Een vergelijking van de synthetische populatieschatting in het LMS/NRM met andere veelgebruikte methoden

Ir. Luuk Brederode
Goudappel Coffeng BV / TU Delft
bdl@goudappel.nl / lbrederode@tudelft.nl

Marten Waanders Bsc.
Student Toegepaste Wiskunde Universiteit Twente
m.j.c.waanders@student.utwente.nl

Bijdrage aan het Colloquium Vervoersplanologisch Speurwerk
21 en 22 november 2013, Rotterdam

Samenvatting

Een vergelijking van de synthetische populatieschatting in het LMS/NRM met andere veelgebruikte methoden

Om het effect van maatschappelijke veranderingen zoals bijvoorbeeld de kredietcrisis in beeld te brengen is een verkeersmodel met onderscheid naar huishoudens- en/of persoons-segmenten noodzakelijk. Voor toepassing van een dergelijk model moet de verdeling over segmenten voor elke modelzone (ook wel synthetische populatie genoemd) geschat worden. Hiervoor zijn vijf methoden onderzocht met als doel om de methode te identificeren die een zo realistisch mogelijke populatie oplevert, makkelijk overdraagbaar is en waarvoor efficiënte implementatie binnen een verkeersmodel-systeem mogelijk is.

Voor de methoden IPF, QUAD (de methode die gebruikt wordt in LMS/NRM) en FBS zijn bestaande implementaties beschikbaar. Door problemen met de FBS software is deze methode niet kwantitatief getest, maar op basis van theorie is bekend dat FBS niet efficiënt en moeilijk overdraagbaar is. De methoden Lagrange multipliers en lineair programmeren zijn in theorie krachtig, maar zijn moeilijk overdraagbaar en er zijn, voor zover bekend bij de auteurs, geen efficiënte implementaties beschikbaar. Deze methoden zijn daarom niet kwantitatief getest. Van de resterende methoden (IPF multizone, IPF zone-by-zone en QUAD) genereren beide typen IPF beduidend realistischer populaties dan QUAD. Bovendien scoort IPF goed op overdraagbaarheid, en zijn er veel en zeer efficiënte implementaties van IPF voorhanden. Daarom is IPF multizone geïdentificeerd als de beste methode binnen de context van dit onderzoek.

Daarnaast zijn een drietal quick wins geïdentificeerd in de QUAD implementatie binnen de LMS/NRM systematiek. Deze zijn tevens geïmplementeerd in Matlab en getest op een MON-dataset van geheel Nederland. De synthetische populatie gegenereerd door de aangepaste versie van QUAD benaderen het realisme van de synthetische populatie van IPF multizone, waarmee de aangepaste versie van QUAD het LMS/NRM bij de state of the art modellen zou laten horen op het gebied van synthetische populatieschatting. De verbetering in prestaties door de quick wins kan bijna in zijn geheel worden bereikt zonder aanpassingen aan de software van het LMS/NRM. Hoe de verbeterde QUAD implementatie zich vertaald in aantallen verplaatsingen, de modal split, de herkomst-bestemmings patronen, de verdeling over vertrekperioden en uiteindelijk wegvakintensiteiten, lijnbelastingen en emissies van stoffen berekend door het LMS/NRM is nog niet onderzocht, maar het effect zal ongetwijfeld substantieel zijn. Een vervolgonderzoek waarin de effecten op het LMS of een NRM op genoemde modeluitvoer in beeld worden gebracht is maatschappelijk zeer relevant, terwijl het een relatief geringe inspanning vergt.

1. Inleiding

1.1 Context en probleemstelling

Maatschappelijke veranderingen veroorzaakt door bijvoorbeeld de kredietcrisis hebben voor bijna iedereen gevolgen voor zijn of haar mobiliteitsgedrag. Hoe het mobiliteitsgedrag precies veranderd is sterk afhankelijk van de beschouwde persoon of het beschouwde huishouden. Mogelijke effecten als gevolg van de kredietcrisis zijn bijvoorbeeld:

- studenten besluiten langer door te leren (dus meer woon-school verplaatsingen en minder woon-werk verplaatsingen);
- hoog opgeleiden geconfronteerd met een inkomensdaling kiezen voor vakantie in eigen land (dus kortere recreatieve verplaatsingen);
- niet leaserijders vervangen hun auto later (dus minder sterk dalende uitstoot); en
- huishoudens met een lager inkomen kiezen vaker voor discounters naast hun reguliere supermarkt (dus andere bestemmingskeuze binnen het winkel motief).

Om het totale effect van dergelijke maatschappelijke veranderingen goed in beeld te brengen zal een verkeersmodel, naast ritmotieven, ook onderscheid moeten maken naar persoons- en/of huishoudenstypen die verklarend zijn voor het mobiliteitsgedrag. Dit onderscheid werkt door op *alle* keuzes die in strategische verkeersmodellen doorgaans het mobiliteitsgedrag beschrijven: ritkeuze, vervoerwijzekeuze, bestemmingskeuze, routekeuze en vertrektijdstipkeuze, maar bijvoorbeeld ook rijbewijsbezit, autobeschikbaarheid en vestigingsplaatskeuze. Hiermee vormt het een fundamentele bouwsteen van verkeersmodellen.

In Nederland bevatten de meeste stedelijke en regionale verkeersmodellen een dergelijk onderscheid niet of nauwelijks; autobeschikbaarheid is vaak de enige variabele op huishoud of persoonsniveau. Het LMS/NRM bevat wel persoons- en huishoudtypen, evenals de meeste landelijke, stedelijke en regionale modellen in de rest van de westerse wereld (gedesaggregeerd modelleren). In de VS is het zelfs gebruikelijk om vervoersvraag af te leiden op basis van eigenschappen van individuele huishoudens en personen (activity based modelleren).

Om onderscheid te kunnen maken in huishoud- en persoonstypen moet bekend zijn hoe de verdeling over huishoud- en persoonstypen er binnen elke modelzone uit ziet. Dergelijke data is vrijwel nooit voorhanden op zonaal niveau en moet daarom worden afgeleid van data die wel op zonaal niveau beschikbaar is in combinatie met data die op bovenzonaal niveau beschikbaar is. Dit wordt geïllustreerd aan de hand van Tabel 1. In dit voorbeeld zijn een tweetal *segmentvariabelen* gedefinieerd (geslacht en leeftijd) elk met twee *klassen* (man of vrouw en <18 of >=18). Deze indeling leidt tot vier *persoonssegmenten* (cursief gedrukt in Tabel 1), welke verklarend zijn voor de mobiliteitskeuze die we willen beschrijven[1]. De verdeling over deze persoonssegmenten is alleen bekend op gemeentelijk niveau vanuit het CBS (de *apriori-segmentverdeling*). Op het zonale niveau zijn wel de *randtotalen* (het aantal mannen en het aantal vrouwen) beschikbaar vanuit CBS wijken en buurten.

[1] In de context van Nederlandse stedelijke en regionale modellen gaat het hier vooral om vervoerwijze- en bestemmingskeuze, maar het concept geldt voor elke mobiliteitskeuze of model (dus bijvoorbeeld ook ritkeuze, routekeuze, rijbewijsbezitsmodel, vertrektijdstipkeuze, stationskeuze, lijnkeuze, etc)

Aantal mannen <18	*Aantal mannen >=18*	Aantal mannen
Aantal vrouwen <18	*Aantal vrouwen >=18*	Aantal vrouwen
Aantal personen <18	Aantal personen >=18	

Tabel 1: voorbeeld van 4 persoonssegmenten op basis van twee segmentvariabelen met elk twee klassen

De gezochte verdeling over persoons- of huishoudsegmenten op zonaal niveau is feitelijk een synthetische populatie voor alle modelzones die zo goed mogelijk voldoet aan de eigenschappen (correlatie tussen verschillende segmenten en randtotalen) van de werkelijke populatie. Het schatten van een dergelijke synthetische populatie op basis van randtotalen en een apriori-segmentverdeling ('population synthesis') bestaat normaliter uit twee stappen (Mueller en Axhausen 2010). In de eerste stap ('fitting') wordt de apriori-distributie aangepast zodat deze op zonaal niveau voldoet aan de randvoorwaarden afkomstig van de randtotalen. In de tweede stap ('allocation') worden individuele huishoudens of personen geselecteerd en toegewezen aan de modelzones. De tweede stap is alleen noodzakelijk voor activity based modellen, voor gedesaggregeerde modellen is de verdeling over segmenten uit de eerste stap voldoende. Deze paper focust daarom alleen op de 'fitting' stap. In het resterende deel van deze paper wordt over *agents* en *segmenten* gesproken in plaats van personen of huishoudens en persoons- en huishoudsegmenten. Voor alle methoden geldt dat op zowel personen als huishoudens als een combinatie gefit kan worden.

In generieke vorm is het fitting probleem een optimalisatieprobleem met randvoorwaarden waarin het verschil tussen twee (multi-dimensionale) matrices moet worden geminimaliseerd onder de randvoorwaarde dat de randtotalen voor alle dimensies gelijk moeten zijn aan bepaalde opgegeven randtotalen.

1.2 Onderzoeksdoel

Het primaire doel van dit onderzoek is om de methode die het beste geschikt is voor toepassing op gedesaggregeerde stedelijke en regionale vervoersmodellen te identificeren en te implementeren zodat deze gebruikt kan worden binnen stedelijke en regionale modellen van Goudappel Coffeng BV. Criteria hierbij zijn op de eerste plaats de kwaliteit van de synthetische populatie, op de tweede plaats de mate waarin de methode overdraagbaar is naar verschillende (typen) gebieden binnen Nederland en op de derde plaats de beschikbaarheid van bestaande implementaties die zorgen voor kennis omtrent de methode en kunnen dienen als basis voor een eigen implementatie.
Een mogelijk nog interessanter bijproduct van het onderzoek zijn een aantal gevonden quick wins voor de methode QUAD die gebruikt wordt in LMS/NRM. Deze quick wins kunnen zorgen voor de aansluiting van LMS/NRM bij de state-of-the-art van oplossingsmethoden voor het fittingprobleem.

1.3 Oplossingsmethoden in de literatuur

Het genoemde optimalisatieproblemen met randvoorwaarden uit paragraaf 1.1 wordt eerst vertaald naar de context van population synthesizers om het aantal te beschouwen methoden uit de literatuur te beperken. Het probleem luidt dan: minimalisatie van de afwijking tussen de synthetische segmentverdeling en een gegeven apriori-segmentverdeling onder de randvoorwaarde dat voor elke klasse van elke

segmentvariabele het synthetische randtotaal op zonaal niveau gelijk is aan een gegeven randtotaal. Bij het zoeken naar oplossingmethoden wordt uit gegaan van deze context, dus methoden die minder data gebruiken, zoals de proportional correction method (Quenneville et all 2008) en de statistical correction method (Tilanus 1968), of juist meer data nodig hebben, zoals nearest neighbor matching (Rässler 2004), regression imputation (Moriarity and Scheuren 2004) en Bayesian finite population imputation (Reiter 2009), worden niet beschouwd.

Op basis van deze aannamen zijn vijf relevante methoden uit de literatuur geselecteerd: (1) iteratief proportioneel fitten (IPF), (2) quadratische optimalisatie (QUAD, gebruikt in LMS/NRM), (3) fitness based synthesis (FBS), (4) methode met Lagrange multipliers en (5) lineaire optimalisatie.

2. Vergelijking methoden uit literatuur

Bij het kwalitatief vergelijken van methoden uit de literatuur is, naast de criteria genoemd in paragraaf 1.2 gelet op het type optimalisatiefuncties en randvoorwaarden dat gebruikt kan worden en de mate waarin convergentie naar de optimale oplossing gegarandeerd is. Hieronder worden de verschillende methoden besproken.

2.1 Iteratief proportioneel fitten (IPF)

IPF kent toepassingen in veel disciplines. Zo wordt IPF in transportmodellering ook wel de Fratar of de Furness methode, in de statistiek biproportioneel fitten, in de economie het RAS algoritme en in de informatica de matrix raking of matrix scaling methode genoemd.

IPF biedt geen vrijheid in de keuze van optimalisatiefunctie, het minimaliseert altijd de relatieve entropie (Csiszár 1975), ook bekend als het discriminatieve informatie criterium en Kullback-Leibler divergentie. IPF kan op zonale basis (*zone by zone*) worden toegepast of op alle zones tegelijk (*multizone*). Zone by zone IPF beschouwd alleen de correlatie tussen segmenten op zonaal niveau, multizone IPF beschouwd ook de correlatie tussen segmenten over de gehele populatie (dus ook correlaties tussen zones).

IPF convergeert altijd naar het globale optimum, zolang de randtotalen over de verschillende segmentvariabelen consistent zijn en er geen nulcellen in de apriori-segmentverdeling en randtotalen zitten. Een probleem waarvan de invoerdata niet aan deze eigenschappen voldoet is niet realistisch, waardoor in de context van population synthesis IPF altijd convergeert.

2.2 Kwadratische optimalisatie (QUAD)

Het programma QUAD voert binnen het LMS/NRM een kwadratische optimalisatie uit om het fitting probleem op te lossen (Significance 2011). Hiervoor wordt de Newton-Raphson methode gebruikt. In theorie kan Newton-Raphson het optimum vinden voor elke functie die tweemaal differentieerbaar is, in de praktijk is hij bijna alleen maar bruikbaar voor kwadratische functies, aangezien andere functies een implementatie erg complex zo niet onmogelijk maken. Newton-Raphson is een zeer efficiënte methode, welke echter niet kan om gaan met randvoorwaarden. In de context van het fitting probleem heeft dit als gevolg dat Newton-Raphson niet kan voldoen aan de randvoowaarde dat randtotalen

overeen moeten komen met waargenomen randtotalen en dat de oplossing negatieve waarden kan bevatten. Beide problemen worden hieronder toegelicht.

Omdat de randtotalen per klasse per segmentvariabele niet als harde randvoorwaarde kunnen worden meegenomen minimaliseert QUAD niet alleen de kwadratische afwijkingen per segment maar ook de kwadratische afwijking per klasse per segmentvariabele. Hierbij geldt voor beide componenten dat deze worden uitgedrukt als fractie ten opzichte van het totaal aantal huishoudens of personen. Feitelijk vormt kwadratische optimalisatie hiermee geen oplossing voor het fitting probleem zoals geformuleerd in 1.3 omdat niet volledig wordt voldaan aan de randvoorwaarde. De oplossing bevat dus (geminimaliseerde) afwijkingen tussen de gevonden en gegeven randtotalen. Het totaal aantal agents in de oplossing wijkt daarmee af van het totaal aantal agents uit de gegeven randtotalen.

In het LMS/NRM wordt dit ondervangen door de resultaten uit QUAD te schalen op het hoektotaal, waardoor het totaal aantal agents in het gehele studiegebied gelijk is aan die uit de randtotalen. Het is onduidelijk waarom niet een schaling per zone wordt gehanteerd, aangezien dit in ieder geval consistentie zorgt tussen de resultaten en het aantal agents per zone. In beide gevallen zorgt een dergelijke correctie voor een minder goede beschrijving van de onderlinge correlatie tussen zowel de verschillende segmenten en tussen de randtotalen.

Alhoewel de optimalisatiefunctie kwadratisch moet zijn om gebruik te kunnen maken van Newton-Raphson, kan het proces wel worden gestuurd door middel van een gewicht dat de relevantie van elk van de klasse per segmentvariabele beïnvloed.

Om te voorkomen dat de oplossing negatieve resultaten bevat wordt de Newton-Raphson methode in het LMS/NRM iteratief toegepast, waarbij het resultaat van de vorige iteratie de startoplossing van de volgende iteratie vormt. Hierbij worden negatieve segmentfracties in de startoplossing telkens op 0 gesteld, dit proces stopt wanneer twee opeenvolgende iteraties tot hetzelfde (niet negatieve) resultaat leiden. In theorie kunnen hiervoor erg veel iteraties nodig zijn, in de praktijk blijkt dat in de context van het fitting probleem meestal niet meer dan 5 tot 6 iteraties nodig zijn. Convergentie van QUAD toepassingen op deze manier is gegarandeerd zolang de doelfunctie kwadratisch is.

2.3.Fitness based synthesis (FBS)

Fitness based synthesis is een vorm van combinatorial optimization. FBS type methoden worden soms ook vernoemd naar de zoek methode die gebruikt wordt binnen de methodiek, zoals hill climbing, simulated annealing en genetische algoritmes.

FBS construeert een synthetische populatie door een subset uit de apriori distributie te trekken die zo goed mogelijk overeenkomt met randtotalen en/of apriori distributie. Hierbij wordt gebruik gemaakt van één van de genoemde zoekmethoden en een doelfunctie die het verschil ten opzichte van randtotalen en/of apriori distributie beschrijft.

Het concept is als volgt. FBS start met een willekeurige subset van de apriori distributie met grootte gelijk aan het aantal agents in de zone (bekend uit de randtotalen). Vervolgens wordt één agent uit de zonale subset vervangen voor een agent van een

ander segment uit de apriori-distributie waarna de waarde van de doelfunctie wordt geëvalueerd. Als de doelfunctiewaarde kleiner is geworden, wordt de nieuwe populatie gehandhaafd, anders wordt terug gegaan naar de oude populatie. Dit proces herhaalt zich tot de doelfunctiewaarde kleiner is dan een bepaalde streefwaarde of het maximum aantal permutaties bereikt is.

Door zoekmethoden toe te voegen wordt bovenstaand proces slimmer gemaakt. Hill climbing of steepest descent is daarbij het meest eenvoudige zoekalgoritme. Hierbij wordt vanuit de bestaande populatie van alle zoekrichtingen bepaald voor welke permutatie de doelfunctie het laagste wordt en wordt die permutatie gekozen. Het nadeel van deze methode is dat de gevonden oplossing slechts een lokaal minimum is, die afhankelijk van de startoplossing kan zijn. Alleen bij convexe problemen is dit gegarandeerd de optimale oplossing.
Simulated annealing kan hier beter mee omgaan omdat dit zoekalgoritme ook een verslechtering van de doelfunctie kan accepteren in de zoektocht. Hierdoor blijft het algoritme niet hangen in een lokaal optimum, maar kan het hier uit komen om een lager, mogelijk globaal optimum te vinden. In theorie vindt simulated annealing altijd de optimale oplossing, ongeacht de startoplossing. Omdat het in essentie een lokale optimalisatiemethode is, is de hoeveelheid rekentijd die nodig is om deze eigenschap werkelijk te laten gelden zo hoog dat in de praktijk geldt dat de startoplossing wel degelijk zeer bepalend is voor de uiteindelijk gevonden oplossing.
Globale zoekalgoritmen zoals evolutionaire algoritmen, waarvan genetische algoritmes de meest gebruikte zijn, zoeken daarom direct in de totale oplossingsruimte. Deze algoritmen kennen een startpopulatie verspreid in de gehele oplossingsruimte en zijn gebaseerd op survival of the fittest en genetische manipulatoren om nieuwe oplossingen te bouwen in volgende iteraties.

Alhoewel de hierboven beschreven zoekmethoden het proces versnellen en zorgen voor een grotere kans op het vinden van het globale optimum zijn ze zeer gevoelig voor de kwaliteit van de startoplossing, de parameterwaarden en convergentiecriterea. In principe kan fitness based synthesis omgaan met elke willekeurige doelfunctie, maar functies met één enkele extreme waarde zijn gewenst om het probleem van lokale minima te omzeilen.

2.4 Lagrange multipliers

Lagrange multipliers vormen een manier om een optimalisatieprobleem met randvoorwaarden om te schrijven tot een optimalisatieprobleem zonder randvoorwaarden. Van dit optimalisatieprobleem kan vervolgens worden bepaald bij welke distributie over segmenten de afgeleide nul is, wat duidt op een optimum. Binnen de context van population synthesis wordt normaliter een (gewogen) kleinste kwadraten functie als doelfunctie gebruikt. Dit is de reden dat deze methode ook wel de kleinste kwadraten methode genoemd wordt.

Omdat de randvoorwaarde uit het originele optimalisatieprobleem een vergelijking is (geen ongelijkheid) wordt voldaan aan de Karush-Kuhn-Tucker condities. Bovendien is de randvoorwaarde lineair, waardoor tevens voldaan wordt aan de Slater condities. Dit maakt dat voor het fitting probleem zoals beschreven in paragraaf 1.3 en uitgaande van

een (gewogen) kleinste kwadraten doelfunctie geldt dat het gebruik van Lagrange multipliers als oplossingsmethode altijd leidt tot het globale optimum.
Problemen met een andere (differentieerbare) doelfunctie en/of randvoorwaarden kunnen in theorie ook worden opgelost met Langrange multipliers, maar het vinden van de oplossing kan, afhankelijk van het type doelfunctie, een stuk complexer worden. Daarbij kan voor niet convexe doelfuncties niet worden gegarandeerd dat het globale optimum gevonden wordt.

2.5 lineaire optimalisatie
Bij lineaire optimalisatie worden de doelfunctie en randvoorwaarden hergeformuleerd als lineair optimalisatie probleem, zodat zeer snelle oplossingsalgoritmes zoals Simplex gebruikt kunnen worden. Aangezien in de context van population synthesis de randvoorwaarden al lineair zijn hoeft alleen de doelfunctie maar benaderd te worden door een (combinatie van) lineaire functie(s). Dit kan bereikt worden door de functie op te knippen in een aantal deel-domeinen. Binnen elk van deze domeinen kan een lineaire functie gefit worden welke tezamen de doelfunctie vormen. Binnen een lineair optimalisatieprobleem kunnen eenvoudig extra randvoorwaarden worden toegevoegd, waardoor bijvoorbeeld aanvullende data (naast de randtotalen en apriori-segmentverdeling) ook verwerkt kan worden. Omdat de lineaire benadering leidt tot een convexe, maar niet strikt convexe functie kunnen er meerdere oplossingen bestaan.

3. Vergelijking van meest gebruikte methoden: casestudy
Het doel van de methoden beschreven in hoofdstuk 2 is om de verdeling over segmenten binnen elke modelzone zo goed mogelijk te beschrijven. Om de prestaties van de verschillende methoden in beeld te brengen worden daarom geschatte verdelingen op zonaal niveau vergeleken met waargenomen verdelingen.

3.1 Aanpak
Op het niveau van modelzones in stedelijke en regionale modellen zijn de segmentverdelingen onbekend (en juist de reden voor het bestaan van population synthesis). Met andere woorden: op het niveau van modelzones binnen stedelijke en regionale modellen is geen data voor handen waaraan synthetische resultaten kunnen worden getoetst. Daarom is op basis van het MON een dataset geconstrueerd die geheel Nederland beschrijft, terwijl de twaalf provincies dienst doen als 'modelzones'. Van deze dataset vormen de randtotalen per provincie en de verdeling over geheel Nederland invoer van het fitting probleem. Voor deze data is het fitting probleem opgelost met de verschillende methoden waarbij de verdeling binnen de provincies dus als onbekend is verondersteld. De resulterende verdelingen per provincie kunnen dan worden vergeleken met de werkelijke verdelingen. Op deze manier ontstaat een beeld over hoe de verschillende methoden zullen presteren in een realistische toepassing. Merk op dat alleen voor QUAD ook nog kan worden gekeken naar de verschillen tussen werkelijke en gesynthetiseerde randtotalen, aangezien randtotalen binnen de andere methoden als harde randvoorwaarde meegenomen zijn en dus per definitie gelijk zijn aan de werkelijke randtotalen. Deze analyse is ook voor QUAD binnen dit onderzoek achterwege gelaten. Bij het interpreteren van resultaten moet daarom in het achterhoofd gehouden worden dat de randtotalen van QUAD afwijken van de gegeven randtotalen, terwijl de andere methoden per definitie geen afwijkingen kennen.

3.2 Segmentering en dataset

De gehanteerde segmentering is geïnspireerd op de segmentering van de module TOURFREQ binnen het NRM2004 en zou daarmee verklarend moeten zijn voor de ritkeuze van personen met woon-werk als ritmotief. De segmentering is iets vergrofd, zodat deze toepasbaar is in combinatie met CBS data beschikbaar op wijk- en buurtniveau. Hiermee is de segmentering toepasbaar voor stedelijke en regionale modellen in heel Nederland.

De volgende segmentvariabelen en –klassen zijn onderscheiden:

- **Auto beschikbaarheid:** Altijd beschikbaar (AB), Soms beschikbaar (SAB), Geen auto beschikbaar (NAB), Potentiële passagier (PASS).
- **Werk:** Geen betaald werk, full time, part time.
- **Rijbewijs bezit:** Heeft een rijbewijs, Heeft geen rijbewijs
- **Leeftijd:** 0-24 jaar, 25-44 jaar, 45 jaar of ouder.
- **Huishoud grootte:** 1 persoon, 2 personen, 3 personen, 4 personen, 5 of meer personen.

In theorie zouden er 360 (4*3*2*3*5) segmenten mogelijk zijn, maar niet alle combinaties zijn realistisch. Zo heeft een persoon die valt binnen autobeschikbaar-heidsklasse AB of SAB per definitie een rijbewijs, terwijl een persoon die valt binnen de klasse PASS per definitie geen rijbewijs heeft. Daarnaast zijn niet alle combinaties verklarend voor het ritkeuze gedrag van individuen. Om deze reden worden de segmentvariabelen leeftijd en huishoudgrootte bij een aantal combinaties van autobeschikbaarheid-, werk- en rijbewijsbezitsklassen niet gebruikt. Uiteindelijk blijven er zo 38 valide segmenten over.

De dataset is opgebouwd door de beschreven persoonskenmerken van individuen in het MON/OViN tussen 2004 en 2010 te stapelen. Deze stapeling is nodig om binnen elk segment en elke provincie voldoende waarnemingen te hebben voor validatie doeleinden.

3.3 geteste methoden

Vanwege een beperkt tijdsbudget zijn niet alle methoden beschreven in hoofdstuk 2 getest in de case study. In de praktijk wordt wereldwijd met name IPF en FBS gebruikt voor population synthesis. Daarnaast wordt in het LMS/NRM QUAD Gebruikt. Daarom is voor deze drie methoden de case study doorgerekend. Voor FBS bleek dat de software implementatie voorhanden (Abraham et all 2011) niet convergeerde op onze dataset. Hoogstwaarschijnlijk is dit op te lossen door de parameters van het zoekproces te optimaliseren, dit is echter niet gedaan binnen dit onderzoek. Hierdoor zijn alleen IPF (multizone en zone-by-zone) en QUAD kwantitatief vergeleken.

De implementatie van QUAD in LMS/NRM software bleek lastig standalone te gebruiken. Daarom is ervoor gekozen om QUAD zelf te implementeren binnen Matlab. Dit bood tevens de mogelijkheid om een tweetal verbeteringen aan de methodiek door te voeren: (1) er zijn segmentspecifieke gewichten aan de doelfunctie toegevoegd en (2) er wordt op zonale totalen geschaald in plaats van het hoektotaal. Beide toevoegingen bieden mogelijkheden om de resultaten van de methodiek te verbeteren. NB: wanneer alle segmentspecifieke gewichten op 1 worden gesteld is de doelfunctie exact gelijk aan die uit de QUAD implementatie van het LMS/NRM. De mogelijkheid om op hoektotalen in plaats van zonale totalen te schalen is niet geïmplementeerd.

3.4 Resultaten

Na het toepassen van multizone IPF, zone-by-zone IPF en QUAD op de dataset beschreven in paragraaf 3.1 zijn per methode de verschillen tussen de gesynthetiseerde en waargenomen segmentering inzichtelijk gemaakt door de absolute afwijking per segment en per provincie te bepalen. Gesommeerd over alle provincies levert dit het resultaat op zoals weergegeven in Tabel 2.

	Totale Absolute Afwijking (TAA)	Index (IPF multizone = 100)
IPF multizone	567758	100.00
IPF zone-by-zone	568738	100.17
QUAD (zonaal geschaald, LMS/NRM gewichten)	1259155	221.78

Tabel 2: totale absolute afwijking (gesommeerd over provincies en segmenten) per methode

Hieruit blijkt dat IPF een synthetische populatie genereert die veel beter lijkt op de werkelijke populatie dan een door QUAD gegeneerde populatie. Daarbij doet multizone IPF het nog iets beter dan zone-by-zone IPF, maar het verschil is klein vergeleken bij het verschil ten opzichte van QUAD. Houd in het achterhoofd dat de implementatie van QUAD binnen de LMS/NRM software op hoek- in plaats van zonale totalen schaalt; de TAA zou op basis van de LMS/NRM implementatie nog groter zijn.

Indexcijfers van afwijkingen per provincie zijn weergegeven in Tabel 3, waarin de overall beste methode (IPF multizone) op index 100 is gesteld. Uit Tabel 3 wordt duidelijk dat de verschillen tussen de IPF methoden en QUAD zich op alle provincies voordoet. Wanneer de verschillen tussen de twee IPF methoden beschouwd worden is duidelijk dat de multizone aanpak voor tien van de twaalf provincies beter scoort dan zone-by-zone. Voor deze provincies voegt de landelijke correlatiestructuur blijkbaar verklarende waarde toe, waar deze voor de provincies Flevoland en Zuid Holland juist een negatieve uitwerking heeft; hier is de index van zone-by-zone lager (dus beter) dan multizone.

	Drenthe	Zeeland	N-Holland	Flevoland	Friesland	Limburg	Overijssel	Groningen	Nrd-Brabant	Z-Holland	Gelderland	Utrecht
IPF multizone	100	100	100	100	100	100	100	100	100	100	100	100
IPF zone-by-zone	101	101	100	99	100	103	101	100	101	96	102	102
QUAD (zonaal geschaald, LMS/NRM gewichten)	248	230	207	257	185	307	191	172	191	277	235	181

Tabel 3: indexcijfers totale absolute afwijking (gesommeerd over segmenten) per provincie per methode.

Om een beeld te krijgen bij de variatie in verschillen tussen waargenomen en gesynthetiseerde segmentaantallen over zowel segmenten als provincies zijn in bijlagen 2a, 2b en 3a scatterplots opgenomen voor multizone IPF, zone-by-zone IPF respectievelijk QUAD zonaal geschaald met NRM parameter instellingen. Hieruit blijkt dat er geen specifieke segmenten aan te wijzen zijn die verantwoordelijk zijn voor het slechter presteren van QUAD in vergelijking tot beide IPF methoden.

4 Quick wins mogelijk voor het LMS en NRM

Op basis van de theorie van de verschillende methoden beschreven in hoofdstuk 2 zijn een aantal quick wins geïdentificeerd die QUAD in staat stellen om de resultaten van IPF multizone te benaderen. De eerste quick win is het schalen van de resultaten naar zonale in plaats van studiegebied-brede inwoneraantallen. De tweede quick win betreft het toevoegen van segment specifieke gewichten aan de doelfunctie van QUAD. De derde quick win betreft het instellen van alle gewichten in de doelfunctie op optimale waarden. Hiervoor is geen aanvullende data of methodiek nodig, het betreft puur het aanpassen van parameterinstellingen gegeven de apriori-distributie en randtotalen.

Deze quick wins zijn getest op de dataset uit hoofdstuk 3 met behulp van de eigen Matlab implementatie. De resultaten zijn te vinden in Tabel 4, Tabel 5 en bijlage 2b. Hieruit blijkt dat de quick wins er voor zorgen dat QUAD de resultaten van IPF multizone zeer dicht benaderd. Dit effect wordt bereikt in alle provincies en segmenten.

	Totale Absolute Afwijking (TAA)	Index (IPF multizone = 100)
IPF multizone	567758	100.00
IPF zone-by-zone	568738	100.17
QUAD (zonaal geschaald, optimale gewichten)	**568010**	**100.04**
QUAD (zonaal geschaald, LMS/NRM gewichten)	1259155	221.78

Tabel 4: totale absolute afwijking per methode (inclusief QUAD met optimale gewichten)

	Drenthe	Zeeland	N-Holland	Flevoland	Friesland	Limburg	Overijssel	Groningen	Nrd-Brabant	Z-Holland	Gelderland	Utrecht
IPF multizone	100	100	100	100	100	100	100	100	100	100	100	100
IPF zone-by-zone	101	101	100	99	100	103	101	100	101	96	102	102
QUAD (zonaal geschaald, optimale gewichten)	**96**	**101**	**98**	**99**	**99**	**95**	**103**	**102**	**101**	**101**	**102**	**102**
QUAD (zonaal geschaald, LMS/NRM gewichten)	248	230	207	257	185	307	191	172	191	277	235	181

Tabel 5: indexcijfers totale absolute afwijking per provincie per methode (incl QUAD met opt. gewichten)

Op basis van uitgevoerde gevoeligheidsanalyses blijkt dat alle drie de quick wins bijdragen aan een betere fit, maar dat de derde quick win (het optimaliseren van de gewichten) zorgt het overgrote gedeelte van de verbetering. Aangezien hiervoor geheel geen aanpassing aan de LMS/NRM software voor nodig is (écht een quick win dus) is een test waarin gebruik gemaakt wordt van de LMS/NRM software en data en segmentering van het LMS of een bepaald NRM eenvoudig uit te voeren.

5. Conclusie en vervolg

Om het effect van maatschappelijke veranderingen zoals de kredietcrisis in beeld te brengen is een verkeersmodel met onderscheid naar huishoudens- en/of persoons-segmenten noodzakelijk. Voor toepassing van een dergelijk model moet de verdeling over segmenten voor elke zone bekend zijn. Omdat deze meestal niet beschikbaar is

wordt een synthetische populatie geschat op basis van data die wel beschikbaar is. Hiervoor zijn vijf veel gebruikte methoden onderzocht met als doel om de methode te identificeren die een realistische populatie oplevert, makkelijk overdraagbaar is en waarvoor reeds implementaties bestaan die zorgen voor kennis van de methode en de mogelijke efficiëntie.

Alle vijf onderzochte methoden zijn toepasbaar binnen de geschetste context. Voor IPF, QUAD en FBS zijn bestaande implementaties binnen verkeersmodelsystemen beschikbaar. Lagrange multipliers en lineair programmeren zijn in theorie krachtige methoden maar hiervan zijn, voor zover bekend bij de auteurs, geen implementaties binnen verkeersmodelsystemen beschikbaar. Bovendien kennen deze methoden een moeilijk overdraagbare implementatie, aangezien deze afhangt van de gekozen doelfunctie. Deze methoden zijn daarom niet kwantitatief getest.

Van de kwantitatief geteste methoden genereert IPF duidelijk de beste synthetische populatie, beide typen IPF (zone-by-zone en multizone) leveren beduidend realistischer populaties dan QUAD. Door problemen met de parameterinstellingen van FBS is deze methode niet kwantitatief getest, maar in theorie zou FBS met de juiste parameters tot een populatie beter dan IPF moeten kunnen leiden. IPF scoort echter beter op overdraagbaarheid, aangezien het geen parameters kent, waarmee het zonder kalibratie toepasbaar op elke dataset die consistent is en geen nulcellen bevat. Bovendien is IPF veruit de meest gebruikte methode in verkeersmodelsystemen waardoor er zeer efficiënte implementaties voorhanden zijn, terwijl de FBS methode van nature vele malen minder efficiënt is. Daarom wordt IPF als de beste oplossingsmethode voor het fitting probleem binnen de context van dit onderzoek aangemerkt.

Bij het testen van QUAD op basis van de implementatie binnen de LMS/NRM systematiek zijn een drietal quick wins geïdentificeerd, geïmplementeerd en getest. Synthetische populaties gegenereerd door de aangepaste versie van QUAD benaderen het realisme van de synthetische populaties van IPF multizone. Met de aangepaste versie van QUAD zou het LMS/NRM bij de state of the art modellen horen als het gaat om de oplosmethode gebruikt voor het fitting probleem. In dit onderzoek is tevens naar voren gekomen dat deze verbetering in prestaties bijna in zijn geheel kan worden bereikt zonder aanpassingen te doen aan de software van het LMS/NRM. Op basis van dit onderzoek wordt geconcludeerd dat de synthetische populatie binnen het LMS/NRM hierdoor veel beter de werkelijke populatie zou beschrijven. Hoe zich dit vertaald in aantallen verplaatsingen, de modal split, herkomst-bestemmings patronen, de verdeling over vertrekperioden en uiteindelijk wegvakintensiteiten, lijnbelastingen en emissies van stoffen is nog niet onderzocht, maar het effect zal ongetwijfeld substantieel zijn. Een vervolgonderzoek waarin de effecten op het LMS of een NRM op genoemde modeluitvoer in beeld worden gebracht is maatschappelijk zeer relevant, terwijl het een relatief geringe inspanning vergt.

Een andere richting voor vervolgonderzoek zou zijn om FBS met betere instellingen voor het simulated annealing schema te testen en om Lagrange multipliers en lineair programmeren te implementeren en te testen. Dit is nodig om de prestaties van deze methoden inzichtelijk te maken, aangezien zonder een implementatie, of minimaal een definitie van een specifieke doelfunctie is het moeilijk is in te schatten hoe deze

methoden presteren op een realistische dataset. Dit zijn echter geen triviale taken en de onzekerheid of er resultaten beter dan multizone IPF gevonden gaan worden is zeer groot.

Nawoord

Het achterliggend onderzoek van deze paper is in de vorm van een stage van de tweede auteur bij Goudappel Coffeng uitgevoerd. Een uitgebreid verslag van dit onderzoek met daarin o.a. wiskundige formuleringen van het optimalisatieprobleem, de invulling van dit probleem in de verschillende methoden, uitgewerkte numerieke voorbeelden en uitgebreider vergelijking van resultaten is op aanvraag beschikbaar.

Literatuur

Abraham, Hunt and Stefan (2011) *SYNTHESIZER Program HB aspect*, technische handleiding, gedateerd juni 2011

Csiszár (1975). *I-Divergence of Probability Distributions and Minimization Problems*. Annals of Probability 3 (1), pp. 146–158.

Moriarity and Scheuren (2004) *Regression based statistical matching: recent developments*. Proceedings of the Survey Research Methods Section, American Statistical Association (2004) pp. 4050-4057.

Müller & Axhausen (2010) Population synthesis for microsimulation: State of the art. *Conference proceedings of the 10th Swiss Transport Research Conference,* Monte Verita Ascona, Switzerland.

Quenneville, Fortier and Gagné (2008) A nonparametric iterative smoothing method for benchmarking and temporal distribution, *presentation on the 5th Eurostat colloqium on modern tools for business cycle analysis* , Luxembourg.

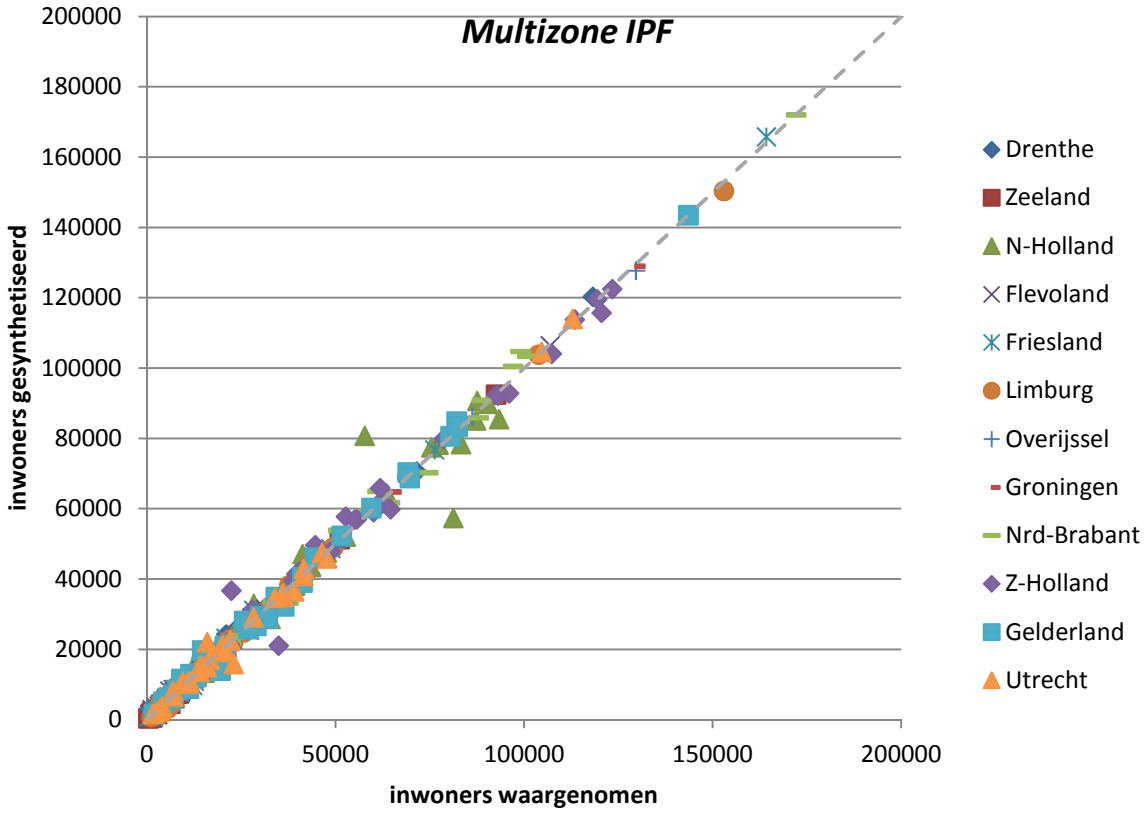
Rässler (2004) *Data fusion: identification problems, validity and multiple imputation*. Austrian journal of statistics Vol 33, pp. 153-171.

Reiter (2009) *Bayesian Finite Population Imputation for Data Fusion*. Online working paper of Duke Population Research institute, Durham, England.

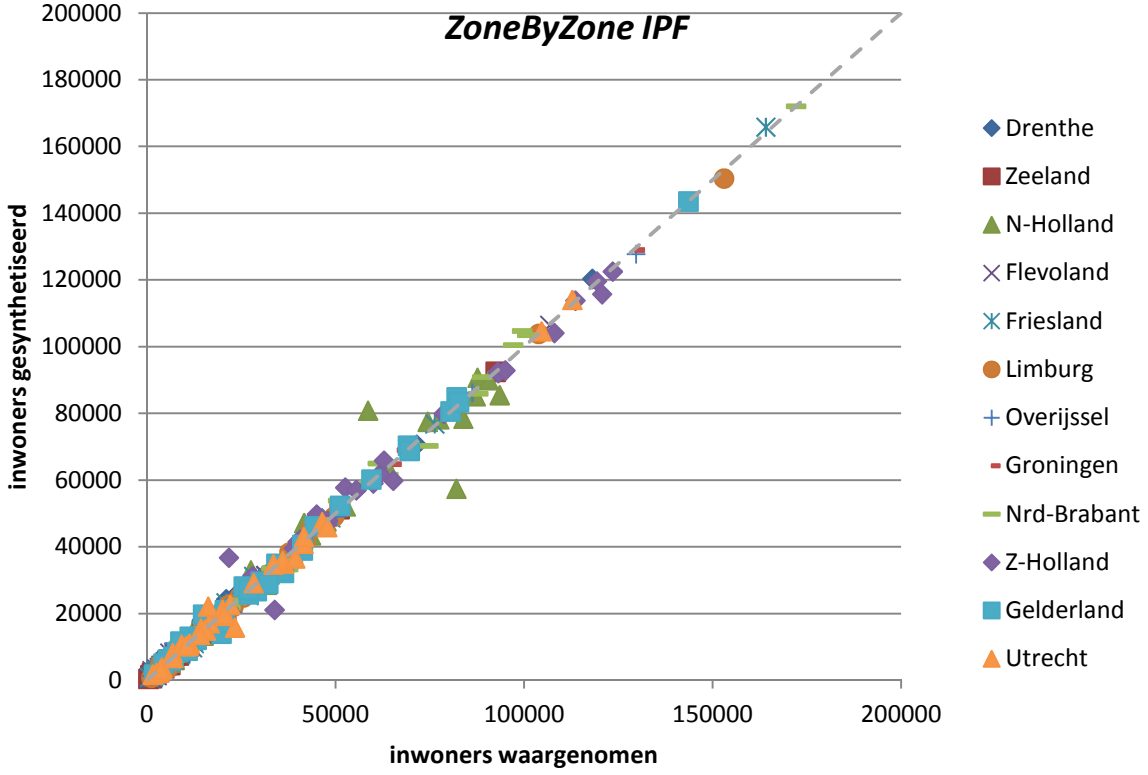
Significance (2011) *Documentatie van GM 2011-Deel D4-1: Programma QUAD*, technische handleiding, gedateerd 20 juni 2011

Tilanus (1968) *Input-Output Experiments: the Netherlands, 1948-1961,* Rotterdam: Rotterdam University Press.

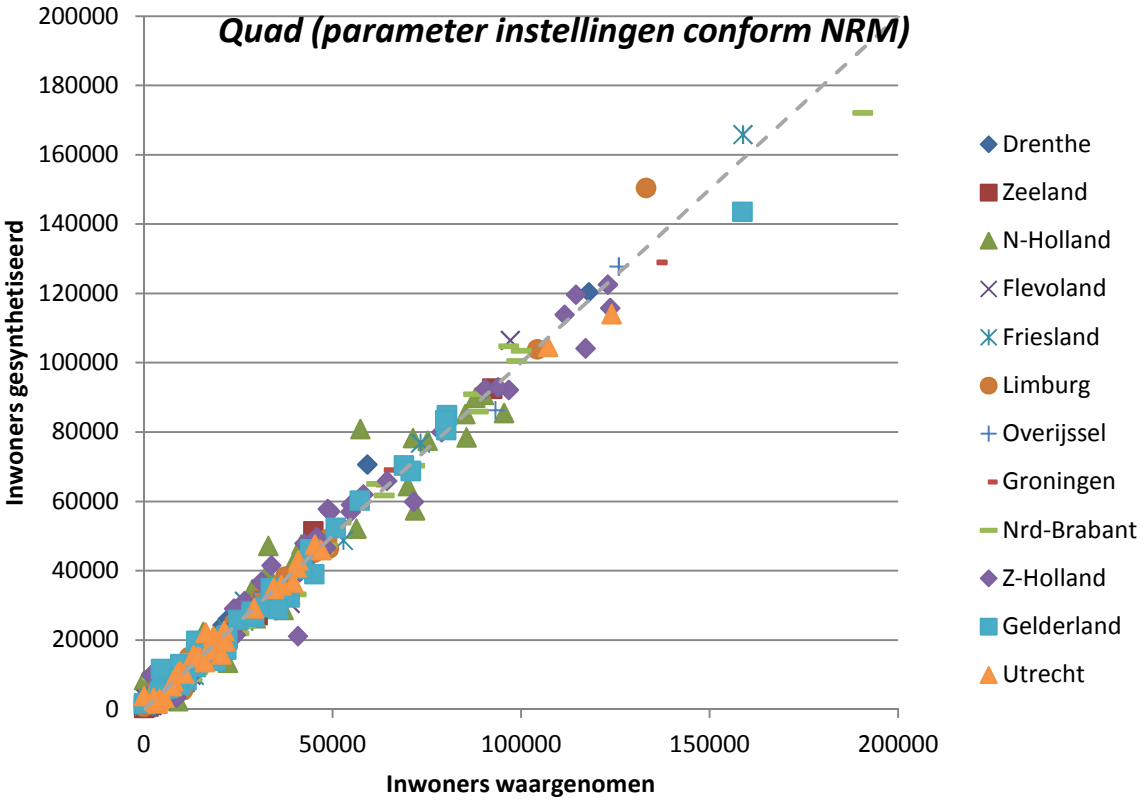
Bijlage 1a: waargenomen versus gesynthetiseerd aantal inwoners per segment per provincie (multizone IPF)



Bijlage 1b: waargenomen versus gesynthetiseerd aantal inwoners per segment per provincie (zone-by-zone IPF)



Bijlage 2a: waargenomen versus gesynthetiseerd aantal inwoners per segment per provincie (QUAD met parameter instellingen zoals in LMS/NRM en schaling per zone)



Bijlage 2b: waargenomen versus gesynthetiseerd aantal inwoners per segment per provincie (QUAD met optimale parameterinstellingen en schaling per zone)