Weekdagen	Maandag	19	0
	Dinsdag	13	0
	Woensdag	12	0
	Donderdag	15	1
	Vrijdag	13	0
# Weekend-dagen	Zaterdag	15	1
	Zondag	19	1

Bijlage 3: socio-demografische kenmerken per cluster voor de dagsequentie

Tabel 47: socio-demografische kenmerken van cluster1 en cluster2 horende bij de dagsequentie

		Cluster1	Cluster2
# pers.		92	75
Man (%)		67	61
Gem. # werkuren		2133	2212
Partners (%)		63	65
Gem. # kinderen		1,3	1,1
5 meest voorkomende leeftijdscat. (#)		35-39 (16)	35-39 (12)
		40-44 (14)	50-54 (11)
		50-54 (13)	40-44 (10)
		45-49 (12)	30-34 (8)
		55-59 (11)	65-69 (8)
Inkomen	<750	4	2
	750-1250	11	11
	1250-1750	24	18
	1750-2250	19	17
	2250-2750	10	6
	>2750	6	3
4 meest voorkomende diploma's (#)		HogerNietUniversitair (40)	HogerNietUniversitair (31)
		HMOTechnischBeroeps (16)	HMOTechnischBeroeps (13)
		HogerUniversitair (14)	HMOAlgemeenVormend (9)
		HMOAlgemeenVormend (10)	HogerUniversitair (8)
5 meest voorkomende beroepen (#)		BediendeGeenKader (36)	BediendeGeenKader (19)
		Gepensioneerd (15)	Gepensioneerd (16)
		Ambtenaar (12)	Ambtenaar (13)
		BediendeKader (9)	BediendeKader (7)
		Arbeider (8)	Zelfstandige (4)
# Weekdagen	Maandag	24	28
	Dinsdag	25	19
	Woensdag	28	20
	Donderdag	28	21
	Vrijdag	19	30
# Weekend-dagen	Zaterdag	21	32
	Zondag	25	36

Tabel 48: socio-demografische kenmerken van cluster3 en cluster4 horende bij de dagsequentie

		Cluster3	Cluster4
# pers.		86	61
Man (%)		73	70
Gem. # werkuren		2185	2192
Partners (%)		80	61
Gem. # kinderen		1	1,1
5 meest voorkomende leeftijdscat. (#)		40-44 (17)	50-54 (13)
		50-54 (14)	45-49 (8)
		45-49 (13)	40-44 (7)
		60-64 (9)	55-59 (7)
		55-59 (8)	30-34 (6)
Inkomen	<750	6	7
	750-1250	9	5
	1250-1750	23	19
	1750-2250	15	8
	2250-2750	4	5
	>2750	4	3
4 meest voorkomende diploma's (#)		HogerNietUniversitair (32)	HogerNietUniversitair (28)
		HogerUniversitair (20)	LMOTechnischBeroeps (9)
		HMOTechnischBeroeps (14)	HMOAlgemeenVormend (8)
		HMOAlgemeenVormend (8)	HMOTechnischBeroeps (7)
5 meest voorkomende beroepen (#)		BediendeGeenKader (19)	BediendeGeenKader (16)
		Gepensioneerd (15)	Gepensioneerd (10)
		Arbeider (11)	Arbeider (8)
		Ambtenaar (9)	Ambtenaar (7)
		BediendeKader (9)	BediendeKader (5)
# Weekdagen	Maandag	23	11
	Dinsdag	24	14
	Woensdag	31	14
	Donderdag	21	11
	Vrijdag	23	16
# Weekend-dagen	Zaterdag	25	27
	Zondag	28	30

Tabel 49: socio-demografische kenmerken van cluster5 horende bij de dagsequentie

		Cluster5
# pers.		86
Man (%)		65
Gem. # werkuren		2317
Partners (%)		77
Gem. # kinderen		0,7
5 meest voorkomende leeftijdscat. (#)		55-59 (15)
		60-64 (14)
		45-49 (13)
		50-54 (9)
		40-44 (7)
Inkomen	<750	3
	750-1250	9
	1250-1750	26
	1750-2250	16
	2250-2750	5
	>2750	6
4 meest voorkomende diploma's (#)		HogerNietUniversitair (35)
		HogerUniversitair (16)
		HMOTechnischBeroeps (14)
		LMOTechnischBeroeps (8)
5 meest voorkomende beroepen (#)		Gepensioneerd (25)
		BediendeKader (12)
		Ambtenaar (11)
		BediendeGeenKader (11)
		Arbeider (4)
# Weekdagen	Maandag	23
	Dinsdag	27
	Woensdag	28
	Donderdag	25
	Vrijdag	23
# Weekend-dagen	Zaterdag	35
	Zondag	37

Bijlage 4: legende bij de beslissingsboom

Tabel 50: legende beroep

na	a
Ambtenaar	b
AndereNietBeroepsactief	c
AndereWelBeroepsactief	d
Arbeider	e
Arbeidsongeschikt	f
BediendeGeenKader	g
BediendeKader	h
Gepensioneerd	i
Huishouden	j
Scholier	k
VrijBeroep	l
Werkloos	m
Zelfstandige	n

Tabel 51: legende dagen

Maandag	a
Dinsdag	b
Woensdag	c
Donderdag	d
Vrijdag	e
Zaterdag	f
Zondag	g

Tabel 52: legende inkomenscategorieën

14-19	a
20-24	b
25-29	c
30-34	d
35-39	e
40-44	f
45-49	g
50-54	h
55-59	i
60-64	j
65-69	k
70-74	l
75-79	m

Tabel 53: legende partner

Nee	a
Ja	b

Tabel 54: legende inkomen

<750	a
750-1250	b
1250-1750	c
1750-2250	d
2250-2750	e
>2750	f

Bijlage 5: interpretatie bij de lezing van de beslissingsbomen

In onderstaande regels worden de knopen van de boom correct beschreven. De 2de tak wordt telkens als verduidelijking gebruikt.
Als eerste wordt het nummer van de knoop beschreven: 2)
Dan wordt de beslissing die in de knoop wordt gemaakt beschreven: Aantal kinderen in een huishouden kleiner dan 0,5 (of geen kinderen in dit geval), ja. Was dit nee dan gaat men naar knoop3
Het eerste cijfer bij knoop 2, 246, geeft het aantal personen weer dat in deze cluster zitten. Het tweede cijfer, 941,80, speelt voor ons geen belang.
Het derde cijfer, 1, geeft weer in welke cluster deze personen zouden komen als de boom na deze knoop zou stoppen.
De cijfers tussen de haken geven de kansen weer van elke cluster in deze knoop. In dit geval betekent dat er 19,9% kans is voor cluster1, 8,5% kans voor cluster2, 14,2% kans voor cluster3,… De hoogste kans wordt dan ook de cluster die de knoop uiteindelijk zal beschrijven of het derde cijfers in de regel.

Beslissingsboom van scenario 3 bij activiteitensequentie uitgeschreven

1) root 435 1674.00 1 (0.20460 0.12184 0.16782 0.14253 0.11724 0.11494 0.13103 0.00000)
2) NumberChildren < 0.5 246 941.80 1 (0.19919 0.08537 0.14228 0.17480 0.10976 0.13415 0.15447 0.00000)
4) Agec: 1,4,6,7,10,11,12 135 502.40 1 (0.26667 0.08889 0.16296 0.08148 0.09630 0.17778 0.12593 0.00000)
8) Income: 0,1,4 40 117.20 1 (0.42500 0.07500 0.05000 0.05000 0.05000 0.32500 0.02500 0.00000)
16) Prof: AndereNietBeroepsactief,AndereWelBeroepsactief,Arbeider,Huishouden,Werkloos 17 55.22 6 (0.23529 0.11765 0.05882 0.05882 0.05882 0.41176 0.05882 0.00000) *
17) Prof: BediendeGeenKader,BediendeKader,Gepensioneerd 23 56.04 1 (0.56522 0.04348 0.04348 0.04348 0.04348 0.26087 0.00000 0.00000) *
9) Income: 2,3,5 95 360.20 3 (0.20000 0.09474 0.21053 0.09474 0.11579 0.11579 0.16842 0.00000)
18) Prof: Ambtenaar,BediendeGeenKader,Gepensioneerd,Zelfstandige 79 298.70 1 (0.22785 0.08861 0.12658 0.11392 0.10127 0.13924 0.20253 0.00000)

36) Agec: 1,4,6,12 25 75.48 1 (0.32000 0.00000 0.16000 0.00000 0.16000 0.28000 0.08000 0.00000) *

37) Agec: 7,10,11 54 200.40 7 (0.18519 0.12963 0.11111 0.16667 0.07407 0.07407 0.25926 0.00000)

74) Income: 2 26 95.28 1 (0.23077 0.19231 0.15385 0.03846 0.11538 0.07692 0.19231 0.00000) *

75) Income: 3,5 28 94.37 7 (0.14286 0.07143 0.07143 0.28571 0.03571 0.07143 0.32143 0.00000) *

19) Prof: Arbeider,Arbeidsongeschikt,BediendeKader 16 33.31 3 (0.06250 0.12500 0.62500 0.00000 0.18750 0.00000 0.00000 0.00000) *

5) Agec: 2,3,5,8,9 111 409.50 4 (0.11712 0.08108 0.11712 0.28829 0.12613 0.08108 0.18919 0.00000)

10) Prof: Ambtenaar,AndereWelBeroepsactief,Arbeider,BediendeGeenKader,Zelfstandige 29 83.98 5 (0.10345 0.00000 0.06897 0.31034 0.41379 0.03448 0.06897 0.00000) *

11) Prof: AndereNietBeroepsactief,Arbeidsongeschikt,BediendeKader,Gepensioneerd,Werkloos 82 292.20 4 (0.12195 0.10976 0.13415 0.28049 0.02439 0.09756 0.23171 0.00000)

22) Income: 0,1,2 59 216.30 4 (0.15254 0.13559 0.16949 0.22034 0.01695 0.13559 0.16949 0.00000)

44) Agec: 2,5,9 28 92.76 4 (0.07143 0.07143 0.25000 0.32143 0.03571 0.03571 0.21429 0.00000) *

45) Agec: 8 31 108.10 1 (0.22581 0.19355 0.09677 0.12903 0.00000 0.22581 0.12903 0.00000)

90) Weekdag: 0,1,2,6 16 51.27 1 (0.37500 0.18750 0.06250 0.18750 0.00000 0.06250 0.12500 0.00000) *

91) Weekdag: 3,4,5 15 47.60 6 (0.06667 0.20000 0.13333 0.06667 0.00000 0.40000 0.13333 0.00000) *

23) Income: 3,5 23 58.63 4 (0.04348 0.04348 0.04348 0.43478 0.04348 0.00000 0.39130 0.00000) *

3) NumberChildren > 0.5 189 715.40 1 (0.21164 0.16931 0.20106 0.10053 0.12698 0.08995 0.10053 0.00000)

6) NumberChildren < 1.5 61 231.30 1 (0.22951 0.18033 0.11475 0.09836 0.09836 0.16393 0.11475 0.00000)

12) Income: 0,2,3 44 164.30 1 (0.27273 0.13636 0.15909 0.11364 0.11364 0.06818 0.13636 0.00000)

24) Agec: 0,4,7,8,9 21 70.81 3 (0.09524 0.19048 0.33333 0.14286 0.09524 0.00000 0.14286 0.00000) *

25) Agec: 1,2,3,5,6 23 72.86 1 (0.43478 0.08696 0.00000 0.08696 0.13043 0.13043 0.13043 0.00000) *

13) Income: 1,4,5 17 50.22 6 (0.11765 0.29412 0.00000 0.05882 0.05882 0.41176 0.05882 0.00000) *

7) NumberChildren > 1.5 128 474.30 3 (0.20312 0.16406 0.24219 0.10156 0.14062 0.05469 0.09375 0.00000)

14) NumberChildren < 2.5 80 288.40 1 (0.26250 0.12500 0.26250 0.08750 0.07500 0.06250 0.12500 0.00000)

28) Agec: 0,1,3,4,6,8 45 151.50 1 (0.42222 0.06667 0.11111 0.06667 0.11111 0.06667 0.15556 0.00000)

56) Agec: 0,3,4 30 90.78 1 (0.50000 0.10000 0.16667 0.03333 0.10000 0.06667 0.03333 0.00000)

112) Income: 0,1,4 15 32.33 1 (0.60000 0.00000 0.13333 0.06667 0.20000 0.00000 0.00000 0.00000) *

113) Income: 2,3 15 43.78 1 (0.40000 0.20000 0.20000 0.00000 0.00000 0.13333 0.06667 0.00000) *

57) Agec: 1,6,8 15 43.10 7 (0.26667 0.00000 0.00000 0.13333 0.13333 0.06667 0.40000 0.00000) *

29) Agec: 2,5,7,9 35 109.70 3 (0.05714 0.20000 0.45714 0.11429 0.02857 0.05714 0.08571 0.00000)

58) Income: 0,2 19 35.22 3 (0.00000 0.05263 0.73684 0.10526 0.00000 0.05263 0.05263 0.00000) *

59) Income: 1,3,4,5 16 56.13 2 (0.12500 0.37500 0.12500 0.12500 0.06250 0.06250 0.12500 0.00000) *

15) NumberChildren > 2.5 48 170.10 5 (0.10417 0.22917 0.20833 0.12500 0.25000 0.04167 0.04167 0.00000)

30) Weekdag: 0,4,5,6 30 107.50 5 (0.13333 0.20000 0.06667 0.16667 0.30000 0.06667 0.06667 0.00000)

60) Prof: Ambtenaar,BediendeKader,Huishouden 15 47.60 5 (0.13333 0.06667 0.00000 0.20000 0.40000 0.06667 0.13333 0.00000) *

61) Prof: AndereWelBeroepsactief,Arbeider,BediendeGeenKader,Zelfstandige 15 50.24 2 (0.13333 0.33333 0.13333 0.13333 0.20000 0.06667 0.00000 0.00000) *

31) Weekdag: 1,2,3 18 48.10 3 (0.05556 0.27778 0.44444 0.05556 0.16667 0.00000 0.00000 0.00000) *

Beslissingsboom van scenario 3 bij dagsequentie uitgeschreven

1) root 556 1780.00 3 (0.20863 0.19964 0.22302 0.15108 0.21763)

2) NumberChildren < 0.5 271 859.30 5 (0.16974 0.18450 0.17712 0.17712 0.29151)

4) Prof: Ambtenaar,AndereNietBeroepsactief,AndereWelBeroepsactief,Arbeidsongeschikt,Gepensioneerd,Huishouden,VrijBeroep,Werkloos,Zelfstandige 209 633.00 5 (0.09091 0.21531 0.15789 0.17703 0.35885)

8) Agec: 3,5,6,7,8,9,10,12 186 547.80 5 (0.09140 0.24194 0.15054 0.12366 0.39247)

16) Agec: 3,5,6,8,9,12 117 324.50 5 (0.08547 0.17094 0.14530 0.11111 0.48718)

32) Prof: Ambtenaar,AndereWelBeroepsactief,Arbeidsongeschikt,Huishouden,Werkloos 36 65.60 5 (0.08333 0.22222 0.02778 0.00000 0.66667)

64) Weekdag: 0,2,3,4 21 29.78 5 (0.09524 0.14286 0.00000 0.00000 0.76190) *

65) Weekdag: 1,5,6 15 31.88 5 (0.06667 0.33333 0.06667 0.00000 0.53333) *

33) Prof: AndereNietBeroepsactief,Gepensioneerd,VrijBeroep,Zelfstandige 81 238.80 5 (0.08642 0.14815 0.19753 0.16049 0.40741)

66) Agec: 3,9,12 51 140.70 5 (0.03922 0.21569 0.27451 0.07843 0.39216)

132) Weekdag: 0,5,6 30 78.61 5 (0.00000 0.30000 0.26667 0.10000 0.33333) *

133) Weekdag: 1,2,3,4 21 54.77 5 (0.09524 0.09524 0.28571 0.04762 0.47619) *

67) Agec: 6,8 30 78.97 5 (0.16667 0.03333 0.06667 0.30000 0.43333)

134) Weekdag: 0,1,3,6 15 35.13 5 (0.13333 0.00000 0.06667 0.33333 0.46667) *

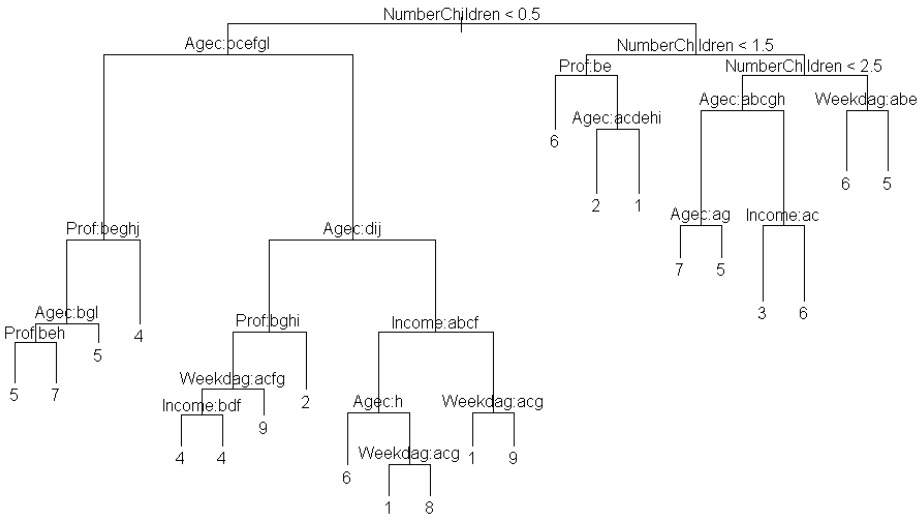
135) Weekdag: 2,4,5 15 42.06 5 (0.20000 0.06667 0.06667 0.26667 0.40000) *
17) Agec: 7,10 69 208.60 2 (0.10145 0.36232 0.15942 0.14493 0.23188)
34) Weekdag: 0,1,2,3 37 104.30 2 (0.10811 0.45946 0.21622 0.10811 0.10811)
68) Weekdag: 0,1 17 42.04 2 (0.00000 0.41176 0.35294 0.11765 0.11765) *
69) Weekdag: 2,3 20 54.37 2 (0.20000 0.50000 0.10000 0.10000 0.10000) *
35) Weekdag: 4,5,6 32 94.21 5 (0.09375 0.25000 0.09375 0.18750 0.37500)
70) Agec: 7 15 35.13 5 (0.06667 0.33333 0.00000 0.13333 0.46667) *
71) Agec: 10 17 53.19 5 (0.11765 0.17647 0.17647 0.23529 0.29412) *
9) Agec: 4,11 23 48.70 4 (0.08696 0.00000 0.21739 0.60870 0.08696) *
5) Prof: Arbeider,BediendeGeenKader,BediendeKader 62 172.60 1 (0.43548 0.08065 0.24194 0.17742 0.06452)
10) Weekdag: 0,4,5,6 38 105.80 3 (0.31579 0.10526 0.34211 0.21053 0.02632)
20) Agec: 1,4,6,7 21 59.23 1 (0.38095 0.09524 0.19048 0.28571 0.04762) *
21) Agec: 2,3,5,8 17 40.14 3 (0.23529 0.11765 0.52941 0.11765 0.00000) *
11) Weekdag: 1,2,3 24 55.35 1 (0.62500 0.04167 0.08333 0.12500 0.12500) *
3) NumberChildren > 0.5 285 895.40 3 (0.24561 0.21404 0.26667 0.12632 0.14737)
6) NumberChildren < 1.5 93 294.80 3 (0.18280 0.22581 0.25806 0.12903 0.20430)
12) Prof: Ambtenaar,AndereWelBeroepsactief,BediendeKader,Gepensioneerd,Huishouden,Werkloos,Zelfstandige 53 154.20 2 (0.13208 0.37736 0.07547 0.13208 0.28302)
24) Agec: 1,2,6,10 15 22.94 2 (0.13333 0.73333 0.00000 0.00000 0.13333) *
25) Agec: 3,5,7,8,9 38 115.80 5 (0.13158 0.23684 0.10526 0.18421 0.34211)
50) Prof: Ambtenaar,Gepensioneerd 17 44.87 4 (0.23529 0.11765 0.00000 0.35294 0.29412) *
51) Prof: AndereWelBeroepsactief,BediendeKader,Huishouden,Werkloos,Zelfstandige 21 56.27 5 (0.04762 0.33333 0.19048 0.04762 0.38095) *
13) Prof: Arbeider,Arbeidsongeschikt,BediendeGeenKader,Scholier 40 102.00 3 (0.25000 0.02500 0.50000 0.12500 0.10000)

26) Agec: 0,1,4,5,7 23 29.57 3 (0.08696 0.04348 0.82609 0.04348 0.00000) *
27) Agec: 2,3,6,8 17 40.88 1 (0.47059 0.00000 0.05882 0.23529 0.23529) *
7) NumberChildren > 1.5 192 595.20 1 (0.27604 0.20833 0.27083 0.12500 0.11979)
14) NumberChildren < 2.5 125 372.60 3 (0.26400 0.22400 0.32000 0.12800 0.06400)
28) Agec: 0,3,4,8,10 46 93.49 2 (0.32609 0.50000 0.00000 0.17391 0.00000)
56) Prof: Ambtenaar,AndereNietBeroepsactief,Gepensioneerd,Scholier 16 32.77 4 (0.31250 0.18750 0.00000 0.50000 0.00000) *
57) Prof: Arbeider,Arbeidsongeschikt,BediendeGeenKader,BediendeKader,Werkloos,Zelfstandige 30 38.19 2 (0.33333 0.66667 0.00000 0.00000 0.00000)
114) Weekdag: 0,2,5 15 15.01 2 (0.20000 0.80000 0.00000 0.00000 0.00000) *
115) Weekdag: 1,3,4,6 15 20.73 2 (0.46667 0.53333 0.00000 0.00000 0.00000) *
29) Agec: 1,2,5,6,7 79 208.60 3 (0.22785 0.06329 0.50633 0.10127 0.10127)
58) Prof: ,BediendeGeenKader,Scholier,VrijBeroep,Werkloos 35 105.10 1 (0.31429 0.08571 0.28571 0.20000 0.11429)
116) Agec: 1,2,7 16 35.31 1 (0.56250 0.00000 0.12500 0.25000 0.06250) *
117) Agec: 5,6 19 56.07 3 (0.10526 0.15789 0.42105 0.15789 0.15789) *
59) Prof: Ambtenaar,Arbeider,BediendeKader,Huishouden,Zelfstandige 44 87.83 3 (0.15909 0.04545 0.68182 0.02273 0.09091)
118) Agec: 5,7 28 39.60 3 (0.03571 0.07143 0.82143 0.03571 0.03571) *
119) Agec: 6 16 33.39 3 (0.37500 0.00000 0.43750 0.00000 0.18750) *
15) NumberChildren > 2.5 67 209.80 1 (0.29851 0.17910 0.17910 0.11940 0.22388)
30) NumberChildren < 3.5 52 155.80 1 (0.34615 0.17308 0.11538 0.09615 0.26923)
60) Prof: Ambtenaar,AndereWelBeroepsactief,BediendeGeenKader,Huishouden 22 52.68 1 (0.59091 0.04545 0.18182 0.09091 0.09091) *
61) Prof: Arbeider,BediendeKader,Scholier,Zelfstandige 30 85.70 5 (0.16667 0.26667 0.06667 0.10000 0.40000)
122) Weekdag: 0,3,4 15 43.10 2 (0.13333 0.40000 0.06667 0.13333 0.26667) *

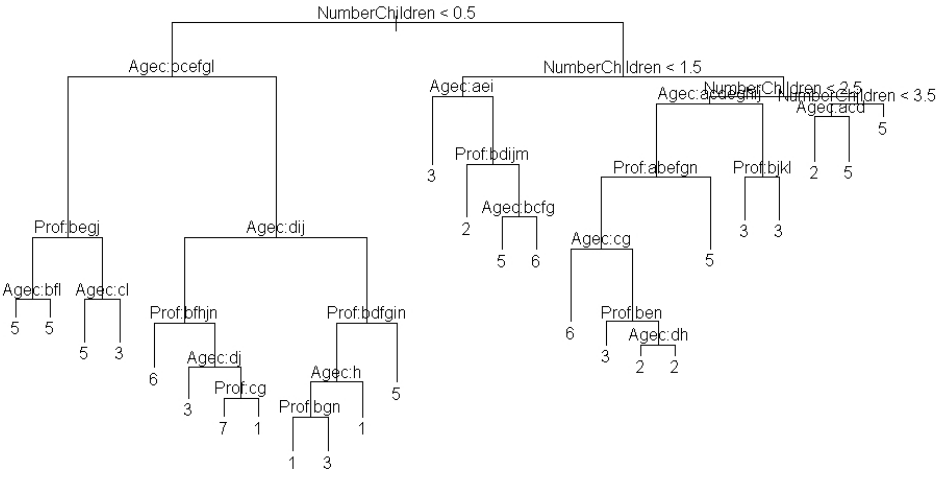
123) Weekdag: 1,2,5,6 15 38.61 5 (0.20000 0.13333 0.06667 0.06667 0.53333) *

31) NumberChildren > 3.5 15 43.78 3 (0.13333 0.20000 0.40000 0.20000 0.06667) *

Bijlage 6: beslissingsbomen voor scenario 1 en scenario 2 bij de activiteitensequentie

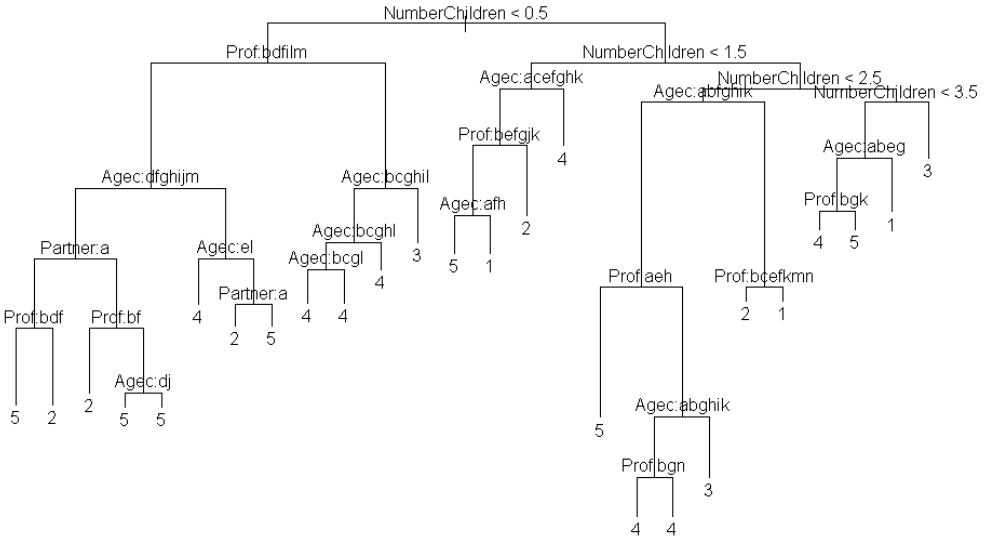


Figuur 10: beslissingsboom voor scenario 1 bij activiteitensequentie

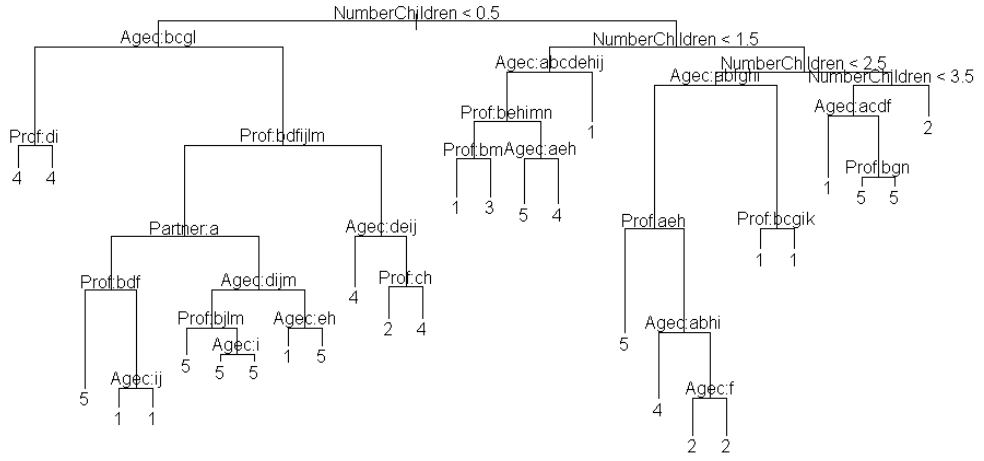


Figuur 11: beslissingsboom voor scenario 2 bij activiteitensequentie

Bijlage 7: beslissingsbomen voor scenario 1 en scenario 2 bij de dagsequentie



Figuur 12: beslissingsboom voor scenario 1 bij dagsequentie



Figuur 13: beslissingsboom voor scenario 2 bij dagsequentie